Protein Structure and Enzyme Catalysis: Knowledge-Based Protein Loop Prediction and *Ab Initio* Equilibrium Constant Estimation

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Quoc-Vu Ha Ngoc

aus Bonn

Köln, April 2008

1. Berichterstatter: Universitätsprofessor Prof. Dr. D. Schomburg

2. Berichterstatter: Universitätsprofessor Prof. Dr. H. W. Klein

Tag der mündlichen Prüfung: 25. April 2008

Danksagung

Prof. Dr. Dietmar Schomburg danke ich besonders für die Möglichkeit diese Dissertation durchführen zu können. Seine kompetente und herzliche Betreuung waren für mich eine große Bereicherung. Für seine Unterstützung und sein Vertrauen über die Jahre möchte ich mich herzlich bedanken.

Sehr herzlich möchte ich auch Prof. Dr. Helmut W. Klein für die freundliche Übernahme des Zweitgutachtens danken.

Dr. Gerd Wohlfahrt danke ich für die angenehme und erfolgreiche Zusammenarbeit bei der Loopvorhersage.

Mein Dank gilt auch Dr. Kai Hartmann in dem ich neben einem zuverlässigen stets hilfsbereiten Kollegen auch einen guten Freund gefunden habe, der mir bei der Vorhersage der Gleichgewichtskonstanten mit Rat und Sachverstand zur Seite stand.

Bedanken möchte ich mich auch bei Dr. Lars Packschies für die persönliche Betreuung bei der Bedienung der quantenmechanischen Software. Seine Hilfsbereitschaft, seine anregenden Ideen und sein Engagement haben wesentlich zum Gelingen dieses Projektes beigetragen.

Besonders bedanke ich mich bei meinen Eltern für ihre Geduld, ihre beständige Unterstützung, und dass sie stets für mich da waren.

Abstract

Prediction methods in the field of bioinformatics can be divided into *ab initio* and knowledge-based methods. The work in this thesis investigates the importance of anchor group positioning in knowledge-based protein loop prediction as well as the *ab initio* estimation of equilibrium constants using Density Functional Theory (DFT).

The maximum possible prediction quality of knowledge-based loop prediction was examined for 595 insertions and 589 deletions with respect to gap length, fragment length, amino acid type, secondary structure and relative solvent accessibility while applying all possible anchor group positions for the fitting of loops between 3 and 12 residues in length. It was possible to predict 74.3 % of insertions and 83.7 % of deletions within an RMS deviation of < 1.5 Å between template and target structure using a knowledge-based fragment databank based on structures of the Protein Databank (PDB). The analysis showed that the importance of anchor group positioning increases with gap length and that medium fragments with lengths between 5-8 residues perform better than shorter or longer fragments. In addition, better predictions were obtained when anchor groups consisted of hydrophobic residues, were located within secondary structures such as helices and beta sheets, or had low relative solvent accessibilities. A test based on targeted anchor group selection using a combination of the above criteria showed an improvement in prediction quality compared to a random selection of anchor groups.

Density Functional Theory (DFT) with a b3lyp/6-311g++ (d,p) basis set was used in combination with a preceding molecular mechanics conformational search to estimate the standard transformed Gibbs free energies of reaction ($\Delta G_r'^{\circ}$) for a set of 45 enzyme-catalyzed reactions at standard biochemical conditions (pH 7 and 298.15 K). For reactions from EC group 1 and EC groups 5 and 6, the calculated $\Delta G_r'^{\circ}$ values deviated from the experimental

values by an average of 2.49 kcal/mol and 5.50 kcal/mol, respectively. This data was comparable to the values calculated using group contribution method by Mavrovouniotis (Mavrovouniotis, J.Biol.Chem 1991; 266:14440-45), where the mean error was 2.76 kcal/mol for reactions from EC group 1 and 4.76 kcal/mol for reactions from EC groups 5 and 6. The mean error for the entire set of reactions was 10.30 kcal/mol. These results are very promising, considering that purely structural information was used, and the method can be improved by further optimization.

Zusammenfassung

Vorhersagemethoden auf dem Gebiet der Bioinformatik lassen sich unterscheiden zwischen *ab initio* und wissensbasierten Methoden. In dieser Dissertation wird sowohl der Einfluss der Ankergruppenpositionierung auf die Qualität der wissensbasierten Loopvorhersage untersucht, sowie eine *ab initio* Abschätzung von Gleichgewichtskonstanten mithilfe der Dichte Funktional Theorie (DFT) vorgenommen.

Für die wissensbasierte Loopvorhersage von 595 Insertionen und 589 Deletionen wurde die maximal mögliche Vorhersagequalität in Abhängigkeit von Gaplänge, Fragmentgröße, Aminosäuretyp, Sekundärstruktur und relativer Lösungsmittel-zugänglichkeit ermittelt. Dabei wurden alle Ankergruppenpositionen berücksichtigt, die bei einer Modellierung von Loops zwischen 3 und 12 Aminosäureresten möglich waren. 74.3 % der Insertionen und 83.7 % der Deletionen könnten mit einer RMS Abweichung von unter 1.5 Å zwischen Leit- und Zielstruktur anhand einer PDB-Struktur basierten Fragmentdatenbank vorausgesagt werden. Die Untersuchungen ergaben, dass der Einfluss der Ankergruppenpositionierung mit Länge der Gaps zunimmt, und dass mittellange Fragmente zwischen 5 und 8 Aminosäurereste bessere Vorhersageergebnisse erzielen, als kurze oder lange Fragmente. Ausserdem wurden bessere Vorhersagen erreicht, wenn die Ankergruppen entweder aus hydrophoben Aminosäureresten bestanden, innerhalb von Sekundärstrukturen wie Helices oder Beta-Faltblätter lagen, oder eine niedrige Lösungsmittelzugänglichkeit besaßen. In einem Test wurden die Ankergruppen durch Kombination der oben genannten Kriterien gezielt ausgewählt, wodurch, im Vergleich zur zufälligen Ankergruppenwahl, eine deutliche Verbesserung der maximalen Vorhersagequalität erzielt wurde.

Für 45 Enzymreaktionen unter Standardbedingungen (pH 7 und 298.15K) wurden die freien Reaktionsenthalpien ($\Delta G_r°'$) über quantenmechanische Berechnung der freien Enthalpien der Metabolite bestimmt, und die Vorhersagequalität durch Vergleich mit den experimentell ermittelten Gleichgewichtskonstanten untersucht. Die Berechnung der freien Enthalpien der Metabolite erfolgte nach molekularmechanischer Konformations-minimierung unter Anwendung der Dichte Funktional Theorie (DFT) mit dem b3lyp/6-311g++ (d,p) Basissatz. Die berechneten freien Reaktionsenthalpien unterschieden sich im Durchschnitt von den experimentellen Werten um 2.49 kcal/mol bei Reaktionen der EC Gruppe 1, und um 5.50 kcal/mol bei Reaktionen der EC Gruppen 5 und 6. Diese Werte waren vergleichbar mit denen, die durch Anwendung der Inkrementmethode von Mavrovouniotis (Mavrovouniotis, J.Biol.Chem 1991; 266:14440-45) erzielt wurden. Dort lag der Durchschnittsfehler bei 2.76 kcal/mol für Reaktionen der EC Gruppe 1, und 4.76 kcal/mol für Reaktionen der EC Gruppen 5 und 6. Für den gesamten Satz der Reaktionen betrug der Vorhersagefehler im Durchschnitt 10.30 kcal/mol. Diese Resultate können als sehr vielversprechend gewertet werden, da ausschliesslich reine Strukturinformationen verwandt wurden, und sie können durch weitere Optimierung der Methode noch verbessert werden.

List of Abbreviations and Constants

$\Delta G°'_{tot}$	- Total standard transformed Gibbs free energy (298.15K, I=0, 1M, pH7)
$\Delta G°_{tot}$	- Total standard Gibbs free energy (298.15K, I=0, 1M)
$\Delta G_f°$	- Standard Gibbs free energy of formation (298.15K, I=0, 1M)
$\Delta G_f°'$	- Standard transformed Gibbs free energy of formation (298.15K, I=0, 1M, pH7)
$\Delta G_r°$	- Standard Gibbs free energy of reaction (298.15K, I=0, 1M)
$\Delta G_r°'$	- Standard transformed Gibbs free energy of reaction (298.15K, I=0, 1M, pH7)
a.u.	- Atomic units (1 a.u. = 1 Hartree = 2625.5 kJ/mol)
COSMO	- Conductor-like Screening Model
DFT	- Density Functional Theory
EC	- Enzyme Commission
FSSP	- Families of Structurally Similar Proteins
I	- Ionic Strength (mol/l)
K	- Equilibrium Constant
K'	- Apparent Equilibrium Constant
kcal and kJ	- Kilocalories and Kilojoules (1 kcal = 4.184 kJ)
MMFF	- Merck Molecular Force Field
NBS	- National Bureau of Standards
NIST	- National Institute of Standards and Technology
PDB	- Protein Databank
PM3	- Parametrized Model Number 3
R	- Gas Constant (R= 8.314472 J K^{-1} mol^{-1})
SCOP	- Structural Classification of Proteins

List of Figures

List of Tables

Table of Contents

CHAPTER 1

1 INTRODUCTION

1.1 Purpose and Motivation

Prediction methods in the field of bioinformatics can be divided into *ab initio* and knowledge-based methods. This thesis investigates the importance of anchor group positioning in knowledge-based protein loop prediction as well as the *ab initio* estimation of equilibrium constants using Density Functional Theory (DFT).

The prediction of protein loops around insertions and deletions represents one of the major challenges in protein structure prediction. In knowledge-based structure prediction, loop modeling creates the second largest source of error next to template-target alignment. The quality of loop prediction is dependent upon several factors such as the algorithm for fragment selection, the completeness of the fragment databank, the fitting/optimization procedure, and the choice of anchor groups. The present thesis will investigate the effect of anchor group selection on loop prediction quality with respect to a variety of criteria including gap length, fragment length, amino acid type, secondary structure, and relative solvent accessibility. Finally, a combination of the criteria will be used for selecting optimal

anchor groups for a loop prediction scenario, and the prediction quality compared to a dataset with randomly chosen anchor groups.

Biochemical reactions are catalyzed by highly specific enzymes which, by lowering the activation energy, allow reactions to run at highly increased rates. However, the feasibility and direction of a biochemical reaction are determined by its equilibrium. Since equilibrium constants are usually not available for any random biochemical reaction without major experimental efforts, methods have been developed to predict them independently from experimental measurements. One such method is the group contribution method by Mavrovouniotis [68][69]. The method is, however, limited to reactions at standard biochemical conditions (pH 7 and 298.15K) and biased towards the biochemical metabolites from which the set of contributions were derived. It has therefore been desirable to find a methods such as, for example, a *ab initio* molecular quantum mechanical calculations which work independently from any experimental data. However in the past, the *ab initio* approach has always been problematic, as computers and theories were not sufficiently developed to generate energy data in a manageable time frame and with acceptable accuracy. Density Functional Theory (DFT) method [58] has, so far, offered the best compromise between accurate results and acceptable calculation times. The COSMO solvation model [57] has also shown to deliver fast and accurate calculations with respect to solvation energies. This project aims at using the recent developments in hardware, software, and quantum mechanical methods in an attempt to develop a procedure for estimating experimental biochemical equilibrium constants in a timely manner and independently from empirical data. The standard transformed Gibbs free energies of reactions ($\Delta G_r°'$) will be determined from the total standard Gibbs free energies ($\Delta G°_{tot}$) of the metabolites, which were calculated by using Density Functional Theory (DFT). The calculated values will then be compared with experimental equilibrium constants provided by the National Institute of Standards and Technology (NIST) Database of Enzyme Reactions.

1.2 Specific Aims

1.2.1 Anchor Group Positioning in Knowledge-Based Loop Prediction

Specific Aim 1: Creation of a test dataset of insertions and deletions by 3-D alignment of protein pairs followed by identification of all possible anchor groups for each insertion/deletion.

Specific Aim 2: Fitting of loop fragments using a fragment databank made from experimental structures of the Protein Databank (PDB) [43] followed by determination of global RMS deviation between template and target and identification of the best fitting fragment for each anchor group combination.

Specific Aim 3: Evaluation of maximum prediction quality dependent on gap length, fragment length, and anchor group properties such as amino acid type, secondary structure, and relative solvent accessibility, followed by a test using a combination of the above criteria for determining optimal anchor groups.

1.2.2 *Ab Initio* Equilibrium Constant Estimation using DFT

Specific Aim 1: Retrieval of biochemical reactions, reaction conditions, and equilibrium constants from the NIST Database of Enzyme Reactions [22], followed by the creation of a database containing only data from reactions at standard biochemical conditions (298.15 K and pH 7).

Specific Aim 2: Identification of the appropriate percentage distribution for the charge isomers of each metabolite at pH 7 by using the pka prediction tool *MarvinSketch* [36], followed by a global conformational search using *Spartan 06* [47], followed by energy minimization and total standard Gibbs free energy ($\Delta G°_{tot}$) calculation using *Gaussian 03* [39].

Specific Aim 3: Evaluation of the transformed standard Gibbs free energies of reaction ($\Delta G_r°'$) by subtracting the total standard Gibbs free energies ($\Delta G°_{tot}$) of products minus reactants in each reaction, followed by an error estimation between values calculated by DFT [58] and experimental values obtained from the NIST Database of Enzyme Reactions [22].

2 BACKGROUND

2.1 Protein Structure

2.1.1 Introduction

Proteins play a role in almost all processes within a living organism. They are involved in the duplication and expression of genetic material, they take part in signal transduction and storage of particles, and they arrange into structural elements such as muscle, bone, tendons, hair, and nails. As antibodies, proteins constitute a major part of the immune system, and as enzymes they allow the progression of life-sustaining chemical reactions at acceptable rates under physiologic conditions.

The functional diversity of proteins is rooted in their structure. Proteins basically consist of one or more unbranched chains of amino acids which fold into a three-dimensional topology. The enormous variety of three-dimensional structures can be attributed to the chemical diversity of the twenty proteinogenic amino acids (Figure 2.1) as well as the possible number of sequences by which they can be arranged. Knowledge about the exact structure of proteins plays a key role in understanding their function. Proteins basically operate by binding specifically and tightly to other molecules.

The structure of proteins can be organized into hierarchical levels also referred to as primary, secondary, tertiary, and quarternary structure. The primary structure of a protein is defined by its sequence of amino acids, while the secondary structure describes local regular conformations of the backbone. Tertiary structure represents the three-dimensional shape of a protein and quarternary structure characterizes the aggregation of several polypeptide strands into multi-domain complexes. The primary sequence of amino acids contains all the information necessary for generating a stable three-dimensional protein structure [3].

2.1.2 Amino Acids and Primary Structure

Proteins are polymers made of linear chains of amino acids. A protein can consist anywhere from 50 to 25000 amino acid residues, with most proteins averaging between 200 to 300 residues [91]. Amino acids consist of an amino and a carboxyl group bonded to a central carbon atom also referred to as the C_α atom. The C_α atom also carries one of 20 different amino acid side chains and can be found in the L-configuration in almost all proteins. The relative frequency in which each amino acid is found in a polypeptide chain is rather constant across natural proteins (Table 2.1). Some variations are found in membrane proteins, where the fraction of hydrophobic residues is increased, or in special proteins such as collagen which contains repetitive patterns of glycine and proline residues [19].

Polypeptide chains are formed by condensation of the carboxyl and amino groups of successive amino acid residues, thus creating the protein backbone (Figure 2.2). Due to resonance, the peptide bond exerts about 40% double bonded character, so that the six surrounding atoms have a coplanar geometry in which neighboring C_α atoms are mostly found in the sterically favored *trans* (180°) conformation [85]. Rotations within the backbone are restricted to the two backbone torsion angles (φ and ψ) around the C_α atom (Figure 2.2).

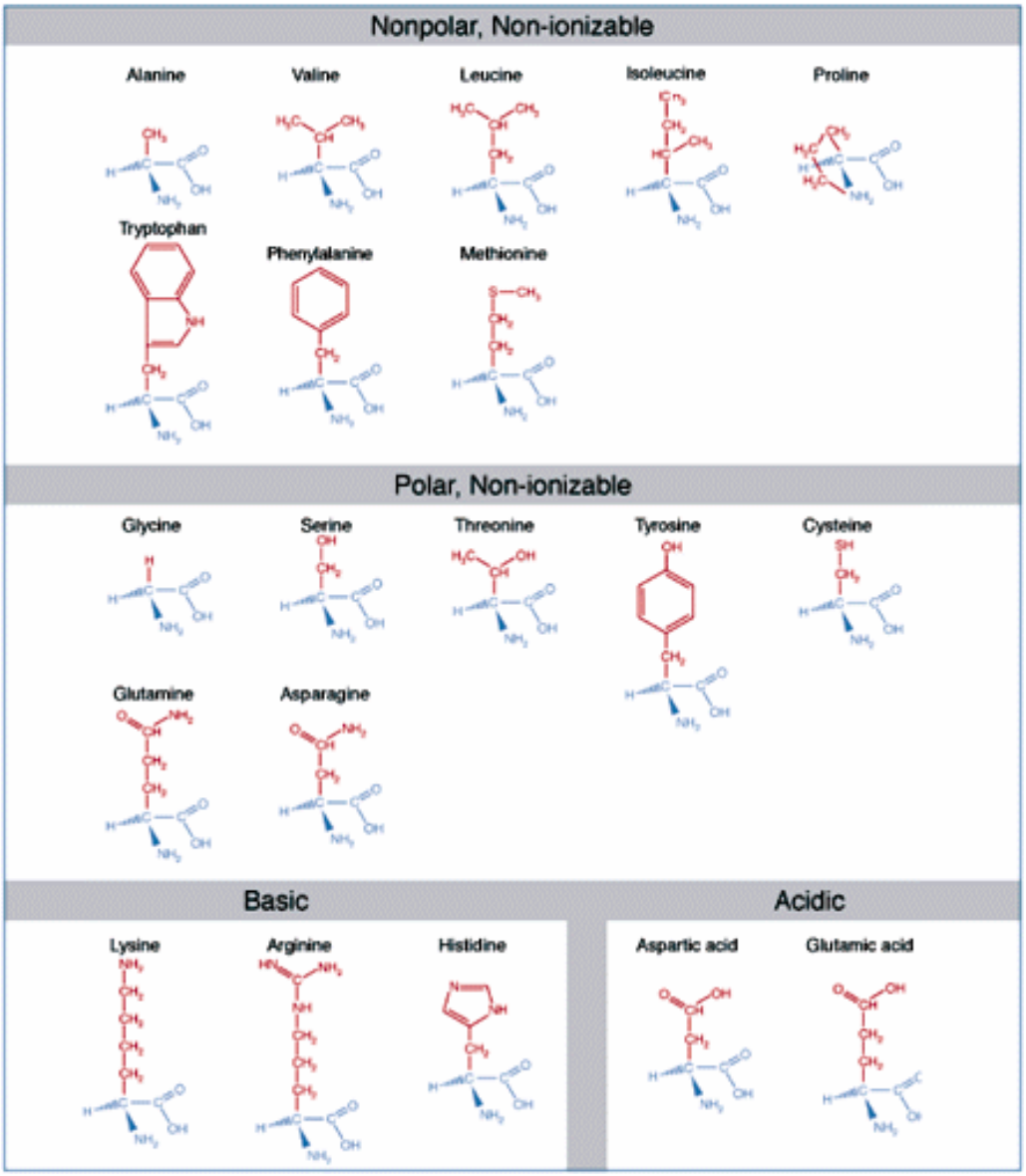


Figure 2.1: **The 20 Naturally Occurring Proteinogenic Amino Acids.** Amino acids are grouped by chemical properties of their side chains. Except for the non-chiral glycine, amino acids are found in the L-configuration in almost all proteins.

---

source: http://trc.ucdavis.edu/biosci10v/bis10v/week2/2webimages/ch5-amino-acids.jpg

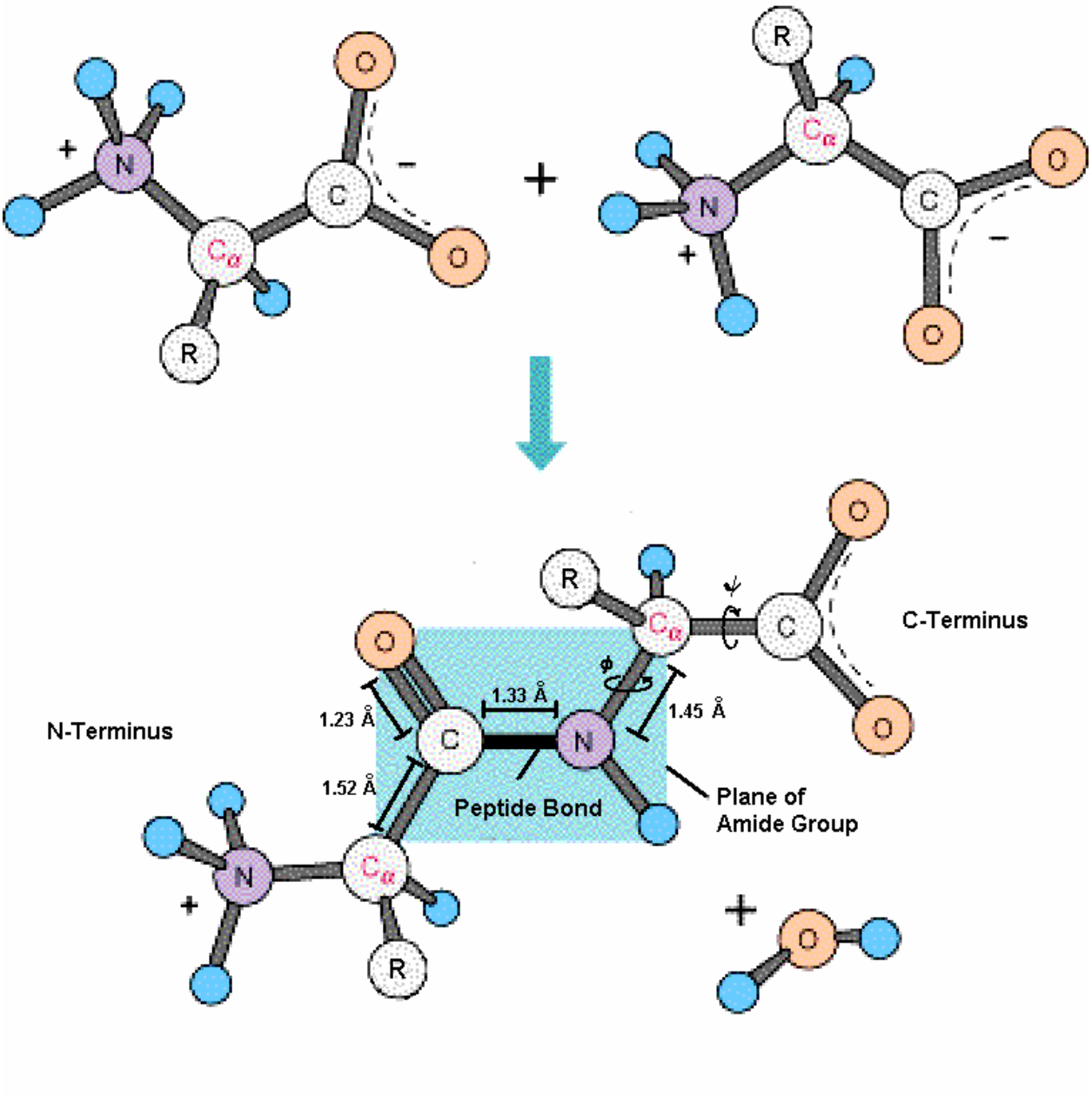


Figure 2.2: **Peptide Bond Formation.** Two amino acids are joined by condensation of adjacent amino and carboxyl groups. The peptide bond has a length of 1.33 Å which lies between the average C-N single bond (1.45 Å) and C=N double bond (1.25 Å) [84]. Planar configuration restricts backbone rotation to the two torsion angles φ and ψ.

---

source: http://hykim.cbnu.ac.kr/lectures/cellbio/3/3.htm

Amino acids differ in their structure and chemical properties. While most amino acids can be classified by general chemical characteristics such as size, charge, and polarity (Figure 2.3), some residues possess additional unique features. Glycine, for example, has no side chain and is therefore a sterically highly flexible residue. Cysteine is special in that two residues can create disulfide bridges by oxidation. Proline is an imino acid and the only residue that can form stable peptide bonds in *cis*-conformation. Proline can often be found in loops and turns (Chapt. 2.1.3).

The primary sequence of a protein can be analyzed by chemical methods. The amino acid composition is routinely identified by the complete hydrolysis of all backbone peptide bonds using 6 M HCL at about 110°C for 24-72 hours followed by chromatographic analysis of the released amino acids [27]. The most successful chemical method for determining the exact protein primary sequence has been a procedure known as Edman Degradation [14]. This procedure identifies the amino acid sequence beginning from the N-terminal residue of a polypeptide. The free N-terminus reacts with phenylisothiocyanate in basic medium followed by cleavage of the residue with trifluoroacetic acid to give the phenylthiocarbamyl (PTC) peptide. The released PTC peptide rearranges in aqueous solution to the phenylthiohydantoin (PTH) derivative and can be analyzed by chromatographic methods. By repeatedly subjecting the remaining shortened polypeptide to this procedure, one eventually obtains the entire amino acid sequence. As of today, more than 5.1 million protein sequences have been determined and are currently stored in the UniProt/TrEMBL databank [38].

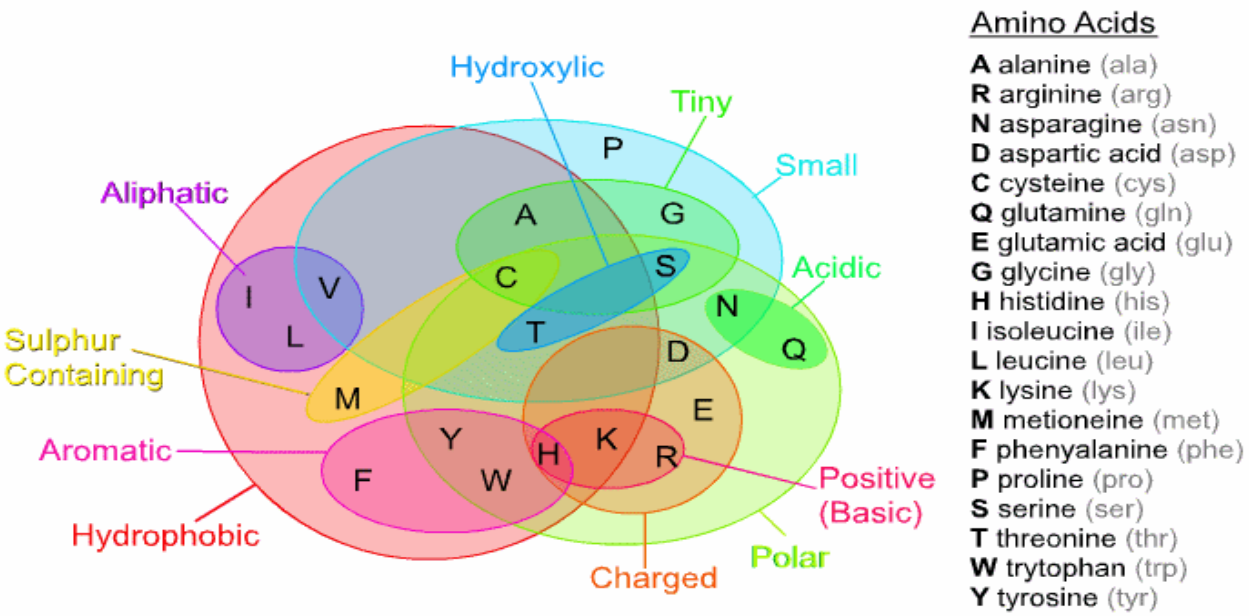


Figure 2.3: **Venn Diagram of Amino Acid Properties.**

source: http://www.dreamingintechnicolor.com/InfoAndIdeas/AminoAcids.gif

Table 2.1: **Frequency of Occurrence of Amino Acids.**

Amino Acid Residue	Mass (Daltons)	Frequency in Proteins (%)*
Ala (A)	71.09	8.3
Arg (R)	156.19	5.7
Asn (N)	114.11	4.4
Asp (D)	115.09	5.3
Cys (C)	103.15	1.7
Gln (Q)	128.14	4.0
Glu (E)	129.12	6.2
Gly (G)	57.05	7.2
His (H)	137.14	2.2
Ile (I)	113.16	5.2
Leu (L)	113.16	9.0
Lys (K)	128.17	5.7
Met (M)	131.19	2.4
Phe (F)	147.18	3.9
Pro (P)	97.12	5.1
Ser (S)	87.08	6.9
Thr (T)	101.11	5.8
Trp (W)	186.21	1.3
Tyr (Y)	163.18	3.2
Val (V)	99.14	6.6

*Frequency was determined across 1021 unrelated proteins of known sequence.

source: P. McCaldon and P. Argos, *Proteins* 4:99-122, 1988

2.1.3 Secondary Structure

The structure of a protein is primarily determined by its backbone conformation. Due to the partially double bonded character of the peptide bond, the protein backbone can be considered a chain of successive coplanar peptide units which are joined together at the C_α atoms and rotate around the two torsion angles φ and ψ (Chapt. 2.1.2). Since these two angles of rotation represent the only degrees of freedom for the protein main chain, its conformation can be completely characterized when the torsion angles φ and ψ for all residues are known. Steric collisions between side chain and backbone atoms lead to a restriction in the number of allowed conformational angles for φ and ψ. The permitted values of φ and ψ were first determined by Ramachandran using hard-sphere models with fixed bond lengths and recorded on a two-dimensional map known as the Ramachandran plot [86]. The authors observed that the flexibility of alanine residues was quite limited, with fully allowed regions occupying about 7.5% and partially allowed regions occupying about 22.5% of the total plot area (Figure 2.4). Plots for the other amino acids look similar with the exception of glycine which has more rotational freedom due to its lack of a chiral side chain.

The allowed regions of the Ramachandran plots contain specific torsion angle combinations which are repetitively found along stretches within natural proteins and are also referred to as regular conformation or secondary structure. Regular conformations are primarily stabilized by hydrogen bonding among polar atoms of the protein backbone. In a secondary structure, all amino acid residues have close to identical torsion angles and create a helical pattern characterized by a fixed number of residues per turn and translation distance per residue. Parameters of some common secondary structures are listed in Table 2.2, and their torsion angles are also found within the fully regions of the Ramachandran plot (Figure 2.4).

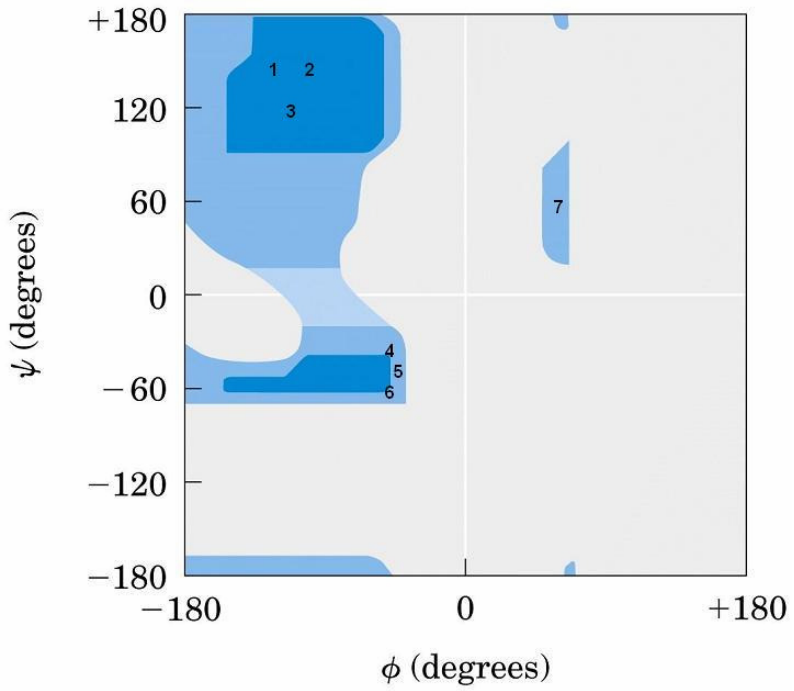


Figure 2.4: **Ramachandran Plot**. Permitted values of φ and ψ torsion angles determined using a model of alanine with hard-sphere atoms and fixed bond geometries. Fully allowed regions are dark-shaded while partially allowed regions are light-shaded. Regular conformations are marked and include anti-parallel β-sheet **(1)**, polyproline I/II and polyglycine **(2)**, parallel β-sheet **(3)**, 3_{10}-helix **(4)**, right-handed α-helix **(5)**, π-helix **(6)**, and left-handed α-helix **(7)** [86].

source: http://fmc.unizar.es/people/fff/Jsancho1/ramachandran.jpg

The most commonly observed secondary structure in natural proteins is the right-handed α-helix. In globular proteins, 31% of residues are located in α-helices [50]. The right-handed α-helix has 3.6 residues per turn, a translation of 1.50 Å per residue, and torsion angles of φ = -57° and ψ = -47° (Table 2.2). Most α-helices are between 10-15 residues in length and they are stabilized through hydrogen bonding of the backbone between the C=O

of the i^{th} residue and the -NH of the $i+4^{th}$ residue further down the chain (Figure 2.5 B). The dipole moment of the hydrogen bonds is pointed parallel and in the direction of the dipole moment of the peptide groups so that they complement each other. The entire α-helix can thus be seen as a macrodipole with a negatively charged carboxyl end and positively charged amino end (Figure 2.5 A), with the absolute value of the helix dipole moment being proportional to the number of residues. The side chains of the α-helix residues point outward the helical cylinder with some α-helices having primarily non-polar residues located along one side and polar residues along the opposite side. These α-helices are also known as amphiphatic helices (Figure 2.5 C&D) and they tend to aggregate into larger structures like helix bundles or coiled coils (Chapt. 2.1.4). Other known helices include the 3_{10} helix, the left-handed α-helix, and the π-helix. The 3_{10} helix can be found at the finishing stretches of right-handed α-helices, while the other two helix types are almost never observed in natural proteins. Features of the polyproline and polyglycine helices can be found as part of the collagen triple helix [77].

Table 2.2: Parameters for Common Regular Polypeptide Conformations

Regular Conformation	Bond Angle (degrees)			Residues per Turn	Translation per Residue (Å)
	Φ	ψ	ω		
Anti-parallel β-sheet	**-139**	**+135**	**-178**	**2.0**	**3.4**
Parallel β-sheet	**-119**	**+113**	**180**	**2.0**	**3.2**
Right-handed α-helix	**- 57**	**- 47**	**180**	**3.6**	**1.5**
3_{10} –helix	**- 49**	**- 26**	**180**	**3.0**	**2.0**
π – helix	**- 57**	**- 70**	**180**	**4.4**	**1.15**
Polyproline I (right-handed)	**- 83**	**+158**	**0**	**3.33**	**1.9**
Polyproline II (left-handed)	**- 78**	**+149**	**180**	**3.0**	**3.1**
Polyglycine	**- 80**	**+150**	**180**	**3.0**	**3.1**

source: Ramachandran and Sasikharan, *Adv. Protein Chem.* 23:283-437 (1968)

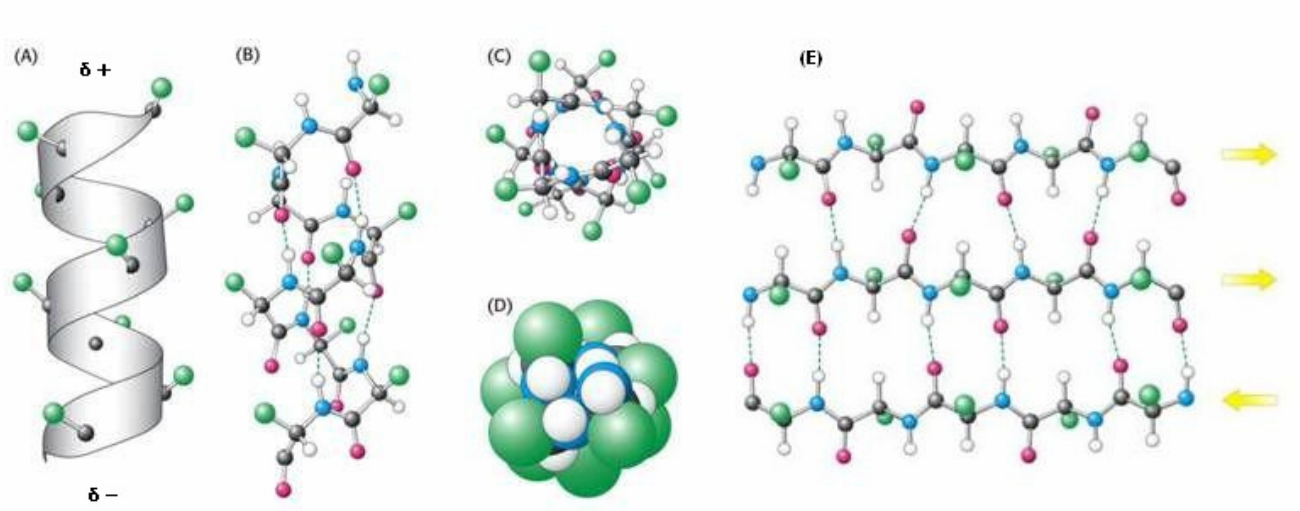


Figure 2.5: **Right-Handed α-Helix and Parallel/Anti-Parallel β-Sheet**. **(A)** α-helix with side chains tilted toward positively charged amino terminal. **(B)** Stabilization by H-bond between C=O of the i^{th} residue to –NH of $i+4^{th}$ residue along the chain. **(C,D)** Helical wheel representation of an α-helix with 3.6 residues per turn or 100° per residue. **(E)** Stabilizing H-bonds in parallel and anti-parallel β-sheets.

source: http://www.food-info.net/uk/protein/structure.htm

β-sheets are the second most commonly observed secondary structure in proteins. In globular proteins, 31% of residues are located in β-sheets [50]. β-sheets are made of multiple β-strands in adjacent arrangement to each other. In their extended form, each β-strand has 2.0 residues per turn and a translation of 3.4 Å per residue (Table 2.2). β-strands are about 5-10 residues in length and are stabilized by hydrogen bonding to a neighboring β-strand (Figure 2.5 E). When multiple β-strands are aligned adjacent to each other, a parallel, anti-parallel, or mixed β-sheet can result, depending on the relative direction of the strands to each other. The most stable type of β-sheet is the anti-parallel β-sheet, due to the short distance and parallel arrangement of its hydrogen bonds. In natural proteins, pure

antiparallel β-sheets are the most commonly found type of β-sheet, while pure parallel β-sheets occur least frequently [50]. β-sheets are also known as 'pleated' sheets due to the alternating positions of the Cα atoms above and below the β-sheet plane. The amino acid side chains within a β-sheet follow a similar pattern, pointing above and below the sheet in an alternating fashion. Side chains in β-sheets can interact with side chains of neighboring β-sheets or α-helices. In most proteins, β-sheets are not planar and flat but slightly right-twisted .

Loops and turns are often regarded as the third type of secondary structure. Unlike α-helices and β-strands, they do not display a regular conformation in terms of having a constant number of residues per turn or a fixed translation distance per residue. Instead, they serve as connecting regions between α-helices and β-sheets. Due to their general lack of regular intramolecular hydrogen bonds, loops have an irregular and flexible conformation, so that they may alter the direction of the polypeptide chain, thus permitting the formation of globular proteins. Loop regions are preferably found on protein surfaces where they frequently serve as enzyme active sites or antigen binding sites. Loops are often rich in charged and polar hydrophilic residues and can be identified using prediction schemes on amino sequences. A well known loop structure is the so-called hairpin loop which connects two antiparallel β-strands and can often be found within variable regions of immunoglobulins.

When comparing homologous protein sequences, it has been observed that residues within loops are far less conserved than in core regions. In evolutionary sense, insertions and deletions that are located in loop regions at the protein surface allow a variation in the type and number of residues within a protein without affecting the structural stability of its core. Knowledge-based prediction of the three-dimensional structure of protein loop regions using a loop fragment database is a central topic of this thesis.

2.1.4 Supersecondary Structure

In natural proteins, secondary structure elements have often been found to appear in specific arrangements called motifs. Motifs can be made of pure α-helices, pure β-strands or a mixture of both, and they are often referred to as supersecondary structure.

One example of a supersecondary structure made purely of α-helices is known as the helix-loop-helix motif. It contains two perpendicular α-helices connected by a loop region and has been found to function in repressor proteins as a recognition site for DNA or in muscle proteins as a binding site for calcium. Four adjacent α-helices connected by loops forms a motif referred to as a four-helix bundle. This motif is found in transport proteins such as the electron carrier cytochrome b 562, the O_2 carrier myohemerythrin, or the Fe^{2+} carrier ferritin. The transported particle is buried inside the hydrophobic core between the four helices.

A coiled coil is a structure created by two or more right-handed α-helices in parallel arrangement. The helices are wound around each other forming a left-handed superhelix. The contact surface consists of hydrophobic side chains along each of the helical cylinders. This arrangement is achieved by an amino acid sequence pattern called the heptad where a hydrophobic residue, often leucine, appears every seven residues [60]. Therefore, the two-stranded coiled coil is also known as a leucine zipper (Figure 2.6 A). Leucine zippers represent the DNA binding site in some transcription factors. Other two-stranded coiled coils are found as intermediate filaments and muscle myosin. Three-stranded coiled coils can be found in α-keratin and a transmembrane protein known as gp41 (Figure 2.6 B). Gp41 is located on the outer membrane of HIV viruses and mediates the attachment and injection of the virus DNA into the target cell [9].

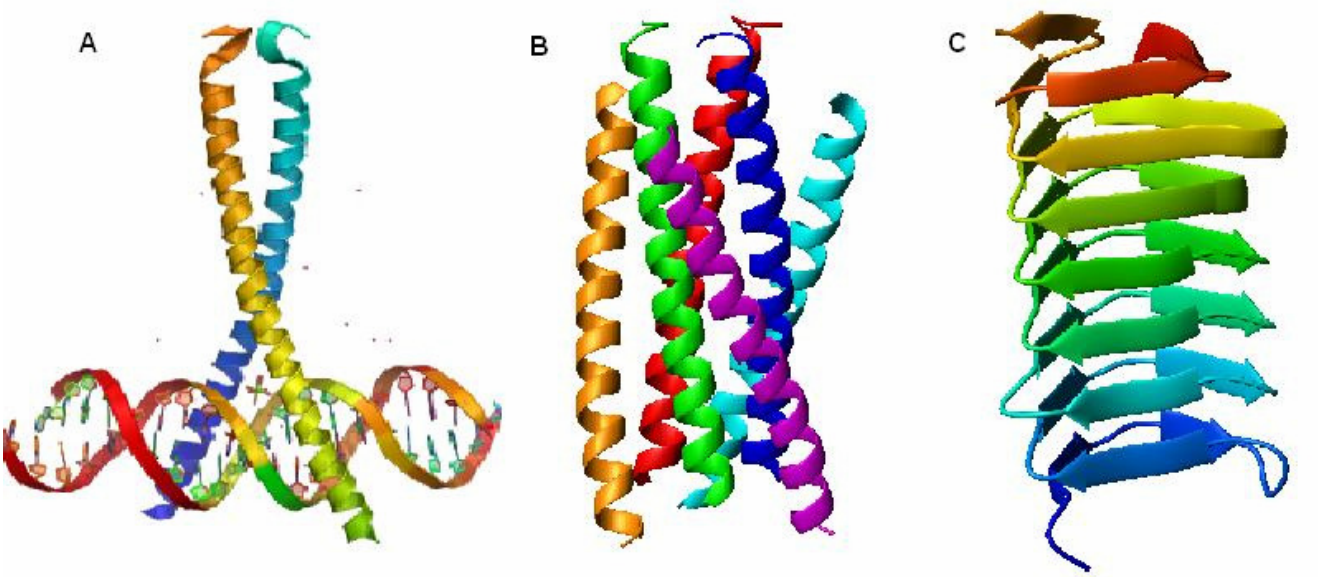


Figure 2.6: Supersecondary Structure Elements. (A) Leucine zipper as DNA binding site of transcription factors [60]. **(B)** Coiled coil hexamer of Gp41 protein of HIV [9]. **(C)** Three-faced left-handed β-helix [56].

source: http://en.wikipedia.com

Supersecondary structures can also be made from β-strands. The β-strand analog of the helix-loop-helix motif is the β-hairpin which consists of two antiparallel β-strands connected by a haipin loop (Chapt. 2.1.3). Another β-strand motif is called the β-meander. It is made from a series of multiple antiparallel β-strands connected by loops. If made of six or more β-strands, the β-meander recoils in space to form a β-barrel. β-barrels are, for example, found in porins which serve as molecular transporters in membranes or in lipocalins, like retinol binding protein (RBP) [76], which serve as extracellular transporters. Another pure β-strand motif is called the Greek key which is named after a pattern found on Greek ornamental artwork. This motif consists of four antiparallel β-strands connected by hairpin loops. Two Greek keys in succession can form a β-barrel.

A helical superstructure made of β-strands is known as a β-helix. β-helices are made of consecutive β-strands which twirl and associate in a helical pattern. They can be either two or three faced and have been found in both left and right-handed orientation. Three-faced β-helices have a shape resembling a triangular prism (Figure 2.6 C) and are found as tailspike protein of bacteriophage P22 [95] or in aggregated form as β-amyloid in Alzheimer's disease [53].

β-α-β motifs are created when two parallel β-strands are connected by an α-helical region. Several of these motifs in succession can result in the formation of α/β barrels such as the TIM-barrel which is named after the enzyme triose phosphate isomerase where the barrel was first discovered [105]. β-α-β motifs also occur in open-twisted sheets which are formed by a parallel β-sheet surrounded on both sides by α-helices. Open-twisted sheets serve as ATP-binding sites for kinases like hexokinase [96] and adenylate kinase or as NAD-binding sites for dehydrogenases like lactate dehydrogenase (LDH) [1] and alcohol dehydrogenase (LADH) [15]. NAD-binding sites consist of two symmetrical domains each made by a pair of β-α-β motifs, also known as Rossman folds [88]. Each Rossman fold binds one of the two nucleotides of NAD.

2.1.5 Tertiary Structure and Folding

Tertiary structure describes the arrangement of a polypeptide chain into its three-dimensional conformation also known as a domain. Domains can be defined as structural units which can independently fold into a stable three-dimensional structure. They can also be regarded as functional units where each unit carries out a distinct biochemical function, or they can be seen as evolutionary units where each unit can be duplicated or undergo recombination. Domains are built of several secondary elements and motifs, and they can be classified into α domains which only contain α-helices or β domains which are purely made

of β-sheets. Domains having a mixture of both α-helices and β-sheets are called α/β domains, while those containing separate α-helix and β-sheet regions are called α+β domains.

The domains of globular proteins generally have hydrophobic residues located on the inside of the protein core while the polar and charged residues are found on the protein surface. Proteins can consist of a single domain or contain several domains which aggregate into a multimeric molecule. If the domains lie on separate polypeptide chains, the protein is said to have a quarternary structure. Multiple domains of a protein may also originate from the same polypeptide chain.

Protein folding is a process which is not completely understood. Anfinsen's renaturation experiment [1] has demonstrated that the amino acid sequence contains all the necessary information for a protein to spontaneously fold into a stable three-dimensional structure. Whether a protein goes from the unfolded to the folded conformation depends on the difference in free energy between the two states. The folded conformation is stabilized by van de Waals forces and intramolecular hydrogen bonding, leading to a decrease in enthalpy. The unfolded conformation has more conformational freedom causing an increase in entropy. Whether the folded or unfolded conformation is ends up as the favored state primarily depends on external conditions such as temperature, pH, ionic strength (I) and polarity of the solvent. At standard conditions (298.15 K, pH 7, I=0), the folded state of hen lysozyme in aqueous solution is only about 16 kcal more stable than the unfolded conformation ($\Delta G°_{folded} - \Delta G°_{unfolded} = -16$ kcal/mol) [79].

The folding and unfolding processes each follow a different mechanism. Unfolding of proteins usually happens at a much higher rate than folding. Due to the cooperativity of stabilizing interactions, a breakage of one intramolecular hydrogen bond will lead to the weakening of all the other neighboring bonds, so that ultimately, the protein unfolds suddenly in a single step. The folding process is more complicated and occurs at a slower rate. In unfolded polypeptide chains, the peptide bonds are equally stable in both *cis* and

trans conformation, leading to a heterogeneous mixture of protein conformations. Since folded proteins primarily contain *trans* peptide bonds (Chapt. 2.1.2), the folding process is usually preceded by a *cis* to *trans* peptide bond isomerization step [7].

Several folding mechanisms have been proposed in the past [104][25][52]. The models all share a pre-folding state followed by a rate-limiting step. The pre-folding state resembles the state of a molten globule [81]. The molten globule has roughly the size of the final folded state and is characterized by the presence of some secondary structural elements. The transition from the unfolded state to the molten globule is fast, while the transition from the molten globule to the final native conformation is slow and cooperative.

The rate of protein folding lies between 0.1 and 10^3 seconds, suggesting that protein folding is not a trial-and-error process in which the protein goes through the entire range of possible conformations. In a thought experiment known as the Levinthal's paradox [65], Levinthal calculated that a 100 residue polypeptide with 10 conformations per residue and 10^{-13} seconds per conformation would lead to a folding time of 10^{77} years which would exceed the age of the universe. Therefore, protein folding must be a somewhat directed process.

Folding models suggest that the process of folding is not only ruled by thermodynamic considerations but also by kinetic aspects. Mutation studies were able to demonstrate that some amino acid residue exchanges may prevent the protein from completing the folding process without affecting the stability of the already folded state [23]. Thus, one may conclude that the final native conformation does not necessarily have to be the thermodynamically most stable but could simply be the most stable kinetically accessible conformation, so that there may be stable folded conformations for which no folding pathways exist [80].

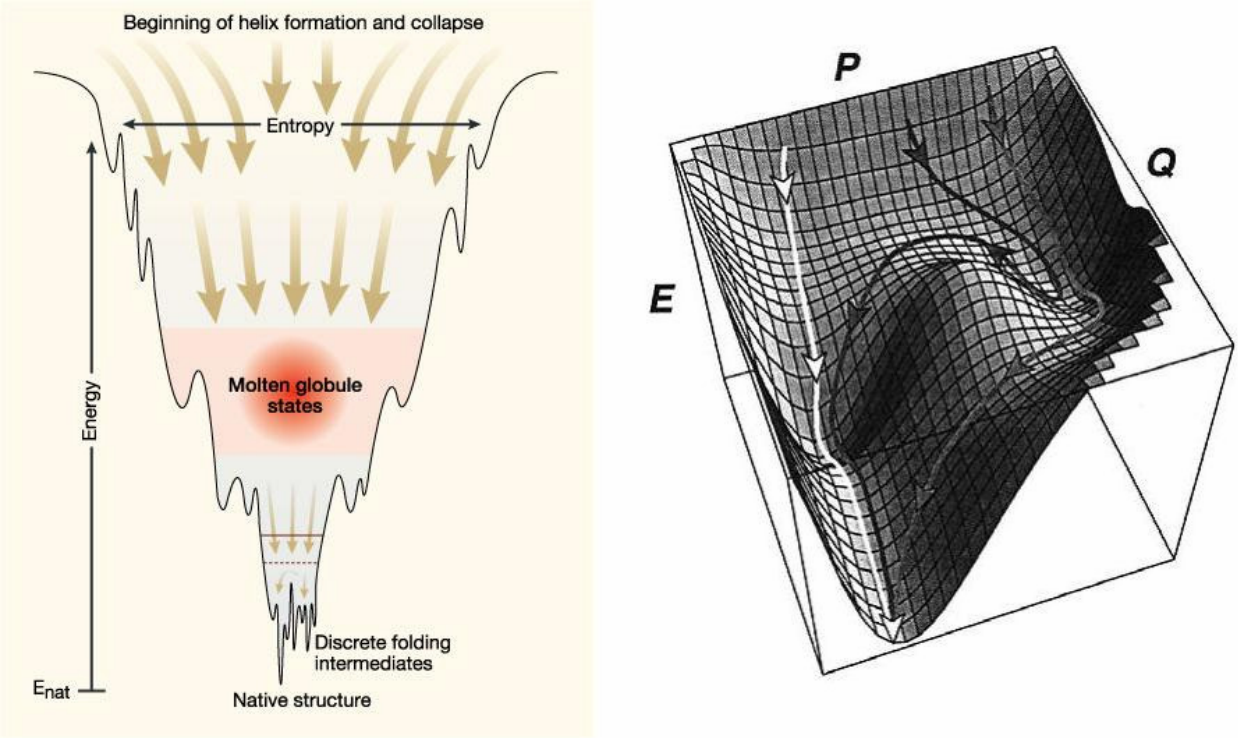


Figure 2.7: **Energy Landscape of Protein Folding.** Folding funnel showing steps towards completion of native conformation via molten globule states and various folding intermediates (left). Energy landscape displays alternate pathways leading to same native structure at bottom of the funnel (right).

source: http://www.nature.com/horizon/proteinfolding/images/summ_f1.jpg

Protein folding can be visualized by an energy landscape resembling a folding funnel where the most stable conformation is located at the bottom tip of the funnel (Figure 2.7). The folding process can thus be compared to water moving down the funnel moving towards the final native conformation which could either be the tip of the funnel (global minimum) or one of the humps in between (local minima). This model also shows that the same native conformation may be reached by different kinetic routes. Folding of large proteins with multiple domains contain the additional problem of aggregation and precipitation. *In vitro*, the folding of such large proteins must be carried out at very low protein concentrations. *In vivo*, protein folding is often assisted by molecular chaperones which temporarily bind to

unfolded parts of the polypeptide chain to prevent them from aggregation. Other folding-related enzymes include prolyl peptide isomerase and protein disulfide isomerase. Prolyl peptide isomerase accelerates the *cis-trans* isomerization ahead of proline residues [17], while protein disulfide isomerase catalyzes the rearrangement of disulfide bonds after they have been formed [66].

Even though the number of theoretically possible three-dimensional folded protein conformations is astronomical, the actual number of different folds occurring in natural proteins is rather small and estimated to include about 1,000 [43]. This could be explained by the fact that during evolution, tertiary structure has been much more conserved than primary structure. Therefore, proteins which originate from evolutionary related species often have very similar folded conformations despite larger differences in amino acid sequence. The cytochrome and serine protease families represent such examples [82]. In these proteins, large variations in the primary sequence mostly occur at the protein surface while residue changes in the protein core occur in a way that preserves torsion angles. Residues around the active site of the protein are highly conserved. Homology has not only been observed among different proteins but also within the same polypeptide chain. Examples for internal homology have been observed for Ferredoxin, parvalbumin, and some immunoglobulins. These proteins usually consist of two or more domains where one domain is thought to have developed from the other by gene duplication [98][70].

2.1.6 Homology

Protein homology is an ambiguous term and can either be understood as evolutionary proximity, similarity in function, similarity in tertiary structure, or sequence identity. Which of the above definitions applies usually depends on the context in which the term is used.

Generally, two proteins are considered homologous when their sequence is identical above a certain percentage value which depends on the length of the alignment. The degree of sequence identity is determined by the root mean square deviation (rmsd) value which equals to the number of identical residues divided by the length of the shorter protein sequence. For alignments of large proteins, sequence identity of 30% is generally sufficient to assume homology.

It has been shown that sequence identity is highly correlated to structural similarity. Chotia and Lesk [10] showed that the difference in the structure of two proteins increases as the sequence identity decreases, while Rost [89] demonstrated that for the alignment of long sequences, a sequence identity of 40% and higher guarantees structural similarity. Some proteins have similar folds or similar functions, yet are very low in sequence identity. This includes several mononucleotide-binding domains, also known as Rossman folds [15] (Chapt. 2.1.4). These domains differ substantially in their primary sequence and have no proven evolutionary relationship, yet still have been found to be similar in three-dimensional structure and function. The question of whether these conformational similarities have arisen from convergent or divergent evolution or happened by pure chance is still open to debate [82]. Examples for functional identity without structural similarity are given by trypsin proteases and serine carboxypeptidases. These enzymes have similar functions but no structural similarities other than their active sites. In this case, functional similarity is thought to have developed from convergent evolution [82].

2.1.7 Structure Classification

Proteins can be classified by a variety of classification systems. The most widely used systems include SCOP [33], CATH [30], and FSSP [37]:

SCOP (Structural Classification of Proteins) is a completely manual classification system which orders proteins in a hierarchy using four levels with increasing specificity [33]:

- **Class:** General structural architecture of domains (α-helix, β-sheet, α/β, α+β, multi-domain, membrane proteins, etc.) (Chapt. 2.1.5)
- **Fold:** Similar arrangement of secondary structure with or without evolutionary relationship
- **Superfamily:** Probable common evolutionary relationship with or without sequence similarity
- **Family:** Clear evolutionary relationship either based on sequence identity (30% or greater) or common structure / function.

The **CATH (Class, Architecture, Topology, Homologous Superfamily)** system [30] is built up in a similar manner using four categories (Figure 2.8):

- **Class :** Total of four classes grouped by general secondary structure content (α-helix, β-sheet, α/β-mixed, few secondary structure)
- **Architecture:** Total of 35 architectures grouped by similar shapes and structures
- **Topology:** Similar structure and arrangement of secondary structure without evidence of homology. Comparable to 'fold' category in SCOP system (see above)
- **Homologous Superfamily**: Probable evolutionary relationship without sequence homology (similar to 'superfamily' category in SCOP system)

All levels are assigned automatically according to structure or sequence similarity, except for 'architecture' which is assigned manually.

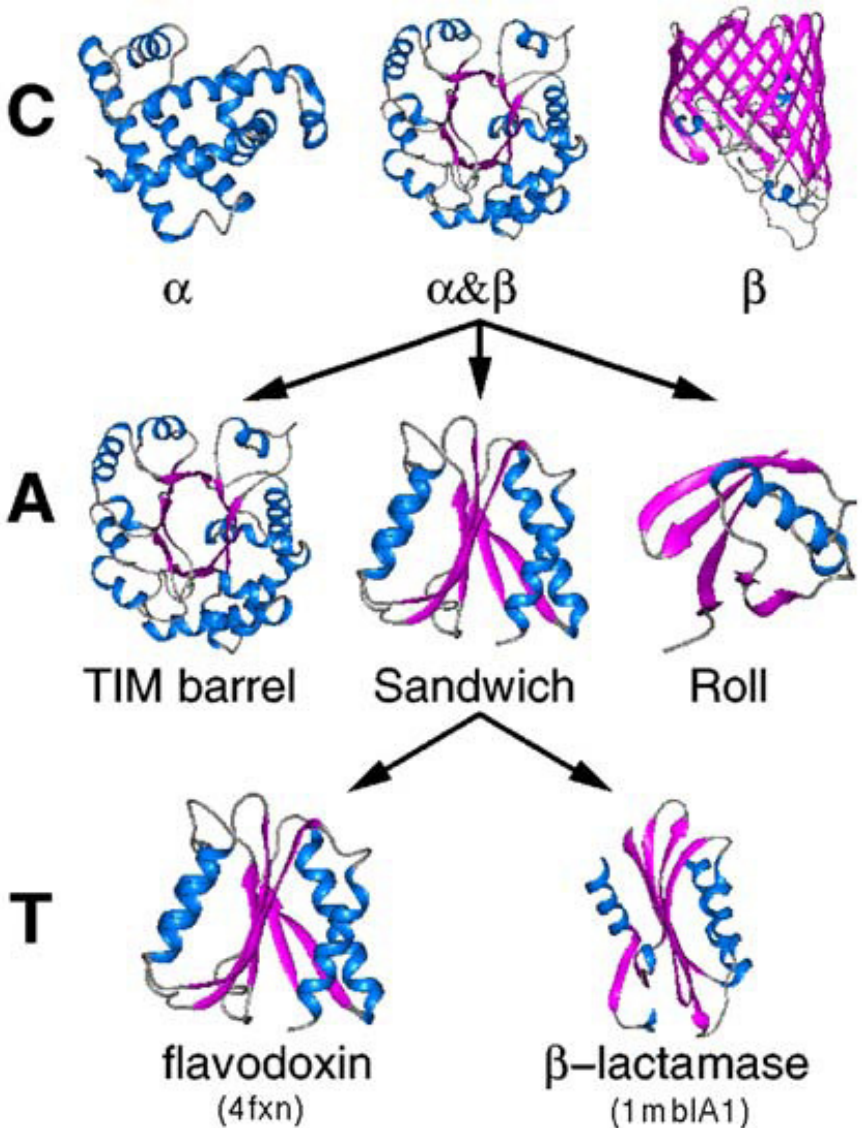


Figure 2.8: **CATH Protein Classification System.** First three levels of CATH protein classification system. Levels are ordered in increasing specificity. Class (C) and topology (T) are assigned by automatic methods while architecture (A) is manually assigned.

---

source: http://en.wikipedia.com

FSSP (Families of Structurally Similar Proteins) is a fully automated system which is based on the comparison of polypeptide chains rather than protein domains. This system uses the DALI [29] algorithm to classify protein structures by their level of homology. Close homologs (>70% sequence identity) are represented by a single structure while medium homologs (30% - 70% sequence identity) are grouped into structural families [37]. This process results in a reduced subset of representative sequences which can be used for searching remote homologous proteins. The FSSP system has the advantage of providing immediate access to structural alignments. In addition, it reports the degree of sequence identity based on root-mean-square deviation (Chapt. 2.1.6). The disadvantage of this system is that it may produce misleading results especially with regard to multi-domain proteins, since important pieces of information on structural domains are neglected.

2.1.8 Structure Determination Methods

X-ray crystallography is the most widely used method for analyzing protein structures. More than 85 % of structures stored in the Protein Data Bank (PDB) have been determined by x-ray diffraction techniques [45]. In order for a protein to be measured by this method, a large and well ordered protein crystal needs to be grown. Building protein crystals is a slow, tedious and not always successful process. Crystals are grown in supersaturated solutions by slowly decreasing the solubility of the protein while empirically varying a number of external parameters such as temperature, pH, and the concentration of additives, in an attempt to find optimal conditions for crystal growth. Crystals may then develop in a setup such as a hanging drop after an initial nucleation step and usually remain high in solvent content after crystallization (on average between 40-60%). Thus, the three-

dimensional structure of crystallized proteins has a high resemblance to the native structure in solution.

A protein crystal suitable for measurement contains about 10^{15} molecules. During measurement, the crystal is irradiated with x-rays while the two-dimensional diffraction patterns representing 'slices' of the crystal are being recorded on photographic film or a charged coupled device (CCD) image sensor [32]. While the crystal is rotated in small steps around slightly more than 180° (Ewald Sphere), diffraction patterns are recorded from all angles of the crystal. A diffraction pattern contains multiple reflections where each reflection has a specific intensity and position recorded in Miller indices (*h,k,l*). Miller indices are lattice coordinates which represent points the original crystal lattice in reciprocal space. The inverse relationship between crystal lattice and diffraction pattern coordinates is reflected in Bragg's Law of diffraction [6]:

$$n\lambda = 2d \sin\theta$$

In the above equation, λ is the x-ray wavelength, while d stands for the distance in lattice points of the crystal lattice in real space, and θ is the reflection angle. Braggs Law shows that an interference pattern based on constructive interference of the scattered waves occurs whenever the phase shift is a multiple of 2π.

The final goal of x-ray diffraction analysis is the construction of an electron density map which displays the position of each atom in a protein crystal. Electron densities of each position in real space ρ(x,y,z) are calculated by summation of the intensity and positional information on all the recorded diffraction patterns using a method known as Fourier synthesis or Fourier transform:

$$\rho_{(x,y,z)} = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l} F_{(h,k,l)} \exp\left[-2\pi \cdot i(hx + ky + lz)\right]$$

The above equation shows that the Fourier transform requires values for $F_{(h,k,l)}$. These are the so-called structure factors which contain information about the amplitude and phase for each reflection in reciprocal space. Structure factor $F_{(h,k,l)}$ is a vector with amplitude $|F_{(h,k,l)}|$ and phase $e^{i\phi(h,k,l)}$:

$$F_{(h,k,l)} = \left|F_{(h,k,l)}\right| e^{i\Phi(h.k.l)}$$

The amplitude $|F_{(h,k,l)}|$ can be directly obtained from the intensity $I_{(h,k,l)}$ of the reflection of the diffraction patterns as one is proportional to the square of the other ($I_{(h,k,l)} \sim |F_{(h,k,l)}|^2$) . The phase information $e^{i\phi(h,k,l)}$ cannot be obtained experimentally, and this is also known as the phase problem in x-ray crystallography [100]. Estimations of phases can be performed by introducing heavy atoms into or in between the crystal atoms. This process alters the reflections of the crystal atoms by interference with the diffracted waves of the crystal points while leaving its lattice structure intact:

- **Single / Multiple isomorphous replacement (SIR / MIR):**
  Introduction of heavy atoms such as uranium or platinum into the crystal by soaking in heavy atom solution, co-crystallization or site-directed mutagenesis [11]
- **Single-/Multi-wavelength anomalous diffraction (SAD / MAD):**
  Incorporation of anomalous scattering atoms such as selenium into the crystal structure by expressing the protein with a methionine auxotroph using a medium containing seleno-methionine [26]

By using the above methods, the diffraction patterns of both crystals with and without heavy atoms are recorded and compared. The way in which the intensities of the reflections of the crystal are affected by the reflections of the heavy atoms depends on their relative phases. Therefore, if the phase and intensity of the heavy atoms are known, they can

be subtracted from the combined signal to yield the net phases of the crystal atom reflections. The phases and intensities of the heavy atoms are determined by first finding their position in the protein crystal. This is done by using the Patterson function, which is essentially a Fourier transform of the intensities rather than the structure factors, since $I_{(h,k,l)} \sim |F_{(h,k,l)}|^2$:

$$P(u,v,w) = \sum_{hkl} |F_{hkl}|^2 \; e^{-2\pi i(hu+kv+lw)}.$$

The Patterson function allows the construction of a Patterson map which is a map containing the distance vectors of all atoms relative to each other normalized to the origin position. These Patterson vectors can then be used to calculate the original position of the heavy atoms. The Patterson map of the heavy atoms is actually a Difference Patterson map, because the structure factor amplitudes of the heavy atoms were estimated as the difference between the crystal containing both the protein and the heavy atoms minus the protein alone without the heavy atoms:

$$\left|F_{h,k,l(Heavy)}\right| \approx \left|F_{h,k,l(\mathrm{Protein+Heavy})}\right| - \left|F_{h,k,l(\mathrm{Protein})}\right|$$

Besides the above method, alternative phase estimation methods have also been in use:

- **Molecular Replacement:** Deriving phase information by superimposing the Patterson map of a known homologous protein on top of the unknown protein crystal followed by refinement of the remaining unknown crystal.
- **Direct Methods (*ab initio* phasing)**: This method uses statistical phase relationships between certain groups of reflections and is used only for small proteins (<1000 atoms) [101]

Once the phases have been determined, an initial electron density map is generated and used as a model to refine the phases iteratively in subsequent steps. During each step, the R-factor represents a measure of the correlation between the structure factor amplitude of the observed experimental diffraction data and the calculated model:

$$R = \frac{\sum_{h,k,l} \left\| F_{h,k,l(obs)} \right| - \left| F_{h,k,l(calc)} \right\|}{\sum_{h,k,l} \left| F_{h,k,l(obs)} \right|}$$

The R-factor of the initial model usually ranges around 0.4 - 0.5 and should be refined to around 0.1 for proteins. Each refinement step involves the correction of the atomic positions as well as the improvement of the temperature factor (Debye-Waller factor) [103] which balances disorders due to harmonic thermal vibrations of the atoms. Resolution is another important criterion for the quality of an x-ray structure determination. The resolution is defined as the minimum interatomic spacing that gives rise to the reflections in the diffraction pattern and should be at least 3 Å for protein crystals. The structure determination of proteins, as opposed to small molecules, rarely yields electron density maps showing all the individual atoms. Therefore, many atoms have to be fitted using standard geometries for the protein backbone (Chapt. 2.1.3). Consequently, knowledge of the primary structure is almost a requirement for successful protein structure determination.

Alternative methods for determining protein structures include neutron diffraction, electron diffraction, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy. Neutron and electron diffraction techniques are rarely used for protein structure determination, while electron microscopy is only performed to obtain low resolution information (max. 20 Å) of large protein complexes and cell organelles.

NMR spectroscopy is the second most used experimental technique for protein structure determination (Chapt. 2.1.9). This method works by measuring the nuclear spin

transitions of ^{1}H, ^{13}C, ^{15}N, etc. and has the advantage of determining the structure of proteins in concentrated solution. The experiments aim at finding molecular restraints which include distance restraints, angle restrains and orientation restraints. Distance restraints represent the range of proximity between measured nuclei as determined by a number of correlation spectroscopy (COSY) experiments. Similarly, angle restraints refer to ranges in torsion angles and are generated from coupling constants using the Karplus equation [51]. The experimentally determined restraints are combined with general properties of proteins such as bond lengths and angles and then converted into energy terms. The energy terms are minimized using algorithms, resulting in a number of structural models in solution. The disadvantage of using NMR spectroscopy for protein structure determination lies in the overlap of signals. Even though the experiments are multidimensional and performed in combination with isotope labeling, interpretation of the results is extremely difficult. Therefore, the technique is usually limited to the structure determination of smaller proteins. As of today, approximately 7000 structures in the Protein Data Bank (Chapt. 2.1.9) have been determined by NMR spectroscopy.

2.1.9 Protein Data Bank (PDB)

The Protein Databank (PDB) represents publicly available collection of experimentally determined protein structures. The databank was established in 1971 at the Brookhaven National Laboratory and originally contained 7 structures. Since 1998, the PDB has been managed by the Research Collaboratory for Structural Bioinformatics (RCSB). So far, the size of the database has risen exponentially and currently holds more than 48,000 structures (Table 2.3).

Table 2.3: Protein Data Bank (PDB) Statistics of February 2008.

Experimental Method	Proteins	Nucleic Acids	Protein/Nucleic Acid Complexes	Other	Total
X-Ray	**38541**	**1016**	**1770**	**24**	**41351**
NMR	**6080**	**802**	**137**	**7**	**7026**
Electron Microscopy	**112**	**11**	**41**	**0**	**164**
Other	**87**	**4**	**4**	**2**	**97**
Total	**44820**	**1833**	**1952**	**33**	**48638**

source: http://pdbbeta.rcsb.org/pdb/statistics/holdings.do

Even though the number of submitted structures increases every year, the number of unique structures is comparably low. After eliminating structures with more than 90 % similarity using BLAST, the number of sequences reduces to about 18,000 [44], which means that about 60% of protein sequences in the databank are redundant. When using a filter of sequence similarity of less than 30%, the number structures reduces further down to about 10,000, which demonstrates that only slightly more than 20% of structures in the PDB can be considered unique structures. In comparison, the amount of available protein primary sequences has been increasing steadily due to the latest amount of genome projects. Currently, about 5.1 million known protein sequences are stored in the UniProt/Trembl databank [38], so that the gap between known protein sequences and solved structures can be expected to further increase in the following years.

Structural genomics is a novel field in which large amounts of protein structures are currently being determined and submitted to the PDB on a genome-wide scale, using a combination of experimental and computational methods. The work is performed by centers of structural genomics, making protein structural determination a cost-effective procedure

compared to regular laboratories. This field provides access to a vast amount of new proteins most of which, however, are of unknown function and lack corresponding publications.

2.1.10 Enzymes

Enzymes are proteins that catalyze biochemical reactions by converting substrates to products. Enzymes allow an increase in the rates of the reactions by lowering their activation energy. Almost all biochemical reactions in a cell which would normally not occur at acceptable rates under physiologic conditions are catalyzed by enzymes. In addition, enzymes are specific for certain substrates. Similar to other catalysts, enzymes do not alter the equilibrium of a chemical reaction. However, their activity may be affected by environmental factors such as temperature, pH, substrate concentration, or by other molecules which is also known as inhibition.

Except for a small number of RNA-enzymes, also referred to as ribozymes [99], almost all enzymes are made of globular proteins. Enzymes normally have a specific binding site for substrates and an active site made of around 3-4 residues, which carries out the catalysis. Some enzymes contain additional binding sites for cofactors which are needed for the catalytic process. Binding sites of enzymes can be highly substrate specific and often display a combination of stereospecificity, regiospecificity, and chemoselectivity. The activity of enzymes can be inactivated by denaturation of its structure by increases in temperature or changes in pH (Chapt. 2.1.5). Several enzymatic domains can form enzyme multi-domain complexes which may perform a successive combination of reactions on a given substrate.

Enzymes are catalysts which speed up biochemical reactions by lowering their activation energy. This process can be achieved in several ways. The enzyme can create an environment which stabilizes the transition state of the reaction either by distorting the

molecular shape of the substrate or by charge stabilization using residues with an opposite charge distribution compared to the transition state. Other mechanisms include the formation of a temporary reaction intermediate called enzyme substrate (ES) complex or simply the approximation of two substrates together in the correct orientation thereby causing a reduction in entropy.

Enzymes have the ability to catalyze up to several million reactions per second. The reaction rates of an enzyme can be measured in enzyme assays where these rates are determined for different substrate concentrations, while temperature, pH, and ionic strength of the solution are kept constant. While the concentration of the substrate is being increased, the respective reaction rates are measured until they reach a constant maximum value. At this maximum rate (V_{max}), all enzyme binding sites are saturated with substrate, so that the concentration of free enzyme reaches zero while the concentration of the enzyme substrate (ES) complex equals to the initial enzyme concentration. The Michaelis-Menten [63] constant (K_m), also called global dissociation constant, is defined as the concentration of substrate at which the reaction reaches half of its maximum velocity (½ V_{max}). K_m values are characteristic for each enzyme and their value corresponds to the binding affinity between enzyme and substrate. Another important quantity is the specificity constant k_{cat}/K_m , where k_{cat} is the turnover number. The turnover number equals to the number of substrate molecules converted into products per second and active site and reflect the enzyme's catalytic ability. The specificity constant k_{cat}/K_m, on the other hand, can be regarded as the efficiency of the enzyme in converting substrates into products. Theoretically, the maximum specificity lies at about 10^8-10^9 $M^{-1}s^{-1}$ which is also called he diffusion limit. At this rate, the generation of product no longer depends on the rate of the enzyme itself but is limited by the rate of diffusion of the substrate. Examples for these so-called kinetically perfect enzymes are triose-phosphate isomerase, carbonic anhydrase, catalase, and superoxide dismutase [20].

Some enzymes have been observed to display reaction rates above the diffusion limit, leading to quantum tunneling as a proposed mechanism for enzyme catalysis [67].

Enzymes are classified by their Enzyme Commission (EC) number which is a classification system for enzymes based on the reactions they catalyze. It is also a nomenclature system in which every EC number is accompanied by a recommended enzyme name. The nomenclature scheme was first introduced in 1961 by the Enzyme Commission which was established at the 1955 International Congress of Biochemistry and today includes a total of 3196 different enzymes [42]. The numbering system consists of the letters 'EC' followed by four numbers separated by periods. The numbers represent an increasingly more detailed classification of the enzyme. The first number in the EC system categorizes the enzyme into one of six top-level enzyme categories also known as EC groups (Table 2.4). EC groups include oxidoreductases (EC 1), transferases (EC 2), hydrolases (EC 3), lyases (EC 4), isomerases (EC 5), and ligases (EC 6). The second number refers to the type of bond the enzyme acts (e.g. peptide bond), while the third and fourth number point to specifics with respect to the atomic location of the molecular site of action.

Table 2.4: Top Level EC Numbers (EC Groups).

EC Group	Reactions Catalyzed	Examples by trivial name
EC 1 (Oxidoreductases)	**redox reactions, electron transfer**	**Dehydrogenase, Oxidase**
EC 2 (Transferases)	**transfer of groups between molecules**	**Kinase, Transaminase**
EC 3 (Hydrolases)	**hydrolysis, condensation**	**Lipase, Peptidase, Amylase**
EC 4 (Lyases)	**bond cleavage, double bond formation**	**Carboxylase**
EC 5 (Isomaerases)	**intramolecular rearrangement**	**Isomerase, Mutase**
EC 6 (Ligases)	**formation of single bond using ATP**	**Synthetase**

source: http://en.wikipedia.org/wiki/EC_number

2.2 Protein Structure Prediction

2.2.1 Introduction

The idea that the folding of proteins is dictated by their primary structure (Chapt. 2.1.5) leads to the possibility of predicting folded conformations beginning from the sequence of amino acids. The current amount of available protein structures is still small compared to the number of known primary sequences (Chapt. 2.1.9), so that this gap may be closed by solving structures using computational techniques.

As of today, the complete prediction of a tertiary structure from its primary sequence is still elusive. The astronomical amount of possible conformations, the uncertainties about the folding process, and the redundancy and flexibility in the correlation between sequence and folded conformation all contribute to this problem (Chapt. 2.1.5). Nevertheless, some general rules still apply, such as the tendency towards a packed interior hydrophobic core paired with a polar surface (Chapt. 2.1.5), as well as the preference for favorable torsion angles (Chapt. 2.1.3). These rules serve to minimize the free energy of the folded conformation leading to either a three dimensional structure of lowest possible energy or kinetically most accessible structure (Chapt. 2.1.5).

Protein structure prediction comprises an array of methods including *ab initio* modeling, knowledge-based prediction (fold recognition and homology modeling), loop prediction, and side chain placement. In an actual modeling situation of an unknown protein structure, *ab initio* and knowledge-based methods are sometimes used in combination. Model quality assessment refers to the accuracy of a predicted protein structure compared to the experimentally solved structure.

2.2.2 *Ab Initio* Modeling

Ab initio modeling or *de novo* protein structure prediction methods attempt to build protein tertiary structures from their primary sequence based only on amino sequence combined with physical and chemical principles. The modeling process involves either the application of energy functions for global optimization or a mimicking of the protein folding process.

The method of *ab initio* modeling of protein structures can be divided into geometry representation, application of an energy function, and employing of a search method. Geometry representation allows a decrease in computational costs down to an acceptable level, which can be achieved either by simplifying the atomic model or by down scaling the space coordinate system. Atomic models can be simplified down to one representative atom per residue while the search space can be reduced to lattice models. Energy functions are used to assess the thermodynamic stability of protein conformations and include either physical or statistical knowledge-based potentials. These functions should be able to numerically distinguish folds which resemble the native structure from ‘misfolds‘. Due to the enormous amount of possible conformations (Chapt. 2.1.5), algorithms have been used to guide the search leading to the energetically most stable structure. The most widely used search method is the Monte-Carlo algorithm [18].

Ab initio structure prediction by itself has not been successful due to the astronomical number of possible conformations and the inaccuracy of energy functions in distinguishing stable structures from ‘misfolded’ conformations. Therefore, this method is usually used in combination with knowledge based methods such as fold recognition, where an *ab initio* step is used for refining the torsion angles of an initial model.

2.2.3 Fold Recognition Modeling

In fold recognition modeling, the structure of a given primary sequence is constructed by using with a library of known protein folds. In this method, the backbone of the primary structure is superimposed onto the folding model by using an algorithm which searches for the optimal alignment between target sequence and template fold. Subsequently, scoring functions based on empirical energy functions and statistics of known structures are used to evaluate the 'goodness of fit' and establish a ranking of suitable models.

Fold libraries can be constructed based on protein classification schemes such as SCOP, CATH, and FSSP (Chapt. 2.1.7). To make the library more efficient, representative folds may be clustered out to reduce redundancy and save computation time. Scoring functions evaluate the compatibility of a certain primary sequence residue to its environment in the three-dimensional structure. Compatibility criteria may include secondary structure, side chain overlaps, and solvent accessibility. Fold recognition methods can be subdivided into structural (3D-1D) profile methods, threading methods, sequence profile methods, and mapping methods.

2.2.4 Homology Modeling

In homology modeling, the unknown three-dimensional structure of a given target primary sequence is determined by alignment of the amino sequence against a known structural template from a homologous protein. The growing number of available protein structures in the PDB (Chapt. 2.1.9) has made this method increasingly successful for determining protein structures. Homology modeling can be divided into the following steps: selection of a suitable protein template, alignment of template structure and target sequence,

initial model building, loop prediction, side chain prediction, structural refinement and model quality assessment.

Identification of one or more suitable protein structure templates can be performed either by sequence alignment or fold recognition methods (Chapt. 2.2.3). Here, sequence identities of 30% or greater are usually sufficient to assume structural similarity (Chapt. 2.1.6). PSI-BLAST [2], the alignment algorithm used for this purpose, is able to incorporate evolutionary relationships between proteins by using protein profiles of protein families. Additional criteria may be used including the use of a phylogenetic tree, the consideration of environmental factors such as ligands, solvents, pH, and quaternary interactions, and the verification of the experimental x-ray crystal structure quality with respect to R-factor and resolution (Chapt. 2.1.8).

For the modeling process, alignment of the target sequence to the template protein is the most important step. For template-target identities of 40 % and above, the alignment shows good results. However, for identities of 30 % and below, the alignment accuracy drops significantly, leading to a major source of error in the homology modeling process [94]. Usually, a more than one alignment is used for the prediction process, since sub-optimal alignments may sometimes still lead to yield good prediction models.

After an appropriate target to template alignment has been defined, the initial protein model is built by copying the structurally equivalent residues of the target sequence onto the 3-D coordinates of the template. At this point, the side chains are ignored and only the backbone coordinates are transferred. The gaps or non-matching residues in the alignment represent insertions and deletions which are also known as structural variable or loop regions (Chapt. 2.1.3).

The modeling of loop regions is a crucial part in the prediction procedure, since structurally variable regions are mostly located on the protein surface, often representing important binding sites or defining the functional specificity of the protein. On the other

hand, loop regions also contain the most amino acid substitutions (insertions and deletions) and have no defined secondary structure, so that the prediction of loop regions is considered the second largest source of error in homology modeling. In this thesis, loop prediction represents a central part and will be discussed in more detail in the following chapter.

In a side chain placement step, side chain atoms are positioned onto the backbone by either selecting the most frequent side chain conformation from a library of preferred rotamers or by directly copying the side chain of the homologous residue from the template structure. Refinements of the final structure may then be performed using energy minimization functions.

The modeling process is finally concluded by a step called model quality assessment. Model quality assessment is conducted by using scoring functions which identify the best model among a set of alternative conformations. Scoring functions can be categorized into physics-based energy functions, knowledge-based scoring functions, and statistical potentials.

2.2.5 Loop Prediction

Loop prediction or loop modeling aims at remodeling structurally variable regions in protein structures. Loop regions are of specific importance in protein prediction because they comprise about one third of secondary structure in globular proteins where they serve as connecting elements for α-helices and β-strands (Chapt. 2.1.3). Often, loops serve as binding sites for ligands and cofactors (Chapt. 2.1.4) or even form the active site of a protein.

In globular proteins, most loops regions are located on the protein surface where they display a large variety of conformations. They also represent the regions where most mutations leading to insertions and deletions in the primary sequence can be found.

Therefore, loops need to be modeled in a separate step following the initial alignment procedure. Due to the inherent complexity of loop region modeling compared to the modeling of α-helical and β-sheet regions, loop prediction has evolved as a separate field.

Similar to protein structure determination methods, loop prediction methods can be divided into *ab inito* and knowledge-based methods. *Ab initio* methods attempt the prediction of loop regions by conducting a conformational search using energy functions. Algorithms include the sampling of energetically favorable torsion angles, random tweak methods, analytical methods, molecular dynamics simulations, and Monte Carlo search methods. Knowledge-based methods predict loop regions by using a fragment databank of known loop structures. Several loop databanks have been created using experimentally determined structures of the PDB (Chapt. 2.1.9). In order for the prediction method to be effective, loop databanks are clustered to reduce redundancy, and loops are classified by criteria such as length, torsion angles, and solvent accessibility. Compared to *ab initio* methods, knowledge-based methods have the advantage of predicting loop structures based on conformations which are physically reasonable since they have been extracted from native protein structures.

The first step in the protein loop prediction process involves the determination of anchor groups. This crucial step is performed in both *ab initio* and knowledge based loop modeling. Anchor groups represent the two amino acid residues which form the beginning and the end of the loop segment which has been identified by the initial template structure to target sequence alignment (Chapt. 2.2.4). Anchor groups are usually selected as being the first and last commonly aligned residue flanking the insertion or deletion, but can also be taken a few residues away from the loop region. The work in this thesis shows that choosing the correct anchor group has an effect on the quality of the entire protein model.

After the appropriate anchor groups have been selected, the loop modeling process takes place. In *ab initio* modeling, the loop is constructed stepwise one residue at a time

either starting from one anchor group residue to the other or from both anchor groups simultaneously with a closing step in the middle. Ranking of fitted loops is subsequently performed by the use of scoring functions which are primarily based on molecular mechanics force fields [78]. In knowledge-based loop prediction, loops are selected from a loop fragment databank made from loop regions of known structures from the PDB and fitted to closely match the geometry of the anchor group residues. Fitted loops may then be ranked by criteria such as 'goodness of fit' between loop and anchor region geometry, sequence similarity between database fragment and modeled loop section, or distant-dependent statistical potentials.

The accuracy of the loop prediction can be expressed by using either local or global root mean square deviation (RMSD). Local RMSD calculates the deviation between the modeled versus the native loop, while global RMSD takes into account the deviation of the entire protein structures. Global RMSD is the preferred measure as it is stricter and takes into account the orientation of the loop within the rest of the protein topology.

Presently, loop modeling still represents a major source of error in protein structure prediction. *Ab initio* methods suffer from the large number of possible conformations so that only smaller loops can be modeled effectively at reasonable computation times [8]. The prediction quality for knowledge-based methods, on the other hand, is dependent on the completeness its fragment databank. For larger loops the number of possible sequences and conformations increases exponentially, which leads to an incompleteness of the loop databank with increasing fragment length. Therefore, knowledge-based and *ab initio* methods both have problems in modeling larger loops, even though the reasoning differs for both methods. Nevertheless, due to the recent increase in experimentally solved structures (Chapt. 2.1.9), the coverage for loop databases has been improving over the years. Currently, databanks have been established with coverage of greater than 95% for loop fragments up to 10 residues [16].

Loop prediction methods are generally tested in 'self-prediction' experiments where loops of experimentally solved structures are removed, remodeled, and the prediction quality evaluated by RMS deviation between model and original native structure. These experiments have led to good results with regard to the modeling of loops with identical length in template and target structure [12] . However, loop prediction for regions with insertions and deletions has been insufficiently inaccurate. In this thesis, a real modeling situation is simulated by aligning pairs of homologous native structures where unaligned regions represent the loop regions to be modeled. By interchanging template and target structures, the loops can be used as both insertions and deletions. This setup allows the evaluation of the accuracy of modeled loops and will be used to derive a set of rules for the positioning of anchor groups in order to improve the quality of knowledge-based protein loop prediction.

2.3 Chemical Equilibrium

2.3.1 Introduction

Biochemical reactions in living cells are catalyzed by enzymes which have the ability to accelerate reactions by lowering their activation energy (Chapt. 2.1.10). However, enzymes cannot change the feasibility or alter the direction of a reaction. The direction in which a reaction proceeds is determined in most part by its equilibrium constant and the relative concentrations of substrates and products. Generally, both chemical and biochemical reactions follow the principle of Le Châtelier, which states that any deviation from equilibrium stimulates a process that tends to restore the system to equilibrium. Thus, when the reactants in a biochemical reaction are in excess of their equilibrium concentration, the net reaction will proceed in the forward direction. Analogously, when products are in excess, the net reaction will proceed in the reverse direction.

2.3.2 Equilibrium Constant (K)

According to the second law of thermodynamics, all processes, including biochemical reactions, spontaneously strive to achieve a state of maximum entropy. In a thermodynamically closed system where no products or substrates can escape, and where temperature, pH and ionic strength of the solution are kept constant, the entropy of the system depends entirely on the concentrations of the reactants and products of the reaction. A reaction which is left to occur spontaneously under such conditions will gravitate towards a steady-state concentration. The stage of the reaction at which no more net changes in either reactant or product concentration takes place is called the reaction equilibrium. At this stage, the reaction has also reached its state of maximum entropy. The exact position of the equilibrium is given by the equilibrium constant which is calculated by dividing the concentrations of the products by the concentrations of the reactants. For a sample biochemical reaction a A + b B = c C + d D, where a, b, c, and d represent the molar coefficients of species A to D, the equilibrium constant K is defined as:

$$K = \frac{(a_C)^c (a_D)^d}{(a_A)^a (a_B)^b} \qquad \text{where} \qquad a_i = \gamma_i c_i$$

Here a_i is the activity of reactant i, which equals to the product of the molar concentration c_i times the activity coefficient γ_i . Activity coefficients are functions of ionic strength, and they are close to unity for neutral molecules in aqueous solutions. Therefore, equilibrium constants for neutral metabolites can be written as K_C and approximated by using molar concentrations instead of activities:

$$K_c = \prod_{i=1}^{N} (c_i)_{eq}^{vi}$$

The value of the equilibrium constant tells us the spontaneous direction of a biochemical reaction assuming a starting concentration of 1M for all metabolites in the reaction. The reaction occurs in the forward direction if the equilibrium constant is larger than 1 and in the backward direction if it is less than 1.

Equilibrium constants can also be expressed using the standard Gibbs free energy of reaction ΔG^0_r. For a chemical reaction a A + b B = c C + d D, the Gibbs free energy change of the reaction is given by:

$$\Delta G_r = \Delta G_r^0 + RT \ln \frac{[C]^c [D]^d}{[A]^a [B]^b}$$

Here, ΔG_r is the Gibbs free energy change of reaction with respect to the concentrations of reactants a, b, c, and d. The factor R represents the gas constant (8.314472 J K^{-1} mol^{-1}), while T is the absolute temperature (298.15 K), and ΔG^0_r stands for the standard Gibbs free energy of reaction, which practically represents the Gibbs free energy change of the reaction at standard conditions and reactant and product concentrations of 1M. At chemical equilibrium, no more changes in concentrations take place, so that ΔG_r equals 0, and the above equation reduces to the following expression:

$$\Delta G_r^0 = -RT \ln K_c$$

This equation relates the equilibrium constant K_c to the standard Gibbs free energy of reaction ΔG^0_r.

Thus, reactions can be characterized either by their equilibrium constant K_c or by their standard Gibbs free energy of reaction (ΔG^0_r):

- For $K >> 1$ or $\Delta G_r^0 << 0$: The reaction is spontaneous and irreversible in the forward direction
- For $K \sim 1$ or $\Delta G_r^0 \sim 0$: The reaction may occur in both directions
- For $K \sim 0$ or $\Delta G_r^0 > 0$: The reaction is spontaneous in the backward direction

The advantage of this relationship lies in the fact that equilibrium constants may not just be determined experimentally by measuring individual concentrations but can also be calculated by subtracting the standard Gibbs free energies of formation (ΔG_f^0) of the reactants from the products:

$$\Delta G_r^0 = \sum_j \nu_j \Delta G_f^0(j)_{prod} - \sum_j \nu_j \Delta G_f^0(j)_{react}$$

Standard Gibbs free energies of formation (ΔG_f^0) of each reactant can be obtained by using listed values or calculated by applying group contribution methods (Chapt. 2.3.6).

This thesis shows that equilibrium constants can also be determined by calculating the standard Gibbs free energy of reaction (ΔG_r^0) using quantum mechanical methods.

2.3.3 Temperature Dependence of Equilibrium Constants

The value of equilibrium constants is temperature dependent. This can be shown by combining the equation for standard Gibbs free energy $\Delta G° = \Delta H° - T\Delta S°$ with the relationship $\Delta G° = - RT \ln K$ as follows:

$$\boxed{\ln K = \frac{-\Delta H°}{R}\left(\frac{1}{T}\right) + \frac{\Delta S°}{R}}$$

Here, $\Delta H°$ and $\Delta S°$ represent the enthalpy and entropy of the reaction at standard state. This relationship permits the determination of the values $\Delta H°$ and $\Delta S°$ from measurements of K at two or more different temperatures. This is achieved by a plot of ln K versus 1/T, known as a van't Hoff plot which yields a straight line of slope $-\Delta H°/R$ and y-intercept $\Delta S°/R$. This estimation approach assumes that $\Delta H°$ and $\Delta S°$ are independent from temperature, which is true to a reasonable extent.

2.3.4 Apparent Equilibrium Constant (K')

Equilibrium constants for biochemical reactions are not only dependent on the concentrations of the reactants but also on the properties of the aqueous environment in which they are measured. The value of K depends largely on solvent properties such as ionic strength (I), pH, temperature (T), and the concentration of specific ions such as magnesium (pMg). An equilibrium constant taking all these parameters into account is called the apparent equilibrium constant (K'). This constant is adjusted towards a specified T, pH, pMg, and I.

Similar to the relationship between equilibrium constant (K) and standard Gibbs free energy of reaction (ΔG_r^0) (Chapt. 2.3.2), the apparent equilibrium constant (K') is related to the standard transformed Gibbs free energy of reaction ($\Delta G_r^{0'}$) in the following manner:

$$\boxed{\Delta G_r^{o'} = -RT \ln K'}$$

Here, the standard transformed Gibbs free energy of reaction ($\Delta G_r^{0'}$) is defined at standard conditions for biochemical reactions in aqueous solution with 298.15K, pH 7, and I=0 (Chapt. 2.3.5). It is important to note that K' = K for reactions at pH 7, where reactants are neutral non-electrolytes such as glucose and fructose, or where the acid dissociation constants (pK_a) of products and reactants are equal in value.

2.3.5 Standard State Convention in Biochemical Reactions

The definition of standard state in physical chemistry defines a reactant at unit activity (~ 1 M) at 298.15 K and 1 atm. For biochemical reactions, where most reactions occur in dilute aqueous solutions near pH 7, the standard state convention for biological systems includes some additional conditions:

- The activity of water is taken to be unity (1 M) despite the fact that its concentration is 55.5 M. That means that for reactions in dilute aqueous solution, the concentration term for water [H_2O] can be ignored in the equilibrium constant, as it is considered as incorporated.
- The hydrogen ion activity is defined as unity at pH 7 rather than at pH 0.

- The standard state of a substance that can undergo acid-base reactions is defined in terms of its naturally occurring ion mixture at pH 7. For example, ATP as a reactant in a biochemical equation at pH 7 represents a mixture of mostly ATP^{4-} and $HATP^{3-}$.
- Standard transformed Gibbs free energies ΔG°' are valid only at pH 7.

The above conventions have an effect on the calculation of equilibrium constants and standard Gibbs free energies of reaction:

- For reactions where all reactants are neutral solutes and that do not contain water in the chemical equation, the standard transformed Gibbs free energy of reaction ($\Delta G_r^{0'}$) equals to the standard Gibbs free of reaction (ΔG_r^0). This means that for these reactions, neither the dilute aqueous environment nor the pH have an effect on the Gibbs free energy of reaction or the equilibrium constant.
- For reactions that involve water in the chemical equation, we can calculate $\Delta G_r^{0'}$ from ΔG_r^0 if we adjust the actual concentration of water in aqueous solution (55.5 M) to the standard state convention of unity (1M).

  Thus, for the following reaction with non-ionizing solutes

  a A + b B = c C + d D + **n H_2O** at 298.15 K, pH 7 and [H_2O] = 1 M we get

$$\Delta G_r^{o'} = -RT \ln K' = -RT \ln\left(\frac{[C]^c[D]^d[1]^n}{[A]^a[B]^b}\right) = -RT \ln\left(\frac{[C]^c[D]^d}{[A]^a[B]^b}\right)$$

  If we compare this to the standard Gibbs free energy of reaction (ΔG_r^0), where [H_2O] = 55.5 M, we get

$$\Delta G_r^o = -RT \ln K = -RT \ln\left(\frac{[C]^c[D]^d[H_2O]^n}{[A]^a[B]^b}\right)$$

So that a combination of both equation yields the relationship:

$$\boxed{\Delta G_r^{o\prime} = \Delta G_r^o + nRT \ln[H_2O] = \Delta G_r^o + nRT \ln[55.5M]}$$

or

$$\boxed{\Delta G_r^{o\prime} = \Delta G_r^o + n(9.96kJ/mol)}$$

Thus, if we wanted to calculate the standard transformed Gibbs free energy of reaction $\Delta G_r^{0\prime}$ at pH 7 and 298.15 K for a biochemical reaction which involves water as a reactant, we must add or subtract 9.96 kJ/mol to the standard Gibbs free energy of reaction (ΔG_r^0) for each mol of water appearing as a reactant or product in the chemical equation. The NIST database of enzyme reactions [22] contains experimental equilibrium constants of reactions involving water as a reactant, which have been adjusted in the manner described.

- Calculations similar to the above apply for the occurrence of hydrogen ions in reactions that involve dissociable species such as acids or bases.

2.3.6 Group Contribution Method

For most compounds and biotransformations, standard Gibbs free energies are not readily available without experimental effort. However, many thermodynamic properties can be estimated based on the structure of a particular compound. A property is estimated by decomposing the compound into functional groups and by using a table where each group has an assigned partial value of the desired property. The total property value for that particular compound is then calculated by taking the sum of all contributing groups in addition to an 'origin' value which is constant for all compounds. In some cases, additional

characteristics such as aromaticity or interactions between certain chemical groups must be considered by the use of special corrections.

A group contribution method for the estimation of standard transformed Gibbs free energies ($\Delta G°'$) and equilibrium constants (K') of biochemical reactions at pH 7 and 298.15 K and has been presented by Mavrovouniotis [68][69]. This method was developed using a table of contributions which was derived from several sources covering a large number of biochemical compounds [28][62]. The contributions were estimated by multiple linear regression using compounds and groups in aqueous solution at the standard state of pH 7 and temperature 298.15 K. Theoretically, the method can be applied to any organic compound in aqueous solution. However, the method may be less accurate for non-biochemical compounds since the data used to create the contribution table was heavily biased in favor of biochemical compounds and reactions. In addition to a standard table built from molecular fragments, the author added a table which allows the consideration of special molecular properties such as aromaticity and multiple ring structures [69]. For example, the standard transformed Gibbs free energy of formation ($\Delta G_f^{0'}$) for glutamate in dilute aqueous solution can be estimated after decomposing the molecule into functional groups (Figure 2.9) followed by the summation of the energy contributions of each group (Table 2.5).

Using this method, not only standard transformed Gibbs free energies of formation ($\Delta G_f^{0'}$) for individual compounds, but also standard transformed Gibbs free energies of reaction ($\Delta G_r^{0'}$) for entire reactions may be computed, allowing the prediction of equilibrium constants (K') for a certain biochemical reaction. For calculations involving reactions, it is sufficient to only consider the differences in contributions between the reactant and product sides, since the net number of most group occurrences in a reaction equal to zero. For reactions involving special pairs of compounds such as NAD(P)/NAD(P)H or oxidized/reduced coenzyme A, the special contributions associated with these transformations may be applied.

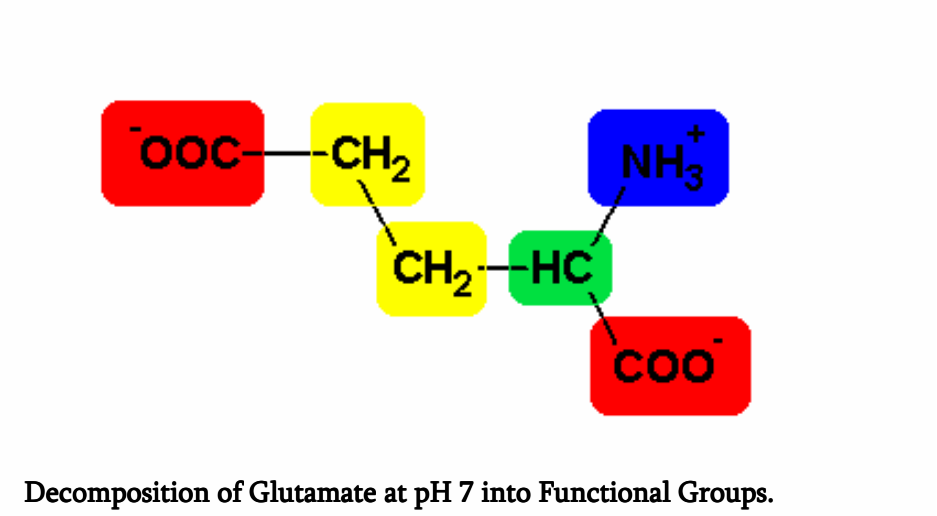


Figure 2.9: **Decomposition of Glutamate at pH 7 into Functional Groups.**

Table 2.5: **Calculation of ΔG_r°' of Glutamate using Group Contributions.**

Group or correction	No. of occurrences	Contribution (kcal/mol)	Total contribution
Origin	1	- 24.7	- 24.7
- NH_3^+	1	4.8	4.8
- COO^-	2	- 71.4	- 142.8
- CH_2 -	2	1.7	3.4
- CH <	1	- 5.4	- 5.4
Total			- 164.7

The group contribution method by Mavrovouniotis has shown to perform quite well for estimating of standard transformed Gibbs free energies of formation in aqueous solution (ΔG_f°') of individual molecules. For most biochemical compounds, the standard transformed Gibbs free energy of formation was estimated within 1-2 kcal/mol from their reported literature value, even though deviations of 5 kcal/mol or higher were also present [68].

Computer programs have been designed to implement the group contribution method for predicting standard transformed Gibbs free energies of formation ($\Delta G_f^{0'}$) for individual molecules, and an improved version of that software for Linux platforms has recently been developed by our research group. The program *Gibbspredictor* written by Kai Hartmann [41] uses the structural information of chemical compounds provided in MOL files. After

correcting the protonation state of the molecule according to pH 7 conditions, the molecule is split into group fragments and the contributions are added and printed out as a single energy value.

The group contribution method by Mavrovouniotis lacks accuracy for sufficiently evaluating exact equilibrium constants but has shown to be useful in predicting the feasibility and reversibility of biochemical reactions and pathways. In this thesis, attempts will be made to predict standard transformed Gibbs free energies of reaction ($\Delta G_r'^\circ$) by using *ab initio* quantum mechanical methods.

2.4 *Ab Initio* Computational Quantum Mechanics

2.4.1 Introduction

The hydrogen atom is the only system for which an exact solution of the Schrödinger equation exists. For all other non-ideal systems, one can only find approximate solutions. *Ab initio* quantum mechanical calculations are characterized by the fact that only natural constants such as electronic charge e, electron mass m_e, Planck's constant $\hbar$ and the exact masses of the atoms are used as initial data sources. *Ab initio* does not mean that the calculation is an exact method but that the calculations are performed by pure quantum mechanics and are based on close-to exact Hamilton operators. Semi-empirical methods, on the other hand, use approximations with regard to starting constants and Hamilton operators, so that the calculation of certain integrals is omitted or certain values are replaced by empirical data. Often, calculations are reduced to a subset of electrons.

Quantum mechanical calculations for larger molecules can require an enormous amount of computational effort. Therefore, several approximations have been required to simplify the calculation process:

- **Born-Oppenheimer approximation**: Separation of the electronic from the nuclear motion, due to the large difference in mass between nucleus and electron
- **Limiting the number of basis functions**: An exact solution requires an endless set of basis functions. Calculations are simplified by restricting the number of basis functions to a limited set.

2.4.2 Schrödinger Equation and Born-Oppenheimer Approximation

The time-independent non-relativistic Schrödinger equation [90] for a particle of mass m is given by

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r})\right]\psi(\mathbf{r}) = E\psi(\mathbf{r}).$$

where the term in brackets represents the Hamiltonian operator Ĥ.

For a random molecule with total wavefunction Φ and Hamiltonoperator $\hat{H}_{tot}$, the general Schrödinger equation can be written as:

$$\hat{\mathcal{H}}_{tot}\left|\Phi\right\rangle = \mathcal{E}\left|\Phi\right\rangle$$

For a molecule with N electrons and M nuclei, the Hamiltonian becomes as a sum of five terms:

$$\hat{\mathcal{H}}_{tot} = \hat{T}_{el} + \hat{T}_{nucl} + \hat{V}_{nucl-el} + \hat{V}_{el-el} + \hat{V}_{nucl-nucl}$$

$$= -\sum_{i=1}^{N}\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N-1}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A=1}^{M-1}\sum_{B>A}^{M}\frac{Z_A Z_B}{R_{AB}}$$

The first term represents the kinetic energy of the electrons, followed by the kinetic energy of the nuclei, the electron-nucleus potential energy, the electron-electron potential energy, and the nucleus-nucleus potential energy.

The Born-Oppenheimer approximation is based on the idea that the mass ratio between nucleus and electron is very large (e.g. H-Atom: m_p/m_e = 1832), so that their movements can be separated from each other. The movement of the electrons is much faster than that of the nucleus, so that the nucleus can be considered fixed, while the movement of the electrons can be characterized by a separate Schrödinger equation. As a result, the nuclear kinetic energy term can be omitted, while the nucleus-nucleus potential energy term is considered constant and can be separated from the Hamilton operator. This reduces the Hamilton operator to the following expression

$$\hat{\mathcal{H}}_{elec} = -\sum_{i=1}^{N} \frac{1}{2}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + \sum_{i=1}^{N-1}\sum_{j>i}^{N} \frac{1}{r_{ij}}$$

and the Schrödinger equation can be simplified to

$$\hat{\mathcal{H}}_{elec} \left| \psi_{elec} \right\rangle = \mathcal{E}_{elec} \left| \psi_{elec} \right\rangle$$

where

$$\psi_{elec} = \psi_{elec}\left(\{\mathbf{r}_i\}; \{\mathbf{R}_A^0\} \right)$$

The electronic wavefunction ψ_{elec} describes the motion of the electrons and is dependent only on the electron coordinates $\{r_i\}$ as variables. Here, $\{R^0_A\}$ represents the set of coordinates for the fixed nuclei which do not act as variables in the calculation, but will influence the initial conditions of the Hamilton operator by potentially altering the electronic wavefunction itself. Another feature of the Born-Oppenheimer approximation is

that it allows the total wavefunction Φ to be estimated as a product of a nuclear wavefunction (χ_{nucl}) and an electronic wavefunction in which the nuclei have fixed coordinates (R_A):

$$\left|\Phi\right\rangle_{BO}\left(\{\mathbf{r}_i\},\{\mathbf{R}_A\}\right) \cong \psi_{elec}\left(\{\mathbf{r}_i\};\{\mathbf{R}_A\}\right)\cdot\chi_{nucl}\left(\{\mathbf{R}_A\}\right)$$

This relationship is also known as the adiabatic approximation.

2.4.3 Basis Functions

According to the LCAO (Linear Combination of Atomic Orbitals) principle, MOs (molecular orbitals) can be approximated by linear combination of AOs (atomic orbitals):

$$f\left(\vec{r}\right)=\sum_{i=1}^{\infty}a_i\psi_i\left(\vec{r}\right)$$

AOs consist of wavefunctions ψ_i, and a complete set of functions would allow the exact description of any MO. Mathematically complete sets are for example the Laguerre-functions which are used in the description of the hydrogen atom. Practically, however, basis sets are always limited to a finite number of basis functions {ψ_i, i = 1, 2,…k) and this limitation represents one of the major errors in quantum mechanical MO calculations.

In MO calculations, AOs are primarily used as basis functions. The most frequently employed functions are the Slater-type orbitals (STO) [92] and the Gauss-type orbitals (GTO) [5]. STOs resemble wave functions of the hydrogen atom, where the Laguerre polynomials of the radial term have been replaced by a simpler linear function. Therefore, Slater polynomials lack nodal spheres and when employed in MO calculations, need to be used in linear combinations.

Gauss-type orbitals are easier to integrate, but have shapes which are less similar to the 'correct' AOs than the Slater-type orbitals. In an attempt to reduce this effect, STOs (Φ^{SF}) have been approximated by linear combination of GTOs, the so called primitive Gauss functions (Φ^{GF}), resulting in the generation of a new 'contracted' Gauss function (Φ^{CGF}):

$$\phi_\mu^{SF}\left(\vec{r}-\vec{R}_A\right) \approx \sum_{p=1}^{L} d_{p\mu} \cdot \phi_p^{GF}\left(\alpha_{p\mu}, \vec{r}-\vec{R}_A\right) \equiv \phi_\mu^{CGF}$$

Examples:

- **STO-2G** : Slater-type orbital reproduced by linear combination of 2 primitive GTOs
- **STO-6G**: Slater-type orbital reproduced by linear combination of 6 primitive GTOs

2.4.4 Basis Sets

Minimal basis sets

Minimal basis sets are mostly used for demonstration or test purposes. In minimal basis sets, each orbital is described by a single basis function which in most cases is a contracted Gauss function (CGF). Minimal basis sets for individual elements are:

H, He: $1s$

Li bis Ne: $1s,\ 2s, 2p_x, 2p_y, 2p_z$

Na bis Ar: $1s,\ 2s, 2p_x, 2p_y, 2p_z,\quad 3s, 3p_x, 3p_y, 3p_z$

K, Ca: $1s,\ 2s, 2p_x, 2p_y, 2p_z,\quad 3s, 3p_x, 3p_y, 3p_z,\quad 4s, 4p_x, 4p_y, 4p_z$

Sc bis Kr: $1s,\ 2s, 2p_x, 2p_y, 2p_z,\quad 3s, 3p_x, 3p_y, 3p_z,\ 3d_{z^2}, 3d_{x^2-y^2}, 3d_{xy}, 3d_{xz}, 3d_{yz},$

Note: For the elements Li and Be, experience has shown that it is necessary to include the unoccupied 2p functions into the MO calculations.

Double Zeta (DZ)

Each basis function in the minimal basis set is replaced by two functions, doubling the number of variable parameters:

$$\psi_k(\vec{r}) = c_{k1}\phi_1^{CGF} + c_{k2}\phi_2^{CGF} = c_{k1}\left(\sum_i^L d_{i,k1}\phi_i^{GF}\right) + c_{k2}\left(\sum_j^L d_{j,k2}\phi_j^{GF}\right)$$

Triple Zeta (TZ)

Each basis function in the minimal basis set is replaced by three functions, tripling the number of variable parameters.

$$\begin{aligned}\psi_k(\vec{r}) &= c_{k1}\phi_1^{CGF} + c_{k2}\phi_2^{CGF} + c_{k3}\phi_3^{CGF} \\ &= c_{k1}\left(\sum_j^L d_{j,k1}\phi_j^{GF}\right) + c_{k2}\left(\sum_j^L d_{j,k2}\phi_j^{GF}\right) + c_{k3}\left(\sum_j^L d_{j,k3}\phi_j^{GF}\right)\end{aligned}$$

Split-valence basis sets

In split-valence basis sets, the inner-core orbitals are described by a single basis function (minimal basis set), while the valence orbitals are described by two (DZ) or three functions (TZ). Table 2.6 provides examples for the notation of split-valence basis sets:

Table 2.6: Nomenclature for Split-Valence Basis Sets by Pople.

3-21G	Core orbitals	one contracted GTO made of 3 primitives
	Valence orbitals (DZ)	first contraction made of 2 primitives
		second contraction made of 1 primitive
6-31G	Core orbitals	one contracted GTO made of 6 primitives
	Valence orbitals (DZ)	first contraction made of 3 primitives
		second contraction made of 1 primitive
6-311G	Core orbitals	one contracted GTO made of 6 primitives
	Valence orbitals (TZ)	first contraction made of 3 primitives
		second contraction made of 1 primitive
		third contraction made of 1 primitive

First digit refers to core, following digits refer to valence orbitals. G stands for Gauss-type orbital (GTO)[59].

Polarization Functions

Polarization functions are essential in describing polar molecules, aromatic ring systems, and intermolecular interactions (H-bonds). The introduction of polarization functions to a particular basis set allows a shift of electron density by the otherwise centralized basis set orbitals. Polarization functions were introduced by the Pople group [59] and have an angular quantum number (l, where l = 0,…, n-1) greater than the functions which they polarize. Figure 2.10 illustrates the manner in which polarization functions can cause a shift in the electron density of the lower orbitals. Examples for basis sets extended by polarization functions are provided as follows:

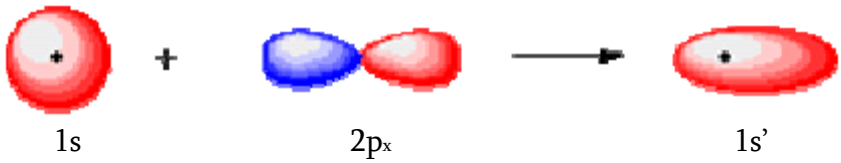
- **6-31G(d) or 6-31G*** : 6-31G basis set extended by 6 d-type functions on heavy atoms
- **6-31G(d,p) or 6-31G**** : 6-31G basis set extended by 6 d-type functions on heavy atoms and 3 p-type functions on H or He atoms

Diffuse Functions

Diffuse functions are introduced when the electron density is distributed across the entire molecule, as for example in excited states and anions. These spread out basis functions may be modeled by GTOs with small exponents. These additional basis functions are called diffuse functions. They are added as single uncontracted GTOs. As an example, diffuse functions can be added to the 6-31G basis set as follows:

- **6-31+G :** adds a set of diffuse s and p-orbitals to the atoms in the first and second rows (Li- Cl).
- **6-31++G :** adds a set of diffuse s and p-orbitals to the atoms in the first and second rows (Li- Cl) *and* a set of diffuse s-functions to hydrogen.

Polarization of a 1s-orbital by a $2p_x$ polarizing orbital:



Polarization of a 2p-orbital by a 3d polarizing orbital:

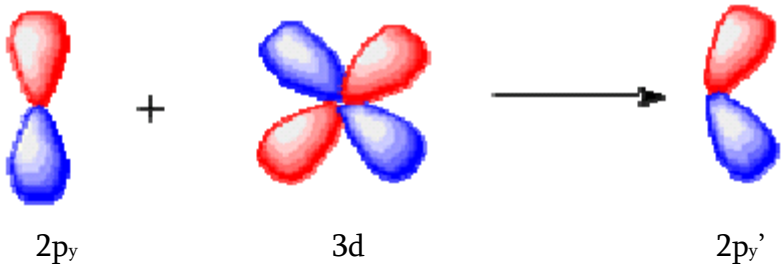


Figure 2.10: **Effect of Polarization Functions on Neighboring Orbitals.** Polarization Functions shift electron densities in neighboring orbitals. As polarization functions of higher angular quantum number are added, electron densities of basis set orbitals shift away from center.

.

2.4.5 Quantum Mechanical Calculations

Single Point Calculations

This procedure simply calculates the energy, wave function and other requested properties at a single fixed geometry. Single point calculations are usually done at the beginning of a study on a new molecule to gain an insight into the nature of the wave function. They are also frequently carried out after a geometry optimization. Compared to the geometry optimization, they are performed using a larger basis set and a more superior method. Thus for a very large system, the geometry may be optimized at the HF level

(Chapt. 2.4.7) with the 3-21G basis set (Chapt. 2.4.4), but energy differences between isomers are then explored with the MP method (Chapt. 2.4.7) and the 6-31G (d,p) basis set .

Geometry Optimization Calculations

Experience has shown that it is essential to find the geometry of a molecule accurately by geometry optimization. The procedure calculates the wave function and energy at a starting geometry and then proceeds to a new geometry which gives a lower energy. This process is then repeated until a local minimum in the vicinity of the starting geometry has been found. Ideally, this procedure calculates the forces on the atoms by evaluating the gradient (first derivative) of the energy with respect to the atomic coordinates. In some cases, gradients may be estimated or sophisticated algorithms may be used for selecting new geometries, resulting in a more rapid convergence towards the local minimum. It is important to recognize that this procedure will not necessarily find the geometry of lowest energy, i.e. the global minimum.

Finding all local minima, and therefore the global minimum, for a particular set of atoms is a complex task. The optimization procedure sometimes ends in a saddle point which typically represents a transition structure. This will occur particularly if the symmetry and degrees of freedom are purposely restricted. For example, the optimized geometry for a restricted planar NH_3 molecule actually represents the transition structure for the "umbrella-like" flipping of the molecule from one pyramidal structure to the other. It is always a good idea to begin a geometry optimization with a small basis set and a poor method before proceeding to the more sophisticated basis set and method of choice for a particular problem. A geometry optimization can be started from geometries which were generated by a poorer approach.

Frequency Calculations

Frequency calculations allow the prediction of I.R. and Raman frequencies and their intensities through force constants while assuming the model of a harmonic oscillator. Vibrational frequencies are obtained by determining the second derivatives of the energy with respect to the Cartesian nuclear coordinates and then transforming them to mass-weighted coordinates. This transformation is only valid at a stationary point. Frequencies calculated at an optimized local or global minimum have all real and positive values. Frequencies at a stationary point other than a minimum (e.g. saddle point) have one or more complex values. These transition structures have 'imaginary frequencies' which are printed out as negative numbers.

By applying statistical thermodynamics, frequency calculations also allow the computation of thermodynamic quantities such as zero-point energy, entropy, heat capacity, and Gibbs free energy at a particular temperature. The calculations are performed by evaluating the translational, rotational, and vibrational partition functions at a specific temperature using standard expressions for an ideal gas.

2.4.6 Solvation Models

Polarizable Continuum Model (PCM)

The Polarizable Continuum Model (PCM) by Tomasi and coworkers [71] is one of the most frequently used continuum solvation methods and has appeared in numerous variations over the years. The PCM model calculates the molecular free energy in solution as the sum over three terms ($G_{sol} = G_{es} + G_{dr} + G_{cav}$). These terms represent the electrostatic (G_{es}) and the dispersion-repulsion (G_{dr}) contributions to the free energy as well as the cavitation energy

(G_{cav}). All three energy terms are calculated using a cavity created by interlocking van-der-Waals spheres which are centered at the atomic positions (Figure 2.11).

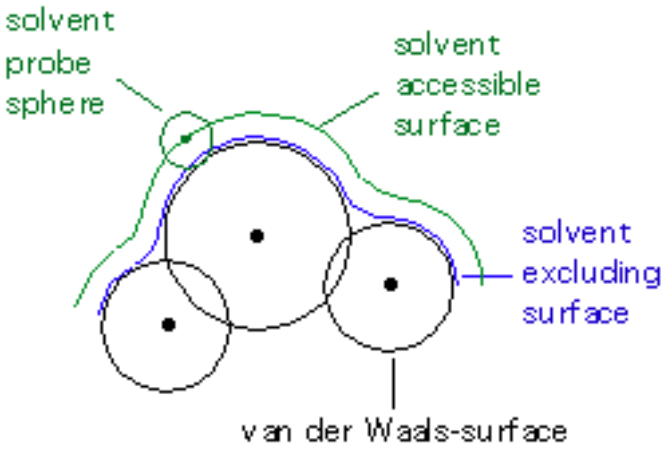


Figure 2.11: **Solvent Accessible Surface.** Figure displays solvent accessible surface for three neighboring atoms with overlapping van-der-Waals surfaces. Surfaces are generated by using a solvent probe sphere mostly represented by a single water molecule.

While calculation of the cavitation energy G_{cav} uses the surface defined by the van-der-Waals spheres, the solvent accessible surface is used to calculate the dispersion-repulsion contribution (G_{dr}) to the solvation free energy. The latter differs from the former through additional consideration of the (idealized) solvent radius. The electrostatic contribution to the free energy in solution (G_{es}) uses an approximate version of the solvent excluding surface constructed through scaling all radii by a constant factor (e.g. 1.2 for water) and then adding some more spheres not centered on atoms in order to arrive at a smoother surface. Localization and calculation of the surface charges is approached by systematic division of the spherical surface into small regions of known area, followed by calculations involving a one-point charge per surface element.

COSMO (Conductor-like Screening Model)

The **Co**nductor-like **S**creening **Mo**del (COSMO) is a continuum approach developed by Klamt and Schürmann [57] which, while more complicated, is computationally quite efficient. The expression for the total screening energy is simple enough to allow the first derivatives of the energy with respect to atomic coordinates to be easily evaluated. The COSMO procedure generates a conducting polygonal surface around the system (ion or molecule), at the van-der-Waals distance. By introducing an ε-dependent correction factor into the expressions for the screening energy and its gradient, the theory can be extended to various dielectric constants while maintaining a small error:

$$f(\varepsilon) = \frac{(\varepsilon - 1)}{(\varepsilon + \frac{1}{2})},$$

2.4.7 Quantum Mechanical Methods

As of today, a wide range of *ab initio* quantum mechanical calculation methods have been developed. However, the vast majority of calculations are carried out using only a particular sub class of methods also known as molecular orbital methods. The earliest and most widely used molecular orbital method is the Hartree-Fock method.

Hartree-Fock Method (HF)

Hartree-Fock calculations belong to the oldest *ab initio* ansatz. They are based on a few principles and do not employ any experimental data. The method works in principle by picking one electron and approximating the interaction between this single and all other electrons by a mean value that is determined from their probability densities. This approach ignores the correlated movement of the electrons caused by their repulsion through equal

electric charges. Despite this deficiency known as the "electron correlation error", the Hartree Fock method provides accurate results in many cases and is not limited to a particular class of chemical compounds. However, Hartree Fock calculations require considerable computer time.

Semi-Empirical Methods

Semi-empirical calculations can be categorized somewhere in between *ab initio* and molecular mechanics methods. Semi-empirical methods determine molecular orbitals within the LCAO model and are based on the variation principle in which most integrals along these calculations are estimated from experimental values. Thus, for chemical compounds that lie outside the classes for which these estimations are parametrized, results may be less accurate. In contrast, however, semi-empirical calculations can be parametrized in detail for specific cases such as spectroscopic properties. Most semi-empirial programs make use of the zero differential overlap (ZDO) approximation, which defines the overlap between different basis functions as zero. The various ZDO models can be grouped according to their approximations for one- and two-electron integrals:

- **CNDO**: complete neglect of differential overlap
- **INDO**: intermediate neglect of differential overlap model
- **NDDO**: neglect of diatomic differential overlap model
- **MINDO/3**: modified INDO
- **MNDO** : modified neglect of diatomic overlap
- **AM1**: Austin Model 1 , analogue of MNDO
- **PM3 [97]**: third parametrization of MNDO, AM1 , analogue of MNDO

Configuration Interaction (CI)

Together with the Coupled Cluster (CC) and Møller-Plesset perturbation theory (MP) methods, configuration interaction (CI) belongs to the class of post-Hartree Fock methods. Calculations of this type target at a determination of the electron correlation based on the variation principle and MO-ansatz. They need a lot of computer time and storage and are therefore mostly applied to small molecules. Furthermore, they need a discrete choice of orbitals to be included, which, in most cases, has to be manually performed.

Møller-Plesset-Calculations (MP)

These methods target at determining the electron correlation based on perturbation theory. MP calculations are limited by the highest degree used for the perturbation expansion and are characterized by the fact that the variation theorem is not valid for a finite highest degree. Within a specific expansion degree, they do not require additional choices, which is one of the reasons for their popularity.

Density Functional Theory (DFT)

Density Functional Theory (DFT) methods were developed by Kohn and Sham [58] and are often considered to be *ab initio* methods for determining the molecular electronic structure, even though many of the most common functionals use parameters derived from empirical data or from more complex calculations. In DFT, the total energy is expressed in terms of the total electron density rather than the wave function. In this type of calculation, there is an approximate Hamiltonian and an approximate expression for the total electron density so that DFT methods can be very accurate for little computational cost. The drawback is, that unlike pure *ab initio* methods, there is no systematic way to improve the method by extending the form of the functional basis, because the electronic energy of the

ground state of a system is entirely described by the electron density. The energy itself is expressed as a functional which is practically a function of a function of the electron density. Analogous to the wave function approach, the functional can be split into three terms:

$$E[\rho] = T[\rho] + E_{ne}[\rho] + E_{ee}[\rho] \; .$$

- $T[\rho]$: functional of the kinetic energy,
- $E_{ne}[\rho]$: functional of the nucleus-electron-interaction,
- $E_{ee}[\rho]$: functional of the electron-electron interaction.

  ($E_{ee}[\rho]$ can be split in a Coulomb part $\mathbf{J}[\rho]$ and a exchange part $\mathbf{K}[\rho]$)

Computation of $T[\rho]$ and $\mathbf{K}[\rho]$ can be carried out with the assumption of a homogeneous electron gas with non-interacting particles. Kohn and Sham opened up DFT for use in computational chemistry by the introduction of orbitals. $T[\rho]$ is split in an exactly computable term $T_s[\rho]$ and a small correction term. The calculation of $T_s[\rho]$ is carried out under the assumption of non-interacting particles, i.e. the orbital occupancy is expected to be 0 or 1, resulting in an error because partially occupied orbitals are not described in this ansatz.

Frequently Used Potentials

The Local Density Methods (LDA): The density is regarded as a local and homogeneous electron gas. In open shell systems it is called LSDA (Local Spin Density Approximation).

Gradient Corrected Methods (GGA): These methods assume an inhomogeneous electron gas. Therefore, E_{xc} is not only dependent on the density but also on the derivatives of the density (non-local methods).

Perdew-Wang (PW86): $E_x^{PW86} = E_x^{LDA} (1 + ax^2 + bx^4 + cx^6)^{1/15}$
with $x = [(|\rho|)/(\rho^{4/3})]$; a, b, c as constants.

Becke (B88): This functional corrects the LSDA exchange energy, which describes the correct asymptotic behaviour of the energy density E_x, but not of the exchange potential.

Perdew-Wang (PW91): This functional is similar to Becke's, but it uses the gradient of the orbitals instead of the density gradient.

Lee-Young-Paar (LYP): This is an independent functional, not merely a correction to LDA. The parameters are obtained by fitting to data of the He atom.

Hybrid Functionals

Hybrid Functionals represent a combination of multiple of the above functionals. **B3LYP**, a combination of the Becke and the Lee-Young-Paar functional is currently the most famous and popular among hybrid functionals.

3 MATERIALS AND METHODS

3.1 Anchor Group Positioning in Knowledge-Based Loop Prediction

3.1.1 Fragment Data Bank

The fragment data bank was based on all X-ray structures in the 2/98 release of the Brookhaven Protein Data Bank (Chapt. 2.1.9) which had resolutions of smaller or equal to 2.0 Å and sequence identities of less than 95% determined by the Smith-Waterman algorithm [93] using standard gap penalties. After fitting N-, C-alpha, and C-carbonyl atoms of two ending residues, fragments were eliminated showing an RMS deviation below 0.25 Å considering all backbone atoms (Table 3.1). The RMS fit was performed following the procedure by Diamond [13]. The limit of 0.25 Å was chosen according to the estimated standard error in X-ray analysis.

As a geometric pre-filter for comparisons, the distance between anchoring group atoms was determined for each fragment. This distance was defined by the distance between the middle of the C-alpha to C-carbonyl atom bond of the N-terminal anchoring residue and the middle of the N- to C-alpha atom bond of the C-terminal anchoring residue. The

fragments were considered structurally distinct, when the difference in anchoring group distance between two corresponding fragments exceeded 0.5 Å.

Table 3.1: Fragment Databank Based on all Structures from PDB 2/98.

Fragment Length	Number of fragments in PDB	Number of fragments in Databank
3	184,157	13,285
4	183,031	53,853
5	181,929	98,919
6	180,835	122,077
7	179,750	133,082
8	178,671	141,165
9	177,596	148,336
10	176,527	153,982
11	175,461	158,225
12	174,403	161,501

Fragments sorted by length before and after clustering using 0.25 Å RMSD cutoff using all backbone atoms

3.1.2 Test Data Set of Aligned Protein Pairs

All examples for insertions and deletions were derived from structurally aligned protein pairs. First, proteins from the Brookhaven Protein Data Bank release 2/98 with less than 50% sequence identity were chosen using the algorithm by Smith and Waterman with standard gap penalties. Then, selected proteins were compared with each other according to structural similarity using the method of Lessel and Schomburg [64]. All proteins with at least 35 matching C-alpha atoms and at least 40% structural similarity were grouped into the same family, resulting in 132 classes where each contained more than one member. From each of these families, the protein pair with the highest structural similarity within the class was selected. Some of these families were further subdivided, since some groups of protein

pairs showed higher similarity to each other than to other members or subclasses within the family. In those cases, representative protein pairs for each subclass were chosen, resulting in a total of 170 protein pairs. The selection was performed in order to avoid biases within examples, e.g. to prevent the occurrence of the same globin surface loop in several variations.

Then, sequence alignments corresponding to the global structural fit were created for these 170 protein pairs using the method of Lessel and Schomburg [64]. These sequence alignments were systematically searched for appropriate insertions and deletions under the following condition: Blocks of at least three structurally aligned residues in a row had to be located at both ends of an insertion or deletion. Structurally aligned was equivalent to an RMSD for C-alpha atoms below 1.8 Å. The number of residues between these blocks was greater for insertions than for deletions, while they were equal for loops with zero-length difference. A maximum of ten separating residues was allowed, since the length of the fragments in the loop data bank was limited to twelve residues (two anchoring plus up to ten separating residues). By exchanging template and target protein, each example was used as an insertion as well as a deletion. This procedure resulted in 544 deletions and 550 insertions (Table 3.2). Additionally, 45 examples with zero-length difference but folding differences in loops with lengths between one and eight residues were chosen (i.e. differences in flexible loops). These examples were included for comparison purposes.

3.1.3 Anchor Group Positioning

In order to test the effect of anchor group positioning on loop prediction, all possible anchor group combinations for each example in the test set using loop fragments ranging from 3 to 12 residues were produced. For a one-residue insertion, for example, a total of 55 different anchor group combinations were generated (i.e. 1 position for the 3-residue

fragment, 2 possible positions for a 4-residue fragment,…,10 possible positions for a 12-residue loop). Their positions range from 10 residues before to 10 residues behind the gap. For a 10-residue insertion, only one combination of anchor groups exists because of the fragment limit of 12. For the 550 insertions and 544 deletions of varying loop lengths, we generated 22,771 and 22,186 anchor group combinations using the above permutation, respectively (Table 3.2) The 45 zero-length examples resulted in 3,902 anchor group pairs with the 'closest allowed' distance of the anchoring residues ranging between 1 to 8 residues. Gaps situated close to either termini of the protein resulted in fewer anchor groups than mathematically possible.

Table 3.2: Test Data Set of Loops with all Possible Anchor Group Positions.

Gap Length	Insertions (orig.)	Insertions (permut.)	Deletions (orig.)	Deletions (permut.)
0	45	3,902	45	3,902
1	273	14,871	274	14,588
2	107	4,200	105	4,044
3	48	1,665	48	1,621
4	37	978	37	946
5	36	669	36	641
6	18	211	18	198
7	12	111	12	100
8	7	42	7	35
9	7	19	7	13
10	5	5	--	--
Total	595	26,673	589	26,088

3.1.4 Loop Modeling and Ranking

After template and target proteins were fitted globally using the 3D-alignment procedure of Lessel and Schomburg [64], each appropriate loop from the fragment databank

was inserted into the templates and compared with the target structures for each problem in the test data set. For the selection of appropriate fragments from the loop data banks, a pure geometric criterion based on the anchoring groups was used, since greater differences in the ending groups caused distorted backbone folds. As a first step, distances between anchoring groups in the template protein were determined as described for the geometric pre-filter in data bank creation. Then, all fragments of the data bank with intramolecular distances 'similar' to the template were fitted onto the anchoring groups of the template protein using the RMS fit procedure of Diamond [13]. During this process, N-, C-alpha, and C-atoms of both anchoring residues were considered. We considered fragment deviations of less than 0.5 Å as 'similar'. This limit was set according to the elimination and clustering procedures during data bank creation (Chapt. 3.1.1). Finally, all tested fragments were sorted in order of increasing root mean square deviation between the fitted atoms. The RMSD value for the loops was derived by comparing the complete structures of template and target rather than by simply using the short loop fragments. It would not be sufficient to determine an RMSD value between solely the inserted loop and the target loop, since an incorrect orientation with respect to the target protein (of a correct loop conformation) would not be identified.

3.1.5 Data Evaluation and Correlation to Anchor Groups

3-D fit of Lessel and Schomburg [64] was applied to all anchor group pairs (Chapt. 3.1.2) During this process, loops in the Brookhaven Data Bank were fitted onto a the template protein and the RMS deviation to the target was determined (Chapt. 3.1.4). The best fit (smallest RMSD) was determined for each protein pair, which resulted in a data set representing the maximum prediction quality. All anchor group residues with known RMS deviations were classified by sequence distance of the insertion/deletion (gap length) and

length of inserted loop fragment (loop length). The secondary structure of anchoring residues was determined by using SSTRUC by David Smith [49]. 3_{10}-, α-, and π-helices were grouped into one general helix class labeled as H (Helix), β-sheets and extended conformations were classified as B (Beta), while all other turns and non-regular structures were classified as O (Other). The relative solvent accessibilities of amino acid residues were calculated using the method of Lee and Richards [61] implemented in the PSA program. The accessibilities of both anchoring residues on either side of the loop fragments were averaged.

3.2 *Ab Initio* Equilibrium Constant Estimation using DFT

3.2.1 Retrieval of Reactions from NIST Enzyme Database

Reaction data was obtained from the NIST Online Database 'Thermodynamics of Enzyme-Catalyzed Reactions' [22]. This searchable database contains a collection of thermodynamic data from a total of 1440 enzyme catalyzed reactions which had been published by Goldberg and Tewari [21]. The program *NIST2MySQL* was used to parse the data from the html pages of the NIST Database and to store its contents into two MySQL format database files. One database file named 'statics' contained 1440 enzyme reactions listed by Reference ID, source, enzyme E.C.number, buffer concentration, method of measurement, and accuracy of measurement (Rating range A-F, with A being best). The second database named 'thermodata' contained thermodynamic data on 4075 experiments with detailed information on experimental conditions including temperature, pH, ionic strength, concentration of ions in solution, enthalpy of reaction, and equilibrium constant. A primary key was assigned as a label for each reaction and used as a link between the tables.

3.2.2 Database Format Conversion and Data Processing

The MySQL formatted databases 'statics' and 'thermodata' which were generated by the program *NIST2MySQL* were opened using *PhPMyAdmin* and then exported as .cvs tables. The .cvs tables were then converted to .xls spreadsheet files using *Microsoft Excel XP*. Both databases 'statics.xls' and 'thermodata.xls' were then processed further under *Microsoft Excel XP.*

The database 'statics.xls' contained 1440 chemical equations of the enzyme reactions together with information on EC numbers, buffer type and concentration, method of measurement, and quality rating (Rating range A-F, with A being best). The database 'thermodata.xls' contained 4075 experiments with information on equilibrium constants, buffer pH, temperature, and other experimental conditions such as ion concentrations wherever included. *Microsoft Excel XP* was used to convert experimental equilibrium constants (K') into standard Gibbs free energies of reaction ($\Delta G_r°'_{exp}$). Both 'statics.xls' and 'thermodata.xls' databases were then imported into *Microsoft Access XP* for further processing.

A *Microsoft Access XP* database file was created which contained both 'statics.xls' and 'thermodata.xls' as tables. 'Thermodata.xls' was subjected to a 298.15 ± 1 K and pH 7 ± 0.1 filtering query which isolates all enzyme reactions near standard biological conditions for which an equilibrium constant was reported. The query resulted in a total of 89 enzyme reactions at standard conditions. The experimental data information for those reactions was then linked to 'statics.xls' by a common primary key. This resulted in a table named 'ph7.xls' containing the 89 reactions together with all the information from the original two tables including experimental equilibrium constants, chemical equation, and reaction parameters. From this list, a total of 45 representative reactions were chosen from all six EC groups for the estimation of equilibrium constants using Density Functional Theory (Table 6.2).

3.2.3 Calculation of Reaction Equilibrium using Group Contributions

Equilibrium constants for reactions at pH 7 were determined by first calculating the standard transformed Gibbs free energies of formation ($\Delta G_f°'$) of all metabolites in a reaction using the program *Gibbspredictor* by Kai Hartmann. Molecular information for each metabolite was obtained as MOL format structure files from either the BRENDA or KEGG databases. The MOL file was stored to the local PC, and the structure of each MOL file was visually verified for structural errors and corrected where necessary using the molecular editor of *Gaussview*. MOL files were then used as input files into *Gibbspredictor* which then calculated the standard transformed Gibbs free energy of formation ($\Delta G_f°'$) for the respective metabolite. $\Delta G_f°'$ for metabolites listed and calculated in the publication by the author were obtained directly from the original article [68][69].

Standard transformed Gibbs free energies of reaction ($\Delta G_r°'_{calc}$) were then determined by subtracting the sum of the standard transformed Gibbs free energies of formation ($\Delta G_f°'$) of reactants minus products of the reaction. Reactions which involved pairs of compounds such as NAD(P)/NAD(P)H (Chapt. 3.2.8) or oxidized/reduced Coenzyme A were adjusted using the ($\Delta G_f°'$) values specifically listed for such transformations in the original publication.

3.2.4 Conformational Space Search using *Spartan 06*

In an attempt to find the conformation with the lowest energy minimum, the conformational search tool in *Spartan 06* by Wavefuncion [47] was used. Molecule data was either imported from the Spartan Molecular Database (SMD) library provided by the software package or drawn manually. For the manual construction of molecules, molecular graphics editors by either *Spartan 06* or *Gauss View* were used. . The conformational search was performed using the Monte Carlo algorithm [18] at the molecular mechanics level also

known as the Merck Molecular Force Field (MMFF94) method [24]. Calculations were set up by selecting the conformer distribution operation for molecules at ground state. Charge and multiplicity were set individually for each molecule. The multiplicity of a molecule equals to the total number of electron lone pairs plus 1. Since none of the metabolites were radicals, the multiplicity always equaled to 1 for all metabolites.

The Monte Carlo algorithm [18] works by initially considering the molecule to be in a high-temperature system. This allows the molecule to freely move between low and high energy conformations, which is important, since the global minimum conformation may often be hidden by multiple local minima. As more conformations are explored, the temperature is decreased, making the molecule less able to move out of low energy conformations. At the end of the search, the lowest energy conformations are kept and, depending on the ordering criteria, listed according to lowest energy *in vacuo* or lowest energy in aqueous solution. The energetically most stable conformation from in each list was extracted as .SDF file for further processing. Conformational distributions were determined separately for each charge isomer of a particular metabolite at pH 7 (Chapt. 3.2.5).

3.2.5 Estimation of pKa using *MarvinSketch*

In order to accurately determine the total standard transformed Gibbs free energy ($\Delta G°'_{tot}$) for a particular metabolite (Chapt 3.2.7), the percentage distribution for each charge isomer of that metabolite at pH 7 needs to be determined. In the equilibrium constant estimation by group contribution method (Chapt. 3.2.3), this task was in part performed by the program *Gibbpredictor* [41] which used the Chemaxon molecular library [36] to select the most prevalent charge isomer of a metabolite at pH 7. The charge isomer distribution of a molecule is dependent on the number of acidic hydrogen atoms and their pKa values. For

example, phosphoenolpyruvate has a pKa value of 6.3 so that its charge isomer distribution at pH 7 is 39.5 % for the -2 charged isomer and 60.5 % for the -3 charged isomer. Isomer distributions for all metabolites were estimated using the pKa prediction function of the java-based online application MarvinSketch by Chemaxon [36]. Molecule files were stored as MOL2 files for further processing by *GaussView* [31].

3.2.6 Quantum Mechanical Calculations using *Gaussian 03*

Calculation jobs in *Gaussian 03* [39] require an input file which contains the information necessary for an *ab initio* calculation job. *GaussView* [31] is a graphical platform which can create such an input file. These input files contain a header with several lines of parameters followed by a z-matrix with the molecular coordinates (see below).

After determining all charge isomers for a particular metabolite (Chapt. 3.2.5) and selecting the most stable conformer from the conformer distribution list for each isomer (Chapt. 3.2.4), the MOL2 files of each conformationally most stable charge isomer were imported into *GaussView*. Command parameters were set using the 'calculations' tab and files saved as Gaussian input files (.com or .gjf). After generation of the input files, commands were modified where needed, using a text editor. The Gaussian input file for the initial calculation typically has the following format:

%chk = filename.chk	(Checkfile used for storage of temporary data during job)
%mem= 1000MB	(Memory allocation for job)
%nproc=4	(Number of CPUs used)
# opt freq b3lyp/6-311++g(d,p) scrf=(cpcm,solvent=water) geom.=connectivity	(Command)
Pyruvate	(Name of Molecule File)
-2 1	(Charge and Multiplicity)
C etc.	(Atomic information)

Table 3.3: Input Commands for *Gaussian 03*.

opt	Geometry optimization (Chapt. 2.4.5)
freq	Frequency calculation and Thermodynamics (Chapt. 2.4.5)
b3lyp	DFT with hybrid functional (Chapt. 2.4.7)
6-31g, or 6-311g	Basis sets (Chapt. 2.4.4)
+/++	Inclusion of Diffuse functions (Chapt. 2.4.4)
(d,p)	Inclusion of Polarization Functions (Chapt. 2.4.4)
scrf = (cpcm, solvent=water)	COSMO Solvation Model (Chapt. 2.4.6)
radii = uff	UFF atomic model, used when default (UAO) was unsuccessful
sphereonh = N	Adds an extra spheres on hydrogen atom N (for UAO model)
geom = connectivity	Uses connectivity (atom bond) information provided at end of file
geom = allcheck	Extracts information from chkfile for continuing an unfinished job

Table 3.3 contains a list with the Gaussian commands used during this project. *Gaussian 03* jobs were primarily performed on either 'suns15k' or 'cliot' (Chapt. 6.3.1) both of which are central servers of the University of Cologne Rechenzentrum (RRZK) [31]. Jobs on servers were started as queue jobs. The queue files used to start the jobs included information such as the amount of allocated memory as well as the name of the Gaussian input file. The Gaussian input files were copied from the PC and pasted into the *emacs* text editor of the central server and saved as .com files. During the calculation, an output file with the extension .out is generated by *Gaussian 03*, which contains all the energy information on the geometry optimization and frequency calculation (Chapt. 2.4.5). Jobs which were interrupted due to limits in computation time were continued by using the command geom = allcheck (Table 3.3) which retrieves information from the previous job stored in the .chk files.

3.2.7 Determination of Gibbs Free Energies of Metabolites ($\Delta G°'_{tot}$)

The input files of the energetically most stable conformations of all charge isomers of each metabolite were generated (Chapt. 3.2.6) and a geometry optimization followed by

frequency calculation was performed using density functional theory (DFT) with the b3lyp hybrid functional and 6-311++g (d,p) basis set (Chapt. 3.2.6). This basis set was the most sophisticated one available for the software package and has also been employed in energy calculations of several sugars [73]. In addition, the solvation energies were determined using the COSMO solvation model (Chapt. 2.4.6) by adding the command 'scrf=(cpcm,solvent=water)' (Table 3.3). The finished calculations yielded the following output:

```
Variational C-PCM results
 ========================
 <psi(f)|   H    |psi(f)>                    (a.u.) =   -1254.009750
 <psi(f)|H+V(f)/2|psi(f)>                    (a.u.) =   -1254.346418
 Total free energy in solution:
  with all non electrostatic terms           (a.u.) =   -1254.327822
 ---------------------------------------------------------------------
 (Polarized solute)-Solvent              (kcal/mol) =    -211.26
 ---------------------------------------------------------------------
 Cavitation energy                       (kcal/mol) =      31.56
 Dispersion energy                       (kcal/mol) =     -21.20
 Repulsion energy                        (kcal/mol) =       1.31
 Total non electrostatic                 (kcal/mol) =      11.67
```

The solvation energy can be regarded as a sum of an electrostatic and a non-electrostatic portion (Chapt. 2.4.6). The first line of the output ('< psi(f) | H | psi(f) >') summarizes the total electronic energy without solvation energy of the molecule in atomic units (Hartrees or a.u.). The second line ('< psi(f) | H+V(f)/2 | psi(f) >') represents the total electronic energy plus electrostatic solvation energy, while the energy value displayed in the third line ('Total free energy in solution with all non-electrostatic terms') includes the total electronic energy plus electrostatic solvation energy plus non-electrostatic solvation energy. The energy value displayed in that third line was used as total thermal energy of the metabolite in our calculations. The electrostatic and non-electrostatic solvation energies are separately displayed in kilocalories (kcal/mol) in lines '(Polarized solute)-Solvent' and 'Total non-electrostatic'.

In order to determine the total standard Gibbs free energy ($\Delta G°_{tot}$) of a partiular metabolite, the entropic contribution had to be added to the above thermal energy value. This was achieved by using the 'freq' command (Chapt. 2.4.5) resulting in the following energy output:

```
Zero-point correction=                          0.198343 (Hartree/Particle)
Thermal correction to Energy=                   0.213944
Thermal correction to Enthalpy=                 0.214888
Thermal correction to Gibbs Free Energy=        0.155958
Sum of electronic and zero-point Energies=        -1254.148075
Sum of electronic and thermal Energies=           -1254.132474
Sum of electronic and thermal Enthalpies=         -1254.131530
Sum of electronic and thermal Free Energies=      -1254.190460
```

Here, the entropy portion is separately displayed under 'Thermal correction to Gibbs Free Energy'. For each metabolite, the 'Total free energy in solution with all non-electrostatic terms' (-1254.327822 Hartrees) (see above) was added to the 'Thermal correction to Gibbs free energy' (0.155958 Hartrees), resulting in a total standard Gibbs free energy for that metabolite (-1254.171864 Hartrees). In our calculations, all energy values were converted from atomic units (Hartrees/a.u.) to kilojoules per mol (kJ/mol) by multiplying the energy values in the Gaussian output with 2625.5 as a conversion factor (-3292828.229 kJ/mol).

For charged metabolites, the total standard Gibbs free energy ($\Delta G°_{tot}$) for each charge isomer were evaluated separately and then multiplied with their percentage distribution at pH 7, determined using the pKa prediction tool in *MarvinSketch* (Chapt. 3.2.5), and then summed up to obtain the energy of the total standard transformed Gibbs free energy ($\Delta G°'_{tot}$) of the molecule in solution at pH 7. For sugar molecules, the relative free energies of the alpha and beta anomers were used to determine their respective percentage distribution. The total standard transformed Gibbs free energy ($\Delta G°'_{tot}$) for each sugar was then calculated by multiplying the percentage ratio of the respective anomers and adding together the total standard transformed Gibbs free energies ($\Delta G°'_{tot}$) of the alpha and beta anomers. The

following example demonstrates the determination of the total standard transformed Gibbs free energy ($\Delta G°'_{tot}$) for D-ribose-5-phosphate. This molecule can be considered to be made of alpha and beta anomers. Each anomer has a phosphate group with a pKa of 6.77, as determined through *MarvinSketch*, resulting in a charge isomer distribution of 37.22 % for the (-1) charged and 62.78 % for the (-2) charged species:

α-D-ribose-5-phosphate $^{-1}$ (37.22 %) $\Delta G°_{tot}$ = -2993389.029 kJ/mol
α-D-ribose-5-phosphate $^{-2}$ (62.78 %) $\Delta G°_{tot}$ = -2992162.053 kJ/mol
α-D-ribose-5-phosphate (average) $\Delta G°_{tot}$ = (-2993389.029) x 37.22% + (-2992162.053) x 62.78 %

$\Delta G°'_{tot}$ (α-anomer) = -2992618.733 kJ/mol

β-D-ribose-5-phosphate $^{-1}$ (37.22 %) $\Delta G°_{tot}$ = -2993390.508 kJ/mol
β-D-ribose-5-phosphate $^{-2}$ (62.78 %) $\Delta G°_{tot}$ = -2992162.624 kJ/mol
β-D-ribose-5-phosphate (average) $\Delta G°_{tot}$ = (-2993390.508) x 37.22% + (-2992162.624) x 62.78 %

$\Delta G°'_{tot}$ (β-anomer) = -2992619.642 kJ/mol

The anomeric distribution can be determined by the difference in energy between the two anomers: $\Delta G°'_{tot}$ (α-β) = (-2992618.733 kJ/mol)–(-2992619.642 kJ/mol) = 0.909 kJ/mol
This energy value was then converted into percentage distribution using the formula:

$$\%(\beta) = \frac{1}{e^{\Delta G^{o}_{tot}(\alpha-\beta)/RT}}$$

with RT = $8.3144*10^{-3}$ (298.15) kJ/mol

For the above example, the resulting anomeric ratio for D-ribose5-phosphate would be 69.3 % β-D-ribose-5-phosphate to 30.7 % α-D-ribose-5-phosphate. The final total standard transformed Gibbs free energy ($\Delta G°'_{tot}$) for D-ribose-5-phosphate would then be calculated as: 30.7 % (-2992618.733 kJ/mol) + 69.3 % (-2992619.642 kJ/mol) = -2992619.363 kJ/mol

For comparison, the standard transformed Gibbs free energy of formation ($\Delta G_f°'$) for each metabolite was also calculated using group contribution method (Chapt. 3.2.3)

3.2.8 Calculation of Gibbs Free Energies of Reaction ($\Delta G_r°'$)

Standard transformed Gibbs free energies of reaction ($\Delta G_r°'$) were determined by subtracting the total standard transformed Gibbs free energies ($\Delta G°'_{tot}$) of reactants from products, thus canceling out the electronic energies on both sides of the equation leaving only the difference in Gibbs free energy of the metabolites. Errors between calculated and experimental standard transformed Gibbs energies of reaction were determined as absolute differences between experimental and calculated standard transformed Gibbs free energies of reaction (Error = $|\Delta G_r°'_{exp} - \Delta G_r°'_{cal}|$).

Reactions from EC group 1 (oxidoreductases) involve the oxidation of a particular metabolite accompanied by reduction of the coenzyme $NAD(P)^+$ to NAD(P)H and H^+. The NAD(P)/NAD(P)H molecules were too large for *ab initio* calculations, so that the experimental value of standard transformed Gibbs free energy of reaction at pH 7 of the alcohol dehydrogenase reaction of ethanol to acetaldehyde from the NIST database of enzyme reactions [22] was used for reference. The standard transformed Gibbs free energy of reaction for $NAD(P)^+$ to NAD(P)H at pH 7 was then obtained by subtracting the standard Gibbs free energies of formation ($\Delta G_f°$) of ethanol and acetaldehyde from the standard transformed Gibbs free energy of reaction ($\Delta G_r°'$) of the alcohol dehydrogenase reaction. These $\Delta G_f°$ values were obtained from the National Bureau of Standards (NBS) tables of chemical thermodynamic properties [102]. Even though $\Delta G_f°$ values provided in the NBS tables were measured at pH 0, they can be used in the case of ethanol and acetaldehyde since they have the same neutral charge at both pH 0 and pH 7:

	Ethanol + NAD^+ → Acetaldehyde + NADH + H^+	$\Delta G_r°'$ = + 24.41 kJ/mol [22]
	Ethanol	$\Delta G_f°$ = -180.75kJ/mol [102]
	Acetaldehyde	$\Delta G_f°$ = -139.75kJ/mol [102]
→	**NAD^+ → NADH + H^+**	**$\Delta G_r°'$= + 24.41 - (-139.75) + (-180.75) = - 16.59 kJ/mol**

In order to account for the energy difference of two hydrogens between reduced reactant and oxidized product, the total standard Gibbs free energy ($\Delta G°_{tot}$) of the hydrogen atom was evaluated using DFT under the same conditions as all the metabolites (Chapt. 3.2.7) and added together with the standard Gibbs free energy of formation of the hydrogen atom ($\Delta G_f°_{H\text{-}atom}$) obtained from the NBS tables [102] to obtain the standard transformed Gibbs free energy of reaction ($\Delta G_r°'$)for the particular oxidoreductase reaction.

For example, the $\Delta G_r°'_{calc}$ for the EC 1.1.1.27 reaction of S-lactate to pyruvate involving NAD^+/NADH as a cofactor was calculated in the following manner:

S-lactate + NAD^+	**→**	**pyruvate + NADH + H^+**	**$\Delta G_r°'_{exp}$ = 26.48 kJ/mol**
$\Delta G°_{tot}$ (S-lactate)	=	- 901151.7688 kJ/mol	(Table 6.3)
$\Delta G°_{tot}$ (pyruvate)	=	- 898017.0534 kJ/mol	(Table 6.3)
$\Delta G°_{tot\ (pyruvate)} - \Delta G°_{tot\ (S\text{-}lactate)}$	=	3134.7154 kJ/mol	
$\Delta G_r°'$ ($NAD^+ \rightarrow$ NADH+H^+)	=	- 16.5900 kJ/mol	(Chapt. 3.2.8)
2 x $\Delta G°_{tot}$ (H-atom)	=	2 x -1346.6478 kJ/mol	(Table 6.3)
2 x $\Delta G_f°$ (H-atom)	=	2 x -203.2470 kJ/mol	(NBS tables [102])
$\Delta G_r°'_{calc}$ (DFT)	=	**18.34 kJ/mol**	

Similarly, the $\Delta G_r°'_{calc}$ was determined for the same reaction using group contribution method by Mavrovouniotis [69] as follows:

S-lactate + NAD^+	**→**	**pyruvate + NADH + H^+**	**$\Delta G_r°'_{exp}$ = 26.48 kJ/mol**
$\Delta G_f°'$ (S-lactate)	=	- 520.490 kJ/mol	(Table 6.3)
$\Delta G_f°'$ (pyruvate)	=	- 480.323 kJ/mol	(Table 6.3)
$\Delta G_f°'_{pyruvate} - \Delta G_f°'_{S\text{-}lactate}$	=	40.167 kJ/mol	
$\Delta G_r°'$ ($NAD^+ \rightarrow$ NADH)	=	19.83216 kJ/mol	(MAV [69])
$\Delta G_f°'$ (H^+)	=	- 39.748 kJ/mol	(MAV [69])
$\Delta G_r°'_{calc}$ (MAV)	=	**20.25 kJ/mol**	

4 RESULTS

4.1 Anchor Group Positioning in Knowledge-Based Loop Prediction

4.1.1 Maximum Prediction Quality

The maximum prediction for the test data set was determined by inserting appropriate loops from the fragment data bank into the templates and determining the global RMS deviation to the target using the 3D-alignment procedure of Lessel and Schomburg [64] (Chapt. 3.1.4). Fitted loop fragments were ranked according to RMS deviation and the anchor group combination with the best fit, i.e. showing the lowest RMSD, represented the maximum prediction quality for that particular insertion or deletion. Out of the 595 insertions and 589 deletions, a total of 369 insertions (62.0 %) and 413 deletions (70.1 %) had best fits with an RMS deviation of 1 Å or below (Figure 4.1 and Table 4.1), i.e. could be successfully predicted. When applying a quality criteria of 1.5 Å or less, the ratio of successful predictions increased to 74.3 % for insertions and 83.7 % for deletions.

Table 4.1: **Maximum Prediction Quality for Test Data Set.**

RMSD range [Å]	No. of Insertions	Percentage	No. of Deletions	Percentage
< 0.5	128	21.5 %	135	22.9 %
0.5 – 1.0	241	40.5 %	278	47.2 %
1.0 – 1.5	85	14.3 %	80	13.6 %
1.5 – 2.0	61	10.3 %	37	6.3 %
> 2.0	80	13.4 %	59	10.0 %
Total	595	100.0 %	589	100.0 %

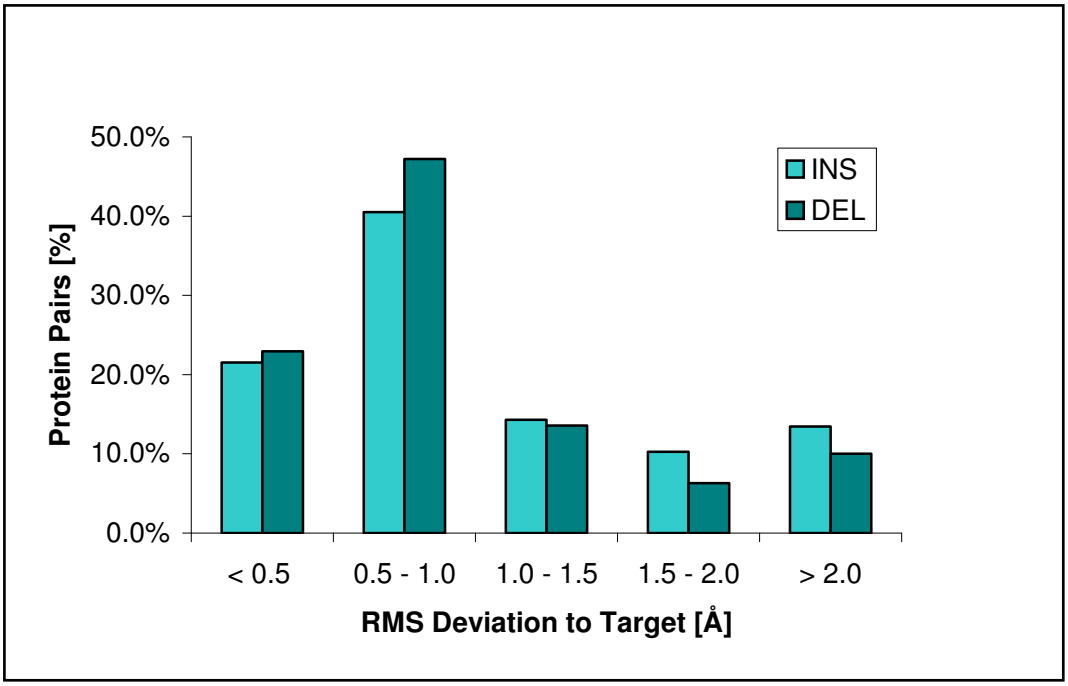


Figure 4.1: **Maximum Prediction Quality.** Maximum prediction quality for the test data set of 595 of insertions and 598 deletions. Graph shows distribution of RMSD range between template and target protein for the best anchor group combination in each protein pair.

When categorizing the data set by gap length, the distribution of the fraction of best fitting anchor groups that allowed an RMSD of < 1.0 Å was highest for shorter gaps and steadily decreased with increasing gap length. The fraction of successfully predicted loops by using the best fitting anchor groups for 1-residue gaps was 75.5% for insertions and 80.3 % for deletions (Table 4.2). For both insertions and deletions, this ratio steadily dropped to 22.2 % and 11.5 % for gaps of 6 residues, and 6.5 % and 11.5 % for gaps of 7 residues or more, respectively (Figure 4.2). In comparison, the ratio of best fitting anchor groups with an RMSD of < 1.0 Å was 91.1 % for zero residue gaps (Table 4.2).

Table 4.2: Maximum Prediction Quality sorted by Gap Length.

Gap Length	*INSERTIONS*			*DELETIONS*		
	Total No.	RMSD < 1Å	Percentage	Total No.	RMSD < 1Å	Percentage
0	45	41	91.1 %	45	41	91.1 %
1	273	206	75.5 %	274	220	80.3 %
2	107	69	64.5 %	105	78	74.3 %
3	48	23	47.9 %	48	27	56.3 %
4	37	14	37.8 %	37	19	51.4 %
5	36	10	27.8 %	36	17	47.2 %
6	18	4	22.2 %	18	8	44.4 %
7+	31	2	6.5 %	26	3	11.5 %
Total	595	369	62.0 %	589	413	70.1 %

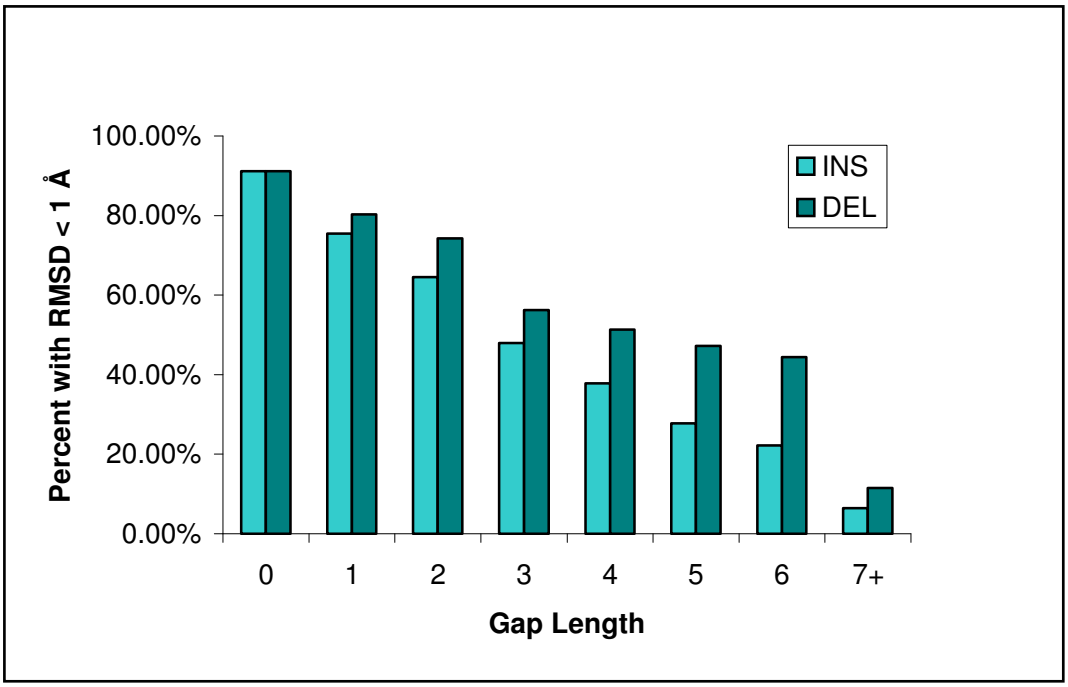


Figure 4.2: **Maximum Prediction Quality sorted by Gap Length.** Fraction of best fitting anchor groups in test data set which allowed the fitting of fragments with RMSD < 1.0 Å. Anchor groups were categorized by gap length.

4.1.2 Influence of Loop Fragment Length

In order to asses whether the length of the fitted loop had an effect on the quality of prediction, the entire data set of best fits for all possible pairs of anchor groups (Chapt. 3.1.3) was sorted according to length of loop fragment. The total data set included 26673 anchor group combinations for insertions and 26088 for deletions. The odds ratio was chosen as a measure of predictive quality, Odds ratios were calculated by taking the ratio between fits with an RMSD < 1 Å and fits with an RMSD ≥ 1 Å for each inserted loop length and dividing this number by the same ratio for the total data set. For example, the odds ratio

for 3-residue loops for insertions (Table 4.3) was determined by dividing 64/245 over 6050/20623. For 4-residue loops the odds ratio was 163/562 divided by 6050/20623, and so on. Odds ratios above 1 represented a higher likelihood of obtaining fits with an RMSD below 1 Å, while odds ratios below 1 represented a lower than average probability of obtaining a good fit. The data showed that medium length loop fragments between 5 to 9 residues performed better than short or long ones (Figure 4.3).

Table 4.3: Prediction Quality sorted by Length of Loop Fragments.

Loop Length	*INSERTIONS*			*DELETIONS*		
	RMSD <1Å	RMSD ≥1Å	Odds Ratio	RMSD <1Å	RMSD ≥1Å	Odds Ratio
3	64	245	0.89	184	852	0.55
4	163	562	0.99	386	1210	0.81
5	319	884	1.23	622	1546	1.02
6	486	1239	1.34	859	1853	1.17
7	590	1699	1.18	1006	2207	1.15
8	677	2193	1.05	1082	2469	1.11
9	799	2668	1.02	1095	2703	1.03
10	903	3174	0.97	1079	2796	0.98
11	979	3714	0.90	868	2499	0.88
12	1070	4245	0.86	208	564	0.93
Total	6050	20623	1.00	7389	18699	1.00

Odds ratio of anchor groups resulting in fitting of loops with RMSD < 1Å. Odds Ratios were determined by taking the ratio of fits with RMSD < 1Å to fits with RMSD ≥ 1Å for each length category divided by the same ratio for the total database.

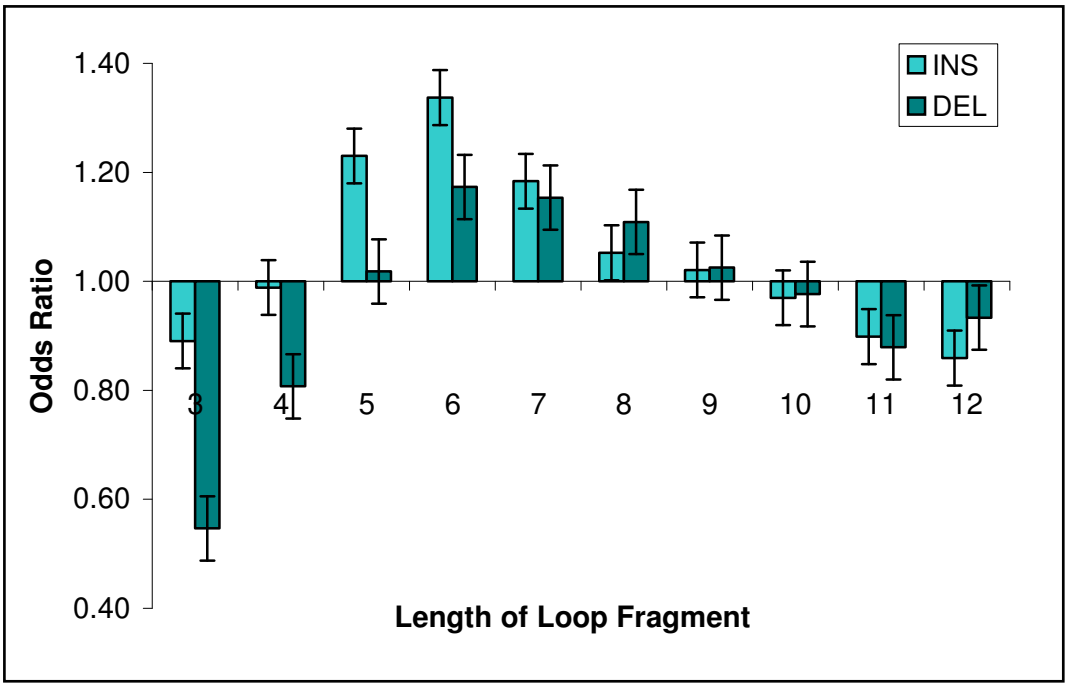


Figure 4.3: **Influence of Loop Fragment Length.** Odds ratio of anchor groups in test data set allowing the fitting of fragments with RMSD < 1.0 Å. Anchor groups were categorized by length of loop fragment. Error bars represent standard error of the mean.

4.1.3 Influence of Amino Acid Type

The odds ratios for the performance of the 20 different amino acids as anchor groups (Table 4.4) showed good performances for tyrosine, leucine, and valine as well as for cysteine and methionine, with the latter two amino acids being low in frequency. Overall, hydrophobic residues revealed a tendency for good predictions, while glycine and proline which are often found in loop regions resulted in low performance. When grouping the amino acids into categories, a similar trend was observed (Table 4.5). Good performances

were achieved with anchor groups of residues with aromatic and hydrophobic side chains, while charged and polar residues performed weakly (Figure 4.4). Glycine and Proline resulted in the lowest prediction quality.

Table 4.4: Prediction Quality sorted by Individual Amino Acids.

Amino Acid Side Chain	*INSERTIONS* RMSD <1Å	RMSD ≥1Å	Odds Ratio	*DELETIONS* RMSD <1Å	RMSD ≥1Å	Odds Ratio	Frequency in PDB
ALA (A)	963	3347	0.98	1131	2815	1.02	8.4 %
ARG (R)	504	1834	0.94	625	1592	0.99	4.9 %
ASN (N)	494	2159	0.78	653	1974	0.84	4.4 %
ASP (D)	812	2664	1.04	829	2441	0.86	5.8 %
CYS (C))	265	656	1.38	329	598	1.39	2.1 %
GLN (Q)	413	1521	0.93	387	1170	0.84	3.7 %
GLU (E)	634	2052	1.05	742	1893	0.99	6.8 %
GLY (G)	1008	4070	0.84	1084	3110	0.88	7.6 %
HIS (H)	259	783	1.13	299	693	1.09	2.2 %
ILE (I)	653	1997	1.11	793	1818	1.10	5.5 %
LEU (L)	1041	2831	1.25	1236	2925	1.07	8.1 %
LYS (K)	636	2363	0.92	862	2191	1.00	6.8 %
MET (M)	199	609	1.11	324	607	1.35	2.2 %
PHE (F)	545	1703	1.09	746	1608	1.17	3.8 %
PRO (P)	509	2056	0.84	577	1979	0.74	4.5 %
SER (S)	769	2710	0.97	874	2678	0.83	5.8 %
THR (T)	716	2775	0.88	1007	2657	0.96	5.7 %
TRP (W)	200	641	1.06	254	578	1.11	1.4 %
TYR (Y)	528	1638	1.10	771	1380	1.41	3.5 %
VAL (V)	953	2837	1.15	1255	2691	1.18	6.9 %
Total	12100	41246	1.00	14778	37398	1.00	100.0 %

Odds ratio of anchor groups resulting in fitting of loops with RMSD < 1Å. Odds Ratios were determined by taking the ratio of fits with RMSD < 1Å to fits with RMSD ≥ 1Å for each amino acid type divided by the same ratio for the total database.

Table 4.5: Prediction Quality sorted by Amino Acid Type.

Amino Acid Category	*INSERTIONS* RMSD <1Å	RMSD ≥1Å	Odds Ratio	*DELETIONS* RMSD <1Å	RMSD ≥1Å	Odds Ratio
ARO	1532	4765	1.10	2070	4259	1.23
HYD	4074	12277	1.13	5068	11454	1.12
NEG	1446	4716	1.05	1571	4334	0.92
POS	1140	4197	0.93	1487	3783	0.99
POL	2391	9165	0.89	2921	8479	0.87
G+P	1517	6126	0.84	1661	5089	0.83
Total	12100	41246	1.00	52176	14778	1.00

Odds ratio of anchor groups resulting in fitting of loops with RMSD < 1Å. Odds Ratios were determined by taking the ratio of fits with RMSD < 1Å to fits with RMSD ≥ 1Å for each amino acid type divided by the same ratio for the total database. Amino acid residues were merged into the following categories: aromatic (**ARO**: F, H, W, Y), hydrophobic (**HYD:** A, C, I, L, M, V), negative (**NEG**: D, E), positive (**POS**: K, R), polar (**POL**: N, Q, S, T) and glycine/proline (**G+P**: G, P).

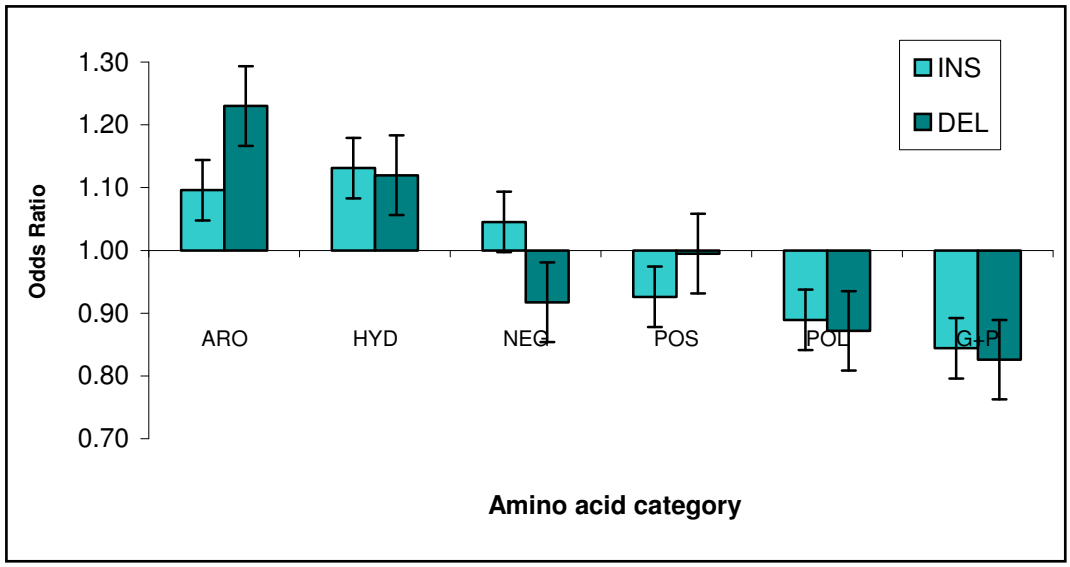


Figure 4.4: Influence of Amino Acid Type. Odds ratio of anchor groups in test data set which allowed the fitting of fragments with an RMSD < 1.0 Å. Anchor groups were categorized by amino acid type.

4.1.4 Influence of Secondary Structure

Anchor groups were grouped into three classes including helix (H), beta sheets (B), and other (O) (Chapt. 3.1.5). Approximately 22 % of the 120300 different anchor groups in the entire data set were helices (H), while 30 % belonged to class 'B' and 48 % to class 'O'. The probability for a successful prediction with an RMSD below 1 Å using different combinations of these three classes can be seen in Table 4.6 and Figure 4.5. Loops connecting two β-sheets (BB) showed the highest probability of being predicted correctly, while loops connecting anchor groups made of two non-regular structure residues (OO) resulted in low performance.

Table 4.6: Prediction Quality sorted by Secondary Structure Combination.

Secondary Structure	*INSERTIONS*			*DELETIONS*		
	RMSD <1Å	RMSD ≥1Å	Odds Ratio	RMSD <1Å	RMSD ≥1Å	Odds Ratio
BB	1072	2060	1.77	1357	1847	1.86
HH	494	1090	1.54	635	1042	1.54
BH	629	1592	1.35	865	1678	1.30
HO	1459	4242	1.17	1729	4152	1.05
BO	1362	6234	0.74	1716	5245	0.83
OO	1034	5405	0.65	1087	4735	0.58
Total	6050	20623	1.00	7389	18699	1.00

Odds ratio of anchor groups resulting in a fitting of loops with RMSD < 1Å. Odds Ratios were determined by taking the ratio of fits with RMSD < 1Å to fits with RMSD ≥ 1Å for each secondary structure combination divided by the same ratio for the total database. Secondary structures were categorized as Helix (**H:** 3_{10}-, α-, π-helices), Beta (**B:** β-sheets and extended conformations), and Other (**O:** turns and other non-regular structures).

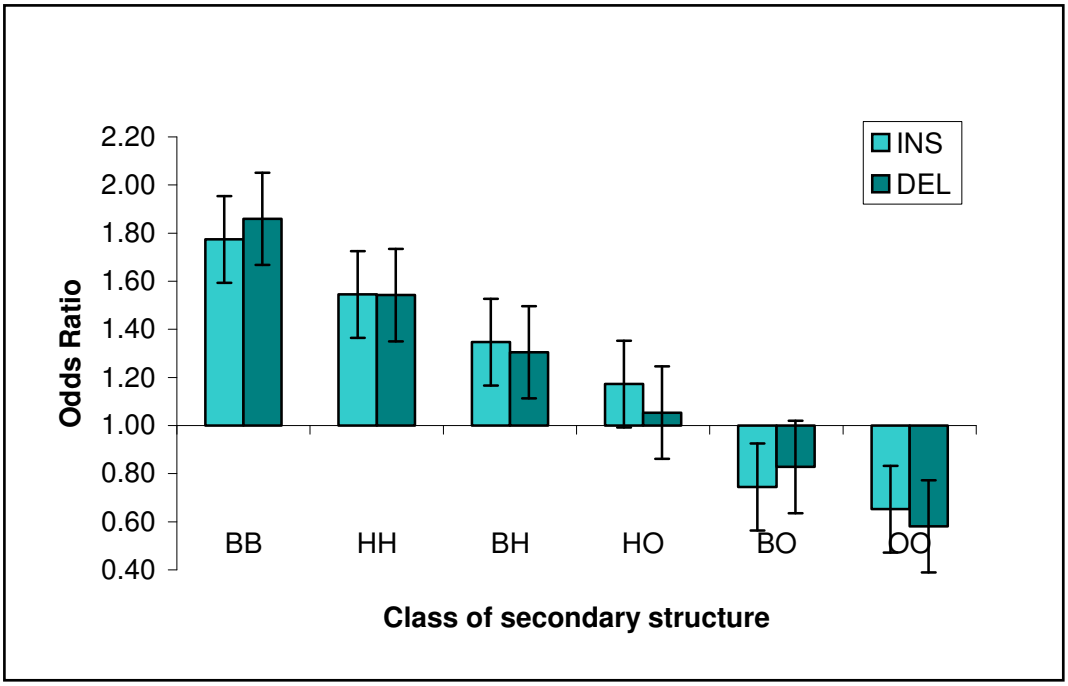


Figure 4.5: **Influence of Secondary Structure.** Odds Ratio of anchor groups in test data set which allowed the fitting of fragments with RMSD < 1.0 Å. Anchor groups were categorized by secondary structure combination. Secondary Structures were categorized into Helix (**H:** 3_{10}-, α-, π-helices), Beta (**B:** β-sheets and extended conformations), and Other (**O:** turns and other non-regular structures). Error bars represent standard error of the mean.

4.1.5 Influence of Solvent Accessibility

For each pair of anchor groups, the relative solvent accessibilities were averaged and grouped into ranges and the odds ratios for the prediction with an RMSD below 1 Å were determined for each group (Table 4.7). The odds ratios showed a clear relationship between relative solvent accessibility of anchor group and prediction quality (Figure 4.6). The highest odds ratio was achieved for anchor groups with 0% relative solvent accessibility with 2.18 for insertions and 2.32 for deletions (Table 4.7). About one third of all anchor group residues had relative solvent accessibilities of below 20 % resulting in a higher than average probability of finding a good fit (Figure 4.6). Anchor groups with almost complete (>70 %) relative solvent accessibility were few in number and had odds ratios of about 0.2 and less.

Table 4.7: Prediction Quality sorted by Relative Solvent Accessibility.

Relative Solvent Accessibilty	*INSERTIONS* RMSD <1Å	RMSD ≥1Å	Odds Ratio	*DELETIONS* RMSD <1Å	RMSD ≥1Å	Odds Ratio
0 %	108	169	2.18	131	143	2.32
0 – 10 %	1497	3290	1.55	2013	3193	1.60
10 – 20 %	1115	3198	1.19	1525	3099	1.25
20 – 30 %	1111	3712	1.02	1414	3474	1.03
30 – 40 %	1029	3631	0.97	1168	3326	0.89
40 – 50 %	638	2996	0.73	658	2648	0.63
50 – 60 %	365	1949	0.64	373	1587	0.59
≥ 60 %	295	1847	0.54	238	1372	0.44
Total	6050	20623	1.00	7389	18699	1.00

Odds ratio of anchor groups resulting in fitting of loops with RMSD < 1Å. Odds Ratios were determined by taking the ratio of fits with RMSD < 1Å to fits with RMSD ≥ 1Å for each relative solvent accessibility category divided by the same ratio for the total database.

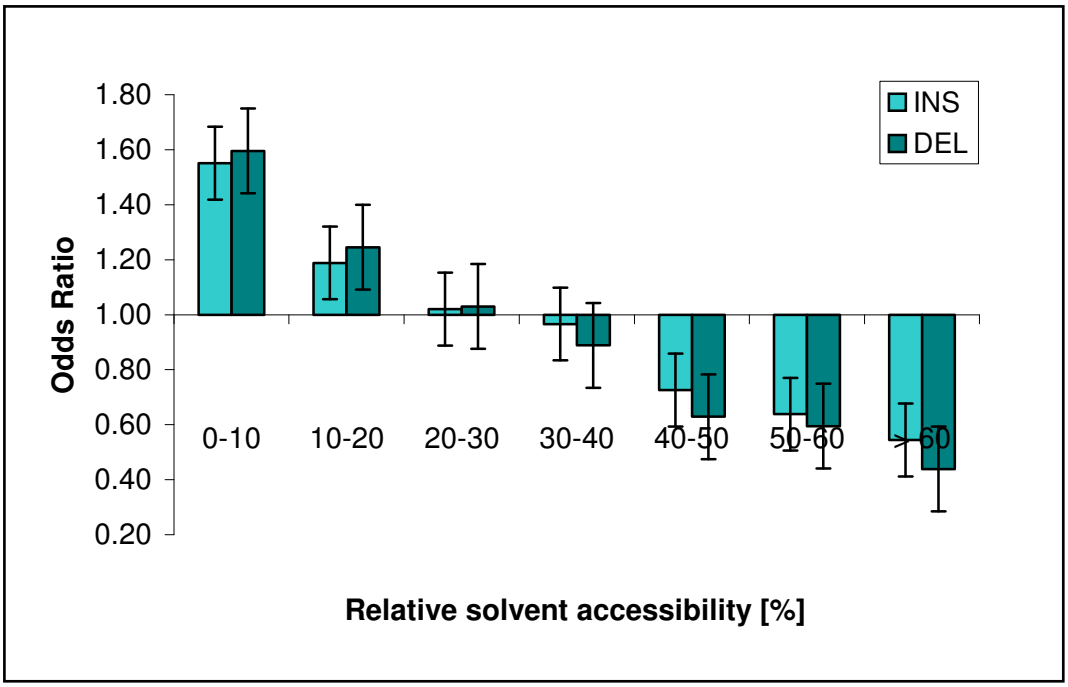


Figure 4.6: **Influence of Relative Solvent Accessibility.** Odds Ratio of anchor groups in test data set which allowed the fitting of fragments with RMSD < 1.0 Å. Anchor groups were categorized by relative solvent accessibility. Error bars represent standard error of the mean.

4.1.6 Prediction using Combination of Criteria

Combined odds ratios were calculated by averaging the odds ratios of loop length, amino acid category, secondary structure, and relative solvent accessibility for each pair of anchor group residues. For each protein pair, the anchor group pair with the highest combined odds ratio was selected and the best fitting loop selected to represent the maximum prediction quality for the chosen anchor group (Table 4.8). The distribution of the maximum prediction quality (Figure 4.7) showed that 26.8 % of the insertions and 34.7 % of the deletions could be predicted with an RMSD below 1 Å. When using 1.5 Å as cutoff

criteria, the fraction of successful predictions increased to 45.3 % and 62.6 %, respectively. A comparison between prediction using combined odds ratios versus random selection of anchor groups showed a clear increase in the quality of prediction for all gap lengths (Figure 4.8). The highest improvement in prediction quality was observed for medium gaps of 3 to 5 residues in length.

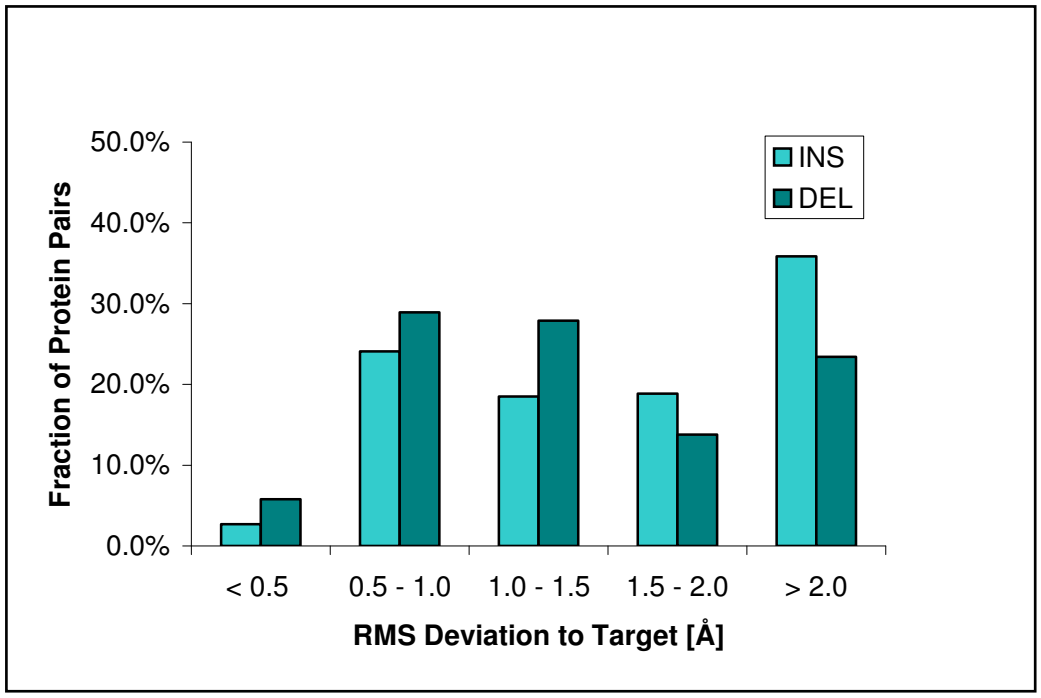


Figure 4.7: **Prediction using Combination of Criteria.** Prediction quality by application of combined anchor group positioning criteria. For each protein pair, the odds ratios assigned by class of loop length, amino acid class, secondary structure, and relative solvent accessibility were averaged to a combined odds ratio and the anchor group combination with the highest combined odds ratio was selected for each protein pair.

Table 4.8: **Prediction using Combination of Criteria.**

RMSD range [Å]	No. of Insertions	% of Total	No. of Deletions	% of Total
< 0.5	16	2.7%	34	5.8 %
0.5 – 1.0	143	24.1 %	170	28.9 %
1.0 – 1.5	110	18.5 %	164	27.9 %
1.5 – 2.0	112	18.9 %	81	13.8 %
> 2.0	213	35.9 %	139	23.4 %
Total	594	100.0 %	588	100.0 %

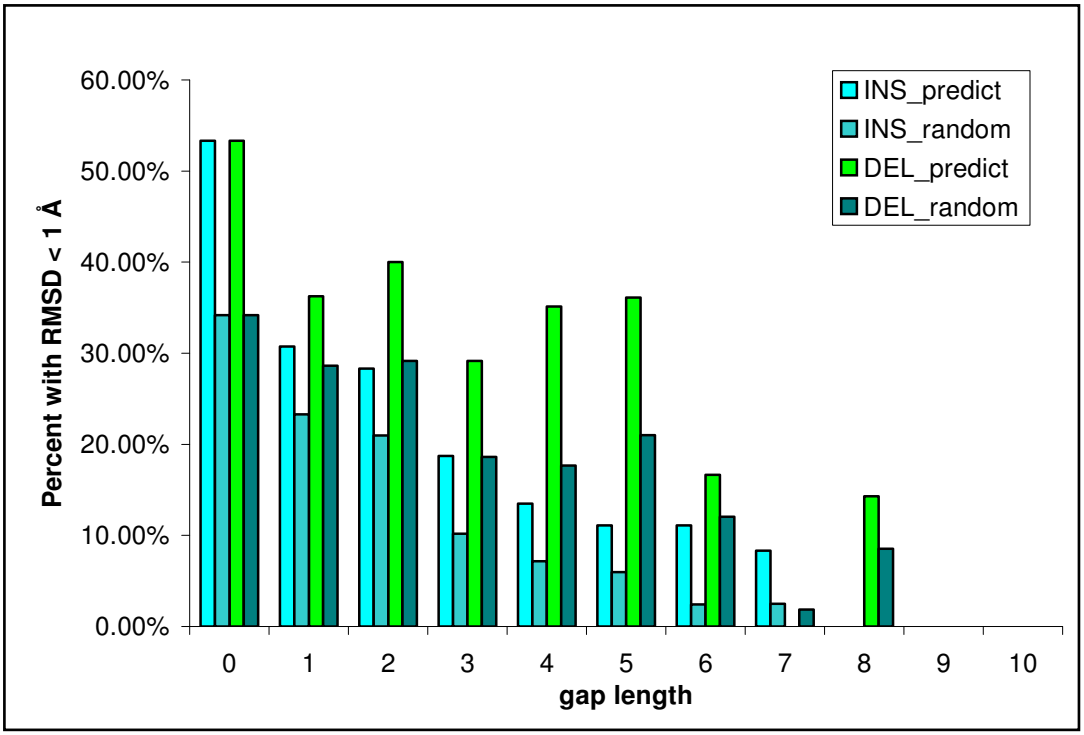


Figure 4.8: **Prediction using Combined Odds Ratios vs. Random Anchor Groups.** Fraction of possible predictions with an RMSD < 1.0 Å by selection of anchor groups using combined odds ratios compared to random anchor group selection. Anchor groups were categorized by gap length.

4.2 *Ab Initio* Equilibrium Constant Estimation using DFT

4.2.1 Effect of Conformational Search Method

Four isomerase reactions (EC Group 5) were chosen from the NIST database of enzyme reactions [22] and a conformational search of all the metabolites was performed using one of the latest semi-empirical methods (PM3) [97] and compared to a commonly used molecular mechanics method (MMFF94) [24]. In each method, the most stable conformer was chosen and used as a starting structure for the *in vacuo* calculation of the standard transformed Gibbs free energy of reaction (ΔG_r'') using DFT (Chapter 3.2.8).

Table 4.9 shows that the conformational search method had a significant effect on the accuracy of the calculated standard transformed Gibbs free energies of reaction (ΔG_r''), and that molecular mechanics (MMFF94) with a mean error of 2.22 kcal/mol was superior to a search by semi-empirical approach using the PM3 method (mean error = 4.50 kcal/mol). This established molecular mechanics (MMFF94) as the conformational search method of choice.

Table 4.9: Effect of Conformational Search on Gibbs Free Energy of Reaction (ΔG_r'')

EC	VACUUM-Semi Empirical (PM3)	ΔG_r'' calc	ΔG_r'' exp	Error DFT	Error MAV
5.3.1.15	D-Lyxose = D-Xylulose	-2.47	3.64	1.46	1.07
5.3.1.4	L-Arabinose = L-Ribulose	-13.65	5.47	4.57	1.51
5.3.1.5	D-Glucose = D-Fructose	-32.57	0.75	7.96	2.78
5.3.1.7	D-Mannose = D-Fructose	-19.49	-2.72	4.01	1.95
		Error (kcal)		4.50	1.83

EC	VACUUM-Molecular Mechanics (MMFF94)	ΔG_r'' calc	ΔG_r'' exp	Error DFT	Error MAV
5.3.1.15	D-Lyxose = D-Xylulose	-4.67	3.64	1.99	1.07
5.3.1.4	L-Arabinose = L-Ribulose	-1.17	5.47	1.59	1.51
5.3.1.5	D-Glucose = D-Fructose	-9.75	0.75	2.51	2.78
5.3.1.7	D-Mannose = D-Fructose	-14.50	-2.72	2.81	1.95
		Error (kcal)		2.22	1.83

4.2.2 Effect of Solvation Model

The same four isomerase (EC group 5) reactions which were used to investigate the effect of conformational search method (Chapt. 4.2.1) were used to analyze the effect of different solvation models on the calculation of the standard transformed Gibbs free energy of reaction (ΔG_r'') using DFT. The most stable conformation of all metabolites was determined by molecular mechanics (MMFF94) (Chapt. 3.2.4) and then used in DFT calculations while comparing the polarized continuum (PCM) solvation model [71] to the COSMO solvation model [57]. Table 4.10 shows that the mean error was slightly lower for the COSMO (1.51 kcal/mol) compared to the PCM (1.64 kcal/mol) solvation model, thus confirming the COSMO solvation model as the method of choice for the remaining calculations (Chapt. 4.2.3). Both calculations involving solvation model were superior to the calculations achieved *in vacuo* (Table 4.9).

Table 4.10: Effect of Solvation Model on Gibbs Free Energy of Reaction (ΔG_r'')

EC	PCM - Solvation Model	ΔG_r'' calc	ΔG_r'' exp	Error DFT	Error MAV
5.3.1.15	D-Lyxose = D-Xylulose	11.38	3.64	1.85	1.07
5.3.1.4	L-Arabinose = L-Ribulose	13.21	5.47	1.85	1.51
5.3.1.5	D-Glucose = D-Fructose	11.21	0.75	2.50	2.78
5.3.1.7	D-Mannose = D-Fructose	-1.21	-2.72	0.36	1.95
		Error (kcal)		1.64	1.83

EC	COSMO - Solvation Model	ΔG_r'' calc	ΔG_r'' exp	Error DFT	Error MAV
5.3.1.15	D-Lyxose = D-Xylulose	6.27	3.64	0.63	1.07
5.3.1.4	L-Arabinose = L-Ribulose	15.75	5.47	2.46	1.51
5.3.1.5	D-Glucose = D-Fructose	7.21	0.75	1.55	2.78
5.3.1.7	D-Mannose = D-Fructose	3.24	-2.72	1.42	1.95
		Error (kcal)		1.51	1.83

4.2.3 Standard Transformed Gibbs Free Energy of Reaction ($\Delta G_r°'$)

The standard transformed Gibbs free energies of reaction ($\Delta G_r°'$) were calculated for all 45 selected enzyme catalyzed reactions at pH 7 using Density Functional Theory (DFT). The mean error for reactions from EC group 1 was 2.49 kcal/mol for the 14 oxidoreductase reactions, which was slightly superior to the error of 2.76 kcal/mol calculated by group contribution method (Table 4.11). Estimation of the standard transformed Gibbs free energies of reaction ($\Delta G_r°'$) for isomerases and ligases (EC groups 5 and 6) yielded mean errors in the same range with 5.50 kcal/mol when using DFT and 4.76 kcal/mol when using group contribution method. The highest deviation was observed for transferase reactions (EC group 3) with an error of 25.24 kcal/mol. The overall mean error for all 45 reactions was 10.30 kcal/mol for DFT and 4.60 kcal/mol for group contribution method (Figure 4.9).

Table 4.11: Standard Transformed Gibbs Free Energies of Reaction ($\Delta G_r°'$).

EC	OXIDOREDUCTASES (EC GROUP 1)	$\Delta G_r°'_{exp}$ (kJ/mol)	$\Delta G_r°'_{cal}$ (kJ/mol)	Error DFT (kcal)	Error MAV (kcal)
1.1.1.-	2-hydroxyglutarate + NAD = 2-oxoglutarate + NADH	27.59	19.25	1.99	1.75
1.1.1.1	ethanol + NAD = acetaldehyde + NADH	24.41	13.58	2.59	0.79
1.1.1.1	2-propanol + NAD = Acetone + NADH	6.49	2.26	1.01	3.29
1.1.1.10	L-xylitol + NADP = L-xylulose + NADPH	20.13	17.63	0.60	0.93
1.1.1.29	(R)-glycerate + NAD = hydroxypyruvate + NADH	32.55	19.82	3.04	2.94
1.1.1.27	(S)-lactate + NAD = pyruvate + NADH	26.48	18.34	1.95	1.49
1.1.1.28	(R)-lactate + NAD = pyruvate + NADH	27.71	18.41	2.22	1.78
1.1.1.30	3-hydroxybutanoate + NAD = 3-oxobutanoate + NADH	10.48	-0.70	2.67	2.34
1.1.1.62	estradiol-17-beta + NAD = estrone + NADH	4.25	-9.27	3.23	1.02
1.1.1.9	ribitol + NAD = D-ribulose + NADH	21.44	23.21	0.42	0.61
1.1.1.8	glycerol-3-phosphate + NAD = dihydroxyacetone-phosphate + NADH	24.14	38.10	3.34	2.53
1.1.1.37	(S)-malate + NAD = oxaloacetate + NADH	28.24	11.68	3.96	1.91
1.1.1.42	isocitrate + NADP + H_2O = 2-oxoglutarate + NADPH + carbonate	0.37	-32.12	7.77	1.85
1.8.1.4	dihydroxy-alpha-lipoate + NAD = alpha-lipoate + NADH	4.54	4.23	0.07	15.45
		MEAN		2.49	2.76

Table 4.11 (cont.): Standard Transformed Gibbs Free Energies of Reaction ($\Delta G_r°'$).

EC	TRANSFERASES (EC GROUP 2)	$\Delta G_r°'_{exp}$ (kJ/mol)	$\Delta G_r°'_{cal}$ (kJ/mol)	Error DFT (kcal)	Error MAV (kcal)
2.4.2.1	adenosine + P_i = adenine + alpha-D-ribose-1-phosphate	12.94	57.54	10.66	4.81
2.6.1.2	L-alanine + 2-oxoglutarate = pyruvate + L-glutamate	1.82	-30.23	7.66	0.33
2.6.1.51	L-alanine + hydroxypyruvate = L-serine + pyruvate	-3.59	-1.00	0.62	0.86
2.6.2.1	L-aspartate + 2-oxoglutarate = oxaloacetate + L-glutamate	4.25	-11.56	3.78	1.02
2.7.1.11	ATP + D-fructose-6-phosphate = ADP + D-fructose-1,6-bisphosphate	-16.57	35.01	12.33	13.86
2.7.1.6	ATP + D-galactose = ADP + alpha-D-galactose-1-phosphate	-8.08	201.11	50.00	9.18
2.7.3.2	phosphocreatine + beta-guanidino-propionate = creatine + phospho-guanidino-propionate	-2.77	1.64	1.06	0.66
2.7.3.2	ATP + creatine = ADP + phosphocreatine	12.34	380.69	88.04	9.03
2.7.4.3	2 ADP = AMP + ATP	0.60	190.81	45.30	0.53
2.7.9.1	ATP + pyruvate + P_i = AMP + phosphoenolpyruvate + PP_i	17.45	-120.76	33.03	21.13
		MEAN		25.24	6.14

EC	HYDROLASES (EC GROUP 3)	$\Delta G_r°'_{exp}$ (kJ/mol)	$\Delta G_r°'_{cal}$ (kJ/mol)	Error DFT (kcal)	Error MAV (kcal)
3.1.3.1	D-glucose-6-phosphate + H_2O = D-glucose + P_i	-13.81	-59.95	11.03	0.50
3.2.1.23	lactose + H_2O = D-galactose + D-glucose	-11.04	-65.67	13.06	5.76
3.2.2.7	adenosine + H_2O = adenine + D-ribose	-9.84	-43.76	8.11	5.35
3.5.1.11	phenylacetylglycine + H_2O = phenylacetic acid + glycine	-0.45	36.38	8.80	4.59
3.5.1.14	N-acetyl-L-alanine + H_2O = acetate + L-alanine	-5.45	31.31	8.79	14.70
3.5.4.5	cytidine + H_2O = uridine + NH_3	-22.91	-8.43	3.46	2.83
		MEAN		8.87	5.62

EC	LYASES (EC GROUP 4)	$\Delta G_r°'_{exp}$ (kJ/mol)	$\Delta G_r°'_{cal}$ (kJ/mol)	Error DFT (kcal)	Error MAV (kcal)
4.2.1.31	(R)-malate = maleate + H_2O	18.90	-21.51	7.57	3.82
4.2.1.35	2-methylmalate = 2-metylmaleate + H_2O	5.80	-38.98	7.95	0.29
4.2.1.85	2,3-dimethylmalate = dimethylmaleate + H_2O	6.00	-61.21	12.71	9.07
4.3.2.2	adenylosuccinate = fumarate + AMP	10.96	-78.04	21.00	1.98
4.6.1.1	ATP = adenosine-3',5'-cyclic-phosphate + PP_i	6.78	-62.22	18.90	10.38
		MEAN		13.63	7.24

Table 4.11 (cont.): Standard Transformed Gibbs Free Energies of Reaction ($\Delta G_r°'$).

EC	ISOMERASES and LIGASES (EC GROUP 5/6)	$\Delta G_r°'_{exp}$ (kJ/mol)	$\Delta G_r°'_{cal}$ (kJ/mol)	Error DFT (kcal)	Error MAV (kcal)
5.3.1.15	D-lyxose = D-xylulose	3.64	6.27	0.63	1.07
5.3.1.4	L-arabinose = L-ribulose	5.47	15.75	2.46	1.51
5.3.1.5	D-glucose = D-fructose	0.75	7.21	1.55	2.78
5.3.1.6	D-ribose-5-phosphate = D-ribulose-5-phosphate	3.31	12.37	2.17	0.09
5.3.1.7	D-mannose = D-fructose	-2.72	3.24	1.42	1.95
5.4.2.2	alpha-D-glucose-1-phosphate = alpha-D-glucose-6-phosphate	-7.02	31.12	9.12	0.52
5.4.2.8	beta-D-glucose-1-phosphate = beta-D-glucose-6-phosphate	-3.24	58.35	14.72	2.00
5.4.2.8	D-mannose-1-phosphate = D-mannose-6-phosphate	0.00	37.34	8.92	29.20
5.1.3.2	alpha-D-galactose-1-phosphate = alpha-D-glucose-1-phosphate	-2.72	30.87	8.03	0.65
6.4.1.1	ATP + pyruvate + carbonate = ADP + P_i + oxaloacetate	0.83	24.06	5.95	7.80
		MEAN		5.50	4.76

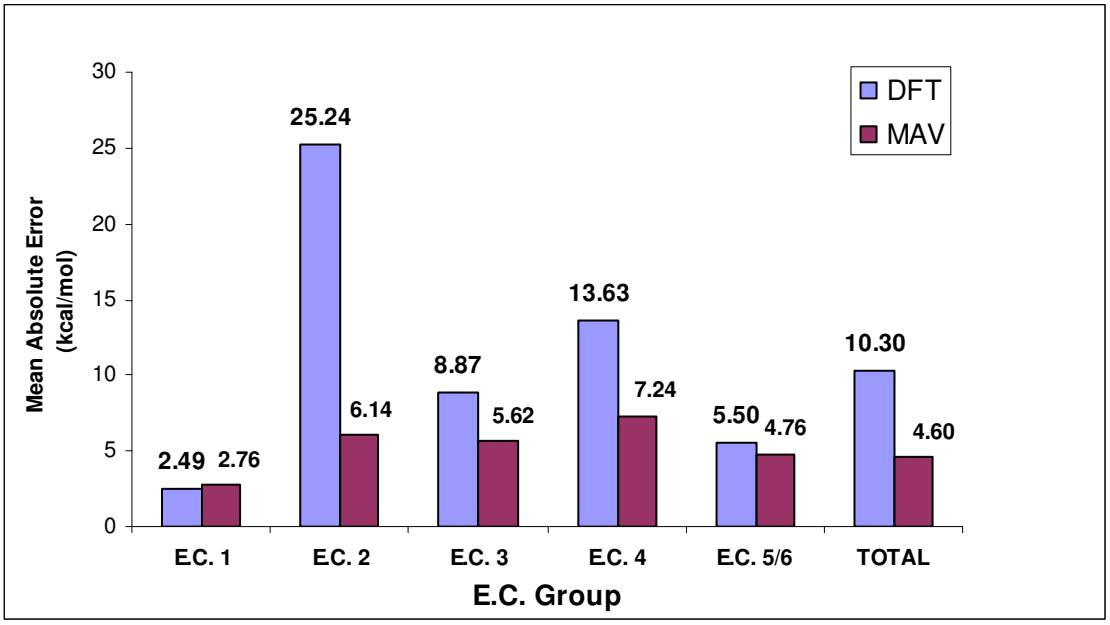


Figure 4.9: **Mean Absolute Error for Estimation of $\Delta G_r°'$.** Mean absolute error for estimation of standard transformed Gibbs free energy of reaction ($\Delta G_r°'$) using DFT compared to group contribution method (MAV) [68].

5 DISCUSSION

5.1 Anchor Group Positioning in Knowledge-Based Loop Prediction

5.1.1 Interpretation of Results

The prediction of protein loops around insertions and deletions is one of the biggest challenges in protein structure prediction. Knowledge-based loop prediction places second in the source of error next to template-target alignment, and its quality is dependent upon multiple factors such as the algorithm for appropriate fragment selection, completeness of the fragment databank, the fitting/optimization procedure, and the choice of anchor groups. The present thesis focuses on the selection of appropriate anchor groups which shall allow sufficient prediction quality for the fragments to be fitted.

By applying the fragment database used in this project, it was possible for about two thirds of the loops in our test data set (62.0 % for insertions and 70.1 % for deletions) to be modeled with an RMSD < 1 Å. This maximum prediction quality could be achieved if good methods for anchor group selection and loop fragment ranking existed. When raising the

strict requirement of a global RMSD < 1 Å up to an RMSD < 1.5 Å, the fraction of loops which could be successfully predicted increased to about 80 % (Chapt. 4.1.1). Even though loops with smaller gap sizes were overrepresented in this data set, this ratio has also been observed in evolution, where the number of examples with longer gaps decreases exponentially [73].

The decrease in maximum prediction quality with increase in gap length (Figure 4.2) can be explained by database incompleteness as longer loops give rise to an exponentially greater number of possible structures. In addition, longer gaps potentially create a more diverse environment with respect to the target so that the anchoring regions alone do not provide enough information about the structure of the whole loop region.

The prediction quality did not correlate with the length of the inserted fragments as one would have expected. Interestingly, the shorter fragments did not perform as well as medium sized fragments, probably due to conformational strain. As fragment length increased, the added residues may have provided additional structural flexibility such that torsional strains from one residue may have been compensated by neighboring residues. It appeared that fragments between 5 and 7 residues showed the best performance by avoiding the steric strain of shorter fragments and the database incompleteness of longer fragments.

The influence of the amino acids of the anchor groups on the prediction quality has shown to favor the hydrophobic type which can mostly be found on the inside of the protein core. At the same time, charged residues and glycine or proline, which are located around loop regions and on the protein surface performed comparably worse. This suggests that conformational stability of the anchor group residues has an effect on the quality of the overall loop prediction.

A similar tendency was observed in the influence of secondary structure. Here, anchor group residues located inside a defined structure such as a helix or beta sheet gave better prediction results than anchor groups in irregular structures and loops. It appears that

an anchor group in a fixed stable conformation provides the correct initial orientation of the loop prediction leading to a better overall prediction quality.

Solvent accessibility showed a trend similar to secondary structure and amino acid type, in that less accessible anchor group residues which are buried inside the protein core clearly showed better results compared to anchor groups on the more solvent accessible protein surface. Again the same interpretation applies in that buried residues tend to be less flexible than surface residues, which favored loop prediction by providing greater conformational stability.

A combination of the above rules lead to a maximum prediction rate of 27 % for insertions and 35 % for deletions, using 1 Å as the RMSD cutoff. When the criteria was expanded to 1.5 Å, the fraction of successfully predicted loops increased to 45.3 % and 62.6 %, respectively. This was a significant improvement compared to a maximum prediction rate of 18 % for insertions and 26 % for deletions when anchor groups were chosen randomly (Figure 4.8). By including additional criteria such as temperature factors or by considering weighting factors for the different criteria, one may further improve the prediction quality. In addition there may be redundancy associated with the different criteria, which means that they may not be treated as independent from each other.

5.1.2 Outlook

Protein loop prediction is still an unsolved problem. The process can be divided into different steps which include the identification of anchor groups that belong to structurally conserved regions of the protein, the selection of loops either by conformational search or from fragment databases, and the ranking and identification of best fitting fragment candidates possibly followed by a final optimization step with side-chain placement. In this

thesis, the importance of anchor group positioning was analyzed with respect to gap length, fragment length, amino acid type, secondary structure, and relative solvent accessibility. It was demonstrated that an improvement in the prediction quality can be achieved by using a combination of the above criteria to locate appropriate anchor groups for each modeled loop region compared to a random choice of anchors. By using an improved scoring method for the criteria in anchor group selection as well as an appropriate algorithm for the ranking of loop fragments, the loop prediction results may be enhanced even further. Additional final optimization steps include the use of energy functions and the correct placement of amino acid side chains.

5.2 *Ab Initio* Equilibrium Constant Estimation using DFT

5.2.1 Interpretation of Results

Biochemical reactions are catalyzed by enzymes which are highly specific, and by lowering the activation energy, enzymes allow catalyzed reactions to run at much higher rates than if non-catalyzed. The feasibility and reversibility of a specific biochemical reaction is determined by its equilibrium constant and the concentrations of its reactants and products. Since equilibrium constants are usually not readily available for any random biochemical reaction, methods have been developed to predict those constants independently from experimental measurements.

This project was designed as an attempt to estimate equilibrium constants of biochemical reactions using Density Functional Theory (DFT). The basis set that was subsequently chosen was the maximally extended 6-311++ G (d,p) set. This basis set resulted in the lowest energies and was the most sophisticated basis set in *Gaussian* 03 for use in DFT

calculations. In addition, this basis set had been used in several works concerning quantum mechanical calculations of carbohydrates, where it had shown to highly correlate with experimental values regarding the estimation of anomeric ratios where the prediction of the anomerizaton ratio of glucose was at around 1 kcal/mol [73].

This study showed that Density Functional Theory (DFT) [58] can be used in combination with a preceding molecular mechanics conformational search (MMFF94) [24] to estimate standard transformed Gibbs free energies of reaction (ΔG_r°') for enzyme catalyzed reactions at standard biochemical conditions (pH 7 and 298.15 K). For reactions from EC group 1 and EC groups 5 and 6, the calculated standard transformed Gibbs free energies of reaction deviated from their experimental values by an average of 2.49 kcal/mol and 5.50 kcal/mol, respectively. These values were comparable to the reaction free energies calculated using group contribution method by Mavrovouniotis, where the mean error was 2.76 kcal/mol for reactions from EC group 1 and 4.76 kcal/mol for reactions from EC groups 5 and 6. Looking at the entire set of 45 reactions, the calculated values deviated from experimental values by an average of 10.30 kcal/mol. For the reactions studied, this result was above the overall deviation of 4.60 kcal/mol determined by group contribution method. However, considering that these were *ab initio* calculations, the results can be considered very satisfactory, and may serve as a good starting point for further optimization.

Improvements may be made by using an improved estimation method for the evaluation of frequencies based on anharmonic potentials rather than harmonic oscillators. Such anharmonic models have recently been developed and may lead to an increase in the accuracy of entropy estimations [4]. Other possible improvements can be made with respect to the solvation model used. SM6 is a continuum solvation model [55] which has shown to be slightly more accurate than COSMO in modeling aqueous solutions. The mean unsigned error (MUE) for SM6 lies at 0.54 kcal/mol, compared to COSMO at 1.11 kcal/mol [55]. SM6 has not yet been implemented in the most recent version of *Gaussian 03*. An improved

solvation model using a combination of actual water molecules as an initial solvation layer combined with a continuum model for the outer layers could also help reduce errors in the solvation energy calculations. Problems, however, may arise in the actual number and configuration of the water molecules to be applied.

As more reactions are being investigated using his method, the results can be used to generate an error function which may reduce the gap between theoretical and experimental equilibrium data. A similar fitting approach to the prediction of free energy calculations has been done by Nanda et. al [75]. By solving such an error term using multiple regression analysis with a chosen set of properties such as size of Van-de-Waals radii or hydrophobicity or charge and weight of the molecules, the discrepancy between theory and experiment might be reduced even to a higher degree than currently achievable with this method.

5.2.2 Outlook

The ability to reliably predict equilibrium constants of enzyme reactions establishes the fundamental groundwork for simulating metabolic flux in biochemical pathways, ultimately leading to the modeling of metabolic networks and the entire cell.

Recent advances in the collection of data in molecular and cellular biology, especially in the field of genome sequencing, have led to the revival of an old vision: the simulation of complete biological systems. However, as of today, the virtual cell is still visionary and its realization depends on the ongoing development of model networks and algorithms as well as the continued generation of large amounts of accurate experimental data. Here, the *in silico* generation of equilibrium constants for biochemical reactions constitutes a crucial feature allowing the construction of metabolic networks despite the lack of available

experimental data. The simulation of such metabolic networks has become one of the main focuses at the Institute of Biochemistry at the University of Cologne.

The actual creation of a metabolic-network model from experimental data consists of several steps. Here, the availability of complete and accurate data is crucial for the modeling of a reliable metabolic network. First, the organism for which the model will be designed needs to be chosen (e.g. *E.coli*, *C. glutamicum*, erythrocyte, etc.). For initial analysis, modeling maybe focused on one or more subsystems (e.g. glycolysis, Krebs cycle). Data acquisition includes reactions and metabolites which can be obtained from the KEGG database [40]. Enzyme parameters can be retrieved from BRENDA [34], while experimental equilibrium constants may be obtained from the NIST Database of Enzyme Reactions [22]. For reactions where no experimental equilibrium constants are recorded, constants may be estimated by group contributions [68][69] or by using *ab initio* quantum mechanical methods as demonstrated through this project.

As of today, the virtual cell is still visionary and its realization depends on the ongoing development of model networks and algorithms as well as the continued generation of large amounts of accurate experimental data. Here, the *in silico* generation of equilibrium constants for biochemical reactions constitutes a crucial feature allowing the construction of metabolic networks despite the lack of available experimental data.

6 APPENDIX

6.1 Databases

6.1.1 Protein Databank (PDB)

The Protein Databank (PDB) [43] represents publicly available collection of experimentally determined protein structures. The databank was established in 1971 at the Brookhaven National Laboratory and originally contained 7 structures. Since 1998, the PDB has been managed by the Research Collaboratory for Structural Bioinformatics (RCSB). So far, the size of the database has risen exponentially and currently holds more than 48,000 structures (Table 2.3). The Worldwide Protein Data Bank (wwPDB) consists of organizations that act as deposition, data processing and distribution centers for PDB data. The founding members are RCSB PDB (USA), PDBe (Europe) and PDBj (Japan). The BMRB (USA) group joined the wwPDB in 2006. The mission of the wwPDB [48] is to maintain a single Protein Data Bank Archive of macromolecular structural data that is freely and publicly available to the global community.

6.1.2 NIST Database of Enzyme Reactions

The NIST Online Database 'Thermodynamics of Enzyme-Catalyzed Reactions' [22] is a searchable collection of thermodynamic data from the National Institute of Standards and Technology and contains a collection of enzyme catalyzed reactions which had previously been published in six separate publications by the same group of authors [21]. The data presented is limited to direct equilibrium and calorimetric measurements performed under *in vitro* conditions. The following information is given for each entry in this database: data reference, chemical equation, enzyme name, Enzyme Commission (EC) number, method of measurement, experimental conditions (temperature, pH, ionic strength, buffers, cofactors), and subjective evaluation. The subjective evaluation was performed by using a rating system: A (high quality), B(good), C (average), or D (low quality). In making these assignments, the authors considered the various experimental details which were provided in the study. These details included the method of measurement, the number of data points determined, and the extent to which the effects of varying temperature, pH, and ionic strength were investigated. A low rating was generally given when few only details of the investigation were reported.

6.1.3 BRENDA (BRaunschweig ENzyme DAtabase)

BRENDA is a publicly open online database containing enzyme functional data [34]. The database is maintained, developed, and hosted by the Institute of Biochemistry at the University of Cologne and is available for academic, non-profit, and commercial users via the internet (www.brenda.uni-koeln.de). The project is the continuation of an attempt to

develop an enzyme data information system which started in 1987 at the German National Research Center for Biotechnology in Braunschweig (GBF).

BRENDA represents a collection of enzyme functional data of at least 83,000 different enzymes from more than 9,800 different organisms collected out of approximately 46,000 references. The data has been systematically arranged and classified according to the EC system of the IUBMB Enzyme Nomenclature Committee into about 4,200 different EC numbers [34]. Data on enzyme function have been extracted directly from the primary literature and critically evaluated by qualified scientists. The original authors' nomenclature for enzyme forms and subunits has been retained and redundant information has been avoided if possible. Enzyme data can be searched according to a variety of search parameters including nomenclature (EC number), structure, stability, substrates, functional parameters (K_m value, pH optimum, turnover number, etc.), organism, and pathology. The database currently develops into a metabolic network information system with links to enzyme expression and regulation information.

6.1.4 KEGG (Kyoto Encyclopedia of Genes and Genomes)

KEGG is a bioinformatics resource developed by the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo [40] The KEGG resource allows the integration of genomic and molecular information as a basis for understanding higher–level biological systems, such as cells, organisms, and their interactions with the environment.

KEGG consists of four main databases also referred to as building blocks. Molecular building blocks include one representing the genomic space (KEGG GENES database) and one representing the chemical space (KEGG LIGAND database). Systemic information is

represented as molecular wiring diagrams in the network space (KEGG PATHWAY database) and ontologies for pathway reconstruction (KEGG BRITE database).

KEGG GENES is a collection of gene catalogs for all complete and some partial genomes from 31 eukaryotes, 235 bacteria, and 23 archae. Each entry contains cross-reference information to outside databases.

KEGG LIGAND is a database which consists of multiple components which include enzyme nomenclature, chemical compound structures, chemical reaction formulas, and glycan structures. The database also contains drug structures with therapeutic categories and target molecule information.

KEGG PATHWAY contains a collection of manually drawn pathway maps for metabolism, genetic information processing, human diseases, and environmental information processing (e.g. signal transduction, ligand-receptor interaction, cell communication, etc.).

KEGG BRITE reflects an attempt to use the hierarchically structured knowledge about the genomic, chemical and network spaces for making functional interpretations as part of the pathway reconstruction process.

The final goal is to develop a complete computer representation of the cell, possibly leading to the computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information.

6.2 Software Packages

6.2.1 *Gaussian 03*

Gaussian 03 [39] is the latest in the *Gaussian* series of electronic structure programs. *Gaussian 03* is used by chemists, chemical engineers, biochemists, physicists and others for

research in established and emerging areas of chemical interest. Starting from the basic laws of quantum mechanics, *Gaussian 03* predicts the energies, molecular structures, and vibrational frequencies of molecular systems, along with numerous molecular properties derived from these basic computation types. It can be used to study molecules and reactions under a wide range of conditions, including both stable species and compounds which are difficult or impossible to observe experimentally such as short-lived intermediates and transition structures. *Gaussian 03* offers the Polarizable Continuum Model (PCM) [71] and COSMO solvation model [57] for modeling systems in solution. These models represents the solvent as a polarizable continuum and place the solute in a cavity within the solvent.

6.2.2 *Spartan '06*

Spartan 06 by Wavefuncion [47], is a quantum mechanics calculation program. Unlike *Gaussian 03,* it integrates the calculational engine into a graphical user interface. In this project, *Spartan 06* was used to search the conformational space of molecules using molecular mechanics (MMFF94) [24] calculations.

6.2.3 *Gibbspredictor*

Gibbspredictor is a program written by Kai Hartmann [41], which allows the automated implementation of the group contribution method by Mavrovouniotis [68][69], in estimating standard transformed Gibbs free energies of formation ($\Delta G_f^{\circ\prime}$) for organic and biochemical compounds. The method is limited to biological standard conditions (pH 7 and 298.15 K). The library is written in Java and uses the Chemistry Development Kit for

reading, writing, storing and mani-pulating molecular structures. It can handle a huge number of input and output formats such as CML, MDL Molfile, MOL2, etc. The current version *Gibbspredictor 2.0* is available through the University of Cologne bioinformatics website: www.hnb-cologne.uni-koeln.de/gibbspredictor/gibbspredictor.html.

6.2.4 *JChem*

To transform the structures to their predominant hydrogenation state at pH 7, the *JChem* library from ChemAxon [36] was used. The library is freely available under an academic license. For purposes of this project, the module was integrated into *Gibbspredictor1.3.0*, allowing an automated conversion of the input molecule into its pH 7 hydrogenation state prior to the standard Gibbs free energy of formation estimation.

6.2.5 *NIST2MySQL*

This module belongs to a software package written by Sebastian Breuers for a project at the Cologne University Bioinformatics Center (http://www.cubic.uni-koeln.de/). The CUBIC project was an attempt to optimize the estimation of standard transformed Gibbs free energies of formation by group contribution method. This module uses regular expression patterns to parse the information contained in the html pages of the NIST Database of Enzyme Reactions (Chapt. 6.1.2). The program extracts the information into MySQL database format and adds primary keys for indexing the data. Data includes enzyme names, reactions catalyzed, experimental equilibrium constants, quality rating, and reaction conditions such as pH, temperature, and solvent.

6.2.6 *GaussView*

GaussView [31] is a graphical user interface designed to help prepare molecular structure files for submission to *Gaussian 03* and to examine graphically the output that *Gaussian 03* produces. Through its advanced visualization facility, *GaussView* allows rapid sketching in even very large molecules, then rotation, translation and zooming in on these molecules through simple mouse operations. It can also import standard molecule file formats such as PDB files. In this project, *GaussView3.0.9* was used to verify the structure of MOL files downloaded from BRENDA (Chapt. 6.1.3) or KEGG (Chapt. 6.1.4) databases for use in standard transformed Gibbs free energy estimations using the program *Gibbspredictor* (Chapt. 6.2.3).

6.2.7 Database Management Software

Databases containing enzyme data and experimental reaction information from the NIST Database of Enzyme Reactions (Chapt. 6.1.2) were managed and analyzed using *PhPMyAdmin, Open Office Calc. 1.1, Microsoft Excel XP, and Microsoft Access XP.*

PhPMyAdmin

PhPMyAdmin is a database administration tool specifically designed for handling databases in MySQL format over the web. The program was written in PhP, an html-embeddable widely used scripting language especially suited for web development. Currently, *PhPMyAdmin* can create and drop databases, create/drop/alter tables, delete/edit/add fields, execute any SQL statement, manage keys on fields, manage privileges, and export data into various formats. In our project, the program was employed as a management tool for the databases created by *NIST2MySQL* (Chapt. 6.2.5). The current

version (*PhPMyAdmin 2.10.0*) is available though the phpmyadmin project homepage under http://www.phpmyadmin.net/home_page/index.php.

Open Office Calc 1.1

Open Office Calc 1.1 is a spreadsheet program similar to *Microsoft Excel* with a roughly equivalent range of features. *OpenOffice.org* aims to compete with *Microsoft Office* and emulate its look and feel where suitable. It can read and write most of the file formats found in Microsoft Office, and many other applications. In this study, the spreadsheet program was used to convert the .cvs database files generated by *PhPMyAdmin* into *Microsoft Excel* .xls files.

Microsoft Excel XP

Microsoft Excel XP is a spreadsheet program written and distributed by Microsoft for computers using the Microsoft Windows operating system and for Apple Macintosh computers. This spreadsheet program belongs to the *Microsoft Office XP* package, and was used as the main tool in combination with *Microsoft Access XP* for processing and analyzing the information extracted from the NIST Database of Enzyme Reactions (Chapt. 6.1.2). In this project, *Microsoft Excel XP* was used for splitting chemical reactions into their individual substrates, for combining equilibrium constants, for calculating standard transformed Gibbs free energies, and for creating graphs representing the analyzed data.

Microsoft Access XP

Microsoft Access XP is a relational database management system by Microsoft which combines the relational Microsoft Jet Database Engine with a graphical user interface, and it is part of the *Microsoft Office XP* system. In this project, the program was used to organize

tables and create links between the metabolite names and the associated codes in BRENDA (Chapt. 6.1.3) and KEGG (Chapt. 6.1.4). In addition, *Microsoft Access XP* allowed the creation of cross-tables and filter queries for data sorting and analysis.

6.3 Hardware and Computer Resources

6.3.1 Servers

The *suns15k* and *cliot* HPC (High Performance Computing) servers (Table 6.1) at the Regional Computing Center (Regionales Rechenzentrum-RRZK) at the University of Cologne [46] were used for performing quantum mechanical calculations with *Gaussian 03* (Chapt. 3.2.6).

Table 6.1: Standard Servers and Software Packages used in this Project.

Server IP-Address	Model	#CPUs	Memory	Software
suns15k.rrz.uni-koeln.de	SUN Fire 15k	72	144 GB	*Gaussian 03*
cliot.rrz.uni-koeln.de	SUN Opteron Cluster	258	772 GB	*Gaussian 03*

6.3.2 Local Workstation

A local PC Pentium 4 with a SUSE Linux 9.0 platform was used for handling database related tasks involving *Open Office Calc* (Chapt. 6.2.7) as well as for storing molecule files and for running the Software *Gibbspredictor* (Chapt. 6.2.3) and *NIST2MySQL* (Chapt. 6.2.5). A Microsoft Windows XP platform was used for data analysis using *Microsoft Office XP* (Chapt. 6.2.7).

6.4 List of Reactions

Table 6.2: List of Reactions for Estimation of Reaction Equilibrium using DFT.

EC	Ref. ID [1]	REACTION	$\Delta G_r^{\circ\prime}{}_{exp}$ [2] (kJ/mol)	Temp (K)	pH	I [3] (mol/l)
1.1.1.-	87BUC/MIL	2-hydroxyglutarate + NAD = 2-oxoglutarate + NADH	27.59	298.15	7	N/A
1.1.1.1	36EUL/ADL	ethanol + NAD = acetaldehyde + NADH	24.41	298.15	7	N/A
1.1.1.1	52BUR	2-propanol + NAD = Acetone + NADH	6.49	298.15	7	N/A
1.1.1.10	59HOL	L-xylitol + NADP = L-xylulose + NADPH	20.13	298.15	7	N/A
1.1.1.29	82GUY	(R)-glycerate + NAD = hydroxypyruvate + NADH	32.55	298.15	7	0.25
1.1.1.27	52NEI	(S)-lactate + NAD = pyruvate + NADH	26.48	298.15	7	N/A
1.1.1.28	86MEI/GAD	(R)-lactate + NAD = pyruvate + NADH	27.71	298.15	7	N/A
1.1.1.30	62KRE/MEL	3-hydroxybutanoate + NAD = 3-oxobutanoate + NADH	10.48	298.15	7	N/A
1.1.1.62	58LAN/ENG	estradiol-17-beta + NAD = estrone + NADH	4.25	298.15	7	N/A
1.1.1.9	59HOL	ribitol + NAD = D-ribulose + NADH	21.44	298.15	7	N/A
1.1.1.8	58YOU/PAC	glycerol-3-phosphate + NAD = dihydroxyacetone-phosphate + NADH	24.14	297.65	7	N/A
1.1.1.37	73GUY/GEL	(S)-malate + NAD = oxaloacetate + NADH	28.24	298.15	7.06	0.25
1.1.1.42	68LON/DAL	isocitrate + NADP + H_2O = 2-oxoglutarate + NADPH + carbonate	0.37	298.15	7.05	0.1
1.8.1.4	85LIE	dihydroxy-alpha-lipoate + NAD = alpha-lipoate + NADH	4.54	298.15	7.03	1.1
2.4.2.1	80CAM/SGA	adenosine + Pi = adenine + alpha-D-ribose-1-phosphate	12.94	298.15	7	N/A
2.6.1.2	45DAR	L-alanine + 2-oxoglutarate = pyruvate + L-glutamate	1.82	298.15	7.15	N/A
2.6.1.51	82GUY	L-alanine + hydroxypyruvate = L-serine + pyruvate	-3.59	298.15	7	0.25
2.6.2.1	45DAR	L-aspartate + 2-oxoglutarate = oxaloacetate + L-glutamate	4.25	298.15	7.15	N/A
2.7.1.11	75BOH/SCH	ATP + D-fructose-6-phosphate = ADP + D-fructose-1,6-bisphosphate	-16.57	298.15	7	N/A
2.7.1.6	61ATK/BUR	ATP + D-galactose = ADP + alpha-D-galactose-1-phosphate	-8.08	298.15	7	N/A
2.7.3.2	86MEY/BRO	phosphocreatine + beta-guanidino-propionate = creatine + phospho-guanidino-propionate	-2.77	298.15	7.06	N/A
2.7.3.2	92TEA/DOB	ATP + creatine = ADP + phosphocreatine	12.34	298.15	6.98	0.25
2.7.4.3	83KHO/KAR	2 ADP = AMP + ATP	0.60	297.15	7	N/A
2.7.9.1	68REE/MEN	ATP + pyruvate + Pi = AMP + phosphoenolpyruvate + PPi	17.45	298.15	7	N/A
3.1.3.1	61ATK/JOH	D-glucose-6-phosphate + H2O = D-glucose + Pi	-13.81	298.15	7	N/A
3.2.1.23	86HUB/HUR	lactose + H2O = D-galactose + D-glucose	-11.04	298.15	7	N/A
3.2.2.7	80CAM/SGA	adenosine + H2O = adenine + D-ribose	-9.84	298.15	7	N/A
3.5.1.11	80SVE/MAR	phenylacetylglycine + H2O = phenylacetic acid + glycine	-0.45	298.15	7	N/A
3.5.1.14	86ROH/ETT	N-acetyl-L-alanine + H2O = acetate + L-alanine	-5.45	298.15	7	N/A

EC	Ref. ID [1]	REACTION	$\Delta G_r^{\circ\prime}{}_{exp}$ [2] (kJ/mol)	Temp (K)	pH	I [3] (mol/l)
3.5.4.5	71COH/WOL	cytidine + H2O = uridine + NH3	-22.91	298.15	7	N/A
4.2.1.31	93WER/TWE	(R)-malate = maleate + H2O	18.90	298.15	7	0.1
4.2.1.35	93WER/TWE	2-methylmalate = 2-metylmaleate + H2O	5.80	298.15	7	0.1
4.2.1.85	93WER/TWE	2,3-dimethylmalate = dimethylmaleate + H2O	6.00	298.15	7	0.1
4.3.2.2	55CAR/COH	adenylosuccinate = fumarate + AMP	10.96	298.15	7	N/A
4.6.1.1	74KUR/TAK	ATP = adenosine-3',5'-cyclic-phosphate + PPi	6.78	298.15	7	N/A
5.3.1.15	65AND/ALL	D-lyxose = D-xylulose	3.64	298.15	7	N/A
5.3.1.4	58HEA/HOR	L-arabinose = L-ribulose	5.47	298.15	7	N/A
5.3.1.5	67TAK	D-glucose = D-fructose	0.75	298.15	7	N/A
5.3.1.6	54AXE/JAN	D-ribose-5-phosphate = D-ribulose-5-phosphate	3.31	298.65	7	N/A
5.3.1.7	67TAK2	D-mannose = D-fructose	-2.72	298.15	7	N/A
5.4.2.2	59ATK/JOH	alpha-D-glucose-1-phosphate = alpha-D-glucose-6-phosphate	-7.02	298.15	7	N/A
5.4.2.8	96OES/SCH	beta-D-glucose-1-phosphate = beta-D-glucose-6-phosphate	-3.24	298.15	7	N/A
5.4.2.8	96OES/SCH	D-mannose-1-phosphate = D-mannose-6-phosphate	0.00	298.15	7	N/A
5.1.3.2	54HAN/CRA	alpha-D-galactose-1-phosphate = alpha-D-glucose-1-phosphate	-2.72	298.15	7.1	N/A
6.4.1.1	66WOO/DAV	ATP + pyruvate + carbonate = ADP + Pi + oxaloacetate	0.83	298.15	7.03	N/A

[1] All thermodynamical data was obtained from NIST Database of Enzyme Reactions [22] 'Thermodynamics of Enzyme Catalyzed Reactions' (http://www.xpdb.nist.gov/enzyme_thermodynamics)

[2] Experimental standard transformed Gibbs free energies of reaction ($\Delta G_r^{\circ\prime}{}_{exp}$) were calculated from apparent equilibrium constants (K') by using $\Delta G_r^{\circ\prime} = -RT \ln K'$ (Chapt. 2.3.4).

[3] N/A= Not available (Information on ionic strength was not obtainable)

6.5 List of Metabolites

Table 6.3: Total Standard Gibbs Free Energies of Metabolites determined by DFT.

Metabolite	Formula[1)]	pKa[2)]	% distrib	Charge	ΔG°_{tot} [3)] (kJ/mol)	ΔG_f°, [4)] (kJ/mol)
Acetaldehyde	**C_2H_4O**		**100.0 %**	**0**	**-403939.264**	**-139.746**
Acetic acid	**$C_2H_4O_2$**		**100.0 %**	**-1**	**-600404.810**	**-366.518**
Acetone	**C_3H_6O**		**100.0 %**	**0**	**-507137.026**	***-146.440***
Adenine	**$C_5H_5N_5$**		**100.0 %**	**0**	**-1227147.900**	**321.331**
Adenosine	**$C_{10}H_{13}N_5O_4$**		**100.0 %**	**0**	**-2530059.800**	**-207.945**
AMP (-1)			37.27 %	-1	-4019729.299	
AMP (-2)		6.77	62.73 %	-2	-4018508.655	
AMP (avg)	**$C_{10}H_{14}N_5O_7P$**		**100.0 %**	**-1.63**	**-4018963.589**	**-208.782**
ADP (-2)			57.10 %	-2	-5509388.685	
ADP (-3)		7.12	42.90 %	-3	-5508167.685	
ADP (avg)	**$C_{10}H_{15}N_5O_{10}P_2$**		**100.0 %**	**-2.43**	**-5508864.876**	**-230.538**
ATP (-3)			84 %	-3	-6999033.443	
ATP (-4)		7.72	16 %	-4	-6997800.128	
ATP (avg) [5)]	**$C_{10}H_{16}N_5O_{13}P_3$**		**100.0 %**	**-3.16**	**-6998836.113**	**-252.714**
Adenosine-(3'5'-cyclic)-P	**$C_{10}H_{12}N_5O_6P$**	**1.83**	**100.0 %**	**-1**	**-3818971.12**	**37.656**
Adenylosuccinate (-3)			38.00 %	-3	-5214149.98	
Adenylosuccinate (-4)		6.78	62.00 %	-4	-5212902.19	
Adenylosuccinate (avg)	**$C_{14}H_{18}N_5O_{11}P$**		**100.0 %**	**-3.62**	**-5213376.35**	***-836.800***
L-Alanine	**$C_3H_7NO_2$**		**100.0 %**	**0**	**-850116.65**	**-369.029**
N-acetyl-alanine	**$C_5H_9NO_3$**		**100.0 %**	**-1**	**-1249790**	***-431.789***
Ammonia	**NH_3**		**100.0 %**	**0**	**-148526.1742**	***-75.730***
α-L-Arabinose			18.3 %	0	-1503725.281	
β-L-Arabinose			81.7 %	0	-1503728.992	
L-Arabinose (avg)	**$C_5H_{10}O_5$**		**100.0 %**	**0**	**-1503728.314**	**-746.007**
L-Aspartate	**$C_4H_7NO_4$**		**100.0 %**	**-1**	**-1344208.909**	**-696.2176**
Carbonic acid	H_2CO_3		10.14 %	0	-696039.7079	-622.998
Hydrogen carbonate	HCO_3^-		89.84 %	-1	-694865.9948	-586.597
Carbonate (avg)			**100.0 %**	**-0.9**	**-694846.0361**	
Creatine	**$C_4H_9N_3O_2$**		**100.0 %**	**0**	**-1240898.169**	**-267.358**
Creatine-P (-1)			9.7 %	-1	-2730550.661	
Creatine-P (-2)		6.07	90.3 %	-2	-2729339.873	
Creatine-P (avg)	**$C_4H_{10}N_3O_5P$**		**100.0 %**	**-1.9**	**-2729457.319**	**-238.906**
Cytidine	**$C_9H_{13}N_3O_5$**		**100.0 %**	**0**	**-2340031.6**	***-589.107***
Dihydroxy-α-lipoate	**$C_8H_{15}O_2S_2$**		**100.0 %**		**-3310489.364**	***-265.266***
Dihydroxyacetone-P (-1)			34.9 %	-1	-2391913.327	
Dihydroxyacetone-P (-2)			65.1 %	-2	-2390679.375	
Dihydroxyacetone-P (avg)	**$C_3H_7O_6P$**		**100.0 %**	**-1.65**	**-2391110.024**	***-444.341***
(2R,3S)-Dimethylmalate	**$C_6H_{10}O_5$**	**5.4**	**100.0 %**	**-2**	**-1601528.31**	***-836.382***
Dimethylmaleate	**$C_6H_8O_4$**	**5.88**	**100.0 %**	**-2**	**-1400812.7**	***-555.635***
Estradiol-17-Beta	**$C_{18}H_{24}O_2$**		**100.0 %**	**0**	**-2233581.866**	***-88.282***
Estrone	**$C_{18}H_{22}O_2$**		**100.0 %**	**0**	**-2230474.752**	***-59.831***
Ethanol	**C_2H_6O**		**100.0 %**	**0**	**-407069.226**	**-180.749**

Metabolite	Formula[1)]	pKa[2)]	% distrib	Charge	ΔG°_{tot} [3)] (kJ/mol)	ΔG_f°, [4)] (kJ/mol)
β-D-Fructopyranose			98.4 %	0	-1804465.189	
β-D-Fructofuranose			1.6 %	0	-1804454.901	
D-Fructose (avg)	**$C_6H_{12}O_6$**		**100.0 %**	**0**	**-1804465.03**	**-906.673**
β-D-Fructofuranose-6-P (-1)			37.22 %	-1	-3294093.794	
β-D-Fructofuranose-6-P (-2)		6.77	62.78 %	-2	-3292880.974	
β-D-Fructopyranose-6-P (-1)			35.96 %	-1	-3294104.341	
β-D-Fructopyranose-6-P (-2)		6.75	64.04 %	-2	-3292858.718	
β-D-Fructofuranose-6-P (avg)			**100.0 %**	**-1.63**	**-3293332.385**	
β-D-Fructopyranose-6-P (avg)			**0.0 %**	**-1.64**	**-3293306.644**	
β-D-Fructose-6-P (avg)	**$C_6H_{13}O_9P$**		**100.0 %**	**-1.63**	**-3293332.385**	**-907.510**
β-D-Fructose-1,6-P (-2)			13.39 %	-2	-4783774.94	
β-D-Fructose-1,6-P (-3)		6.46	23.83 %	-3	-4782595.094	
β-D-Fructose-1,6-P (-3)			22.58 %	-3	-4782551.622	
β-D-Fructose-1,6-P (-4)		7.06	40.20 %	-4	-4781336.281	
β-D-Fructose-1,6-P (avg)	**$C_6H_{14}O_{12}P_2$**		**100.0 %**	**-3.27**	**-4782237.217**	**-888.263**
Fumarate	**$C_4H_4O_4$**	**4.41**	**100.0 %**	**-2**	**-1194489.66**	**-608.772**
α-D-Galactose			51.0 %	0	-1804470.632	
β-D-Galactose			49.0 %	0	-1804470.537	
D-Galactose (avg)	**$C_6H_{12}O_6$**		**100.0 %**	**0**	**-1804470.586**	**-895.794**
α-D-Galactose-1-P (-1)			24.80 %	-1	-3294123.047	
α-D-Galactose-1-P (-2)		6.52	75.20 %	-2	-3292907.973	
α-D-Galactose-1-P (avg)	**$C_6H_{13}O_9P$**		**100.0 %**	**-1.75**	**-3293209.311**	**-887.426**
α-D-Glucose			17.3 %	0	-1804469.027	
β-D-Glucose			82.7 %	0	-1804472.912	
D-Glucose (avg)	**$C_6H_{12}O_6$**		**100.0 %**	**0**	**-1804472.242**	**-895.794**
α-D-Glucose-1-P (-1)			24.03 %	-1	-3294128.951	
α-D-Glucose-1-P (-2)		6.5	75.97 %	-2	-3292877.865	
β-D-Glucose-1-P (-1)			24.03 %	-1	-3294127.497	
β-D-Glucose-1-P (-2)		6.5	75.97 %	-2	-3292879.091	
α-D-Glucose-1-P (avg)			**44.2 %**	**-1.76**	**-3293178.442**	
β-D-Glucose-1-P (avg)			**55.8 %**	**-1.76**	**-3293179.025**	
D-Glucose-1-P (avg)	**$C_6H_{13}O_9P$**		**100.0 %**	**-1.76**	**-3293178.768**	**-887.426**
α-D-Glucose-6-P (-1)			20.83 %	-1	-3294139.938	
α-D-Glucose-6-P (-2)		6.42	79.17 %	-2	-3292886.24	
β-D-Glucose-6-P (-1)			20.83 %	-1	-3294142.358	
β-D-Glucose-6-P (-2)		6.42	79.17 %	-2	-3292851.95	
α-D-Glucose-6-P (avg)			**100.0 %**	**-1.79**	**-3293147.324**	
β-D-Glucose-6-P (avg)			**0.0 %**	**-1.79**	**-3293120.679**	
D-Glucose-6-P (avg)	**$C_6H_{13}O_9P$**		**100.0 %**	**-1.79**	**-3293352.861**	**-896.631**
L-Glutamate	**$C_5H_9NO_4$**		**100.0 %**	**-1**	**-1447387.643**	**-689.105**
(R)-Glycerate	**$C_3H_5O_4$**		**100.0 %**	**-1**	**-1098715.828**	**-669.022**
Glycerol-3-P (-1)			37.2 %	-1	-2395046.271	
Glycerol-3-P (-2)			62.8 %	-2	-2393801.422	
Glycerol-3-P (avg)	**$C_3H_9O_6P$**		**100.0 %**	**-1.62**	**-2394264.506**	**-477.813**
Glycine	**$C_2H_5NO_2$**		**100.0 %**	**0**	**-746942.48**	**-374.886**
Guanidinopropionate	**$C_4H_9N_3O_2$**		**100.0 %**	**0**	**-1240965.188**	**-288.278**
P-Guanidinopropionate (-1)			9.8 %	-1	-2730621.948	
P-Guanidinopropionate (-2)		6.09	90.2 %	-2	-2729403.265	
P-Guanidinopropionate (avg)	**$C_4H_{10}N_3O_5P$**		**100.0 %**	**-1.90**	**-2729522.696**	**-259.826**

Metabolite	Formula[1)]	pKa[2)]	% distrib	Charge	ΔG°_{tot} [3)] (kJ/mol)	ΔG_f°, [4)] (kJ/mol)
Hydrogen atom	**H**		**100.0 %**	**0**	**-1346.648**	**-203.247**
3-Hydroxybutanoate	**$C_4H_7O_3$**		**100.0 %**	**-1**	**-1004328.714**	***-513.795***
2-Hydroxyglutarate	**$C_5H_8O_5$**		**100.0 %**	**-2**	**-1498393.45**	***-840.984***
Hydroxypyruvate	**$C_5H_4N_4O$**		**100.0 %**	**-1**	**-1095579.628**	**-628.855**
Isocitrate	**$C_6H_8O_7$**		**100.0 %**	**-3**	**-1992430.428**	**-1163.152**
(R)-Lactate	**$C_3H_6O_3$**		**100.0 %**	**-1**	**-901151.8474**	**-520.490**
(S)-Lactate	**$C_3H_6O_3$**		**100.0 %**	**-1**	**-901151.7688**	**-520.490**
α-D-Lactose			21.1%	0	-3408111.8	
β-D-Lactose			78.9%	0	-3408115	
D-Lactose (avg)	**$C_{12}H_{22}O_{11}$**		**100.0 %**	**0**	**-3408114.4**	**-1519.629**
α-Lipoate	**$C_7H_{11}O_2S_2$**		**100.0 %**		**-3307368.75**	***-305.432***
α-D-Lyxose			78.2 %	0	-1503726.645	
β-D-Lyxose			21.8 %	0	-1503723.478	
D-Lyxose (avg)	**$C_5H_{10}O_5$**		**100.0 %**	**0**	**-1503725.955**	**-746.007**
(R)-Malate	**$C_4H_6O_5$**	**5.13**	**100.0 %**	**-2**	**-1395219.51**	**-848.097**
(S)-Malate	**$C_4H_6O_5$**	**5.13**	**100.0 %**	**-2**	**-1395218.699**	**-848.097**
Maleate	**$C_4H_4O_4$**	**5.61**	**100.0 %**	**-2**	**-1194469.47**	***-608.354***
α-D-Mannose			0.8 %	0	-1804456.407	
β-D-Mannose			99.2 %	0	-1804468.36	
D-Mannose (avg)	**$C_6H_{12}O_6$**		**100.0 %**	**0**	**-1804468.265**	**-895.794**
α-D-Mannose-1-P (-1)			24.03 %	-1	-3294099.27	
α-D-Mannose-1-P (-2)		6.5	75.97 %	-2	-3292883.401	
β-D-Mannose-1-P (-1)			24.03 %	-1	-3294116.867	
β-D-Mannose-1-P (-2)		6.5	75.97 %	-2	-3292881.343	
α-D-Mannose-1-P (avg)			**25.4 %**	**-1.76**	**-3293175.517**	
β-D-Mannose-1-P (avg)			**74.6 %**	**-1.76**	**-3293178.182**	
D-Mannose-1-P (avg)	**$C_6H_{13}O_9P$**		**100.0 %**	**-1.76**	**-3293177.504**	**-887.426**
α-D-Mannose-6-P (-1)			20.83 %	-1	-3294124.638	
α-D-Mannose-6-P (-2)		6.42	79.17 %	-2	-3292881.229	
β-D-Mannose-6-P (-1)			20.83 %	-1	-3294118.707	
β-D-Mannose-6-P (-2)		6.42	79.17 %	-2	-3292857.92	
α-D-Mannose-6-P (avg)			**100.0 %**	**-1.79**	**-3293140.17**	
β-D-Mannose-6-P (avg)			**0.0 %**	**-1.79**	**-3293120.48**	
D-Mannose-6-P (avg)	**$C_6H_{13}O_9P$**		**100.0 %**	**-1.79**	**-3293140.163**	**-1009.599**
(R)-2-Methylmalate	**$C_5H_9NO_4$**	**5.3**	**100.0 %**	**-2**	**-1498379**	***-842.239***
2-Methylmaleate	**$C_5H_6O_4$**	**5.79**	**100.0 %**	**-2**	**-1297643.67**	***-600.822***
Oxaloacetate	**$C_4H_4O_5$**		**100.0 %**	**-2**	**-1392090.644**	**-807.930**
3-Oxobutanoate	**$C_4H_6O_3$**		**100.0 %**	**-1**	**-1001213.031**	***-473.629***
2-Oxoglutarate	**$C_5H_4O_5$**		**100.0 %**	**-2**	**-1495257.82**	**-800.818**
P_i (-1)			38.69 %	-1	-1690461.2	
P_i (-2)		6.8	61.31 %	-2	-1689225.2	
P_i (avg)	**H_3O_4P**		**100.0 %**	**-1.61**	**-1689703.4**	**-249.366**
Phenylacetic acid	**$C_8H_8O_2$**		**100.0 %**	**-1**	**-1206987.4**	***-210.874***
Phenylacetylglycine	**$C_{10}H_{11}NO_3$**		**100.0 %**	**-1**	**-1753203.4**	***-329.281***
PP_i (-2)			31.6 %	-2	-3180115.962	
PP_i (-3)		6.68	68.4 %	-3	-3178901.449	
PP_i (avg)	**$H_4O_7P_2$**		**100.0 %**	**-2.68**	**-3179285.197**	**-240.162**
2-Propanol	**C_3H_8O**		**100.0 %**	**0**	**-510255.6653**	***-186.606***
Pyruvate	**$C_3H_4O_3$**		**100.0 %**	**-1**	**-898017.0534**	**-480.323**

Metabolite	Formula[1)]	pKa[2)]	% distrib	Charge	ΔG°_{tot} [3)] (kJ/mol)	ΔG_f°, [4)] (kJ/mol)
Ribitol	**$C_5H_{12}O_5$**		**100.0 %**	**0**	**-1506852.94**	**-790.776**
α-D-Ribose			76.0 %	0	-1503719.2	
β-D-Ribose			24.0 %	0	-1503716.4	
D-Ribose (avg)	**$C_5H_{10}O_5$**		**100.0 %**	**0**	**-1503718.5**	**-753.538**
α-D-Ribose-1-P (-1)			32.88 %	-1	-2993374.082	
α-D-Ribose-1-P (-2)		*6.69*	67.12 %	-2	-2992157.999	
α-D-Ribose-1-P (avg)	**$C_5H_{11}O_8P$**		**100.0 %**		**-2992557.798**	**-745.589**
α-D-Ribose-5-P (-1)			37.22 %	-1	-2993389.029	
α-D-Ribose-5-P (-2)		6.77	62.78 %	-2	-2992162.053	
β-D-Ribose-5-P (-1)			37.22 %	-1	-2993390.508	
β-D-Ribose-5-P (-2)		6.77	62.78 %	-2	-2992162.624	
α-D-Ribose-5-P (avg)			**40.9 %**	**-1.63**	**-2992618.733**	
β-D-Ribose-5-P (avg)			**59.1 %**	**-1.63**	**-2992619.642**	
D-Ribose-5-P (avg)	**$C_5H_{11}O_8P$**		**100.0 %**	**-1.63**	**-2992619.27**	**-754.375**
α-L-Ribulose			44.5 %	0	-1503712.258	
β-L-Ribulose			55.5 %	0	-1503712.805	
L-Ribulose (avg)	**$C_5H_{10}O_5$**		**100.0 %**	**0**	**-1503712.562**	**-746.844**
D-Ribulose-5-P (-1)			37.22 %	-1	-2993383.592	
D-Ribulose-5-P (-2)		6.77	62.78 %	-2	-2992146.426	
D-Ribulose-5-P (avg)	**$C_5H_{11}O_8P$**		**100.0 %**	**-1.63**	**-2992606.899**	**-751.446**
α-D Ribulose			28.8 %	0	-1503709.198	
β-D-Ribulose			71.2 %	0	-1503711.436	
D-Ribulose (avg)	**$C_5H_{10}O_5$**		**100.0 %**	**0**	**-1503710.791**	**-746.844**
L-Serine	**$C_3H_7NO_3$**		**100.0 %**	**0**	**-1047680.222**	**-517.561**
Uridine	**$C_9H_{12}N_2O_6$**		**100.0 %**	**0**	**-2392276.6**	***-784.918***
Water	**H_2O**		**100.0 %**	**0**	**-200762.7994**	**-236.814**
L-Xylitol	**$C_5H_{12}O_5$**		**100.0 %**	**0**	**-1506849.066**	**-790.776**
α-L-Xylulose			2.7 %	0	-1503706.46	
β-L-Xylulose			97.3 %	0	-1503715.303	
L-Xylulose (avg)	**$C_5H_{10}O_5$**		**100.0 %**	**0**	**-1503715.06**	**-746.844**
α-D-Xylulose			6.0 %	0	-1503713.276	
β-D-Xylulose			94.0 %	0	-1503720.092	
D-Xylulose (avg)	**$C_5H_{10}O_5$**		**100.0 %**	**0**	**-1503719.682**	**-746.844**

1) Molecular formula refers to uncharged species.

2) Values in *italics* obtained from Alberty R.A. (2006) *Applications of Mathematica, Wiley New Jersey,* p.235, all others were calculated using *MarvinSketch* (Chapt. 3.2.5).

3) Total standard Gibbs free energies were evaluated by DFT using bl3yp/6-311g++(d,p) basis set and include electrostatic and non-electrostatic solvation energies calculated by COSMO continuum model [57].

4) Standard transformed Gibbs free energies of formation in *italics* were obtained from Mavrovouniotis [68][69], all other values were calculated using *Gibbspredictor* (Chapt. 6.2.3) .

5) The ΔG°_{tot} of ATP^{-4} was used in the calculations.

7 REFERENCES

[1] Adams M. (1970) “Structure of lactate dehydrogenase at 2.8 Å resolution” *Nature* 227: 1098-1103.

[2] Altschul S.F., Madden T.L., Schaeffer A.A., Zhang J., Zhang Z., Miller W. and Lippmann D.J. (1997) „Gapped BLAST and PSI-BLAST: a new generation of protein database search programs“ *Nucl Acid Res. 25*:3389-3402.

[3] Anfinsen C. (1972). "The formation and stabilization of protein structure" *Biochem. J.* 128 (4): 737-49.

[4] Barone V. (2004) J*. Chem. Phys*. 120: 3059.

[5] Boys S.F. (1950) *Proc. R. Soc. London Ser. A* 200: 542.

[6] Bragg W.L. (1914) "The Diffraction of Short Electromagnetic Waves by a Crystal", *Proc. Cambridge Phil Soc* 17:43-57.

[7] Brandts J.F. (1975) „Consideration of the posibility that the slow step in protein denaturation reactions is due to *cis-trans* isomerism of proline residues” *Biochemistry* 14:4953-4963.

[8] Bruccoleri R.E. and Karplus M. (1987) „Prediction of the folding of short polypeptide segments by uniform conformational sampling” Biopolymers 26:137-168.

[9] Chan D.C. and Kim P.S. (1998) “HIV entry and its inhibition” *Cell* 93(5):681-4.

[10] Chotia C. and Lesk A.M. (1986) "The relationship between the divergence of sequence and structure in proteins" *EMBO J.* 5(4):823-826.

[11] DaoPin S. (1987) „Use of site-directed mutagenesis to obtain isomorphous heavy-atom derivatives for protein crystallography; cysteine containing mutants of phage T4 lysozyme *Protein Engin*. 1:115-123.

[12] Deane C.M. and Blundell T.L. (2000) "A novel exhaustive exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. *Proteins* 40:135-144.

[13] Diamond R. (1988) "A note on the rotational superposition problem" *Acta Cryst. A* 44:211-216.

[14] Edman P. and Begg G. (1967) "A protein sequenator" *Eur. J. Biochem*. 1:80-91.

[15] Eklund H. (1976) "Three dimensional structure of horse liver alcohol dehydrogenase at 2.4 Å resolution" *J Mol Biol.* 102: 27-59.

[16] Fernandez-Fuentes N. and Fiser A. (2006) „Saturating representation of loop conformational fragments in structure databanks" *BMC Struct Biol*. 6:15

[17] Fischer G., Bang H., Mech C. (1984) „Nachweis einer Ensymkatalase für die cis-trans-Isomerisierung der Peptidbindung in prolinhaltigen Peptiden" *Biomed Bioim Acta* 43(10):1101-11.

[18] Fishman, G.S., (1995) *Monte Carlo: Concepts, Algorithms, and Applications*, Springer Verlag, New York.

[19] Fraser R.D.B. (1979) *J. Mol Biol*. 129: 463-481.

[20] Garcia-Viloca M.,Gao J., Karplus M. and Truhlar D.G. (2004) "How Enzymes Work: Analysis by Modern Rate Theory and Computer Simulations" *Science* 303: 186 – 195.

[21] Goldberg R. N,. Tewari Y. B, Bell D., Fazio K., and Anderson E. (1993) *J. Phys. Chem. Ref. Data* 22 : 515.

[22] Goldberg RN, Tewari YB, Bhat TN (2004), "Thermodynamics of Enzyme-Catalyzed Reactions -a Database for Quantitative Biochemistry", *Bioinformatics* 20(16):2874-2877.

[23] Goldenberg D.P. (1988) "Genetic studies of protein stability and mechanisms of folding" *Ann. Rev. Biophys Chem*. 17:481-507.

[24] Halgren T (1995) "Merck molecular force field: Basis, form, scope, parameterization, and performance of MMFF94" *J. Comp. Chem* 17:490-519.

[25] Harrison S.C. and Durbin R. (1985) „The jigsaw puzzle model" *Proc. Natl. Acad Sci* 82:4028-4030.

[26] Hendrickson W.A. (1990) "Selenomethionyl proteins produced for analysis by multi-wavelength anomalous diffraction (MAD): a vehicle for direct determination of 3D structure" *EMBO J.* 9:1665-1672.

[27] Hill R.L. (1965) "Hydolysis of Proteins" *Adv. Protein Chem*. 20:37-107.

[28] Hinz H.J. (1986) "Thermodynamic Data for Biochemistry and Biotechnology", *Springer Verlag, New York*.

[29] Holm L., Ouzounis C., Sander C., Tuparev G., Vriend G. (1992). "A database of protein structure families with common folding motifs." *Protein Science* 1, 1691-1698.

[30] http://cathwww.biochem.ucl.ac.uk/latest/index.html

[31] http://compuchem.de/gaussvw.htm

[32] http://en.wikipedia.org/wiki/Charge-coupled_device

[33] http://scop.mrc-lmb.cam.ac.uk/scop/

[34] http://www.brenda-enzymes.info/

[35] http://www.chem.qmul.ac.uk/iubmb/enzyme/

[36] http://www.chemaxon.com/

[37] http://www.ebi.ac.uk/dali/fssp/fssp.html

[38] http://www.ebi.ac.uk/trembl/

[39] http://www.gaussian.com/

[40] http://www.genome.jp/kegg/

[41] http://www.hnb-cologne.uni-koeln.de/gibbspredictor/gibbspredictor.html

[42] http://www.iubmb.unibe.ch/

[43] http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html&tb=false

[44] http://www.rcsb.org/pdb/statistics/clusterStatistics.do

[45] http://www.rcsb.org/pdb/statistics/holdings.do

[46] http://www.uni-koeln.de/rrzk/

[47] http://www.wavefun.com/products/spartan.html

[48] http://www.wwpdb.org/
J. Chim. Phys. 65: 44-45.

[49] Kabsch W, Sander C. (1983) "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." *Biopolymers.* 22:2577–2637.

[50] Kabsch W. and Sander C. (1983) "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features" *Biopolymers* 22:2577-2637.

[51] Karplus M. (1959). "Contact Electron-Spin Coupling of Nuclear Magnetic Moments". *J. Chem. Phys.* 30 (1): 11-15.

[52] Kato S. (1982) „Identification and characterization of the direct folding process of hen egg-white lysozyme" *Biochemistry* 21:38-43.

[53] Kayed R. "Common Structure of Soluble Amyloid Oligomers implies Common Mechanism of Pathogenesis" *Science* 300 : 486-489.

[54] Kelly C., Cramer C. and Truhlar D.G. (2005) *J Chem Theor Comput* 1:1133-1152.

[55] Kelly, C., Cramer, C. and Truhlar D.G. (2005) *J Chem. Theory Comput* 1: 1133-1152.

[56] Kisker C., Schindelin H., Alber B.E., Ferry J.G. and Rees D.C. (1996) "A left-handed β-helix revealed by the crystal structure of a carbonic anhydrase from the achaeon *Methanosarcina thermophile*" *EMBO J.*, 15: 2323-2330.

[57] Klamt A. and Schüümann G. (1993) *J. Chem. Soc. Perkin Transactions* 2: 799-805.

[58] Kohn W. and Sham L.J. (1965) *Phys. Rev.* 140: A1133 – A 1138.

[59] Krishnan R, Binkley J.S., Seeger R.,. Pople J.A (1980) *J. Chem. Phys*. 72: 650-654.

[60] Landschulz W.H., Johnson P.F., McKnight S.L. (1988) "The leucine zipper: a hypothetical structure common to a new class of DNA-binding proteins" *Science* 240:1759-1764.

[61] Lee B., Richards F.M. (1971) „The interpretation of protein structures: estimation of static accessibility" *J Mol. Biol*. 55:379-400.

[62] Lehninger, A.E. (1986*)* "Principles of Biochemistry", *Worth, New York.*

[63] Leonor Michaelis, Maud Menten (1913). Die Kinetik der Invertinwirkung, *Biochem. Z*. 49:333-369.

[64] Lessel U. and Schomburg D. (1994) „Similarities between protein 3D structures" *Protein Eng*. 7:1175-1187.

[65] Levinthal C. (1968) "Are there pathways for protein folding?"

[66] Lundström J, Holmgren A (1990). "Protein disulfide-isomerase is a substrate for thioredoxin reductase and has thioredoxin-like activity". *J. Biol. Chem.* 265 (16): 9114–20.

[67] Masgrau L., Roujeinikova A., Johannissen L. O., Hothi P., Basran J., Ranaghan K. E., Mulholland A. J., Sutcliffe M. J., Scrutton N. S., Leys D. (2006). "Atomic Description of an Enzyme Reaction Dominated by Proton Tunneling". *Science* 312 (5771): 237–241.

[68] Mavrovouniotis, M.L. (1990) *Biotechnol. Bioeng*. 36, 1070-1082.

[69] Mavrovouniotis, M.L. (1991) *J. of Biol. Chem*. 266, 14440-14445.

[70] McLachlan A.D. (1979) „Gene duplication in the structural evolution of chymotrypsin" *J Mol Biol* 128:49-79.

[71] Miertus S., Scrocco E., and Tomasi J. (1981) *Chem. Phys.* 55: 117f

[72] Mizuguchi K., Deane C.M. Blundell T.L., Johnson M.S., Overingto J.P. "JOY:protein sequence-structure representation and analysis." *Bioinformatics* 14:617-623.

[73] Mommany F.A , Appell M., Willett J.L.,Schnupf U., Bosma W.B. (2006) *Carbohydr. Res.* 341: 525-537.

[74] Mommany F.A. and Appell M. (2005) *Carbohydr Res* 340:1638-1655.

[75] Nanda H., Lu N, Woolf T.B. (2005) *J Chem Phys* 122**:** 13411 f.

[76] Newcomer M.E. (1984) "The three dimensional structure of retinol-binding protein" *EMBO J*. 3:1451-1454.

[77] Okuyama K. (1981) "Crystal and molecular structure of a collagen-like polypeptide" *J. Mol. Biol*. 152: 427-443.

[78] Ponder JW and Case DA. (2003) Force fields for protein simulations. *Adv. Prot. Chem.* 66: 27-85.

[79] Privalov P.L. (1979) "Stability of proteins. Small globular proteins" *Adv. Prot Chem*. 33:167-241.

[80] Ptitsyn O.B. (1987) „Protein folding:hypotheses and experiments" *J. Protein Chem*. 6:273-293.

[81] Ptitsyn O.B. (1990) "Evidence for a molten globule state as a general intermediate in protein folding" *FEBS Letters* 262:2024.

[82] Ptitsyn O.B., and Finkelstein A.V. (1980) „Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding?" *Quart Rev Biophys* 13: 339-386.

[83] Qian B. and Goldstein R.A. (2001) "Distribution of indel length" *Proteins* 45:102-104.

[84] Ramachandran G.N. (1974) “The mean geometry of the peptide unit from crystal structure data” *Biochim Biophys Acta* 359: 298-302.

[85] Ramachandran G.N. and Mitra A.K. (1976) “An explanation for the rare occurrence of cis peptide units in proteins and polypeptides” *J. Mol. Biol*. 107: 85-92.

[86] Ramachandran G.N. and Saisekharan V. (1968) “Conformation of Polypeptides and Proteins” *Adv Protein Chem.* 23:283-437.

[87] Rao S. and Rossmann M. (1973). "Comparison of super-secondary structures in proteins" *J Mol Biol* 76 (2): 241-56.

[88] Rossman M., Moras D. and Olsen K. (1974) “Chemical and biological evolution of a nucleotide-binding protein” (1974) *Nature* 250: 194-199.

[89] Rost B. (1999) “Twilight zone of protein sequence alignments.” *Protein Eng* 12(2):85-94.

[90] Schrödinger E. (1926) "An Undulatory Theory of the Mechanics of Atoms and Molecules", *Phys. Rev*. 28: 1049.

[91] Schulz G.E. and Schirmer R.H. (1977) “Principles of Protein Structure” *Springer Verlag,* New York

[92] Slater J.C., (1930) Atomic Shielding Constants, *Phys. Rev.* 36: 57 .

[93] Smith TF and Waterman MS (1981). "Identification of Common Molecular Subsequences". *Journal of Molecular Biology* 147: 195-197

[94] Srinisavan N., Guruprasad K., Blundell T. (1995) “Comparative Modeling of proteins “ *Protein Structure Prediction: A practical approach series Sternberg (ed.):* 111-140.

[95] Steinbacher S., Seckler R., Miller S., Steipe B., Huber R. and Reinemer P. (1994) "Crystal structure of P22 tailspike protein: interdigitated subunits in a thermostable trimer" *Science*, 265:383-386.

[96] Steitz T.A. (1976) „High resolution x-ray structure of yeast hexokinase, an allosteric protein exhibiting a non-symmetric arrangement of subunits" *J Mol Biol.* 104:197-222.

[97] Steward J.P. (1989) „Optimization of parameters for semiempirical methods" *J Comp Chem* 10:209-220.

[98] Tang J. (1978) „Structural evidence for gene duplication in the evolution of the acid proteases" *Nature* 271:618-621.

[99] Tang J. and Breaker R. (1997). "Structural diversity of self-cleaving ribozymes". *Proceedings of the National Academy of Sciences* 97 (11): 5784-5789.

[100] Taylor G. (2003). "The phase problem". *Acta Crystallogr. D Biol. Crystallogr.* 59 (Pt 11): 1881-90.

[101] Usón I, Sheldrick GM (1999). "Advances in direct methods for protein crystallography". *Curr. Opin. Struct. Biol.* 9 (5): 643-8.

[102] Wagman D., Evans H., Parker V., Schumm, R., Halow I., Bailey S., Churney K. and Nuttall R. (1982) "The NBS tables of chemical thermodynamic properties" *J. Phys Chem. Ref. Data* 11 Suppl. 2.

[103] Waller I. (1923). "Zur Frage der Einwirkung der Wärmebewegung auf die Interferenz von Röntgenstrahlen". *Z. Phys.* **17**:398-408.

[104] Wetlaufer D.B. (1973) „ The nucleation, rapid-growth model" *Proc. Natl. Acad Sci* 70: 697-701.

[105] Wierenga R.K. (2001). "The TIM-barrel fold: a versatile framework for efficient enzymes" *FEBS Lett.* 492 (3): 192–198.

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat, dass sie – abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist, sowie dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Dietmar Schomburg betreut worden.

Quoc-Vu Ha Ngoc

Teilpublikationen

Wohlfahrt G., **Hangoc V**., Schomburg D. (2002) Positioning of Anchor Groups in Protein Loop Prediction: The Importance of Solvent Accessibility and Secondary Structure Elements. *Proteins*, **47**, 370-378.