A definitive pharmacophore modelling study on CDK2 ATP pocket binders: tracing the path of new virtual high-throughput screenings

Marco Tutone[1*], Giulia Culletta[1], Luca Livecchi[2], Anna Maria Almerico[1]

[1]Dipartimento di Scienze e Tecnologie Biologiche Chimiche e Farmaceutiche (STEBICEF), Università degli Studi di Palermo, Via Archirafi 32, 90123 Palermo (Italy); [2]Department of Clinical and Pharmaceutical Sciences, University of Hertfordshire, Hatfield, United Kingdom

KEYWORDS: CDK2, ATP pocket, Molecular Dynamics, MM-GBSA, Pharmacophore Modelling, Common Hits Approach

ABSTRACT: Cyclin Dependent Kinases-2 (CDK2) are members of serine/threonine protein kinases family. They play an important role in the regulation events of the eukaryotic cell division cycle, especially during the G1 to S phase transition. Experimental evidences indicate that excessive expression of CDK2s should cause abnormal cell cycle regulation. Therefore, since long time, CDK2s have been considered potential therapeutic targets for cancer therapy. In this work, one-hundred and forty-nine complexes of inhibitors bound in the CDK2-ATP pocket were submitted to short MD simulations (10ns) and free energy calculation. Comparison with experimental data (K_i, K_d and pIC_{50}) revealed that short simulations are exhaustive to examine the crucial ligand−protein interactions within the complexes. Information collected on MD simulations of protein−ligand complexes have been used to perform a molecular modelling approach that incorporates flexibility into structure-based pharmacophore modelling (Common Hits Approach, CHA). The high number of pharmacophore models resulting from the MD simulation was thus reduced to a few representative groups of pharmacophore models. The performance of the models have been assessed by using the ROC curves analysis. This definitive set of validated pharmacophore models could be used to screen *in-house* and/or commercial datasets for detection of new CDK-2 inhibitors. We provide the models to all the researchers involved in this field.

Introduction

The proliferation in mammalian cells is controlled by the cell cycle and protein phosphorylation is a crucial post-translational modification. The cyclin-dependent kinases (CDKs), proteins regulating cell division, are activated by serine/threonine kinases and control critical checkpoints in the G1/S and G2/M phase transitions. Cancer growth is associated with the loss of these checkpoints. This suggests that CDKs are a pivotal target for the development of pharmacologically interesting agents. For this reason, the interest in the development of CDK inhibitors is recently growing. CDKs are relatively small proteins, with molecular weights ranging from 34 to 40 kDa, and contain little more than the kinase domain. They express their activity upon binding a regulatory protein called cyclin; without cyclin, CDK has little kinase activity. They are also involved in other several physiological events such as transcription regulation, mRNA processing, and the nerve differentiation. Given the involvement of CDKs in multiple cellular processes, development of selective small molecule inhibitors for specific CDKs is expected to enhance their therapeutic potential in cancer treatment.

CDK2 together with CDK1, CDK4, and CDK6[1–3], remains the most attractive target for oncology The design of novel chemical scaffold of potent CDK inhibitors was allowed by a large amount of structure-based and computational work. The interest in computer-aided methods' application has significantly increased, as confirmed by considerable increment in the number of available structures of cyclin-dependent kinases. A recent check (November 2018) on Protein Data Bank web page (www.PDB.org) retrieved 527 structures for CDKs, of which 441 published between 2005-2018[4]. Over 300 out of 527 are complexes related to CDK2 bound with an inhibitor in the ATP binding pocket. The ATP-binding pocket is common to all kinases and it is often chosen as a reference target for the research of inhibitors. The CDKs' family presents a conserved structure of the ATP-binding pocket, nevertheless, there are slight differences among them. This allows designing molecules that show a significant specificity for a given subclass of kinases. In CDK2 the ATP binding site is characterized by two amino acids, the Leu83 and the Glu81, both crucial in the binding of the ATP and, therefore, of all its competitive inhibitors. The ribose and phosphate groups form multiple polar interactions, one of which involves coordination to the catalytic magnesium via the phosphate groups along with Asp145 and Asn132 [5-7]. In past, only classical docking procedures and/or pharmacophore modelling were employed. More recent studies often report flexible docking outcomes combined also with MD simulations. Instead, pharmacophore modeling studies have been performed to generate pharmacophore maps based on a set of crystal structures of protein–ligand complexes, without the use of MD simulations. The vast amount of crystallographic data available is certainly an important starting point to use in Virtual High-Throughput

Screenings (VHTSs) [8-9]. It is widely known that proteins and small molecules are dynamic entities which are able to perform a wide sort of movements. For this reason, using a single pose of a dynamic system provides scarce information about the conformational flexibility of the ligand and about the motion of the residues near the binding pocket[10]. Therefore a pharmacophore model generated from a single structure might include artificial features, caused either by crystal packing effects or simply by picking a single set of coordinates of the structure. Incorporating dynamic features in pharmacophore modelling represents a new frontier, and in a recent past some attempts were tried[11-13]. Among them is the "Common Hits Approach (CHA)" , which performs a consensus pharmacophore-based virtual screening on the conformational ensemble of the protein-ligand complexes obtained by means of MD simulations[14]. However, the simulation time of these approaches has not been standardized yet. Generally, the simulations are of medium length (over 20 ns) and repeated several times, in order to fully explore ligand-protein interaction, comporting a great expenditure of calculation time. Therefore, a possible solution to overcome and minimize the "time" issue could be carrying out short molecular dynamics simulations (10ns), calculate the ΔG values obtained from the trajectories (ΔGcalc) using MM-GBSA, and compare them with the experimental activity data (ΔGexp). In fact, if the experimental data can be simulated through short trajectories, this entails that the ligand-protein interaction could be explored exhaustively. Thus, short dynamics would be useful for the purposes of VHTS allowing to save calculation time, but guaranteeing equal effectiveness. , Despite the time factor plays a crucial role in the simulation, a longer simulation is not always necessary to obtain higher prediction accuracy[15]. For this reason, we performed an exhaustive MD simulations study on CDK2/inhibitor complexes in order to obtain definitive pharmacophore models to use in VHTS. Although a large number of allosteric CDK2 inhibitors were under investigation[16], we decided to focus attention on ATP competitive inhibitors.

Results and Discussion

Among the over 300 CDK2/ATP competitive inhibitors complexes, only the ones presenting an experimental activity data, such as K_i, K_d and/or pIC_{50}, were selected (Supporting Information). Experimental data are fundamental to compare ΔGcalc. Thus, we collected 149 CDK2/ATP competitive inhibitors complexes, plus the CDK2/ATP complex (PDB ID:1B39). Starting from X-ray coordinates we performed short (10ns) MD simulations, and each frame was collected to calculate ΔGcalc. Finally, the average value of all the frames was calculated. A table with all the ΔGcalc values is reported in Supporting Information. In order to correlate experimental data and ΔGcalc, the experimental dataset has been separated in three different datasets: a first dataset of 42 complexes with known K_i; a second dataset of 23 complexes with known K_d; a third dataset of 121 complexes with known pIC_{50}. For some inhibitors, more than one experimental data has been considered. K_i and K_d values have been converted in ΔGexp according to the equation 1.

$$\Delta Gexp = - RT \cdot \ln(K) \quad \text{(eq. 1)}$$

where R is the constant of the gases equal to 1.987 cals K^{-1} mol^{-1}, T is the temperature in Kelvin and K it is the constant of the analyzed equilibrium, inhibition (K_i) or dissociation (K_d).

In order to detect outlier data between the two independent variables (ΔGcalc and/or ΔGexp, pIC_{50}), the ratio distribution[17] of the two variables has been calculated. Often the ratio distributions are heavy-tailed, and it could be challenging to work with such distributions to develop an associated statistical test. A method based on the median has been suggested as a "work-around"[18]. According to this method, a value outside of the interval: [Q1-k(Q3-Q1); Q3-k(Q3-Q1)],where Q1 and Q3 are the first quartile and the third quartile, respectively, k is a constant that regulates the width of the interval, is defined as outlier. Normally, the width of the interval assumes the value of 1.5[19]. The detection of outliers reduced the samples to 114 complexes with known pIC_{50}, 41 complexes with known K_i and 22 complexes with known K_d. Calculated binding energies were plotted against pIC_{50}, and/or ΔGexp values for the series. The degree of correlation between the two parameters was evaluated using the Pearson's correlation coefficient, Rp and the Spearman's rank correlation coefficient, Rs, as reported. Rs compares the position of each inhibitor compound when ranked by binding energy to its position when ranked by its pIC50 or ΔGexp values value[20]. The Spearman's rank correlation coefficient is defined as:

$$Rs = 1 - \frac{\mathbf{6}\sum d_i^{\mathbf{2}}}{n(n^{\mathbf{2}} - \mathbf{1})},$$

where d_i is difference in rank for the i^{th} compound under the two different criteria, i.e., binding energy and experimental binding constant, and n is the number of compounds in the series. Significance of Rp and Rs was evaluated by means of t-Test and z-Test.

Plots of binding energy versus pIC_{50} or ΔGexp values for each series are shown in Figure 1. Pearson's correlation coefficient for the plots, as well as Spearman's rank correlation coefficients, are reported for each series in Table 1.

Table 1. Summary of results

	n	**Rp**	**Rs**	**t**	**z**	**DOF**	**A**
pIC50	114	0.520	0.520	6.5	5.52	112	<0.001
Ki	41	0.568	0.490	4.36	3.09	39	<0.001
Kd	22	0.794	0.839	5.83	3.84	20	<0.001

Legenda. t; student t test; z = normal standard distribution test; DOF, degrees of freedom.

In all cases, the outcomes showed realistic correlation and significance (α<0.001) both in terms of Rp and Rs as demonstration that short MD simulations could lead to a reliable interpretation of protein/drug interactions.

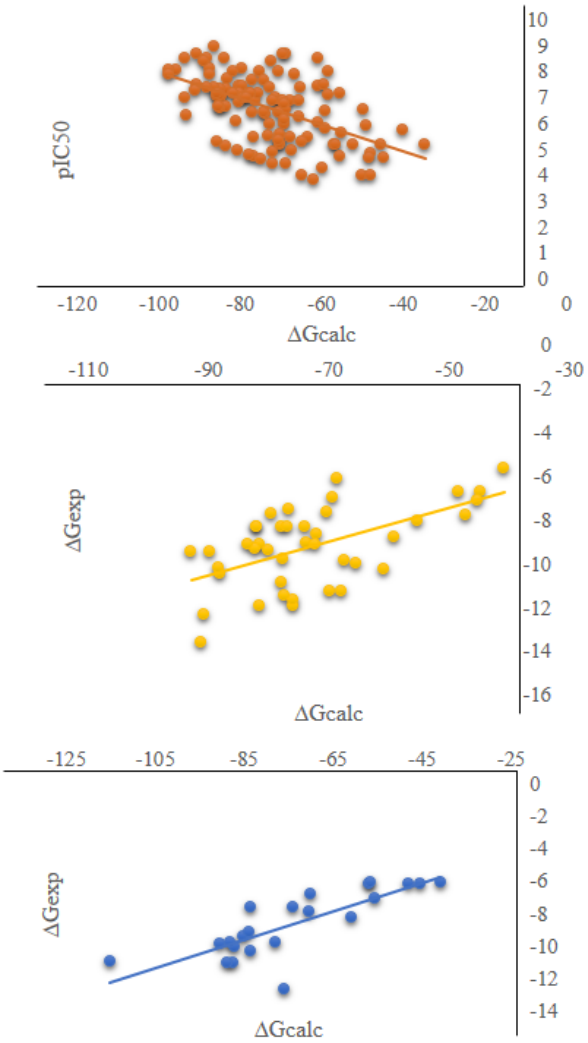

Figure 1. Plots of binding energies (ΔGcalc) versus pIC_{50} (red circles) or ΔGexp values (Ki, cyan circles; Kd, green circles)

Therefore, the snapshots, collected during the MD simulations, and the coordinates of the PDB file were processed according to the CHA. The pharmacophore models are generated starting from each single snapshot, and subsequently a feature vector (represented as a bit string) was generated for each pharmacophore model. This procedure results in 1001 feature vectors per protein−ligand complex (1000 pharmacophore models obtained from the MD simulation, plus the pharmacophore model obtained from the PDB crystallographic file). The feature vectors were aggregated into distinct vectors, counting how many times that particular combination of pharmacophore features was identified during MD simulations, in a process called "appearance count". This reduces the number of relevant vectors, in fact, instead of using 1000 individual feature vectors, a smaller number of distinct feature vectors, observed one or more times during MD simulations, was obtained. Distinct feature vectors observed only once were discarded (and considered random artifacts). Only distinct feature vectors with an appearance count≥2 were considered, named Representative Pharmacophore Models (RPMs). The performance of about 30100 RPMs obtained from 149 ligand-protein complex MD was assessed by means of the ROC (Receiver Operating Characteristic) curves analysis with a validation dataset obtained from the DUD-E site[21] containing molecules (676 Active and 28121 Decoys) generated specifically for CDK-2. For each RPM, a hit-list was collected, so that for each ligand-protein complex several hit-lists were identified. The multiple RPM hit-lists have been combined into a single list named RPM-HIT-LIST, consisting only of unique compounds. The molecules in this list have been ranked according to the number of times they are recognized in the hit-lists. For example, if a molecule is present in many hit-lists it is classified with a higher score than one that appears only in a few hit-lists. (Figure 2)

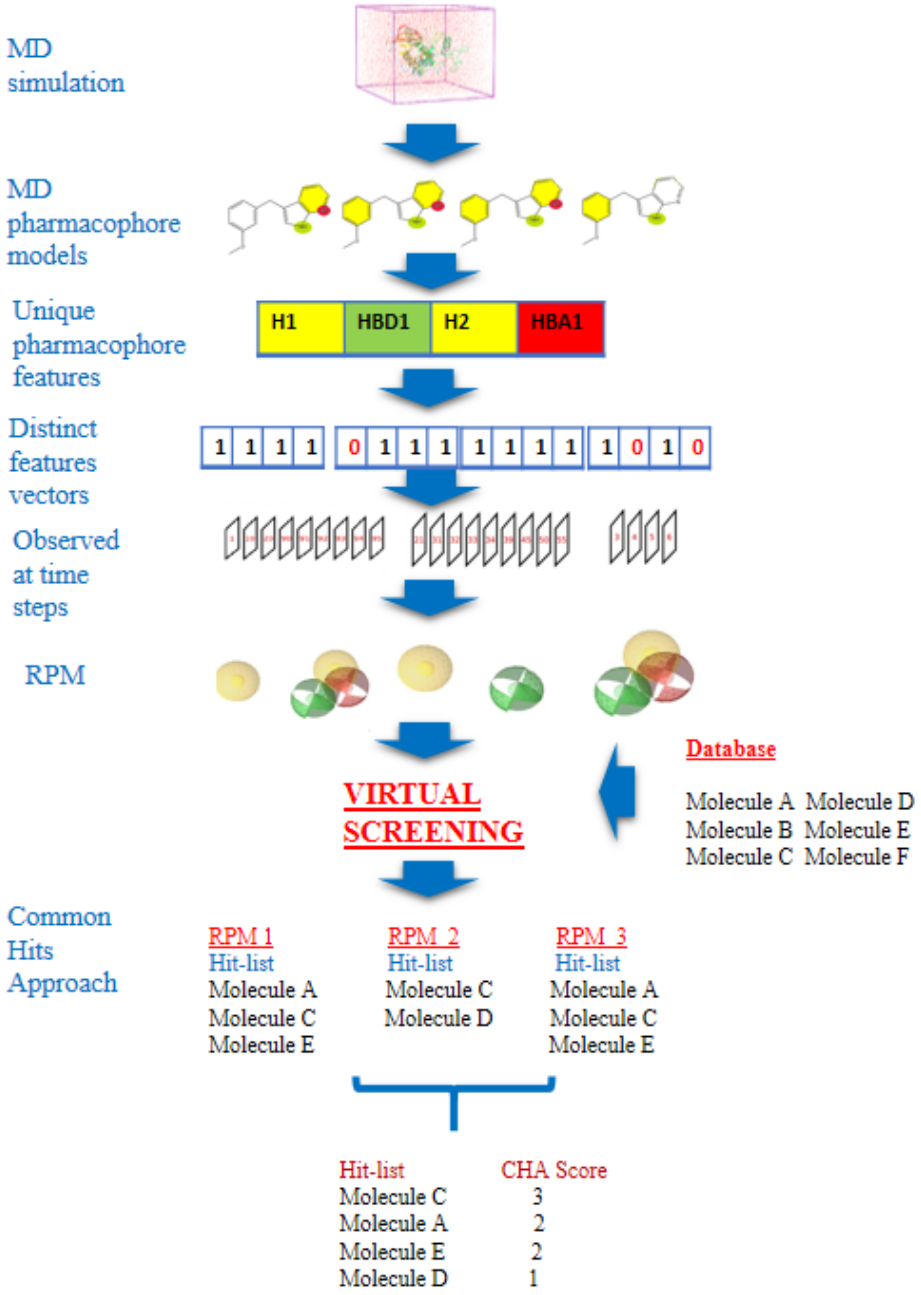

Figure 2. Flowchart of the method from MD to CHA

The ROC curves have been calculated and analyzed plotting the number of True Positives (TPR) on False Positives (FPR). The performance of the RPM models was evaluated using the ROC curve at 2% and 100%, considering for each complexes the first 50 hits. Due to space limits, the table with the AUC values is reported in the supporting information. To improve the accuracy of the approach we carried out a consensus considering actives and molecules in common between the first ten, twenty and thirty RPM-HIT-LISTs, named RPM-CONSENSUS-10 RPM-CONSENSUS-20, and RPM-CONSENSUS-30 respectively (Figure 3 and Table 2).

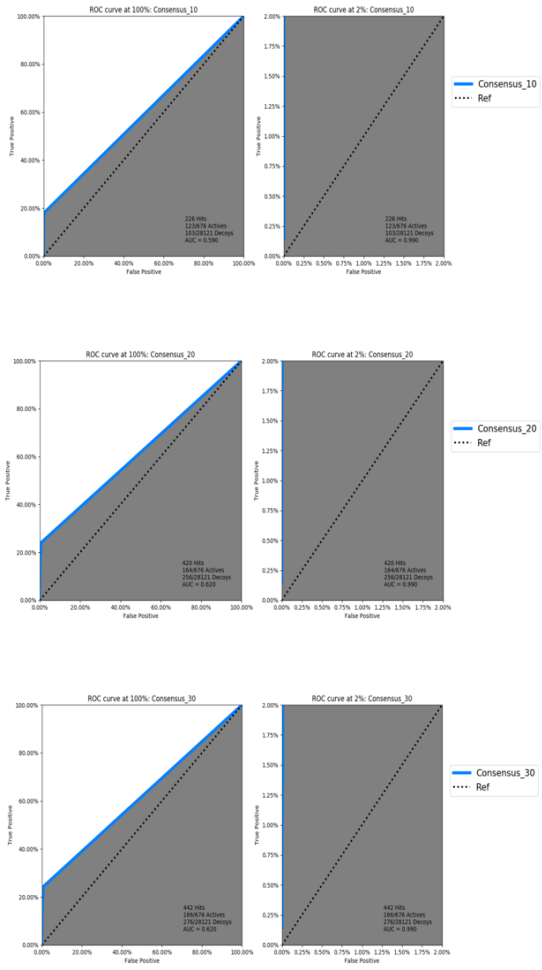

Figure 3. ROC curves at 2% and 100% for RPM-Consensus-10 (up), RPM-Consensus-20 (middle), RPM-Consensus-30 (down)

Table 2.

RPM Consensus	*Hits*	*Actives*	*Actives/hits rate*	*AUC 100%*	*AUC 2%*
10	226	123	54%	0.590	0.990
20	420	164	39%	0.620	0.990
30	442	166	37%	0.620	0.990

Analyzing the results, it is worth noting the identical value of AUC 2% for all the consensus, whilea slight difference can be observed when considering the AUC 100% values.. We decided to perform a further analysis on the RPM-CONSENSUS-10 list, due to its higher rate of actives/hits, compared to the other two consensus lists. The RPM-CONSENSUS-10 list consists of 226 hits, of which 123 actives, and 552 RPMs. The 552 RPMs have been clustered according to feature similarity, with a maximum distance between them of 0.5 Å. 59 RPMs were obtained and re-submitted to the CHA. The CHA-HIT-LIST cluster obtained was compared with the actives resulting in a good matching for 18 RPMs out of 59. Finally, the original DUD-E dataset was screened against the 18 RPMs selected, and the AUC 2% and 100% were calculated. At the end, 11 RPMs showed AUC 2% > 0.90 (Table 3) proving to be the most consistent pharmacophore models among the ones considered, and valid for performing VHTS in the search of new CDK2 ATP pocket binders inhibitors. These pharmacophore models consist of a number of features between 4 and 7. In particular, the identified features are hydrophobic moiety, H-bond acceptors and donors, and aromatic rings. In Figure 4, a representation of the superimposed features of all the selected 11 models can be observed; a detailed report of each pharmacophore feature for single models is attached in Supporting information. Moreover the *.pml* files will be available for all the researchers who are involved in this field.

Table 3.

RPM-cluster	Actives	Decoys	AUC 100%	AUC 2%
10	**66**	**160**	**0.55**	**0.98**
13	**42**	**26**	**0.53**	**0.97**
18	**31**	**13**	**0.52**	**0.96**
5	**35**	**191**	**0.52**	**0.95**
1	**44**	**182**	**0.53**	**0.94**
15	**37**	**189**	**0.52**	**0.94**
9	**35**	**191**	**0.52**	**0.94**
14	**31**	**195**	**0.52**	**0.94**
6	**34**	**192**	**0.52**	**0.92**
3	**22**	**43**	**0.52**	**0.92**
8	**19**	**2**	**0.51**	**0.92**
2	21	205	0.51	0.87
11	13	19	0.51	0.87
16	10	8	0.51	0.83
12	14	212	0.51	0.78
17	7	8	0.51	0.78
4	4	11	0.50	0.69
7	9	217	0.50	0.64

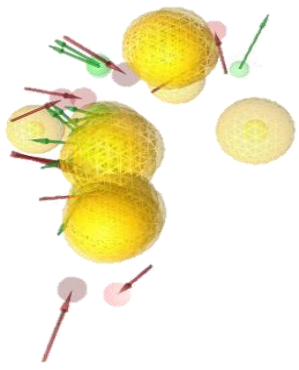

Figure 4. Super-positioning of the 11 selected pharmacophore models

Conclusion

In summary, CDK2s, serine/threonine kinases involved in the cell cycle regulation and tumorigenesis, are to-date a very challenging target for the researchers involved in drug discovery. In literature, numerous attempts of molecular modelling studies, aimed to CDK2 inhibitors discovery and development, were reported. Most of these studies are limited to docking and MD of few molecules. At best of our knowledge, attempts of extensive molecular modelling studies regarding MD and pharmacophore modelling have not been performed. For this reason, we exploited the big amount of data available in the PDB to build definitive pharmacophore models with the aim to improve virtual screenings of new chemical entities. Short MD simulations and free energy calculation were applied on 149 CDK2 structures complexed with an ATP pocket binder; comparing ΔGcalc values obtained with experimental activity data revealed that these short simulations are exhaustive to explore all the host-guest interactions. The MD trajectories snapshots were then processed by means of the CHA. The representative pharmacophore models obtained (RPMs) were re-processed to validate them and to identify the most reliable. At the end of the study, we proposed 11 pharmacophore models showing AUC 2% > 0.92. They consist of 4-6 features (H-bond donors and acceptors, hydrophobic moiety and aromatic ring), and could be considered as the ultimate models to perform VHTS. The 3D features coordinates of each model (Supporting Information) will be available for every researcher involved in this field, interested in testing new chemical entities.

ASSOCIATED CONTENT

Supporting Information

Methods (SI S2)
Preparation of the Proteins (SI S2)
Molecular Dynamics Simulations (SI S2)
MM-GBSA free energy calculations (SI S2)
Conversion of MD trajectories (SI S2)
Pharmacophore Models Generation by means of CHA and Virtual Screenings (SI S3)
Pharmacophore Models (SI S5)
PDB IDs of CDK2/ATP competitive inhibitors and related experimental data (SI S7)
Table 1. Average ΔGcalc values and standard deviation for all the CDK2 complexes. (SI S10)
Table 2. ΔGcalc and pIC50 values (SI S12)
Table 3. ΔGcalc and ΔGexp derived from Ki values (SI S14)
Table 4. ΔGcalc and ΔGexp derived from Kd values (SI S15)
Table 5. ROC values of CHA hit-list RPMs (SI 16)
References (SI S19)

AUTHOR INFORMATION

Corresponding Author

*Marco Tutone email: marco.tutone@unipa.it

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

ACKNOWLEDGMENT

CG would like to thank Dr. Arthur Garon, and Prof T. Langer University of Vienna, for the use of facilities during her leave of absence from the University of Palermo.

REFERENCES

(1) Chohan T A, Qian H, Pan Y, Chen J.-Z. Cyclin-Dependent Kinase-2 as a Target for Cancer Therapy: Progress in the Development of CDK2 Inhibitors as Anti-Cancer Agents. Curr Med Chem 2015; 22 (2): 237–263.

(2) Sánchez-Martínez C, Gelbert L M, Lallena M J, De Dios A. Cyclin Dependent Kinase (CDK) Inhibitors as Anticancer Drugs. Bioorg Med Chem Lett 2015; 25 (17): 3420–3435.

(3) Bose P, Simmons G L, Grant S. Cyclin-Dependent Kinase Inhibitor Therapy for Hematologic Malignancies. Expert Opin Investig Drugs 2013; 22 (6): 723–738.

(4) Tutone M, Almerico A M. Recent Advances on CDK Inhibitors: An Insight by Means of in Silico Methods. Eur J Med Chem 2017; 142: 300-315.

(5) Tripathi S K, Muttineni R, Singh S K. Extra Precision Docking, Free Energy Calculation and Molecular Dynamics Simulation Studies of CDK-2 Inhibitors. J Theor Biol 2013; 334: 87-100.

(6) Noble M E M, Endicott J A. Chemical Inhibitors of Cyclin-Dependent Kinases Insights into Design from X-Ray Crystallographic Studies. In Pharmacology and Therapeutics 1999; 82: 269-278.

(7) Wyatt P G, Woodhead A J, Berdini V, *et al.* Identification of *N*-(4-Piperidinyl)-4-(2,6-Dichlorobenzoylamino)-1*H*-Pyrazole-3-Carboxamide (AT7519), a Novel Cyclin Dependent Kinase Inhibitor Using Fragment-Based X-Ray Crystallography and Structure Based Drug Design. J Med Chem 2008; 51 (16): 4986–4999.

(8) Abdulghani H, Sliman F. Virtual Screening and Molecular Docking Studies for the Discovery of Novel CDK2 Inhibitors. Int J Pharm 2018; 50 (05): 25–33.

(9) Zou J, Xie H, Yang S, Chen J, Ren J, Wei Y. Towards more accurate pharmacophore modeling : Multicomplex-based comprehensive pharmacophore map and most-frequent-feature pharmacophore model of CDK2. J Mol Graph & Model 2008; 27 (4): 430–438.

(10) Mirjalili V, Feig M. Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles. J Chem Theory Comput 2013; 9 (2): 1294-1303.

(11) Wieder M, Perricone U, Boresch S, Seidel T, Langer T. Evaluating the Stability of Pharmacophore Features Using Molecular Dynamics Simulations. Biochem Biophys Res Commun 2016; 470 (3): 685–689.

(12) Perricone U, Wieder M, Seidel T, *et al*. A Molecular Dynamics-Shared Pharmacophore Approach to Boost Early-Enrichment Virtual Screening: A Case Study on Peroxisome Proliferator-Activated Receptor α. Chem Med Chem 2017; 12 (16): 1399-1407.

(13) Tutone M, Pantano L, Lauria A, Almerico A M. Molecular Dynamics, Dynamic Site Mapping, and Highthroughput Virtual Screening on Leptin and the Ob Receptor as Anti-Obesity Target. J Mol Model 2014; 20 (5): 2247.

(14) Wieder M, Garon A, Perricone U, *et al*. Common Hits Approach: Combining Pharmacophore Modeling and Molecular Dynamics Simulations. J Chem Inf Model 2017; 57 (2): 365-385.

(15) Wang W, Donini O, Reyes C M, Kollman P A. Biomolecular Simulations: Recent Developments in Force Fields, Simulations of Enzyme Catalysis, Protein-Ligand, Protein-Protein, and Protein-Nucleic Acid Noncovalent Interactions. Annu Rev Biophys Biomol Struct 2001; 30: 211-243.

(16) Richardson C M, Nunns C L, Williamson D S, *et al*. Discovery of a Potent CDK2 Inhibitor with a Novel Binding Mode, Using Virtual Screening and Initial, Structure-Guided Lead Scoping. Bioorg Med Chem Lett 2007; 17 (14): 3880-3885.

(17) Geary R C. The Frequency Distribution of the Quotient of Two Normal Variates. J R Stat Soc 1930; 93: 442-446.

(18) Brody J P, Williams B A, Wold B J, Quake S R. Significance and Statistical Errors in the Analysis of DNA Microarray Data. Proc Natl Acad Sci 2002; 99 (20): 12975–12978.

(19) Devore J. Statistics for Business and Economics. Am Stat 2006;

60 (4): 342-343.
(20) Wei H Y, Tsai K C, Lin T H. Modeling Ligand-Receptor Interaction for Some MHC Class II HLA-DR4 Peptide Mimetic Inhibitors Using Several Molecular Docking and 3D QSAR Techniques. J Chem Inf Model 2005; 45 (5): 1343-1351.
(21) Mysinger M M, Carchia M, Irwin J J, Shoichet B K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. J Med Chem 2012; 55 (14): 6582-6594.

Graphical Abstract