E19-2006-98

R. A. Selwyne[1,2], Kh. T. Kholmurodov[1,3,*], N. A. Koltovaya[1,3]

HOMOLOGY MODELING AND MOLECULAR DYNAMICS OF CYCLIN-DEPENDENT PROTEIN KINASE

Submitted for the book review entitled «Emerging Multi-Core Technologies Theories and Implementations» (Universities Press, c/o Orient Longman Private Ltd., Hyderabad, India)

[1] Laboratory of Radiation Biology, Joint Institute for Nuclear Research, 141980 Dubna, Moscow region, Russia
[2] Department of Botany, Bharahthiar University, Coimbatore-641046, TamilNadu, India
[3] International University «Dubna», 141980 Dubna, Moscow region, Russia
[*] E-mail: mirzo@jinr.ru

Селвайн Р. А., Холмуродов Х. Т., Колтовая Н. А. E19-2006-98
Гомологичное моделирование и молекулярная динамика
циклинзависимых протеинкиназ

Представлен обзор методов биоинформатики, важных при анализе белков. Также представлены поиск по базе данных, сравнение последовательностей и структурные прогнозы. Даны ссылки на страницы всемирной паутины (WWW) по каждой теме. Базы данных с биологической информацией рассмотрены с акцентом на базы с нуклеотидными последовательностями, геномами, аминокислотами и трехмерными структурами. Описаны «widespread»-методы для сравнения последовательностей, выравнивания множественных последовательностей и вторичного прогноза структуры. Представлено гомологичное моделирование и молекулярно-динамический (МД) анализ свойств структурных конформаций циклинзависимых киназ дрожжей и человека CDC28 и CDK2. На основе гомологичного моделирования с использованием кристаллической структуры CDK2 человека предсказана структура киназы человека при помощи пакета MODELLER. Далее проводилось МД-моделирование с использованием пакета AMBER8.0 в течение 2 нс и исследовалось конформационное поведение кристаллической решетки как дрожжей CDC28, так и CDK2/циклина А/АТФ-Mg^{2+}/субстрата человека при физиологической температуре $T = 300$ К. Основываясь на МД-моделировании, мы исследовали молекулярный механизм регуляции фосфорилирования и структурных изменений киназы.

Работа выполнена в Лаборатории радиационной биологии ОИЯИ.
Препринт Объединенного института ядерных исследований. Дубна, 2006

Selwyne R. A., Kholmurodov Kh. T., Koltovaya N. A. E19-2006-98
Homology Modeling and Molecular Dynamics of Cyclin-Dependent Protein Kinase

An overview of bioinformatics techniques of importance in protein analysis is given. These include database searches, sequence comparisons and structural predictions. Links useful to World Wide Web (WWW) pages are given in relation to each topic. Databases with biological information are reviewed with emphasis on databases for nucleotide sequences, genomes, amino acid sequences, and three-dimensional structures. Furthermore the widespread methods for sequence comparisons, multiple sequence alignments and secondary structure predictions were described. The homology modeling and molecular dynamics (MD) simulation analysis of structural conformation properties for yeast and human cyclin-dependent kinases CDC28 and CDK2 have been performed. Based on the homology modeling a structure of yeast CDC28 is predicted using a lattice crystal structure of human CDK2 using the MODELLER software. Further MD simulations run using AMBER8.0 package for 2 ns and the conformation behavior of crystal lattice for both yeast CDC28 and human CDK2/cyclin A/ATP-Mg^{2+}/substrate at physiological temperature $T = 300$ K have been investigated. Based on the MD simulation results we discuss the molecular mechanism regulation of phosphorylation and the structural changes of kinases.

The investigation has been performed at the Laboratory of Radiation Biology, JINR.

Preprint of the Joint Institute for Nuclear Research. Dubna, 2006

1. INTRODUCTION

The word «bioinformatics» refers to the application of information technology (IT) in molecular biology, and exists in the similar areas of study: computational (molecular) biology, biocomputing or biocomputation, computational genomics, *in silico* biology, and computational proteomics. In the present day scenario any endeavor in modern life has been enhanced by the application of information technology, made biologists easy. We may consider bioinformatics to comprise the study of *information pathways* in biology. And information pathways crucially required for the existence of any organism. DNA and protein sequences form the major proportion of the information pathways in molecular biology. These sequences are nothing but a set of four alphabets for DNA, and twenty ones for proteins. Thus all the tools and techniques that have been developed to analyze these sequences which carry information regarding the physiological mechanisms through the process digital information. Thus, bioinformatics is intimately connected with theoretical computer science, especially such topics as natural language processing, machine learning, computational linguistics and digital pattern recognition. Ideas and methods have been incorporated from these sciences and effectively applied in bioinformatics to obtain useful biological information.

1.1. Introduction to Molecular Biology. Before the invention of modern molecular biology, biological systems were thought to be based upon an unknown principle that set them apart from non-living matter. The understanding of the function of each separate portion so gained is put together, bit by bit, to build an understanding of the entire biological system. The development of the discipline of bioinformatics is just one manifestation of this success.

1.1.1. Genetic Information. The chief molecules involved in the information transfer pathway are deoxyribonucleic acids or DNA, ribonucleic acids or RNA, and proteins. Another common feature shared by all three is that they are polymeric molecules. The difference between RNA and DNA is the presence, in RNA, of an extra oxygen atom on the sugar ring. In the case of the nucleic acids, DNA and RNA, this is the sugar-phosphate backbone. In the case of proteins this is the polypeptide backbone.

DNA is a double-stranded molecule, consisting of two nucleic acid strands that run in opposite directions, and are wound around each other to form a double helix. One end is referred to as the 5' end and the other as the 3' end. The negatively charged phosphate-sugar backbones of the two strands are outside of the double helix, while the planar, nitrogenous and hydrophobic bases are inside of the double helix, away from the aqueous solvent molecules. The two strands

stick to each other through hydrogen bonds that form between the bases in a highly specific manner [1]. Since there are four types of bases, one may think of the information as being represented by four symbols, namely A, T, G and C.

RNA is similar to DNA but the alphabet consists of the four bases A, T, G and C, in RNA the base T is not found and instead we have the base U. RNA has three different functions in the process of information storage and transfer.

A *protein* chain is represented again by a string of symbols, this time chosen from an alphabet of 20 letters, representing the 20 different amino acids (see Fig. 1).

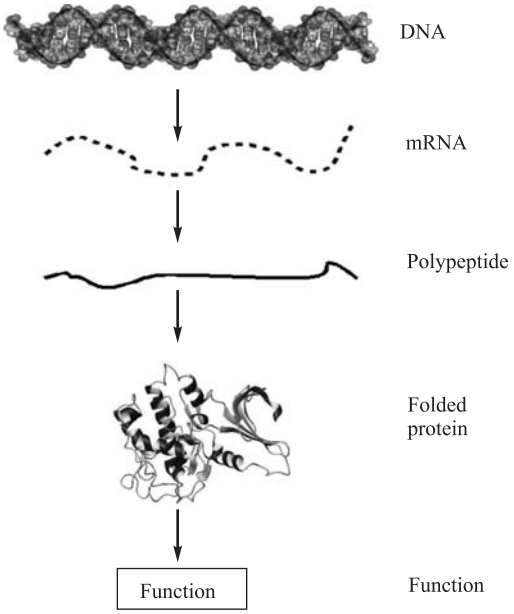


Fig. 1. Central dogma of molecular biology

1.2. What Does Bioinformatics Mean? Bioinformatics includes computer simulation of individual metabolic processes, as well as the more ambitious simulation of a whole cell or even a whole organism. Many of the processes involved in the development of new drugs, such as lead discovery, lead optimization through molecular modeling, design and analysis of the laboratory and clinical trials, are all considered as a part of bioinformatics.

1.2.1. Functions involved in bioinformatics. The analyses of DNA, RNA and protein sequences and structures may generally be broken down into the following elemental tasks.

• Searching for patterns within a sequence;

• Obtaining statistical information on a sequence;

• Searching for similarities between two sequences, or performing sequence alignment for a pair of sequences;

• Searching for similarities among many sequences, or performing multiple sequence alignment;

• Constructing phylogenetic trees based on sequences;

• Predicting and analyzing the secondary structures on basis of the sequence;

• Predicting and analyzing tertiary structure and folding for protein and RNA sequences.

The first task, i. e. searching for patterns, is the one that is the most difficult, as well as the one most often required. The second task in the list above relates to the statistical information on a single sequence, such as the base or amino acid composition. The third elemental task is searching for sequence similarities [2]. The fourth task in the list, multiple sequence comparison and alignment, is also very important for functional annotation. The last two tasks mentioned in the list relate to the structures of the molecules. To understand completely the function of any physical system, in the case of the biologically active molecule, it is necessary to know its structure.

2. MOLECULAR BIOLOGY DATABASES

An extensive data description method has been devised and implemented, such that the database can accept, store, search and analyze all the relevant types of data, including textual descriptions, images, three-dimensional structures, molecular interactions, molecular complexes, networks of interactions, physical locations within the cell, etc.

2.1. Data Types in Molecular Biology. Sequences and structures are dealt with in the later sections of their own. In this section we describe the other types of databases in molecular biology (see Fig. 2.).

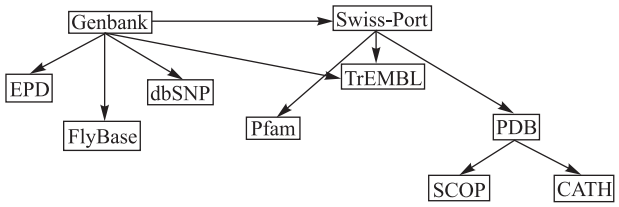


Fig. 2. A portion of the network of molecular biology databases available

2.1.1. Expression data. The Gene Expression Omnibus (GEO) at the NCBI site is a database for expression data obtained using a variety of methods including gene chips and Serial Analysis of Gene Expression (SAGE). The *Saccharomyces* Genome Database (SGD) (http://genome-www.stanford.edu/Saccharomyces) is an address where data regarding yeast genome expression may be accessed in this format [3].

2.1.2. Metabolic pathways and molecular interactions. The Kyoto Encyclopedia of Genes and Genomes (KEGG: http://www.genome.ad.jp/kegg) is one of the well-known databases for metabolic and regulatory pathways [4]. RegulonDB is a database of transcriptional regulation and operon organization for *E.coli* at http://www.cifn.unam.mx/Computational_Biology/regulondb [5].

2.1.3. Mutations and polymorphisms. The Online Mendelian Inheritance in Man (OMIM: http://www.ncbi.nlm.nih.gov/Omim) is a computerized catalogue of human genes and the genetic mutations and changes that lead to clinical disorders [6].

dbSNP (http://www.ncbi.nlm.nih.gov/SNP) is a database of SNPs, defined, in spite of the name, both the changes of a single base nucleotide, as well as short deletion and insertion polymorphisms [7].

The Protein Mutant Database (PMD: http://pmd.ddbj.nig.ac.jp) is a collection of information on mutant proteins that includes natural as well as artificial mutants [8].

2.2. Sequence Databases. Sequence and structure databases may be classified into two types, viz. primary or raw databases and secondary or derived databases.

2.2.1. Primary nucleotide sequence repositories — GenBank, EMBL, DDBJ. These are the three chief databases that store and make available raw nucleic acid sequences. GenBank is physically located in the USA and is accessible through the NCBI portal over the Internet. EMBL (stands for European Molecular Biology Laboratory) is in the UK, at the European Bioinformatics Institute, and DDBJ (DNA DataBank of Japan) is in Japan [9, 10, 11].

2.2.2. Primary protein sequence repositories. The PIR-PSD is now a comprehensive, non-redundant, expertly annotated, fully classified and extensively cross-referenced protein sequence database in the public domain [12]. It is available at http://pir.georgetown.edu/pirwww.

The other well-known and extensively used protein sequence database is SWISS-PROT (http://www.expasy.ch/sprot) [13]. The core data consists of the sequence entered in the common single letter amino acid code, and the related references and bibliography. The annotations contain information on the function or functions of the protein, post-translational modifications such as phosphorylation, acetylation, etc., functional or structural domains and sites, such as calcium binding regions, ATP-binding sites, zinc fingers, etc., known secondary structural features as, for example, alpha helix, beta sheet, etc., the quaternary structure of the protein.

2.2.3. Derived or secondary databases of nucleotide sequences. FlyBase or The Berkeley Drosophila Genome Project (http://www.fruitfly.org) gives the information on the entire genome of the fruit fly *D. melanogaster* to a high degree of completeness and quality [14].

The Eukaryotic Promoter Database (EPD: http://www.epd.isb-sib.ch) is one such a collection [15]. It contains the sequences and annotations of eukaryotic promoters recognized by RNA polymerase II, i. e. POL II.

2.2.4. Derived or secondary databases of amino acid sequences: patterns and signatures. PROSITE is one such a pattern database, which is accessible at http://www.expasy.ch/prosite [16]. The protein motifs or patterns are encoded as «regular expressions».

BLOCKS database (http://blocks.fhcrc.org/blocks) is automatic process of identifying patterns [17].

A database containing profiles built using the hidden Markov models is called Pfam (http://www.sanger.ac.uk/Software/Pfam) [18].

2.3. Primary Structure Databases

2.3.1. The primary structure databases — PDB. PDB (http://www.rcsb.org) stands for Protein Data Bank. In spite of the name, PDB archives the three-dimensional structures of not only proteins but also all biologically important molecules, such as nucleic acid fragments, RNA molecules and large peptides [20]. Structures determined by X-ray crystallography and NMR experimemts form the large majority of the entries.

2.3.2. Derived or secondary databases of biomolecular structures. The SCOP database (Structural Classification Of Proteins: http://scop.mrc-lmb.cam.ac.uk/scop/) is a manual classification of protein structures in a hierarchical scheme with many levels [20]. CATH (http://www.biochem.ucl.ac.uk/bsm/cath) stands for Class, Architecture, Topology and Homologous superfamily [21].

3. SEQUENCE ALIGNMENT

Bioinformatics provided the first successful transplants of algorithms from the realm of computer science into biology. It continues to attract the attention of mathematicians, who try to devise ever-newer algorithms to match strings of symbols, in general.

3.1. Sequence Search

3.1.1. Why align sequences? The reason we align sequences is to look for a common or related pattern amongst them. If we discover such sequence similarities, we may infer biological similarity between the two sequences. This could be a structural, functional or evolutionary relationship.

3.1.2. Scoring schemes – briefly. Here we will introduce in Table 1 the PAM100 scoring scheme [22]. The PAM series of matrices are 20×20 matrices

also known as «substitution» matrices. Each element of the matrix tells the score we have to use if, in an alignment, we find the residue pair labeling that element as matching residues. This matrix is called the substitution matrix $s(x, y)$, where x and y represent amino acids. We could use sophisticated functions to reflect various biological realities, but in the example below we will use a very simple linear function, which is given as

$$G = k \times n,$$

where k is a constant set to –8 in the examples below, and n is the number of gaps.

Table 1. The PAM100 substitution matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	–3	–1	–1	–3	–2	0	1	–3	–2	–3	–3	–2	–5	1	1	1	–7	–4	0
R	–3	7	–2	–4	–5	1	–3	–5	1	–3	–5	2	–1	–6	–1	–1	–3	1	–6	–4
N	–1	–2	5	3	–5	–1	1	–1	2	–3	–4	1	–4	–5	–2	1	0	–5	–2	–3
D	–1	–4	3	5	–7	0	4	–1	–1	–4	–6	–1	–5	–8	–3	–1	–2	–9	–6	–4
C	–3	–5	–5	–7	9	–8	–8	–5	–4	–3	–8	–8	–7	–7	–4	–1	–4	–9	–1	–3
Q	–2	1	–1	0	–8	6	2	–3	3	–4	–2	0	–2	–7	–1	–2	–2	–7	–6	–3
E	0	–3	1	4	–8	2	5	–1	–1	–3	–5	–1	–4	–8	–2	–1	–2	–9	–5	–3
G	1	–5	–1	–1	–5	–3	–1	5	–4	–5	–6	–3	–4	–6	–2	0	–2	–9	–7	–3
H	–3	1	2	–1	–4	3	–1	–4	7	–4	–3	–2	–4	–3	–1	–2	–3	–4	–1	–3
I	–2	–3	–3	–4	–3	–4	–3	–5	–4	6	1	–3	1	0	–4	–3	0	–7	–3	3
L	–3	–5	–4	–6	–8	–2	–5	–6	–3	1	6	–4	3	0	–4	–4	–3	–3	–3	0
K	–3	2	1	–1	–8	0	–1	–3	–2	–3	–4	5	0	–7	–3	–1	–1	–6	–6	–4
M	–2	–1	–4	–5	–7	–2	–4	–4	–4	1	3	0	9	–1	–4	–3	–1	–6	–5	1
F	–5	–6	–5	–8	–7	–7	–8	–6	–3	0	0	–7	–1	8	–6	–4	–5	–1	4	–3
P	1	–1	–2	–3	–4	–1	–2	–2	–1	–4	–4	–3	–4	–6	7	0	–1	–7	–7	–3
S	1	–1	1	–1	–1	–2	–1	0	–2	–3	–4	–1	–3	–4	0	4	2	–3	–4	–2
T	1	–3	0	–2	–4	–2	–2	–2	–3	0	–3	–1	–1	–5	–1	2	5	–7	–4	0
W	–7	1	–5	–9	–9	–7	–9	–9	–4	–7	–3	–6	–6	–1	–7	–3	–7	12	–2	–9
Y	–4	–6	–2	–6	–1	–6	–5	–7	–1	–3	–3	–6	–5	4	–7	–4	–4	–2	9	–4
V	0	–4	–3	–4	–3	–3	–3	–3	–3	3	0	–4	1	–3	–3	–2	0	–9	–4	5

3.2. BLAST. BLAST has assumed almost iconic status, and has become representative not only of sequence matching and comparisons, but very nearly of all of bioinformatics [2]. BLAST performs sequence search and comparison algorithms. BLAST do fast searches through large databases for matches to the query sequence, and then do more detailed alignments of the query sequences with the matches. BLAST compares a DNA sequence against DNA database, translated (in all six frames) version of a DNA sequence against translated (six-frame) version of the DNA database, translated (six-frame) version of a DNA sequence against protein database, a protein sequence against translated (six-frame) version of a DNA database, or a protein sequence against a protein database

[2]. The BLAST algorithm uses a word-based heuristic to execute an approximate version of the Smith–Waterman algorithm known as the «maximal segment pairs» algorithm. BLAST is the most frequently used over the Internet on the BLAST server (http://www.ncbi.nlm.nih.gov/BLAST/).

PSI-BLAST, stands for Position Specific Iterated BLAST [24]. This algorithm returns more distantly related sequences from the database than BLAST. PHI-BLAST stands for Pattern-Hit Initiated BLAST [24]. This is a search program for which the input is not only a query DNA or protein sequence, but also a pattern.

4. MULTIPLE SEQUENCE ALIGNMENT

One of the common goals of building multiple sequence alignments is to characterize protein and/or gene families, and identify shared regions of homology. In general, MSA therefore helps to establish phylogenetic relationships between sequences, and by extension, between the parent organisms. MSA helps to predict the secondary and tertiary structures for new sequences, and identify templates for threading and homology modeling, which are methods for 3D structure prediction.

4.1. Scoring MSA. CLUSTAL is a popular program for MSA that uses an extensively modified version of the Feng-Doolittle algorithm [25, 26]. The CLUSTAL algorithm builds up the MSA by using such profiles wherever appropriate. Every time an alignment is made, a profile is generated, and in the subsequent steps of the MSA construction, the profile is used, instead of the individual sequences. Thus, we have sequence–sequence comparisons, sequence–profile comparisons and profile–profile comparisons.

4.2. Substitution Matrices

4.2.1. What are substitution matrices? A matrix of values that is used to score residue replacements or substitutions is called a substitution matrix. Every element of this matrix then represents the score when the residue corresponding to the column index replaces the residue corresponding to the row index of the element.

4.2.2. BLOSUM substitution matrices. BLOSUM stands for BLOcks SUbstitution Matrices. In 1992, Henikoff and Henikoff devised the BLOSUM family of substitution matrices.

4.2.3. Gap penalties. A gap is a consecutive run of spaces in a single sequence of an alignment. It corresponds to an insertion or deletion of a subsequence. Gap penalties are also part of the scoring scheme, and must be chosen along with the substitution scores.

4.2.4. Phylogenetic trees. Phylogeny refers to the evolutionary relationships among species. Speciation is the process through which one species becomes

divided into two or more new species. The pattern of evolutionary relationships among species is called their phylogeny [27].

5. PROTEIN STRUCTURE PREDICTIONS AND PROTEIN FOLDING

The three-dimensional structure of a molecule is considered as known when the precise location of each and every atom in it is known. The structure of a protein may be described at four major levels. The amino acid sequence of the polypeptide chain is called its *primary structure*. The next level of the arrangement of atoms in the protein is called its *secondary structure* (see Fig. 3).

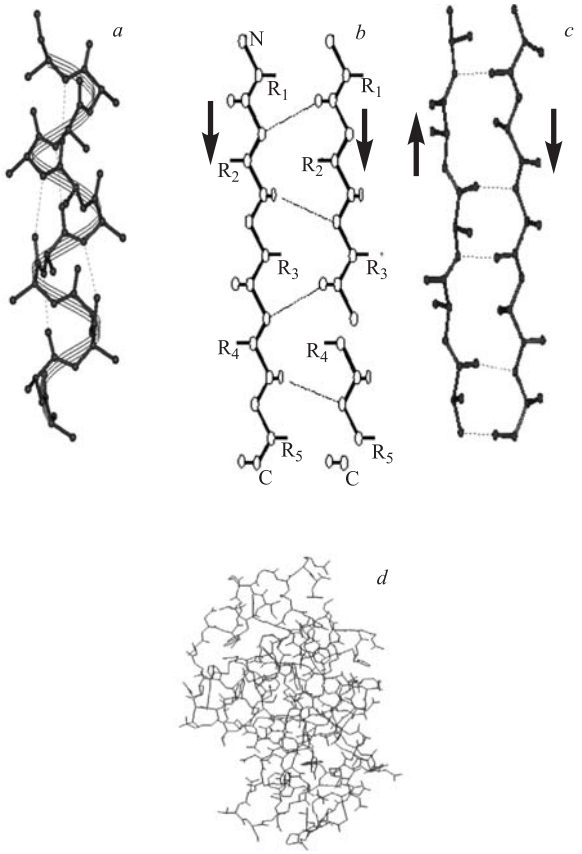


Fig. 3. *a*) The α-helix. *b*) β-strands. Two or more such strands may come together to form β-sheets in parallel (*b*) or antiparallel (*c*) orientation. *d*) The tertiary structure is shown as a line diagram

A helix is a structure with a typical repetition after 3.6 amino acids (with exception of the 3–10 helix, which is tighter and has a repetition after 3 amino acids; it is found in 3.4% of all helices). The beta sheet can be parallel (the same direction of the chain) or antiparallel, it may contain two or more chains and it has a typical length of 5–10 amino acids.

Tertiary Structure is the native state, or folded form, of a single protein chain. Tertiary structure of a protein includes the coordinates of its residues in three-dimensional space.

Quaternary Structure is the structure of a protein. Some proteins form a large assembly to function. This form includes the position of the protein subunits with respect to each other.

5.1. Protein Secondary Structure Prediction. For the purposes of prediction, every residue in a protein chain is always considered to exist in one of three (or four) secondary structural states. These are: helix, usually represented as H; beta strand, represented as B or E (for «extended»); and random coil, signified as C. Lower case letters are used when the prediction is not very certain. Often an additional state is also predicted, namely turn, signified as T. The output of most secondary structure prediction algorithms and programs is the sequence of the protein along with one of the above symbols for each residue.

5.1.1. Neural networks in secondary structure prediction – PHD. Computer-based artificial neural networks (ANNs) thus consist of units analogous to neurons that receive input information from other units and send output signals to others. The connections are many-to-one and one-to-many [29] (see Fig. 4).

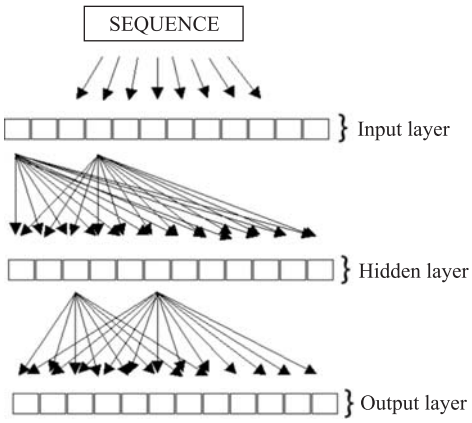


Fig. 4. A schematic diagram of a feed-forward multilayered artificial neural network (ANN). For clarity, only some of the interconnections between the layers are shown. But, in fact, every element of a layer is connected to every element of the next layer

Here we describe one of the most successful and commonly used implementations, namely PHD. PHD consists of several steps other than the ANN. First, the input sequence is compared with available sequences in the database and a multiple sequence alignment with all similar sequences is constructed [30]. In order to obtain the structure prediction at each residue position, i. e., each column in this multiple alignment is considered and the following information is extracted and fed into the ANN: the profile of amino acid substitutions; the weights for each amino acid type compiled for all the columns in the alignment; the numbers of insertions/deletions (indels) in each column; the position of the window with respect to the entire sequence; and the amino acid composition and length of the protein. All this information has been incorporated into the gating function and the ANN has been constructed and trained with four units in the hidden layer and three units in the output layer.

5.2. Protein Tertiary Structure Prediction. In general, the techniques can be divided into three broad categories: homology modeling, threading techniques, and *ab initio* structure prediction.

5.2.1. Homology modeling. This is also known as comparative protein modeling or knowledge based modeling. Broadly the technique consists of four steps: selecting the template, alignment of target with template, building the model, and evaluating the model.

The first step, selecting the template, is the most important one. One may then be confident that they share a common function and therefore a common structure. Regions of the protein that normally have divergent structures, such as loops and turns have similar structures only when the sequence identity is greater than 50%. Also the number of insertions and/or deletions increases as sequence identity decreases.

Template selection is facilitated by the availability of a several sequence and structure databases and efficient software for matching the target sequence with these. Programs like BLAST, FASTA, etc., when used on databases such as the PDB, CATH, etc., swiftly identify possible templates. A refinement of this technique is the use of multiple sequence alignments. The target sequence is aligned with families of sequences that are already categorized as possessing similar structures and functions.

After selecting the template, the second step in the modeling procedure is to align the target sequence with the template sequence. It is better to perform multiple sequence alignment using programs such as CLUSTAL, or a variation of it. All possible templates are first multiply aligned and profile constructed. Other sequences belonging to the same family can also be added to the profile, which is then aligned to the target. The best possible template is then chosen as the initial model for the target.

The third step is to build the model, based on the target-template alignment. Building the model is to use the template to calculate restraints to be applied on

the target, such as inter-residue distances and angles, specific disulphide bonds, stacking interactions between aromatic residues, etc. Thus model building would consist of first building the backbone, then placing the side chains, and finally optimizing the entire structure.

The final step in the modeling process is evaluation of the model. The Ramachandran plot is a very good way of checking the geometry of the model and programs such as PROCHECK are available to carry out these tasks [31] (see Fig. 5).

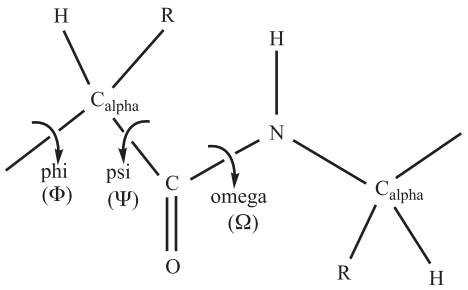


Fig. 5. Ramachandran plot showing the two torsion angles Φ and Ψ

Models have been used to identify active sites. A particular use of homology models is in drug design, where frequently small sequence changes in the certain crucial regions of the protein lead to loss of effect for the drug. Such models can be used for detailed studies, for example, of the docking of small ligands or to define and study antibody epitopes.

5.2.2. Threading. Threading generalizes the technique of homology modeling, and aligns the unknown sequence, not to another sequence of known structure, but to a likely structure built from families of structures with sequences similar to the target. Threading is therefore also known as «fold recognition» algorithm, or «inverse folding», since we have a library of folds, and are looking to see which one best fits or «threads» the target sequence [32].

5.2.3. Ab initio structure prediction. Ab initio algorithm uses only the sequence of the protein, and the well-established laws and principles of physics and chemistry, to determine its three-dimensional structure. From the principles of physics is it clear that the final folded form of the protein is its minimum energy state. In order to be useful in structure prediction, the chief property that this function should possess is that its global minimum should represent the native structure of the protein. By global minimum, we mean that for all possible allowed structures of the protein [33].

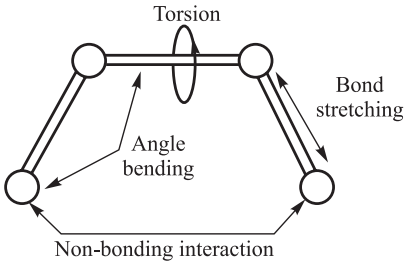
6. MOLECULAR SIMULATIONS ON PROTEIN STRUCTURES

The principles of force fields (also known as molecular mechanics) are based upon Newtonian mechanics. The basic idea is that bond lengths, valence and torsional angles have «natural» values depending on the involved atoms and that molecules try to adjust their geometries to adopt these values as closely as possible. Additionally, steric and electrostatic interactions, mainly represented by Van der Waals and Coulomb forces, are included in the so-called potential. These parameters are optimized to obtain the best of experimental values, as geometries, conformational energies and spectroscopic properties.

6.1. Force Fields

6.1.1. Energy calculation:

$$E_{\text{total}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{torsion}} + E_{\text{non-bonding}}$$



Many of the molecular modeling force fields in use today can be interpreted in terms of a relatively simple four-component picture of intra- and intermolecular forces within the system [34].

Bond-energy. The energy between two bonded atoms increases, when the bond is compressed or stretched. The potential is described by an equation based on Hooke's law for springs [35]

$$E_{\text{bond}} = \sum_{\text{bonds}} k_b (r - r_0)^2,$$

where k_b is the force constant, r is the actual bond length and r_0 is the equilibrium length. This quadratic approximation fails as the bond is stretched towards the point of dissociation.

Angle-energy. Energy increases if the equilibrium bond angles are bent. Again the approximation is harmonic and uses Hooke's law [35]

$$E_{\text{angle}} = \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2,$$

where k_θ controls the stiffness of the angle; θ is the current bond angle and θ_o — the equilibrium angle. Both, the force and the equilibrium constant have to be estimated for each triple of atoms.

Torsion energy. Intra-molecular rotations (around torsions or dihedrals) require energy as well:

$$E_{\text{torsion}} = \sum_{\text{torsions}} \frac{V_n}{2}\left(1 + \cos\left(n\omega - \gamma\right)\right).$$

V_n controls the amplitude of this periodic function, n is the multiplicity, and the so-called phase factor, shifts the entire curve along the rotation angle axis z. Again the parameters V_n, n and γ for all combinations of four atoms have to be determined [35].

Non-bonding energy. The simplest potential for non-bonding interactions includes two terms, Van der Waals and Coulomb terms [35]

$$E_{\text{non-bonding}} = \underbrace{\sum_i \sum_{j>i} \left(\frac{A_{ij}}{r_{ij}^6} - -\frac{B_{ij}}{r_{ij}^{12}} \right)}_{\text{Van der Waals}} + \underbrace{\sum_i \sum_{j>i} \frac{q_i q_j}{r_{ij}}}_{\text{Coulomb}}.$$

The Van der Waals term accounts for the attraction and the Coulomb term for electrostatic interaction. The shown approximation for the Van der Waals energy is of the Lennard–Jones 6–12 potential type.

6.1.2. Molecular dynamics. Molecular dynamics employs the so-called united atom method, where atom groups with nonpolar hydrogen atoms are treated as an ensemble. The inclusion of the solvent can be done explicitly where the solute is immersed in a cubic box of solvent molecules. The use of non-rectangular periodic boundary conditions, stochastic boundaries and «solvent shell» can help to reduce the number of solvent molecules required and therefore accelerate the molecular dynamics simulation [36]. When using implicit solvent models in molecular dynamics simulations, there are two additional effects to bear in mind. The solvent also influences the dynamical behavior of the solute: 1) via random collisions, and 2) by imposing a frictional drag on the motion of the solute through the solvent. While explicit solvent calculations include these effects automatically, it is also possible to incorporate these effects of solvent without requiring any explicit specific solvent molecules to be present. The Langevin equation of motion is the starting point for these stochastic dynamics models [37]

$$F_i(t) = m_i a_i(t) = m_i \frac{\partial^2 r_i(t)}{\partial t^2}, \quad \text{whereas} \quad F_i(t) = -\frac{\partial E_{\text{tot}}}{\partial r_i},$$

$$m_i \frac{\partial^2 r_i(t)}{\partial t^2} = F_i\left(r_i(t)\right) - \gamma_i m_i \frac{\partial r_i(t)}{\partial t} + R_i(t).$$

The first component is due to interactions between the particle and other particles. The second force arises from the motion of the particle through the solvent and is equivalent to the frictional drag on the particle due to the solvent. γ_i is often referred to as a friction coefficient. The third contribution, the force $R_i(t)$, is due to random fluctuations caused by interactions with solvent molecules.

First, the number of non-bonded interactions in a molecule grows as $n(n-1)/2$, where n is the number of atoms in the molecule. Second, this non-bonded interaction term must include the solvation effects, because biomolecules are usually solvated in water. This solvation has a major influence on the electrostatic forces [38]. The most accurate way for describing this solvation is including the solvent and counter-ions explicitly. Such an «explicit solvent» approach increases the number of particles considerably, because a lot of solvent molecules are need for an accurate description of salvation [39]. General to all these force fields are simple approaches for bond, angle and torsion potentials, to reduce the calculation time for the energy function and the gradient. The most prominent of these force fields is the Cornell force field of AMBER. One of the most widely used force fields is AMBER (Assisted Model Building with Energy Refinement). It is suitable for the calculation of two the most important types of macromolecules in biochemistry, namely peptides and nucleic acids. The current version of the package, AMBER8.0 is comprised of several modules that fulfill specific tasks.

There are four major input data to AMBER modules:

1) Cartesian coordinates for each atom in the system;

2) «Topology»: connectivity, atom names, atom types, residue names and charges;

3) Force field: Parameters for all of the bonds, angles, dihedrals and state parameters desired;

4) Commands: The user specifies the procedural option and state parameters desired. The modules can be divided into three categories.

Preparatory programs. LEaP is the primary program to create the AMBER specific topology file prmtopand and the coordinate file prmcrd.

Energy programs. SANDER is the energy minimizer and molecular dynamics module, GIBBS is the free energy perturbation program, NMODE is the normal mode analysis program and ROAR — a module, where parts of the molecule can be treated quantum mechanically and others with molecular mechanics.

Analysis programs. ANAL is created for analyzing single conformations, CARNAL — to examine molecular dynamics simulations. The AMBER force field, or better the Cornell force field, consists of five potential terms

$$
\begin{aligned}
E_{\text{total}} = & \sum_{\text{bonds}} K_b (r - r_0)^2 + \\
& + \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 + \\
& + \sum_{\text{torsions}} \frac{V_n}{2} \left(1 + \cos\left(n\omega - \gamma\right)\right) + \\
& + \sum_i \sum_{i<j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}} \right] + \\
& + \sum_{\text{H}-\text{bonds}} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{10}} \right].
\end{aligned}
$$

The most critical term, even for biomolecules, are the non-bonded interactions.

7. HOMOLOGY MODELING AND MOLECULAR DYNAMICS OF CYCLIN-DEPENDENT KINASES

Although protein function is the best determined experimentally [40, 41], it sometimes can be predicted by matching the unknown sequence of a protein with proteins of known function [41–43]. Sequence-based predictions of function can be improved by considering three-dimensional (3D) structure of proteins. This is possible because similar protein sequences tend to have similar functions, although exceptions also occur [44]. In addition, because evolution tends to conserve function, which depends more directly on structure than on sequence, structure is more conserved in evolution than sequence and the net result is that patterns in space are frequently more recognizable than patterns in sequence [45]. Among all current theoretical approaches, modeling is the only method that can reliably generate a 3D model of a protein (target) from its amino acid sequence [46, 47]. The fraction of the known protein sequences that have at least one segment related to one or more known structures varies with a genome, currently ranging from 20 to 50% [48–55]. To gain a three-dimensional fabrication for the unknown sequence one must have at least one experimentally solved 3D structure (template) that has a significant amino acid sequence similarity to the target sequence. The idea of an easy-to-use, automated modeling facility with integrated expert knowledge was first implemented 50 years ago by Peitsch et al. [56–58]. The prediction process consists of search for structural homologs, target–template alignment, model building, and model assessment and structure validation.

The cyclin-dependent kinases (CDKs) belong to the serine/threonine-specific protein kinases subfamily. The enzymes catalyze the transfer of γ-phosphate in adenosine triphosphate (ATP) to a protein substrate. CDKs are crucial regulators in timing and coordination of eukaryotic cell cycle events. Transient activation of

these kinases at specific cell stages is believed to trigger the principal cell cycle transitions, including the DNA replication and the entry into mitosis. In yeast, transition events are controlled by a single CDK (CDK1/CDC28 in *Saccharomyces cerevisiae* [88]) and several cyclins, while in human, cell cycle progression is governed by several CDKs and cyclins. In particular, CDK4-cyclin D is required to pass through G1, CDK2-cyclin E for the G1 to S phase transition, CDK2-cyclin A to progress through the S phase and CDC2-cyclin B to reach the M phase. Cell cycle-dependent oscillations in CDK activity are induced by complex mechanisms that include binding to positive regulatory subunits (cyclins) and phosphorylation at positive and negative regulatory sites. After cyclin binding occurs, a separate protein kinase, known as the CDK-activating kinase, phosphorylates the CDK catalytic subunit on a threonine residue (T160 in human CDK2 and T169 in yeast CDC28) in T-loop. Under some circumstances, CDK2 can also be negatively regulated by phosphorylation on Y15 and T14 in G-loop or binding to inhibitor.

The CDK2 and CDC28 proteins have been extensively studied. The central role that CDKs play in cell division timing, in cell cycle regulation and repair together with the high incidence of genetic alteration of CDKs or deregulation of CDK inhibitors observed in several cancers, made CDC28 a very attractive model for structural and functional CDK studies. Crystallographic studies of several eukaryotic protein kinases have shown that they all share the same fold and tertiary structure. The crystal structure of the human CDK2 [89, 90] has been served as a model for the catalytic core of other CDKs, including CDC28 [91, 92]. But correctness of such approximation is under the question.

7.1. Materials and Methods

Search for structural homolog. In this study the three-dimensional structure for the yeast cyclin-dependent kinase CDC28 (Uniprot accession number: P00546) is modeled using the MODELLER program (see Fig. 7). The primary structure for the yeast CDC28 has 298 amino acids and can be obtained from SWISS PROT database (http://cn.expasy.org/sprot/) server. The modeling step can be carried out by searching the yeast CDC28 sequence against the databases of well-defined template sequences derived from Protein Data Bank entry (http://www.rcsb.org/pdb/)

The MODELLER program calculates the three-dimensional structure for the query sequence by searching for the related matching structures using satisfaction of spatial restraints [59] (see Fig. 6). The spatial restraints include: (i) homology-derived restraints on the three-dimensional geometrical information including the distances and dihedral angles in the unknown query sequence, obtained from its alignment with the template structures [59]; (ii) stereochemical restraints such as bond length and bond angle preferences, obtained from the CHARMM-22 molecular mechanics force field [60]; (iii) statistical preferences for dihedral angles and non-bonded interatomic distances, obtained from a representative set of known protein structures [61]; and (iv) optional manually curate restraints, such as

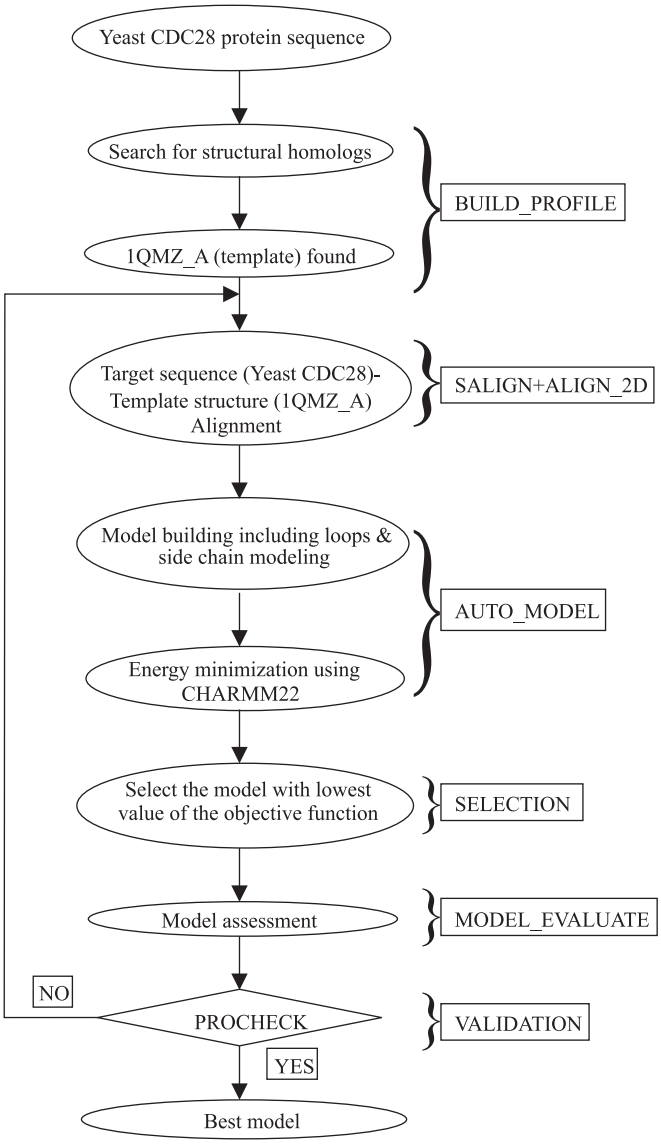


Fig. 6. Flowchart for homology modeling of yeast CDC28

those from NMR spectroscopy, rules of secondary structure packing, cross-linking experiments, fluorescence spectroscopy, image reconstruction from electron microscopy, site-directed mutagenesis and intuition. The spatial restraints, expressed

as probability density functions, are combined into an objective function that is optimized by a combination of conjugate gradients and molecular dynamics with simulated annealing.

The MODELLER searches the templates used for model building, which is a representative of multiple structure alignments can be obtained from DBALI [62]. Sequence profiles are defined as the sequence position — specific scoring matrix. This scoring matirx is designed for both the yeast CDC28 protein (target) sequences and the 1QMZ_A sequence (template) by searching in contrast with the Swiss-Prot/TrEMBL database of sequences. The BUILD PROFILE module of the MODELLER executes this sequence profile construction. The BUILD_PROFILE command has many options. Unrecognized residues are ignored. In this study the structural homolog search is set to use the BLOSUM62 similarity matrix inbuilt in the MODELLER program itself. Consequently, the parameters for the gap penalties are set to the appropriate values for the BLOSUM62 matrix. A match is reported if its exponential value falls below the threshold set. Lower exponential value thresholds are more stringent, and report fewer matches. Many hits were displayed on the basis of the sequence identity and exponential value between the protein sequences. The query sequence found 64.11% identity and value $E = 0$ with PDB entry: 1QMZ (phosphorylated CDK2-cyclin A-substrate peptide complex) of the human species by running the MODELLER program (Fig. 7). The matching part of the PDB entry: 1QMZ chain-A derived from the significant hit was used as the template structure for the model building.

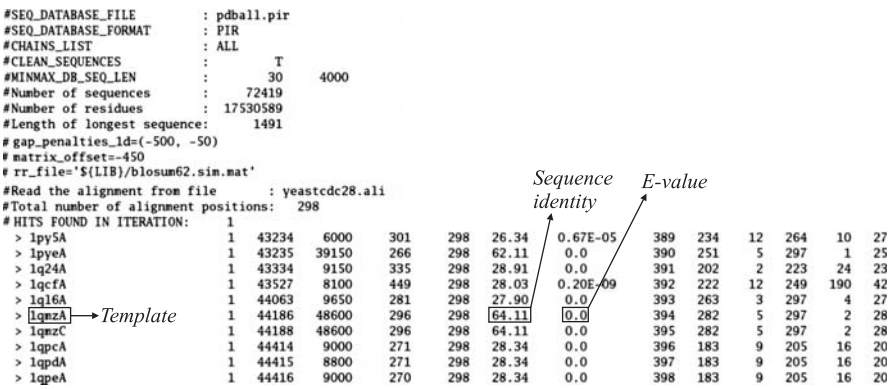


Fig. 7. Sequence identity and exponential value between the protein sequences. The query sequence found 64.11% identity and value $E = 0$ with PDB entry: 1QMZ (phosphorylated CDK2-cyclin A-substrate peptide complex) of the human

Target–template alignment. The alignments between the yeast CDC28 and 1QMZ_A is executed by the SALIGN module of the MODELLER, which relies on a multiple structure alignment method similar to that in the COMPARER program [63]. Target sequence and template structure matches are determined by aligning the target sequence profile against the template profiles, using local dynamic programming in the SALIGN module which is similar to that of PSI-BLAST [64] and COMPASS [65]. This alignment tends to be more accurate than the PSI-BLAST alignment because (i) it engages all the sequences and structures that are qualified known to be matching with the target sequence, (ii) it incorporates a structure-dependent gap penalty function to position gaps in a group of related structures, and (iii) it finds the matching part of the complete structural domains as obtained from the known template structures.

In order to analyze the close relation between the target and template protein sequences, we carry out the comparative modeling procedure. Comparative modeling requires the information on target–template alignment. Now the matching parts of the template structure and the unknown sequence were realigned by the use of the ALIGN-2D command of the MODELLER program [45]. This command executes a global dynamic programming method for comparison between the target–template sequences and also relies on the observation that evolution tends to place residue insertions and deletions in the regions that are solvent exposed, curved, outside secondary structure segments, and between two Cα positions close in space [66]. Gaps are included between the target–template alignment, in order to get maximum correspondence between the protein sequences. Gaps in these regions of high correspondence are favored by variable gap penalty function that is executed from the template structure alone. In principle, the errors between the target–template alignment is greatly minimized almost by one-third relative to the present day sequence alignment methods (Fig. 8).

Models are calculated for each of the sequence-structure matches using the MODELLER program [59]. Nevertheless, there is clearly a need for even more accurate sequence-structure alignments and for using multiple template structures, so that more accurate models are obtained [26]. The resulting models are then evaluated by a composite model quality criterion that depends on the compactness of a model, the sequence identity of the sequence-structure match and statistical energy Z-scores [67].

Model building. Here we are discussing about the generation of the three-dimensional structure for the unknown yeast CDC28 protein sequence (target) with PDB: 1QMZ_A (template) as its suitable structural homolog. There are a few steps in construction of the three-dimensional model. MODELLER builds the model for the unknown sequence using spatial restraints. Initially, spatial restraints parameters including the distance and dihedral angles on the yeast CDC28 sequence is obtained by the alignment with the 1QMZ_A (template) structure. Next, the alignments between the yeast CDC28 sequence vs. 1QMZ_A is searched in the

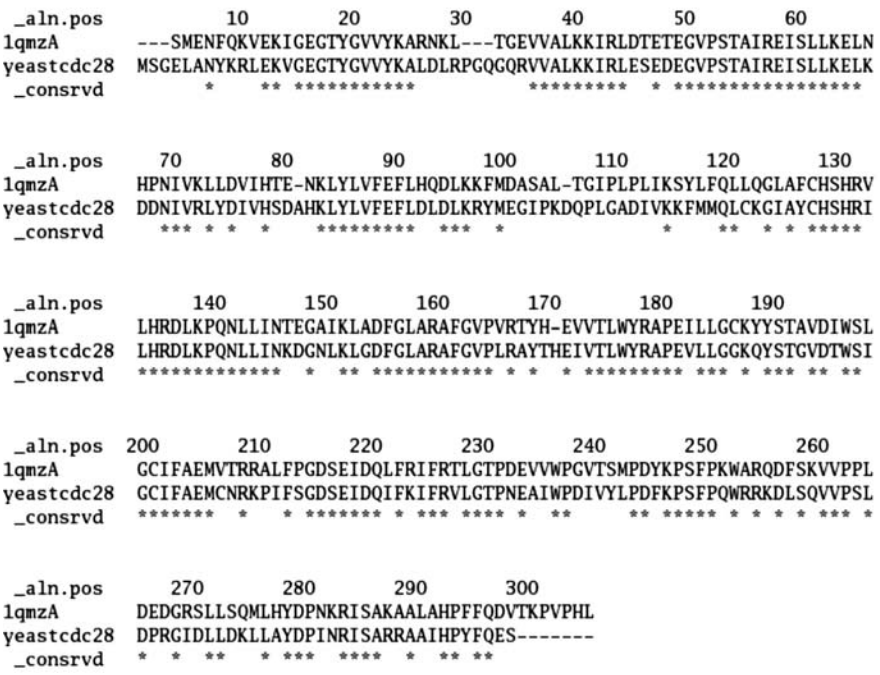


Fig. 8. Target–template alignment. «∗» shows the matching between the residues and «–» shows the gaps

database of alignments using the AUTO_MODEL module of the MODELLER program. The output of this module displays many restraints parameters between the target–template alignments including the distances, main chain dihedral angles, side chain dihedral angles, disulphide dihedral angle, NMR distant restraints and non-bonded restraints between these two proteins [59]. These relationships are expressed as conditional probability density functions (pdf's) and can be used directly as spatial restraints. The spatial restraints and the CHARMM22 force field terms enforcing proper stereochemistry [68] are combined into an objective function. These template derived restraints parameters are combined with the most of the CHARMM energy terms [68, 69] to obtain a full objective function. Then the model with the lowest value of the objective function is selected and assessed using the MODEL_EVALUATE module of the MODELLER program.

Loop modeling. It is expected that target sequences often have inserted/deleted (indels) residues with respect to the chosen template structures or some distinguishable regions where there is a high degree of variation in the structural information between these two proteins. These regions are generally addressed

as loops. Loops often have to play a leading role in describing the functional specificity, forming the active and binding sites. The MODELLER algorithm for the construction of the loops provided the use of the information based on spatial restraints. To simulate comparative modeling problems, the loop modeling procedure was evaluated by predicting loops of known structure in only approximately correct environments, which were obtained by distorting the anchor regions, corresponding to the three residues at either end of the loop, and all the atoms within 50 Å of the native loop conformation for up to 2–3 Å by molecular dynamics simulations [59].

Side chain modeling. Geometry of representing a side chain conformation is determined based on the steric or energy considerations and from similar structures, i. e., from the suitable templates [70, 71]. The construction of the disulphide bridges for the query sequence is built from disulphide bridges in existing protein structures [72, 73] and from relevant disulfide bridges in closely related structures with respect to the unknown sequence [74]. The disulphide bridges for yeast CDC28 are built with reference to the experimental structure.

Selecting the appropriate model for the yeast CDC28. A three-dimensional model was generated by a GA341 score (DOPE energy, see Fig. 9) that is the parameters including Z-score (Zs) calculated with a statistical potential function [75],

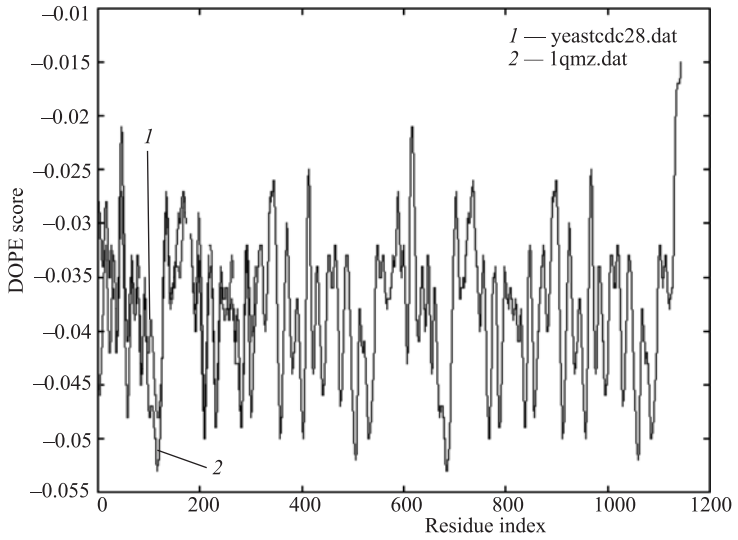


Fig. 9. Graph plot between the DOPE energy and residues for both 1QMZ and yeast CDC28. The overlapping structure shows the A chain, i. e., the kinase part of 1QMZ is correlated with modeled CDC28 showing high homology

target–template sequence identity (Si) and a measure of structural compactness (Sc) [75, 66]. The GA341 score is defined as

$$\mathrm{GA341} = 1 - -[\cos(\mathrm{Si})]^{(\mathrm{Si+Sc})/\exp(\mathrm{Zs})}.$$

Sequence identity is defined as the fragments of positions with identical residues in the yeast CDC28 (target) – 1QMZ_A (template) alignment. Structural compactness is the ratio between the sum of the standard volumes of the amino acid residues in the protein and the volume of the sphere with the diameter equal to the largest dimension of the model.

The Z-score is calculated for the combined statistical potential energy of the generated model, using the mean and standard deviations of the statistical potential energy of random sequences with the same composition and structure as the model [75]. Finally, from the set of five generated models for the yeast CDC28 sequence the model with lowest energy is selected (Figs. 9 and 10).

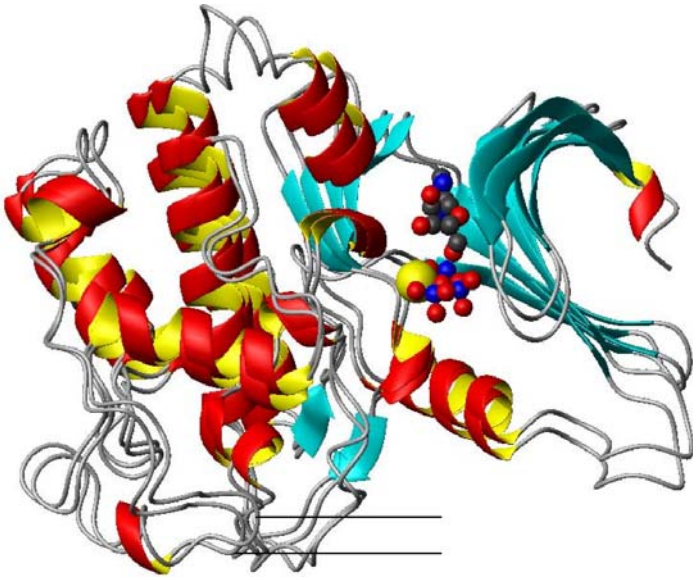


Fig. 10. Superposition of the target structure (PDB: 1QMZ A Chain) and the modeled template structure (yeast kinase CDC28). ATP complex is represented by ball models. Magnesium ion is shown as large sphere

Assessment of the model. This is necessary to allow the MODELLER to calculate correctly the energy, and additionally allows for the possibility that the PDB file has atoms in a non-standard order, or has different subsets of atoms (e. g., all atoms including hydrogen, while the MODELLER uses only heavy

atoms, or vice versa). The final correctness of the artificially generated three-dimensional model for the yeast CDC28 was determined by comparison with the corresponding to the high similarity structure 1QMZ_A extracted from the Protein Data Bank (PDB) [76]. The root mean square deviation (RMSD) between the corresponding Cα atoms of the artificially generated three-dimensional model and the native structure, i. e., 1QMZ_A were calculated upon rigid body least squares superposition of all the Cα toms. Next, the percentage of high matching regions between the yeast CDC28 and the 1QMZ_A was defined in terms of the percentage of the Cα atoms in the model that are located within the proximity of 5 Å of the corresponding atoms in the superposed structure (Figs. 11 and 12). In order to enhance the best model the MODELLER program finally incorporates corresponding alignment through a comparison with the structure-based alignment produced by the CE program [77]. The percentage of high matching positions was defined as the percentage of positions in the structure-based alignment between

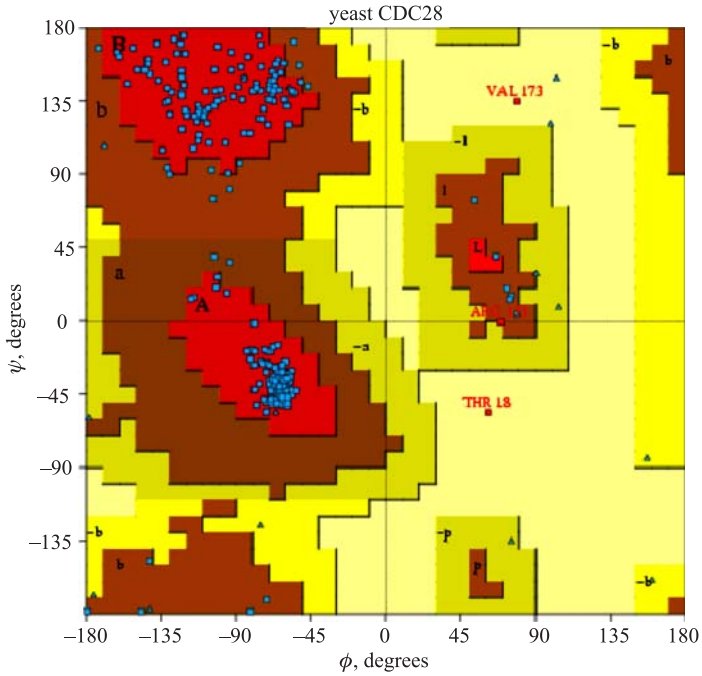


Fig. 11. Ramachandran plot for yeast CDC28. Most favoured regions = 236 (number of residues), 91.5% (percentage). Additional allowed regions = 19 (number of residues), 7.4% (percentage). Generously allowed regions = 1 (number of residues), 0.4% (percentage). Disallowed regions = 2 (number of residues), 0.8% (percentage). Non-glycine and non-proline residues = 258, 100.0% (percentage). End-residues (excl. Gly and Pro) = 2. Glycine residues = 21. Proline residues = 17. Total number of residues = 298

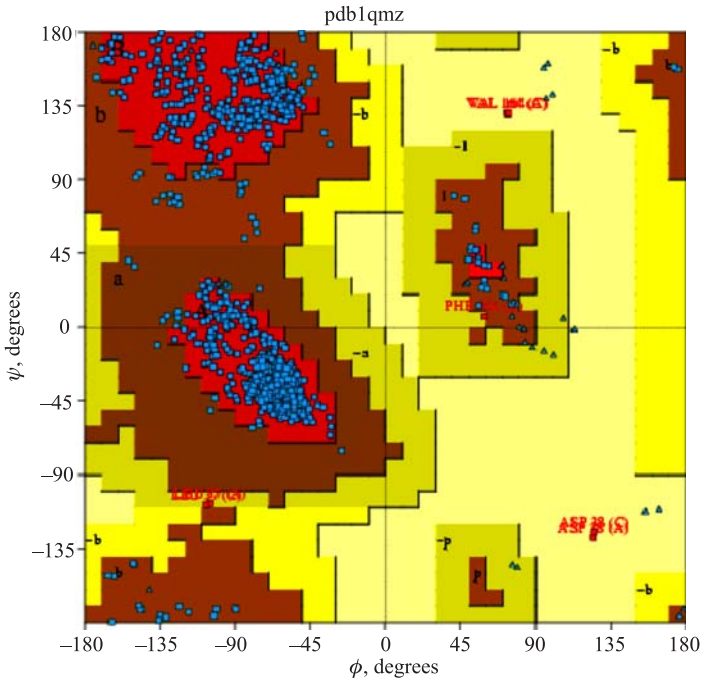


Fig. 12. Ramachandran plot for pdb 1QMZ. Most favoured regions = 906 (number of residues), 91.1% (percentage). Additional allowed regions = 81 (number of residues), 8.1% (percentage). Generously allowed regions = 3 (number of residues), 0.3% (percentage). Disallowed regions = 4 (number of residues), 0.4% (percentage). Non-glycine and non-proline residues = 994, 100.0% (percentage). End-residues (excl. Gly and Pro) = 12. Glycine residues = 50. Proline residues = 68. Total number of residues = 1124

the yeast CDC28 and 1QMZ_A structure. The residues that are matching with the gap positions are neglected in this operation.

Structure validation. Validation refers to the procedure for assessing the quality of deposited atomic models (structure validation) and for assessing how well these models fit the experimental data. Validation parameters include the covalent bond distances and angles, stereochemical validation, atom nomenclature are taken care. Moreover, all the distances between the atoms including the water oxygen atoms and all polar atoms (oxygen and nitrogen) of the macromolecules, ligands and solvent are calculated. The results are displayed along with the PROCHECK server (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/) and Ramachandran plot.

Ramachandran plot displays the phi (Φ) and psi (Ψ) backbone conformational angles for each residue in a protein. The phi (Φ) angle is the angle of right-hand

rotation around N-Cα bond and the psi (Ψ) angle is the angle of right-hand rotation around Cα–C bond. Φ and Ψ angles are also used in the classification of some secondary structure elements such as alpha helix and beta turns.

In a Ramachandran plot, the core or allowed regions indicates the preferred areas for Ψ/Φ angle pairs for all residues in a protein. If the determination of protein structure is reliable, most pairs will be in the favored regions of the plot and only a few will appear in the «disallowed» regions. The score for the crystal structure 1QMZ is 91.1% (Fig. 12). The score of 91.5% for yeast CDC28 (Fig. 11) lays in the allowed region and confirms good homology prediction.

7.2. Molecular Dynamics Simulations. For the MD simulations, the SANDER modules of the program package AMBER8.0 [78] and of the modified version of AMBER7.0 for a special-purpose computer MDGRAPE-2 [79] were used. The starting geometries for the simulations were prepared using X-ray structures from the Brookhaven Protein Data Bank (http://www.pdb.org). The all-atom force field [80] was used in the MD simulations. A system was solvated with TIP3P molecules [81], generated in a spherical (non-periodic) water bath. The system temperature was kept constant by the Berendsen algorithm with 0.2 ps coupling time [82]. Only bond lengths involving hydrogen atoms were constrained using the SHAKE method [83]. The integration time step in the MD simulations was 1 fs. The simulation procedures were the same in all calculations [84]. Firstly, a potential energy minimization was performed for each system on an initial state. Then, the MD simulation was performed on the energy-minimized states. The temperatures of the considered systems were gradually heated to 300 K and then kept at 300 K for the next 2 million time steps [85]. The trajectories at 300 K for 2 ns were compared and studied in detail. The simulation data and images of simulated proteins results were analyzed by RasMol [86] and MOLMOL [87] packages. Complete data flow in AMBER is shown in Fig. 13.

Root mean square deviation. A very popular quantity used to express the structural similarity is the root mean square distance (RMSD) calculated between equivalent atoms in two structures, defined as

$$\mathrm{rmsd} = \sqrt{\frac{\sum_i d_i^2}{n}},$$

where d is the distance between each of the n pairs of equivalent atoms in two optimally superposed structures. The RMSD is 0 for identical structures, and its value increases as the two structures become more different. RMSD values are considered as reliable indicators of variability when applied to very similar proteins, like alternative conformations of the same proteins. In other words, RMSD is a good indicator for structural identity, but not so good for structural divergence. The RMS deviation of the MD structures from the crystal 1QMZ and the modeled structure of the yeast cyclin-dependent kinase CDC28 vs. time is calculated.

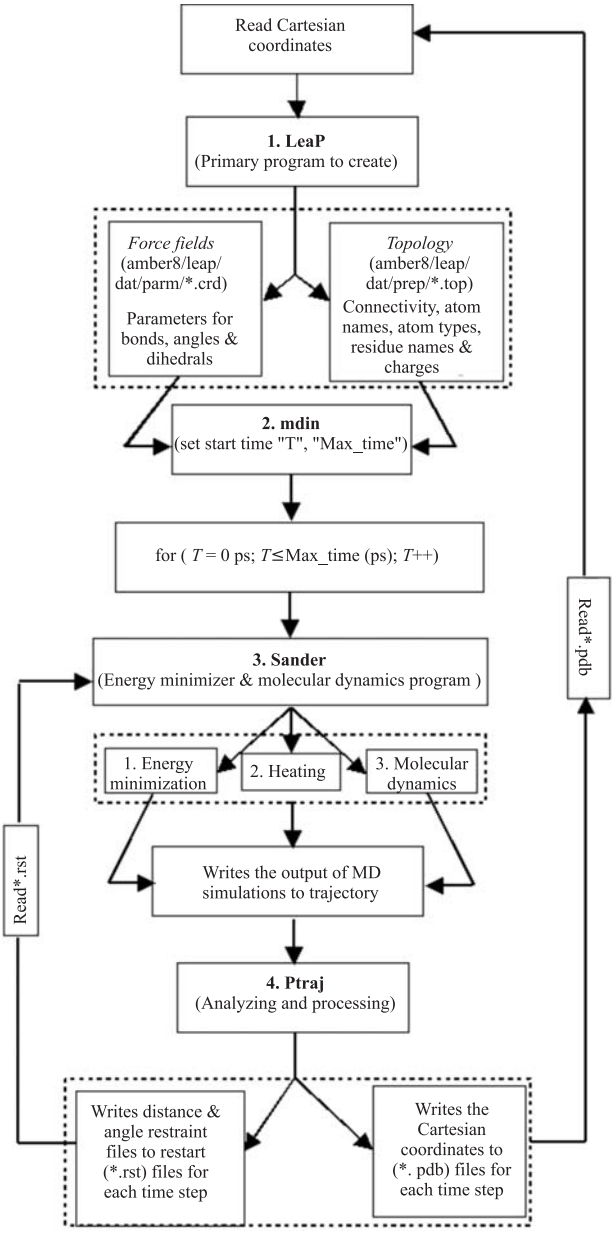


Fig. 13. Flowchart of data flow in the AMBER program

This relatively small deviation indicates that the dynamic structure of the 1QMZ and CDC28 remained in the realm of the crystal geometry during the course of the simulation and is further inherent stability of the model. The main discrepancy is found in the β regions. During the MD simulation the whole structure relaxed from its initial model structure with increasing RMSD and finally RMSD remained stable around an average of 2.0 Å over a considerable period of the latter part of the trajectory. This indicates that the structure has reached a stable average one. Many of the features are common for both plots and for much structures the values are well matched (Fig. 14).

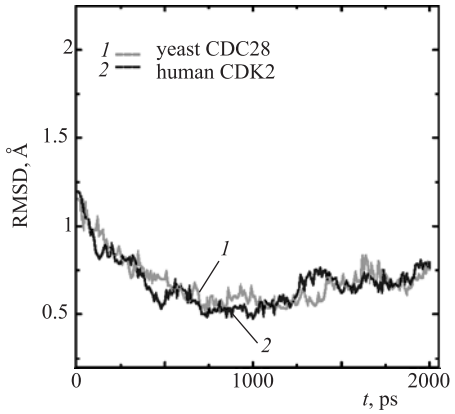


Fig. 14. Averaged RMSD (root mean square deviations) for crystal structure 1QMZ and the modeled structure of the yeast cyclin-dependent kinase CDC2 8 vs. time is calculated

Root mean square fluctuation. Root Mean Square Fluctuation (RMSF) at time t of atoms in a molecule with respect to the average structure is defined as

$$\rho_i^2 = \langle \Delta r_i^2 \rangle = \frac{1}{N} \sum_{k=1}^{N} \Delta r_i^2,$$

where Δr_i — atomic displacement from average position, N — total number of structures.

By observing the graph of RMSF for the 1QMZ crystal structure, it is shown that the sharp peaks arise due to the presence of beta sheets. During the course of MD trajectory these beta strands are prone to have more fluctuations than the alpha helices. These beta strands are more flexible due to the presence of hydrogen bonds. The regions corresponding to the residues Phe5 to Glu13 (FQKVEKIGE), Val18 to Asn24 (VVYKARN), Val30 to Lys34 (VVALK), Leu67 to Ile71 (LLDVI), Tyr78 to Glu82 (YLVFE), Val124 to Leu125,

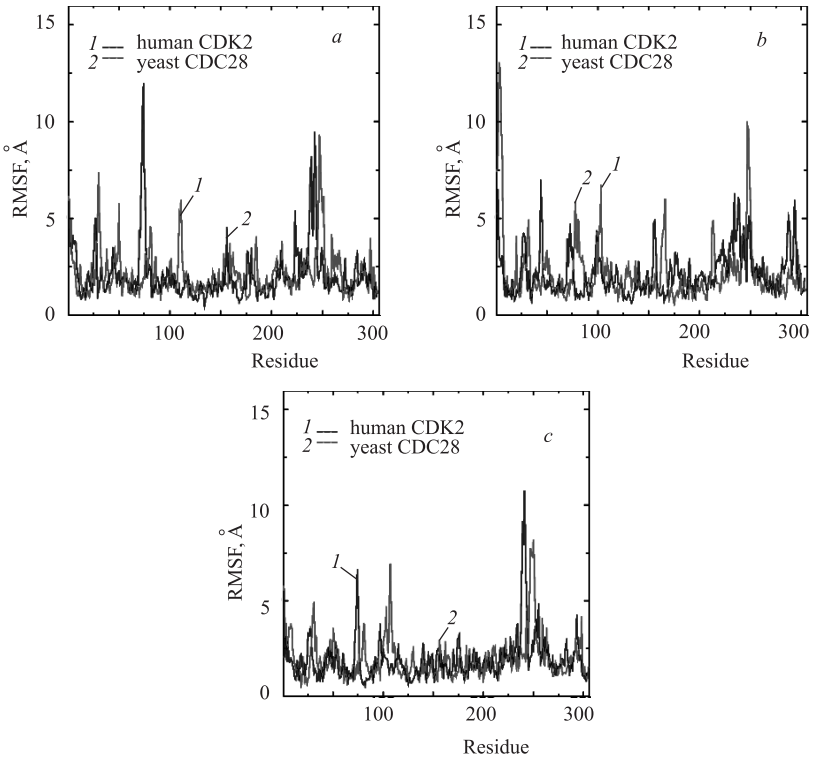


Fig. 15. Various behavior of Root Mean Square Fluctuation (RMSF) for: *a*) 15 ps, *b*) 1 ns and *c*) 2 ns for CDK2 and CDC28, respectively

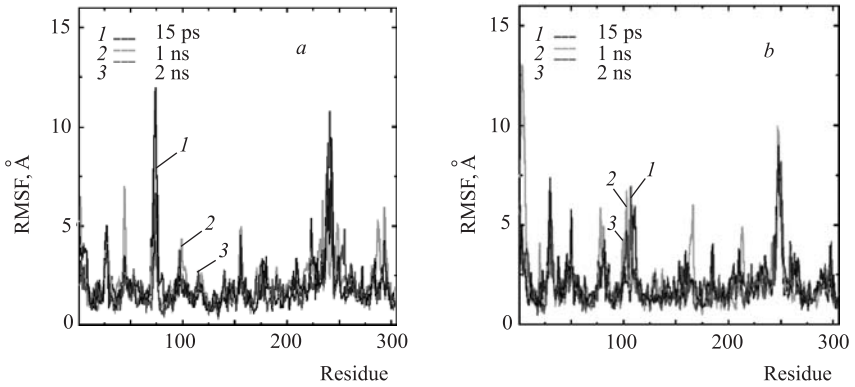


Fig. 16. Summary of various behavior of Root Mean Square Fluctuation (RMSF) for CDK2 (*a*) and CDC28 (*b*), respectively

Leu134 to Asn137 (LLIN), Ala141 to Leu144 (AIKL), Arg151 to Ala152 (RA) are the regions of beta strands which showed fluctuation during the trajectory. Similarly in the structure of modeled CDC28 yeast the residues are between the Tyr8 to Glu16 (YKRLEKVGE), Val21 to Asp27 (VVYKALD), Val36 to Ile42 (VVALKKI), Leu73 to Val77 (LYDIV), Leu84 to Glu89 (LYLVFE), Leu93 to Asp94 (LD), Ile132 to Leu133, Leu143 to Asn145 (LIN), Asn149 to Lys151 (NLK), Arg159 to Ala160 (RA) (Figs. 15 and 16).

Dynamic cross correlation map. The dynamic characteristics of the protein in MD simulation can be analyzed to yield information about correlated motion. Correlated motions can occur among proximal residues composing well-defined domain regions of secondary structure and also regions between the domains — domain communication. The extent of the correlated motion is indicated by magnitude of the corresponding correlation coefficient. The cross correlation coefficient for the displacement of any two atoms i and j is given by

$$\Delta C_{ij} = \langle \Delta r_i \Delta r_j \rangle / \sqrt{\langle \Delta r_i^2 \rangle \langle \Delta r_j^2 \rangle},$$

where Δr_i is the displacement of the mean position of the ith atom. The elements of Cij can be collected as in matrix form and displayed as three-dimensional cross correlation matrix (DCCM) map. The Cij are computed as averages over the successive backbone of N, Cα and C atoms to give one entry per pair of amino acid residues. There is time scale implicit in Cij as well. The intensity of the shading is proportional to the magnitude of the coefficient. The positive correlations are given in the upper triangle and the negative correlations are given in the lower triangle. Regions of regular secondary structures are expected to move in concert.

The DCCM map for each structure of 1QMZ and CDC28 was plotted over the time (15 ps, 1 ns, 2 ns), respectively (Figs. 17 and 18). The major cross peaks are found in the DCCM map in the areas of residues belonging to 1QMZ between 5–13, 18–24, 94–102, 153–164, 198–208, 233–255 and 281–295 from the in-

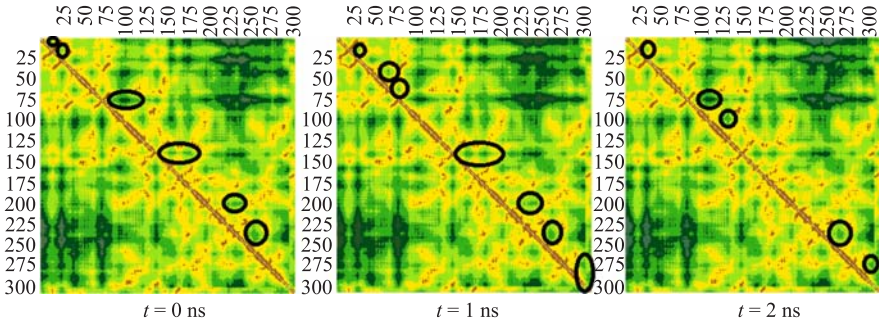


Fig. 17. Dynamic cross correlation map (DCCM) for 1QMZ structure for $t = 0$ ns, $t = 1$ ns and $t = 2$ ns. The black circles are the regions of β-sheets showing cross peaks

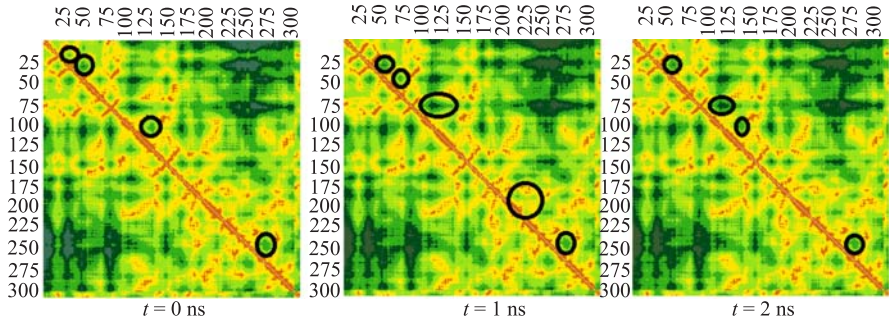


Fig. 18. Dynamic cross correlation map (DCCM) for yeast CDC28 structure for $t = 0$ ns, $t = 1$ ns and $t = 2$ ns. The black circles are the regions of β-sheets showing cross peaks

teraction of non-contiguous residues which fold to form the parallel β-sheets. Similarly the major cross peaks were also observed in CDC28 model of the residues between 21–27, 36–42, 103–110, 130–143, 208–217 and 242–256.

The CDK2/ATP and CDC28/ATP structural conformations. First, the inactive complex CDK2/ATP was analyzed. Analysis of the CDK2/ATP binary complex [89] indicates that ATP localizes in the cleft between the two lobes. Two loops, G-loop in small lobe and T-loop in large lobe, can be used to estimate a cleft width, which is very important for localization of ATP. G16 and T160 can serve as markers of G-loop and T-loop, respectively.

Simulated CDK2 structure (Figs. 19 and 20) was compared to the CDC28 after conformational change evaluations. The resulting wild-type CDK2 and CDC28 structural conformations are shown. The picture displays the initial (left)

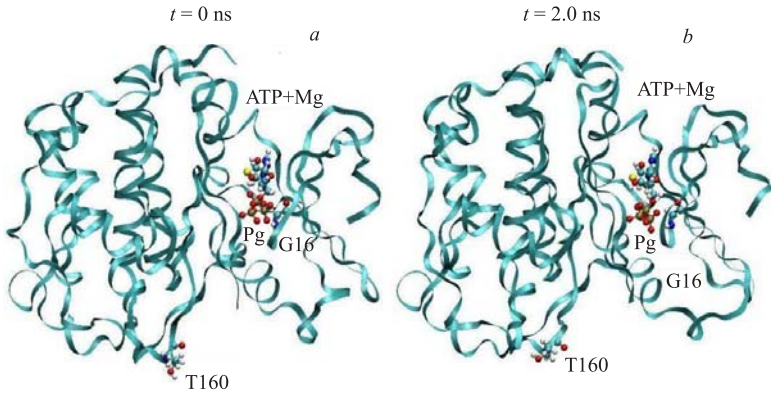


Fig. 19. The initial (a) and final (b) (2-ns state) structures of the CDK2/ATP complex. The ATP molecule and residue 16 of the G-loop are represented by ball models

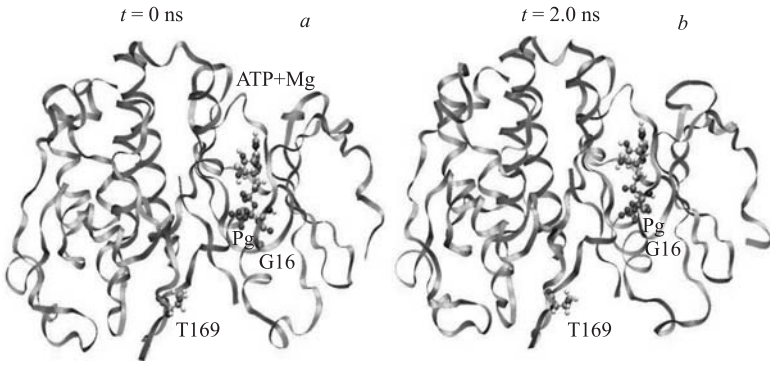


Fig. 20. The initial (a) and final (b) (2-ns state) structures of the CDC28/ATP. The ATP molecule and residue 16 of the G-loop are represented by ball models

and the final (right) 2-ns states. Positional changes between the ATP, residue G16 in G-loop and T160 in T-loop (the latter covers a left bottom α-helix) are shown in Fig. 21, a, b. Comparing initial and final states of CDK2 and CDC28 structures, small difference was visually observed. So, for the protein structures the original state is kept in a relatively stable conformation.

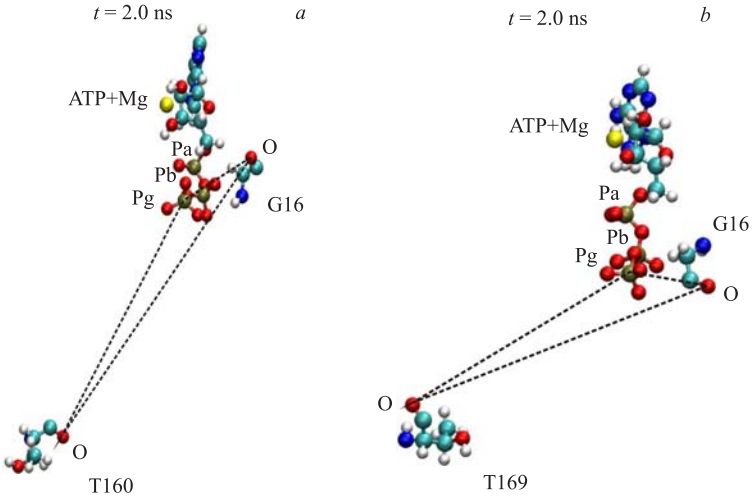


Fig. 21. a) The relative positions of the T160, ATP and res16 (an «activation triangle») are shown for CDK2. b) The relative positions of the T169, ATP and res16 (an «activation triangle») are shown for CDC28. The ATP molecule, residues T160 and 16 are represented by ball models

An activation triangle around the ATP. The T160, ATP and G16 positions (an «activation triangle») of CDK2 structure in the final (2-ns) state are represented in Figs. 22 and 23 aiming to estimate (although indirectly) the possibility of the hydrogen bond formation in the ATP and G-loop region. Similarly the

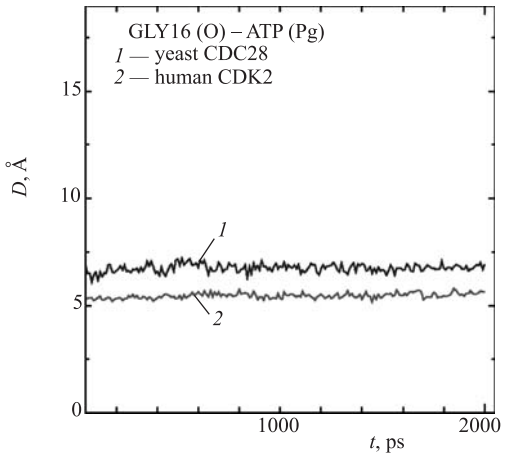


Fig. 22. The time dependences of the res16-ATP distance are shown for the CDK2 and CDC28, respectively, in accordance to the «activation triangle»

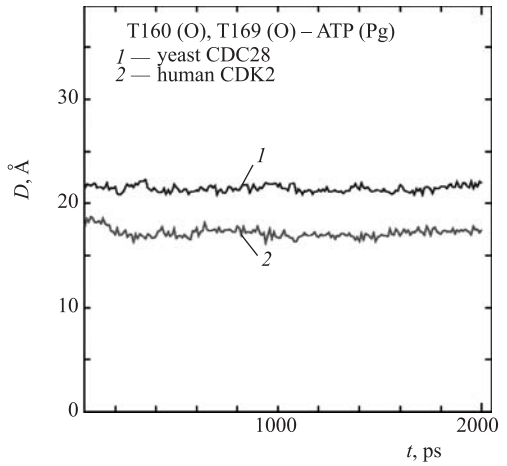


Fig. 23. The time dependences of the T160-ATP, T169-ATP distances are shown for the CDK2 and CDC28, respectively, in accordance to the «activation triangle»

T169, ATP and G20 positions (an «activation triangle») of CDC28 structure in the final (2-ns) state are represented in Figs. 19, *b* and 20, *b*, respectively. The ATP-res16 and ATP-res20 distances for the CDK2 and CDC28 structures, respectively, show a completely different behaviors (Fig. 24). The ATP-res16 and ATP-res20 distances in the CDK2 and CDC28 structures, respectively, evidently lay within $\sim$ 5.0 Å and $\sim$ 5.5 Å during the all 2-ns dynamical changes. Thus, all hydrogen bond networks in the ATP-res20 for the binding site vary between the CDK2 and CDC28 structures.

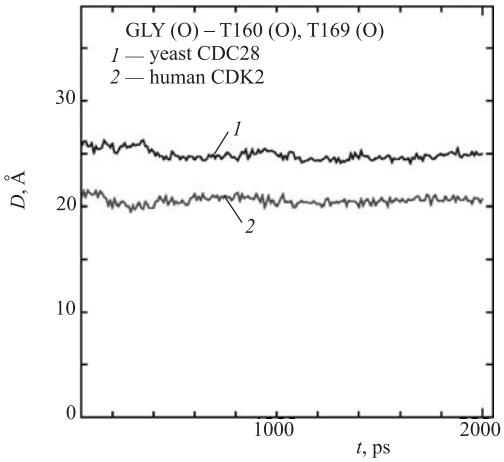


Fig. 24. The time dependences of the T160-res16, T169-res16 distances are shown for the CDK2 and CDC28, respectively, in accordance to the «activation triangle»

The amino acid residues around phosphorylated regulatory site. The CDK2, CDC28/ATP' dynamical peculiarities in the neighbor of phosphorylation site (T160 in CDK2, T169 in CDC28) were analyzed in detail. They are showed by snapshots and animation movies at all the amino acid positions in the T-loop. From the «activation triangle» described above, the T160 (T169)-res16 distances for the CDK2-G16/ATP and CDC28-G20/ATP were estimated. The T160-res20 distance in the CDK2-G20/ATP structure is significantly larger than T169-res16 in the wild-type CDC28-G16/ATP one (Fig. 24).

8. DISCUSSION

Only sequences and structures are among the several different types of data required in the practice of modern molecular biology. The primary nucleotide sequence database is the trio GenBank/EMBL/DDBJ. The primary proteins sequence database is the duo SWISS-PROT/PIR. Derived or secondary databases

may be divided into specialized subcollections of data and collections of information gathered by analysis of the primary databases. Patterns, signatures or motifs of protein sequences are to be found in databases such as PROSITE, PRINTS, BLOCKS, Pfam, etc., where they are stored as alignments, regular expressions, the hidden Markov models, consensus sequences or profiles. The primary structure database is the PDB. This contains all the experimentally determined structures of biological macromolecules, i. e., proteins, nucleic acids, and their complexes. Derived databases of protein motifs and patterns include SCOP and CATH. These databases cluster protein structures together with an increasingly distant hierarchy of structural similarity. Such databases are immensely useful in identifying the function of a newly sequenced protein.

The Neeedleman–Wunsch algorithm is a global alignment method, while the Smith–Waterman method is a local alignment one. BLAST is most frequently used over the Internet on the BLAST server (http://www.ncbi.nlm.nih.gov/BLAST/). CLUSTAL is a popular program for multiple sequence alignment that uses an extensively modified version of the Feng–Doolittle algorithm. A matrix of values that is used to score residue replacements or substitutions is called a substitution matrix. The two most popular statistically derived matrices are the PAM and BLOSUM matrices. PAM (Percentage Accepted Mutation) matrices are based on a Markovian model of evolutionary change in the sequences.

PHD is currently one of the most successful secondary structure prediction programs. It uses artificial neural networks to carry out the predictions. Methods to predict the tertiary structure of proteins may be divided into three broad categories — *homology modeling*, *threading* and *ab initio* methods. Homology modeling (MODELLER) is used when the unknown sequence, called the target, bears a sufficiently strong sequence similarity with another sequence, called the template, for which the structure is already known. Threading generalizes the technique of homology modeling, and aligns the unknown sequence to a likely structure, which can be built from families of structures with sequences similar to the target. An *ab initio* algorithm uses only the sequence of the protein, and the well-established laws and principles of physics and chemistry, to determine its three-dimensional structure.

To reach a deeper understanding of their function it is necessary to perform various geometrical calculations, such as bond lengths and angles, torsion angles, plane calculations, etc. The Ramachandran plot is a very good way of checking the geometry of the model and programs such as PROCHECK are available to carry out these tasks. Dynamic Cross Correlation Map (DCCM) computed as averages over the successive backbone of N, Cα and C atoms to give one entry per pair of amino acid residues. There is time scale implicit in C_{ij} as well. The intensity of the shading is proportional to the magnitude of the coefficient. The positive correlations are given in the upper triangle and the negative correlations are given in the lower triangle.

RMSD and RMSF values are considered as reliable indicators of variability when applied to very similar proteins, like alternative conformations of the same protein. Regions known to be of greater physiological importance including the structural conformations of protein kinases; the activation loops around ATPs; the amino acid residues inside the phosphorylated regulatory sites were described.

9. CONCLUSION

Information retrieval is important in various biomedical research fields. This work covers the theoretical background and the state of the art and future trends in biomedical information retrieval. Techniques for literature searches, genomic information retrieval and database searches are discussed. Literature search techniques cover name entity extraction, document indexing, document clustering and event extraction. Genomic information retrieval techniques are based on sequence alignment algorithms. This paper also briefly describes widely used biological databases and discusses the issues related to the information retrieval from these databases. Information retrieval technology has been used to gather information from biological sequence data, as well as from functional and structural descriptions of biomaterials. To handle the complex nature of the biological data, intelligent data analysis approaches such as sequence alignment, document clustering, and terminology systems, are used to facilitate the retrieval of semantically related information that would not be retrieved through keyword-based searches. Current information retrieval techniques are enabling the retrieval of information from digital libraries. Advances in computational biology and information retrieval are enabling the prediction of homologous gene or proteins whose function can be similar to the input query sequence, and then attempt to determine the function of this sequence based on the annotation of the homologous sequences and molecular dynamics calculations. Detailed analysis of the data obtained from structure prediction methods and molecular dynamics calculations confirms high degree of similarity between yeast protein kinase CDC28 and human kinase CDK2. Through this InSilico approach one can understand the conformation behavior [91, 92] between the important conserved regions including the G- and T-loops of kinases, ATP–Mg^{2+} ion complex and substrate component in correlation with the physiological properties between these structures.

REFERENCES

1. *Crick F.* Central Dogma of Molecular Biology // Nature. 1970. V. 227. P. 561–563.
2. *Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J.* Basic Local Alignment Search Tool // J. Mol. Biol. 1990. V. 215. P. 403–10.

3. *Christie K. R.* Saccharomyces Genome Database (SGD) Provides Tools to Identify and Analyze Sequences from *Saccharomyces cerevisiae* and Related Sequences from Other Organisms // Nucleic Acids Res. (Database issue). 2004. V. 32. D311–D314.

4. *Kanehisa M.* KEGG: Kyoto Encyclopedia of Genes and Genomes // Nucleic Acids Res. 2000. V. 28(1). P. 27–30.

5. *Huerta A. M.* RegulonDB: A Database on Transcriptional Regulation in *Escherichia coli* // Nucleic Acids Res. 1997. P. 55–60.

6. *Hamosh A.* Online Mendelian Inheritance in Man (OMIM), a Knowledge Base of Human Genes and Genetic Disorders // Nucleic Acids Res. 2002. V. 30, No. 1. P. 52–55.

7. *Sherry S. T., Ward M. H., Kholodov M., Baker J., Phan L., Smigielski E. M., Sirotkin K.* dbSNP: the NCBI Database of Genetic Variation // Nucleic Acids Res. 2001. V. 29(1). P. 308–311.

8. *Kawabata T.* The Protein Mutant Database // Nucleic Acids Res. 1998. V. 27. P. 355–357.

9. *Pruitt K. D.* NCBI Reference Sequence (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins // Nucleic Acids Res. 2005. V. 33. Database issue D501–D504.

10. *Kanz C.* The EMBL Nucleotide Sequence Database // Nucleic Acids Res. 2005. V. 33. Database Issue D29–D33.

11. *Miyazaki S.* DDBJ in the Stream of Various Biological Data // Nucleic Acids Res. 2004. V. 32. Database issue D31–D34.

12. *Wu C. H. et al.* The Protein Information Resource // Nucleic Acids Res. 2003. V. 31(1). P. 345–7.

13. *Bairoch A.* The SWISS-PROT Protein Sequence Database and Its Supplement TrEMBL in 2000 // Nucleic Acids Res. 2000. V. 28, No. 1. P. 45–48.

14. The FlyBase Consortium. The FlyBase Database of the Drosophila Genome Projects and Community Literature // Nucleic Acids Res. 2002. V. 30(1). P. 106–108.

15. *Perier R. C.* The Eukaryotic Promoter Database (EPD): Recent Developments // Nucleic Acids Res. 1998. P. 307–309.

16. *Hulo N.* The PROSITE Database // Nucleic Acids Res. 2006. V. 34. Database issue D227–D230.

17. *Henikoff J. G.* Blocks. NAR Molecular Biology Database Collection (Entry number 203). 2006.

18. *Finn R. D.* Pfam: Clans, Web Tools and Services // Nucleic Acids Res. 2006. V. 34. Database issue D247–D251.

19. *Kouranov A.* The RCSB PDB Information Portal for Structural Genomics // Nucleic Acids Res. 2006. V. 34. Database issue D302–D305.

20. *Murzin A. G.* SCOP — Structural Classification of Proteins. NAR Molecular Biology Database Collection (Entry number 282). 2006.

21. *Pearl F. M.* CATH. NAR Molecular Biology Database Collection (Entry number 258). 2006.

22. *Mullan L.* Pairwise Sequence Alignment — It's All About Us // Bioinform. 2006. V. 7(1). P. 113–115.

23. *Altschul S. F. et al.* Gapped BLAST and PSI-BLAST. A New Generation of Protein Database Search Programs // Nucleic Acids Res. 1997. V. 25(17). P. 3389–402.

24. *Zheng Zhang.* Protein Sequence Similarity Searches Using Patterns as Seeds // Nucleic Acids Res. 1998. V. 26, No. 17. P. 3986–3990.

25. *Oliver T.* Using Reconfigurable Hardware to Accelerate Multiple Sequence Alignment with ClustalW // Bioinform. 2005. V. 21(16). P. 3431–3432.

26. *Feng D.-F., Doolittle R. F.* Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees // J. Mol. Evol. 1987. V. 25. P. 351–360.

27. *Thompson J. D., Higgins D. G., Gibson T. J.* CLUSTALW: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position Specific Gap Penalties, and Weight Matrix Choice // Nucleic Acids Res. 1994. V. 22. P. 4673–4680.

28. *Henikoff S., Henikoff J. G.* Amino Acid Substitution Matrices from Protein Blocks // Proc. of the National Academy of Sciences. 1992. V. 89. P. 10915–10919.

29. *Hansen L. K., Salamon P.* Neural Network Ensembles // IEEE Trans. Pattern Anal. Machine Intel. 1990. V. 12. P. 993–1001.

30. *Rost B.* PHD: Predicting One-Dimensional Protein Structure by Profile Based Neural Networks // Meth. Enzymol. 1996. V. 266. P. 525–539.

31. *Lovell S. C. et al.* Structure Validation by Cα Geometry: Φ, Ψ and β Deviation // Proteins. 2003. V. 50. P. 437–450.

32. *Panchenko A., Marchler-Bauer A., Bryant S. H.* Threading with Explicit Models for Evolutionary Conservation of Structure and Sequence // Proteins. Suppl. 1999. V. 3. P. 133–140.

33. *Ortiz A. R. et al.* Ab Initio Folding of Proteins Using Restraints Derived from Evolutionary Information // Proteins. Suppl. 1999. V. 3. P. 177–185.

34. *Allen M. P., Tildesley D. J.* Computer Simulation of Liquids. N. Y.: Oxford University Press, 1987.

35. *Haile J. M.* Molecular Dynamics Simulations: Elementary Methods. N. Y.: Wiley, 1992.

36. Computer Simulations of Biomolecular Systems: Theoretical and Experimental Applications. V. 1 / Ed. W. Van Gunsteren, P. Weiner. Leiden: ESCOM, 1989; V. 2, V. 3 / Ed. W. Van Gunsteren, P. Weiner, A. T. Wilkinson. Leiden: ESCOM, 1993; 1996. [Good series on biomolecular simulations, covering both algorithms and applications].

37. *Gould H., Tobochnik J.* An Introduction to Computer Simulation Methods: Applications to Physical Systems. Parts 1 and 2. Addison–Wesley, 1988.

38. *Frenkel D., Smit B.* Understanding Molecular Simulations. From Algorithms to Applications. San Diego: Academic Press, 1996.

39. *Brooks C. L., Karplus M., Pettitt B. M.* A Theoretical Perspective of Dynamics, Structure and Thermodynamics. N. Y.: Wiley Interscience, 1988.

40. *Oliver S. G.* From DNA Sequence to Biological Function // Nature. 1996. V. 379. P. 597–600.

41. *Koonin E. V., Mushegian A. R.* Complete Genome Sequences of Cellular Life Forms: Glimpses of Theoretical Evolutionary Genomics // Curr. Opin. Gen. Dev. 1996. V. 6. P. 757–762.

42. *Dujon B.* The Yeast Genome Project: What Did We Learn? // Trends Genet. 1996. V. 12. P. 263–270.

43. *Orengo C. A., Jones D. T., Thornton J. M.* Protein Domain Superfolds and Superfamilies // Nature. 1994. V. 372. P. 631–634.

44. *Sanchez R., Sali A.* Large-Scale Protein Structure Modeling of the *Saccharomyces Cerevisiae* Genome // Proc. Nat. Acad. Sci. USA. 1998. V. 95. P. 13597–13602.

45. *Tramontano A., Leplae R., Morea V.* Analysis and Assessment of Comparative Modeling Predictions in CASP4 // Proteins. 2001. V. 45. P. 22–38.

46. *Marti-Renom M. A. et al.* Comparative Protein Structure Modeling of Genes and Genomes // Ann. Rev. Biophys. Biomol. Struct. 2000. V. 29. P. 291–325.

47. *Guex N., Diemand A., Peitsch M. C.* Protein Modelling for All // Trends Biochem. Sci. 1999. V. 24. P. 364–367.

48. *Fischer D., Eisenberg D.* Assigning Folds to the Proteins Encoded by the Genome of *Mycoplasma genitalium* // Proc. Nat. Acad. Sci. USA. 1997. V. 94. P. 11929–11934.

49. *Sanchez R., Sali A.* Assigning Folds to the Proteins Encoded by the Genome of *Mycoplasma genitalium* // Proc. Nat. Acad. Sci. USA. 1998. V. 95. P. 13597–13602.

50. *Rychlewski L., Zhang B., Godzik A.* Fold and Function Predictions for *Mycoplasma genitalium* Proteins // Fold. Des. 1998. V. 3. P. 229–238.

51. *Huynen M. et al.* Homology-Based Fold Predictions for *Mycoplasma genitalium* Proteins // J. Mol. Biol. 1998. V. 280. P. 323–326.

52. *Grandori R.* Systematic Fold Recognition Analysis of the Sequences Encoded by the Genome of *Mycoplasma pneumoniae* // Protein Eng. 1998. V. 11. P. 1129–1135.

53. *Teichmann S. A., Park J., Chothia C.* Structural Assignments to the *Mycoplasma genitalium* Proteins Show Extensive Gene Duplications and Domain Rearrangements // Proc. Nat. Acad. Sci. USA. 1998. V. 22. P. 14658–14663.

54. *Jones D. T.* GenTHREADER: An Efficient and Reliable Protein Fold Recognition Method for Genomic Sequences // J. Mol. Biol. 1999. V. 287. P. 797–815.

55. *Peitsch M. C., Jongeneel C. V.* A 3D Model for the CD40 Ligand Predicts That It Is a Compact Trimer Similar to the Tumor Necrosis Factors // Intern. Immunol. 1993. V. 5. P. 233–238.

56. *Peitsch M. C.* Protein Modelling by E-mail // BioTechnology. 1995. V. 13. P. 658–660.

57. *Peitsch M. C.* ProMod and Swiss-Model: Internet-Based Tools for Automated Comparative Protein Modelling // Biochem. Soc. Trans. 1996. V. 24. P. 274–279.

58. *Sali A., Blundell T. L.* Comparative Protein Modelling by Satisfaction of Spatial Restraints // J. Mol. Biol. 1993. V. 234. P. 779–815.

59. *MacKerell A. D. J. et al.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins // J. Phys. Chem. B. 1998. V. 102. P. 3586–3616.

60. *Sali A., Overington J. P.* Derivation of Rules for Comparative Protein Modeling from a Database of Protein Structure Alignments // Protein Sci. 1994. V. 3. P. 1582–1596.

61. *Marti-Renom M. A., Ilyin V. A., Sali A.* DBAli: A Database of Protein Structure Alignments // Bioinformatics. 2001. V. 17. P. 746–747.

62. *Sali A., Blundell T. L.* Definition of General Topological Equivalence in Protein Structures. A Procedure Involving Comparison of Properties and Relationships through Simulated Annealing and Dynamic Programming // J. Mol. Biol. 1990 V. 212. P. 403–428.

63. *Altschul S. F. et al.* Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs // Nucleic Acids Res. 1997. V. 25. P. 3389–3402.

64. *Sadreyev R., Grishin N.* COMPASS: A Tool for Comparison of Multiple Protein Alignments with Assessment of Statistical Significance // J. Mol. Biol., 2003. V. 326. P. 317–336.

65. *Sanchez R., Sali A.* Large-Scale Protein Structure Modeling of the *Saccharomyces cerevisiae* Genome // Proteins. 1997. V. 1. P. 50–58.

66. *Melo F., Sanchez R., Sali A.* Statistical Potentials for Fold Assessment // Protein Sci. 2002. V. 11. P. 430–448.

67. *MacKerell A. D.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins // J. Phys. Chem. B. 1998. V. 102. P. 3586–3616.

68. *Braun W.* Calculation of Protein Conformations by Proton–Proton Distance Constraints: A New Efficient Algorithm // J. Mol. Biol. 1985. V. 186. P. 611–626.

69. *Sali A.* Modelling Mutations and Homologous Proteins // Curr. Opin. Biotech. 1995. V. 6. P. 437–451.

70. *Vasquez M.* Modeling Side-Chain Conformation // Curr. Opin. Str. Biol. 1996. V. 6. P. 217–221.

71. *Thornton J. M.* Disulphide Bridges in Globular Proteins // J. Mol. Biol. 1981. V. 151. P. 261–287.

72. *Jung S. H., Pastan I., Lee B.* Design of Interchain Disulfide Bonds in the Framework Region of the Fv Fragment of the Monoclonal Antibody B3 // Proteins. 1994. V. 19. P. 35–47.

73. *Sali A., Overington J. P.* Derivation of Rules for Comparative Protein Modeling from a Database of Protein Structure Alignments // Protein Sci. 1994. V. 3. P. 1582–1596.

74. *Goffeau A. et al.* Life with 6000 Genes // Science. 1996. V. 274. P. 563–567.

75. *Bratley P., Fox B. L., Schrage L. E.* A Guide to Simulation. N. Y.: Springer-Verlag, 1987.

76. *Berman H. M. et al.* The Protein Data Bank // Nucleic Acids Res. 2000. V. 28. P. 235–242.

77. *Shindyalov I. N., Bourne P. E.* Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path // Protein Eng. 1998. V. 11. P. 739–747.

78. *Case D. A. et al.* AMBER 8.0. University of California, 2003.

79. *Okimoto N. et al.* // Chem. Bio. Inform. J. 2003. V. 3(1). P. 1–11;
Narumi T. et al. 2000 MDM Version of AMBER.

80. *Cornell W. D. et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids and Organic Molecules // J. Am. Chem. Soc. 1995. V. 117. P. 5179–5197.

81. *Jorgensen W. L., Chandrasekhar J., Madura J. D.* Comparison of Simple Potential Functions for Simulating Liquids // J. Chem. Phys. 1983. V. 79(2). P. 926–935.

82. *Berendsen H. J. C. et al.* Molecular Dynamics with Coupling to External Bath // J. Chem. Phys. 1984. V. 81(8). P. 3684–3690.

83. *Ryckaert J. P., Ciccotti G., Berendsen H. J. C.* Numerical Integration of the Cartesian Equations of a System with Constraints // J. Comput. Phys. 1997. V. 23. P. 327–341.

84. *Kholmurodov Kh.* Molecular Dynamics Simulations of Rhodopsin and Prion Proteins: The Effect of Disease-Related Amino Acid Mutations on the Structural Conformations // Particles and Nuclei. 2005. V. 36, No. 2. P. 1–16;
Kholmurodov Kh., Hirano Y., Ebisuzaki E. MD Simulations on the Influence of Disease-Related Amino Acid Mutations into Human Prior Protein // Chem. Bio. Inform. J. 2003. V. 3, No. 2. P. 86–95.

85. *Kholmurodov Kh. T. et al.* Methods of Molecular Dynamics for Simulation of Physical and Biological Processes // Particles and Nuclei. 2003. V. 34(2). P. 474–501.

86. *Sayle R. A., Milner-White E. J.* RasMol: Biomolecular Graphics for All // Trends in Biochem. Sci. 1995. V. 20. P. 374–376.

87. *Koradi R., Billeter M., Wuthrich K.* MOLMOL: A Program for Display and Analysis of Macromolecular Structure // J. Mol. Graphics. 1996. V. 4. P. 51–55.

88. *Mendenhall M. D., Hodge A. E.* Regulation of CDC28 Cyclin-Dependent Protein Kinase Activity During the Cell Cycle of the Yeast // Microbiol. Mol. Biol. Rev. 1998. V. 62. P. 1191–1243.

89. *De Bondt H. L. et al.* Crystal Structure of Cyclin-Dependent Kinase 2 // Nature. 1993. V. 363. P. 595–602.

90. *Jeffrey P. D. et al.* Mechanism of CDK Activation Revealed by the Structure of a Cyclin A – CDK2 Complex // Nature. 1995. V. 373. P. 313–320.

91. *Koltovaya N. A., Guerasimova A. S., Kretov D. A., Kholmurodov Kh. T.* Sequencing Analysis of Mutant Allele *cdc28-srm* of Protein Kinase CDC28 and Molecular Dynamics Study of Glycine-Rich Loop in Wild Type and Mutant Allele G16S of CDK2 as Model. ISBN: 1-59454-912-5. Nova Science Publishers, 2006.

92. *Kretov D. A., Kholmurodov Kh. T., Koltovaya N. A.* MD Simulations on Human Kinase Protein: the Influence of a Conserved Glycine by Serine Substitution in G-Loop of a CDK2 Active Complex // Mendeleev Commun. 2006. V. 4 (in press).

Received on July 10, 2006.

Корректор *Т. Е. Попеко*

Подписано в печать 15.11.2006.
Формат 60 × 90/16. Бумага офсетная. Печать офсетная.
Усл. печ. л. 2,95. Уч.-изд. л. 4,15. Тираж 220 экз. Заказ № 55559.

Издательский отдел Объединенного института ядерных исследований
141980, г. Дубна, Московская обл., ул. Жолио-Кюри, 6.
E-mail: publish@jinr.ru
www.jinr.ru/publish/