Molecular modelling and bioinformatics studies of CDK4 and related proteins

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by

Muhammad Imtiaz Shafiq

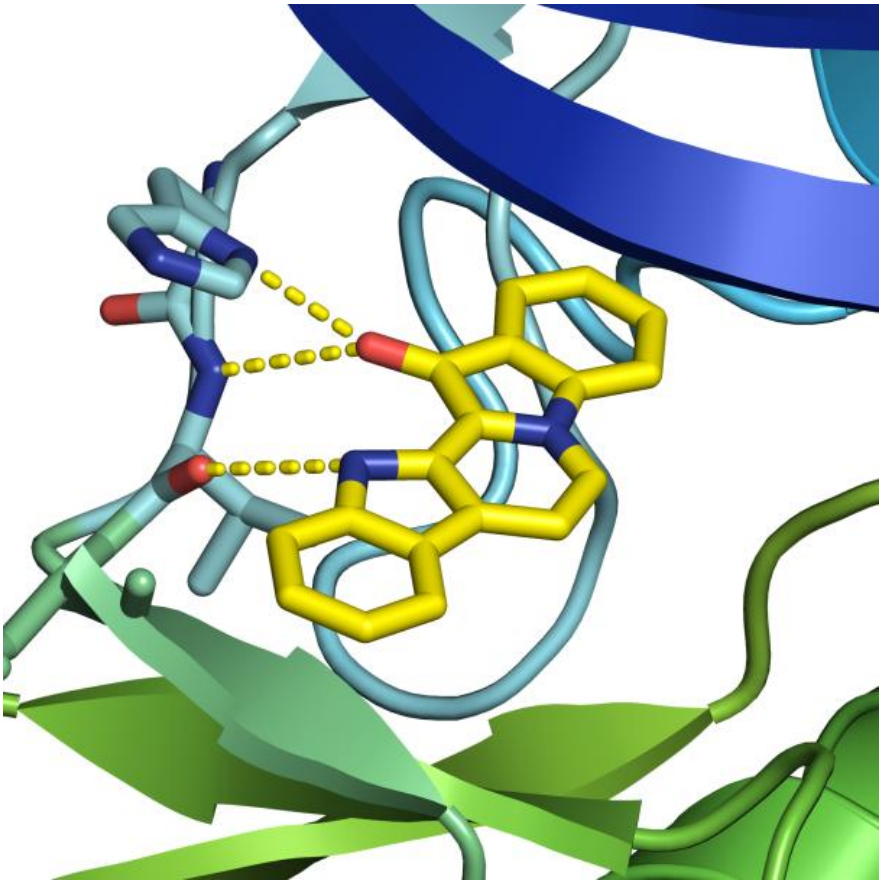
Department of Biochemistry
University of Leicester

December 2010

Dedication

For my family, who offered me unconditional love and support throughout my PhD

Molecular modelling and bioinformatics studies of CDK4 and related proteins



Muhammad Imtiaz Shafiq

Department of Biochemistry
University of Leicester

December 2010

Acknowledgments

First and foremost, my thanks and gratitude goes to my supervisor Dr. Ralf Schmid for his supervision, generosity and moral support. Without his strong encouragement and motivation, I may not have been able to finish this work. My sincere gratitude also goes to my committee members, Professor Dr. Andrew Fry, Professor Dr. Mark Carr, and Dr. Paul Jenkins for their valuable suggestions and feedback on my research work. I would also like to extend my thanks to Dr. Thomas Steinbrecher for providing computational facilities and his help to run thermodynamics integration experiments at Rutgers, The State University of New Jersey, USA.

I owe special thanks to Dr. Wei-Cheng Huang for his valuable suggestions from time to time. I also thanks to my group fellow Abbas Alameer for his motivation and support for me during the long hours of work. I am indebted to all the friends and colleagues who have prayed for me and for the successful completion of my PhD

I would like to thanks to all my family members and in particular to my wife and kids for their love and for accompanying me here in UK far away from home. I have no words to say thanks to my parents for their support, encouragement and prayers which enabled me to finish my work.

Muhammad Imtiaz Shafiq

Title: Molecular modelling and bioinformatics studies of CDK4 and related proteins
Author: Muhammad Imtiaz Shafiq

Abstract

Cyclin-dependent kinases play a key role in the regulation of the eukaryotic cell cycle. CDK4 regulates the G1/S phase transition and the entry into the S-phase of the cell cycle. The activity of CDK4 is misregulated in many human cancers. The natural product fascaplysin inhibits CDK4 specifically, and is considered as a lead compound for specific CDK4 inhibitors. In the present work the structural features of the active sites of CDKs are compared, the evolution of CDKs is studied and homology models of CDK4 are generated and used to gain insights into its sequential and structural features. Also the CDK4-ligand interactions of fascaplysin and its tryptamine based derivatives are predicted and the fascaplysin specificity for CDK4 is at least partially explained using thermodynamic integration.

CDK4 homology models were generated based on CDK2 templates. However, after the availability of experimentally determined X-ray structures of CDK4 in an inactive form, CDK4 models were built in a putative active form by incorporating the structural information from both CDK4 and CDK2 for its later use in molecular modelling.

Docking studies on fascaplysin with CDK4 predict a polar contact between $His95^{CDK4}$ and fascaplysin in addition to bidentate hydrogen bonds with Val96. This interaction partly explains the selectivity for CDK4 compared to CDK2. The effect of the positive charge of fascaplysin on specificity is studied in thermodynamic integration MD simulations by the isoelectronic substitution of the positively charged nitrogen into a carbon atom. From these thermodynamics integration calculations it is concluded that fascaplysin shows a preference for CDK4 due to better stabilization of the positive charge.

ChemScores for tryptamine based derivatives docked into CDK4 show a weak correlation with experimental IC_{50} values. This indicates that the ChemScores can be used as a weak predictor for relative affinities of CDK4 inhibitors. A new class of α-carboline based inhibitors is proposed, and based on docking studies, predicted to have improved binding affinities for CDK4.

Abbreviations

3D	Three-dimensional
Å	Angstrom = 1.0×10^{-10} metres
aa	Amino acid
ADP	Adenosine diphosphate
AGC	Protein kinases A, G and C
AMBER	Assisted model building and energy refinement
aPKs	Atypical kinases
ATP	Adenosine triphosphate
BLAST	Basic local alignment search tool
blastn	Nucleotide BLAST
blastp	Protein BLAST
BRAF	v-raf murine sarcoma viral oncogene homolog B1
CAK	CDK-activating kinase
CAMK	Ca2+/calmodulin-dependent protein kinases
CDC2	Cell-division cycle 2 kinase
CDK	Cyclin dependent kinase
CDK4	Cyclin dependent kinase 4
CDKL	Cyclin dependent kinase like protein
CHARMM	Chemistry at Harvard macromolecular mechanics
CK1	Casein kinase 1
CKI	CDK inhibitor proteins
Cl^-	Chlorine ion
CMGC	Cyclin-dependent kinases, mitogen-activated protein kinase, glycogen synthase kinase 3, and casein kinase 2),
CML	Chronic myeloid leukaemia
CTD	C-Terminal Domain
DFG	Aspartic acid, phenylalanine and glycine
DOPE	Discrete optimized protein energy
E2F	A genetic transcription factor
EMBL	European molecular biology laboratory

ePKs	Eukaryotic protein kinases
EST	Expressed sequence tag
E_{TOT}	Total energy
FDA	Food and drug administration
FGFR	Fibroblast growth factor receptor
GOLD	Genetic optimization for ligand docking
HMM	Hidden Markov model
IC_{50}	The half maximal inhibitory concentration
JAK	Janus kinase or just another kinase
LPC	Ligand polar contact
MD	Molecular dynamics
Me	Methyl (CH3-)
MEGA	Molecular evolutionary genetics analysis.
MEK	Mitogen-activated protein kinase
MSA	Multiple sequence alignment
MUSCLE	Multiple sequence comparison by log-expectation
NMR	Nuclear magnetic resonance
NTRK2	Neurotrophic tyrosine kinase receptor type 2
OPLS	Optimized potentials for liquid simulations
p16	The p16 protein
PDB	Protein data bank
PDB xxxx	Protein databank code, here xxxx is a four letter alpha-numeric code
Pdfs	Probability density functions
PERL	Practical extraction and report language
PHDK	Pyruvate dehydrogenase kinases
PIKK	Phosphatidylinositol 3′ kinase-related kinases
PKA	Protein Kinase A
PKC	Protein Kinase C
pmemd	Particle Mesh Ewald Molecular Dynamics
pRb	Retinoblastoma protein
PSK-J3	Cyclin-dependent kinase 4
Rb	Retinoblastoma gene

RGC	Receptor guanylate cyclases
RIO	Right open reading frame
RMSD	Root mean square deviation
RMSD[a]	All atoms RMSD
RMSD[b]	Backbone RMSD
Sander	Simulated annealing with NMR derived energy restraints
STE	Homologs of yeast sterile 7, sterile 11, sterile 20 kinases
STI571	Imatinib mesylate
TI	Thermodynamic integration
TK	Tyrosine kinases
TKL	Tyrosine kinase-like kinases
TrEMBL	Translated EMBL
VEGFR	Vascular endothelial growth factor receptor

Table of contents

Chapter one

INTRODUCTION

"The universe is full of magical things, patiently waiting for our wits to grow sharper."

Eden Phillpotts

Chapter 1 Introduction

1.1 Life and cell division

About 3.8 billion years ago life arose on our planet in the form of simplest cells from the pre-biotic building blocks (Mojzsis *et al.*, 1996; Oró and Lazcano, 1997; Orgel, 1998; Krakauer and Sasaki, 2002; Davies and Lineweaver, 2005; Isabelle *et al.*, 2007). The replication of cells is one of the fundamental characteristic of living systems (Koshland, 2002). The living organisms on earth are categorized into three phylogenetic domains, which are Eukarya, Archaea, and Bacteria (Woese *et al.*, 1990; Pace, 2009). Bacteria and Archaea represent the simplest form of life, which is unicellular in nature, in contrast to eukaryotes which represent complex and multicellular organisms. Life took many years for its evolution from the unicellular to the multicellular form (Cavalier-Smith, 2006). The continued survival of any living system is dependent on its ability to reproduce. The process by which single parent cells can divide or reproduce into two identical daughter cells is known as cell division.

1.2 The eukaryotic cell division cycle

The cell cycle or cell division cycle comprises a series of different stages, which a cell undergoes during its division. The eukaryotic cell cycle is complex when compared to the prokaryotic cell cycle as eukaryotic cells contain multiple chromosomes within a nucleus. For most eukaryotic cells the cell cycle can be divided into four distinct phases (Figure 1-1) which are known as Gap1 (G_1), Synthesis (S), Gap2 (G_2), and Mitosis (M) (Alberts *et al.*, 2008). Following a previous successful cell division the dividing cells first enter the G1 phase, which corresponds to the growth of cells. During this phase the cell integrates mitogenic and growth inhibitory signals and makes the decision to proceed, pause, or exit the cell cycle.

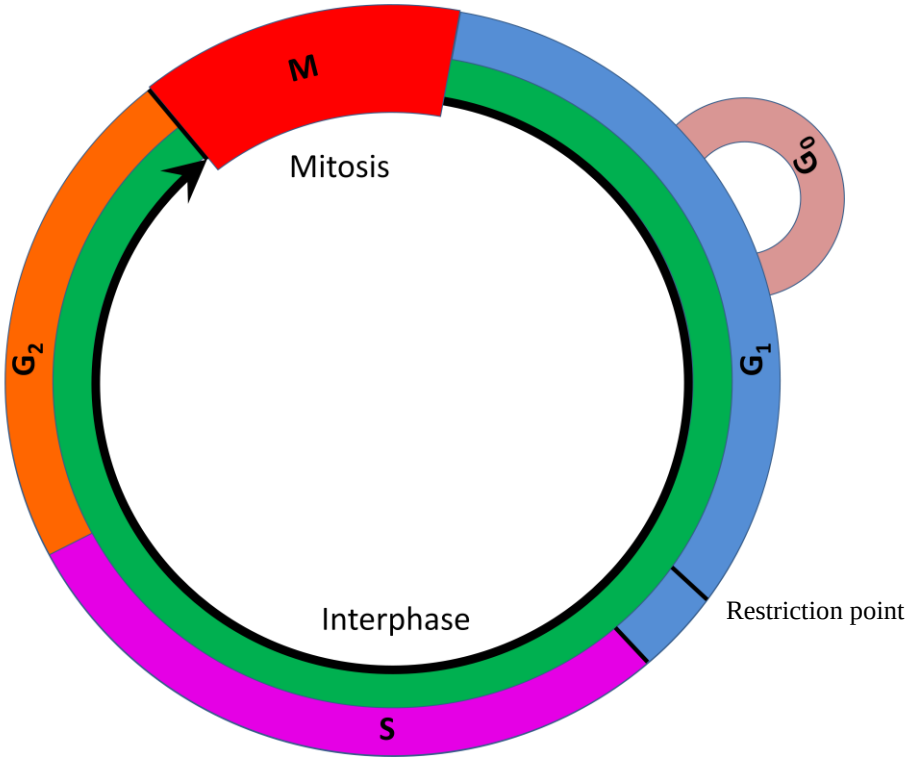


Figure 1-1 The eukaryotic cell cycle. The eukaryotic cell cycle is divided into four distinct phases, which are G_1, S, G_2 and M, respectively. In the absence of a signal for cell division cells enter into a quiescence state known as G_0. Different phases of the cell cycle are regulated by different cyclin dependent kinases.

The G_1 phase is followed by a period of DNA synthesis (S phase) where the replication of genetic material takes place. During the S phase chromatids are also duplicated and this duplication of chromatids keeps genetic material conserved for each daughter cell. The cells then enter the G_2 phase during which cells undergo rapid growth to prepare for mitosis. Finally during the M phase separation of the duplicated chromatids takes place into two sets of chromosomes followed by a cell division resulting in two identical daughter cells (Nasmyth, 2002; Alberts *et al.*, 2008; Cooper and Hausman, 2009). The G_1, S and G_2 phases are collectively known as interphase. Cells spend the majority of their time in the interphase. A typical human cell divides approximately every 24 hours (Alberts *et al.*, 2008). In the absence of a signal for cell

division a cell enters into a quiescence phase also known as G_0 (Nasmyth, 2002; Alberts *et al.*, 2008; Cooper and Hausman, 2009).

1.3 Cell cycle regulation

A complex regulatory network that controls the order and time steps for different phases of the cell cycle is named as cell cycle control system. Different phases of the cell cycle are controlled by different cyclin dependent kinases (CDKs) and associated cyclins. In order to understand the cell cycle regulation it is necessary to understand structure and function of kinases in general and CDKs with reference to their particular role in the cell cycle.

1.3.1 Kinases

Kinases are a class of enzymes that catalyze the transfer of a phosphate group from energy rich donor molecules, such as ATP, to specific substrates in the cell in a process known as phosphorylation. Phosphorylated proteins are ubiquitous both in prokaryotes and eukaryotes. About 30% of proteins encoded by human genome are phosphorylated (Cohen, 2002a). Protein phosphorylation is a reversible signalling process in the regulation of proteins and is carried out by a group of kinases known as protein kinases (Hunter, 2000; Cohen, 2002a; Cohen, 2002b). Protein kinases represent a large family of proteins in eukaryotes, which plays an important role in the regulation of metabolism, differentiation, transcription, cytoskeletal rearrangement, apoptosis, and a wide variety of signal transduction processes in the cell (Ullrich and Schlessinger, 1990; Kato *et al.*, 1993; Cohen, 2002b; Scheeff and Bourne, 2005). Eukaryotic protein kinases can be divided into two main groups, which are eukaryotic protein kinases (ePKs) and atypical kinases (aPKs). These two groups can be further divided into 12 families (Miranda-Saavedra and Barton, 2007; Martin *et al.*, 2009). ePK are one of the largest protein families, comprising 1.5–2.5% of all eukaryotic genes (Manning *et al.*,

2002a; Manning *et al.*, 2002b). The ePKs represent eight families of protein kinases namely the AGC, CAMKs, CK1, CMGC, RGC, STE, TK and TKL families. The aPKs represent a relatively small group of protein kinases, which exhibit protein kinase activity but do not have sequence similarity with ePKs (Koike *et al.*, 2001). The aPKs are divided into four groups, namely alpha kinases, PIKK, RIO and PHDK (Miranda-Saavedra and Barton, 2007; Martin *et al.*, 2009).

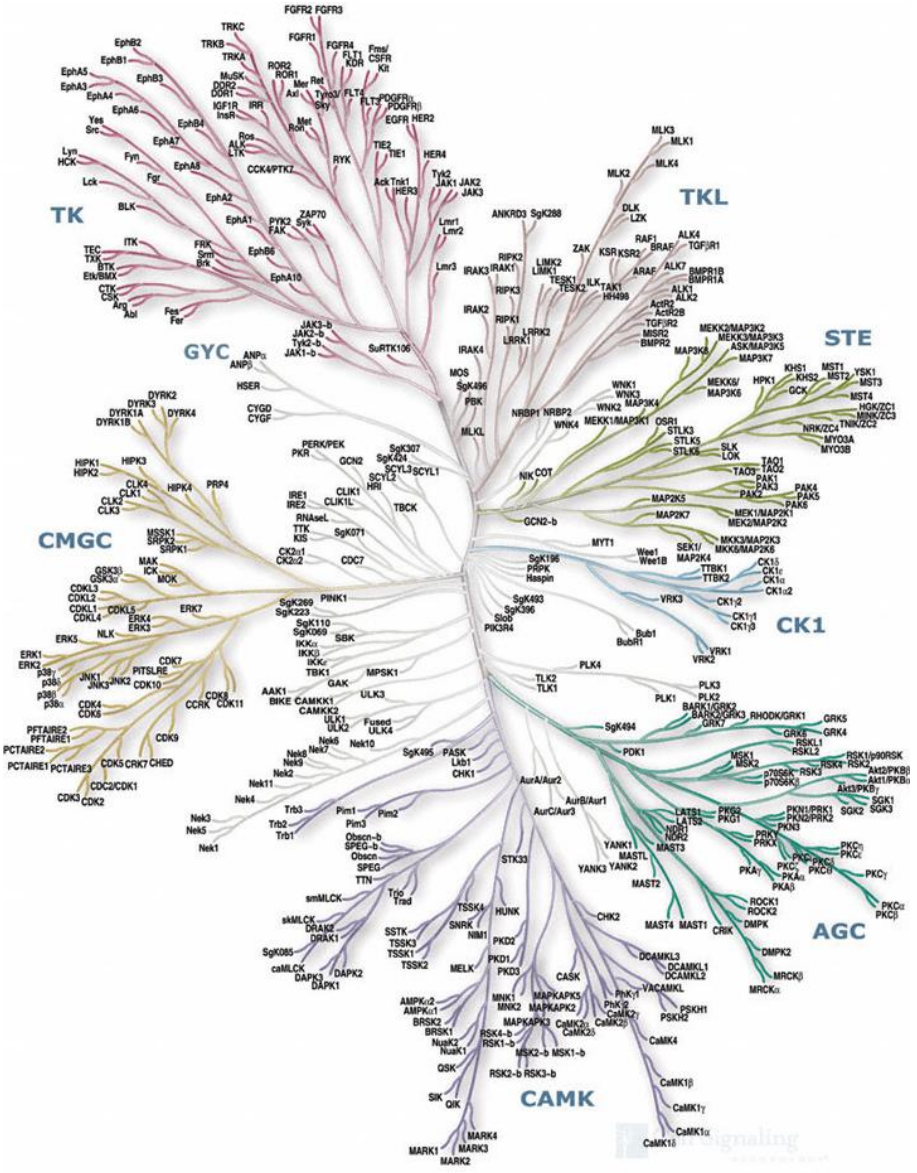


Figure 1-2 Overview of human kinome. The human kinome is comprises 478 ePKs and 40 aPKs based on the sequence similarities in the catalytic domain (Manning *et al.*, 2002b). Kinome Illustration reproduced courtesy of Cell Signaling Technology, Inc. (www.cellsignal.com)

The protein kinase complement of the human genome is named as human kinome (Figure 1-2). There are about 518 kinases in the human genome out of which 478 are classified as ePK and 40 as aPK based on the sequence similarities in the catalytic domain (Manning *et al.*, 2002b).

Protein kinases have a bilobal structure namely the N and C-lobe, with a catalytic core and ATP binding site located in the cleft between these lobes. The adenine ring of ATP forms hydrogen bonds with the kinase hinge and the γ-phosphate of ATP is oriented outwards of the binding pocket. Kinases have a conserved set of secondary structure elements arranged into 12 sub-domains (see Table 1-1 and Figure 1-3) indicated by roman numerals I–XI, with VI being divided into VIa and VIb (Bossemeyer, 1995; Hanks and Hunter, 1995; Zhang *et al.*, 2009). Within these sub-domains certain residues are recognized as being invariant in the kinase superfamily (Hanks and Hunter, 1995). Sub-domains I, II, VIB, and VII are considered as the most important because of their direct role in ATP binding and activation of kinases. Sub-domain I consist of a β-sheet with a GXGXXG motif essential for ATP binding and sub-domain II includes a catalytic lysine residue, which binds to ATP. Sub-domain VIB contains an invariant aspartate, which mediates transfer of a phosphate group from ATP to the substrate (Bossemeyer, 1995; Hanks and Hunter, 1995; Zhang *et al.*, 2009).

Table 1-1 : Subdomain descriptions of kinase domain in protein kinase A (PKA) and cyclin dependent kinase 2 (Hanks and Hunter, 1995).

Subdomain	**Residue Number**		**Subdomain**	**Residue Number**	
	PKA	**CDK2**		**PKA**	**CDK2**
Subdomain I	43-64	4-25	Subdomain VIB	161-177	122-138
Subdomain II	65-83	26-43	Subdomain VII	178-193	139-157
Subdomain III	84-98	44-58	Subdomain VIII	194-210	158-175
Subdomain IV	99-113	59-73	Subdomain IX	211-240	176-209
Subdomain V	114-137	74-98	Subdomain X	241-260	210-254
Subdomain VIA	138-160	99-121	Subdomain XI	261-297	255-282

The activation loop of the protein kinases is located in the region between kinase sub-domains VII and VIII. The activation loop contains a conserved DFG (aspartic acid, phenylalanine and glycine) motif in subdomain VII, which is common in all kinases. The activation loop plays an important role in the regulation of kinase activity by adopting active and inactive conformations. The active conformer is usually phosphorylated. The inactive conformer blocks the binding of substrate to the kinase binding site.

Ray Erikson investigated the role of protein kinases as early as in disease in 1978 (Collett and Erikson, 1978; Cohen, 2002b). Since then protein kinases have been found to be involved in mechanisms of many diseases. In cells almost all signal transduction and regulatable process are regulated by a phosphotransfer reaction catalyzed by kinases (Bossemeyer, 1995; Zhang *et al.*, 2009). Any deregulations of the activities of kinases can cause many disorders which include cancer, immunological, metabolic, neurological and infectious diseases (Zhang *et al.*, 2009). The development of kinase inhibitors has made the kinome one of the key target families for the development of new therapeutics. The kinases are now the second largest drug target for the pharmaceutical companies after the G-protein coupled receptors family (Cohen, 2002b; Manning *et al.*, 2002b; Yan *et al.*, 2006).

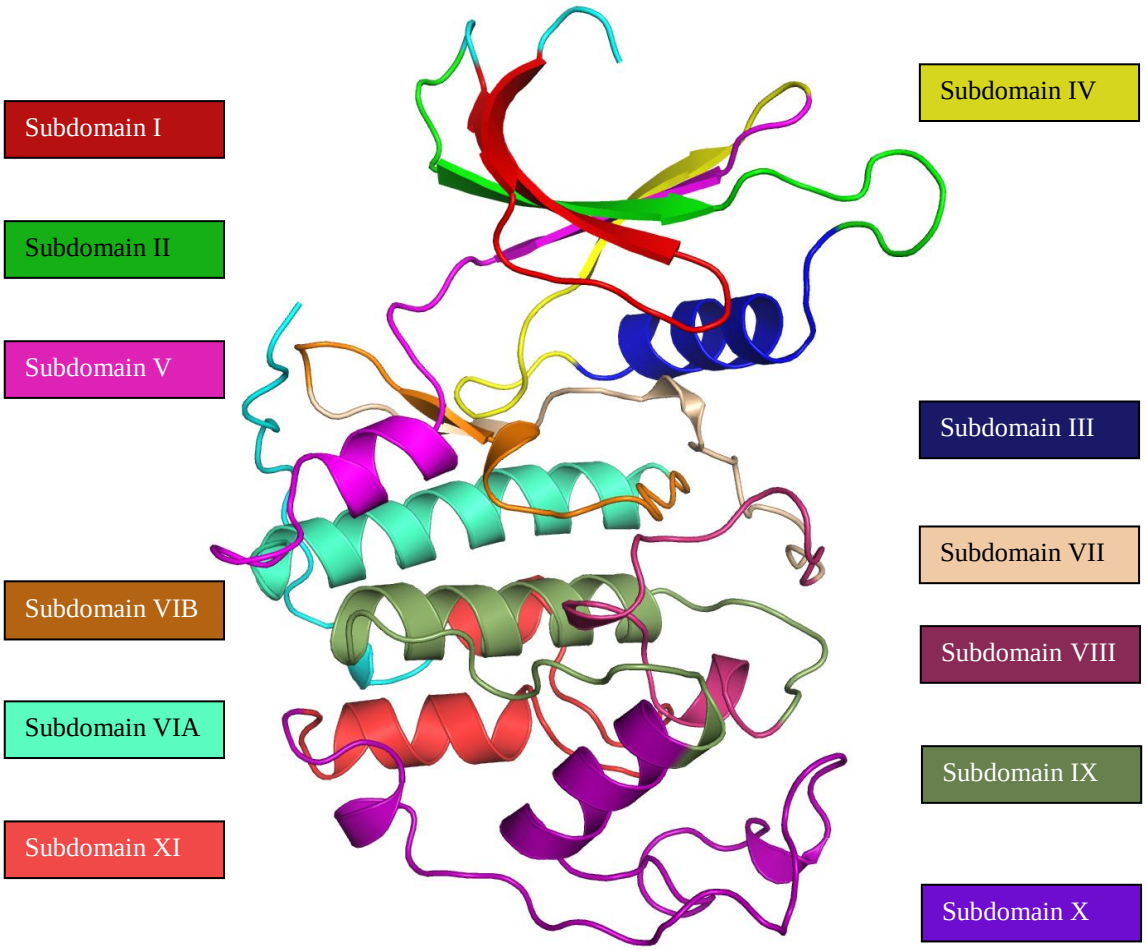


Figure 1-3 Twelve sub-domains of CDK2. Kinases have a conserved set of secondary structure elements arranged into twelve sub-domains, a breif description of these sub-domain residues is given in Table 1-1. Subdomains I, II, VIB, and VII are considered most important because of their direct role in ATP binding and activation of kinases.

Approximately 30 kinase targets are in Phase I clinical trial and many of these targets are being investigated for their role in cancer treatment. The US Food and Drug Administration Authority has already approved 11 kinase inhibitors for cancer treatment (Grant, 2009; Zhang *et al.*, 2009). The development of kinase inhibitors is facing many challenges such as selectivity and efficacy of the kinase inhibitors and the difficulty to validate a particular kinase as a drug target for a particular disease, which may be due to high degree of conservation at ATP binding site (Cohen, 2002b).

Kinases can be divided into three classes based on their role in tumorigenesis. The kinases in the first class are supposed to be involved in the process leading to

tumorigenesis and are considered as oncogenic e.g. BRAF and JAK. The second class of kinases is required for the survival and proliferation of cancer cells, this class includes MEK1, MEK2 and cyclin dependent kinases (CDKs). The third class represents kinases, which are expressed in the tumour or nearby tissues and are needed for the formation and maintenance of different stages of a tumour. These kinases are also required in developing and sustaining the tumour blood supply e.g. NTRK2, VEGFR and FGFR kinases (Zhang *et al.*, 2009).

1.3.2 Cyclin dependent kinases (CDKs)

The cyclin-dependent kinases (CDKs) (EC 2.7.1.37) represent a large family of serine/threonine protein kinases (~34-40 kDa), that play a well established role in the regulation of the eukaryotic cell division cycle (Norbury and Nurse, 1992; Morgan, 1995; Morgan, 1997; Harper and Adams, 2001). CDKs have also been implicated in the control of gene transcription and in the processes that integrate extracellular and intracellular signals for the coordination of the cell cycle in response to environmental change (Morgan, 1997; Malumbres and Barbacid, 2005; Suzek *et al.*, 2007). They also play a role in apoptosis (Morgan, 1997; Murray, 2004).

The first member of the CDK family, CDK1 (also known as CDC2) was identified by the Nobel Laureate Sir Paul Nurse while studying the cell division in *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* (Russell and Nurse, 1986). The second member of the CDK family designated as CDK2 was identified in *S. cerevisiae* c28 mutants (Elledge and Spottswood, 1991). Human CDC2 and CDK2 share homology with fission yeast CDC2 and budding yeast CDC28. There is about 60-65% amino acid sequence identity between human CDC2 and CDK2 and their yeast counterparts (Norbury and Nurse, 1992). Other CDK family members have 35%-65% sequence identity to the prototypes CDC2 and CDC28 (Morgan, 1997). In contrast to

yeast where a single CDK controls all the major cell cycle transitions, the higher eukaryotes and human cells require more CDKs. Five cell cycle CDKs in mammals regulate the cell cycle indicating specialization of CDKs for particular phases of the cell cycle (Liu and Kipreos, 2000a). CDK1, CDK2, CDK4 and CDK6 are involved directly in the cell cycle control system, as discussed in the following sections. CDK7 plays a key role in the activation of other CDKs and is also known as CDK-activating kinase (CAK). Other CDKs do not play a direct role in the cell cycle regulation, but some of these may have an auxiliary role. There are 11 classical CDKs according to the criteria establish by the "Cold Spring Harbor Symposium on Cell Cycle" in 1991 (Malumbres and Barbacid, 2005). According to this criterion no kinase can be named as a 'CDK' unless its activating partner molecule is cyclin or a cyclin-like regulatory subunit (Malumbres and Barbacid, 2005; Malumbres *et al.*, 2009). After the establishment of a unifying nomenclature for cyclin dependent kinases CDC2 was renamed to CDK1, PSK-J3 was assigned the name CDK4, PSSALRE became CDK5 and PLSTIRE was renamed to CDK6 (Malumbres and Barbacid, 2005). A new classification of CDKs has very recently been proposed (Malumbres *et al.*, 2009). The nomenclature for CDKs based on this proposal suggests that there are 26 CDKs and CDK like proteins known for human. According to current classification some CDKs do not have a partner cyclin (Malumbres *et al.*, 2009).

1.3.2.1 CDK2

CDK2 binds to cyclin E at the onset of the S-phase of the cell cycle and induces the initiation of DNA synthesis. When a cell passes through the S-phase of the cell cycle CDK2 interacts with cyclin A, and the CDK2/cyclin A complex plays a role in progression through DNA synthesis. CDK2 is one of the structurally best-characterized kinases with more than 190 experimentally solved structures (see Appendix 1.1)

deposited in the Protein Data Bank (PDB). A 3D-representation of the structure of human CDK2, whose major structural features are probably conserved in all CDKs, is shown in Figure 1-4. It contains a smaller N-terminal lobe that is dominated by a β-sheet, and a larger C-terminal lobe that is primarily α-helical. The N-terminal lobe consists of five anti-parallel β-strands and a α-helix (α1) also known as PSTAIRE helix. The C-terminal lobe comprises six α-helices (α2-α7) and two β-strands (β6-β7) (De Bondt *et al.*, 1993). The adenine base of ATP fits into a hydrophobic pocket within the cleft between the smaller N-terminal lobe and the larger C-terminal lobe. The cyclin protein binds at the entrance of the ATP binding site in a position that allows the γ-phosphate of ATP to face the hydroxylated side chain on the protein substrate, which presumably binds along the upper surface of the large lobe at the entrance of the cleft when the kinase is in its active confirmation (De Bondt *et al.*, 1993). The ATP molecule interacts with several residues in the binding pocket. The phosphates of ATP interact with backbone amides of the glycine rich loop GxGxxG (residues 11-16) by making ionic and hydrogen bonding interactions with them. In addition to these interactions ATP also interacts with Lys 33 and Asp 145.

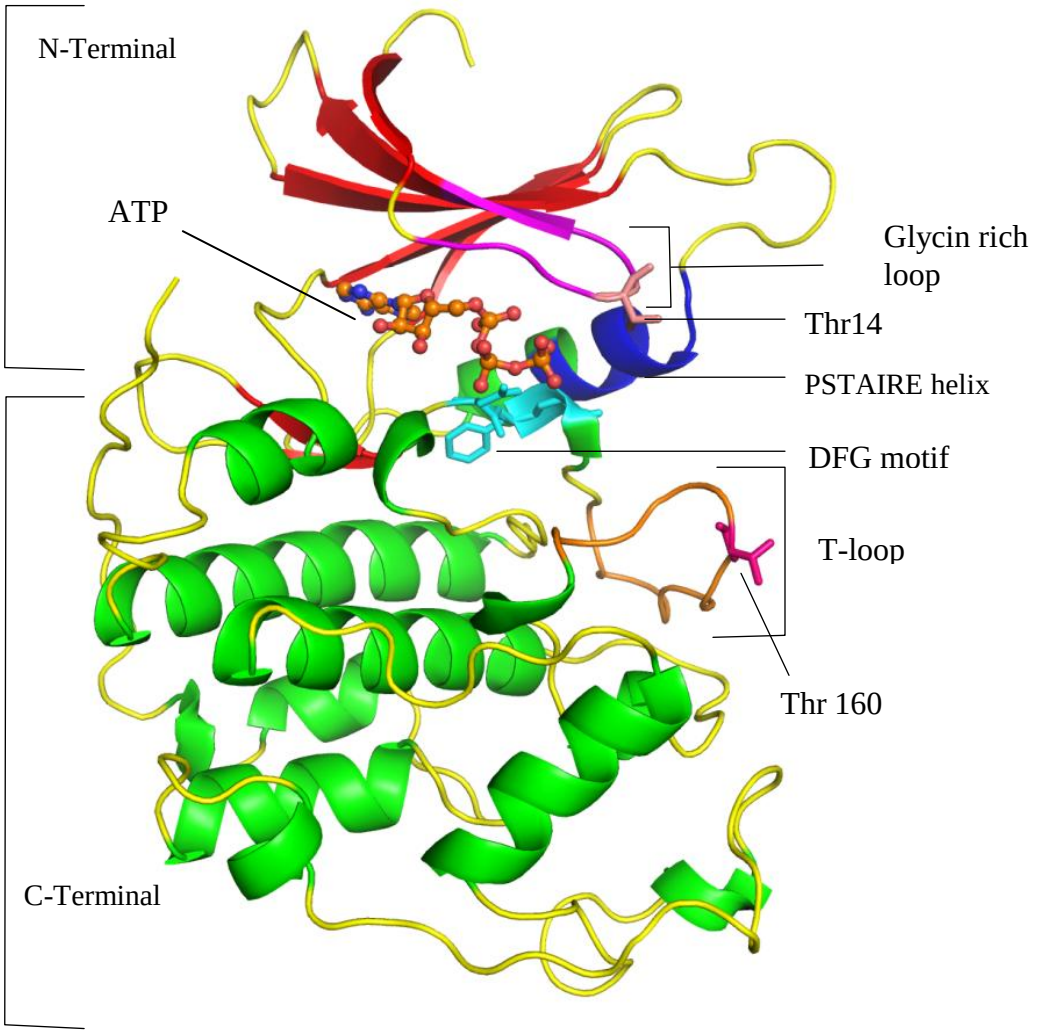


Figure 1-4 A cartoon representation of the structure of human CDK2 with ATP. The structure of CDK2 [PDB ID: 1FIN (Jeffrey *et al.*, 1995b)] shows the N-terminal β sheet and the PSTAIRE helix. ATP is shown in sticks representation in the active site of the CDK2. The inhibitory phosphorylation sites Thr14 and activating phosphorylation site Thr 160 are shown in sticks. T-loop and is PSTAIRE helix are highlighted in brown and blue.

Kinase activity of the CDK2 is regulated by cyclin A/E binding and phosphorylation. The binding of a corresponding cyclin molecule or phosphorylation alone does not fully activate CDK2 (Gu Y, 1992; Connell-Crowley *et al.*, 1993; Morgan, 1997; Kontopidis *et al.*, 2006; Larochelle *et al.*, 2007). The CDK2 cyclin complex becomes fully active by its phosphorylation at Thr160 by CDK-activating kinase (CAK). CDK2 can also be negatively regulated to some extent by its phosphorylation at Thr14 (De Bondt *et al.*, 1993; Morgan, 1997; Larochelle *et al.*,

2007). Thr160 is located on the T-loop as shown in Figure 1-4, cyclin binding exposes this residue to the solvent allowing phosphorylation by CAK.

1.3.2.2 CDK4 and CDK6

CDK4 and CDK6 are homologous proteins that integrate mitogenic and antimitogenic extracellular signals with the cell cycle (Sherr, 1996; Morgan, 1997). CDK4 plays a very important role in regulation of the G_0-G_1 phase of the cell cycle i.e. the emergence of the cell from quiescence. CDK4 activity is also required in the regulation of G1/S phase transition and entry into the S-phase of the cell cycle. CDK4/CDK6 with their partner D-type cyclins are responsible for the phosphorylation of the retinoblastoma gene product (Sherr, 1996). The retinoblastoma protein acts as negative regulator of E2F family of transcription factors. The phosphorylation of the retinoblastoma protein (pRb) inactivates it and thus results in the release of the E2F family of transcription factors (i.e. E2Fs 1–3). The E2F factors activate the expression of the S-phase genes enabling the cell to pass through the restriction point and resulting in the onset of the S-Phase (Sherr, 1996; Dyson, 1998; Attwooll *et al.*, 2004). CDK4 and CDK6 do not fully overlap and compensate each other's role in different cells as both are required for the normal functioning of many tissues and cells type (Rane *et al.*, 1999; Tsutsui *et al.*, 1999; Jirawatnotai *et al.*, 2004). It has been reported that in the embryonic development of mice CDK4/6 and their associated cyclin counterpart are dispensable (Malumbres *et al.*, 2004). However CDK6 cannot compensate for CDK4 in CDK4/CDK2 double knockout mouse models (Berthet and Kaldis, 2006). Cellular proliferation in many vertebrate cells requires external growth factors for the activation of CDK4 and its close relative CDK6. To control cellular proliferation inhibition of CDK4/CDK6 is necessary. The activity of CDK4 in normal cells is controlled by the

Ink4 family CDK Inhibitors (CKIs) such as P16 (Morgan, 1997; Soni *et al.*, 2000). The role of CDK4 as a drug target is outlined in Section 1.5.

1.3.3 Other CDKs

CDK1 (also known as CDC2) is believed to play a role in the control and preparation for mitosis. Mitosis is initiated by the CDK1-cyclin B complex, which is also known as M-phase promoting factor (MPF) (Nigg, 1995; Edgar and Lehner, 1996; Morgan, 1997). CDK3 is closely related to CDK2 in terms of sequence similarity and it also plays a role in the G-S transition in mammalian cells (Hofmann and Livingston, 1996). CDK5 plays a role in the central nervous system and is required during neural differentiation. It is expressed in post-mitotic cells of the central nervous system (Ohshima *et al.*, 1996; Wei and Tomizawa, 2007). The CDK7/Cyclin H complex (CAK) does not only plays a role in the activation of CDKs by their phosphorylation but also play a role in transcription. In humans this complex is also associated with TFIIH (RNA polymerase II general transcription factor) and phosphorylate C-terminal domain (CTD) of RNA polymerase II during transcription. The CDK7 homologue in budding yeast *S. cerevisiae* is Kin28. Kin 28 plays its role only in the regulation of transcription and does not carry any CAK activity (Cismowski *et al.*, 1995). The closest CDK7 relatives in fission yeast *S. pombe* are Mcs6 and Csk1 which both show CAK activity (Saiz and Fisher, 2002). In addition to cell cycle CDKs, CDK9 inhibition also contributes to the anticancer activity (Wang and Fischer, 2008; Nowicki and Walkinshaw, 2010).

1.3.4 Activation, deactivation and regulation of CDKs

CDK activity in most CDKs is controlled via four different mechanisms. These are activation by binding to corresponding cyclins, inhibition by binding to cyclin-

dependent kinase inhibitors (CKIs), inhibitory phosphorylation of the CDK, and activating phosphorylation of the CDK (Morgan, 1997; Liu and Kipreos, 2000b). The active site of the CDKs is blocked in the absence of cyclin (De Bondt *et al.*, 1993; Morgan, 1997). The first step in the mechanism of CDK activation is the binding of a cyclin (Gu Y, 1992; Connell-Crowley *et al.*, 1993; Jeffrey *et al.*, 1995a). However, CDKs do not become fully active on binding with corresponding cyclins, but some CDKs show a partial activation on binding with Cyclin (Desai *et al.*, 1992; Gu Y, 1992; Connell-Crowley *et al.*, 1993). The activity of CDKs is further regulated by phosphorylation as complete activation of most CDKs requires phosphorylation of the CDK at a conserved threonine residue by CAK (CDK7) (Morgan, 1997; Liu and Kipreos, 2000b).

The activities of the CDKs can be suppressed in the cell either by cyclin destruction, decreased cyclin gene expression, inhibition of CDK activity by CDK inhibitor proteins (CKIs) that bind and inactivate cyclin-CDK complexes or by small molecule inhibitors competing with ATP (De Bondt *et al.*, 1993; Morgan, 1997).

1.3.5 Cyclins

Sir R. Timothy Hunt (Nobel laureate in Physiology 2001) discovered cyclins in 1982 while studying the cell cycle of sea urchins (Evans *et al.*, 1983). Cyclins are named cyclin because of cyclic synthesis and destruction of members of this family of proteins during different phases of cell cycle. Cyclins play a significant role in cell cycle regulation by binding and activating the corresponding cyclin dependent kinases. More than 16 mammalian cyclins have been indentified and all of these share a common homologous region known as cyclin box (Nugent *et al.*, 1991). The cyclin box is a relatively conserved stretch of 150 amino acid residues and it is used to bind and activate CDKs (Malumbres and Barbacid, 2005). In human cells different cyclins are

expressed during the four phases of the cell-cycle (Johnson and Walker, 1999). The concentration of these cyclins varies over time during these four stages of the cell cycle (Morgan, 1997). Cyclin D1, D2 and D3 belong to the G1-phase of the cell cycle and bind to CDK4 and CDK6. D-type cyclins do not oscillate during the cell cycle in contrast to many other cyclins (Johnson and Walker, 1999). The S-phase cyclin, cyclin E binds to and activates CDK2. The CDK2 activation initiates DNA replication. Cyclin A binds to CDK2 during S-phase and is required for the cell to progress through the S-phase. During the M phase cyclin B is expressed which binds to and activates CDK1. The activation of CDK1 is required for various mitotic activities.

The active complex of CDK4/Cyclin D1 plays a regulatory role at the G1/S checkpoint as shown in Figure 1-5. The active retinoblastoma protein (pRb) binds to and inactivates the E2F protein. A threonine residue (Thr826) of the pRb is phosphorylated by the active complex of CDK4/Cyclin D1 and thus making it inactive (Zarkowska and Mittnacht, 1997; Takaki *et al.*, 2005). The phosphorylated pRb releases the E2F protein, which then induces the S-phase gene transcription.

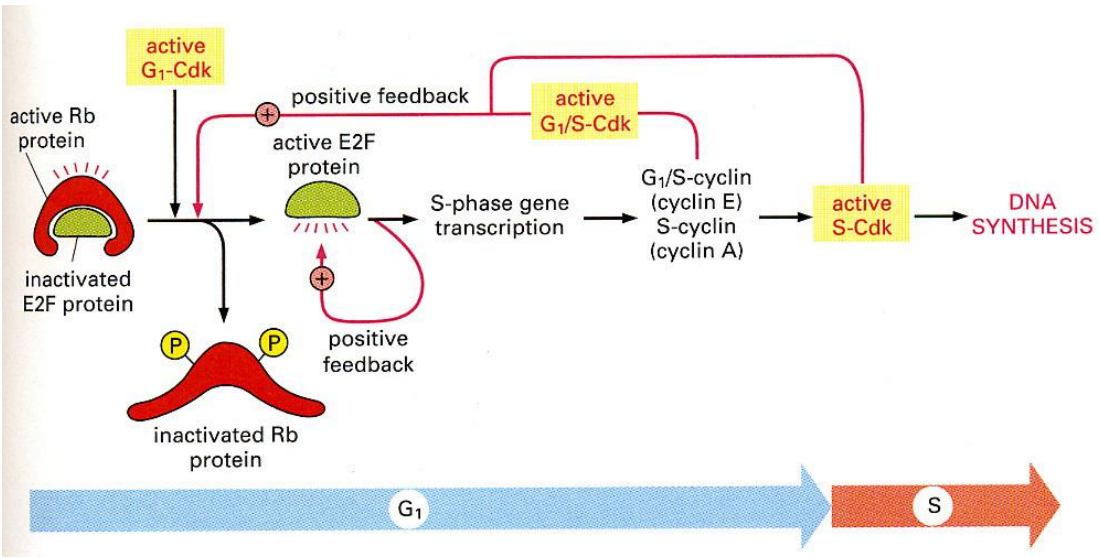


Figure 1-5 The G1/S phase checkpoint (From Molecular Biology of the Cell 4E by Alberts *et al.* reproduced with permission of Garland Science/Taylor and Francis LLC)

1.4 Cancer

Cancer is a disease that represents the accumulation of genetic alterations related to cell cycle control and cellular proliferation. Alterations in the machinery that controls the decisions in cell division can prove fatal to an organism (Hanahan and Weinberg, 2000; Kastan and Bartek, 2004). Normal cells proliferate in a very controlled fashion, only when required to do so in response to growth signals and cellular needs. Cancer cells exhibit a deregulated cell cycle control and homeostasis. An uncontrolled cellular proliferation no longer subject to inhibitory influences results in an accumulation of unwanted cells that leads to the growth of a neoplasm or tumour. The mutated genes that help or contribute in the transformation of normal cells into malignant cancers are known as oncogenes. In 1978 an oncogene was found to be a protein kinase that functions as a transforming factor of the Rous sarcoma virus (Collett and Erikson, 1978; Cohen, 2002b). Mutation of different oncogenes and tumors suppressors such as Ras, Rb, p53, p16, P13K effect the cellular proliferation and developmental process and contribute toward the development of cancer.

The tumorigenesis in human is a multiple stage process with four to seven rate-limiting, stochastic events in which cells evolve from normalcy to invasive cancers via pre-malignant states (Renan, 1993; Bergers *et al.*, 1998; Hanahan and Weinberg, 2000). The process of conversion of normal human cells into cancer cells is analogous to the Darwinian evolution as it involves random mutations (Hanahan and Weinberg, 2000). There are more than 100 distinct types of cancer, however despite this diversity most or perhaps all of human cancers display a manifestation of six essential hallmarks (Figure 1-6) in cell physiology that collectively dictate cellular proliferation and malignancy. These six hallmarks in cancer cell physiology are self-sufficiency in growth signals, resistance to antiproliferative signals, anti apoptotic behaviour, deregulation of the cell

replicative potential, sustained angiogenesis, and invasion of surrounding tissues and metastasis to the distal organs (Hanahan and Weinberg, 2000).

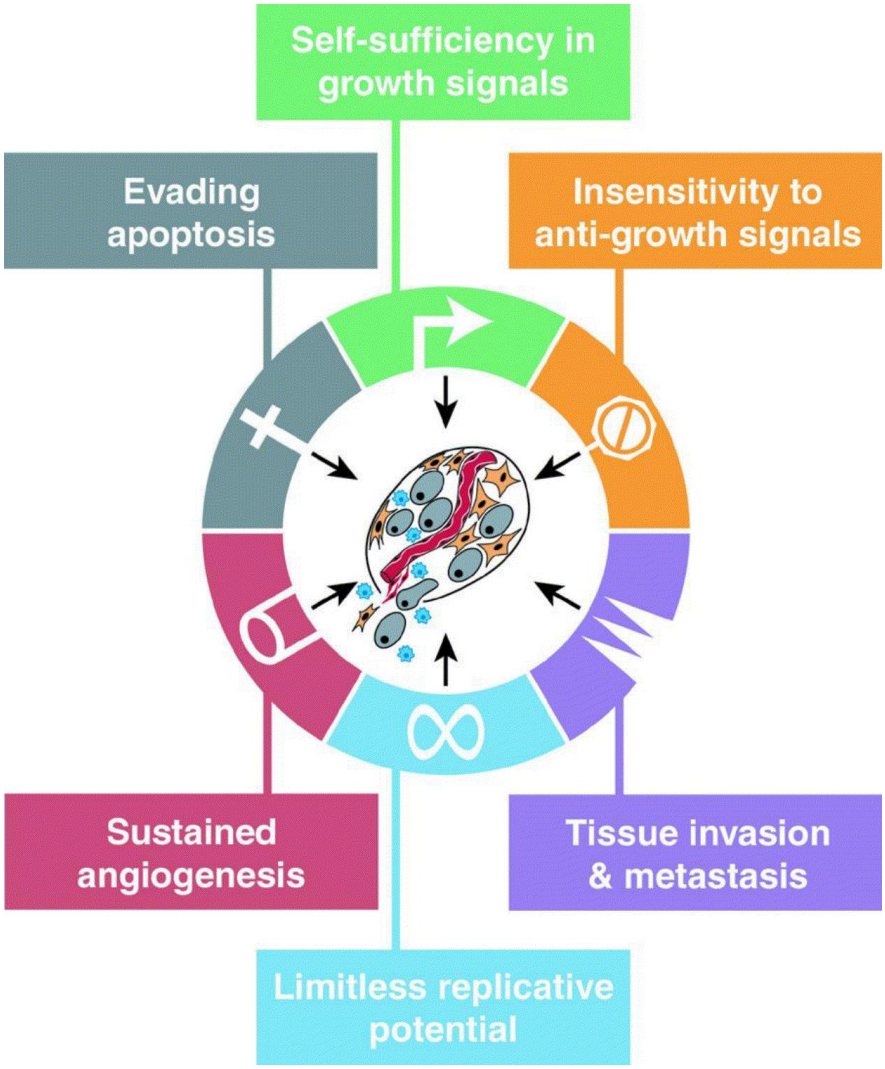


Figure 1-6 Six essential hallmarks of cancers. Self sufficiency in growth signals, anti growth signals, metastasis, limitless replactive potential, sustained angiogenesis and evading apoptosis are the acquired capbilities of most of the cancers (Hanahan and Weinberg, 2000). Reprinted with permission from Elsevier.

1.4.1 Principle of cancer therapies

Cancer treatment is a one of the most complex aspects of medical care due to complexity of the mutations and diversity of alterations in different type of cancers (Luo 2010). There has been a tremendous progress in the cancer research during the last

few decades. This has helped to understand this disease and the development of anticancer strategies. Traditionally cancer can be treated either by surgical removal of tumours or destruction of cancer cells either by chemotherapeutic drugs or by radiotherapy. Surgical removal, chemotherapy and radiotherapy are also used in different combinations to treat different types of cancers. The decision to use any of these of a combination of traditional treatments depends on the type and staging of a cancer. The effective use of cancer chemotherapy requires a thorough understanding of the principles of functional nodes in the oncogenic network. Anticancer treatments mostly rely on certain molecular, biochemical, and cellular features of cancerous or tumor cells, which distinguish them from normal cells e.g. the higher rates of cell division taking place in cancerous cells compared to normal cells. Anticancer treatments target cancer cells after identifying certain oncogenic features to bring a system failure and apoptosis of cancer cells (Luo 2010). There are different therapies used to treat cancer that include use of alkylating agents, protein kinase inhibitors, hormone therapies and monoclonal antibodies (Crosignani, 2003) (Green *et al.*, 2000) (Dubowchik and Walker, 1999). The role of protein kinase inhibitors in treatment of cancer is described in Section 1.6.

1.5 CDK4 as a target for cancer drug discovery

The activity of CDK4 is considered as a direct or indirect target of genetic alterations in cancer (Ortega *et al.*, 2002a; Malumbres and Barbacid, 2006). CDK4 is found as constitutively activated in many human cancers either due to over-expression of cyclin D, mutation(s) in CDK4 catalytic subunit or due to deletion of the p16 protein, which acts as a CDK4 inhibitor in the cell (Okamoto *et al.*, 1994; Palmero and Peters, 1996). The activity of CDK4 is misregulated in 60-70% of human cancers (Soni *et al.*,

2000). CDK4/CyclinD1 has been validated as an anti cancer drug target in MCF-7 breast cancer cells (Grillo *et al.*, 2006).

The G1/S transition is dependent on external growth factors and is initiated by the CDK4-cyclin D1 complex phosphorylating the retinoblastoma protein (Weinberg, 1995; Bartek *et al.*, 1996a; Bartek *et al.*, 1996b). The alterations and lack of growth control in cancer cells is a result of changes in the regulatory pathways involved in cell cycle control (Ortega *et al.*, 2002b). The transition from G1 to S phase is deregulated in many cancers (Deshpande *et al.*, 2005a). The restriction points in G1/S transition play a decisive role for the proliferation of a cell (Blagosklonny and Pardee, 2002). The external mitogenic factors do not influence the cellular proliferation after G1/S transition (Blagosklonny and Pardee, 2002). The CDK4-cyclin D and retinoblastoma pathway is among the frequently disrupted pathway in breast carcinomas, sarcomas, gliomas, human hepatoma cell lines and hepatocellular carcinoma (Wei *et al.*, 1999; Laurent-Puig and Zucman-Rossi, 2006; Graf *et al.*, 2009; Rivadeneira *et al.*, 2010). The retinoblastoma functionality is missing at a relatively high frequency in hepatocellular carcinoma. Inhibition of CDK4-cyclin D results in pRb hypophosphorylation, which prevents cell proliferation. Studies have shown that inhibition of CDK4/6 with small molecule inhibitors significantly controls the hepatocyte proliferation (Rivadeneira *et al.*, 2010). The CDK4 knock out in mice has shown a resistance to tumor development (Rodriguez-Puebla *et al.*, 2002). CDK4 and CDK6 both are involved in oral cancer; however CDK4 is supposed to play its role in the early event of carcinogenesis of oropharyngeal squamous cell carcinoma (Poomsawat *et al.*, 2010).

The ongoing research in this area has made the CDK4-cyclin D complex an attractive molecular target for cancer therapy and supports the concept that inhibition of CDK4 activity may have therapeutic value for the cancer patients. Different research

groups have reported a large number of CDK4 inhibitors (Honma *et al.*, 2001; Ikuta *et al.*, 2001; García *et al.*, 2006; Jenkins *et al.*, 2008; Slamon *et al.*, 2010).

1.6 Protein kinase inhibitors

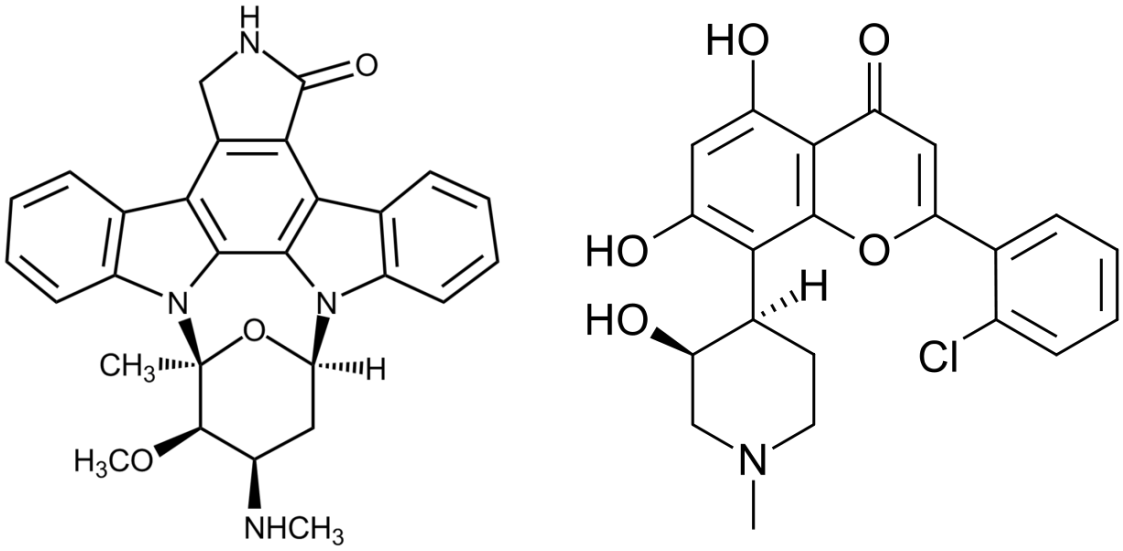
The first protein-kinase inhibitors were reported in 1980s (Hidaka *et al.*, 1984; Cohen, 2002b). Hidaka *et al*. observed that naphthalene sulphonamides which were developed as antagonists of the calcium binding protein calmodulin also inhibit several kinases at higher concentration (Hidaka *et al.*, 1984; Cohen, 2002b). These early protein kinase inhibitors were not very specific as they exhibit inhibition for several protein kinases (Hidaka *et al.*, 1984; Cohen, 2002b). One of the early kinase inhibitors, the compound AT877 (fasudil hydrochloride) was selected for human clinical trials in 1990 (Shibuya *et al.*, 1990) and it was approved in Japan for the treatment of cerebral vasospasm in 1995 (Asano *et al.*, 1998). Glivec (STI571, imatinib) was first kinase inhibitor successfully used for the treatment of cancer and was approved for clinical use in May 2001. It is also being marketed as Gleevec by Novartis (USA). Glivec is used to treat chronic myeloid leukaemia which is characterised by a mutant chromosome (the Philadelphia chromosome) formed as a result of fusion of chromosome 9 and 22 (Capdeville *et al.*, 2002). It is used to block the activity of a mutant tyrosine kinase encoded by the mutant Philadelphia chromosome. Glivec treatment provides a success in the cancer treatment (Capdeville *et al.*, 2002; Druker, 2002; Levitzki, 2002). Imatinib is dominant in the CML market and in less than a decade imatinib reached about $1.1 billion of sales in US, with global sales approaching $3.95 billion (Storey, 2009; Aggarwal, 2010).

1.7 Small molecule inhibitors for CDKs

A large number of small molecule inhibitors of CDKs have been synthesised and some of these have already been approved for different stages of clinical trials (Grant and Roberts, 2003; Grant, 2009). CDK inhibitors can be divided into two categories based on their selectivity, i.e. selective and non selective inhibitors. The non selective compounds inhibit different CDKs and other unrelated serine/threonine and tyrosine kinases in similar concentrations. Staurosporine is an example of non selective inhibitor of CDKs. Purine analogs olomoucine, roscovitine, and purvalanols represent relatively specific inhibitors of CDKs. These compounds are very useful lead compounds in anti-cancer drug discovery. Inhibitors with improved specificity for individual CDKs are required to enhance their therapeutic potential in cancer treatment.

1.7.1 Staurosporine

Staurosporine is a natural product (Figure 1-7) isolated from *Streptomyces staurosporeus* and has antimicrobial, anti fungal, anti-hypertensive and protein kinase inhibition activity (Omura *et al.*, 1977; Ruegg and Burgess, 1989). Staurosporine is an example of a naturally occurring ATP competitive inhibitor and was originally identified for its activity to inhibit protein kinase C (Tamaoki *et al.*, 1986). Funato *et al.* elucidated the X-ray structure of staurosporine in 1994 (Funato *et al.*, 1994). Its structure overlaps well with the adenosine group of ATP and for this reason it is a potent, but unselective inhibitor of protein kinases and CDKs (Toledo and Lydon, 1997).



a) Staurosporine b) Flavopiridol

Figure 1-7 Molecular structures of protein kinase inhibitors. a) staurosporine isolated from *Streptomyces staurosporeus* is an unselective inhibitor of kinases **b)** flavopiridol inhibits multiple CDKs such as CDK1, CDK2, and CDK4

After the discovery of staurosporine the pharmaceutical industry has shown a great interest in synthesizing derivatives of staurosporine, some of these compounds e.g. bisindolyl maleimides and N-benzoyl staurosporine (PKC412) have progressed to human clinical trials (Cohen, 2002b).

1.7.2 Flavopiridol

Flavopiridol is the first and most successful flavonoid (Figure 1-7) anti cancer agent. It inhibits multiple CDKs such as CDK1, CDK2, and CDK4 (Carlson *et al.*, 1996; Senderowicz, 1999; Christian *et al.*, 2009). Flavopiridol is not a selective inhibitor of CDKs and equally inhibits CDK2 and CDK4 (Ikuta *et al.*, 2001). It was originally identified as a tyrosine kinases inhibitor (Losiewicz *et al.*, 1994). Flavopiridol (Alvocidib) is under clinical development for the treatment different metastatic cancers (Lin *et al.*, 2007; Lin *et al.*, 2008; Carvajal *et al.*, 2009).

1.7.3 Fascaplysin as a lead compound

Fascaplysin, (Figure 1-8) a red pigment isolated from the sponge *Fascaplysinopsis Bergquist* sp. is a pentacyclic quaternary salt with antimicrobial activity (Roll *et al.*, 1988). Fascaplysin inhibits specifically CDK4 and CDK6 activity with a tenfold higher specificity for CDK4 compared to CDK6 and one thousand fold higher specificity for CDK4 compared to CDK2 (Soni *et al.*, 2000; Mahale *et al.*, 2006a). The IC_{50} for CDK4 is 0.41 μM while the IC_{50} for CDK2 has been reported as >250 μM (Mahale *et al.*, 2006a). Fascaplysin cannot be used as an anti-cancer drug and to control CDK4 activity because of its highly toxic nature. It is suggested that due to its planar structure fascaplysin intercalates with DNA (Hormann *et al.*, 2001).

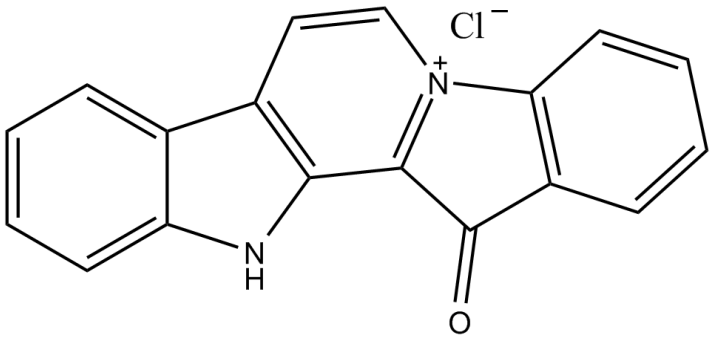


Figure 1-8 Molecular structure of fascaplysin. Fascaplysin is a red pigment isolated from the sponge *Fascaplysinopsis Bergquist sp*. is a pentacyclic quaternary salt with antimicrobial activity

A promising approach is to design and synthesise new CDK4 inhibitors based on the lead structure of fascaplysin with no intercalation with the DNA. In order to minimise the undesirable toxicity of fascaplysin and to improve its affinity for CDK4 several derivatives and non planar analogues of fascaplysin have been reported (Aubry *et al.*, 2004; Aubry *et al.*, 2006; Mahale *et al.*, 2006a; Mahale *et al.*, 2006b; Aubry *et al.*, 2009; Kuzmich *et al.*, 2010). The structural information about the interaction of

these compounds with CDK4 is very limited as inhibitor bound crystal structures of CDK4 are not yet available.

1.7.4 Other compounds

There are many other CDKs inhibitors synthesized by different pharmaceutical companies including SNS-032 (BMS-387032), AT-7519, P276–00 and PD-0332991 in clinical trials (Malumbres *et al.*, 2008). Seliciclib (CYC202) a trial drug by Cyclacel inhibits CDK2, CDK7 and CDK9 (Lacrima *et al.*, 2005). The marine natural product variolin B inhibits CDK2/Cyclin A, CDK2/Cyclin E, CDK1/Cyclin B and to a lesser extent CDK7/Cyclin H (Simone *et al.*, 2005). Butyrolactone I isolated from Aspergillus species F-25799 selectively inhibits CDK2 and CDC2 kinase (Kitagawa *et al.*, 1993).

Rational design of CDK inhibitors have been greatly aided by different bioinformatics techniques (Wu *et al.*, 2003; Hamdouchi *et al.*, 2005; Aubry *et al.*, 2006; Corsino *et al.*, 2009). An introduction to different bioinformatics techniques used in present study is described in the Section 1.8 to 1.12.

1.8 Homology Modelling

The gap between the number of experimentally determined structures and the known sequences for the proteins is on its increase as the total number of sequences in the public sequence databases is increasing each year almost double to the number of structures being solved experimentally (Jacobson M., 2004). Homology or comparative modelling techniques are useful for generating structural models of proteins whose experimental structures have not yet been solved or are very difficult to solve. Homology modelling predicts the three-dimensional structure of a protein based on the sequence alignment between the target sequence and a template selected from a closely related protein for which an experimental structure is available (Bajorath *et al.*, 1993;

Marti-Renom *et al.*, 2000; Baker and Sali, 2001). Structural similarity is pronounced than sequence similarity which implies that homology modelling for a target protein is possible using 3D structure of closely related protein as a template (Chothia, 1986; Marti-Renom *et al.*, 2000). Different web servers and computer programs such as SWISS-MODEL (Arnold *et al.*, 2006), SCWRL (Canutescu *et al.*, 2003), ICM Homology (Abagyan *et al.*, 1994) and MODELLER (Sali and Blundell, 1993; Sanchez and Sali, 2000; Eswar *et al.*, 2007) are available for homology modelling studies. Homology modelling comprises four main steps (Figure 1-9). The first step is finding a suitable template. Template identification is based on the percentage of sequence identity between target protein and the template. In order to build a reliable homology model target-template sequence identity should be higher than 30% (Marti-Renom *et al.*, 2000; Yang and Honig, 2000; Schwede *et al.*, 2003). The high sequence identity between the target and the template ensure the quality of a homology model. Sequence identity more than 30% is fairly good predictor of model accuracy (Marti-Renom *et al.*, 2000). The modelling procedure bases the construction of a new protein model on the information from the sequence alignment and coordinates of already known template, therefore higher the resolution of the template crystal structure the better may be the quality of the model. Second step is generating a target template alignment. To build a reliable model an optimal alignment between the sequence of the target protein and the sequence of the known structure is essential. This alignment is used as an input to calculate a model. After getting the alignment the third step is building the models for the target protein and finally the assessment and validation of the models (Marti-Renom *et al.*, 2000; Jacobson M., 2004; Narayanan Eswar, 2006).

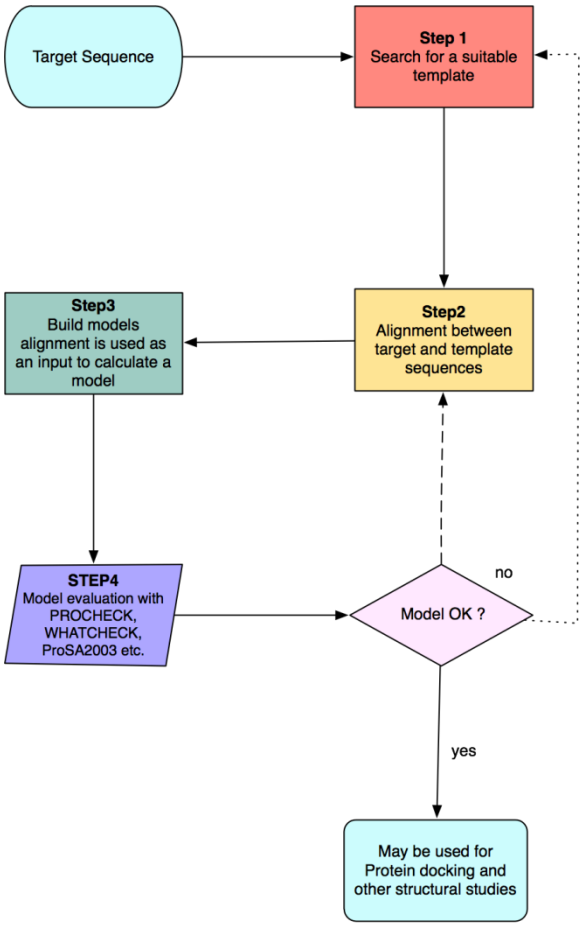


Figure 1-9 Schematic diagram of a homology modelling experiment. Homology modelling can be done in four steps that start with the identification of a suitable template, followed by the alignment between target and template sequences, model building and finally the evaluation of the model.

1.9 Fundamentals of protein ligand binding

The binding and interactions of proteins with small ligands in a specific manner is a very important feature of biological systems. An understanding of the interactions of proteins and ligands provids a basis to explain protein ligand specificity at a molecular level. The binding of a ligand with a protein in aqueous solution is an exchange process where ligand and protein bind with each other after losing their interaction with water. This protein ligand binding brings a favourable free energy. The energetics of protein

ligand binding play a key role in understanding the structural basis of protein ligand interactions and specificity of proteins for certain ligands.

Experimental data of ligand binding properties are usually described by the dissociation constant (K_D) of a ligand or inhibition constant (K_i) in case ligand is an inhibitor. To establish the relationship of these variables with the free energy of binding of ligand ($\Delta G°_{Bind}$) the following relation is used:

$$\Delta G°_{Bind} = - RT \ln(K_D) \quad \textbf{1-1}$$

Where R = 8.314 $JK^{-1}mol^{-1}$ (the ideal gas constant) and T is the absolute temperature

The extent by which an inhibitor inhibits a target protein is often measured by the IC_{50}. This represents the concentration of an inhibitor that is required for 50% inhibition of its target. The relationship between IC_{50} and K_i is defined by the Cheng-Prusoff equation (Yung-Chi and Prusoff, 1973).

$$K_i = \frac{IC_{50}}{1+\frac{[S]}{K_m}} \quad \textbf{1-2}$$

Where [S] is the substrate concentration, K_m is the concentration of substrate at which enzyme activity is at half maximal

1.10 Molecular Docking

In priniciple molecular docking attempts to predict the structure and binding affinity of a complex formed between protein and ligand (Lybrand, 1995; Halperin *et al.*, 2002; Brooijmans and Kuntz, 2003). Docked structures are used to study and compare structural features of protein inhibitor complexes (Shoichet *et al.*, 2002).

Protein ligand docking programs play a very important role in virtual screening, in which large libraries of many thousands of small molecules compounds are docked into a target-binding site. Virtual screening can provide novel leads for drug targets. In addition to the virtual screening molecular docking is also used in structure based drug design and facilitate in medicinal chemistry projects to prioritize the synthesis of new compounds. The accurate prediction of ligand orientations in the binding pocket of receptors is very important in structure based drug design and optimization of existing lead (Kuntz, 1992).

Docking methods can be divided into two parts that are a ligand docking algorithm to calculate the possible poses for a ligand in a receptor binding pocket and a scoring function to evaluate the resulting poses against each other and to estimate the binding free energy of the complex so formed (Kavraki, 2007). The docking problem is hard to solve because of a large number of degrees of freedom, relative orientation of two molecules and scoring problem. There is a no general scoring function available and the scoring problem is one of the major challenges in the development of docking programs (Gohlke and Klebe, 2002).

A docking experiment can be performed either as rigid docking or as flexible docking. Rigid body docking involves the docking of both protein and ligand as rigid bodies without any flexibility in their structure, but this is rarely successful. Most of the docking programs take protein molecules as rigid bodies although proteins change their conformation in response to the ligand binding (Davis *et al.,* 2003). Flexible protein-ligand docking has been evolved to a level where full flexibility on the ligand is commonly employed allowing exploration of torsional degrees of freedom in the docking process (Brooijmans and Kuntz, 2003). To further extend this some docking

programs also allow limited flexibility of some protein residues during a docking experiment (Verdonk *et al.*, 2003).

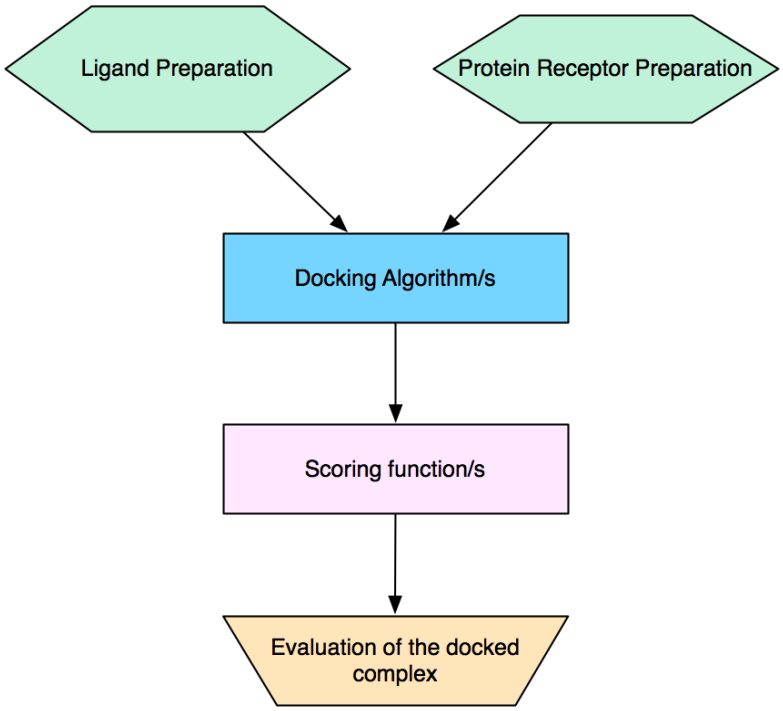


Figure 1-10 A schematic diagram of a general set up of protein ligand docking. A docking experiment require ligand and receptor prepration and use of appropriate docking algorithm and scoring functions to generate docked complex.

A docking experiment requires the structure preparation of the ligand and the receptor molecule (Figure 1-10). This includes the selection of a target structure, addition of hydrogen atoms, analysis of histidine, aspartate, glutamate and cysteine protonation states, and assignment of partial charges and the atom types to the ligand molecule.

The results of the docking experiment are ranked according to the docking score. The success of a docking algorithm in predicting a ligand-binding complex is often measured in terms of the root-mean-square deviation (RMSD) between the experimentally observed heavy-atom positions of the ligand and the one(s) predicted by the algorithm

Although there has been great advancement in the docking protocols over the last few years and different docking tools are available, there are limitations to this approach, for example how to introduce protein flexibility. Also the principles that guide ligand binding in proteins are not thoroughly understood (Sousa *et al.*, 2006).

1.10.1 Docking Algorithms

As stated above the first part of a docking method is to find the possible poses of a ligand in a protein active site. Different docking algorithms based on chemistry and geometry of atoms are being used (Gohlke and Klebe, 2002). The docking search algorithms use three general approaches; The first approach breaks down a ligand into fragments and incrementally rebuilds it in the environment of the binding pocket after placing a base fragment (Rarey *et al.*, 1996), the second approach relaxes the entire ligand in the binding site (Jones *et al.*, 1997; Morris *et al.*, 1998), the third that rigidly positions an ensemble of pregenerated ligand conformers into the binding site (McGann *et al.*, 2003).

DOCK (Lang *et al.*, 2009), AutoDock (Morris *et al.*, 2008) FlexX (Rarey *et al.*, 1996), eHits (Zsoldos *et al.*, 2007), Glide (Repasky *et al.*, 2007) and GOLD (Jones *et al.*, 1997; Verdonk *et al.*, 2003) are some examples of protein ligand docking programs commonly used. In the current work GOLD (Genetic optimization for ligand docking) has been used for molecular docking. GOLD utilizes a genetic algorithm (GA) to explore not only the full flexibility of the ligand, but also partial flexibility of the protein (Jones *et al.*, 1997). Genetic algorithms are search algorithms based on an evolutionary strategy. The genetic algorithm generates a set of possible solutions to the docking problem. Each possible solution is known as a chromosome and the set of solutions is termed as population. The genetic algorithm applies genetic operators such as crossover and mutation to evolve the solutions in order to find the best solutions (Jones *et al.*, 1997).

1.10.2 Scoring functions

The scoring functions for a docking method are used to evaluate the ranking of different possible poses of ligands against each other. There is a large and ever increasing number of scoring functions available (Taylor *et al.*, 2002; Kellenberger *et al.*, 2004; Sousa *et al.*, 2006; Warren *et al.*, 2006). Scoring functions can be divided into three main categories namely force field scoring functions, empirical scoring functions and knowledge based scoring functions.

1.10.2.1 Force field scoring functions

Force field scoring functions use molecular mechanics force fields such as AMBER (Weiner and Kollman, 1981; Cornell *et al.*, 1995), OPLS (Jorgensen *et al.*, 1996) or CHARMM (Brooks *et al.*, 1983) to quantify the interaction energy between receptor and the ligand and the internal energy of the ligand (Sousa *et al.*, 2006). Different force field scoring functions are based on different parameters, but these are functionally similar. The binding free energy of protein ligand complexes is accounted through a sum of van der Waals and electrostatic interactions. The Lennard Jones potential is often used for the van der Waals energy term. D-Score (Kramer *et al.*, 1999), G-Score (Kramer *et al.*, 1999), GoldScore (Verdonk *et al.*, 2003), and the AutoDock 3.0 (Morris *et al.*, 1998) scoring function are example of the force field scoring functions.

The GoldScore is provided in the GOLD (Genetic Optimization for Ligand Docking) docking suite. It takes into account factors such as H-bonding energy, van der Waals energy and ligand torsion strain for the prediction of ligand binding positions.

$$\text{GoldScore} = S_{\text{hb\_ext}} + S_{\text{vdw\_ext}} + S_{\text{hb\_int}} + S_{\text{vdw\_int}} \quad \textbf{1-3}$$

The GoldScore is represent by equation 1-3 where S_{hb_ext} is the protein-ligand hydrogen bond score, S_{vdw_ext} is the protein-ligand van der Waals score, S_{hb_int} is the score from intramolecular hydrogen bonds in the ligand and S_{vdw_int} is the score from intramolecular strain in the ligand.

1.10.2.2 Empirical scoring functions

The types of interactions that may be included in empirical scoring functions are hydrogen bonds, electrostatic interactions, hydrophobic contacts and solvent exclusion volumes. Such scores often express the binding free energy by a weighted sum of various types of interactions between the two binding partners (Sousa *et al.*, 2006). Training sets of experimentally determined complexes are used to determine the coefficients for the various terms (Kitchen *et al.*, 2004; Sousa *et al.*, 2006). Examples of empirical scoring functions include ChemScore (Eldridge *et al.*, 1997), LigScore (Krammer *et al.*, 2005), LUDI (Bohm, 1994), F-Score (Rarey *et al.*, 1996) and X-Score (Wang *et al.*, 2002).

The ChemScore is a fitness function that estimates the total free energy change that occurs on ligand binding and is trained by regression against binding affinity data. ChemScore is used in PRO_LEADS (Baxter *et al.*, 1998) and GOLD (Verdonk *et al.*, 2003). The ChemScore function estimates the free energy of binding of a ligand to a protein as shown in equation 1-4 and 1-5.

$$\Delta G_{binding} = \Delta G_0 + \Delta G_{hbond} S_{hbond} + \Delta G_{metal} S_{metal} + \Delta G_{lipo} S_{lipo} + \Delta G_{rot} H_{rot} \quad \textbf{1-4}$$

$$\text{ChemScore} = \Delta G_{binding} + E_{clash} + E_{int} + E_{cov} \quad \textbf{1-5}$$

These equations contain S_{hbond}, S_{metal}, and S_{lipo} that are scores for hydrogen bonding, acceptor-metal, and lipophilic interactions, respectively. H_{rot} is a score representing the loss of conformational entropy of the ligand upon binding to the protein.

1.10.2.3 Knowledge based scoring functions

Knowledge based scoring functions are derived from a statistical analysis of experimentally solved ligand protein complex in PDB. These scoring functions are trained to reproduce experimentally determined structure or binding data (Gohlke *et al.*, 2000; Muegge, 2001). PMF (Muegges's Potential of Mean) (Muegge *et al.*, 1999) and DrugScore (Gohlke *et al.*, 2000) are the examples of knowledge based scoring functions.

1.10.2.4 Consensus scoring functions

The use of a combination of any of the above mentioned scoring function is termed as consensus scoring (Charifson *et al.*, 1999). There is a conceptual problem establishing a correlation and in scaling of different scoring functions; despite this drawback consensus scoring methods have shown some success (Charifson *et al.*, 1999; Sousa *et al.*, 2006). X-CSCORE (Wang *et al.*, 2002) is an example of consensus scoring which combines OMP, ChemScore and FLexX scoring functions.

1.11 Molecular dynamics

Molecular dynamics (MD) is a form of computer simulation where atoms and molecules are allowed to interact for a period of time under known laws of physics. MD simulations are frequently used to study proteins and other chemical or biomolecular systems. MD simulations can provide atomic level understanding of different chemical and biological systems by providing details of individual particles motion as a function

of time. In addition to the study of motional properties of biological systems MD simulations can also be used to calculate thermodynamics properties of a system.

In order to create an aqueous environment MD simulations normally include solvent either explicitly or implicitly (Simonson *et al.*, 2004). Molecular dynamics simulations using explicit solvent are computationally expensive, therefore sometimes implicit water models (Tsui and Case, 2000) which include solvent effects within the force field equations are used. Implicit models represent the solvent as a uniform dielectric medium, which neglects the geometrical, bonding patterns (such as water's hydrogen bonds) details of the solvent.

Different computer programs such as CHARMM (Brooks *et al.*, 1983; Brooks *et al.*, 2009), AMBER (Case *et al.*, 2005) and GROMACS (Hess *et al.*, 2008) with a great range of capabilities are available to study proteins and other biomolecular systems. MD simulations of proteins may complement the experimental data by refining the protein structure obtained by the X-Ray or NMR. Protein molecular dynamics study can also provide information about the role of solvent in protein dynamics, any conformation changes or folding of a protein over a period of time.

1.12 Thermodynamics integration (TI) and free energy calculations

Study of energetics and thermodynamic properties of a biomolecular system provide invaluable information for its better understanding. The most important thermodynamic quantities are standard Gibbs binding free energy ($\Delta G°_{bind}$) (see equation 1-1) describing the tendencies of a biomolecular system to associate. The predication of binding free energy ($\Delta G°$) of protein ligand interactions play a very important role in rational drug discovery by identifying novel molecules that can bind more effectively to the target protein (Lipinski *et al.*, 1997; Hopkins and Groom, 2002). With the advancement in the computational techniques different molecular dynamics

methods including thermodynamic integration methods have been developed to study the free energies of molecular systems (Quirke and Jacucci, 1982; Ravishanker *et al.*, 1986; van Gunsteren and Berendsen, 1987; Kollman, 1993; Dominy, 2008).

Thermodynamic integration calculates the difference in free energy of a system between a reference state (state A) and the state of interest (state B). The free energy difference between the two states is calculated by coupling them via a parameter λ that serves as an additional, nonspatial coordinate (van Gunsteren and Berendsen, 1987; Gouda *et al.*, 2003). The free energy is determined as the work necessary to change the system from A to B over a reversible path. If the path is reversible, work done is taken equal to change in the free energy.

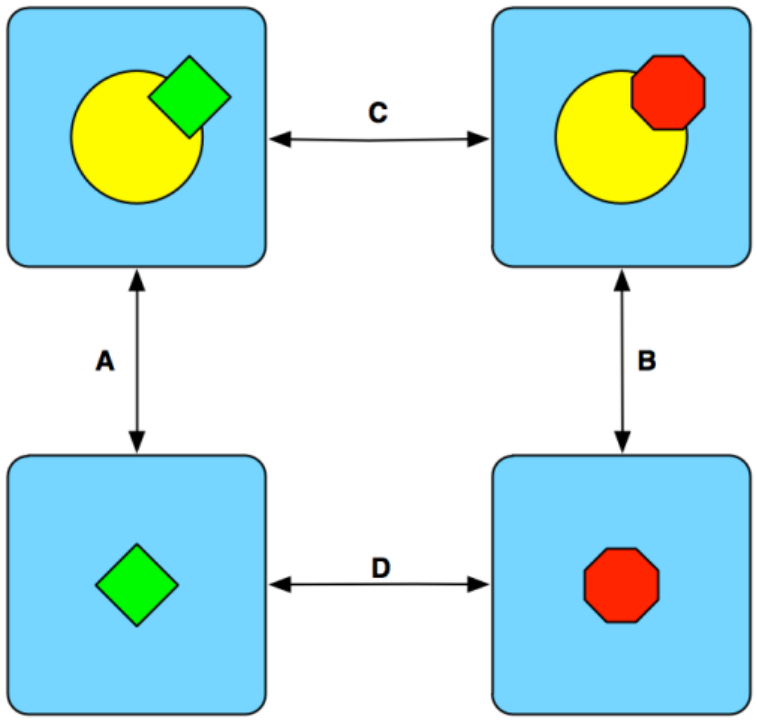


Figure 1-11 An overview of thermodynamics integration experiment. In this picture the processes A and B represent the binding of two different ligands to a protein, while processes in the process C and D representation the transformations from one ligand to the other with and without bound to a protein (This picture is adopted from http://ambermd.org)

The Helmholtz free energy (F^0) of a system is a thermodynamic potential which measures the "useful" work obtainable from a closed thermodynamic system at a constant temperature and volume and it can be calculated by the equation 1-6. In this equation k_B denotes Boltzmann's constant and Z represent the configuration integral expressed by the equation 1-7.

$$F^0 = -k_B T \ln Z \quad \textbf{1-6}$$

$$Z = \sum_i (-E_i / k_b T) \quad \textbf{1-7}$$

The Helmholz free energy of a system can be assumed as equal to the Standard Gibbs free energy (equation 1-8).

$$G^0 \approx F^0 = -k_B T \ln Z \quad \textbf{1-8}$$

The free energy difference between the two states A and B (equation 1-9) can be calculated using Zwanzig formula (Zwanzig, 1954) and equation 1-8.

$$\Delta G^0 \left(A \rightarrow B\right) = G_B - G_A = -k_B T \ln \langle e^{-[V_B - V_A]/k_B T} \rangle_A \quad \textbf{1-9}$$

$$\Delta G^0 = G_B - G_A = -\beta^{-1} \ln \langle e^{(-\beta \Delta V)} \rangle_A = -\beta^{-1} \ln \langle e^{(-\beta \Delta V)} \rangle_B \quad \textbf{1-10}$$

Where $\beta = 1/k_B T$ and $\Delta V = V_B - V_A$

In the equation 1-9 V_A and V_B denote the potential functions of state A and B and $\langle\ \rangle_A$ denotes a MD generated ensemble average of $\Delta V = V_B - V_A$ that is sampled using V_A potential or in other words it indicate the exponential term should be

evaluated over a Boltzmann-weighted ensemble average generated according to potential function of the state A. In equation 1-10 it is assumed that the configurational sampling (or partition function) is carried out under constant temperature and pressure conditions. The $\Delta G^0\left(A \rightarrow B\right)$ in equation 1-9 can also be denoted with ΔG^{FEP} where FEP stands for free energy of perturbation. A calculation according to equation 1-9 can only be useful and converge if a reasonable probability of configurations sampled on the potential V_A also exists for V_B as shown in equation 1-10. This means a considerable overlap of thermally accessible regions of the state A and state B. In order to solve this equation a multistep approach is adopted by breaking down the transition from one state into the other in multiple smaller reversible steps and introducing a set of intermediate potential energy functions. In thermodynamic integration the two states are coupled with via a parameter λ_m that serves as an additional, non-physical coordinate as shown in equation 1-11 where λ_m varies from 0 to 1.

$$Vm = (1 - \boldsymbol{\lambda}_m)V_A + \boldsymbol{\lambda}_m V_B \qquad \textbf{1-11}$$

Now equation 1-10 will take the form as follow

$$\Delta G^0 = G_B - G_A = -\beta^{-1} \sum_{m=1}^{n-1} \ln\left\langle e^{[-\beta(V_{m+1} - V_m)]}\right\rangle_m \qquad \textbf{1-12}$$

If we define $\Delta\boldsymbol{\lambda}_m = \boldsymbol{\lambda}_{m+1} - \boldsymbol{\lambda}_m$ and combine this with equation 1-11 and 1-12 we get

$$\Delta G^0 = -\beta^{-1} \sum_{m=1}^{n-1} \ln\left\langle e^{[-\beta\Delta V \Delta\boldsymbol{\lambda}_m]}\right\rangle_m \qquad \textbf{1-13}$$

with $\Delta V = V_B - V_A$.

From the definition of V_m (equation 1-11) the relations $\Delta V = \partial V_m / \partial \boldsymbol{\lambda}_m$ and $V_{m+1} - V_m = \frac{\partial V_m}{\partial \boldsymbol{\lambda}_m} \Delta \boldsymbol{\lambda}_m$ can be derived with a condition that $\Delta\lambda$ steps are infinitely small. Based on this the equation 1-12 takes the form

$$\Delta G^0 = -\beta^{-1} \sum_{m=1}^{n-1} \ln \left\langle e^{\left[-\beta \frac{\partial V_m}{\partial \boldsymbol{\lambda}_m} \Delta \boldsymbol{\lambda}_m\right]} \right\rangle_m \qquad \textbf{1-14}$$

By solving the exponent and the logarithm the above equation can be simplified to as follow

$$\Delta G^0 = \sum_{m=1}^{n-1} \left\langle \frac{\partial V_m}{\partial \boldsymbol{\lambda}_m} \right\rangle \Delta \boldsymbol{\lambda}_m \qquad \textbf{1-15}$$

When λ is approaching zero ($d\lambda \rightarrow 0$) the above equation can be written in an integral form as below.

$$\Delta G_{TI}^0 = \int_0^1 \left\langle \frac{\partial V(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \right\rangle_{\boldsymbol{\lambda}} d\boldsymbol{\lambda} \qquad \textbf{1-16}$$

The equation 1-16 is referred as thermodynamic integration (TI) formula for free energy calculations and it indicates that the free energy difference between two states

can be calculated by integrating the Boltzmann-weighted λ derivative of the mixed potential function over λ.

1.13 Molecular phylogenetics

Phylogenetic analysis provides information about the evolutionary relationship between different individuals, species, families, organisms, genes and biomolecules such as proteins and discloses the descendents from a common ancestor (Nei, 1996).

Molecular phylogenetics relies on the information obtained by multiple sequence alignments of DNA/RNA or proteins (Nei, 1996; Whelan *et al.*, 2001; Blair and Murphy, 2010). Prior to molecular phylogenetics only classical methods based on phenotypic and physiological assessment of different organisms were used to determine the evolutionary relationship between them. In classical phylogeny the evolutionary history of organisms is inferred using methods of comparative morphology and comparative physiology, however this approach does not always produce a clear picture of the evolutionary history because of the complex nature of morphological and physiological characters (Nei, 1996; Nei and Kumar, 2000).

Recent advances in molecular biology techniques and availability of DNA and protein sequence data have led to the origin of molecular phylogenetics (Brocchieri, 2001; Blair and Murphy, 2010). The basic principle of molecular phylogeny is that sequence similarity reveals common ancestry. Comparison of DNA and protein data provides useful and accurate information about the evolutionary relationship between different organisms that is otherwise difficult to obtain (Nei and Kumar, 2000; Brocchieri, 2001). Molecular phylogeny has made it easy to study the evolution of genes and proteins in different organisms. The results of phylogenetic analyses are usually expressed with phylogenetic trees. A phylogenetic tree is a graphical representation of the evolutionary relationship among different taxonomic groups or

other biological entities that are known to have a common ancestor (Trooskens *et al.*, 2005). In a phylogenetic tree, each node with descendants represents the most recent common ancestor of the descendants (Figure 1-12). In some phylogenetic trees the length of the edges is an estimate of evolutionary time. The external nodes at the ends of the branches represent extant taxa and are often termed as operational taxonomic units (OTUs). The internal nodes also known as hypothetical taxonomic units (HTUs) represent hypothetical progenitors of the OTUs.

Phylogenetic trees can be rooted or unrooted. In an unrooted tree relatedness of individual taxa relative to each other is illustrated without indicating the direction of the evolutionary path. In case of a rooted tree one of the nodes is used as a root representing the common ancestor of all the other external nodes (OTUs) from which a unique evolutionary path leads to other nodes. The most distant sequence within a tree (outgroup) is used for rooting a tree (Colless, 1985; Graham *et al.*, 2002).

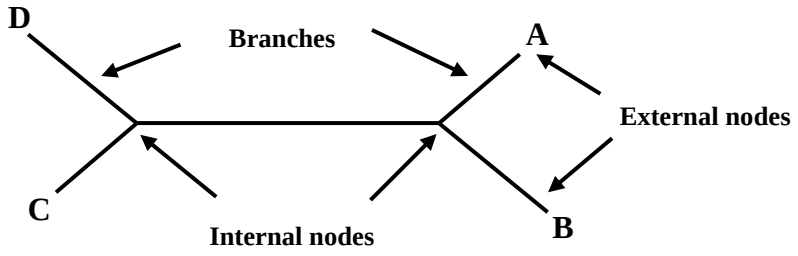


Figure 1-12 An unrooted phylogenetic tree showing 4 taxonomic units. An unrooted tree display the relatedness of different taxonomic units and does not provide an eveloutionary path.

There are different algorithms, procedures and computer programs available for molecular phylogenetic analysis from the sequence data (Felsenstein, 1996; Whelan *et al.*, 2001; Whelan, 2007; Kumar *et al.*, 2008). Phylogenetic tree of a group of sequences obtained by these methods do not necessarily represent the "true" phylogenetic tree

(Penny *et al.*, 1990; Edwards, 2009). In the present work the Bayesian inference (Huelsenbeck and Ronquist, 2005) and Neighbor-joining methods (Saitou and Nei, 1987) are used for the phylogenetic analysis of CDKs and related proteins (see Section 2.5).

1.14 Aims and objectives of the present study

CDKs have been identified as important targets for therapeutic intervention to control the cell cycle and in the anti-cancer field (De Bondt *et al.*, 1993; Huwe, 2003; Rizzolio *et al.*, 2010). Different CDKs are involved in different phases of the cell cycle and have different roles. In order to control these CDKs individually there is a need to find specific inhibitors as these may exhibit a better side effect profile by only inhibiting the target enzyme. There is a significant need to develop novel inhibitors of CDK4 with a better selectively and IC_{50} in order to control the activity of CDK4 for a systemic therapy of cancer.

Fascaplysin (Figure 1-8) isolated from the sponge *Fascaplysinopsis Bergquist sp.* specifically inhibits CDK4 compared to CDK2. Fascaplysin inspired inhibitors with a high specificity for CDK4 over CDK2 have been synthesized (Aubry *et al.*, 2006; García *et al.*, 2006; Mahale *et al.*, 2006a; Mahale *et al.*, 2006b; Jenkins *et al.*, 2008; Aubry *et al.*, 2009). However, the ligand and protein interactions contributing to the CDK4 selectively are poorly understood in the absence of any ligand bound crystal structure for CDK4.

The present study is designed to understand similarities and variations between active sites of different CDKs, to explore phylogenetic relationships of CDKs family and to gain an insight into CDK4 structural information using homology modelling. CDK4 homology modelling based on CDK2/CDK6 was one of the initial objectives of this study, however after the availability of the X-ray structures of CDK4 this objective

was expanded to incorporate the structural information from the X-ray structures into homology modelling.

The present work is focused on learning CDK4 interactions with fascaplysin and its derivatives using molecular docking approaches, an analysis of docked models to generate a hypothesis how to improve binding selectivity and IC_{50} of the CDK4 inhibitors and make suggestions to prioritise the synthesis of new inhibitors with selectivity and improved affinity for CDK4. This study is aimed at addressing the question of fascaplysin specificity for CDK4 compared with CDK2 using bioinformatics techniques such as molecular docking and thermodynamic integration.

Chapter Two

MATERIALS AND METHODS

Chapter 2 Materials and Methods

2.1 Sequence retrieval and sequence databases

The sequences of all CDKs, CDKL proteins and cyclins (for sequence identifiers see Chapter 3) were retrieved from UniProtKB (Bairoch *et al.*, 2007; Apweiler *et al.*, 2010), UniRef100 (Suzek *et al.*, 2007) and NCBI's non-redundant protein sequence (nr) databases. UniProt is a central repository of protein data created by joining the information contained in Swiss-Prot, TrEMBL, and PIR. Swiss-Prot is a curated protein sequence database that provides a high level of annotation, a minimal level of redundancy and a high level of integration with other databases (Bairoch and Apweiler, 2000). TrEMBL stands for translated EMBL database and is a computer-annotated protein sequence database complementing the Swiss-Prot. The UniRef100 database combines identical sequences and sub-fragments from the same source organism (species) as a single entry displaying the sequence of a representative protein and all of its accession numbers linked to UniProt entries (Suzek *et al.*, 2007). The "nr" database by NCBI (National Centre for Biotechnology Information) contains non-identical sequences from GenBank CDS translations, PDB, Swiss-Prot, PIR, and PRF. The NCBI dbEST database (Boguski *et al.*, 1993) was used to retrieve all expressed sequence tags (ESTs) for *Gallus gallus* (see Section 3.3.2). ESTs are single pass reads of cDNA sequences (usually 200 to 500 nucleotides long) that are generated by sequencing either one or both ends of an expressed gene.

In addition to the above-mentioned databases FASTA-formatted PDB sequence databases obtained from the Dunbrack Lab and NCBI were also used. The Dunbrack Lab offers an up-to-date PDB sequence database as part of their PISCES PDB culling server (Wang and Dunbrack, 2005). This database contains a complete set of sequences representing each entry in the PDB. The NCBI also maintains a PDB sequence database

but this database does not represent all the PDB structures as it uses redundancy filters and is therefore only a subset of Dunbrack Lab PDB sequence database.

2.2 BLAST similarity searches

The Basic Local Alignment Search Tool (BLAST) is an algorithm to find regions of local similarity between sequences of amino acids and nucleotides (Altschul *et al.*, 1990). BLAST compares a sequence query with those contained in nucleotide and protein databases and identifies the sequences that resemble the query sequence by finding regions of local alignment between the query and a database sequence. BLAST is a collection of different programs e.g. blastp for protein-protein similarity search and blastn which searches a nucleotide database using a nucleotide query.

BLAST sequence similarity searches were carried out using protein-protein BLAST (blastp) both at NCBI and against locally installed databases. A local installation of BLAST is required for ease of access, automation with Perl scripts and in order to perform the BLAST searches against customized databases. The BLAST search algorithm requires sequence databases to be presented in a special data format as FASTA formatted database files are not compatible for the BLAST. UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, UniRef100, *Gallus gallus* ESTs and PDB sequence (from NCBI and Dunbrack Lab) data were formatted as BLAST searchable databases and installed locally using the formatdb command. The BLAST formatdb parameter *-o* for indexes was set to true (*T)* in order to parse SeqID and create indices. formatdb is a program for formatting BLAST databases from FASTA format input. An example for protein database formatting is shown below. The parameter *–p* takes two arguments either *F* for formatting nucleotide sequences or *T* for formatting protein sequences.

```
formatdb –i database –p T –o T
```

The sequences of the CDKs, CDKL proteins and cyclins, were subjected to BLAST sequence similarity searches using protein-protein BLAST v2.2.17 (blastp) both at NCBI and against locally installed databases. An example command line for blastp is given below.

```
blastall -p blastp -d database -i query -o output
```

The sequences with required cut-offs (as discussed in following results chapters) were retrieved using fastacmd or via download from BLAST web server. The "fastacmd" is a part of the BLAST software package from the NCBI and it retrieves FASTA formatted sequences from a BLAST database. The command used for fastacmd is shown below.

```
fastacmd –d database –i list
```

In this command –i parameter specify an input file with GIs or accessions numbers for a batch retrieval of corresponding FASTA sequences.

2.3 Multiple sequence alignments

Comparison of proteins sequences allow us gaining insight not only into proteins' sequence similarity and variation, but also in protein evolution and function. Multiple sequence alignments (MSA) of proteins provide information about the patterns of amino acid conservation, which can be used to infer phylogenetic relationships. Multiple sequence alignments for all CDKs, CDKL proteins, cyclins, CDK2/CDK3 and CDK4/CDK6 sequences (obtained from BLAST results) were generated using a local installation of MUSCLE v3.6 (Edgar, 2004). MUSCLE stands for multiple sequence comparison by log-expectation; it is a multiple sequence alignment tool for protein and

nucleotide sequences. Multiple sequence alignments were obtained using MUSCLE with default parameters. MUSCLE is a relatively new MSA tool, that is faster and performs better (Edgar, 2004) than older tools such as Clustal W. An example command line is given below where input.fasta and output.fasta are input and output files in FASTA file format.

```
muscle -in input.fasta -out output.fasta
```

Profile versus profile alignments are used to align two existing multiple sequence alignments while keeping the features of individual alignments. Two multiple sequence alignments, one for CDK2 and the other for CDK4/CDK6 were obtained by using MUSCLE v3.6. These alignments were used for profile versus profile alignment using MUSCLE v3.6 and Clustal W v2.0 (Larkin *et al.*, 2007).

All the alignments obtained with MUSCLE and profile vs profile alignment were further optimized by manual adjustment in Jalview (Clamp *et al.*, 2004; Waterhouse *et al.*, 2009). The Jalview program is used for visualisation and manual editing of multiple sequence alignments. The manual adjustment and optimization of multiple sequence alignments involves removal of any redundancy and incomplete sequence, adjustment of conserved motifs for any misalignments, and inspection and adjustment of gaps, insertions and deletions. The multiple sequence alignments obtained by MUSCLE were also used for phylogenetic analysis as described in Section 2.5.

2.4 Hidden Markov Model search

In sequence analysis a Hidden Markov Model (HMM) is a statistical model which describes subtle patterns to define families of homologous sequences. HMMs offer a powerful detection of distinct relatives of a protein family. Hidden Markov Model also give more sensitive results than BLAST by selecting amino acids at successive

positions from a position specific probability distribution of the input multiple sequence profiles while BLAST's pair wise sequence comparison method assumes that all amino acid positions are equally important (Eddy, 1996; Eddy, 1998a).

HMMER is an implementation of a profile HMM method for sensitive database searches using multiple sequence alignments as queries (Eddy, 1998b). Multiple sequence alignment of CDK2/CDK4 and CDK4/CDK6 were used to build a profile HMM using the hmmbuild program. The profile HMMs were calibrated using the hmmcalibrate program. Searches against the UniRef100 database was carried out using hmmsearch program to distinguish between CDK2/CDK3 and CDK4/CDK6 and in order to find any distinct relatives of CDK2/CDK3 and CDK4/CDK6 that may be have been missed with BLAST search. hmmbuild, hmmcalibrate and hmmsearch are part of the HMMER v2.3.2 software package.

2.5 Phylogenetic analysis

Phylogenetic analysis provides information about the evolutionary relationship of different organisms, species and biological molecules i.e. gene and proteins. Molecular phylogenetic studies can be performed using DNA or protein sequences. Amino acid sequences are useful for evolution studies and provide information on long-term phylogenetic relationship of genes or species (Nei and Kumar, 2000). There are different methods and algorithms available to carry on molecular phylogenetic analysis. In the present study the Bayesian inference and Neighbor-joining methods were used to carry out phylogenetic analysis of CDKs, CDK like proteins and cyclins.

2.5.1 Phylogenetic analysis using Bayesian inference

Bayesian inference is statistical inference, which is based on the notion of posterior probabilities of a phylogenetic tree. MrBayes is a program for Bayesian inference of phylogeny (Huelsenbeck and Ronquist, 2005). Phylogenetic trees for all

human CDKs, CDK2/CDK3 and CDK4/CDK6 and cyclins were generated with MrBayes v3.1.2. MrBays requires multiple sequence alignments in a stringent Nexus format to be used as an input. To get data into MrBayes multiple sequences alignments generated as described in Section 2.3 were converted from FASTA into Nexus format using Clustal W v2.0 followed by manual editing. In order to allow so-called model jumping between fixed-rate amino acid models, the "prior" for the amino acid model was set to "mixed" by using the "prset aamodelpr = mixed" command as first step of the phylogenetic analysis. This setting for "aamodelpr" parameter permits jumping between amino acid substitution models. The models included in "mixed" are Jones (Jones *et al.*, 1994), Dayhoff (Dayhoff *et al.*, 1978), Mtrev (Adachi, 1996), WAG (Whelan and Goldman, 2001), Mtmam (Yang *et al.*, 1998), Rtrev (Dimmic *et al.*, 2002), Cprev (Adachi *et al.*, 2000),Vt (Muller and Vingron, 2000) and Blosum (Henikoff and Henikoff, 1992). MrBayes was initially run for 10000 generations using "mcmc ngen=10000 samplefreq=10" to ensure at least thousand samples from the posterior probability distribution. Analysis was continued until the standard deviation of split frequencies was below 0.01 (default). The trees were summarized using "sumt burnin" command using default parameters. Finally, phylogenetic trees were obtained in Nexus tree format and read by the tree drawing program TreeView for visualisation (Page, 1996).

2.5.2 Phylogenetic analysis using neighbor joining method

It is a good practice to use more than one method for a phylogenetic analysis in order to find a consensus for phylogenetic trees. Molecular phylogenetic analysis of CDKs and cyclins was also carried out with neighbor-joining method (Saitou and Nei, 1987) using MEGA v4.0 (Tamura *et al.*, 2007; Kumar *et al.*, 2008). MEGA stand for Molecular Evolutionary Genetics Analysis.

The multiple sequence alignments for all human CDKs, CDK2/CDK3 and CDK4/CDK6 and cyclins generated as described in Section 2.3 were used as an input for the MEGA alignment explorer and were converted into a MEGA data format (.meg file). Phylogenetic trees were constructed using the neighbour-joining (NJ) method with 1000 bootstrap replications. Evolutionary distances were computed using the Poisson correction method. All positions containing gaps and missing data were eliminated from the dataset using the default complete deletion option. Phylogenetic trees obtained from MEGA were visualised by the built in TreeExplorer.

2.5.3 Mapping the tree of life

The Tree of Life (ToL) is a graphical representation of the relationship between all forms of life on Earth. The root of the Tree of Life represents a universal common ancestor to all life on Earth. The closer two organisms are on the tree; the more closely they are related to each other in an evolutionary context.

The tree of life corresponding to the selected species of super-phylum deuterostome and order actiniaria (see Chapter 3) was built after obtaining the taxonomic Ids of CDKs and cyclin sequences from UniProt Taxonomy Database (Bairoch *et al.*, 2007; Apweiler *et al.*, 2010) and mapping these IDS with the tree of life (Maddison and Schulz, 2007)

2.6 Ligand-Protein Contacts analysis

Ligand protein contacts of all available CDK2 structures bound with ligands were derived with the LPC software (Sobolev *et al.*, 1999). LPC performs a detailed analysis of interatomic contacts such as hydrogen bonds, hydrophobic-hydrophobic contacts, hydrophobic-hydrophilic contacts, contact surface area and distances between atoms of the ligand and protein residues.

Care was taken to fix any formatting and misnumbering issues in the PDB files. Some of the PDB files e.g. 1JVP required manual fixing. There are two tautomeric forms of the ligand in 1JVP which are marked as 1 and 2 in the 17th position of the each coordinate line for the ligand. LPC software cannot take into account for such PDB files. Only one form of the ligand was kept and the position 17th in this PDB file was made empty as in standard PDB files.

The LPC output for each CDK2 structure was parsed with a Perl script to obtain a list of amino acids with a nearest distance cut-off <=4 Å and for contact surface area >= 25 $Å^2$ cut-off between protein and ligand molecule. These lists of amino acids for all CDK2 structures were combined and the contact frequency for each amino acid was calculated to define the active site residues of CDK2.

2.7 Analysis of active site volume and shape

Ligand-receptor interactions are known to take place inside the active site clefts or cavities for most of the proteins; therefore characterization of the active site in term of its volume is very important to determine the possible accommodation and interactions of a ligand within a active site cleft. The active site volume of inactive and activated forms CDK2 and CDK4 was measured with fpocket program (Le Guilloux *et al.*, 2009). The shape of the active site clefts of CDK2 and CDK4 were analysed with Caver program (Petřek *et al.*, 2006).

2.8 Clustering of protein structures

The BLAST search for CDK2 (P24941) against the Dunbrack PDB database (see Section 2.2) revealed the presence of more than 190 structures of CDK2 (Appendix 1.1) in the PDB. To take advantage of this plethora of structural information about CDK2 all the CDK2 PDB files were downloaded and stored in a single directory using a PERL

script (*pdbDownload.pl*) written for this purpose (Appendix 2.1). MaxCluster was used to align and perform the clustering of the CDK2 structures. MaxCluster is a tool for protein structure comparison and clustering (Herbert and Sternberg, 2008) with a capability to process thousands of structures either against a single reference protein or in an all-versus-all fashion. MaxCluster works only on single chains and does not allow multiple occupancy of residues. Perl scripts (Appendix 2) were written in order to pre-process and modify the PDB files to be used as input files by removing extra chains and multiple occupancy of the residues from the PDB files.

2.9 Homology modelling

Homology modelling techniques are important to generate structural models of proteins for which experimental structures have not yet been solved. MODELLER 9v1 (Narayanan Eswar, 2006) was used for homology modelling. Three dimensional models with MODELLER are obtained by optimally satisfying spatial restraints. These restraints are derived from the sequence alignment between the target template sequence and template structure, and expressed as probability density functions (pdfs) for the features restrained (Sali and Blundell, 1993).

The modelling procedure begins with the preparation of the required input files which are the alignment of the sequence of the protein to be modelled with the sequence(s) of related template structure(s), the PDB file containing the 3D structures of the template(s) and a customized MODELLER script written in Python.

The alignment files for MODELLER were generated via the alignment strategy described in Section 4.3. The resulting alignments were manually converted into MODELLER compatible PIR format. Alignments and selected templates were used as an input for MODELLER to generate a model. MODELLER uses a special form of the PIR format where information about sequence numbering and chain codes is written

into the 'description' line between the PIR protein tag (>P1) and the actual alignment. The description line of each entry contains information necessary to extract atomic coordinates of the segment from the original PDB coordinate set.

Criteria for template selection were the protein crystallized in the active form, complexed with an inhibitor, low R_{free} value and high resolution for the X-ray structure. Based on these criteria PDB structure 2CCH (Cheng *et al.*, 2006) was chosen as a template for CDK4. PDB structure 2CCH is crystallized with ATP at a X-ray resolution of 1.70 Å and R_{free} 0.182. The python script for MODELLER contains all the parameters required and references to the input files (see Appendix 3). The number of models to be generated was set to 50. A sample script is given below.

```
from modeller import *
from modeller.automodel import *

env = environ()
a = automodel(env, alnfile='cdk4-2cch.ali',
              knowns='2cch', sequence='cdk4',
              assess_methods=(assess.DOPE))
a.starting_model = 1
a.ending_model = 50
a.make()
```

Finally the homology models were generated using the following command to run the python script with MODELLER 9v1.

```
mod9v1 model_cdk4_based_on_cdk2.py
```

CDK2 and CDK6 both were also used to construct CDK4 models based on two templates. After CDK4 crystal structures (inactive form) becoming available in April

2009 a CDK4 “active form model” was prepared based on active form of CDK2 and CDK4 crystal structure.

2.10 Validation of homology modelling

Homology models were validated for the correctness of the overall fold/structure, stereochemical parameters such as bond lengths, angles and dihedrals, and also to check any errors over localized regions. Modeller built in checks, ProSa2003 (Wiederstein and Sippl, 2007), PROCHECK (Laskowski *et al.*, 1993) and WHAT_CHECK (Hooft *et al.*, 1996) programs were used for the validation of CDK4 homology models build by MODELLER. In addition to these validations control models of CDK6 were also generated as discussed below. The final selection of the best model of CDK4 was based on the protein quality assessments by PROCHECK and WHAT_CHECK and also MODELLER and ProSa2003 score.

2.10.1 MODELLER built in checks

MODELLER built in checks were used to find any errors in the protein models generated by MODELLER. First of all the log file generated by the MODELLER was inspected for any error or restraint violations. In addition to that the models were evaluated by MODELLER with the DOPE (**D**iscrete **O**ptimized **P**rotein **E**nergy) potential and comparison was made with the templates DOPE potential. Three models were selected with best MODELLER and DOPE potential for their further evaluation.

2.10.2 Protein models evaluation with ProSa2003

ProSa2003 assists in the evaluation of protein structures and models (Sippl, 1993; Wiederstein and Sippl, 2007). It makes use of knowledge-based potentials (z-score) to analyse the quality of protein structures. The z-score indicates overall model

quality. The z-score as implemented in ProSa2003 for a protein structure or a model is a statistical measure which quantifies the distance (expressed in units of standard deviation) a data point is from the mean of a data set of all NMR and X-ray structures. The z-score value of an input structure or homology model is displayed in a plot that contains the z-scores of all experimentally determined NMR and X-ray structures. If the z-score of the input structure is outside the z-scores plot for experimentally determined structures the structure under investigation probably contains errors. The ProSa2003 z-score for CDK4 models was calculated by the ProSa2003 and it was compared with the z-score obtained for templates.

2.10.3 Protein model validation with PROCHECK and WHAT_CHECK

PROCHECK checks the stereochemical quality of a protein structure or model and generates a number of plots including a Ramachandran plot (Ramachandran *et al.*, 1990). A Ramachandran plot displays the possible conformations of ψ and φ angles for a polypeptide or a protein molecule. The ψ and φ angles cluster into distinct regions in the Ramachandran plot where each region corresponds to a particular secondary structure. Ramachandran plots were obtained for the protein models and were further analysed for protein quality and compared with the template. The homology models were also subject to WHAT_CHECK (Hooft *et al.*, 1996) analysis to identify problematic regions.

2.10.4 Control models

Control models were also used to validate the homology modelling strategy. CDK6 was modelled based on a CDK2 template using the same approach as for CDK4, but not using any information from the structure of CDK6. The resulting model for CDK6 was then compared with available structures for the CDK6 in the PDB. An

overlay of the CDK6 structure and model was carried out in PyMOL and RMSD between these was calculated.

2.11 Molecular visualization and structural alignment

PyMOL (DeLano, 2002), Chimera (Pettersen, 2004) and Silver (Verdonk *et al.*, 2003) were used for the visualization of protein structures from protein database files (PDB). Chimera and Silver facilitate the preparation of receptor protein and ligand molecules to be used in docking experiments. In addition GoldMine (Verdonk *et al.*, 2003) was used for the visualization of docked structures.

PyMOL offers a feature for the structural alignment of proteins and the preparation of publication quality figures. The active site analysis of CDK2, CDK6 and other CDKs was performed in PyMOL (as shown in Figure 3-2, 3.4 and 3.6). The PDB files containing the structural information for the corresponding active sites were created with Perl script (Appendix 2.2) by removing all residues other than the active site residues. Active site residues were selected using the Ligand-Protein Contacts (LPC) software (Sobolev *et al.*, 1999) as discussed above in section 2.4.

2.12 Molecular docking

Molecular docking attempts to predict the structure of a complex formed between protein and ligand. A docking experiment starts with the structure preparation of the ligand and receptor molecules followed by ligand docking and analysis of docked complexes as described below.

2.12.1 Ligand preparation

The input structures of all ligands used in this work were prepared in HyperChem 8.0 (Hypercube Inc, 2008). Partial charges and atom types were assigned to the ligand molecules. Automated assignments of atom type and partial charges by HyperChem were checked manually and corrected when necessary as in case of aromatic and non

aromatic nitrogen. These ligand molecules were energy minimized using the steepest descent algorithm until conversion with a termination condition of RMS gradient of 0.1 kcal/ (Å mol). Finally the ligand coordinates were saved in MDL mol and Tripos mol2 format to be used in docking experiments.

2.12.2 Preparation of receptor

Receptor molecules were prepared for docking by the addition of hydrogen atoms, analysing the states of relevant histidines, glutamines and asparagines, removal of existing inhibitor present in the PDB in case of CDK2/CDK6. In some docking experiments His95 and Lys35 of CDK4 were treated as flexible using Dunbrak rotamers (Dunbrack and Karplus, 1993) in GOLD. The redundant chains in the PDB structures were also removed. *In-silico* variants for CDK2 and CDK4 were prepared in Chimera (Pettersen, 2004) by substitution of the active site residue Phe 82 from CDK2 with His 85 from CDK4 and vice versa to mimic CDK2 active site with CDK4 and CDK4 active site with CDK2. These *in-silico* variants were prepared to understand the possible role of CDK4 His95 in its selectivity toward fascaplysin.

2.12.3 Ligand docking

GOLD was used for ligand docking in this work (Verdonk *et al.*, 2003). The GOLD docking suite was selected based on the criteria such as its availability, trained for kinases and other related proteins, and availability of ChemScore (see Section 1.10.2) which is taken as an estimate of binding energy. The PDB coordinates of CDK2, CDK4 and CDK6 after receptor preparations were loaded into the GOLD v4.0 interface Hermes for docking. Acceptable protein input file formats for GOLD are PDB and MOL2. All hydrogen atoms, including those necessary to define the correct ionisation and tautomeric states of Asp, Glu and His were added to the receptors

proteins and water molecules were removed. The active site was defined with a radius of 10 Å around Leu83 for CDK2, Val96 for CDK4 and Val101 for CDK6 by a manual selection in the visualiser. Ligand molecules to be docked with receptor proteins were loaded into GOLD. The genetic algorithm search rate was set to slow with a search efficiency set to very flexible. The default parameters were used for population size, selection pressure, niche size, mutation frequency and the number of operations. The default option for early termination of docking experiment was unselected in order to get most accurate solution. The GoldScore and ChemScore with kinase template params file were used.

2.12.4 Analysis of docked complexes

GoldMine and Hermes were used with GOLD (Verdonk *et al.*, 2003) for initial visual analysis of docking results. Detailed analysis including the visualization of ligand protein hydrogen bonds and individual atom distance calculations of docking solutions were carried out in PyMOL. All the images of docking solutions were also prepared using PyMOL. RMSD analysis between the docked solutions from GOLD output was performed using a PERL script.

2.13 Molecular dynamics studies

The Amber 10 (Assisted Model Building with Energy Restraints) package (Case *et al.*, 2008) was used to perform molecular dynamics experiments. Amber is a collection of different programs, which collectively form a molecular dynamics simulation suite. The Amber suite is divided into two packages, which are Amber tools that consisting of *antechamber, leap* (preparatory programs), *& ptraj* (analysis program) and Amber 10 containing *sander* (**S**imulated **A**nnealing with **N**MR **D**erived **E**nergy **R**estraints) and *pmemd* (**P**article **M**esh **E**wald **M**olecular **D**ynamics) which are

the main simulation programs. *amber* is a different term from Amber and represents a empirical force field which is being used by many simulation packages.

2.13.1 Coordinates preparation

The atomic coordinates of proteins are required as an initial input in a molecular dynamics study and can be obtained either from the PDB (NMR or X-ray data) or from homology modelling in case where a experimentally solved structure is not available. These coordinates are used to create the input files required for *sander*. The CDK2 structure (1FIN) was selected from the PDB and its coordinate were prepared for tLeap by removing the extra chains and HETATOM entries while keeping the one chain for CDK2. A hybrid model of CDK4 (as described in Chapter 4) was used for CDK4 molecular dynamics experiments.

2.13.2 Force field selection and parameterization

A protein force field contains parameters for all of the bonds, angles, dihedrals and atom types in a system. In the present study ff99SB and General Amber Force Field (GAFF) (Wang *et al.,* 2004) provided with Amber 10 were used. The force filed parameters for CDK2 and CDK4 were obtained from ff99SB. The GAFF was used in the parameterization of carbo-fascaplysin and fascaplysin.

The Amber force fields do not provide parameters for small ligand molecules such as carbo-fascaplysin and fascaplysin. So carbo-fascaplysin and fascaplysin molecules needed to be parameterized in order to use them for molecular dynamics simulation with Amber.

The antechamber tool was used to create parameter files for carbo-fascaplysin and fascaplysin using an established strategy (Wang *et al.,* 2006). The molecular structure of carbo-fascaplysin and fascaplysin was prepared and energy minimized in

HyperChem (Hypercube Inc, 2008). The resultant structures were saved in mol2 format. The mol2 file generated was converted into Gaussian input format using Open Babel (Guha *et al.*, 2006). Gaussian 03 is needed for *ab initio* calculation of partial charges (Frisch *et al.*, 2004). The Gaussian 03 runs were carried out with the help of Dr. Ralf Schmid using following charge methods and parameters in the Gaussian route line.

```
# HF/6-31G* SCF=tight Test Pop=MK iop(6/33=2) iop(6/42=6) opt
```

Hartree-Fock Hamiltonian (HF) is used with a level/basis set 6-31G*. The SCF=tight is used for most precise convergence. The keyword "Test" used in the command suppresses the automatic creation of an archive entry. The Pop=MK setting produce charges fit to the electrostatic potential at points selected according to the Merz-Singh-Kollman scheme. MerzKollman and ESP are synonyms for MK. Gaussian write out the grid points for calculation of the electrostatic potential using iop(6/33=2). The iop(6/42=6) is used to controls the density of the grid point. The keyword "opt" is used to performs a geometry optimization according to the Berny algorithm.

Gaussian calculates the atomic orbitals and electrostatic potential, as well as an optimization of molecule geometry before the potential computation. The Gaussian output was used with antechamber to generate the frcmod and lib files for carbo-fascaplysin and fascaplysin as given below.

```
antechamber -i fasc.gout -fi gout -o fasc.mol2 -fo mol2 -c bcc -s 2 -nc 1
parmchk -i fasca.mol2 -f mol2 -o fasca.frcmod
```

Here the –i fasc.gout specifies the Guassian output file for fascaplysin. The -o fasc.mol2 specifies the name of output file in mol2 format (-fo mol2). The -c bcc option

is used for the AM1-BCC charge model in order to calculate the atomic point charges. The -s 2 option defines the verbosity of antechamber status information. The –nc flag was only used in case of fascaplysin to address the charge on fascaplysin.

2.13.3 Topology and parameters files generation

In order to set up a molecular dynamics simulation with sander topology and parameters files are required. The prmtop file contains information about the molecular topology and the necessary force field parameters, the inpcrd file provides a description of the atom coordinates and also the current periodic box dimensions. CDK2, CDK4 and their complex with fascaplysin and carbofascaplysin were first neutralize with counterions then solvated with a 12 Å solvent box using TIP3P and TIP4P-Ew water models, and their prmtop and inpcrd files were generated using tLeap.

2.13.4 Relaxing the system prior to molecular dynamics

In many cases protein 3-dimensional coordinates obtained from the PDB or from homology modelling may not correspond to the actual minima in the energy, therefore a short energy minimization is required to relax the system before molecular dynamics. In absence of a minimization the proteins bad contacts may lead to instabilities during the molecular dynamics run. The energy minimization step works towards the closest local minimum.

In all CDK2 and CDK4 experiments minimization was carried out with sander by selecting 1000 steps of minimization with a default value of nonbonded cutoff (8.0 Å). The steepest descent algorithm was selected for the first 500 steps before switching to the conjugate gradient algorithm for the remaining steps. The sander program supports different minimization algorithms of which commonly used are steepest descent and conjugate gradient algorithms. The steepest descent algorithm is good for

quickly removing the largest strains in the system but converges slowly compared to the conjugate gradient method which is more efficient for convergence.

2.13.5 Equilibration of the solvated complex

To equilibrate following short energy minimization of CDK2 and CDK4, 50ps of heating was carried out from the initial temperature of 0K to 300K at constant volume in 25000 MD steps. After the heating step 50ps of density equilibration with weak restraints was carried out in 25000 steps at 300K at constant pressure with isotropic position scaling and 1ps pressure relaxation time. Restraints were used for all protein residues with a restraint weight of 2.0 kcal/mol–$Å^2$ during heating and density equilibration. Finally 500ps of constant pressure equilibration simulations at 300K were done in 250000 steps at 300K with a pressure relaxation time 2ps. All simulations were run with "shake" on hydrogen atoms and Langevin dynamics for temperature control. During these equilibration runs bonds involving the hydrogen were kept constrained and bond interactions involving hydrogen atom were omitted by using the ntc = ntf = 2 settings variables. The default value of nonbonded cut-off (8.0 Å) was kept during equilibration simulations. The molecular dynamics output (i.e. mdout) and coordinates (i.e. mdcrd) were written after every 500 steps for 50ps heating and density equilibration simulations and after every 1000 steps for 500ps constant pressure equilibration.

2.13.6 Production run

Production runs were carried out with Sander in 250000 steps (this was run 10 times to obtain 5ns duration) at constant pressure with isotropic position scaling and 2ps pressure relaxation time. The time step was adjusted to 0.002ps for these MD simulations. The output coordinates were obtained after every 10ps by setting

ntwx=5000. Same conditions were set as the final phase of equilibration i.e. default value of nonbonded cut-off (8.0 Å) was kept and temperature was kept at 300K with langevin dynamics for temperature control. Bonds involving the hydrogen were kept constrained and bond interactions involving hydrogen atom were omitted. Sample input files are given in the Appendix 3.

2.13.7 Analysis of trajectories

The analysis of equilibrium and production runs for temperature, density, total energy was carried out with the *process_mdout.pl* script. The backbone RMSD analysis to study the conformational stability during the equilibration and production run was carried out with ptraj program provided with Amber suite.

2.14 Thermodynamic Integration

The advancement in molecular dynamics simulations and computational methods for biochemical systems had made it possible to estimate the free energy changes of a system. Thermodynamic integration (TI) estimates the free energy changes of a system between two states A and B by coupling them via lambda (λ) that serves as an additional, non spatial coordinate (see Section 1.11.1). TI experiments were carried out for transformation of carbofascaplysin to fascaplysin (Figure 6-11) in CDK2, CDK4 and water.

CDK2 structure 1FIN and the hybrid model of CDK4 were used for the thermodynamic integration. For each thermodynamic experiment two runs of tLeap were carried out to produce four sets of parameter and restart files containing both ligands in the protein bound and solvated states. In the first run ff03 and gaff force fields were loaded into tLeap program. The TIP4P-Ew water model was selected for thermodynamic integration experiments therefore frcmod.tip4pew was loaded into the

leap program after setting WAT=T4E. The carbofascaplysin bound with CDK2 (or CDK4) was loaded to tLeap and was solvated with 12 Å solvent box. The charge on the system was set to zero by adding the Cl- counter ions. The first leap run produced the two PDB files of the solvated and neutralized carbofascaplysin complex and of the carbofascaplysin ligand in water. The solvated PDB files for the fascaplysin complex and ligand were produced by manually replacing carbon atom with nitrogen atom. In the second run of tLeap these four pdb files were used to generate the *.prm and *.rst files yielding four parameter and four rst files. Similarly *.prm and *.rst files were also obtained for ligands only. These parameters and rst files were used for the transformation of carbofascaplysin into fascaplysin in the thermodynamic integration run. The thermodynamic integration run was carried out using the parameters shown in the table 2-1.

Table 2-1 : An overview of the thermodynamic integration parameters

Input file	Changing CRB into FAS (19 λ windows 0.05 to 0.95, at 5 ns each)
Process 0 (V0)	prmtop: cdk2_fas crgmask=':FAS@N2' scmask=':FAS@N2' ifsc=1
Process 1 (V1)	prmtop: cdk2_crb crgmask=':CRB@C0,N2' scmask=':CRB@C0,N2' ifsc=1

* scmask specifies the unique atoms for this process. This, along with crgmask, is the only parameter that will frequently be different in the two mdin files for V0 and V1. The ifsc flag is used for soft core potentials

All systems were minimized with 500 steps of steepest descent minimization. 50 ps of density equilibration were carried out after minimization. Equilibrated rst and prm files generated after a preliminary thermodynamics run were sent to Rutgers, The State University of New Jersey, USA. The final thermodynamic integration experiments with 19 λ points/windows (0.05 to 0.95, at 5 ns each window) for transition from carbofascaplysin to fascaplysin were carried out in specialized computing cluster in Rutgers. The 5ns thermodynamic integrations simulations were divided into 25 steps each with 200ps simulations. The dV/dλ data was collected for each step to calculate an

estimate of the relative binding free energy carbofascaplysin and fascaplysin with CDK2 and CDK4. The value of ΔG for each step was obtained by numerically integrating (see Section 1.12) dV/dλ values with a PERL script.

2.15 PERL programming and computational resources

Perl stands for **p**ractical **e**xtraction and **r**eport **l**anguage (Wall *et al.*, 2000; Tisdall, 2001) and it is a very powerful scripting tool. The following Perl scripts (See Appendix 2) were written for the data extraction, analysis and to facilitate the overall work.

- pdbDownload.pl: A tool to download PDB files from the PDB server. This script was used to download all the CDKs PDB files.
- activesiteExtract.pl: A tool to create PDB files containing only active sites for the corresponding original PDB files based on the data from the LPC server.
- pdbParser.pl: A tool to retrieve the summary information from the PDB files.
- maxPrep.pl: A tool to prepare PDB files for MaxCluster to work on these files.

A Linux computer cluster (x86_64) with 9 nodes (each containing four 2.0 GHz Intel® Xeon™ processors) was used for the molecular dynamics simulations described in Section 2.13.6. The average computational time for a 1ns production run on this cluster was approximately 53 hours. The LAM environment or universe (Burns *et al.*, 1994) was created on each node using lamboot command before running the MD simulation. The parallel version of the sander module i.e. sander.MPI was invoked using mpirun command and specifying the number of CPUs/processes to run on with the -np option.

mpirun -np <number of processes> sander.MPI <arguments>

The mpirun is a program which is required to execute MPI (message passing interface) programs in parallel on LAM nodes.

Chapter Three

ACTIVE SITE COMPARISON AND CDK EVOLUTION

Chapter 3 Active site comparison and CDK evolution

3.1 Introduction

This chapter combines two aspects of sequence and structural comparison of CDKs to study active site variations and CDK evolution. Active site analysis and phylogenetic studies of CDKs help to better understand their evolutionary relationship, structural and functional conservation and the role in cell cycle.

Active site analysis plays a significant role in understanding of the contribution specific residues in the active site to substrate recognition and overall function of a particular protein. A structural and primary sequence comparison of the active site of CDK2, CDK4 and CDK6 was carried out to understand potential differences in these proteins. This study was extended to a comparative analysis of active sites of all known human CDKs.

The phylogenetic study presented here starts with the construction of a phylogenetic tree for all known human CDKs. A better understanding of CDKs requires an updated knowledge of growing list of CDKs and related proteins. Therefore, the construction of phylogenetic trees was extended to CDK like proteins and cyclins. The phylogenetic relationship between CDK2/CDK3 and CDK4/CDK6 was also analyzed among different species of the super-phylum deuterostome and the order actiniaria. Deuterostomes are bilaterians animals (bilaterally symmetrical) that are characterized by their embryonic development as their first opening appearing in the embryo become anus. This super-phylum includes craniates, cephalo-chordates, tunicates, echinoderms and hemichordates (Blair and Hedges, 2005; Swalla and Smith, 2008). The actiniaria order represent sea anemones which are found in all marine habitats (Daly *et al.*, 2008).

3.2 Active site analysis

An analysis of active sites of all known human CDKs and CDK like proteins was carried out to better understand the structural properties, conservation and variations within the active sites of all human CDKs. One particular question to be addressed is the CDK4 specificity toward different ligands e.g. fascaplysin and its derivatives.

3.2.1 Defining the active site residues of CDK2

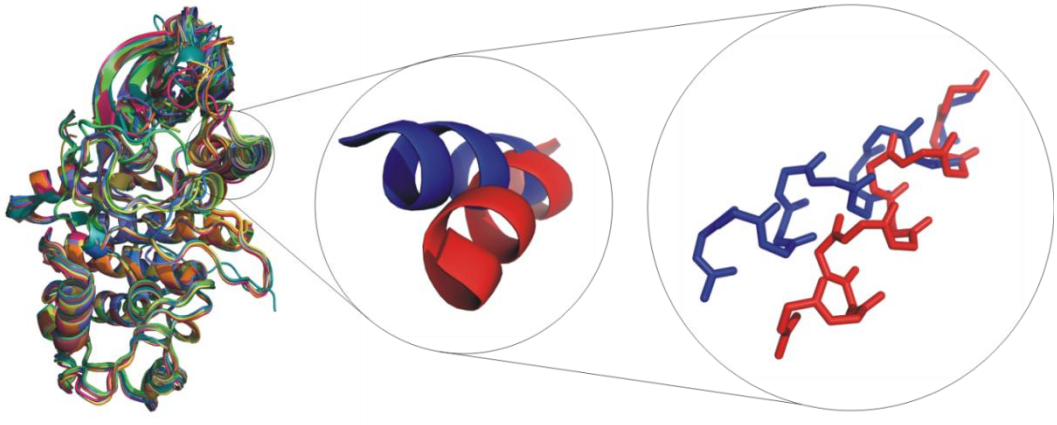
Structure based analysis of active site starts with CDK2 because this is the best characterized CDK in structural terms with more than 190 PDB structures available (see Appendix 1.1). The recognition of CDK2 active site residues interacting with ligands was based on "Ligand-Protein Contacts" (LPC) which were derived from 122 CDK2 PDB structures bound with different small molecule ligands and seven CDK2 structures bound with ATP (X-ray resolution of 2.5 Å or better) using a local installation of the LPC software (Sobolev *et al.*, 1999) and using a nearest distance cut-off of 4 Å and a 25 $Å^2$ cut-off for the contact surface area (as described in Chapter 2). The results of the LPC analysis for each individual structure out of the 129 structures were combined and parsed with a Perl script to obtain the CDK2 active site residues, which are shown in Table 3-1. Based on this analysis 21 residues (IGEGVAKVFEFLHQDKQNLAD) with a contact frequency cut-off 44 were selected as a CDK2 active site signature.

Table 3-1: Combined result of the LPC analysis. 129 PDB structures were selected with a resolution of 2.5 Å or better and bound with ATP or other small molecule ligands. LPC analysis was performed on each individual PDB structure to obtain a list of amino acids with a nearest distance cut-off <=4 Å and for contact surface area >= 25 $Å^2$ cut-off between protein and ligand molecule. This list of amino acids for each PDB was combined and the contact frequency for each amino acid was calculated. Leu83, Ala31, Leu134 were found to be in contact with the ligands in all 129 structures.

Residue No.	Amino Acid	Frequency	Residue No.	Amino Acid	Frequency
83	LEU	129	64	VAL	87
31	ALA	129	85	GLN	84
134	LEU	129	131	GLN	83
10	ILE	127	33	LYS	75
81	GLU	126	89	LYS	69
82	PHE	120	144	ALA	57
80	PHE	116	132	ASN	56
18	VAL	111	13	GLY	48
84	HIS	100	12	GLU	47
86	ASP	98	11	GLY	44
145	ASP	89			

3.2.2 Structure based clustering and active site analysis of CDK2

To distinguish the active and inactive form of CDK2 structures systematically MaxCluster (Herbert and Sternberg, 2008) was used for structure based clustering of CDK2 structures. 146 CDK2 structures (X-Ray resolution of 2.5 Å or better) out of total 190 PDB structures were analysed. Two clusters were found, the first cluster (108 PDB structures) represent the inactive form of CDK2 and the second cluster contains the active form of the CDK2 (38 PDB structures) with phospho-threonine at position 160 in 26 structures. An overlay of representative active and inactive clusters of CDK2 is shown in Figure 3-1



(a) **(b)** **(c)**

Figure 3-1 Overlay of active and inactive clusters of CDK2. (a) A cartoon representation of the overlay of active and inactive clusters of CDK2 **(b)** Cartoon highlighting the different orientations of the C-α helix in the active (blue) and inactive (red) form of CDK2 structure **(c)** A stick representation of carbon backbone of the C-α helix as shown in b)

A structural comparison of the active site residues between the active sites of two clusters of CDK2 reveals structural differences within these clusters. There are two distinct orientations of His84 as shown in the Figure 3-2. The surface accessible side chain of His84 is pointing outward from the binding cavity and it does not seem to have any direct polar interaction with any of the ligands. The two orientations of His84 are more prominent in the active form of CDK2 as shown in Figure 3-2 (Panel B) compared with the inactive form of CDK2 structures Figure 3-2 (Panel A) indicating a possible role of this residue in the binding properties of this protein. In the active form cluster of 38 structures 23 active structures have His84 oriented downward while in 15 structures it is in upward direction. In the inactive cluster all His84 are oriented downward except only two structures (1AQ1 and 1OIR), which have His84 oriented in opposite direction.

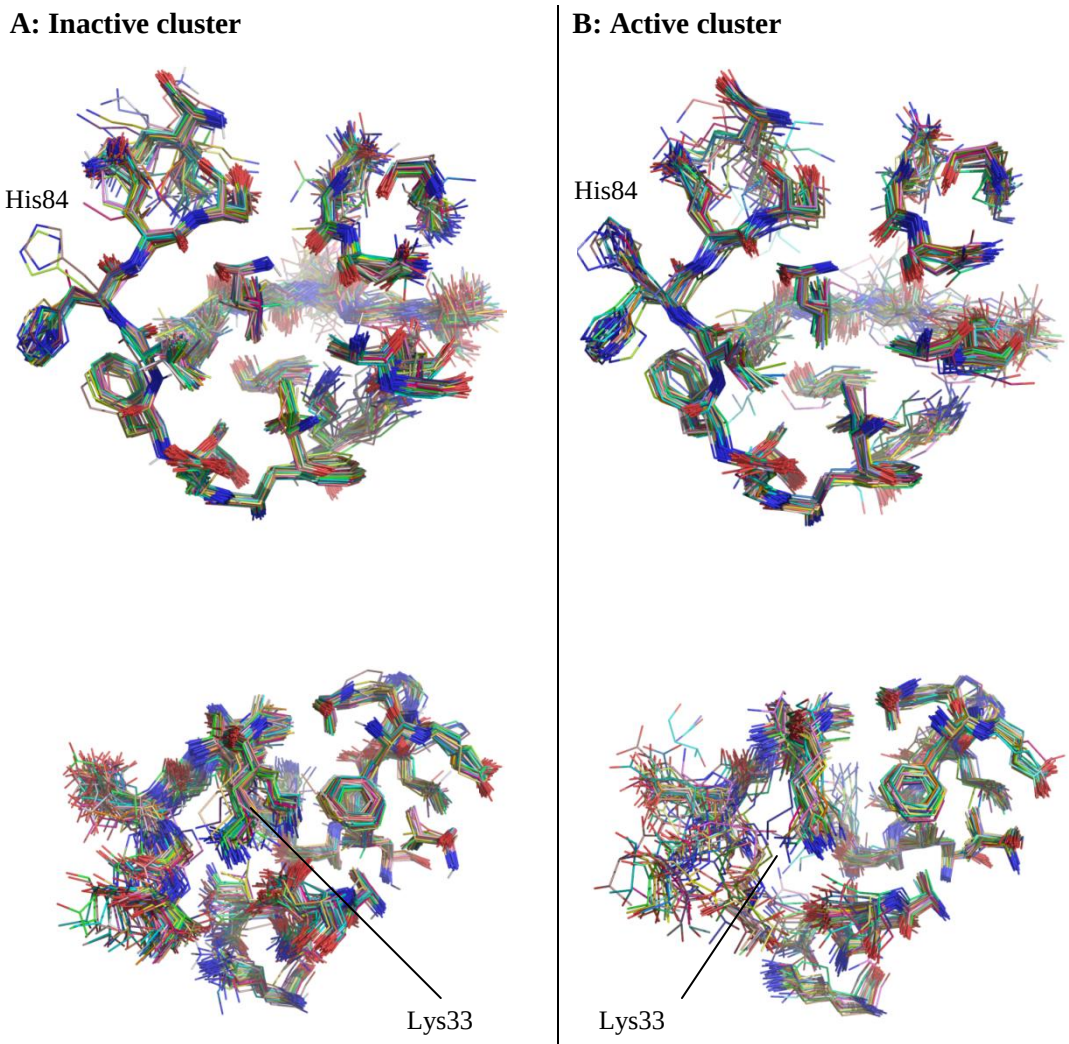


Figure 3-2 Structural variations in active site residues of CDK2. A line representation of the active site residues of the inactive form (panel A) and active form (panel B) structures of CDK2 showing the variation in active site residue orientations in space. The picture highlights the two distinct orientations of the His84, which is more prominent in right panel. There are also variations in the orientation of Lys33 in the inactive form of CDK2 (panel A bottom). However the active form has Lys33 in one orientation only (panel B bottom).

In addition to His84 (Figure 3-2 bottom panels) Lys33 shows two distinctive orientations. All the active structures have a similar orientation of the Lys33 side chain. The inactive structures have two different orientations of the Lys33. This Lys33 forms a salt bridge with Glu51 of the C-α helix in the active form of CDK2 compared with a salt bridge with Asp145 of T-loop in the inactive form of CDK2 (Figure 3-3). Side

chain of Lys33 also make interactions with ATP and different inhibitors of CDK2 such as indirubin-5-sulphonate in PDB 1E9H (De Bondt *et al.*, 1993; Davies *et al.*, 2001; Meijer *et al.*, 2003).

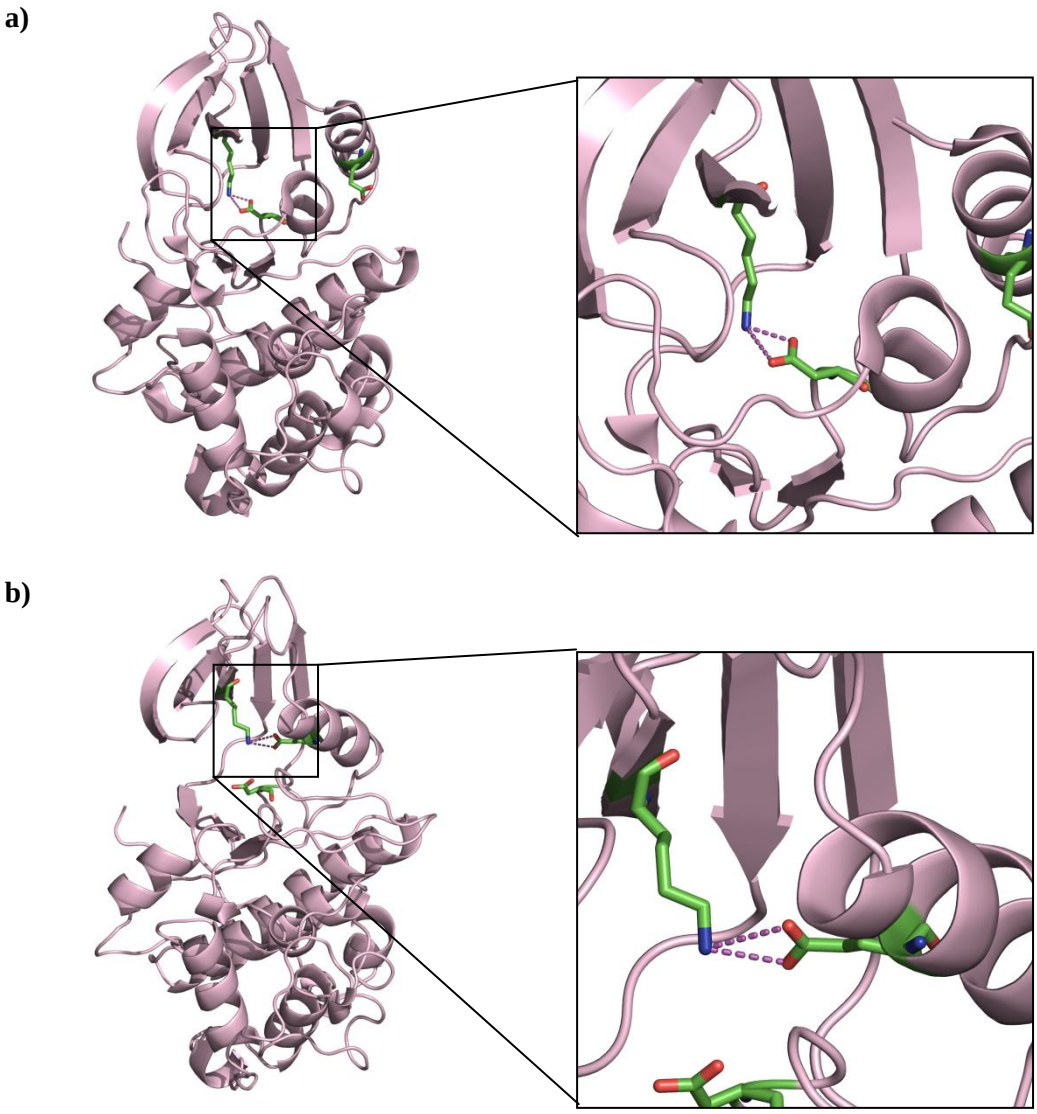


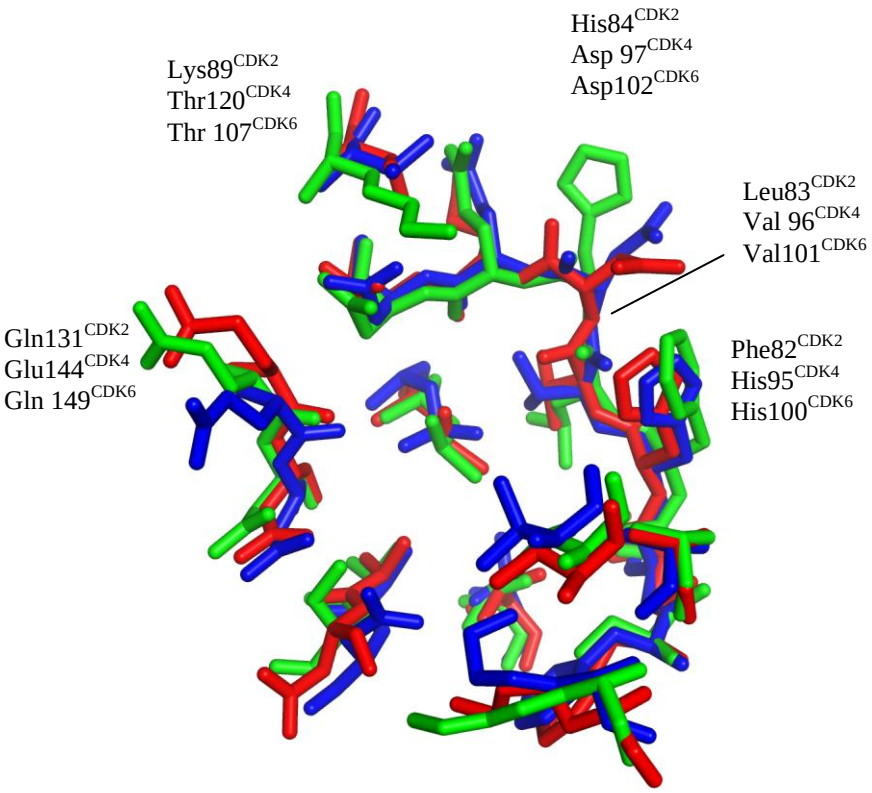
Figure 3-3: Salt Bridges in CDK2. a) In the inactive form of CDK2 (PDB ID 2R3I) the salt bridge is formed between the NZ of Lys33 and OD1/OD2 of Asp 145 (T-Loop) **b)** The salt bridge in the active form (PDB ID 1FIN) of CDK2 is formed between the NZ of Lys33 and OE1/OE2 of Glu 51 (Cα helix).

The CDK2 active site comparison provided identification about the variations of CDK2 residues and this information may be used to introduce protein residue flexibility in molecular docking experiments.

3.2.3 Active site comparison of CDK2, CDK4 and CDK6

A structural comparison of active site residues was originally performed between the structures of CDK2 and CDK6, but it was extended to include CDK4 after the availability of a CDK4 structure. Figure 3-4 highlights a distinct difference in the active sites of CDK2 and CDK6 as a phenylalanine ($Phe82^{CDK2}$) in CDK2 is replaced by a histidine residue ($His100^{CDK6}$) in CDK6. Similarly, CDK4 also has a histidine ($His95^{CDK4}$) in this position. These results have significance for ligand binding of CDK4 since $His95^{CDK4}$ is oriented toward the interior of the active site. This will affect binding of inhibitors and could explain differences in the binding specificity of different inhibitors of CDK2 and CDK4. In addition to $Phe82^{CDK2}$ some other residues which display differences are shown in Figure 3-4 (enclosed in boxes). CDK2 and CDK6 both have glutamine at position 131 and 149, respectively, compared to Glu144 in CDK4. This Glu144 may also have a significant role toward the specificity of CDK4 (see Section 5.3). This active site analysis provided a base to explore the potential role of His95 of CDK4 toward fascaplysin specificity using molecular docking approach as discussed in Chapter 5.

(a)



(b)

	Human CDK2																
12	18	31	33	64	80	81	**82**	**83**	**84**	85	86	**89**	**131**	132	134	144	145
E	V	A	K	V	F	E	**F**	**L**	**H**	Q	D	**K**	**Q**	N	L	A	D
	Human CDK4																
14	20	32	34	72	93	94	**95**	**96**	**97**	98	99	**102**	**144**	145	147	157	158
V	V	A	K	V	F	E	**H**	**V**	**D**	Q	D	**T**	**E**	N	L	A	D
	Human CDK6																
21	27	41	43	77	98	99	**100**	**101**	**102**	103	104	**107**	**149**	150	152	162	163
E	V	A	K	V	F	E	**H**	**V**	**D**	Q	D	**T**	**Q**	N	L	A	D

Figure 3-4 Overlay of CDK2, CDK4 and CDK6 active sites. (a) Stick representation of the active site of CDK2 overlayed with the active sites of CDK4 and CDK6. The CDK2 structure is shown in green (PDB ID 2CCH) , CDK4 in blue (PDB ID 2W96) and CDK6 in red (PDB ID 1G3N). The amino acid residues for the active site were obtained from LPC (Sobolev *et al.*, 1999) as described in Section 2.6. The first four residues of the CDK2, CDK4 and CDK6 active site (IGEG residue 10-13 in CDK2) are ommited for a better representation.
(b) A comparison of the active site residues of CDK2, CDK4 and CDK6, The overall active site is conserved with differences at six residues. The residues enclosed in the box indicate difference in the CDK2, CDK4 and CDK6 active sites. The NH and carbonyl group of Leu83^{CDK2} forms hydrogen bonds with most of the known inhibitors for CDK2. CDK4 and CDK6 have His95 and His100 corresponding to the Phe82 in CDK2. CDK4 differs from both CDK2 and CDK6 by having Glu144 corresponding to Gln131 in CDK2 and Gln149 in CDK6.

3.2.4 Active site analysis of all known human CDKs and CDK like proteins

The sequence based active site analysis of human CDKs (CDK1-CDK11) is shown in Figure 3-5. The selection of active site residues for all human CDKs was obtained from a multiple sequence alignment based on the corresponding twenty-one residues in the CDK2 active site as shown in

Table 3-1. This analysis indicates the overall conservation of the active site among these CDKs and also shows some variations. The active site residues corresponding to Lys33 in CDK2 (described in Chapter 3.2.2) are fully conserved in all CDKs. This suggests that the salt bridge is a feature of all CDKs. CDK8, CDK10 and CDK11 has a tyrosine residue corresponding to His85 and His100 in CDK4 and CDK6. All other CDKs have phenylalanine at this position. Glu144 is specific to CDK4. The distribution of different amino acids in the CDK1-CDK11 active sites is illustrated in Figure 3-6 (a) with a sequence logo (Schneider and Stephens, 1990; Crooks *et al.*, 2004).

The active site analysis of human CDKs was further extended to include all the CDKs and CDK like proteins as displayed in Figure 3-6 . Malumbres *et al*. have proposed a new extended nomenclature of the CDKs that has been agreed by the HUGO gene nomenclature committee, therefore the new nomenclature is used in Figure 3-5 along with classical names of CDK11-CDK20 and uniprot ids of these proteins (Malumbres *et al.*, 2009). Two highly similar genes CDCL1 and CDCL2, which originated by gene duplication, encode CDK11. The active site residues of both CDK11 (A and B) are identical.

$Lys33^{CDK2}$ is fully conserved over the entire extended family of CDKs including the CDK like proteins. In addition to this the downstream neighbors of lysine Gly-Val-

Ala are also conserved in the entire family of CDKs and CDK like proteins forming a fully conserved GVAK block.

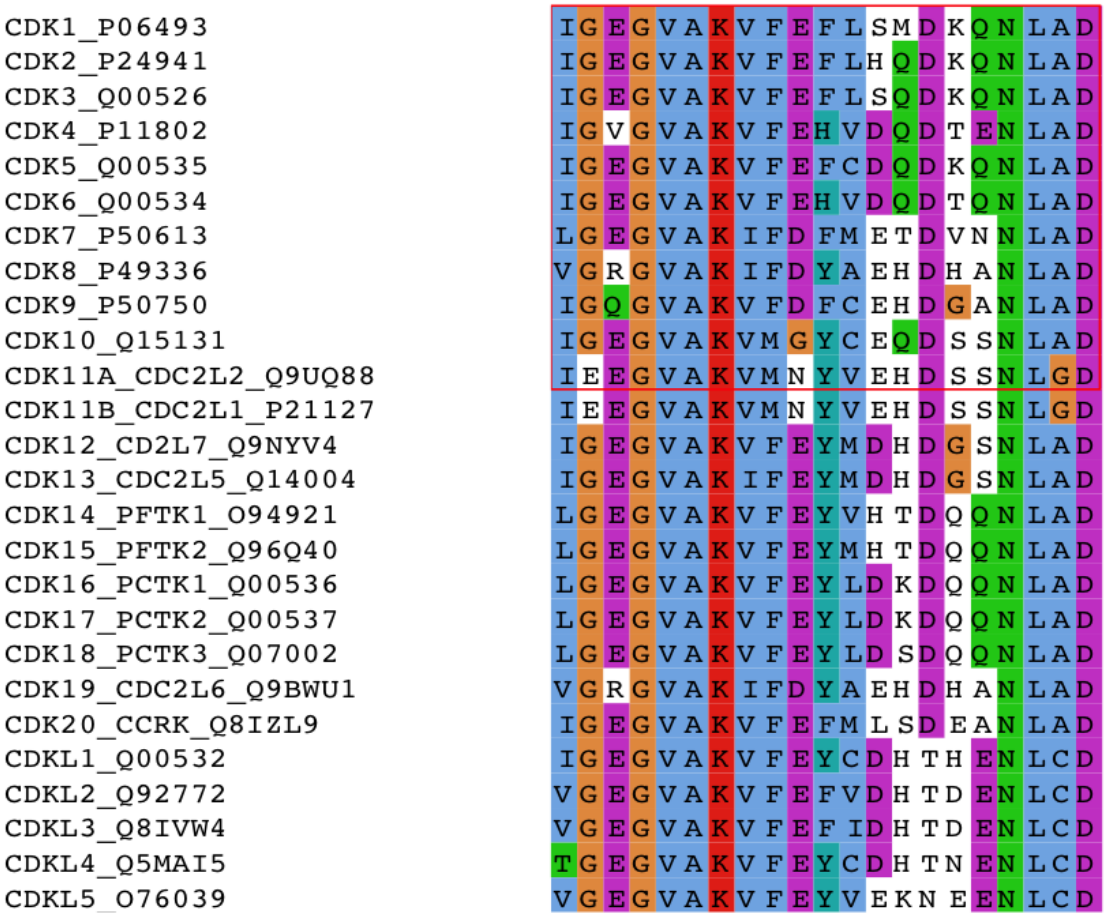


Figure 3-5 Active site analysis of all known human CDKs and CDK like proteins (CDK1CDK11 enclosed by a red border). The Gly-Val-Ala-Lys region is fully conserved in all members of extended family of CDKs and CDK like proteins. The Glu144 is unique to CDK4 compared with all other CDKs. However, all CDKL proteins also have a Glu at the corresponding position in the active site.

Glu144 is a key residue in the active site of CDK4 that does not follow the phylogenetic distribution (Figure 3-11). As a consequence this residue may play a significant role in the specificity for different ligands of CDK4. Only CDK like proteins (CDKL1-5) have Glu residue corresponding to the Glu144 CDK4 in the active site while all other members of the extended family of CDKs have different amino acids at this position which are Gln (CDK1-CDK3, CDK5, CDK6 and CDK14-CDK18), Asn (CDK7), Ala (CDK8, CDK9, CDK19 and CDK20) and Ser (in CDK10-CDK13).

Glu144 of CDK4 is considered to be involved in determination of CDK4 specificity compared to CDK2 for different inhibitors (McInnes *et al.*, 2004), which is further discussed in Chapter 5.

A sequence logo for the active site of all members of extended family of CDKs and CDK like proteins is shown in Figure 3-6 (b). In this sequence logo the height of each letter is proportional to its presence in CDKs and CDK like proteins. The GVAK block of consensus sequence (position 4-7) as described earlier is also prominent in this sequence logo.

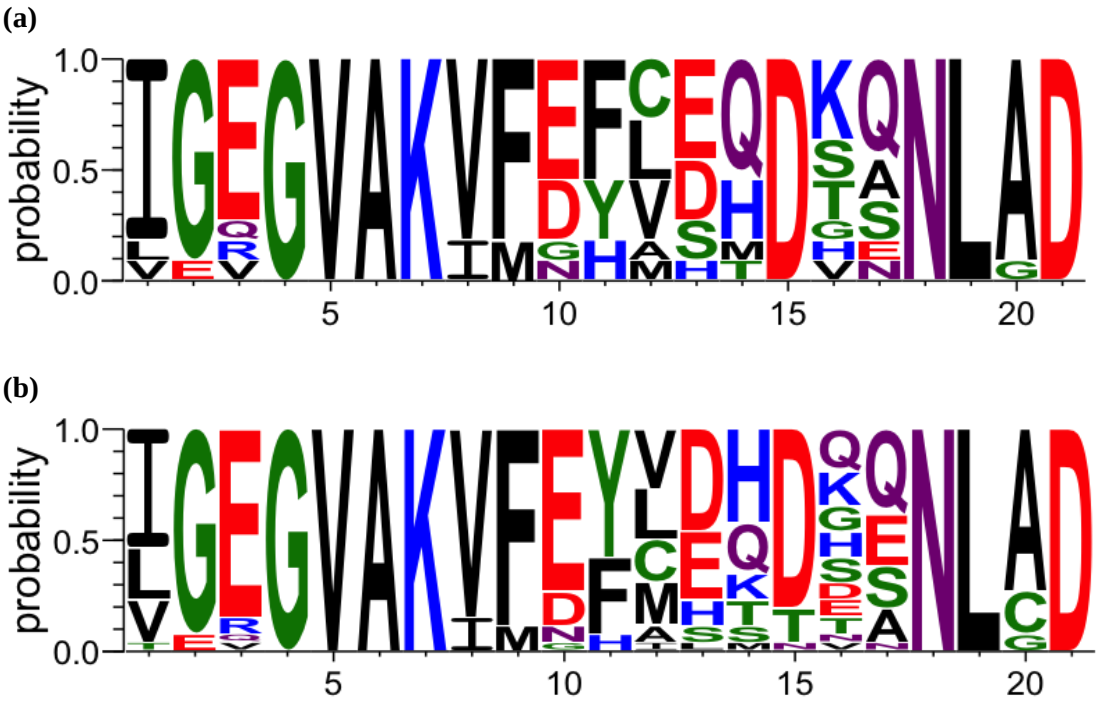


Figure 3-6 Sequence logos of CDKs active site (a) A sequence logo obtained for active site analysis of human classical CDKs (CDK1-CDK11) indicating a consensus region
(b) Consensus sequences in active site of all human CDKs and CDK like proteins are highlighted with a sequence logo.

A BLAST search for all CDKs sequences against the Dunbrack PDB database (pdbaa) reveals that protein crystal structures are available for CDK2, CDK4, CDK5, CDK6, CDK7 and CDK9 (last updated April 16, 2010). A structural comparison of the active sites of all available CDKs is shown in Figure 3-7. All CDKs shown in the

picture have a phenylalanine residue corresponding to the Phe82 of CDK2 with the exception of CDK4 and CDK6 which have a histidine residue shown in blue and red. As discussed earlier the side chains of $His85^{CDK4}$ and $His100^{CDK6}$ are oriented toward the binding pocket, therefore these residues may have some role in the differential selectivity of CDK4 and CDK6 for certain ligands compared with CDK2. The Lys89 residue is only present in CDK2 and CDK5 (shown in Figure 3-7). The corresponding residues in CDK4 and CDK6 are Thr, Val in CDK7 and Gly in CDK9.

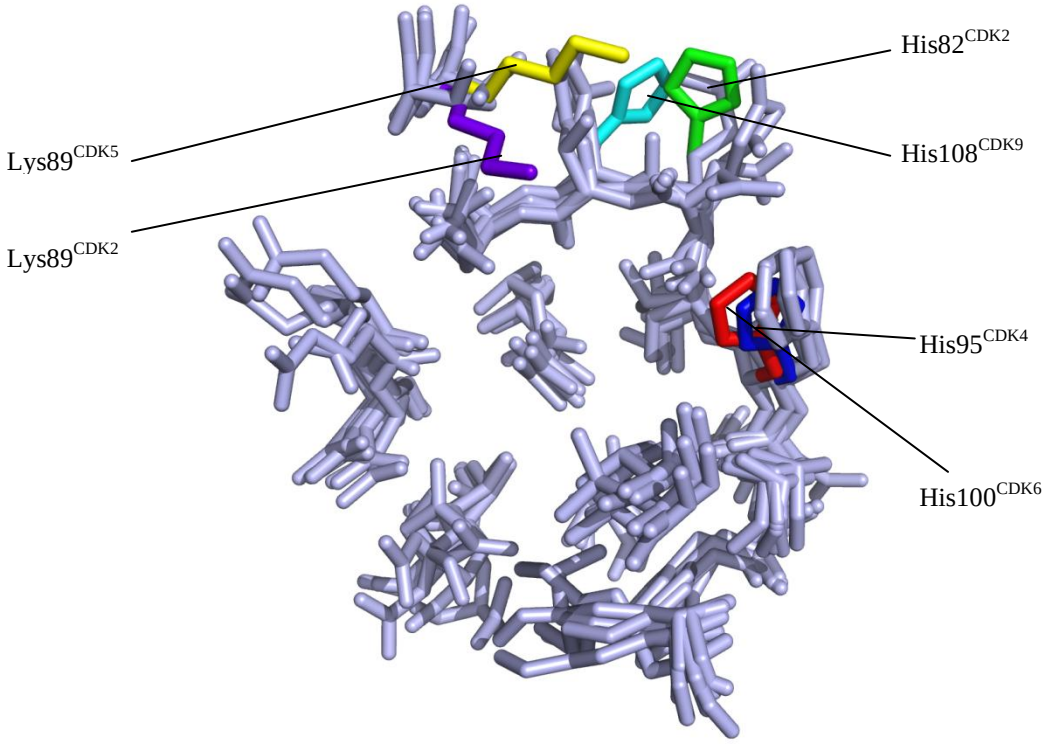


Figure 3-7: A stick representation of the structural comparison of active sites of CDK2, CDK4, CDK5, CDK6, CDK7 and CDK9. The His82 of CDK2 structure (PDB ID 2CCH) is shown in green. The His 95 of CDK4 (PDB ID 2W99) and His100 of CDK6 (PDB ID 1BI7) are shown in blue and red. The Lys89 of CDK2 and CDK5 (PDB ID 1UNL) are shown in purple and yellow. The His108 of CDK9 (PDB ID 3BLR) is shown in cyan.

The potential role of $Lys89^{CDK2}$ has also been reported to explain CDK2 selectivity and in structure-based design of new inhibitors by different groups (Ikuta *et al.*, 2001; Schoepfer *et al.*, 2002; McInnes *et al.*, 2004; Alzate-Morales *et al.*, 2007). The structural and sequential analysis of active site variations of CDKs highlight potential residues which may play a key role toward the selectivity of these CDKs and

the information obtained with this study may guide the structure-based design of selective inhibitors of CDK4.

3.3 Phylogenetic Analysis

3.3.1 Phylogenetic analysis of Human CDKs

The sequences of human CDKs were obtained from the UniProtKB/Swiss-Prot database (Bairoch *et al.*, 2007; Apweiler *et al.*, 2010). The multiple sequence alignment generated for human CDKs (CDK1-CDK11 UniRef IDs given in Table 3-2) is shown in Figure 3-8 (see Section 2.3). It is apparent from the multiple sequence alignment that only CDK4 has a glycine rich loop (residue 42–48 consisting of 7 glycines) as it is not present in other CDKs. It is also apparent that CDK7 has an insertion of Ser, Glu and Asp in the corresponding region. Phylogenetic trees analysis of human CDKs based on this multiple sequence alignment (Figure 3-9) provide information about the evolution and clustering of these CDKs. These trees are generated with MrBayes and MEGA programs using Bayesian inference of phylogeny and neigbor joining methods, respectively. The phylogenetic trees of these CDKs are rooted using CDKL1 (Q00532) as an out-group. The phylogenetic trees for CDKs generated using MrBayes and MEGA agree on the overall clustering of CDKs. This clustering of CDKs may also correspond to the different biological roles of these CDKs in the cell. Both trees show that CDK1 to CDK6 are grouped in one cluster while CDK7 to CDK11 are grouped separately in a second cluster. It is also evident from these trees that CDK2 and CDK3 are grouped together closely suggesting that these may have evolved from a common gene ancestor. In a similar way CDK4 and CDK6 are also shown to originate from a single common gene ancestor. CDK1, CDK2, CDK4 and CDK6 are directly involved in regulation of different phases of cell cycle. CDK3 is closely related to CDK2 in terms

of sequence similarity and it is proposed to be involved in the G1/S phase of cell cycle in mammals (Hofmann and Livingston, 1996; Hengstschlager *et al.*, 1999), however, CDK3's role in cell cycle is yet not fully understood (Deshpande *et al.*, 2005b). It is believed that CDK5 is not involved in the cell cycle, rather this kinase plays a key role in sensory pathways and neuron biology (Tsai *et al.*, 1993; Paglini and Caceres, 2001; Wei and Tomizawa, 2007) and in brain development (Hawasli *et al.*, 2009).

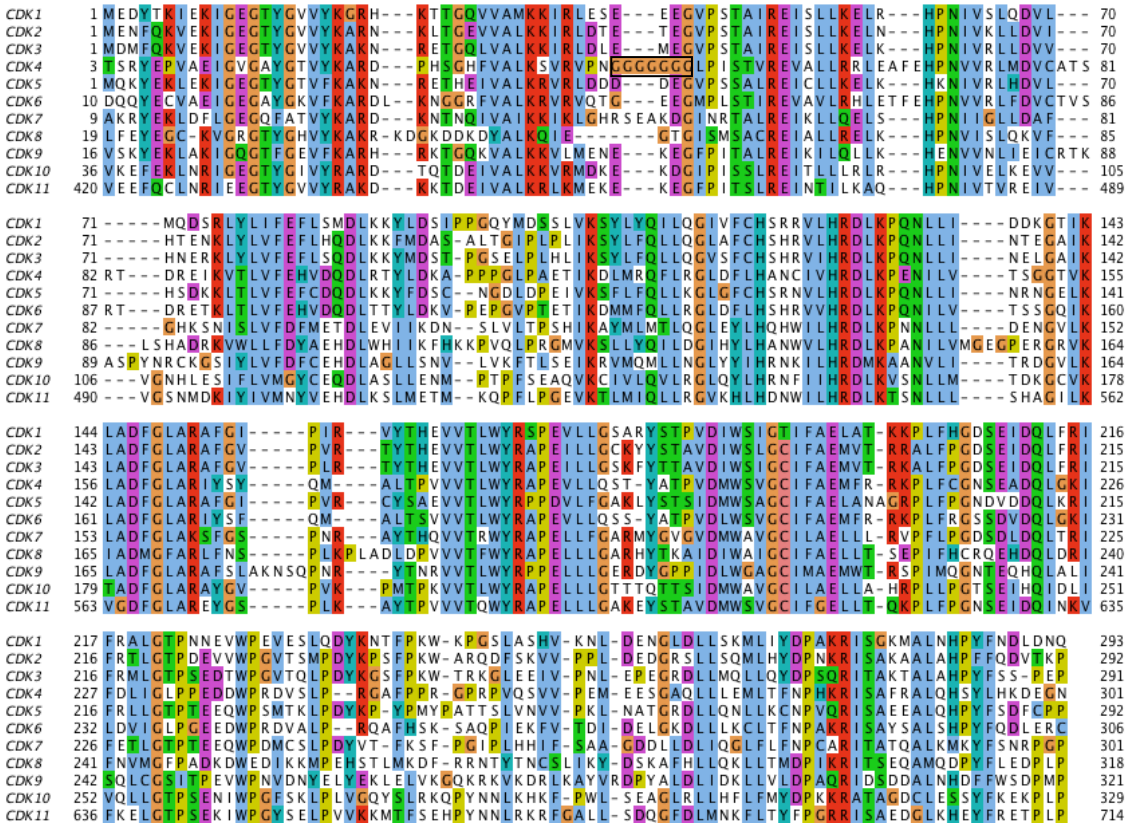


Figure 3-8: Multiple sequence alignment of human CDKs (CDK1-CDK11). These protein sequences were aligned with MUSCLE.

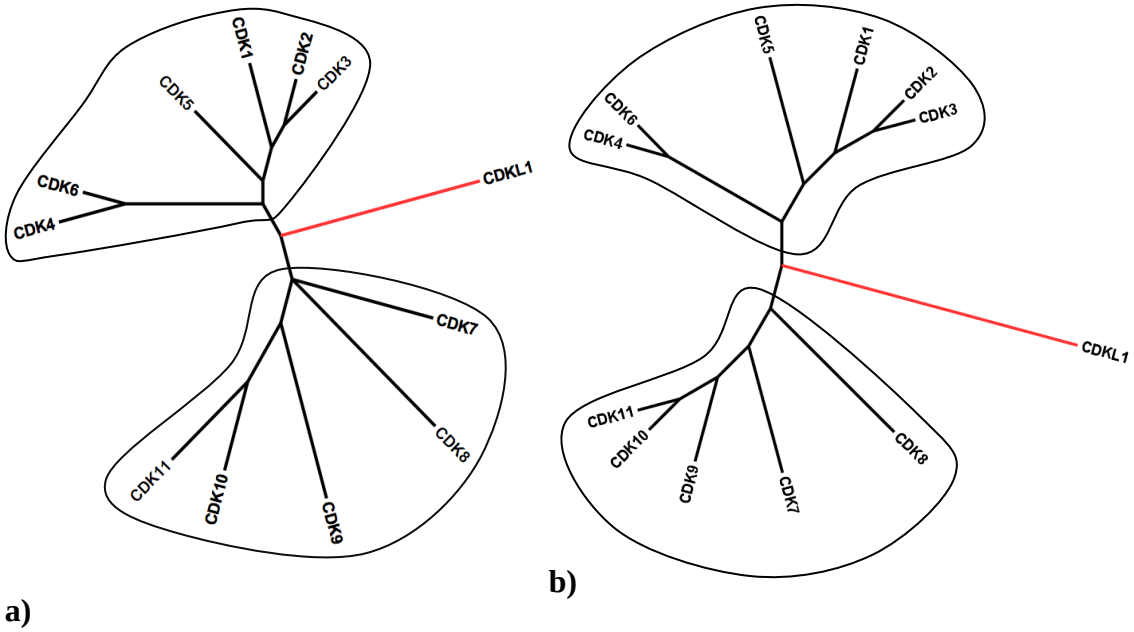


Figure 3-9: Phylogenetic tree of eleven human CDKs. a) Phylogenetic tree generated with MrBayes (Huelsenbeck and Ronquist, 2001), which uses Bayesian Inference of Phylogeny. This tree is rooted at CDKL1. **b)** Phylogenetic tree generated with MEGA 4.0 using the neighbor joining method.

To study how newly classified CDKs (CDK12-CDK20 as discussed in Chapter 3.2.4) and CDKLs fit into the evolution of CDKs the phylogenetic analysis was extended to include all the members of CDKs and CDKLs. The multiple sequence alignment of all human CDKs and CDKL proteins (UniRef IDs given in Table 3-2) is shown in Figure 3-10. This alignment also shows that a glycine rich loop as described earlier is a characteristic feature of human CDK4 and this loop is absent in all other CDKs and CDKL proteins. The phylogenetic trees of the extended family of CDKs and CDKL proteins based on this multiple sequence alignment are shown in Figure 3-11. The trees obtained with Bayesian inference of phylogeny and neigbor joining methods are in consensus for the distribution of all CDKs and CDKL. This result shows that CDK2/CDK3 and CDK4/CDK6 are grouped together similar to as shown in Figure 3-8.

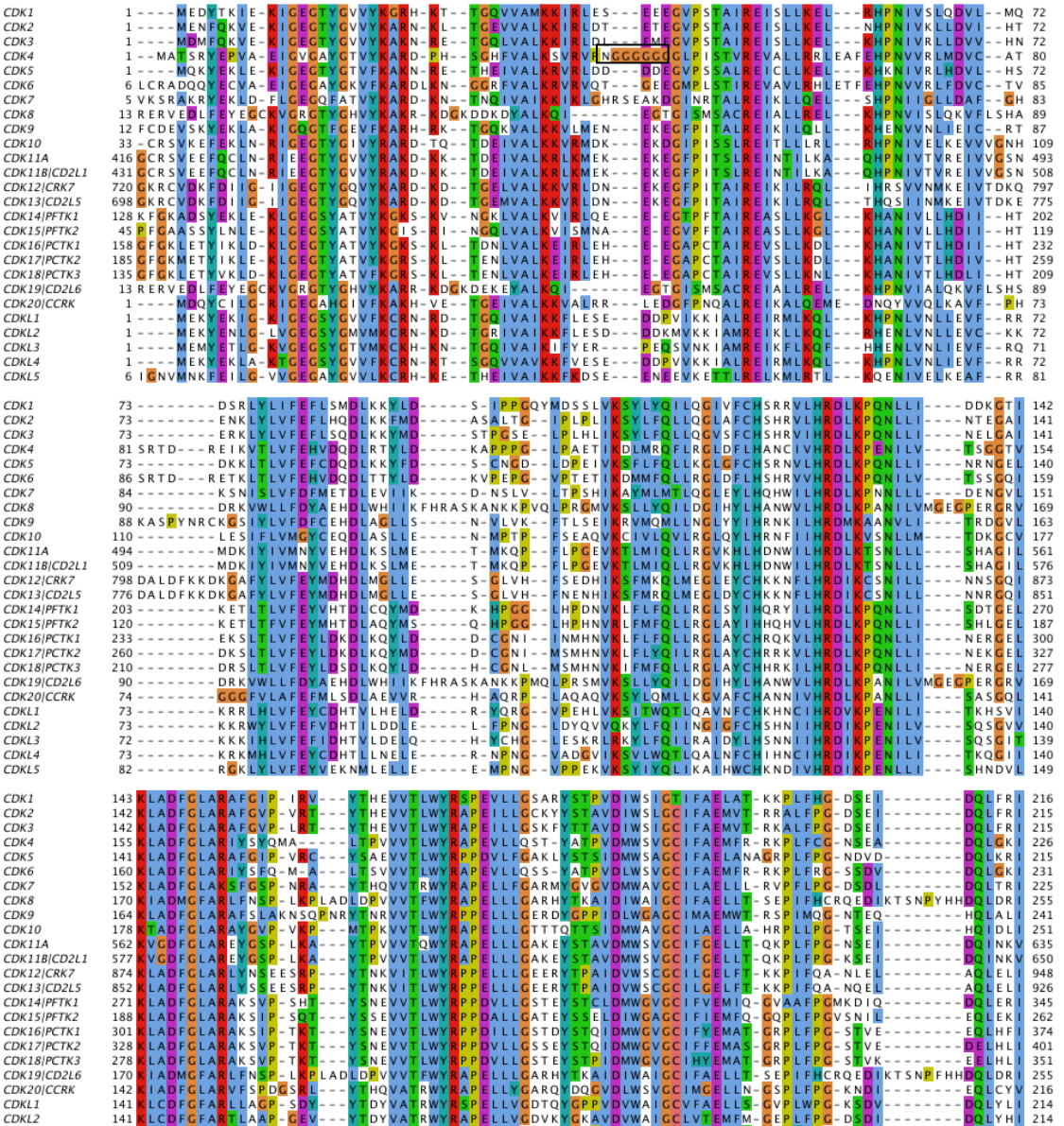


Figure 3-10: Multiple sequence alignment of all members of an extended family of CDKs and CDK like proteins. This MSA also displays that a glycin rich loop is characteristic of CDK4 only and is not present in other CDKs and CDKLs.

Table 3-2: A list of Human CDKs and CDKL and their corresponding UniRef identifiers.

Symbol	UniRef ID	Symbol	UniRef ID	Symbol	UniRef ID
CDK1	P06493	CDK10	Q15131	CDK18	Q07002
CDK2	P24941	CDK11A	Q9UQ88	CDK19	Q00534
CDK3	Q00526	CDK11B	P21127	CDK20	Q8IZL9
CDK4	P11802	CDK12	Q9NYV4	CDKL1	Q00532
CDK5	Q00535	CDK13	Q14004	CDKL2	Q92772
CDK6	Q00534	CDK14	O94921	CDKL3	Q8IVW4
CDK7	P50613	CDK15	Q96Q40	CDKL4	Q5MAI5
CDK8	P49336	CDK16	Q00536	CDKL5	O76039
CDK9	P50750	CDK17	Q00537		

CDK5 is shown to be closely related to CDK14 (PFTK1), CDK15 (PFTK2) and CDK16-CDK18 (PCTK1-PCTK3). CDK14 is reported to interact with cyclin D3 and it play a role in cell proliferation by acting as a CDK (Shu *et al.*, 2007). It is also reported to form a complex with cyclin Y and promoting Wnt signalling linked to the cell cycle (Davidson and Niehrs, 2010).

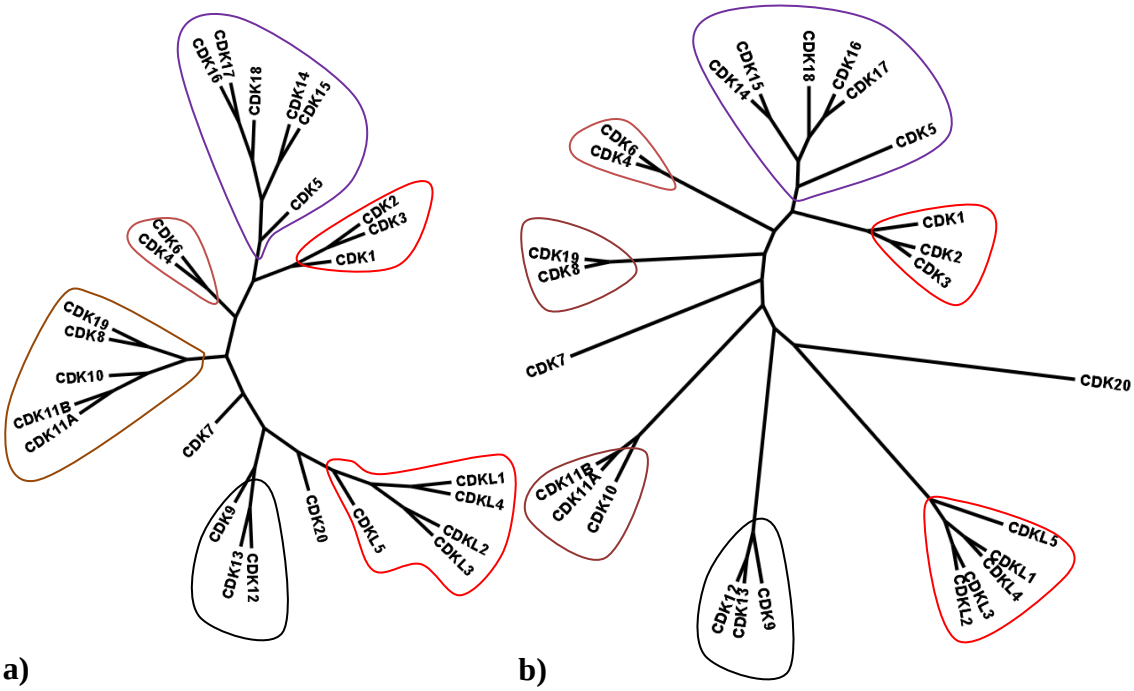


Figure 3-11: Phylogenetic tree displaying evolutionary relationship among all members of the human CDKs and CDKL proteins. a) Phylogenetic tree generated with MrBayes using Bayesian Inference of Phylogeny **b)** Phylogenetic tree generated with MEGA 4.0 using neigbor joining.

CDK15 forms a complex with the cyclin Y-like-1 protein and shares substrate specificity with CDK14 (Davidson *et al.*, 2009; Kaldis and Pagano, 2009). The functions of CDK16-18 are yet not fully understood (Morgan, 1997; Malumbres *et al.*, 2009). The CDK-like proteins CDKL1-CDKL5 are clustered together, but separated from the all other CDKs in the phylogenetic tree. Little is known about the biological role of CDKLs and their association with cyclins, however, they are believed to be involved in neuro-developmental processes (Gomi *et al.*, 2010).

The sequence and phylogenetic analysis of the CDK family of proteins provides valuable information about the relatedness of different members of this family which can be used to study the compensatory functions of closely related CDKs and their potential importance as therapeutic targets in human disease.

3.3.2 Cross-species phylogenetic analysis of CDK4 and CDK6

Phylogenetic analysis of CDK4 and CDK6 was carried out to study their presence and absence among different species. A BLAST similarity search with human CDK4 (P11802) against Uniref100 with a cut-off (E-value[*] 5e[†]-89 and score 331) resulted in 54 sequences of CDK4 and CDK6 which were retrieved using "fastacmd". These 54 sequences were subjected to a multiple sequence alignment and a sequence profile of CDK4/6 was obtained which was then confirmed with Hidden Markov Model search using HMMER for detection of any missing remote homolog of CDK4/CDK6 (see Section 2.4). The CDK4/CDK6 profile was further refined manually with the removal of 22 redundant and incomplete sequences from the 54 sequences in total. The multiple sequence alignment of the remaining 32 sequences (Figure 3-12) was used for phylogenetic analysis. These 32 selected sequences correspond to different species of

[*] The E−value is a measure of the reliability of the S score.
[†] In BLAST notation a numeric value such as e^{-130}, the "e" means to the 10th power i.e. 10^{-130}

the super-phylum deuterostome and order actiniaria. The multiple sequence alignment and the sequence logo (Figure 3-13) highlights certain conserved amino acids which distinguish CDK4 from CDK6. The MSA reveals that Val96^{CDK4} in the CDK4 replaced with Ile residue in Mouse (*Mus musculus*) and Rat (*Rattus norvegicus*) CDK4.

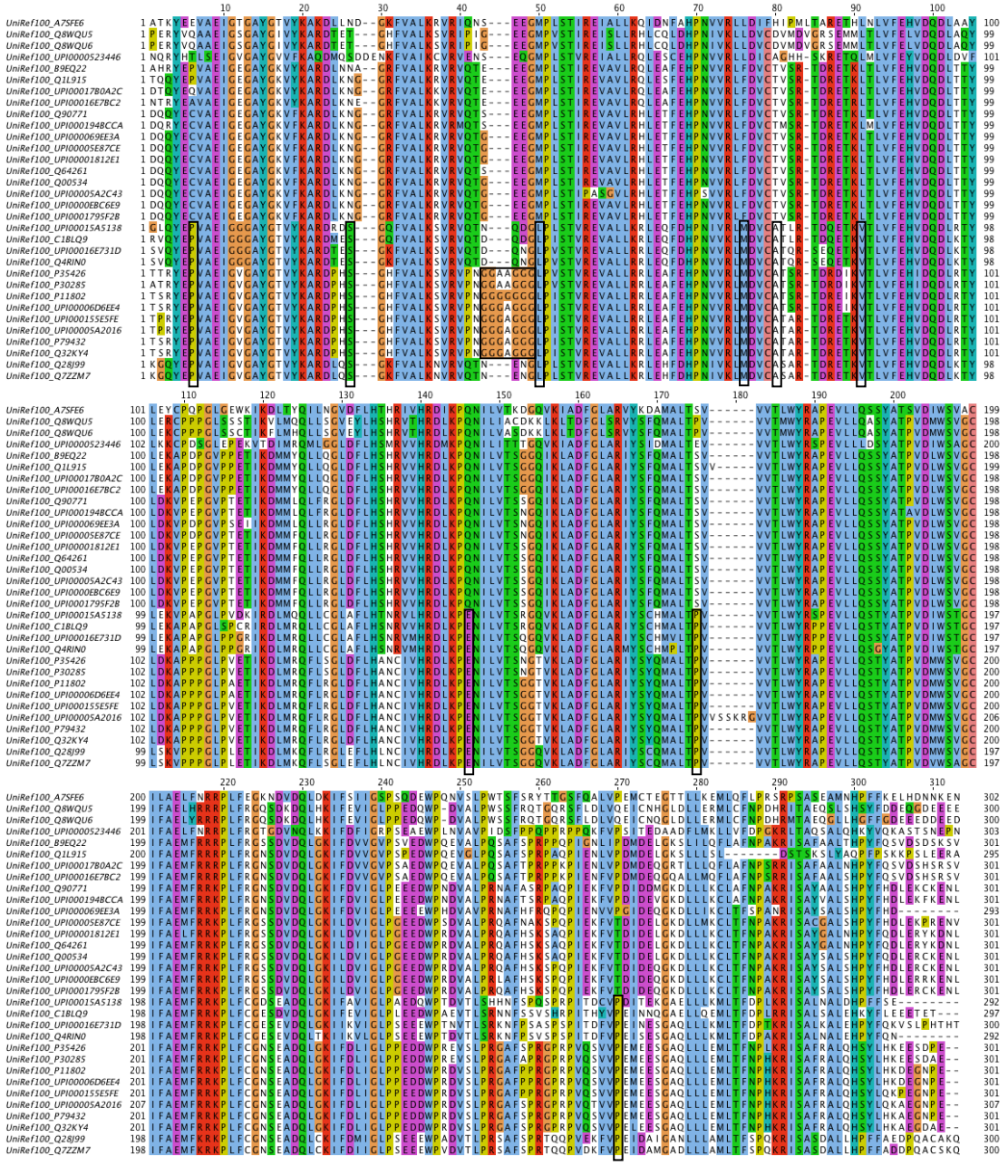


Figure 3-12: Multiple sequence alignment of CDK4/CDK6 for different species of super-phylum deuterostomes and order actiniaria. The above MSA display 32 sequences. The sequences for CDK4/CDK6 were obtained by a BLAST similarity search with human CDK4 (P11802) against the Uniref100. The initial selection of these sequences were confirmed with HMMER profile search and then manually adjusted to remove redundant and incomplete sequences.

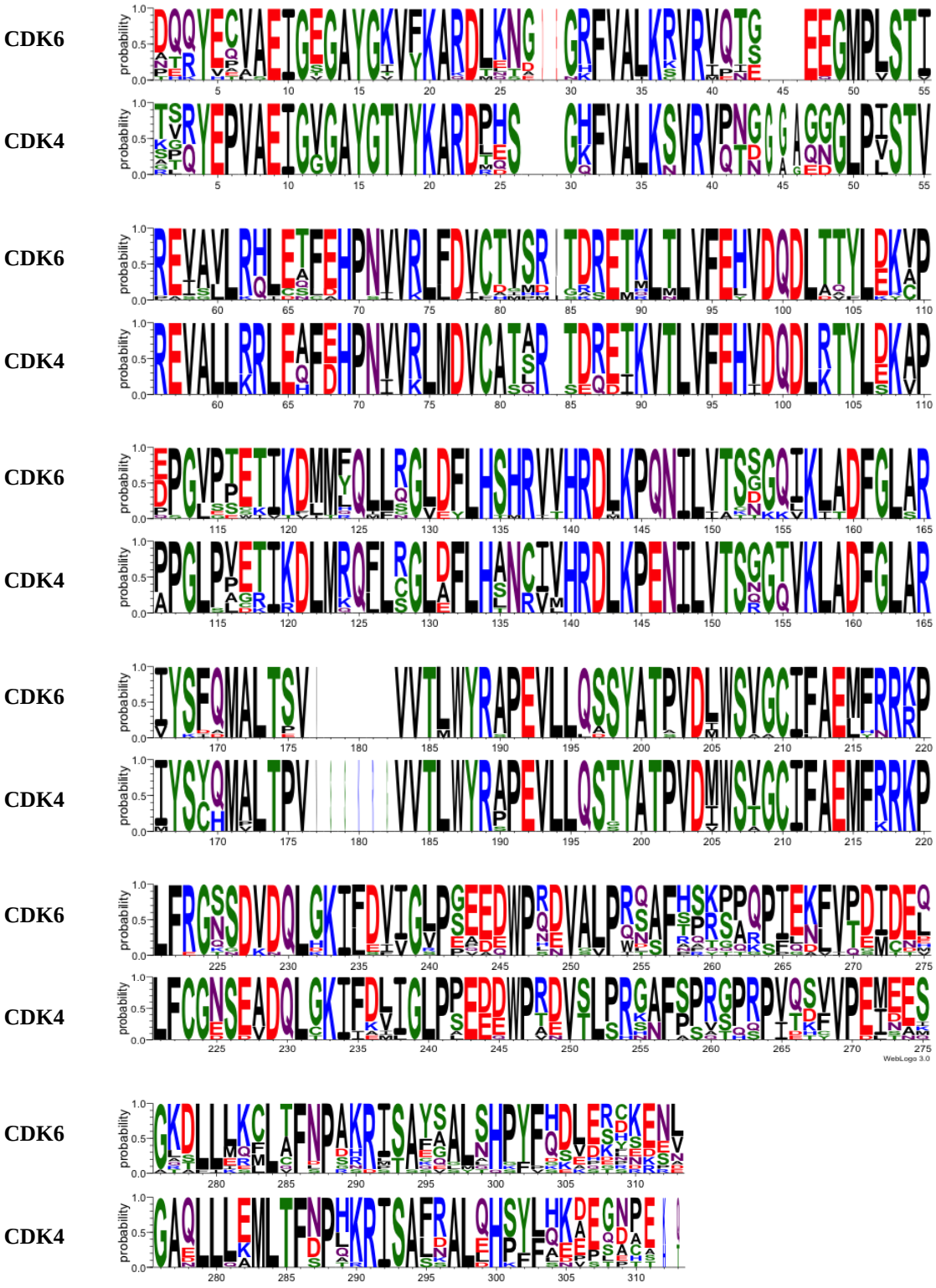


Figure 3-13 **Sequence logos of CDK4/CDK6.** A sequence logo of CDK4/CDK6 obtained based on the multiple sequence alignment of 32 sequences of CDK4/6 (Figure 3-12)

The CDK4/CDK6 multiple sequence alignment indicates that the glycine rich loop (residue 42–48 consisting of 7 glycine) present in human CDK4 is not found in all

species. In addition to human this loop is observed only in mammals with an Ala insertion as shown in Figure 3-12. Based on these results it is suggested that this glycine rich region is a characteristic of mammalian CDK4 only. As discussed in pervious section this poly glycine rich region is also absent in all other CDKs and CDKLs. CDK4 sequences for the representative species have conserved Pro, Ser, Leu, Ala, Val, Glu, and Pro residue as highlighted in the Figure 3-12.

A phylogenetic tree of CDK4 and CDK6 proteins is shown in Figure 3-14. The overall evolution of CDK4 and CDK6 among different species is consistent with the tree of life (Figure 3-15). CDK4 and CDK6 both are present in mammals and in fishes. The phylogenetic tree for CDK4/CDK6 suggests absence of CDK4 in chicken and other avian species. A BLAST similarity search with human CDK4 (P11802) vs. chicken genome (*Gallus gallus*) had also shown absence of CDK4 in chicken genome.

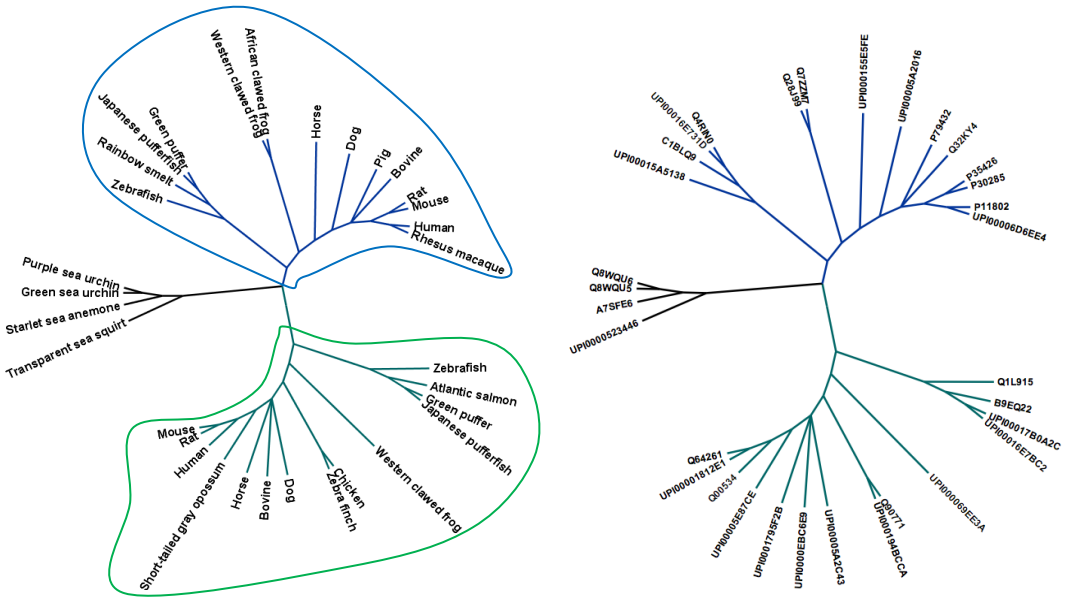


Figure 3-14: Phylogenetic tree of CDK4 and CDK6 in different species. The phylogentic tree was constructed using MrBayes. The clade shown in blue represents CDK4 and clade shown in green represent CDK6

To further investigate the presence or absence CDK4 in chicken a BLAST similarity search against chicken ESTs (locally compiled database consisting 600075

chicken ESTs last updated on 28 Dec, 2009) was carried out. This search supports that CDK4 may be absent in chicken as the closest hits retrieved represent CDK6 and CDK3. A BLAST similarity search against Zebra Finch (*Taeniopygia guttata*) genome had also shown that CDK4 seems not present in its genome. The binding partners of CDK4 (Cyclin D1, D2 and D3), however are present in both Chicken and Zebra Finch as revealed by BLAST searches.

Based on these results it may be concluded that CDK4 is not present in some or all avian species and its role may be compensated for by CDK6. It may be interesting to find the presence and absence of all other proteins which interact with CDK4 and to find how the function of CDK4 is compensated in birds. This seems to be an example of gene loss in the avian linage.

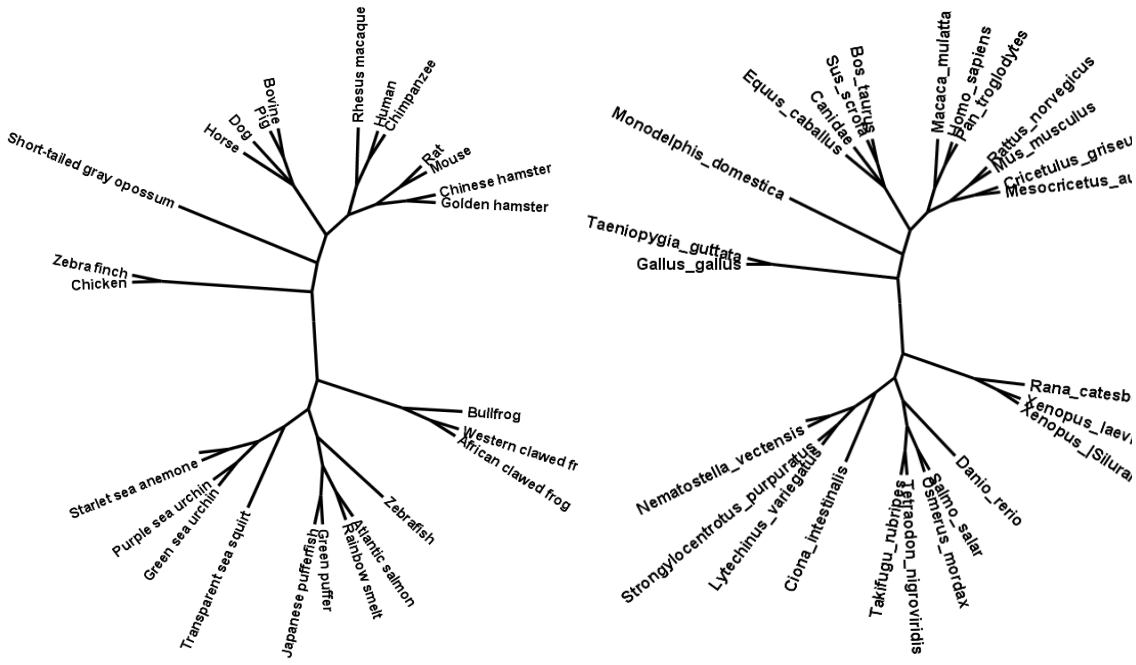


Figure 3-15: A sub-set of the tree of life. This tree is displaying species shown in phylogenetic trees for CDK2, CDK3, CDK4 and CDK6

3.3.3 Cross-species phylogenetic analysis of CDK2 and CDK3

BLAST similarity search with human CDK2 (P24941) used as a query sequence against Uniref100 with a cut off (E-value e-129 and score 465) resulted in 37 sequences of CDK2 and CDK3. These 37 sequences were subjected to a multiple sequence alignment, which was then used to build a sequence profile of CDK2/3. This profile was then also confirmed with Hidden Markov Model search using HMMER (see Section 2.4). The CDK2/CDK3 profile was further refined manually with the removal of redundant and incomplete sequences. In total 37 sequences 10 sequences were removed. The final multiple sequence alignment comprised of 27 sequences (Figure 3-16) and was used for phylogenetic analysis. There is high (76.0%) sequence identity between CDK2 and CDK3 calculated for human CDK2 and CDK3. All the CDK3 sequences have a C-terminal extension compared with CDK2. The multiple sequence alignment highlights certain conserved amino acids which distinguish CDK2 from CDK3 (Figure 3-16). There is a Gln residue in all CDK3 corresponding to the Glu28 of human CDK2. There is a conserved pattern containing Thr-Glu-Thr (residue 39-41 in human CDK2) in all sequences of CDK2 while only Glu is found conserved in CDK3. There is a conserved Lys, Glu, Val and Ile in all sequences of CDK3 corresponding to Asn59, Gly98, Leu115 and Leu124 of human CDK2, respectively. Leu115 is a part of α-helix located in subdomain VIA of CDK2 while all other residues constitute different loops. Some of these distinguishing features can be used to identify CDK2 and CDK3 from unannoted sequences.

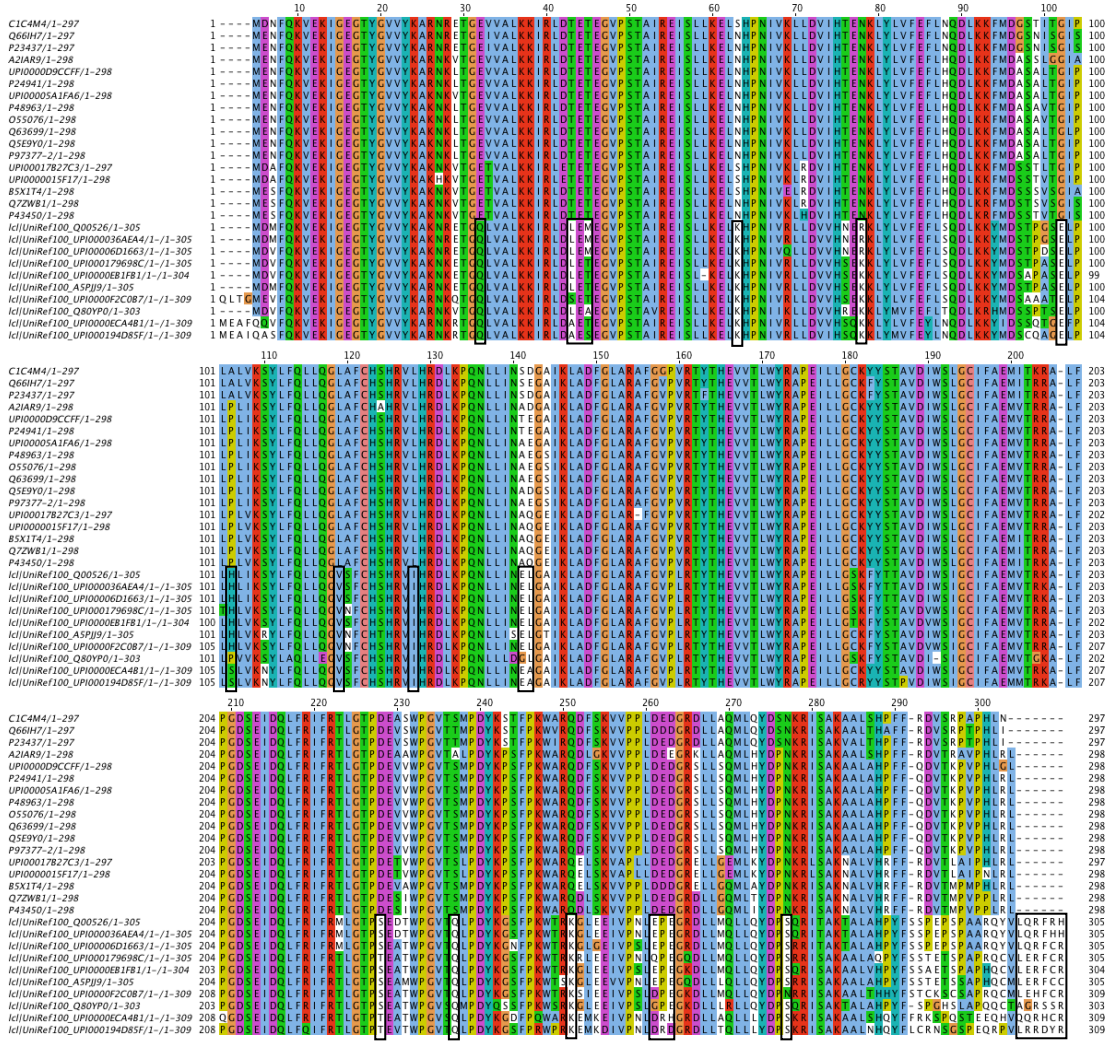


Figure 3-16: Multiple Sequence Alignment of 27 sequences of CDK2 and CDK3 among different species. The sequences for CDK2/CDK3 were obtained by a BLAST similarity search with human CDK2 Sequence (P24941) against the Uniref100. The initial selection of these sequences was confirmed with HMMER profile search (See Methods) and then manually adjusted to remove redundant and incomplete sequences.

A phylogenetic tree for CDK2 and CDK3 among different species is displayed in Figure 3-17. Similar to CDK4 and CDK6 the overall distribution of CDK2 and CDK3 among different species is also consistent with the tree of life (Figure 3-15). CDK2 and CDK3 both are found in mammals and birds. CDK3 is not found in frogs and fishes, and it might have evolved later.

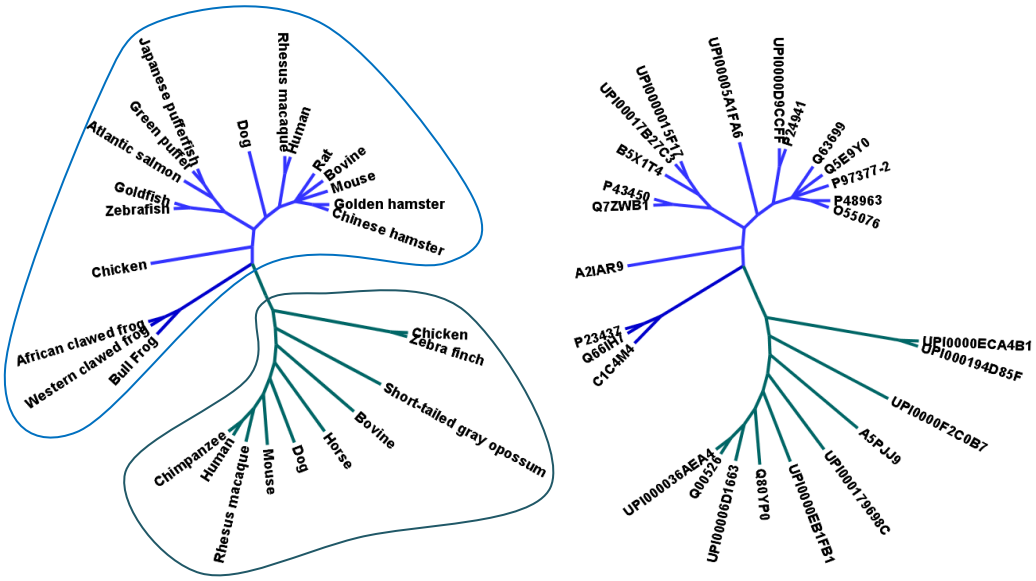


Figure 3-17: Phylogenetic tree of CDK2 and CDK3. The phylogentic tree was constructed using MrBayes. The clade shown in blue represents CDK2 and clade shown in green represent CDK3

3.3.4 Phylogenetic analysis of cyclins

Cyclins associated with CDKs are important regulators of the cell cycle. Cyclins coordinate the timing of different events necessary for cell cycle progression by activating cyclin-dependent kinases (CDKs). There are twenty nine cyclins and cyclin like proteins. The complex network of protein-protein interactions (interactome) and phylogenetic relationship contains a lot of biological information. There are different experimental and computational methods available to study the interaction of proteins (Jansen *et al.*, 2003; Shoemaker and Panchenko, 2007). Comparison of evolutionary histories (phylogenetic trees) of proteins can be used to get an insight into protein interactions and their co-evolution (Juan *et al.*, 2008; Pazos *et al.*, 2008).

The study of phylogenetic analysis was extended to cyclins and cyclin like proteins family to investigate the co-evolution of CDKs and corresponding cyclins. The phylogenetic tree (Figure 3-18) shows that cyclins also are branched together according to their biological role. All the cyclins involved in the cell cycle are shown to be evolved from a single common ancestor and these belong to a same clade. Cyclin D1,

D2 and D3 are closely related to each other and interact with the closely related pair of CDK4 and CDK6. Similarly, the interacting partners (cyclin A1, A2, B1, B2, B3, E1, and E2) of closely related CDK1, CDK2 and CDK3 show a close evolutionary relationship.

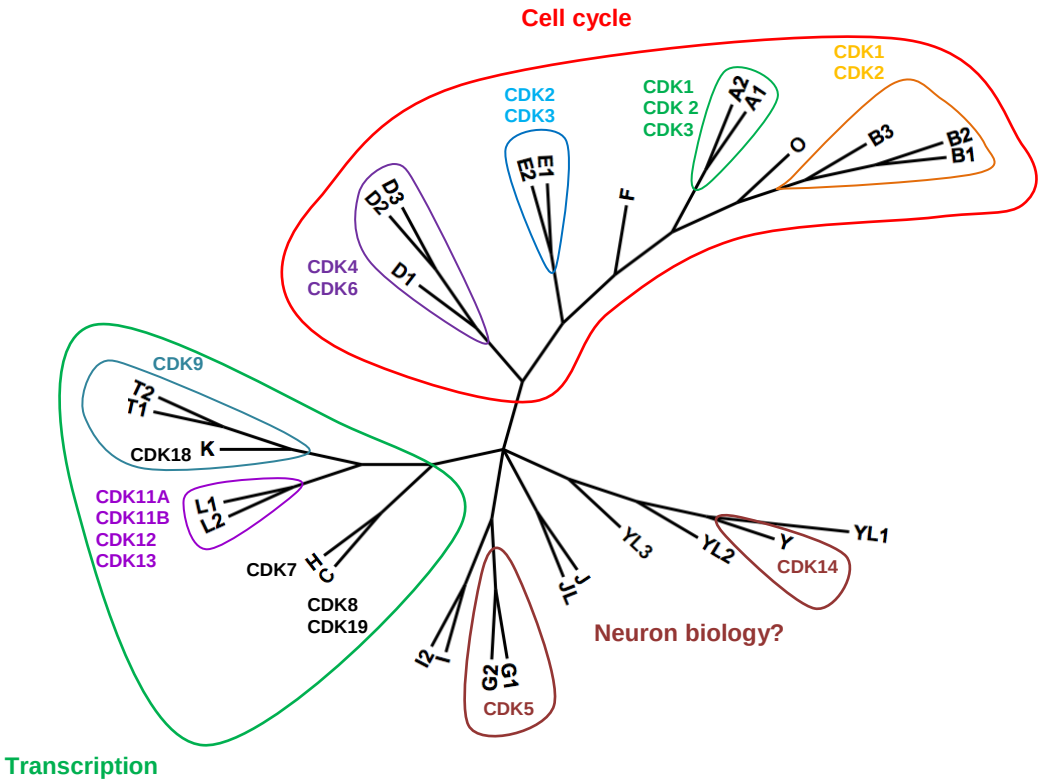


Figure 3-18: Phylogenetic tree among all human cyclins and cyclin like proteins. The phylogenetic tree was generated using MrBayes**.**

Cyclin F is an orphan cyclin as it does not bind or activate any cyclin-dependent kinases (CDKs) (Bai *et al.*, 1994; Fung and Poon, 2005). Cyclin F shares about 40% sequence identity with cyclin A. The phylogenetic conservation of cyclin F as shown in Figure 3-18 suggests that it may have some role in cell cycle events. It has been reported that cyclin F is expressed ubiquitously and Cyclin F protein levels fluctuate during the cell cycle (Bai *et al.*, 1994). The level of cyclin F oscillates during the cell cycle similar to cyclin A (Fung and Poon, 2005). Tetzlaff et al have reported cell cycle

defects in mouse embryonic fibroblasts (MEFs) lacking cyclin F (Tetzlaff *et al.*, 2004). Centrosomal and mitotic abnormalities are also reported to be associated with the depletion of cyclin F in G2 phase of the cell cycle (D'Angiolella *et al.*, 2010). The cyclins phylogenetic tree also predicts the relatedness of cyclin O with the cell cycle. The human Cyclin O has 350 amino acids and shares about 28% of identity with human Cyclins A2 and B1. Limited information is available about the role of cyclin O. There is evidence that cyclin O interact with both CDK1 and CDK2 and preferentially activate CDK2 (Huguet, 2008). The interaction of cyclin O with cell cycle CDKs (CDK1 and CDK2) is also supported by its phylogenetic distribution (Figure 3-18). Cyclin C, H, K, L1, L2, T1 and T2 are linked with transcription (Malumbres *et al.*, 2009) and these are cluster together on the phylogenetic tree. Cyclin C is also believed to interact with CDK3 and to play a role in cell cycle (Sage, 2004). Cyclin H and K interact with CDK7 and CDK9, respectively (Fu *et al.*, 1999; Saiz and Fisher, 2002). Cyclin G1, G2 and cyclin Y play a role in neuron biology (Jiang *et al.*, 2009; Malumbres *et al.*, 2009). Cyclin I, I2, J, cyclin J like protein and cyclin Y like protein are not fully characterized yet however there phylogenetic distribution predicts that these may also have some role in the neuron biology.

The phylogenetic distribution of cyclins corresponds to phylogenetic distribution of most of the CDKs (Figure 3-11) indicating a possible co-evolution of these interacting partners, a further detailed analysis of phylogenetic distribution of CDKs and cyclin in different species may provide valuable information.

3.4 Conclusion

Active site analysis plays a major role in understanding the involvement of particular residues to substrate recognition and the overall function of a particular protein. In this work the active site of CDK2 is defined based on protein ligand

interactions of 129 high resolution X-ray structures using the LPC tool (Sobolev et al., 1999). A comparison of the active site residues of the active and inactive form of CDK2 shows alternative conformations of His84 and Lys33. The side chain of $Lys33^{CDK2}$ is known to make salt bridge in CDK2 and also interact with different ligands (De Bondt et al., 1993; Davies et al., 2001; Meijer et al., 2003). Lys33CDK2 is found as conserved in the entire family of CDKs and CDKLs. The residue corresponding to $Lys33^{CDK2}$ in CDK4 is Lys35. Sequential and structural analysis of the active sites of all the available CDKs show a high degree of sequence identity; however there are variations in some residues which may be the basis of specificity for individual CDK. These variations in the active side include Glu144, and Thr102 in CDK4, in CDK2 the corresponding residues are Gln131 and Lys89. Glu144 is specific to CDK4, all other CDKs have Gln, Ala or Ser at this position. In addition to CDK4, all CDKL proteins also have Glu residue at the equivalent position in the active site. His95 in CDK4 and His100 in CDK6 corresponding to the Phe82 of CDK2. Based on the orientation and position of His95 in the active site of CDK4 it is hypothesized that $His95^{CDK4}$ may have a role for CDK4 specificity for fascaplysin and other compounds. The information gained from the active site analysis is very useful for assigning flexibility in molecular docking studies (see Chapter 5).

Cross species phylogenetic analysis of CDK4/6 and CDK2/3 with the available dataset indicates the absence of CDK4 in chicken and zebra finch, CDK3 in frogs and fishes and may be explained by linage specific gene loss. A glycine-rich loop (residue 42–48) which has been thought as a characteristic for CDK4 is found in CDK of mammals only. Phylogenetic analysis of CDKs and cyclins show a close relationship of these interacting partners. The phylogenetic clustering of cyclin O and cyclin F with cell cycle cyclins predict that cyclin O and cyclin F may have some direct or indirect

role in cell cycle. The complete understanding of biological relationships and interactions between CDKs and cyclin is essential for understanding their exact role in different disease mechanisms and in the development of new therapeutic agents.

Chapter Four

Homology modelling of CDK4

Chapter 4 Homology modelling of CDK4

4.1 Introduction

Structural information of a protein is often more valuable than sequence alone in the studying of protein function, dynamics, interactions with other proteins and ligands, and in drug discovery (Hillisch *et al.*, 2004). Homology or comparative modelling can provide three-dimensional structural models for proteins ("targets") based on evolutionary related proteins for which experimental structures have been solved ("templates"). The prediction process consists of the identification of a suitable template, target-template alignment, model-building and the evaluation of models (Marti-Renom *et al.*, 2000; Sanchez and Sali, 2000). This chapter describes the generation of CDK4 homology models as in the beginning of this research project (April 2007) no structural information for CDK4 was available. Structural models for CDK4 based on CDK2 sequence similarity are described in Section 4.2 to 4.6.3. Two research groups published crystal structures of human CDK4 in its inactive form in March 2009 (Day *et al.*, 2009; Takaki *et al.*, 2009). The availability of X-ray structures provided an excellent opportunity to validate the homology modelling work and to use CDK4 structural information to generate a CDK4 model in a putative active form. A comparison of CDK4 structures and homology models is presented in Section 4.7.

4.2 Template Selection

The BLAST results (blastp) for CDK4 against the nr (non-redundant) protein database obtained from NCBI show 45% sequence identity with CDK2 and 68% with CDK6. The percentage sequence identity between CDK6 and CDK2 is 49%. The above results indicate that CDK2 and CDK6 structures are suitable templates for homology modelling of CDK4. The BLAST results for human CDK2 with the Dunbrack PDB database (March 26, 2008) revealed that 153 PDB structures have already been solved

and are present in the PDB, and in principle which could be used as potential templates for CDK4 modelling. Since March 2008 the list of available CDK2 structures has grown to more than 190 (Figure 4-1 and Appendix 1.1).

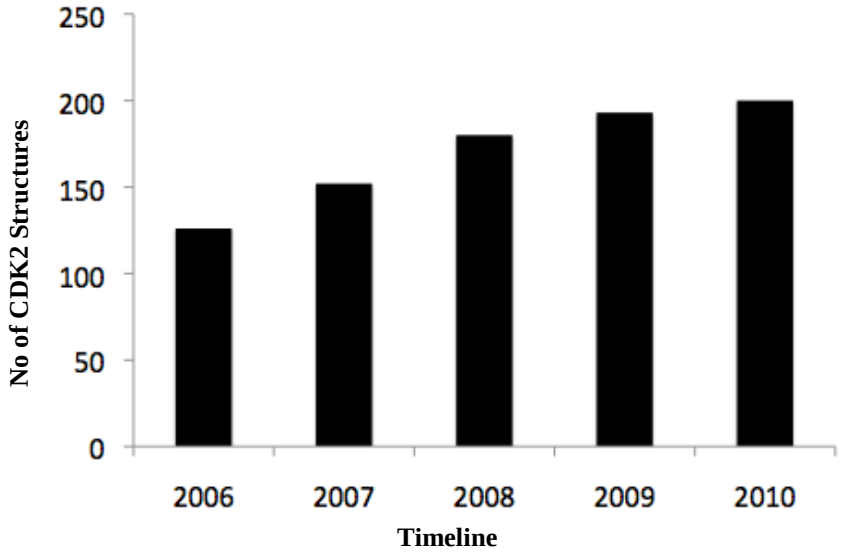


Figure 4-1 Growth of CDK2 structures during last five years. Over past five years (2006-November 28, 2010) total number of CDK2 structures has grown from 126 to 200

There are only eight structures of CDK6 available in PDB which are solved at relatively low X-ray resolution and with some missing regions (see Appendix 1.4). The quality of a homology model is limited by the quality of the chosen templates (Baker and Sali, 2001). A huge number of potential templates allowed considering additional criteria for template selection. Generally, template selection should take into account the following points in addition to the sequence homology. A good template for CDK4 models to be used in ligand docking studies should have been crystallized in the active form, be complexed with an inhibitor, have a high resolution of the structure and have a low R_{free} value. Based on these criteria, CDK2 crystal structure (PDB ID 2CCH) solved in an active form at 1.7Å resolution was selected to get a representative CDK4 model.

4.3 Target template alignment

BLAST searches were performed with CDK2, CDK4 and CDK6 against the combined UniProtKB/Swiss-Prot and UniProtKB/TrEMBL databases. From these BLAST results 25 sequences were selected for CDK2 similarity with an E-value cut off e-130 and score 467 and 22 sequences were selected for CDK6/CDK4 with an E-value cut off of e-112 and score 407. Multiple sequence alignments for CDK2 and CDK6/CDK4 were performed with MUSCLE and these were saved as profile alignment for CDK2 and CDK6/CDK4, respectively. CLUSTAL-W was used for the "Profile vs. Profile" alignment between CDK2 and CDK6/CDK4. These results indicate a high degree of similarity and conservation between these CDKs. The results of multiple sequence alignment are shown in Figure 4-2 which shows three sequences for the human CDK2 (P24941), CDK4 (P11802), and CDK6 (Q00534) taken from the 47 sequences Profile vs. Profile alignment. All other sequences were removed for this figure for the sake of clearity.

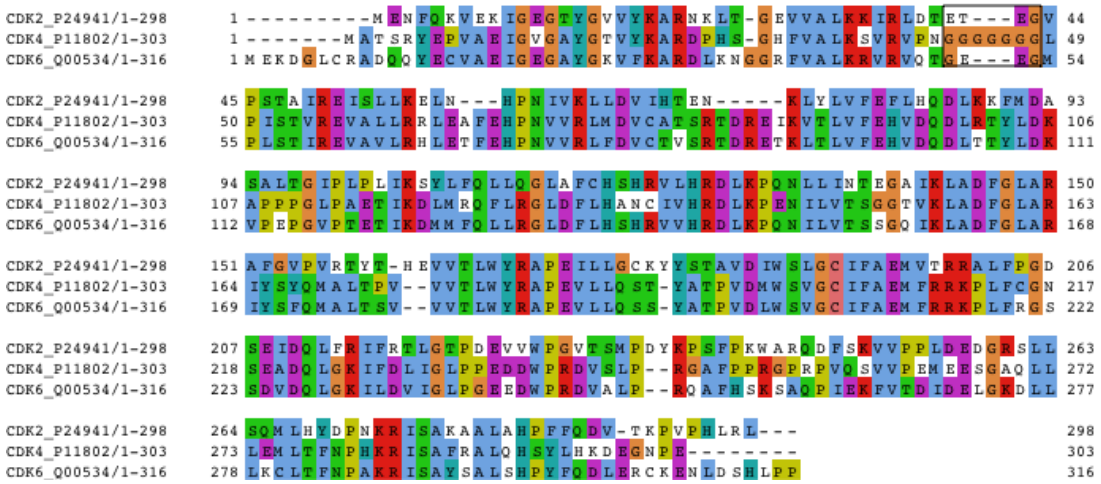


Figure 4-2: Multiple Sequence Alignment between CDK2, CDK6 and CDK4. The alignment shown is a part of a 47 sequences alignment obtained by the profile vs profile alignment using MUSCLE, CLUSTAL W and manual editing in Jalview. Only the human CDK2, CDK4 and CDK6 sequences are shown.

4.4 Validation of Modelling Strategy: CDK6 Model Based on CDK2

The experiment described in this section was performed to verify the modelling strategy for CDK4 homology modelling. To estimate the accuracy of the modelling procedure for a particular protein family it is good idea to generate control models. An already known structure of CDK6 was modelled with MODELLER (Narayanan Eswar, 2006; Eswar *et al.*, 2007) using CDK2 as a template, but without using any information from CDK6 X-rays structure. The overlay of the CDK6 model based on CDK2 and the experimentally determined X-ray structure of CDK6 (Figure 4-3) shows a high similarity between these two. The structural alignment and overlay in PyMOL (DeLano, 2002) gets an RMSD value of 1.26 Å between the experimentally determined CDK6 structure and the CDK6 model based on the CDK2 template. The CDK6 model was analysed by ProSa2003 (Wiederstein and Sippl, 2007) and WHAT_CHECK (Hooft *et al.*, 1996) validations.

Figure 4-3: A cartoon representation of the Model of CDK6 (shown in yellow) based on CDK2 template (PDB ID 2CCH). The model was built using the MODELLER. An overlay of the CDK6 X-ray structure (PDB ID 1XO2 (Lu *et al.*, 2005) shown in red)

The Z-Score (Figure 4-4) for the CDK6 model obtained from ProSa2003 is -7.08 compared with -7.57 for the experimentally determined structure of CDK6 (PDB ID 1XO2 (Lu *et al.*, 2005)). The Z-Score (explained in Section 2.10.2) for the CDK6 model is shown to be well in range for structures of a similar size in the protein database. No potential problems in CDK6 model were found with WHAT_CHECK analysis. The generation and validation of CDK6 suggests that a reasonably correct CDK4 model can be achieved via the homology modelling route.

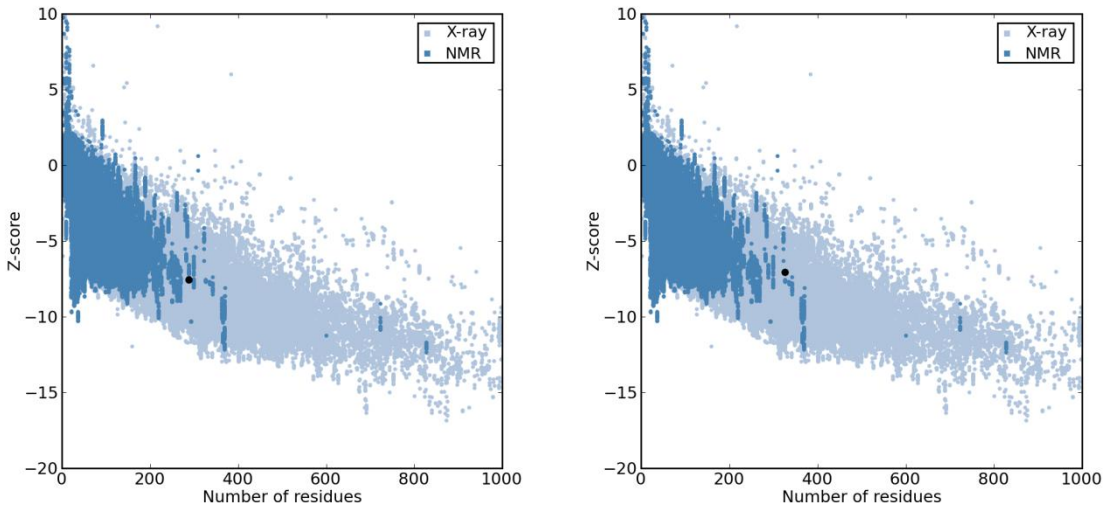


Figure 4-4: A graphical representation of the ProSa2003 score for CDK6 model (right) and the CDK 6 (PDB ID 1XO2 (Lu *et al.*, 2005)) X-ray structure (left). Each dot in the above graph represents a PDB structure. The dots shown in dark blue colour represent PDB entries by NMR and light blue dots indicate the X-Ray results. The X-axis represents the number of protein residues and the Y-axis represents the Z-Score

4.5 CDK4 Model Based on CDK2

Three dimensional models of CDK4 with MODELLER (Narayanan Eswar, 2006; Eswar *et al.*, 2007) were obtained by optimally satisfying spatial restraints derived from the alignment and expressed as probability density functions (pdfs) for the features restrained (Sali and Blundell, 1993). 50 models of CDK4 based on CDK2 (PDB ID 2CCH) template were generated. The models generated contain all main chain and side chain non-hydrogen atoms. Each model generated is also assigned with an energy score known as modeller objective function (molpdf). The molpdf score is a sum of all the restraint functions and has no units. The molpdf score can be used for model assessment. The best model was selected by picking the model with lowest value of the molpdf (which in this case is 1604.69). The values obtained for the modeller objective function does not provide an absolute evaluation of the model; it is only used to rank the different models generated by MODELLER from the same alignment (Sali and Blundell, 1993; Fiser and Sali, 2003; Eswar *et al.*, 2007).

The CDK4 homology model based on the CDK2 template is shown in Figure 4-5. As expected, the CDK4 model has two lobes similar to CDK2 with an N-terminal lobe consisting of antiparallel β–sheets (shown in yellow) and a C-terminal lobe consisting of α-helices (shown in red). The conserved PISTVRE region of CDK4 is shown in blue.

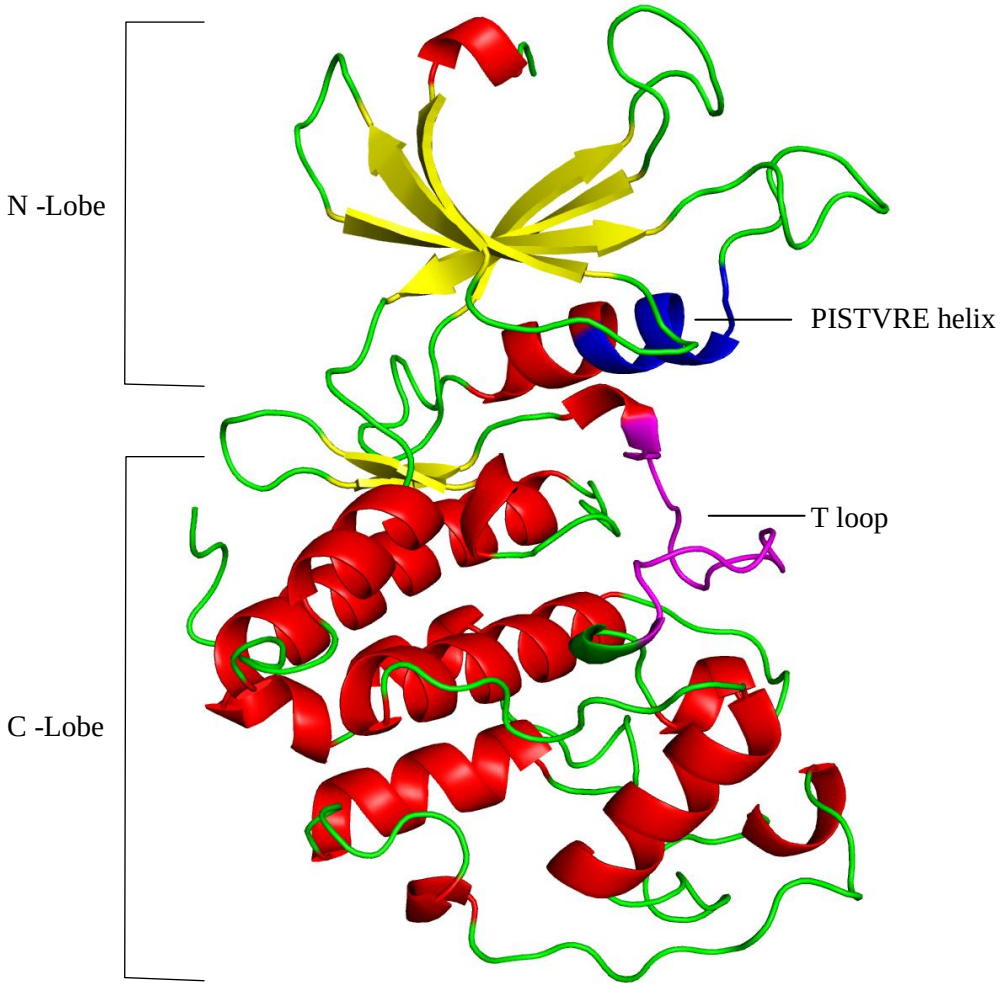


Figure 4-5: A cartoon representation of the CDK4 model based on a CDK2 template (PDB ID 2CCH) The underlying sequence alignment between CDK4 and CDK2 is shown in Figure 4-2. CDK4 N terminal lobe consisting of antiparallel β–Sheets is shown in yellow and a C-terminal comprising α-helices is shown in red. The activation loop or T-Loop is shown in magenta.

4.6 Evaluation of CDK4 Model

4.6.1 Modeller built in checks

The selected model of CDK4 was subjected to validity analysis by Modeller built in checks before any external evaluation. MODELLER build in checks evaluate an input model with the DOPE (Discrete Optimized Protein Energy) potential. The DOPE score profile per residues of the model (CDK4) and template (CDK2) obtained from MODELLER is shown in Figure 4-6. A comparison of the DOPE scores provides an idea about the quality of the input alignment for homology modelling. Gaps in the plot can be seen corresponding to the gaps in the target-template alignment. DOPE scores do not represent an absolute measure for a comparison of model and template quality. The DOPE is a distance dependent statistical potential based on probabilistic theory (Shen and Sali, 2006). The extended peak above the zero line in region 42–48 of the CDK4 model highlights that they may be some error in the raw model. This region 42–48 corresponds to the glycine rich loop of CDK4 for which there is no structural information available from the template. A comparison of different models revealed that this region is highly variable in different models.

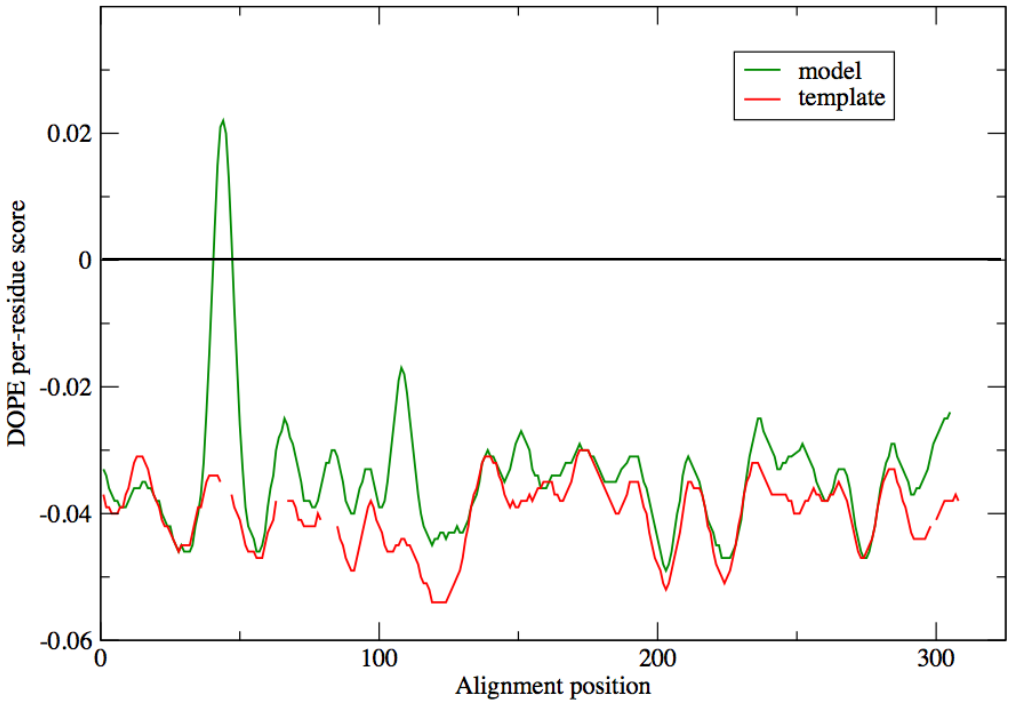


Figure 4-6 DOPE score profile of the CDK4 model and template obtained from MODELLER. The DOPE profile of template and model is shown in red and green, respectively. The model and template were evaluted using standard *evaluate_model.py* and *evalute_template.py* MODELLER scripts

4.6.2 Validation of CDK4 Model by ProSa2003 and WHAT_CHECK

The CDK4 model was validated with ProSa2003 (Wiederstein and Sippl, 2007). The overall quality of the CDK4 model was estimated by calculating its ProSa2003 energy Z-score and comparing it to template Z-score. The CDK4 model has a Z-Score -7.37 compared with -7.12 for CDK2 (PDB ID 2CCH). The Z-Score for the CDK4 model is very similar to the Z-Score of the template and is shown to be well in range for structures of similar size in the protein database (Figure 4-7). The CDK4 model was also evaluated for correctness of the overall fold/structure, stereochemical parameters such as bond lengths, angles and dihedral using WHAT_CHECK (Hooft *et al.*, 1996). No potential problems in the CDK4 model were found in the WHAT_CHECK analysis.

The backbone conformation analysis gives a score (Z-Score 4.247) that is normal for well refined protein structures.

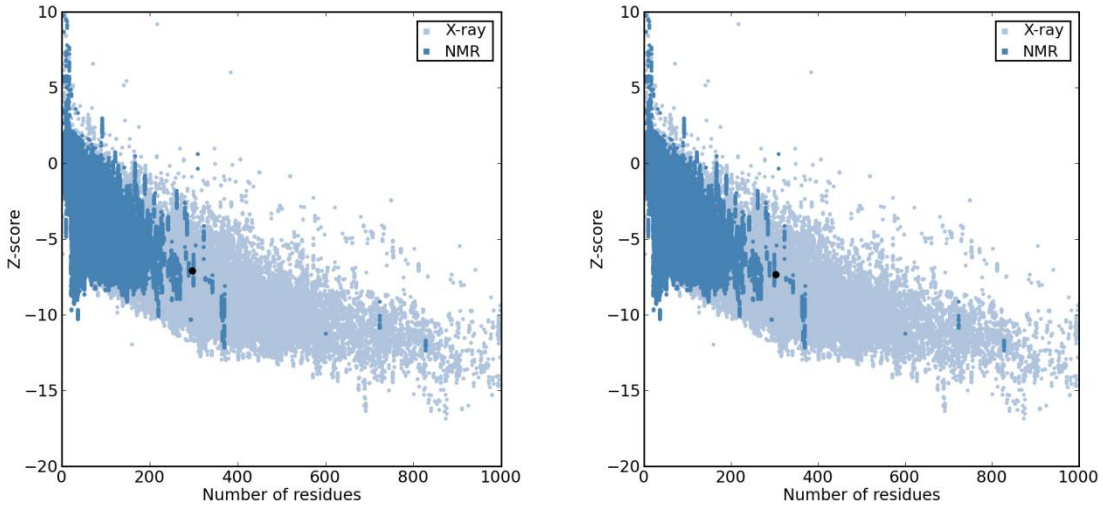
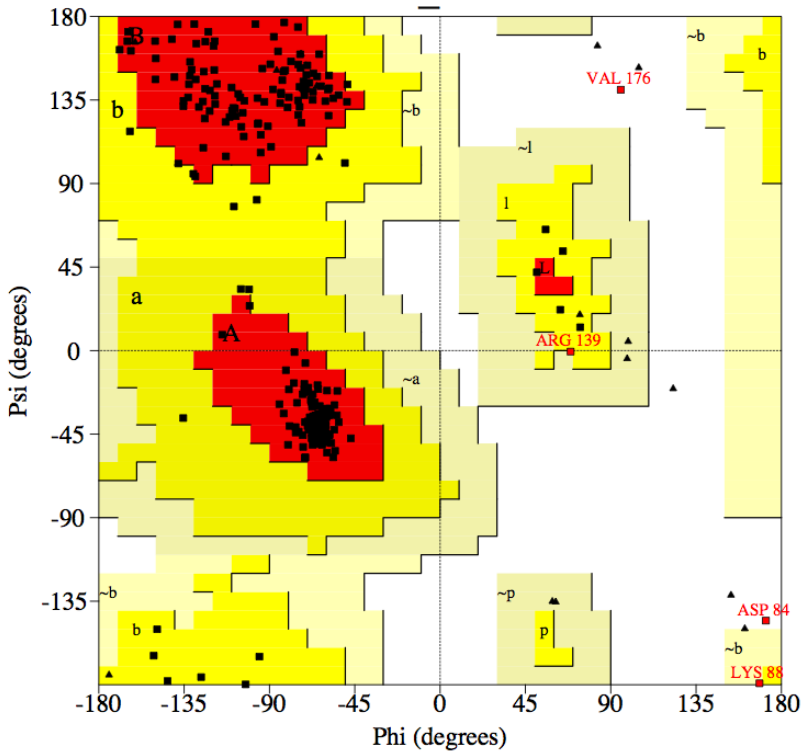


Figure 4-7 : A graphical representation of the ProSa2003 output for CDK4 (right) model and the CDK 2 (left) template. Each dot in the above graph represents a PDB structure. The dots shown in dark blue colour represent PDB entries by NMR and light blue dots indicate the X-Ray results. The one black dot in the graph represent CDK4 model (right) and CDK2 template (left) suggesting the validity of the model as compared to the all PDB structures. The X-axis represents the number of protein residues and the Y-axis represents the Z-Score

4.6.3 Quality assessment of CDK4 model with Ramachandran plot

A Ramachandran plot obtained with PROCHECK for the CDK4 model indicates that 91.3% of residues are in the most favoured regions A, B and L (Figure 4-8). Only 2 residues Val176 and Asp84 are in disallowed regions. All other residues are in additional and generously allowed regions. A good quality model is expected to have over 90% in the most favoured regions (Laskowski *et al.*, 1993). This analysis along with ProSa2003 and WHAT_CHECK analysis (4.6.2) indicates that CDK4 model is of good quality and can be used for docking and other structural studies.



Plot statistics

Residues in most favoured regions [A,B,L]	230	91.3%
Residues in additional allowed regions [a,b,l,p]	18	7.1%
Residues in generously allowed regions [~a,~b,~l,~p]	2	0.8%
Residues in disallowed regions	2	0.8%
	----	----
Number of non-glycine and non-proline residues	252	100%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	24	
Number of proline residues	25	

Total number of residues	303	

Figure 4-8: Ramachandran plot for the CDK4 Model. Plot statistics with PROCHECK indicate 91.3% residues are in the most favoured regions A, B and L. There are 7.1% residues in additional and 0.8% in generously allowed regions. Only 2 residues (0.08%) Asp84 and Val176 are in disallowed regions. A good quality model is expected to have over 90% in the most favoured regions (Laskowski *et al.*, 1993).

4.7 Comparison of CDK4 Model with CDK4 X-ray Structure

The X-ray structures of CDK4/cyclin D3 and CDK4/cyclin D1 were made available by two research groups in 2009 (Day *et al.*, 2009; Takaki *et al.*, 2009). The original attempts by Day *et al* to crystallize phosphorylated CDK4 co-expressed with Cyclin D in an active form were not successful (Day *et al.*, 2009). To facilitate crystallization they introduced some modifications into the CDK4 and corresponding

Cyclin D1 sequences. These modifications include the replacement of the distinctive glycine rich loop (residues 42-48) of CDK4 with the corresponding GEEG sequence from CDK6 (Day *et al.*, 2009). Takaki *et al* (Takaki *et al.*, 2009) reported a non-phosphorylated CDK4/cyclin D3 complex (PDB ID 3G33) at a resolution of 3.0Å. The availability of X-Ray structures (see Appendix 1.2) provided an excellent opportunity to test the quality of the homology modelling work described earlier (see Section 4.5). A structural overlay between the homology model and X-ray structures (Figure 4-9) revealed overall similarity of secondary structural folds.

.

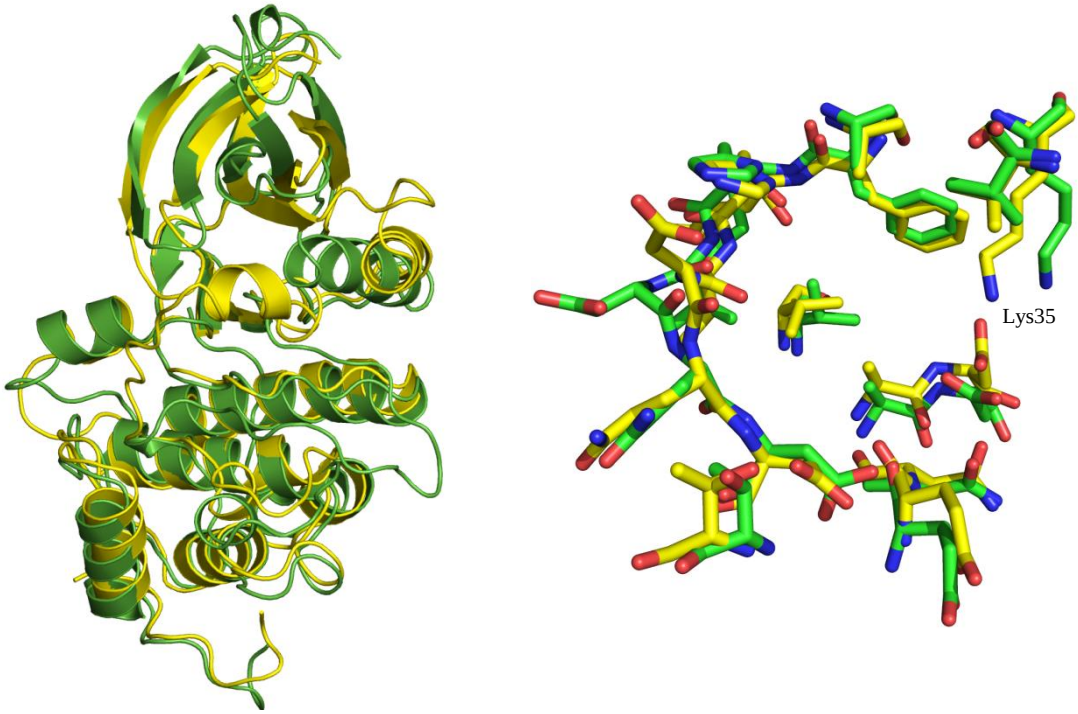


Figure 4-9: a) A cartoon representation of the CDK4 model (green) based on CDK2 template (PDB ID 2CCH) compared with PDB structure (yellow) 2W96 (published on March 12, 2009). a) The crystal structure of CDK2 is in inactive conformation compared to the active confirmation of CDK4 model. **b)** An active site overlay of the CDK4 model (green) based on CDK2 template (PDB ID 2CCH) compared with PDB structure 2W96 (yellow)

Some differences found in the homology model and the X-ray structures are due to the fact that the X-ray structures are not obtained in an active form while the CDK4 model was built on a template crystallized in the active form. An overlay of active sites between the homology model and X-ray structure (Figure 4-9) also shows a different orientation of Lys33. The T-loop and C-α helix of CDK4 structures resembles the inactive structures of CDK2. RMSD values of 1.88 Å, 1.92 Å, 1.97 Å, 1.92 Å and 1.64 Å are obtained for the Cα atoms between the CDK4 model and CDK4 PDB structures 2W96, 2W99, 2W9F, 2W9Z and 3G33, respectively (Table 4-1). A structural overlay of all CDK4 structures (Figure 4-10) revealed a RMSD value ranging from 0.56 Å to 1.09 Å between all CDK4 PDB structures. The high RMSD values between CDK4 homology model and PDB structures is not surprising as a higher RMSD value (e.g. 1.49 Å between 2CCH and 2R3I) are also found between the structures of the active and inactive forms of CDK2.

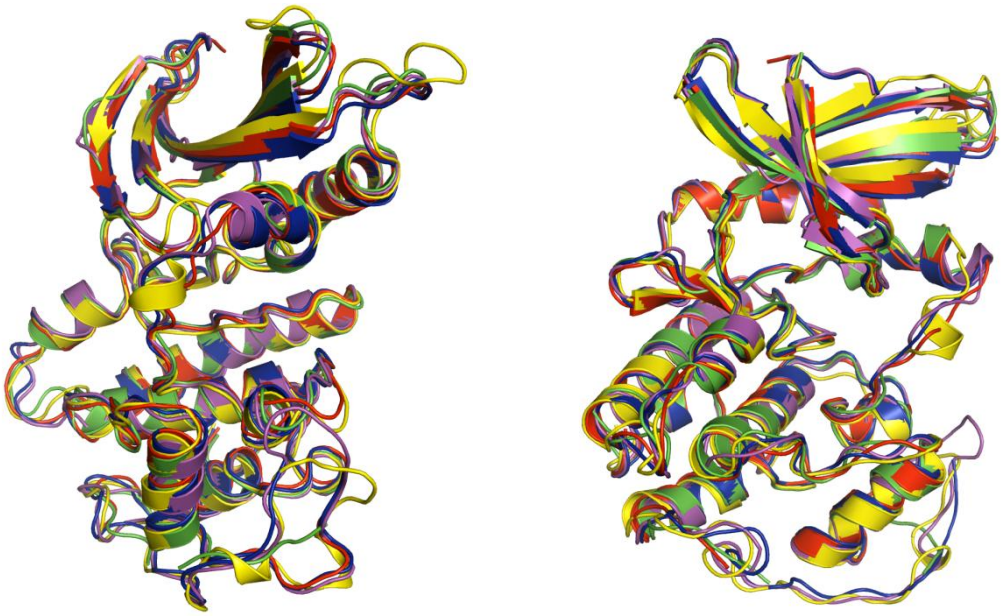


Figure 4-10: An overlay of all the available PDB structures of CDK4. CDK4 structures overlay is shown in two different orientations. All of these PDB structures reprsent an inactive confirmation of the CDK4.

RMSD calculations between active sites of CDK4 model and PDB structures also show a higher value compared to the RMSD comparison between the PDB structures alone (Table 4-1). The RMSD between 3G33 and 2W99 active site is found relatively higher with a value of 0.83Å, This is due to some backbone variations in residues Phe93, Gln98, Asp99 and Glu144.

Table 4-1: RMSD matrix of CDK4 model and X-ray structures. RMSD values are obtained for the Cα atoms between the CDK4 model and CDK4 PDB structures (shown in orange) and their active sites (shown in blue).

		RMSD[a] MATRIX between CDK4 Model and PDB structures					
		CDK4	**2W96**	**2W99**	**2W9F**	**2W9Z**	**3G33**
RMSD[a] between Active Site Residues	**CDK4**	0.00	1.88	1.92	1.97	1.92	1.64
	2W96	0.94	0.00	0.76	0.77	0.56	0.86
	2W99	1.06	0.64	0.00	0.70	0.66	1.04
	2W9F	1.2	0.65	0.63	0.00	0.80	0.92
	2W9Z	1.18	0.53	0.51	0.55	0.00	0.94
	3G33	1.33	0.59	0.83	0.76	0.66	0.00

As all of the CDK4 PDB structures are in an inactive confirmation, and some of these do not correspond to the wild type CDK4 there is still a need for good quality CDK4 model in an active confirmation.

4.8 CDK4 Model based on CDK4 X-Ray Structure

After the availability of CDK4 structures in the inactive conformation the modelling strategy was updated to incorporate information from CDK4 X-ray structures into a new CDK4 model which should be in an active conformation. According to the new strategy CDK4 modelling was carried out based on two templates i.e. CDK2 PDB-

ID 2CCH and CDK4 PDB-ID 2W96. The newly constructed models based on two templates carries all the information from the active form of CDK2 for the regions involved in activation and inactivation and all other regions from CDK4 X-ray structure. 20 models were generated and best model was selected based on molpdf score. This model based on two templates may be a good representative of CDK4 structure in an active form.

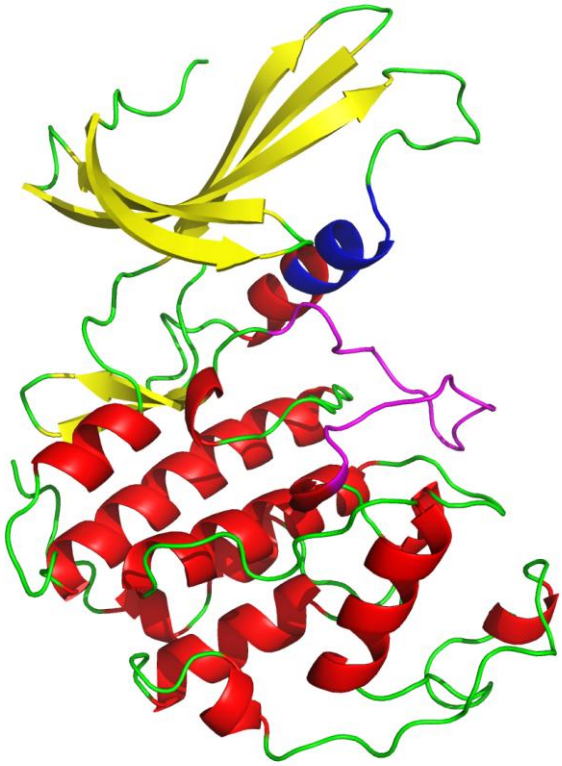


Figure 4-11 : CDK4 model based on CDK2 and CDK4 templates. PDB structure 2CCH for CDK2 and 2W96 for CDK4 was used to generate this model to generate an active confirmation of CDK4.

The CDK4 model based on CDK2 and CDK4 template was also subjected to model evaluation. The ProSa2003 Z-score for CDK4 model is found -7.84 compared to -7.96 for CDK4 (PDB ID 2W96) and -7.12 for CDK2 (PDB ID 2CCH). CDK4 model evaluation with PROCHECK and WHAT_CHECK tools also ensured the quality of the CDK4 model based on CDK2/CDK4 templates.

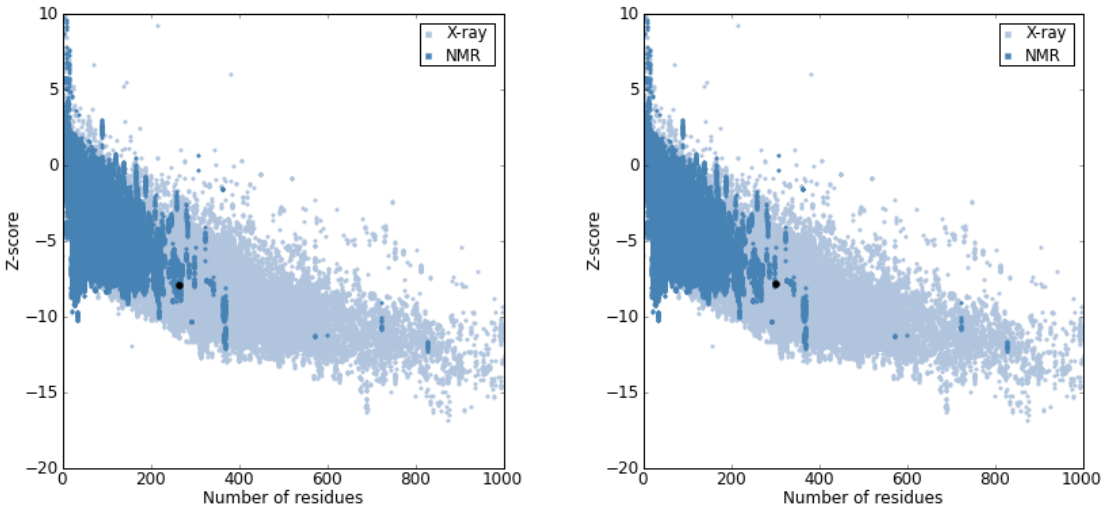


Figure 4-12 A graphical representation of the ProSa2003 Z-score for CDK4 model (right) and CDK4 PDB 2W96 (left). The ProSa Z-score for 2nd template CDK2 PDB ID 2CCH is shown in Figure 4-7

The quality assessment of CDK4 model ensures that it can be used to perform molecular docking and molecular dynamics studies (Chapters 5 & 6). The homology modeling technique provides a useful approach to bridge the so called "sequence–structure gap" until complete experimental structures of pharmacologically important proteins are available. The usefulness of CDK4 homology models in molecular docking and structure based drug design has been reported by different research groups (Aubry *et al.*, 2004; Hillisch *et al.*, 2004; McInnes *et al.*, 2004; Aubry *et al.*, 2006).

4.9 Conclusion

Structural information for a protein is essential to understand its interactions with different inhibitors. In the absence of an experimental structure of CDK4 homology models were generated with the closely related CDK2 as a template. The choice of sequence alignment strategy plays a critical role in generating accurate homology models (Martin *et al.*, 1997; Forrest *et al.*, 2006). In the present work "Profile vs. Profile" alignment between CDK2 and CDK6/CDK4 is used.

An important aspect of homology modelling experiments is assessment of the stereochemical quality and reliability of the model. The selected model was validated with different validation tools such as ProSa2003 (Wiederstein and Sippl, 2007), PROCHECK (Laskowski *et al.*, 1993) and WHAT_CHECK (Hooft *et al.*, 1996). After the availability of CDK4 X-ray structures (Day *et al.*, 2009; Takaki *et al.*, 2009) in an inactive form, a new CDK4 model was build in a putative active form by incorporating the structural information both from CDK4 and active CDK2. This CDK4 model based on two templates (CDK4 and active CDK2) is used for molecular docking and molecular dynamics studies in the following chapters.

Chapter Five

Molecular docking and structure based design of CDK4 inhibitors

Chapter 5 Molecular docking and structure based design of CDK4 inhibitors

5.1 Introduction

Molecular docking of inhibitors with crystal structures or homology models is used as a tool in drug discovery research and in the optimization of existing drugs (Brooijmans and Kuntz, 2003; Cavasotto and Orry, 2007). The molecular docking problem can be divided into two parts. The first one is the prediction of the accurate pose of the ligand in the active site of the receptor. The second is the accurate calculation of binding affinity or a score representing the strength of ligand binding. An essential requirement for a useful application of docking methods is their ability to find the correct pose of a ligand.

Docking software is usually trained and calibrated with ligand complexes from the Protein Data Bank (PDB). The performance of a docking program in predicting the accurate pose is measured by its ability to reproduce the experimentally solved ligand-binding modes of a protein ligand complex. This type of validation is usually carried out with the native protein conformers. In the present work the GOLD docking suite (Verdonk *et al.*, 2003) was selected for the docking studies of CDK2 and CDK4. GOLD had previously been tested using different data sets of kinases (Hartshorn *et al.*, 2007; Verdonk *et al.*, 2008). In this work CDK2 docking experiments were carried out both with the native (Section 5.2.1) and non native conformer (Section 5.2.2) of CDK2 in order to find the propensity of GOLD docking algorithm to reproduce the X-ray pose of ligand-protein binding for different small molecule inhibitors of CDK2.

Molecular docking of fascaplysin and its derivatives into CDK2 and CDK4 was carried out in an attempt to explain the specificity of CDK4 for fascaplysin and its derivatives compared with CDK2 (Section 5.3). To design highly selective ligands computational techniques are very helpful. Based on the docking results of tryptamine

based inhibitors of CDK4 (Section 5.4) new compounds with better predictive binding properties for CDK4 were rationally designed (Section 5.5) and some of these compounds have been synthesized by Dr. Paul Jenkins group in the Department of Chemistry, University of Leicester and determination of their binding affinities and specificity is in progress.

5.2 Testing Gold performance on CDK2

The performance of GOLD was tested for CDK2 using the native conformer docking (self docking approach) and non-native conformers docking (cross docking approach). In the native conformer docking each ligand is docked back into the CDK2 conformation from the structure that contained that ligand. In the cross docking approach the ligand obtained from a particular CDK2/inhibitor structure is used to dock against other CDK2 structures.

5.2.1 Native conformer docking

A set of 21 CDK2 protein-ligand complexes from the Protein Data Bank (PDB) (Berman *et al.*, 2000) with a resolution 1.8 Å or better was selected for the GOLD self docking performance test. The chosen set (Table 5-1) represents the highest quality ligand-protein complexes of CDK2 for which electron density maps are also available. The electron density map for the inhibitors allows a comparison between experimental structure and predicted pose based on experimental data. All the ligands from the CDK2 complexes were prepared and re-docked into the corresponding protein structure using GOLD as described in Section 2.12. For each chosen PDB structure of CDK2 two sets of GOLD docking poses were generated using the GoldScore and the ChemScore scoring functions. The results of the self-docking approach of CDK2 are summarized in Table 5-1.

Table 5-1. Molecular docking of CDK2. Twenty one highest quality PDB structures of CDK2 were subjected to a self docking experiment using GOLD. ChemScore and GoldScore were used for ranking and RMSD between the reference pose and the GOLD top rank and best solution is calculated.

PDB ID	Ligand ID*	X-ray Resolution (Å)	GoldScore	ChemScore	rRMSD (Å)		bRMSD (Å)	
					GoldScore	ChemScore	GoldScore	ChemScore
2R3I	SCF	1.28	63.16	30.69	1.4	0.6	0.6	0.6
1GZ8	MBP	1.30	48.54	19.01	1.3	1.2	1.3	1.2
2R3Q	5SC	1.35	77.36	34.44	2.0	1.4	0.4	1.4
2R3R	6SC	1.47	61.50	33.47	1.2	0.9	0.9	0.8
2R3H	SCE	1.50	47.82	28.34	0.6	2.9	0.6	2.6
2R3F	SC8	1.50	62.51	32.91	2.2	0.7	2.2	0.7
1JVP	LIG	1.53	67.35	42.62	1.5	0.8	1.4	0.8
2R3G	SC9	1.55	59.04	32.82	4.6	1.2	4.3	1.0
1H00	FAP	1.60	63.82	24.66	3.9	3.5	3.9	3.0
1OIT	HDT	1.60	63.03	35.62	1.1	1.5	1.0	1.1
1URW	I1P	1.60	69.21	38.21	1.7	2.5	1.5	1.7
2R3N	SCZ	1.63	70.59	38.79	1.1	0.8	1.1	0.6
2R3J	SCJ	1.65	59.21	33.78	1.6	1.2	1.6	1.2
2R3L	SCW	1.65	59.59	35.77	1.2	0.8	1.2	0.8
2R3P	3SC	1.66	67.33	31.79	1.9	1.1	1.9	1.1
2R3K	SCQ	1.70	56.67	34.90	1.0	0.8	0.8	0.5
2R3M	SCX	1.70	66.41	28.83	1.7	0.9	0.7	0.6
1H08	BWP	1.80	69.23	36.15	2.1	1.1	2.1	1.1
2C6I	DT1	1.80	70.88	26.93	1.0	0.7	1.0	0.7
2CLX	F18	1.80	45.44	22.04	2.5	4.5	2.1	3.4
2R3O	2SC	1.80	64.37	35.44	1.5	1.4	1.5	1.1

rRMSD = RMSD between the first ranked solution as obtained by GOLD and the experimental confirmation of the ligand.
bRMSD = RMSD between the best pose of the docking solution and the experimental confirmation of the ligand
* The abbreviated names used for ligands are as defined in the Appendix 1.8.

The ability of the GOLD docking program to reproduce the binding mode of a ligand in its corresponding crystal structure is measured by calculation of the root-mean-square distance (RMSD) between the non-hydrogen atoms of the ligand in the

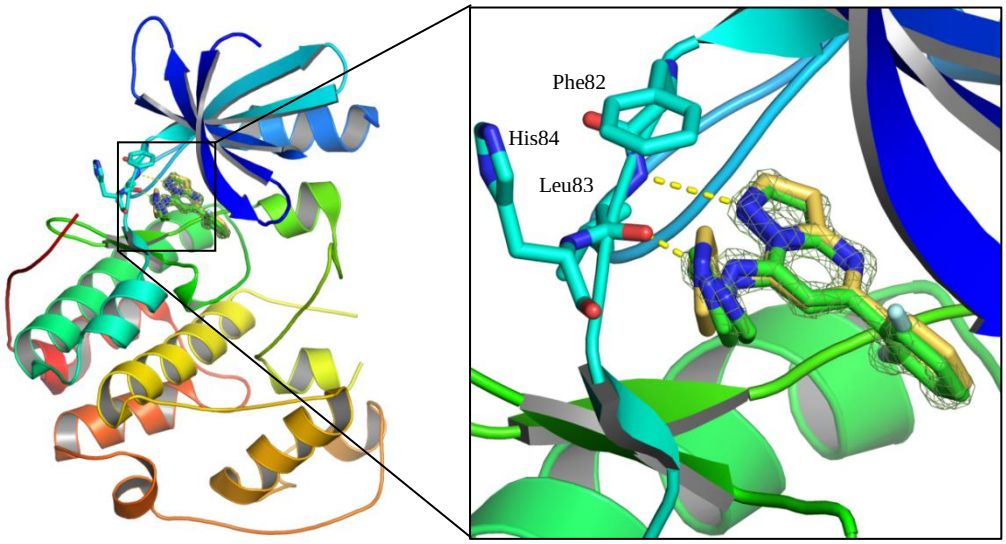
crystal structure and the corresponding atoms in the docked poses. A RMSD threshold value of 2.0 Å is widely accepted as a indicator for successful docking in such experiments (Verdonk *et al.*, 2008). The RMSD between top-ranked poses as obtained by GOLD and best pose of the docking solutions based on lowest value of RMSD against the experimental confirmation of the ligand in the following will be referred as rRMSD and bRMSD, respectively.

In the cases studied the GOLD docking success rate is 76% with top ranked solutions within 2.0 Å rRMSD of the experimental binding mode using GoldScore and 81% with ChemScore. The docking performance reported in the literature lies between 70-80% for the native conformers docking (Friesner *et al.*, 2006; Hartshorn *et al.*, 2007; Verdonk *et al.*, 2008). Verdonk *et al* have reported a GOLD docking performance of ~80% with native conformer docking (Verdonk *et al.*, 2008). The results of the present study of CDK2 native conformer docking are in line with Verdonk *et al.* results. ChemScore slightly outperforms the GoldScore; therefore ChemScore was selected in all the follow up docking experiments presented in this Chapter. A selected set of CDK2-self docking examples is discussed in details in the following section to highlight strengths and weaknesses of CDK2 docking experiments.

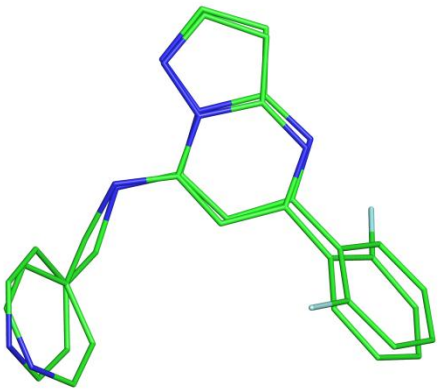
5.2.1.1 2R3I — an example of successful ligand docking

The CDK2 self docking (PDB: 2R3I) with the small molecule inhibitor SCF (5(2-fluorophenyl)-N-(pyridin-4-ylmethyl)pyrazolo[1,5-a]pyrimidin-7-amine) successfully reproduced the experimental binding pose of the ligand as shown in Figure 5-1. The PDB coordinates for the inhibitors (SCF) show two possible binding poses of the inhibitor with ~180^0 of rotation for the fluorophenyl ring. As typical for CDK2 inhibitors SCF adopts hydrogen bonds from the backbone NH of Leu83 and N1 of the ligand and between the carbonyl group of Leu83 and HN (at the 7-position) of the

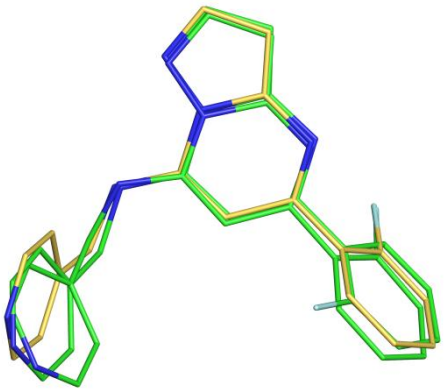
ligand (Figure 5-1). The RMSD difference between the two native poses of ligands is 1.4 Å. The rRMSD difference between the heavy atoms of docking pose and the two native poses is 0.6 Å and 1.4 Å. Ligand docking reproduces experimental structure very well and finds all the relevant polar interactions. In this example re-docking works very well.



a)



b)



c)

Figure 5-1 **Molecular docking of CDK2 PDB ID 2R3I. a)** A cartoon representation of CDK2 (PDB ID 2R3I) complexed with the inhibitor SCF (5-(2-fluorophenyl)-N-(pyridin-4-ylmethyl)pyrazolo[1,5-a]pyrimidin-7-amine) covered in mesh representation of the electron density map. The ligand PDB coordinates are shown in green, the docked ligand is shown in yellow. The data for the electron density map are obtained from the Uppsala electron density server (Kleywegt *et al.,* 2004). **b)** A line representation of the native poses of SCF. The RMSD between these poses is 1.4 Å, c) Overlay of experimental coordinates of SCF with predicted coordinates. The RMSD between the heavy atoms of docked solution and the two native poses is 0.6 Å, and 1.4 Å, respectively.

5.2.1.2 1GZ8 & 2R3R — more examples of successful docking

The docking of 1GZ8 with inhibitor MBP (1-[(2-amino-6,9-dihydro-1h-purin-6-yl)oxy]- 3-methyl-2-butanol) and 2R3R with inhibitor 6SC (3-bromo-5-phenyl-N-(pyridin-4-ylmethyl)pyrazolo[1,5-a]pyrimidin-7-amine) are also examples for successful redocking with a rRMSD 1.2 and 0.9 Å, respectively. Similar to the example described earlier (2R3I) the docking solutions occupy the electron density map very similar to the experimental pose of the ligand and reproduce the polar contacts between the ligands and the receptor (Figure 5-2).

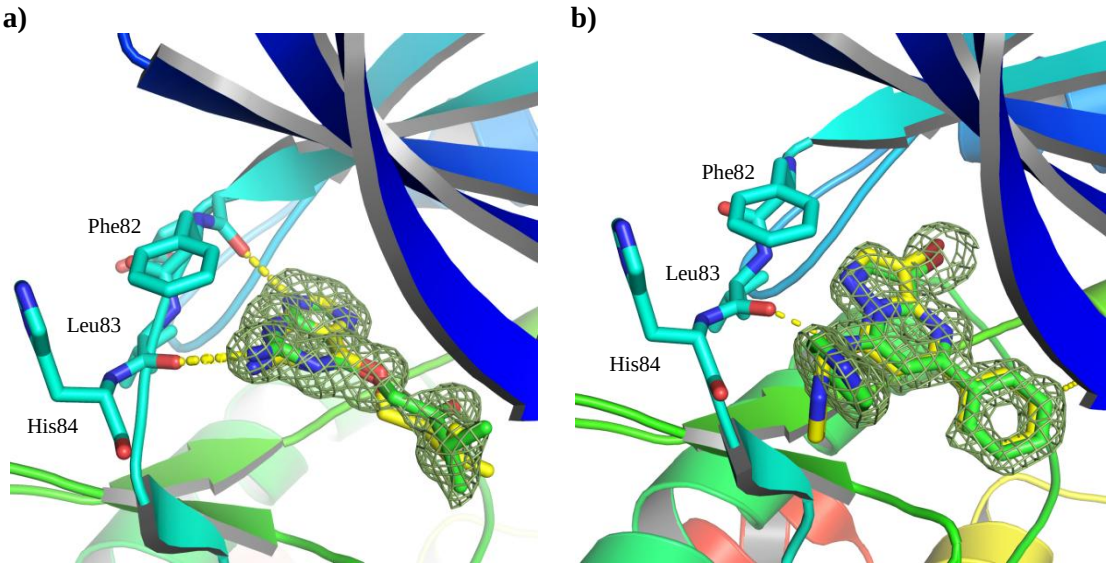


Figure 5-2 Molecular docking of native conformer of 1GZ8 and 2R3R a) The PDB native coordinates of 1GZ8 ligand binding mode (green) overlayed with the docked solution (yellow) covered in electron density map **b)** Overlay of protein ligand docking (yellow) of 2R3R with the native confirmation of the ligand (green)

5.2.1.3 1URW — successful docking poses with RMSD > 2.0 Å

While a RMSD greater than 2.0 Å is usually considered as incorrect, there are some examples where a high RMSD value may be obtained even if essential binding features are present and docked pose is identified as essentially correct on manual inspection. Such cases arise if a moiety that is less relevant to the binding mode (e.g., a

solvent-exposed group) deviates substantially from the crystal structure and results in a higher RMSD as shown in the IURW docking results (Figure 5-3). A rRMSD of 4.5 Å is observed despite the main core of the ligand (n-[3-(dimethylamino)propyl]-4-[(4-imidazo[1,2-b]pyridazin-3-yl-2-pyrimidinyl)amino]benzenesulfonamide) in the docking pose occupying the same core of the electron density map as the experimentally solved conformation. Also the docking reproduced the experimentally known polar contacts between the ligand and the CDK2. The reason for the higher RMSD and the discrepancy between the docking and experimental pose of the ligand can be attributed to the substantial deviation of solvent exposed dimethylaminopropyl group of the ligand (Figure 5-3) and lack of clear electron density for this region. Also the B-factor value for the dimethylaminopropyl group of the ligand is higher than the B-factor of the main core of the ligand.

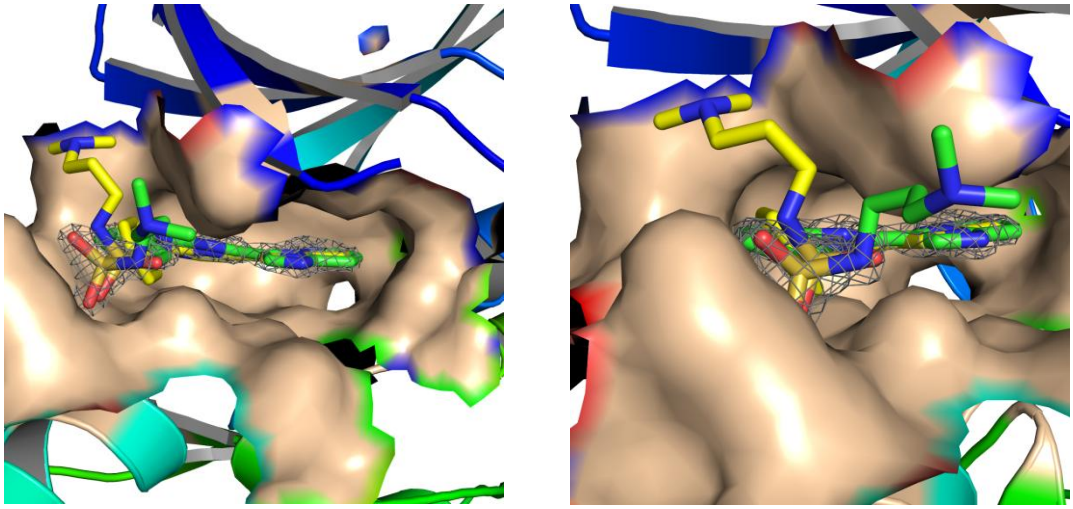


Figure 5-3 Molecular Docking of CDK2 PDB ID 1URW. The experimently solved coordinates of 1URW ligand binding mode (green) overlayed with the GOLD docked solution (yellow) covered in electron density map. The solvent-exposed group deviates substantially from the crystal structure and result in a higher RMSD value between the docking pose and the native cooerdinates.

Although in small molecule docking experiments a heavy atom RMSD 2.0 Å or smaller is consider as the main performance indicator, the above example illustrates that the RMSD approach has shortcomings in some cases which may lead to

misclassification of correct and incorrect poses. Kroemer *et al.* 2004 have reported examples where hydrogen bond interactions are not preserved in the docking solutions compared to the reference structure despite having low RMSD (Kroemer *et al.*, 2004). It is therefore important to take into account the essential interactions between ligands and proteins in addition to RMSD criteria in such experiments.

5.2.1.4 2CLX — potential problems in an experimental ligand pose

One has to ensure that benchmark sets for evaluation of docking are accurate. Occasionally even high resolution X-ray structures have problems with their ligand pose. The docking results for the CDK2 structure 2CLX show a higher RMSD compared to experimental pose than the performance threshold of 2.0 Å. The docking results with GoldScore suggest a pose where the inhibitor is placed further away from the Leu83 of CDK2 while maintaining a pose similar to the experimental pose (Figure 5-4b). The best docking pose obtained with ChemScore is oriented some what differently as shown in Figure 5-4c. PyMOL visualisation reveals unusually short interatomic distance of 1.98 Å between the ligand atom N17 and the carbonyl oxygen of Leu83 of the CDK2 in 2CLX. An analysis with the WHAT IF program (Vriend, 1990) indicated that this distance is 0.72 Å shorter than expected and represents an unlikely high energy pose. Also the occupancy of the ligand in the PDB is 0.70 with an alternative occupancy 0.30 for water molecules (residue identifiers 2250-2253) in place of ligand, this indicate some difficulties in placing the ligand into the electron density. Due to questionable reliability of the ligand pose in 2CLX this structure was omitted from the performance evaluation of GOLD CDK2 re-docking.

Further insight into how well a pose fits the electron density may be gained using the real space R-factor (RSR) criterion (Yusuf *et al.*, 2008). PDB structures with

potential problems are not a good choice in preparing a validation set for a docking program as these do not provide meaningful information for a comparison.

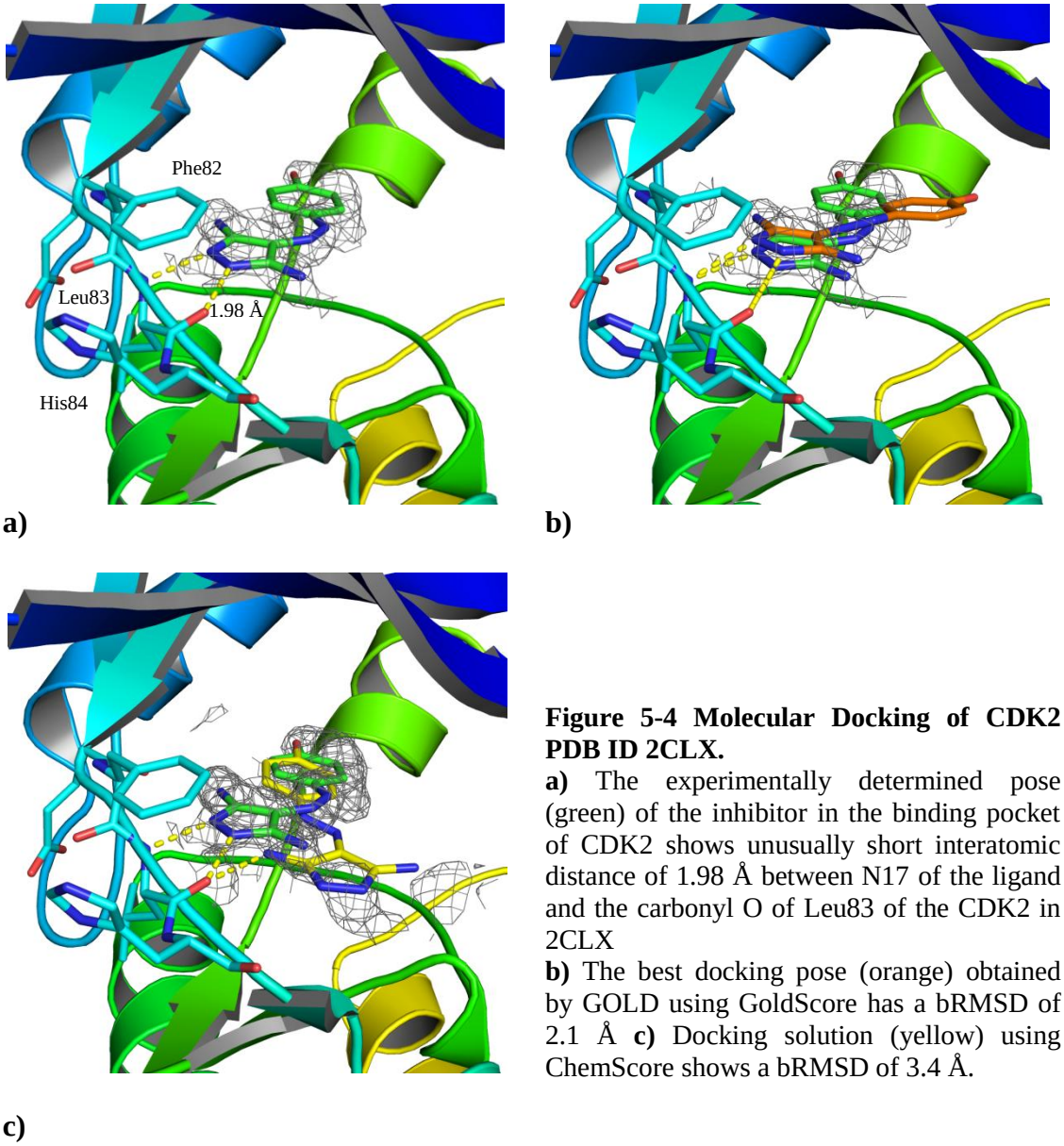


Figure 5-4 Molecular Docking of CDK2 PDB ID 2CLX.
a) The experimentally determined pose (green) of the inhibitor in the binding pocket of CDK2 shows unusually short interatomic distance of 1.98 Å between N17 of the ligand and the carbonyl O of Leu83 of the CDK2 in 2CLX
b) The best docking pose (orange) obtained by GOLD using GoldScore has a bRMSD of 2.1 Å **c)** Docking solution (yellow) using ChemScore shows a bRMSD of 3.4 Å.

The CDK2 docking results are very encouraging as in most of the cases docked ligand is occupying the same electron density as of the ligand in the experimentally known structure. The major objective of CDK2 docking study with GOLD was to gain

confidence for its later use for CDK4 and other CDKs. The success of ligand docking for CDK2 using GOLD assures that it may be used for CDK4 and other CDKs.

5.2.2 Non native conformer docking

The non-native conformer docking experiments involve the docking of fifty-eight ligands obtained from different CDK2 ligand-complexes with a cutoff 2.0 Å or better against one structure of CDK2. 2R3I was chosen for this experiment. The ligands selected from different CDK2 PDB structures were all prepared as described in Section 2.12. Out of fifty-eight ligands forty one ligands docked very well into 2R3I and reproduced all the essential binding features of the ligands with CDK2. One example of a ligand which does not dock into the 2R3I conformer is LS5 (1KE9). However, it docks very well into its own native conformer and reproduces the binding features typical of CDK2 as discussed earlier. A structural comparison of the 2R3I and 1KE9 reveals that both structures have the same conformation of the c-α helix and both belong to the inactive cluster (see Section 3.2.2); however the conformation for residue Lys33 in the active site is different in these two structures. The variations in the conformation of Lys33 in CDK2 (IKE9 and 2R3I) structures are illustrated in Figure 5-5. This issue of docking LS5 (1KE9) into 2R3I was addressed in the follow up experiment allowing the side-chain flexibility of Lys33 to occupy different rotamers. Lys33 of CDK2 is also involved in salt bridge making with Glu51 and Asp145 (Figure 3-3) and its orientational flexibility is highlighted in Figure 3-2. This amino acid is conserved in the extended family of CDKs and CDK like proteins (see Section 3.2.4).

This example illustrates that different ligands may induce protein conformational changes in CDK2 such as different orientations of Lys33. Therefore, it is important to consider flexibility of Lys33 for CDK2 docking experiments.

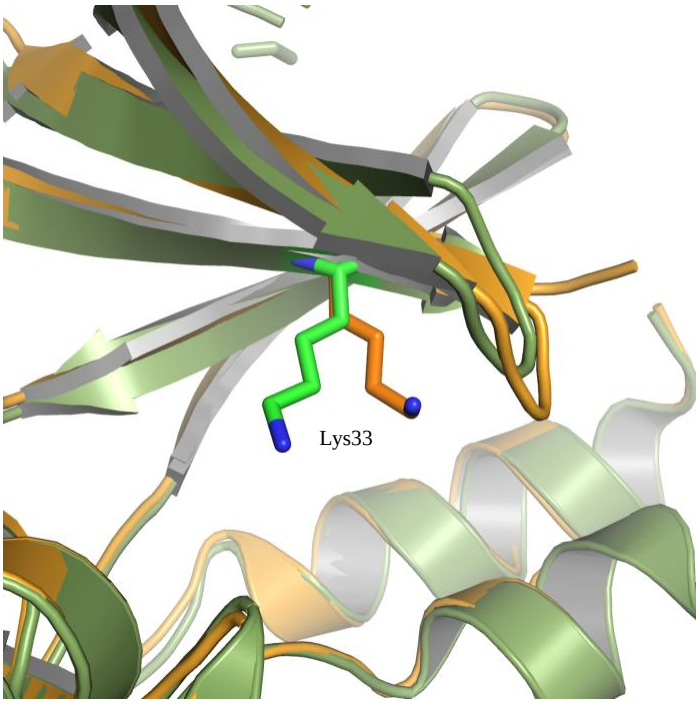


Figure 5-5. Different orientation of Lys33 in the CDK2 active site of 2R3I and 1KE9. The ligand from PDB structure 1KE9 (orange) dose not dock into 2R3I (green) due to different orientation of Lys33.

The application of docking methods in screening for drug discovery mostly involves non-native docking experiments to predict the binding modes of newly designed compounds or libraries of compounds against a protein structure. Based on CDK2 non-native docking results it is concluded that if required the Lys35 of CDK4 corresponding to $Lys33^{CDK2}$ should be treated as flexible using GOLD.

5.3 Molecular Docking of Fascaplysin into CDK2, CDK4 and CDK6

Fascaplysin (Figure 1-8) was docked into CDK4 in order to predict and study the binding mode of fascaplysin with this protein and particularly to understand the huge difference in binding affinity between the CDK4-fascaplysin and CDK2-fascaplysin complexes. CDK2 and CDK6 were also docked with fascaplysin in order to investigate the CDK4 selectivity for fascaplysin. This docking experiment was performed in

GOLD using PDB structure 1FIN for CDK2, CDK4 Model (generated as described in methods) and CDK6 PDB structure 1XO2. The results of the CDK2, CDK4 and CDK6 docking with fascaplysin (Figure 5-6) show a similar docking pose of fascaplysin in the binding pocket of all three proteins. Polar interactions between protein and fascaplysin are found at $Leu83^{CDK2}$, $Val96^{CDK4}$ and $Val101^{CDK6}$ which are in-line with the known polar interactions between CDK2 and different inhibitors in PDB (also shown in Section 5.2.1). This docking experiment successfully solves the first part of docking problem by predicting the likely structure of protein ligand complex. However, the structural basis of CDK4 potency is not evident in the docking score (Table 5-2) in this early docking result. The ChemScore value obtained for CDK4 and CDK2 is 31.29 and 29.65, respectively.

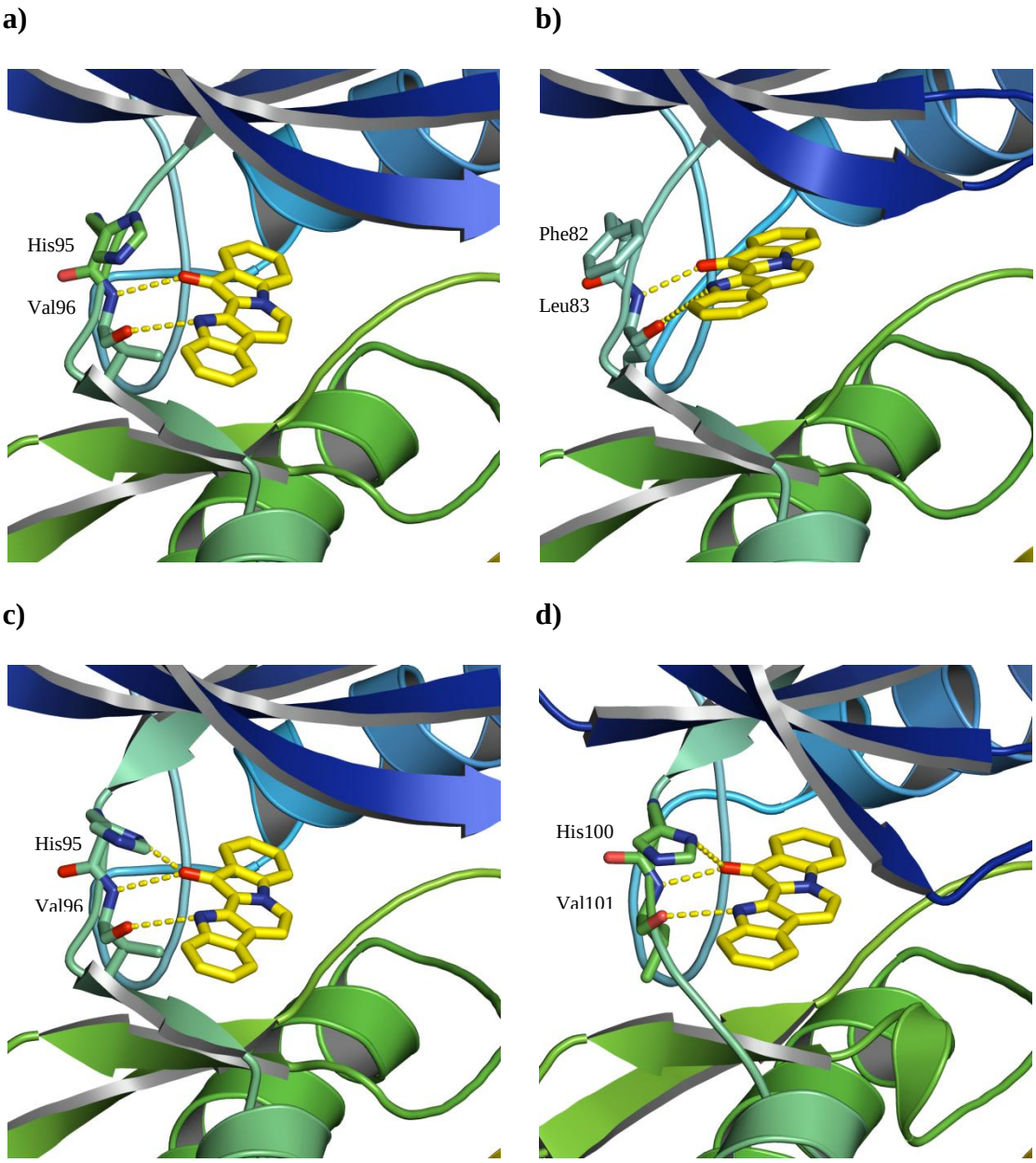


Figure 5-6 Molecular Docking of CDK2, CDK4 and CDK6 with Fascaplysin
a) A cartoon representation of the CDK4 model complexed with fascaplysin. The docking result shows the hydrogen bonds between NH and carbonyl groups of fascaplysin and the carbonyl and NH groups of Val96 **b)** CDK2 structure complexed with the fascaplysin inhibitor. It shows a hydrogen bond pair between NH and carbonyl groups of fascaplysin and the carbonyl and NH groups of Leu83. **c)** CDK4 model after His95 adjustment complexed with fascaplysin. The docking result shows an additional hydrogen bond between carbonyl group of fascaplysin and hydrogen on the delta nitrogen of the imidazole side chain in $His95^{CDK4}$. **d)** CDK6 complexed with fascaplysin. A donor-acceptor hydrogen bond pair between NH and carbonyl groups of fascaplysin and the carbonyl and NH groups of Val101 is shown.

There are some differences in the sequence of the binding pocket between CDK4 and CDK2 (as described in Section 3.2.3) for example $His95^{CDK4}$ replaces the $Phe82^{CDK2}$. The corresponding residue to $His95^{CDK4}$ in CDK6 is His100. Structural overlays of the corresponding $His100^{CDK6}$ from all the available PDB structures of CDK6 indicate conformational flexibility for this histidine (Figure 5-7). The visualisation of His95 in PyMol shows that this residue may have some polar interactions with the ligand if some flexibility is allowed. The orientation of $His95^{CDK4}$ was adjusted (Figure 5-8) based on the information obtained from $His100^{CDK6}$ using Dunbrack rotamer library (Dunbrack and Karplus, 1993).

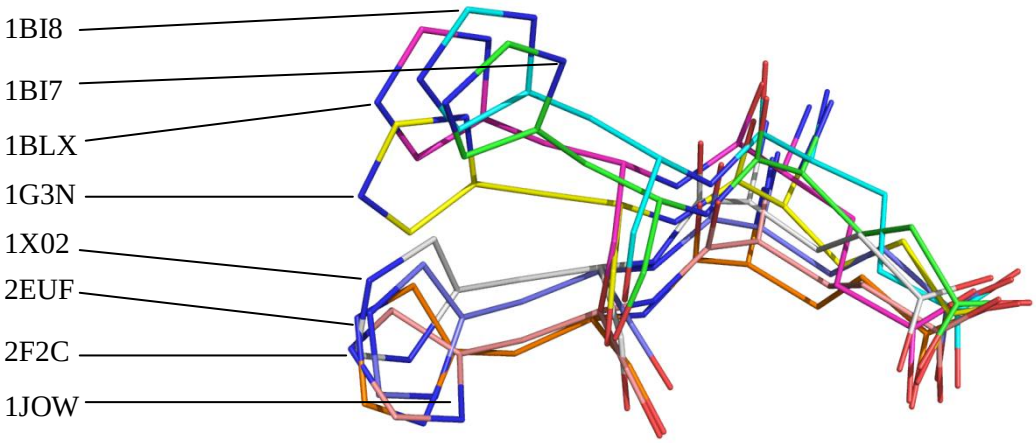


Figure 5-7 : Structural overlay of His100CDK6. There are eight PDB structures for CDK6 (1BLX, 1G3N, 2F2C, 1BI8 1XO2, 2EUF, 1JOW and 1BI7). This structural overlay of all $His100^{CDK6}$ shows orientation flexibility for this residue.

The docking results have shown a donor-acceptor hydrogen bond pair between the indolyl-NH and carbonyl groups of fascaplysin with the backbone NH and carbonyl group $Val96^{CDK4}$. Followed by $His95^{CDK4}$ orientation adjustment CDK4 docking resulted in an additional polar interaction between the carbonyl group of fascaplysin and delta nitrogen of imidazole side chain in $His95^{CDK4}$ and an improvement in the

docking scores from 31.29 to 33.37 (Table 5-2). The bidentate pattern of the fascaplysin binding with Val96 of CDK4 is consistent with the binding mode reported earlier (McInnes *et al.*, 2004; Aubry *et al.*, 2006) however, the possibility of polar interaction with $His95^{CDK4}$ is reported for the first time in this study.

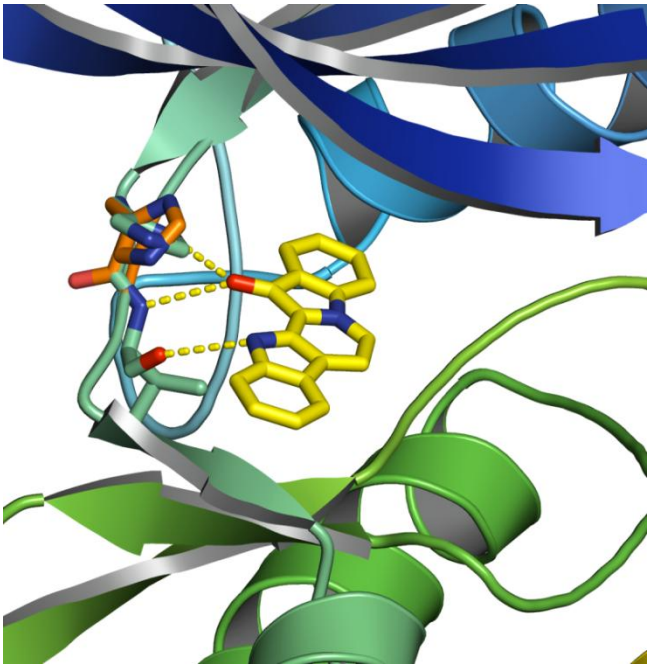
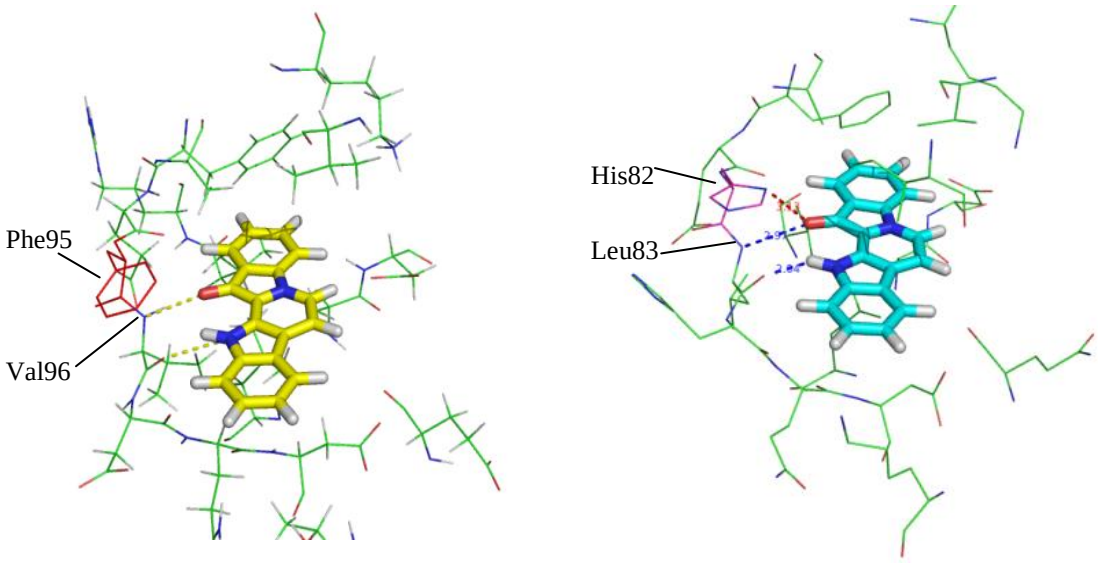


Figure 5-8 Overlay of different orientation of His95 of CDK4.

To investigate the impact of $His95^{CDK4}/Phe^{CDK2}$ substitution in the binding pocket of CDK2/CDK4 on the docking scores (and potentially K_D), *in-silico* mutations were carried out replacing the corresponding amino acid in CDK4 (His95) with that of CDK2 (Phe82) and vice versa (Figure 5-9). A change in ChemScore for these *in-silico* mutants is observed (Table 5-2). The ChemScore for $CDK4^{Phe95}$ is decreased from 33.37 to 32.04 and ChemScore for $CDK2^{His82}$ is increased from 29.65 to 34.40 indicating a possible role of this $His95^{CDK4}$ toward the CDK4 specificity. However the $\Delta\Delta G$ obtained for CDK4 and CDK2 with fascaplysin docking ChemScore (Table 5.2) is 3.7 kJ compared with the $\Delta\Delta G^0$ 17.6 kJ calculated from already known

IC_{50} values (Table 5.3). Therefore $His95^{CDK4}$ alone can, at the most partly explain the CDK4 selectivity indicating a contribution of additional factors.



a) **b)**

Figure 5-9: Docking of CDK2 and CDK4 *in-silico* mutants a) Line representation of CDK4 active site with Phe95 (*in-silico* mutation) shown in red docked with *fascaplysin*. **b)** Line representation of CDK2 active site with His82 (*in-silico* mutation) shown in magenta docked with fascaplysin.

Table 5-2: ChemScore (top hits) for CDK2 and CDK4 with and without *in-silico* mutation. The PDB structure 1FIN was used for CDK2 docking. The formula $\Delta G^0 = -RT\ln K_D$ was used for calculation of estimated K_D. Here ΔG^0 represents the free binding energy, R ideal gas constant with a value of 8.314 $JK^{-1}mol^{-1}$, T = 298 K, K_D dissociation constant and $\Delta\Delta G^0$ represents the change in binding energy. ChemScore is used as an estimate for ΔG^0.

	ChemScore	**Estimated K_D (µM)**	**$\Delta\Delta G^0$ (kJ)**
CDK4 with His95 flexibility adjustment	33.37	1.4	3.7
CDK2 PDB Structure	29.65	6.34	
CDK4. with Phe95 *in-silico* mutation	32.04	2.4	1.5
CDK2 with His82 *in-silico* mutation	34.40	0.9	

Table 5-3: Estimated ΔG^0 (free binding energy) calculated from the known IC_{50} values (Jenkins *et al.*, 2008) of fascaplysin with CDK2 and CDK4 using the equation $\Delta G_0 = -RT\ln K_D$. Here ΔG stands for binding energy, R ideal gas constant with a value of 8.314 JK−1mol−1, T 298 K and K_D dissociation constant. K_D values are taken as approximately equal to IC_{50}.

	K_D (µM)	**Estimated ΔG^0 (kJ)**	**$\Delta\Delta G^0$ (kJ)**
CDK2	500	18.8	17.6
CDK4	0.41	36.4	

McInnes *et al* had proposed that CDK4 selectivity is based on residues $E144^{CDK4}$ and $T102^{CDK4}$ which are replaced with $Q131^{CDK2}$ and $K89^{CDK2}$ in CDK2 (McInnes *et al.*, 2004). They proposed that this two-unit increase in the formal charge of the binding pocket of CDK2 relative to CDK4 destabilises the CDK2 fascaplysin complex relative to CDK4 fascaplysin complex. To further investigate this in addition to $His95^{CDK4}$ mutation (shown in Figure 5-9) *in-silico* mutations at $Asp97^{CDK4}$ and $Glu144^{CDK4}$ and equivalent residues in CDK2 were also carried out. The role of these residues at least in *in-silico* mutations docking experiments were found insignificant (results not shown). The GOLD program does not incorporate long-range electrostatics so such effects by these residues ($E144^{CDK4}$ and $T102^{CDK4}$) cannot be ruled out. Thermodynamics integration method (Straatsma and Berendsen, 1988) is chosen to test the role of formal positive charge on fascaplysin toward its specificity for CDK4. In order to further explore selectivity for CDK4 calculations of free energy difference using thermodynamics integration (see Section 2.14) were carried out as described in Chapter 6.

5.4 Molecular docking of tryptamine based inhibitors of CDK4.

Over 100 fascaplysin inspired inhibitors with a high affinity for CDK4 over CDK2 have been synthesized (Aubry *et al.*, 2004; Aubry *et al.*, 2006; García *et al.*, 2006; Mahale *et al.*, 2006a; Mahale *et al.*, 2006b; Jenkins *et al.*, 2008; Aubry *et al.*, 2009). The tryptamine based bi-phenylcarbonyl and (biphenylcarbonyl)-tetrahydro-β-carboline compounds (Figure 5-10) related to fascaplysin have shown an increased affinity for CDK4 compared to CDK2 (Jenkins *et al.*, 2008).

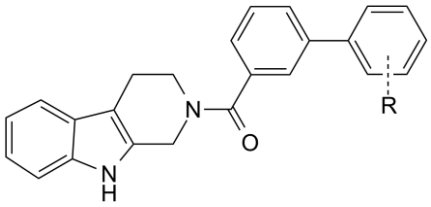
The molecules **1a**, **1b**, **1c**, **2a**, **2b** and **2c** (Figure 5-10) have shown that a change in the position of a methyl group between ortho, meta and para positions changes the

specificity of the CDK4 as shown Table 5-4. In order to understand the binding interactions of these compounds (**1a**, **1b**, **1c**, **2a**, **2b** and **2c**) with CDK4 molecular docking studies were carried out. There are two main goals of this study. First, to predict the ligand binding poses for further improvement of the design. Second, an analysis of scores in order to find if and how far these scores are predictive for CDK4 inhibitors ranking. The above mentioned compounds were docked both with CDK4 model and in the X-ray structure (PDB ID: 2W96) of CDK4. The results of docking are summarized in the Table 5-4. These compounds did not dock into the X-ray structure (PDB ID: 2W96) of CDK4 and a typical kinase inhibitor binding pose was absent. The kinase binding pose is defined as present if a hydrogen bond between the ligand and any one of the three residues (Glu94, Val96 and Asp97) is detected in the docking solutions.

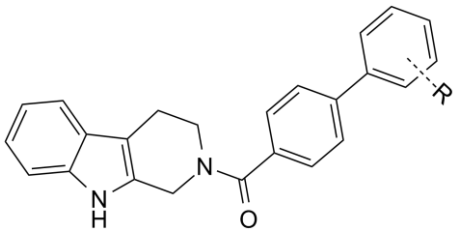
Table 5-4 Sumary of molecular docking of tryptamin based bi-phenyl inhibitors of CDK4. Kinase binding pose is taken as present if a hydrogen bond between the ligand and any one of the three residues (GLU94, Val96 and ASP97) is detected in the docking solutions.

Ligand	**ChemScore***		**Kinase binding pose**		**Predicted K_D (µM)**		**CDK4 measured IC_{50}/µM (Jenkins *et al.*, 2008)**
	CDK4 Model	**PDB 2W96**	**CDK4 Model**	**PDB 2W96**	**CDK4 Model**	**PDB 2W96**	
1a	35.25	25.60	Yes	No	0.66	32	9±0.8
1b	32.46	29.60	Yes	No	2.04	6.0	25±3
1c	30.66	29.60	Yes	No	4.22	6. 0	32±2
2a	30.05	30.48	Yes	No	5.40	4.5	24±2
2b	32.13	30.72	Yes	No	2.33	4.1	11±1
2c	32.05	30.66	Yes	No	2.40	4.2	75±3

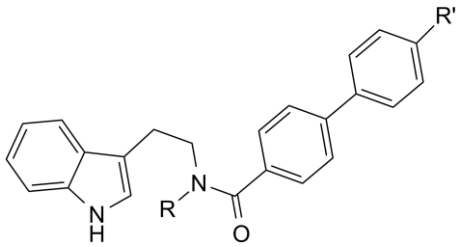
* ChemScore is taken as an estimate of binding affinity to predict the value of K_D using the equation $\Delta G_0 = -RT\ln K_D$.



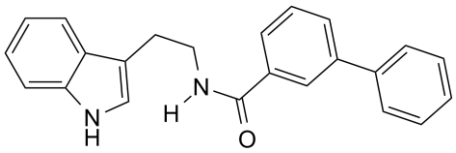
1a: R = p-Me
1b: R=m-Me
1c: R' = o-Me1
1d: R = p-F
1e: R = m-F
1f: R = o-F
1g: R' = m-OMe
1h: R' = o-OMe



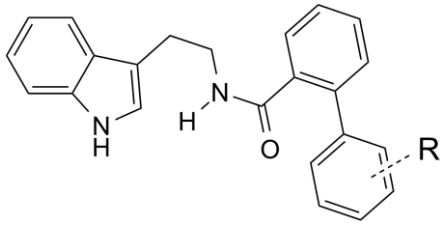
2a: R = p-Me2
2b: R = m-Me2
2c: R = o-Me2
2d: R = p-F
2e: R = m-F
2f: R = o-F
2g: R = m-OMe
2h: R = o-OMe



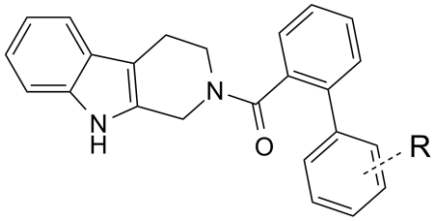
3a: R = Me, R' = H
3b: R = Me, R' = F
3c: R = Me, R' = Me
3d: R = Me, R' = tBu
3e: R = Me, R' = OMe
3f: R =Me, R' = Ph
3g: R = H, R' = H
3h: R = H, R' = F
3i: R =H,R' =Me
3j: R = H, R' = tBu
3k: R = H, R' = OMe



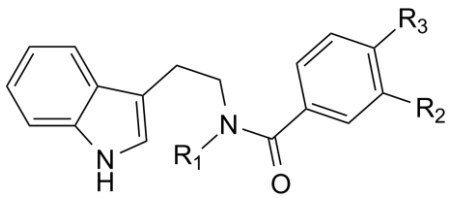
4



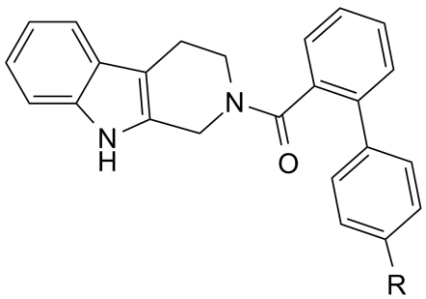
5a: R' = p-F
5b: R' = p-tBu
5c: R' = p-OMe
5d: R' = m-OMe
5e: R' = o-OMe
5f: R' = m-F
5g: R' = m-Me



6a: R' = p-F
6b: R' = m-F
6c: R' = o-F
6d: R' = m-Me



7: R1 = Me, R2 = H, R3 = 4-pyridyl
8: R1 = H, R2 = 3-pyridyl, R3 = H



9a: R = 4-pyridyl
9b: R = 3-pyridyl

Figure 5-10 Molecular structure of tryptamin based bi-phenyl inhibitors of CDK4. IC_{50} values of these structures are given in Table 5-6

This reason why these compounds do not dock well with 2W96 could be attributed to the fact that CDK4 X-ray structure is in an inactive state, thus provide less space to accommodate these ligand in the binding cavity. An analysis of the active sites volume (see Section 2.7) revealed that active site volume of CDK4 X-ray structures is smaller than that of CDK4 homology models (Table 5-5). A comparison of inactive and active form of CDK2 also shows similar results.

Table 5-5 Active site volume of CDK2 and CDK4.

CDK2			CDK4		
PDB ID	**Active/Inactive**	**Pocket Volume (Å)**	**PDB ID**	**Active/Inactive**	**Pocket Volume (Å)**
2R3I	Inactive	2147.21	2W96	Inactive	2216.17
1FIN	Active	2821.34	CDK4 Model	Active	2916.27

The compound **1a** is predicted to be a strongest binder (Table 5-4) with a ChemScore value of 35.25, which is consistent with lowest experimental IC_{50} of compound **1a** (Table 5-4). A higher value of ChemScore is taken as an estimate of higher binding affinity.

The docking pose of ligand **1a** is shown in Figure 5-11. This pose shows three polar interactions between carbonyl group of the ligand and backbone NH of Val96, between carbonyl group of the ligand and delta nitrogen of imidazole side chain in His95 and between indole group nitrogen and backbone carbonyl group of Asp97. The terminal ring of the ligand **1a** is predicted to make π-stacking interactions with Phe93 of CDK4.

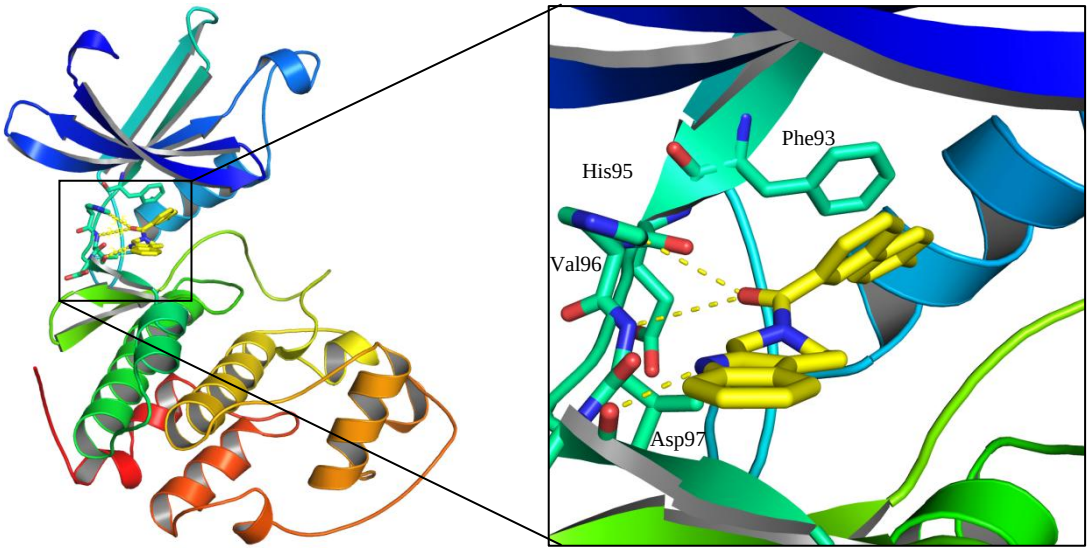
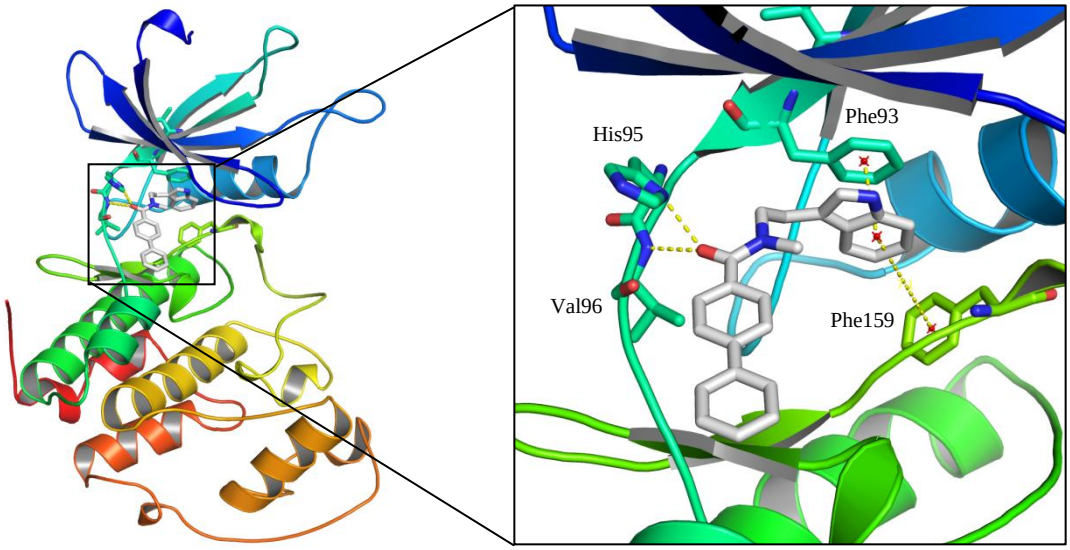


Figure 5-11 Molecular docking of tryptamin based bi-phenyl inhibitors of CDK4. A cartoon representation of the CDK4 model complexed with **1b** inhibitor. The docking results shows a donor-acceptor hydrogen bond pair between NH and carbonyl groups of ligand and the carboxyl and amino groups of Val96 and His 95 of CDK4

The most active compound (**3a**) biphenyl-4-carboxylic acid [2-(1H-indol-3-yl)-ethyl]-methyl-amide has an IC_{50} of 6 µM for CDK4/cyclin D1 and 521 µM for CDK2/Cycline A (Jenkins *et al.*, 2008). The compound **3a** was also docked with both, CDK4 model and all the available PDB structures of CDK4. The PDB structures of CDK4 (e.g. 2W96) did not dock well with compound **3a** as kinase binding pose and polar interactions were absent.

The docking of **3a** with CDK4 model predicted that the biphenyl moiety of inhibitor is projected toward the surface (Figure 5-12). Two hydrogen bonds are predicted to form between backbone delta nitrogen of the His95 and backbone NH group of Val96 with carbonyl group of the ligand. The indole ring participates in aromatic parallel-displaced π-stacking interaction with Phe93 and an edge-to-face π-stacking with Phe159. The binding mode predicted in the present study is inconsistent with the previously published work (Aubry *et al.*, 2006). Aubry *et al*. have predicted π-stacking interactions between the terminal benzoid ring of compound **3a** with Phe93 and Phe159, together with polar interactions between the N atom of the indol group and

backbone carbonyl group of Val96 (Aubry *et al.*, 2006). In order to investigate this matter fully it may be interesting to measure the activity of inhibitor with a substitution of N atom of indol group with a carbon or an addition of the methyl group at this position, which will interrupt the polar interaction between the N atom of the indole group and backbone carbonyl group of Val96. Similarly an addition of hydrophilic group to the terminal ring of **3a** may provide some further insight into its binding pose.



a) **b)**

Figure 5-12 : Molecular docking of compound 3a with CDK4. The biphenyl ring is projected toward surface of CDK4. The indole ring participates in aromatic parallel-displaced pi stacking interaction residues Phe93 and and an edge-to-face π-stacking with Phe159

Each of the tryptamine based biphenyl inhibitors of CDK4 (Figure 5-10) was docked into the CDK4 model in order to study the correlation between ChemScore and the experimental IC_{50} of these compounds. ChemScore fitness function was chosen, because unlike the GoldScore it has been parameterized against binding affinities (Eldridge *et al.*, 1997). The value of the top-ranked docking solution is given in Table 5-6. A comparison between ChemScore values (taken as an estimate of binding affinity)

and the experimental log IC_{50} values shows a correlation with a regression coefficient of r^2=0.54 (Figure 5-13).

Table 5-6 Molecular docking of tryptamin based bi-phenyl inhibitors of CDK4. Experimental IC_{50} (Jenkins *et al.*, 2008) and ChemScore values for each of the compounds shown in (Figure 5-10).

Compound	ChemScore[a]	IC_{50} µM	Compound	ChemScore[a]	IC_{50} µM
1a	35.25	9	3g	39.83	16
1b	32.46	25	3h	39.12	15
1c	30.66	32	3i	31.10	18
1d	32.92	12	3j	35.22	9.2
1e	34.38	21	3k	32.29	48
1f	32.68	17	4	36.96	86
1g	34.73	28	5a	28.85	38
1h	32.95	22	5b	34.29	38
2a	30.05	24	5c	30.08	79
2b	32.13	11	5d	28.93	53
2c	32.05	75	5e	28.51	26
2d	35.48	25	5f	29.46	20
2e	32.42	27	5g	30.83	15
2f	31.12	18	6a	29.33	23
2g	29.63	33	6b	29.82	35
2h	36.96	79	6c	30.59	19
3a	34.94	6	6d	31.16	26
3b	33.05	21	7	30.33	26
3c	38.30	24	8	31.19	30
3d	35.76	11	9a	25.18	39
3e	33.47	49	9b	25.18	36
3f	34.37	35			

[a] Top ranked of 10 docked orientations.

The regression coefficient r^2 is a measure of goodness of fit of linear regression. The value of r^2 lies between 0 and 1, it has no units. The value of r^2=1.0 means knowing the value of X will predict the Y perfectly. The regression coefficient value (r^2=0.54) obtained in present study is not very good, but it can be used to discriminate between the tight and weak binding compounds and to predict binding of novel inhibitors (Gohlke and Klebe, 2001).

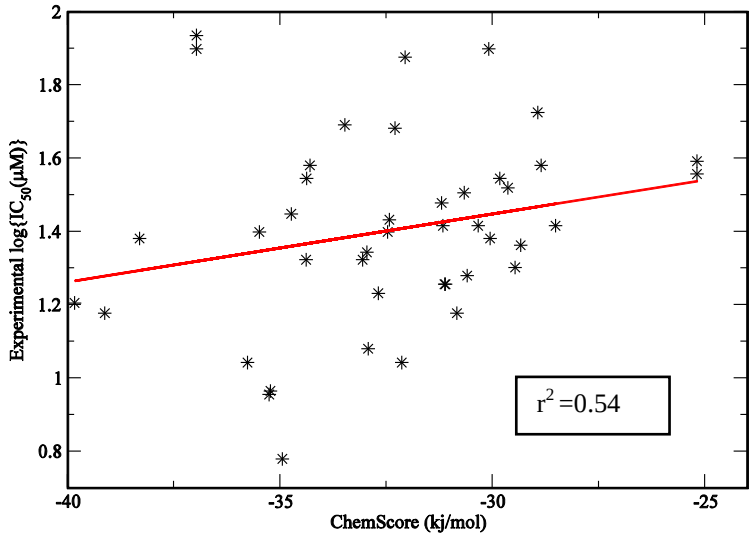


Figure 5-13 Correlation between the value of the ChemScore and IC_{50}. Correlation between the ChemScore fitness function for the top ranked docked solution and the experimental log IC_{50} value for each of the compounds shown in (Figure 5-10) is obtained with a regression coefficient of r^2=0.54.

The accuracy of ligand protein docking may be affected by different factors such as treatment of water mediated protein ligand interactions, the solvent effect, receptor flexibility and choice of scoring function (Mohan *et al.*, 2005; Mpamhanga *et al.*, 2005; Sousa *et al.*, 2006; Roberts and Mancera, 2008). The presence of water molecules in the active site is very important due to their ability to mediate protein ligand interactions with hydrogen bond formation (Poornima and Dean, 1995).

Due to limited information available about waters in the active site of CDK4 docking studies were carried out in the absence of water molecules. The identification of water molecules that may have a potential role in protein ligand docking in the active site of CDK4 is challenging due to absence of ligand bound crystal structures of CDK4. In addition there are no conserved locations of the water molecules observed in X-ray structures of CDK4 (PDB ID 2W96, 2W99, 2W9F and 2W9Z). One of the crystal structure of CDK4 (PDB ID 3G33) does not provide any coordinates of the water molecules. When a ligand binds to protein, water molecules in the active site are either displaced into the bulk solvent or stabilize the protein ligand complex by mediating

interaction between ligand and protein with hydrogen bonds (Barillari *et al.*, 2007; Amadasi *et al.*, 2008). Also a tendency for the occupancy of conserved positions of water molecules is observed in proteins structures obtained under different conditions, bound with different ligands and also in structurally related proteins (Sreenivasan and Axelsen, 1992; Chung *et al.*, 1998; Bottoms *et al.*, 2006). It has been suggested to include water molecules whenever possible in a ligand−protein docking as it may increase the accuracy of molecular docking (Roberts and Mancera, 2008; Thilagavathi and Mancera, 2010).

5.5 Structure-based design of new inhibitors of CDK4

The information obtained from the active site analysis of CDK4 and molecular docking results (as discussed in pervious section) was used to design a new class of inhibitors of CDK4. The intramolecular distance between the indolyl-NH and carbonyl groups of fascaplysin is 2.9 Å (Figure 5-14 a). As described earlier fascaplysin is predicted to forms bidentate hydrogen bonds with CDK4. The distance between NH and amide carbonyl group of Val96 involved in bidentate hydrogen bond with fascaplysin is 2.6 Å. The corresponding intramolecular distance for tryptamine based bi-phenyl ligands is 4.6 Å (Figure 5-14 b) therefore these cannot make bidentate hydrogen bonds with Val96 as it is predicted for fascaplysin. In order to maintain the binding pose of fascaplysin with Val96 a new class of compounds (Figure 5-15) is proposed by shifting the position of nitrogen at the third ring of β-carbolines from meta to ortho position thus making them α-carbolines. To investigate binding poses of these newly proposed compounds molecular docking was carried out.

The docking result shows that newly designed compounds (α-carboline) *in-silico* are able *silico* are able to form bidentate hydrogen bonds with Val96 of CDK4 (Figure 5-14 d), and these compounds also have improved ChemScores (Table 5-7)

Table 5-7compared to β-carboline compounds. Based on this newly proposed class eleven compounds (α-carbolines and α-carbolines N-oxides) (Figure 5-16) have been synthesized in the group of Dr. Paul Jenkins in Chemistry Department and the determination of their properties is in progress.

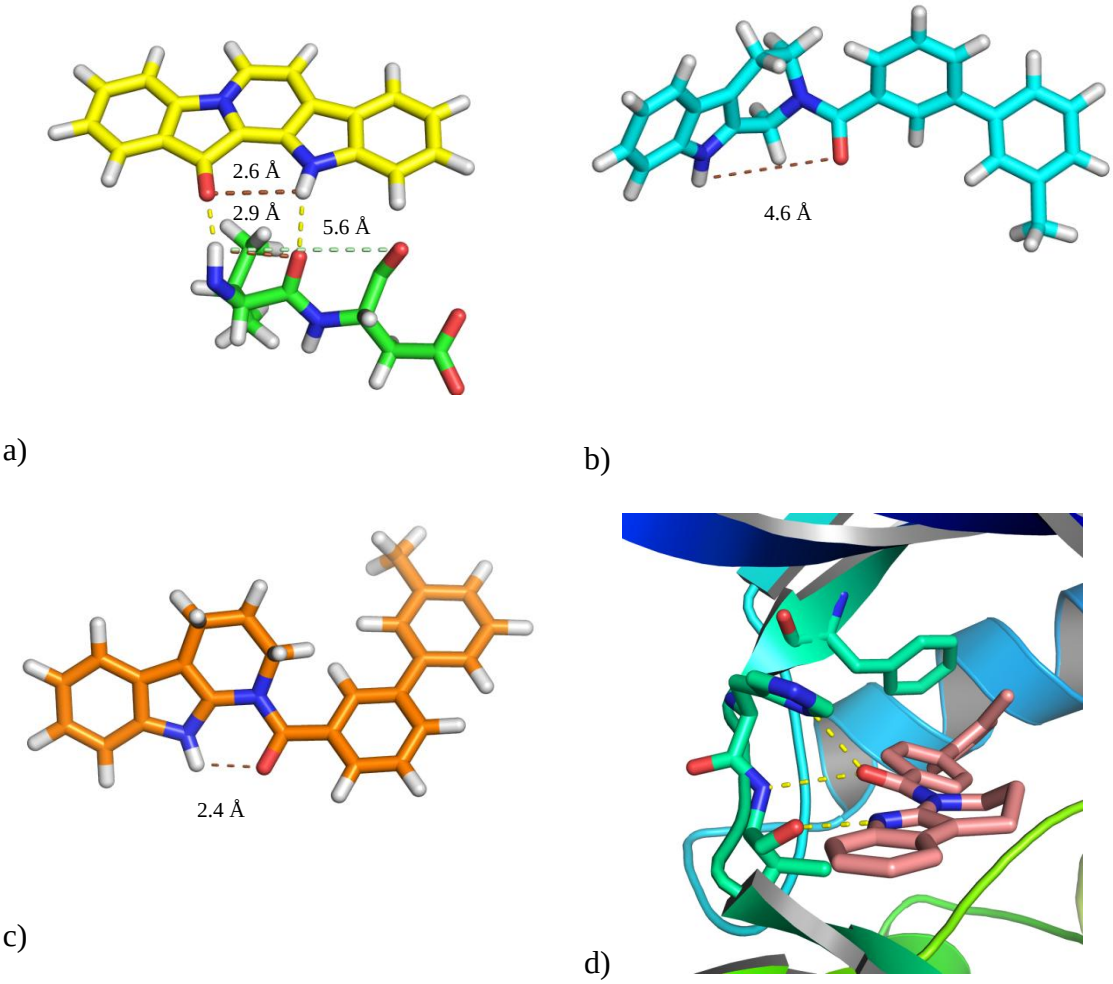
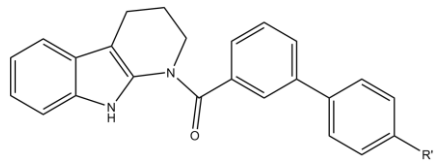
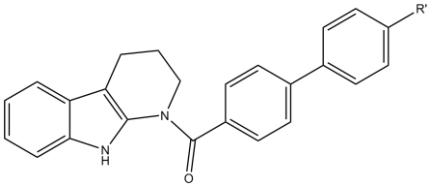


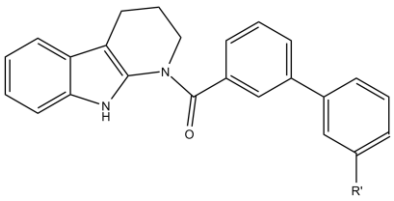
Figure 5-14 . Structure-based design of new inhibitors of CDK4. a) Stick representation of fascaplysin complexed with CDK4. The dashed lines in yellow represent hydrogen binding, brown and green lines are used to represent interamolecular distance calculations. The distance between indolyl-NH and carbonyl groups of fascaplysin is 2.9 Å. The corresponding distance in CDK4 active site between the NH and amide carbonyl group of Val96 is 2.6 Å. The distance between the NH group of Val96 and the amide carbonyl group of neighboring Asp97 is 5.6 Å. **b)** The corresponding intramolecular distance in Tryptamin based bi-phenyl ligands for CDK4 between the atoms involved in possible polar interactions with CDK4 is 4.6 Å. **c)** Newly proposed class of molecules where the position of the nitrogen is shifted from the Tryptamin based bi-phenyl ligands in an attempt to match the corresponding distance of fascaplysin **d)** Newly proposed molecule docked with CDK4 model.



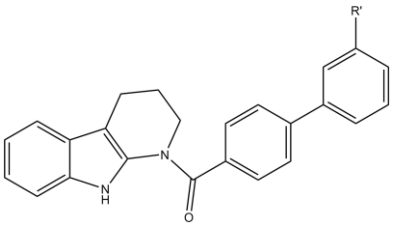
1a-n



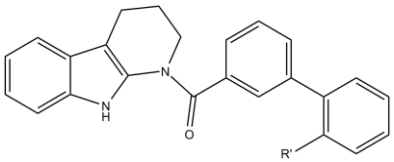
2a-n



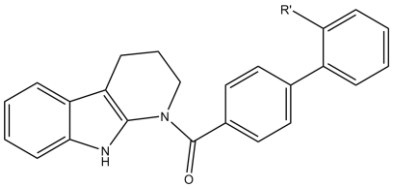
1b-n



2b-n



1c-n



2c-n

1 a-n: R'=p-Me
1 b-n: R'=m-Me
1 c-n: R'=o-Me

2 a-n: R'=p-Me
2 b-n: R'=m-Me
2 c-n: R'=o-Me

Figure 5-15 Molecular structure of newly proposed alpha carbolines. Structure based design of newly proposed of alpha carbolines similar to the tryptamin based bi-phenyl inhibitors of CDK4 as shown in Figure 5.10

Table 5-7. Summary of molecular docking of newly proposed α-carbolines with CDK4.

Ligand	ChemScore*	Predicted KD (μM)	Ligand	ChemScore*	Predicted KD (μM)
1a-n	38.13	0.20	**2a-n**	35.18	0.68
1b-n	36.74	0.36	**2b-n**	32.11	2.35
1c-n	34.80	0.79	**2c-n**	30.60	4.32

*ChemScore is taken as an estimate of binding affinity to predict the value of K_D using the equation ΔG_0=-RTlnK_D.

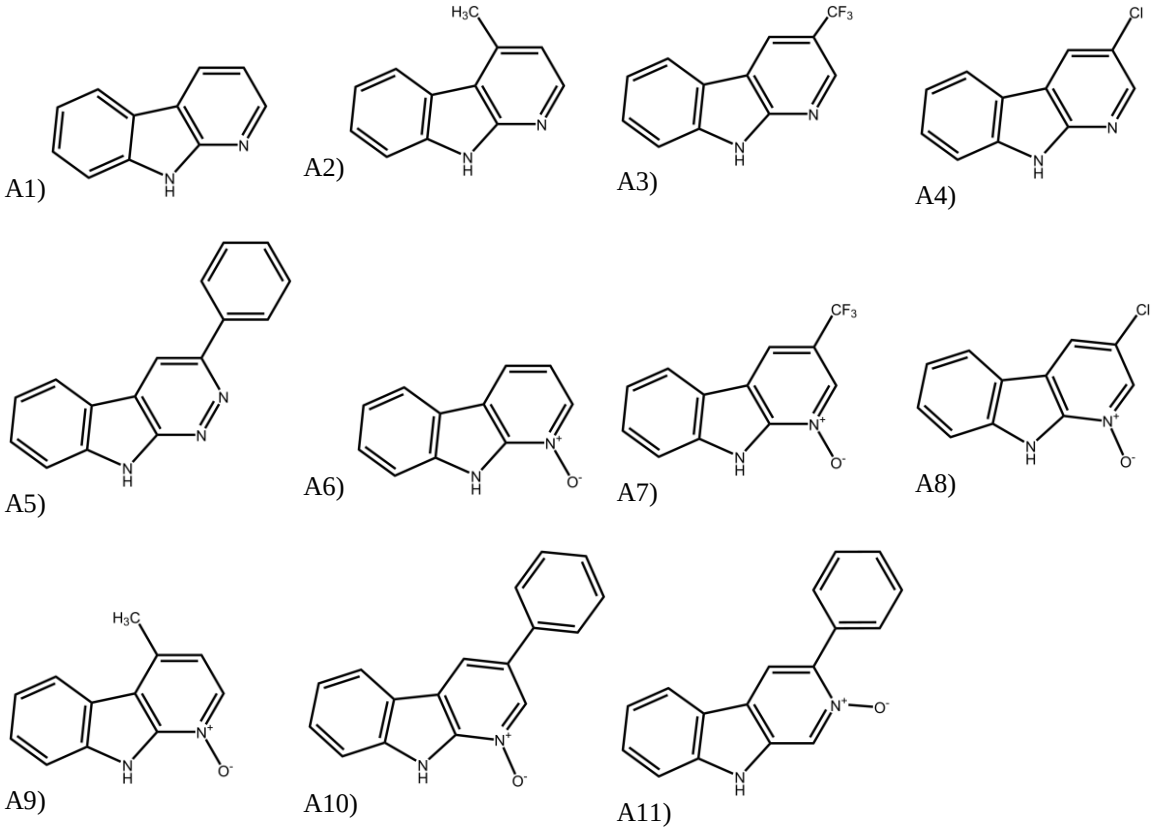


Figure 5-16 Chemical structures newly proposed compunds. Compound A1-A5 represent α-carboline and its derivatives and compound A6-A11 α-carboline N-oxide.

Table 5-8 Molecular docking of newly proposed alpha-carboline and its derivatives.

Ligand	CDK4-Model		PDB structure 2W96	
		Kinase binding pose		Kinase binding pose
A1	30.18	Yes	25.00	No
A2	30.77	Yes	26.79	No
A3	29.71	Yes	25.34	No
A4	30.52	Yes	27.02	No
A5	31.75	Yes	26.25	No
A6	25.50	Yes	23.67	No
A7	26.47	Yes	24.41	No
A8	26.6	Yes	24.38	No
A9	25.6	Yes	24.51	No
A10	35.36	Yes	26.38	No
A11	29.85	Yes	26.91	No

The docking results of these compounds predict that α-carbolines and α-carbolines N-oxides have a different binding pattern. The compound 1 makes polar interactions with Val96, His95 and Glu94. The binding pattern observed for the core structure remains conserved with the addition of the CF3, chlorine, methyl and phenyl substituents (compound 2-5). The core structure of α-carbolines N-oxide (compound 6) shows a different binding pattern than any other compound and it make a hydrogen bond with His95. The compound 7-11 maintains the bidentate hydrogen bonds with Val96 similar as predicted for fascaplysin. The addition of phenyl ring to the core structure of α-carbolines N-oxide (compound 10 and 11) seems to stabilize the inhibitor complex via π-stacking with $Phe93^{CDK4}$. This suggests that these compounds may be more potent inhibitors of CDK4 than the previously synthesized β-carbolines.

In a structure based design of new inhibitors usually the focus is optimization of a ligand affinity. The structure based design and development of CDK4 inhibitors has been focus of different research groups (Ikuta *et al.*, 2001; Aubry *et al.*, 2004; Horiuchi *et al.*, 2009). However the design and synthesis of novel inhibitors of CDK4 to achieve a nano-molar activity is still in progress.

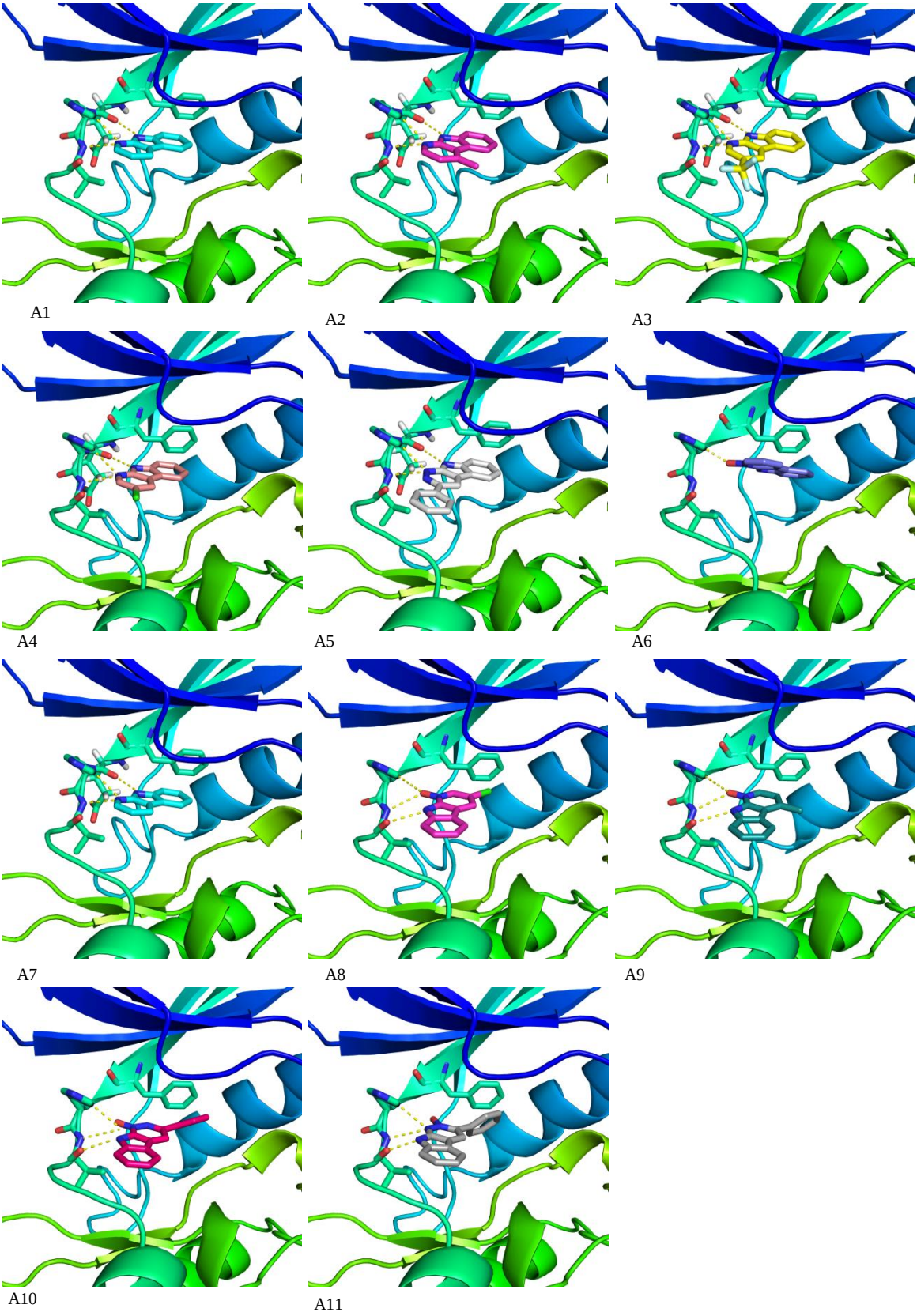


Figure 5-17 **Molecular docking of newly proposed α-carboline derivatives.** α-carbolines (A1-A5) and α-carbolines N-oxides (A6-A11) make polar interactions with Val96, His95 and Glu94.

5.6 Conclusion

Knowledge of the three-dimensional structure of protein ligand complexes in the form of high-resolution crystal structures plays a key role in structure-based design or rational design of ligands with increased binding affinity for the binding site of a particular target protein (Congreve *et al.*, 2005). To date no ligand bound structure of CDK4 is available. Therefore, docking studies are used to learn how different ligands may interact with CDK4. Molecular docking is one of the commonly used techniques in drug design and in *in-silico* lead optimization. In the present work a test set of 21 high quality CDK2–ligand complexes was used for the purpose of validating the docking strategy and to find its utility for CDK4 docking. The performance of GOLD docking suite with CDK2 shows a docking success rate of 76% and 81% using GoldScore and ChemScore, respectively.

Docking studies on fascaplysin with CDK4 predict a polar contact between $His95^{CDK4}$ and fascaplysin in addition to the bidentate hydrogen bond with Val96. The interaction between the $His95^{CDK4}$ and fascaplysin partly explain the selectivity of CDK4 compared to CDK2. The docking of tryptamine based derivates with CDK4 model also show a polar interaction between these compounds and $His95^{CDK4}$. The tryptamine based derivatives used in the present study did not dock well with CDK4 experimental structures (e.g. PDB ID: 2W96), probably because these are in an inactive conformations and appeared to have less space in the binding site to accommodate these ligands.

A comparison of experimental log IC_{50} values for tryptamine based inhibitors and their estimated binding free energies (ChemScore) shows a correlation with a regression coefficient of $r^2 = 0.54$. Although not perfect, this correlation may be useful for ranking and predicting relative affinities for novel inhibitors of CDK4. The

information obtained by the active site analysis of CDK2/4 and docking studies was used to propose a new class of α-carboline inhibitors, *In-silico* these compounds have improved binding properties for CDK4, remains to be seen if and how far this can be confirmed experimentally.

Chapter Six
Thermodynamic integration studies of the CDK2 and CDK4 fascaplysin complexes

Chapter 6 Thermodynamic integration studies of CDK2 and CDK4 fascaplysin complexes

6.1 Introduction

The binding of inhibitors (ligands) to protein receptors with high affinity and specificity is central to structure-based drug design applications. The quest for the calculation of binding affinity remains one of the main goals of modern computational methods. Fascaplysin (Figure 1-8) specifically inhibits CDK4 compared with CDK2. The structural bases of fascaplysin specificity are poorly understood. Molecular docking studies on fascaplysin with CDK2 and CDK4 only partly explain selectivity based on the possible role of $His95^{CDK4}$ (see Section 5.3). An alternative hypothesis is that CDK4 selectivity is caused by differential electrostatic stabilization based on the difference in the formal charge of the binding pockets of CDK2 and CDK4. CDK4 has an acidic residue Glu144 and a neutral residue Thr102 compared to $Gln131^{CDK2}$ and positively charged $Lys89^{CDK2}$ in the equivalent positions of CDK2 (Figure 3-4). Based on this hypothesis the formal positive charge on fascaplysin brings selectivity for CDK4 vs. CDK2 due to electrostatic effects. In order to test this hypothesis and to explain fascaplysin selectivity for CDK4 a thermodynamic integration experiment is designed (see Section 2.14) to calculate the ΔG for the conversion of fascaplysin into carbofascaplysin by isoelectronically replacing nitrogen atom (positively charged) to a carbon atom (neutral) (Figure 6-11). Thermodynamic integration (TI) method is used to evaluate free energy differences of two given phases by molecular dynamics simulations (see Section 1.12). Thermodynamic integration is an ideal method to computationally study the impact of small changes in inhibitor structures and properties on the binding affinity.

6.2 Molecular dynamics simulations and stability of the trajectories

CDK2 structure 1FIN (Jeffrey *et al.*, 1995b) and the CDK4 hybrid homology model (as discussed in Chapter 4) was used for the molecular dynamics studies. 1FIN was selected from more than 190 structures available for CDK2 (see Appendix 1.1) based on the fact that it represents the active form of CDK2 and has no missing regions in PDB coordinates corresponding to the CDK2 full-length sequence (298aa). Also the CDK4 hybrid homology model represents the putative active state of CDK4, has no missing regions and it also incorporates the structural information from the CDK4 X-ray structure.

It is a prerequisite to thermodynamic integration to perform molecular dynamics simulations to test the stability of a system with a chosen set of parameters and water models. Molecular simulations of free and ligand (fascaplysin and carbofascaplysin) bound CDK2 and CDK4 were performed as described in Section 2.13. Different trajectories were collected during these molecular dynamics simulations, which are presented and discussed here.

6.2.1 Trajectories of CDK2 MD simulations with different water models

A series of five nanosecond (5ns) MD simulations were carried out to study the effect of different water models. MD simulations were performed with the established TIP3P (Jorgensen *et al.*, 1983) and the more recently developed TIP4P-Ew (Horn *et al.*, 2004) water models in presence and absence of buried crystal waters. TIP3P and TIP4P-Ew significantly increases the complexity of the simulations and the time required to run the simulations compared to implicit water models (Tsui and Case, 2000). TIP4P-Ew water model employs the re-parameterization of TIP4P (Jorgensen *et*

al., 1983) water model with Ewald summation for the calculation of large range electrostatic interactions (Horn *et al.*, 2004).

The MD simulations were initially carried out in absence of any crystal waters using a TIP3P solvent box. In order to identify a possible effect of buried crystal waters on the stability of the protein simulations and also to study the dynamics of water molecules found in the PDB structure CDK2 simulations were performed with a TIP3P solvent box while keeping all the 136 water molecules from the PDB structure (1FIN). In addition to the TIP3P runs, conformational stability of CDK2 was also tested with TIP4P-Ew water model while keeping all the PDB waters.

During the system preparation (before MD run) 16760, 16650, 16640 water molecules were added using tLeap (see methods Section 2.13) to solvate the CDK2 structure for the above mentioned three experiments, respectively. Following a standard Amber "good practice" protocol each solvated structure of CDK2 was first equilibrated with a short energy minimization, 50ps of heating and 50ps density equilibration followed by 500ps of constant pressure equilibration at 300K with weak restraints on the CDK2 (see Section 2.13) before the 5ns production runs. The total charge on the CDK2 was found to be +4.00, therefore 4 Cl^- were added to neutralize the system. The trajectories for CDK2 density, temperature, total energy and RMSD[b] (for the initial 500 ps equilibration run) for all three experiments are shown in Figure 6-1. These trajectories show stability for density, total energy (ETOT) and temperature.

The average density of the system during first equilibration run using TIP3P is 1.01 g/cm^3 at 300 K, which is almost equal to the density of the water at 300 K which is 0.996 g/cm^3. The temperature of the system remained at 300 K. The average energy of the system for 600 ps is calculated to be -1.40×10^5 kcal with a rise in energy after

first 50 ps of minimization. The RMSD curve of CDK2 is shown in Figure 6-1 (d) with an average RMSD[b] of 1.15 Å.

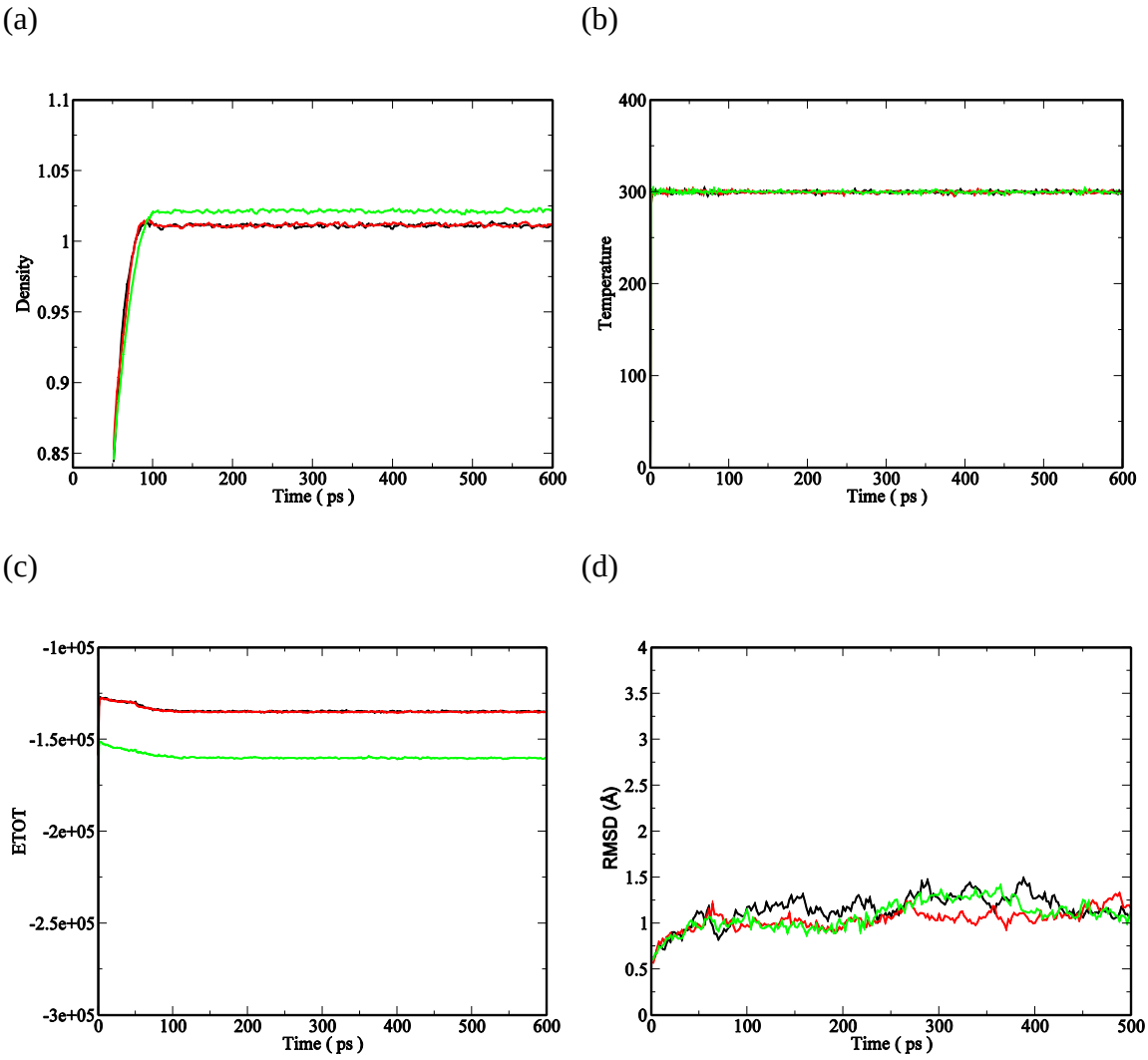


Figure 6-1: Trajectory curves for CDK2 equilibration simulations. CDK2 equilibration simulations displaying stability of density **(a)**, temperature **(b)**, total energy **(c)** for initial 600 ps and RMSD **(d)** for 500 ps equilibration. PDB structure 1FIN (Jeffrey *et al.*, 1995b) was used for CDK2. These equilibration simulations were carried out with a short minimization, 50ps of heating and 50ps density equilibration followed by 500ps of constant pressure equilibration at 300K with weak restraints on the CDK2. Black and red lines represent the CDK2 simulations in TIP3P solvent box with and without crystal waters, respectively. CDK2 simulations in TIP4P-Ew with crystal waters shown in green.

In the second experiment the average density of the system using TIP3P solvent box while keeping all the crystal waters during MD simulation is 1.01 g/cm^3 and total energy of the system is -1.34×10^5 kcal/mol with an average RMSD[b] value of 1.04 Å

for the equilibration run. In the third experiment the average density of the system with TIP4-PEw water model is 1.02 g/cm^3 and total energy of the system is - 1.59x10^5 kcal with an average RMSD[b] value of 1.08 Å during equilibration.

Overlay of the three RMSD curves of 5ns CDK2 production run MD simulations are shown in Figure 6-2. This indicates that system attained an equilibrium indicating overall stability of the system over 5ns run for all the three experiments.

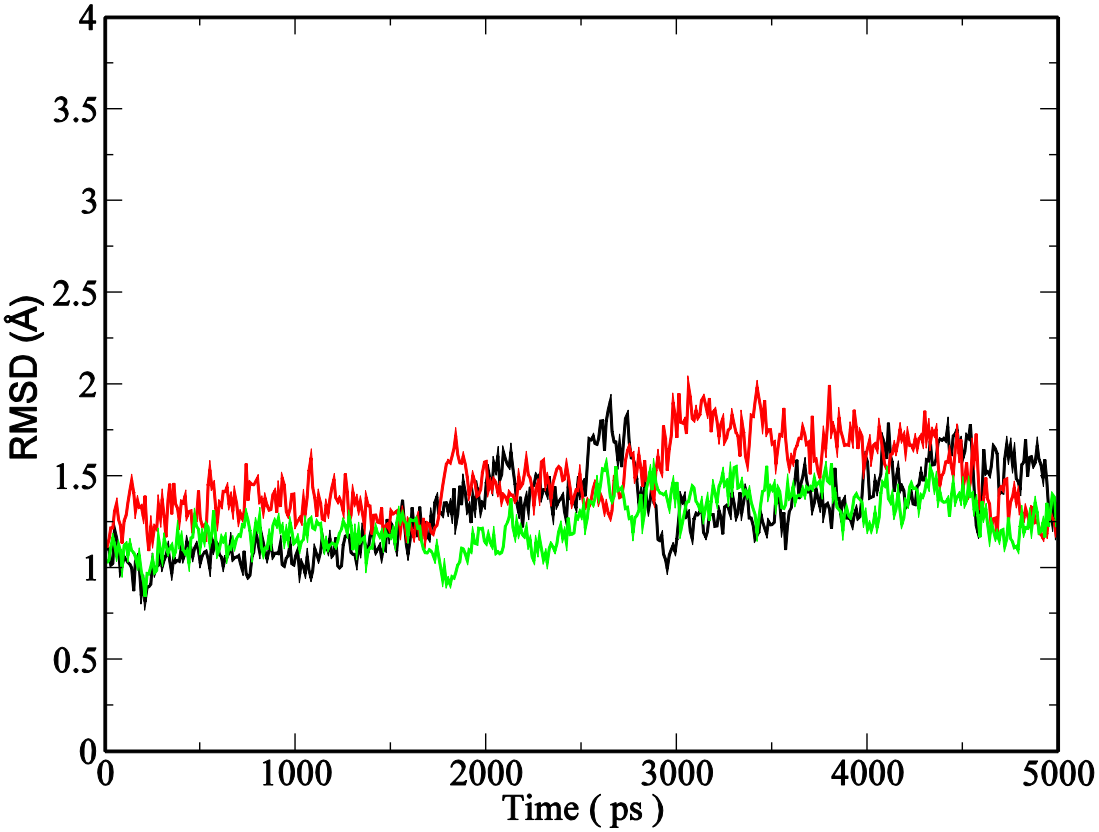


Figure 6-2: RMSD curves overlay for 5ns MD simulations of CDK2. Black and red lines represent the CDK2 simulations in TIP3P solvent box with and without crystal waters, respectively. CDK2 simulations in TIP4P-Ew with crystal waters shown in green.

A comparison of average density, RMSD[b] and total energy 5ns simulations is shown in Table 6-1. The average RMSD[b] calculated for the three runs is 1.31Å, 1.48Å and 1.25Å, respectively. The comparison between these three runs indicates CDK2 simulations with TIP4P-Ew are the most stable at lower energy compared to CDK2 simulations in TIP3P solvent box (with and without crystal waters).

Table 6-1: A comparison of average density, temperature and total energy for three 5ns MD simulations of CDK2.

Production run (5ns)	Avg. Density (g/cm^3)	Avg.ETOT Kcal/mol	Avg. RMSD (Å)
CDK2 with TIP3P solvent	1.01	$-1.34x10^5$	1.31
CDK2 with TIP3P solvent and crystal waters	1.01	$-1.35x10^5$	1.48
CDK2 with TIP4P-Ew solvent and crystal waters	1.02	$-1.60x10^5$	1.25

The difference shown in Table 6-1 may be related to precise temperature dependent properties of pure water supplied by TIP4P-Ew model compared with TIP3P model. The TIP4P-Ew is reported as the best performing water model to reproduce the density of water as a function of temperature, and also account appropriately for electrostatic interactions (Horn *et al.*, 2004; Abascal and Vega, 2007). The slightly higher density than the density of water using TIP4P-Ew solvent box seems reasonable for a protein solution (Horn *et al.*, 2004). This slightly higher average density of TIP4P-Ew compared to TIP3P is consistent with the previously published work (Shirts and Pande, 2005).

6.2.2 Dynamics of CDK2 buried crystal waters

The dynamics of CDK2 buried crystal waters was studied over the 5ns simulations. All 136 waters as found in the PDB file are shown in Figure 6-3 (a, b). The visualization of CDK2 waters in PyMOL (DeLano, 2002) demonstrates that eleven waters are deeply buried in the core of CDK2 as shown by the cartoon and surface representation in Figure 6-3 (c, d). An analysis of CDK2 simulations with VMD (Humphrey *et al.*, 1996) revealed that eleven water molecules remain buried in the CDK2 structure after 50ps of initial equilibration simulations. Two of these waters left the CDK2 pockets during the next equilibration run. Nine water molecules remained buried in CDK2 for the full 5 ns MD simulations.

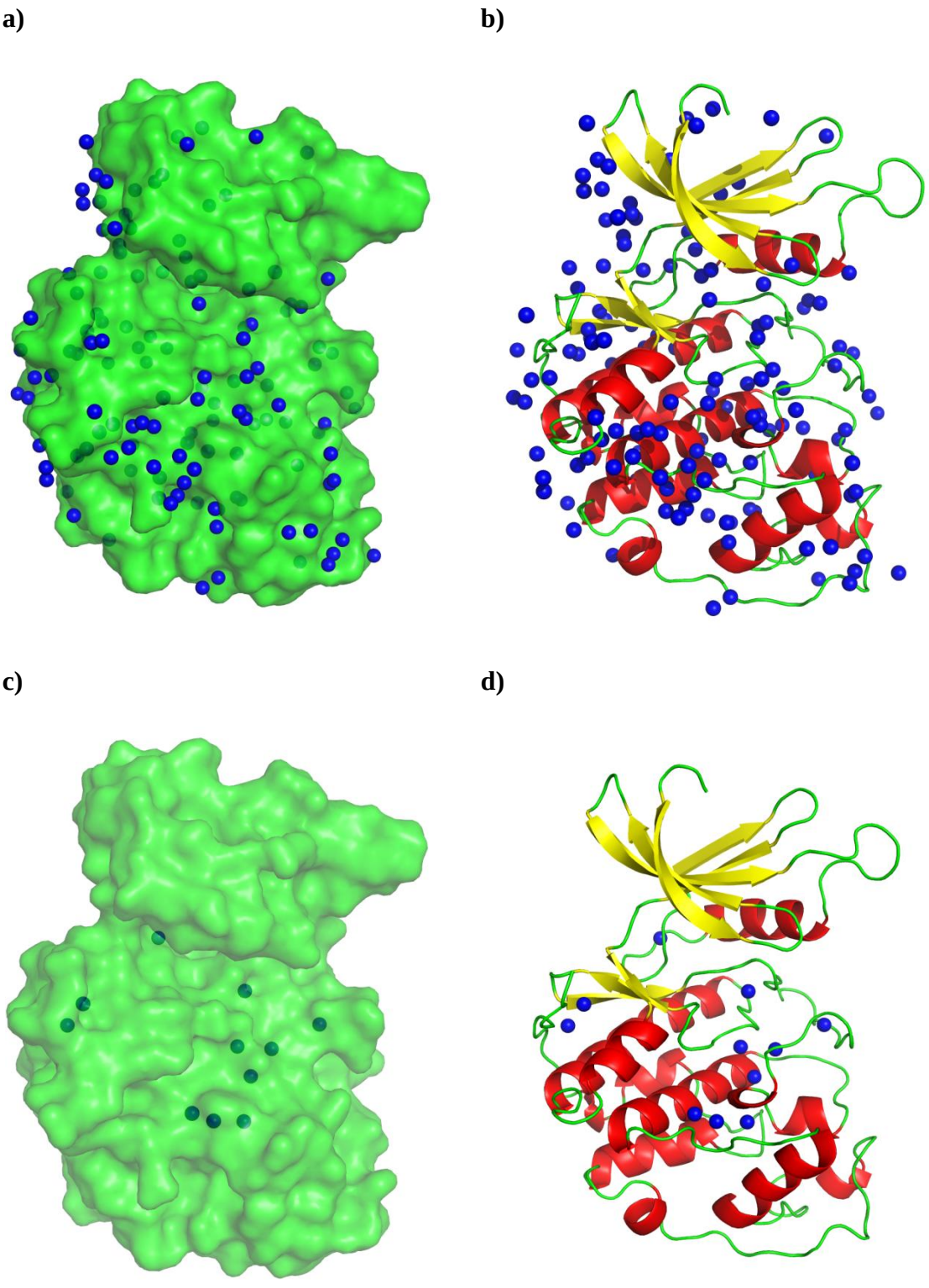


Figure 6-3: Crystal waters in CDK2 (1FIN) before MD simulations a) CDK2 surface representation (green) displaying all 136 water (blue) molecules **b)** Cartoon representation of CDK2 displaying the crystal waters **c)** & **d)** A selection of twelve water molecules found to be buried in the crystal structure shown in CDK2 surface and cartoon representation

A follow-up analysis in VMD revealed that these nine water molecules are the same water molecules as found in original CDK2 structure before simulation. No exchange for these waters is observed.

a) **b)**

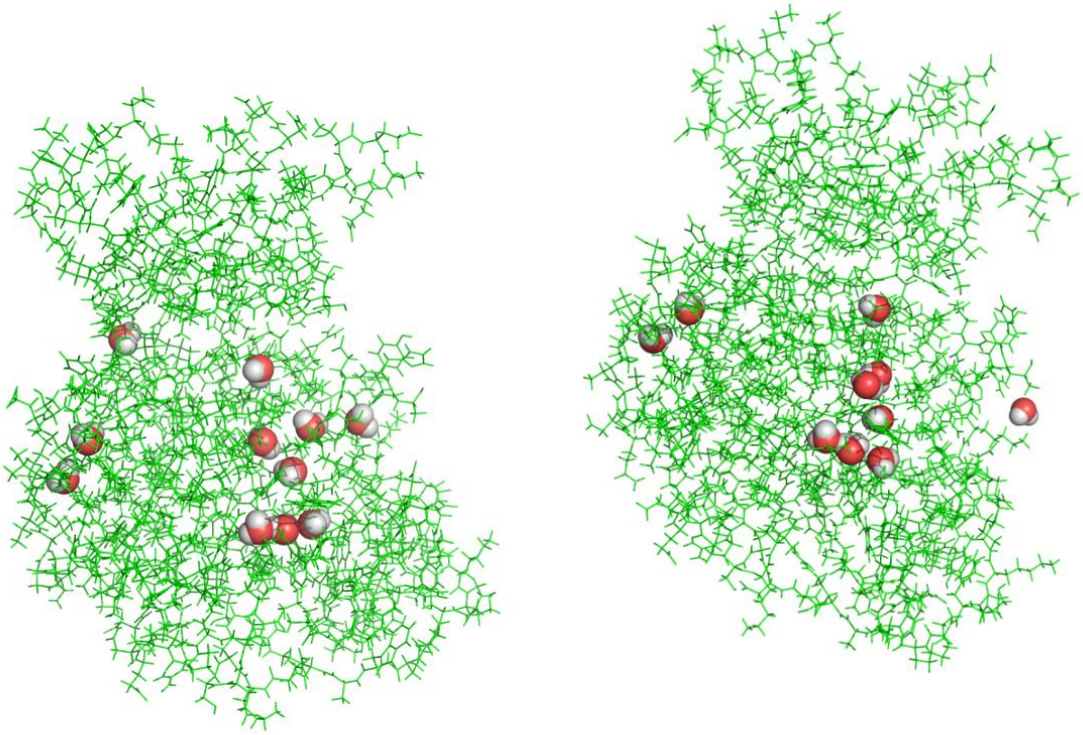


Figure 6-4: Crystal waters in CDK2 (1FIN) after MD simulations. CDK2 structure with all 136 crystal structures was used for molecular dynamics using TIP4P-Ew solvent box. **a)** eleven water molecules buried in the CDK2 structure after 50ps equilibration simulations **b)** nine water are found to be buried in CDK2 after 5 ns MD simulations. These nine molecules are the same water molecules as were in original CDK2 structure before simulation.

Based on these results it was concluded to keep these nine waters in the CDK2 structure for further molecular dynamics studies and thermodynamic integration. Also a decision was made to use the TIP4P-Ew water model for the thermodynamic integration experiments.

6.2.3 Molecular Dynamics Simulations and conformational stability of CDK2/fascaplysin and CDK2/carbofascaplysin complexes

A series of five nanoseconds MD simulations were performed for the CDK2/fascaplysin and CDK2/carbofascaplysin both with TIP3P and TIP4P-Ew water models. The docked complexes of CDK2/fascaplysin and CDK2/carbofascaplysin were prepared as described in Section 2.12. Carbofascaplysin is a hypothetical compound based on fascaplysin (see Section 6.3 and Figure 6-11). The nine waters molecules buried in the CDK2 structure as shown in Figure 6-4 b) were kept for both complexes. These complexes were solvated using tLeap (see methods Section 2.13). The solvated structures of CDK2 complexes were equilibrated before the 5ns production run using the same steps as described earlier for free CDK2 (see Section 2.13). The trajectories for CDK2/fascaplysin and CDK2/carbofascaplysin complexes displaying density, temperature, total energy and RMSD[b] for initial 500 ps equilibration run both with TIP3P and TIP4P-Ew water models are shown in Figure 6-5.

These trajectories show that all four systems attained equilibrium after some initial minor fluctuations. The average density of the CDK2/fascaplysin and CDK2/carbofascaplysin systems with TIP4-PEw water model is 1.02 g/cm^3 compared to 1.01 g/cm^3 with TIP3P water model. The slightly higher density with TIP4P-Ew seems reasonable for protein solution.

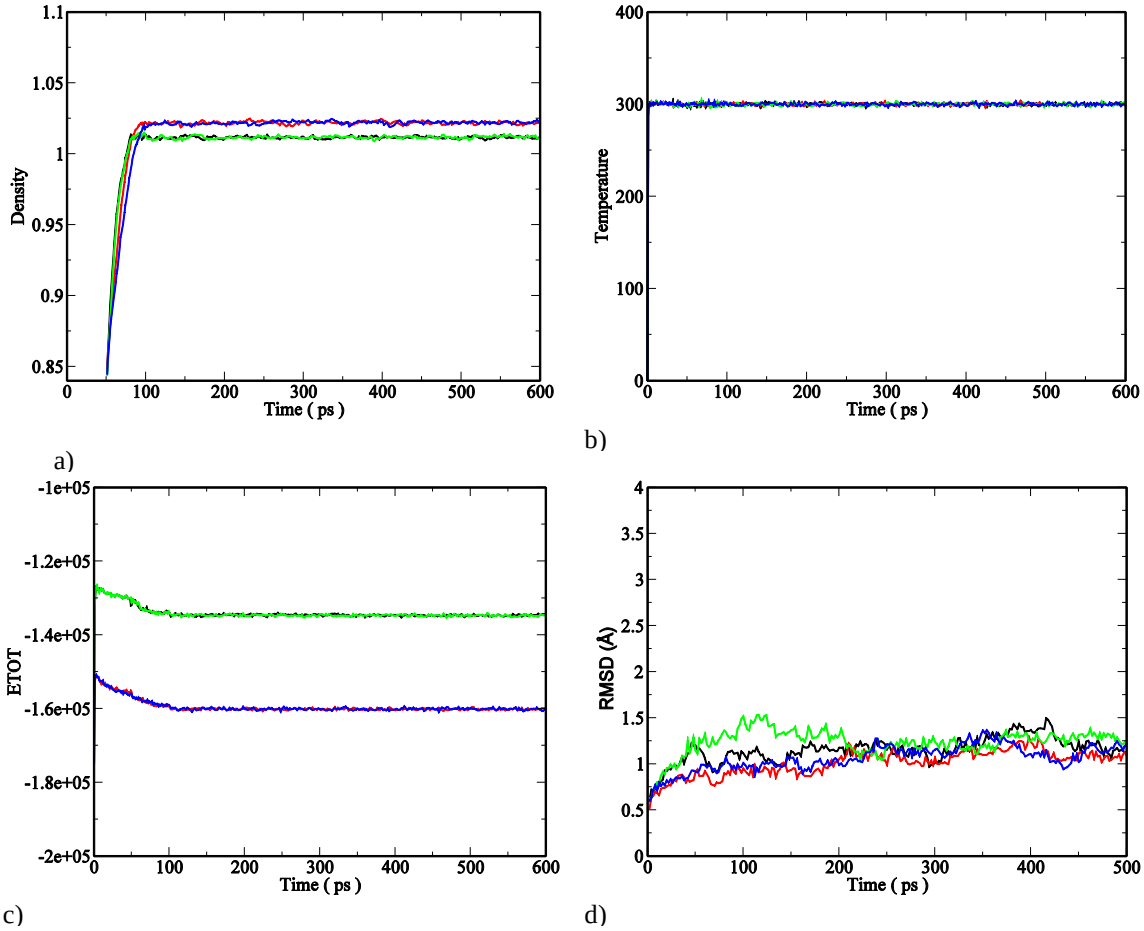


Figure 6-5: Trajectory curves for CDK2/Fascaplysin and CDK2/Carbofascaplysin equilibration simulations. These equilibration simulations were carried out with a short minimization, 50ps of heating and 50ps density equilibration followed by 500ps of constant pressure equilibration at 300K with weak restraints on the CDK2. Black and red lines represent the CDK2/carbofascaplysin simulations and green and blue lines represent CDK2/fascaplysin simulations with TIP3P and TIP4P-Ew models, respectivly.

The overlay of the 5ns production runs for CDK2/fascaplysin and CDK2/carbofascaplysin indicates that both complexes have attained an equilibrium and stability over the 5ns for all the four runs (Figure 6-6); however CDK2 simulations with carbofascaplysin appear slightly more stable compare to fascaplysin.

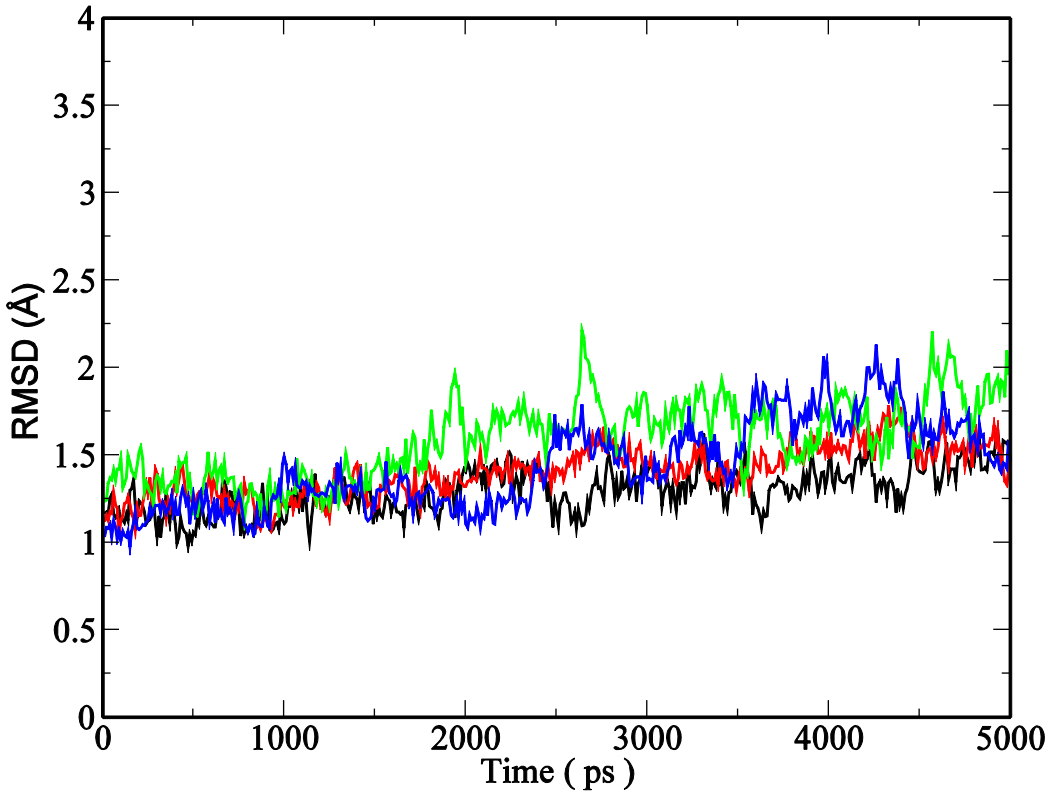


Figure 6-6: An overlay of 5ns CDK2/fascaplysin and CDK2/carbofascaplysin simulations. Red and black curves represent the CDK2/carbofascaplysin simulations and green and blue curves represent CDK2/fascaplysin simulations with TIP3P and TIP4P-Ew models, respectivly.

6.2.4 Molecular dynamics simulations and conformational stability of free CDK4

Conformational stability of free CDK4 was initially tested using TIP3P and TIP4P-Ew solvent boxes without crystal waters. In the follow up simulations the crystal waters corresponding to crystal buried waters situated in the CDK2 X-ray structure (see Section 6.3.2) were modelled in the CDK4 model.

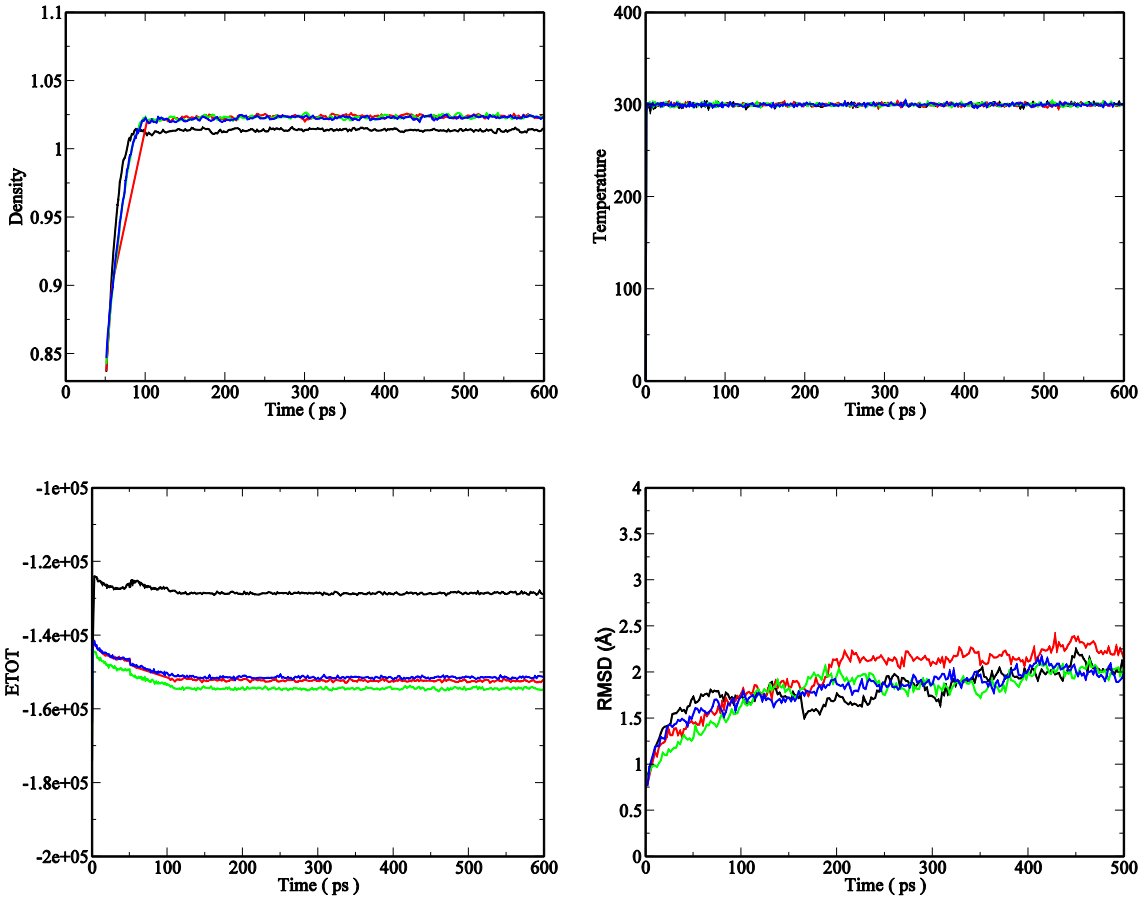


Figure 6-7 Trajectory curves for CDK4 equilibration simulations. Trajectory curves displaying stability of density, temperature, total energy for CDK4 model. These equilibration simulations were carried out with a short minimization, 50ps of heating and 50ps density equilibration followed by 500ps of constant pressure equilibration at 300K with weak restraints on the CDK2. Black and red lines represent the CDK4 simulations in TIP3P and TIP4P-Ew solvent box without crystal waters, respectively. CDK4 simulations in TIP4P-Ew are shown grean and blue with crystal waters.

The CDK4 structure (2W96) also has four buried water molecules at positions to the CDK2 water. The CDK4 simulations were carried out using TIP4P-Ew water model and crystal buried waters modelled into CDK4. The trajectories for CDK4 equilibration and production run, from four different experiments are shown in Figure 6-7 and Figure 6-8, respectively. The CDK4 equilibration and production run using TIP3P and TIP4P-Ew solvent box without crystal waters (shown in black and red), show a stability of the system. The trajectories of CDK4 MD simulations with TIP4P-Ew using crystal waters (shown in green Figure 6-8) show an increasing trend in the RMSD between during 3 to

5ns of the production run. An analysis in VMD reveals that the increased in RMSD is almost exclusively due to some variation and flexibility in the last eight residues of the C terminus of CDK4. In a follow up experiment of CDK4 MD simulations were carried out after removing these eight terminal residues. The trajectories (shown in blue Figure 6-8) for this system appears to be more stable. This system was finally selected for thermodynamic integration experiment. The average RMSD for selected run of CDK4 is 2.51 Å which seems very reasonable for a homology model (Park *et al.*, 2004).

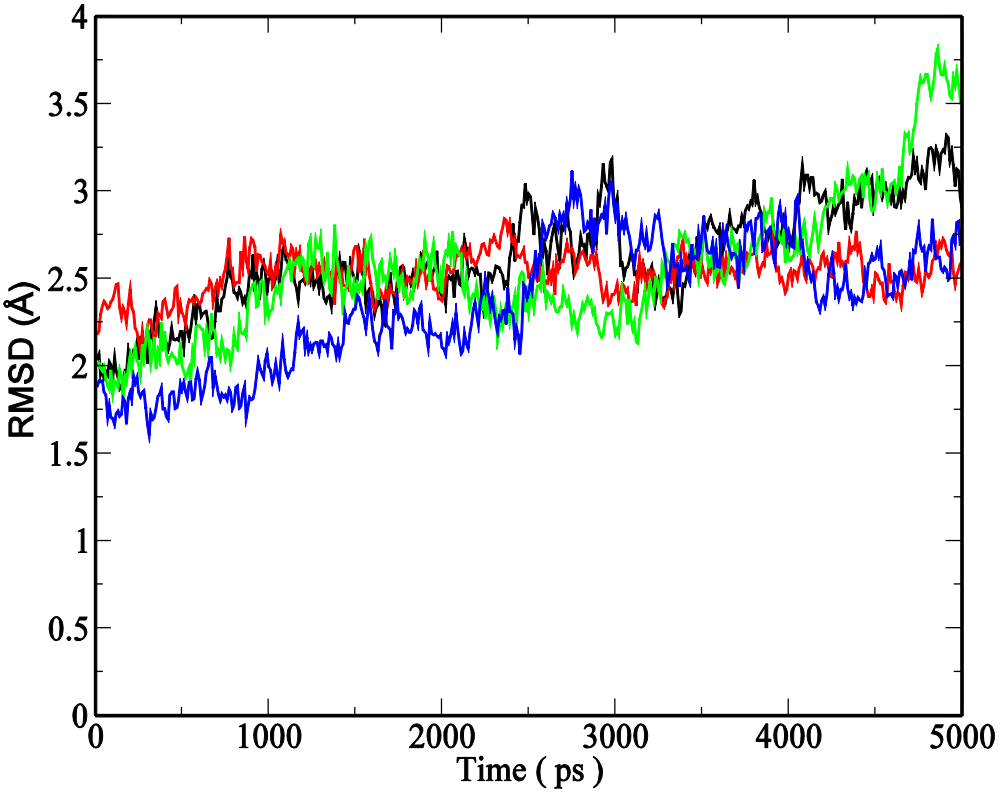


Figure 6-8: RMSD curves overlay for 5ns MD simulations of CDK4. Black and red lines represent the CDK4 simulations in TIP3P and TIP4P-Ew solvent box without crystal waters, respectively. CDK4 simulations in TIP4P-Ew using crystal waters are shown grean and blue.

6.2.5 Molecular Dynamics Simulations and conformational stability of CDK4/Carbofascaplysin and CDK4/Fascaplysin complexes

Similar to the earlier experiments 5ns of MD simulations were performed for CDK4/fascaplysin and CDK4/carbofascaplysin both with TIP4P-Ew solvent box. The docked complexes of CDK4/fascaplysin and CDK4/carbofascaplysin were prepared as described in Section 2.12. The nine waters crystal waters were kept in CDK4 model as described earlier.

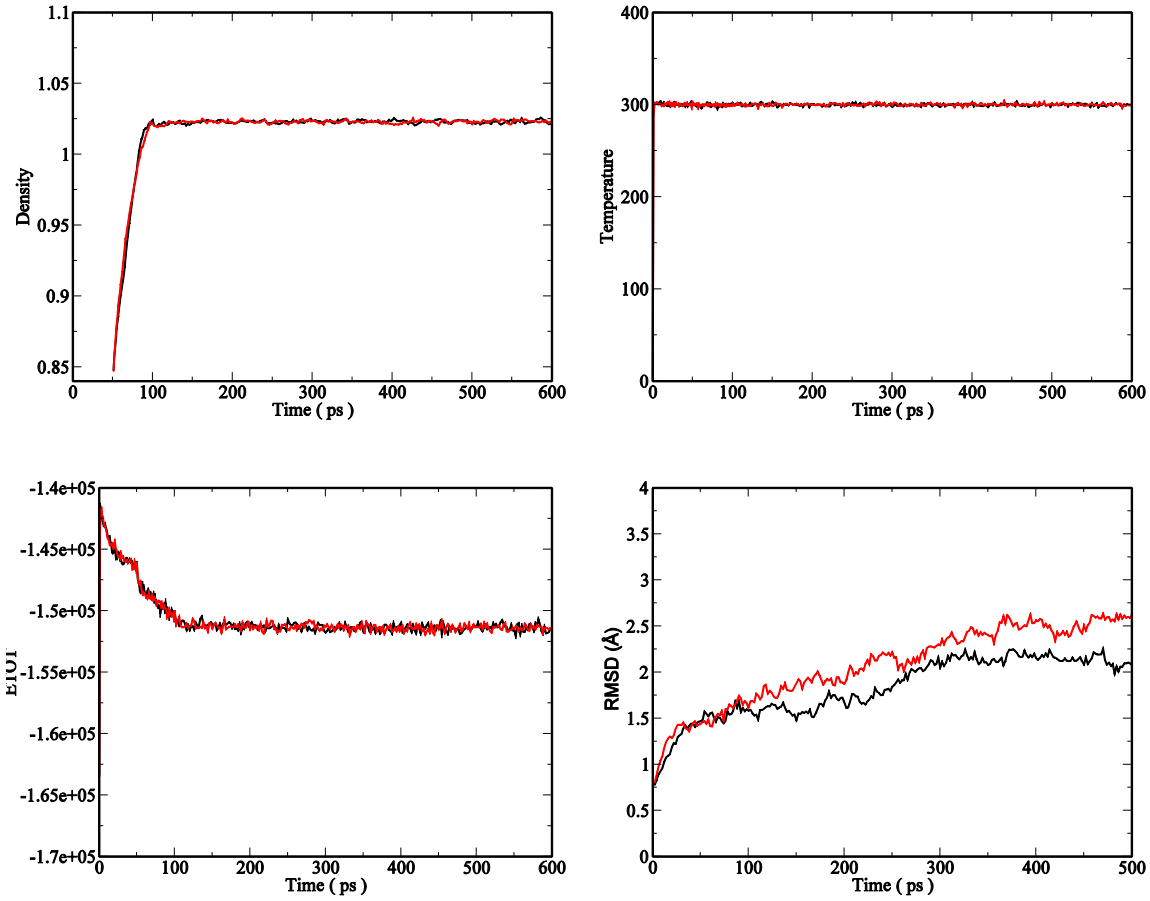


Figure 6-9 Trajectory curves for CDK4/Fascaplysin and CDK4/Carbofascaplysin equilibration simulations. These equilibration simulations were carried out with a short minimization, 50ps of heating and 50ps density equilibration followed by 500ps of constant pressure equilibration at 300K with weak restraints on the CDK4. Black and red lines represent the CDK4/carbofascaplysin CDK4/fascaplysin simulations with TIP4P-Ew models, respectivly.

The CDK4/fascaplysin and CDK4/carbofascaplysin complexes were solvated using tLeap (Section 2.13). Solvated structure of CDK4 complexes were equilibrated (Figure 6-9) before the 5ns production runs using the same steps as described earlier in details for free CDK2 complexes (Section 6.2.3). An overlay of 5ns production runs between CDK4/fascaplysin and CDK4/carbofascaplysin display indicates that the MD runs are stable.

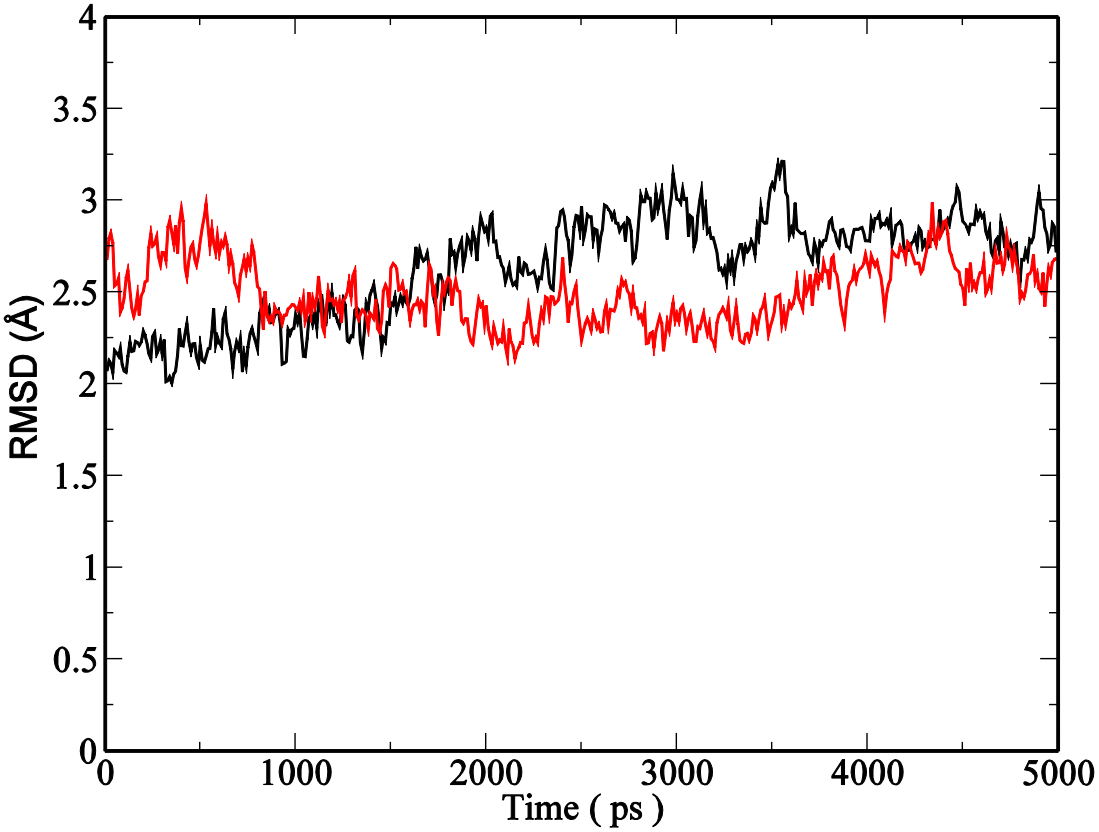


Figure 6-10: An overlay of 5ns CDK4/carbofascaplysin and CDK4/fascaplysin MD simlutions. Red and black curves represent the CDK4/carbofascaplysin CDK2/fascaplysin simulations with TIP4P-Ew, respectivly.

6.3 Calculation of free-energy differences by thermodynamic integration

Thermodynamic integration molecular dynamics simulations (see Section 2.14) were used to calculate the relative difference in the free energy of binding of fascaplysin and carbofascaplysin to CDK2 and CDK4. The hypothesis is that a positive charge accounts, at least to an extent, for fascaplysin selectivity for CDK4 over CDK2. Fascaplysin carries a formal positive charge at N11, hence a hypothetical neutral compound, carbofascaplysin, was prepared by isoelectronically replacing N11 with C11 to directly study the effect of positive charge on selectivity (Figure 6-11). The two ligands were sketched and parameterized with gaff atom types and resp charges were generated using antechamber on gaussian03 output files (see Section 2.13.2). The CDK2 X-ray structure 1FIN and the CDK4 homology model (see Section 2.9 and 4.8) both carrying buried crystal waters as described earlier were used for TI calculations, during the thermodynamic integration experiment 19 values for λ (0.05 to 0.95, 5 ns each window) were used to mix potential functions. The transition from carbofascaplysin to fascaplysin was studied for three systems, unbound ligands and ligands complexed with CDK2 and CDK4, respectively (Figure 6-14).

Figure 6-11 Fascaplysin and carbofascaplysin. Carbofascaplysin is a hypothetical compound with no charge which is sketeched by replacing N11 of fascaplysin with C11.

After running the MD simulations requested for thermodynamic integration a tremendous amount of data is obtained. Due to the small step size of changes during TI

calculations it is assumed that the system does not undergo significant conformational changes during the transformation. The stability curves for the TI transformations are shown in Figure 6-12 which displays 19 window simulations.

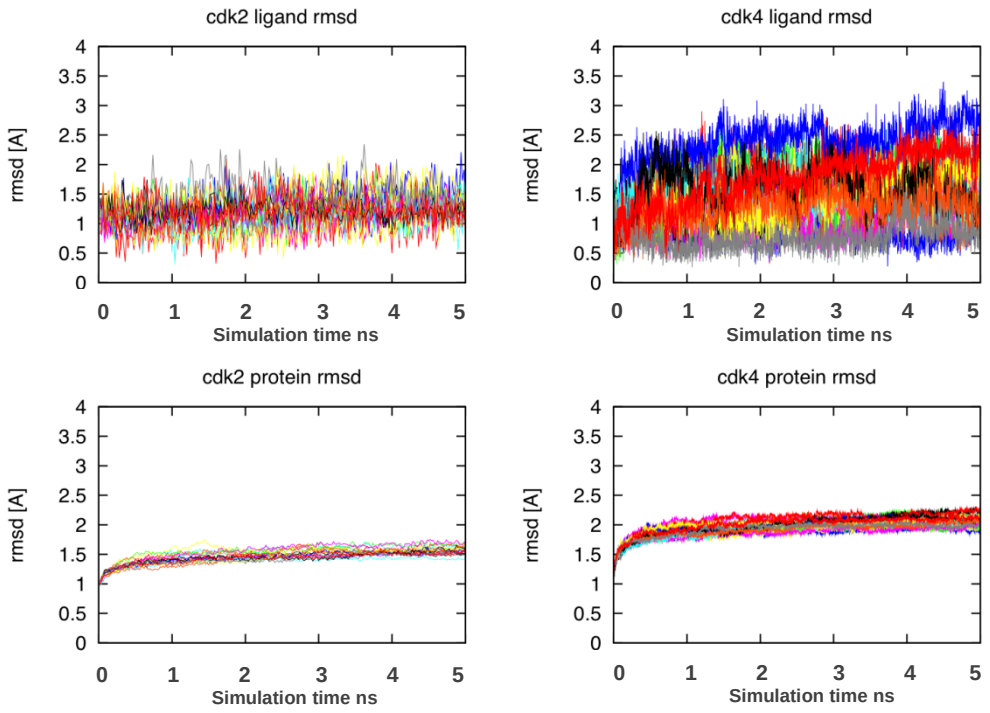


Figure 6-12 RMSD curves for transformations. These curves show the stability of the system during the thermodynamic integration

In order to obtain an estimate of the change of the relative binding free energy of the two ligands (fascaplysin and carbofascaplysin) the values of dV/dλ are obtained from the amber mdout files for transformations of carbofascaplysin to fascaplysin, carbofascaplysin to fascaplysin complexed with CDK2 and carbofascaplysin to fascaplysin complexed with CDK4. The plot of dV/dλ curves obtained from 19 λ points each for 200ps simulation (first step) are shown in Figure 6-13. The curves for transformations of carbofascaplysin to fascaplysin are smoother than those for the ligand change in the protein, which is due to the more

complex conformational landscape of a protein-ligand complex. Since thermodynamics integrations were run for 5ns there are 25 steps each with 200ps simulations.

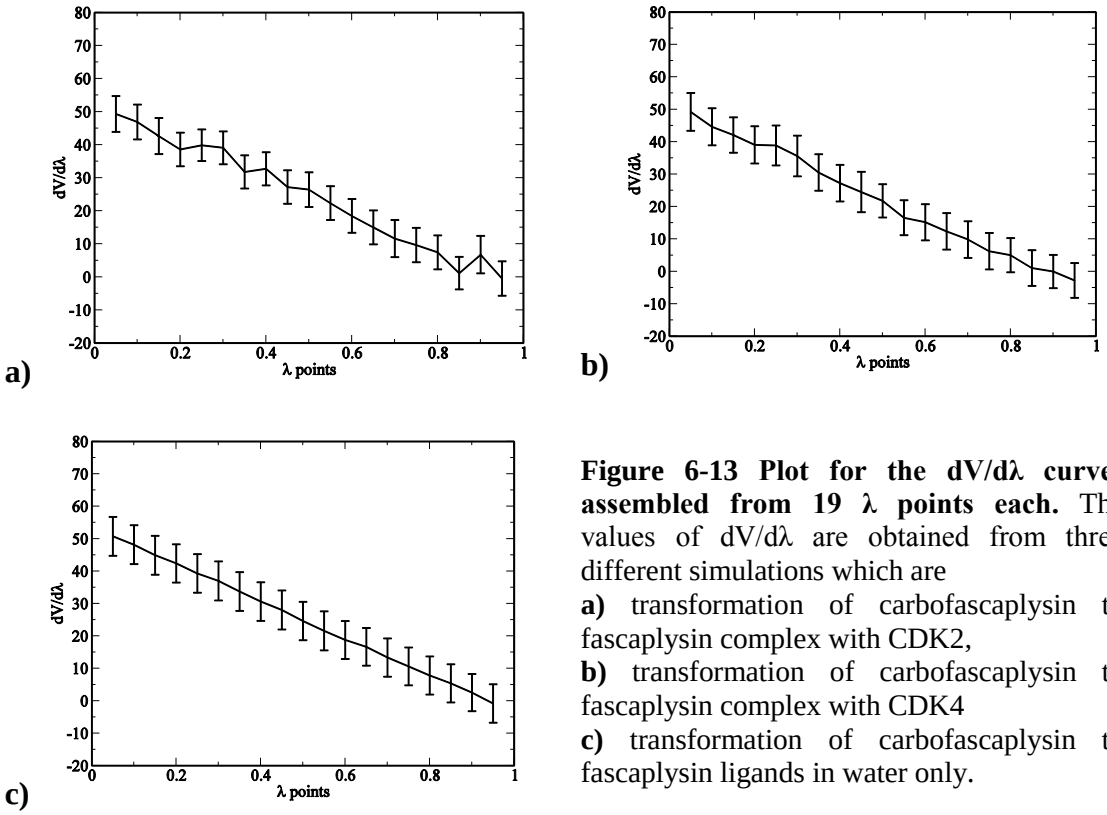


Figure 6-13 Plot for the dV/dλ curves assembled from 19 λ points each. The values of dV/dλ are obtained from three different simulations which are
a) transformation of carbofascaplysin to fascaplysin complex with CDK2,
b) transformation of carbofascaplysin to fascaplysin complex with CDK4
c) transformation of carbofascaplysin to fascaplysin ligands in water only.

The free energy change (Figure 6-14) associated with each of the steps was calculated by numerically integrating dV/dλ values for λ=0-1 (19 points). The ΔG values calculated by numerically integrating dV/dλ values of the first step (shown in Figure 6-13) are found as 24.431 kcal/mol, 22.0618 kcal/mol, and 25.009 kcal/mol for the transformation of carbofascaplysin to fascaplysin in CDK2, the transformation of carbofascaplysin to fascaplysin in CDK4 and transformation of carbofascaplysin to fascaplysin ligands only (Figure 6-14). Twenty five values of ΔG are derived from the integration of the dV/dλ from 19 λ points for each 200ps steps of 5ns thermodynamics integration simulations (Figure 6-15).

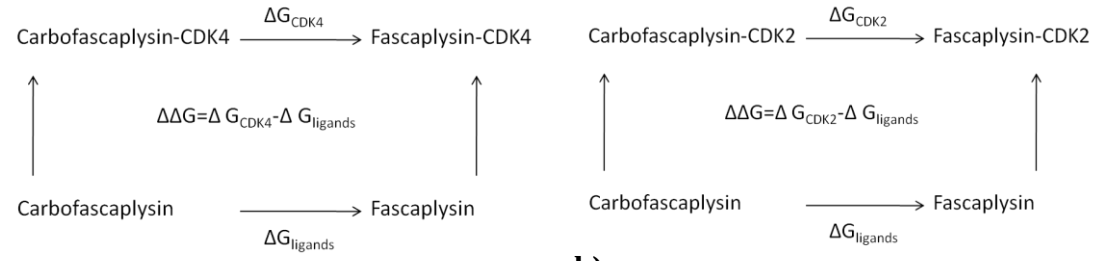


Figure 6-14 Thermodynamic cycle. a) Difference of binding free energy between the CDK2 complexes and carbofascaplysin and fascaplysin **b)** Difference of binding free energy between the CDK4 complexes and two ligands.

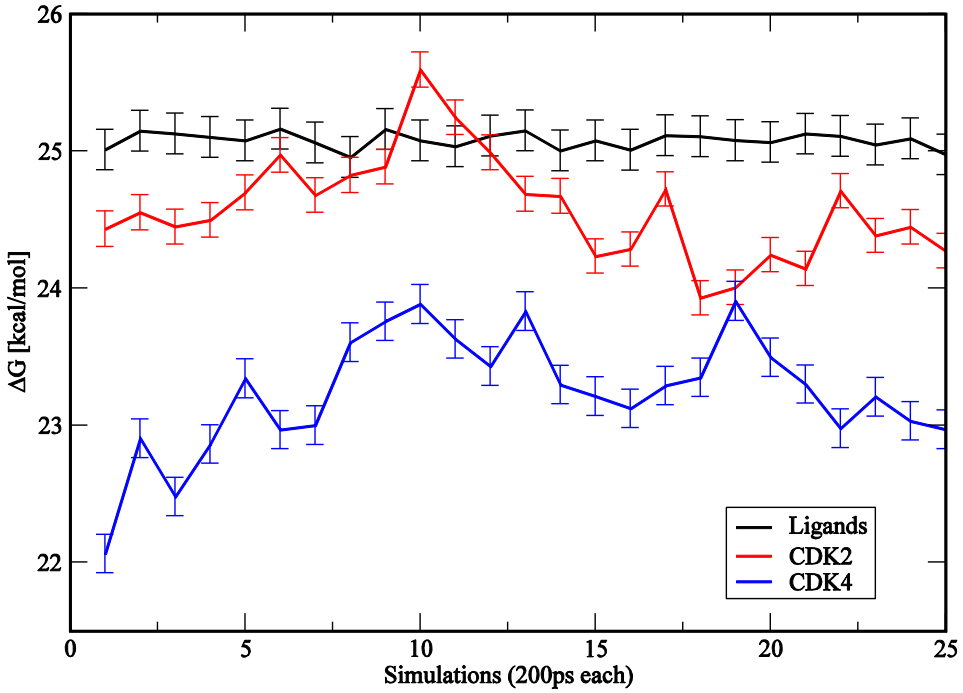


Figure 6-15 Free energy changes from three transformation. These curves represent the transformation of carbofascaplysin (ligand) to fascaplysin both in unbound and bound state with CDK2 and CDK4. Each point on these curve is obtained by the numerical integration of the dV/dλ from 19 λ points

The thermodynamic effect of transition from carbofascaplysin to fascaplysin is shown in Figure 6-15. It is apparent from the Figure 6-15 that transforming carbofascaplysin into fascaplysin is easier on CDK4 than on CDK2. The total free energy for the complexes is subject to some fluctuations, but both curves are clearly separated all the time. Total ΔΔG for the carbofascaplysin to fascaplysin transformation

complexed with CDK4 is 23.36 kcal/mol compared to 24.59 kcal/mol with CDK2. The ΔΔG for unbound ligand is calculated to be 25.08 kcal/mol. The ΔΔG for the relative preference of fascaplysin relative to carbofascaplysin for CDK4 is calculated as -1.72 kcal/mol and -0.49 kcal/mol for CDK2. These results indicate that CDK4 compared to CDK2 stabilises the "positive charge" by more than 1kcal/mol ΔG. Fascaplysin is therefore predicted to bind stronger to CDK4 than carbofascaplysin, confirming that fascaplysin is specific for CDK4. While these results do not fully explain the experimental IC_{50} data, however these results reveal the trend of relatively strong binding of fascaplysin the CDK4.

The results of the present study are consistent with the McInnes *et al* hypothesis that positively charged inhibitor show specificity for CDK4 (McInnes *et al.,* 2004). The calculation of relative free energies to predict the relative binding of enzyme and inhibitor binding has also been successfully tested with experimental results (Merz and Kollman, 1989; Ferguson *et al.*, 1991).

6.4 Conclusion

Molecular dynamics simulations try to predict the accurate physical properties of molecules for biological systems. The applications of molecular dynamics simulation play an important role in understanding the interactions of proteins with different inhibitors. The main objective of the present work was to understand the specificity of fascaplysin toward CDK4 compared to CDK2. The thermodynamic integration (TI) method is used to test the hypothesis that a positive charge on fascaplysin is responsible for its specificity toward CDK4. Prior to the TI experiment the stability of the CDK2, CDK4 and their complexes (with fascaplysin and carbofascaplysin) was tested with different water models in the presence and absence of crystallographic water molecules.

Energy, temperature, density and RMSD values of all the systems were stable during the 5ns MD simulations. Nine waters in the CDK2 are found as structurally conserved.

The calculation of difference of free energies explains the preferences of fascaplysin toward CDK4 due to a formal positive charge on fascaplysin. The ΔΔG for fascaplysin and carbofascaplysin binding for CDK4 is calculated as -1.72 kcal/mol and -0.49 kcal/mol for CDK2. Therefore CDK4 compared to CDK2 stabilises the "positive charge" by more than 1kcal/mol ΔG. Based on thermodynamics integration a result finding it is also proposed that new compounds with a positive charge may show increased selectivity toward CDK4.

Chapter Seven

CONCLUSION

Chapter 7 Conclusion

Cyclin dependent kinases and their activating partners' cyclins play a key role in the regulation of the cell division cycle. The possible role of CDKs in cancer and tumour development has prompted scientific efforts to identify, design and synthesize specific inhibitors for these CDKs that can be used in the treatment of different cancers. CDK4 is considered as a major target for cancer drug discovery as it controls the entry into the cell division cycle (Buolamwini, 2000; Malumbres and Barbacid, 2006). Some anti-cancer agents have the ability to inhibit multiple CDKs such as CDK1, CDK2 and CDK4 (Webster, 1998). However, there is a need to develop specific inhibitors to control the activity of individual CDK in different disease mechanisms and to minimize undesirable side effects by inhibiting the unrelated CDKs.

Sequential and structural information of CDKs active sites can help to design specific inhibitors. To achieve this, sequence and structure based analysis of active sites of all the available CDKs was performed. The active site of CDK2 was defined based on protein-ligand interaction information obtained from the plethora of CDK2 PDB structures. The comparison of the active site residues of all available PDB structures of CDK2 revealed two distinct orientations of His84 and some variation in the conformation of Lys33. Lys33 is an important residue in the CDK2 active site, as it is involved in salt bridge formation. In the active and inactive forms of CDK2 Lys33 forms a salt bridge with Glu51 and Asp145, respectively. Further more, active site analysis shows $Lys33^{CDK2}$ is fully conserved in all CDKs and CDK like proteins. The evaluation of the CDK2 and CDK4 active sites highlights His95 in CDK4 corresponding to Phe82 in CDK2. Apart from CDK4, only CDK6 has a histidine residue (His100) at the corresponding position. Glu144 is a key residue in the active site of CDK4 that is considered as responsible for CDK4 specificity toward positively

charged inhibitors (McInnes *et al.*, 2004). All members of the CDK family lack this residue at the corresponding position in their active sites; however a Glu residue is present in all CDKLs at this position in their active sites.

Understanding the CDK4 binding mode for different ligands and 3D structural information is very important for designing new inhibitors for CDK4. The utility of CDK4 homology model to design specific inhibitors has also been demonstrated (Aubry *et al.*, 2006; Horiuchi *et al.*, 2009). In the absence of any experimentally solved structures of CDK4, homology models based on CDK2/CDK6 templates were generated for later use in molecular modelling studies. It has been demonstrated that a target template sequence identity higher than 30% is indicative of good quality model (Martí-Renom *et al.*, 2000). The higher sequence identity (45%) between CDK2 and CDK4 suggest feasibility of this approach. After CDK4 X-ray structures became available (which only represents the inactive form) the structural information from CDK4 X-ray structures was incorporated into a hybrid model of CDK4 in a putative active form of CDK4.

Molecular docking studies provide three dimensional coordinates of ligand-receptor complex. While the ligand docking approach is widely used, it is still not without complications. Docking problem is divided into two parts which are pose prediction and scoring of these poses. The first part can be addressed to a greater extent by introducing full and partial flexibility of the ligand and protein, respectively, however ranking the poses generated by docking is still a challenging task. The performance of GOLD was evaluated against a selected set of 21 CDK2 structures from PDB to find the best set of parameters prior to its use in molecular docking of CDK4.

Fascaplysin inhibits CDK4 specifically. Fascaplysin was docked in order to explore the binding poses and in particular to study the role of $His95^{CDK4}$ toward

specificity of CDK4 compared with CDK2. This analysis revealed that $His95^{CDK4}$ can make an addition polar contact with fascaplysin while keeping the bidentate hydrogen bonds with Val96. However $His95^{CDK4}$ can only partly account for specificity toward CDK4. Fascaplysin inspired tryptamine derivatives were docked into CDK4 homology model. The docking results show a typical kinase inhibitor binding pose of these inhibitors. The current approach of combining homology modelling with molecular docking provides a useful tool for the predication of the binding pose. However good pose prediction does not always correlate with good scores. A comparison of ChemScore values obtained by docking of tryptamine derivatives with the experimental log IC50 values show a weak correlation with a regression coefficient r2=0.54. Therefore accurate ranking of CDK4 inhibitors according to their binding affinity using ChemScore remains a challenging task. It is proposed that in future work different docking algorithms, scoring functions or using a combination of different scoring functions (consensus scoring) could be tried.

Focusing on the optimization of CDK4 inhibitors, a rearrangement of the nitrogen in the third ring of β-carbolines from meta to ortho position is proposed to form α-carbolines. *In-silico* studies shows that these new compounds maintain a bidentate hydrogen pattern similar to fascaplysin and have *in-silico* improved binding properties.

In order to investigate the role of fascaplysin charge toward its binding affinity for CDK4, a thermodynamic integration experiment was designed. Facaplysin was transformed into the hypothetical compound carbofascaplysin with isoelectronic substitution of the charged nitrogen into an uncharged carbon atom. This transformation was carried out in the CDK2 complex, CDK4 complex and water. The calculation of difference of free energies changes during this transformation shows that fascaplysin

binds tightly to CDK4 compared to carbofascaplysin due to a formal positive charge on fascaplysin. The ΔΔG for fascaplysin and carbofascaplysin binding for CDK4 is calculated as -1.72 kcal/mol and -0.49 kcal/mol for CDK2. Therefore CDK4 compared to CDK2 stabilises the "positive charge" better by more than 1kcal/mol ΔG. Based on these finding it is concluded that affinity for CDK4 inhibitors can be increased by introducing positive charge in CDK4 inhibitors.

The present work has successfully established a system for thermodynamic integration experiments on CDK2 and CDK4. It will be important to further study the role of $Glu144^{CDK4}$ toward fascaplysin specificity using thermodynamics integration method. This can be achieved by transforming $Glu144^{CDK4}$ into Gln residue and calculating the free energy changes during this transformation. Similarly the role of $His95^{CDK4}$ can also be confirmed by mutating it to a Phe residue using thermodynamic integration. These future studies will provide a better understanding of CDK4 structural properties.

A complete understanding of structural and functional properties of CDK4 and related proteins will be useful to design novel inhibitors to control CDK4 activity during a disease mechanism.

References

References

Abagyan, R., Totrov, M. & Kuznetsov, D. 1994. ICM - A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. ***Journal of Computational Chemistry***, 15**,** 488-506.

Abascal, J. L. & Vega, C. 2007. Dipole-quadrupole force ratios determine the ability of potential models to describe the phase diagram of water. ***Phys Rev Lett***, 98**,** 237801.

Adachi, J. 1996. MOLPHY version 2.3 : programs for molecular phylogenetics based on maximum likelihood. ***Comput Sci Monogr***, 28**,** 1-150.

Adachi, J., Waddell, P. J., Martin, W. & Hasegawa, M. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. ***J Mol Evol***, 50**,** 348-58.

Aggarwal, S. 2010. Targeted cancer therapies. ***Nat Rev Drug Discov***, 9**,** 427-8.

Alberts, B., Wilson, J. H. & Hunt, T. 2008. *Molecular biology of the cell*, New York, Garland Science.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic Local Alignment Search Tool. ***Journal of Molecular Biology***, 215**,** 403-410.

Alzate-Morales, J. H., Contreras, R., Soriano, A., Tunon, I. & Silla, E. 2007. A computational study of the protein-ligand interactions in CDK2 inhibitors: using quantum mechanics/molecular mechanics interaction energy as a predictor of the biological activity. ***Biophys J***, 92**,** 430-9.

Amadasi, A., Surface, J. A., Spyrakis, F., Cozzini, P., Mozzarelli, A. & Kellogg, G. E. 2008. Robust classification of "relevant" water molecules in putative protein binding sites. ***J Med Chem***, 51**,** 1063-7.

Apweiler, R., Martin, M. J. & Uniprot Consortium 2010. The Universal Protein Resource (UniProt) in 2010. ***Nucleic acids research***, 38**,** D142-D148.

Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. ***Bioinformatics***, 22**,** 195-201.

Asano, T., Ikegaki, I., Satoh, S., Seto, M. & Sasaki, Y. 1998. A protein kinase inhibitor, fasudil (AT-877): A novel approach to signal transduction therapy. ***Cardiovascular Drug Reviews***, 16**,** 76-87.

Attwooll, C., Lazzerini Denchi, E. & Helin, K. 2004. The E2F family: specific functions and overlapping interests. ***EMBO J***, 23**,** 4709-16.

Aubry, C., Jenkins, P. R., Mahale, S., Chaudhuri, B., Marechal, J. D. & Sutcliffe, M. J. 2004. New fascaplysin-based CDK4-specific inhibitors: design, synthesis and biological activity. ***Chemical Communications,*** 1696-1697.

Aubry, C., Wilson, A. J., Emmerson, D., Murphy, E., Chan, Y. Y., Dickens, M. P., García, M. D., Jenkins, P. R., Mahale, S. & Chaudhuri, B. 2009. Fascaplysin-inspired diindolyls as selective inhibitors of CDK4/cyclin D1. ***Bioorganic & Medicinal Chemistry***, 17**,** 6073-84.

Aubry, C., Wilson, A. J., Jenkins, P. R., Mahale, S., Chaudhuri, B., Marechal, J. D. & Sutcliffe, M. J. 2006. Design, synthesis and biological activity of new CDK4-specific inhibitors, based on fascaplysin. ***Organic & Biomolecular Chemistry***, 4**,** 787-801.

Bai, C., Richman, R. & Elledge, S. J. 1994. Human cyclin F. ***EMBO J***, 13**,** 6087-98.

Bairoch, A. & Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. ***Nucleic Acids Res***, 28**,** 45-8.

Bairoch, A., Bougueleret, L. & Uniprot Consortium 2007. The universal protein resource (UniProt). ***Nucleic acids research***, 35**,** D193-D197.

Bajorath, J., Stenkamp, R. & Aruffo, A. 1993. Knowledge-based model building of proteins: Concepts and examples. ***Protein Sci***, 2**,** 1798-1810.

Baker, D. & Sali, A. 2001. Protein structure prediction and structural genomics. ***Science***, 294**,** 93-6.

Barillari, C., Taylor, J., Viner, R. & Essex, J. W. 2007. Classification of water molecules in protein binding sites. ***J Am Chem Soc***, 129**,** 2577-87.

Bartek, J., Bartkova, J. & Lukas, J. 1996a. The retinoblastoma protein pathway and the restriction point. ***Curr Opin Cell Biol***, 8**,** 805-14.

Bartek, J., Lukas, L. & Strauss, M. 1996b. [Control mechanisms of cell transition from the G1 phase to the S phase--the R point]. ***Cas Lek Cesk***, 135**,** 634-5.

Baxter, C. A., Murray, C. W., Clark, D. E., Westhead, D. R. & Eldridge, M. D. 1998. Flexible docking using Tabu search and an empirical estimate of binding affinity. ***Proteins-Structure Function and Genetics***, 33**,** 367-382.

Bergers, G., Hanahan, D. & Coussens, L. M. 1998. Angiogenesis and apoptosis are cellular parameters of neoplastic progression in transgenic mouse models of tumorigenesis. ***Int J Dev Biol***, 42**,** 995-1002.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. 2000. The Protein Data Bank. ***Nucleic Acids Research***, 28**,** 235-242.

Berthet, C. & Kaldis, P. 2006. Cdk2 and Cdk4 cooperatively control the expression of Cdc2. ***Cell Div***, 1**,** 10.

Blagosklonny, M. V. & Pardee, A. B. 2002. The restriction point of the cell cycle. ***Cell Cycle***, 1**,** 103-10.

Blair, C. & Murphy, R. W. 2010. Recent Trends in Molecular Phylogenetic Analysis: Where to Next? ***J Hered***.

Blair, J. E. & Hedges, S. B. 2005. Molecular phylogeny and divergence times of deuterostome animals. ***Mol Biol Evol***, 22**,** 2275-84.

Boguski, M. S., Lowe, T. M. & Tolstoshev, C. M. 1993. dbEST--database for "expressed sequence tags". ***Nat Genet***, 4**,** 332-3.

Bohm, H. J. 1994. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. ***J Comput Aided Mol Des***, 8**,** 243-56.

Bossemeyer, D. 1995. Protein-Kinases - Structure and Function. ***Febs Letters***, 369**,** 57-61.

Bottoms, C. A., White, T. A. & Tanner, J. J. 2006. Exploring structurally conserved solvent sites in protein families. ***Proteins***, 64**,** 404-21.

Brocchieri, L. 2001. Phylogenetic inferences from molecular sequences: review and critique. ***Theor Popul Biol***, 59**,** 27-40.

Brooijmans, N. & Kuntz, I. D. 2003. Molecular recognition and docking algorithms. ***Annu Rev Biophys Biomol Struct***, 32**,** 335-73.

Brooks, B. R., Brooks, C. L., 3rd, Mackerell, A. D., Jr., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L.,

Wu, X., Yang, W., York, D. M. & Karplus, M. 2009. CHARMM: the biomolecular simulation program. ***J Comput Chem***, 30**,** 1545-614.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. 1983. CHARMM : A program for macromolecular energy, minimization, and dynamics calculations. ***Journal of Computational Chemistry***, 4**,** 187-217.

Buolamwini, J. K. 2000. Cell cycle molecular targets in novel anticancer drug discovery. ***Curr Pharm Des***, 6**,** 379-92.

Burns, G., Daoud, R. & Vaigl, J. Year. LAM: An Open Cluster Environment for MPI. *In:* Proceedings of Supercomputing Symposium, 1994. 379-386.

Canutescu, A. A., Shelenkov, A. A. & Dunbrack, R. L., Jr. 2003. A graph-theory algorithm for rapid protein side-chain prediction. ***Protein Sci***, 12**,** 2001-14.

Capdeville, R., Buchdunger, E., Zimmermann, J. & Matter, A. 2002. Glivec (ST1571, Imatinib), a rationally developed, targeted anticancer drug. ***Nature Reviews Drug Discovery***, 1**,** 493-502.

Carlson, B. A., Dubay, M. M., Sausville, E. A., Brizuela, L. & Worland, P. J. 1996. Flavopiridol induces G1 arrest with inhibition of cyclin-dependent kinase (CDK) 2 and CDK4 in human breast carcinoma cells. ***Cancer Res***, 56**,** 2973-8.

Carvajal, R. D., Tse, A., Shah, M. A., Lefkowitz, R. A., Gonen, M., Gilman-Rosen, L., Kortmansky, J., Kelsen, D. P., Schwartz, G. K. & O'reilly, E. M. 2009. A Phase II Study of Flavopiridol (Alvocidib) in Combination with Docetaxel in Refractory, Metastatic Pancreatic Cancer. ***Pancreatology***, 9**,** 404-409.

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B. & Woods, R. J. 2005. The Amber biomolecular simulation programs. ***Journal of Computational Chemistry***, 26**,** 1668-88.

Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Crowley, M., Walker, R. C., Zhang, W., Merz, K. M., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossvary, I., Wong, K. F., Paesani, F., Vanicek, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V. & Kollman, P. A. 2008. *AMBER 10*, University of California.

Cavalier-Smith, T. 2006. Cell evolution and Earth history: stasis and revolution. ***Philos Trans R Soc Lond B Biol Sci***, 361**,** 969-1006.

Cavasotto, C. N. & Orry, A. J. 2007. Ligand docking and structure-based virtual screening in drug discovery. ***Curr Top Med Chem***, 7**,** 1006-14.

Charifson, P. S., Corkery, J. J., Murcko, M. A. & Walters, W. P. 1999. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. ***J Med Chem***, 42**,** 5100-9.

Cheng, K.-Y., Noble, M. E. M., Skamnaki, V., Brown, N. R., Lowe, E. D., Kontogiannis, L., Shen, K., Cole, P. A., Siligardi, G. & Johnson, L. N. 2006. The Role of the Phospho-CDK2/Cyclin A Recruitment Site in Substrate Recognition. ***J. Biol. Chem.***, 281**,** 23167-23179.

Chothia, C. L., A. 1986. The relation between the divergence of sequence and structure in proteins. ***EMBO J.***, 5**,** 823-826.

Christian, B. A., Grever, M. R., Byrd, J. C. & Lin, T. S. 2009. Flavopiridol in chronic lymphocytic leukemia: a concise review. ***Clin Lymphoma Myeloma***, 9 Suppl 3**,** S179-85.

Chung, E., Henriques, D., Renzoni, D., Zvelebil, M., Bradshaw, J. M., Waksman, G., Robinson, C. V. & Ladbury, J. E. 1998. Mass spectrometric and thermodynamic

studies reveal the role of water molecules in complexes formed between SH2 domains and tyrosyl phosphopeptides. ***Structure***, 6**,** 1141-51.
Cismowski, M. J., Laff, G. M., Solomon, M. J. & Reed, S. I. 1995. KIN28 encodes a C-terminal domain kinase that controls mRNA transcription in Saccharomyces cerevisiae but lacks cyclin-dependent kinase-activating kinase (CAK) activity. ***Mol Cell Biol***, 15**,** 2983-92.
Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. 2004. The Jalview Java alignment editor. ***Bioinformatics***, 20**,** 426-427.
Cohen, P. 2002a. The origins of protein phosphorylation. ***Nat Cell Biol***, 4**,** 127-130.
Cohen, P. 2002b. Protein kinases - the major drug targets of the twenty-first century? ***Nat Rev Drug Discov***, 1**,** 309-315.
Colless, D. H. 1985. On the Status of Outgroups in Phylogenetics. ***Systematic biology***, 34**,** 364-366.
Collett, M. S. & Erikson, R. L. 1978. Protein kinase activity associated with the avian sarcoma virus src gene product. ***Proc Natl Acad Sci U S A***, 75**,** 2021-4.
Congreve, M., Murray, C. W. & Blundell, T. L. 2005. Structural biology and drug discovery. ***Drug Discov Today***, 10**,** 895-907.
Connell-Crowley, L., Solomon, M. J., Wei, N. & Harper, J. W. 1993. Phosphorylation independent activation of human cyclin-dependent kinase 2 by cyclin A in vitro. ***Mol. Biol. Cell***, 4**,** 79-92.
Cooper, G. M. & Hausman, R. E. 2009. *The cell : a molecular approach*, Washington, D.C. Sunderland, Mass., ASM Press ; Sinauer Associates.
Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. 1995. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. ***Journal of the American Chemical Society***, 117**,** 5179-5197.
Corsino, P., Horenstein, N., Ostrov, D., Rowe, T., Law, M., Barrett, A., Aslanidi, G., Cress, W. D. & Law, B. 2009. A novel class of cyclin-dependent kinase inhibitors identified by molecular docking act through a unique mechanism. ***J Biol Chem***, 284**,** 29945-55.
Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. 2004. WebLogo: a sequence logo generator. ***Genome Res***, 14**,** 1188-90.
Crosignani, P. G. 2003. Breast cancer and hormone-replacement therapy in the Million Women Study. ***Maturitas***, 46**,** 91-92.
D'angiolella, V., Donato, V., Vijayakumar, S., Saraf, A., Florens, L., Washburn, M. P., Dynlacht, B. & Pagano, M. 2010. SCFCyclin F controls centrosome homeostasis and mitotic fidelity through CP110 degradation. ***Nature***, 466**,** 138-142.
Daly, M., Chaudhuri, A., Gusmão, L. & Rodríguez, E. 2008. Phylogenetic relationships among sea anemones (Cnidaria: Anthozoa: Actiniaria). ***Mol Phylogenet Evol***, 48**,** 292-301.
Davidson, G. & Niehrs, C. 2010. Emerging links between CDK cell cycle regulators and Wnt signaling. ***Trends in Cell Biology***, 20**,** 453-460.
Davidson, G., Shen, J., Huang, Y. L., Su, Y., Karaulanov, E., Bartscherer, K., Hassler, C., Stannek, P., Boutros, M. & Niehrs, C. 2009. Cell cycle control of wnt receptor activation. ***Dev Cell***, 17**,** 788-99.
Davies, P. & Lineweaver, C. 2005. Finding a second sample of life on Earth. ***Astrobiology***, 5**,** 154-163.

Davies, T. G., Tunnah, P., Meijer, L., Marko, D., Eisenbrand, G., Endicott, J. A. & Noble, M. E. 2001. Inhibitor binding to active and inactive CDK2: the crystal structure of CDK2-cyclin A/indirubin-5-sulphonate. ***Structure***, 9**,** 389-97.

Davis, A. M., Teague, S. J. & Kleywegt, G. J. 2003. Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. ***Angew Chem Int Ed Engl***, 42**,** 2718-36.

Day, P. J., Cleasby, A., Tickle, I. J., O'reilly, M., Coyle, J. E., Holding, F. P., Mcmenamin, R. L., Yon, J., Chopra, R., Lengauer, C. & Jhoti, H. 2009. Crystal structure of human CDK4 in complex with a D-type cyclin. ***Proceedings of the National Academy of Sciences***, 106**,** 4166-4170.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. 1978. A model of evolutionary change in proteins. ***Atlas of protein sequence and structure***, 5**,** 345-351.

De Bondt, H. L., Rosenblatt, J., Jancarik, J., Jones, H. D., Morgant, D. O. & Kim, S.-H. 1993. Crystal structure of cyclin-dependent kinase 2. ***Nature***, 363**,** 595-602.

Delano, W. L. 2002. The PyMOL Molecular Graphics System ***DeLano Scientific, Palo Alto, CA, USA.***

Desai, D., Gu, Y. & Morgan, D. O. 1992. Activation of Human Cyclin-Dependent Kinases Invitro. ***Molecular Biology of the Cell***, 3**,** 571-582.

Deshpande, A., Sicinski, P. & Hinds, P. W. 2005a. Cyclins and cdks in development and cancer: a perspective. ***Oncogene***, 24**,** 2909-15.

Deshpande, A., Sicinski, P. & Hinds, P. W. 2005b. Cyclins and cdks in development and cancer: a perspective. ***Oncogene***, 24**,** 2909-2915.

Dimmic, M. W., Rest, J. S., Mindell, D. P. & Goldstein, R. A. 2002. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. ***J Mol Evol***, 55**,** 65-73.

Dominy, B. N. 2008. Molecular recognition and binding free energy calculations in drug development. ***Current Pharmaceutical Biotechnology***, 9**,** 87-95.

Druker, B. J. 2002. Imatinib and chronic myeloid leukemia: validating the promise of molecularly targeted therapy. ***European Journal of Cancer***, 38**,** S70-S76.

Dubowchik, G. M. & Walker, M. A. 1999. Receptor-mediated and enzyme-dependent targeting of cytotoxic anticancer drugs. ***Pharmacology & Therapeutics***, 83**,** 67-123.

Dunbrack, R. L., Jr. & Karplus, M. 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. ***J Mol Biol***, 230**,** 543-74.

Dyson, N. 1998. The regulation of E2F by pRB-family proteins. ***Genes Dev.***, 12**,** 2245-2262.

Eddy, S. 1996. Hidden markov models. ***Current Opinion in Structural Biology***.

Eddy, S. 1998a. Profile hidden Markov models. ***Bioinformatics***.

Eddy, S. R. 1998b. Profile hidden Markov models. ***Bioinformatics***.

Edgar, B. A. & Lehner, C. F. 1996. Developmental control of cell cycle regulators: a fly's perspective. ***Science***, 274**,** 1646-52.

Edgar, R. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. ***Nucleic acids research***, 32**,** 1792.

Edwards, S. V. 2009. Natural selection and phylogenetic analysis. ***Proc Natl Acad Sci U S A***, 106**,** 8799-800.

Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V. & Mee, R. P. 1997. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. ***J Comput Aided Mol Des***, 11**,** 425-45.

Elledge, S. J. & Spottswood, M. R. 1991. A new human p34 protein kinase, CDK2, identified by complementation of a cdc28 mutation in Saccharomyces cerevisiae, is a homolog of Xenopus Egl. ***The EMBO journal***.

Eswar, N., Webb, B., Marti-Renom, M., Madhusudhan, M., Eramian, D., Shen, M., Pieper, U. & Sali, A. 2007. Comparative protein structure modeling using Modeller. ***Curr Protoc Protein Sci***.

Evans, T., Rosenthal, E. T., Youngblom, J., Distel, D. & Hunt, T. 1983. Cyclin: A protein specified by maternal mRNA in sea urchin eggs that is destroyed at each cleavage division. ***Cell***, 33**,** 389-396.

Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. ***Methods Enzymol***, 266**,** 418-27.

Ferguson, D. M., Radmer, R. J. & Kollman, P. A. 1991. Determination of the relative binding free energies of peptide inhibitors to the HIV-1 protease. ***J Med Chem***, 34**,** 2654-9.

Fiser, A. & Sali, A. 2003. Modeller: generation and refinement of homology-based protein structure models. ***Methods Enzymol***, 374**,** 461-91.

Forrest, L. R., Tang, C. L. & Honig, B. 2006. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. ***Biophys J***, 91**,** 508-17.

Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., Sanschagrin, P. C. & Mainz, D. T. 2006. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. ***J Med Chem***, 49**,** 6177-96.

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E. & Robb, M. A. 2004. *Gaussian 03, Revision B.03,* Wallingford, CT 06492 USA, Gaussian, Inc.

Fu, T.-J., Peng, J., Lee, G., Price, D. H. & Flores, O. 1999. Cyclin K Functions as a CDK9 Regulatory Subunit and Participates in RNA Polymerase II Transcription. ***Journal of Biological Chemistry***, 274**,** 34527-34530.

Funato, N., Takayanagi, H., Konda, Y., Toda, Y., Harigaya, Y., Iwai, Y. & Omura, S. 1994. Absolute Configuration of Staurosporine By X-Ray Analysis. ***Tetrahedron letters***, 35**,** 1251-1254.

Fung, T. K. & Poon, R. Y. 2005. A roller coaster ride with the mitotic cyclins. ***Semin Cell Dev Biol***, 16**,** 335-42.

García, M. D., Wilson, A. J., Emmerson, D. P. G., Jenkins, P. R., Mahale, S. & Chaudhuri, B. 2006. Synthesis, crystal structure and biological activity of beta-carboline based selective CDK4-cyclin D1 inhibitors. ***Organic & biomolecular chemistry***, 4**,** 4478-84.

Gohlke, H., Hendlich, M. & Klebe, G. 2000. Knowledge-based scoring function to predict protein-ligand interactions. ***J Mol Biol***, 295**,** 337-56.

Gohlke, H. & Klebe, G. 2001. Statistical potentials and scoring functions applied to protein-ligand binding. ***Curr Opin Struct Biol***, 11**,** 231-5.

Gohlke, H. & Klebe, G. 2002. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. ***Angew Chem Int Ed Engl***, 41**,** 2644-76.

Gomi, H., Sassa, T., Thompson, R. F. & Itohara, S. 2010. Involvement of cyclin-dependent kinase-like 2 in cognitive function required for contextual and spatial learning in mice. ***Front Behav Neurosci***, 4**,** 17.

Gouda, H., Kuntz, I. D., Case, D. A. & Kollman, P. A. 2003. Free energy calculations for theophylline binding to an RNA aptamer: MM-PBSA and comparison of thermodynamic integration methods. ***Biopolymers***, 68**,** 16-34.

Graf, F., Koehler, L., Kniess, T., Wuest, F., Mosch, B. & Pietzsch, J. 2009. Cell Cycle Regulating Kinase Cdk4 as a Potential Target for Tumor Cell Treatment and Tumor Imaging. ***Journal of Oncology***, 2009**,** 1-13.

Graham, S. W., Olmstead, R. G. & Barrett, S. C. 2002. Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. ***Mol Biol Evol***, 19**,** 1769-81.

Grant, S. 2009. Therapeutic Protein Kinase Inhibitors. ***Cellular and Molecular Life Sciences***, 66**,** 1163-1177.

Grant, S. & Roberts, J. D. 2003. The use of cyclin-dependent kinase inhibitors alone or in combination with established cytotoxic drugs in cancer chemotherapy. ***Drug Resist Updat***, 6**,** 15-26.

Green, M. C., Murray, J. L. & Hortobagyi, G. N. 2000. Monoclonal antibody therapy for solid tumors. ***Cancer Treatment Reviews***, 26**,** 269-286.

Grillo, M., Bott, M. J., Khandke, N., Mcginnis, J. P., Miranda, M., Meyyappan, M., Rosfjord, E. C. & Rabindran, S. K. 2006. Validation of cyclin D1/CDK4 as an anticancer drug target in MCF-7 breast cancer cells: Effect of regulated overexpression of cyclin D1 and siRNA-mediated inhibition of endogenous cyclin D1 and CDK4 expression. ***Breast Cancer Res Treat***, 95**,** 185-94.

Gu Y, R. J., Morgan Do 1992. Cell cycle regulation of CDK2 activity by phosphorylation of Thr160 and Tyr15. ***EMBO J.***, 11**,** 3995-4005.

Guha, R., Howard, M. T., Hutchison, G. R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J. & Willighagen, E. L. 2006. Blue Obelisk - Interoperability in chemical informatics. ***Journal of Chemical Information and Modeling***, 46**,** 991-998.

Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. 2002. Principles of docking: An overview of search algorithms and a guide to scoring functions. ***Proteins***, 47**,** 409-43.

Hamdouchi, C., Zhong, B., Mendoza, J., Collins, E., Jaramillo, C., De Diego, J. E., Robertson, D., Spencer, C. D., Anderson, B. D., Watkins, S. A., Zhang, F. & Brooks, H. B. 2005. Structure-based design of a new class of highly selective aminoimidazo[1,2-a]pyridine-based inhibitors of cyclin dependent kinases. ***Bioorg Med Chem Lett***, 15**,** 1943-7.

Hanahan, D. & Weinberg, R. A. 2000. The hallmarks of cancer. ***Cell***, 100**,** 57-70.

Hanks, S. K. & Hunter, T. 1995. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. ***Faseb Journal***, 9**,** 576-596.

Harper, J. W. & Adams, P. D. 2001. Cyclin-dependent kinases. ***Chem Rev***, 101**,** 2511-26.

Hartshorn, M. J., Verdonk, M. L., Chessari, G., Brewerton, S. C., Mooij, W. T., Mortenson, P. N. & Murray, C. W. 2007. Diverse, high-quality test set for the validation of protein-ligand docking performance. ***J Med Chem***, 50**,** 726-41.

Hawasli, A. H., Koovakkattu, D., Hayashi, K., Anderson, A. E., Powell, C. M., Sinton, C. M., Bibb, J. A. & Cooper, D. C. 2009. Regulation of hippocampal and behavioral excitability by cyclin-dependent kinase 5. ***PLoS One***, 4**,** e5808.

Hengstschlager, M., Braun, K., Soucek, T., Miloloza, A. & Hengstschlager-Ottnad, E. 1999. Cyclin-dependent kinases at the G1-S transition of the mammalian cell cycle. ***Mutat Res***, 436**,** 1-9.

Henikoff, S. & Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. ***Proc Natl Acad Sci U S A***, 89**,** 10915-9.

Herbert, A. & Sternberg, M. J. E. 2008. *MaxCluster - A tool for Protein Structure Comparison and Clustering* [Online]. Available: http://www.sbg.bio.ic.ac.uk/~maxcluster [Accessed].

Hess, B., Kutzner, C., Van Der Spoel, D. & Lindahl, E. 2008. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. ***Journal of Chemical Theory and Computation***, 4**,** 435-447.

Hidaka, H., Inagaki, M., Kawamoto, S. & Sasaki, Y. 1984. Isoquinolinesulfonamides, novel and potent inhibitors of cyclic nucleotide-dependent protein kinase and protein kinase C. ***Biochemistry***, 23**,** 5036-5041.

Hillisch, A., Pineda, L. F. & Hilgenfeld, R. 2004. Utility of homology models in the drug discovery process. ***Drug Discov Today***, 9**,** 659-69.

Hofmann, F. & Livingston, D. M. 1996. Differential effects of cdk2 and cdk3 on the control of pRb and E2F function during G1 exit. ***Genes Dev***, 10**,** 851-61.

Honma, T., Yoshizumi, T., Hashimoto, N., Hayashi, K., Kawanishi, N., Fukasawa, K., Takaki, T., Ikeura, C., Ikuta, M., Suzuki-Takahashi, I., Hayama, T., Nishimura, S. & Morishima, H. 2001. A Novel Approach for the Development of Selective Cdk4 Inhibitors: Library Design Based on Locations of Cdk4 Specific Amino Acid Residues. ***Journal of medicinal chemistry***, 44**,** 4628-4640.

Hooft, R. W., Vriend, G., Sander, C. & Abola, E. E. 1996. Errors in protein structures. ***Nature***, 381**,** 272.

Hopkins, A. L. & Groom, C. R. 2002. The druggable genome. ***Nature Reviews Drug Discovery***, 1**,** 727-730.

Horiuchi, T., Nagata, M., Kitagawa, M., Akahane, K. & Uoto, K. 2009. Discovery of novel thieno[2,3-d]pyrimidin-4-yl hydrazone-based inhibitors of cyclin D1-CDK4: synthesis, biological evaluation and structure-activity relationships. Part 2. ***Bioorg Med Chem***, 17**,** 7850-60.

Hormann, A., Chaudhuri, B. & Fretz, H. 2001. DNA binding properties of the marine sponge pigment fascaplysin. ***Bioorganic & Medicinal Chemistry***, 9**,** 917-921.

Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L. & Head-Gordon, T. 2004. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. ***J Chem Phys***, 120**,** 9665-78.

Huelsenbeck, J. & Ronquist, F. 2005. Bayesian analysis of molecular evolution using MrBayes. ***Statistical methods in molecular evolution. New York: Springer,*** 183–232.

Huelsenbeck, J. P. & Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. ***Bioinformatics***, 17**,** 754-5.

Huguet, R. R. 2008. *Study of the regulation and signalling of Cdk2-cyclin O complexes during apoptosis*. Ph.D, Universitat Pompeu Fabra.

Humphrey, W., Dalke, A. & Schulten, K. 1996. VMD: Visual molecular dynamics. ***Journal of Molecular Graphics***, 14**,** 33-&.

Hunter, T. 2000. Signaling—2000 and beyond. ***Cell***.

Huwe, A. 2003. Small molecules as inhibitors of cyclin-dependent kinases. ***Angewandte Chemie - International Edition***, 42**,** 2122-2138.

Hypercube Inc 2008. HyperChem(TM) Professional 8.014 Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA.

Ikuta, M., Kamata, K., Fukasawa, K., Honma, T., Machida, T., Hirai, H., Suzuki-Takahashi, I., Hayama, T. & Nishimura, S. 2001. Crystallographic approach to identification of cyclin-dependent kinase 4 (CDK4)-specific inhibitors by using CDK4 mimic CDK2 protein. ***J Biol Chem***, 276**,** 27548-54.

Isabelle, D., Philippe, O. & Roland, W. 2007. Origins of Life and Biochemistry under High-Pressure Conditions. ***ChemInform***, 38.

Jacobson M., S. A. 2004. Comparative Protein Structure Modeling and Its Applications to Drug Discovery. ***Annual Reports in Medicinal Chemistry***, 39**,** 259-276.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F. & Gerstein, M. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. ***Science***, 302**,** 449-53.

Jeffrey, P. D., Russo, A. A., Polyak, K., Gibbs, E., Hurwitz, J., Massague, J. & Pavletich, N. P. 1995a. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. ***Nature***, 376**,** 313-320.

Jeffrey, P. D., Russo, A. A., Polyak, K., Gibbs, E., Hurwitz, J., Massagué, J. & Pavletich, N. P. 1995b. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. ***Nature***, 376**,** 313-20.

Jenkins, P. R., Wilson, J., Emmerson, D., Garcia, M. D., Smith, M. R., Gray, S. J., Britton, R. G., Mahale, S. & Chaudhuri, B. 2008. Design, synthesis and biological evaluation of new tryptamine and tetrahydro-beta-carboline-based selective inhibitors of CDK4. ***Bioorganic & Medicinal Chemistry***, 16**,** 7728-39.

Jiang, M., Gao, Y., Yang, T., Zhu, X. & Chen, J. 2009. Cyclin Y, a novel membrane-associated cyclin, interacts with PFTK1. ***FEBS Lett***, 583**,** 2171-8.

Jirawatnotai, S., Aziyu, A., Osmundson, E. C., Moons, D. S., Zou, X., Kineman, R. D. & Kiyokawa, H. 2004. Cdk4 Is Indispensable for Postnatal Proliferation of the Anterior Pituitary. ***Journal of Biological Chemistry***, 279**,** 51100-51106.

Johnson, D. G. & Walker, C. L. 1999. Cyclins and cell cycle checkpoints. ***Annu Rev Pharmacol Toxicol***, 39**,** 295-312.

Jones, D. T., Taylor, W. R. & Thornton, J. M. 1994. A mutation data matrix for transmembrane proteins. ***FEBS Lett***, 339**,** 269-75.

Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. 1997. Development and validation of a genetic algorithm for flexible docking. ***J Mol Biol***, 267**,** 727-48.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. 1983. Comparison of Simple Potential Functions for Simulating Liquid Water. ***Journal of Chemical Physics***, 79**,** 926-935.

Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. 1996. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. ***Journal of the American Chemical Society***, 118**,** 11225-11236.

Juan, D., Pazos, F. & Valencia, A. 2008. Co-evolution and co-adaptation in protein networks. ***FEBS Lett***, 582**,** 1225-30.

Kaldis, P. & Pagano, M. 2009. Wnt signaling in mitosis. ***Dev Cell***, 17**,** 749-50.

Kastan, M. B. & Bartek, J. 2004. Cell-cycle checkpoints and cancer. ***Nature***, 432**,** 316-23.

Kato, H., Faria, T. N., Stannard, B., Roberts, C. T., Jr. & Leroith, D. 1993. Role of tyrosine kinase activity in signal transduction by the insulin-like growth factor-I (IGF-I) receptor. Characterization of kinase-deficient IGF-I receptors and the action of an IGF-I-mimetic antibody (alpha IR-3). ***J Biol Chem***, 268**,** 2655-61.

Kavraki, L. 2007. Protein-Ligand Docking, Including Flexible Receptor-Flexible Ligand Docking. ***Connexions***.

Kellenberger, E., Rodrigo, J., Muller, P. & Rognan, D. 2004. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. ***Proteins-Structure Function and Bioinformatics***, 57**,** 225-242.

Kitagawa, M., Okabe, T., Ogino, H., Matsumoto, H., Suzuki-Takahashi, I., Kokubo, T., Higashi, H., Saitoh, S., Taya, Y., Yasuda, H. & Et Al. 1993. Butyrolactone I, a selective inhibitor of cdk2 and cdc2 kinase. ***Oncogene***, 8**,** 2425-32.

Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. 2004. Docking and scoring in virtual screening for drug discovery: methods and applications. ***Nat Rev Drug Discov***, 3**,** 935-49.

Kleywegt, G. J., Harris, M. R., Zou, J. Y., Taylor, T. C., Wahlby, A. & Jones, T. A. 2004. The Uppsala Electron-Density Server. ***Acta Crystallogr D Biol Crystallogr***, 60**,** 2240-9.

Koike, A., Nakai, K. & Takagi, T. 2001. The origin and evolution of eukaryotic protein kinases. ***GENOME INFORMATICS SERIES*****,** 392-393.

Kollman, P. 1993. Free-Energy Calculations - Applications to Chemical and Biochemical Phenomena. ***Chemical Reviews***, 93**,** 2395-2417.

Kontopidis, G., Mcinnes, C., Pandalaneni, S. R., Mcnae, I., Gibson, D., Mezna, M., Thomas, M., Wood, G., Wang, S., Walkinshaw, M. D. & Fischer, P. M. 2006. Differential binding of inhibitors to active and inactive CDK2 provides insights for drug design. ***Chem Biol***, 13**,** 201-11.

Koshland, D. E. 2002. Special essay. The seven pillars of life. ***Science***, 295**,** 2215-6.

Krakauer, D. C. & Sasaki, A. 2002. Noisy clues to the origin of life. ***Proc Biol Sci***, 269**,** 2423-8.

Kramer, B., Rarey, M. & Lengauer, T. 1999. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. ***Proteins***, 37**,** 228-41.

Krammer, A., Kirchhoff, P. D., Jiang, X., Venkatachalam, C. M. & Waldman, M. 2005. LigScore: a novel scoring function for predicting binding affinities. ***J Mol Graph Model***, 23**,** 395-407.

Kroemer, R. T., Vulpetti, A., Mcdonald, J. J., Rohrer, D. C., Trosset, J. Y., Giordanetto, F., Cotesta, S., Mcmartin, C., Kihlen, M. & Stouten, P. F. 2004. Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations. ***J Chem Inf Comput Sci***, 44**,** 871-81.

Kumar, S., Nei, M., Dudley, J. & Tamura, K. 2008. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. ***Briefings in Bioinformatics***, 9**,** 299-306.

Kuntz, I. D. 1992. Structure-based strategies for drug design and discovery. ***Science***, 257**,** 1078-82.

Kuzmich, A. S., Fedorov, S. N., Shastina, V. V., Shubina, L. K., Radchenko, O. S., Balaneva, N. N., Zhidkov, M. E., Park, J. I., Kwak, J. Y. & Stonik, V. A. 2010. The anticancer activity of 3- and 10-bromofascaplysins is mediated by caspase-8, -9, -3-dependent apoptosis. ***Bioorg Med Chem***, 18**,** 3834-40.

Lacrima, K., Valentini, A., Lambertini, C., Taborelli, M., Rinaldi, A., Zucca, E., Catapano, C., Cavalli, F., Gianella-Borradori, A., Maccallum, D. E. & Bertoni, F. 2005. In vitro activity of cyclin-dependent kinase inhibitor CYC202 (Seliciclib, R-roscovitine) in mantle cell lymphomas. ***Ann Oncol***, 16**,** 1169-76.

Lang, P. T., Brozell, S. R., Mukherjee, S., Pettersen, E. F., Meng, E. C., Thomas, V., Rizzo, R. C., Case, D. A., James, T. L. & Kuntz, I. D. 2009. DOCK 6: combining techniques to model RNA-small molecule complexes. ***RNA***, 15**,** 1219-30.

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. & Higgins, D. G. 2007. Clustal W and clustal X version 2.0. ***Bioinformatics***, 23**,** 2947-2948.

Larochelle, S., Merrick, K. A., Terret, M. E., Wohlbold, L., Barboza, N. M., Zhang, C., Shokat, K. M., Jallepalli, P. V. & Fisher, R. P. 2007. Requirements for Cdk7 in the assembly of Cdk1/cyclin B and activation of Cdk2 revealed by chemical genetics in human cells. ***Mol Cell***, 25**,** 839-50.

Laskowski, R. A., Macarthur, M. W., Moss, D. S. & Thornton, J. M. 1993. Procheck - a Program to Check the Stereochemical Quality of Protein Structures. ***Journal of Applied Crystallography***, 26**,** 283-291.

Laurent-Puig, P. & Zucman-Rossi, J. 2006. Genetics of hepatocellular tumors. ***Oncogene***, 25**,** 3778-86.

Le Guilloux, V., Schmidtke, P. & Tuffery, P. 2009. Fpocket: an open source platform for ligand pocket detection. ***BMC Bioinformatics***, 10**,** 168.

Levitzki, A. 2002. Tyrosine kinases as targets for cancer therapy. ***European Journal of Cancer***, 38**,** S11-S18.

Lin, T. S., Fischer, B., Blum, K. A., Andritsos, L. A., Jones, J. A., Moran, M. E., Broering, S., Heerema, N. A., Lozanski, G., Schaaf, L. J., Mahoney, L. S., Johnson, A. J., Smith, L. L., Wagner, A. J., Raymond, C. A., Phelps, M., Dalton, J. T., Grever, M. R. & Byrd, J. C. 2007. Preliminary results of a phase II study of flavopiridol (Alvocidib) in, relapsed chronic lymphocytic leukemia (CLL): Confirmation of clinical activity in high-risk patients and achievement of complete responses (CR). ***Blood***, 110**,** 913a-913a.

Lin, T. S., Heerema, N. A., Lozanski, G., Fischer, B., Blum, K. A., Andritsos, L. A., Jones, J. A., Flynn, J. M., Moran, M. E., Mitchell, S., Johnson, A. J., Phelps, M. A., Grever, M. R. & Byrd, J. C. 2008. Flavopiridol (Alvocidib) Induces Durable Responses in Relapsed Chronic Lymphocytic Leukemia (CLL) Patients with High-Risk Cytogenetic Abnormalities. ***Blood***, 112**,** 23-24.

Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. ***Advanced Drug Delivery Reviews***, 23**,** 3-25.

Liu, J. & Kipreos, E. T. 2000a. Evolution of cyclin-dependent kinases (CDKs) and CDK-activating kinases (CAKs): differential conservation of CAKs in yeast and metazoa. ***Mol Biol Evol***, 17**,** 1061-1074.

Liu, J. & Kipreos, E. T. 2000b. Evolution of Cyclin-Dependent Kinases (CDKs) and CDK-Activating Kinases (CAKs): Differential Conservation of CAKs in Yeast and Metazoa. ***Molecular Biology and Evolution***, 17**,** 1061-1074.

Losiewicz, M. D., Carlson, B. A., Kaur, G., Sausville, E. A. & Worland, P. J. 1994. Potent inhibition of CDC2 kinase activity by the flavonoid L86-8275. ***Biochem Biophys Res Commun***, 201**,** 589-95.

Lu, H., Chang, D. J., Baratte, B., Meijer, L. & Schulze-Gahmen, U. 2005. Crystal structure of a human cyclin-dependent kinase 6 complex with a flavonol inhibitor, fisetin. ***J Med Chem***, 48**,** 737-43.

Lybrand, T. P. 1995. Ligand-protein docking and rational drug design. ***Curr Opin Struct Biol***, 5**,** 224-8.

Maddison, D. R. & Schulz, K.-S. 2007. *The Tree of Life Web Project. Internet address* [Online]. Available: http://tolweb.org [Accessed].

Mahale, S., Aubry, C., James Wilson, A., Jenkins, P. R., Maréchal, J.-D., Sutcliffe, M. J. & Chaudhuri, B. 2006a. CA224, a non-planar analogue of fascaplysin, inhibits Cdk4 but not Cdk2 and arrests cells at G0/G1 inhibiting pRB phosphorylation. ***Bioorganic & Medicinal Chemistry Letters***, 16**,** 4272-8.

Mahale, S., Aubry, C., Jenkins, P. R., Maréchal, J.-D., Sutcliffe, M. J. & Chaudhuri, B. 2006b. Inhibition of cancer cell growth by cyclin dependent kinase 4 inhibitors synthesized based on the structure of fascaplysin. ***Bioorg Chem***, 34**,** 287-97.

Malumbres, M. & Barbacid, M. 2005. Mammalian cyclin-dependent kinases. ***Trends in biochemical sciences***, 30**,** 630-41.

Malumbres, M. & Barbacid, M. 2006. Is Cyclin D1-CDK4 kinase a bona fide cancer target? ***Cancer Cell***, 9**,** 2-4.

Malumbres, M., Harlow, E., Hunt, T., Hunter, T., Lahti, J. M., Manning, G., Morgan, D. O., Tsai, L.-H. & Wolgemuth, D. J. 2009. Cyclin-dependent kinases: a family portrait. ***Nature cell biology***, 11**,** 1275-1276.

Malumbres, M., Pevarello, P., Barbacid, M. & Bischoff, J. R. 2008. CDK inhibitors in cancer therapy: what is next? ***Trends in Pharmacological Sciences***, 29**,** 16-21.

Malumbres, M., Sotillo, R. O., Santamar1 a, D., Galán, J., Cerezo, A., Ortega, S., Dubus, P. & Barbacid, M. 2004. Mammalian Cells Cycle without the D-Type Cyclin-Dependent Kinases Cdk4 and Cdk6. ***Cell***, 118**,** 493-504.

Manning, G., Plowman, G. D., Hunter, T. & Sudarsanam, S. 2002a. Evolution of protein kinase signaling from yeast to man. ***Trends in Biochemical Sciences***, 27**,** 514-520.

Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. 2002b. The Protein Kinase Complement of the Human Genome. ***Science***, 298**,** 1912-1934.

Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F. & Sali, A. 2000. Comparative protein structure modeling of genes and genomes. ***Annual Review of Biophysics and Biomolecular Structure***, 29**,** 291-325.

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F. & Sali, A. 2000. Comparative protein structure modeling of genes and genomes. ***Annual Review of Biophysics and Biomolecular Structure***, 29**,** 291-325.

Martin, A. C., Macarthur, M. W. & Thornton, J. M. 1997. Assessment of comparative modeling in CASP2. ***Proteins***, Suppl 1**,** 14-28.

Martin, D. M. A., Miranda-Saavedra, D. & Barton, G. J. 2009. Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. ***Nucleic Acids Research***, 37**,** D244-D250.

Mcgann, M. R., Almond, H. R., Nicholls, A., Grant, J. A. & Brown, F. K. 2003. Gaussian docking functions. ***Biopolymers***, 68**,** 76-90.

Mcinnes, C., Wang, S., Anderson, S., O'boyle, J., Jackson, W., Kontopidis, G., Meades, C., Mezna, M., Thomas, M., Wood, G., Lane, D. P. & Fischer, P. M. 2004. Structural determinants of CDK4 inhibition and design of selective ATP competitive inhibitors. ***Chemistry & Biology***, 11**,** 525-34.

Meijer, L., Skaltsounis, A. L., Magiatis, P., Polychronopoulos, P., Knockaert, M., Leost, M., Ryan, X. P., Vonica, C. A., Brivanlou, A., Dajani, R., Crovace, C., Tarricone, C., Musacchio, A., Roe, S. M., Pearl, L. & Greengard, P. 2003. GSK-3-selective inhibitors derived from Tyrian purple indirubins. ***Chem Biol***, 10**,** 1255-66.

Merz, K. M. & Kollman, P. A. 1989. Free energy perturbation simulations of the inhibition of thermolysin: prediction of the free energy of binding of a new inhibitor. ***Journal of the American Chemical Society***, 111**,** 5649-5658.

Miranda-Saavedra, D. & Barton, G. J. 2007. Classification and functional annotation of eukaryotic protein kinases. ***Proteins***, 68**,** 893-914.

Mohan, V., Gibbs, A. C., Cummings, M. D., Jaeger, E. P. & Desjarlais, R. L. 2005. Docking: successes and challenges. ***Curr Pharm Des***, 11**,** 323-33.

Mojzsis, S. J., Arrhenius, G., Mckeegan, K. D., Harrison, T. M., Nutman, A. P. & Friend, C. R. 1996. Evidence for life on Earth before 3,800 million years ago. ***Nature***, 384**,** 55-9.

Morgan, D. 1997. CYCLIN-DEPENDENT KINASES: Engines, Clocks, and Microprocessors. ***Annual Review of Cell and Developmental Biology***, 13**,** 261-291.

Morgan, D. O. 1995. Principles of CDK regulation. ***Nature***, 374**,** 131-4.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K. & Olson, A. J. 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. ***Journal of Computational Chemistry***, 19**,** 1639-1662.

Morris, G. M., Huey, R. & Olson, A. J. 2008. Using AutoDock for ligand-receptor docking. ***Curr Protoc Bioinformatics***, Chapter 8**,** Unit 8 14.

Mpamhanga, C. P., Chen, B., Mclay, I. M., Ormsby, D. L. & Lindvall, M. K. 2005. Retrospective docking study of PDE4B ligands and an analysis of the behavior of selected scoring functions. ***J Chem Inf Model***, 45**,** 1061-74.

Muegge, I. 2001. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state.

Muegge, I., Martin, Y. C., Hajduk, P. J. & Fesik, S. W. 1999. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. ***J Med Chem***, 42**,** 2498-503.

Muller, T. & Vingron, M. 2000. Modeling amino acid replacement. ***J Comput Biol***, 7**,** 761-76.

Murray, A. W. 2004. Recycling the cell cycle: cyclins revisited. ***Cell***, 116**,** 221-34.

Narayanan Eswar, B. W., Marc A. Marti-Renom, M.S. Madhusudhan, David Eramian, Min-Yi Shen, Ursula Pieper, and Andrej Sali 2006. *Comparative Protein Structure Modeling Using Modeller*, John Wiley & Sons, Inc. .

Nasmyth, K. 2002. Segregating sister genomes: the molecular biology of chromosome separation. ***Science***, 297**,** 559-65.

Nei, M. 1996. Phylogenetic analysis in molecular evolutionary genetics. ***Annu Rev Genet***, 30**,** 371-403.

Nei, M. & Kumar, S. 2000. Molecular evolution and phylogenetics. 333.

Nigg, E. A. 1995. Cyclin-dependent protein kinases: key regulators of the eukaryotic cell cycle. ***BioEssays***, 17**,** 471-80.

Norbury, C. & Nurse, P. 1992. Animal cell cycles and their control. ***Annual Review of Biochemistry***, 61**,** 441-468.

Nowicki, M. W. & Walkinshaw, M. D. 2010. CDK9 inhibitors push cancer cells over the edge. ***Chem Biol***, 17**,** 1047-8.

Nugent, J. H. A., Alfa, C. E., Young, T. & Hyams, J. S. 1991. Conserved Structural Motifs in Cyclins Identified by Sequence-Analysis. ***Journal of Cell Science***, 99**,** 669-674.

Ohshima, T., Ward, J. M., Huh, C. G., Longenecker, G., Veeranna, Pant, H. C., Brady, R. O., Martin, L. J. & Kulkarni, A. B. 1996. Targeted disruption of the cyclin-dependent kinase 5 gene results in abnormal corticogenesis, neuronal pathology and perinatal death. ***Proc Natl Acad Sci U S A***, 93**,** 11173-8.

Okamoto, A., Demetrick, D. J., Spillare, E. A., Hagiwara, K., Hussain, S. P., Bennett, W. P., Forrester, K., Gerwin, B., Serrano, M., Beach, D. H. & Et Al. 1994. Mutations and altered expression of p16INK4 in human cancer. ***Proc Natl Acad Sci U S A***, 91**,** 11045-9.

Omura, S., Iwai, Y., Hirano, A., Nakagawa, A., Awaya, J., Tsuchya, H., Takahashi, Y. & Masuma, R. 1977. A new alkaloid AM-2282 OF Streptomyces origin. Taxonomy, fermentation, isolation and preliminary characterization. ***J Antibiot (Tokyo)***, 30**,** 275-82.

Orgel, L. E. 1998. The Origin of Life–How Long did it Take? ***Origins of Life and Evolution of Biospheres***.

Oró, J. & Lazcano, A. 1997. Comets and the origin and evolution of life. ***Comets and the Origin and Evolution of Life***.

Ortega, S., Malumbres, M. & Barbacid, M. 2002a. Cell Cycle and Cancer: The G1 Restriction Point and the G1 / S Transition. ***Current Genomics***, 3**,** 245-263.

Ortega, S., Malumbres, M. & Barbacid, M. 2002b. Cyclin D-dependent kinases, INK4 inhibitors and cancer. ***Biochim Biophys Acta***, 1602**,** 73-87.

Pace, N. R. 2009. Mapping the Tree of Life: Progress and Prospects. ***Microbiology and Molecular Biology Reviews***, 73**,** 565-576.

Page, R. D. M. 1996. TreeView: An application to display phylogenetic trees on personal computers. ***Computer Applications in the Biosciences***, 12**,** 357-358.

Paglini, G. & Caceres, A. 2001. The role of the Cdk5--p35 kinase in neuronal development. ***Eur J Biochem***, 268**,** 1528-33.

Palmero, I. & Peters, G. 1996. Perturbation of cell cycle regulators in human cancer. ***Cancer Surv***, 27**,** 351-67.

Park, H., Yeom, M. S. & Lee, S. 2004. Loop flexibility and solvent dynamics as determinants for the selective inhibition of cyclin-dependent kinase 4: comparative molecular dynamics simulation studies of CDK2 and CDK4. ***Chembiochem***, 5**,** 1662-72.

Pazos, F., Juan, D., Izarzugaza, J. M., Leon, E. & Valencia, A. 2008. Prediction of protein interaction based on similarity of phylogenetic trees. ***Methods Mol Biol***, 484**,** 523-35.

Penny, D., Hendy, M. D., Zimmer, E. A. & Hamby, R. K. 1990. Trees from sequences: panacea or pandora's box? ***Australian Systematic Botany***, 3**,** 21-38.

Petřek, M., Otyepka, M., Banáš, P. & Košinová, P. 2006. CAVER: a new tool to explore routes from protein clefts, pockets and cavities. ***BMC ….***

Pettersen, E. F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. 2004. A Visualization System for Exploratory Research and Analysis. ***J. Comput. Chem.***, 25**,** 1605-1612.

Poomsawat, S., Buajeeb, W., Khovidhunkit, S. O. & Punyasingh, J. 2010. Alteration in the expression of cdk4 and cdk6 proteins in oral cancer and premalignant lesions. ***J Oral Pathol Med***, 39**,** 793-9.

Poornima, C. S. & Dean, P. M. 1995. Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions. ***J Comput Aided Mol Des***, 9**,** 500-12.

Quirke, N. & Jacucci, G. 1982. Energy Difference Functions in Monte-Carlo Simulations - Application to (1) the Calculation of the Free-Energy of Liquid-Nitrogen, (2) the Fluctuation of Monte-Carlo Averages. ***Molecular Physics***, 45**,** 823-838.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. 1990. Stereochemistry of Polypeptide-Chain Configurations. ***Current Science***, 59**,** 813-817.

Rane, S. G., Dubus, P., Mettus, R. V., Galbreath, E. J., Boden, G., Reddy, E. P. & Barbacid, M. 1999. Loss of Cdk4 expression causes insulin-deficient diabetes and Cdk4 activation results in beta-islet cell hyperplasia. ***Nat Genet***, 22**,** 44-52.

Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. 1996. A fast flexible docking method using an incremental construction algorithm. ***Journal of Molecular Biology***, 261**,** 470-489.

Ravishanker, G., Mezei, M. & Beveridge, D. L. 1986. Conformational Stability and Flexibility of the Ala Dipeptide in Free Space and Water - Monte-Carlo Computer-Simulation Studies. ***Journal of Computational Chemistry***, 7**,** 345-348.

Renan, M. J. 1993. How many mutations are required for tumorigenesis? Implications from human cancer data. ***Mol Carcinog***, 7**,** 139-46.

Repasky, M. P., Shelley, M. & Friesner, R. A. 2007. Flexible ligand docking with Glide. ***Curr Protoc Bioinformatics***, Chapter 8**,** Unit 8 12.

Rivadeneira, D. B., Mayhew, C. N., Thangavel, C., Sotillo, E., Reed, C. A., Grana, X. & Knudsen, E. S. 2010. Proliferative suppression by CDK4/6 inhibition: complex function of the retinoblastoma pathway in liver tissue and hepatoma cells. ***Gastroenterology***, 138**,** 1920-30.

Rizzolio, F., Tuccinardi, T., Caligiuri, I., Lucchetti, C. & Giordano, A. 2010. CDK inhibitors: from the bench to clinical trials. ***Curr Drug Targets***, 11**,** 279-90.

Roberts, B. C. & Mancera, R. L. 2008. Ligand-protein docking with water molecules. ***J Chem Inf Model***, 48**,** 397-408.

Rodriguez-Puebla, M. L., Miliani De Marval, P. L., Lacava, M., Moons, D. S., Kiyokawa, H. & Conti, C. J. 2002. Cdk4 deficiency inhibits skin tumor development but does not affect normal keratinocyte proliferation. ***Am J Pathol***, 161**,** 405-11.

Roll, D. M., Ireland, C. M., Lu, H. S. M. & Clardy, J. 1988. Fascaplysin, an unusual antimicrobial pigment from the marine sponge Fascaplysinopsis sp. ***J. Org. Chem.***, 53**,** 3276-3278.

Ruegg, U. T. & Burgess, G. M. 1989. Staurosporine, K-252 and UCN-01: potent but nonspecific inhibitors of protein kinases. ***Trends Pharmacol Sci***, 10**,** 218-20.

Russell, P. & Nurse, P. 1986. Schizosaccharomyces pombe and saccharomyces cerevisiae: A look at yeasts divided. ***Cell***, 45**,** 781-782.

Sage, J. 2004. Cyclin C Makes an Entry into the Cell Cycle. ***Developmental cell***, 6**,** 607-608.

Saitou, N. & Nei, M. 1987. The Neighbor-Joining Method - a New Method for Reconstructing Phylogenetic Trees. ***Molecular Biology and Evolution***, 4**,** 406-425.

Saiz, J. E. & Fisher, R. P. 2002. A CDK-activating kinase network is required in cell cycle control and transcription in fission yeast. ***Curr Biol***, 12**,** 1100-5.

Sali, A. & Blundell, T. 1993. Comparative Protein Modelling by Satisfaction of Spatial Restraints. ***Journal of Molecular Biology***, 234**,** 779-815.

Sanchez, R. & Sali, A. 2000. Comparative protein structure modeling. Introduction and practical examples with modeller. ***Methods Mol Biol***, 143**,** 97-129.

Scheeff, E. D. & Bourne, P. E. 2005. Structural evolution of the protein kinase-like superfamily. ***Plos Computational Biology***, 1**,** 359-381.

Schneider, T. D. & Stephens, R. M. 1990. Sequence logos: a new way to display consensus sequences. ***Nucleic Acids Res***, 18**,** 6097-100.

Schoepfer, J., Fretz, H., Chaudhuri, B., Muller, L., Seeber, E., Meijer, L., Lozach, O., Vangrevelinghe, E. & Furet, P. 2002. Structure-based design and synthesis of 2-benzylidene-benzofuran-3-ones as flavopiridol mimics. ***J Med Chem***, 45**,** 1741-7.

Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. 2003. SWISS-MODEL: An automated protein homology-modeling server. ***Nucleic Acids Res***, 31**,** 3381-5.

Senderowicz, A. M. 1999. Flavopiridol: the First Cyclin-Dependent Kinase Inhibitor in Human Clinical Trials. ***Investigational New Drugs***, 17**,** 313-320.

Shen, M. Y. & Sali, A. 2006. Statistical potential for assessment and prediction of protein structures. ***Protein Sci***, 15**,** 2507-24.

Sherr, C. J. 1996. Cancer Cell Cycles. ***Science***, 274**,** 1672-1677.

Shibuya, M., Suzuki, Y., Sugita, K., Saito, I., Sasaki, T., Takakura, K., Okamoto, S., Kikuchi, H., Takemae, T. & Hidaka, H. 1990. Dose Escalation Trial of a Novel Calcium-Antagonist, At877, in Patients with Aneurysmal Subarachnoid Hemorrhage. ***Acta Neurochirurgica***, 107**,** 11-15.

Shirts, M. R. & Pande, V. S. 2005. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. ***J Chem Phys***, 122**,** 134508.

Shoemaker, B. A. & Panchenko, A. R. 2007. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. ***PLoS Comput Biol***, 3**,** e42.

Shoichet, B. K., Mcgovern, S. L., Wei, B. & Irwin, J. J. 2002. Lead discovery using molecular docking. ***Current Opinion in Chemical Biology***, 6**,** 439-446.

Shu, F., Lv, S., Qin, Y., Ma, X., Wang, X., Peng, X., Luo, Y., Xu, B. E., Sun, X. & Wu, J. 2007. Functional characterization of human PFTK1 as a cyclin-dependent kinase. ***Proc Natl Acad Sci U S A***, 104**,** 9248-53.

Simone, M., Erba, E., Damia, G., Vikhanskaya, F., Di Francesco, A. M., Riccardi, R., Bailly, C., Cuevas, C., Fernandez Sousa-Faro, J. M. & D'incalci, M. 2005. Variolin B and its derivate deoxy-variolin B: new marine natural compounds with cyclin-dependent kinase inhibitor activity. ***Eur J Cancer***, 41**,** 2366-77.

Simonson, T., Carlsson, J. & Case, D. A. 2004. Proton binding to proteins: pK(a) calculations with explicit and implicit solvent models. ***Journal of the American Chemical Society***, 126**,** 4167-4180.

Sippl, M. J. 1993. Recognition of errors in three-dimensional structures of proteins. ***Proteins: Structure, Function, and Bioinformatics***, 17**,** 355-362.

Slamon, D. J., Hurvitz, S. A., Applebaum, S., Glaspy, J. A., Allison, M. K., Dicarlo, B. A., Courtney, R. D., Kim, S. T., Randolph, S. & Finn, R. S. 2010. Phase I study of PD 0332991, cyclin-D kinase (CDK) 4/6 inhibitor in combination with letrozole for first-line treatment of patients with ER-positive, HER2-negative breast cancer. . ***Journal of Clinical Oncology***, 28.

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. & Edelman, M. 1999. Automated analysis of interatomic contacts in proteins. ***Bioinformatics***, 15**,** 327-332.

Soni, R., Muller, L., Furet, P., Schoepfer, J., Stephan, C., Zumstein-Mecker, S., Fretz, H. & Chaudhuri, B. 2000. Inhibition of Cyclin-Dependent Kinase 4 (Cdk4) by Fascaplysin, a Marine Natural Product. ***Biochemical and Biophysical Research Communications***, 275**,** 877-884.

Sousa, S. F., Fernandes, P. A. & Ramos, M. J. 2006. Protein-ligand docking: current status and future challenges. ***Proteins***, 65**,** 15-26.

Sreenivasan, U. & Axelsen, P. H. 1992. Buried water in homologous serine proteases. ***Biochemistry***, 31**,** 12785-91.

Storey, S. 2009. Chronic myelogenous leukaemia market. ***Nat Rev Drug Discov***, 8**,** 447.

Straatsma, T. P. & Berendsen, H. J. C. 1988. Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by

molecular dynamics simulations. ***The Journal of Chemical Physics***, 89**,** 5876-5886.

Suzek, B. E., Huang, H., Mcgarvey, P., Mazumder, R. & Wu, C. H. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. ***Bioinformatics***, 23**,** 1282-8.

Swalla, B. J. & Smith, A. B. 2008. Deciphering deuterostome phylogeny: molecular, morphological and palaeontological perspectives. ***Philos Trans R Soc Lond, B, Biol Sci***, 363**,** 1557-68.

Takaki, T., Echalier, A., Brown, N. R., Hunt, T., Endicott, J. A. & Noble, M. E. M. 2009. The structure of CDK4/cyclin D3 has implications for models of CDK activation. ***Proceedings of the National Academy of Sciences***, 106**,** 4171-4176.

Takaki, T., Fukasawa, K., Suzuki-Takahashi, I., Semba, K., Kitagawa, M., Taya, Y. & Hirai, H. 2005. Preferences for phosphorylation sites in the retinoblastoma protein of D-type cyclin-dependent kinases, Cdk4 and Cdk6, in vitro. ***Journal of Biochemistry***, 137**,** 381-386.

Tamaoki, T., Nomoto, H., Takahashi, I., Kato, Y., Morimoto, M. & Tomita, F. 1986. Staurosporine, a Potent Inhibitor of Phospholipid/Ca++Dependent Protein-Kinase. ***Biochemical and Biophysical Research Communications***, 135**,** 397-402.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. ***Molecular Biology and Evolution***, 24**,** 1596-1599.

Taylor, R. D., Jewsbury, P. J. & Essex, J. W. 2002. A review of protein-small molecule docking methods. ***J Comput Aided Mol Des***, 16**,** 151-66.

Tetzlaff, M. T., Bai, C., Finegold, M., Wilson, J., Harper, J. W., Mahon, K. A. & Elledge, S. J. 2004. Cyclin F disruption compromises placental development and affects normal cell cycle execution. ***Mol Cell Biol***, 24**,** 2487-98.

Thilagavathi, R. & Mancera, R. L. 2010. Ligand-protein cross-docking with water molecules. ***J Chem Inf Model***, 50**,** 415-21.

Tisdall, J. D. 2001. *Beginning Perl for bioinformatics*, Beijing ; Sebastopol, CA, O'Reilly.

Toledo, L. M. & Lydon, N. B. 1997. Structures of staurosporine bound to CDK2 and cAPK--new tools for structure-based design of protein kinase inhibitors. ***Structure***, 5**,** 1551-6.

Trooskens, G., De Beule, D., Decouttere, F. & Van Criekinge, W. 2005. Phylogenetic trees: visualizing, customizing and detecting incongruence. ***Bioinformatics***, 21**,** 3801-2.

Tsai, L. H., Takahashi, T., Caviness, V. S., Jr. & Harlow, E. 1993. Activity and expression pattern of cyclin-dependent kinase 5 in the embryonic mouse nervous system. ***Development***, 119**,** 1029-40.

Tsui, V. & Case, D. A. 2000. Theory and applications of the generalized Born solvation model in macromolecular Simulations. ***Biopolymers***, 56**,** 275-291.

Tsutsui, T., Hesabi, B., Moons, D. S., Pandolfi, P. P., Hansel, K. S., Koff, A. & Kiyokawa, H. 1999. Targeted disruption of CDK4 delays cell cycle entry with enhanced p27(Kip1) activity. ***Mol Cell Biol***, 19**,** 7011-9.

Ullrich, A. & Schlessinger, J. 1990. Signal transduction by receptors with tyrosine kinase activity. ***Cell***, 61**,** 203-12.

Van Gunsteren, W. F. & Berendsen, H. J. 1987. Thermodynamic cycle integration by computer simulation as a tool for obtaining free energy differences in molecular chemistry. ***Journal of computer-aided molecular design***, 1**,** 171-6.

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. 2003. Improved protein-ligand docking using GOLD. ***Proteins***, 52**,** 609-23.

Verdonk, M. L., Mortenson, P. N., Hall, R. J., Hartshorn, M. J. & Murray, C. W. 2008. Protein-Ligand Docking against Non-Native Protein Conformers. ***Journal of Chemical Information and Modeling***, 48**,** 2214-2225.

Vriend, G. 1990. WHAT IF: a molecular modeling and drug design program. ***J Mol Graph***, 8**,** 52-6, 29.

Wall, L., Christiansen, T. & Orwant, J. 2000. *Programming Perl (3rd Edition)*, O'Reilly.

Wang, G. & Dunbrack, R. L. 2005. PISCES: recent improvements to a PDB sequence culling server. ***Nucleic Acids Research***, 33**,** W94-W98.

Wang, J. M., Wang, W., Kollman, P. A. & Case, D. A. 2006. Automatic atom type and bond type perception in molecular mechanical calculations. ***Journal of Molecular Graphics & Modelling***, 25**,** 247-260.

Wang, J. M., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. 2004. Development and testing of a general amber force field. ***Journal of Computational Chemistry***, 25**,** 1157-1174.

Wang, R., Lai, L. & Wang, S. 2002. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. ***J Comput Aided Mol Des***, 16**,** 11-26.

Wang, S. & Fischer, P. M. 2008. Cyclin-dependent kinase 9: a key transcriptional regulator and potential drug target in oncology, virology and cardiology. ***Trends Pharmacol Sci***, 29**,** 302-13.

Warren, G. L., Andrews, C. W., Capelli, A. M., Clarke, B., Lalonde, J., Lambert, M. H., Lindvall, M., Nevins, N., Semus, S. F., Senger, S., Tedesco, G., Wall, I. D., Woolven, J. M., Peishoff, C. E. & Head, M. S. 2006. A critical assessment of docking programs and scoring functions. ***J Med Chem***, 49**,** 5912-31.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. 2009. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. ***Bioinformatics***, 25**,** 1189-1191.

Webster, K. R. 1998. The therapeutic potential of targeting the cell cycle ***Expert Opinion on Investigational Drugs***, 7**,** 865-887(23).

Wei, F. Y. & Tomizawa, K. 2007. Cyclin-dependent kinase 5 (Cdk5): a potential therapeutic target for the treatment of neurodegenerative diseases and diabetes mellitus. ***Mini Rev Med Chem***, 7**,** 1070-4.

Wei, G., Lonardo, F., Ueda, T., Kim, T., Huvos, A. G., Healey, J. H. & Ladanyi, M. 1999. CDK4 gene amplification in osteosarcoma: reciprocal relationship with INK4A gene alterations and mapping of 12q13 amplicons. ***Int J Cancer***, 80**,** 199-204.

Weinberg, R. A. 1995. The retinoblastoma protein and cell cycle control. ***Cell***, 81**,** 323-30.

Weiner, P. K. & Kollman, P. A. 1981. AMBER - Assisted model building with energy refinement. A general program for modeling molecules and their interactions. ***Journal of Computational Chemistry***, 2**,** 287-303.

Whelan, S. 2007. New approaches to phylogenetic tree search and their application to large numbers of protein alignments. ***Syst Biol***, 56**,** 727-40.

Whelan, S. & Goldman, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. ***Mol Biol Evol***, 18**,** 691-9.

Whelan, S., Lio, P. & Goldman, N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. ***Trends Genet***, 17**,** 262-72.

Wiederstein, M. & Sippl, M. J. 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. ***Nucleic Acids Research***, 35**,** W407-10.

Woese, C. R., Kandler, O. & Wheelis, M. L. 1990. Towards a Natural System of Organisms - Proposal for the Domains Archaea, Bacteria, and Eucarya. ***Proceedings of the National Academy of Sciences of the United States of America***, 87**,** 4576-4579.

Wu, S. Y., Mcnae, I., Kontopidis, G., Mcclue, S. J., Mcinnes, C., Stewart, K. J., Wang, S., Zheleva, D. I., Marriage, H., Lane, D. P., Taylor, P., Fischer, P. M. & Walkinshaw, M. D. 2003. Discovery of a novel family of CDK inhibitors with the program LIDAEUS: structural basis for ligand-induced disordering of the activation loop. ***Structure***, 11**,** 399-410.

Yan, S. F., King, F. J., Zhou, Y., Warmuth, M. & Xia, G. 2006. Profiling the kinome for drug discovery. ***Drug Discovery Today: Technologies***, 3**,** 269-276.

Yang, A. S. & Honig, B. 2000. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. ***J Mol Biol***, 301**,** 665-78.

Yang, Z., Nielsen, R. & Hasegawa, M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. ***Mol Biol Evol***, 15**,** 1600-11.

Yung-Chi, C. & Prusoff, W. H. 1973. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. ***Biochemical Pharmacology***, 22**,** 3099-3108.

Yusuf, D., Davis, A. M., Kleywegt, G. J. & Schmitt, S. 2008. An alternative method for the evaluation of docking performance: RSR vs RMSD. ***J Chem Inf Model***, 48**,** 1411-22.

Zarkowska, T. & Mittnacht, S. 1997. Differential phosphorylation of the retinoblastoma protein by G(1)/S cyclin-dependent kinases. ***Journal of Biological Chemistry***, 272**,** 12738-12746.

Zhang, J. M., Yang, P. L. & Gray, N. S. 2009. Targeting cancer with small molecule kinase inhibitors. ***Nature Reviews Cancer***, 9**,** 28-39.

Zsoldos, Z., Reid, D., Simon, A., Sadjad, S. B. & Johnson, A. P. 2007. eHiTS: a new fast, exhaustive flexible ligand docking system. ***J Mol Graph Model***, 26**,** 198-212.

Zwanzig, R. W. 1954. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. The Journal of Chemical Physics, 22, 1420-1426.

Appendices

Appendix 1 Lists of CDKs PDB structures

A 1.1 List of Human CDK2 Structures in PDB (updated Nov, 28, 2010)

PDB ID	**Resolution** (Å)	**R-Free**	**PDB**	**Resolutin** (Å)	**R-Free**
2R3I	1.28	0.202	3DDQ	1.8	0.229
1GZ8	1.3	0.185	3IG7	1.8	0.257
2R3Q	1.35	0.202	3IGG	1.8	0.257
2R3R	1.47	0.211	1E1X	1.85	0.281
2R3F	1.5	0.236	1H07	1.85	0.235
2R3H	1.5	0.219	2B54	1.85	0.263
1JVP	1.53	0.253	2B55	1.85	0.256
2R3G	1.55	0.212	2BTR	1.85	0.272
1H00	1.6	0.242	2FVD	1.85	0.235
1URW	1.6	0.253	2VU3	1.85	0.246
1OIT	1.6	0.24	2XNB	1.85	0.293
2R3N	1.63	0.235	2B52	1.88	0.275
2R3J	1.65	0.236	1H0V	1.9	0.234
2R3L	1.65	0.226	2BHE	1.9	0.322
2R3P	1.66	0.226	2C6K	1.9	0.254
2VTT	1.68	0.227	2C6M	1.9	0.27
2CCH	1.7	0.182	2VTH	1.9	0.24
2R3K	1.7	0.232	2VTQ	1.9	0.243
2R3M	1.7	0.216	2VTR	1.9	0.261
1H01	1.79	0.219	2VTS	1.9	0.26
1H08	1.8	0.242	2W05	1.9	0.292
2C6I	1.8	0.275	1HCK	1.9	0.272
2CLX	1.8	0.231	2VV9	1.9	0.228
1HCL	1.8	0.254	2XMY	1.9	0.271
1YKR	1.8	0.271	3EZR	1.9	0.243
2EXM	1.8	0.27	3FZ1	1.9	0.259
2R3O	1.8	0.227	1OIR	1.91	0.241

PDB ID	Resolution (Å)	R-Free	PDB ID	Resolutin (Å)	R-Free
1E1V	1.95	0.269	3LE6	2	0.248
1PW2	1.95	0.249	3LFQ	2.03	0.241
1PXI	1.95	0.251	2W06	2.04	0.233
2A0C	1.95	0.244	1CKP	2.05	0.26
2C68	1.95	0.249	1W8C	2.05	0.256
1JSV	1.96	n/a	2UUE	2.06	0.244
1PXO	1.96	0.235	1H0W	2.1	0.287
2BTS	1.99	0.261	1OI9	2.1	0.276
3EZV	1.99	0.244	2C5N	2.1	0.259
1OIU	2	0.253	2C5O	2.1	0.256
2IW9	2	0.24	2C69	2.1	0.28
2VTA	2	0.28	2C6O	2.1	0.288
2VTI	2	0.199	1B39	2.1	0.27
2VTL	2	0.234	1DM2	2.1	0.267
1AQ1	2	0.26	1H1P	2.1	0.258
1B38	2	0.25	3BHV	2.1	0.229
1GII	2	0.259	3MY5	2.1	0.219
1H1R	2	0.294	1W98	2.15	0.246
1H1S	2	0.286	1Y91	2.15	0.295
1KE6	2	0.222	2VTP	2.15	0.247
1KE7	2	0.235	2W17	2.15	0.3
1KE8	2	0.228	2W1H	2.15	0.308
1KE9	2	0.238	2VTO	2.19	0.241
1PYE	2	0.252	1H27	2.2	0.269
1R78	2	0.262	1WCC	2.2	0.252
1Y8Y	2	0.272	2V0D	2.2	0.286
2B53	2	0.273	2VTJ	2.2	0.262
2DS1	2	0.29	2VTN	2.2	0.25
3BHT	2	0.22	1DI8	2.2	0.226

PDB ID	Resolution (Å)	R-Free	PDB ID	Resolutin (Å)	R-Free
1FVT	2.2	0.232	1OIY	2.4	0.309
1GIJ	2.2	0.286	2BPM	2.4	0.271
1KE5	2.2	0.224	2UZE	2.4	0.243
1QMZ	2.2	0.28	2UZL	2.4	0.263
1W0X	2.2	0.27	1OKV	2.4	0.278
2DUV	2.2	0.31	2A4L	2.4	0.27
1VYZ	2.21	0.28	3F5X	2.4	0.283
1P5E	2.22	0.257	3LFS	2.4	0.25
1H26	2.24	0.268	1E9H	2.5	0.273
2C5Y	2.25	0.295	1H24	2.5	0.27
2VTM	2.25	0.286	1H25	2.5	0.266
3LFN	2.28	0.268	1OKW	2.5	0.253
2C6L	2.3	0.281	1F5Q	2.5	0.278
2CJM	2.3	0.265	1H1Q	2.5	0.332
2IW6	2.3	0.287	1P2A	2.5	0.28
2IW8	2.3	0.255	1PXL	2.5	0.249
2UZN	2.3	0.297	1PXN	2.5	0.277
2UZO	2.3	0.291	2G9X	2.5	0.265
1FIN	2.3	n/a	2J9M	2.5	0.294
1JSU	2.3	0.26	2WIH	2.5	0.25
1PKD	2.3	0.268	1PF8	2.51	0.306
1PXJ	2.3	0.271	2WPA	2.51	0.261
1PXP	2.3	0.273	1PXM	2.53	0.28
1VYW	2.3	0.258	2WFY	2.53	0.258
2R64	2.3	0.257	1OGU	2.6	0.26
2WEV	2.3	0.242	1OL2	2.6	0.29
3BHU	2.3	0.24	2BHH	2.6	0.325
1OIQ	2.31	0.271	2BKZ	2.6	0.259
1V1K	2.31	0.269	1BUH	2.6	0.25

PDB ID	Resolution (Å)	R-Free	PDB ID	Resolution (Å)	R-Free
1JST	2.6	n/a	1H28	2.8	0.322
1URC	2.6	0.254	1FVV	2.8	0.26
2V22	2.6	0.277	1GIH	2.8	0.275
2WMB	2.6	0.296	1PXK	2.8	0.266
2WXV	2.6	0.243	2I40	2.8	0.222
1G5S	2.61	0.242	2WIP	2.8	0.258
2C6T	2.61	0.257	2WMA	2.8	0.285
1GY3	2.7	0.313	2C5V	2.9	0.282
2C4G	2.7	0.25	2C5X	2.9	0.264
2CCI	2.7	0.321	1OL1	2.9	0.284
2UZB	2.7	0.265	2WHB	2.9	0.263
3DDP	2.7	0.27	1FQ1	3	0.313
3DOG	2.7	0.248	3EID	3.15	0.261
2UZD	2.72	0.269	3EOC	3.2	0.238
2X1N	2.75	0.254	3EJ1	3.22	0.259

A 1.2 List of Human CDK4 Structures in PDB

PDB ID	Resolution (Å)	R-Free	PDB ID	Resolution (Å)	R-Free
2W96	2.30	0.259	2W9F	2.85	0.300
2W9Z	2.45	0.272	3G33	3.00	0.314
2W99	2.80	0.270			

A 1.3 List of Human CDK5 Structures in PDB

PDB ID	Resolution (Å)	R-Free	PDB ID	Resolution (Å)	R-Free
1UNL	2.20	0.219	1UNH	2.35	0.230
1UNG	2.30	0.225	1H4L	2.65	0.287

A 1.4 List of Human CDK6 Structures in PDB

PDB ID	Resolution (Å)	R-Free	PDB ID	Resolution (Å)	R-Free
2F2C	2.80	0.301	1G3N	2.90	0.262
1BLX	1.90	0.253	1JOW	3.10	0.323
1BI7	3.40	0.330	1XO2	2.90	0.313
2EUF	3.00	0.306	1BI8	2.80	0.308

A 1.5 List of Human CDK7 Structures in PDB

PDB ID	Resolution (Å)	R-Free	PDB ID	Resolution (Å)	R-Free
1UA2	3.02	0.288			

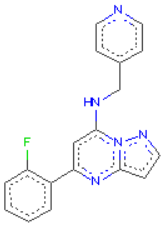
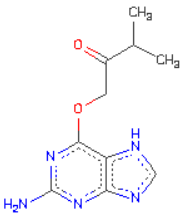
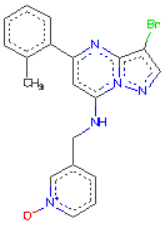
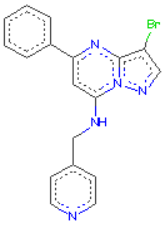
A 1.6 List of Human CDK9 Structures in PDB

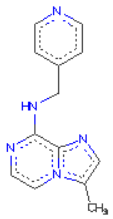
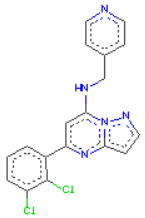
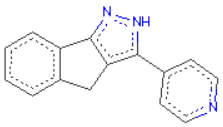
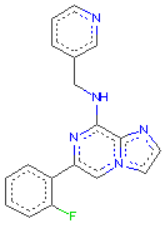
PDB ID	Resolution (Å)	R-Free	PDB ID	Resolution (Å)	R-Free
3BLH	2.48	0.221	3BLQ	2.90	0.234
3BLR	2.80	0.228			

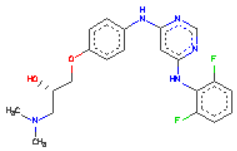
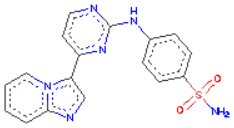
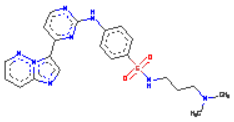
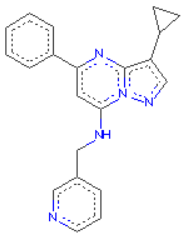
A 1.7 List of human cyclin UniRef identifiers

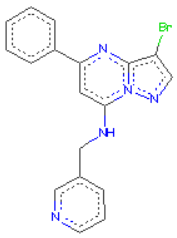
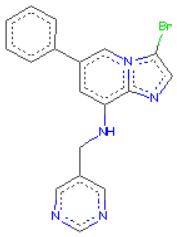
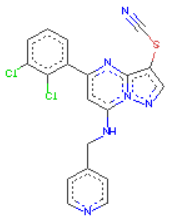
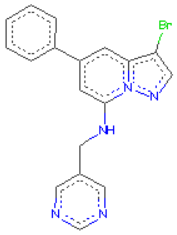
Symbol	UniRef ID	Symbol	UniRef ID	Symbol	UniRef ID
Cyclin-A1	P78396	Cyclin-E2	O96020	Cyclin-K	O75909
Cyclin-A2	P20248	Cyclin-F	P41002	Cyclin-L1	Q9UK58
Cyclin-B1	P14635	Cyclin-G1	P51959	Cyclin-L2	Q96S94
Cyclin-B2	O95067	Cyclin-G2	P51946	Cyclin-O	P22674
Cyclin-B3	Q8WWL7	Cyclin-H	P51946	Cyclin-Y	Q8ND76
Cyclin-C	P24863	Cyclin-I	Q14094	Cyclin-JL	Q8IV13
Cyclin-D1	P24385	Cyclin-I2	Q6ZMN8	Cyclin-YL1	Q8N7R7
Cyclin-D2	P30279	Cyclin-T1	O60563	Cyclin-YL1	Q5T2Q4
Cyclin-D3	P30281	Cyclin-T2	O60583	Cyclin-YL3	P0C7X3
Cyclin-E1	P24864	Cyclin-J	Q5T5M9		

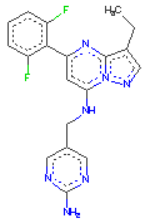
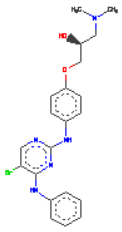
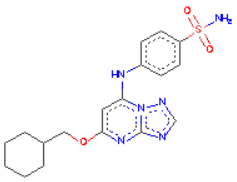
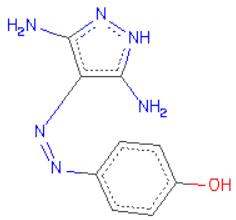
A 1.8 CDK2 PDB structures and ligands used in self socking experiments

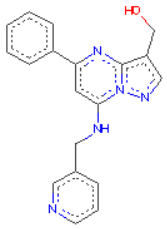
PDB ID	Ligand ID	Ligand name and structure
2R3I	SCF	5-(2-fluorophenyl)-N-(pyridin-4-ylmethyl)pyrazolo[1,5-a]pyrimidin-7-amine  C18 H14 F N5
1GZ8	MBP	1-[(2-AMINO-6,9-DIHYDRO-1H-PURIN-6-YL)OXY]-3-METHYL-2-BUTANOL  C10 H13 N5 O2
2R3Q	5SC	3-((3-bromo-5-o-tolylpyrazolo[1,5-a]pyrimidin-7-ylamino)methyl)pyridine 1-oxide  C19 H16 Br N5 O
2R3R	6SC	3-bromo-5-phenyl-N-(pyridin-4-ylmethyl)pyrazolo[1,5-a]pyrimidin-7-amine  C18 H14 Br N5

PDB ID	Ligand ID	Ligand name and structure
2R3H	SCE	3-methyl-N-(pyridin-4-ylmethyl)imidazo[1,2-a]pyrazin-8-amine  C13 H13 N5
2R3F	SC8	5-(2,3-dichlorophenyl)-N-(pyridin-4-ylmethyl)pyrazolo[1,5-a]pyrimidin-7-amine  C18 H13 Cl2 N5
1JVP	LIG	3-pyridin-4-yl-2,4-dihydroindeno[1,2-c]pyrazole  C15 H11 N3
2R3G	SC9	6-(2-fluorophenyl)-N-(pyridin-3-ylmethyl)imidazo[1,2-a]pyrazin-8-amine  C18 H14 F N5

PDB ID	Ligand ID	Ligand name and structure
1H00	FAP	(2S)-1-[4-({6-[(2,6-DIFLUOROPHENYL)AMINO]PYRIMIDIN-4-YL}AMINO)PHENOXY]-3-(DIMETHYLAMINO)PROPAN-2-OL  C21 H23 F2 N5 O2
1OIT	HDT	4-[(4-imidazo[3,2-a]pyridin-3-ylpyrimidin-2-yl)amino]benzenesulfonamide  C17 H14 N6 O2 S
1URW	I1P	N-(3-dimethylaminopropyl)-4-[(4-imidazo[2,3-f]pyridazin-3-ylpyrimidin-2-yl)amino]benzenesulfonamide  C21 H24 N8 O2 S
2R3N	SCZ	3-cyclopropyl-5-phenyl-N-(pyridin-3-ylmethyl)pyrazolo[1,5-a]pyrimidin-7-amine  C21 H19 N5

PDB ID	Ligand ID	Ligand name and structure
2R3J	SCJ	3-bromo-5-phenyl-N-(pyridin-3-ylmethyl)pyrazolo[1,5-a]pyrimidin-7-amine  C18 H14 Br N5
2R3L	SCW	3-bromo-6-phenyl-N-(pyrimidin-5-ylmethyl)imidazo[1,2-a]pyridin-8-amine  C18 H14 Br N5
2R3P	3SC	5-(2,3-dichlorophenyl)-N-(pyridin-4-ylmethyl)-3-thiocyanatopyrazolo[1,5-a]pyrimidin-7-amine  C19 H12 Cl2 N6 S
2R3K	SCQ	3-bromo-5-phenyl-N-(pyrimidin-5-ylmethyl)pyrazolo[1,5-a]pyridin-7-amine  C18 H14 Br N5

PDB ID	Ligand ID	Ligand name and structure
2R3M	SCX	N-((2-aminopyrimidin-5-yl)methyl)-5-(2,6-difluorophenyl)-3-ethylpyrazolo[1,5-a]pyrimidin-7-amin  C19 H17 F2 N7
1H08	BWP	(2S)-1-{4-[(4-ANILINO-5-BROMOPYRIMIDIN-2-YL)AMINO]PHENOXY}-3-(DIMETHYLAMINO)PROPAN-2-OL  C21 H24 Br N5 O2
2C6I	DT1	4-{[5-(CYCLOHEXYLMETHOXY)[1,2,4]TRIAZOLO[1,5-A]PYRIMIDIN-7-YL]AMINO}BENZENESULFONAMIDE  C18 H22 N6 O3 S
2CLX	F18	4-[(E)-(3,5-DIAMINO-1H-PYRAZOL-4-YL)DIAZENYL]PHENOL  C9 H10 N6 O

PDB ID	Ligand ID	Ligand name and structure
2R3O	2SC	(5-phenyl-7-(pyridin-3-ylmethylamino)pyrazolo[1,5-a]pyrimidin-3-yl)methanol  C19 H17 N5 O

Appendix 2 Perl Scripts

A 2.1 A tool to download multiple PDB files from a given list of PDB identifiers

pdbDownload.pl

```
#!/usr/bin/perl -w
#########################################################################
#                                                                       #
#  A tool to download PDB files with PDB identifiers                    #
#                                                                       #
# by Muhammad Imtiaz Shafiq                                             #
#                                                                       #
#########################################################################
# June 2007
# Last Modified August 27, 2009 14:20 GMT

# Print instructions on screen for user input of file name

print "please write down the file where you have your PDB identifiers stored \n";

#Record user input and store in a variable using <STDIN>

$myanswer = "<STDIN>";

# defined a new veriable for file name

$myfilename = $myanswer;

#file extension stored in a variable

$fileend = ".pdb";

# open the file if available in the directory otherwise die and print the
instructions on the screen for any possible typo by the user
open( SP, "$myfilename" )
  or die
"Your desired file $myfilename is not found in the current directory,
please correct either correct the typo, or search a different directory\n";

# loop to read line by line
while ( $pdb = <SP> ) {
    chomp $pdb;
    # changing all uper case to lower case
    #$pdb =~ tr/A-Z/a-z/;
    $url = "http://www.rcsb.org/pdb/files/$pdb$fileend";
    system("wget -N $url");
}
exit;
```

A 2.2 A tool to extract active site residues from multiple PDB file and save them into a new PDB file.

activesiteExtract.pl

```
#!/usr/bin/perl -w
########################################################################
########################################################################
##                                                                    ##
##                                                                    ##
##    Active site extract version 2.0                                 ##
##    Author: Muhammad Imtiaz Shafiq                                  ##
##            Department of Biochemistry,                             ##
##            University of Leicester,                                ##
##            Leicester, UK                                           ##
##                                                                    ##
########################################################################
########################################################################
##    Date: August 1, 2008                                            ##
########################################################################

use strict;

if ($ARGV[0]){
active();
}else{

use Term::ANSIColor qw(:constants);
    $Term::ANSIColor::AUTORESET = 1;
    print BOLD RED "Usage: please type number of residues in the
    active site at command line as scriptname.pl 10 11 12 13 \n";
    print BOLD BLUE "To continue further run script again\n";
}


###########subs###########

#~~~~~~~~~~active~~~~~~~~~~~~~~~~#
sub active{
#open the directory and than read all the files with *.pdb
my @all_pdb_files=openDir('pdb');

#reterive desired data from the PDB file
foreach (@all_pdb_files){

open (SP, "$_");
my $test;
my $pattern = join '|', @ARGV;
while (<SP>){
   #match PDB Id in the file header
   if ($_=~/^HEADER[\s\S]+(....).............$/){
   $test=$1;
   #chomp $test;
   #print "$test";
   $test= "$test" . "-pdb.as";
   open (OUTPUT,">$test");
   while (<SP>) {
            if (/^HEADER[\s\S]+/){
               print OUTPUT;
            }elsif (/^ATOM\s+\S+\s+\S+\s+\S+\s+\S+\s+\S\s+($pattern)\s/){
               print OUTPUT;
            }
```

```
       }
    }
  }
}
}
#~~~~~~~~~~active~~~~~~~~~~~~~~~~#

#~~~~~~~~~~openDir~~~~~~~~~~~~~~~#
sub openDir{
my($end)=@_;
print "$end";
use Cwd;
my $dir = cwd;
opendir(DIR,"$dir") or die "$!";
my @allFiles;
@allFiles  = grep {/\.$end$/} readdir DIR;
return @allFiles;
}
#~~~~~~~~~~openDir~~~~~~~~~~~~~~~#
```

Appendix 3 Homology modelling scripts

A 3.1 model-single.py script

```
from modeller import *
from modeller.automodel import *

env = environ()
a = automodel(env, alnfile='2cch-cdk4.ali',
              knowns='2cch', sequence='CDK4')
a.starting_model = 1
a.ending_model = 50
a.make()
```

A 3.2 evaluate-model.py script

```
from modeller import *
from modeller.scripts import complete_pdb
log.verbose() # request verbose output
env = environ()
env.libs.topology.read(file='$(LIB)/top_heav.lib') # read topology
env.libs.parameters.read(file='$(LIB)/par.lib') # read parameters
# read model file
mdl = complete_pdb(env, 'CDK4.B99990016.pdb')
# Assess with DOPE:
s = selection(mdl) # all atom selection
s.assess_dope(output='ENERGY_PROFILE NO_REPORT', file='CDK4.profile',
normalize_profile=True, smoothing_window=15)
```

A 3.3 evaluate-template.py script

```
from modeller import *
from modeller.scripts import complete_pdb
log.verbose() # request verbose output
env = environ()
env.libs.topology.read(file='$(LIB)/top_heav.lib') # read topology
env.libs.parameters.read(file='$(LIB)/par.lib') # read parameters
# read model file
mdl = complete_pdb(env, '2CCH.pdb')
# Assess with DOPE:
s = selection(mdl) # all atom selection
s.assess_dope(output='ENERGY_PROFILE NO_REPORT', file='2CCH.profile',
normalize_profile=True, smoothing_window=15)
```

Appendix 4 Input files for Molecular Dynamics

A 4.1 Input files for equilibration simulations

min.in

```
Minimise CDK4
 &cntrl
  imin=1, maxcyc=1000, ncyc=500,
  cut=8.0,ntb=1,
  ntc=2,ntf=2, ntpr=100,
  ntr=1, restraintmask=':1-303',
  restraint_wt=2.0
 /
```

heat.in

```
heat CDK4
 &cntrl
  imin=0,irest=0,ntx=1,
  nstlim=25000,dt=0.002, ntc=2,ntf=2,
  cut=8.0, ntb=1, ntpr=500, ntwx=500,
  ntt=3, gamma_ln=2.0,
  tempi=0.0, temp0=300.0,
  ntr=1, restraintmask=':1-303',
  restraint_wt=2.0,
 /
 &wt TYPE='TEMP0', istep1=0,
istep2=25000,
  value1=0.1, value2=300.0, /
 &wt TYPE='END' /
```

density.in

```
Density CDK4
 &cntrl
  imin=0,irest=1,ntx=5,
  nstlim=25000,dt=0.002,
  ntc=2,ntf=2,
  cut=8.0, ntb=2, ntp=1, taup=1.0,
  ntpr=500, ntwx=500,
  ntt=3,
  gamma_ln=2.0,
  temp0=300.0,
  ntr=1, restraintmask=':1-303',
  restraint_wt=2.0,
 /
```

equil.in

```
Equil CDK4
 &cntrl
  imin=0,irest=1,ntx=5,
  nstlim=250000,dt=0.002,
  ntc=2,ntf=2,
  cut=8.0, ntb=2, ntp=1,
  taup=2.0,
  ntpr=1000, ntwx=1000,
  ntt=3, gamma_ln=2.0,
  temp0=300.0,
 /
```

A 4.2 Input files for production run

prod.in

```
prod CDK4
 &cntrl
  imin=0,irest=1,
  ntx=5,
  nstlim=250000,
  dt=0.002,
  ntc=2,ntf=2,
  cut=8.0, ntb=2,
  ntp=1, taup=2.0,
  ntpr=5000,
  ntwx=5000,
  ntt=3, gamma_ln=2.0,
  temp0=300.0,
 /
```

Appendix 5 Input files for thermodynamic integration

A 5.1 Minimization

```
COMMENT: This script is for minimization
 &cntrl
  imin = 1,     ntx = 1,
  maxcyc=500,
  ntpr = 100,
  ntf = 2,      ntc = 2,
  ntb = 1,      cut = 9.0,
  icfe=1,       clambda = 0.${X},
  ifsc=0,
  crgmask='${mask0}',
 &end
```

A 5.2 Equilibration

```
COMMENT: This script will run density equilibration
 &cntrl
  imin = 0,     ntx = 1,        irest = 0,
  ntpr = 2500,  ntwr = 10000,   ntwx = 0,
  ntf = 2,      ntc = 2,
  ntb = 2,      cut = 9.0,
  nstlim = 25000,       dt = 0.002,
  temp0 = 300.0,        ntt = 3,        gamma_ln = 5,
  ntp = 1,      pres0 = 1.0,    taup = 0.2,
  icfe=1,       clambda = 0.${X},
  ifsc=0,
  crgmask='${mask0}',
 &end
```

A 5.3 Production

```
COMMENT: This will script is for production run
 &cntrl
  imin = 0,     ntx = 5,        irest = 1,
  ntpr = 10000, ntwr = 100000,  ntwx = 10000,
  ntf = 2,      ntc = 2,
  ntb = 2,      cut = 9.0,
  nstlim = 100000,      dt = 0.002,
  temp0 = 300.0,        ntt = 3,        gamma_ln = 2,
  ntp = 1,      pres0 = 1.0,    taup = 2.0,
  icfe=1,       clambda = 0.${X},
  ifsc=0,
  crgmask='${mask0}',
 &end
```