OTTO-VON-GUERICKE UNIVERSITY MAGDEBURG

Visualisation Toolkit for Contact Density Potentials within Amino Acid Neighbourhoods in Protein Structures

by

Corinna Vehlow

supervised by

Prof. Dr. Bernhard Preim and Dr. Michael Lappe

A diploma thesis submitted in partial fulfillment for the degree of Graduate Engineer of Computational Visualistics

in the

Faculty of Computer Science

Department of Simulation and Graphics

October 2010

Declaration of Authorship

I, Corinna Vehlow, declare that this thesis titled, "Visualisation Toolkit for Contact Density Potentials within Amino Acid Neighbourhoods in Protein Structures" and the work presented in it are my own. I confirm that only the sources and means cited have been used. Parts that are direct quotes or paraphrases are identified as such. The same is true of tables and figures that have been used. This thesis, or any part of it has not previously been submitted for a degree or any other qualification at this University or any other institution.

Signed: _____

Date: _____

OTTO-VON-GUERICKE UNIVERSITY MAGDEBURG

Abstract

Faculty of Computer Science
Department of Simulation and Graphics

Graduate Engineer of Computational Visualistics

by Corinna Vehlow

The prediction of the folding of protein structures is of great value for the investigation of causes for and development of cancer as well as for the purpose of drug design. This thesis will introduce a developed software, which extracts and visualises orientation propensities of amino acids within proteins and discusses how to slip the derived information into the process of structure prediction. Therefore, first of all an introduction into the biological background, particularly the structure of proteins and experimental techniques used to derive a protein's folding, will be given. This will be followed by some related work regarding possible visualisation techniques. After a comprehensive description of the concept and the implemented visualisation, the latter will be evaluated with respect to its usability. Different possibilities, describing how to make use of the derived information, will be discussed in the context of sampling methods and energy functions as well as scoring functions for the process of structure prediction. Finally, a conclusion will be drawn.

Acknowledgements

Sincere thanks to Prof. Dr. B. Preim, who was supervising me and helping with words and deeds during the my studies, especially during my final project up to the present day and supported me with lots of advice.

Furthermore, I would like to thank Dr. M. Lappe, who was guiding me through the project of this thesis on site from the very first beginning. Thanks for all the advice and help in all respects and for stimulating my research. Many thanks as well to H. Stehr and M. Winkelmann, who often gave me some inspiring input and support. I would also like to thank the rest of the "Bioinformatics Research Group" of M. Lappe for their help during the validation process, particularly for the time they spent on evaluation procedure. Many thanks to Lars Petzold, who provided me with some illustrations.

Finally many thanks to my family, especially my parents, and my friends, who are always backing me with what I am doing and have confidence into my plans. Without those people I would not have become the kind of person I am.

Contents

Abbreviations

AA	**A**mino **A**cid
BB	**B**ack**b**one
CASP	**C**ritical **A**ssessment of Protein **S**tructure **P**rediction
CATH	**C**lassification of
CCD	**C**harged **C**oupled **D**evice
COSY	**Co**rrelation **S**pectroscop**y**
DBSCAN	**D**ensity **B**ased **S**patial **C**lustering of **A**pplications with **N**oise
DNA	**D**eoxyribo**n**ucleic **a**cid
DOI	**D**egree **o**f **I**nterest
EDM	**E**lectron **D**ensity **M**ap
GDT	**G**lobal **D**istance **T**est
LOS	**L**og-**O**dds-**S**core
MAD	**M**ultiwavelength **A**nomalous **D**iffraction
MIR	**M**ultiple **I**somorphous **R**eplacement
MMDS	**M**etric **M**ulti-**d**imensional **S**caling
NBH	**N**eigh**b**our**h**ood
NMR	**N**uclear **M**agnetic **R**esonance
NOE	**N**uclear **O**verhauser **E**ffect
PDB	**P**rotein **D**ata **B**ank
RCC	**R**ank **C**orrelation **C**oefficient
RF	**R**adio **F**requency
RMSD	**R**oot **M**ean **S**quare **D**eviation
SC	**S**ide**c**hain
SCOP	**S**tructural **C**lassification **o**f **P**roteins
SS	**S**econdary **S**tructure
SSE	**S**econdary **S**tructure **E**lement
SHS	**S**pherical **H**armonic **S**ynthesis

Symbols

λ	angle (latitude)	$^\circ$
ϕ	angle (longitude)	$^\circ$
ψ	angle	$^\circ$
θ	angle	$^\circ$
Å	Ångström	
C	carbon atom	
H	hydrogen atom	
O	oxygen atom	
N	nitrogen atom	

Chapter 1

Introduction

Protein structure prediction is of high importance for various scientific areas. Chemists use information about the binding sites for the purpose of drug design. In contrast, biochemists are much more interested in the three-dimensional structure of proteins, which gives insights into its particular *cellular function*. The position of certain sequence features of a protein within the 3D structure is of substantial and practical value. The knowledge about the spatial location of certain residues within the structure can be used for site directed mutagenesis. The rate at which protein structures are solved has increased, whereas to predict the 3D structure from sequence still remains a challenge. Up to date more than 60.000 high-resolution protein structures are available in public protein data banks (PDB). For less than 1% of the known protein sequences the 3D structure was determined experimentally, i.e. with the help of X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy.

There exist different approaches for structure prediction. As lots of proteins are quite similar regarding their sequence, one approach is to use template PDB structures and alignment methods, like homology modelling or protein threading. Besides comparative protein modelling, ab initio or de novo protein modelling are used, especially for those sequences, where no template PDB structures could be found. The aim of protein modelling methods is to identify the most plausible structure, which is usually done via energy functions and potentials of mean force.

Many different approaches to derive potentials for interaction determination and structure prediction can be found in the literature. Residues of a protein's primary sequence interact with each other, if they are connected via hydrogen bonds and situated close to each other, i.e. their distance deceeds a given threshold. Most of those methods, used to determine interactions, rely on the assumption that known X-ray or NMR resolved protein structures represent classical equilibrium states. They are used to derive some kinds of standards and principles. Statistical methods, like the Boltzmann device, are used quite often (e.g. by [Sip90]) to obtain *distance dependent mean force potentials*. Many studies, like that of Tobi et al. [TE00, TSLE00] use only distance constraints.

Such distance constraints for the residues of the protein define, how far or close certain amino acids of the chain are situated from each other. These constraints are further used for structure prediction, as they give information on how the residues might be situated to each other. Nevertheless, such constraints do not allow to determine structures unambiguously. This includes e.g. the occurrence of spiegelmers, i.e. mirror-inverted 3D structures of an amino acid chain.
Another structural, quite important parameter that could be used to determine how chains fold to proteins is the relative packing of side chains within proteins. The way how a chain folds and how neighbouring residues of a chain are placed to one another can be described through two angles (phi and psi). Those dihedral angles pose an important determinant of local geometry, i.e. secondary structure, as well as three-dimensional topology, i.e. tertiary structure. Bahar and Jernigan were the first, who captured the existence of relative orientation probabilities [BJ96]. To show that the *preferred orientation of side chains depends* on the nature of the residue, they used a simple body fixed coordinate system. Their results implied that statistical potentials are sensitive to orientation.
The backbone (BB) atoms of amino acids (AA) itself are relatively fixed within a plane, whereas the attached side chain (SC) can have different orientations. Neighboured BBs of the amino acid sequence are connected over a certain angle to each other. The way, how certain AAs are placed to one another, is defined by the dihedral angle pair mentioned above. Those angles for all involved contacts define the tertiary structure of the protein. The aim in terms of protein fold recognition and structure prediction is to obtain accurate residue-dependent potentials of spatial arrangement. Besides those angles between BB elements, also the rotamers of attached side chains are of interest. Rotamers are favoured conformations, i.e. orientations of hydrogen bonds between atoms of side chains that occur more frequently as they are energetically more favourable [IntroProteinStructure].
Summarising, the analysis of contacts and side chain packing, offering distance and *orientation constraints* supports the development of knowledge-based potentials. To overcome ambiguities in structure prediction, which occur when using distance constraints alone, the definition of at least some orientation constraints between residues might help.

1.1 Objective

The aim of this project is to develop a graphical support to detect spatially preferred orientations of specific residue contacts. Based on the data analysis of available resolved protein structures, the potential spatial distribution of residue contacts can be extracted. This includes the alignment of contacts along an orientation and translation invariant framework.
The visualisations should highlight preferred dihedral angles and thus help to define

several orientation constraints for a subset of the residue contacts. Together with the distance constraints those orientation constraints could be handed over to the reconstruction method, which will hopefully produce more precise models of the 3D structure. In the best case, the problem of spiegelmers will be removed, thus giving the correct of the two possible structures. The definition of orientation constraints might also help to analyse the side chain packing, i.e. how good the side chains are intermeshed. In the long run, it could expand the conditions for scoring functions that are used to assess models, produced by different prediction methods.

The following paragraphs will give an overview about the biological background of this thesis. The structure of proteins will be discussed, after giving an introduction about amino acids and its properties. This will be followed by a short résumé about the two experimental techniques, X-ray and NMR spectroscopy, used for structure prediction. Afterwards, related work in the field of visualisation of protein structures will be summarised and discussed. This will be complemented through the discussion of various general visualisation techniques, which could be used for the aim of this thesis. Based on the introduced visualisation possibilities the concept will be presented and discussed. Afterwards the implementation will be presented and evaluated, followed by a discussion about future perspectives and potential improvements.

Chapter 2

Biological Background

This chapter explains the structure of proteins and methods to determine those. As proteins are build of amino acids, these will be introduced first. The structure of a protein will be explained in different structural levels, including the primary, secondary, tertiary and quaternary structure. Afterwards the two main experimental procedures to obtain a protein's three-dimensional structure will be introduced and compared. Finally, the processing of such experimental data will be explained.

2.1 Amino Acids

There are twenty different amino acids, which are the building blocks of protein structures [BT99]. They can be combined in manifold ways into more proteins of variable length (average length of a protein is around 300 amino acids). In comparison, the DNA comprises only four different nucleic acids as basic units, which differ from those of proteins. All amino acids are composed of a *main chain*, which is the same for every AA, and a *side chain* (amino acid residue). The latter is different for each AA and thus defines its properties. The main chain is built of the central carbon atom (C_α), to which attached are a hydrogen atom (H), an amino group (NH_2), a carboxyl group ($COOH$) (see Figure 2.1(a)), as well as the side chain.

Amino acids form chains via peptide bonds, i.e. the neighboured units are connected via a peptide bond between the C'-atom of one residue and the nitrogen atom of the next residue (see Figure 2.1(b)). Thus, through the peptide bonds the main chain atoms form the base $NH{-}C_\alpha H\text{-}C'{=}O$. The chain of these amino acid bases is called *backbone* of the protein.

Respective their chemical properties, amino acids can be divided into three main classes: *hydrophobic*, *charged* and *polar* amino acids. Amino acids can be ranked respectively by their hydrophobicity, i.e. their capacity of reaction within water. To date, about 56 different hydrophobicity scales have been suggested. Other differentiations are e.g.

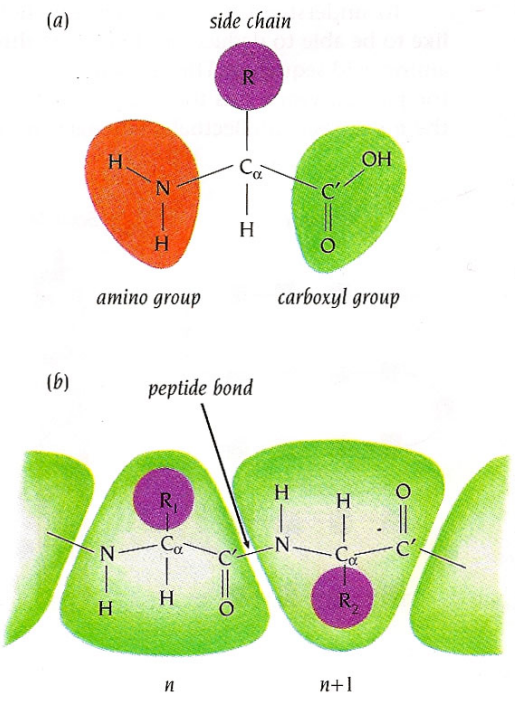


FIGURE 2.1: Schematic diagram of an amino acid (a) and polypeptide chain (b) [BT99, p.4].

aromatic vs. non-aromatic ring systems, aliphatic vs. non-aliphatic AAs, small vs. tiny AAs or in case they are charged positive vs. negative AAs. An amino acid is aliphatic, if its side chain contains only carbon or hydrogen atoms. An amino acid is aromatic, if it contains an aromatic ring system.

2.2 Protein Structures

The *primary structure* of a protein is defined by its amino acid sequence, i.e. the types and order of occurring amino acids within the protein's polypeptide chain [ZB07][BT99]. It can therefore be handled as simple string of 1D single letter coded amino acids. The *secondary structure* (SS) describes the local shape of the protein structure (see Figure 2.2 including different schematic views of common protein secondary structure elements). Certain parts of protein chains fold to form generic structures, which can be found in all proteins. This includes *alpha-helices* and *beta-sheets*. The secondary structure is to some extent determined by the amino acid sequence. There are certain amino acid sequences that favour either alpha helices or beta strands, whereas others prefer the formation of loop regions.
Secondary structure elements (SSEs) arrange themselves in simple *motifs*, i.e. a series of helices and/or sheets connected via loops. Some motifs are associated with a specific

function, whereas others do not have a specific biological function alone but contribute to larger structural and functional assemblies. Some of the frequently occurring motifs are the helix-turn-helix, calcium binding, helix-loop-helix, hairpin beta, greek key and beta-alpha-beta motif. Motifs are formed through the dense packing of side chains of adjacent alpha helices and beta strands. Further, they can be combined to more complex motifs and in formations with other motifs build up globular structures, called domains. *Domains* are folding-based units of function, i.e. they represent fundamental units of the *tertiary structure*. Such parts of a polypeptide chain can fold independently into stable tertiary structure and are often associated with different functions. On the other hand, similar domain structures frequently occur in different proteins with different functions and completely different amino acid sequence. Proteins may include a single up to several dozen domains. Domain structures can be classified into alpha domains, beta domains and alpha-beta-domains.
The intrinsic physical and chemical properties of proteins give rise to some striking regularities, e.g. in way of assembling secondary structures and in the topologies of polypeptide chains. These regularities provide a good basis to classify proteins respective the way they fold. An extensive description of such structural as well as evolutionary relationships of proteins of known protein structure is supplied by the SCOP (structural classification of proteins) database [MBHC95]. It includes the possibility to search for homologue sequences, that have a significant level of sequence similarity compared to a given sequence. As the number of known structures increases rapidly with more than 100 new structures each month, it becomes impracticable to classify the proteins manually. Orengo et al. [OMJ+97] introduced a semi-automatic procedure to derive a novel hierarchical classification of protein domain structures (CATH). The classification is thereby based on four main levels: protein class, architecture, topology and homologous superfamily.
Further folding and packing together of such secondary structure elements and motifs is described by the tertiary structure. The arrangement of atoms in the third dimension considers all physical interactions within the molecule and with its surrounding. Protein molecules that fold from a single chain are called monomeric proteins. Functional proteins are often formed by more than one protein chain. They are therefore called multimeric molecules. The *quaternary structure* describes the composition and arrangement of such individual chains, called protein subunits, to multi-subunit proteins.

2.2.1 Secondary Structure

The most frequent secondary structure elements are *alpha-helices* and beta strands. The latter can be further divided into *parallel* and *anti-parallel beta sheets* [ZB07][BT99].
The alpha helix is one of the classical elements of protein structures, first described in 1951 by Linus Pauling. With respect to the Ramachandran plot it is found in proteins, when a stretch of consecutive residues all have the phi-psi (ϕ - ψ) pair approximately

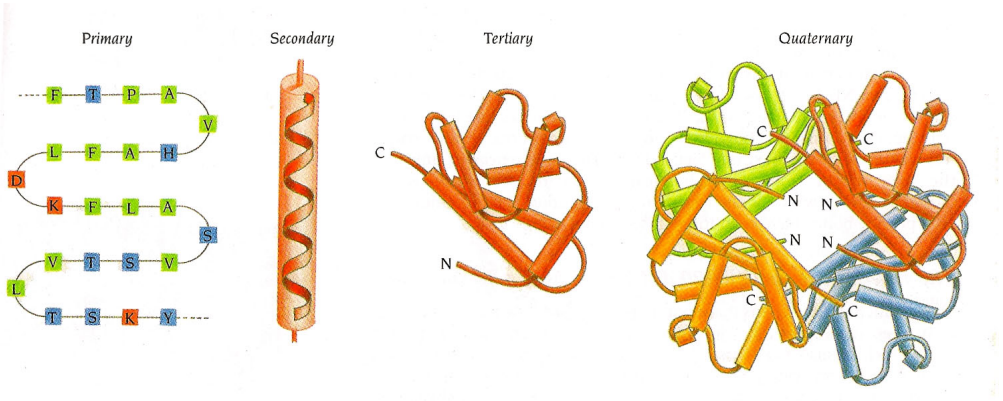


FIGURE 2.2: Schematic views of primary, secondary, tertiary and quaternary structure [BT99, p.3].

-60°to -50°. Alpha helices contain on average 3.6 residues per turn. Hydrogen bonds between the $C'{=}O$ group of residue n and NH-group of residue $n+4$ stabilise the helices. Such bonds exist for all NH and $C'{=}O$ groups, except the first and the last one. This is why the ends of helices are polar and almost always at the surface of a protein molecule. Sometimes the alpha helix is more loosely or more tightly coiled. In this case the hydrogen bonds are stretched from residue n to $n+3$ to $n+5$. Alpha-helices vary considerably in length, especially in globular proteins, from five to forty residues (ten residues on average). Although some side chains have weak but definite preference either for or against being in a helix, this preference is not strong enough to give an accurate prediction.

Whereas helices build up from one continuous region, beta sheets build up from a combination of several regions of the polypeptide chain. These regions, called beta strands, are generally five to ten residues long. They are aligned adjacent to each other and connected over hydrogen bonds between $C'{=}O$ groups of one beta strand and NH groups of an adjacent beta strand and vice versa. Beta sheets can be formed of several such beta strands in two ways: parallel or anti-parallel. In parallel beta-sheets the amino acids within the aligned strands all run into the same biological direction, from amino to carboxy terminal. In contrast, the amino acids in successive strands of anti-parallel beta sheets have alternating directions, from amino to carboxy terminal followed by carboxy to amino terminal and so on. Beta strands can also be combined into mixed beta sheets, with some parallel and some anti-parallel beta strands.

Most protein structures build up from combinations of *SS elements*, i.e. alpha-helices and beta-sheets, which are *connected by loop regions* of various length and irregular shape. These loop regions are positioned at the surface of a molecule and occur in various forms. The preferred loop structures are hairpin loops, short hairpin loops (reverse or simple turns) and long loop regions. Hairpin loops mainly connect two adjacent anti-parallel beta-strands. Long loop regions have an influence on the function of the protein and might therefore switch between open and closed conformations.

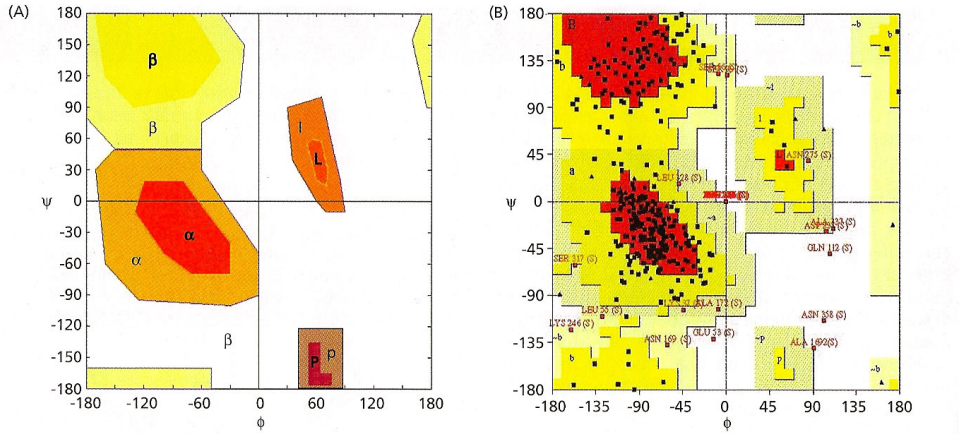


FIGURE 2.3: Ramachandran plots of sterically allowed regions for different secondary structure types [ZB07, p.34]. The angle ϕ is thereby plotted on the x and ψ on the y-axis. *(A)* An ideal R. plot, showing preferred conformations for β-sheet residues, α-helices, left-handed helices (L) and the epsilon structure (P). The darker the colour, the more favourable the conformation. *(B)* A R. plot calculated for a model of a kinase enzyme.

2.2.2 Tertiary Structure

The tertiary native structure is stabilised by a complex network of non-covalent interactions, hold by disulfide bridges [BT99] [ZB07]. Sometimes such disulfide bridges even hold different polypeptide chains together. The polypeptide main chain can be divided into repeating peptide units (one for each amino acid) running from one C_α to the next C_α. Thus, every C_α-atom, besides the first and last one, belongs to two of such units. This partitioning helps to describe the spatial orientation of neighboured chain elements to each other, i.e. the folding. All atoms of such a unit (NH and C'=O), excluding the side chain, are fixed within a plane with similar bond length and angles. The peptide units (rigid groups) are linked into a chain by covalent bonds attached to the C_α-atom (see Figure 2.1(b)). The rotation around those bonds (*phi* (ϕ) angle around C_α-N and *psi* (ψ) angle around C_α-C') poses the only degrees of freedom respective the spatial orientation. Consequently, each amino acid residue is associated with two such conformational angles ϕ and ψ. Thus, the folding of the whole main chain depends on all single angles (for all residues of the sequence).
However, those angles can have values of restricted ranges. The reason therefore are steric collisions between the main chain and side chain, attached to the Calpha atom, or between atoms in different peptide groups, which might occur otherwise. Atoms within molecules occupy a certain amount of space, which leads to steric effects in case they are brought too close together. If the electron clouds of different atoms overlap, this may change the preferred shape and reactivity of a molecule.
The Ramachandran plot (developed by the Indian biophysicist G. N. Ramachandran, who first made calculations of sterically allowed regions) shows the angles ϕ and ψ plotted

against each other (see Figure 2.3) [ZB07]. The only exception is constituted by Glycine, which can adopt to a much wider range of conformations. As the side chain only contains a single hydrogen atom, this amino acid is much more flexible and thus plays an important role, allowing unusual main-chain conformations. The Ramachandran plot shows *sterically allowed regions for conformational angles* for right-handed alpha helices, beta-strands of parallel and anti-parallel beta-sheets and left-handed alpha-helices. There exist several different conformations, but some are energetically more favourable than others. The most energetically favoured arrangement of two tetrahedrally carbon atoms are the so-called staggered (not-aligned) conformations. Most side chains have one or few conformations that occur more frequently than the other possible staggered conformations. The existence of such rotamers was exposed through the analysis of already accurately determined protein structures. Collections of such favourable conformations are saved within different rotamer libraries.
Amino acids, situated in the interior of a protein, have almost exclusively hydrophobic side chains. The driving force for folding water-soluble globular protein molecules leads to a hydrophobic core and hydrophilic surface. Consequently, the distribution of hydrophobic and hydrophilic residues plays an important role in structure prediction, as it gives information on possible spatial distributions and orientations. The hydrophobic side chains are packed quite densely inside the interior of the molecule. This is achieved through the formation of regular SS, i.e. alpha-helices and beta-sheet, within the core. These SS elements build a rigid stable framework, leading to relative little flexibility of amino acids with respect to each other. This restriction makes them the best defined parts of protein structures, determined by X-ray and NMR techniques. The functional groups of the peptide chain are mostly attached to the core in loop regions that connect sequentially adjacent secondary structure elements.

2.2.3 Protein Structure Prediction

Based on the sequencing of the human genome, biologists get a wider range of opportunities to understand the molecular basis of physiological processes and disease states. The necessary level of detail is provided by the *three-dimensional structure*, which is unique to a protein. Although the structure and function of a protein are determined completely by its amino acid sequence, the folded structure cannot be predicted from sequence alone [BT99] [ZB07]. This is because the rules for folding are not yet understood. In fact, the structure has to be *determined experimentally* with the help of techniques like, *X-ray crystallography* or *nuclear magnetic resonance techniques*. Both techniques will be explained in more detail within the next subchapters.
The knowledge about a protein's tertiary structure is a prerequisite for a proper understanding and engineering of its function. From so far determined amino acid sequences (which are more than 500.000) more than 6.000 have been solved within the past 30

years. But although recent significant technology advances have been made, the experimental determination of tertiary structure is still slow in comparison to the rate with that sequence data is accumulated, i.e. the structure determination lags behind protein sequence determination. The degree of whole structure coverage of the human proteome (genome), based on experimental structures, is about 25% [XB05]. Hence, the *folding problem* is the major *unsolved* problem in structural molecular biology and thus central to a rapid progress in post-genomic biology. There are different approaches to derive the tertiary structure: *ab initio modelling methods* and *comparative modelling methods* (homology modelling or protein threading). *Homology modelling* is often used, as there are lots of homologous proteins, i.e. proteins that have a quite similar structure and function. Homology in terms of protein structures means the similarity of structure, physiology, development and evolution of organisms based upon common genetic factors. Homologous proteins have conserved structural cores, i.e. they always contain a similar core region, and variable loop regions. Those core regions contain mainly SS elements that are built up in the interior of the protein. Conditioned by the similarity in sequence, homologous proteins also have a similar three-dimensional structure, whereby the greater the sequence identity, the more closely related the scaffold structure is.
A basic requirement for structure prediction is the knowledge of secondary structure. But even if the composition of SS elements is known, there are lots of possible conformations that can occur. The aim of structure prediction is to extract those possible conformations, which are free of collisions and most likely, as they need the lowest energy to remain stable. There are always several of those stable possible conformations among which the structure changes. Consequently, the three-dimensional structure is not consistent, but wobbles around within a certain degree of possible motion (molecular motion). Not just similar, but also unrelated sequences may have the same folds, giving a similar three-dimensional structure.
To derive how certain proteins are folded in the third dimension, still remains a challenge but that is at the same time an important and necessary step to define a protein's function. Particularly, which residues, i.e. which part of the primary sequence, lie(s) within the core of the 3D structure or at the surface, gives clues about the function of the protein.

2.3 X-ray Crystallography and Nuclear Magnetic Resonance

The primary structure of proteins can be obtained by biochemical methods either via direct determination of the AA sequence from the protein or indirectly from the nucleotide sequence of the corresponding gene (cDNA). The quaternary structure of large proteins and virus particles etc. can be determined with the help of electron microscopy. These methods are relatively easy and straight forward but offer structured information with

low resolution. To get more detailed information of the arrangement of atoms within the protein, it is necessary to predict the secondary and tertiary structure. The former can be predicted with 80% certainty from the primary sequence. Nuclear magnetic resonance techniques or X-ray crystallography are often used additionally to increase the degree of certainty [BT99].
The structure of individual proteins can be measured experimentally via NMR or X-ray crystallography. The latter was used for about 86% of the derived protein structures. As NMR techniques have been developed in recent years and are only suitable for small molecules, this technique makes up a relative small part of about 13%. The remaining 1% was derived from electron microscopy, which will not be further investigated [OGF+10].

2.3.1 X-ray Crystallography

Crystallography describes the determination of atomic structure of a crystal through the *diffraction of certain radiation at the crystal grid*. The most often used type of radiation is monochromatic X-ray radiation. The oldest and most precise method is single-crystal X-ray crystallography. In this process a beam of X-ray strikes a single crystal, thereby producing scattered beams, i.e. the X-ray beam diffracts into many specific directions (see Figure 2.4). These arising patterns of spots (reflections) can be observed on a screen behind the crystal. During X-ray crystallography not the position of atoms, but the distribution of electrons within a unit cell is determined, as the electrons interact with the radiation. X-ray crystallography thus *produces electron density maps*.
The average resolution of X-ray crystallography is very high with 2.5-1.2Å [Rho06].
The procedure of X-ray crystallography can be divided into three main steps:

- Crystal Generation
- Striking and Recording
- Data Analysis

Crystal Generation The first difficult step is to obtain an adequate crystal of the material of interest. The crystal should be sufficiently large, pure in composition (monocrystal) and regular in structure (homogenous). If these factors are given, the structure can be derived with a resolution within a few Ångström (Å).

Striking and Recording For the measurement procedure, the crystal is placed on a rotatable bail within an intense beam of X-rays. In most cases monochromatic X-rays of a single wavelength are used. Hence, the X-ray beam is filtered first and further collimated to a single direction, before it strikes the crystal. The brightest and most

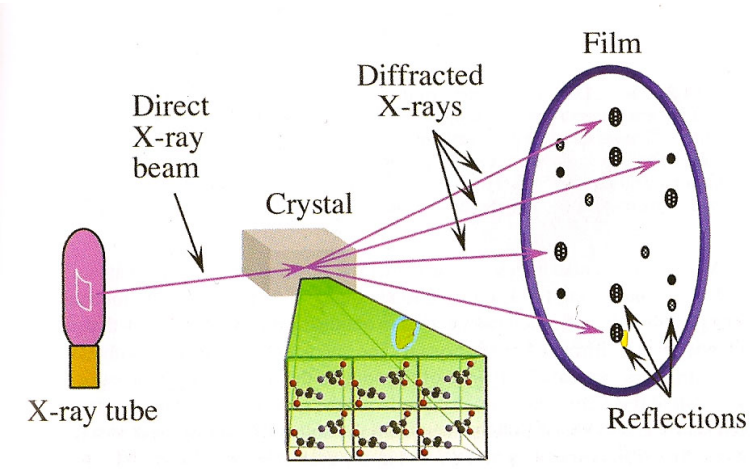


FIGURE 2.4: Schematic setup of X-ray crystallography [Rho06, p.13]. The X-ray beam diffracts at the protein crystal into many different directions, thereby producing distinct reflections on the film.

powerful X-ray sources are synchrotrons, which have much higher luminosity and thus offering a better resolution than other sources.
The resulting diffraction patterns (see Figure 2.5) of regularly spaced spots (reflections) are produced for different directions, as one image is insufficient to derive the 3D structure. Therefore, the mounted crystal is rotated piece-wise. The diffraction patterns are recorded either on image planes (reusable films) or with the help of electronic detectors (area detectors of charge-coupled devices (CCDs)). When using the first one, the diffraction patterns are scanned and stored on the PC before the films are erased again. In comparison, electronic detectors (electronic films) directly transfer the patterns in digitised form to the PC. The relative intensities provide information about the arrangement of molecules in atomic detail.
The rule for diffraction is given by Bragg's law:

$$2 * d * \sin\theta = \lambda \tag{2.1}$$

where θ is the reflection angle, d the distance between planes and λ the wavelength. The geometry of unit (primitive) cells of the crystal grid can be derived completely by means of the angles for those spots, for which the maxima of diffraction occur. To calculate the crystal structure, the strength and angles of the detected spots are used.

Data Analysis The series of two-dimensional images for the different rotation angles are converted into a *three-dimensional model of electron densities via the Fourier transformation*. Each diffraction beam recorded as spot on the image is defined by three properties: amplitude, wavelength and phase. To determine the positions of atoms giving rise to the diffracted beams, all properties are necessary. The amplitude is given by the intensity of the spot and the wavelength is set by the X-ray source. Thus, the only undefined property is the phase. Phase determination poses the major problem

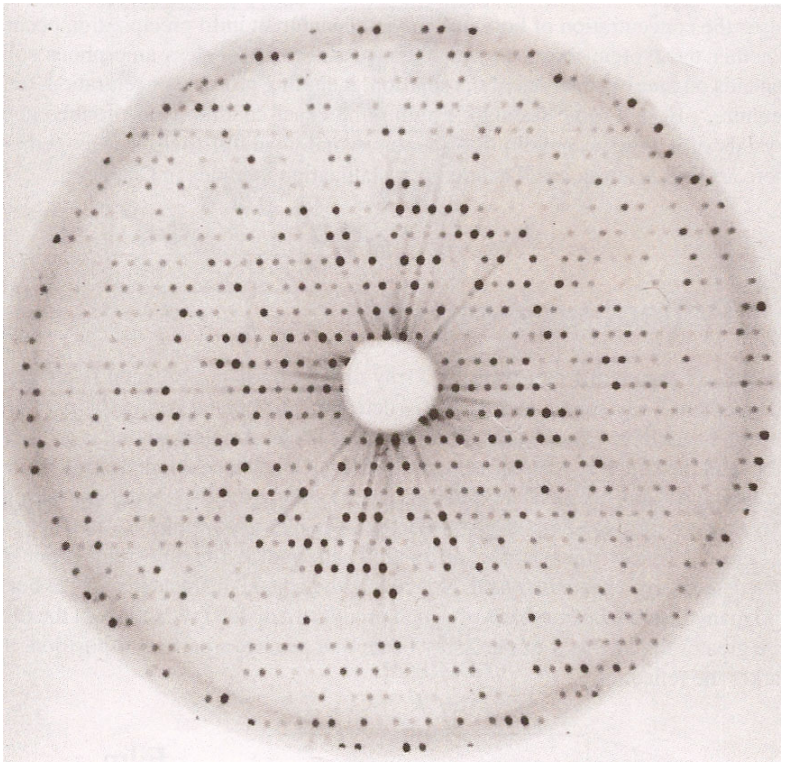


FIGURE 2.5: Diffraction pattern of X-ray crystallography, caused by diffraction of radiation at the crystal grid. Reflections are distributed in a regular pattern, whereas their intensities can vary [Rho06, p.14]. A set of such diffraction pattern recorded for different directions can be used to calculate an electron density map (via Fourier transformation).

in crystallography. The phase can be estimated via multiple isomorphous replacement (MIR) or multiwavelength anomalous diffraction (MAD) experiments.

To analyse how each spot contributes to the electron density the following steps are necessary: indexing, merging and scaling and phasing. The indexing phase is used to determine which variation corresponds to which spot. During the next step the relative strength of the spots in the different images are merged and scaled. Finally the phasing handles the problem of how the variations should be combined to yield the total electron density.

The crystal structures are finally calculated by fitting the atomic model of the molecule into the electron-density map (EDM) or isosurface. For this purpose, an initial model is determined via a trial and error process during the initial phase, i.e. a model of the primary sequence is determined that fits into the electron-density map thereby avoiding steric collisions. This initial model is further improved during the refinement phase, which removes errors. Thereby, the model is refined at some atomic positions to fit the observed diffraction data. Afterwards the model is fit to the new electron density map, followed by a further round of refinement. This *alternating process of refinement and fitting* is repeated until the correlation between the diffraction data and the model is maximised [BT99][Rho06].

2.3.2 Nuclear Magnetic Resonance Spectroscopy

Nuclear magnetic resonance techniques give information about the *distribution of certain nuclei*, more precisely the distances between atoms in a molecule. Such distances can be used to derive a 3D model of the molecule. Whereas for small molecules it is acceptable to look at hydrogen (H) atoms, for larger proteins it is necessary to introduce carbon (C) and nitrogen (N) atoms to get sufficient data about the side chain conformations. Hence, the examined proteins grown in media are enriched with these isotopes. During this technique, the protein molecules are placed within a strong magnetic field, which leads to an alignment of the water atoms along the field. To move a nucleus from its equilibrium to an exited state, it is exposed to radio frequency (RF) pulses. During the alignment, as well as reversion into the default state, the atom spins. While the nucleus reverts into the equilibrium state, it emits (induces) RF radiation which can be measured. The frequency of the emitted radiation depends on the molecular environment of the nucleus and is thus different (unique) for each atom, except for those that are chemically equivalent.
During the procedure the chemical shifts, measured through the frequency, are obtained relatively to the reference signal and plotted in a diagram (chemical shift vs. chemical shift). The diagonal (see Figure 2.6) of the resulting two-dimensional NMR spectrum corresponds to the normal one-dimensional NMR spectrum. The off-diagonal peaks represent the interactions between hydrogen atoms that are close to each other in space. Thus, there is a strong parallel to the contact map of protein structures. By varying the nature of the applied RF pulses, the peaks can reveal different types of interactions. Different types of NMR spectra are used for the analysis. The spectrum derived during the *COSY* (correlation spectroscopy) experiment gives peaks between pairs of hydrogen atoms, which are covalently connected through one or two other atoms. In contrast, the peaks within the *NOE* (nuclear Overhauser effect) spectrum occur for pairs of hydrogen atoms that are close together in space (distance less than 5Å), even for those amino acid residues that are quite distant within the primary sequence. These spectra give information on the secondary and tertiary structure of a protein and can be interpreted by methods of sequential assignment. These methods are based on the differences in the number of hydrogen atoms and their covalent connectivity within the different amino acid residues. Each residue type leads to a specific combination of cross peaks in the COSY spectrum. These fingerprints are further combined with information gained from the NOE spectrum. During sequence-specific assignment of the NOE spectrum, the signals (of the NOE spectrum) are used to determine which fingerprint in the COSY spectrum comes from which residue adjacent to one previously identified.[BT99]

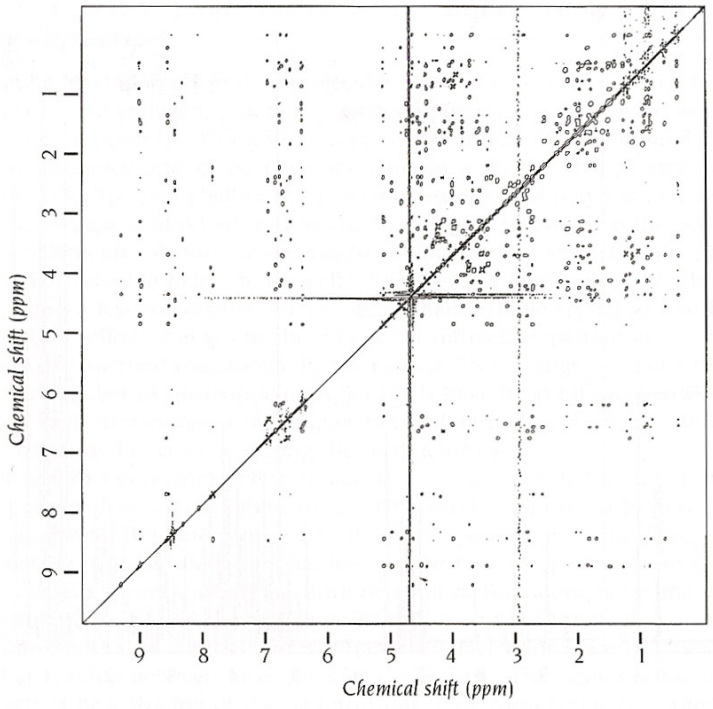


FIGURE 2.6: NOESY spectrum used to obtain information on secondary and tertiary structure [BT99, p.388]. The peaks along diagonal correspond to the one-dimensional NMR spectrum. The off-diagonal peaks represent interactions between hydrogen atoms, which are closer than a certain distance (5Å).

2.3.3 Comparison of X-ray Crystallography and NMR

For both methods the knowledge about the amino acid sequence is essential, when using NMR for the correct interpretation of the spectra and when using X-ray crystallography for the interpretation of the electron density maps. X-ray crystallography produces an image of the 3D model of the protein molecule, whereas NMR spectroscopy directly derives distance constraints for different hydrogen atoms along the polypeptide chain [BT99]. Both can be used to derive possible structures (called models) of a protein.
The resolution of NMR resolved structures is lower than for X-ray crystallography. NMR thus has larger RMSD values (root mean square deviation). In best case, the main-chain deviations are maximum 0.4A and side-chain RMSD values no greater than 1.0A. The high RMSD values result from the lack of sufficient distance constraints [Rho06]. The reason therefore is the spectral resolution, which is not sufficient to resolve all couplings, as these might overlap. When using crystallography, highly mobile parts of the molecule are often missing in the electron density map and thus from the molecular model [OGF+10]. This is because current methods are based on the concept of a rigid structure. Recently some methods have been proposed, which account the molecular motion and produce structural ensembles similar to NMR structures.
Besides the possibility of dynamically tracking specific reaction in living cells, when using NMR spectroscopy, another advantage is that there is no necessity of molecule

crystallisation. On the contrary, NMR is applicable only for molecules with an upper limit of size and produces structures less precisely than X-ray does.

2.4 Data Extraction

The resolved protein structures, measured via NMR or crystallography, are saved within public protein databanks and can be used for further data analysis. To visualise and manipulate distance constraints, used for reconstruction, contact maps are used. Hence, the use of such contact maps will be introduced first, before the preparation of data to extract spatial (orientational) constraints will be explained.

2.4.1 Contact Map

Within a contact map the amino acid sequence of a protein is plotted against each other, showing if and how frequent the residues are in contact, i.e. they interact. A contact map contains only distance information (based on a hard cutoff) and no information about orientation. By definition, two residues interact, if they are less distant from each other than a given threshold (like 5 or 8Å) (see Figure 2.7). The distance of two residues is therefore defined by the distance between its C_α atoms. These contacting residues define the tertiary structure of a protein. Various studies of Sathyapriya et al. showed that not all of those contacts are necessary to predict the 3D structure [SDS+09]. A subset of about 10% of contacts is sufficient, if the right ones (i.e. an appropriate set) are chosen. This is because, when thinking of hydrogen bonds for contacts, it is not necessary to keep all of them to preserve the folding. Thus, the other way around, we don't need all, but the relevant contacts to obtain the folding. Sathyapriya et al. investigated that intuitive algorithms, based on long-range or short-range contacts and other approaches, do not necessarily lead to better results than random selections. Thus, a better selection algorithm was necessary. They introduced a selection method, called Cone peeling algorithm, which leads to better reconstruction results than random selections do. This method uses a common neighbourhood condition to remove outliers, i.e. it filters false positives, in terms of reconstruction. The distance constraints of the selected contacts are finally used for the reconstruction (by reconstruction tools). As previously mentioned, those constraints could be supplemented through relative spatial (orientational) constraints for certain residue contacts. To define such angle annotations first of all a general orientation and alignment of residues and their neighbourhood (NBH) is necessary to allow comparisons and detect (statistical) spatial preferences. For this purpose a rotation and translation invariant framework, which will be explained in more detail within the next subchapter, was developed by Lappe et al. [LBF+09].

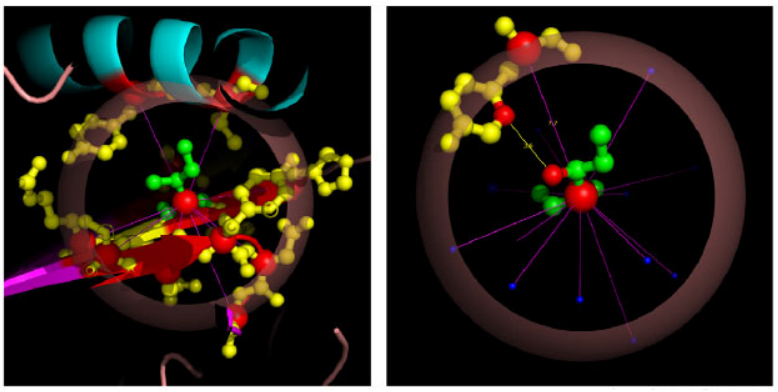


FIGURE 2.7: Cartoon representation of PDB 1a1m. Left: zoom onto residue I95 with all neighbouring residues. The C_α carbons are highlighted as enlarged red spheres. This illustrates how a single residue participates in the cooperative formation of the complex interaction network defining the overall structure. Right: Pairwise interaction between two neighbouring residues. When calculating pairwise propensities, all other contacts (now reduced to blue spheres) are ignored and assumed as statistically independent [LBF+09].

As distance is quite important for the purpose of contact definition, the distance definition should be introduced at this point. The *distance classification of residues into short-, middle- and long-range* can be defined either over the sequence distance of the spatial distance. Considering the sequence distance, long-range contacts occur between residues that are far in sequence (i.e. $|inum - jnum| > 9$) but close in space and short-range contacts are close in sequence (i.e. $|inum - jnum| <= 9$) and space. Whether two residues that are in contact, are far or close in sequence is quite important e.g. for the purpose of folding. In comparison, after Buchete et al. [BST04a] short-range is defined as distance range 2.0-5.6Å, middle-range 5.6-9.2Å and long-range 9.2-12.8Å. This definition will be used later to compute spatial contact potentials for different radius shells, i.e. different distance ranges.
The local neighbourhood of a residue is defined as set of all residue that are closer than the given contact distance-cutoff. It can therefore be defined within the contact map quite easily as the set of contacts along the row and/or column (see Figure 2.8). The neighbourhood is often written as String, called neighbourhood-string, of the one-letter-codes of the residues contained within the neighbourhood. The local neighbourhood will be analysed in the later context with the help of neighbourhood representing traces.

2.4.2 Rotation and Translation Invariant Framework

Naive approaches, used to compare the local structural environment of all residues contained in the sequence, align all residues in a pair-wise manner and hence derive the appropriate statistics. This is computationally very expensive, regarding the number of proteins contained in protein data banks, and thus not practicable. The PDB *cullPDB20*

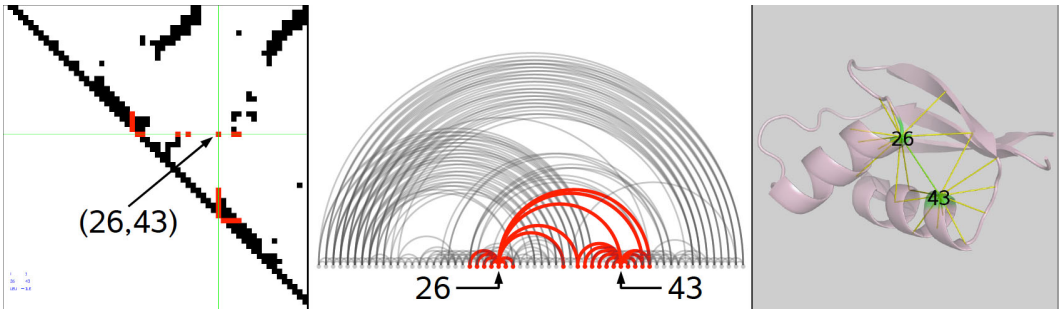


FIGURE 2.8: Neighborhood selection: The green cross haires point to the contact between residue 26 and 43 selected in the neighborhood selection mode which also highlights in red all contacts incident to the anchor residues 26 and 43. Subfigure (b) gives an alternative representation of the contact map. This graph clarifies the set of neighbouring residues of the residues 26 and 43. Note, that this representation is not implemented in CMView. Sending this selection to PyMol yields in a figure (c). Loaded structure: chain A of 1BXY. [prepared by Lars Petzold]

is a non-redundant subset of pdb's, which contains only pdb's of the PDB, which have a sequence identity of maximal 20% [WD03]. It contains approximately 1.200 structures (the PDB contains about 60.000) and altogether around 790.000 residues. Thus, when doing pairwise comparison of global structure, about $144*10^6$ comparisons are necessary. The pairwise comparison of local structure includes about $624 * 10^9$.

An efficient comparison of local structural environments across the database can be achieved via the use of a local instead of global coordinate system. Thereby, each residue with its local environment is transformed into a *standardised geometric framework*, which enables a *rotation and translation invariant analysis* of residue environments across the protein (in $O(log(n))$). Particularly, an amino acid is translated in this way that the C_α atom lies in the origin of the coordinate system. The amino acid structure is further rotated in two steps so that the C_α-C bond lies along the x-axis and the nitrogen atom lies in the x-z-plane. These two rotations are defined by two angles (*theta* and *phi* or *phi* and *lambda*).

The geometric operations are applied on every residue of the protein structure and all its neighbourhood. The new coordinates of each contacting residue in relation to the central residue can be defined via spherical coordinates (r, λ, φ) or Cartesian coordinates (x, y, z) (see subsection 3.3.2 for the definition of those coordinate systems). The neighbourhood of a residue of the protein is defined by all residues that are (according to the contact definition) spatially close to the central residue. For the purpose of spatial distribution patterns, this procedure of geometric operations is performed not only for a single protein structure but a set of non-redundant protein structures of a PDB.

The translated data, i.e. the coordinates for all contacting residues of protein structures, are saved for further analysis of contact potentials and other purposes within databases. These contain various tables to save all necessary information about the nodes of the protein structures (the basic residue characteristics) and its edges (basic properties of residues at either end of every edge in Cartesian and polar coordinates). They further provide information about the central coordinates, i.e. the coordinates of every atom

according to the new framework. In addition, so called neighbourhood-strings, describing the sequence, and transformation matrices used for the translation and compounded rotation are saved.
For the purpose of extracting contact residue potentials, the edges table will be most important. This table contains all necessary information to extract spatial distribution clusters for specific residue contacts. It includes information about all contacting residue pairs, which are closer than a given distance-cutoff, with the help of different attributes (*Graph_ID, Accession_Code, cid, i_num, i_res, i_sstype, j_num, j_res, j_sstype, dist_ij, j_atom, x, y, z, r, theta, phi*). The most important information for the extraction of orientation dependent contact potentials are the distances of these residues and the coordinates of residue J with respect to residue I (as central residue) and vice versa. The neighbourhood-string table contains the attributes *Graph_ID, i_num and nbhString*. It will be used to extract amino acid sequences of certain type (with a neighbourhood similar to a specific nbhString).
Through the transformation in a standard framework, the relative spatial distribution of amino acid residues and their neighbouring contacts becomes visible. Obviously the relative distribution deviates far from the uniform random distribution, as spatial clusters are visible for different residue types and their contacting residues. The frequencies are collected from a data set comprising several hundreds of different structures. Thus, the occurring relative distribution patterns of residue atoms and their contacts pose general features of protein structures.

2.4.3 Empirical Potential Energy Formalism

The resolved protein structures are used for statistical analysis and further for the derivation of knowledge-based potentials, which are more successful than the physics-based ones. Physic based potentials are derived from quantum mechanical calculations. In contrast, knowledge-based potential are empirical, statistical potentials, described by an energy function. Hence, knowledge-based potentials convert such residue contact potentials into energy-like quantities, weights or scores. The empirical potential functions are based on the relation of observed properties of known structures to (randomly) expected properties. The pairwise contact potential between residue $iRes$ (i) and $jRes$ (j) is defined as:

$$E_{i,j} = \log\left(\frac{P(observed)}{P(expected)}\right), \tag{2.2}$$

where $P(Obs)$ represents the probability of interactions between residues $iRes$ and $jRes$ ($\#observed\_contacts$ (under specific condition) divided by $\#all\_contacts$) and $P(Exp)$ represents the probability of interactions in the reference state ($\#expected\_contacts$ divided by $\#all\_contacts$). Thus, E_{ij} defines a statistical score that could be transferred into energy values between the residues $iRes$ and $jRes$ [LBF+09].

Chapter 3

Related Work

Within the next chapters various visualisations for proteins and their residue interactions will be introduced. Afterwards, an overview about different map projection techniques will be given. Furthermore, different general visualisation approaches and techniques, which could be used for the purpose of this project, will be discussed. Finally, a rough overview about clustering methods will be given and one will be explained in more detail.

3.1 Visualisation of Amino Acid Contact Data

One of the first visualisation of steric effects of protein folding was done by Ramachandran et al. in 1963 [RRS63] (see Figure 3.1). As mentioned before, the conformation of polypeptide chains can be described conveniently in terms of dihedral angles of relatively free rotation about two single bonds at each alpha-carbon atom. The Ramachandran plot clarifies, that only certain of these angles are possible (occur) within certain secondary structure types. Within right-handed alpha helices the residues are placed in succession with angles ϕ and ψ that are near -57°and -48°. In contrast, for pleated sheet structures formed between chains all angles are close to -139°and +135°. This plot is quite close to the analysis of spatial propensities, which is part of this thesis. Thereby, the analysis of bonded dihedral angles between two residues is extended towards non-bonded contacts (residue pairs). Besides, some filtering is applied for the secondary structure as well as the residue types to enhance properties within the plot.

Also Machin and Phillips [Phi69] used simple diagrams to visualise contact information for proteins. Within these diagrams standard secondary structure types become visible through concentrations of density thickening at the diagonal (for alpha-helices) or at right angle to the diagonal (for anti-parallel beta sheets). They also show other centres in which side chains and thus contacts are densely packed. Polypeptide chains are visualised in form of stereodiagrams (see Figure 3.2), in which the C_α positions of the amino acids are represented as open circles, numbered respective their type and connected via

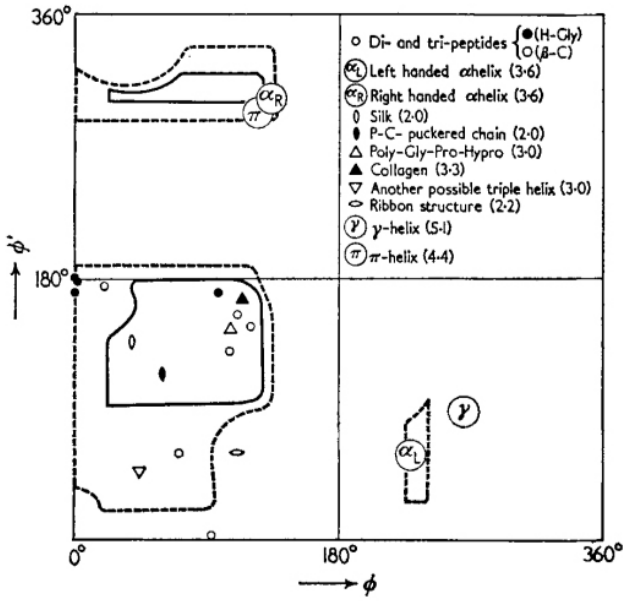


FIGURE 3.1: ϕ-ψ-Plot showing the boundaries of fully allowed and outer limit regions [RRS63].

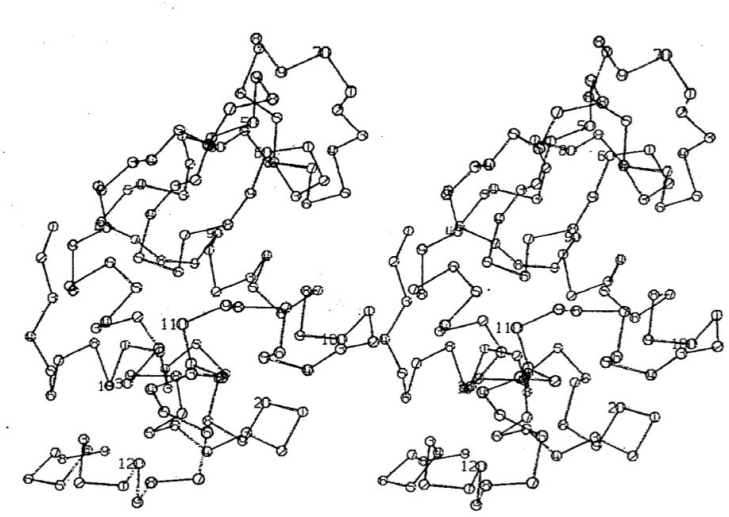


FIGURE 3.2: Stereodiagram of the main polypeptide chain in lysozyme [Phi69].

lines. Stereo diagrams pose a precursor of 3D visualisations, plotting the nodes at their positions and thereby ignoring the depth information. This makes it difficult to imagine the 3D arrangement. Nevertheless, such diagrams seem quite suitable for the spatial representation of amino acid sequences.

One of the first classical contact maps was published by Wyckoff et al. in 1967 (see Figure 3.3). Within classical contact maps the sequence of amino acids is plotted against itself and contacts between two residues are marked, if their distance lies below a certain threshold.

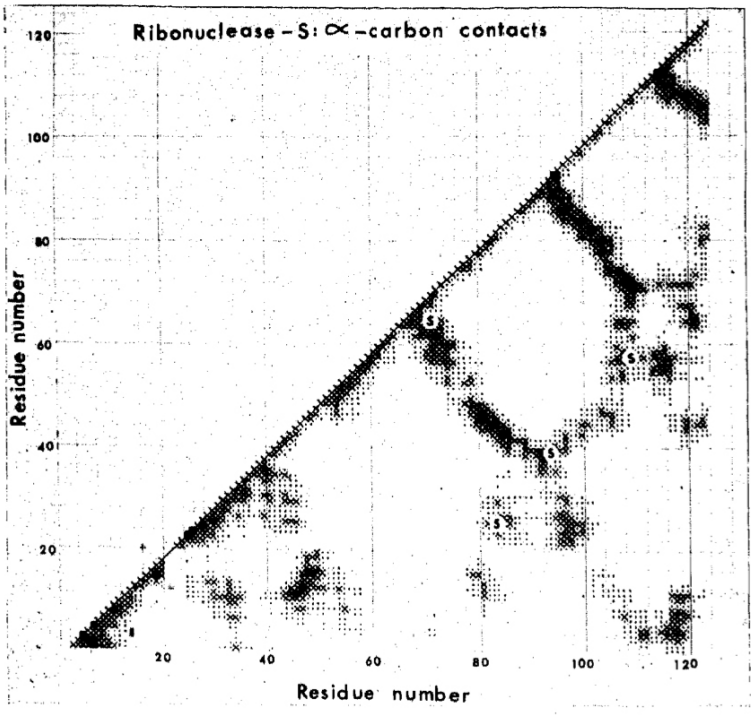


FIGURE 3.3: Map of contacts between C_α-atoms in ribonuclease-S; Black squares indicate contacts < 5Å, crosses indicate contacts between 5-10Å, dots indicate contacts between 10-15Å. [Phi69]

An even more simple approach to visualise residue interactions within a protein can be achieved with adjacency matrices [JAV07]. These matrices can be used to rank residues sited in a given chain. Vriend et al. made use of ϕ-ψ-plots as indicator for the quality of structures in the context of directional atomic contact analysis [VS93]. Besides the relative spatial distribution of residue-residue contacts, also atom-atom interactions have been investigated. The contact probability distributions are calculated for each voxel and represented as density maps (see Figure 3.4). Contacts are thereby characterised by the fragment type, atom type and the 3D position of an atom relative to the local frame of fragment. The approach of visualising areas of high densities as 3D n-gons was originally introduced by Rosenfield et al. [JSJ+84]. The representation through n-gons gives a good overview about the relative spatial distribution, but as interactions within 3D space are usually not that straight forward, this might be used just in addition to other visualisations.

The idea to use statistical knowledge-based potentials for the purpose of protein folding simulations was already seized by Buchete et al. [BST03b]. They introduced a method to obtain anisotropic, i.e. distance and orientation dependent, statistical potentials for coarse-grained representations of groups of atoms such as side chains. To derive quantitative parameters for orientation dependence for certain residue-residue contacts, local reference frames for each amino acid were defined. The statistical potentials were derived via the Boltzmann device. The anisotropic pair distribution $P_{ij}(r,\phi,\psi)$ is defined as probability that a residue of type I has a contact to a residue of chain type J within

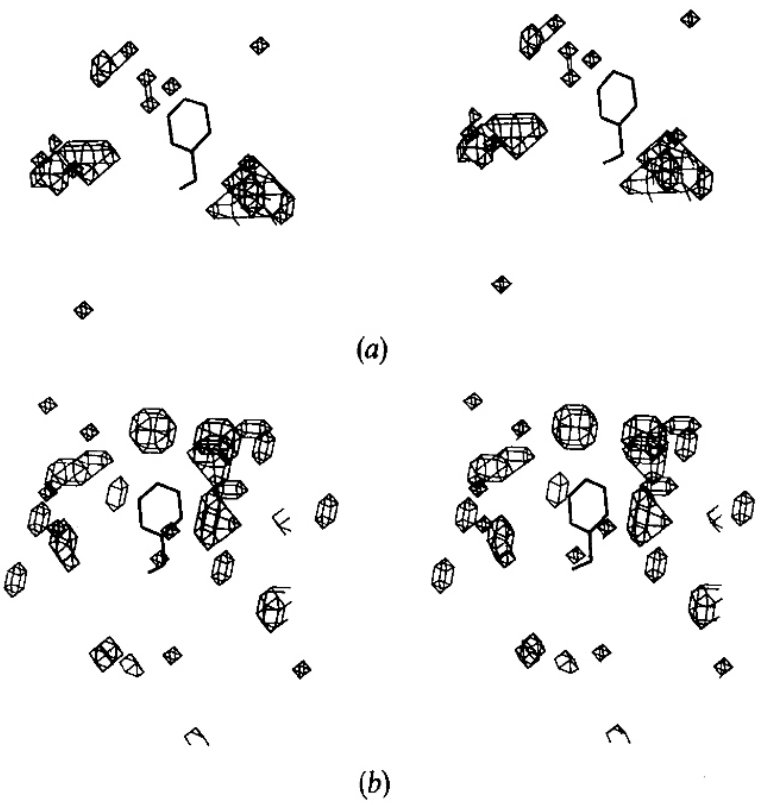


FIGURE 3.4: Probability distribution for positively charged nitrogen atoms around the phenylalanine side chain, contoured at an intermediate level (based on probability threshold) [VS93].

the volume element defined by r, ϕ and ψ. For the analysis of orientation dependence, distance dependent shells were used. These include short-range, middle-range and long-range distance shells. The statistical potential fields are visualised through the plotting as equiangular grids or as 3D representation (see Figure 3.5 and 3.6). The latter was expanded via superimposing the relative position of the peptide group of the central residue. It was further changed such that the magnitude of the potential is not only plotted onto colour but in addition onto the radius of the sphere. For both types of visualisation smooth potentials have been computed via spherical harmonic synthesis (SHS) (see Figure 3.6b). The representation of potentials at certain angles (spherical coordinates) via projection them onto the surface of a sphere is very straight forward. It poses a promising attempt to derive orientation constraints.

Another approach was suggested by Rakshit and Ananthasuresh [RA08], who used metric multi-dimensional scaling (MMDS) maps, based on inter-residue contact energies, using the Miyazawa-Jernigan matrix (see Figure 3.7). The MMDS map shows which amino acid residues are similar to one another in terms of statistic-based contact energies. On those maps, each amino acid of the protein is represented as a point on the MMDS map and the distance between two such points quantifies the dissimilarity in contact energies, whereby larger distance leads to larger dissimilarity. Residues that

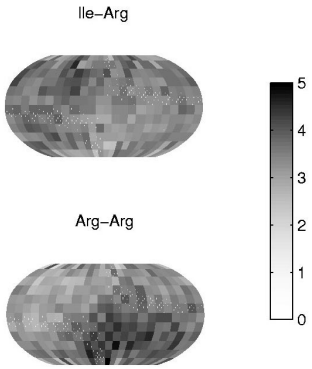


FIGURE 3.5: Orientational probability density maps for Arginine(Arg) interactions with Isoleucine (Ile) and Arg. The probability amplitudes correspond to the scale shown in the scale bar, in units of 10^{-3} [BST03b].

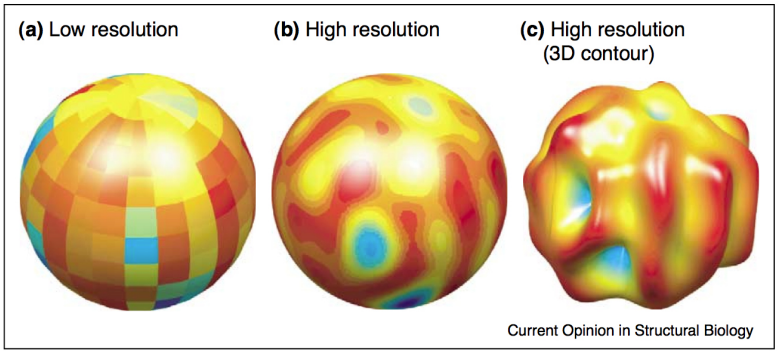


FIGURE 3.6: 3D representations of the SHS-reconstructed short-range residue-based orientation-dependent potentials for Ala-Gly interactions (a) on a 12×24 angular grid of the same size as the grid used for collecting the statistical data and (b,c) on a grid with a resolution that is ten times higher. In (c), the magnitude of the potential is proportional to both the distance from the interaction centre of the sidechain and the colour scale. [BST04b]

have a positive log-odds-score in the appropriate matrix are connected via double-ended arrows. Amino acids are coloured respective their chemical properties and thus classified into five groups. A further grouping of amino acids is achieved through a hierarchical clustering method. Such MMDS maps seem appropriate to compare contact energies of certain residues, but not for the analysis of spatial propensities (information), as the location and further distance is used as indicator for similarity or difference.

The most detailed and common form to represent tertiary structure is the ball-and-stick model, within that all atoms are displayed [ZB07]. There exist various simplifications of this model, i.e. computer generated diagrams at different degrees of simplification). As it is hard to interpret a flat picture of a 3D model, secondary structure types are highlighted. Alpha helices are generally displayed as cylinders and beta strands as arrows, simultaneously giving the direction of the strand from amino to carboxy terminus. The remaining parts are usually represented by ribbons.

Donoghue et al. [OGF+10] summarised various visualisations of proteins and residue contacts. Most of them are 3D surface-based or volume-based approaches. These 3D

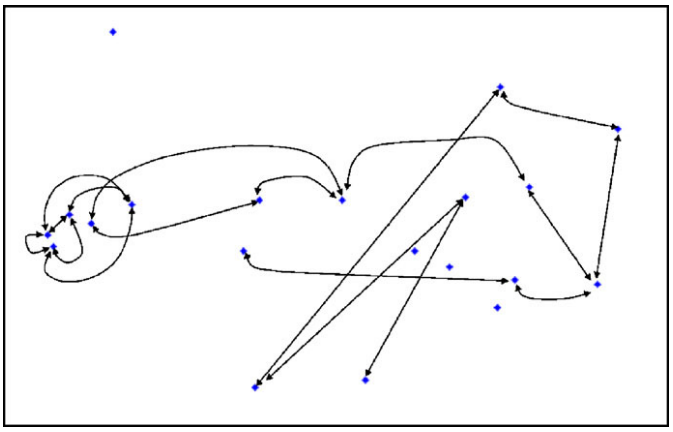


FIGURE 3.7: Amino acid map constructed using the metric multi-dimensional scaling method and the modified Miyazawa-Jernigan matrix as the proximity matrix [RA08].

visualisations are not based on volume data but a set of 3D coordinates for all atoms of the protein structure, which are used to position standard SST specific shapes or ball-and-stick models or to calculate and illustrate the surface (like a convex hull) of the protein. With the molecular graphics viewer, local properties, like hydrogen bonding ability, hydrophobicity and electrostatics, are mapped onto the surface using colour codings. However, volume-based approaches are used to analyse the space around the target molecule and to highlight regions with strong interactions e.g. with small molecules. Hence, the 3D datasets are rendered to show areas that are most favourable for interactions. Maps of atomic affinity can be rendered as isocontours, text-mapped clipping planes or volume rendering.
However, the last represented 3D visualisation techniques can only be applied on resolved protein structures, more precisely models that contain the three-dimension structure of a protein. They are thus good to compare different computed models and to check, whether adding orientation constraints to the distance constraints changed the model so that it is closer to the target model. Therefore different models can be loaded and superimposed.

The techniques, that are most promising to use for the problem of representing spatial propensities, are the Ramachandran plot and the spherical representations by Buchete et al.. Both are used to show spatial tendencies, i.e. preferred orientations, of some structures. What seems most adequate to visualise amino acid traces is the stereogram representation by Machin and Phillips.
The other introduced visualisations by Wyckoff et al. and Jha et al., do not seem suitable or applicable for the problem of visualising geometric orientation propensities. These techniques are much more suitable to analyse pure distance dependencies.

3.2 Visualisation of Amino Acid Sequences

Amino acids sequences are usually simply displayed as sequence of letters, whereas some letters might be highlighted through size or colour to demonstrate some properties of residues. The amino acids within the chain are mostly coloured respective their type, the motif they belong to, their secondary structure type or some biochemical properties (see Figures 3.8). This representation is suitable to show various properties of amino acids, but does not include any spatial information.

(a)

(b)

FIGURE 3.8: Representations of amino acid sequences [ZB07]

Wu et al. [WLYY03] introduced a two-dimensional representation of the DNA sequence, called dual-base-curve. It displays two of four DNA bases at a time on a plan, which emphasises repetitive structures. As proteins are build on much more than just four different bricks, it might not be possible or suitable to transfer this method for proteins.
A non-graphical approach to visualise the primary sequence of the DNA was suggested by Randić [Ran00], who developed a condensed representation. Instead of matrices, plotting the number of nucleic bases against each other and assigning each row/column to an individual nucleic bases, reduced matrices are used. These have the size 4 times 4, assigning rows and columns to the total number of nucleic acids of the same kind. This representation could be easily conveyed onto amino acid contact visualisation of proteins but does not include any information about orientation.

3.3 Map Projections

Map projections are used to adapt 3D features to flat models, thus representing an attempt to portray the surface of the earth or at least a portion of it on a flat surface

[Dan00]. Map projections describe, how to transform points from almost spherical objects onto flat sheets. There exist different projections and distortion technique, based on certain map properties, which will be summarised within the next paragraphs. The focus will therefore lie especially on those projection types, that are most suitable for the purpose of of defining orientation constraints.

3.3.1 Map Properties and Distortions

For the context of map projections, the most important map properties are area, shape, direction (conformality), bearing, distance and scale. On map projections at least one of those properties is distorted inevitably, depending on the type of projection. Consequently, distortion poses the limiting factor in the process of map projection. The untrue representation of area, linear dimension, angle or shape often leads to misinterpretation. Through distortion effects area might get enlarged or diminished, lines might become curved lines, shapes might be sheared or curved and points can become lines (like south and north pole) [Pea90]. Whereas, some projections minimise the distortion of some properties, at the same time they maximise the distortion of others [Dan00].
Conformality in terms of map projections means that the scale of the map at any point of the map is the same in any direction, whereby meridians (longitudes) and parallels (latitudes) intersect at right angles. A map is equidistant, if it portrays distances from the centre of projection to any other place of the map, if it preserves length along one of the base directions. A map preserves direction, if the angles from a point on a line to another point (called azimuths) are portrayed correctly in all directions. On equal-area maps, all portrayed areas over the entire map have the same proportional relationship to areas on earth that they represent.
A correct scaling of a map assumes that the relationship between the distance portrayed on the map and the same distance on the earth is constant over the whole projection. In general, we can distinguish linear and area scale, whereas the former can be further differentiated into nominal (principle) and local linear scale. The nominal linear scale correlates the ratio of true distance on the map to the equivalent distance on the model. In contrast, the local linear scale of a representation for a given point along a given direction is defined by the ratio of length of the segment on the projection to that of the corresponding segment on the surface of the sphere. The ratio between local and nominal linear scale of a map is called scale factor μ [BS95]. In those areas of the map where distortion appears, the local scale is larger or smaller than the principle scale.
The two main distortions on maps are the distortion of angles and the distortion of length. The latter occurs on all projections except conformal and equal area projections.

3.3.2 Coordinate Systems

For describing a position or point in 3D space, different coordinate systems can be used [BS95]. These comprise the Cartesian, spherical and cylindrical coordinate system. To define the position of points on a sphere, usually the (geographical) spherical coordinate system is used. It is defined by the zenith (polar) axis (usually the z-axis), an orthogonal (equatorial) plane (usually the x-y-plane), which divides the sphere into two halves, and a reference axis (x-axis) within that plane that goes through the polar axis (see Figure 3.9(a)). Within this coordinate system each point on the sphere is uniquely identified by the spherical (polar) coordinates radius, inclination angle and azimuthal angle [Wei10]. The radius is given by the Euclidian distance between the origin O of the sphere and the point P on the sphere. The inclination angle, also called zenith, polar or elevation angle, describes the smallest angle between the zenith-axis and the line segment OP (see Figure 3.9(b)). The azimuthal angle is measured from the reference axis to the orthogonal projection of the line segment OP onto the reference plane.
In the context of geometry, the inclination angle is often referred to as latitude (λ), but is given through the smallest angle of the axis through the surface point and central point of the sphere towards the equatorial plane, i.e. the latitude is equal 90°minus inclination angle. The azimuthal angle is equal the longitude (ϕ) in geographic context [Fur09].
Respective the literature, different denotations for those angles are possible. In this work, the symbols for radius, zenith and azimuthal angle coordinates are taken as r, ϕ (phi) and λ (lambda) with respect to the standard definition in the geographic context. Other definitions often use ϕ (phi) for azimuth and θ (theta) for inclination angle.
These spherical coordinates have restricted ranges [Wei10]:

$$r \geq 0 \tag{3.1}$$

$$0 \leq \phi \leq 180^\circ = \pi \tag{3.2}$$

$$0 \leq \lambda \leq 360^\circ = 2\pi \tag{3.3}$$

Considering the standard convention of the geographical latitude and longitude, these restriction ranges are [-180°:+180°] and [-90°:+90°].
Spherical coordinates can be transformed into Cartesian coordinates and the other way round with the following formulas:

$$r = \sqrt{x^2 + y^2 + z^2} \tag{3.4}$$

$$\phi = \arccos \frac{z}{r} \tag{3.5}$$

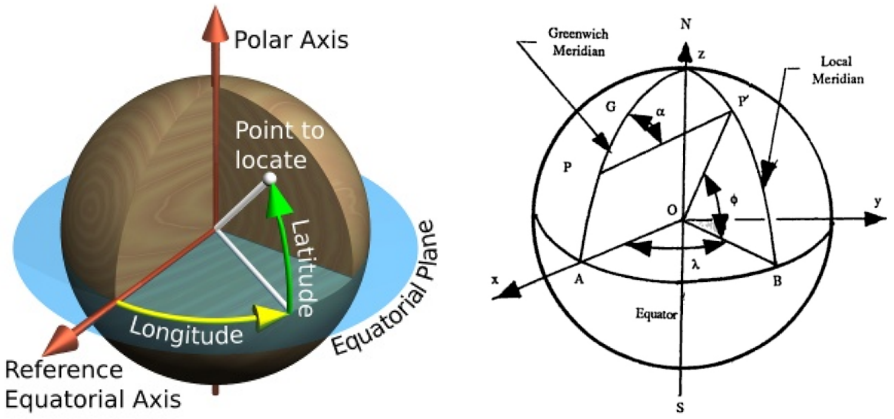
$$\lambda = \arctan \frac{y}{x} \tag{3.6}$$

$$x = r * \cos\lambda * \sin\phi \tag{3.7}$$

$$y = r * \sin\lambda * \sin\phi \tag{3.8}$$

$$z = r * \cos\phi \tag{3.9}$$

The spherical grid of coordinate lines over the spherical surface is called graticule. The circles situated on planes normal to the north-south axis, called parallels, do not cross one another. In contrast, the semi-circular arcs with the north-south axis as core, called meridians, meet at the geographic poles. Parallels and meridians cross at right angles.



(a) Definition of spherical reference system (b) Definition of spherical coordinates [Pea90] [Fur09]

FIGURE 3.9: Demonstration of the spherical coordinate system.

3.3.3 Projection Forms

As mentioned before, different types of projection surface can be used for map projections: cylinder, cone and plane [Wes74]. What further defines the projection is how the surface is placed relative to the sphere, i.e. at which direction the surface is plotted either tangent or secant to the model [Pea90]. Depending on the orientation at which the surface touches the sphere, we can differentiate normal (equatorial), transverse (polar) and oblique projections.
Another classification can be made according to the properties the projection type does preserve. This might be direction (azimuthal projections), local shape (conformal or orthomorphic projections), area (equal area projections), distance (equidistant projections) or shortest route (gnomonic projections) [BS95].

The projection types of the twentieth century [Sny93][Dan00][BS95] are:

- Cylindrical & pseudo-cylindrical projections
- Azimuthal & pseudo-azimuthal projections
- Stereographic projections
- Conic & pseudo-conic projections

The next paragraphs will give an overview about cylindrical and azimuthal projections, as these will be used in the later context. Stereographic, conic and pseudo-conic projections seem to be less suitable for interactions within the map and the definition of angle ranges. Especially conic projections might lead to irritations and misinterpretation of location (angle), because of their variation in stretching at north and south pole, which leads to a strong distortion of angles.

3.3.3.1 Cylindrical Projections

Cylindrical projections result from a projection of the spherical surface onto a cylinder. Within these projections the meridians are mapped to equally spaced vertical lines and circles of latitude, also called parallels, are mapped to horizontal lines. Thus, the straight meridians and parallels are always intersecting at right angles. Whereas the stretch distance from east to west is the same at any latitude, it changes from north to south in direct ratio to the corresponding difference in longitude (given by angle phi ϕ). By this projection transformation neither angular distortion nor distortion of length along the central meridian is introduced [Dan00][BS95].
We can differentiate cylindrical equal-area projections and cylindrical equidistant projections. The former is calculated by the transformation equations:

$$x = (\lambda - \lambda_s) * \cos \phi_s \tag{3.10}$$

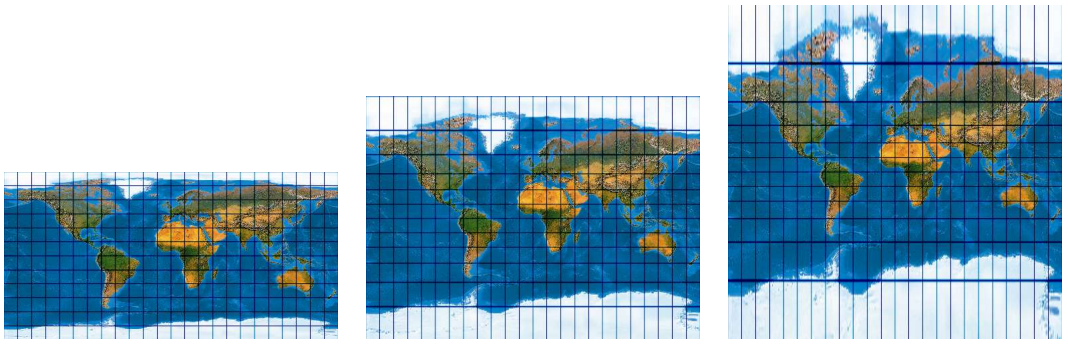
$$y = \sin \phi * \sec \phi_s \tag{3.11}$$

where λ is the longitude, λ_s is the standard longitude (horizontal centre of projection), ϕ is the latitude and ϕ_s is the so-called standard latitude. The latter is calculated via the transformation formulas:

$$x = (\lambda - \lambda_s) * \cos \phi \tag{3.12}$$

$$y = \phi \tag{3.13}$$

whereas for the equirectangular projection ϕ_s=0°.



(a) Carreé (b) Miller (c) Mercator

FIGURE 3.10: Cylindrical map projections [Boe10]

The most common equidistant cylindrical projections are the Platé Carreé, different Miller (ϕ_s=37°, 43°or 50°) and Mercator projections (see overview in Figure 3.10 and 3.11). The equidistant Platé Carreé projection is the conceivably simplest type of projection with a north-south distance that is neither stretched nor compressed. Within the Miller cylindrical projection the north-south stretching grows with the latitude, even if less quickly than the east-west stretching. The conformal cylindrical Mercator projection distorts areas excessively in high latitudes, as the north-south stretching is equal to the east-west stretching. In comparison, the transverse Mercator projection results from a projection of the sphere onto a cylinder tangent to the central meridian. Thus, it portrays areas with a larger north-south than east-west extent.
The Lambert (ϕ_s=0°), Behrmann (ϕ_s=30°), Gall orthographic (45°) and Peters (ϕ_s= 44.138°) projections (see Figure 3.11) belong to the equal area projections [Wei10]. Their north-south compression is exactly the reciprocal of their east-west stretching [Dan00][BS95].

3.3.3.2 Pseudo-cylindrical Projections

Pseudo-cylindrical projections are quite similar to cylindrical projections, with a straight central longitude and straight, equally spaced latitudes. In contrast, all the other meridians are curved, whereby the degree of curvature depends on the parallel the point is situated on and the distance from the central meridian. On the one hand these projections are conformal (angle preserving), but on the other hand they lead to distortions of shape or area, especially at polar regions.
The most common pseudo-cylindrical projections are the Mollweide, Collignon, Sinussoidal-equal-area, Sanson-Flamsted (Mercator-Sanson), Robinson, Kavrayskiy and Eckert projections (see Figure 3.12).
The Mollweide projection, also called elliptical projection or homolographic equal-area projection, is often used for world maps with elliptical longitudes, straight but unequally spaced parallels and poles represented as points [YST02][Wei10]. The transformation is given by:

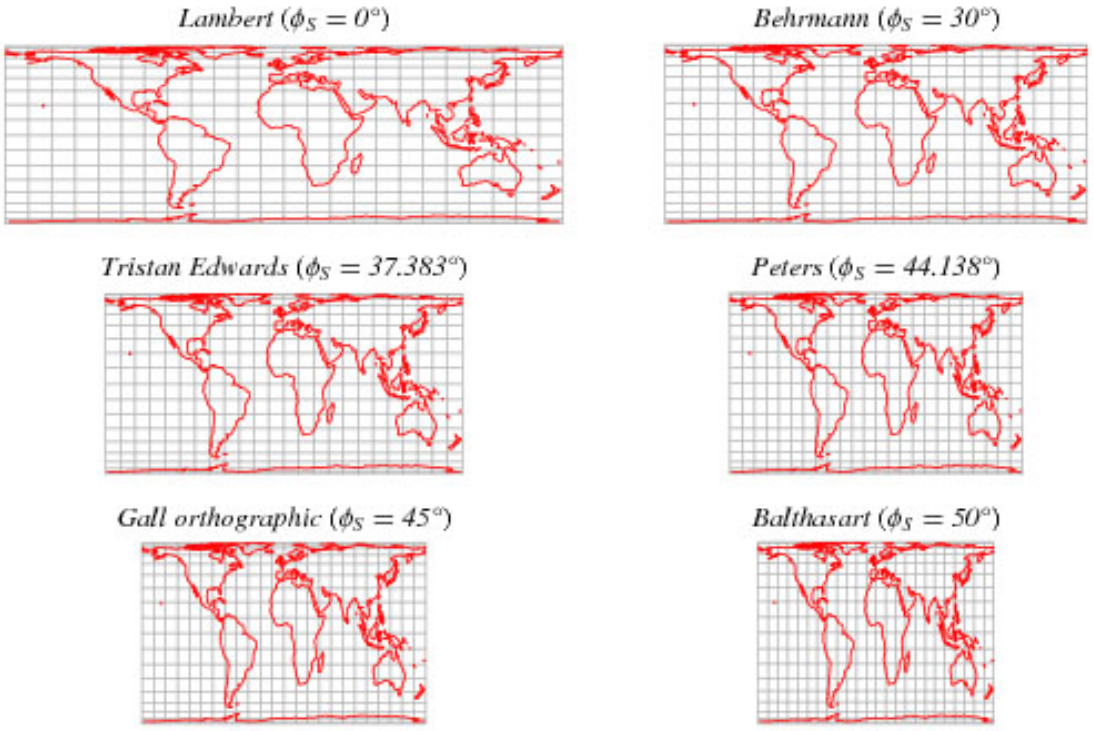


FIGURE 3.11: Cylindrical map projections [Wei10].

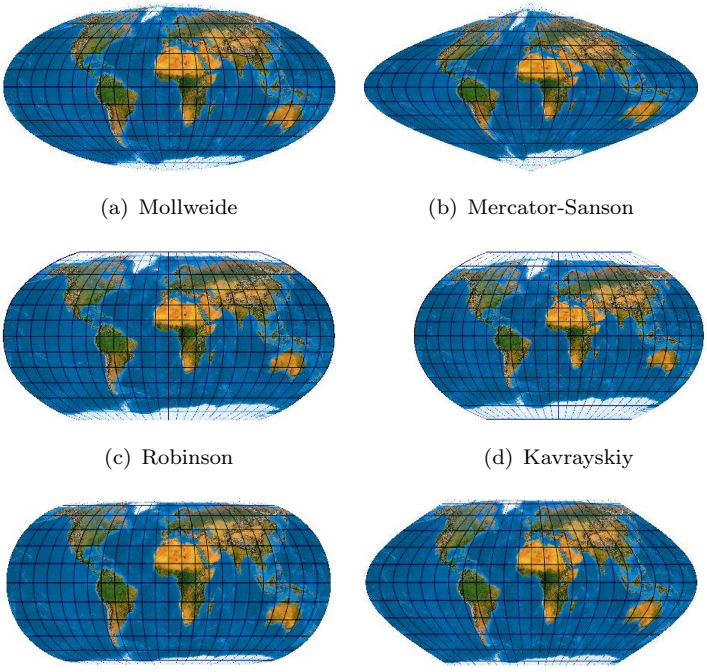
$$x = \frac{2 * \sqrt{2} * (\lambda - \lambda_0) * \cos\phi}{\pi} \tag{3.14}$$

$$y = \sqrt{2} * \sin\phi \tag{3.15}$$

The Collignon projection represents each meridian as two straight line segments, running from a pole to the equator. What makes the Robinson projection special is that it is based on lookup-tables of coordinates instead of mathematical formulas, like all the other projection types. This projection tries to balance the errors of projection properties via a simple uninterrupted graticule [Fur09]. The table defines x and y-coordinate values for increments of latitude of about five degrees. All other points in between those coordinates have to be interpolated.
Within both, the Mercator-Sanson and the Kavrayskiy projection, all meridians are represented as sine curves, whereas within the former the poles are illustrated as points and within the latter as lines [YST02]. Like the Robinson projection, the Kavrayskiy projection intends to produce good quality with low overall distortion. The projection is defined as:

$$x = \frac{3\lambda}{2\pi} * \sqrt{\frac{\pi^2}{3} * \phi^2} \tag{3.16}$$

$$y = \phi \tag{3.17}$$



(a) Mollweide (b) Mercator-Sanson

(c) Robinson (d) Kavrayskiy

(e) Eckert IV (f) Eckert VI

FIGURE 3.12: Pseudo-cylindrical map projections [Boe10]

There are several Eckert projections, all with straight central meridians and flat poles, both half as long as the equator [Fur09]. Whereas all even-numbered projections compress the vertical scale near the poles and stretch it near the equator to preserve area, all odd-numbered ones do better present the overall shape, as they have equally spaced latitudes only. The best known Eckert projections are IV and VI, which can be derived with the following formulas:
Eckert IV:

$$x = \frac{2}{\sqrt{\pi * (4 + \pi)}} * (\lambda - \lambda_0) * (1 + \cos\phi) \tag{3.18}$$

$$y = 2 * \sqrt{\frac{\pi}{4 + \pi}} * \sin\phi \tag{3.19}$$

EckertVI:

$$x = \frac{(\lambda - \lambda_0) * (1 + \cos\phi)}{\sqrt{2 + \pi}} \tag{3.20}$$

$$y = \frac{2 * \phi}{\sqrt{2 + \pi}} \tag{3.21}$$

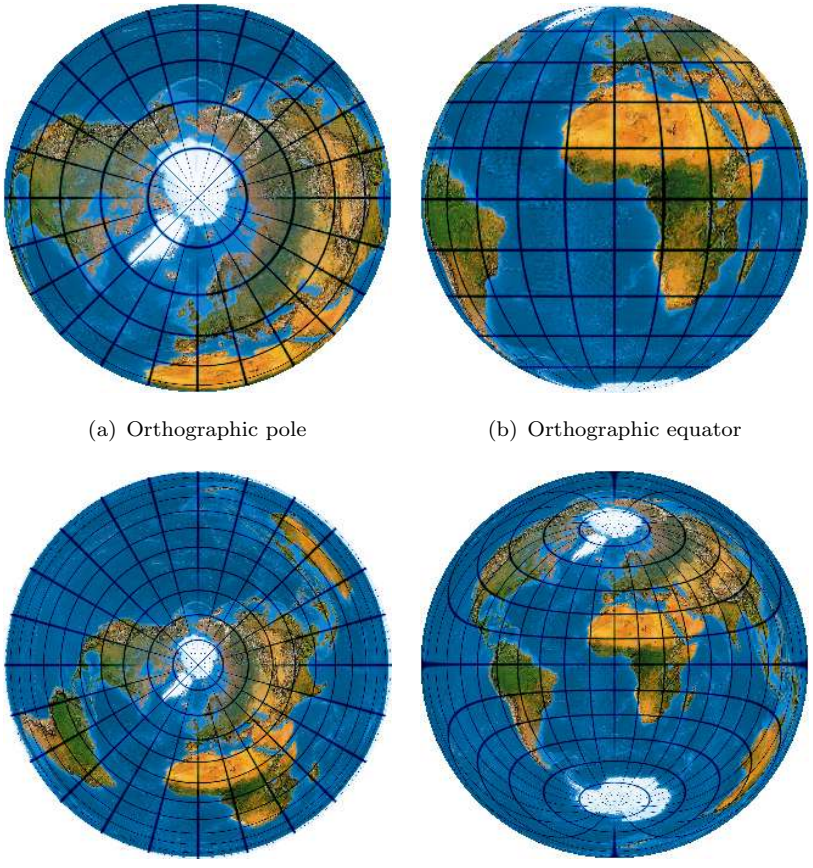
The Eckert VI projection is quite similar to the Kavrayskiy projection, with meridians represented as sine curves, besides they are very sharp near the equator. Within this projection, the distortion of shape increases at the poles.

3.3.3.3 Azimuthal Projections

For azimuthal projections the spherical surface is projected onto a plane. Azimuthal projections can be divided into true perspective projections and non-true perspectives. To the group of true perspective projections, among others belong the Gnomonic projection, the general perspective projection, the orthogonal (orthographic) projection and the azimuthal conformal projection (see examples Figure 3.13). The orthographic projection, which preserves neither area nor angle, is given by:

$$x = \cos\phi * \sin(\lambda - \lambda_0) \tag{3.22}$$

$$y = \cos\phi_s * \phi - \sin\phi_s * \cos\phi * \cos(\lambda - \lambda_0) \tag{3.23}$$



(a) Orthographic pole (b) Orthographic equator

(c) Lambert pole (d) Lambert equator

FIGURE 3.13: Azimuthal map projections [Boe10]

Forms like the azimuthal equidistant (neither equal-area nor conformal), Lambert azimuthal equal-area and logarithmic azimuthal equal area projections do not have true

perspective. The Lambert projection is given by the formulas:

$$x = k' * \cos\phi * \sin(\lambda - \lambda_s) \tag{3.24}$$

$$y = k' * [\cos\phi_s * \sin\phi - \sin\phi_s * \cos\phi * \cos(\lambda - \lambda_s)] \tag{3.25}$$

Based on the spherical coordinate system, within azimuthal projections the circles of equal distance can be projected onto circles concentric about the centre, whereby verticals can be projected onto straight lines passing through the centre of these circles.

3.3.3.4 Evaluation of Projection Types

Different cylindrical, pseudo-cylindrical and azimuthal projection types have been introduced. All three projection classes are suitable to visualise spatial propensities of residue contacts and define orientation constraints. Users might tend to use one of the classes. However, it is in the eye of the beholder, which one is most appropriate. Some users also might like to switch between the projections. Hence, different projections types should be available. As the different types of one class vary just slightly, it should be sufficient to offer at least one projection type of each class. The cylindrical projection will be implemented as one the most common equidistant cylindrical projections, the Platé Carreé projection. As representative for the pseudo-cylindrical projections, the Kavrayskiy projection will be implemented, producing good quality with low overall distortion. The orthographic orthogonal projection might be the most convenient one, as it is well known and should be familiar to all users, showing the topview on the globe.

3.3.4 Map Projections in Medicine

Map projections have been used for different purposes in medicine. Maps were widely used for colon flattening in the context of virtual endoscopy or colonoscopy [HGQ+06, BWK+01, HAK00]. The 3D colon surface, which was derived via segmentation techniques first, is thereby mapped onto a 2D representation.
Other approaches use map projection to map brain functions onto a 2D representation. Dum et al. [DS03] created a unfolded map of the dentate, thereby plotting experimental results onto an unfolded map of the cerebral cortex.
In the context of cardiology, map projections can be used to visualise cardial perfusion defects in subregions of the endocardium. Oeltze et al. [OGH+06] combine myocardium related results with a coronary artery analysis. They introduced a visualisation of perfusion data, which allows to relate perfusion data to morphological image data. For this purpose, a so-called "Bull's-Eye-Plots" are used, which show the perfusion information in an abstract way. The segments of the plot are coloured respective perfusion at rest and under stress. Each of the 17 segments thereby represents a different region of the

endocardium.

Rieder et al. [RWS+10] used map projections two produce a 2D image of the tumor surface. The tumor map is used during post-interventional assessment to illustrating and further evaluate the results of radiofrequency ablation therapy. The tumor surface is thereby coloured with respect to the coagulation zone it cuts. In particular, three zones have been defined, including the zone of complete cell destruction (green), the zone of cell destruction outside of the safety margin (yellow) and the zone of tumor cells which could not be ablated (red). The 3D surface is than flattened into a 2D map via spherical parametrization. Different map projections have been implemented, e.g. the Mollweide projection (see Figure 3.14).

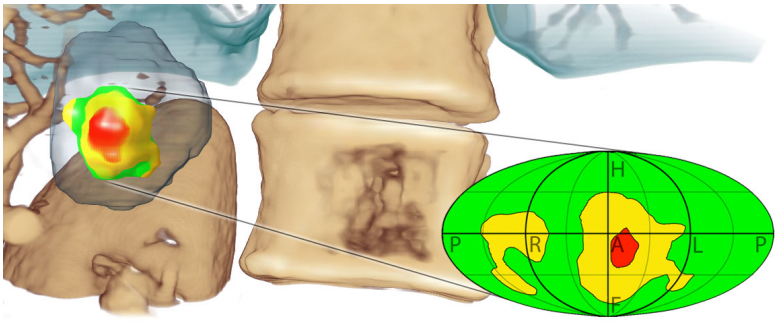


FIGURE 3.14: Tumor cell destruction is colour-coded onto the tumor surface [RWS+10]: red: no ablated tumor tissue, yellow: tissue outside of margin, green: tissue inside of margin. The surface can be transformed into a 2D tumor map (here: Mollweide projection).

Neugebauer et al. [NGB+09] introduced a visualisation of near wall flow data of cerebral aneurysms, thereby combining a 2D and 3D view to offer a complete overview. The simulated data values for the wall shear stress are mapped onto the 3D model, which poses the focus view of the representation. It is further surrounded by a multi-perspective 2D projection map, which provides the contextual overview, with respect to the spatial relation. Hence, the 3D surface data is projected onto a map, thereby only including those surface regions that are occluded.

In analogy, the contact density potentials could be mapped onto the model of a sphere (sphere surface). The model could be visualised via rendering (as 3D model) or via orthographic azimuthal projection. The backside of the projection could be visualised through multi-perspective maps as well, through the projection of cake slice shaped parts of the back around the focus view. Depending on the projection used for the cake slices, the overall representation would look like a flower or star.

3.4 Network Visualisation

Network or graph visualisation deals with the problem or *representing relational structures* of non-linear data [BW04]. Such data might also include the characteristic of enclosure (containment), i.e. hierarchical organisation (mostly represented by trees) [Maz09]. Network visualisation is used for various analysis problems of network data in the context of social structures, UML-diagrams, bio-chemical reaction chains, money flows, traveling patterns and more [JH04]. The challenge for network visualisations is to display graphs in order that structures, patterns or also outliers become visible. The specific graphical representation depends on the source and the properties of the underlying data [BW04]. Within most common layouts, vertices are illustrated as graphical symbols and edges as straight, curved or rectangular lines. Respective the kind of relation, they can also be represented through enclosing or touching of nodes.

Graphs consist of nodes and edges, which constitute the correlation between two

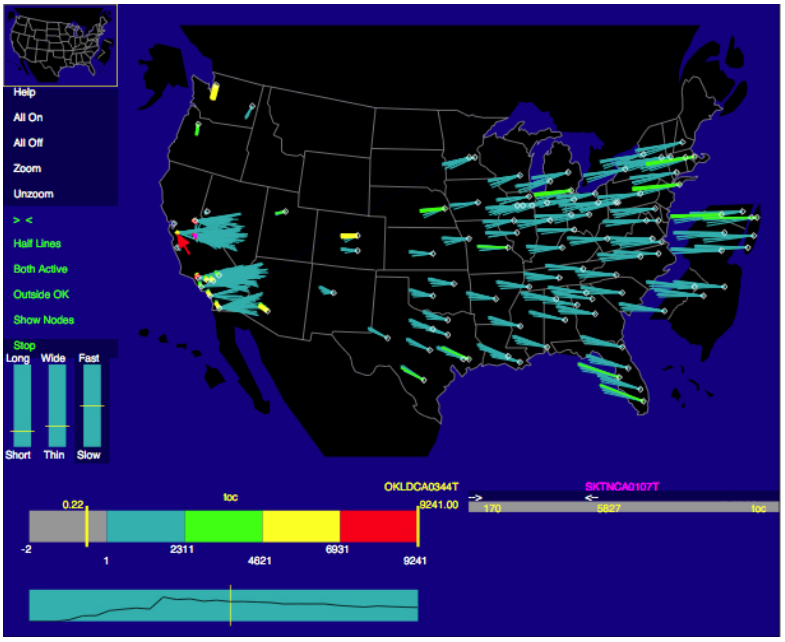


FIGURE 3.15: Network-wide overload. Instead of node-to-node edges, line shortening is used (half-lines are drawn only part way between the nodes that they connect) [BEW95].

entities. Nodes, also called vertices, represent instances of the data, i.e. physical or non-physical objects. If the graph contains just a single type of relation among nodes, it is called simplex network, whereas multiplex networks contain more than just one relation. To illustrate, how two entities are related, different edges are used. Edges can be directed, i.e. originate in a source node and end in a target node, or undirected thus representing a bonded tie between the nodes. They also might have an associated strength or weight, which can be of nominal, binary, signed or ordinal type. Additionally, edges can have labels, i.e. textual descriptions that depict relation properties. Networks consist of subgroups, which determine its macrostructure [JH04]. Thereby, the smallest subgroup is given by two nodes connected via an edge.

To overcome the problem of the high density of nodes and edges concentrated in small space, which result from huge amounts of data, various *graph layouts and graph drawing principles* have emerged [Maz09]. These layouts try to produce nicely arranged, reduced but more readable graphs. With an increasing number of crossing edges, it becomes more difficult, if not even impossible to perceive a graphs general structure. Thus, one of the principles is to reduce the number of crossing edges. Other graph drawing principles ask for edges of more or less equal length and other aesthetic properties like uniform distribution, minimal distance between unconnected nodes and some symmetry [BW04]. Based on these principles we can define layout techniques as optimisation techniques that try to find optimal positions for all nodes. These techniques include e.g. spring-embedders, gradient methods (force-directed techniques) or evolutionary methods. Besides an optimal positioning, another technique to improve the readability is to reduce the complexity of the graph. This can be achieved via filtering certain nodes and edges, e.g. considering their importance (weight) thus eliminating redundant edges. The number of visualised nodes could also be diminished through the clustering, i.e. grouping of "similar" nodes by combining them to one "pseudo-node".

In case that graphs are used to describe topology of networks, the position of a node is given by a spatial or geographic component. Within such graphs, map properties like volume or traffic between two locations are often mapped onto edge properties, like colour or width (see Figure 3.15)).
Graph layouts present one of the most fundamental tools for interpreting molecular interaction data via visualising nodes and edges as two-dimensional networks. Hence, some specific network visualisation systems for the context of biological data will be introduced. Breitkreutz et al. [BST03a] developed a software platform, called Osprey, to visualise and manipulate complex interaction networks of genes. They used a colour-coded representation of gene function and experimental interaction data. Thereby, genes are displayed as nodes and interactions between them as edges. To display the huge amount of interaction data adequately, data can be filtered regarding function or source of the gene, the used experimental system (NMR or X-ray crystallography) or connectivity. This approach could be transferred onto the visualisation of the neighbourhood of certain residues within the sequence. Nodes would thereby represent residues, particularly one specific atom of the residue, and edges the interactions (contacts) between them.
Pavlopoulos et al. [PWS08] gave a review on visualisation tools that are currently available to visualise biological networks. These tools can be used to analyse and interpret the relation of certain biological entities, like sequences, protein structures and families, proteomics data and others. The aim for all of them is to gain insight into the complexity and dynamics of biological systems. To do so, the data is mapped onto two-dimensional graphs, which visualise biological interactions and relationships between entities. These relations might be of different type, e.g. evolutionary relation, a shared protein domain

or the same protein family. Biological components might be linked by more than one relation, displayed via multiple edges (multi-edge networks), each with different meaning.

Becker et al. [BEW95] introduced a network visualisation system for geographical relationships that can further be adjusted with the help of display parameters, i.e. via interactive controls. To reduce the amount of data and thus information to plot, three methods are introduced: aggregation (for a large number of links and nodes), averaging (over a number of time periods) and thresholding (to detect changes over time). Different edge representations are suggested to display statistics, e.g. directional edges can be shown as arrows or bisecting segments (see Figure 3.16). To avoid an overload within areas where lots of edges start or end, either half lines are used (see Figure 3.15) or the most important ones are highlighted. This at the same time makes it hard to see where exactly the edges end. Nodes are displayed as glyphs and related properties are mapped onto its colour, shape or size. As long as the number of properties and thus variation in colour, shape and size is manageable, this is a very intuitive attempt to visualise different properties.

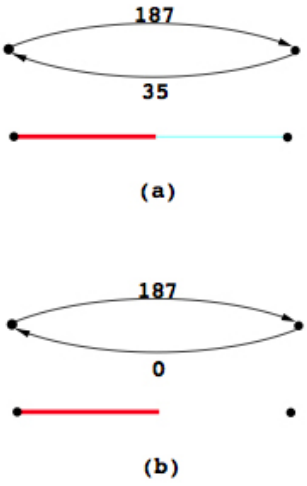


FIGURE 3.16: Representing link data: (a) the upper illustration is a conventional arrow diagram with numbers to show link statistics; the lower illustration uses line thickness and colour to convey the same information more compactly; (b) shows how half-lines represent statistics with value zero. [BEW95]

Shannon et al. [SMO+03] developed a software for the integration of biomolecular interaction networks and states, which constitutes a general-purpose modelling environment. Network graphs are used to visualise biomolecular interaction, whereas molecular species are represented as nodes and molecular interactions as edges between them. The network graph can be adapted via filtering, choice of attributes (single predicates of nodes or edges) and annotations (hierarchical classifications). Filtering is first of all used to reduce the complexity of a large molecular interaction network to a subset of

nodes and edges. Different graph layouts have been implemented including the spring-embedded layout, hierarchical layout and circular layout.
Besides Cytoscape, there are lots of other software tools that can be used for the characterisation and visualisation of molecular interactions and states, e.g. graph viewers as Pajek [BBM03], Graphlet [Gra00] and da-Vinci [Dav05].

3.5 Exploration of Large Data Rooms

The amount of data that needs to be displayed gets more and more, thus often leading to a lack of space when trying to display everything at once but still clear and interpretable. This physical problem directly leads us to the *need for interactions* like [Tid05]

- navigation,
- sorting and rearrangement of data,
- browsing,
- searching or
- filtering.

The common aim of these techniques is the supply of a *global overview*, while providing *specific details* at the same time [Maz09].
Shneiderman suggested the following three tasks of graphical interfaces, pooled as "information seeking mantra", provide an *overview*, offer *zoom* and *filtering* opportunities and present *details on demand*.
Interactive visual representations can either be static, manipulable or transformable [Maz09]. Whereas static representations do not allow any interaction at all, manipulable representations at least offer the possibility to change the view via zooming, rotating or panning. In contrast, transformable representations allow the user to manipulate the data during the preprocessing phase of the visualisation process, e.g. via filtering of the input data.
With the help of navigation and browsing techniques, points of interest can be displayed in context with the rest of the data. The most common of such techniques are

- zooming (often combined with scrolling and panning)
- overview and detail
- and focus and context techniques.

When using zooming in combination with panning, the user can get lost easily. To prevent this, a global view, i.e. the context, should be available at the same time. This can be realised via overview and detail techniques, which provide a global view either within an extra window or the corner of the main screen, or focus and context techniques. The latter unify both, detailed and context information, within one display with the help of distortion techniques (via transfer functions). This poses the main advantage in comparison to overview and detail techniques, within which the information is broken into two displays leading to a decrease in performance caused by the necessary visual search. Within focus and context techniques, the central part of the screen is used to represent the information of interest and the contextual information is represented within the peripheral parts of the screen. A selective reduction of information within the contextual area can thereby be achieved by:

- geometric distortion,
- filtering (i.e. elimination of details),
- selective aggregation,
- micro-macro-readings
- or highlighting [CMS99].

Geometric distortion means the relative change of numbers of pixels that are dedicated to objects in space. Whereas during filtering a selection of cases with respect to certain attributes is chosen, during selective aggregation new cases are produced via aggregation of others. Micro-macro readings are graphics within that the detail cumulates into larger coherent structures. During highlighting, individual items are marked or made visually distinctive from others.
To define which items are part of the focus of attention and which do not, so-called *"degree of interest"* (DOI) functions are used [CMS99]. Among others, the bifocal view, the fisheye view and the perspective wall are forms of this technique. Whereas within the bifocal view the contextual information is distorted outside of two vertical axes, the perspective wall uses the three-dimensional perspective to distort the context. Both techniques are characterised by a clear border between focus and context. In comparison, within fisheye views the detailed information fades increasingly with the distance from the centre of interest. The idea of these *"attention-warped"* displays was first introduced by Furnas in 1986 [Fur86].

What is sometimes even more important than the navigation through the visual representation to gain insights and understand relationships among various attributes, is the transformation of the underlying data. Transformation techniques, like *filtering, data recording, dynamic queries* [Maz09] or the *magic lenses* [BSP+93] remove irrelevant

parts from large multivariate data sets. Filtering is applied during the preprocessing phase to eliminate all data items, whose attributes are not relevant and should thus not be considered by the visual representation. Furthermore, an explorative analysis of just a subcollection of data should be done to ease the drawing of conclusions. The filtering process can also be applied during the recording phase, thus reducing the amount of data right from the beginning. When performing dynamic (database) queries the graphical representation should change instantly to offer direct feedback. In comparison to those techniques, when using the *magic lens*, filtering is just applied on that part of the data that is covered by the lens, thus affecting the appearance of these data. Another technique, called *data brushing*, uses filtering and highlights a subset of data within several graphics at once.

Overview and detail techniques as well as focus and context techniques are both powerful tools to explore large data rooms. Although by analogy to map views of the earth, where these techniques are used quite often, for the visualisation of spatial contact propensities they would not bring any support in defining orientation constraints. In contrast, filtering techniques might pose a good chance to extract propensities for certain contacts of specific secondary structure types. Even though focus and context might not be in demand, at least panning could be of interest, as it can be used to keep the area (dihedral angle range) of attention in the centre of the map projection.

3.6 Clustering Techniques

3.6.1 Overview on Clustering Methods

Clustering techniques are used to analyse a heterogeneous sum of objects, pursuing the aim of identifying homogeneous subsets of objects. The underlying data usually consists of many data points (objects), which are comparable. This also necessitates the definition of similarity measures or distance measures between objects as numeric values. Often used distance measures are e.g. the Euclidean, Hamming or Manhattan distance, or the maximum norm [DAK09].
There are different ways (possibilities) to discriminate clustering techniques. The mayor distinction (distinguishes) separates hierarchic from non-hierarchic methods. The former generates clusters as nested structures in hierarchical fashion, whereas the latter result in a set of unnested clusters [BEPW08].
Hierarchical clustering methods can further be divided into agglomerative (bottom-up) and divisive (top-down) methods. Bottom-up methods start with the finest partitioning, i.e. the single objects, collecting and grouping elements together stepwise. In contrast, top-down approaches start with the grossest partition, i.e. the group containing all

elements, and splitting it up stepwise during the refinement procedure.The most commonly used non-hierarchical clustering methods are the partitional clustering algorithms. These start on a given initial grouping fragmentation of objects. During the clustering process, single elements of the initial clusters are rearranged between groups with the help of exchange algorithms until a given objective function reaches its maximum.
Other differentiations distinguish non-overlapping (hard) from overlapping ("soft") methods. When using hard clustering methods, like k-means or spectral clustering, every data point is assigned to one cluster. In contrast, during "soft" clustering methods, like EM-algorithm or Fuzzy C-means algorithm, every data point is assigned a probability for each cluster.
The main disadvantage of those methods is that they expect a fixed number of clusters, which thus needs to be set in advance. They also do not consider noise and allocate each object to a cluster, even if just in terms of probabilities. Thus, they are not suitable for clustering nodes of amino acid sequence traces, as the plotting of these shows that there are many outliers.
In comparison to that, density based algorithms regard clusters as dense regions of objects within the data space, which are separated by regions of low density, i.e. noise. Thus, these cluster techniques allow the determination of arbitrary clusters. Among others the DBSCAN (density based spatial clustering of applications with noise), DENCLUE and the mean-shift algorithm are density based clustering techniques. The DBSCAN, which is the most common density based algorithm, relies on the density-based notion of clusters and tries to group objects into meaningful subclasses, thereby considering noise. As this algorithm seems to be the most suitable one for this context, it will be further described within the next section.

3.6.2 DBSCAN

The DBSCAN allows the detection of several clusters of arbitrary shape, while ignoring noise. Thereby data points, that are dense enough, i.e. they fulfil a density criterion, are grouped together to one cluster [DAK09]. Two data points are density connected, if there exists a chain of dense objects that connects those points. The core objects of a cluster itself are dense, i.e. they have a minimal number of neighbours (elements) that are closer than a given threshold. In contrast, density reachable objects are connected to the core object, but not dense themselves, as they do not fulfil the density criterion. They are usually situated at the border of a cluster. Noise points are not assigned to any density connected cluster and returned separately. There are two parameters, which describe the density criterion: Epsilon describing the closeness, i.e. a threshold for a maximal distance two points can have to be reachable from each other, and a second threshold which defines the minimal amount of neighbours, thus defining the neighbourhood.

The advantage of the DBSCAN is that it does not require the number of clusters in advance. It is further deterministic and allows the application of various distance functions.

for all $unvisitedP \in DataSet(P)$ **do**
  $P \leftarrow visited$
  $Nbh \leftarrow getNeighbourhood(P, eps)$
  if $sizeof(Nbh) \leq minNumPts$ **then**
    $P \leftarrow noise$
  else
    $C\_ID \leftarrow C\_ID + 1$
    $expandCluster(P, C\_ID, Nbh, eps, minNumPts)$
  end if
end for

Algorithm 1: DBSCAN

Asign P to cluster C_ID
for all $P' \in Nbh(P')$ **do**
  if P' not visited yet **then**
    $P' \leftarrow visited$
    $Nbh' \leftarrow getNeighbourhood(P', eps)$
    if $sizeof(Nbh) > minNumPts$ **then**
      $Nbh \leftarrow Nbh \cup Nbh'$
    end if
  end if
  if P' not yet assigned to any cluster **then**
    Assign P' to cluster C_ID
  end if
end for

Algorithm 2: expandCluster

Chapter 4

Concept

The aim of this thesis is to offer a graphical support for the detection of spatially preferred orientations of specific residue contacts. Therefore, the potential spatial distribution of residue contacts needs to be extracted, based on the data analysis of available resolved protein structures. This includes the alignment of contacts along an orientation and translation invariant framework (which was explained in subsection 2.4.2). The visual representation should highlight preferred dihedral angles and thus help to define several orientation constraints for a subset of the residue contacts. To provide decision support for the definition of preferred spatial orientations, two approaches of visualisation will be pursued. To facilitate the definition of preferred dihedral angles, both visualisations should use respective colouring for log-odds-scores and amino acid types as well as other techniques to highlight important properties.
First of all, a comprehensive description of the extraction of the necessary statistical background information will be given. This will be followed by a detailed description of the two visualisation approaches, including the main representations of orientational (geometric) contact potentials (the map of log-odds-scores) and neighbourhood traces. Thereby, the potential and the benefit of certain visualisation approaches and techniques, introduced within chapter 3, as well as general decisions made for implementation will be discussed.

4.1 Derivation of the Statistical Background Information

The first visualisation should represent the anisotropic contact density potential for a selected residue contact. This potential can be described as the probability that an amino acid of a certain residue type is located at some orientation (defined by two angles) with respect to another residue of the chain. Such probability values can be derived from specific knowledge based potentials that are based on the relation of observed and expected values. Particularly, these potentials will be calculated referring to the

Bayes'theorem:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{4.1}$$

where $P(A)$ is the probability that a contact of any type is located at the spherical position (r, λ, ϕ):

$$P(A) = \frac{\#edges(r, \phi, \lambda)}{\#edges()} \tag{4.2}$$

and $P(B)$ is the probability that there is a contact between residue I ($iRes$), whereby I is of secondary structure type $iSStype$, and residue J ($jRes$) at any position:

$$P(B) = \frac{\#edges(iRes, jRes, iSSType)}{\#edges()} \tag{4.3}$$

and P(A|B) is the probability that there is a contact between $iRes$ with $iSStype$ and $jRes$ at a certain spherical position in space, we name sphoxel:

$$P(A|B) = \frac{\#edges(r, \phi, \lambda, iRes, jRes, iSStype)}{\#edges()} \tag{4.4}$$

Such orientation propensities can be derived as LOSs for certain volume elements, named *sphoxel* (spherical voxel). We defined a sphoxel, as volume element in spherical space, which is bounded by the spherical coordinates: (r_{min}, ϕ_{min}, λ_{min}) and (r_{max}, ϕ_{max}, λ_{max}), whereas $\phi \in [0 : \pi]$ and $\lambda \in [0 : 2\pi]$.
The number of observed contacts arises from P(A|B) :

$$P(observed) = P(A|B) \tag{4.5}$$

and the number of expected contacts:

$$P(expected) = P(A) * P(B) \tag{4.6}$$

This leads to the final log-odds-score:

$$LOS = \log(\frac{\#edges(r, \phi, \lambda, iRes, jRes, iSStype) * \#edges}{(\#edges(r, \phi, \lambda) * \#edges(iRes, jRes, iSStype))}) \tag{4.7}$$

The number of contacts ($\#edges$) for the different conditions are determined via SQL-queries on the table ($edges$) of a database, based on the orientation and translation invariant framework. This table contains all contacts of the *cullPDB20*, which is a selection of various resolved PDB's. The queries for the different edge-counts can be phrased like:

- #edges() = SELECT COUNT(*) FROM edges;

- #edges($iRes, jRes, iSSType$) = SELECT COUNT(*) FROM edges WHERE $i\_res = iRes$ AND $j\_res = jRes$ AND $i\_sstype = iSSType$;

- #edges(r, ϕ, λ) = SELECT COUNT(*) FROM edges WHERE $r >= rMin$ AND $r < rMax$ AND $theta >= pMin$ AND $theta < pMax$ AND $phi >= lMin$ AND $phi < lMax$;

- #edges($r, \phi, \lambda, iRes, jRes, iSSType$) = SELECT COUNT(*) FROM edges WHERE $i\_res = iRes$ AND $j\_res = jRes$ AND $i\_sstype = iSSType$ AND $r >= rMin$ AND $r < rMax$ AND $theta >= pMin$ AND $theta < pMax$ AND $phi >= lMin$ AND $phi < lMax$;

These log-odds-scores determine, if a contact is over- or underrepresented at a specific dihedral angle, i.e. it occurs more or less often than expected. Orientations within that residues occur statistically more often can be interpreted as preferred orientations. To detect these preferred orientations, it is necessary to compute the LOSs for the whole range of λ and ϕ. The resolution and thus sphoxel-size was thereby set to 2.5°for λ and ϕ, whereas Δr was set to 3.6Å. The latter follows the definition for distance shells, introduced in subsection 2.4.1. A resolution of 2.5°for the two angles should be more than sufficient, as the aim is to analyse the overall distribution of orientation propensities. A further decrease in resolution would not lead to any further information. This resolution produces a grid of 72×144 LOSs.

The second visualisation should represent the whole local neighbourhood of the central residue of interest ($iRes$), whereas $jRes$ (of a selected contact $iNum\_jNum$) is part of this local neighbourhood. The neighbourhood of a residue is defined as the set of all contacting residues, i.e. all residue that are closer than a given distance cutoff (usually set to 7-10Å). Regarding the Contact Map the NBH for $iRes$ ($jRes$) is given by all contacts, which lie in the same row (for $iRes$) or column (for $jRes$).
This local neighbourhood can be described by a String, more exactly the chain of one-letter-codes for the residues ordered by their sequence number. This template String can be used to extract neighbourhoods from resolved structures contained in the *cullPDB20*, which are built in a similar fashion.
More exact, the SQL-query:

- SELECT $graph\_id, i\_num$ FROM nbhstrings WHERE nbhstring LIKE "ADxCL";

can be used to extract the pdb's, which contain a residue or even residues (defined by i_num) with a similar NBH given by a string (e.g. "ADxCL"). The small letter 'x' thereby stands for the central residue i_res. Based on the extracted pdb's (defined by

the *graph_id*) and residue numbers (*i_num*)) the contacting residues can be extracted from the table *edges* via database queries like:

- SELECT $j\_num, j\_res, r, theta, phi$ FROM edges WHERE $graph\_id = id$ AND $i\_num = num$;

4.2 Sphoxel-Map Representation

The contact densities within the three-dimensional neighbourhood of a specific residue can be visualised either in 3D, via volume or surface rendering, or in 2D with the help of two-dimensional plots. Such three-dimensional visualisation techniques might give a better feeling for the spatial distribution of contact density potentials than two-dimensional ones do, at least while rotating the view. They might be more suitable to give an overall view about how clusters are situated around the central amino acid. Also the information about the radius gets lost within two-dimensional plots. On the other hand two-dimensional views are more suitable for interaction to define preferred dihedral angles. In addition, it is also easier to read the exact angle values from a scale in two-dimensional views.
The focus of the visual representation lies on the detection and further selection of the dihedral angle ranges for lambda (λ) and phi (ϕ), whereas r is not of much interest, as it is already given by the distance constraints. This is why I decided to first of all use a two-dimensional visualisation. Volume and surface rendering techniques could be used later on to complement the map projections and will therefore be discusses within section 4.6.
For the sphoxel map representation, the contact potentials of the surrounding neighbourhood of some residue, bounded by given shells, are plotted onto the surface of a sphere (restricting the neighbourhood) via summing up the values within the shells following the concept of Buchete et al. (represented in section 3.1). To convert the spherical, still three-dimensional representation into a two-dimensional one, map projection techniques are necessary.

Respective the "sphoxel" map projection, we can choose between a vast number of different projection methods. Comparing the different general classes of map projections, pseudo-cylindrical projections might be the most suitable ones for this purpose. The reason therefore is that those projections preserve angles, which are of our main interest. Particularly, the Kavrayskyi projection will be implemented. Although this projection class seems to fit best, also other projections should be offered as alternative to determine angle pairs. These contain the equirectangular cylindrical projection and an orthographic azimuthal projection including front and back sight.
To support the readability and assignment of exact angle values, rulers could be plotted for both dimensions of the map (see Figure 4.1and 4.2), further supplemented by a cross-hair drawn for the current mouse position. In addition, longitudes and latitudes

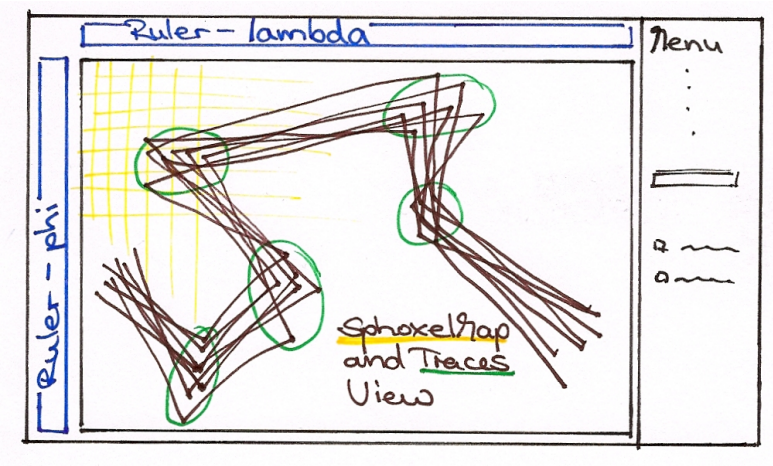


FIGURE 4.1: Draft Interface: SphoxelMap (implied by the yellow grid) will be superimposed by the NBHString traces (brown). Rulers of the two spherical angles λ and ϕ will be added for orientation and the menu on the right for adjustment of settings and parameters.

could be drawn within the map projections for steps of 90°and 45°. Furthermore, the spherical zero point, where both angles are zero, could be surrounded by a circle.
For those sphoxels, where the number of observed contacts is smaller than the number of expected contacts the scores are negative and vice versa. To visualise the value of the scores, these should be mapped onto an appropriate colour scale and tiles should be coloured respective the LOSs. A tile here stands for the surface representation of a sphoxel. For this purpose, different colour scales are conceivable:

- colour scales based on two colours
- or colour scales that include a whole range of colours.

The former could be used to plot negative LOSs onto the one and positive scores onto the other colour, whereas the absolute value would determine the brightness or saturation of the colour. The latter would map the scores on a linear scale including several colours, e.g. the rainbow colour scale. For the mapping of LOSs onto colour, the LOSs have to be normalised first. Thereto, the minimum and maximum occurring values are used to normalise the (negative and positive) values and transfer them into a range from -1 to +1. The scores may vary drastically for different residue types and orientations. To get rid of outliers, the user should be able to set the thresholds, used for the normalisation, by hand. If these thresholds are applied instead of the minimal and maximal value, all scores that are below the lower threshold are set to -1 and all above the upper threshold to +1.

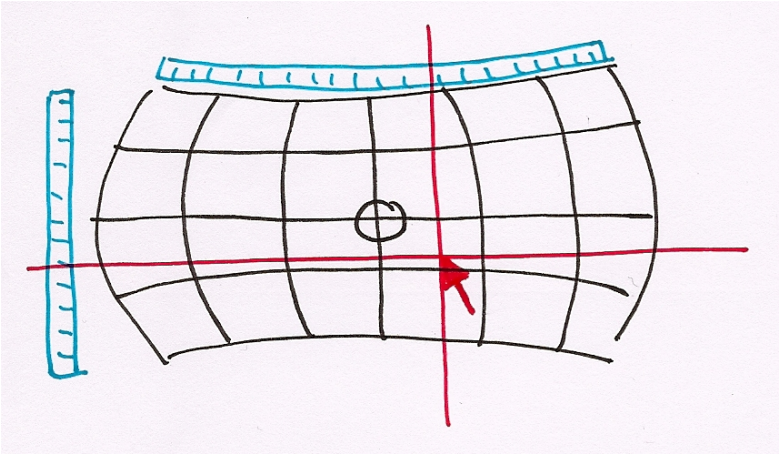


FIGURE 4.2: Draft Rulers: SphoxelMap will be superimposed by some main longitudes and latitudes. Rulers of the two spherical angles λ and ϕ will be added for orientation as well as the crosshair.

4.3 Neighbourhood-Traces

Derivation of Template NBHStrings In accordance with the homology approach, residues with similar neighbourhood, represented through the neighbourhood string, should be oriented in similar fashion. The local neighborhood of a residue could be represented as abstraction of the backbone, i.e. the connectivity along the polypeptide chain. Particularly, it will be illustrated as pathway (amino acid chain) via connecting the residues contained in the neighborhood ordered by their residue number via edges. Therefore, it is necessary to extract and plot the location of them with respect to the central residue (see Figure 4.1). In order to get some statistical information about similar local neighbourhoods, it is necessary to extract other sequences, which contain a similar neighbourhood (as described within section 4.1).
The original template neighbourhood string (*nbhString*), derived from the contact map, usually does not provide lots of output traces, when looking for similar neighbourhoods. At that, it is necessary to reduce the *nbhString* and remove some residues to increase the number of output traces. This could be done either interactively through the user via toggling on or off certain residues or through the selection of one of some suggested substrings. The latter option would be easier in terms of usability. The substrings can be extracted via queries on a database, which contains all occurring neighbourhoods of the *cullPDB20*, and the respective *result vectors* for the 20 residue types. The *result vector* contains all residue types ordered by their number of occurrances as central residue (x) within the neighbourhood and those states which residue types are supported best within such kind of neighbourhood. The query extracts nbhStrings that have a high support and what is even more important a preferred position of the residue of type *i_res* as central residue.
All extracted neighbourhoods could be plotted as traces. Occurring clusters of nodes

within these traces should help to define certain angle ranges, within that a contact between two specific residues is situated with high probability. These clusters should further coincide with areas within the first visualisation, where this residue contact is over-represented.

NBHs as Graphs Looking at amino acid sequences as traces, there are certain parallels to graphs and networks. They can be seen as some kind of directed acyclic graphs, whereas each two entities that are situated next to each other within the sequence are connected via an edge. Thus, each residue (besides the first and the last of the sequence) is connected to two other residues. Further, the position of each node is given by its geographical position (the dihedral angles), which is why we do not have to deal with the problem of positioning the nodes to see the general structure of the graphs, which is much more familiar (through the linear sequence of residue). Rather, it is required to see visual clusters of nodes in some areas and locations, i.e. the spatial distribution and arrangement of nodes.

Nodes of NBH Traces As the position of nodes is given, only other properties of the amino acids need to be mapped. These properties include the amino acid (residue) type, residue number and the secondary structure type of the residue. The supply of necessary detailed information about the residue, i.e. secondary structure type, residue type and residue number can be simply achieved via labels and visual mapping parameters like colour, size or shape (see Figure 4.3). Whereas colour will be used to classify the residues considering their secondary structure type, the number of residue types (20) and possible residue numbers are to high to enable an adequate mapping onto colour and shape. That is why labels will be used to communicate these data values. The colouring, shape as well as labelling of nodes should be explained to the user within legends.

Edges of NBH Traces Colouring could not only be used for the nodes but also for the edges (traces). These could be coloured with increasing residue number (from the first to the last residue), quite similar to the rainbow colouring used by other programs for protein data visualisation like Pymol [DeL] (which uses a colour gradient from N- to C-terminus). The colour of the edge would therefore depend on the sequence separation of the residue with respect to the central residue I, i.e. the difference in residue number. The colour gradient along a trace would thereby clarify where and how far in sequence the residues are situated.
The traces could be plotted in two ways either including the central residue or excluding the central residue. The central residue is situated in the centre of the sphere and could therefore be mapped onto any position on the surface of the sphere. If we imagine a transparent sphere, independently from which direction we are looking at it, the central

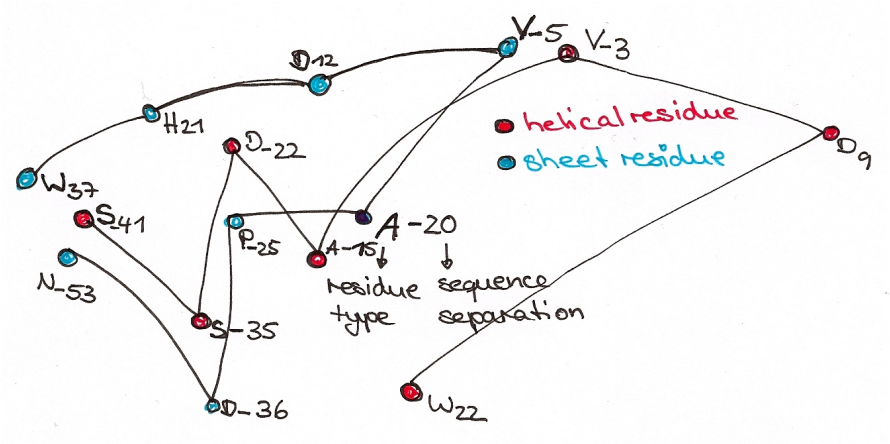


FIGURE 4.3: Draft "node labels": nodes (residues) are plotted and connected, ordered by their residue number. The colour of the node could indicate the secondary structure type, whereas the residue type is given by the one-letter code within the label. The label also includes the sequence separation of the residue to the central residue $iRes$, i.e. $kNum - iNum$ where $kRes$ is the k'th residue of the NBH.

node would always be in the centre of the sphere (of the orthographic projection). In analogy to the orthographic projection, the central residue is also drawn in the centre of the projection within the other map projections.
As the node positions are handled in form of spherical coordinates, also the edges between them should be treated in terms of a spherical coordinate system. Thus, it might be more appealing to draw the edges as geodesics instead of straight lines (which represent the Euclidean path). A geodesic represents the shortest path between two points on a sphere. A geodesic is defined as the shortes path along the great circle, that contains both points [Wei10]. The great circle around a sphere containing two points is given by that circle, that has the same radius as the sphere. On map projections they look like arcs.

4.4 Clustering of Nodes and Edges

The aim of the NBH trace representation is to illustrate visual clusters of nodes in some areas and locations, i.e. the spatial distribution and arrangement of nodes. What might be useful and of interest is to identify these areas of node clusters and highlight them. The clusters could be extracted via clustering methods like the DBSCAN algorithm (explained in subsection 3.6.2). This technique could be adjusted to extract clusters of nodes of the neighbourhood traces, i.e. groups of nodes which are dense. As the algorithm considers noise, all nodes that were not assigned to any cluster are returned as group of noise points. The clustering output can be controlled via the two parameters ϵ (*epsilon*) and $minNumPts$. The parameter ϵ is used for the density criterion, to decide whether two nodes are density connected or not. The distance of two nodes is derived

as geodesic (great-circle) distance d:

$$d = r * \Delta\sigma \tag{4.8}$$

$$\Delta\sigma = \arccos\left(\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos\Delta\lambda\right) \tag{4.9}$$

whereas r is the radius of the sphere and $\Delta\sigma$ is the angle that spans between the two lines that run through the origin of the sphere and the two points on the surface of the sphere. The latter can be determined with the help of the two spherical points $P_1(r, \lambda_1, \phi_1)$ and $P_2(r, \lambda_2, \phi_2)$. As the geodesic distance is proportional to the angle, it is sufficient to use this as density constraint. Thus, ϵ (in degree) should define the maximum angle between the lines through the two points. The minimal number of points ($minNumPts$), i.e. minimal number of close neighbours, defines if a node is a core node.

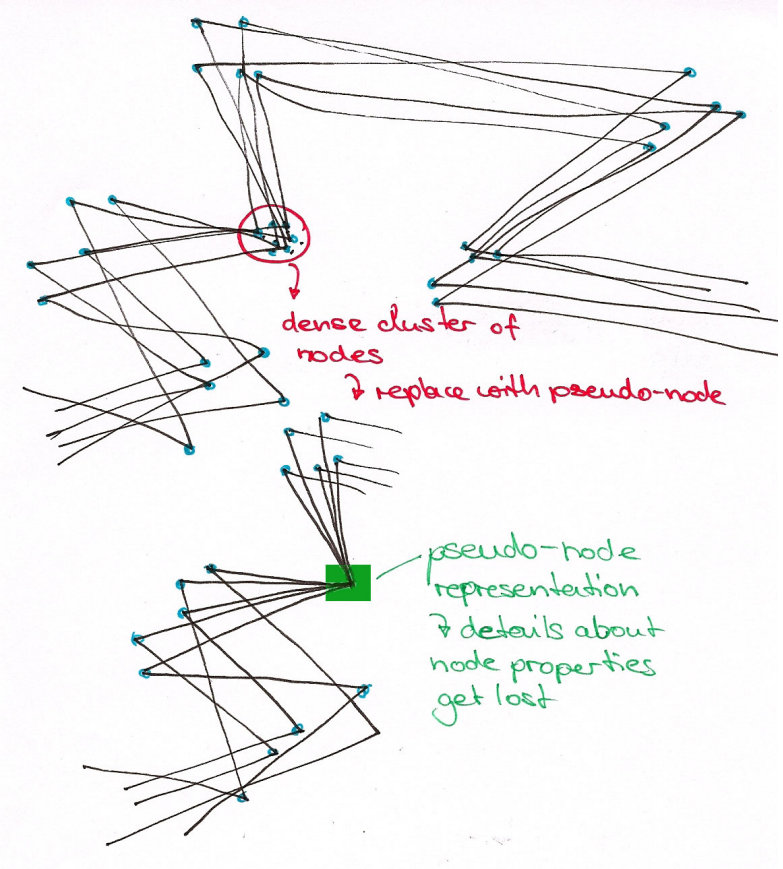


FIGURE 4.4: Draft "pseudo-nodes": dense clusters of nodes can be condensed and replaced by "pseudo-nodes". The extent of the rectangle indicates the variation of node position.

The extracted clusters could then be highlighted via size, colour or shape or replace by "pseudo-nodes" (see Figure 4.4). Hence, nodes would have to be grouped respective their closure (distance) to each other. The problem when using pseudo-nodes would be, that the specific distribution within this cluster area would no longer be visible to

the user. In fact, these "pseudo-nodes" could simply represent the average position and maybe the variance of location (with the help of a rectangle). In addition, such clustering methods are only suitable if a minimum number of traces are generated (extracted from the database) and only for some locations. Even for a huge number of traces, clusters become visible for only certain areas, whereas within other regions nodes are relatively equally spaced.
Instead of replacing the whole cluster with a pseudo-node, it might be better to highlight the single nodes and further visualise the extent of the cluster with the help of a rectangle or other shape (see Figure 4.1).
What would be more helpful in terms of visualising clusters is to extract the bundles of edges that start or end in those node clusters. If the majority of edges branches of a cluster of edges into the same direction, this bunch of edges could be represented or at least extended through an arrow that implies the average direction (see Figure 4.5).

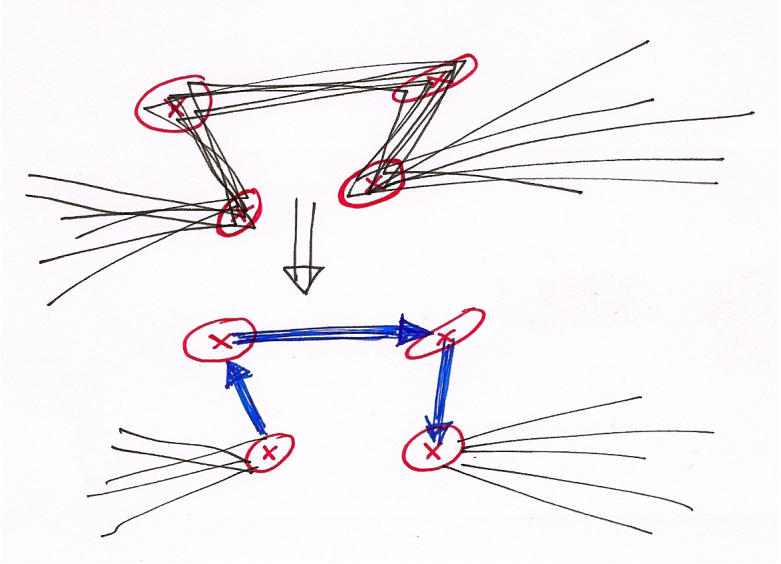


FIGURE 4.5: Draft "edge bundles": dense clusters of edges can be condensed and replaced by arrows that point into the average direction.

4.5 Exploration of Large Data Rooms and Filtering

A lot of network visualisation techniques use filtering as method to reduce the amount of data to plot. As the edges within the traces all have the same importance, it might be difficult to find appropriate filter attributes that do further not lead to gaps within the traces. A reduction would only be suitable, if whole sequences (traces) are removed from the display. This could be done based on the degree of similarity between the template neighbourhood-string and this of the sequence. In this case it would be necessary to first of all define how to measure this similarity.

Talking about the problem of huge amounts of data (traces), one might think of various possibilities to explore huge data rooms. For the representation of contact density potentials and the routes of amino acid sequences it is of much higher importance to keep the overall overview about the distribution of areas of over- and underrepresented contacts and contact clusters. Thus, there is no need for overview and detail or focus and context techniques. A fisheye view might be the most suitable technique, allowing the display of certain area of a screen, e.g. a cluster of nodes, in greater resolution and detail. Nevertheless, this technique would not give any additional necessary information, as it is more of an interest to identify where those clusters are situated and not how they are composed. Even then a histogram would be more suitable to illustrate the distribution of certain residues of certain type within a selected area of the map (see Figure 4.6). Another histogram could show the distribution of over- and underrepresented contacts within the same area, thereby plotting the distribution of LOSs within the selected angle range in comparison to the whole map.

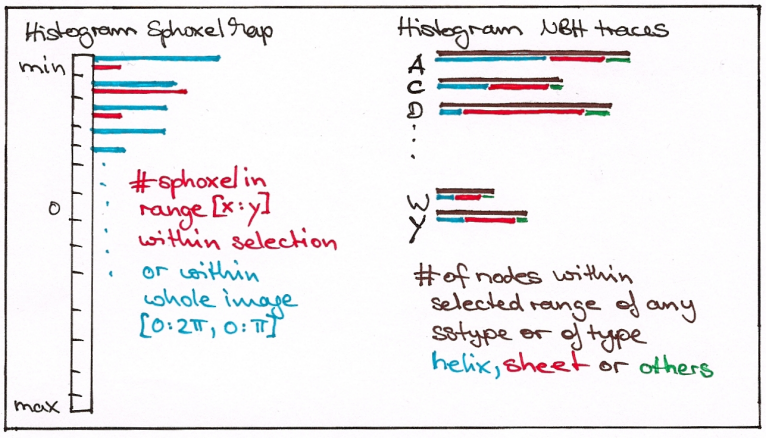


FIGURE 4.6: Draft "histogram": left: histogram of the LOS distribution within the whole map or a selected range of λ and ϕ, right: histogram of the distribution of residues of certain type within the same selected range [λ,ϕ].

What is much more suitable, than context and detail techniques, is a manipulable representation that offers the possibility to scroll, particularly to move the projection this way that the area of interest lies in the centre of projection. Thereby the overall view will always be visible, as still the whole map is shown.
The use of content based filtering seems much more promising than distortion techniques. Based on filtering the user should be able to analyse the potential distribution and orientation of some residue-residue contacts of specific secondary structure type within certain distance shells. It might be helpful for interpretation to remove outliers of very high or low contact potentials via thresholds. Also the display of amino acid sequences should adjust with reference to the chosen template neighbourhood (sequence) string.

4.6 3D Visualisation

Three-dimensional visualisations can give a good impression and overview about the distribution of contacts in space around the central amino acid including information about the distance (radius), which gets lost during the projection onto spherical surfaces at certain shells. To produce a 3D visualisation, first of all the LOSs need to be extracted for voxels around the central residue. The scores could be derived via queries similar to these of section 4.1, but including cartesian instead of spherical cooridnates:

- #edges() = SELECT COUNT(*) FROM edges;

- #edges($iRes, jRes, iSSType$) = SELECT COUNT(*) FROM edges WHERE $i\_res = iRes$ AND $j\_res = jRes$ AND $i\_sstype = iSSType$;

- #edges(x, y, z) = SELECT COUNT(*) FROM edges WHERE $x >= xMin$ AND $x < xMax$ AND $y >= yMin$ AND $y < yMax$ AND $z >= zMin$ AND $z < zMax$;

- #edges($x, y, z, iRes, jRes, iSSType$) = SELECT COUNT(*) FROM edges WHERE $i\_res = iRes$ AND $j\_res = jRes$ AND $i\_sstype = iSSType$ AND $x >= xMin$ AND $x < xMax$ AND $y >= yMin$ AND $y < yMax$ AND $z >= zMin$ AND $z < zMax$;

The scores do further need scaling, so that the negative scores lie within a grey-value range $[0 : 100]$, whereby the lowest (most negative) values were mapped onto 100. In contrast, the positive scores were mapped onto the range $[120 : 220]$.
The derived volumes could be visualised in 3D either via volume rendering techniques with the help of a transfer function or iso-surface rendering.

Chapter 5

Implementation

The whole program was integrated into the Software *CMView*, which stands for Contact Map View. Thus, it is possible to apply filtering with the help of the contact map, used to select the residue-residue contact (see Figure 5.1), which is under investigation, and simultaneously extract the neighbourhood describing String (*nbhString*). The visualisation toolkit, used for the analysis of statistical residue contact potentials and orientation dependencies can thereby be started via right click (the drop down menu) on a residue contact within the contact map.

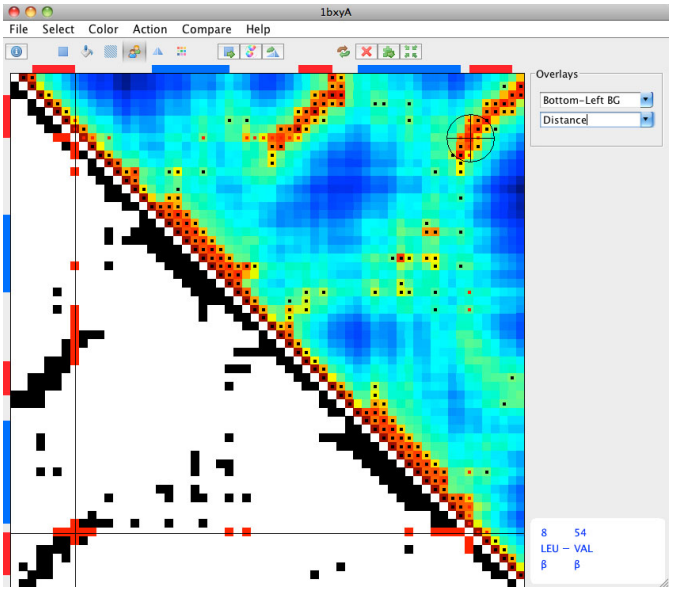


FIGURE 5.1: CMView: neighbourhood selection of contact within contact map (bottom left) and distance map (top right) of a model for the protein 1bxy, protein chain A.

The Software is implemented in *Java* as platform independent application. Besides the standard Java libraries, the library *vecmath* is used for data handling. For the interface and visualisation *java.Swing* and *java.awt.Graphics2D* have been used, which include a

huge amount of methods for the handling of interface elements, drawing methods, colour handling etc. To increase the performance and interactivity, when moving the mouse over the visualisation and defining orientation constraints (via drawing squares), the whole background visualisation is drawn to and saved within a screen buffer. This background includes the sphoxel map, the traces and clusters as well as the longitudes and latitudes. Each time the scene is repainted, the background is loaded from the screen buffer. Only the currently selected angle ranges, the crosshair and the information panel are updated during the mouse movement. The background is only updated, if it needs to be changed due to settings or panning.

The following sections will describe, which ideas of the concept have been realised, starting with the extraction of the necessary statistical background information that will be visualised. This will be followed by the description of the main representations of orientational (geometric) contact potentials (the map of log-odds-scores) and neighbourhood traces. Moreover, various additional options, that can be used to ease the interpretation and definition of orientation constraints, will be introduced. These options include a clustering method, which can be performed on the neighbourhood traces, various histogram views, a colour scaling view and interaction possibilities.
Finally a 3D visualisation of the log-odds-scores (LOS) will be explained, which was realised externally, i.e. independently from CMView.

5.1 Derivation of the Statistical Background Information

As introduced within the concept, the anisotropic contact density potentials are calculated following the Bayes'theorem. Several queries are necessary to compute the LOSs for the sphoxel map representation. These can be applied with the help of the package *java.sql*.

As it appears from the queries listed in section 4.1, four queries which include several constraints are necessary for each LOS. Based on these queries, the calculation of a whole map of LOSs for a contact of certain type is computationally very expensive with an hour per map. The computation time could be broken down to 10 up to 15 minutes with the help of indexing of the table(s) and the creation of temporal tables with intermediate results for the residue and SS types, as well as the ranges of r, ϕ and λ. As this was still not sufficient for interactive visualisation and filtering, the maps for all residue combinations (20×20) and SST (H, S, O or A for any) as well as three different distance shells have be precomputed and saved as *csv*-files. The radii of the shells used for the background information are (as already introduced in subsection 2.4.1) $2.0 - 5.6$Å, $5.6 - 9.2$Å and $9.2 - 12.8$Å. The overall 4800 tables were stored within an archive, which can be used to load the background information for the respective

filtering options very fast.
In comparison to the map of LOSs the NBH traces can be extracted via database queries on demand within maximum 5 seconds. The computation time thereby depends on the number of similar neighbourhoods, found for the given template *nbhString*.

5.2 Sphoxel-Map Representation

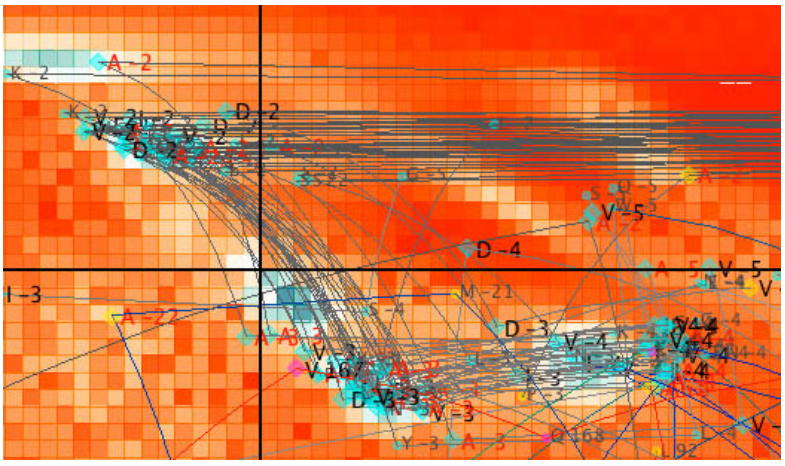
As indicator for the residue contact probabilities at specific positions around the central residue, so-called log-odds-scores have been computed. The positions are defined by the two dihedral angles $\lambda(lambda)$ and $\phi(phi)$ and the *radii* restricting the shell, which can be either close, middle or far shell. The user can switch between those shells with the help of a slider. Based on the mapping of the computed LOSs onto a spherical surface, different map projections have been implemented:

- equidistant cylindrical Platé Carreé projection
- pseudo-cylindrical KavrayskiyVII projection
- orthographic azimuthal projection

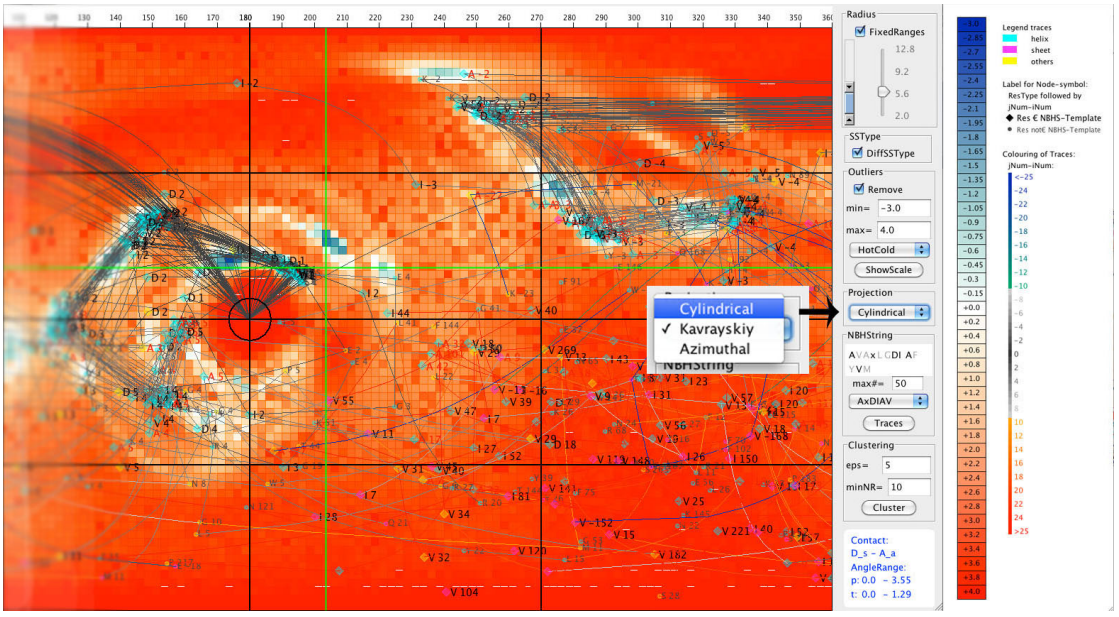
The user can switch between these projections interactively, at the same time changing the rulers, showing exact angle values (see Figures 5.2). Not only the idea of rulers and a cross-hair, but also the concept of longitudes and latitudes has been implemented. Further, all exact information about the selected contact (about residues I and J) as well as the mouse position (coordinate $[\lambda,\phi]$) are shown within a small information panel at the bottom of the menu (see right bottom of the interface in Figure 5.6). Respective the idea of colour-coded tiles to represent negative and positive scores, the values (LOSs) are mapped onto one of the implemented colour scales:

- Blue-Red Scale
- RGB Scale
- Hot-Cold Scale

The scale can be selected by the user via a drop-down box within the menu. The default blue-red scale maps negative scores onto blue, whereas the brightness of the colour is determined by its absolute value. Positive scores are thereby mapped onto red with an brightness depending on the value. When using the RGB scale, negative values are mapped onto colours from blue to green and positive scores from green to red. In



(a) Section of (b): Nodes are labeled with the one-letter-code of the amino acid type and sequence distance. Bold, black labels for residues of the $NBHString$, red labels for residues of same type as $jRes$ and smaller grey labels for all others.



(b) Left: SphoxelMap view, Middle: menu, Right: Colour-Scale-View.

FIGURE 5.2: Cylindrical projection of contact potentials between Aspartic Acid and Alanine with a scaled hot-cold colour-scale, i.e. thresholds (-3,+4) were applied. The SphoxelMap representation is superimposed with traces for the template String AxDIAV, including the position of the central residue.

comparison, the hot-cold scale maps negative values onto cold colours only, from blue to green, and positive values onto warm colours, from yellow to red (see e.g. Figure 5.2). The latter both have the advantage that the opacity still could be used for an additional property, e.g. the distribution in depth (radius). A colour-scale-view is provided to the user and can be opened with the help of the menu. This view shows how the LOS values are mapped onto colour (see Figure 5.2) and automatically adjusts, when thresholds or the colour-scale are changed. It further contains a legend, explaining the shape, colour and labels of nodes as well as the colouring of the edges.

As introduced within the concept, the scores are first normalised to a fixed scale [-1:1]

before they are mapped onto colour. Furthermore, cutoffs can be defined by the user via the interface, which is not only useful to remove outliers, but also to use consistent scaling (normalisation) and make different maps for different contact types comparable. Figures 5.2 and 5.5 indicate, how the representation changes, when cutoffs are applied. As the values can be mapped threshold-based or non-threshold-based, it is important to present the user how the values are mapped onto colour. Thereto, the user can open a Scale-View, which represents the colour scale provided with labels representing the LOSs (see Figure 5.2 or 5.3).
Considering the LOSs, the ssType of the contacting (central) residue is included in the query (as explained within chapter 4). To disregard this constraint from the query, the user can toggle on or off a check box.

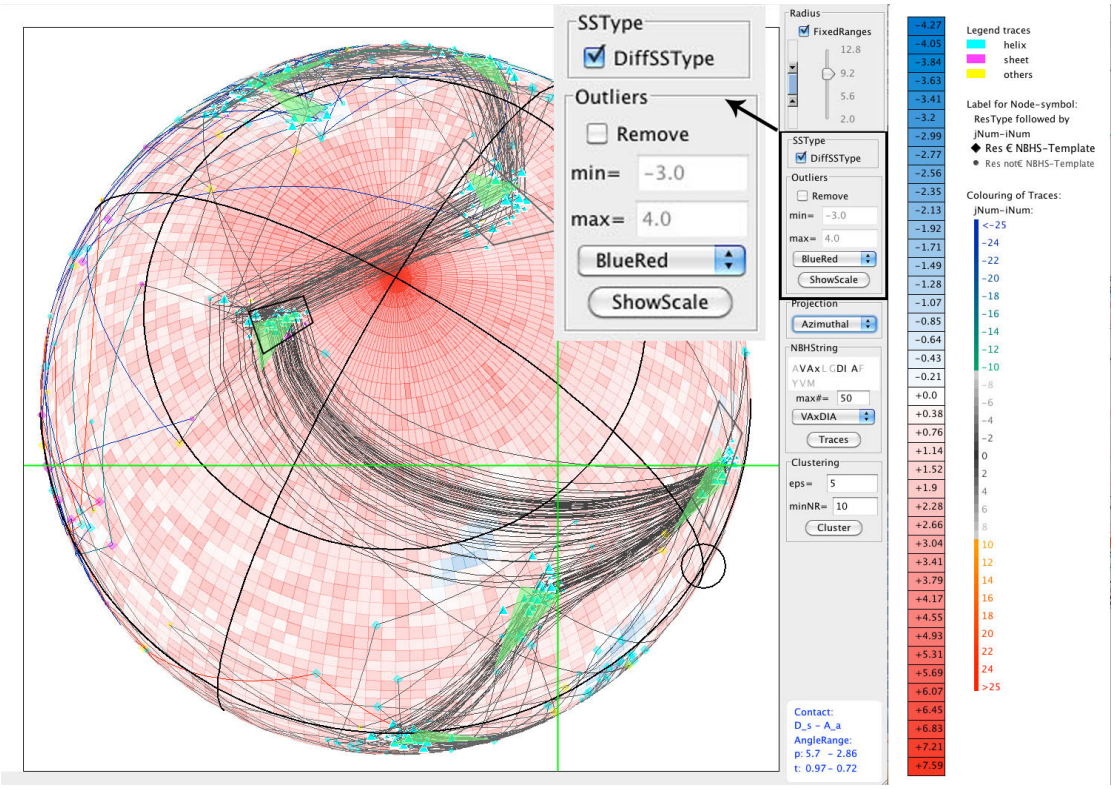


FIGURE 5.3: Azimuthal projection of contact potentials between Aspartic Acid and Alanine with a blue-red colour-scale. The Neighbourhood-Traces are plotted for the template String VAxDIA, excluding the position of the central residue. DBScan was applied and the average direction of outgoing edges from clusters is represented through green arrows. Already selected dihedral angle ranges are surrounded by a grey or black(for the currently selected contact) rectangle. Right: Colour-Scale-View.

5.3 Neighbourhood-Traces

Derivation of template NBHStrings As the original template neighbourhood string (*nbhString*) usually does not provide lots of output traces, it is necessary to reduce the *nbhString* and remove some residues to increase the number of output traces.

The user can therefore toggle on or off certain residues interactively (via clicking on the letters within the *NBHString* selection panel) or via selecting one of the suggested substrings (see Figure 5.4). These substrings are extracted via queries on a database, as described within section 4.3.
The ten best strings, i.e. with the highest support for *iRes* are offered to the user within a drop down box, from which they can be selected. Thus, it is possible to select a template string from the drop down box and further change it manually. The shorter the *nbhString*, the more traces are extracted by the query. Nevertheless, to stem the amount of traces for a well scoring NBH template, a maximal number of traces is extracted and visualised. This number can be adjusted by the user via the interface (see Figure 5.4). The user can further filter the bundles of NBH traces for the secondary structure type (any ssType, helix, strand, loop or others) via another drop down box to analyse the local environments of only these residues that are of some specific *ssType*.

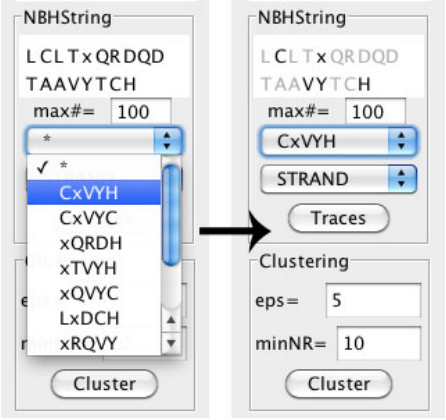


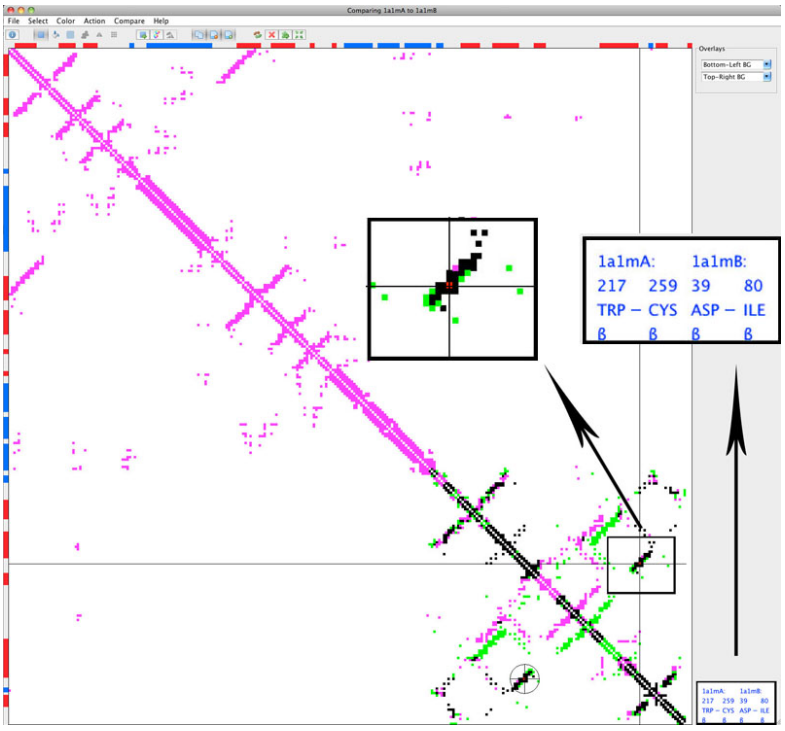
FIGURE 5.4: Selection of template-neighbourhood-String (*nbhString*) via drop-down box. Single residues can be toggle on/off via clicking on them. Enabled residues of the original (full) template nbhString are coloured black and disabled grey.

Nodes of NBH Traces As already explained in the concept, similar neighbourhoods will be visualised via plotting the contacting residues at their position with respect to the central residue (see Figure 5.2). They are connected via edges corresponding their residue number (from lowest to highest number). Thus we get a path (trace) for each extracted neighbourhood.
The contacting residues are plotted as nodes of different shape, either as sphere (default representation) or as rhombus, if the residue is element of the template *nbhString*. If the node is element of a cluster (based on the cluster analysis which will be explained within section 5.4), it is represented as triangle. Not only shape, but also colour and size are used for highlighting. Thus, the triangles (nodes of a cluster) and rhombuses (residues of the neighbourhood) are bigger. The triangles are further highlighted through a white border. The colour is used to represent the secondary structure type of the residue, which might be either helix (cyan), sheet (magenta) or any other SStype (yellow) (see

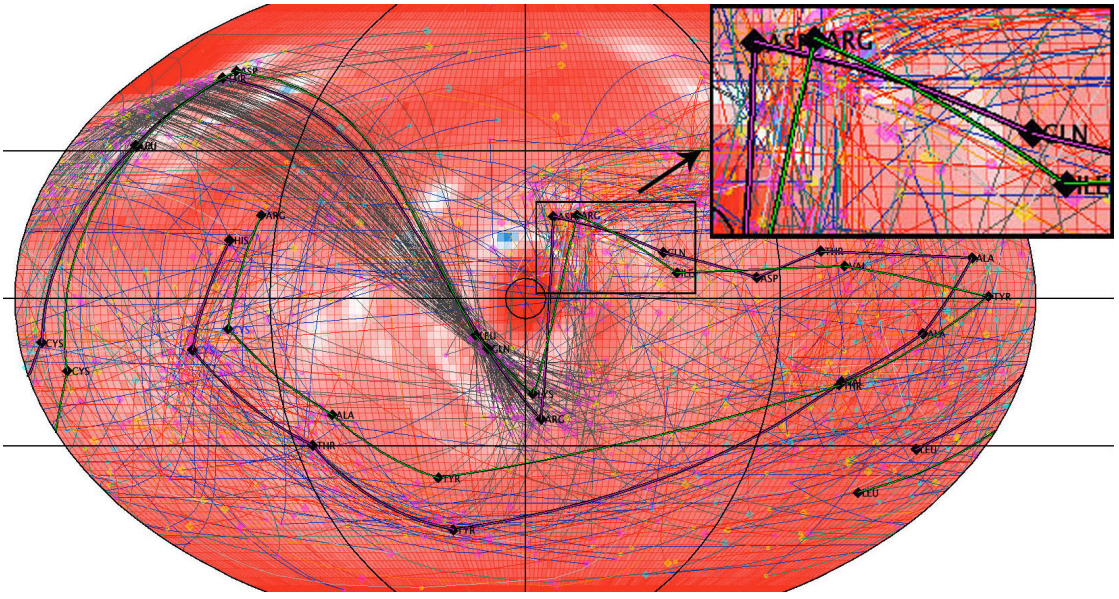
Colour-Scale-View in Figures 5.2 and 5.3). The nodes are further labeled with the one-letter-code of the residue type and the difference in residue number ($jRes - iRes$), which represents the distance within the sequence. Thus, all residues that are situated within the sequence before the central residue are labeled with a negative number and those behind with a positive number. These captions can be switched on or off to reduce the amount of represented information.

Edges of NBH Traces As introduced within the concept, a colour gradient could be used for the edges, to visualise where and how far in sequence the residues of the neighbourhood are situated. Particularly, the edges between residues, which have a sequence distance smaller than zero, are coloured with the cold part of the hot-cold-scale (from blue to green), whereas those with positive distance are coloured with the warm colours of the same scale (yellow to red). This colourscheme is just applied onto the long-range contacts, i.e. those contacts that are far in sequence (i.e. $|jRes - iRes| > 9\mathring{A}$) but close in space. The edges between the short-range contacts are coloured based on a grey scale.
The traces can be plotted in two ways either including the central residue or excluding the central residue (see Figure 5.2 or 5.5). In case the user toggled on the option for the central residue, the node is included in the trace and plotted at the centre of the map projection. The edges are thereby implemented and drawn as geodesics instead of straight lines.

Template NBH Traces The extracted neighbourhood traces represent statistical information. To analyse how far the local neighbourhood of the selected contact complies those statistical spatial propensities, the trace for the template *nbhString* is highlighted and drawn with bold edges (see Figure 5.5), if the option is toggled on.
In case that two models are loaded in CMView for the purpose of comparison, two traces are plotted within the representation (see Figure 5.5). This should help to visualise and analyse the local differences in folding of a specific residue neighbourhood between the two models of the same protein. The bold traces for the two model are coloured appropriate the contacts within the contact map (see Figure 5.5). As contacts, that are contained only within the first/second model are coloured pink/green within the contact map, the related traces are coloured respectively. This visualisation could be used e.g. in the context of CASP (Critical Assessment of protein Structure Prediction), to evaluate the accuracy of predicted structures in comparison with the resolved target structures. Thereby, the two models need to be loaded and aligned first. To show the neighbourhood traces, the user then needs to select a contact, which both models have in common (one of the black contacts).



(a) Contact map of 1a1m chain A and B, whereas common contacts are black, and those contained only in A pink and in B green.



(b) Kavrayskiy projection of contact potentials between Tryptophan and Cystein with a scaled blue-red colour-scale, i.e. thresholds (-4,+6) where applied. The Neighbourhood-Traces are plotted for the template String CxVY, excluding the position of the central residue and the template Traces itself (bold and pink/green for chain A/B). Traces are filtered for sheet-type.

FIGURE 5.5: Possibilities within the comparing mode of CMView.

5.4 Interaction

Change of Views The representation of the LOSs (Sphoxel-Map) as well as of the neighbourhood traces can be changed via interactions, particularly via filtering (changing of query parameters). As mentioned before, the user can change the Sphoxel-Map-colouring via switching between the available colour-scales and the specification of thresholds to remove outliers. Considering the traces, the user can change the nbhString-template used for the query and limit the number of traces to a maximum.
The map view can also be changed via panning, i.e. the user can move the area of interest into the centre of the screen with the help of the mouse or the arrow-keys. The navigation within the cylindrical and pseudo-cylindrical map is therefore limited to one dimension, allowing to rotate the sphere around the azimuthal (vertical) axis. In contrast, when using the azimuthal map projection, rotation can be performed around the centre of the sphere (see Figure 5.3). This allows us to put any point (also a pole) into the centre of the screen.

Derivation of Orientation Constraints Besides filtering and navigation, interaction is also necessary to define certain orientation constraints. Thereby, the user needs to select a range with the help of the rectangle or cluster selection tool. The latter will be explained within the next paragraph. When using the rectangle selection mode, the user can define a range via clicking on a point and dragging the mouse, while the rectangle becomes visible. To support the user during this selection procedure, the view is overlayed by a crosshair marking the mouse position and the exact angle position is shown within the *information panel* (see right bottom of the interface in Figure 5.6) while the mouse moves over the map. The position, at which the user clicks is shown on the left and the dragging position on the right of the information panel. This panel also holds information about the currently selected contact, i.e. $iResidue$ and $jResidue$ and their $SSType(secondaryStructureTypes)$.
In case, that a restriction range has already been assigned for the currently selected contact, this is replaced by the newest selection. All assigned ranges are visualised within the map via (projected) rectangles (see Figure 5.3), whereas the rectangle for the currently chosen contact is drawn in different colour (black and the rest in grey). A selected range can be deleted via right-click onto the selection and choosing the option *DeleteSelectedRange*. It is further possible to open a histogram view for a specific range via the drop-down menu (see paragraph Histogram View).
As described before, the Sphoxel-Map as well as the neighbourhood representation both depend on the selected contact within the contact map of CMView. Thus, if another contact is clicked within the map, contact constraints and template nbhString change. Thereby, both representations change respectively, as they rely in on these properties,

while so far selected ranges are saved for certain contacts. This way it is possible to define various orientation constraints after each other with the help of certain background information depending on the contact.

5.5 DBSCAN on Neighbourhood-Traces

Clustering Method As suggested within the concept, the DBSCAN algorithm was implemented as clustering technique to extract clusters of nodes of the neighbourhood traces. The clustering output can be controlled via the two parameters ϵ (*epsilon*) and $minNumPts$ by the user (see Interface Clustering, e.g. within Figure 5.6). As described within the concept, ϵ should be declared in degree, preferably within a range from 3°to 8°. The minimal number of points ($minNumPts$) should have a value within the range of $[1 : numberOfTraces]$.

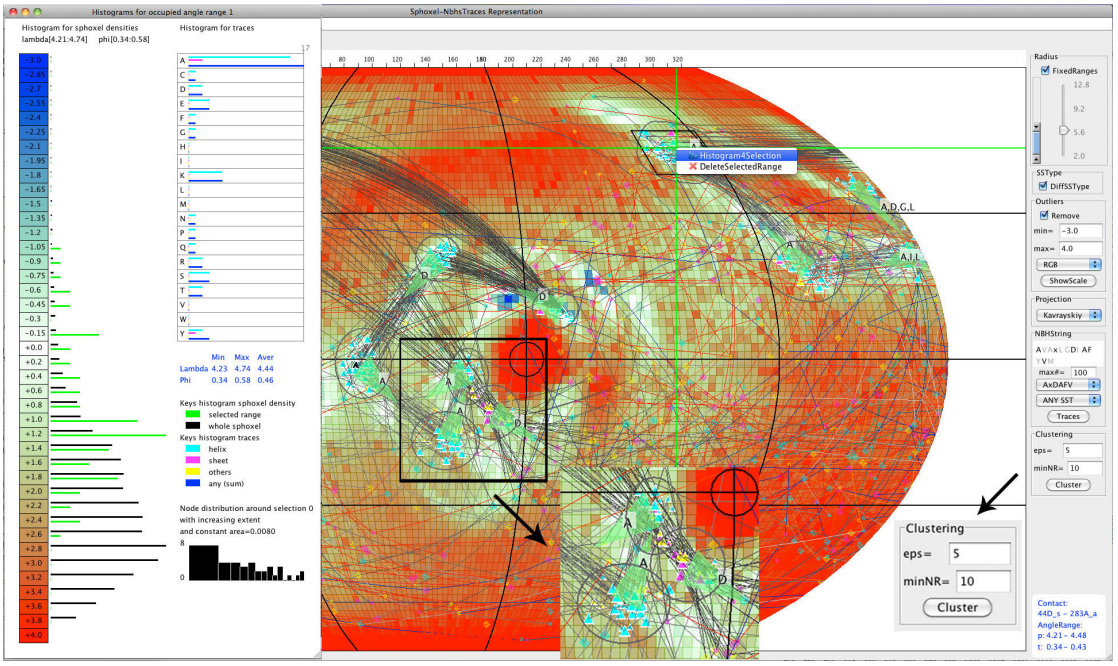


FIGURE 5.6: Kavrayskiy projection of contact potentials between Aspartic Acid and Alanine with a scaled RGB colour-scale. The Neighbourhood-Traces are plotted for the template String AxDAFV, excluding the position of the central residue. DBScan was applied, resolved clusters are highlighted, predominant residue types are plotted on semitransparent white ground and the average direction of outgoing edges from clusters is represented through green arrows. The currently selected dihedral angle ranges is surrounded by a black rectangle. The histogram view for that range is opened via selection within a drop-down menu. Left: Histogram-View for selected angle range.

Visualisation of Extracted Clusters The clustering method should help to extract orientation constraints and to determine ranges for the dihedral angles *lambda* and *phi*. For this purpose, all nodes that are contained within a cluster are highlighted.

As already mentioned in the last paragraph, they are drawn as triangles, surrounded by a white border. To better convey the range of λ and ϕ they comprise, each cluster is surrounded by an ellipse with an extent with respect to the angle ranges (see Figure 5.6).

Based on this cluster result, a further analysis of the variation in node positions within the cluster is performed, i.e. the minimal, maximal and average values of λ and ϕ are printed within the histogram view (see next paragraph). In addition, the average direction of the outgoing edges is computed for each cluster. For this purpose first of all the direction of each edge, which starts in any of the nodes contained in the cluster and ends within a node of higher residue number, are determined. To circumvent the influence of outliers, i.e. single edges that run into completely different direction than the majority, the set of edges is divided into four sets, one for each quadrant. The average edge direction will be computed based on that set (for one quadrant), that contains the majority of edges. The average direction is determined with the help of the angle bisectors, i.e. via vector addition and normalisation. It is displayed via thick green semitransparent arrows (see Figure 5.6).

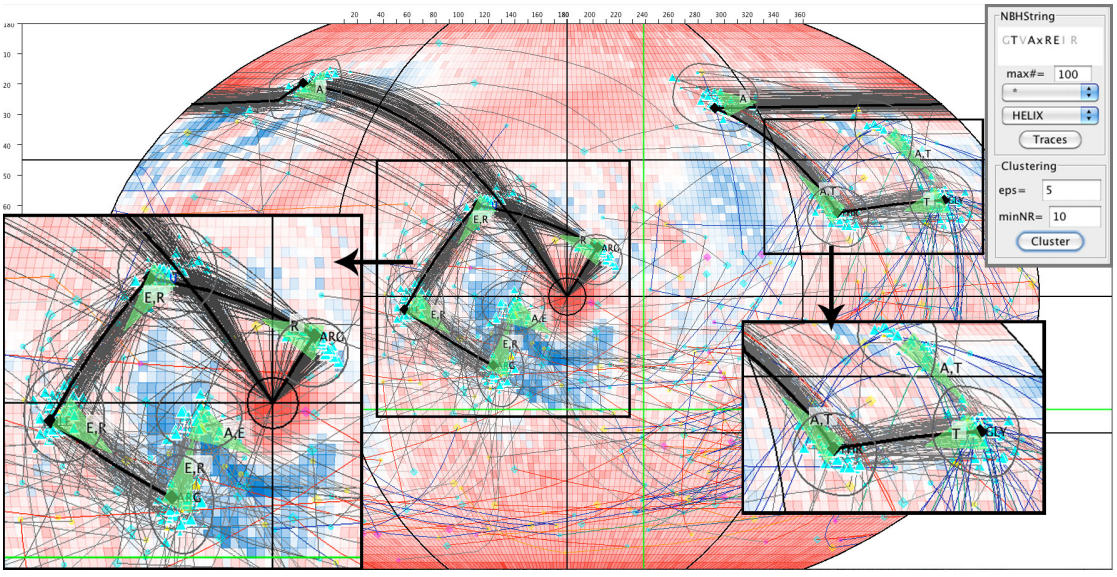


FIGURE 5.7: Kavrayskiy projection of contact potentials between Leucine and Glutamic Acid in Blue-Red colour-scale. The Neighbourhood-Traces are plotted for the template String TAxRE of type Helix, including the position of the central residue. The template NBHString is derived for the contact from protein 1kx5. DBScan was applied, resolved clusters are highlighted, predominant residue types are plotted on semitransparent white ground and the average direction of outgoing edges from clusters is represented through green arrows.

Besides the average direction, also the predominant residue types of each cluster are visualised (see Figure 5.7). Along with it, the average number of occurrances of a cluster, as well as maximal number are derived first. The two values are used to define a threshold for the minimal number of occurrances (t_{occ}) within the cluster, which is

given by the mass centre between the two values. The set of predominant residue types consists of all residues, which have a number of occurrances greater than the threshold (t_{occ}).
The derived set of clusters allows visual support to determine orientation constraints on the one hand side, but can also be used to automatically derive angle ranges (see section 5.4 Interaction). For the latter, the user has to switch to the cluster-selection mode. A click on any cluster then automatically sets the angle range for the currently selected contact to the range stretched by the minimal and maximal positions of nodes occurring within the cluster. The rectangle that defines the dihedral angle range will than span around the cluster.

5.6 Histogram-View

The histogram-view can be opened for any determined dihedral angle range via right-click within a rectangle and choosing the option *Histogram4Selection*. The histogram-view contains various information and different histograms for the LOSs as well as neighbourhood traces, based on the selected range (see Figures 5.6 and 5.8).

Histogram of LOSs The histogram for the spatial propensities plots the distribution of LOSs within the selected angle range (green, see legend within histogram-view) in comparison to the whole map (black), i.e. $\lambda[0 : 2\pi]$ and $\phi[0 : \pi]$. Thereby, the numbers of occurring pixels, which lie within a certain value range, are determined and plotted via bars next to the colour-scale view. The length of the bars is proportional to the ratio of occurring number and maximal number within the selection (for green bars) or whole map (for black bars). Thus, it is not possible to compare the exact numbers of occurrances within the selected range and the whole map, but the general distribution of them. For easier interpretation, the bars are not just plotted along a labeled axis for the LOSs, but along a colour-scale that is superimposed with the exact LOS values. If the colour scale coding was changed within the main interface, the colour scale of the histogram changes respectively. The frequency of elements of the histogram helps to determine how favourable this dihedral angle range is.

Histogram of NBH Trace Nodes The second histogram (see right upper histogram of Figures 5.6 and 5.8) is based on the neighbourhood traces. Within that histogram the numbers of occurring nodes of certain residue type and certain secondary structure type within the selected range are plotted as bars. Similar to the first histogram, the length of the bar is proportional to the maximal occurring number, which is plotted at the top right of the histogram. The bars are coloured respective the secondary structure type, similar to the colour coding used for the nodes itself. This histogram can be used

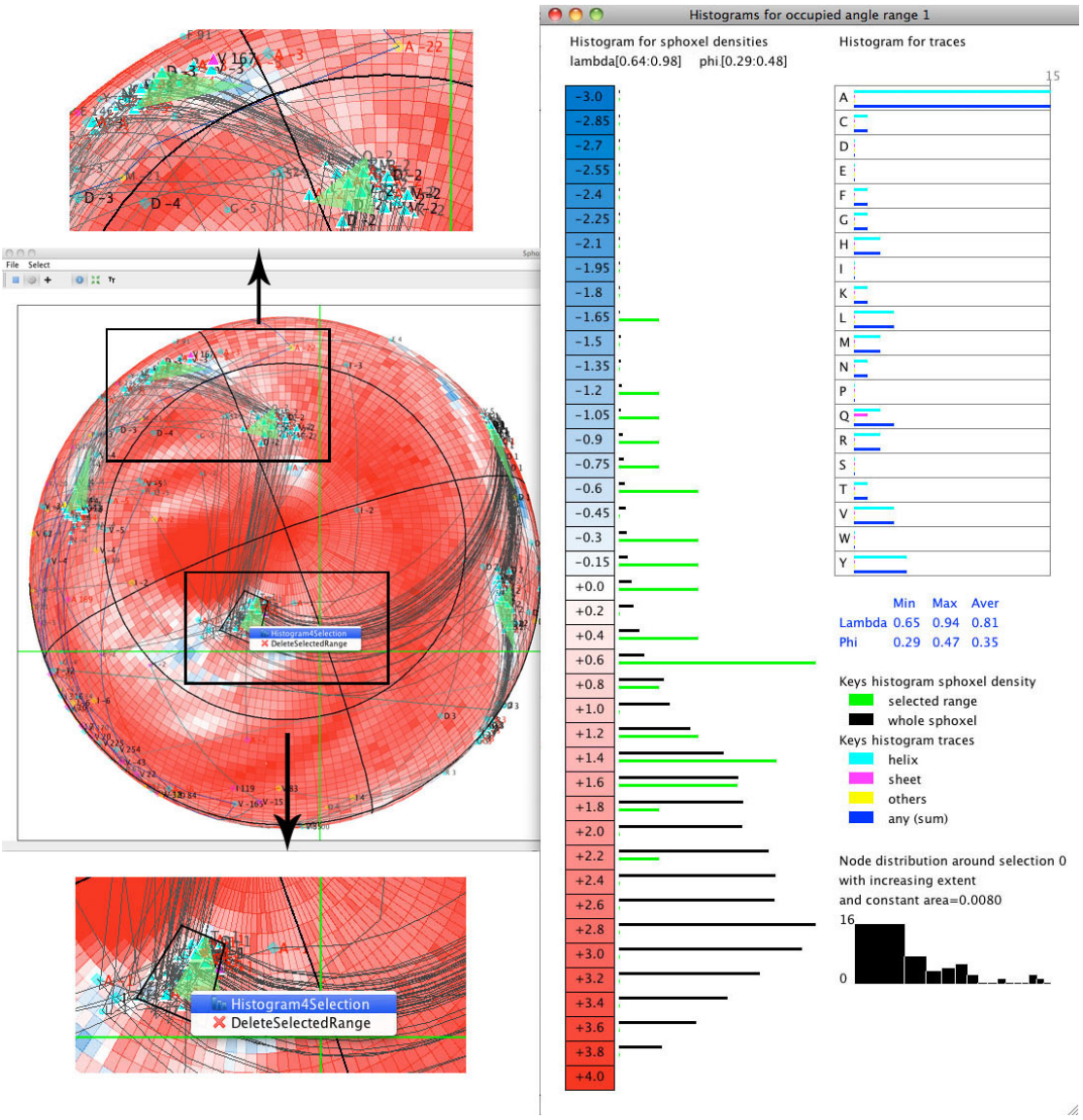


FIGURE 5.8: Azimuthal projection of contact potentials between Aspartic Acid and Alanine with a scaled blue-red colour-scale. The Neighbourhood-Traces are plotted for the template String AxDIAV, excluding the position of the central residue. DBScan was applied, resolved clusters are highlighted and the average direction of outgoing edges from clusters is represented through green arrows. The currently selected dihedral angle ranges is surrounded by a black rectangle. The histogram view for that range is opened via selection within a drop-down menu. Right: Histogram-View for selected angle range. Trace-Histogram shows strong preference towards Alanine contacts with selected range.

to determine the predominant residue types within the cluster. Thus it can be used to check whether the selected angle range poses a good constraint for the currently selected contact between $iResidue$ and $jResidue$.

Distribution of Trace Nodes Below the second histogram, a table gives information about the variance in node positions, that lie within the angle range. The table includes the minimal, maximal and average values for λ and ϕ.

The third histogram visualises the distribution of nodes within and around the close surrounding of the selected range. For this purpose, the number of occurring nodes within certain annuli (the inner annulus is a sphere) is counted. The radii for the annuli are set this way, that the area stays constant (equal to that of the inner sphere). The numbers are plotted within the histogram via bars, whereas the high (length) of the bar is proportional to the number and the width corresponds to the radii, i.e. the bar starts at the inner and ends at the outer radius of the annulus.

5.7 Further Options

The Sphoxel-Map overlayed with the traces can be saved as image (png-file), but also exported separately as csv-files. The former contains the LOSs within rows (for ϕ) and columns (for λ), whereas the latter contains columns for $id_{Protein}$, $num_{iResidue}$, $num_{jResidue}$, θ, ϕ, $residueType$ and $ssType$. The number of rows is thus equal to the number of nodes.
It is further possible to save all settings and selected ranges within a file, that can be loaded later on to continue working on the same protein model. These settings include all determined dihedral angle ranges for selected contacts as well as all user defined parameters, like the chosen $nbhString$, $maxNumTraces$, clustering parameters ϵ and $minNumPts$, the $centreOfProjection$ (for the panning position) and the $thresholds$ to remove outliers.

5.8 3D Visualisation

The three-dimensional visualisations can give a good impression and overview about the distribution of contacts in space around the central amino acid. Therefore, I decided to implement two 3D views at least externally, i.e. independently from CMView, which can be used in addition to the two-dimensional view. The scores have been derived for voxels around the central residue within a range of +-8Å in each dimension. The resolution (stepsize of x, y and z) was thereby set to 0.5Å. Before the volume was exported, the scores were scaled as described within the concept. The volume was exported as *tiff-volume* and *raw-volume*, which can be imported with various different software.

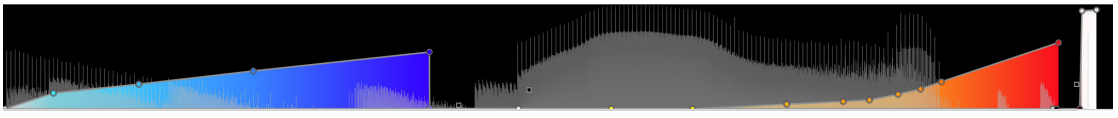
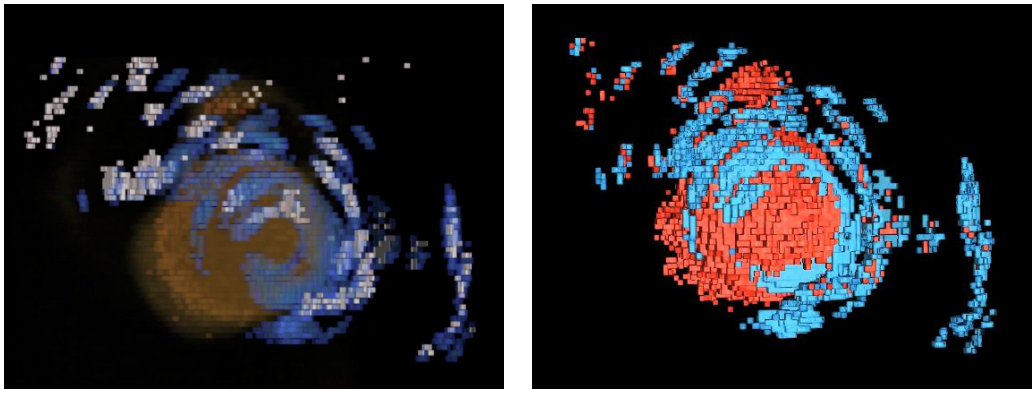


FIGURE 5.9: Transferfunction for Volume Rendering of LOSs.

I used Osirix to load the exported volume, to perform volume rendering and to produce iso-surfaces. For the former I created a new transfer-function (see Figure 5.9) to map the scaled grey values, representing the LOSs respectively. Similar to the Blue-Red-Scale of the 2D view, the transfer-function maps negative values onto blue colour and positive values onto red tints, whereas transparency decreases with increasing value. Maximal grey values are mapped onto opaque white colour. Figure 5.10(a) shows an example of a volume rendering of LOSs for a specific contact with the help of the transferfunction. The LOSs-volume also includes the raw counts (in addition to the LOSs). Thereby, all voxels within that the number of raw counts exceeds a threshold are set to the maximal grey value (255).
The iso-surface rendering gives quite similar results (visualisations) (see Figure 5.10(b)). The iso-surface for the LOSs is computed via thresholding, whereas the first threshold is set for the negative scores and rendered with semitransparent colour. The second threshold is set for the positive scores, which are rendered with another, but opaque colour.



(a) Volume Rendering (b) Iso-Surface

FIGURE 5.10: 3D representation of contact potentials between Lysine and Arginine. Positive LOSs are coloured red and negative ones blue.

Chapter 6

Validation & Evaluation

For the evaluation and validation, two concepts have been designed. These will be presented first, before the accomplishment and results of at least one of them will be presented. Afterwards, the possibility to make use of orientation constraints for the purpose of scoring functions will be discussed as well as further possibilities to improve the usability of the implemented tool.

6.1 Concepts

Both concepts rely on resolved structures as basis for the evaluation of the predicted model. A suitable and comprehensive dataset of predicted models is available in the context of CASP [MFK+09], which is a worldwide experiment for protein structure prediction. Within that contest a number of about 120 recently resolved protein structures are offered as target in form of their primary sequence. Thereby, the three-dimensional structures have not been published yet and remain unknown to the participating groups. Thus, they can test their methods to identify a proteins three-dimensional structure from its amino acid sequence. The submitted predicted models (by different groups) as well as the rankings and evaluations are available to the participants at the end of each CASP competition. For the *ranking* of the submitted models scores like *global distance test* (GDT) score or *root mean square deviation* (RMSD) are used. These scores define how close a predicted model is compared to the resolved structure, whereas the former is intended as more accurate. This is because the RMSD is more sensitive to outlier regions. The GDT identifies sets of residues, which deviate from the target by less than a specified distance cutoff, thereby using many superpositions. The GDT score is therefore defined as the largest set of C_α atoms within the model deviating from the target by no more than a specified distance cutoff.

The **first concept** is based on the evaluation of the visualisation tool with respect to its *suitability to define the quality of a model* in a subjective way. Thereby, a set of different submitted models, two for each selected target, will be used as basis. The n targets and respective $2 * n$ models will be divided into three classes respective the difference in scoring values between model A and B. The three classes comprise models that have a scoring difference (i.e. a difference in GDT-score) either:

- greater than 70% (i.e. delta-GDT-score>0.7)
- greater than 50% and smaller than 70%
- or greater than 30% and smaller than 50%.

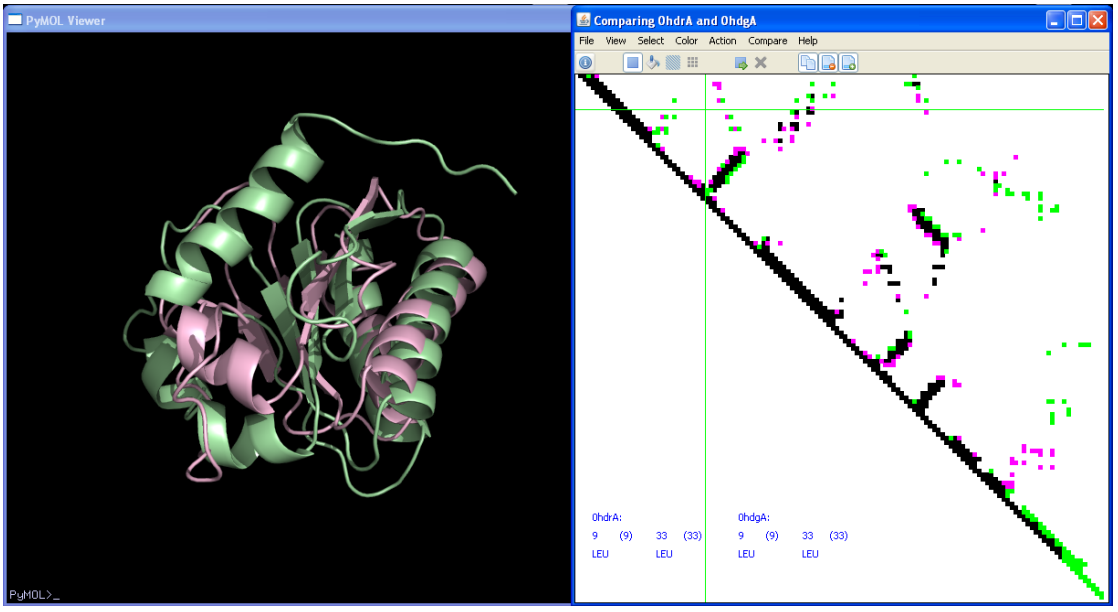


FIGURE 6.1: Comparison Mode: first Model is superimposed by second model. Black contacts are element of both, pink only of the first and green only of the second structure.

Several users will have to evaluate these models, i.e. they have to decide which of the two models is the better one. For this purpose, the two models for one target need to be loaded within CMView, which offers the possibility to compare two protein structures (see Figure 6.1). After loading the second model (structure), it is necessary to align both models.
The subject will then have to look at the "Sphoxel-Map" and "Neighborhood-Traces", which are superimposed by the trace of the "Neighbourhood Template-Traces" for both models. They are constraint to analyse the views (neighbourhoods) of several (about 10) contacts. For each selected contact and its neighbourhood-presenting trace (two traces for model A and B), the user should rate how good the trace fits into the statistical distribution and orientation of similar neighbourhoods. Based on the analysis of the statistical geometric assembly within the neighbourhood/environment (i.e. the bundle

of traces) for several contacts, the user should be able to decide which one of the two models is closer to the target structure, i.e. it has a higher GDT-score. For the better model more traces should be in accord with the background information.
For the analysis the subjects should not only document which contacts, *NBHStrings* and *filtering options* they used, but also the properties of the selected contacts. Contacts might be of type *"common"*, i.e. they are element of both models, or contained just within one of the two models (*"single"*). Contacts can sometimes be assigned to a SST, e.g. helical (H) or strand (sheet) (S) contact. It is further possible to differentiate SR and LR contacts, i.e. contacts between residues that are close or far in sequence.
Finally the numbers and ratios of correctly classified models can be derived for each subject and each model. This evaluation method should give some indication of the suitability for the evaluation of a predicted model. It should thereby help to answer the following questions:

- Does the visualisation of the statistical geometric orientation help to discriminate good and bad models?
- How certain can this differentiation be conducted?
- Are there certain contact types that improve the chance of correct evaluation, e.g. long-range vs. short-range contacts, helix vs. sheet contacts, common contacts vs. "single contacts"?

Referring the last question we can assume, that the local environments of LR contacts contain more information and are more suitable than SR contacts. Another hypothesis can be made with respect to the containment of the contact within the two models. It might be that "single" contacts, only contained within model A or B, mislead to wrong conclusions.

The **second concept** was designed to derive, whether orientation (geometric) constraints can be used to improve the folding simulation and thus prediction of three-dimensional structure. So far, solely distance constraints have been used for sampling methods, to decrease the conformation space and thus number of possible conformations that need to be scored via energy functions. The question, whether the extension of these constraints through orientation constraints improves the predicted model, arises. To decide whether the models are closer to the original structure, they would have to be computed based on distance constraints alone and based on distance and orientation constraints. To score the quality of the predicted models RMSD or GDT scores could be used. The RMSD-value should be smaller and the GDT score accordingly higher for models derived based on both constraints.
The modelling via energy functions can be applied with the help of tools like TINKER. Thereby, several parameters and distance constraints can be handed over to simulate

the folding.
The distance constraints are given by the contact map, which defines a maximal distance between the contacting residues. In contrast, the orientation constraints need to be defined by the user with the help of the "Sphoxel Map and Neighbourhood Traces Visualisation Tool". Thereto, the user needs to analyse the statistical preferred orientation of several contacts. The propensity of a contact between two specific residues being located at some position, is represented through the colour-coded visualisation of the LOS's ("Sphoxel-Map"). The statistically preferred orientation of the whole neighbourhood is thereby visualised via the "Neighbourhood Traces". These two visualisations as well as the histograms can be used to derive the preferred orientation, i.e. dihedral angle range within that the contact will most likely be located. Based on these steps the user should derive several orientation constraints, particularly one preferred orientation range for each selected contact. The angle ranges can be dedicated for each selected contact via interacting (drawing) within the "Sphoxel Map". These derived constraints would then have to be handed over to TINKER, or some similar tool, together with the distance constraints, given through the contact map.
To remove user dependent influences several subjects would have to conduct the same test for the same set of PDBs. In addition, it would be interesting to analyse how many constraints, i.e. for how many contacts, are necessary to improve the model. There might be a minimal number of necessary constraints to get a significant change of the RMSD. Also we can assume, that similar to distance constraints, a maximum number of orientation constraints is sufficient and further constraints do not lead to any significant improvement of the model. The limitation of distance constraints and reduction to a minimal subset of contacts was analysed by Lappe et al. [SDS+09] (recall section 2.4.1). Similar to the influence of the choice of distance constraints, the choice of contacts for which the orientation constraints are defined might have an impact on the improvement of the model. Thus, it might be more effective to derive orientation constraints for long-range contacts than for short-range contacts.
This evaluation method should give some indication of the potential in improving simulated foldings with the help of statistically derived orientation preferences between residues and their neighbourhood. It should thereby help to answer the following questions:

- How many orientation constraints are necessary to improve the model?
- Does the improvement increase linearly with the number of constraints or is it convergent up to a maximal number of constraints?
- Does the choice of contacts have an impact on the improvement?
- Which contact types are most suitable to derive constraints?

6.2 Results

The second concepts requires the integration of orientation constraints into an energy function used for the protein modelling, e.g. with the help of modelling software like TINKER. As this could not be managed and prepared in time, only the first concept was feasible.
For the accomplishment of this concept, a set of 20 targets has been chosen. Particularly two models have been selected for each target and anonymised. Thus, they could not be identified by the users and allocated to the available CASP8 results. As mentioned within the concept description, the targets were classified with respect to the difference in GDT-values between the two selected models (see Table 6.1). The models for 4 of the selected targets have a difference in GDT-score of more than 70%, 8 of more than 50% and another 8 of more than 30%.

ClassID	TargetIDs	ΔGDT
1	T0396(75), T0423(78), T0523(74), T0594(72)	$> 70\%$
2	T0431(57), T0440(53), T0451(51), T0469(51) T0473(53), T0492(53), T0493(56), T0499(58)	$> 50\%$
3	T0389(32), T0401(35), T0411(32), T0417(32) T0427(31), T0457(32), T0462(32), T0474(33)	$> 30\%$

TABLE 6.1: Classification of selected CASP targets respective difference in GDT score. The exact values for ΔGDT are put into brackets after each targetID.

The targets have been analysed by 6 subjects, whereas the first 10 targets have been handled by subjects 1-4 and the other 10 targets by subject 1,5 and 6. The subjects were asked to analyse the local environments of about 10 contacts and to try to derive for each of them, which of the two NBHs coincides better with the statistical background information. For the majority of selected contacts a decision for either model A or B could be made, whereas for some contacts (about 17%) it was not possible to make a choice. This is because some environments are quite similar, i.e. either both fit into the statistical background information or both deviate drastically compared to bundles of traces.

Table 6.2 gives an overview about the correctly differentiated models (targets). The targets within table 6.2 are grouped with respect to the target classes and colour-coded in relation to the performance, i.e. how good the target was handled. Thus, the table emphasises that models with higher difference in GDT-score ($\Delta GDT > 50\%$) on average have been differentiated better than models with a difference about 30%. On average 6.7 of the 10 targets were processed successfully, i.e. almost 70% of the better models were correctly identified.
The percentage of correct estimates is not only presented per target (within table 6.2),

TargetID	S1	S2	S3	S4	S5	S6	Correct Identified (%)
T0396	✓	✓	✓	✓			100
T0423	✓	✓	✓	✓			100
T0523	✓				✓	✓	100
T0594	✓				×	×	33
T0431	✓				✓	✓	100
T0440	✓	✓	×	✓			75
T0451	✓				✓	✓	100
T0469	✓	×	×	×			25
T0473	✓	✓	✓	✓			100
T0492	✓	×	×	✓			50
T0493	✓	✓	✓	✓			100
T0499	×	×	×	×			0
T0389	✓	✓	×	×			50
T0401	×				×	×	0
T0411	✓	×	×	×			25
T0417	✓				✓	✓	100
T0427	✓				✓	✓	100
T0457	×				✓	✓	66
T0462	✓				×	×	33
T0474	✓				✓	✓	100
Correct Identified (%)	85	60	40	60	70	70	

TABLE 6.2: Overview about results: checkmarks (crosses) illustrate correct (wrong) differentiated targets.

but also per subject. These numbers illustrate the user dependent differences, which clarify that some persons found it easier to interpret the views than others. The comparatively high percentage of correct estimates for subject 1 (S1, which was myself), can be traced to the comprehensive knowledge about the software. Concerning the other users, most decisions about the model quality have been made solely based on the view of the statistical background information. Features like the histogram view have been used very rarely if at all, but do support the decision for the statistically better neighbourhood of a contact. This leads to the assumption that a more comprehensive introduction and use of the available features, like clustering, filtering and histogram views, would lead to a better certainty regarding the evaluation of models.

One of the assumptions made before was that the choice of contact has an impact on the analysis of the contact environment, i.e. the NBH of residue I of contact (I_J). For the examination of this hypothetical dependencies, all decisions (for each selected contact) were sorted respective different contact type properties. The analysis of the containment property ("common" or "single" contact) showed no obvious coherence. In relation, wrong decisions have been made for neighbourhoods of "common" contacts just as much as for "single" contacts. The same applies to right estimates about the model quality.

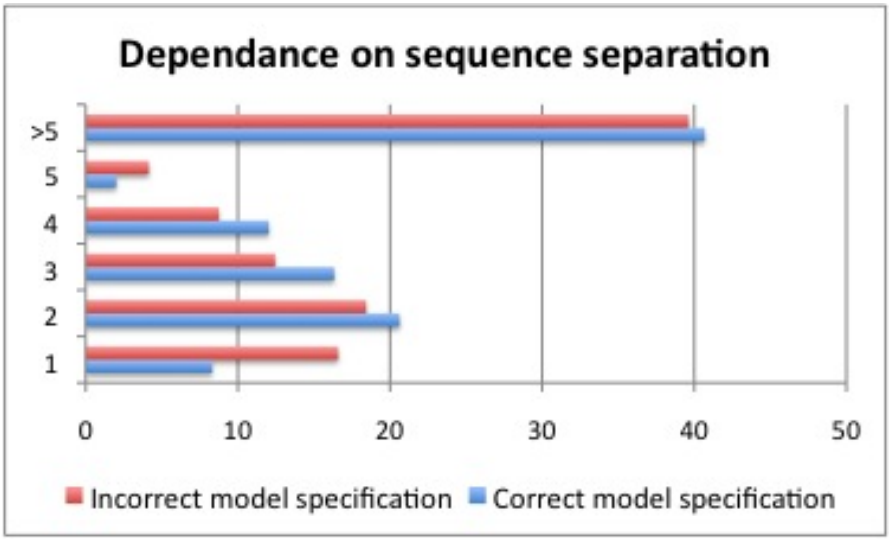


FIGURE 6.2: Dependance of evaluation certainty on the sequence separation between the contact residues: the percentage of contacts with $|iNum - jNum| = n, n[1:5]$ or $n > 5$ is plotted for correct (blue) and incorrect (red) model specifications for the respective neighbourhoods.

In comparison, the contrasting juxtaposition of right and wrong estimates against the sequence separation of the contacting residues showed some tendencies towards the incapability of direct neighbours. The sequence separation is defined by the difference in residue numbers between the two contacting residues $iRes$ and $jRes$. It is used for the distinction of short-range and long-range contacts, which was explained in subsection 2.4.1. The chart within Figure 6.2 plots the percentage of right and wrong estimates for contacts with different sequence separations of about 1 up to 5 or more than 5. It shows that contacts, which have a sequence separation of about 1 (direct neighbours) on average lead twice as often to a wrong than to a right decision. These contacts can be assumed to be less suitable for the prediction of the quality of the local contact environment. All contacts that have a sequence separation greater than one seem to be likewise/equally convenient. Thus, we can not extrapolate from the results of chart (Figure 6.3) to an advantage of choosing LR contacts, which have a sequence separation of more than nine.

When looking at the percentages of correct estimation for all targets within table 6.2, it becomes obvious that some targets were easier to handle than others. For a further analysis of right and wrong decisions about the better model that have been made by the participants, the native structure of the target can be loaded and plotted as trace as well. The native structure is usually terminated by NMR or Crystallography and used as reference state for the evaluation of the submitted model. The neighbourhoods of selected contacts, which lead to a misjudgement, have been analysed again. The focus was thereby especially on those targets for which the models were unrecognised by most participants. For this analysis, the native structure was loaded in addition to the two models and the trace for the respective NBH plotted as simple bold black line (see Figure 6.4).

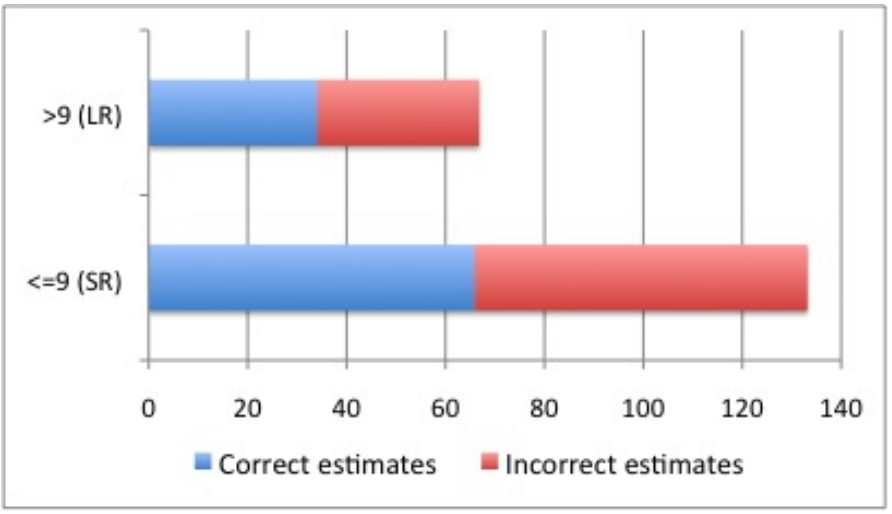


FIGURE 6.3: Dependance of evaluation certainty on the sequence separation between the contact residues: the percentage of LR and SR contacts is plotted for correct (blue) and incorrect (red) model specifications for the respective neighbourhoods.

The investigation of the mentioned NBHs showed that there are various factors or reasons that do sometimes lead to a wrong estimate. For many contacts and appropriate environment the traces for the two models were both aligned quite well with the clusters and bundles of edges, whereas one lies within the cluster of helical environments and the other within the cluster for strands (see Figure 6.4). Traces that include mainly clusters of nodes of helical type (coloured in cyan) illustrate helical environments, whereby node clusters of type strand (sheet) build the anchor points of strand traces. In case that one trace runs in helical and one in strand like fashion, the user intuitively decides for that trace and thus secondary structure type with higher support. Hence, the decision often goes for that trace that lies within the stronger cluster. The strength of support for the SS types is more or less obvious for different NBHs.
It is partly possible to identify the SS within the contact map. Parallel and antiparallel sheets come up as clusters of contacts that lie parallel or vertical to the diagonal. In comparison, helices stick out of the contact map in form of clusters along the diagonal. As a result, for contacts which lie close to the main diagonal of the contact map, i.e. that have a small sequence separation, the support is mostly higher for helices. This often leads to a wrong decision for the helical and against the strand environment.

Another factor that is often misleading concerning the rating is the length of the trace. The fewer contacts *iRes* has, the shorter the trace. But the number of contacts is not necessarily a good indicator for the properness of the model. In case that the environments of *iRes*, i.e. the traces, are quite similar for the two model, but on trace was significantly shorter, the participants mostly tended to the model that produces the more complex environment, i.e. longer trace (see Figure 6.5).

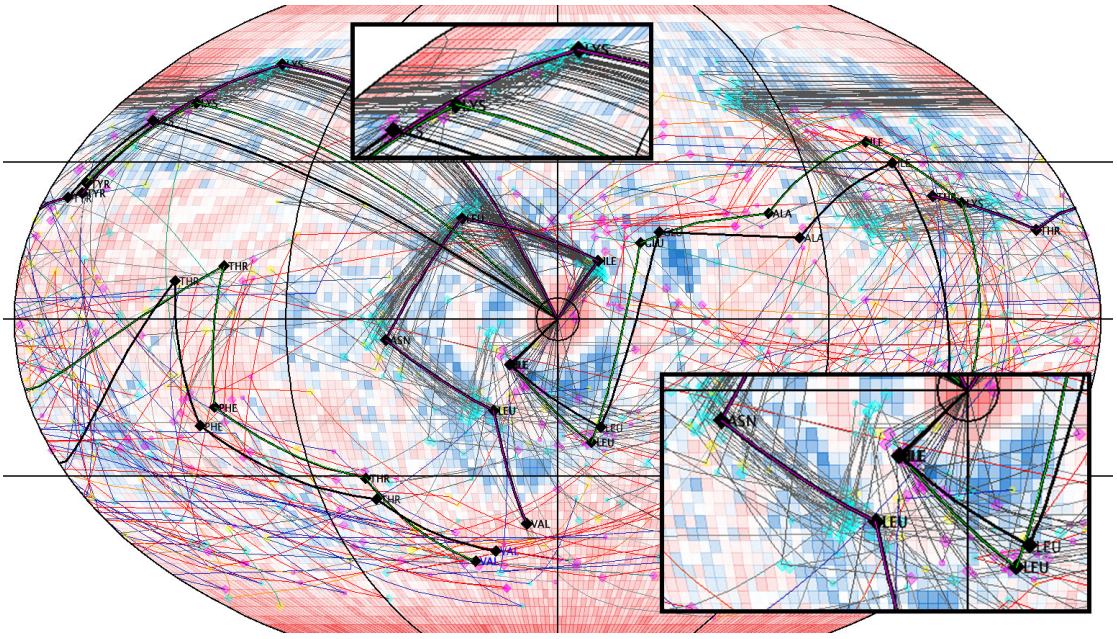


FIGURE 6.4: Neighbourhood traces for two models of target 499 for the contact between residue 15(Glutamic Acid) and 16(Alanine). The background information has been resolved for the NBH template string "AKExAK". The trace for the first model (pink) runs along the helical trace bundles, whereas the second model (green) covers clusters of strand nodes. The decision for the better model was made for the pink one, but the comparison with the native structure (black trace) shows that the second model is the better one.

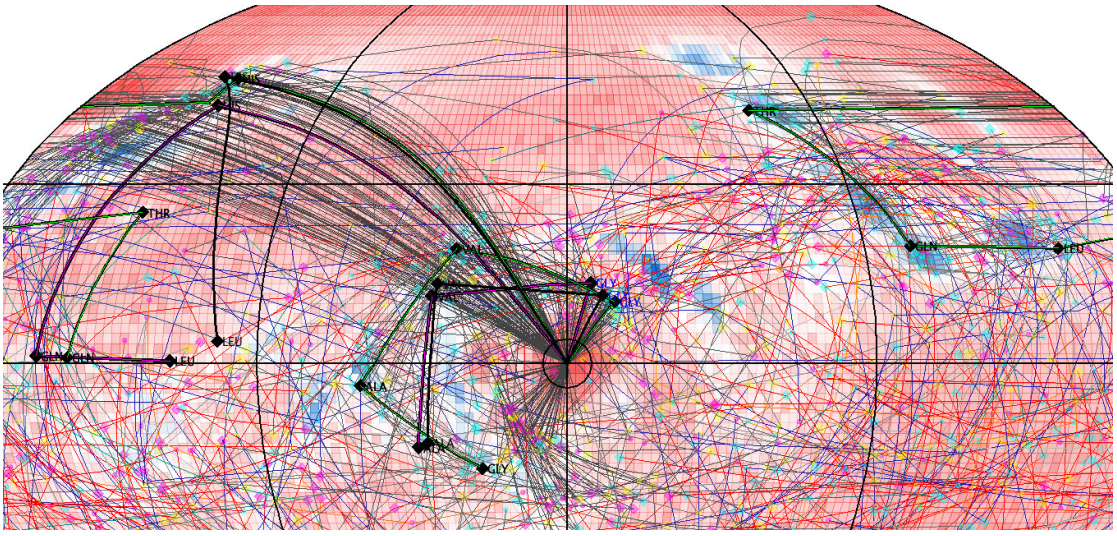


FIGURE 6.5: Neighbourhood traces for two models of target 594 for the contact between residue 47(Serine) and 55(Glycine). The background information has been resolved for the NBH template string "HxGVA". Both traces run along helical trace bundles, but the trace for the first model (pink) is much shorter than that for the second model (green). The decision for the better model was made for the green one, but the comparison with the native structure (black trace) shows that the first model is the better one.

.

The investigation of residue environments for falsely scored models showed, that the poorer model is sometimes locally closer to the native structure (see Figure 6.6). Based on the statistical support (the background information), for some contacts the decision falls for the overall worse model but locally better model. If of all contacts these are selected and evaluated, this of course might spoil the overall estimation of the better model if only few contacts were selected and analysed.

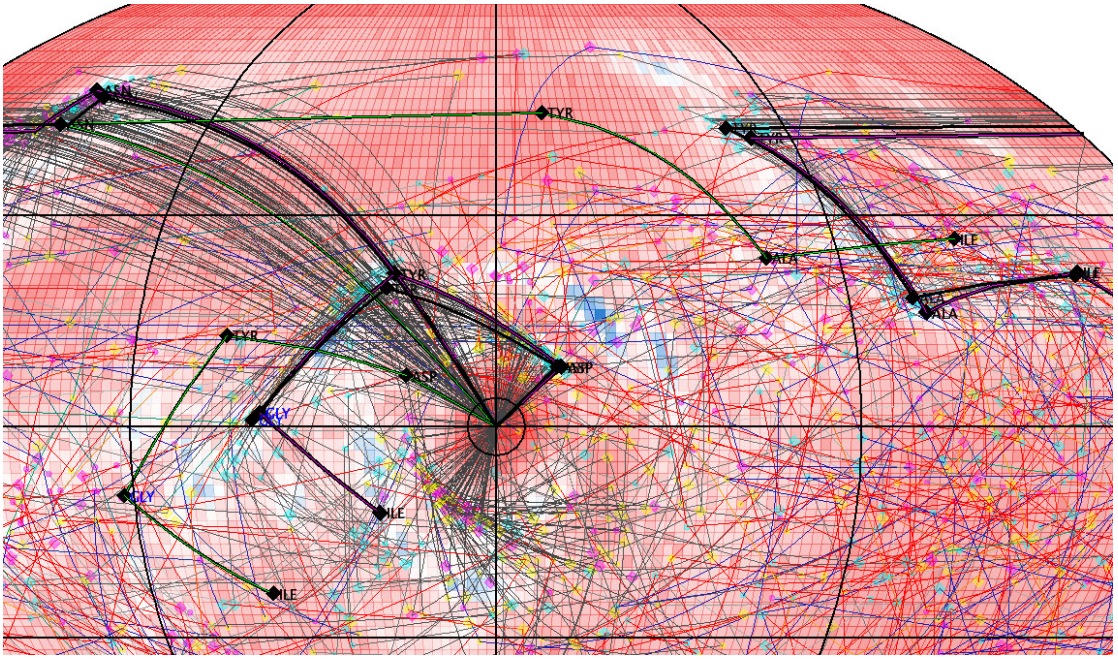


FIGURE 6.6: Neighbourhood traces for two models of target 401 for the contact between residue 66(Histidine) and 69(Glycine). The background information has been resolved for the NBH template string "YxDYG". The first model (pink) is much closer to the native structure, although the second model (green) is the overall better one. The decision for the first model for this local environment is thus reasonable.

6.3 Improvements and Extensions

There are two main directions for potentials extensions. On one side the visualisation itself could be extended, i.e. the way the statistical information is presented as well as the interactivity. The latter involves fast responses from 3D embeddings and simulations. Apart from that the extracted statistical information could be used for sampling methods to decrease the conformation space as well as scoring functions, i.e. energy functions for structure modelling.

Visualisation What would be of high value and importance to improve the visualisation and particularly the interactivity of the tool is the enhancement of data access and

thus computation of statistical information. The cullPDB20, used for the derivation of propensities, comprises 7-8 Mio. environments. Considering the huge amount of data, contained within the database, dynamic queries are difficult. Hence, pre-calculations respective residue and secondary structure types as well as orientations are necessary. Approaches to achieve a faster computation could be based on more effective indexing and the computation and storage of tables with intermediate results based on basic filtering constraints. Furthermore, the use of hash functions and parallelisation (on different clusters) could help to improve the performance. This would enlarge the options for interactive filtering, which implies a computation in real time. Consequently, the "Sphoxel Map" could be calculated on demand also for others than the default distance shells. Also the traces could be extracted faster, thereby facilitating options like magic lenses [BSP+93] or lenses in general. Such lenses could be used to interactively highlight different properties of the traces. The area within the lens would thereby show a different representation of the nodes covered by the lens. The lens could be moved over the view, thereby updating its representation. This could be of interest, if e.g. "pseudo-nodes" were implemented. In this case, the lens could be used to show the real distribution of nodes, while moving over the "pseudo-node".

Although the priority lied on the identification of the overall distribution of clusters and preferred orientations, it might nevertheless be an interesting option to include zooming, especially when it comes to the analysis of densely packed clusters of nodes. This could be achieved either via complete zooming into the map, or with the help of distortion lenses. The latter could be even combined with the previously mentioned magic lens, whereas it would make more sense within a representation that is not abstracted (without "pseudo-nodes" or "pseudo-edges").

Another approach could include a different representation of the clusters of NBH traces in the style of the hierarchical edge bundles of Holten and Wijk [HCvW07, Hol06]. Thereby, edges that run from one to another node cluster could converge in the middle, i.e. between the node clusters. Each edge would still start and end at the real position of the node instead of a "pseudo-node", thereby still illustrating how the nodes are distributed. Only the edges would look more compact, whereas the distribution within the cluster would be vaguely perceptible via the distribution of nodes. A disadvantage of this visualisation would affect the evaluation of models, as it would not be possible to see whether a template trace lies within the more or less dense part of a cluster. Therefore, it is obvious that this visualisation should be implemented as option, selectable by the user.

Also brushing and linking could be used to enhance the visualisation. This approach can be used to simultaneously highlight items within a representation, which are selected within another representation. Within the current representation of NBH traces, residues that are contained within the template $NBHString$ are highlighted only with bigger labels and a different node shape. It is conceivable that a stronger highlighting of only one residue type at a time would improve the perception. At that, only that

residues of the same type as pointed on with the mouse within the $NBHString$ selection panel (which is used to toggle on or off single residues) could be highlighted. All other nodes would thereby look the same, even if they are generally contained within the neighbourhood.

Scoring of Protein Structures The visualisation toolkit enables users to make an educated judgement about the quality of a predicted model. Moreover, it can be used to define orientation constraints for the prediction and scoring of models via energy functions. It is conceivable to extend the tool towards an automatic scoring of models. Besides the GDT score and RMSD, there are various scoring functions to rank protein models. As the GDT-score used in CASP currently represents the gold standard, it is obvious to use this scoring function as reference to derive the quality or suitability of other scoring function. The correlation between the rankings based on different scoring functions can be determined by rank correlation coefficients (RCC) like *Spearman's RCC* (ρ) or *Kendall's tau RCC* (τ). The RCC lies within the range [-1:+1] and defines how far two rankings agree. A value of one thereby poses perfect agreement, whereas zero means that there is no correlation at all. A value of -1 clarifies that one ranking is perfectly anti-correlated to the second. The Spearman's Rho RCC assumes the scale values to be element of a monotonic function, whereas Kendall's tau is based simply on ordinal information.
Scoring functions usually sum up score values, e.g. error values, for a number (or even all) contacts of the model. Depending on the type of scoring value, the final score needs to be normalised by the number of contacts.
A first suggestion for a scoring function was to use the LOS's for the spatial propensities of residue contacts. A score for a given contact $(iRes, jRes, iSSType)$ and its location (r, ϕ, λ), i.e. the position of residue J with respect to I (based on the rotation and translation invariant framework), would therefore be defined as:

$$LOG(dPos) = \log(\frac{\#edges(r, \phi, \lambda, i, j, s) * \#edges}{(\#edges(r, \phi, \lambda) * \#edges(i, j, s))}) \tag{6.1}$$

The delta value for the position ($dPos$) defines the range around the position (r, ϕ, λ), i.e. the minimum and maximum values for the spherical coordinates that are necessary for the queries (see Chapter 4). These scores can than be added up for all contacts, besides direct neighbours or neighbours that are close within the sequence. Thus, only the scores for contacts that have a sequence separation greater than 2 ($|iNum - jNum| > 2$) are included into the scoring function. As LOS's are already normalised, a further normalisation is not required.
This "geometric" scoring function was used to score the submitted server models for several targets of CASP8. To compare this score with the GDT score (used for the CASP ranking), the "Kendall's tau B" RCC (τ_B) was computed for several targets and

thus rankings. The correlation coefficient (τ_B) was on average about 0.29. It was thereby better than the widely used score Rosetta [KLD+10] with a correlation of about 0.12, but worse than the currently best scoring function DOPE(discrete optimized protein energy) [SS06] with a correlation of about 0.43.
An improved scoring function, based on statistical information, should not only use LOS's for the purpose of scoring but also the clusters of NBH traces. These traces could be computed for a substring of the $nbhString$, which gives a good statistical support, e.g. more than 150 neighbourhoods and thus traces. Such a substring could be derived automatically via selection from the set of optimal substrings, which was explained within chapter 5.2. In case, the selected substring still does not produce enough traces, it can further be shortened through the removal of (a) single residue(s), e.g. at the end of the string. The DBSCAN clustering method could be applied with default values for ϵ and $minNumPts$.
The score for each relevant contact (regarding sequence separation) would thereby be composed of the LOS and a score which declares how well the template trace for the NBH of the contact fits the bundle of NBH traces. The latter value would depend on different factors:

- the number of nodes of the NBH trace that lie within any cluster
- the number of nodes which lie in that cluster, where its residue type occurs predominantly
- the position and further cluster coverage for the node ($jRes$) of the contact

The second factor is based on the histogram analysis of the clusters and the determination of predominant residue types within each cluster (explained in chapter 5.4). Depending on which of these factors (that might be even all of them) are used an appropriate way of summing up and normalisation needs to be defined.

Chapter 7

Conclusion

The determination of the three-dimensional structure based on the sequence information is of great value for a number of reasons. The knowledge about the structure, which defines a protein's function, gives insights into the molecular basis of different diseases, e.g. cancer or Alzheimer's. Further, it introduces advanced possibilities for the purpose of drug design.
The speed at which protein structures are predicted lacks behind the determination of new sequences. This is because currently used experimental techniques, like NMR and X-ray crystallography, are quite difficult and time consuming and can not be applied to every protein of interest. The prediction of 3D structures solely based on the primary sequence still poses a problem considering accuracy. There are different approaches based on homology modelling, which can produce models with an RMSD of about $1 - 2$Å for models with high sequence identity of about 70%. At least, such techniques can be used to produce hypotheses about a protein's function. Besides the template selection and alignment of the target with a protein template, the process of homology modelling also includes the model construction and further model assessment. The latter can be performed via the methods statistical potentials or physics-based energy calculation. Such statistical potentials comprise empirical methods based on observed residue-residue frequencies among known protein structures. On the basis of such statistics various constraints can be derived, which can be used to decrease the amount of possible conformations or at least to rate possible states. Nevertheless, there is still a high degree of flexibility concerning the 3D structure, the distance constraints do not determine the structure bijectively. Rather, there exists an ensemble of views that is energetically possible.
That is where the extraction and further visualisation of statistically preferred orientations of specific residues comes into play. These geometric propensities can be used to support the process of model assessment. The visualisation of these propensities clarifies, that there exist clear tendencies for residues to be located at some predominant position with respect to each other. It can therefore directly illustrate how good or bad

a model is. For this purpose, the gained information only needs to be evaluated and integrated in an appropriate fashion.
Above all, the displays can be utilised to estimate the quality of a predicted model via analysing how well it fits into the statistical background information. This of course requires some user interaction and embodies a rather subjective evaluation. On the other hand, the gained knowledge about preferred residue dependent orientations can be used to calculate quantitative scores about the quality of a predicted model. This presupposes an adequate and well-defined scoring function.
Besides the assessment of the quality of a predicted model, there still lies a potential in the use of geometric orientation propensities for the definition of orientation constraints. These can be formulated as dihedral angles which could be taken into account by sampling methods used to reduce the amount of possible conformations during the phase of model assessment, thereby extending these methods towards a more funded basis. The definition of dihedrals requires a high degree of interaction through the user. This is justified by the necessity to evaluate and analyse several contacts in the context of their residue environments (NBHs). At the same time, the implemented options for clustering, cluster and histogram analysis can help within the decision making process. Especially the extraction of predominantly occurring residue types within the clusters supports the determination of the preferred orientations of residues with respect to each other.
To what extent there exist a potential in improving the predicted model, via extending the set of constraints for sampling methods, might be better assessable after the development of a scoring (energy) function and the comparison of this with a reference score like the GDT score. For this purpose, the scoring function should evaluate solely in how far the position of $jRes$ with respect to $iRes$ suits the statistical distribution of nodes, i.e. whether or not it coincides with a cluster that preferably contains residues of the same type. In case that a developed scoring function, based on such statistical information, correlates well with other established scores, we can assume that the derived geometric orientation constraints will yield an increase in accuracy.
The determination and extraction of statistical residue and orientation dependent probabilities facilitates to gain further insights into the mechanisms of protein folding. These information can be used and integrated into the process of homology-based structure modelling in general and the phase of model assessment in particular.

Bibliography

[BBM03] Vladimir Batagelj, Vladimir Batagelj, and Andrej Mrvar. Pajek: analysis and visualization of large networks. In *Graph Drawing Software*, pages 77–103. Springer, 2003.

[BEPW08] Klaus Backhaus, Bernd Erichson, Wulff Plinke, and Rolf Weiber. *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung.* Springer, Berlin, 12., vollständig überarbeitete auflage. edition, 2008.

[BEW95] Richard A. Becker, Stephen G. Eick, and Allan R. Wilks. Visualizing network data. *IEEE Transactions on Visualization and Computer Graphics*, 1(1):16–28, 1995.

[BJ96] I. Bahar and R.L. Jernigan. Coordination geometry of nonbonded residues in globular proteins. *Folding & design*, 1(5):357–70, Jan 1996.

[Boe10] R. Boehm. Die ganze kartennetzentwurfslehre kurzgefasst, Mai 2010.

[BS95] L.M. Bugayevskiy and J.P. Snyder. *Map Projections: A reference manual.* CRC Press, first edition edition, June 1995.

[BSP+93] Eric A. Bier, Maureen C. Stone, Ken Pier, William Buxton, and Tony D. DeRose. Toolglass and magic lenses: the see-through interface. In *SIGGRAPH '93: Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pages 73–80, New York, NY, USA, 1993. ACM.

[BST03a] B.-J.J. Breitkreutz, C. Stark, and M. Tyers. Osprey: a network visualization system. *Genome biology*, 4(3), 2003.

[BST03b] N.-V. Buchete, J.E. Straub, and D. Thirumalai. Anisotropic coarse-grained statistical potentials improve the ability to identify nativelike protein structures. *J Chem Phys*, 118(16):7658–7671, Jan 2003.

[BST04a] N.-V. Buchete, J.E. Straub, and D. Thirumalai. Continuous anisotropic representation of coarse-grained potentials for proteins by spherical harmonics synthesis. *J Mol Graph Model*, 22(5):441–50, May 2004.

[BST04b] N.-V. Buchete, J.E. Straub, and D. Thirumalai. Development of novel statistical potentials for protein fold recognition. *Curr Opin Struct Biol*, 14(2):225–32, Apr 2004.

[BT99] C.-I. Branden and J. Tooze. *Introduction to Protein Structure: Second Edition*. Garland Publishing, second edition, January 1999.

[BW04] Ulrik Brandes and Dorothea Wagner. Netzwerkvisualisierung. *it - Information Technology*, 46(3):129–134, 2004.

[BWK+01] A. Vilanova Bartrol, R. Wegenkittl, A. Knig, E. Grller, E. Sorantin, and Tiani Medgraph. Virtual colon flattening, 2001.

[CMS99] Stuart K. Card, J. D. Mackinlay, and Ben Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Academic Press, London, 1999.

[DAK09] Swagatam Das, Ajith Abraham, and Amit Konar. *Metaheuristic Clustering*. Springer Publishing Company, Incorporated, 2009.

[Dan00] Peter H. Dana. Map projections, 10 2000.

[Dav05] da vinci. http://www.informatik.uni-bremen.de/uDrawGraph/en/index.html, 2005.

[DeL] W.L. DeLano. *The PyMOL User's Manual*. DeLano Scientific, Palo Alto, CA, USA.

[DS03] Richard P Dum and Peter L Strick. An unfolded map of the cerebellar dentate nucleus and its projections to the cerebral cortex. *Journal for Neurophysiology*, 89(1):634–639, 2003.

[Fur86] G. Furnas. Generalized fisheye views. *CHI '86: Proceedings of the SIGCHI conference on Human factors in computing systems*, Apr 1986.

[Fur09] C.A. Furuti. Cartographical map projections, July 2009.

[Gra00] Graphlet toolkit. http://www.infosun.fim.uni-passau.de/Graphlet/, 2000.

[HAK00] Steven Haker, Sigurd Angenent, and Ron Kikinis. Nondistorting flattening maps and the 3d visualization of colon ct images. *IEEE Trans. on Medical Imaging*, 19:665–670, 2000.

[HCvW07] D. Holten, B. Cornelissen, and J.J. van Wijk. Trace visualization using hierarchical edge bundles and massive sequence views. pages 47 –54, jun. 2007.

[HGQ+06] Wei Hong, Xianfeng Gu, Feng Qiu, Miao Jin, and Arie Kaufman. Conformal virtual colon flattening. In *SPM '06: Proceedings of the 2006 ACM symposium on Solid and physical modeling*, pages 85–93, New York, NY, USA, 2006. ACM.

[Hol06] Danny Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):741–748, 2006.

[JAV07] A.N. Jha, G.K. Ananthasuresh, and S. Vishveshwara. Protein sequence design based on the topology of the native state structure. *J Theor Biol*, 248(1):81–90, Sep 2007.

[JH04] C. Johnson and C. Hansen. *Visualization Handbook*. Academic Press, Inc., Orlando, FL, USA, 2004.

[JSJ+84] Richard E. Rosenfield Jr, Stanley M. Swanson, Edgar F. Meyer Jr, Horace L. Carrell, and Peter Murray-Rust. Mapping the atomic environment of functional groups: turning 3d scatter plots into pseudo-density contours. *Journal of Molecular Graphics*, 2(2):43 – 46, 1984.

[KLD+10] Kristian W. Kaufmann, Gordon H. Lemmon, Samuel L. DeLuca, Jonathan H. Sheehan, and Jens Meiler. Practically useful: What the rosetta protein modeling suite can do for you. *Biochemistry*, 49(14):2987–2998, 2010. PMID: 20235548.

[LBF+09] M. Lappe, G. Bagler, I. Filippis, H. Stehr, J.M. Duarte, and R. Sathyapriya. Designing evolvable libraries using multi-body potentials. *Current Opinion in Biotechnology*, 20(4):437 – 446, 2009. Protein technologies / Systems and synthetic biology.

[Maz09] Riccardo Mazza. *Introduction to Information Visualization*. Springer, 1 edition, March 2009.

[MBHC95] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, Apr 1995.

[MFK+09] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, and A. Tramontano. Critical assessment of methods of protein structure predictionround viii. volume 77, pages 1–4. Wiley Subscription Services, Inc., A Wiley Company, August 2009.

[NGB+09] M. Neugebauer, R. Gasteiger, O. Beuing, V.Diehl, M. Skalej, and B. Preim. Map Displays for the Analysis of Scalar Data on Cerebral Aneurysm Surfaces. In *Computer Graphics Forum (EuroVis)*, volume 28 (3), pages 895–902, Berlin, 10.-12. Juni 2009.

[OGF+10] S.I. O'Donoghue, D.S. Goodsell, A.S. Frangakis, F. Jossinet, R.A.Laskowski, M. Nilges, H.R. Saibil, A. Schafferhans, R.C. Wade, E.Westhof, and A.J. Olson. Visualization of macromolecular structures. *Nature methods*, 7(3 Suppl), March 2010.

[OGH+06] Steffen Oeltze, Frank Grothues, Anja Hennemuth, Anja Ku, and Bernhard Preim. Integrated Visualization of Morphologic and Perfusion Data for the Analysis of Coronary Artery Disease. In *IEEE/Eurographics Symposium on Visualization*, Informatik aktuell, pages 131–138. Springer, 2006.

[OMJ+97] Ca Orengo, Ad Michie, S Jones, Dt Jones, and Mb Swindells. Cath:a hierarchic classification of protein domain structures. *Structure*, 5:1093–1108, 1997.

[Pea90] Frederic Pearson. *Map Projections: Theory and Applications*. CRC Press, second edition edition, March 1990.

[Phi69] D.C. Phillips. The development of crystallographic enzymology. *Biochemical Society Symposium*, 30:11–28, Nov 1969.

[PWS08] G. Pavlopoulos, A.L. Wegener, and R. Schneider. A survey of visualization tools for biological network analysis. *BioData Mining*, 1(1):12+, November 2008.

[RA08] S. Rakshit and G.K. Ananthasuresh. An amino acid map of inter-residue contact energies using metric multi-dimensional scaling. *J Theor Biol*, 250(2):291–7, Jan 2008.

[Ran00] M. Randić. Condensed representation of dna primary sequences. *Journal of Chemical Information and Computer Sciences*, 40(1):50–56, Jan 2000.

[Rho06] Gale Rhodes. *Crystallography made crystal clear - A guide for users of macromolecular models*. Academic Press, Complementary Science Series, third edition, February 2006.

[RRS63] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Molecular Biology*, 7(1):95–99, Jan 1963.

[RWS+10] Christian Rieder, Andreas Weihusen, Christian Schumann, Stephan Zidowitz, and Heinz-Otto Peitgen. Visual support for interactive post-interventional assessment of radiofrequency ablation therapy. volume 29, pages 1093–1102. Blackwell Publishing Ltd, June 2010.

[SDS+09] R. Sathyapriya, J.M. Duarte, H. Stehr, I. Filippis, and M. Lappe. Defining an essence of structure determining residue contacts in proteins. *PLoS Comput Biol*, 5(12):e1000584, 12 2009.

[Sip90] Manfred J. Sippl. Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol*, 213(4):859–83, Jun 1990.

[SMO+03] P Shannon, A Markiel, O Ozier, N S Baliga, J T Wang, D Ramage, N Amin, B Schwikowski, and T Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, November 2003.

[Sny93] J.P. Snyder. *Flattening the Earth: Two Thousand Years of Map Projections*. University of Chicago Press, 1993.

[SS06] Min-yi Shen and Andrej Sali. Statistical potential for assessment and prediction of protein structures. *Protein Science*, 15(11):2507–2524, November 2006.

[TE00] D. Tobi and R. Elber. Distance-dependent, pair potential for protein folding: results from linear optimization. *Proteins*, 41(1):40–6, Oct 2000.

[Tid05] J. Tidwell. *Designing Interfaces : Patterns for Effective Interaction Design*. O'Reilly Media, Inc., November 2005.

[TSLE00] D. Tobi, G. Shafran, N. Linial, and R. Elber. On the design and analysis of protein folding potentials. *Proteins*, 40(1):71–85, Jul 2000.

[VS93] G. Vriend and C. Sander. Quality-control of protein models - directional atomic contact analysis. *J Appl Crystallogr*, 26:47–60, Jan 1993.

[WD03] Guoli Wang and L. Dunbrack. Pisces: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, Aug 2003.

[Wei10] E.W. Weisstein. Wolframmathworld the web's most extensive mathematics resource, April 2010.

[Wes74] Westermann, editor. *Diercke Weltatlas*. Georg Westermann Verlag, Braunschweig, 180. auflage (92. auflage der neubearbeitung) edition, 1974.

[WLYY03] Y.H. Wu, A.W.C. Liew, H. Yan, and M.S. Yang. Db-curve: a novel 2d method of dna sequence visualization and representation. *Chem Phys Lett*, 367(1-2):170–176, Jan 2003.

[XB05] Lei Xie and Philip E. Bourne. Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Comput. Biol*, 1(3):31, August 2005.

[YST02] Qihe. Yang, John P. Snyder, and Waldo R. Tobler. *Map projection transformation : principles and applications*. Taylor & Francis, 2002.

[ZB07] M. Zvelebil and J.O. Baum. *Understanding Bioinformatics*. Garland Science, first edition, August 2007.