Structural Bioinformatics: From the Sequence to Structure and Function

Article *in* Current Bioinformatics · January 2009

DOI: 10.2174/157489309787158170

CITATIONS
12

READS
1,678

1 author:

Marco Wiltgen
Medical University of Graz
73 PUBLICATIONS **618** CITATIONS

SEE PROFILE

Structural Bioinformatics: From the Sequence to Structure and Function

Marco Wiltgen*

Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria

Abstract: Proteins are the molecules of life which are involved in cellular processes. The functional specificity of a protein is linked to its structure. A great section of bioinformatics deals with the prediction, analysis and visualization of protein 3D structures. High-throughput methods for the determination of protein structures provide the information needed to build structure-activity relationships. The accessibility of these structural data together with genomic and clinical data is of crucial importance for the application of bioinformatics in medical research. The experimental methods are supplemented by homology modelling, where new protein structures are predicted by exploiting structural information from known configurations. Computer visualization of protein models provide insights into biological processes which can not be adequately explained otherwise. For the analysis of protein-protein interactions, Voronoi tessellations are used to quantify the macromolecular interfaces. Details at the atomic and electronic levels of the protein molecules, needed for a deeper understanding of properties that remain unrevealed after structural elucidation, are provided by methods based on quantum theoretical calculations. Many proteins are of immediate medical and pharmacological relevance. The structural analysis is therefore of special interest to understand diseases at a molecular level, which is the prerequisite for new developments in diagnosis and therapy.

Keywords: Protein structures, protein data base, protein visualization, homology modelling, voronoi tessellation, density functional theory, molecular medicine.

1. INTRODUCTION

Proteins are not only of fundamental biological importance, they also play an important role in many diseases. Because the function of a protein is linked to its structure, the structural analysis is of great importance for understanding pathological processes in order to provide new ways in diagnostic and therapeutic medicine [1]. Therefore structural bioinformatics has a great impact to biomedical science and drug discovery [2]. The focus of this paper lies in molecular medicine. The influence of structural bioinformatics to clinical research is highly dependent on the accessibility of genomic, proteomic and clinical data and the development of methods for connecting this information with structural data [3]. The visualization of the structural data is of crucial importance. Procedures such as: inhibition and neutralization of enzymes; interaction of a hormone with its receptor and antigen-antibody interactions cannot be adequately explained without the optical advantage obtained with such visualization methods. The structural analysis is complemented by quantum theoretical methods providing details at the electronic level, for example during the binding of a substrate to an enzyme, that remain unrevealed after structural elucidation. Briefly, thanks to these attempts at molecular modelling and visualization the construction and synthesis of drug and inhibitor molecules are very much enhanced and help to determine new trends in the treatment of diseases.

In this introductory paper, an overview of several methods in structural bioinformatics is presented and explained. Starting from the experimental determination of protein structures, these include: structural classification of proteins; accessibility of structural data in databases; visualization of protein structures by considering sequence-structure and structure-function relationships; database aided identification of sequences; determination of protein structures by homology modelling; structural alignments and the analysis of protein-protein interactions. Special attention is given to quantum theoretical methods, which complement the structural methods by delivering insights into the dynamical behaviour of protein reactions. The methods are illustrated by means of selected proteins, which are of immediate medical relevance. The importance and influence of bioinformatics algorithms in medical research is accentuated. This paper also achieves a practical aspect by presenting a number of URLs, where information and tools suitable for structural analysis can be found and applied.

A lot of calculations and visualizations in this review paper have been done with the Swiss-PdbViewer (http://-spdbv.vital-it.ch), Voro3D program (Voronoi tessellations, http://www.lmcp.jussieu.fr/~mornon/voronoi.html), SWISS-MODEL and MODELLER (homology modelling, http://-swissmodel.expasy.org/, http://salilab.org/modeller/), HyperChem (quantum calculations, http://www.hyper.com) and several protein analysis software tools from the ExPASy server (http://au.expasy.org/tools/). The data files for the visualized protein structures were retrieved from the PDB database (http://www.rcsb.org/pdb).

2. PROTEIN BIOSYNTHESES

Proteins are used by the cell to read and translate the genomic information into other proteins for performing and controlling cellular processes such as: metabolism (decomposition and biosynthesis of molecules); physiological signalling; energy storage and conversion; formation of cellular

*Address correspondence to this author at the Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerplatz 2, A-8036 Graz, Austria; Tel: ++43 316/385-3587;
Fax: ++43 316/385-3590; E-mail: marco.wiltgen@meduni-graz.at

1574-8936/09 $55.00+.00

Amino acids	Quantity	%
Ala (A)	51	8.7
Arg (R)	24	4.1
Asn (N)	21	3.6
Asp (D)	32	5.5
Cys (C)	15	2.6
Gln (Q)	22	3.8
Glu (E)	34	5.8
Gly (G)	46	7.9
His (H)	15	2.6
Ile (I)	27	4.6
Leu (L)	55	9.4
Lys (K)	38	6.5
Met (M)	25	4.3
Phe (F)	26	4.4
Pro (P)	27	4.6
Ser (S)	34	5.8
Thr (T)	28	4.8
Trp (W)	12	2.1
Tyr (Y)	21	3.6
Val (V)	32	5.5

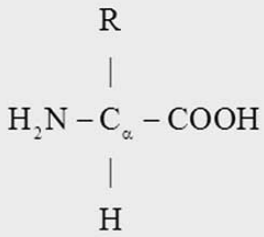


Fig. (1). Proteins are built up as polymers of up to 20 different amino acids. The table shows the percentage of the different amino acid types constituting the enzyme glutamic acid decarboxylase. Each amino acid consists of an amino group, a carboxyl group and a variable side chain (R), which determines the type of amino acid.

structures etc. Proteins are synthesized by means of the genetic code which is inherent through replication.

By considering the chemical composition of a protein, one obtains the, for organic molecules, typical chemical elements in different quantities as illustrated in the case of glutamic acid decarboxylase (GAD):

$C_{2925}H_{4532}N_{780}O_{844}S_{40}$

However, the chemical composition delivers little information about proteins. A better understanding of the protein molecules is obtained by considering their modular composition. Proteins contain amino acids as subunits. Amino acids consist of an amino group, a carboxyl group and a variable side chain (Fig. **1**). The proteins are built up as polymers of the amino acids. The amino acids are connected by a peptide bond between the basic amino group and the acidic carboxyl group of each amino acid in the polypeptide chain. 20 different amino acids are involved as residues in protein sequences. Each of the 20 amino acids differs by its side chain, which is responsible for the physicochemical properties such as: electric charge; polar, non polar; acidic; basic; hydrophobic etc. Some of the side chains have ring like structures (aromatic), while others consist of straight carbon chains (aliphatic). The different amino acids are represented in different quantities in a protein (Fig. **1**). Hypothetically, were one to mix the amino acids in the right quantities in a test tube, one would never obtain the desired protein. The protein and its structure are not only determined by the percentage of the different amino acids but also the information about the sequence of the amino acids in the polypeptide chain is necessary. The information encoded in the amino acid sequence determines the 3D structure of a protein and in consequence its function.

The protein biosyntheses is determined by the genetic code in the DNA (deoxyribonucleotid acid) through the intermediate RNA (ribonucleotid acid). During gene expression the DNA sequence of the gene is copied into messenger RNA (mRNA). The mRNA then serves as template for the protein biosyntheses (Fig. **2**). It contains a sequence of 4 nucleotides: adenine, cytosine, guanine and uracil, which determine the copied genetic code. To encode an amino acid,

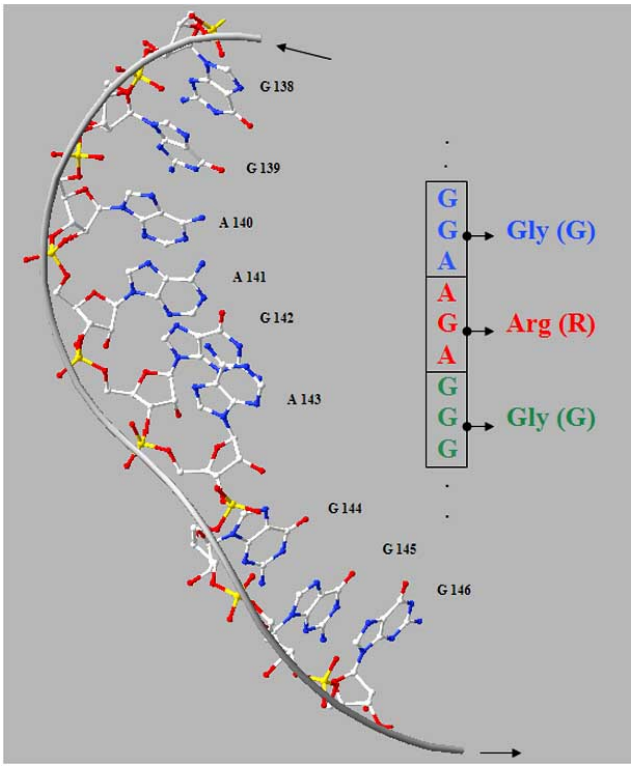


Fig. (2). Messenger RNA (mRNA) contains the information needed for protein biosyntheses. The information is determined by a sequence of 4 nucleotides: adenine (A), cytosine (C), guanine (G) and uracil (U). An amino acid is encoded by a triplet of nucleotides. This encoding is redundant in that sense, that one amino acid can be encoded by different triplets.

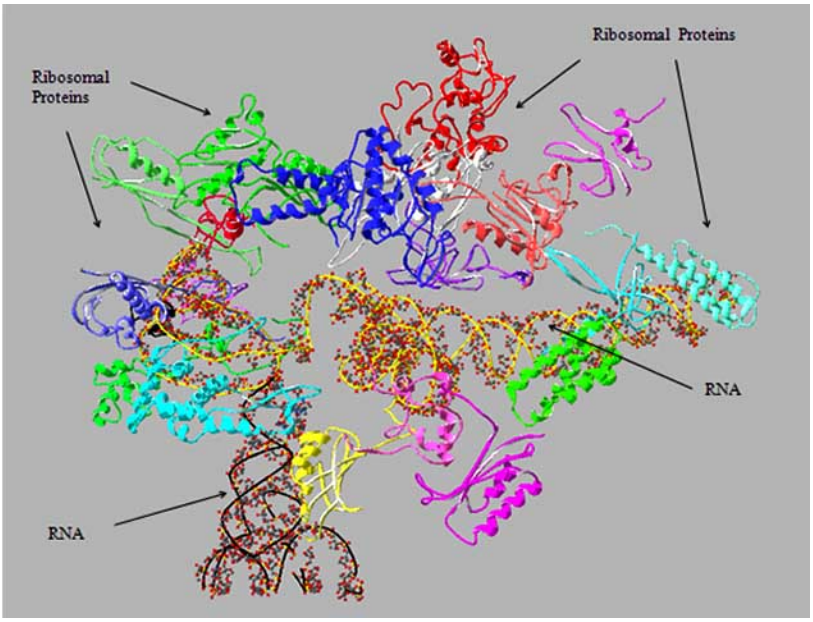


Fig. (3). Parts of the 3OS ribosomal subunit of the whole ribosomal complex with transfer RNA (tRNA) and mRNA (resolution 6.46 angstroms, R-value = 0.354). The ribosome is the biological unit in the cell where the protein biosyntheses takes place. It is a complex of different ribosomal proteins and ribonucleotid acids (RNA). The mRNA serves as template for the protein syntheses. Each tRNA bounds an amino acid. The amino acids are then connected according to the mRNA sequence.

3 nucleotides are combined in a triplet (codon). Because 64 possible combinations of the 4 nucleotides in triplets are encoding 20 different amino acids the code is redundant. The ribosome is the biological unit in the cell where the copy of the genetic code in the mRNA is translated into the amino acid sequence and the protein is build up (Fig. **3**) [4]. The ribosome is a complex of different ribosomal proteins and RNA. During the translation the condons of the mRNA are recognized by the complementary nucleotide triplets (anticondons) of the transfer RNA (tRNA) and a condon-anticondon bonding results. Every tRNA molecule transports an amino acid, where the type of the amino acid is dependent from the anticodon. During the interaction of the mRNA and the tRNA the amino acids are connected according to the encoded information in the mRNA sequence.

Once an amino acid sequence is synthesized in the cell, it folds together to a well defined and, for its sequence, unique 3D structure (Fig. **4**). It can be differentiated between the primary structure (the sequence of the residues), the secondary structure (α-helices, β-sheets and loops) and the tertiary structure (folding of the secondary structure elements into a three dimensional structure). The polypeptide chain forms secondary structures via hydrogen bonds (interaction between the hydrogen of the amino group and the oxygen of the carboxyl group) between amino acids in the chain. The secondary structure elements are then folded, through further

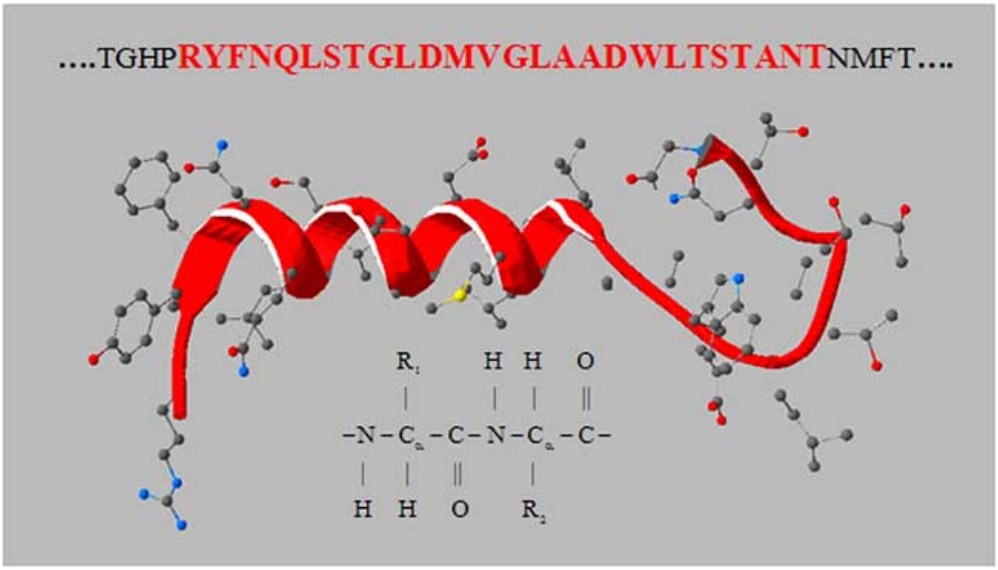


Fig. (4). The 3D structure of a protein is uniquely determined by its amino acid sequence. In the polypeptide chain the amino acids are connected via the respective carboxyl group and amino group.

interactions between the residues, into the three dimensional structure. The main driving forces in protein folding are hydrophobic and electrostatic interactions. Domains are compact regions of structure within the larger protein structure. Every domain has its own hydrophobic core and is connected to the rest of the protein via the polypeptide backbone. Normally a domain is collinear with a defined sequence part and the protein structure is divided into several domains. Many proteins are oligomeric, which means they are composed of more independent polypeptide chains (quaternary structure: three dimensional arrangements of the monomers). The individual polypeptide chains in the quaternary structure can interact with electrostatic forces and hydrogen bonds.

Functional specificity of a protein is linked to its structure. Due to the folding structure each of the critical residues responsible for the protein function is brought into a precise geometric arrangement. They are building the active site: a localized combination of amino acids within the tertiary structure that acts with other molecules and provides the protein with biological activity. Then interactions of a protein with other molecules are determined by residues which are close in 3D space but may be very distant in the amino acid sequence. The active site is then often found only in a small part of the structure and the rest of the protein structure is mainly necessary to enable and maintain the correct spatial position between the amino acids on the active site. Therefore, to understand a protein function, the 3D structure of the protein reveals far more information, than its sequence.

The knowledge of the 3D structure is essential to understand the function of proteins and the interactions between proteins within the organism resulting in metabolism, reproduction and form. Classes of proteins are: enzymes; hormones; regulators; signal receptors; ion channels; antibodies etc.

3. PROTEINS IN MEDICINE

Many protein structures are of immediate medical and pharmacological relevance. This is especially valuable for: human proteins; eukaryotic model organisms and human pathogen micro organisms. Among the great amount and diversity of human proteins the following are of special medical importance:

- Enzymes which are involved in drug and xenobiotica metabolism.
- Proteins involved in signal transduction processes (growing factors).
- Ion channels.

The detection of structural changes at the protein level will enable a better understanding of certain disorder processes and their reasons.

Cytochrome-P450

Cytochrome-P450 is a membrane associated enzyme involved in many processes such as: the transformation of xenobiotica; metabolism of drugs; biosynthesesis of steroid hormones [5,6]. In the mitochondrial matrices of the suprarenal gland, the enzyme cytochome-P450 catalyses the conversion of cholesterol into an initial stage (pregnenolon) of stereoid hormones (Fig. **5**).

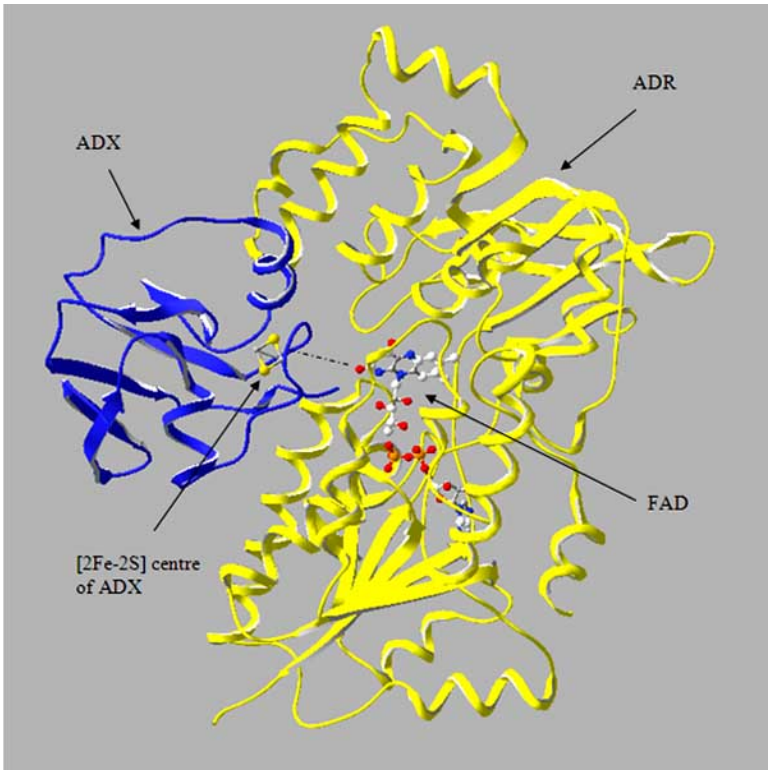


Fig. (5). Cytochrome-P450 is an enzyme involved in drug and xenobiotica metabolism. The cytochrome-P450 system is a complex (only two chains are visualized, resolution: 2.3 angstrom, R-value = 0.222) between adrenodoxin reductase (ADR) and adrenodoxin (ADX). Adrenodoxin reductase transfers an electron, needed for the hydroxilisation of the cholesterol side chains, via flavinadenindinucleotide (FAD) to adrenoxidin. Probably, the electron migrates along covalent bonds and hydrogen bonds.

The electrons, necessary for the hydroxilisation of the cholesterol side chains, are delivered to the enzyme by an electron transfer system [7]. Mutations in the cytochrom-P450 system may cause defects in drug metabolism. For example, antibiotics and chemo therapeutics are not decomposed because of the enzyme mutation, which leads to toxic damage or marrow insufficiency.

Growing Factors

Growing factors are extra cellular proteins with hormonal function, which regulate the gene expression by activating receptors localized in the cell membrane. The structure biological questions concern the specification of the ligands and the mechanism of the signal transduction. Four different classes of structures can be considered:

- Cystine node growing factor — Vascular endothelia growth factor
- 4-helices bundle structure — Erythropoietin
- Beta-clover-leaf — Fibroblast growing factor
- Chemochines — Interleukin-8

Growing factors are involved in the differentiation of stem cells into specialized cells such as: leucocytes; erythrocytes; fibroblasts etc.

Cystine node growing factors are building up as homo dimmers [8,9]. Every homomer contains 4 central β-sheets. Two sulphur bridges are building a kind of circle like structure, connecting sheets next to each other (Fig. **6**). A third sulphur bridge penetrates the circle structure and connects the remaining sheets. A representative of the class of cystine node growing factors is the vascular endothelial growth factor. Defects can cause an increase of endothelial cells. This can be the reason for tumours or brittle vessels. In the 4 helices bundle structure, 4 α-helices are arranged in parallel to each other (Fig. **6**). Representatives of this class of growing factors are erythropoietin (EPO), which regulates the differentiation of spinal stem cells into erythrocytes cells, and the leukaemia-inhibitory-factor (LIF) [10,11]. The beta-clover-leaf growing factors are monomers with a repetitive (3x) motive of β-sheets (Fig. **6**). A representative is the fibroblast growing factor (GF) [12,13]. This growing factor causes the differentiation of stem cells into fibroblast cells, which play a role in healing of wounds. Chemocines are homo dimer growing factors consisting of β-sheets and crosswise arranged α-helices (Fig. **6**). A representative is leucine, which plays a role in the differentiation of leucocytes [14,15]. Defects can cause leukaemia.

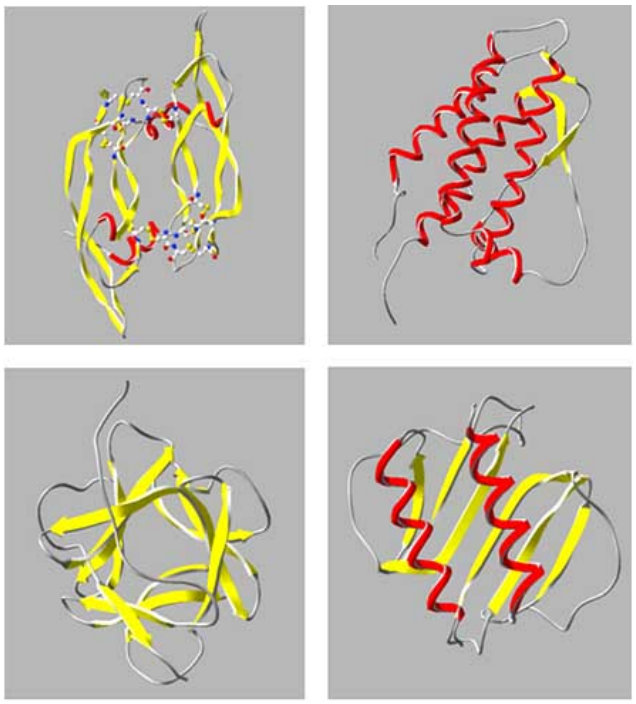


Fig. (6). Growing factors are extra cellular proteins with hormonal function. They are involved in the differentiation of stem cells into specialized cells. Top: Left picture: Cystine node growing factor. Right picture: 4 helices bundle structure. Bottom: Left picture: Beta-clover-leaf growing factor. Right picture: Chemocine.

The corresponding receptors are building up of an extra cellular part, a short transmembran segment and an intra cellular part [16]. Hormone-receptor interactions are of fundamental biological importance. For example: Depending on the conditions, interferon-γ stimulates or suppresses the production of antibodies (Fig. **7**). Oxytocin is a sexual hormone which bounds to its carrier protein neurophysin (Fig. **7**). The structure analysis is therefore of special interest for understanding these processes. Hormone-receptor interactions may also be involved in pathological processes.

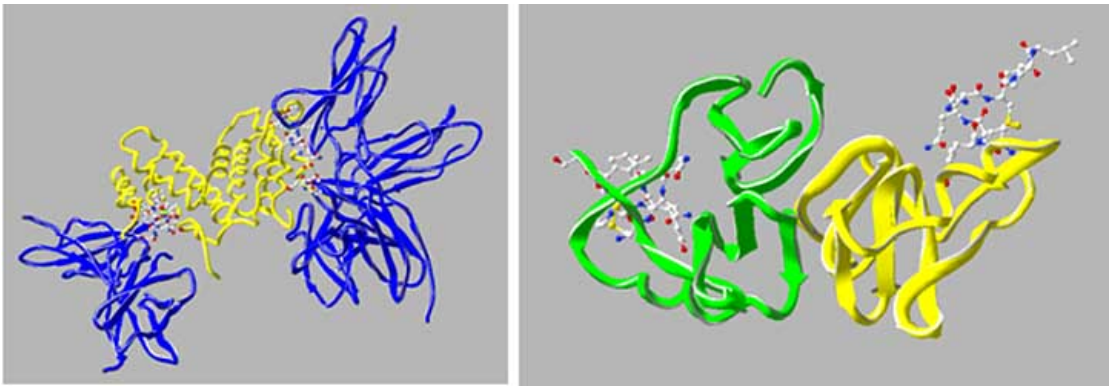


Fig. (7). Left: Interferon-γ (yellow) interacting with its receptor (blue). Depending on the conditions interferon-γ stimulates or suppresses the production of antibodies (resolution: 2.9 angstrom, R-value = 0.237). Right: Structure of neurophysin (ribbon) in complex with oxytocin (resolution: 3.0 angstrom, R-value = 0.182

Ion Channels

Progress has been made in understanding neurodegenerative diseases at a molecular level by studying ion channel proteins. Ion channels provide narrow, channel-like pathways used by ions to cross the cell membrane. The membrane is an electrical insulator and blocks the ions in the absence of ion channel proteins. Ion channels are ion selective. For each of the different ions, a different protein has to be present to cross the membrane. Some ion channels are responsible for action potentials in neurons. Action potentials are small voltage changes across the cell membrane in neurons and are basic signalling processes of the brain. Repetitive action potentials form firing patterns in individual nerve cells. The signalling mode of neurons is enabled due to a combination of three different classes of ion channels. Many of these channel proteins are involved in neurodegenerative diseases. Therefore the relationship between structure and function of ion channels are of special importance and has greatly improved the understanding of Alzheimer's, Parkinson's and Huntington's diseases [17-22].

The structure and function of an ion channel is illustrated in the case of aquaporin, a channel which enables water molecules to enter the cells (Fig. **8**) [23].

Fig. (8). Aquaporin is a membrane spanning protein that allows only water molecules to pass. Aquaporin consists of six hydrophobic α-helices which stretch through the cell membrane (resolution: 3.7 angstrom, R-value = 0.361). The key to the function of aquaporin is its hourglass-shaped pore with the smallest width of 3 angstrom. For the simulation of the aquaporin function, a water molecule has been added. The figure shows the hydrogen bonding (dotted lines) of the water molecule to two asparagines residues (at positions 76 and 192 in the sequence), which extend their side chains to form the narrowest part of the channel, allowing only water molecules to pass.

4. EXPERIMENTAL DETERMINATION OF PROTEIN STRUCTURES FROM DIFFRACTED X-RAYS

The atomic details of molecules are not visible using microscopes (the diffracted X-rays in crystallographic methods can not be focused by a lens). The protein structures can only be determined indirectly by special detection methods whereby the shapes of the molecules are derived from the experimental data with mathematical methods. We concentrate our attention on the determination of protein structures with crystallographic methods (alternative methods are the nuclear magnetic resonance spectroscopy and the cryo- electron microscopy).

In the X-ray crystallography, X-rays are reflected by the electron clouds surrounding the atoms in a protein crystal [24,25]. In a protein crystal, individual protein molecules are arranged in a regular lattice. The proteins are crystallized out by extraction of the soluble. For X-ray crystallography, sufficiently large and unflawed protein crystals are needed. Inside the crystals the molecules are still surrounded by water molecules. The X-ray reflections scattered from the protein crystal build up a regular diffraction pattern on a detector, which is analyzed to produce an electron density map of the protein (Fig. **9**). To this purpose the intensity and position of each diffracted X-ray on the detector are measured. The intensities are proportional to the square of the structure factor amplitude:

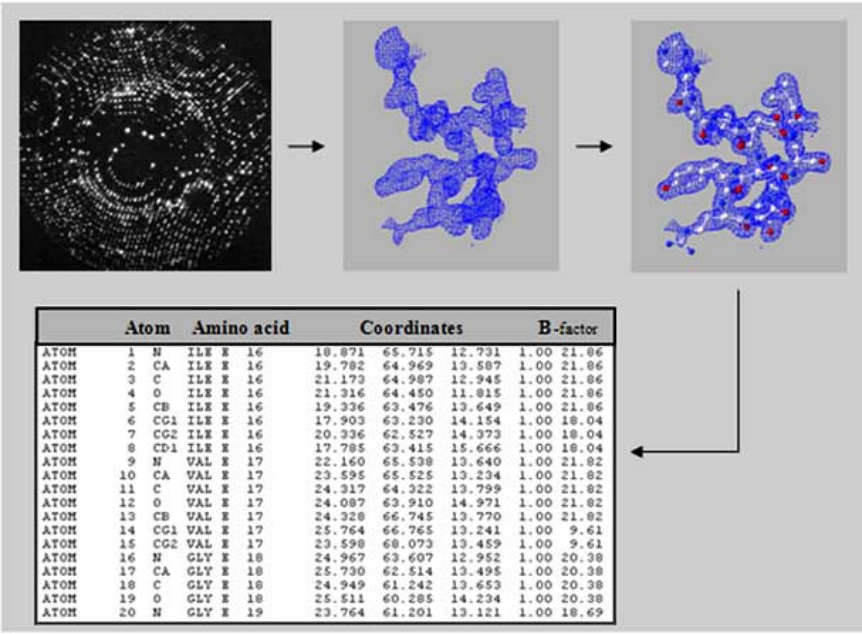
$$I_{hkl} \propto F^2_{h,k,l}$$

h,k,1 are the Miller indices. By Fourier transformation the electron density at every point of the protein crystal is calculated:

$$\rho_{xyz} = \frac{1}{V} \sum_h \sum_k \sum_l F_{hkl}\, e^{ia_{hkl}} e^{-2\pi i(hx+ky+lz)}$$

V is the volume of the crystal elementary cell. The volume and the structure factor amplitude can be determined experimentally. Unfortunately, the diffraction pattern shows no information about the relationship among the phases $a_{h,k,1}$ of all the scattered X-rays. Therefore in a first step the phases must be estimated, allowing the calculation of a fuzzy electron density map. Then the phases that will be shown on this rudimentary model are computed. From these phases and the original data from the diffraction pattern, a new electron density map, showing more detail, is calculated. This bootstrapping process (model building, calculating phases, calculating new electron density map, rebuilding) is repeated until the process converges, this means that the latest model gives the same phases as the previous one. Computer algorithms have been developed for the automatic generation, interpretation and refinement of electron density maps [26-29]. Protein atomic coordinates are determined by modelling the best possible way for the atoms to fit into the electron density map (Fig. **9**).

Electron density maps are available from the Electron Density Server: http://fsrv1.bmc.uu.se/eds/ at the Uppsala University and can be used to evaluate and ameliorate protein structures by optimal fitting of residues into the electron density map [30]. The determined protein structure is not an exact representation of the atomic positions in the crystal, but it is the model that best fits the electron density map of



	Atom		Amino acid			Coordinates			B-factor	
ATOM	1	N	ILE	E	16	18.871	65.715	12.731	1.00	21.86
ATOM	2	CA	ILE	E	16	19.782	64.969	13.587	1.00	21.86
ATOM	3	C	ILE	E	16	21.173	64.987	12.945	1.00	21.86
ATOM	4	O	ILE	E	16	21.316	64.450	11.815	1.00	21.86
ATOM	5	CB	ILE	E	16	19.336	63.476	13.649	1.00	21.86
ATOM	6	CG1	ILE	E	16	17.903	63.230	14.154	1.00	18.04
ATOM	7	CG2	ILE	E	16	20.336	62.527	14.373	1.00	18.04
ATOM	8	CD1	ILE	E	16	17.785	63.415	15.666	1.00	18.04
ATOM	9	N	VAL	E	17	22.160	65.538	13.640	1.00	21.82
ATOM	10	CA	VAL	E	17	23.595	65.525	13.234	1.00	21.82
ATOM	11	C	VAL	E	17	24.317	64.322	13.799	1.00	21.82
ATOM	12	O	VAL	E	17	24.087	63.910	14.971	1.00	21.82
ATOM	13	CB	VAL	E	17	24.328	66.745	13.770	1.00	21.82
ATOM	14	CG1	VAL	E	17	25.764	66.765	13.241	1.00	9.61
ATOM	15	CG2	VAL	E	17	23.598	68.073	13.459	1.00	9.61
ATOM	16	N	GLY	E	18	24.967	63.607	12.952	1.00	20.38
ATOM	17	CA	GLY	E	18	25.730	62.514	13.495	1.00	20.38
ATOM	18	C	GLY	E	18	24.949	61.242	13.653	1.00	20.38
ATOM	19	O	GLY	E	18	25.511	60.285	14.234	1.00	20.38
ATOM	20	N	GLY	E	19	23.764	61.201	13.121	1.00	18.69

Fig. (9). The X-ray reflections, scattered from the protein crystal, build up a regular diffraction pattern, characterized by the diffraction angles and intensities, on a detector. The reflections are analyzed to produce an electron density map of the protein. The electron density map defines surfaces with constant electron densities. The atomic coordinates are determined by modelling in the best possible way for the atom centres to fit into the electron density map.

the protein. To evaluate the quality of a protein structure, the structure factor amplitude F'_{hkl} is calculated from the electron density by inverse Fourier transformation. This calculated structure factor amplitude is compared with the experimentally determined structure factor amplitude. The so called R-factor indicates how well the model fits the original data:

$$R = \frac{\sum_h \sum_k \sum_l \left| F_{hkl} - F'_{hkl} \right|}{\sum_h \sum_k \sum_l F_{hkl}}$$

The R-factor should be as small as possible. For proteins, a desirable R-factor is 0.2 for a resolution of 2.5 angstrom. Another factor, the B-factor indicates the precision of each atom position. Atom positions can be uncertain due to motion (at temperatures above absolute zero, there is always thermal motion) of the atoms or alternative side-chain conformations. B-factor values of 5 and 20 correspond to uncertainties of 0.25 respectively 0.5 angstroms. The resolution is of special importance for the evaluation of a structure regarding to specific biological questions. Three different levels of structure can be distinguished:

- low resolution: few details are visible other than the shape of the molecule
- medium resolution: the fold of a protein (how the protein chain twists through space) can be determined
- high resolution: most or all of the protein atoms are resolved, which means that their Cartesian coordinates are known with an accuracy of the order of a bond length

It is important to realise that all three levels can help answer biological questions, although in order to answer questions at the level of atoms such as: enzyme catalysis; ligand-binding; protein-protein interaction and protein-nucleic acid interaction, high resolution is obviously the most useful.

Once the atomic coordinates of the protein structure have been determined, a table of these coordinates is deposited into the protein database (PDB).

5. STRUCTURAL CLASSIFICATION OF PROTEINS

The evolution of proteins is restricted due to physical chemical reasons. Solvent accessibility and hydrophobicity are playing an important role in protein folding. Proteins always exist in aqueous solution and they constantly interact with water molecules. The side chains of the amino acids are divided into two groups: the hydrophilic and the hydrophobic side chains. Water molecules have polar charges and liquid water consists of an uninterrupted lattice of hydrogen bonded molecules. A non polar protein molecule dissolved in water interrupts the regular hydrogen bond lattice, which is energetically unfavourable because the water molecules are forced to work around the globular protein and they form a kind of cage. Therefore to diminish the energy costs of solvating the protein, it is favourable to interact with the water molecules and the amino acids with hydrophilic properties have to be arranged solvent exposed at the protein surface, whereas the hydrophobic amino acids are mainly concentrated in the protein core. The number of possible occurring structure motifs (folds) is therefore limited, due to these physical reasons, and they are repetitively used in different proteins. This physical aspect is used for a hierarchical classification of protein structures. Proteins are grouped according to what kind of secondary structure elements they have. Within those larger classes, subclasses are based on the kind of arrangement of the secondary structure elements in the protein. The focus in protein classification is the arrangement of proteins that have similar structural architectures; it doesn't matter if their amino acid sequences are related. The

same physical reasons, leading to the limited number of different occurring folds in proteins and their hierarchical classification, are the basis for the success of many algorithms in structural bioinformatics such as: the structural alignments for the search of homologue proteins and the homology modelling.

The Structural Classification of Proteins (SCOP) is a database maintained by the MRC Laboratory of Molecular Biology UK (http://scop.mrc-1mb.cam.ac.uk/). At the top level of SCOP the proteins are classified by their secondary structure characteristics into all-alpha (α), all-beta (β) or mixed alpha-beta (α/β, $\alpha+\beta$) structures (Class level) [31,32]. In the next level (fold) the succession and orientation of the secondary elements is taken into consideration (globin like, alpha-beta barrel, helix bundle etc.). Beyond the fold level, proteins are divided into Superfamilies (similar structures with no or poor sequence similarity) and Families (similar structures and similar sequences).

Proteins that "look" grossly the same, in structural terms, are classified as more closely. This plays an important role for the search and detection of homologies, common functions and evolutionary relationships.

6. PROTEIN STRUCTURE DATABASE

Protein structural information is publicly available at the protein database (PDB), an international repository for 3D structure files [33]. The PDB database was initiated at the Brookhaven National Laboratory and is now handled by the RCSB (Research Collaboratory for Structural Biology) at the Rutgers University and the UC (University of California) San Diego. PDB is the most important source of protein structures. At the moment PDB contains more than 52.000 protein structures. Before a new structure of a protein is inserted into the database, a time consuming and careful examination of the data must be carried out to guarantee the quality of the structure. Access to the database is enabled via the Internet.

Access to the Structural Data

The entry point to the structural protein data is the PDB web site: http://www.rcsb.org/pdb. The search for a particular protein structure of interest can be initiated by entering the 4 letter PDB identification code at the PDB main page. Alternatively, the PDB can be searched by use of keywords. A comfortable access to the PDB database is enabled by the integrated ENTREZ (www.ncbi.nlm.nih.gov/Entrez) search and retrieval system of the NCBI (National Centre for Biotechnology Information) [34]. This can be done by searching the structural database using specific keywords such as: the name of the protein or organism or other identifying features such as: domains; folds; structures of the protein bound to a substrate etc. The keywords are used to search for entries in the most important fields in the PDB data header. The advantage of the access via ENTREZ is the availability of several public domain tools like BLAST (Basic Local Alignment Search Tool) which enables the user to compare an uncharacterized protein sequence with the whole sequence database or VAST (Vector Analysis Search Tool), which is useful for the determination of structural neighbours of a given protein structure.

After the successful access to the protein of interest in the database, the appropriate information is available through the Information Portal to Biological Macromolecular Structures interface (Fig. **10**) which offers, among others, the opportunity to:

- View protein structures with simple 3D display tools
- Display PDB file and Header in HTML format and download structure files

Fig. (10). A large number of protein structures are deposited in the PDB structure database. The structural information is stored in the PDB data files as atomic coordinates. Information about protein structures is available through the Information Portal to Biological Macromolecular Structures interface. In the database, the proteins are identified by a 4 letter code. The available information includes: sequence of the protein; experimental parameters; structural properties such as dihedral angles; bond angles; bond lengths and others.

- Display geometrical properties, such as: dihedral angles; bond lengths and bond angles, in tabular format.
- Display and download sequences in FASTA format [35].

Additionally, links for many tool databases for protein structure analysis and protein classification databases (similar to SCOP) are offered. The retrieved protein structure is available as a structure file which contains the atomic coordinates of the protein crystal structure.

Data Format

The Brookhaven PDB format is the most commonly used protein structure data format [36,37]. This format is divided into a section containing miscellaneous information and a section containing the atom records. The main parts of the PDB data format are:

- HEADER: contains name and source of the protein, resolution, description of experimental conditions and details, names of the authors, crystallographic parameters including R-factor, sequence information, secondary structure information, literature citations etc.
- SEQRES: this part contains the protein sequence.
- ATOM: this part contains the atomic coordinates and the B-factor values as part of the protein chain.
- HETATM: coordinates of cofactor molecules, substrates, other groups that are not covalently bound to the protein.

PDB files are used as input for protein visualization and they offer the necessary information for the calculation of protein properties, manipulation and representation of the structures. Other commonly used formats are: mmCIF (macromolecular Crystallographic Information Format) and the ASN.1 format [38,39].

From the PDB database the structure files can be downloaded and the protein visualized at the local place (Fig. **11**).

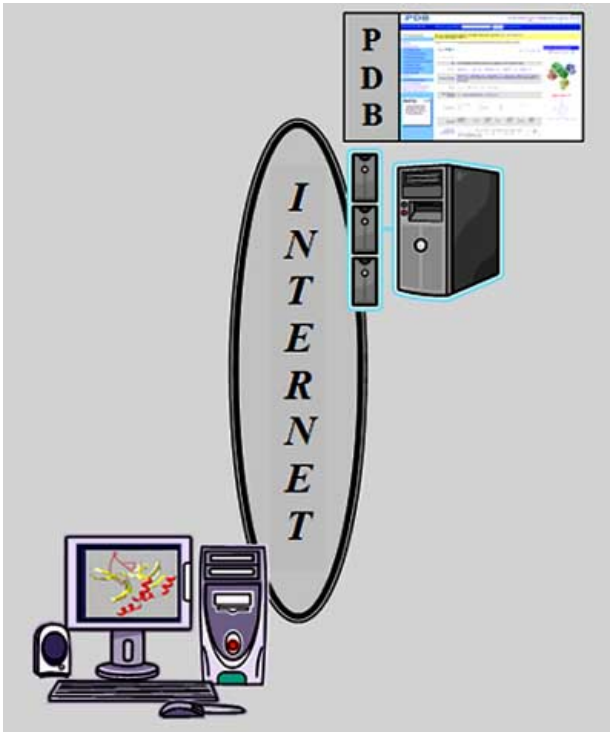


Fig. (11). The PDB database can be accessed through the Internet. The PDB data files are the input files commonly used, for protein visualization. The data files are downloaded from the PDB database, and the protein visualized, locally.

7. VISUALIZATION OF PROTEIN STRUCTURES

The first computer representations of macromolecular structures were made during the 1960's. Researchers at the MIT developed a system based on an oscilloscope for displaying rotating wireframe representations of macromolecular structures [40]. At the same time, a program to produce stereoscopic drawings of molecular structures with a pen plotter was developed [41]. In the mid 1970's a protein structure was visualized entirely with computers for the first time [42]. During the 1980's computer systems showing wireframe renderings of the amino acid chain, which could be rotated in real time become very popular for crystallographers [43]. In the 1990's first programs running on Macintosh and PC brought molecular visualization to a large number of scientists [44]. Nowadays, a number of protein viewers are freely available from referenced Web sites, which run on most computer platforms and operating systems including Microsoft Windows, Macintosh and UNIX X-Windows.

Such viewer programs convert the atomic coordinates into a view of the protein. Common viewers provide ways to manipulate the protein by rotating, translating and zooming the molecule and enable features such as: distance measurements; analysis of bond angles and torsions angles etc. Commonly used viewers are: Swiss-PdbViewer; RasMol; Chime; Cn3D; JMOL; KiNG etc. [45-50].

Protein Models

In the world of protein molecules the application of visualization techniques is particularly rewarding, for it allows us to depict phenomena that cannot be seen by any other means. Being able to "see" the 3D structure of a protein and to analyze its shape is of crucial importance for understanding protein properties and interactions (Fig. **12**). Obviously, it is not possible to see a protein for example by a microscope with X-rays focussing lenses. Therefore there exists no real image, like a microscopic view from cell, of a protein. Instead a model, based on atomic coordinates resulting from the optimal fitting into the electron density map, must be used. (To be precise: the real experimental results in X-ray crystallography are the diffraction patterns on the detector, the electron density map is already derived from these data). This means that there is always an element of interpretation of the experimental data that leads to a model - a hypothesis about the structure that gave rise to the experimental data that we collected. This model of a structure is a three-dimensional representation of a protein that contains information about the spatial arrangements of groups of atoms. The model reflects the experimental data in a consistent way. With the aid of the model, experimental results concerning structural properties can be interpreted and further hypothesis about the protein interactions be formulated.

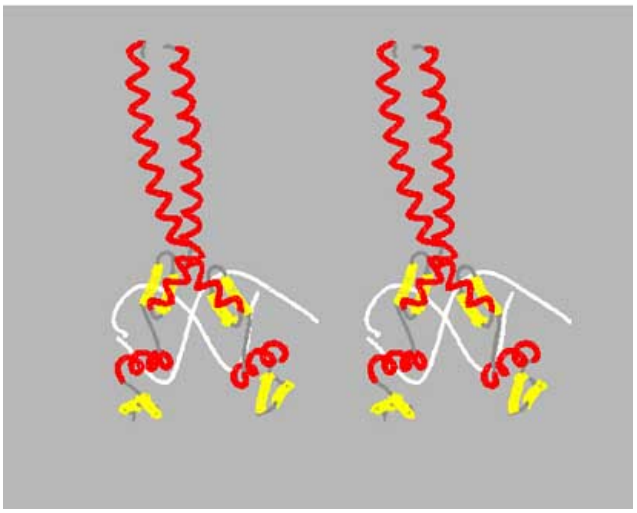


Fig. (12). Stereo view of a zinc finger domain (involved in the specific modulation of gene expression) interacting with a DNA double strand (resolution: 1.5 angstrom, R-value = 0.216). The two images of the protein are rotated around the vertical axis by 2 degrees. The principle of seeing a 3 dimensional image is based on the fact that each eye sees a slightly different side of the protein and the 2 views are combined mentally to a 3D object. By fixing the left image with the left eye and the right image with the right eye a 3D impression results (good luck).

However, simply plotting the coordinates is not suitable for visualization, for this, the connectivity between atoms has to be taken into account. This can be done by the "chemistry rule method". This method uses some of the rules of chemical bond lengths to connect atoms. For example, all neighbouring carbon atoms at a distance of 1.5 angstrom will be connected by a chemical bond. If such rigid chemistry rules are implemented in the viewing program, then no complete bonding information is necessary in the structure file. This approach is based on the PDB files. Another method to connect atoms in a protein is the "chemical graph method". By this method residue dictionaries, containing a list of the amino acids with their respective standard atoms and bond lengths, are used to construct a chemical graph by considering sequence information. Then the viewing program uses the information from the dictionary to connect the atoms. Structure files (MMDB: Molecular Modelling Database type) containing residue dictionaries are derived from PDB files.

Representation of Protein Structures

The information stored in the structure file is then visible as 3D graphic [51-56]. The proteins are so complex, that the 3D structures are difficult to interpret visually. The first problem in the visualization and interpretation of a protein structure is the appropriate representation [57-60]. The human eye can interpret 3D solids but has difficulties with topologically complex 3D data sets. The number of amino acids in proteins ranges from 50 to 2000 residues. For example: the human serum albumin protein contains 4902 atoms. Because of the complex protein structure, special simplified representations are necessary. Although the representations are constructed, they are based on experimental data and therefore they represent real aspects of proteins. Successively higher grades of abstraction of the representations are reached.

The basic representation of protein structures is the ball-and-stick model where atomic details are visualized (Fig. **13**). The covalent bonds are represented as sticks between atoms, which are represented as spheres or simply coloured junction points of the sticks [61]. Non covalent bonds such as hydrogen bonds are represented by doted lines. Structurally important covalent bonds such as sulphur bonds are highlighted by colour encoding (yellow in the case of sulphur bonds). The atoms are generally coloured according to the CPK colour scheme: white for carbon; red for oxygen; blue for nitrogen and yellow for sulphur.

Additionally there are a number of conventionally simplified representations of protein structures that allow the visualization of overall topology, without the confusion of atomic details. A common feature of all proteins is that they are composed of a linear polymer chain, the backbone. In the native state of a protein the polymer chain is folded in 3D space. The first step is the representation of the backbone as a string showing the spatial course, from N-terminus to C-terminus, of the protein backbone (Fig. **13**). This allows the identification of simple topological properties (for example, whether the protein has a spherical shape or is stretched) and facilitates the relation of the sequence of the protein to its spatial structure.

Another simplification of the representation and visualization of proteins is based on ribbons and helices, allowing

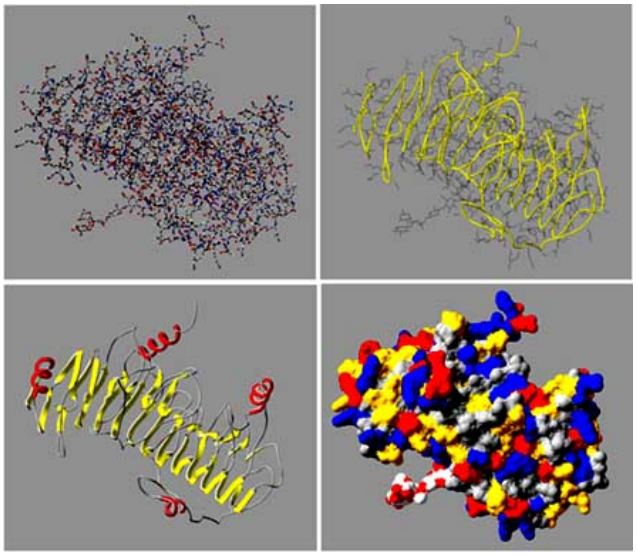


Fig. (13). Representations of protein structures illustrated in the case of chondroitinase B (enzyme that cleaves the glycosaminoglycan, dermatan sulphate) with the beta solenoid domain (resolution: 1.7 angstrom, R-value = 0.19). Top: Left picture: Ball-and-stick representation. The atoms of the backbone and the side chains are connected according to chemistry rules. Right picture: Representation of the backbone as a string. Bottom: Left picture: Representation of the secondary structure elements as ribbons: α-alpha helices (red) and β- sheets (yellow) connected by loops in grey colour. Right picture: The molecular surface is coloured by chemical type. Non polar residues are gray, polar residues are yellow, basic residues are blue and acidic residues are red.

the user to visualize essential features of the protein topology as secondary structure elements (Fig. **13**). Protein coordinate data sets don't come labelled with secondary structure classifiers. Secondary structures can be detected by their hydrogen bond patterns and their backbone torsion angles. Secondary structure elements can also be represented as icons (cartoon representation), whereas the α-helices are represented as cylinders and the β-sheets as arrows. Visualisation and analysis of the secondary elements plays, amongst other things, an important role in protein classification.

The proteins can further be represented by their molecular surface (Fig. **13**). The molecular surface is the boundary of the molecule volume within which no other molecule can enter. Molecular surfaces are helpful for the description of the topological properties of the protein shape. The surface is defined by the van der Waals radii of the atoms, which is the radius of the sphere around the atom centre, inside which another atom's spherical boundary can't pass. Then the surface depends on the atomic van der Waals radii and the coordinates of the atoms in the molecule. The protein surface is not a quantity defined by a unique property. Surfaces can be constructed as molecular or accessible surfaces. The first kind of surface is defined as the contact surface of a spherical probe (water molecule) with the protein molecule, the second kind as the locus of the centre of the probe rolled around the molecule [62-64]. The default probe radius for each type of surface is 1.4 angstroms.

An important task is the visualization of non-structure-based functional annotations in protein 3D structures. Bioinformatics is a science with centralized data banks, combining sequence information to functional and structural information, which are interconnected through the Internet (Fig. **14**). An adequate representation of structural properties together with the mapping of biological and physicochemical information from different databases helps to understand proteins visually showing context and connection.

Representation of Sequence-Structure Relationships

To investigate the structure and function of a protein, researchers create increasingly sophisticated computational models for transforming genetic sequencing data into comprehensive information at the protein level to better understand biological processes.

Special attention has been done in the development of viewers that can present structure and sequence data in a unified interface [65-72]. 3D viewers and alignment editors are connected together allowing the rapid refinement of sequence-structure alignments by taking advantage of the immediate visualization of resulting insertions/deletions and strict conservations in their structural context. Multiple linked visualization of protein structure and sequence view show context maintenance by fish eye sequence view and magic lens [68]. Additionally, web applications for the visualization of multiple protein sequences and structures, allowing the visualization of conservative and variable residues of active and binding sites in protein families were developed. Visualization of the relationship between exon distribution in the gene and protein structural elements and the association of sequence signals to 3D structure provide a valuable visualization environment for gene organization, gene evolution, protein folding and protein structure classification [69]. The

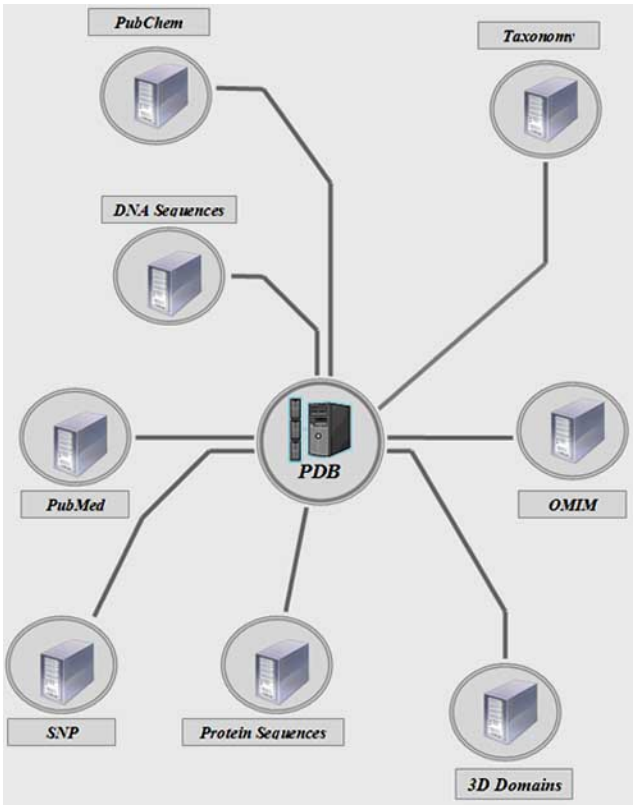


Fig. (14). The PDB database is interconnected with different public databases. The compound of data bases includes molecular, biological and literature databases which are additionally interconnected to further databases. The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq. Protein sequences result from a variety of sources, including SwissProt, PIR, PRF. PubMed is an archive of biomedical and life sciences journals. The PubChem substances database contains descriptions of chemical samples. OMIM (Online Mendelian Inheritance in Man) contains a catalogue of human genes and genetic disorders. The taxonomy database contains the names of all organisms that are represented in the genetic databases with at least one nucleotide or protein sequence. SNP (Single Nucleotide Polymorphism) contains gene variations. The 3D Domain database contains compact structural domains. The access from one data base to another is done by active links.

location of the exon boundaries and the intron phase are mapped onto a multiple structural alignment [67]. Comparative analysis of exon/intron organization of genes and their resulting protein structures are important for understanding evolutionary relationships between species, rules of protein organization and protein functionality.

Physicochemical properties like: the isoelectric point; the estimated half-life of the protein; its instability index and the average of hydropathicity can be calculated out of the amino acid sequence [73-75]. This can be done for example with the program ProtParam (http://www.expasy.ch/tools/protparam.html), which is additionally used for the calculation of the molecular mass and the percentage of amino acid composition of the protein (Fig. **1**). The courses of the hydropathicity values along the amino acid sequence can be

determined and plotted by ProtScale (http://www.expasy.org/cgi-bin/protscale.pl) [76]. Several programs are used to predict secondary structures of proteins from their amino acid sequence. For example the program PHD (Profile network prediction HeiDelberg) predicts the secondary structure elements out of multiple sequence alignments [77]. By use of neural networks, 3 different secondary structures (helix, strand and loop) are predicted (http://cubic.bioc.columbia.edu/predictprotein/). The neural networks are first trained with proteins of known structures.

Representation of Structure-Function Relationships

Special procedures are used for exploring structure-function relationships. Therefore, sequence-based functional annotations need to be mapped to the corresponding part of the protein structure. Currently there are two types of annotations, the primary database (knowledgebase): SWISS-PROT with sequence annotations (http://ftp.expasy.ch/sprot) and the derived databases (cross referenced by SWISS-PROT): PROSITE (http://ftp.expasy.ch/prosite) with biological significant sequence patterns; PRINTS (http://umber.sbs.man.ac.uk/cgi-bin/dbbrowser/sprin) with fingerprint motifs of proteins; PFAM (http://www.sanger.ac.uk/cgi-bin/Pfam) contains for every family a multiple sequence alignment as well as a hidden Markov model; Blocks (http://blocks.fhcrc.org/) a database for multiple alignments of motifs; Highly conserved domains in the structures can be found in the CDD: Conserved Domain Database (http://www.ncbi.nlm.nih.gov/Structure/cdd) etc. [78-87].

A motif is a locally conserved region in a sequence or a short sequence pattern shared by a set of sequences. Often motifs are localized in active sites or binding sites for substrates and coenzymes. Motifs are common in protein families where the requirements of the active site mean a restriction of the protein evolution. Motifs are derived from multiple sequence alignments of members of a protein family. Common motifs in proteins often indicate a common function of proteins, even when no other sequence similarity can be found. Therefore motifs are useful for the prediction of functions of an unknown protein. PROSITE is a database which contains biologically significant patterns responsible for the function of a protein family [88,89]. The PROSITE motifs consist of highly conserved residues which are relevant for:

- Catalytic sites for enzymatic function
- Binding sites for molecules such as: DNA, ATP or other proteins
- Binding sites for ions.

These motifs are described by the PROSITE syntax [90]:

[FY]-x(6)-C-C-x(7)-C-[LFY]-x(6)-[LIVMFYW].

This pattern represents a common motif in the albumin family [91-93]. The amino acids are represented in the one letter code. x(n) means an pattern of n arbitrary residues. Letters between brackets (for example: [LFY]) mean that one of the involved amino acid must be present at the specific position. One letter (C) in the pattern means that exactly this amino acid and no other (very highly conserved) must be present at the specific position. This motif pattern is used to search and visualize the corresponding residues in the 3D structure of human serum albumin (Fig. **15**). Albumin is synthesized in the liver and serves among others as transport protein for water unsolvable substances. It binds fatty acids, hormones, bilirubin and drugs [94].

Essential properties of proteins, responsible for their functional behavior, can be described by use of single physicochemical parameters. This simplifies the understanding and computation of specific protein properties. For example: properties of transmembran proteins can be described by hydrophobic parameters; electrostatic interactions between an enzyme and a substrate (resulting in catalytic activities) can be described by an dielectric constant. The combination of physicochemical and structural properties of proteins may suggest: the location of catalytic sites; the location of interaction sites; identification of targets for site directed mutagenesis and the identification of membrane spanning domains. Therefore a mapping of physicochemical properties onto the protein structure is performed. To connect information about physicochemical properties with the topology, colour encoding according to the specific parameter is used.

Many other important protein features can be obtained via some recent developed bioinformatics tools. For instance, at the web server "Cell-PLoc" at http://chou.med.harvard.edu/bioinf/Cell-PLoc/ subcellular locations of proteins in various organisms can be identified, at the web server "MemType-2L" at http://chou.med.har-vard.edu/bioinf/-MemType/ membrane protein types are identified, and

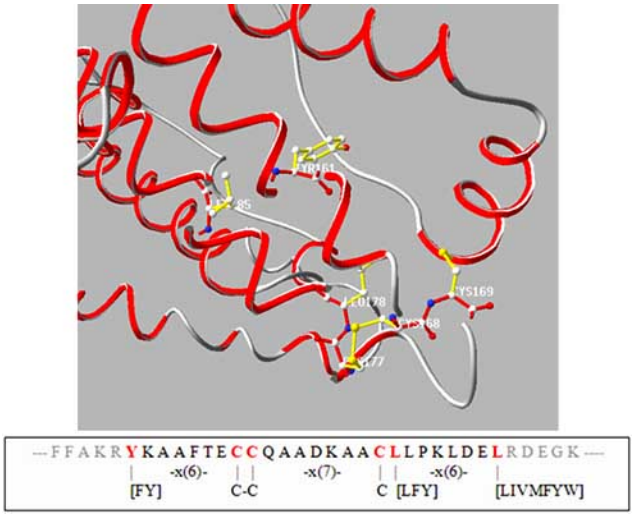


Fig. (15). Connections of the human serum albumin structure with functional information (resolution: 2.7 angstrom, R-value = 0.251). The figure shows a common biological significant residue sequence pattern (motif) of the albumin family and its location in the albumin 3D structure. The expressions between brackets show alternatively allowed amino acids at the specific position in the motif pattern. Some amino acids of the active site are distant in the sequence but close in 3D space. In the sequence the residues are represented in the one letter code, whereas in the 3D the residues are represented in the three letter code. (Y: Tyr: Thyrosin, C: Cys: Cystein, L: Leu: Leucin). The motif sequence pattern results from the PROSITE database, but only its visualization in the 3D structure is the precondition for the analysis and understanding of its function.

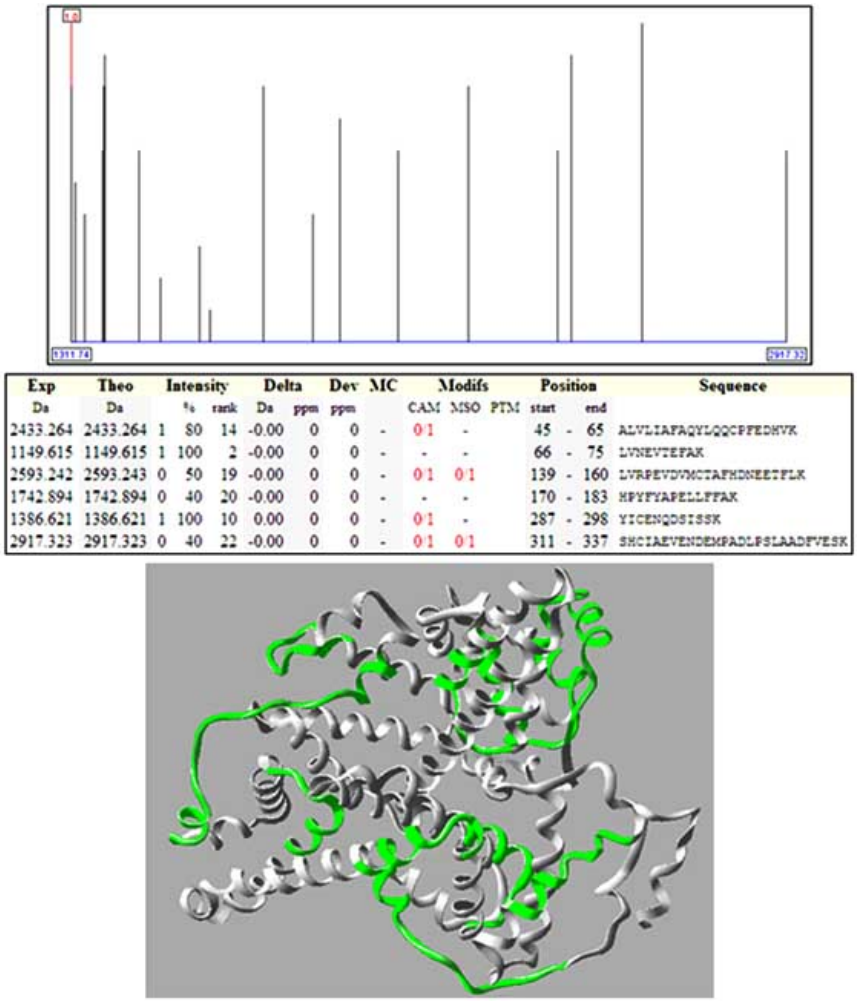
by use of the web servers "Signal-CF" at http://chou.-med.harvard.edu/bioinf/Signal-CF/ or "Signal-3L" at http://-chou.med.harvard.edu/bioinf/Signal-3L/ signal peptides of proteins in various organisms are identified [95-97]. These kinds of information are very useful for basic research and drug design.

How bioinformatics contributes to human health care will be dramatically influenced by the degree to which the tools are adopted and used as a routine aspect of biomedical research. The choice of problem based representations will improve the usability and importance of visualization tools for medical research. This will broaden the use of genomic and proteomic information in medical research and practice while enabling new insight into biological processes.

8. DATABASE AIDED IDENTIFICATION OF PEPTIDE FRAGMENTS

In order to identify a protein, it is necessary first to determine its amino acid sequence. Large proteins are decomposed into peptide fragments by an enzyme (for example: trypsin). The masses (in Dalton units) of the peptide fragments are determined by mass spectroscopy. In mass spectroscopy, the peptide fragments are ionized and then separated according to their mass using several techniques. One technique is the SELDI-TOF "Surface enhanced laser desorption ionisation - time of flight" method, where the various masses of peptide fragments are determined according to their time of flight (TOF) [98]. Mass spectroscopic methods deliver, for every protein, a characteristic peptide mass fingerprint [99,100]. From this fingerprint, the peptide mass spectrum, the peptide fragments can be identified by data base aided methods which allow conclusions to be drawn about the sequences of the peptide fragments and the identification of the protein (Fig. **16**). A lot of software tools for protein analysis are available at http://au.expasy.org/-tools/ [101]. One of these tools, the program Aldente enables an automatically comparison of the determined mass spectrum with the masses of all possible amino acid combinations [102]. It takes advantage of the Hough transformation for spectra recalibration and outlier exclusion. The output is a list of the peptide fragments with their sequences and position in the protein (Fig. **16**).

Once the sequences of the peptide fragments have been determined they are linked together to the whole protein se-



Exp	Theo	Intensity			Delta		Dev	MC	Modifs			Position		Sequence
Da	Da		%	rank	Da	ppm	ppm		CAM	MSO	PTM	start	end	
2433.264	2433.264	1	80	14	-0.00	0	0	-	01	-		45	65	ALVLIAFAQYLQQCPFEDHVK
1149.615	1149.615	1	100	2	-0.00	0	0	-	-	-		66	75	LVNEVTEFAK
2593.242	2593.243	0	50	19	-0.00	0	0	-	01	01		139	160	LVRPEVDVMCTAFHDNEETFLK
1742.894	1742.894	0	40	20	-0.00	0	0	-	-	-		170	183	HPYFYAPELLFFAK
1386.621	1386.621	1	100	10	0.00	0	0	-	01	-		287	298	YICENQDSISSK
2917.323	2917.323	0	40	22	-0.00	0	0	-	01	01		311	337	SHCIAEVENDEMPADLPSLAADFVESK

Fig. (16). Connection between the clinical laboratory and bioinformatics. Human serum albumin is decomposed into peptide fragments by trypsin. Mass spectrometric methods deliver for every protein a characteristic peptide mass fingerprint: the peptide mass spectrum (top). From this fingerprint, the peptide fragments can be identified by data base aided methods which enable an automatic comparison of the mass spectrum with the mass spectrum of all possible amino acid combinations (middle). A part of the identified peptide fragments (green) is visualized on the albumin 3D structure (bottom). The masses range from 1000 - 3000 Dalton.

quence. If the structure of a protein of known amino acid sequence has not yet been determined experimentally, it can be determined by homology modelling.

9. HOMOLOGY MODELLING

Because of the abundant number of known protein sequences, which exceed by far the number of experimentally determined structures, the relatively slow and expensive experimental methods have been supplemented by theoretical methods. Homology modelling is a form of protein tertiary structure prediction, based on the assumption that proteins that are homologous in sequence are similar in structure [103-108]. In homology modelling a protein sequence with an unknown structure is aligned with one or more protein sequences with known structures. The necessary condition for successful homology modelling is a detectable similarity between the amino acid sequences (more than 30%) allowing the construction of a correct alignment.

Principle

Homology modelling is a knowledge-based prediction of protein structures. The method uses parameters extracted from existing structures to predict a new structure from the sequence. This approach is possible because changes in the amino acid sequence usually cause only small structural changes. Therefore homology modelling first involves the finding of already known protein structures and then building the query sequence onto the homologous template structures (Fig. **17**). The steps in homology modelling are:

- The sequence with the unknown structure (the target) is used as a query to find similar sequences with known structures (the templates).
- The sequences are brought into an optimal alignment.
- The backbones of the templates are used to model the protein backbone of the target sequence.
- Loop-modelling procedures are used to fill gaps in the alignment.
- The side chains are added and their positions are arranged.
- The obtained structure is optimized by energy minimization or knowledge-based optimization.

These steps are found in common in every method of homology modelling [109,110]. A crucial issue in successful homology modelling is the quality of the alignment. The sequence alignment is used to determine the equivalent residues in the target and the template protein. Then the structure of the target protein is constructed by exploiting the information from the template structure. Some algorithms rely on rules of spacing between atoms, bonds lengths and angles etc. from observed values in known protein structures. These data are extracted from the template in form of spatial restraints and the structure of the target is constructed by satisfying all the restraints as well as possible. Other algorithms build the core of the model by averaging the backbone atom positions of the template structures, whereas the templates are weighted by the sequence similarity to the target sequence. The side chains are built by iso-sterically replacement of side chains in template structures. Of course, the predicted structure in homology modelling is not exactly the same as the template structures because of variations in the amino acid sequences. But structural deviations are allowed only in the range of differences found between homologous structures.

Distortions in the structure, which have been introduced by rigid modelling, are regulated by energy minimization of a molecular force field (chapter 10). Energy minimization is

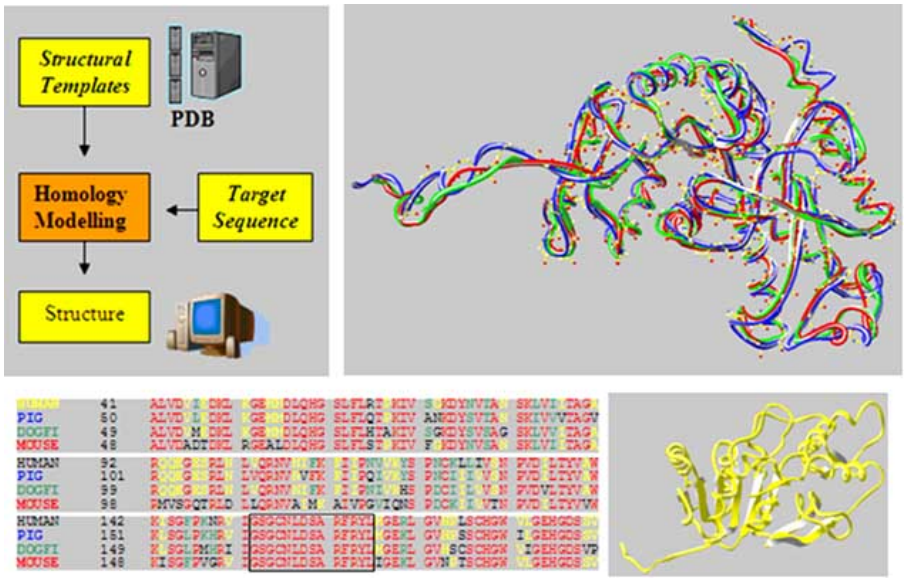


Fig. (17). Principles of homology modelling: In homology modelling a protein sequence with an unknown structure (the target) is used to find protein sequences with known structures (the templates). First the sequences are brought into an optimal alignment (bottom left). A part of the active site was determined by detecting the pig lactate dehydrogenase residues in the environment of the bounded pyruvic acid and highlighted (square) in the alignment. After the homology modelling procedure, the target sequence is "wrapped" around the superposed template structures (top right). The structure of the target is constructed by satisfying all the restraints, taken from the template structures, as well as possible (bottom right).

a technique that changes the conformation of a molecule in order to decrease its energy as much as possible. Energetically unreasonable features in conformations are for example: atoms that clashes with each other; unfavourable long bonds or bond angles etc. The predicted structure is optimized by energy minimization allowing the structure to settle into a lower energy conformation as similar as possible to the modelled conformation

The chance to find a homologous sequence with a known structure in the PDB continually increases because the number of experimentally determined structures is growing rapidly and because physical constrictions limit the number of occurring structural folds (chapter 5). It is estimated that in a few years for every new protein sequence, at least one member with known structure of the same family will be available in the database. This shows that the usefulness and success of homology modelling is steadily increasing. The different homology modelling programs carry out single steps by proprietary algorithms. Two well known programs, which are free for academic research, are MODELLER (http://-salilab.org/modeller/) and SWISS-MODEL (http://swissmodel.expasy.org/) [111-116].

Example: Homology Modelling of Human Lactate Dehydrogenase

In order to illustrate the principles, homology modelling with the human lactate dehydrogenase sequence is carried out [117]. Lactate dehydrogenase is the enzyme of the glycolysis which catalyses the reduction of pyruvic to lactate. Lactate is the salt of the lactic acid which can cause stiffness. First, the human lactate dehydrogenase sequence is used as a query to search the PDB database for homologous sequences of known structures. To this purpose, the program BLAST (Basic Local Alignment Search Tool) was used, where the query sequence is entered via a simple web form (http://blast.ncbi.nlm.nih.gov/Blast.cgi) [118]. This program performs pairwise comparisons of sequences. As a result, the sequences of lactate dehydrogenase from pig, mouse and dogfish were found as homologues. The alignment with the human lactate dehydrogenase sequence shows an E-value (expection-value) of e-142 for the sequence from pig-lactate dehydrogenase, e-136 from mouse-lactate dehydrogenase and e-132 from dogfish-lactate dehydrogenase [119-121]. The E-value reflects the likelihood that a given sequence alignment is just random (occurred by chance). Alignments with low E-values are very significant, it means there is a high probability that the sequences are homologous. The three structures were selected as templates for the homology modelling (Fig. **17**). The output from the homology modelling process is also a structure file in PDB format, containing the atomic coordinates of the calculated structure.

Evaluation of the Predicted Protein Structures

The quality of the structure prediction can be evaluated with ProCHECK, a program that compares the geometry of the predicted structure with well known high resolution structures (http://www.biochem.ucl.ac.uk/~roman/procheck/-procheck.html). Important estimation criteria (stereochemical parameters) are: bond lengths; bond angles; torsion angles and planarity [122]. The so called *G*-factor provides a measure of how realistic a given structure is. The *G*-factor is based on the observed distributions of the stereochemical parameters. A low *G*-factor indicates a low-probability conformation.

Applications of Molecular Modelling

Homology modelling remains, at present, the most efficient technique to provide accurate structural models of proteins. Typical uses of homology modelling are: designing of site-directed mutants to test hypothesis about function; identifying binding sites; improving inhibitors for a given binding site etc. The structures of medical relevant proteins involved in malaria, nicotinic acetylcholine receptor, bacteria inhabiting the gastrointestinal tract, anticancer drugs, hematopoietic stem/progenitor cell selection, autoimmunity and many more, have been determined by homology modelling [123-129].

10. MOLECULAR FORCE FIELDS

Molecular force fields are used to calculate the energy of a given molecular structure [130]. A task is the finding of the optimal geometry of stable molecules or of different possible confirmations. This can be done for example after a structure determination with homology modelling. The problem is then reduced to determine energy minima on the potential energy surface. Force fields are based on the observation that molecules are composed of units that are similar in different molecules. For example all C-H bonds have roughly constant lengths and their stretch vibrations are similar in different molecules. The empirical experience that molecules are composed of group of atoms, which are similar in a variety of molecules, is implemented in molecular force fields as concept of atom types. The atom type is defined by its chemical environment which can be distinguished by: the atomic charge; the nearest neighbours; the hybridization etc. The interactions are then calculated according to the atom type. Therefore force field methods are restricted to classes of molecules where information already exists. In molecular force fields the electronic energy is included as parametric function of the atomic coordinates where the parameters are fitted to experimental data. The building blocks in molecular force fields are atoms; electrons are not included as individual particles. The quantum aspects of the nuclear motion are neglected and the dynamics of the atoms in the molecule is treated by classical mechanics.

Formulation of Molecular Force Fields

The molecular force field describes a molecule as a collection of interacting atoms, with Cartesian coordinates r, that are expressed by simple analytical functions. The different interactions can be studied separately. Molecular force fields describe the potential energy surfaces. A general formulation of molecular force fields is as follows:

$$V = \sum_{bonds} K_r \left(r - r_0\right)^2 + \sum_{bends} K_\Theta \left(\Theta - \Theta_0\right)^2 + \sum_{dihedrals} V_n \left(1 + \cos(n\Phi)\right)$$
$$+ \sum_i \sum_j \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + \sum_i \sum_j \frac{q_i q_j}{r_{ij}} + \sum_i \sum_j \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right)$$

The first three terms describe the energetically situation of covalent bonds (Fig. **18**). The main parameters are: bond length ($r - r_0$), bond angle (Θ) and the dihedral angles (ϕ). The first part describes the extension (stretching) of the covalent bonds, the second part the distortion of the bond angles, the third part the distortion of the dihedral angles (an-

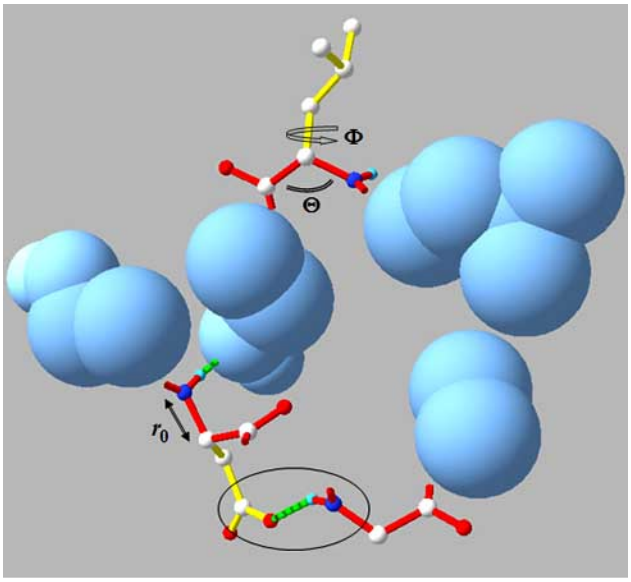


Fig. (18). A molecular force field describes the energy of a molecular system with respect to: bond lengths; bond angles; van der Waals forces; electrostatic forces and hydrogen bonds. The energy of covalent bonds depends from bond length (r_o), bond angle (Θ) and the dihedral angles (ϕ). Due to the stretching of bond lengths and the distortions of bond and dihedral angles, energetically unfavourable situations may occur in a protein molecule. Steric hindrance results from overlapping of the, with the residues associated, van der Waals spheres (blue), leading to strong repulsive forces. Hydrogen bonds (green dotted lines) are stabilizing forces arising between a proton donator and a proton acceptor. The peptide backbone is coloured in red, whereas the side chains are yellow.

gles of rotation of the bonded atom pair around the central bond). The constant expressions (K_r, K_Θ, V_n) are dependent on the kind of covalent bond. The bond lengths are determined by the types of involved atoms and the number of the shared electrons between the atoms. An important covalent bond, contributing to protein structure stability is the disulphide bond between the sulphur atoms of two side chains of cystein residues. Bond angles are constrained by the structure of the electron orbital. The dihedral angles are constrained mainly by steric hindrance.

The fourth term describes the van der Waals forces. Due to the movement of the electrons around the atomic nucleus, an atom can be considered as an electric dipole. The dipole of an atom polarises the neighbour atom resulting in a transient attractive force between the atoms. Conversely, at short ranges a repulsive force between the electrons of both atoms arises. The radius at which the repulsive force begins to increase sharply is called the van der Waals radius. Steric interactions arise when the van der Waals spheres of two non bonded atoms are approaching and entering each other (Fig. **18**). The parameters A_{ij} and B_{ij} depend on the type of the involved atoms.

The fifth term describes the electrostatic forces arising between atoms carrying a positive ionic charge and atoms carrying a negative ionic charge. Charge-charge interactions between ions are called salt bridges. They occur between the side chains of the amino acids with opposite charges. They play a significant role in protein structure stabilization. There are other, weaker interactions that occur between dipolar residues (carrying partial positive and partial negative charge) and other partial or ionic charges: dipole-dipole, charge-dipole, charge-nonpolar and dipole-nonpolar.

The sixth term describes hydrogen bonds arising between a proton donator and a proton acceptor (Fig. **18**). These are weak chemical/electrostatic interactions between two atoms. They arise from the interaction of two polar groups containing a proton donor (amino group) and proton acceptor (carboxyl group). The donor group contains a hydrogen atom, covalently bonded with an electronegative atom, with partially positive charge. The electron of the proton is partially shifted to the bonding partner. The acceptor group has a partial negative charge with no attached proton. The parameters C_{ij} and D_{ij} depend on the types of the proton donator and acceptor atoms. The strength of hydrogen bonds depends on the distance and the bond angle. The overlapping of the electronic orbitals of the involved atoms in hydrogen bonds is usually small. Hydrogen bonds are identified if a proton donor atom and a proton acceptor atom of such groups are interacting at distance equal to or less than 3.6 angstroms. Hydrogen bonds can be formed between atoms of side-chains and/or peptide backbone atoms. They are one of the most stabilizing forces in proteins (mainly the secondary structures like α-helices and β-sheets) and are responsible for the binding of the protein to substrates.

Applications of Molecular Force Fields

Energy minimization of the molecular force field establishes physically realistic structural configurations of molecules (for example after homology modelling). Equilibrium geometry of a molecule, with respect to bond lengths, angles, non overlapping van der Waals spheres etc., describes the atomic coordinates at a minimum on the potential energy surface. Different minima are connected with different conformations and saddle points represent transition states. Monte Carlo methods are used to simulate equilibrium properties and transitions between different states. For the study of molecular mechanics (MM), molecular force fields provide information about the time-dependent behaviour of molecules resulting in interconversions of different conformations and chemical reactions. Force field methods are however inherently not suitable to describe electron rearrangements and transfers in molecules, which must be done by quantum theoretical methods. An advantage of molecular force fields method is the speed with which calculations can be performed, enabling the application to large bio molecules. Even with moderate computer power, molecules with thousands of atoms can be optimized. This facilitates the molecular modelling of proteins and nucleotide acids which is nowadays done by most pharmaceutical companies.

There are several common force fields in use in professional molecular modelling programs. They are simplified for the treatment of macromolecules, for example, by not considering hydrogen atoms explicitly (united atom approach). The force fields are calibrated against experimental data like: bond lengths, bond angles etc. Common molecular force fields are: GROMOS, CHARMM, AMBER, MM+, BIO+ and OPLS force fields [131-136].

11. STRUCTURAL ALIGNMENTS AND HOMOLOGIES

Similar protein structures are determined by direct comparison of 3D protein structures with the VAST (Vector Alignment Search Tool) algorithm [137]. VAST search can be started at: http://www.ncbi.nlm.nih.gov/Structure/-VAST/vastsearch.html. The neighbour structures are often examples of remote homology, undetectable by sequence comparison. As such they may provide insights into structure, function and evolution of a protein family. The principle of VAST's significance calculation is based on the definition of the unit of tertiary structure similarity as pairs of secondary structure elements (SSE) that have similar type, relative orientation, and connectivity. Similarities of small substructures that occur by chance are not considered. VAST detects via structural alignment similarities between proteins that can not detected by sequence alignments because of poor similarities between the sequences.

To compare two proteins, the structures are superposed (Fig. **19**). The difference between two protein structures is expressed by the root mean squared deviation (RMS) of the respective atomic positions in the two structures.

$$RMS = \sqrt{\frac{\sum_i d_i}{N}}$$

d_i is the distance between two corresponding atoms and N is the number of considered atoms. Mostly the differences of the positions of the respective C_α atoms of the backbones in the two structures are measured. RMS values up to 1 angstrom show a high similarity between the two structures. The protein structures are described in Cartesian coordinates and each structure has a build-in orientation in its proper coordinate system. To compare the two structures, one structure serves as reference and the other structure must be superimposed. This allows the evaluation of which parts of the structures are showing a good or high similarity and where significant structural deviations are located allowing to examine the patterns of structural conservation and change within a protein family.

12. PROTEIN INTERACTIONS

Processes inside and outside cells are described as networks of interacting proteins. The amino acids on the surface of a protein interact with other molecules such as substrates, ligands, receptors etc. The shape of a protein is of special

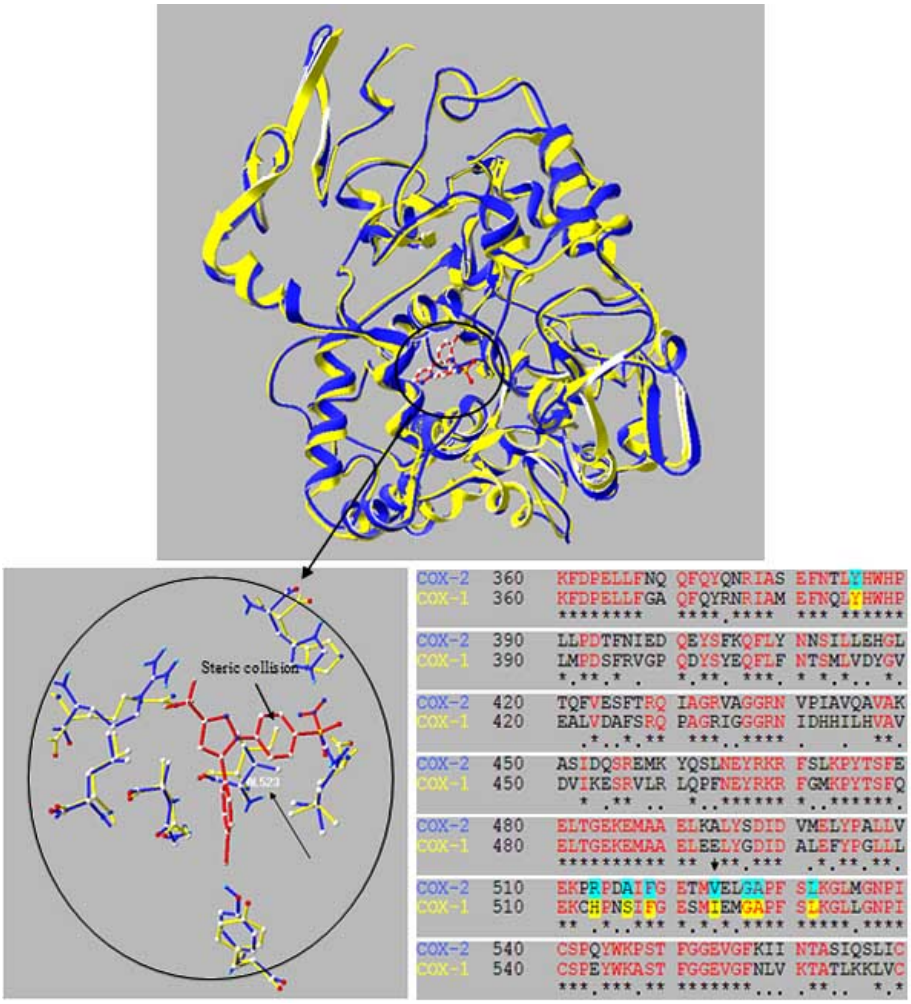


Fig. (19). The structure of COX-2 (blue) is superimposed with the structure of COX-1 (yellow). RMS=0.93. The structural alignment (only a part is displayed) shows a high portion of similar residues in both chains and the corresponding residues of the active sites. The active centre, where COX-2 bounds the selective inflammation inhibitor SC-558 (red), is shown. The interchanging of valine by isoleucine at position 523 leads to a steric collision with SC-558 in the case of COX-1.

importance for intermolecular interactions. These interactions can be described in terms of locks and keys (Fig. **20**). To enable an interaction, the shape of the lock (for example the enzyme) must be complementary to the shape of the key (the substrate). In the immune system the shape of the antibodies must be complementary to the shape of the antigens to initiate an immune response. The body develops antibodies with the right shape to attack specific antigens (for example a virus). Therefore the shape of the molecular surface of a protein offers information on how two proteins in a metabolic pathway interact with each other or why an enzyme is specific to a particular substrate [138,139].

Experimental Determination of Protein-Protein Interactions

The yeast two-hybrid (Y2H) method was developed to study experimentally the protein-protein interactions [140]. The principle of the method is as follow: In a medium without, for example, histidine, yeast can only grow if the corresponding gene (reporter gene) for the production of histidine is switched on. First, the transcription factor responsible for the expression of the reporter gene is cut into 2 parts: the DNA binding domain and the activation domain which stimulates the RNA polymerase to begin transcription. The protein of interest (protein A) is fused with the DNA binding domain part. Its potential counterpart (protein B) for the protein-protein interaction is fused with the activation domain. If both proteins are interacting they form together with the DNA binding and activation domain an active promoter complex which initiates the transcription of the reporter gene and yeast is growing (if they don't interact, the reporter gene stays inactive). This situation occurs when 2 yeast stems, where each carries one of the two fusion proteins on a plasmid, are merged (mating). Information and results about protein interaction can be found on the yeast protein interaction map project homepage (www://depts.washington.edu/sfields/yplm/data/index.html).

Description of the Macromolecular Interface

The molecular surface of a protein M comprising N atoms with radii α_1 and coordinates r_1 can be considered analytically as iso-contours of a sum of exponential functions:

$$G(r, M, d) = \sum_{1}^{N} e^{-(|r-r_i|-\alpha_i)/d}$$

d is an adjustable parameter [141,142]. The sum of the exponential functions approximately represents the electron density distribution. Surfaces similar to the solvent accessible surface can be derived as well as those corresponding to its van der Waals surface [143]. The distance functional, reflecting the distance to the molecular surface, is defined by:

$$D(r, M, d) = -d \ln(G(r, M, d))$$
$$= -d \ln(\sum_{1}^{N} e^{-(|r-r_i|-\alpha_i)/d})$$

This definition is formally equivalent to an iso-contour value of the sum of exponential functions. At any point r the major contribution to the sum will be the distance $(|r - r_i| - \alpha_i)$ to the closest atom. The value of the parameter d defines the relative amount of the close atoms to the sum.

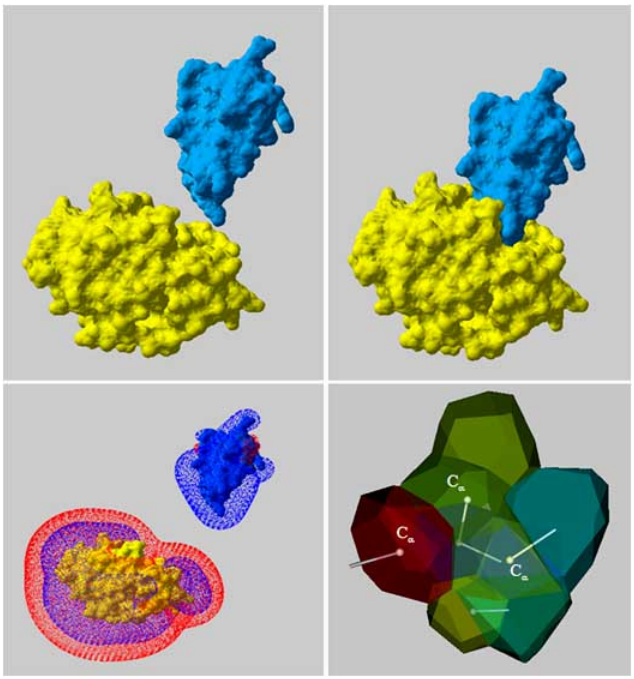


Fig. (20). Trypsin (yellow) and its inhibitor (blue). Trypsin is a digestive enzyme of mammals which catalyses the breaking up (hydrolysis) of peptide bonds. Proteins fit together in geometrically, specific ways, so the shape of the lock has to be complementary to the shape of the key (top). The macromolecular interface is quantified by Voronoi tessellation. In 3D, the Voronoi cell is a polyhedron (bottom right). Each polyhedron is uniquely associated with one of the residues and the starting point is the alpha-carbon atom. The polyhedron is characterized by its number of edges and faces where common faces define the contacts to nearest neighbours. The electrostatic potential plays a role in protein recognition and interaction (bottom left). Positive values are blue encoded, negative values red.

As the value of d decreases, the number of contributing atoms decreases where at d = 0 only the closest atoms contribute. Thus the distance functional reflects largely the distance to the closest atoms. Then the surface of the protein M is defined by:

$$D(r, M, d) = 0$$

If all α_i are atomic van der Waals radii, then the condition for the surface defines an approximation to the van der Waals surface. If all α_i are atomic van der Waals radii incremented by the radius of the solvent probe then the solvent accessible surface is approximated. (If d increases then the surface is getting smoother). The macromolecular interface between two proteins M_1 and M_2 is defined by use of the distances to the surfaces of the first and second protein:

$$D(r, M_1, d) = D(r, M_2, d)$$

The contribution of the amino acids to the macromolecular interface and the total area of the interface can be quantified by Voronoi tessellation.

Quantification by Voronoi Tessellation

In Voronoi tessellation the space, containing a set of discrete points, is subdivided into non overlapping regions

[144]. Each region is associated with an element of the set of points. For a given set of points P in a space:

$$P = \{P_1, ..., P_n\}$$

the regions V_l are defined by:

$$V_l = \{x \mid \|P_l - x\| < \|P_k - x\|, l \neq k\}$$

Every region V_1 contains space points x, which have the shortest distance to the associated point P_1 (the distance is shorter than to the other points of the set). For a given set of points the Voronoi decomposition is unique. The regions V_1 are called Voronoi cells. The set of Voronoi cells is called a Voronoi diagram, which defines the topological relations of the set of discrete points.

In 3D space, the Voronoi cells are polyhedrons (Fig. **20**). The Voronoi tessellation describes then the space filled by a packing of solid polyhedrons, connected by their faces, without empty space between them. The tessellations are mostly performed on the alpha carbon atoms (as point P_1) of the structure [145-148]. Then the polyhedron, sharing a particular alpha carbon C_α as a vertex, defines its closest neighbourhood. The circumscribed sphere of each polyhedron does not contain any other alpha carbon. Each polyhedron is characterized by its number of edges and faces. Two residues are in direct contact when their corresponding polyhedrons share a common face. The macromolecular interface is characterized by the common faces of polyhedrons, joining the two proteins, or in other words: by the common interfacial faces. The contribution of an amino acid to the interface is characterized by the number and size of the faces exposed to the other chain in the macromolecular interface. The total area of the Voronoi polyhedron V_1 of the l-te residue is denoted by A_1. The area of the i-te common face to a residue of the other chain is: a_i. N_a is the number of common faces. Thus a measure for the exposure of the l-te residue to the residues of the other chain is the relative exposure value:

$$E_l = \frac{1}{A_l} \sum_{i=1}^{N_a} a_i$$

The total area of the macromolecular interface of a protein is defined by summing up the areas of the common faces for all the involved residues:

$$F_I = \sum_j E_j A_j$$

Therefore the contribution of the l-te residue to the interface is given by (relative contribution):

$$I_l = \frac{E_l A_l}{\sum_j E_j A_j}$$

The identified residues, which contribute mostly to the macromolecular interface, are used for comparing different interfaces and the evaluation of energetically effects. Detecting similar protein surfaces provides an important route for discovering unrecognized or novel functional relationships between proteins. Voronoi tessellations on proteins can be performed with the Voro3D program (http://www.lmcp.jussieu.fr/~mornon/voronoi.html).

Benefits from the Analysis of the Macromolecular Interface

Starting with two separate unbound proteins, or a protein and a substrate, protein docking determines a bound structure complex. Shape complementary can be used to find thigh match between the molecular surfaces of the two interacting structures. This is the key to understand molecular recognition in different biological and medical relevant processes such as: enzyme-substrate interaction; hormone-receptor interaction and protein-DNA interaction. There exist several approaches and methods for studying macromolecular interfaces such as: the web resource iPfam, allowing the investigation of protein interactions in the PDB structures at the level of (Pfam) domains and amino acid residues (http://www.sanger.ac.uk/Software/Pfam/iPfam); MolSurfer, which establish a relation between a 2D Map (for navigation) and the 3D molecular surface (http://projects.villa-bosch.de/mcm/software/molsurfer); and the approach based on the interactive connection of the interface contact matrix with the 3D structure [149-152].

13. QUANTUM THEORETICAL METHODS

High-throughput methods for the determination of protein structures and the structure of protein-ligand complexes provide a lot of information to build the structure-activity relationships (SAR). Details at the atomic and electronic levels of the protein active sites, needed for deeper understandings of the processes that remain unrevealed after structural elucidation, are provided by methods based on quantum theoretical calculations. Such details are for example: electro negativities; polarities; electron transfer; molecular electrostatic potentials; exited energy states and bond breaking. Therefore the information provided by structural analysis is complemented by the information resulting from quantum mechanical (QM) calculations. The behaviour of atoms and molecules are beyond the realm of our everyday experience and they are not directly accessible, due to principle reasons, for us. Modern computers give us the possibility to visualize these phenomena and let us observe events that cannot be witnessed by any other means. However, one has to be aware of the fact that the visualizations depict the mathematical objects describing reality, not reality itself. Computer visualizations of such energetically phenomena in proteins are of special interest, allowing insights into the dynamical behaviour of molecules.

There exist two alternative ways for the calculation of molecular properties by quantum theoretical methods: the Hartree-Fock (HF) method based on the electron state functions and the density functional theory (DFT) method based on the electron density. Both methods use the concept of independent electrons.

QM/MM Partition

Because proteins contain thousands of atoms, a full treatment on the quantum level is beyond the realm of the available computers. Therefore mixed quantum/classical (QM/MM) methods are chosen [153,154]. These calculations mix a quantum mechanical procedure of the selected part with a classical description (often based on molecular force fields) of the rest of the molecule. The goal of the QM/MM approach is the portioning of the molecule into a small re-

gion of interest, where a quantum description is required (for example active sites) and the bulk of the system which is treated classically (Fig. **21**).

Electron Dynamics in Molecules

Electronic energy can not be treated by classical mechanics calculations. Electronic energy must be computed by solving the quantum mechanical Schrödinger equation [155-158]. For a molecule with N electrons and K nuclei, described by the state function Ψ, the Schrödinger equation is given by:

$$H\Psi = E\Psi$$

The Hamiltonian operator H involves the kinetic energy of the electrons and the nuclei, the interaction of the electrons with the nuclei, the electron-electron interactions and the nuclei-nuclei interactions:

$$H = -\sum_{i=1}^{N}\frac{\hbar^2}{2m}\Delta_i - \sum_{a=1}^{K}\frac{\hbar^2}{2M_a}\Delta_a - \sum_{i=1}^{N}\sum_{a=1}^{K}\frac{Z_a e^2}{r_{ai}} + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j>i}}^{N}\frac{e^2}{r_{ij}} + \sum_{a=1}^{K}\sum_{\substack{b=1\\ b>a}}^{K}\frac{Z_a Z_b e^2}{R_{ab}}$$

The first two terms represent the kinetic energy of the electrons (mass: m) and the nuclei (mass: M_a). ($\hbar = h/2\pi$, h is Planck's constant). The next three terms represent the potential energies of the electron-nuclei interactions (Z_a is the nuclear charge in atom a, e is the electron charge), the electron-electron interactions and the nuclei-nuclei interactions (Fig. **22**).

The first step in solving the Schrödinger equation is the separation of the coupling between the nuclei and electronic motion (Born-Oppenheimer approximation). Due to this approximation the Schrödinger equation is separated in an electronic part and a nuclear part. Then the electronic properties can be calculated with the nuclear positions as parameters where the nuclei coordinates are provided by the PDB data files. The resulting major part of the further calculation remains in solving the electronic Schrödinger equation for a given set of nuclear coordinates. The Hamiltonian operator for the electronic Schrödinger equation is then given by:

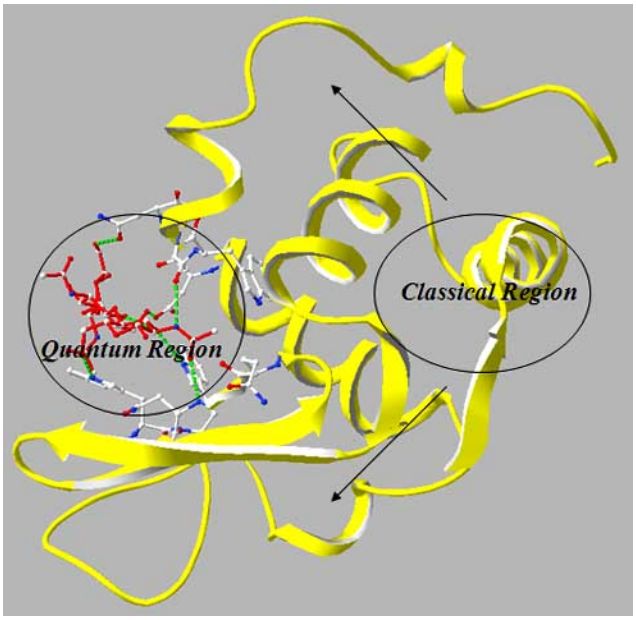


Fig. (21). Lysozyme (yellow) complexed with the inhibitor Tri-N-Acetylchitotriose (red). Resolution 1.75 angstrom, R-value = 0.229. Lysozyme is an enzyme with a bactericidal action in blood plasma. The molecules of the complex are partitioned into quantum and classical regions. The functional residues, involved in the binding of the inhibitor, and the inhibitor molecule lies within the quantum region and the backbone atoms are treated classical (often by using molecular force fields).

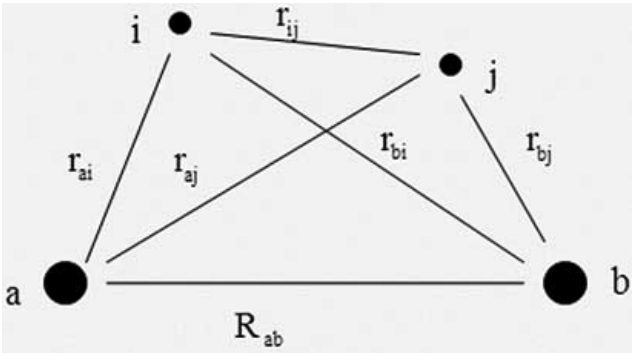


Fig. (22). The interactions between the different components (electrons i and j, nuclei a and b) are shown. The electrons interact with the nuclei (r_{ai}, r_{bi}, r_{aj}, r_{bj}) and with each other (r_{ij}), just as the nuclei (R_{ab}).

$$H = -\sum_{i=1}^{N}\frac{\hbar^2}{2m}\Delta_i - \sum_{i=1}^{N}\sum_{a=1}^{K}\frac{Z_a e^2}{r_{ai}} + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j>i}}^{N}\frac{e^2}{r_{ij}}$$

The only possible values for an observable (for example the energy) in quantum theory are the eigenvalues of the corresponding operator (in case of the energy, the Hamiltonian operator) in the eigenstate equation (described by eigenfunctions). The predicted values of the observable are then the expectation values (mean values) of the operator in pure (eigenstates) or mixed states (linear combination of eigenstates). The energy of a molecular system, described by an appropriate normalized state function, is therefore calculated as the expectation value of the Hamiltonian operator (in Dirac bra-ket notation):

$$E = \langle\Psi|H|\Psi\rangle = \int\Psi^*(r_i)H\Psi(r_i)dV$$

To simplify the expression, the coordinate of the i-the electron (r_i) is written as: (i). Ψ(i) is the state function (wave function), which is an element of a complex Hilbert-space $\mathcal{H}$. One important property of the state functions is that they are orthogonal:

$$\langle\Psi_k(i)\,|\,\Psi_l(i)\rangle = \delta_{kl}$$

The physical interpretation of the state function is that $\Psi^*(i)\Psi(i)dV$ gives the probability of finding the electron i in the space volume element dV. It is therefore the square of the state function rather than the state function itself that is related to a physical measurement. The probability of finding the electron in a volume V is given by:

$$\langle i\,|\,i\rangle = \int_V\Psi^*(i)\Psi(i)dV$$

The electron interactions in a many electron system are complicated and require sophisticated computational methods.

Independent Electron Model

A significant simplification can be obtained by introducing an independent-particle model [155-158]. In this model, the dynamic of an electron is considered to be independent of all other electrons and the interaction is taken into account in an average fashion. That means that the electron moves independently in the electric field of the nuclei and the field of all other electrons. In the independent-particle model, the total state function of a molecule with N electrons is the product of the N single electron state functions:

$$\Psi = \prod_{n=1}^{N} \Psi_n(n)$$

Because the electrons are indistinguishable and due to their fermionic nature (half spin value) the total state function must be antisymmetric. This leads automatically to the Pauli Exclusion Principle. By taking into consideration that the state functions are orthogonal, the energy of the molecular system is given by:

$$E = \sum_{k=1}^{N} \left\langle \Psi_k(i) \middle| -\frac{\hbar^2}{2m}\Delta_i - \sum_{a=1}^{K} \frac{Z_a e^2}{r_{ai}} \middle| \Psi_k(i) \right\rangle + \sum_{k=1}^{N} \sum_{\substack{l=1 \\ l>k}}^{N} \left[\left\langle \Psi_k(i)\Psi_l(j) \middle| \frac{e^2}{r_{ij}} \middle| \Psi_k(i)\Psi_l(j) \right\rangle - \left\langle \Psi_k(i)\Psi_l(j) \middle| \frac{e^2}{r_{ij}} \middle| \Psi_l(i)\Psi_k(j) \right\rangle \right]$$

To shorten the expression and making it clearly arranged, it can be written:

$$E = \sum_{k=1}^{N} \left\langle \Psi_k(i) \middle| h_i \middle| \Psi_k(i) \right\rangle + \sum_{k=1}^{N} \sum_{\substack{l=1 \\ l>k}}^{N} [\left\langle \Psi_k(i)\Psi_l(j) \middle| h_{ij} \middle| \Psi_k(i)\Psi_l(j) \right\rangle - \left\langle \Psi_k(i)\Psi_l(j) \middle| h_{ij} \middle| \Psi_l(i)\Psi_k(j) \right\rangle]$$

The first term describes the mean of the kinetic energy of the electrons and their potential energy in the electrostatic field of the nuclei:

$$\left\langle \Psi_k(i) \middle| h_i \middle| \Psi_k(i) \right\rangle = \int \Psi_k^*(i) \left[-\frac{\hbar^2}{2m}\Delta_i - \sum_{a=1}^{K} \frac{Z_a e^2}{r_{ai}} \right] \Psi_k(i)\, dV_i$$

The second term describes the Coulomb interaction between 2 electrons, where one electron is located in the k-orbital and the second in the l-orbital (the charge is distributed inside the orbitals):

$$\left\langle \Psi_k(i)\Psi_l(j) \middle| h_{ij} \middle| \Psi_k(i)\Psi_l(j) \right\rangle = \iint \Psi_k^*(i)\, \Psi_k(i) \frac{e^2}{r_{ij}} \Psi_l^*(j)\, \Psi_l(j)\, dV_i\, dV_j$$

The third term describes the exchange interaction resulting from the indistinguishability of the electrons (this is a pure quantum effect and has no classical analogue):

$$\left\langle \Psi_k(i)\Psi_l(j) \middle| h_{ij} \middle| \Psi_l(i)\Psi_k(j) \right\rangle = \iint \Psi_k^*(i)\, \Psi_l^*(j) \frac{e^2}{r_{ij}} \Psi_l(i)\, \Psi_k(j)\, dV_i\, dV_j$$

To solve the Schrödinger equation for a molecular system, the best state functions (molecular orbitals) are determined by a variation principle to find the energy minimum:

$$\delta \left[\left\langle \Psi \middle| H \middle| \Psi \right\rangle - \sum_{k=1}^{N} \varepsilon_k \left\langle \Psi_k(i) \middle| \Psi_k(i) \right\rangle \right] = 0$$

From the variation results the Hartree-Fock (HF) equation system, which describes the energy ε_k of single independent electrons:

$$\left\{ -\frac{\hbar^2}{2m}\Delta_i - \sum_{a=1}^{K} \frac{Z_a e^2}{r_{ai}} + \sum_{\substack{l=1 \\ l\neq k}}^{N} \int \frac{e^2 \Psi_l^*(j)\Psi_l(j)}{r_{ij}} dV_j \right\} \Psi_k(i) - \left\{ \sum_{\substack{l=1 \\ l\neq k}}^{N} \int \frac{e^2 \Psi_l^*(j)\Psi_k(j)}{r_{ij}} dV_j \right\} \Psi_l(i) = \varepsilon_k \Psi_k(i)$$

This is formally a one electron eigenvalue equation for a molecular system. The HF equations describe independent electrons moving in an effective potential provided by the nuclei and the other electrons (one term describes the Coulomb interaction, the second term results from the exchange interaction). Solutions of the HF equation system are the one electron molecular orbitals $\Psi_k(i)$.

HF- Linear Combination of Atomic Orbitals

The molecular orbitals are expanded in terms of the basis functions, the atomic orbitals:

$$\Psi_n(i) = \sum_{\mu=1}^{M} c_{n\mu} \phi_\mu(i)$$

This expansion is called "Linear Combination of Atomic Orbitals" (LCAO) [155-158]. Until now the molecules were considered as a system of electrons and nuclei, the expansion in atomic orbitals reflects the chemical view where molecules are building up of single atoms. The coefficients $c_{n\mu}$ define the contribution of the atomic orbitals to the molecular orbitals and are the weight of the μ-the atomic orbital in the n-the molecular orbital. These coefficients have to be determined to find the molecular orbitals. By multiplying the Hartree-Fock equations from left by a specific basis function and integrating out yields the HF-LCAO equation system (also called Roothan-Hall equations) for the coefficients:

$$\sum_{\beta=1}^{M} \left(H_{\alpha\beta} - \varepsilon_n S_{\alpha\beta} \right) c_{n\beta} = 0$$

These are the HF equations in the atomic orbital basis in matrix notation. The solutions of the HF-LCAO equation system yield the energy $\varepsilon_n; (n = 1,...,M)$ and the coefficients $c_{n\beta}; (\beta = 1,...,M)$ for the molecular orbitals $\Psi_n; (n = 1,...,M)$. The matrix elements of the Hamiltonian are given by:

$$H_{\alpha\beta} = \left\langle \phi_\alpha(i) \middle| h_i \middle| \phi_\beta(i) \right\rangle + \sum_{\rho=1}^{M} \sum_{\sigma=1}^{M} P_{\rho\sigma} \left[\left\langle \phi_\alpha(i)\phi_\rho(j) \middle| h_{ij} \middle| \phi_\beta(i)\phi_\sigma(j) \right\rangle - \frac{1}{2} \left\langle \phi_\alpha(i)\phi_\rho(j) \middle| h_{ij} \middle| \phi_\sigma(i)\phi_\beta(j) \right\rangle \right]$$

$S_{\alpha\beta}$ is the overlap matrix:

$$S_{\alpha\beta} = \left\langle \phi_\alpha \middle| \phi_\beta \right\rangle$$

$P_{\alpha\beta}$ is the density matrix, which is defined by:

$$P_{\alpha\beta} = 2 \sum_{n=1}^{N/2} c_{n\alpha}^* c_{n\beta}$$

The matrix elements $H_{\alpha\beta}$ contain one-electron integrals:

$$\left\langle \phi_\alpha(i) \middle| h_i \middle| \phi_\beta(i) \right\rangle = \int \phi_\alpha^*(i) \left[-\frac{\hbar^2}{2m}\Delta_i - \sum_{a=1}^{K} \frac{Z_a e^2}{r_{ai}} \right] \phi_\beta(i) dV_i$$

and two-electron (electron interactions: Coulomb and exchange) integrals:

$$\left\langle \phi_\alpha(i)\phi_\rho(j)\left|h_{ij}\right|\phi_\beta(i)\phi_\sigma(j)\right\rangle = \iint \phi_\alpha^*(i)\,\phi_\rho^*(j)\frac{e^2}{r_{ij}}\phi_\beta(i)\,\phi_\sigma(j)\,dV_i dV_j$$

$$\left\langle \phi_\alpha(i)\phi_\rho(j)\left|h_{ij}\right|\phi_\sigma(i)\phi_\beta(j)\right\rangle = \iint \phi_\alpha^*(i)\,\phi_\rho^*(j)\frac{e^2}{r_{ij}}\phi_\sigma(i)\,\phi_\beta(j)\,dV_i dV_j$$

The HF-LCAO equations are solved iteratively (Self-Consistent Field theory: SCF). This is done by, first guess a set of coefficients $c_{n\mu}$, form $H_{\alpha\beta}$, calculate a new set of coefficients, form a new $H_{\alpha\beta}$, and so on until the coefficients remain the same (they are consistent).

Once the coefficients $c_{n\mu}$ are calculated, the molecular orbitals Ψ_n can be built up by linear combination of the atomic orbitals ϕ_μ. The molecular orbitals are arranged in order of increasing energy ε_n (Fig. **23**). Then the electrons are assigned to the orbitals beginning with the lowest energy. As a result, in a system with N electrons, the N/2 lowest molecular orbitals are occupied by respectively up to two electrons with opposite spins (Fig. **23**). The remaining M-N/2 (virtual) orbitals are unoccupied in the ground state and are occupied in excited electronic states. Physically, the square of the orbitals represents the electron spatial density distribution or in other words: the probability for finding the electron in a certain space volume element. The calculated molecular orbitals (canonical orbitals) are delocalized; by a linear combination of the molecular orbitals, localized orbitals are obtained which describe chemical bonds such as σ-bonds and hybrid bonds (sp^3) in the convenient chemical notation.

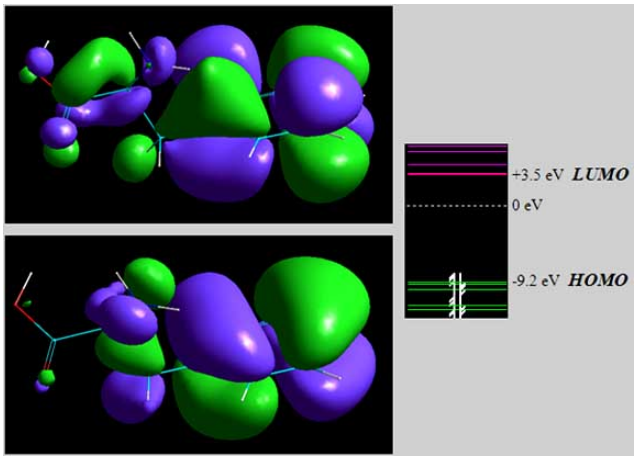


Fig. (23). 3D visualization of two individual molecular orbitals of the amino acid phenylalanine. Phenylalanine relates to the central nervous system and is used to treat schizophrenia, Parkinsons's diseases, migraines, depression and others. Positive values of the molecular orbitals are green encoded, negative values violet. The highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) are shown, which are the most important in organic chemistry. Since the square of the orbitals represents the electron density distribution, changes in the shape of the molecular orbitals show charge transfer when the molecule is excited. Additionally a part of the energy levels of the bounded electrons and of the excited states are shown. By exciting the electrons, the energy differences determine the electronic spectrum of the molecule. The ab initio calculation was done by the HF-LCAO method with a 3-21G basis set.

The HF method averages over electron repulsions and every electron moves in the averaged field of the remaining electrons. This is of course an approximation because the electron repulsion is correlated and every individual electron interacts with each single electron. The error resulting from this approximation (the real energetically ground state lies deeper than the calculated) can be corrected by sophisticated and time consuming methods like the configuration interaction expansion (CI) where the occupied molecular orbitals are systematically replaced by excited ones. The difference between the true ground state and the calculated is the correlation energy.

A principal difficulty in solving the HF-LCAO equations is the large number (~M^4) of two-electron integrals. These integrals can involve up to four atomic centres. Therefore ab initio calculations for large molecules require enormous computer power. Semi-empirical methods reduce the computational cost by reducing the number of these integrals.

Semi Empirical Methods

A great part of semi-empirical methods are based on the Zero Differential Overlap (ZDO) approximation, neglecting all products of atomic orbitals depending on the same electron coordinates which are located on different centres:

$$\phi_\alpha^*(i)\,\phi_\beta(i)\,dV_i = \phi_\alpha^*(i)\,\phi_\beta(i)\,\delta_{\alpha\beta}\,dV_i$$

Then all three-centre and four-centre two-electron integrals, which represent the greatest part of these integrals, are neglected [155-158]. Additionally, the overlap matrix is reduced to the unit matrix. A further approximation is that only valence electrons (C, N, O: 2s, 2p orbitals, H: 1s orbital) are taken into consideration. To solve the remaining integrals, the atomic orbitals are expressed as Slater type orbitals (STO) in polar coordinates:

$$\phi(r,\vartheta,\varphi) = Nr^{n-1}e^{-\zeta_l r}Y_l^m(\vartheta,\varphi)$$

(n=1,2…). N is a normalization factor and $Y_l^m(\vartheta,\varphi)$ are the spherical harmonic functions (1 =0 for an s orbital, 1= 1 for a p orbital, m=-l,..,+l). The Slater exponents ξ_l are fitted on experimental data. The Slater type orbitals approximate the exact orbitals of the hydrogen atoms but, in contrast to these exact orbitals, they have no radial nodes. The STO's are suitable for methods where all three-centre and four-centre integrals are neglected. To compensate the error resulting from this rigorous integral approximation, the one electron parts are replaced by formulas with empirical parameters.

A suitable ZDO based method for large bio molecules is the NDDO (Neglect of Diatomic Differential Overlap) approximation, where the two electron integrals satisfy the condition:

$$\delta^{AB}\delta^{CD}\iint \phi_\alpha^{*A}(i)\phi_\beta^{B}(i)\frac{e^2}{r_{ij}}\phi_\rho^{*C}(j)\phi_\sigma^{D}(j)\,dV_i\,dV_j$$

That means an overlap is only allowed between orbitals on the same centres (A,B,C,D). NDDO retains all one-centre overlap terms when Coulomb and exchange integrals are calculated (Fig. **24**).

Density Functional Theory

In the HF formalism the problem of finding the ground state of a molecule consist of finding the lowest energy value

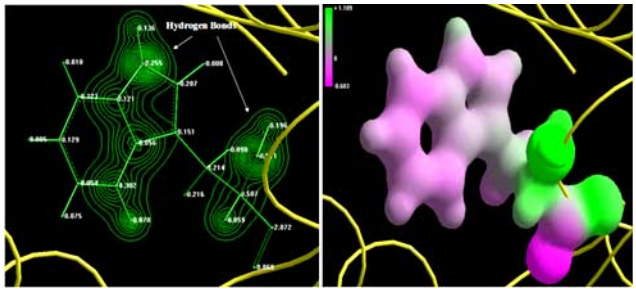


Fig. (24). The electron density of the active site residue tryptophan (Trp 63) of the lysozyme molecule is visualized as a 3D plot (right picture) and as a contour plot (left picture). The contour plot shows the values of the spatial electron density on a plane that is parallel to the plot. The plane is specified through the centre of mass of the residue. Atomic charges are included. The electron density in the 3D plot is colour encoded according to its electrostatic potential. The surface shows where in the spatial distribution the electron density has a value of $0.135\ e/a_0^3$.

and the corresponding molecular orbital [155-157]. In density functional theory (DFT) a different approach, based on the electron density, is followed. Instead of the abstract state function Ψ, the electron density ρ, related to a physical interpretation, is used. Formally the N electron problem with 3N spatial variables is reduced to the formulation of the overall electron density depending only on 3 variables. The basis of DFT is the Hohenberg-Kohn theorem, which states that, the ground state energy E of an electronic system is uniquely determined by the ground state electron density ρ:

$$E = E[\rho]$$

The functional connecting these two quantities is not generally known [159,160]. The principle of the DFT method is to design functionals connecting the energy with the electron density. The total electronic energy of a molecular system, expressed as functional of the electron density, is given by the sum of the kinetic part, the electron-nuclei interaction part and the electron-electron interaction part:

$$E[\rho] = E_T[\rho] + E_V[\rho] + E_{ee}[\rho]$$

As in the case of the HF method the electron-electron interaction is divided in a Coulomb part and an exchange part. Additionally in DFT the electron correlation energy functional is included. This makes one of the main differences to HF. Normally the exchange and the correlation part are joined together in an exchange-correlation functional. Then in DFT the total energy can be written as:

$$E[\rho] = E_T[\rho] + E_V[\rho] + E_J[\rho] + E_{XC}[\rho]$$

The first term is the single particle kinetic energy of the non interacting electrons, the second term is the potential energy of the electron-nuclei interaction, the third term is the Coulomb energy part of the electron-electron interaction and the fourth part is the combined exchange energy and correlation energy part of the electron-electron interaction. The electron kinetic energy is calculated from an auxiliary set of orbitals which is used for representing the electron density:

$$E_T[\rho] = \sum_{k=1}^{N} \int \psi_k^*(i) \left[-\frac{\hbar^2}{2m} \Delta_i \right] \psi_k(i)\, dV_i$$

These orbitals are called Kohn-Sham orbitals. Again, in order to simplify the expression, the coordinate of the i-the electron (r_i) is written as: (i). The Kohn-Sham orbitals define the electron density as (b_k is the orbital occupation number):

$$\rho(i) = \sum_k b_k e \psi_k^*(i) \psi_k(i)$$

Then the terms for the potential energy of the electron-nuclei interaction and the Coulomb energy of the electron-electron interaction can by expressed explicitly as functional of the electron density by:

$$E_V[\rho] = Z_a e \int \frac{\rho(i)}{r_i} dV_i$$

and:

$$E_J[\rho] = \frac{1}{2} \iint \frac{\rho(i)\rho(j)}{r_{ij}} dV_i dV_j$$

The term $E_{XC}[\rho]$ is the exchange-correlation energy functional. This functional represents the main problem in DFT. The exact form of the functional is unknown and therefore approximations are necessary. The local density approximation (LDA) assumes that the exchange and correlation energy of an electron at a point r_i depends only on the electron density $\rho(i)$ at that point (homogeneous electron gas). Better results are obtained by including the gradient of the electron density. This allows a better description of inhomogeneous electron densities (due to different electrostatic potentials) in atoms and molecules. A lot of such gradient-corrected functionals have been proposed during the last decades. Briefly the exchange-correlation functionals are of the general form:

$$E_{XC}[\rho] = \int \rho(i)\, \varepsilon_{XC}\left(\rho(i), \nabla\rho(i)\right) dV_i$$

Today, hybrid functionals, connecting several single functionals, are used in practice. Commonly used hybrid functionals are the B3LYP and B3PW91 functionals [156].

By minimization of the total energy with respect to the Kohn-Sham orbitals (same procedure as in the case of HF), results the Kohn-Sham (KS) equation system:

$$\left\{ -\frac{\hbar^2}{2m}\Delta_i - \sum_{a=1}^{K} \frac{Z_a e^2}{r_{ai}} + \sum_{\substack{l=1 \\ l \neq k}}^{N} e^2 \int \frac{\psi_l^*(j)\psi_l(j)}{r_{ij}} dV_j + U_{XC}[\rho] \right\} \psi_k(i) = \varepsilon_k \psi_k(i)$$

These are one electron pseudo eigenvalue equations for the one electron energy ε_k. Solutions of the KS equation system are the one electron Kohn-Sham molecular orbitals. The exchange-correlation potential is given by:

$$U_{XC}[\rho] = \frac{\partial E_{XC}[\rho]}{\partial \rho}$$

The KS equations describe electrons in Kohn-Sham orbitals moving in an effective field consisting of the electron nuclei interaction, the mean Coulomb interaction with the other electrons and the exchange-correlation potential. Under the action of the effective potential the non-interacting electrons acquire the same density as true interacting electrons. The Kohn-Sham equations are the DFT equivalence to the Hartree-Fock equations.

Similar to the HF-LCAO method the Kohn-Sham orbitals are expanded in terms of atomic orbitals. The same procedure as in the case of HF yields an equation system (DFT-

LCAO) for the coefficients $c_{n\mu}$ of the expansion. The one-electron part (kinetic energy part and nuclei potential energy) and the Coulomb energy part of the resulting Hamilton matrix $H_{\alpha\beta}$ are identical to the corresponding HF-LCAO parts. The exchange-correlation part of the matrix makes the difference and is given in terms of the electron density and its derivative.

$$U_{\alpha\beta} = \int \phi_\alpha^*(r) U_{XC}[\rho(r), \nabla\rho(r)] \phi_\beta(r)\, dV$$

This integral cannot be calculated analytically because the functional depends implicitly on the integration variables via the electron density. Therefore the exchange-correlation part must be evaluated numerically on a lattice.

$$U_{\alpha\beta} \approx \sum_l^G U_{XC}[\rho(r_l), \nabla\rho(r_l)] \phi_\alpha^*(r_l) \phi_\beta(r_l) \Delta v_l$$

For an infinity number of lattice points G, the approximation becomes exact. In practise the selected number of lattice points r_1 is based on the desired accuracy. Typically 1000-10000 lattice points are used for each involved atom.

As atomic orbitals generally Gaussian basis functions are used for the calculation. Gaussian-type orbitals (GTO) are similar to Slater-type orbitals but they have an exponential factor that goes as the square of the distance between the electron and the orbital centre:

$$\phi(r) = N x^u y^v z^w e^{-\alpha r^2}$$

($u + v + w = 0$ corresponds to an s orbital, $u + v + w = 1$ corresponds to a p orbital, etc.). Mostly linear combinations of Gauss functions are used like 3-21G: the inner orbitals are described by 3 Gauss functions which are combined to one and the valence orbitals are also described by 3 functions where 2 are combined to one and the third is used singly.

Formally the DFT-LCAO calculation runs like a HF-LCAO calculation and like those are solved in a self-consistent (SCF) procedure; a major difference is that the numerical integration of the exchange-correlation part has to be done at the end of each cycle.

Of the many quantum theoretical methods available, DFT has over the past decade become a key method for bio molecules. The advantage over the HF method is that no costly configuration interaction expansion is necessary because in DFT the electron correlation is already included explicitly in the exchange-correlation functional. The exchange-correlation energy is the only unknown functional in DFT. Because the exchange-correlation energy is only a relatively small part of the total energy, even crude approximations for this term provide quite accurate results. Today, DFT is the "routine" method for the calculation of molecular properties.

Electronic Properties of Molecules

Properties of bio molecules derived by quantum theoretical calculations are: electron density distribution; atomic charges; electric dipoles; molecular electrostatic potentials; electronic spectra and vibration spectra. The properties can be calculated out of the molecular orbitals, in HF or in DFT, once these are known. Their values are the expectation values of the corresponding operator.

A first example of a molecular property is the electric dipole moment, which is a measure for the polarity of a molecule or residue:

$$\mu = \sum_n \int \Psi_n^*(r)\, r\, \Psi_n(r)\, dV$$

The electron density distribution represents the probability of finding an electron at a point r in space. It is calculated by:

$$\rho(r) = \sum_n \Psi_n^*(r) \Psi_n(r) = \sum_\alpha \sum_\beta P_{\alpha\beta} \phi_\alpha(r) \phi_\beta(r)$$

The electron density is usually expressed in units of e/a_0^3 where a_0^3 is the Bohr hydrogen atom radius (0.5292 angstrom). For the visualization, surfaces with constant density values are shown (Fig. **25**). The detection of regions with high electron densities in molecules enables the evaluation of polarities, electro negativities etc.

By definition, the electrostatic potential is calculated by:

$$V(r) = \sum_a \frac{Z_a}{|R_a - r|} - \int \frac{\rho(r')}{|r' - r|} dr'$$

In a molecule, the nuclei are represented as positive point charges Z_a surrounded by a continuous distribution of electron charge $\rho(r')$. The electrostatic potential is expressed as potential map, showing constant contouring values (Fig. **25**). The electrostatic potential is a measure of the force exerted by the protein on other molecules (proteins, ligands etc.), enabling an understanding of intermolecular interactions. Computing the electrostatic potential of a protein allows the calculation of quantities such as: individual amino acid pKa values (pH value at which an acidic or basic residue loses or gains a proton and therefore creating the chemical conditions for a specific function); intermolecular binding energies and specific protein-protein (protein-substrate) recognition. Unusually strong electrostatic potentials (or unusual pKa values) indicate regions of catalytic importance.

From the HF-LCAO or DFT-LCAO equation system, the coefficients $c_{n\alpha}$, determining the molecular orbitals, are calculated out. The charge distribution of the electrons in a molecule is determined by the Mulliken population analysis. The brute population of a single atomic orbital (ϕ_α) is defined by.

$$n_\alpha = \sum_{n=1}^M b_n c_{n\alpha}^2 + \sum_{n=1}^M b_n \sum_{\beta=1}^M c_{n\alpha} c_{n\beta} S_{\alpha\beta}$$

b_n is the occupation number of the n-te orbital. From the summing up of the coefficients over all the atomic orbitals of an atom, results the brute atomic population. Then the atomic charges are obtained by subtracting the brut atomic population from the number of valence electrons of the neutral atom (Fig. **24**). If there is an access of electrons on the atom then the atomic charge is negative. From a lack of electrons results a positive atomic charge. The second term in the equation defines the overlapping population between two atomic orbitals which can be used as measure for the binding force (σ-bonds and π-bonds) between the two atoms.

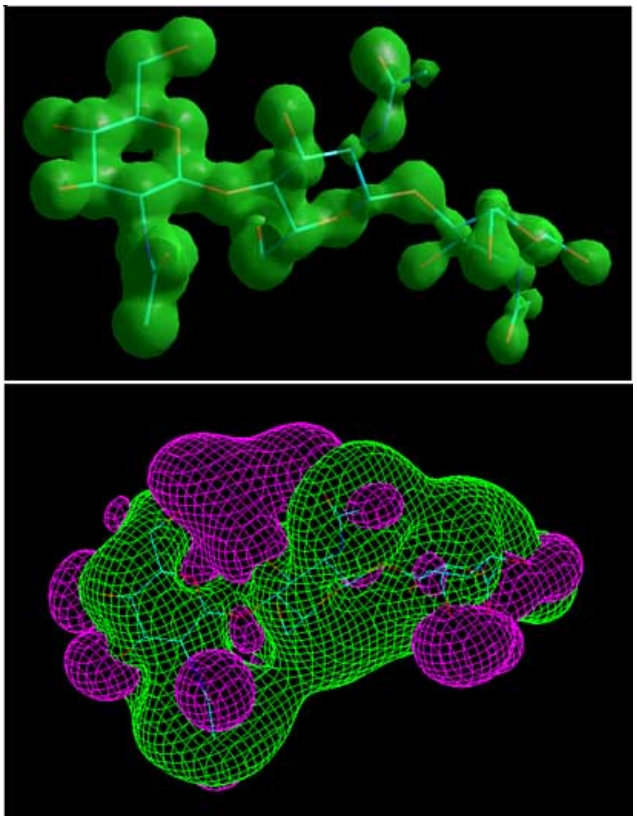


Fig. (25). Top: The electron density distribution of the inhibitor Tri-N-Acetylchitotriose bounded to lysozyme, calculated by density functional theory (DFT) by use of the B3LYP hybrid functional. The surface shows electron density values of 0.08 e/a_0^3. The atoms are coloured according to their types: carbon: white, nitrogen: blue, oxygen: red. Bottom: The spatial distribution of the electrostatic potential surrounding the molecule. Positive values are red encode, negative values green. The surface shows where in 3D the electrostatic potential has a value of 0.015 e/a_0.

The total energy of a molecule is given by the sum of the molecular orbital energies:

$$E = \sum_{n=1}^{M} b_n \varepsilon_n$$

When an electron in an occupied molecular orbital is excited, it changes into an unoccupied orbital and the difference between the energy states determines the electronic spectrum:

$$\Delta\varepsilon_{mn} = \varepsilon_m - \varepsilon_n$$

By comparing the calculated excitation spectra (in the visible light and UV light) with the experimentally measured spectra, it can be evaluated how realistic calculated intermediate conformations, binding adducts and alternative reaction mechanisms are.

Structural Analysis Complemented by Quantum Theoretical Methods

Quantum theoretical calculations aid the exploration of processes that provide the link between structural analysis and physicochemical and biological functions. They are used to understand enzyme mechanism (bond breaking and bond formation), ligand binding (hydrogen bonding), ligands that interfere with ion channels (large electric fields), photochemistry of psoralen compounds (electronic spectra) and many other fundamental mechanism in biological medical context. Based on studies of alternative reaction mechanisms for orbital interactions, binding modes, electron transfer, polarization effects, intermediate conformations and effects on energy barriers, the model that best matches the experimental data is selected. The role of the different parts of the reaction mechanisms can be determined, which is information that normally can not be extracted from structural analysis alone. Structural analysis completed by quantum theoretical methods provided insights into catalytic and binding mechanism for a lot of medical relevant enzymes such as: human aldose reductase; glutathione S-transferase; influenca neuraminidase; human thrombin; uracil-DNA glycosylase; thymidine kinase inhibitors and many more [161-165].

Additionally it has been observed that pure quantum effects, like tunnelling, are significant in the catalytic mechanism of several enzymes. Quantum tunnelling effects occur, for example, in enzymatic proton transfer through the reaction barrier in a number of enzyme systems. In ligands, the C-H bond breakage occurs by extreme quantum tunnelling. If a quantum object, described by the state function Ψ, with a kinetic energy E cannot overpass a potential energy barrier, with height V, it will tunnel through the barrier a height below the maximum. This is possible due to the wave character of the quantum object. The narrower the barrier, the smaller the mass of the particle and the smaller V, the greater the tunnelling probability. Such reactions can be modelled by using the potential energy surface of the molecular system and incorporating a correction factor to account for tunnelling below the saddle point (representing transition states) of the energy surface. Vibrational motions of the protein scaffold play a role in driving proton tunnelling reactions in enzymes. Reaction rates and kinetic effects of proton and hydride tunnelling have been studied in: liver alcohol dehydrogenase; enolase; methylamine dehydrogenase and more [166-169]. Such calculations have provided a detailed atomic level picture of the factors which are playing a role in tunnelling effects in enzymes.

Special interest is nowadays given to catalytic and active centres in proteins containing metal atoms such as: haemoglobin, myoglobin and DNA binding proteins, involved in the control of transcription processes.

There are a number of commercial software tools that perform quantum theoretical calculations on molecules. Two well known packages, which are used in the scientific community for calculations on proteins, are GAUSSIAN (http://-www.gaussian.com) and HyperChem (http://www.hyper.-com). Both are running on PC or UNIX workstation. They can start from PDB data files.

14. MOLECULAR MEDICINE

The progress in knowledge about the molecular mechanism of metabolism and differentiation processes has influenced modern medicine. The reduction of reasons for diseases to processes at the molecular level offers the possibility to systematic interventions providing new ways in preventive, diagnostic and therapeutic medicine. The principle

of molecular medicine and the benefits from structural analysis will be demonstrated in the following on hand of selected examples of enzyme inhibitions, structural deviations, mutations etc.

In clinical bioinformatics, proteomic analysis is of relevant value only when connected to clinical data [170]. It provides biological and medical information to enable for individual healthcare. The influence of bioinformatics on clinical research and routine is highly dependent on the access and retrieval of genomic and proteomic information and the representation of the biomedical information in an appropriate fashion. Information about genes and diseases is available at: http://www.ncbi.nlm.nih.gov/disease.

Non-Steroidal Anti-Inflammatory Drugs

Structural diversities in the proteome (the whole set of proteins encoded by a genome) are determined by several thousands of SNP (Single Nucleotide Polymorphisms) in the genome. This leads to an individualisation of the organism concerning, among others, the metabolism, the receptor and the transport system. This is of special importance for the planning of an individual drug treatment. Better medication can be developed once the structures of binding sites from target proteins, involved in drug metabolism, are known [171-175]. By use of scoring functions, based on an estimation of the free binding energies, a rapid evaluation of protein-drug interactions can be done by docking the drug molecule into the target protein. The beneficial effects of drugs are based on molecular recognition and binding of ligands to the active sites of specific targets, such as: enzymes; receptors and nucleic acids. The effect of binding can be inhibition of enzymatic activity, signal transduction and molecular transport. For example: Aspirin and other non-steroidal anti-inflammatory drugs (NSAID) inhibit two prostaglandin-cyclo-oxygenases: COX-1 and COX-2 [176]. COX-1 is expressed in the gastric mucous membrane. During inflammations COX-2 is induced. A well known undesirable secondary effect of aspirin is the occurrence of bleeding of the stomach wall [177]. Active agents, which inhibit only COX-2 and not COX-1, are therefore desirable [178-180]. The selective inflammation inhibitor SC-558 (1-phenylsulfoamid-3-trifluormethyl-5-parabromophenylpyrazole) inhibits COX-2 and not COX-1. The active centre of COX-1 differs (among others) from the appropriate centre of COX-2 by an interchanging of the residue valanine by isoleucine. The side chain of isoleucine is responsible for a steric collision with SC-558 and therefore no binding is possible (Fig. **19**).

There are many other successful examples in which structural bioinformatics tools were used to timely provide very useful information for developing drugs to treat various diseases, such as Alzheimer disease [181-185], depressing [186], schizophrenia [187], SARS [188,189], diabetes [190], ion channel disorders (such as long QT and chronic pain) [191], AIDS [192], and influenza [193-195]. The structural bioinformatics tools were also utilized to provide useful information for studying caspase inhibitors [196,197], antiviral inhibitors [198] and personalized drug design [199,200]. In computer aided drug design, quantum theoretical methods are widely used tools that are applied: to calculate accurate force field parameters; to describe the electronic structure of a molecule; to perform conformational analysis; to calculate atomic point charges which are used to study binding properties and to investigate charge transfer processes, which can play an important role in molecular recognition [201-204].

Human Immunodeficiency Virus

The HIV (human immunodeficiency virus) virus is the pathogen of the acquired immune deficiency syndrome (AIDS) where patients suffer from neurological syndromes, profuse perspiration, attack of fever etc. [205]. Among other molecules, the HIV virus contains RNA and the reverse transcriptase (Fig. **26**). The reverse transcriptase performs the transcription of the one stranded RNA code into a double stranded DNA which is inserted into the genome of the host cell [206]. Then, via regular transcription, copies of the virus RNA are produced and new virus particles are synthesized.

The HIV protease is one of the three enzymes which are coded by the RNA of the HIV virus and is essential for the proliferation of HIV [207]. The protein is a homo dimmer and consists mainly of β-sheets and the active site which is located in a kind of hole (Fig. **26**). Therapies are based on the inhibition of the HIV protease, so that syntheses of new virus particles are prevented [208-210]. By knowing the structure of the protease molecule, appropriate inhibitor molecules can be designed. The search for possible candidates of inhibitors showed that the well known molecule pepstatine (inhibitor of several proteases) inhibits the HIV protease too. Pepstatine contains the non natural amino acid statine. For the development of inhibitors, this central functional group is retained and the supporting rest structures are varied according to the observed substrate specificity of the HIV protease. Structural analysis of the resulting protease-inhibitor complexes enables an evaluation of the binding forces and delivers indications for the improvement of the inhibitor, for example by engineering additional hydrogen bonds.

The development of the commercially available inhibitors, such as: saquinavir (ROCHE); lopinavir; ritonavir and

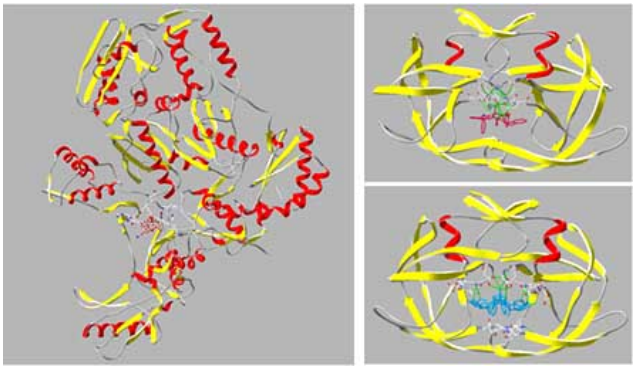


Fig. (26). The HIV reverse transcriptase (left) transforms the RNA of the virus into DNA. The HIV protease (right) is encoded in the RNA. Therapies are based on the inhibition of the HIV protease. The picture on top shows the HIV protease (resolution: 2.7 angstrom, R-value = 0.198) in complex with the inhibitor Bea450 (red). The bottom picture shows the inhibitors saquinavir and darunavir (blue) bounded to the HIV protease (resolution: 1.1 angstrom, R-value = 1.69). The inhibitor molecules are bounded to the protease molecule by hydrogen bonds.

tipranavir; are based on such structural analysis, occasionally combined with quantum theoretical methods [211,212]. The HIV protease database contains structures of the HIV protease and their complexes with inhibitors, together with analysis tools and information about AIDS: http://mcl1.ncifcrf.gov/hivdb/.

Binding of Oxygen, Carbon Monoxide and Nitric Monoxide in Myoglobin

Myoglobin stores oxygen in the muscular tissue. The biological function of myoglobin is the binding and release of oxygen, which occurs in the active centre: the heme (Fig. **27**) [213]. The heme contains an iron-porphyrin (FeP) compound in its centre. The heme iron is covalently bounded to the nitrogen atom of the proximal histidine residue 93 (HIS93). The distal side (face to hystidine 64: HIS64) of the iron-porphyrin is free and ready to bind oxygen (O_2) or carbon monoxide (CO) or nitric monoxide (NO) as ligands. Details of the binding of these diatomic molecules to the heme can not be understood with structural analysis alone. Therefore, the structural analysis is complemented by quantum theoretical calculations based on density-functional theory with LDA approximation, combined with a molecular force field [214-216]. This enables the analysis of dynamic short-timescale processes in the myoglobin.

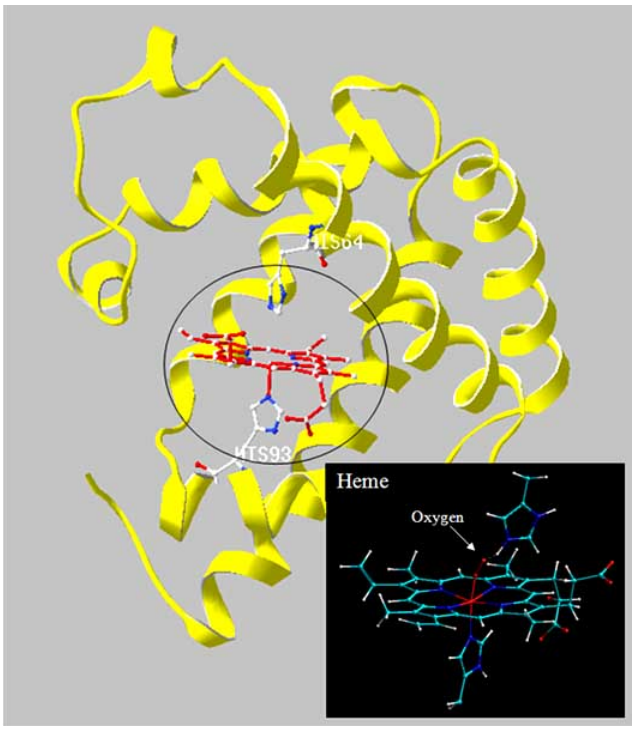


Fig. (27). The biological function of myoglobin is the storage of oxygen in muscles (resolution: 1.25 angstrom, R-value = 0.136). The heme (red) is the active site and bounds the oxygen molecule. The heme is bounded to the proximal histidine (HIS93). On the other side the heme is faced to the distal histidine (HIS64). Quantum calculations based on functional density theory show that the proximal histidine increases the binding strength of $Fe-O_2$ and $Fe-CO$ whereas the contrary is true for $Fe-NO$. The molecules O_2 and CO are stabilized by the interaction with the distal histidine (small picture). The larger O_2 interaction due to hydrogen bonding seems to be the origin of the CO discrimination of myoglobin.

Of special interest is, the quantification of the interplay between the structure, energy and dynamics of the binding of O_2, NO, CO to the heme active centre. The calculation shows that the heme porphyrin substituents do not affect the structural and electronic properties of Fe-ligand bonds. But the proximal hystidin residue increases the binding strength of the Fe-CO and Fe- O_2bonds as opposed to Fe-NO. The lowest energy spin-state was found to be a triplet for FeP, singlet for FeP-CO and FeP- O_2 and a doublet for FeP-NO. The Fe-ligand complexes are characterized by having a curved porphyrin. This distortion reinforces the bonding between the Fe d_{z^2} orbital and the orbital of the diatomic molecule. The heme-CO structure is not affected by the conformation of the distal pocket. Conversely the CO stretch frequency and the strength of the CO-HIS64 interaction are highly dependent on the orientation and tautomerization of the distal hystidin. HIS64 is pronated at the nitrogen atom and O_2 and CO are stabilized by interaction with HIS64. The larger interaction of O_2 leads to the conclusion that hydrogen bonding is the origin of myoglobin discrimination of CO.

The results are in agreement with thermodynamic and spectroscopic data. Briefly the quantum calculations enables an understanding of the structure and dynamics of the Fe-ligand bonds, the role of the proximal and distal hystines and the CO stretch bands in the infrared spectrum of myoglobin.

QT-Syndrome

The QT-syndrome is a disease based on the defect of the rhythm control of the heart. The patients suffer from an accelerated heart rhythm [217,218]. This defect results from a structural deviation of the potassium channel KvLQT1 of the heart muscle, due to a mutation in the corresponding gene [219]. More than 50 genes coding a potassium ion channel have been found in different organisms. The relation of sequence analysis with the structural information plays an important role to understanding such diseases. Other ion channel related diseases are: Charcot-Marie-Tooth disease (Calcium channel) and Myotonie (Chloride channel CIC-2) [220,221].

Sickle Cell Anaemia

SNP's are mutations where the difference between genes is restricted to the exchange of a single base pair. Several SNP's are involved in diseases. A mutation, due to an exchange of the nucleotide adenine by thymine in the gene of β-globine, results in a replacement of the residue glutamic acid through valine in position 6 of the beta chain of the protein. This leads to a "sticky surface" on the haemoglobin molecule resulting in a polymerisation of the molecules. From this originates the sickle cell anaemia where patients suffer under colic like pain attacks and haemolytic crisis [222,223]. Information about disease associated SNP's are available at: http://snp.cshl.org/.

Anti-Cancer Drug

Cisplatin ($PtCl_2(NH_3)_2$ is an anti-cancer drug, which is effective against tumours in the sex glands, head and neck. The target of the drug is cellular DNA, where it distorts the tertiary structure and inhibits replication and transcription

(Fig. **28**). Several intrastrand (binding to 5´-GG, 5´-AG and 5´-GXG, G: guanine; A: adenine; GXG: non-adjacent GG) adducts and interstrand (cross-linking binding between the two strands at GG nucleotides) adducts are formed by binding cisplatin to DNA. To understand the reasons for the efficacy of cisplatin, details of the cisplatin activation and its binding to DNA are necessary.

Studies based on density functional theory investigated several solvation mechanisms, intermediate conformations and the effects of different adducts on reaction energy barriers [224-227]. The activation of cisplatin is a two step process, the first and the second aquation:

$$Pt[Cl]_2[NH_3]_2 + 2H_2O \rightarrow Pt[Cl][NH_3]_2[H_2O]^+ + Cl^- + H_2O$$

$$Pt[Cl][NH_3]_2[H_2O]^+ + Cl^- + H_2O \rightarrow Pt[NH_3]_2[H_2O]_2^{2+} + 2Cl^-$$

The water-substitution reactions of cisplatin proceed by way of trigonal pyramidal transition states. The energy barrier for the first aquation is lower tan for the second aquation. A further question is under which form the drug reaches the DNA and binds to its target, like for example:

$$Pt[NH_3]_2[H_2O]_2^{2+} + G \rightarrow Pt[G][NH_3]_2[H_2O]^{2+} + H_2O$$

$$Pt[G][NH_3]_2[H_2O]^{2+} + G \rightarrow Pt[G]_2[NH_3]_2^{2+} + H_2O$$

For the first and the second substitution different adducts for the binding of cisplatin to DNA have been calculated.

Energetically evaluations give hints about the observed probabilities of the occurring adducts: 65% intrastrand 5´-GG adduct, 25% intrastrand 5´-AG adduct, 6% intrastrand 5´-GXG, 3% interstrand GG adduct. Although the calculations reveal important details of the mechanism, such as solvation energy, optimized conformations of activated cisplatin and energy barriers of different conformations of activated cisplatin with guanine and adenine at the DNA strand, several details are still under discussion.

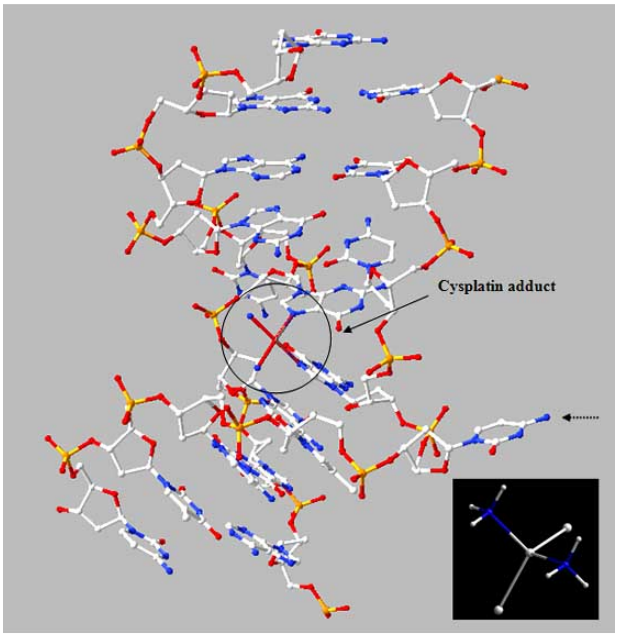


Fig. (28). Cisplatin (small image) is an anti-tumour drug which distorts the tertiary structure of the cell DNA and therefore inhibits the replication and transcription machinery of the tumour cell. The visualization shows the major groove of the intrastrand lesion caused by the binding on two guanine nucleotides (resolution: 1.63 angstrom, R-value = 0.169). From the distortion results an extruding of the complementary cytosine from the helix (dotted arrow). Quantum calculations based on density functional theory provide details about solvation and binding of cisplatin to the DNA strand.

Alzheimer's Disease

Alzheimer's disease is characterized by the deposition and aggregation of the normally soluble amyloid-beta (Abeta) peptide in the extracellular spaces of the brain as parenchymal plaques and in the walls of cerebral vessels as cerebral amyloid angiopathy (CAA). CAA is a common cause of brain haemorrhage in the elderly and is found in most patients with Alzheimer's disease [228]. Symptoms of this disease are disturbance of memory and subsequent neuropsychological symptoms such as disturbance of orientation.

The epsilon4 allele of the apolipoprotein E (APOE) gene is a risk factor for CAA [229,230]. Therefore genetic variability at the apoliprotein E gene is a major determinant of late onset of Alzheimer's disease [231,232] (Fig. **29**). The identification of molecular or epigenetic factors affecting primary molecular mechanisms underlying the disease may ultimately contribute to the development of rational therapy for Alzheimer's disease.

Homolog-Scanning Mutagenesis

Homolog-scanning mutagenesis is used for the identification of functional segments within proteins [233-235]**.** It helps to analyze related protein binding sites to assess the specificity and importance of individual side chain contributions to binding affinity. The principle of homolog-scanning mutagenesis is demonstrated on hand of Glutamic Acid Decarboxylase (GAD 65) which is one of the antigens playing a considerable role in Diabetes mellitus Type 1. The enzyme GAD uses the irreversible α-decarboxylation of L-glutamate as substrate and catalyzes it into γ-aminobutyrate (GABA). GABA and GAD are typically present in several tissues, especially in the pancreatic β-cells [236]. In these cells GABA regulates the glucagon secretion. In autoimmune diabetes, an attack of inflammatory cells to the endocrine pancreatic β-cells leads to their complete destruction, resulting in the inability to produce insulin for the body's requirements.

For an analysis of the neutralization of GAD 65 with antibodies, first the locations of epitopes in the amino acid sequence must be determined. The epitopes are identified by homolog-scanning mutagenesis, where segments of sequences from a homologous molecule (GAD 67) known not to bind to the specific antibodies are systematically substituted throughout the GAD 65 gene [237,238]. A complete or partial loss of antibody reactivity by the resulting mutated molecules (chimeras) suggests that GAD 65 specific sequences required for contact with the antibody has been exchanged by the point mutation. The location of these sequence fragments, which are constituting the epitopes, are highlighted on the GAD 3D structure [239]. The Voronoi tessellation is used to get quantitative values for the exposure of the epitopes. To this purpose the exposure rate (chapter 12) of the epitopes to the surrounding solvent was calculated,

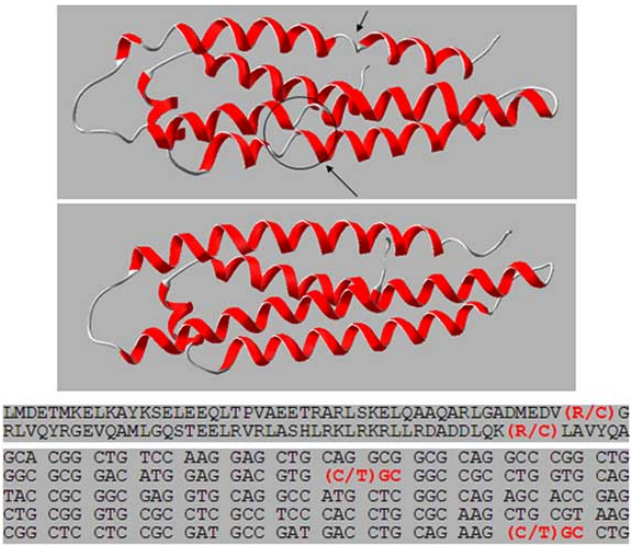


Fig. (29). The protein apolipoprotein E is involved in Alzheimer's diseases (resolution: 1.7 angstrom, R-value = 0.219). A major determinant of the late onset of this disease is the genetic variability at the corresponding gene. The exchange of the nucleotide cytosine (C) through tymine (T) in the DNA sequence results in an exchange of the amino acid arginine (R) by cysteine (C) in the protein sequence. As a consequence structural changes in the 3D structure of the protein occur.

giving hints about the accessibilities of the antibodies to the GAD molecule (Fig. **30**).

Gene Therapy

For the therapy of diseases, based on the malfunction of proteins, new possibilities are obtained by manipulating the corresponding genes [240-242]. In gene therapy, defect genes are either replaced (for the production of the correct protein) or hindered in their expression (HIV therapy) or therapeutically genes are expressed (therapy of malign tumours). The transfer of a gene in the cell nucleus is performed by a vector (for example a virus) which inserts the gene in the cell genome. One procedure for deactivating a gene is based on the insertion of a short DNA fragment which bounds specifically on a certain sequence segment of the gene and therefore inhibits its expression.

15. CONCLUSIONS

In principle, the DNA string holds the template for human development, physiology and certain diseases. The influence of bioinformatics on medical research is highly dependent on the accessibility of genomic and proteomic data and the transformation of such information into understanding, prevention and treatment of diseases. In future, it will be become more and more routine to tailor medical treatments to the protein shape resulting from individual genetic profiles of patients, their pathogens etc. Therefore the determination of sequence-structure-function relationships, complemented by quantum theoretical methods, is of special importance.

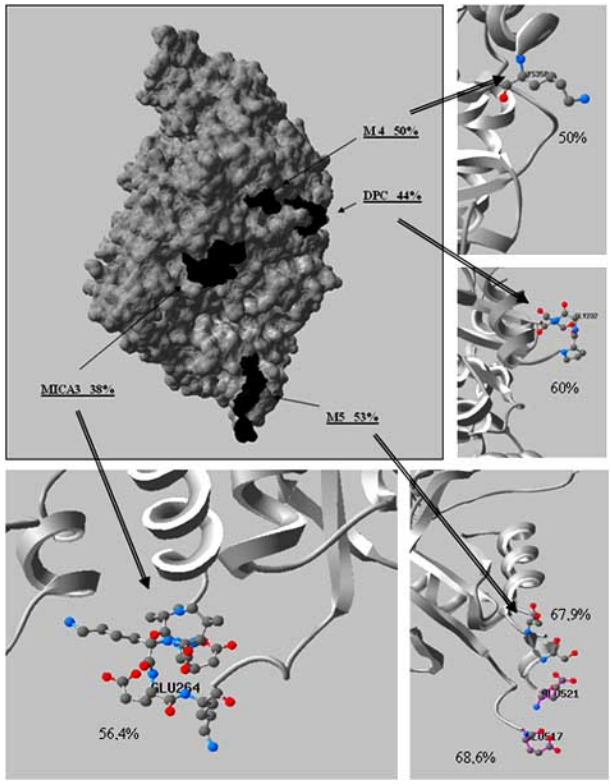


Fig. (30). The central figure show the molecular surface of the GAD 65 molecule (the positions of the epitopes are shaded black), whereas in the surrounding pictures the residues of the epitopes are represented as "balls-and-sticks". The exposure values of the residues, constituting the epitopes, have been calculated by Voronoi tessellation. These values (in %) indicate how much the epitopes are exposed to the surrounding solvent, indicating the accessibility to potential antibodies. In the central figure the mean exposure values of the epitope residues are shown, whereas the small figures display the maximum values of the most exposed residues in the single epitopes.

ACKNOWLEDGEMENTS

We wish to thank Ms. G. Searle for the critical reading of the text. Additionally we thank the reviewers for their valuable comments.

REFERENCES

[1] Chou KC. Review: Structural bioinformatics and its impact to biomedical science. *Curr Med Chem* **2004**; 11: 2105-34.

[2] Chou KC. Structural bioinformatics and its impact to biomedical science and drug discovery, In: Atta-ur-Rahman A, Reitz B Eds. Frontiers in Medicinal Chemistry. Bentham Science Publishers, The Netherlands, 2006; 3: 455-502.

[3] Dong X, Ying X, Uberbacher EC. Computational tools for protein modelling. *Curr Protein Pept Sci* **2000**; 1: 1-21.

[4] Petry S, Brodersen DE, Murphy IV FV, *et al.* Crystal Structures of the Ribosome in Complex with Release Factors RF1 and RF2 Bound to a Cognate Stop Codon. *Cell* **2005**; 123: 1255- 66.

[5] Bernhardt R. Cytochrome P450: structure, function, and generation of reactive oxygen species. *Rev. Pysiol Pharmacol* **1995**; 127: 137-221.

[6] Muller JJ, Lapko A, Bourenkov G, Ruckpaul K, Heinemann U. Adrenodoxin reductase-adrenodoxin complex structure suggests electron transfer path in steroid biosynthesis. *J Biol Chem* **2001**; v276: 2786-89.

[7] Müller A, Müller JJ, Muller YA, Uhlmann H, Bernhardt R, Heinemann U. New aspects of electron transfer revealed by the crystal

structure of a truncated bovine adrenodoxin, Adx(4-108). *Structure* **1998**; 6: 269-80.

[8] Mc Donald NQ, Hendrickson WA. A structural super-family of growth factors containing a cystine knot motif. *Cell* **1993**; 73: 421-24.

[9] Suto K, Yamazaki Y, Morita T, Mizuno H. Crystal structures of novel vascular endothelial growth factors (VEGF) from snake venoms: insight into selective VEGF binding to kinase insert domain-containing receptor but not to fms-like tyrosine kinase-1. *J Biol Chem* **2005**; 280: 2126-31.

[10] Wells JA, de Vos AM. Hematopoitic receptor complexes. *Annu Rev Biochem* **1996**; 65: 609-34.

[11] Syed RS, Reid SW, Li C, *et al*, Efficiency of signalling through cytokine receptors depends critically on receptor orientation. *Nature* **1998**; 395: 511- 16.

[12] Zhu X, Komiya H, Chirino A, *et al*. Three-dimensional structures of acidic and basic fibroblast growth factors. *Science* **1991**; 251: 90-93.

[13] Bernett MJ, Somasundaram T, Blaber M. An atomic resolution structure for human fibroblast growth factor 1. *Proteins* **2004**; 57: 626-34.

[14] Clore GM, Gronenborn AM. Three-dimensional structures of alpha and beta chemokines. *FASEB J* **1995**; 9: 57-62.

[15] Gerber N, Lowman H, Artis DR, Eigenbrot C. Receptor-binding conformation of the "ELR" motif of IL-8: X-ray structure of the L5C/H33C variant at 2.35 A resolution. *Proteins* **2000**; 38: 361-67.

[16] Thiel DJ, le Du MH, Walter RL, *et al*, Observation of an unexpected third receptor molecule in the crystal structure of human interferon-gamma receptor complex. *Structure Fold Des* **2000**; 8: 927-36.

[17] Cheung KH, Shineman D, Müller M, *et al*. Mechanism of Ca2+ disruption in Alzheimer's disease by presenilin regulation of InsP3 receptor channel gating. *Neuron* **2008**; 58(6): 871-83.

[18] Coles B, Wilton LA, Good M, Chapman PF, Wann KT. Potassium channels in hippocampal neurones are absent in a transgenic but not in a chemical model of Alzheimer's disease. *Brain Res* **2008**; 1190: 1-14.

[19] Zhang X, Rueter JK, Chen Y, *et al*. Synthesis of N-pyrimidinyl-2-phenoxyacetamides as adenosine A2A receptor antagonists. *Bioorg Med Chem Lett* **2008**; 18(6): 1778-83.

[20] Zeng J, Wang G, Chen SD. ATP-sensitive potassium channels: novel potential roles in Parkinson's disease. *Neurosci Bull* **2007**; 23(6): 370-76.

[21] Oyama F, Miyazaki H, Sakamoto N, *et al*. Sodium channel beta4 subunit: down-regulation and possible involvement in neuritic degeneration in Huntington's disease transgenic mice. *J Neurochem* **2006**; 98(2): 518-29.

[22] Scatena R, Martorana GE, Bottoni P, Botta G, Pastore P, Giardina B. An update on pharmacological approaches to neurodegenerative diseases. *Expert Opin Investig Drugs* **2007**; 16(1): 59-72.

[23] Ren G, Reddy VS, Cheng A, Melnyk P, Mitra AK. Visualization of a water-selective pore by electron crystallography in vitreous ice. *Proc Natl Acad Sci USA* **2001**; v98: 1398-1403.

[24] Kendrew JC, Dickerson RE, Strandberg BE, et al. Structure of myoglobin. A three- dimensional Fourier sythesis at 2 angstrom resolution. *Nature* **1960**; 185: 422-27.

[25] Perutz MF, Muirhead H, Cox JM, Goaman LC. Three-dimensional Fourier synthesis of horse oxyhaemoglobin at 2.8 angstrom resolution: the atomic model. *Nature* **1968**; 219: 131-39.

[26] De la Fortelle E, Bricogne G. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol* **1997**; 276: 472-94.

[27] Perrakis A, Morris R, Lamzin VS. Automated protein model building combined with iterative structure refinement *Nat Struct Biol* **1999**; 6: 458-63.

[28] Terwilliger TC, Berendzen J. Automated structure solution for MIR and MAD. *Acta Crystallogr D* **1999**; 55: 849-61.

[29] Weeks CM, Miller R. Optimizing shake-and-bake for proteins. *Acta Crystallogr D* **1999**; 55: 492-500.

[30] Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wählby A, Jones TA. "The Uppsala Electron-Density Server". *Acta Cryst* **2004**; D60: 2240-49.

[31] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database fort he investigation of sequences and structures. *J Mol Biol* **1995**; 247: 536-40.

[32] Andreeva A, Howorth D, Chandonia JM, *et al*. Data growth and its impact on the SCOP database: new developments. *Nucl Acid Res* **2008**; 36: 419-25.

[33] Berman HM, Westbrook J, Feng Z, *et al*. The Protein Data Bank. *Nucleic Acids Res* **2000**; 28: 235-42.

[34] Baxevanis AD. Searching the NCBI databases using Entrez. *Curr Protoc Hum Genet* **2006**; Chapter 6: Unit 6.10.

[35] Falkner JA, Hill JA, Andrews PC. Proteomics FASTA archive and reference resource. *Proteomics* **2008**; 8(9): 1756-57.

[36] Sussman JL, Abola EE, Lin D, Jiang J, Manning NO, Prilusky J. The protein data bank. Bridging the gap between the sequence and 3D structure world. *Genetica* **1999**; 106(1-2): 149-58.

[37] Sussman JL, Lin D, Jiang J, *et al*. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* **1998**; 54(Pt 6 Pt 1): 1078-84.

[38] Westbrook JD, Fitzgerald PM. The PDB format, mmCIF, and other data formats. *Methods Biochem Anal* **2003**; 44: 161-79.

[39] McCray AT, Divita G. ASN.1: defining a grammar for the UMLS knowledge sources. *Proc Annu Symp Comput Appl Med Care* **1995**; 868-72.

[40] Levinthal C, Barry CD. Computer Graphics in Macromolecular Chemistry, In: Secrest D, Nievergelt J Eds. Emerging Concepts in Computer Graphics. W. A. Benjamin, NY 1968; 231-53.

[41] Johnson CK. "OR TEP: A FORTRAN Thermal-Ellipsoid Plot Program for Crystal Structure Illustrations". ONRL Report #3794. Oak Ridge National Laboratory, Oak Ridge, Ten, 1968.

[42] Beem KM, Richardson DC, Rajagopalan KV. Metal sites of copper-zinc superoxide dismutase. *Biochemistry* **1977**; 16(9): 1930-36.

[43] Ricardson DC, Richardson JS. Kinemages – Simple macromolecular graphics for interactive teaching and publication. *Trends Biochem Sci* **1994**; 19: 135-38.

[44] Sayle RA, Milner-White EJ. RASMOL: Biomolecular graphics for all. *Trends Biochem Sci* **1995**; 20: 374.

[45] Hogue CW. Cn3D: A new generation of three-dimensional molecular structure viewer. *Trends Biochem Sci* **1997**; 22: 314-16.

[46] Wang Y, Geer LY, Chappey C, Kans JA, Bryant SH. Cn3D: sequence and structure views for Entrez. *Trends Biochem Sci* **2000;** 5(6): 300-02.

[47] Guex N, Peitsch MC. SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modelling. *Electrophoresis* **1997**; 18: 2714-23.

[48] Walther D. WebMol- a Java based PDB viewer. *Trends Biochem Sci* **1997**; 22: 274- 75.

[49] Davis IW, Murray LW, Richardson JS, Richardson DC. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res* **2004**; 1: 32.

[50] Lancashire RJ. The JSpecView Project: an Open Source Java viewer and converter for JCAMP-DX and XML spectral data files. *Chem Cent J* **2007**; 1:31.

[51] Watanabe K, Yasukawa K, Iso K. Graphic display of nucleic acid structure by a microcomputer. *Nucleic Acids Re* **1984**; 12: 801-09.

[52] Bowen JP, Charifson PS, Fox PC, *et al*. Computer-assisted molecular modeling: indispensable tools for molecular pharmacology. *J Clin Pharmacol* **1993**; 33(12): 1149-64.

[53] Smith TJ. MolView: a program for analyzing and displaying atomic structures on the Macintosh personal computer. *J Mol Graph* **1995** ;13(2): 122-25.

[54] Hannon GJ, Jentoft JE. MOLECULAR DESIGNER: an interactive program for the display of protein structure on the IBM-PC. *Comput Appl Biosci* **1985**; 1(3): 177-81.

[55] Light WR, Gaber BP. NanoVision-molecular graphics for the Macintosh. *J Mol Graph* **1994**; 12(3): 172-77.

[56] Merritt EA, Murphy ME. Raster3D Version 2.0. A program for photorealistic molecular graphics. *Acta Crystallogr D Biol Crystallog* **1994**; 50:869-73.

[57] Crivelli S, Kreylos O, Hamann B, Max N, Bethel W. ProteinShop: a tool for interactive protein manipulation and steering. *J Comput Aided Mol De* **2004**; 18(4): 271-85.

[58] Gabdoulline RR, Hoffmann R, Leitner F, Wade RC. ProSAT: functional annotation of protein 3D structures. *Bioinformatics* **2003**; 19(13): 1723-25.

[59] Gabdoulline RR, Ulbrich S, Richter S, Wade RC. ProSAT2—Protein Structure Annotation Server. *Nucleic Acids Res* **2006**; 34: 79-83.

[60] Wiltgen M, Holzinger A. Visualization in Bioinformatics: Protein Structures with Physicochemical and Biological Annotations, In: Zara J, Sloup J Eds. Central European Multimedia and Virtual Reality Conference. CEMVRC, 2005; 69-74.

[61] Huang W, Matte A, Li Y, *et al*. Crystal structure of chondroitinase B from Flavobacterium heparinum and its complex with a disaccharide product at 1.7 A resolution. *J Mol Biol* **1999**; 294: 1257-69.

[62] Lee B, Richards FM. The interpretation of protein structures: Estimation of static accessibility. *J Mo Biol* **1971**; 55: 379-400.

[63] Eisenberg D, Mc Lachlan AD. Solvation energy in protein folding and binding. *Nature* **1986**; 319: 199-203.

[64] Richards FM. Areas, volumes, packing, and protein structure. *Annu Rev Biophys Bioeng* **1997**; 6: 151-76.

[65] Neshich G, Rocchia W, Mancini AL, et al. Java Protein Dossier: a novel web-based data visualization tool for comprehensive analysis of protein structures. *Nucleic Acids Res* **2004**; 32: 595-601.

[66] Golovin A, Oldfield TJ, Tate JG et al. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res* **2004**; 32: 211-16.

[67] Leslin CM, Abyzov A, Ilyin VA. Structural exon database, SEDB, mapping exon boundaries on multiple protein structures. *Bioinformatics* **2004**; 20(11): 1801-03.

[68] Oldfield TJ. A Java applet for multiple linked visualization of protein structure and sequence. *J Comput Aided Mol De* **2004**; 18(4): 225-34.

[69] Vivek G, Tan TW, Ranganathan S. XdomView: protein domain and exon position visualization. *Bioinformatics* **2003**; 19(1): 159-60.

[70] Ilyin VA, Pieber U, Stuart AC, Marti-Renom MA, MCMahan L, Sali A. ModView, visualization of multiple protein sequences and structures. *Bioinformatics* **2003**; 19(1):165-66.

[71] Catherinot V, Labesse G. ViTO: tool for refinement of protein sequence-structure alignments. *Bioinformatics* **2004**; 20(18): 3694-96.

[72] Zhang J, Rowe WL, Struewing JP, Buetow KH. Hapscope: a software system for automated and visual analysis of functionally annotated haplotypes. *Nucleic Acids Res* **2002**; 30(23): 5213-21.

[73] Bjellqvist B, Hughes GJ, Pasquali C, *et al*. The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* **1993**; 14: 1023- 31.

[74] Bachmair A, Finley D, Varshavsky A. In vivo half-life of a protein is a function of its amino-terminal residue. *Science* **1986**; 234: 179-86.

[75] Guruprasad K, Reddy BV, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng* **1990**; 4: 155-61.

[76] Kyte J, Doolitle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **1982**; 157: 105-32.

[77] Rost B. PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods in Enzymology* **1996**; 266: 525-39.

[78] Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* **2003**; 31: 3784-88.

[79] Bairoch A, Apweiler R. The SWISS-PROT protein sequence databank and its supplement TrEMBL in 1998. *Nucl Acids Res* **1998**; 26: 38-42.

[80] Attwood TK, Avison H, Beck ME, *et al*. The PRINTS database of protein fingerprints: a novel information resource for computational molecular biology. *J Chem Inform Comput Sci* **1997**; 37(3): 417- 24.

[81] Attwood TK, Croning MDR, Flower DR, *et al*. PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res* **2000**; 28: 225-27.

[82] Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam protein families database. *Nucleic Acids Res* **2000**; 28: 263-66.

[83] Sammut SJ, Finn RD, Bateman A. Pfam 10 years on: 10,000 families and still growing. *Brief Bioinform* **2008**; 9(3): 210-19.

[84] Marchler-Bauer A, Panchenko AR, Shoemaker BA, et al. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res* **2002**; 30: 281-83.

[85] Marchler-Bauer A, Anderson JB, Derbyshire MK, *et al*. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* **2007**; 35: 237-40.

[86] Henikoff S, Henikoff JG. Automated assembly of protein blocks for database searching. *Nucleic Acids Res* **1991**; 19: 6565-72.

[87] Henikoff S, Henikoff JG, Pietrokovski S. Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* **1999**; 15: 471-79.

[88] Hulo N, Bairoch A, Bulliard V, *et al*. The 20 years of PROSITE. *Nucleic Acids Res* **2008**; 36: 245-49.

[89] Henschel A, Winter C, Kim WK, Schroeder M. Using structural motif descriptors for sequence-based binding site prediction. *BMC Bioinformatics* **2007**; 8(Suppl 4): 5.

[90] Bucher P, Bairoch A. A generalized profile syntax for biomolecular sequences motifs and its function in automatic sequence interpretation, In: Altman R, Brutlag D, Karp P, Lathrop R, Searls D Eds. ISMB-94; Proceedings 2nd International Conference on Intelligent Systems for Molecular Biology. AAAIPress, Menlo Park 1994; 53-61.

[91] Haefliger DN, Moskaitis JE, Schoenberg DR, Wahli W. Amphibian albumins as members of the albumin, alpha-fetoprotein, vitamin D-binding protein multigene family. *J Mol Evol* **1989**; 29: 344-54.

[92] Schoentgen F, Metz-Boutigue M. -H, Jolles J, Constans J, Jolles P. Complete amino acid sequence of human vitamin D-binding protein (group-specific component): evidence of a three-fold internal homology as in serum albumin and alpha-fetoprotein. *Biochim Biophys Actan* **1986**; 871: 189-98.

[93] Lichenstein HS, Lyons DE, Wurfel MM, *et al*. Afamin is a new member of the albumin, alpha-fetoprotein, and vitamin D-binding protein gene family. *J Biol Chem* **1994**; 269: 18149-54.

[94] Lejon S, Frick IM, Bjorck L, Wikstrom M, Svensson S. Crystal structure and biological implications of a bacterial albumin binding module in complex with human serum albumin. *J Biol Chem* **2004**; 279: 42924-28.

[95] Chou KC, Shen HB. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem Biophys Res Comm* **2007**; 360: 339-45.

[96] Chou KC, Shen HB. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem Biophys Res Comm* **2007**; 357: 633-40.

[97] Shen HB, Chou KC. Signal-3L: a 3-layer approach for predicting signal peptide. *Biochem Biophys Res Comm* **2007**; 363: 297-303.

[98] Vorderwülbecke S, Cleverley S, Weinberger SR, Wiesner A. Protein quantification by the SELDI-TOF-MS-based ProteinChip System. *Nature Methods* **2005**; 2,5: 393-95.

[99] Zenobi R, Knochenmuss R. Ion formation in maldi mass spectrometry. *Mass Spectrom Rev* **1998**; 17: 337-66.

[100] Karas M, Glückmann M, Schäfer J. Ionization in matrix-assisted desorption/ionization: singly charged molecular ions are the lucky survivors. *J Mass Spectrom* **2000**; 35: 1-12.

[101] Gasteiger E, Hoogland C, Gattiker A, *et al*. Protein Identification and Analysis Tools on the ExPASy Server, In: Walker JM Eds. The Proteomics Protocols Handbook. Humana Press, 2005; 571-607.

[102] Tuloup M, Hernandez C, Coro I, Hoogland C, Binz P-A, Appel R D. Aldente and BioGraph: An improved peptide mass fingerprinting protein identification environment, In: Understanding Biological Systems through Proteomics. Fontis Media, Basel, Switzerland, 2003; 174-76.

[103] Xu D, Xu Y, Uberbacher EC. Computational Tools for Protein Modeling. *Curr Protein Pept Sci* **2000**; 1: 1-21.

[104] Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL. Knowledge-based proteinmodelling. *CRC Crit Rev Biochem Mol Biol* **1994**; 29: 1-68.

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

[106] Sanchez R, Sali A. Advances in comparative protein-structure modelling. *Curr Opin Struct Biol* **1997**; 7: 206-14.

[107] Schwede T, Diemand A, Guex N, Peitsch MC. Protein structure computing in the genomic era. *Res Microbiol* **2000**; 151: 107-12.

[108] Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modelling. In. Watanabe M, Roux B, MacKerell AD Jr., Becker O Eds, Computational Biochemistry and Biophysics. Marcel Dekker, 2001; 275-312.

[109] Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **1993**; 234: 779-815.

[110] Fiser A, Kinh G, Do R, Sali A. Modeling of loops in protein structures. *Protein Sci* **2000**; 9: 1753-73.

[111] Peitsch MC. Protein modelling by E-Mail. *Bio Technol* **1995**; 13: 658-60.
[112] Peitsch MC. ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem Soc Tran* **1996**; 24: 274-79.
[113] Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web- based environment for protein structure homology modelling. *Bioinformatics* **2006**; 22 2: 195-201.
[114] Schwede T, Kopp J, Guex N, Peitsch M.C. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research* **2003**; 231(13): 3381-85.
[115] Eswar N, Eramian D, Webb B, Shen MY, Sali A. Protein structure modelling with MODELLER. *Methods Mol Biol* **2008**; 426: 145-59.
[116] Eswar N, Webb B, Marti-Renom MA, *et al*. Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci* 2007; Chapter 2: Unit 2.9.
[117] Read JA, Winter VJ, Eszes CM, Sessions RB, Brad RL. Structural basis for altered activity of M- and H-isozyme forms of human lactate dehydrogenase. *Proteins* **2001**; 43: 175-85.
[118] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. "Basic local aligment search tool". *J Mol Biol* **1990**; 215: 403-10.
[119] White JL, Hackert ML, Buehner M, *et al*. A comparison of the structures of apo dogfish M4 lactate dehydrogenase and its ternary complexes. *J Mol Biol* **1997**; 102: 759-79.
[120] Hogrefe HH, Griffith JP, Rossmann MG, Goldberg E. Characterization of the antigenic sites on the refined 3-A resolution structure of mouse testicular lactate dehydrogenase C4. *J Biol Chem* **1987**; 262: 13155-62.
[121] Grau UM, Trommer WE, Rossmann MG. Structure of the active ternary complex of pig heart lactate dehydrogenase with S-lac-NAD at 2.7 A resolution. *J Mol Biol* **1981**; 151: 289-307.
[122] Laskowski R A, MacArthur M W, Moss D S, Thornton J M. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Cryst* **1993**; 26: 283-91.
[123] Adane L, Bharatam PV. Modelling and informatics in the analysis of P. falciparum DHFR enzyme inhibitors. *Curr Med Chem* **2008**; 15(16): 1552-69.
[124] Lambert JM, Siezen RJ, de Vos WM, Kleerebezem M. Improved annotation of conjugated bile acid hydrolase superfamily members in Gram-positive bacteria. *Microbiology* **2008**; 154(8): 2492-2500.
[125] Rocher A, Marchand-Geneste N. Homology modelling of the Apis mellifera nicotinic acetylcholine receptor (nAChR) and docking of imidacloprid and fipronil insecticides and their metabolites. *SAR QSAR Environ Res* **2008**; 19(3-4): 245-61.
[126] Hou S, Li B, Wang L, *et al*. Humanization of an Anti-CD34 Monoclonal Antibody by Complementarity-determining Region Grafting Based on Computer-assisted Molecular Modelling. *J Biochem* **2008**; 144(1): 115-20.
[127] Sgobba M, Degliesposti G, Ferrari AM, Rastelli G. Structural models and binding site prediction of the C-terminal domain of human Hsp90: a new target for anticancer drugs. *Chem Biol Drug Des* **2008**; 71(5): 420-33.
[128] Capitani G, De Biase D, Gut H, Ahmed A, Grütter MG. Structural Model of Human GAD65: Prediction and Interpretation of Biochemical and Immunogenic Features. *Proteins: Struct Funct Bioinform* **2005**; 59: 7-14.
[129] Wiltgen M, Tilz GP. A Basic Molecular analysis of the diabetic antigen Gad by Homology Modelling. Principles of the Method and understanding of Antigenicity and binding sites. *Pteridines* **2007**; 18: 79-94.
[130] Guvench O, MacKerell AD Jr. Comparison of protein force fields for molecular dynamics simulations. *Methods Mol Biol* **2008**; 443: 63-88.
[131] Li Z, Yu H, Zhuang W, Mukamel S. Geometry and Excitation Energy Fluctuations of NMA in Aqueous Solution with CHARMM, AMBER, OPLS, and GROMOS Force Fields: Implications for Protein Ultraviolet Spectra Simulation. *Chem Phys Lett* **2008**; 452(1-3): 78-83.
[132] Komáromi I, Owen MC, Murphy RF, Lovas S. Development of glycyl radical parameters for the OPLS-AA/L force field. *J Comput Chem* **2008**; 29(12): 1999-2009.
[133] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**; 65(3): 712-25.
[134] Christen M, Hünenberger PH, Bakowies D, *et al*. The GROMOS software for biomolecular simulation: GROMOS05. *J Comput Chem* **2005**; 26(16): 1719-51.
[135] Price DJ, Brooks CL. Modern protein force fields behave comparably in molecular dynamics simulations. *J Comput Chem* **2002**; 23(11): 1045-57.
[136] Xu Z, Luo HH, Tieleman DP. Modifying the OPLS-AA force field to improve hydration free energies for several amino acid side chains using new atomic charges and an off-plane charge model for aromatic residues. *J Comput Chem* **2007**; 28(3): 689-97.
[137] Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. *Curr Opin Struct Biol* **1996**; 6(3): 377-85.
[138] Bacon DJ, Moult J. Docking by least-squares fitting of molecular surface patterns. *J Mol Biol* **1992**; 225: 849-58.
[139] Chau PL, Dean PM. Molecular recognition: 3D surface structure comparision by gnomonic projection. *J Mol Graphics* **1987**; 5: 97-100.
[140] Uetz P, Giot L, Cagney G, *et al*. A comprehensive analysis of protein- protein interactions in Saccharomyces cerevisiae. *Nature* **2000**; 403: 623-27.
[141] Duncan B, Olson AJ. Shape analysis of molecular surface. *Biopolymers* **1993**; 33: 231-38.
[142] Duncan B, Olson AJ. Approximation and characterization of molecular surfaces. *Biopolymers* **1993**; 33: 219-29.
[143] Gabdoulline RR, Wade RC. Analytically defined surfaces to analyze molecular interaction properties. *J Mol Graph* **1996**;14(6): 341-53, 374-75.
[144] Voronoi G. Recherche sur les paralleloedres primitifs. *J Reine Angw Math* **1908**; 134: 198-287.
[145] Sadoc JF, Jullien R, Rivier N. The Laguerre polyhedral decomposition: application to protein folds. *Eur Phys J B* **2003**; 33: 355-63.
[146] Dupuis F, Sadoc JF, Mornon JP. Protein Secondary Structure Assignment Through Voronoi Tessellation. *Proteins: Struct Funct Bioinform* **2004**; 55: 519-28.
[147] Angelov B, Sadoc JF, Jullien R, Soyer A, Mornon JP, Chomilier J. Nonatomic Solvent-Driven Voronoi Tessellation of Proteins: A Open Tool to Analyze Protein Folds. *Proteins: Struct Funct Genet* **2002**; 49: 446-56.
[148] Dupuis F, Sadoc JF, Jullien R, Angelov B, Mornon JP. Voro3D: 3D voronoi tessellations applied to protein structures. *Bioinformatics* **2005**; 21(8): 1715-16.
[149] Finn RD, Marshall M, Bateman A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* **2005**; 21(3): 410-12.
[150] Gabdoulline RR, Wade RC, Walther D. MolSurfer: A macromolecular interface navigator. *Nucleid Acids Res* **2003**; 31(13): 3349-51.
[151] Wiltgen M, Holzinger A, Tilz GP. Interactive Analysis and Visualization of Macromolecular Interfaces between Proteins. *Lecture Notes in Computer Sciences* **2007**; 4799: 199-212.
[152] Wiltgen M, Tilz GP. Tumour necrosis factor and its receptor: a basic structural analysis of two counterparts. *Hematology* **2008**; 13,4: 224-29.
[153] Jayapal P, Sundararajan M, Hillier IH, Burton NA. QM/MM studies of Ni-Fe hydrogenases: the effect of enzyme environment on the structure and energies of the inactive and active states. *Phys Chem Chem Phys* **2008**; 10(29): 4249-57.
[154] Perruccio F, Ridder L, Mulholland AJ. Quantum-mechanical/-Molecular-mechanical Methods in Medicinal Chemistry, In: Carloni P, Alber F Eds, Quantum Medicinal Chemistry. WILEY_VCH Verlag GmbH&Co. Weinheim 2003.
[155] Jensen F. Introduction to computational chemistry. John Wiley&Sons Ltd, UK 2007.
[156] Reinhold J. Quantentheorie der Moleküle. B.G. Teubner Verlag, Wiesbaden 2004.
[157] Hinchliffe A. Molecular modelling for beginners. John Wiley&Sons Ltd UK 2003.
[158] Szabo A, Ostlund NS. Modern quantum chemistry. Dover Publications, INC, Mineola, New York 1996.
[159] Cavalli A, Folkers G, Recanatini M, Scapozza L. Density-functional theory applications in computational medicinal chemistry, In: Carloni P, Alber F Eds, Quantum Medicinal Chemistry. WILEY_VCH Verlag GmbH&Co, Weinheim 2003.
[160] Raber J, Liano J, Eriksson LA. Density-functional theory in drug design – the chemistry of the anti-tumour drug cisplatin and photoactive psoralen compounds. In: Carloni P, Alber F. (eds) Quan-

tum Medicinal Chemistry WILEY_VCH Verlag GmbH&Co, Weinheim 2003.
[161] Várnai P, Richards WG, Lyne PD. Modelling the catalytic reaction in human aldose reductase. *Proteins* **1999**; 37(2): 218-27.
[162] Ridder L, Rietjens IM, Vervoort J, Mulholland AJ. Quantum mechanical/molecular mechanical free energy simulations of the glutathione S-transferase (M1-1) reaction with phenanthrene 9,10-oxide. *J Am Chem Soc* **2002**; 124(33): 9926-36.
[163] Mlinsek G, Novic M, Hodoscek M, Solmajer T. Prediction of enzyme binding: human thrombin inhibition study by quantum chemical and artificial intelligence methods based on X-ray structures. *J Chem Inf Comput Sci* **2001**; 41(5): 1286-94.
[164] Dinner AR, Blackburn GM, Karplus M. Uracil-DNA glycosylase acts by substrate autocatalysis. *Nature* **2001**; 413(6857): 752-5.
[165] Sulpizi M, Schelling P, Folkers G, Carloni P, Scapozza L. The rational of catalytic activity of herpes simplex virus thymidine kinase a combined biochemical and quantum chemical study. *J Biol Chem* **2001**; 276(24): 21692-7.
[166] Alhambra C, Sanchez ML, Corchado JC, *et al*. Quantum mechanical tunneling in methylamine dehydrogenase. *Chemi Phys Lett* **2002**; 355: 388-94.
[167] Alhambra C, Corchado J, Sanchez ML, et al. Canonical variational theory for enzyme kinetics with the protein mean force and multidimensional quantum mechanical tunneling dynamics. Theory and application to liver alcohol dehydrogenase. *J Phys Chem B* **2001**; 105(45): 11326-40.
[168] Alhambra C, Corchado JC, Sanchez ML, et al.Quantum dynamics of hydride transfer in enzyme catalysis. *J Am Chem Soc* **2000**; 122(34): 8197-203.
[169] Alhambra C, Gao JL, Corchado JC, et al. Quantum mechanical dynamical effects in an enzyme-catalyzed proton transfer reaction. *J Am Chem Soc* **1999**; 121(10): 2253-58.
[170] Chang PL. Clinical bioinformatics. *Gung Med* . **2005**; 28(4): 201-11.
[171] Snow CD. Hunting for predictive computational drug-discovery models. *Expert Rev Anti Infect Ther* **2008**; 6(3): 291-93.
[172] Wishart DS. Identifying putative drug targets and potential drug leads: starting points for virtual screening and docking. *Methods Mol Biol* **2008**; 443: 333-51.
[173] Chen YP, Chen F. Identifying targets for drug discovery using bioinformatics. *Expert Opin Ther Targets* **2008**; 12(4): 383-89.
[174] Bharatam PV, Patel DS, Adane L, Mittal A, Sundriyal S. Modeling and informatics in designing anti-diabetic agents. *Curr Pharm Des* **2007**; 13(34): 3518-30.
[175] Rainsford KD. Anti-inflammatory drugs in the 21st century. *Subcell Biochem* **2007**; 42: 3-27.
[176] Kopecny D, Sebela M, Briozzo P, *et al*. Mechanism-based inhibitors of cytokinin oxidase/dehydrogenase attack FAD cofactor. *J Mol Biol* **2008**; 380: 886- 99.
[177] Hamera-Słynarska M, Słynarski K. The undesirable side-effects of non-steroidal anti- inflammatory drugs. *Ortop Traumatol Rehabil* **2001**; 3(1): 126-28.
[178] Salinas G, Rangasetty UC, Uretsky BF, Birnbaum Y. The cycloxygenase 2 (COX-2) story: it's time to explain, not inflame. *J Cardiovasc Pharmacol Ther* **2007**; 12(2): 98- 111.
[179] Reddy RN, Mutyala R, Aparoy P, Reddanna P, Reddy MR. Computer aided drug design approaches to develop cyclooxygenase based novel anti-inflammatory and anti- cancer drugs. *Curr Pharm Des* **2007**; 13(34): 3505-17.
[180] Rajakrishnan V, Manoj VR, Subba Rao G. Computer-aided, rational design of a potent and selective small peptide inhibitor of cyclooxygenase 2 (COX2). *J Biomol Struct Dyn* **2008**; 25(5): 535-42.
[181] Chou KC, Watenpaugh KD, Heinrikson RL. A Model of the complex between cyclin- dependent kinase 5(Cdk5) and the activation domain of neuronal Cdk5 activator. *Biochem Biophys Res Commun* **1999**; 259: 420-28.
[182] Chou KC, Howe WJ. Prediction of the tertiary structure of the beta-secretase zymogen. *BBRC* **2002**; 292: 702-08.
[183] Chou KC. Insights from modelling the tertiary structure of BACE2. *J Proteome Res* **2004**; 3; 1069-72.
[184] Chou KC. Modeling the tertiary structure of human cathepsin-E. *Biochem Biophys Res Commun* **2005**; 331: 56-60.
[185] Wei DQ, Sirois S, Du QS, Arias HR, Chou KC. Theoretical studies of Alzheimer's disease drug candidate [(2,4-dimethoxy) benzylidene]-anabaseine dihydrochloride (GTS-21) and its derivatives. *BBRC* **2005**; 338: 1059-64.
[186] Chou KC. Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. *Biochem Biophys Res Commun* **2004**; 316: 636-42.
[187] Chou KC. Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor. *Biochemical and Biophysical Research Communication* **2004**; 319: 433-38.
[188] Chou KC, Wei DQ, Zhong WZ. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. (Erratum: ibid., 2003, Vol.310, 675). *Biochem Biophys Res Comm* **2003**; 308: 148-51.
[189] Zhang R, Wei DQ, Du QS, Chou KC. Molecular modeling studies of peptide drug candidates against SARS. *Med Chem* **2006**; 2: 309-14.
[190] Chou KC. Molecular therapeutic target for type-2 diabetes. *J Proteome Res* **2004**; 3: 1284-88.
[191] Chou KC. Insights from modelling three-dimensional structures of the human potassium and sodium channels. *J Proteome Res* **2004**; 3: 856-61.
[192] Gao WN, Wei DQ, Li Y, *et al*. Agaritine and its derivatives are potential inhibitors against HIV proteases. *Med Chem* **2007**; 3: 221-26.
[193] Wang SQ, Du QS, Chou KC. Study of drug resistance of chicken influenza A virus (H5N1) from homology-modeled 3D structures of neuraminidases. *Biochem Biophys Res Comm* **2007**; 354: 634-40.
[194] Du QS, Wang SQ, Chou KC. Analogue inhibitors by modifying oseltamivir based on the crystal neuraminidase structure for treating drug-resistant H5N1 virus. *Biochem Biophys Res Comm* **2007**; 362: 525-31.
[195] Chou KC, Jones D, Heinrikson RL. Prediction of the tertiary structure and substrate binding site of caspase-8. *FEBS Lett* **1997**; 419: 49-54.
[196] Chou KC, Tomasselli AG. Heinrikson RL. Prediction of the Tertiary Structure of a Caspase-9/Inhibitor Complex. *FEBS Lett* **2000**; 470: 249-56.
[197] Wei H, Zhang R, Wang C, Zheng H, Chou KC, Wei DQ. Molecular insights of SAH enzyme catalysis and their implication for inhibitor design. *J Theor Biol* **2007**; 244: 692-702.
[198] Wang JF, Wei DQ, Li L, Zheng SY, Li YX, Chou KC. 3D structure modeling of cytochrome P450 2C19 and its implication for personalized drug design. *Biochem Biophys Res Commun* (Corrigendum: ibid, 2007, vol. 357, 330). **2007**; 355, 513-19.
[199] Wang JF, Wei DQ, Chen C, Li Y, Chou KC. Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design. *Protein Pept Lett* **2008**; 15: 27-32.
[200] Chou KC, Shen HB. Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat Protoc* **2008**; 3: 153-62.
[201] Raha K, Peters MB, Wang B, *et al*. The role of quantum mechanics in structure-based drug design. *Drug Discov Today* **2007**; 12(17-18): 725-31.
[202] Spiegel K, Magistrato A. Modeling anticancer drug-DNA interactions via mixed QM/MM molecular dynamics simulations. *Org Biomol Chem* **2006**; 4(13): 2507- 17.
[203] Reddy MR, Singh UC, Erion MD. Ab initio quantum mechanics-based free energy perturbation method for calculating relative solvation free energies. *J Comput Chem* **2007**; 28(2): 491-94.
[204] Spiegel K, Magistrato A. Modeling anticancer drug-DNA interactions via mixed QM/MM molecular dynamics simulations. *Org Biomol Chem* **2006**; 4(13): 2507- 17.
[205] Hammer SM, Eron JJ Jr, Reiss P, *et al*. International AIDS Society-USA. Antiretroviral treatment of adult HIV infection: 2008 recommendations of the International AIDS Society-USA panel. *JAMA* **2008**; 300(5): 555-70.
[206] Tucker TJ, Saggar S, Sisko JT, *et al*. The design and synthesis of diaryl ether second generation HIV-1 non-nucleoside reverse transcriptase inhibitors (NNRTIs) with enhanced potency versus key clinical mutations. *Bioorg Med Chem Lett* **2008**; 18: 2959-66.
[207] Liu F, Kovalevsky AY, Tie Y, Ghosh AK, Harrison RW, Weber IT. Effect of flap mutations on structure of HIV-1 protease and inhibition by saquinavir and darunavir. *J Mol Biol* **2008**; 381: 102-15.
[208] Pawiński T, Pulik P, Gralak B, Horban A. Pharmacokinetic monitoring of HIV-1 protease inhibitors in the antiretroviral therapy. *Acta Pol Pharm* **2008**; 65(1): 93-100.
[209] Nair V, Chi G. HIV integrase inhibitors as therapeutic agents in AIDS. *Rev Med Virol* **2007**; 17(4): 277-95.

[210] Sherman W, Tidor B. Novel method for probing the specificity binding profile of ligands: applications to HIV protease. *Chem Biol Drug Des* **2008**;71(5): 387-407.

[211] Almerico AM, Tutone M, Lauria A. Docking and multivariate methods to explore HIV-1 drug-resistance: a comparative analysis. *J Comput Aided Mol Des* **2008**; 22(5): 287-97.

[212] Saen-oon S, Aruksakunwong O, Wittayanarakul K, Sompornpisut P, Hannongbua S. Insight into analysis of interactions of saquinavir with HIV-1 protease in comparison between the wild-type and G48V and G48V/L90M mutants based on QM and QM/MM calculations. *J Mol Graph Model* **2007**; 26(4): 720-27.

[213] Brunori M, Bourgeois D, Vallone B. Structural dynamics of myoglobin. *Methods Enzymol* **2008**; 437: 397-416.

[214] Rovira C, Schulze B, Eichinger M, Evanseck JD, Parrinello M. Influence of the heme pocket conformation on the structure and vibrations of the Fe-CO bond in myoglobin: a QM/MM density functional study. *Biophys J* **2001**; 81(1): 435-45.

[215] Strickland N, Mulholland AJ, Harvey JN. The Fe-CO bond energy in myoglobin: a QM/MM study of the effect of tertiary structure. *Biophys J* **2006**; 90(4): 27-29.

[216] Capece L, Marti MA, Crespo A, Doctorovich F, Estrin DA. Heme protein oxygen affinity regulation exerted by proximal effects. *J Am Chem Soc* **2006**; 128(38): 12455- 61.

[217] Riera AR, Uchida AH, Ferreira C, *et al.* Relationship among amiodarone, new class III antiarrhythmics, miscellaneous agents and acquired long QT syndrome. *Cardiol J* **2008**; 15(3): 209- 19.

[218] Fazio G, Pipitone S, D'Angelo L, *et al*. The long QT syndrome in pediatric age: prognosis and risk factor. *Minerva Cardioangiol* **2008**; 56(4): 387-90.

[219] Yang WP, Levesque PC, Little WA, Conder ML, Shalaby FY, Blanar MA. KvLQTI, a voltage-gated potassium channel responsible for human cardiac arrhythmias. *Proc Natl Acad Sci U.S.A* **1997**; 94: 4017-21.

[220] Sugie M, Ishihara K, Simizu Y, Oono H, Kawamura M. Case report of transient splenium abnormality in Charcot-Marie-Tooth disease. *Rinsho Shinkeigaku* **2008**; 48(5): 359-62.

[221] Jackson CE. A clinical approach to muscle diseases. *Semin Neurol* **2008**; 28(2): 228- 40.

[222] Walke VA, Walde MS. Haematological study in sickle cell homozygous and heterozygous children in the age group 0-6 years. *Indian J Pathol Microbiol* **2007**; 50(4): 901-04.

[223] Lebensburger J, Persons DA. Progress toward safe and effective gene therapy for beta- thalassemia and sickle cell disease. *Curr Opin Drug Discov Devel* **2008**; 11(2): 225- 32.

[224] Zhu C, Raber J, Eriksson LA. Hydrolysis process of the second generation platinum- based anticancer drug cis-amminedichlorocyclohexylamineplatinum(II). *J Phys Chem B* **2005**; 109(24): 12195-205.

[225] Raber J, Zhu C, Eriksson LA. Theoretical study of cisplatin binding to DNA: the importance of initial complex stabilization. *J Phys Chem B* **2005**; 109(21): 11006- 15.

[226] Coste F, Malinge JM, Serre L, *et al*. Crystal structure of a double-stranded DNA containing a cisplatin interstrand cross-link at 1.63 A resolution: hydration at the platinated site. *Nucleic Acids Res* **1999**; 27(8): 1837-46.

[227] Huang H, Zhu L, Reid BR, Drobny GP, Hopkins PB. Solution structure of a cisplatin- induced DNA interstrand cross-link. *Science* **1995**; 270(5243): 1842-45.

[228] Thal DR, Griffin WS, de Vos RA, Ghebremedhin E. Cerebral amyloid angiopathy and its relationship to Alzheimer's disease. *Acta Neuropathol* **2008**; 115(6): 599-609.

[229] Urosevic N, Martins RN. Infection and Alzheimer's disease: the APOE epsilon4 connection and lipid metabolism. *J Alzheimers Dis* **2008**; 13(4): 421-35.

[230] Zappasodi F, Salustri C, Babiloni C, *et al.* An observational study on the influence of the APOE-epsilon4 allele on the correlation between 'free' copper toxicosis and EEG activity in Alzheimer disease. *Brain Res* **2008**; 1215: 183-89.

[231] van der Flier WM, Pijnenburg YA, Schoonenboom SN, Dik MG, Blankenstein MA, Scheltens P. Distribution of APOE genotypes in a memory clinic cohort. *Dement Geriatr Cogn Disord* **2008**; 25(5): 433-38

[232] Raber J. AR, apoE, and cognitive function. *Horm Behav* **2008**; 53(5): 706-15.

[233] Xiong YM, Haas TA, Zhang L. Identification of functional segments within the beta2I-domain of integrin alphaMbeta2. *J Biol Chem* **2002**; 277(48): 46639-44.

[234] Li Y, Lawrence DA, Zhang L. Sequences within domain II of the urokinase receptor critical for differential ligand recognition. *J Biol Chem* **2003**; 278(32): 29925-32.

[235] Pál G, Fong SY, Kossiakoff AA, Sidhu SS. Alternative views of functional protein binding epitopes obtained by combinatorial shotgun scanning mutagenesis. *Protein Sci* **2005**; 14(9): 2405-13.

[236] Baekkeskov S, Landin M, Kristensen JK, *et al*. Antibodies to a 64,000 Mr human islet cell antigen precede the clinical onset of insulin-dependent diabetes. *J Clin Invest* **1987**; 79(3): 926-34.

[237] Myers MA, Fenalti G, Gray R, *et al*. A major diabetes-related conformational epitope on GAD65. *Ann NY Acad Sci* **2003**; 1005: 250-52.

[238] Schwartz HL, Chandonia JM, Kash SF, *et al.* High-resolution autoreactive epitope mapping and structural modelling of the 65 kDa form of glutamic acid Decarboxylase. *J Mol Biol* **1999**; 287: 983-99.

[239] Fenalti G, Law RH, Buckle AM, *et al*. GABA production by glutamic acid decarboxylase is regulated by a dynamic catalytic loop. *Struct Mol Biol* **2007**; v14: 280-86.

[240] Bagley J, Tian C, Iacomini J. Prevention of type 1 diabetes in NOD mice by genetic engineering of hematopoietic stem cells. *Methods Mol Biol* **2008**; 433: 277-85.

[241] Lei P, Andreadis ST. Efficient retroviral gene transfer to epidermal stem cells. *Methods Mol Biol* **2008**; 433: 367-79.

[242] Phillips JE, Garcia AJ. Retroviral-mediated gene therapy for the differentiation of primary cells into a mineralizing osteoblastic phenotype. *Methods Mol Biol* **2008**; 433: 333-54.

Received: August 20, 2008 Revised: September 29, 2008 Accepted: October 30, 2008