Homology Modelling and Molecular Dynamics of Cyclin-Dependent Protein Kinases

Robert A. Selwyne[a,b], **Kholmirzo T. Kholmurodov**[*a,c], **Natalia A. Koltovaya**[a,c]

[a]Laboratory of Radiation Biology, Joint Institute for Nuclear Research, Dubna, 141980
Moscow region, Russia

[b]Department of Botany, Bharathiar University, Coimbatore 641046, Tamil Nadu, India

[c]International University 'Dubna', Dubna 141980,
Moscow region, Russia
*mirzo@jinr.ru

Abstract

This chapter gives an overview of bioinformatics techniques of importance in protein analysis. These include database searches, sequence comparisons and structural predictions. Useful links 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. Further more the chapter describes widespread methods for sequence comparisons, multiple sequence alignments and secondary structure predictions. We have performed the homology modelling and molecular

IT FOR REAL WORLD PROBLEMS, (Universities Press Series in Systems, Models, Informatics and Control), Sree Hari Rao (Ed.), pp. 1—72 ©Universities Press (India) Private Limited 2011, ISBN:978 81 7371 734 5

dynamics (MD) simulation analysis of structural conformation properties of yeast and human cyclin-dependent kinases CDC28 and CDK2. Based on the homology modelling a structure of yeast CDC28 is predicted using a lattice crystal structure of human CDK2 using MODELLER software. Further MD simulations run using the AMBER 8.0 package. For 2 nanoseconds we investigated the conformational behavior of crystal lattice for both yeast CDC28 and human CDK2/cyclin A/ATP-Mg2+/substrate at a physiological temperature T=300K. Based on the MD simulation results we discuss the molecular mechanism of regulation of phosphorylation and the structural changes of kinases.

1 Introduction

The word 'bioinformatics' refers to the application of information technology (IT) to molecular biology. It plays a major role in areas of study such as computational (molecular) biology, biocomputing or biocomputation, computational genomics, *in silico* biology, and computational proteomics. In the present day scenario, any endeavor one undertakes is made easy by the application of information technology. We may consider bioinformatics to comprise the study of *information pathways* in living organisms. DNA and protein sequences form the major components of information pathways in molecular biology. These sequences are nothing but a set of four alphabets for DNA and twenty for proteins. All the tools and techniques that have been developed to analyse these sequences yield information regarding physiological mechanisms through digital information processing technology. Thus, bioinformatics is intimately connected with theoretical computer science, especially with topics such as natural language processing, machine learning, computational linguistics and digital pattern recognition. Ideas and methods have been taken from these sciences and effectively incorporated 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. Darwinian principles of evolution, together with related theories on prebiotic systems, showed it possible to think of Life as consisting of complex interactions among inert chemicals. 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. All three are polymeric molecules. In the case of DNA and RNA, the monomeric units are nucleotides, while proteins are built from amino acids. The difference between RNA and DNA is the presence, in RNA, of an extra oxygen atom on the sugar ring. All three molecules possess a backbone of atoms that repeats monotonously over the entire length. 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 that 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 two strands of DNA are arranged with the 5' end of one strand on the same side as the 3' end of the other, and vice versa. The negatively charged phosphate−sugar backbones of the two strands are on the outside of the double helix, while the planar, nitrogenous and hydrophobic bases are on the 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 [14] . Genetic information is represented by the sequence of bases in the DNA chain. 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. Therefore, a sequence of DNA bases is a genetic message, or a portion of one. Thus, there are sequences that represent proteins and each such sequence is a gene. The entire set of genes and control elements is called the genome of the organism.

RNA is similar to DNA but contains the base U in place of T. RNA has three different functions in the process of information storage and transfer. Its three-dimensional structure may be compact and folded, like a protein, or it may be extended, without any particular structure, depending on the role it plays.

A protein chain is also represented by a string of symbols, this time chosen from an alphabet of 20 letters, representing the 20 different amino acids. The amino acid structures are shown in Figure 1 and their codons are listed in Table 1.

Table 1 *The genetic code. The codons are written 5' to 3' as seen in the mRNA sequence (and not in the DNA sequence). Hence U is used, not T.*

AAA	*Lys*	UAA	*STOP*	GAA	*Glu*	CAA	*Gln*
AAU	*Asn*	UAU	*Tyr*	GAU	*Asp*	CAU	*His*
AAG	*Lys*	UAG	*STOP*	GAG	*Glu*	CAG	*Gln*
AAC	*Asn*	UAC	*Tyr*	GAC	*Asp*	CAC	*His*
AUA	*Ile*	UUA	*Leu*	GUA	*Val*	CUA	*Leu*
AUU	*Ile*	UUU	*Phe*	GUU	*Val*	CUU	*Leu*
AUG	*Met*	UUG	*Leu*	GUG	*Val*	CUG	*Leu*
AUC	*Ile*	UUC	*Phe*	GUC	*Val*	CUC	*Leu*
AGA	*Arg*	UGA	*STOP*	GGA	*Gly*	CGA	*Arg*
AGU	*Ser*	UGU	*Cys*	GGU	*Gly*	CGU	*Arg*
AGG	*Arg*	UGG	*Trp*	GGG	*Gly*	CGG	*Arg*
AGC	*Ser*	UGC	*Cys*	GGC	*Gly*	CGC	*Arg*
ACA	*Thr*	UCA	*Ser*	GCA	*Ala*	CCA	*Pro*
ACU	*Thr*	UCU	*Ser*	GCU	*Ala*	CCU	*Pro*
ACG	*Thr*	UCG	*Ser*	GCG	*Ala*	CCG	*Pro*
ACC	*Thr*	UCC	*Ser*	GCC	*Ala*	CCC	*Pro*

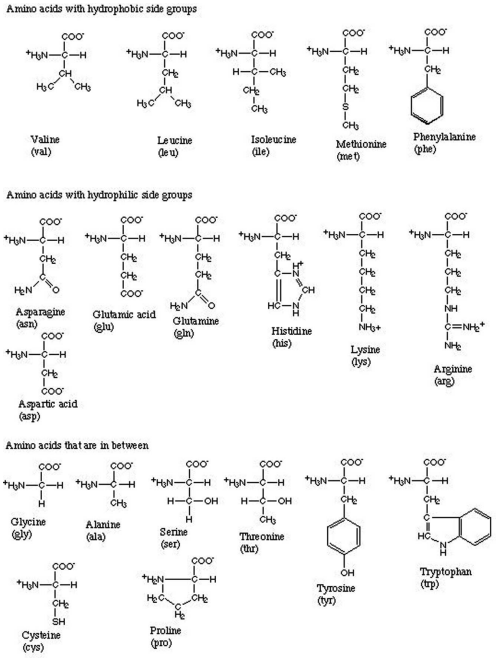

Figure 1 *Twenty different amino acids*

When a cell divides, it passes about half its complement of molecules and structures to each of the two daughter cells. Each daughter cell receives from its parent an identical set of genes. This set of genes is present in the DNA that the cell receives. The genes contain the instructions required by the cell to build new molecules and structures, as and when required in its life cycle. The principal molecular process involved in this pathway of information transfer is the replication of the DNA. Figure 2 shows the transcription of DNA into mRNA and further translation of mRNA into a protein. The correct three-dimensional folding of the protein chain depends crucially on the amino acid sequence. Different sequences lead to different structures which in turn give rise to different functions. The ultimate step in the chain of expression is the synthesis of a functional protein.

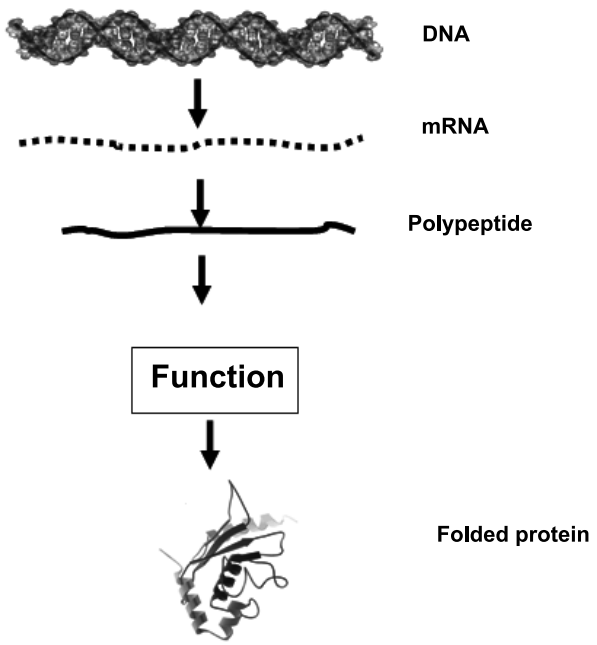

Figure 2 *Central dogma of molecular biology*

1.2 Introduction to 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 the elemental tasks.

1. Searching for patterns within a sequence
2. Obtaining statistical information about a sequence
3. Searching for similarities between two sequences, or performing sequence alignment for a pair of sequences
4. Searching for similarities among many sequences, or performing multiple sequence alignment
5. Constructing phylogenetic trees based on sequences
6. Predicting and analysing the secondary structures on basis of the sequence
7. Predicting and analysing 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. Once the sequence is obtained and validated, coding regions or genes need to be identified along with various transcription signals such as promoters, enhancers, etc. Such gene annotation includes the identification of exons and introns.

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. Sequence comparison and alignment programs such as BLAST and FASTA are among the most frequently used for this purpose [2].

The fourth task in the list, multiple sequence comparison and alignment, is also very important for functional annotation. Further, it usually precedes the fifth task, viz. phylogenetic analysis, which explores relationships between the sequences and therefore between the parent organisms. The analysis helps to identify possible evolutionary relationships among the organisms.

The last two tasks mentioned in the list relate to the structures of the molecules. To understand the function of any physical system completely, in this case the biologically active molecule, it is necessary to know its structure. In the case of proteins, the single polypeptide chain takes up different three-dimensional arrangements such as alpha helices and beta strands, according to its sequence. The next or tertiary level of structure refers to the arrangement of these secondary structures in space. The final three-dimensional structure of the protein is its fully functional form. Within the ambit of structure analyses, we also include molecular modelling, docking, *in silico* mutation analyses and techniques used for drug discovery and protein engineering.

1.2.2 Applications of bioinformatics

Bioinformatics-based techniques are used for the following purposes:

1. To obtain sequences and structures of biological molecules, for example, shotgun sequencing and protein structure prediction
2. To analyse gene expression profiles, especially those based on DNA chip technology
3. To obtain medically important data such as single nucleotide polymorphisms or restriction fragment length polymorphisms (DNA fingerprinting)
4. In the computer simulation of individual metabolic processes, as well as the more ambitious simulation of a whole cell or even a whole organism
5. To aid in processes involved in the development of new drugs, such as lead discovery, lead optimisation through molecular modelling, design and analysis of the laboratory and clinical trials
6. In the construction, maintenance, use and analyses of databases of all types of biological data, even those not directly related to molecular biology such as environmental data, biodiversity data, toxicology, etc.
7. To analyse the three-dimensional structure of biological molecules (structural bioinformatics)

2 Molecular Biology Databases

Nucleic acid and protein sequences and structures form the core raw material of bioinformatics as defined in this book. Due to the central role played by the genome in the life cycle of the cell, it is essential to know and understand the sequence of the DNA it contains. The sequence represents a blueprint for the life of the organism. This view of the genome has lead to the setting up of large sequencing projects. Studies are underway on the genome of other organisms ranging from bacteria to plants to other mammals. Sequence and structure databases may be classified into two types, viz. primary or raw databases and secondary or derived databases. The former consist of just the sequences and structures, and the value addition, if any, is in the annotation and arrangement. Derived databases are usually subsets of the primary databases, and are constituted either to serve a specific subtopic in biology, such as cancer, or have some added value, for example multiple alignment of a set of sequences or protein

domain structures grouped according to structural similarity. Secondary databases are greatly dependent on the algorithms used to make the arrangements as well as on the curator of the database. Therefore it is not surprising that often there is more than one database dealing with the same biological information, for instance, the relationship between protein sequence and structure, but yet have different points of view and different results.

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 (National Centre for Biotechnology Information) portal over the Internet. EMBL (European Molecular Biology Laboratory) is in UK, at the European Bioinformatics Institute, and DDBJ (DNA DataBank of Japan) is in Japan. Historically the three started out as separate databases, with independent means of collecting the data, and with different formats for arranging it. Now the three together form the International Nucleotide Sequence Database Collaboration. They have uniform (but not identical) formats and exchange data on a daily basis, with the result that it is sufficient to connect to any one of them in order to access the sequences present in all three. Here we will describe only one of the data formats, GenBank, in detail, assuming that its differences with the other two are insignificant. In particular, we will describe the implementation of this database at the NCBI site.

The access to GenBank, as to all databases at NCBI, is through the Entrez search program. This front-end search interface allows a great variety of search options. More importantly, it enables searching more than one database at the same time, interpreting the search parameters in an appropriate way for each database. Thus if the user specifies the name of a protein and asks for a 'nucleotide' search, Entrez will look for the corresponding gene sequence. The keyword 'nucleotide' identifies GenBank, as well as other nucleic acid sequence databases. The search therefore occurs across different database formats. This means that the formats in which the data are stored in the databases are necessarily transparent to the user, who would otherwise have to change her query to suit each database. Nevertheless, it would be instructive to study the data types and identifiers used in GenBank. This is especially so for the following two reasons. One, the result of a search would usually be a sequence record arranged to display all the database information, and it is necessary to know what the structure of this record is. And two, users often prefer to download the entire database on to their own local computer systems. The data is made available at NCBI (and the other

sites) for download as flat files[1], and again it is necessary to know just what data is obtained through such a download. Figure 3*a* shows one

```
LOCUS       NM_007272      894 bp      mRNA      lineer      PRI 17-DEC-2004
DEFINITION  Homo sapiens chymotrypsin C (caldecrin) (CTRC), mRNA.
ACCESSION   NM_007272
VERSION     NM_007272.1 GI:11321627
KEYWORDS
SOURCE      Homo sapiens (human)
  ORGANISH  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 894)
  AUTHORS   Tomomura,A., Akiyama,M., Itoh,H., Yoshino,I., Tomomura,M.,
            Nishii,Y., Noikura,T. and Saheki,T.
  TITLE     Molecular cloning and expression of human caldecrin
  JOURNAL   FEBS Lett. 386 (1), 26-28 (1996)
  MEDLINE   96221265
  PUBMED    8635596
COMMENT     PROVISIONAL REFSEQ: This record has yet been subject to final
            NABI review. The reference sequence was derived from S82198.1.
FEATURES             Location/Qualifiers
     Source          1..894
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="1"
                     /map="1p36.21"
     gene            1..894
                     /gene="CTRC"
                     /note="synonym: CLCR"
                     /db_ xref="GeneID:11330"
                     /db_ xref="LocusID:11330"
                     /db_ xref="MIM"601405"
     CDS             1..807
                     /gene="CTRC"
                     /note="caldecrin (serum calcium decreasing factor,
                     Elastase IV); 0TTHUMP00000044526;
                     Go_function: trypsin activity [goid 0004295] [evidence
                     IEA];
```

Figure 3(a) *The output of a search of the nucleotide database in NCBI using the keyword* '*chymotrypsin*'.

[1] Flat files are collections of data and other information arranged so that any general text-processing program can read them. They consist only of the characters found on any computer keyboard (i.e., the ASCII set of characters).

result of a search for the keyword 'chymotrypsin' through the 'nucleotide' database at NCBI. The format of the data is given as plain text, and contains a lot of information besides the gene sequence of chymotrypsin. All the data are grouped under titles, which not only describe the information under that heading, but also serve to define keyword fields that can be used in search programs. The word ACCESSION defines a field containing unique identification numbers. The sequence and other information may be retrieved from the database simply by searching for a given accession number. Taking these field names in order, we have, first of all, the word LOCUS. This is a GenBank title that names the sequence entry. Apart from the accession number, it also specifies the number of bases in the entry, the nucleic acid type, a codeword PRI that indicates that the sequence is from a primate, and the date on which the entry was made. PRI is one of 17 such words that are used to classify the data (Table 2). They were chosen in the early years of GenBank to fairly represent the data then available. Subsequently, other codes such as EST were added to list. However their utility is now not very great and most uses of the GenBank do not have anything to do with them. The next line of the file contains the definition of the entry, giving the name of the sequence. The unique accession number comes next, followed by a version number in case the entry has gone through more than one version. The next item is a list of specially defined keywords that are used to index the entries (Note that in this particular example, no keywords have been specified). This is followed by a set of SOURCE records, which describe the organism from which sequence was extracted. The complete scientific classification is given. This is followed by publication details. In the beginning, sequences were extracted from published literature, and painstakingly entered in the database. Each entry was therefore associated with a publication. Every REFERENCE field would thus contain the title of the article, the names of the authors, and the name, volume, page numbers and publication year of the journal in which the article appeared. With the vast explosion of sequence data, it is now no longer feasible to either print the sequences in the journals, nor to manually enter the sequences into the database. The entries are all electronic, straight from the laboratory to the database. The REFERENCE field now points to the publication containing the biological information obtained from the sequencing project.

Table 2 *GenBank keywords*

1.	PRI	primate sequences
2.	ROD	ordent sequences
3.	MAM	other mammalian sequences
4.	VRT	other vertebrate sequences
5.	INV	invertebrate sequences
6.	PLN	plant, fungal, and algal sequences

7.	BCT	bacterial sequences
8.	VRL	viral sequences
9.	PHG	bacteriophage sequences
10.	SYN	synthetic sequences
11.	UNA	unannotated sequences
12.	EST	EST sequences (expressed sequence tags)
13.	PAT	patent sequences
14.	STS	STS sequences (sequence tagged sites)
15.	GSS	patent sequences
16.	HTG	HTGS sequences (high throughput genomic sequences)
17.	HTC	HTC sequences (hgh throughput cDNA sequences)

The FEATURES table that follows is one of the most important pieces of information accompanying the sequence. It is an annotation of the sequence and describes whatever the contributors know about the sequence. The features (for nucleic acid sequences) include coding regions, exons, introns, promoters, alternate splice patterns, mutations, variations, and a translation into a protein sequence, if it codes for one. Each feature may be accompanied by a cross-reference to another database. After the features table, a single line gives the base count statistics for the sequence. Finally comes the sequence itself. The sequence is typed in the lowercase, and for ease of reading, each line is divided into six columns of ten bases each. A single number on the left numbers the bases. The above description does not cover all the fields used in GenBank, but only the more important ones. While this is one of the chief formats used to deliver the sequences to the user, a variety of other display and file formats are also available.

Finally, it is worth noting that the GenBank is in fact a relational database system, and the data are actually stored as tables, each column corresponding to a field descriptor, and each row corresponding to a sequence entry. In addition, extensive indices of keywords are also a part of the database, making it extremely fast to search through the mountains of data.

2.2 Primary protein sequence repositories

While the three 'mother' databases described above allow access to both nucleic acids and protein sequences, there are a few databases that contain solely protein sequences. The chief ones among them are the PIR-PSD (Protein Information Resource - Protein Sequence Database) at the NBRF (National Biomedical Research Foundation, USA), and the SWISS-PROT at the SBI (Swiss Biotechnology Institute), Switzerland.

The PIR-PSD is a collaborative endeavour between the PIR, the MIPS (Munich Information Centre for Protein Sequences, Germany) and the JIPID (Japan International Protein Information Database, Japan). This database grew out of the work of Margaret Dayhoff and her colleagues

and collaborators at the NBRF. Beginning in the 1960s, before the advent of computerised databases, Dayhoff collected published protein sequences, and on the basis of alignments and sequence comparisons, classified and annotated the collection painstakingly by hand, and made it available to the scientific community in the form of the Atlas of Protein Sequence and Structure, a set of printed volumes issued periodically. (Dayhoff also made two other important contributions to the then nascent field of bioinformatics. The first was her extensive reworking of evolutionary trees on the basis of protein sequences, in particular the sequences of Cytochrome C. The other was the development of the PAM matrices that indicated which amino acid substitution in the sequence was harmless to the function of the protein, and which disastrous. Both these contributions will be discussed at length later on in the book.) The Atlas of Protein Sequence and Structure grew into the PIR-PSD, and continues the former's tradition of high-quality annotation, except that the classification is now carried out by well-validated automatic procedures. Also the database is now fully computerised and uses the Oracle object relational DBMS. The PIR-PSD is now a comprehensive, non-redundant, expertly annotated, fully classified and extensively cross-referenced protein sequence database in the public domain. It is available at http://pir.georgetown.edu/pirwww. A unique characteristic of the PIR-PSD is its classification of protein sequences based on the superfamily concept. Sequences in PIR-PSD are also classified based on homology domains and sequence motifs. Homology domains may correspond to evolutionary building blocks, while sequence motifs represent functional sites or conserved regions. The classification approach allows a more complete understanding of sequence-function-structure relationships. Figure 3*b* shows the result of searching the PIR-PSD database using the keyword 'chymotrypsin'.

Protein Name	Chymotrypsin C (Caldecrin)
Taxonomy	**Danio rerio** (zebra danio) *NCBI Taxon ID:* 7955 *Lineage:* cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Actinopterygii; Actinopteri; Neopterygii; Teleostei; Elopocephala; Clupeocephala; Otocephala; Ostariophysi; Otophysi; Cypriniphysi; Cypriniformes; Cyprinoidea; Cyprinidae; Rasborinae; Danio
Source Organism	Danio rerio (*Taxon ID:* 7955)
Bibliography	View Bibliography information. Submit Bibliography. PubMed: PMID:12477932
Sequence Database	
Protein Sequence	MMKFVVLAVLVVGAYSCGLPTFPPIVTRVVGGVDVRPNSWPWQISLQYKSGSNWYHTCGG SLIDKQWVLTAAHCISSSRTYRVFLGKHSLSQEENGSVAIGAGKIIVHEAWNSFTIRNDI ALIKLETAVTIGDTITPACLPEAGYVLPHNAPCYVTGWGRLYTNGPLADILQQALLPVVD HATCSKSDWWGSQVTTSMVCAGGDGVVAGCDGDSGGPLNCAGSDGAWEVHGIVSFGSGLS CNYNKKPTVFTRVSAYSDWISKNMASY

Figure 3(b) *The output of a search of the PIR-PSD using the keyword 'chymotrypsin'.*

The other well-known and extensively used protein sequence database is SWISS-PROT (http://www.expasy.ch/sprot). Like the PIR-PSD, this curated protein sequence database also provides a high level of annotation. The data in each entry can considered separately as core data and annotation. The core data consists of the sequence entered in the common single letter amino acid code and the related references and bibliography. The taxonomy of the organism from which the sequence was obtained also forms part of this core information. 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, similarities to other proteins if any, and diseases that may be associated with deficiency of the protein or mutations. Any sequence conflicts or variants that may arise due to different authors publishing different sequences for the same protein, or due to mutations in different strains of an organism are also described as part of the annotations. This scheme also works to reduce redundancy as far as possible, with a given protein from a given organism having only one entry. Figure 4*a* gives what is called a 'NiceProt' view of the data, in which the data descriptions are given in easily readable full

```
        10         20         30         40         50         60
MLGITVLAAL LACASSCGVP SFPPNLSARV VGGEDARPHS WPWQISLQYL KNDTWRHTCG

        70         80         90        100        110        120
GTLIASNFVL TAAHCISNTR TYRVAVGKNN LEVEDEEGSL FVGVDTIHVH KRWNALLLRN

       130        140        150        160        170        180
DIALIKLAEH VELSDTIQVA CLPEKDSLLP KDYPCYVTGW GRLWTNGPIA DKLQQGLQPV
```

Figure 4(a) '*Nice-Prot*' *view of the search results of Swiss-Prot with the keyword* '*chymotrypsin*'. *The view has been extensively edited.*

format. In the actual database, however, the data are stored as shown in Figure 4*b*. This closely follows the format of entries in EMBL nucleotide sequence database. Each line begins with a two-letter code that specifies the information to be found in that line, as well as the format of that line. The first one is the ID line that identifies the entry. The subsequent line codes and the information each line contains are given in Table 3. The last of the two letter codes is SQ, which also begins the sequence. The

sequence lines do not have any two-letter code, but the first two columns are blank. The sequence is written in 6 groups of 10 amino acids each per line, i.e., a total of 60 amino acids per line. The entry is terminated by a double slash (//).

```
ID   CLCR_HUMAN          STANDARD;          PRT;    268  AA.
AC   Q99895;        000765;       Q9NUH5;
DT   16-OCT-2001 (Rel.  40,     Created)
DE   Caldecrin precursor (EC 3.4.21.2)  (Chymotrypsin C).
GN   Name=CTRC; Synonyms-CLCR;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE, AND VARIANT TRP-80.
RC   TISSUE=Pancreas;
RX   MEDLINE=96221265; PubMed=8635596 [NCBI, ExPASy, EBI, Israel, Japan];
RA   Tomomura A., Akiyama M., Itoh H., Yoshino I., Tomomura M., Nishii Y.,
RA   Noikura T., Saheki T.;
RT   "Molecular cloning and expression of human caldecrin.";
RL   FEBS Lett. 386:26-28 (1996).
CC   ---------------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation –
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institute as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/)
```

Figure 4(b) *Complete view of the search results of Swiss-Prot with the keyword* '*Chymotrypsin*'. *The view has been extensively edited.*

Both PIR-PSD and SWISS-PROT have software that enables the user to easily search through the database to obtain only the required information. SWISS-PROT has the SRS (Sequence Retrieval System) that searches also through the other relevant databases on the site, such as TrEMBL (Translated EMBL). In addition users may download the entire database for use and manipulation on a local computer system, though of course this would require very large resources. Firstly, SWISS-PROT does not include variants as different entries. Secondly, SWISS-PROT includes only validated entries of protein sequences. Putative protein sequences translated from the DNA or mRNA sequence normally go to TrEMBL, until they are verified.

TrEMBL is a computer-annotated protein sequence database that was released as a supplement to SWISS-PROT. It contains the translations of all coding sequences present in the EMBL Nucleotide Sequence Database, which have not been fully validated. Thus it may contain sequences of proteins that are never expressed and never actually identified in the organisms. These sequences are themselves classified in two ways. SP-TrEMBL sequences are given a SWISS-PROT accession number and are expected to be validated and entered into the main database. REM-TrEMBL entries, on the other hand, may or may not be eventually supported as individual protein sequences by experiments and other analyses. These do not have accession numbers. The format of the TrEMBL entries follows that of SWISS-PROT, except for the changes mentioned above.

Table 3 *Line codes in SWISS-PROT database*

Code	Expansion	Remarks
ID	Identification	Occurs at the beginning of the entry. Contains a unique name for the entry, plus information on the status of the entry. If it has been checked and conforms to SWISS-PROT standards, it is called STANDARD.
AC	Accession numbers	This is a stable way of identifying the entry. The name may change but not the AC. If the line has more than one number, it means that the entry was constituted by merging other entries.
DT	Date	There are three dates corresponding to the creation date of the entry and modification dates of the sequence and the annotations respectively
DE	Description	Lines that start with this identifier contain general description about the sequence.
GN	Gene name	The name of the gene (or genes) that codes for the protein
OS, OG, OC	Organism name, organelle, organism classification	The name and taxonomy of the organism, and information regarding the organelle containing the gene, e.g. mitochondria or chloroplast, etc.

RN, RP, RC, RX, RA, RT, RL	Reference number, position, comments, cross-reference, authors, title and location	Bibliographic reference to the sequence. This includes information (following the code RP) on the extent of work carried out by the authors.
CC	Comments	These are free text comments that provide any relevant information pertaining to the entry.
DR	Database cross-reference	This line gives cross-references to other databases where information regarding this entry is also found. As for example to structural information for the protein in the PDB.
KW	Keywords	This line gives a list of keywords that can be used in indexes. Search programs very often simply go through such indices to identify required information.
FT	Features table	These lines describe regions or sites of interest in the sequence, e.g. post-translational modifications, binding sites, enzyme active sites and local secondary structure.
SQ	Sequence header	This line indicates the beginning of the sequence data and gives a brief summary of its content.

The CluSTr (Clusters of SWISS-PROT + TrEMBL proteins at http://www.ebi.ac.uk/clustr) database offers an automatic classification of the entries in the SWISS-PROT and TrEMBL databases into groups of related proteins. The clustering is based on analysis of all pairwise comparisons between protein sequences. There are two steps in forming the clusters. First all proteins sequences are compared against all other sequences and matrix of similarity scores is constructed. For pairs of sequences, say A and B, which have a similarity score higher than a certain threshold, the next step is taken. This is to compare A with a set of randomly generated sequences that have the same length and amino acid composition as B. An average similarity score M for this comparison is calculated along with the standard deviation σ. The $Z(A, B)$ score may now be calculated as

$$Z(A, B) = (S(A, B) - M)/\sigma$$

where $S(A, B)$ is the similarity score between A and B. Similarly a score $Z(B, A)$ is calculated by comparing B with randomised sequences that have the same length and composition as A. The final Z score between A and B is taken as the minimum of these two. Based upon the Z score between pairs of sequences, a hierarchical scheme is used to cluster

together proteins at different levels of similarity. Note that the clustering is based purely on the sequence. The clustering has been carried out for a few sets of protein sequences, such as the mammalian sequences or the plant proteins, as well as for complete genomes, such as that of yeast.

2.3 Structure databases

Structure databases, like sequence databases, come in two varieties, primary and secondary. Strictly speaking, there is only one database that stores primary structural data of biological macromolecules, namely, the PDB (Protein Data Bank). In the context of this database, the term macromolecule stretches to cover three orders of magnitude of molecular weights from 1000 daltons to 1000 kilodaltons. Small (i.e. molecular weight approximately less than 1000 daltons) biological and organic molecules have their structures stored in another primary structure database, viz. the CSD (Cambridge Structural Database), which is also widely used in biological studies. This contains the three dimensional structures of drugs, inhibitors and fragments or monomers of the macromolecules. Apart from these there is no other primary structure database. In contrast, there are plenty of derived structural databases that store subsets of data extracted from the primary databases. However, most deal with protein structures, and there are very few secondary databases for nucleic acid structures.

2.3.1 The primary structure databases: PDB and CSD

In spite of its name, PDB archives the three-dimensional structures of not only proteins but also all biologically important molecules, such as nucleic acid fragments, RNA molecules, large peptides such as the antibiotic gramicidin, and complexes of proteins and nucleic acids. The database holds data derived from mainly three sources. Structures determined by X-ray crystallography form the large majority of the entries. In number this is followed by structures arrived at by NMR experiments. There are also several structures obtained by molecular modelling. The PDB was created at the Brookhaven National Laboratory in USA, and was known for a long time as the Brookhaven protein data bank. In recent times its curation and maintenance has been taken over by a consortium of three organisations called the Research Collaboratory for Structural Bioinformatics, comprising Rutgers University, the San Diego Supercomputer Centre and the National Institute of Standards and Technology, all in the USA. It is available as a set of CDs supplied free on request, or, more easily, over the Internet at http://www.rcsb.org.

Before we take a more detailed look at how the data is organised, let us consider how a three-dimensional molecular structure is stored in electronic form. While pictures and maps are more accessible to human understanding and also more appealing, they do not reflect the level of accuracy reached by the experimental structure determination techniques. As will be described in a later chapter in this book, the methods of X-ray crystallography and NMR have been developed to an extent where the precise three-dimensional position of every single atom in the molecule can be determined. Each atomic position is represented as a set of three coordinates referred to a right-handed orthogonal system of axes, whose origin, i.e. the coordinate set (0.0, 0.0, 0.0) is chosen arbitrarily, or according to experimental convenience (Figure 5). Such a system of axes is known as the Cartesian coordinate system. Each three-dimensional structure in the PDB is therefore represented as a table of coordinates, each coordinate triple corresponding to one atom in that molecule or molecular complex.

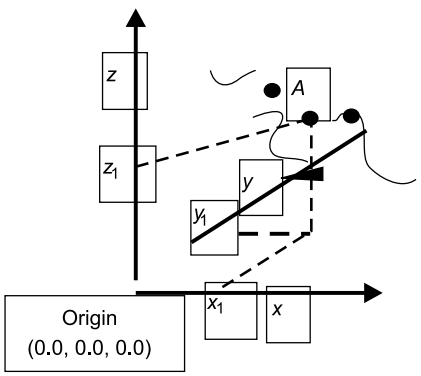

Figure 5 *The Cartesian coordinate system. The coordinates of 'atom' A are* (x_1, y_1, z_1).

In order to represent the entire molecule, there will be as many triples as there are atoms. This way of representing a three dimensional structure has several advantages over pictures. Firstly, and most importantly, pictures are only two-dimensional projections of three-dimensional objects, and as such are only a small portion of the information present in the structure. Secondly, it is always possible to generate any picture of the molecule from the coordinates, either manually, or, as is common now, automatically with computers. This is equivalent to extracting a subset of the total three-dimensional information. In fact software for viewing molecules is designed so that the three-dimensional coordinates are directly loaded into the computer system, so that any view of the molecule can be instantly displayed. Thirdly, coordinates allow the calculation of exact geometrical parameters such as bond lengths, bond angles, hydrogen bond distances and angles, and geometries of interaction between, for example, a protein

and its substrate. And fourthly, coordinates allow scientists to use the structure as a model that can be manipulated and changed by rotating or moving parts of the molecule with respect to the rest, to see what the effect on the function would be. These advantages are in addition to the fact that the experimental techniques cited above naturally present their results in the form of atomic coordinates.

The data in the PDB is organised as flat files, one for each structure, which means that each file contains one molecule, or one molecular complex. The files are named according to their PDB ID code, which consists of four alphanumeric characters, which may very approximately be an abbreviation of the name of the molecule. All files that have the same characters at positions two and three of their names are placed together in a single directory named with that pair of characters. Thus the directory named '00' contains the files Pdb100d, Pdb200d, Pdb200l, Pdb300d and Pdb400d. Totally there are as many files as there are structures in the database. Within each file are to be found the coordinates of the structure, apart from other related information. This information is formatted in the well-recognised technique of placing a line identifier at the beginning of each line to specify the data contained in that line. Figure 6 gives an example of a structure file.

The data format of a PDB file is well established and is widely recognised by almost all software designed to read and manipulate structural data. As mentioned above, and just as in some of the sequence databases (e.g. GenBank), every line in a PDB coordinate file begins with a word that identifies the type of information present in that line.

These keywords or identifiers may be grouped into many sections, viz. title, primary structure, heterogen, secondary structure, crystallographic and coordinate transformations, coordinates and connectivity. In the title section, the keywords point to information about the name, source, etc. of the compound, details of the authors, the references to published literature, and up to 1000 lines of general information about the structure, including all necessary experimental details required to interpret the structure correctly. The section on the primary structure contains keywords pointing to the protein or nucleic acid sequence of the structure in that file. References to sequence databases are also given here. Information on atoms that do not belong to the protein or nucleic acid, such as water, ions or prosthetic groups is given under keywords that are classified as heterogeneous. The secondary structure section specifies the positions of alpha helices, beta sheets and turns in the protein structures. The connectivity annotation section specifies tertiary interactions such as disulphide bonds, hydrogen bonds and salt bridges. The section on crystallographic and coordinate transformations is, of course, specific to crystallographic structures. Here we find the parameters that are required if we wish to transform the coordinates from the Cartesian axial system (in which they occur in the PDB) to the crystallographic coordinate system (in which they

are specified in the original experimental structure determination). Such transformations are essential to study intermolecular interactions between any given molecule and its neighbours in the crystal.

```
HEADER    DEOXYRIBONUCLEIC ACID              25-AUG-04          1XA2
TITLE     COBALT HEXAMMINE INDUCED TAUTOMERIC SHIFT IN Z-DNA: THE
TITLE     2 STRUCTURE OF D(CGCGCA) .D(TGCGCG) IN TWO CRYSTAL FORMS
COMPAND   MOL_ID: 1;
COMPAND   2 MOLECULE: 5'-D(*CP*GP*CP*GP*CP*A)-3';
- - -
SOURCE    MOL_ID: 1;
- - -
KEYWDS    DOUBLE HELIX, Z-DNA
EXPDTA    X-RAY DIFFRACTION
AUTHOR    S.THIYAGARAJAN,S.S.RAJAN,N.GATHAM
REVDAT    1   16-NOV-04 1XA2       0
JRNL            AUTH        S.THITAGARAJAN,S.S.RAJAN,N.GAUTHAM
JRNL            TITL        COBALT HEXAMMINE INDUCED TAUTOMERIC SHIFT IN
- - -
REMARK    1
REMARK    2
REMARK    2   RESOLUTION. 1.71 ANGSTROMS.
REMARK    3
REMARK    3   REFINEMENT.
- - -
REMARK 290      SYMOP    SYMMETRY
REMARK 290     NNNMMM    OPERATOR
REMARK 290       1555    X,Y,Z
REMARK 290       2555    1/2 – X, –Y, 1/2 + Z
REMARK 290       3555    –X, 1/2 + Y , 1/2 – Z
REMARK 290       4555    1/2 + X, 1/2 – Y, –Z
REMARK 290
```

Figure 3.6 *An example of a PDB file. The file has been extensively edited, wherever indicated by the ellipsis (...).*

The most important part of the entry is specified by keywords in the coordinate section. Of these the atomic coordinates of the molecule all begin with the word ATOM. In contrast to the sections discussed till now, which were in 'free format', i.e., had no particular arrangement within

the line, each line that begins with ATOM has the coordinates of one atom of the molecule arranged in a very specific format. In fact, it is this format that is commonly referred to as the PDB format. The lines marked ATOM in Figure 6 are in this format. Each ATOM record starts with an identifier of the atom that includes a serial number, atom name, residue name, polypeptide or nucleic acid chain name and residue number, all between columns 7 and 26. Between columns 31 to 54 are given the three Cartesian coordinates of the atom in Angstrom units[2] from the origin, each coordinate occupying 8 columns. The next number between columns 55 and 60 gives the value of the occupancy of the atom, in other words it specifies the proportion of the molecules in which that particular atom is present in that position instead of in another. The occupancy is usually 1.0, but some portion of the molecule, such as the active site of an enzyme may have disordered atoms or atoms that occupy different positions at different times. These may have an occupancy value less than 1.0. The next number is the temperature factor and describes the degree of thermal vibration suffered by the atom. The larger the number, the greater the vibration of the atom about its mean position as specified by the coordinates. The temperature factor is a positive number that occupies columns 61 to 66. The rest of the line has an identifier for that segment of data, the symbol for the chemical element of the atom and the charge on the atom, if any.

The ATOM records are followed by the HETATM records that give the coordinates of the heteroatoms in exactly the same format. After the terminator keyword TER comes the connectivity records that describe the bonds that connect the atoms to one another. This information is sometime redundant since such information can always be calculated from the coordinates. If the distance between two atoms is less than the sum of their normal atomic radii, it is usually assumed that they are bonded, the nature of the bond corresponding to the actual distance between the two. The file ends with the keyword END.

As mentioned earlier, the PDB may be obtained on CD and implemented on the local computer by the user. However the RCSB website at the address given above has software that helps search the database and obtain the coordinates of only the molecule of interest. The website is also linked to display software that show pictures of the molecule, which may be turned around, or enlarged or reduced (zoomed in or zoomed out) to get a better appreciation of the three dimensional information. It also has software to make most geometrical calculations. Several other websites, such as the NCBI, also have the entire structural database. MMDB (Molecular Modelling Database) is the name of this database at NCBI, but it is not different from PDB, except that it is linked to different display and retrieval software.

[2] 1 Angstrom unit = 10^{-10} metres

The Cambridge Structural Database (CSD) was originally a project of the University of Cambridge, which set up the Cambridge Crystallographic Data Centre (CCDC) to collect together the published three-dimensional structures of small organic molecules. In terms of molecular weight, 'small' refers to any molecule less than about 1000 Daltons. This excludes proteins and medium sized nucleic acid fragments, but small peptides such as neuropeptides, and monomers and dimers of nucleic acids find a place in the CSD. In the beginning, the staff of CCDC would scan through the published material, in particular, the journal Acta Crystallographica, (where, even today, most of the structures are published) and painstakingly enter the coordinates and other data into the database. This was an expensive task and right from the beginning the database has not freely accessible. Users had to pay to obtain information from it. However, since most of the users were, at that time, from the academic community, the arrangement was that each country would pay a certain amount each year to Cambridge, which would then supply the annual update. The nominated institution in each country would then make the database available to all users in its country. In India, the Department of Crystallography and Biophysics at the University of Madras, Chennai, played the role of the nominated centre for a long time. Lately, however, there have been drastic changes in the profile of the users (pharmaceutical and chemical companies are the most interested users of the data), the technology for collecting and disseminating the data, and the organisational structure of the CCDC. Now any individual laboratory or department can obtain a copy of the database directly from the CCDC on CDs on payment. In countract to the other databases discussed here, this database is not available over the Internet. Currently the CSD holds the crystal structure information of about 2.5 lakh organic and metal organic compounds. All of these crystal structures have been obtained using X-ray or neutron diffraction techniques. For each entry in the CSD there are three distinct types of information stored. These are categorised as bibliographic information, chemical connectivity information and the three-dimensional coordinates. The annotation data fields incorporate all of the bibliographic material for information on the particular entry and summarise the structural and experimental information for the crystal structure. The text and numerical information include the authors' names and the full journal references, as well as the crystallographic cell dimensions and space group. Connectivity information is stored as a table comprising atom and bond properties. The coordinates are referred to as crystallographic axial system. This system is the one that naturally describes the unit cell of the crystal and is the most useful one for not only encoding the molecular structures, but also the other molecules within the crystal that are related to the reference by symmetry transformations. The conversion to (and from) the Cartesian coordinate system is a simple one. The CSD is accompanied by a powerful search and display software called Quest, with a more recent version

called ConQuest. Other software for retrieval, display and analysis of the structures are also supplied.

2.3.2 Derived or secondary databases of biomolecular structures

Just as for protein sequences, secondary databases of structures may also be classified as subcollections, such as the NDB, and collections of patterns and motifs, such as SCOP, CATH, PALI, etc.

NDB stands for Nucleic acid Data Base. It is a relational database of three-dimensional structures containing nucleic acids. This encompasses DNA and RNA fragments, including those with unusual chemistry such as phosphothiorate backbones, RNA molecules, and complexes of nucleic acids and proteins or other molecules. The structures are the same as those found in the PDB, and therefore the NDB qualifies to be called a specialised subcollection. However, a substantial amount of value addition has been carried out, and, unlike the PDB, the NDB is much more than just a collection of files. The structures of DNA have been classified into the A, B and Z polymorphic forms, based on information supplied by the authors. Other classes include RNA structures, unusual structures, and protein-nucleic acid complexes. These classes of structures are arranged in the form of an Atlas of Nucleic Acid Containing Structures, which can been browsed and searched to obtain the structure or structures required. Each entry in the atlas has information on the sequence, crystallisation conditions, references, and details of the parameters and figures of merits used in structure solution. The entry has links not only to the coordinates, but also to automatically generated graphical views of the molecule. NDB also has archives of structural geometries calculated for all the structures or for a subset of them. And finally, the database stores average geometrical parameters for nucleic acids obtained by statistical analysis of the structures. These parameters are widely used in computer simulations of nucleic acids and their interactions The NDB may be accessed at http://ndbserver.rutgers.edu/NDB/.

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. The principal classes are the family, the superfamily and the fold. A set of proteins are said to belong to the same family if there exists an evolutionary relationship between them that can be clearly demonstrated on the basis of similarity in the structure. While this usually also implies sequence similarity of greater than 30%, some of families, for instance among the globins, have a very high degree of similarity in structure and function concurrently with less than 15% similarity in sequence. The superfamily is at the next level, and brings together proteins that again have low sequence identity, but with sufficient similarity in the structure and function to suggest a common ancestor. In some cases a domain of a protein (i.e., a structurally autonomous

portion of a large protein) may be a member of a superfamily together with other domains or complete proteins. A superfamily is obviously a superset of families, with each superfamily subsuming many families. The third major class is the fold, which is a superset of superfamilies. Proteins that belong to the same fold have the same secondary structure elements, in the same three dimensional arrangements with the same or similar connections between the secondary structures. No evolutionary relationship is implied between the members of the same fold. SCOP is a searchable and browsable database. In other words, one may either enter SCOP at the top of the hierarchy and examine the different folds and families as one pleases, or one may supply a keyword or a phrase to be used to search the database and retrieve corresponding entries. Once a structure, or a set of structures, has been selected, they may be obtained or viewed either as PDB entries or as graphical images. Each entry also has other annotation regarding function, etc., and links to other databases, including other structural classifications such as CATH.

CATH stands for Class, Architecture, Topology and Homologous superfamily. The name reflects the classification hierarchy used in the database. The structures chosen for classification are a subset of PDB, consisting of those that have been determined to a high degree of accuracy. This choice allows greater reliance to be placed on the resulting classification. The next step in processing the data for classification consists of identifying domain structures in the proteins. Though the programs are automatic, multiple methods are used, and only those domains that have been identified by all of them are accepted as such. Each identified domain is considered an individual protein in the classification procedure. As mentioned above, there are four main levels in the CATH hierarchy. The first and most general of them is class or the C level. This refers to the amount of a particular type of secondary structure present in the structure. Thus a protein may consist entirely of alpha helices (alpha class) or of beta sheets alone (beta class) or a mixture of both (alpha/beta class) or of low secondary structure (coil). The next level of the hierarchy is architecture or A level. Structures classified as having the same architecture have same orientation of their secondary structures, but the order in which they occur along the polypeptide chain may be different. The third level consists of structures classified as having the same topology or T level. This term refers to both the orientation and the connectivity of the secondary structures in the structure. The final level is called the homologous superfamily or H level. Each group under this classification contains proteins that are closely related in structure, and therefore in function, implying a common evolutionary origin. There is yet another level of similarity in the hierarchy, the sequence or S level. Here the structures in each H group are clustered together on the basis of sequence similarity. CATH is accessible at http://www.biochem.ucl.ac.uk/bsm/cath for browsing and for searching. In addition the files of lists may be downloaded

and implemented locally to suit any particular purpose. There are three main files of lists. The first one is the actual classification. Each item in this list consists of an ID that identifies the protein domain, and along with it specifies the identification number of the class, architecture, topology, homologous superfamily and sequence groups in which it belongs. The second list consists of a single protein in each of the different groups, and may therefore be considered as the file that specifies and describes the groups. The third file lists the domains considered in CATH and links them to the PDB entries.

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 constantly try to devise new algorithms to match strings of symbols, in general. This refers to the fact that DNA and protein sequences may be considered strings of symbols chosen from limited sets—four symbols in the case of DNA and twenty in the case of proteins.

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. The common pattern may be a short stretch of one sequence that is similar to a short stretch of the other sequence, or many short similar bits of the two sequences, or a part of one sequence matching a part of the other, or the similarity may be spread all across both sequences. 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 Homologs, heterologs, analogs, orthologs, paralogs, xenologs

According to the definition given in the previous section, all homologous sequences have descended from some common ancestral sequence.

Homologs is a general term to indicate sequences of common origin. The opposite of homologs is *heterologs*. Heterologous sequences may still be similar, but they do not have a common origin nor do they have a common function or activity.

Sequences that have the same function but lack sufficient similarity to imply common origin are said to be *analogs*.

Orthologs are homologs that arise by speciation. The implication is that the two sequences are currently taken from two different species, but they show similarity because they are both derived from the same ancestral sequence that was present in the ancestral organism.

Paralogs are homologs that arise by gene duplication, without this duplication event being followed by a speciation event. That is to say, paralogs are homologous sequences that exist in the same organism, and that have different functions. Paralogs have homologous origin but heterologous function.

Xenologs are homologs resulting from horizontal gene transfer. They are an exception to the rule that homologs are always descended from a common ancestor.

3.1.3 Scoring schemes

Here we will introduce, in Table 3, the PAM100 scoring scheme [54]. The PAM (Percent Accepted Mutation) series of matrices are 20 x 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 labelling that element as matching residues. The unit of measure is 'bits', since the matrices are derived from information theoretical considerations. This matrix is called the substitution matrix $s(x, y)$, where x and y represent amino acids.

Gap penalties are also part of the scoring scheme, and must be chosen along with the substitution scores. Here we briefly remark that there are two aspects to gaps—the number of gaps in the alignment and their respective sizes. Normally, therefore, there is a *gap opening* penalty, which is the basic penalty applied whenever a gap exists. In addition for each gap there is a *gap extension* penalty, which depends on the size of the respective gap. We could use sophisticated functions to reflect various biological realities, but in the example below 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.

3.2 BLAST: Basic Local Alignment Search Tool

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 performs fast searches through large databases for matches

to the query sequence, and then performs more detailed alignments of the query sequences with the matches. BLAST compares a DNA sequence against a DNA database, a translated (in all six frames) version of a DNA sequence against a translated (six-frame) version of the DNA database, a translated (six-frame) version of a DNA sequence against a protein database, a protein sequence against a 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*. A word list is prepared for the query sequence and is searched against the table for the database to identify exact matches [2]. These are the maximal segment pairs or MSPs. MSPs do not allow gaps, and have the very valuable property that their statistics are well understood. Thus, we can readily compute a significant probability for a maximal segment pair alignment. The word size W is chosen as 3 for proteins, and 11 for nucleic acid sequences. This is because, since there are only four nucleotides, the amount of background noise is too large with smaller word sizes. These word hits of size W do not have to be identical; rather, their scores have to be better than some threshold value T. Each double word hit that passes this step then triggers a process called ungapped extension in both directions, such that each diagonal is extended as far as it can, until the running score starts to drop below a pre-defined value within a certain range. The result of this pass is called a High-Scoring segment Pair or HSP. Those gapped alignments with expectation values or E-values better than the user specified cutoff are reported. They are now the 'expectation,' that an alignment with a score better than the one reported may be obtained purely by chance.

BLAST is most frequently used over the Internet on the BLAST server (http://www.ncbi.nlm.nih.gov/BLAST/). There are several versions of BLAST and the home page lists all of them. 'blastn' is 'nucleotide–nucleotide' BLAST, or the matching of a nucleic acid sequence against the nucleic acid sequence database. Similarly 'blastp' is protein-protein BLAST. A different subset of programs performs translated searches, where either the query ('blastx'), or the database ('tblastn'), or both ('tblastx') are nucleic acid sequences translated in six frames into protein sequences. There are other specialised programs such as, for example, 'megablast'. This program uses a different algorithm for nucleotide sequence alignment search. It is optimised for aligning sequences that differ slightly, perhaps as a result of sequencing errors [2].

The sequence has been entered in the so-called FASTA format, in which the first line, always beginning with the symbol '>', contains a brief user-controlled description of the sequence or any other information the user requires. The sequence begins on the next line and proceeds without gaps to the end, with any number of line breaks.

Table 4 *The PAM100 substitution matrix*

A	4	−3	−1	−1	−3	−2	0	1	−3	−2	−3	−3	−2	−5	1	1	1	−7	−4	0
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

This section would be incomplete without a description of two additional algorithms that have been added in 1998 to the BLAST family. The first is called PSI-BLAST, and stands for Position Specific Iterated BLASTPosition Specific Iterated BLAST] [24]. This algorithm returns more distantly related sequences from the database than BLAST. In other words, the sensitivity of the search is improved, without overly compromising on the specificity. PSI-BLAST brings this about this by using a profile to search the database instead of just the query sequence.
The second algorithm referred to above is called PHI-BLAST and stands for Pattern-Hit Initiated BLAST [92]. This is a search program for which the input is not only a query DNA or protein sequence, but also a pattern. PHI-BLAST helps to answer the following question: given a query sequence that contains a particular recognised pattern, what other sequence in the database has the same pattern and is homologous to the query sequence in the neighbourhood of the pattern? In the cases where such patterns are known, PHI-BLAST is particularly useful is filtering out false positives of sequence similarity that may arise by chance.

4 Multiple Sequence Alignment

A multiple sequence alignment, or MSA, may be formally defined as a two-dimensional table in which each row represents a protein or nucleic acid sequence, and the columns are the individual residue positions. The table is obtained by aligning all the sequences being considered simultaneously in order to obtain the best overall score. Such simultaneous alignment of several sequences has led to many important results regarding common sequence patterns or motifs in proteins and nucleic acids. 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. The study of evolution at the molecular level is strongly assisted by establishing such phylogenetic networks, and MSA usually provides the initial information to build the networks. And lastly, MSA is also usually the first step also in building three-dimensional models of protein structure. MSA helps to predict the secondary and tertiary structures for new sequences, and identify templates for threading and homology modelling, which are methods for 3-D structure prediction.

4.1 Scoring an MSA

The most common way of finding the *score* of any given MSA is the so-called *sum-of-pairs* or SP score. We consider the representation of the

MSA as a two-dimensional matrix [57]. In the SP scoring scheme, the score of the alignment is the sum of the scores of each of the columns. The score of column i is given by the expression

$$S_i = \sum\ s(\text{residue}_i^k,\ \text{residue}_i^l)$$

In this expression, the indices k and l refer to the different sequences. Residueik is the residue in the kth sequence and ith column, and likewise residueil is the residue in the lth sequence and ith column. s(residue$_i^k$, residue$_i^l$) is the substitution matrix score for the pair of residues indicated. The summation is made over every pair of residues in the column. In case one of the pair of residues is a gap, a gap score is defined as s(residue,gap) or s(gap,residue) and added to the sum.

CLUSTAL is a popular program for MSA that uses an extensively modified version of the Feng-Doolittle algorithm [17]. The CLUSTAL algorithm builds up the MSA by using such profiles wherever appropriate. The steps in the procedure are as follows. As in the previous algorithm, the first step is to construct a half-matrix of $n(n-1)$ distances between all pairs of n sequences by standard pairwise alignment. Feng and Doolittle calculated the distance between two sequences a and b by the following expression [86].

$$D_{ab} = -\log[(S_{ab} - S_{\text{rand}})/(S_{\max} - S_{\text{rand}})],$$

where S_{ab} is the best similarity score between a and b, S_{rand} is the random score obtained by aligning two sequences with the same length and residue composition, $S_{\max}$ is the maximum possible score, obtained by aligning each of the two sequences to itself and taking the average of the two maximal scores. In the next step of the algorithm, a guide tree is constructed by using the neighbor-joining algorithm for clustering. The final step in the algorithm consists of using the guide tree to progressively add sequences to the MSA, starting for that pair of sequences that are the closest to one another. Everytime 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. There are a variety of such scoring schemes available, constructed on the basis of different principles. We

write the 20 amino acids along the topmost row as well as along the left-most column of the 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 PAM substitution matrices

PAM stands for Percent Accepted Mutation [92]. In an alignment between two protein sequences, if an amino acid in the first is substituted by another in the second, it indicates a point mutation in the sequence. These are the two chief features considered in compiling the PAM matrices, viz. there must a replacement of one amino acid by another; and this substitution must be accepted by natural selection. We may write a matrix A whose elements A_{ij} are the counts of the number of times the residue i has changed to residue j.

4.2.3 BLOSUM substitution matrices

BLOSUM stands for BLOcks SUbstitution Matrices. In 1992, Henikoff and Henikoff devised the BLOSUM family of substitution matrices. Just as in the case of the PAM matrices, the scores are obtained as the logarithms of likelihood ratios [28]. The elements of the transition probability matrix were obtained by the analysis of these blocks of aligned sequences.

4.2.4 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. A single mutation can create a gap — this is perhaps the most common cause. Sometimes unequal crossover during meiosis can lead to insertion or deletion of strings of bases in the DNA sequences. Gaps may occur at three possible locations in an alignment: before the first character of one of the sequences, inside one of the sequences or after the last character of one of the sequences. Gap penalties are also part of the scoring scheme, and must be chosen along with the substitution scores. There are two reasons for applying a penalty whenever a gap is introduced, one practical, and the other biological. The practical reason is that if there were no penalties for gaps and if any number and size of gaps was to be allowed then even two random sequences may be aligned with high scores. The biological reason for allowing gaps is that during the course of evolution related sequences diverge by acquiring insertions and deletions (or 'indels').

4.2.5 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 [86]. It is convenient to represent phylogeny as a tree in which lines represent species. At some places a line splits into two to represent points where ancestral species evolved through speciation into two new species. However, one could discuss phylogeny and phylogenetic trees without reference to evolution and the formation of new species, and simply as the relationships between different individuals being compared. In this sense, a phylogenetic tree is considered as a graph with a set of nodes, and lines joining them.

5 Protein Structure Predictions and Protein Folding

The three-dimensional structure of a molecule is considered 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*, which comprises helices and beta sheets. 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 typical length is 10 amino acids and it is stabilised by hydrogen bonds. It is important to mention that there are only right-handed helices in nature although left-handed helices could be possible if there is availability of. The beta sheet can be parallel (same direction of the chain) or antiparallel, it can 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. This form is also called the functional form. Tertiary structure of a protein includes the coordinates of its residues in three-dimensional space. *Quaternary structure* is the structure of a protein complex. Some proteins form a large assembly to function. This form includes the position of the protein subunits of the assembly with respect to each other. Protein structure prediction operates primarily at the level of the secondary and tertiary structure. The fundamental principle underlying all the methods is that the sequence of the protein, viz. its primary structure, completely specifies the final three-dimensional form of the functional protein. The different kinds of protein structures are shown in Figure 7($a-d$). The different kinds of protein structures are shown in Figure 7($a-b$).

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 its three (or four) secondary structural states.

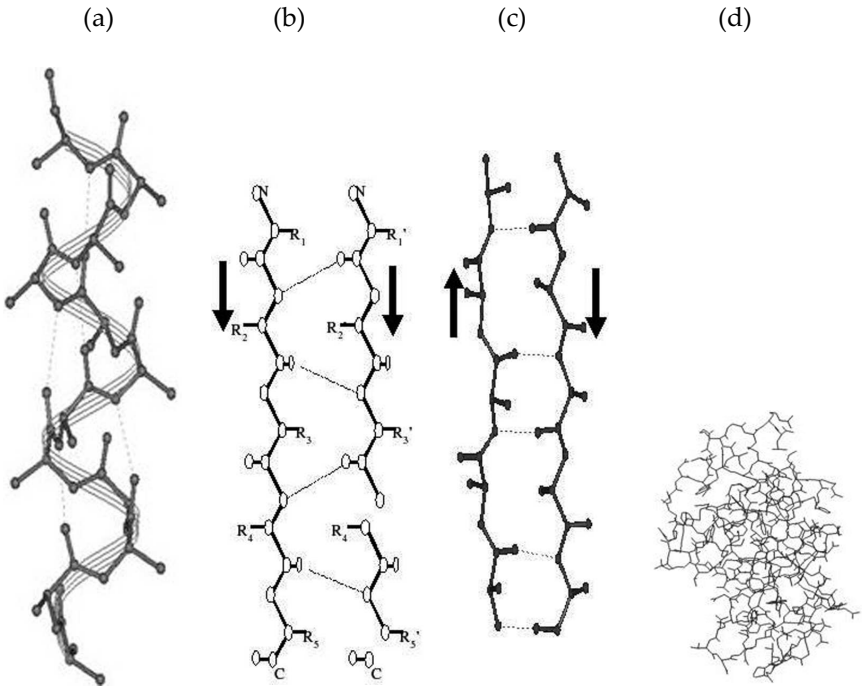

Figure 7 *(a) The α-helix. Two or more such strands coming together to form β-sheets in (b) parallel orientation or (c) antiparallel orientation. (d) The tertiary structure is shown as a line diagram.*

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. Protein secondary structure prediction methods may be broadly classified according to the techniques used as those based on single residue statistics, those based on explicit rules and those based on neural networks.

5.1.1 Neural networks in secondary structure prediction: PHD

Computer-based artificial neural networks (ANNs) consist of units analogous to neurons that receive input information from other units and send output signals to yet others. The connections are many-to-one and one-to-many [27]. This variety allows the use of ANNs in several different fields including computational linguistics, speech recognition, recognising faces, speech synthesis, and so on. Neural networks used in secondary structure prediction are the so-called multi-layered, feed-forward networks.

The units are arranged in layers and the connections are always in the forward direction from the units in the lower layer to the ones in the upper layer. The mathematical function used to process the input at each unit is such as to produce an output signal that is close to one of two values, i.e., 0 or 1 (or sometimes −1 or +1). This allows the ANN to be configured as a classification or recognition system [27]. In the present case, the ANN is used to classify the amino acids as belonging to a specific secondary structure type. Figure 8 shows a three-layered ANN, with the first input layer accepting information from each residue in the sequence of the protein. Each unit in this layer applies the processing function to all the input it receives and, according to the result, sends out an output signal to each unit in the second hidden layer [27]. This procedure is again repeated in this layer and an output signal is sent to each layer in the third or output layer.

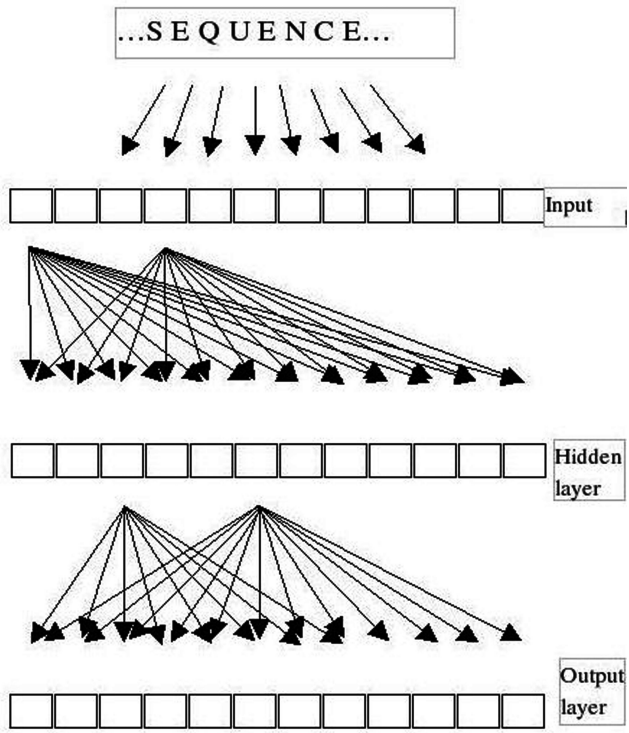

Figure 8 *A schematic diagram of a feed-forward multi-layered artificial neural network (ANN). For clarity, only some of the interconnections between the layers are shown. But, in fact, every element of a layer is connected to every element of the next layer.*

The use of an ANN involves three steps, viz. construction, training and utilisation. The first step involves decisions regarding the number of layers, the number of units in each layer, and the mathematical function, called the gating function, that is to be used in each layer to connect the input to the output [29]. These decisions, in particular the last one, give rise to a number of weights and constants that have to be adjusted in order to obtain correct performance from the ANN. For example, the gating function is usually a weighted sum of all the inputs, multiplied by a sigmoid function that results in a number close to either 0 or to 1. *Training* the ANN carries out the adjustment of the weights by using a set of data where the answers are already known. In the present context, this means a set of sequences for which the secondary structure is already known. This training set of sequences is fed into the newly constructed ANN and for each residue in each sequence the prediction by the ANN is compared with the known structure. This is repeated for all the residues and all the sequences, through several cycles, until the weights do not change and the predictions are as close as possible to the correct structures.

Here we describe one of most successful and commonly used implementations, namely PHD. PHD consists of several steps. First the input sequence is compared with the 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), a window of thirteen residues having six residues on either side is considered. 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 in the output layer.

5.2 Protein tertiary structure prediction

The literature on methods to predict the three dimensional structure of a protein from its sequence alone is vast, and scores of techniques have been devised. In general, the techniques may be divided into three broad categories: homology modelling, threading techniques and *ab initio* structure prediction.

5.2.1 Homology modelling

This is also known as comparative protein modelling or knowledge based modelling, and the word homology, used in this context, does not necessarily imply a strong evolutionary relationship, though such a relationship may well be present. The prime requirement is the existence of a model structure, usually one that has been determined experimentally. The technique 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. 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. A wrong selection at this stage cannot be corrected at a later stage in the modelling and careful attention to a proper choice of the template is amply rewarded by the final results. It is best if the target and the template are actually homologous proteins, in the strict biological sense of sharing a common ancestor. 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 several sequence and structure databases and efficient software that help find a match for the target sequence. 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 categorised as possessing similar structures and functions. Firstly that template is best that has the closest sequence similarity to the target. Secondly, the template structure that we use should, as far as possible, have been determined in the same biochemical conditions of pH, ionic strength, etc., as those in which we desire the structure of the target. And finally, the template structure should be a reliable one.

After selecting the template, the second step in the modelling procedure is to align the target sequence with the template sequence. While sequence alignment precedes template selection, the alignment needs to be repeated with greater rigor once the selection is made. This is especially true when the similarity score is low. We now use pairwise global sequence alignment method. Again if possible, 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 a profile constructed. Other sequences belonging to the same family may 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. There are conceptually two ways this can be achieved. The first is to use the structure of the template directly as the structure of the target, replacing the side chains of the residues that differ. Another way to build 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 optimising the entire structure.

The final step in the modelling process is evaluation of the model. We first ensure that the model satisfies physical constraints and principles, such as non-interpenetration of the atoms. Next in degree of importance comes the stereochemistry, such as bond lengths, angles, 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 [31]. The sequence in which the amino acids occur along the polypeptide chain closely controls how the chain will fold into the three-dimensional structure.

The polypeptide chain is built up when the amino acids join to each other by means of a peptide bond. The peptide (C-N) bond has a partial double bond character, which means that the molecule cannot rotate about it. The six atoms that form the peptide group are constrained to lie on a plane thereby forming a planar peptide group. If the rotational possibilities of the longer side chains are ignored, and the backbone of the protein chain alone is considered, then each residue is described by two torsion angles Φ and Ψ [31]. Figure 9 clearly represents the two torsion angles. The freedom of rotation about even these bonds is not absolute, and is restricted by possible steric hindrances. For example, if the $C^{\alpha}_{i-1} - N_i$ bond is *cis* to the $C^{\alpha}_i - C'_i$ bond when the $C_i - N_{i+1}$ bond is *cis* to the $N_i - C^{\alpha}_i$ bond, i.e. when both Φ and Ψ are 0 degree, then a severe clash between H_{i+1} and O_{i-1} will occur, making this particular pair of values disallowed. The map is correct for eighteen of the twenty side chains. The exceptions are glycine and proline. Glycine has a much larger allowed region owing to the absence of the C^{α} atom. The side chain of proline binds to the peptide nitrogen atom and this limits the value of Φ to -60 Ψ 20 degrees. Finally, biological and biochemical principles are considered. The model should, of course, satisfy all such known concepts [31].

Despite passing all the evaluation tests, the model could still possess one or many of the following errors. There could be errors in the packing of side chain atoms in the core of the protein. There could be errors in the trace of the main chain polypeptide sequence as a consequence of incorrect placement of the side chain atoms. There could be errors that arise due to the insertions or deletions of a long stretch of residues. A model for such a

stretch in the target becomes difficult to build. When the target−template identity is less than 30% especially, the source of error lies in the incorrect alignment of the two sequences When the target−template identity is less than 30%, especially, the source of error lies in the incorrect alignment of the two sequences. This implies that incorrect structures, chosen from the template are assigned to misaligned residues in the target. Using multiple sequences and/or structures to build the model may reduce alignment errors.

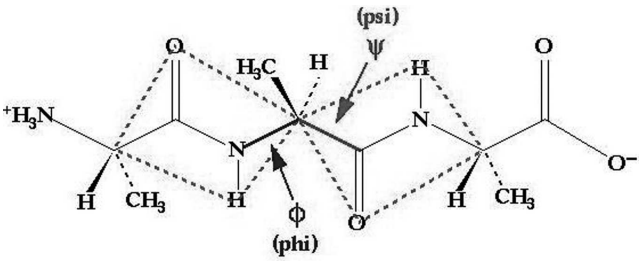

Figure 9 *Ramachandran plot showing the two torsion angles* Φ *and* Ψ

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. Models of the accompanying structural changes may be built to understand the effects, and perhaps design more effective drugs. Highly accurate models based on more than 50% sequence identity have average accuracy nearly that of low resolution of X-ray structures. 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 generalises the technique of homology modelling, 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 optimised to obtain the best of experimental values such as geometries, conformational energies and spectroscopic properties.

6.1 Force fields

6.1.1 Energy calculation

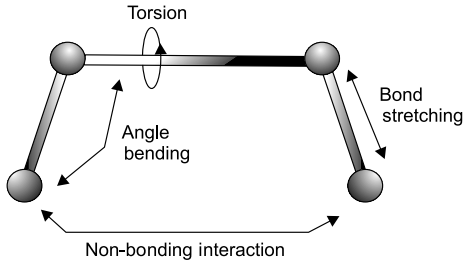

Figure 10 *Forces acting within the system*

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

Many of the molecular modelling force fields in use today can be interpreted in terms of a relatively simple four-component picture (Fig. 3.10) of intra- and inter-molecular 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 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].

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

Here, k_θ controls the stiffness of the angle, θ is the current bond angle and θ_0 the equilibrium angle. Both the force and equilibrium constant have to be estimated for each triple of atoms.

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

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

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, a Van der Waals term and a Coulomb term [35].

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

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 a technique called the united atom method, where atom groups with non-polar 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 reduce the number of solvent molecules required and therefore accelerate the molecular dynamic 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 via (a) random collisions, and by (b) 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].

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

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

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

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 the friction coefficient. The third contribution, the force $Ri(t)$ is due to random fluctuations caused by interactions with solvent molecules.

$$E_{\text{elec}} = \underbrace{\sum_i^N \sum_{j=i+1}^N \frac{q_i q_j}{r_{ij}}}_{\text{Coulomb}} - \underbrace{\frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right) \sum_i^N \sum_{j=i+1}^N \frac{q_i q_j}{f(r_{ij}, a_{ij})}}_{\text{generalized Born}}$$

$$f(r_{ij}, a_{ij}) = \sqrt{r_{ij}^2 + a_{ij}^2 e^{-D}}, \text{ where } a_{ij} = \sqrt{a_i a_j} \text{ and } D = r_{ij}^2 / 2(a_{ij})^2$$

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 needed for

an accurate description of solvation. Other approaches, named 'implicit models', represent the environment (counterion, solvent) around macromolecules as a continuum. Such models must describe the damping of the electrostatic interaction by the solvent in an appropriate way. There are several force field program packages available for biomolecular computation [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 the two most important types of macromolecules in biochemistry, namely peptides and nucleic acids. There is a difference between the AMBER program package and the so-called AMBER force field, which is implemented in the AMBER package, but also in various other programs. The force field is public domain, whereas the package is distributed under license agreement. The current version of the package, AMBER 8. 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 the coordinate file prmcrd.

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

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

$$E_{\text{total}} = \sum_{\text{bonds}} K_b(r - r_0)^2 +$$

$$+ \sum_{\text{angles}} K_\theta (r - r_\theta)^2 +$$

$$+ \sum_{\text{torsions}} \frac{V_n}{2}(1 + \cos(n\omega - \gamma)) +$$

$$+ \sum_i \sum_{i<j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}} \right] +$$

$$+ \sum_{H-\text{bonds}} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{10}} \right] +$$

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

7 Homology Modelling and Molecular Dynamics of Cyclin-Dependent Kinases

7.1 Introduction

Although protein function is 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, modelling 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 modelling facility with integrated expert knowledge was first implemented 50 years ago by Peitsch txi 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 catalyse the transfer of γ-phosphate in adenosine triphosphate (ATP) to a protein substrate. CDKs are crucial regulators in timing and co-ordination 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 humans, cell cycle progression is governed by several CDKs and cyclins. In particular, CDK4-cyclin D is required for the pass through cell cycle to the G1 phase, 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 make 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 served as a model for the catalytic core of other CDKs, including CDC28 [91, 92]. But there is a question about correcting such approximation.

7.2 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 (Figure 11). The primary structure for the yeast CDC28 has 298 amino acids and can be obtained from theSWISS PROT database (http://cn.expasy.org/sprot/) server. The modelling 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/).

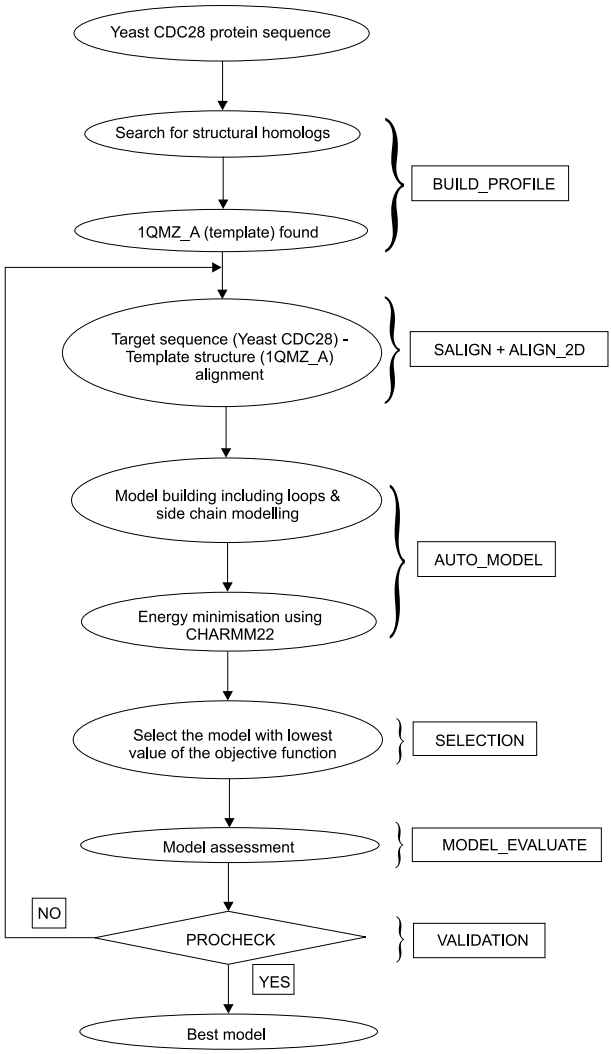

Figure 11 *Flowchart for homology modelling of yeast CDC28*

MODELLER calculates the three dimensional structure for the query sequence by searching for the related matching structures using satisfaction of spatial restraints [59]. 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 CHARMM22 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 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 optimised 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 that 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 MODELLER executes this sequence profile construction. The BUILD_PROFILE command has many options. Unrecognised 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 falls below the threshold set. Lower E value thresholds are more stringent and report fewer matches.

Many hits were displayed on the basis of the sequence identity and E value between the protein sequences. The query sequence found 64.11% identity and E value = 0 with PDB entry (phosphorylated CDK2-cyclin A-substrate peptide complex) of the human species by running the MODELLER program (Figure 12). The matching part of the PDB entry: 1QMZ chain-A is derived from the significant hit was used as the template structure for the model building.

```
#SEQ_DATABASE_FILE         : pdball.pir
#SEQ_DATABASE_FORMAT       : PIR
#CHAINS_LIST               : ALL
#CLEAN_SEQUENCES           :       T
#MINMAX_DB_SEQ_LEN         :      30    4000
#Number of sequences       :   72419
#Number of residues        : 17530589
#Length of longest sequence:    1491
# gap_penalties_1d=(-500, -50)
# matrix_offset=-450
# rr_file='${LIB}/blosum62.sim.mat'                                  Sequence   E-Value
#Read the alignment from file       : yeastcdc28.ali                 Identity
#Total number of alignment positions:  298
# HITS FOUND IN ITERATION:     1
  > 1py5A                     1   43234    6000     301    298   26.34    0.67E-05    389   234    12   264    10    27
  > 1pyeA                     1   43235   39150     266    298   62.11    0.0         390   251     5   297     1    25
  > 1q24A                     1   43334    9150     335    298   28.91    0.0         391   202     2   223    24    23
  > 1qcfA                     1   43527    8100     449    298   28.03    0.20E-09    392   222    12   249   190    42
  > 1q16A                     1   44063    9650     281    298   27.90    0.0         393   263     3   297     4    27
  > 1qmzA  -->Template        1   44186   48600     296    298   64.11    0.0         394   282     5   297     2    28
  > 1qmzC                     1   44188   48600     296    298   64.11    0.0         395   282     5   297     2    28
  > 1qpcA                     1   44414    9000     271    298   28.34    0.0         396   183     9   205    16    20
  > 1qpdA                     1   44415    8800     271    298   28.34    0.0         397   183     9   205    16    20
  > 1qpeA                     1   44416    9000     270    298   28.34    0.0         398   183     9   205    16    20
```

Figure 12 *Sequence identity and E value between the protein sequences. The query sequence found 64.11% identity and E value=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 MODELLER, which relies on a multiple structure alignment method similar to that in the program COMPARER [63]. Target sequence−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 analyse the close relation between the target and template protein sequences, we carry out the comparative modelling procedure. Comparative modelling 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 the variable gap penalty function that is executed from the template structure alone. In principle, the errors between the target−template alignment is greatly minimised almost by one-third relative to the present day sequence alignment methods (Figure 13).

Models are built for each of the sequence−structure matches using MODELLER [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].

```
_aln.pos              10          20          30          40          50          60
1qmzA     ---SMENFQKVEKIGEGTYGVVYKARNKL---TGEVVALKKIRLDTETEGVPSTAIREISLLKELN
Yeastcdc28 MSGELANYKRLEKVGECTYGVVYKALDLRPGQGQRVVALKKIRLESEDEGVPSTAIREISLLKELK
_consrvd              *     **  **********          *********  *  ******************

_aln.pos        70          80          90          100         110         120         130
1qmzA     HPNIVKLLDVIHTE–NKLYLVFEFLHQDLKKFMDASAL–TGIPLPLIKSYLFQLLQGLAFCHSHRV
Yeastcdc28 DDNIVRLYDIVHSDAHKLYLVFEFLDLDLKRYMEGIPKDQPLGADIVKKFMMQLCKGIAYCHSHRI
_consrvd       *** * *   *     *********  ***   *                 *     **   * * *****

_aln.pos            140         150         160         170         180         190
1qmzA     LHRDLKPQNLLINTEGAIKLADFGLARAFGVPVRTYH–EVVTLWYRAPEILLGCKYYSTAVDIWSL
Yeastcdc28 LHRDLKPQNLLINKDGNLKLGDFGLARAFGVPLRAYTHEIVTLWYRAPEVLLGGKQYSTGVDTWSI
_consrvd  ************   *   ** ***********  * *   * **********  ***  *  ***  **  **

_aln.pos  200         210         220         230         240         250         260
1qmzA     GCIFAEMVTRRALFPGDSEIDQLFRIFRTLGTPDEVVWPGVTSMPDYKPSFPKWARQDFSKVVPPL
Yeastcdc28 GCIFAEMCNRKPIFSGDSEIDQIFKIFRVLGTPNEAIWPDIVYLPDFKPSFPQWRRKDLSQVVPSL
_consrvd  *******   *    *  ******** *  *** ****  *  **       **  *****  *  *  *  *  ***  *

_aln.pos        270         280         290         300
1qmzA                 DEDGRSLLSQMLHYDPNKRISAKAALAHPFFQDVTKPVPHL
Yeastcdc28 DPRGIDLLDKLLAYDPINRISARRAAIHPYFQES–––––––
_consrvd  *   *  **    * ***  ****   *  **  **
```

Figure 13 *Target−template alignment. (*) shows the matching between the residues and (-) shows the gaps.*

Model building: In this section 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 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 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 most of the CHARMM energy terms [68, 69] to obtain a full objective function. Then the model with lowest value of the objective function is selected and assessed using the MODEL_EVALUATE module of the MODELLER program.

Loop modelling: 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 high degree of variation in the structural information between these two proteins. These regions are generally addressed as loops. Loops often 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 modelling problems, the loop modelling procedure was evaluated by predicting loops of known structure in only approximately correct environments. Models 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 modelling: The geometry of the 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 experimentally available structural information.

Selecting the appropriate model for the yeast CDC28: A three dimensional model was generated by a GA341 score that is the parameters including Z-score (Zs) calculated with a statistical potential function [75], target--template sequence identity (Si) and a measure of structural compactness (Sc) [75,66]. The GA341 score is defined as:

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

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 deviation 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 (Figures 14−15).

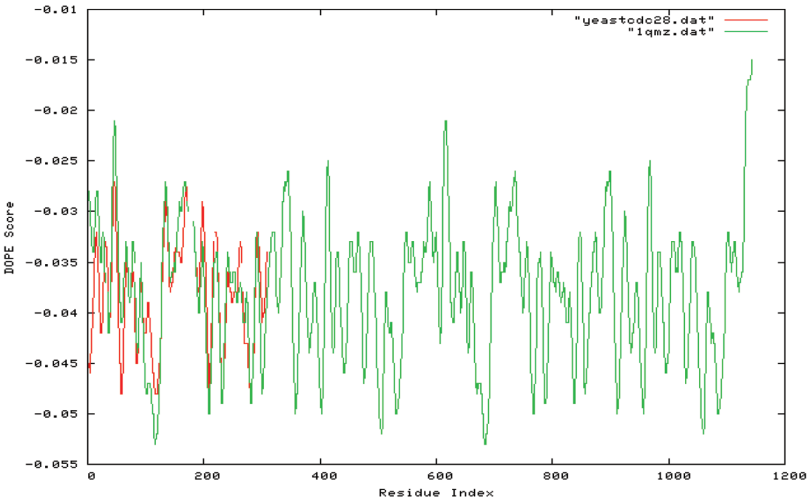

Figure 14 *Graph plot between the DOPE energy and residues for both 1QMZ and yeast CDC28. The overlapping structure shows the chain A. The kinase part of 1QMZ is correlated with modelled CDC28 showing high homology.*

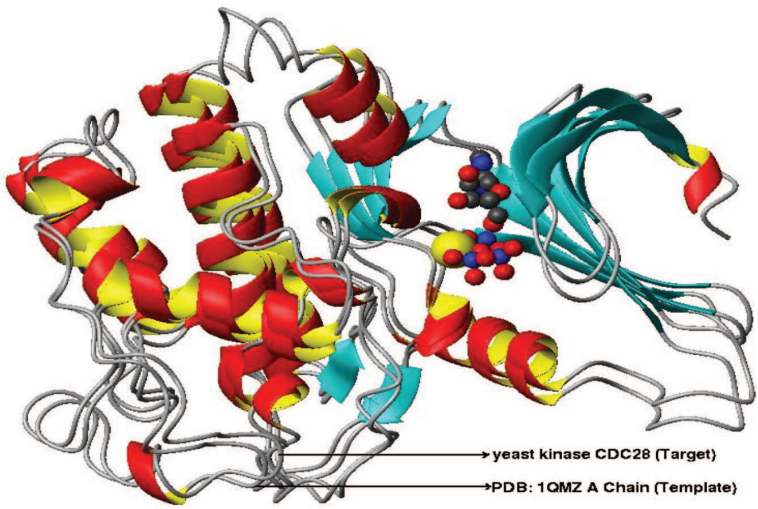

Figure 15 *Superposition of the target structure (PDB: 1QMZ A chain) and the modelled template structure (yeast kinase CDC28). ATP complexes are shown as ball models. Magnesium ion is shown in yellow.*

Assessment of the model: This is necessary in order for MODELLER to correctly calculate 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 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 Ca atoms of the artificially generated three-dimensional model and the native structure i.e., 1QMZ_A was calculated upon rigid body least squares superposition of all the Ca atoms. Next, the percentage of high matching regions between the yeast CDC28 and the 1QMZ_A was defined in terms of the percentage of the Ca atoms in the model that are located within the proximity of 5 Å of the corresponding atoms in the superposed structure (Figure 14). In order to enhance the best model MODELLER 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 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 includes 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 is 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-handed rotation around N-Ca bond and the psi (ψ) angle is the angle of right-handed rotation around Ca-C bond. Phi and psi 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 psi/phi 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 the in 'disallowed' regions. The score for the crystal structure 1QMZ is 91.1% (Figure 17). The score for yeast CDC28 was 91.5% (Figure 16) and lies in the allowed region, which confirms good homology prediction.

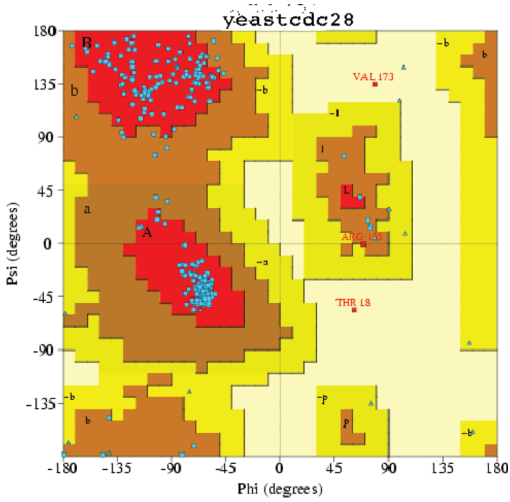

Figure 16 *Ramachandran plot for yeast CDC28 {Most favoured regions = 236 (number of residues), 91.5% (percentage). Additional allowed regions = 19 (number of residues), 7.4% (percentage). Generously allowed regions = 1 (number of residues), 0.4% (percentage). Disallowed regions = 2(number of residues), 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.}*

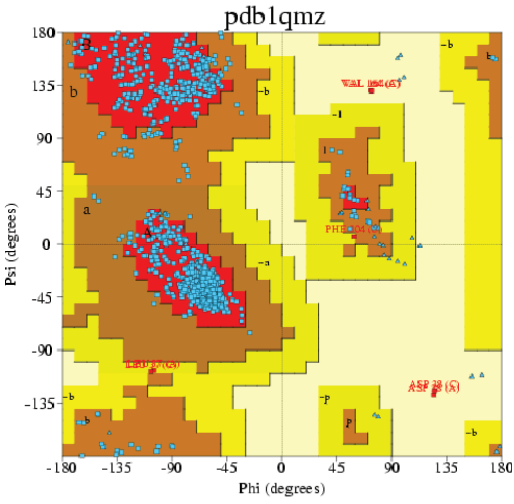

Figure 17 *Ramachandran plot for PDB 1QMZ {Most favoured regions = 906 (number of residues), 91.1% (percentage). Additional allowed regions = 81 (number of residues), 8.1% (percentage). Generously allowed regions = 3 (number of residues), 0.3% (percentage). Disallowed regions= 4 (number of residues), 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.}*

7.3 Molecular dynamics simulations

For the MD simulations, the SANDER modules of the program package AMBER 8.0 [78] and of the modified version of AMBER 7.0 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 the calculations [84]. Firstly, a potential energy minimisation was performed for each system on an initial state. Then, the MD simulation was performed on the energy-minimised 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 analysed by RasMol [86] and MOLMOL [87] packages. Complete data flow in AMBER is shown in Figure 18.

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

$$\text{RMSD} = \sqrt{\frac{\sum_i d_i^2}{n}}$$

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 less so for structural divergence. The RMS deviation of the Molecular Dynamics structures from the crystal structure 1QMZ and the modelled structure of the yeast cyclin-dependent kinase CDC28 vs time is calculated.

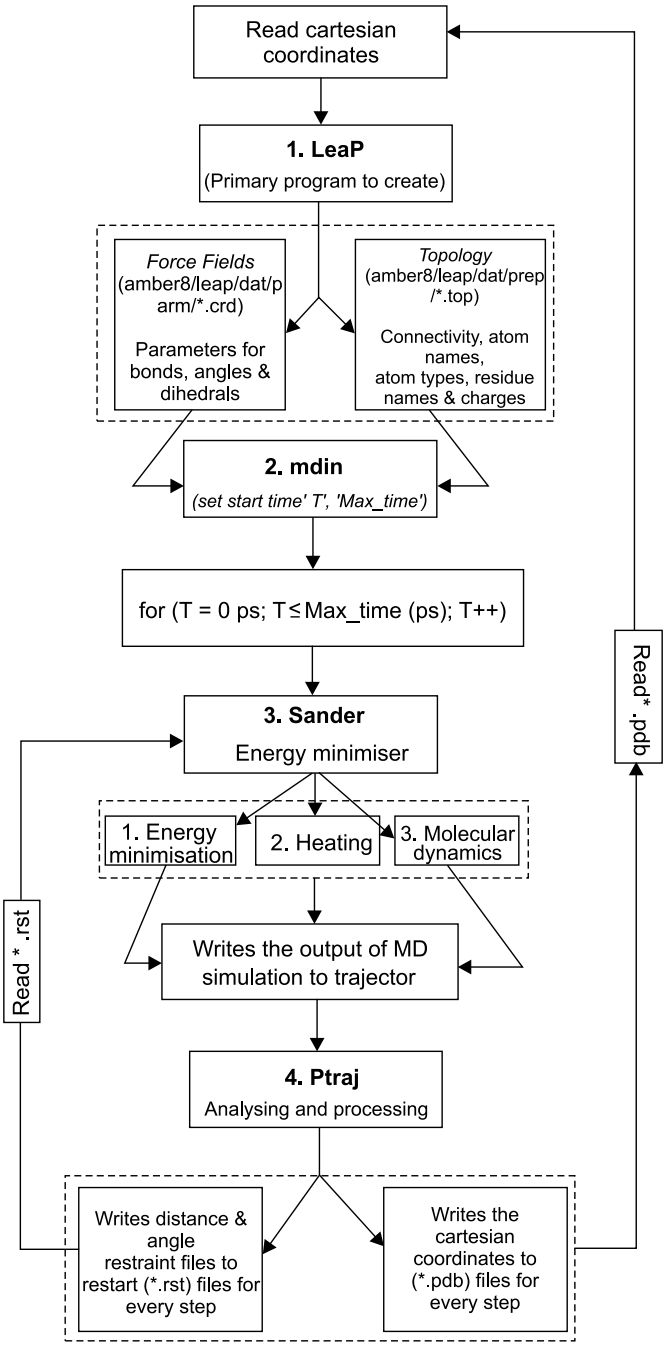

Figure 18 *Flowchart of data flow in AMBER*

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 in found in the b-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 the structure has reached a stable average configuration. Many of the features are common to both the plot and for much for the structures the values are well matched (Figure 19).

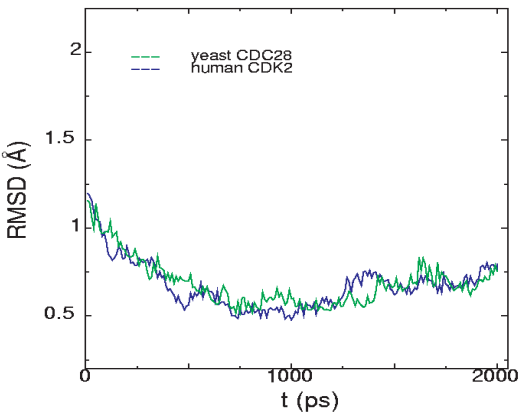

Figure 19 *Averaged RMSD (root mean square deviations) and for crystal structure 1QMZ and the modelled structure of the yeast cyclin dependent kinase CDC2 8 vs time is calculated.*

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

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

Δr_i = atomic displacement from average position, N = Total number of structures.

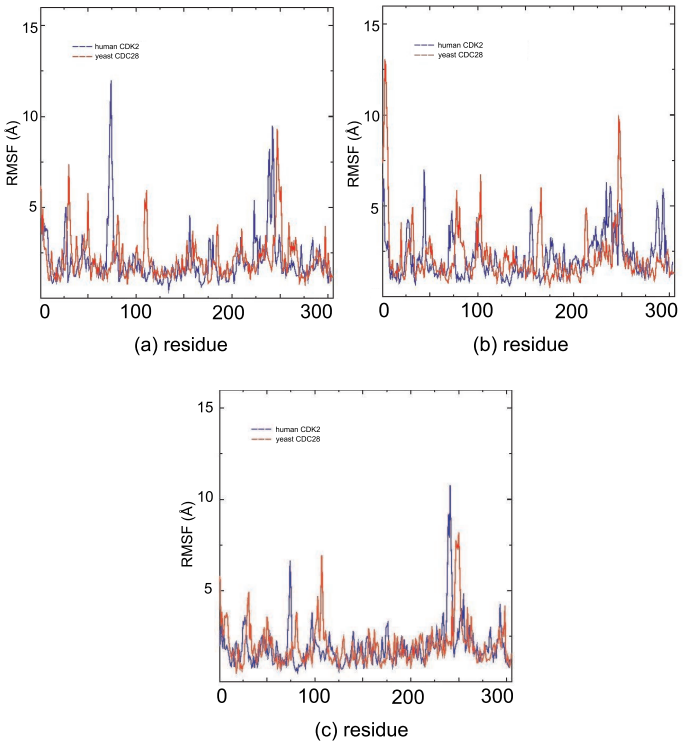

Figure 20 *Various behaviours of root mean square fluctuation (RMSF) for (a) 15 ps (b) 1 ns and (c) 2 ns for CDK2 and CDC28 respectively*

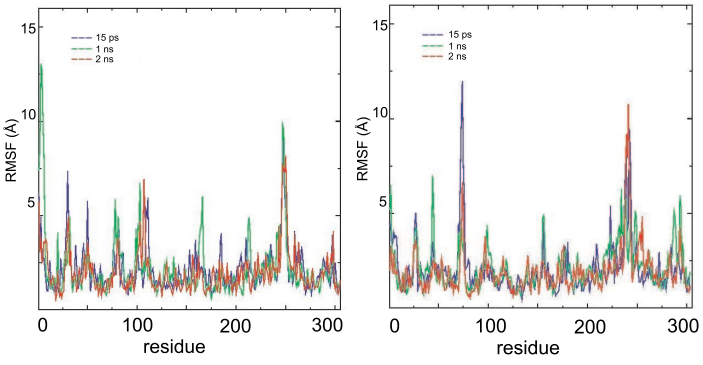

Figure 21 *Summary of various behaviours of root mean square fluctuation (RMSF) for (a) 15 ps (b) 1 ns and (c) 2 ns for CDK2 (left) and CDC28 (right) respectively*

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, 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 modelled yeast CDC28 the residues 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) (Figures 20 and 21).

Dynamic cross correlation map: The dynamic characteristics of the protein in a MD simulation can be analysed 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 domain−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

$$\Delta C_{ij} = < \Delta r_i \Delta r_j > / \sqrt{< \Delta r_i^2 >< \Delta r_j^2}$$

where Δr_i is the displacement of the mean position of the ith atom. The elements of C_{ij} can be collected as in matrix form and displayed as three-dimensional cross correlation matrix (DCCM) map. The C_{ij} are computed as averages over the successive backbone of N, Ca 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. Regions of regular secondary structures are expected to move in concert.

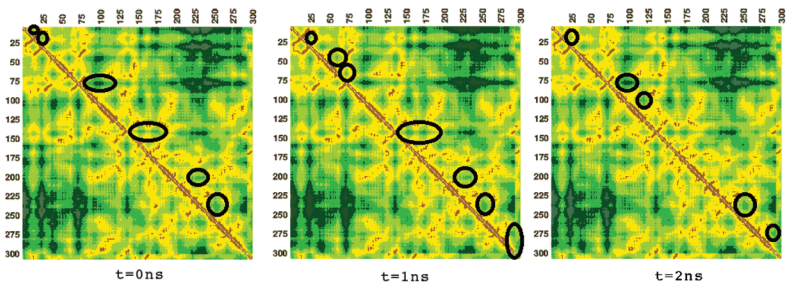

Figure 22 *Dynamic cross correlation map (DCCM) for 1QMZ structure for t=0, t=1 ns and t=2 ns. The black circles are the regions of the b-sheets showing cross peaks.*

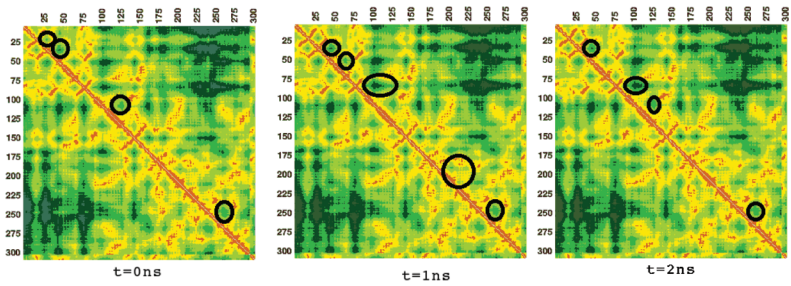

Figure 23 *Dynamic cross correlation map (DCCM) for yeast CDC28 structure for t=0, t=1 ns and t=2 ns. The black circles are the regions of b-sheets showing cross peaks.*

The DCCM map for each structure of 1QMZ and CDC28 was plotted over the time (15 ps, 1 ns, 2 ns) respectively (Figures 22 and 23). 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 interaction of non contiguous residues which fold to form the parallel beta 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 analysed. 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 use to estimate a cleft width, which is very important for localisation of ATP. G16 and T160 can serve as marker of G-loop and T-loop, respectively.

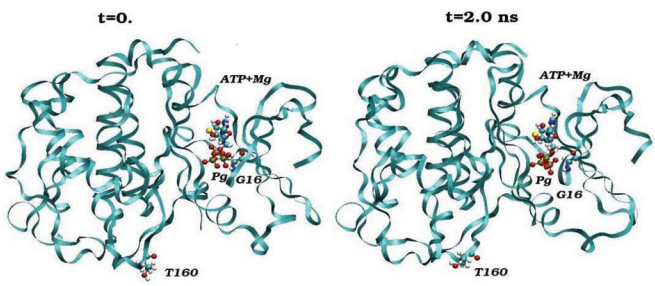

Figure 24 *The initial (a) and final (b) (2-ns state) structures of the CDK2/ATP complex. The ATP molecule and residue 16 of the G-loop are represented by the balls model.*

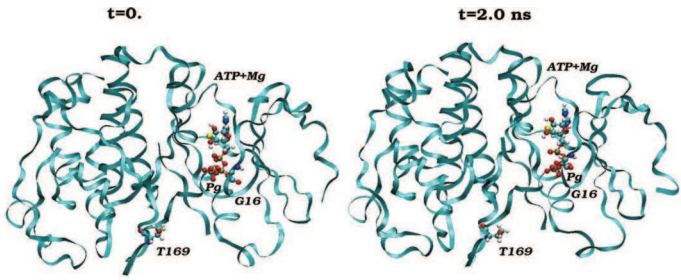

Figure 25 *The initial (a) and final (b) (1-ns state) structures of the CDC28/ATP. The ATP molecule and residue 16 of the G-loop are represented by the balls model.*

Simulated CDK2 structure (Figures 24 and 25) was compared to the CDC28 after conformational changes evaluation. The resulting wild-type CDK2 and CDC28 structural conformations are shown. The picture displays the initial (left) and the final 2-ns (right) states. Positional changes between the ATP, residue G16 in G-loop and T160 in T-loop (the latter covers a left bottom α-helix shown in Figures 26*a* and 26*b*. Comparing initial and final states of CDK2 structure and CDC28 structure, no big difference was visually observed. So, for the protein structures the original state is kept in a relatively stable conformation.

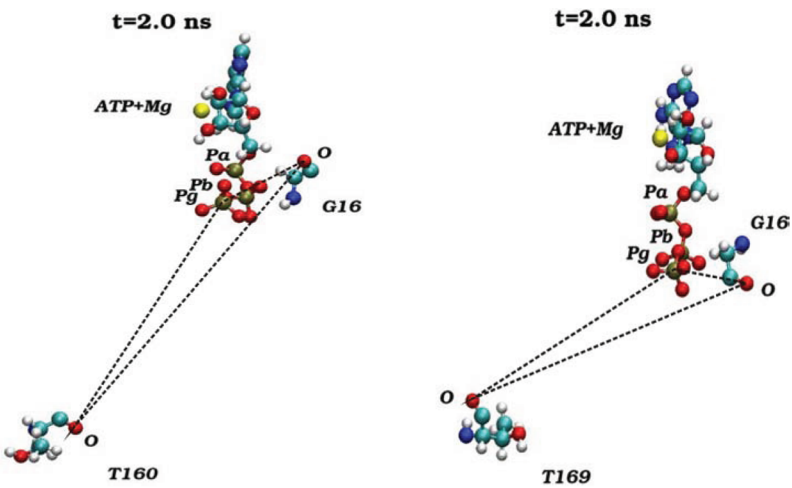

Figure 26(a) *The relative positions of the T160, ATP and res16 (an 'activation triangle') are shown for CDK2. (b). The relative positions of the T169, ATP and res16 (an activation triangle) are shown for CDC28. The ATP molecule, residues T160 and 16 are represented by the balls model.*

An activation triangle around ATP: The T160, ATP and G16 positions (an 'activation triangle') of CDK2 structure in the final (2-ns) state are represented in Figures 23−24 aiming to estimate (although indirectly) the possibility of the hydrogen bond formation in the ATP and G-loop region.

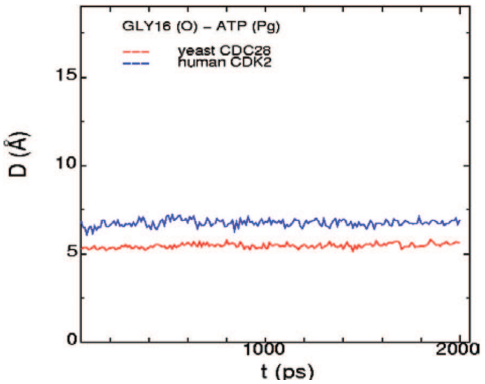

Figure 27 *The time dependence of the res16-ATP distance are shown for the CDK2 and CDC28, respectively, in accordance to the activation triangle.*

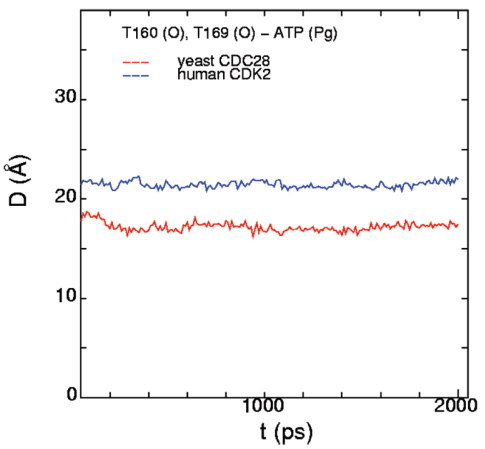

Figure 28 *The time dependence of the T160-ATP, T169-ATP distances are shown for the CDK2 and CDC28, respectively, in accordance to the activation triangle.*

Similarly, the T169, ATP and G20 positions (an activation triangle) of CDC28 structure in the final (2-ns) state are represented in Figures 24*b* and 25*b*, respectively. The ATP-res16 and ATP-res20 distance for the CDK2 and CDC28 structures, respectively, shows a completely different behavior (Figure 29). The ATP-res16 and ATP-res20 distance in the CDK2 and CDC28 structures, respectively, evidently lies within ~5.0 Å and ~5.5

Å during the all 2-ns dynamical changes. Thus, the all hydrogen bond network in the ATP-res20 for the binding site varies between the CDK2 and CDC28 structure.

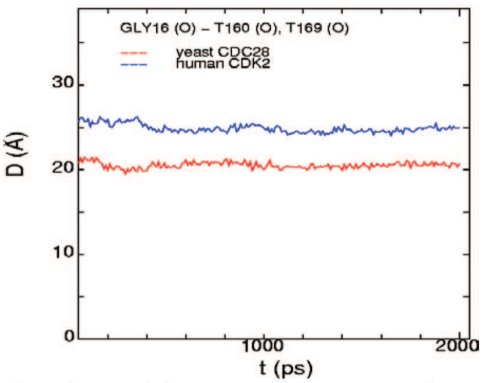

Figure 29 *The time dependence of the T160-res16, T169-res16 distances are shown for the CDK2 and CDC28, respectively, in accordance to the activation triangle.*

The amino acid residues around phosphorylated regulatory site: The CDK2, CDC28/ATP' dynamical peculiarities in the neighbour of phosphorylation site (T160 in CDK2, T169 in CDC28) were analysed in detail, showing by snapshots and animation movies 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 ones (Figure 29).

8 Discussion

The word bioinformatics has several slightly different meanings, but primarily it means the analyses of nucleic acid and protein sequences and structures. The reductionist approach to biology seeks to divide a biological system into several subsystems, understand the subsystems in detail, and then put them together to derive an understanding of the entire system. Information pathways are important in biological systems. Computers and information technology is are required to make sense of all this information. Being based on the manipulation of digital data, computer technology is well suited to deal with DNA and protein sequences. Bioin-

formatics deals with a variety of subjects that include pattern recognition, sequence and structure comparisons, statistical analysis of sequences, and constructing phylogenetic trees.

Sequences and structures are only one among the several different types of data that are required in the practice of modern molecular biology. Other important data types include metabolic pathways and molecular interactions, mutations and polymorphisms in molecular sequences and structures as well as in organelle structures and tissue types, genetic maps, physicochemical data, gene expression profiles, two dimensional DNA chip images of mRNA expression, two dimensional gel electrophoresis images of protein expression, data on restriction nucleases, nomenclature, video images and networks of biological data reflecting their inter-relationships. The primary nucleotide sequence database is the trio GenBank/EMBL/DDBJ. The arrangement of data in each entry in the GenBank is self-explanatory. Besides the DNA sequence, it contains extensive annotations that give all details regarding the sequence. The primary proteins sequence database is the duo SWISS-PROT/PIR. Again each protein sequence entry in the database is accompanied by extensive annotation. Derived or secondary databases may be divided into specialized subcollections of data and collections of information gathered by analysis of the primary databases. Examples of subcollections of nucleotide sequences include the FlyBase, the Comprehensive Microbial Resource, the Eukaryotic Promoter Database, etc. Some of these also store information on patterns obtained by analysis of the sequences, such as the consensus regions in the promoter sequences. Patterns, signatures or motifs of protein sequences can be found in databases such as PROSITE, PRINTS, BLOCKS, Pfam, etc., where they are stored as alignments, regular expressions, 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. Structures are stored as a list of the three-dimensional Cartesian coordinates of each atom in the molecule. Derived databases of protein motifs and patterns include SCOP and CATH. These databases cluster protein structures together in an increasingly distant hierarchy of structural similarity. Such databases are immensely useful in identifying the function of a newly sequenced protein.

One of chief tasks of bioinformatics is to compare DNA and protein sequences and find similarities, or differences, and infer structural, functional or evolutionary relationships. Sequence comparison and alignment is not a trivial task. Any written language may be analysed by almost exactly the same methods used to analyse DNA and protein sequences. Many alignments are possible between any two sequences. To decide which of the alignments is the best, we need a function that helps us to find some figure of merit or score for each alignment. Brute force or trial and error approaches to sequence alignment lead to combinatorial explosion. We say

that two sequences are similar to each other when, after the best alignment, identical (or similar) residues occur at identical (or similar) positions. Similarity could arise by chance, or it could be a convergence towards a common sequence and structure and therefore function, through evolution, or, the similarity could arise from divergent evolution of the two sequences from a common ancestral sequence. The Neeedleman-Wunsch algorithm is a global alignment method, while the Smith-Waterman method is a local alignment method. BLAST is most frequently used over the Internet on the BLAST server (http://www.ncbi.nlm.nih.gov/BLAST/).

A multiple sequence alignment may be defined as a two-dimensional table in which each row represents a protein or nucleic acid sequence, and the columns are the individual residue positions. One of the common goals of building multiple sequence alignments is to characterise protein and/or gene families, and identify shared regions of homology. Multiple sequence alignments may be represented by consensus sequences or profiles, or the entire alignment may be shown with colors or shade to highlight different features of the alignment. The most common way of finding the score of any given multiple sequence alignment is the so-called sum-of-pairs score.

The use of dynamic programming algorithms for multiple sequence alignment is, in practice, not possible for more than three or four sequences, because the computational cost increases exponentially with the number of dimensions, i.e., the number of sequences. The Feng-Dolittle algorithm is a progressive alignment technique that uses a distance matrix to construct a crude phylogenetic tree of relationships between the sequences to guide the alignment. Based on the tree, sequences are progressively added to the multiple sequence alignment, until all sequences are a part of it. 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 matrices and the BLOSUM matrices. PAM (Percentage Accepted Mutation) matrices are based on a Markovian model of evolutionary change in the sequences. Each site, i.e., residue, in the sequence is considered to evolve independently of the other sites.

Structure prediction is carried out at the level of secondary and tertiary structure. PHD is currently one of the most successful secondary structure prediction programs. It uses artificial neural networks to carry out the predictions. Best prediction accuracies are currently at the level of 85 to 90%. Methods to predict the tertiary structure of proteins may be divided into three broad categories—homology modelling, threading and ab initio methods. Homology modelling (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 modelling, and aligns the unknown sequence to a likely structure, which may

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. It is also necessary to calculate the Ramachandran map, which is an important tool to analyse protein structures. Further, we need to recognise and specify the secondary structures in the molecule. 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. Finally biological and biochemical principles are considered. The model should of course satisfy all such known concepts. 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 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.

Conformations and interactions of biomolecules can be most rigorously studied using quantum mechanical methods. These methods solve for the electronic structure of molecules and thus derive the effective Born-Oppenheimer potential for nuclear motion from first principles. However, none of these quantum mechanical methods are currently able to furnish results for larger biomolecules because the calculations on such systems are either time consuming, or rather inaccurate, when carried out at an approximate level of quantum mechanical theory. Instead, force field methods that ignore the electronic motions are used to calculate the energy of systems as a function of the nuclear positions only. Molecular dynamics study on the existing crystal structure and high similarity predicted structures were analysed. RMSD values are considered as reliable indicators of variability when applied to very similar proteins, like alternative conformations of the same protein. In other words, RMSD is a good indicator for structural identity, but less so for structural divergence. Root mean square fluctuation (RMSF) at time t of atoms in a molecule with respect to the average structure is defined as the atomic displacement from average position showed good results. Regions known to be of greater physiological importance including the structural conformations of protein kinases, the activation loops around ATP's and the amino acid residues inside the phosphorylated regulatory sites were described.

9 Conclusion

Information retrieval is important in various biomedical research fields. This chapter covers the theoretical background and 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 chapter also briefly describes the 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 may be similar to the input query sequence. One may 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 dynamic 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.

10 References

1. Crick, Francis. (1970). Central Dogma of Molecular Biology. *Nature* 227: 561—563.

2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman D. J. (1990). Basic local alignment search tool. *J Mol. Biol.* 215: 403—10.

3. Christie, Karen R. (2004). Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.* January 1; 32 (Database issue): D311—D314.

4. Kanehisa, Minoru. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes *Nucleic Acids Res.* January 1; 28(1): 27—30.

5. Huerta, Araceli M. (1997). RegulonDB: a database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res.* 55—60.

6. Hamosh, Ada. (2002). Online Mendelian Inheritance in Man (OMIM), a knowledge base of human genes and genetic disorders. *Nucleic Acids Res.* 30 (1): 52—55.

7. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29 (1):308—311.

8. Tawabata, T. (1998). The Protein Mutant Database. *Nucleic Acids Res.* 355—357.

9. Pruitt, Kim D. (2005). NCBI Reference Sequence (RefSeq): a curated non—redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 33 (Database issue): D501—D504.

10. Kanz, Carola. (2005). The EMBL nucleotide sequence database: *Nucleic Acids Res.* January 1; 33 (Database Issue): D29—D33.

11. Miyazaki, S. (2004). DDBJ in the stream of various biological data. *Nucleic Acids Res.* January 1; 32 (Database issue): D31—D34.

12. Wu, C. H., Yeh, Huang L.S, Arminski H., Castro-Alvear L. J., Chen, Y., Hu Kourtesis Z., Ledley, P, Suzek R. S., Vinayaka, B. E, Zhang, C. R. (2003). The Protein Information Resource. *Nucleic Acids Res.* Jan 1; 31(1): 345—7.

13. Bairoch, Amos. (2000). The SWISS—PROT protein sequence database and its supplement TrEMBL in 2000. Oxford University Press. *Nucleic Acids Res.* 28(1): 45—48.

14. The FlyBase Consortium. (2002). The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res.* January 1; 30 (1): 106—108.

15. Perier, Rouaida Cavin. (1998). The Eukaryotic Promoter Database (EPD): Recent developments. *Nucleic Acids Res.* 307—309.

16. Hulo, Nicolas. (2006). The PROSITE database. *Nucleic Acids Research* 34 (Database issue): 227—230.

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

18. Finn, Robert D. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res.* 34 (Database issue): 247—251.

19. Kouranov, A. (2006). The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.* 34 (Database issue): D302—D305.

20. Murzin. SCOP—Structural Classification of Proteins. *NAR Molecular Biology Database Collection entry number 282*.

21. Pearl, F. M. CATH. *NAR Molecular Biology Database Collection entry number 258*.

22. Mullan, L. (2006). Pairwise sequence alignment—it's all about us. *Briefings in Bioinformatics* 7(1): 113—115.

23. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. Gapped BLAST and PSI—BLAST. (1997). A new generation of protein database search programs. *Nucleic Acids Res.* Sep 1; 25(17): 3389—3402.

24. Zhang, Zheng. (1998). Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res*, 26(17): 3986—3990.

25. Oliver, T. (2005). Using reconfigurable hardware to accelerate multiple sequence alignment with ClustalW. *Bioinformatics* 21(16): 3431—3432.

26. Feng, D. F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol.* 25: 351—360.

27. Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucleic Acids Res.* 22: 4673—4680.

28. Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* 89:10915—10919.

29. Hansen, L. K. and Salamon, P. (1990). Neural Network Ensembles. *IEEE Trans. Pattern Anal. Machine Intel.* 12: 993—1001.

30. Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile based neural networks. *Meth. Enzymol.* 266: 525—539.

31. Lovell, S. C., Davis, I. W., Arendall, W. B., III, de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. and Richardson, D. C. (2003). Structure validation by CFY geometry and Cb deviation. *Proteins* 50: 437—450.

32. Panchenko, A., Marchler—Bauer, A. and Bryant, S. H. (1999). Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins* (Suppl 3): 133—140.

33. Ortiz, A. R., Kolinski, A., Rotkiewicz, P., Ilkowski, B. and Skolnick, J. (1999). Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins* (Suppl 3): 177—185.

34. Allen, M. P and Tildesley, D. J. (1987). *Computer Simulation of Liquids*. Oxford University Press, New York.

35. Haile, J. M. (1992). *Molecular Dynamics Simulations: Elementary Methods*. Wiley, New York.

36. Van Gunsteren, W. and Weiner, P. (eds.) (1989) and Van Gunsteren, W., Weiner, P., and Wilkinson, A. T., (eds.) (1993, 1996). *Computer Simulation of Biomolecular Systems: Theoretical and Experimentational Applications*. 1,2,3. ESCOM, Leiden, The Netherlands.

37. Gould, H. and Tobochnik, J. (1988). *An Introduction to Computer Simulation Methods: Applications to Physical Systems, Part 1 and 2*. Addison—Wesley, Reading, MA.

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

39. Brooks, C. L., Karplus, M. and Pettitt, B. M. (1988). *A Theoretical Perspective of Dynamics, Structure, and Thermodynamics.* Wiley Interscience, New York.

40. Oliver, Stephen G. (1996). From DNA sequence to biological function. *Nature* 379: 597 — 600

41. Koonin, E. V. and Mushegian, A. R. (1996). Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. *Curr Opin Gen Dev.* 6:757—762.

42. Dujon, B. (1996). The yeast genome project: what did we learn? *Trends Genet.* 12: 263—270.

43. Orengo, C. A., Jones, D. T. and Thornton, J. M. (1994). Protein domain superfolds and superfamilies. *Nature* 372: 631—634.

44. Sanchez, R. and Sali, A. (1998). Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc. Natl Acad. Sci. USA* 95: 13597—13602.

45. Tramontano, A., Leplae, R. and Morea, V. (2001) Analysis and assessment of comparative modeling predictions in CASP4. *Proteins* 45 (Suppl. 5): 22—38.

46. Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F. and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29: 291—325.

47. Guex, N., Diemand, A. and Peitsch, M.C. (1999). Protein modelling for all. *Trends Biochem. Sci.* 24: 364—367.

48. Fischer, D. and Eisenberg, D. (1997). Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc. Natl Acad. Sci. USA* 94: 11929—11934.

49. Sanchez, R. and Sali, A. (1998). Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc. Natl Acad. Sci. USA* 95: 13597—13602.

50. Rychlewski, L., Zhang, B. and Godzik, A. (1998). Fold and function predictions for *Mycoplasma genitalium* proteins. *Fold. Des.* 3: 229—238.

51. Huynen, M., Doerks, T., Eisenhaber, F., Orengo,C., Sunyaev, S., Yuan, Y. and Bork, P. (1998). Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.* 280: 323—326.

52. Grandori, R. (1998). Systematic fold recognition analysis of the sequences encoded by the genome of *Mycoplasma pneumoniae*. *Protein Eng.* 11: 1129—1135.

53. Teichmann, S. A., Park, J. and Chothia, C. (1998). Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements *Proc. Natl Acad. Sci. USA* 22: 14658—14663.

54. Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* 287: 797—815.

55. Peitsch, M. C. and Jongeneel, C. V. (1993). A 3-D model for the CD40 ligand predicts that it is a compact trimer similar to the tumor necrosis factors. *Intern. Immunol.* 5: 233—238.

56. Peitsch, M. C. (1995). Protein modelling by e-mail. *BioTechnology* 13: 658—660.

57. Peitsch, M. C. (1996). ProMod and Swiss—Model: Internet-based tools for automated comparative protein modelling. *Biochem. Soc. Trans.* 24: 274—279.

58. Sali, A. and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234: 779—815.

59. MacKerell, A. D. J., Bashford, D., Bellott, R. L., Dunbrack R. L., Jr., Evanseck J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S. et al. (1998). All-Atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* 102: 3586—3616.

60. Sali, A. and Overington, J. P. (1994). Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.* 3: 1582—1596.

61. Marti-Renom, M. A., Ilyin, V. A. and Sali, A. (2001). DBAli: a database of protein structure alignments. *Bioinformatics* 17: 746—747.

62. Sali, A. and Blundell, T. L. (1990). 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.* 212: 403—428.

63. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI−BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389−3402.

64. Sadreyev, R. and Grishin, N. (2003). COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.* 326: 317−336.

65. Sanchez, R. and Sali, A. (1997). Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proteins* 1**:** 50−58.

66. Melo F., Sanchez R. and Sali A. (2002). Statistical potentials for fold assessment. *Protein Sci.* 11: 430−448.

67. MacKerell A.D. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* 102:3586−3616.

68. Braun W. (1985). Calculation of protein conformations by proton−proton distance constraints: A new efficient algorithm. *J. Mol. Biol.* 186:611−626.

69. Sali, A. (1995). Modelling mutations and homologous proteins. *Curr. Opin. Biotech.* 6: 437−451.

70. Vasquez, M. (1996). Modeling side−chain conformation. *Curr. Opin. Str. Biol.* 6: 217−221.

71. Thornton, J.M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.* 151: 261−287.

72. Jung, S. H, Pastan, I. and Lee, B. (1994). Design of interchain disulfide bonds in the framework region of the Fv fragment of the monoclonal antibody B3. *Proteins* 19: 35−47.

73. Sali, A. and Overington, J. P. (1994). Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.* 3: 1582−1596.

74. Goffeau A., Barrell B. G., Bussey H., Davis R. W., Dujon B., Feldman H., Galibert F., Hoheisel, J. D., Jacq C., Johnston M. et al. (1996). Life with 6000 genes. *Science* 274: 563−567.

75. Bratley, P., Fox, B. L. and Schrage, L. E. (1987). *A Guide to Simulation*. Springer-Verlag, StateNew York.

76. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28: 235−242.

77. Shindyalov, I. N. and Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* 11: 739−747.

78. Case, D. A., Pearlman, D. A., Caldwell, J. W., Cheatham III T. E., Ross, W. S., Simmerling, C. L., Darden, T. A., Merz, K. M., Stanton, R. V., Cheng, A. L., Vincent, J. J., Crowley, M., Ferguson, D. M., Radmer, R. J., Seibel, G. L., Singh, U. C., Weiner, P. K., Kollman, P. A. (2003). *AMBER 8.0*. University of California.

79. Okimoto, N., Yamanaka, K., Suenaga, A., Hirano, Y., Futatsugi, N., Narumi, T., Yasuoka, K., Susukita, R., Koishi, T., Furusawa, H., Kawai, A., Hata, M., Hoshino, T., Ebisuzaki, T. (2003). *Chem-Bio Informatics J*. 3 (1): 1—11.

80. Narumi, T., Susukita, R., Furusawa, H., Yasuoka, K., Kawai, A., Koishi, T., Ebisuzaki, T. (2000). *MDM version of AMBER*.

81. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, Jr. 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. *J. Am. Chem. Soc*. 117: 5179—5197.

82. Jorgensen, W. L., Chandrasekhar, J. and Madura, J. D. (1983). Comparison of simple potential functions for simulating liquids. *J. Chem. Phys.* 79 (2): 926—935.

83. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., Haak, J. R. (1984). Molecular dynamics with coupling to external bath. *J. Chem. Phys.* 81 (8): 3684—3690.

84. Ryckaert, J. P., Ciccotti, G., Berendsen, H. J. C. (1997). Numerical integration of the cartesian equations of a system with constraints. *J. Comput. Phys*. 23: 327—341.

85. Kholmurodov, Kh. (2005). *PEPAN (Physics of Particles and Nuclei)* 36 (2): 1—16.

86. Kholmurodov, Kh, Hirano, Y., Ebisuzaki, E. (2003). *Chem-Bio Informatics J*. 3 (2): 86—95.

87. Kholmurodov Kh. T. et al. (2003). Methods of Molecular Dynamics for Simulation of Physical and Biological Processes. *PEPAN (Physics of Elementary Particles and Atomic Nuclei)* 34 (2): 474—501.

88. Sayle, R. A., Milner-White, E. J. (1995). RasMol: Biomolecular graphics for all. *Trends in Biochem. Sci.* 20: 374—376.

89. Koradi, R., Billeter, M., Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structure. *J. Mol. Graphics* 4: 51—55.

90. Mendenhall, M. D. and Hodge, A. E. (1998). Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast. *Microbiol. Mol. Biol. Rev.* 62: 1191—1243.

91. De Bondt, H. L., Rosenblah, J., Jancarik, J., Jones, H. D., Morgan, D. O., Kim, S. H. (1993). Crystal structure of cyclin-dependent kinase 2. *Nature* 363: 595—602.

92. Jeffrey, P. D., Russo, A., Polyak, K., Gibbs, E., Hurwitz, J., Massague, J. and Paoletich, N. P. (1995). Mechanism of CDK activation revealed by the structure of a cyclin A-CDK2 complex. *Nature* 373: 313—320.

93. Koltovaya, N. A., Guerasimova, A. S., Kretov, D. A., Kholmurodov, Kh. T. (2006). *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*. Nova Science Publishers.

94. Kretov, D. A., Kholmurodov, Kh. T., Koltovaya, N. A. (2006). 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* 4.