HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

# BACHELORARBEIT

Herr
**Alexander Eisold**

## Quantitative Acquisition and Feature Selection based of 3D Structure Information for Prediction with SVM of the *cis-trans* Isomerisation of Xaa-Pro in Membrane Proteins

2012

Fakultät **Mathematik/Naturwissenschaften/Informatik**

# BACHELORARBEIT

# Quantitative Acquisition and Feature Selection based of 3D Structure Information for Prediction with SVM of the *cis-trans* Isomerisation of Xaa-Pro in Membrane Proteins

Autor:
**Alexander Eisold**

Studiengang:
Biotechnologie/Bioinformatik

Seminargruppe:
BI09w1-B

Erstprüfer:
Prof. Dr. rer. nat. Dirk Labudde

Zweitprüfer:
Dipl-Inf. (FH) Daniel Stockmann

Mittweida, August 2012

**Bibliografische Angaben**

Eisold, Alexander: Quantitative Acquisition and Feature Selection based of 3D Structure Information for Prediction with SVM of the *cis-trans* Isomerisation of Xaa-Pro in Membrane Proteins, 45 Seiten, 17 Abbildungen, 15 Tabellen, Software CD, Hochschule Mittweida (FH), Fakultät Mathematik/Naturwissenschaften/Informatik

Bachelorarbeit, 2012



Satz: LaTeX

**Referat**

Die Bachelorarbeit befasst sich mit der *cis-trans* Isomerisierung von Xaa-Pro (Xaa ist eine beliebige Aminosäure vor Prolin), deren Quantitative Erfassung und die Auswahl von 3D-Strukturinformationen für die Vorhersage durch eine Support Vektor Maschine (SVM). Die mengenmäßige Erfassung des Auftretens von *cis-*, *trans-* und der *cis/trans* Konformation in Membranproteinen wird untersucht und ausgewertet. Bei den 3D-Strukturinformationen handelt es sich um 12 Merkmale der Aminosäuren rund um und einschließlich des Prolin besitzen. Dazu gehören innen/außen Klassifikation, reale Sekundärstruktur, energetische Betrachtung, sowie fünf weiter Aminosäure Eigenschafen in einem definierten Radius um das Prolin vorkommen. Aus diesen Informationen wurde ein Datensatz für die SVM geschaffen, die als Programm zur Vorhersage von unbekannten und bekannten Xaa-Pro-Isomerien dient. Die Methoden für die Auswertungen wurden mit der plattformunabhängigen Programmiersprache Java umgesetzt. Aus der Arbeit sind zwei Programme hervorgegangen, Xaa-PIPT (Xaa-Proline *cis/trans* Isomerization Prediction Tool) für die Quantitative Erfassung und Extrahieren von strukturellen Informationen und $\mu$Xaa-PIPT zur reinen Vorhersage der Xaa-Pro-Isomerie in Proteinstrukturen. Als Grundlage dienten $389$ Membranproteine der PDB (Protein Data Bank). Die gewonnen Daten wurden zudem Statistisch untersucht und ausgewertet.

The bachelor thesis is about *cis-trans* isomerization of Xaa-Pro (Xaa = any amino acid), their quantitative acquisition and the selection of 3D structure information for the prediction with a support vector machine (SVM). The quantitative detection of occurrence of *cis-*, *trans-* and *cis/trans* conformation in membrane proteins will be examined and evaluated. The 3D structure informations include 12 features, the amino acids around proline and are including of proline. These include the inside/outside classification, the real secondary structure, energy consideration, as well as five further amino acid occur properties within a defined radius of the proline. From this information, a data set was created for the SVM. This program is used for the prediction of unknown and known Xaa Pro Isomerisms. The methods for the analysis were implemented with the platform independent programming language Java. Two programs have emerged from the work to a Xaa PIPT for the quantitative detection and extracting structural information and $\mu$Xaa-PIPT to the pure prediction of Xaa-Pro isomerism in protein structures. $389$ Membrane

proteins from the PDB (Protein Data Bank) served as a basis. The data were also statistically analysed and evaluated.

# I. Contents

# II. List of Figures

# III. List of Tables

# IV. Nomenclature

3D ..... three-dimensional, page 9

5-$HT_3$ ..... 5-Hydroxytryptamine type 3, page 2

AA ..... amino acid, page 15

ASCII ..... American Standard Code for Information Interchange, page 13

CGI ..... Common Gateway Interface, page 1

CSV ..... Comma-separated values, page 13

ePros ..... Energy Profile Suite, page 11

NMR ..... Nuclear magnetic resonance, page 3

PDB ..... Protein Data Bank, page 9

PDBTM ..... Protein Data Bank of Transmembrane Proteins, page 9

PPIases ..... Peptidyl-Prolyl-*cis-trans*-Isomerases, page 2

Pro ..... the amino acid proline in three-letter code, page 2

RCSB ..... Research Collaboratory for Structural Bioinformatics, page 9

SASA ..... solvent accessible surface area, page 5

SVM ..... support vector machine, page 7

Xaa-PIPT ..... Xaa-Proline cis/trans isomerization prediction tool, page 9

# V. Acknowledgment

> „Man muß mit den richtigen Leuten
> zusammenarbeiten, sie achten und motivieren.
> Dauerhafter Erfolg ist nur im Team möglich.“
>
> Klaus Steilmann (*1929)

An dieser Stelle möchte ich den Menschen danken die immer an meiner Seite sind und mich bei meiner Bachelorarbeit unterstützt haben.

Herr Prof. Dr. rer. nat. Dirk Labudde ist ein Beispiel für Arbeitsausdauer. Mein Mentor hat mich auf dieses Thema gebracht. Dabei stand er mir mit wertvollen Ideen zu Seite, so dass wir zusammen stets Themen übergreifende Lösungsansätze finden konnten. Er hat viel für mich möglich gemacht.

Danke auch an Herrn Dipl.-Informatiker (FH) Daniel Stockmann. Er hat die nötige Rechenkapazität verfügbar und die Zugänge darauf möglich gemacht.

Mein besondere Danke geht an Flo. Er ist immer zuverlässig und sehr hilfsbereit. Wir haben uns bei der Zusammenarbeit sehr gut verstanden und oft ergänzt. Seine guten Ideen und seine sehr hohe Arbeitsbereitschaft hat uns nach vorne gebracht. Gerne würde ich in zukünftigen Projekten mit ihm arbeiten und von ihm lernen.

Ich danke meiner Freundin Sarah die mich in den schwierigen Momenten mit aufbauenden Worten wieder auf die Spur gebracht hat. In ihrer Nähe konnte ich mit ihr lachen. Des weiteren möchte ich hier meinen Eltern danken. Sie haben mir immer gut zugesprochen und sind immer für mich da. Außerdem danke ich meinem großem Bruder Robert und meiner kleinen Schwester Elisa. Sie haben mir immer Freude bereitet und ich konnte mich in ihrer Gegenwart immer gut erholen.

# 1 Introduction

## 1.1 Motivation

The tasks in bioinformatics are versatile. The tasks are connected with a high data arise. This is also in the evaluation of *cis-trans* isomerism. Also give many researchers in publications an overview of some functions. Only the amino acid proline is able to switch the conformation, even though all the amino acids to be able change. That is just so interesting prediction of protein structures. There are a variety of algorithms for predictions. To work with one of these is very interesting and informative. It has been researched much on this topic with many approaches. It is a challenge to find a new approach. The subject of *cis-trans* isomerism is very up to date. However, there was still non 3D structure viewing this topic.

Early 2011 I came the first time by Prof. Dr. rer. nat. Dirk Labudde on this topic. He has presented this problem Florian Kaiser and me. We were allowed to take part in the collaboration with Andreas Tzakos from the "Section of Organic Chemistry and Biochemistry" of the University of Ioannina in Greece. Our task was to make a statistical analysis at the time. Our task was to make a statistical analysis at the time. These included the sequence pattern analysis, the role of the hydrogen bonds, physicochemical properties of adjacent residues and the solvent accessibility specific to the *cis-trans* isomerism. For these evaluations we have developed a tool in the Common Gateway Interface (CGI) scripting language Perl.

In different publications was researched with predictions of Xaa-Pro *cis-trans* isomerism. However, still no 3D structure was considered analysis. We made us the task to develop a software that makes a prediction based on 3D structure data. For that, a prediction accuracy is required. Could give a prediction for the Xaa-Pro *cis-trans* isomerism and may be wrong solved protein structures right predictions.

## 1.2 Membrane Protein and Proline

Membrane proteins are components build of as which pass biological membrane in all organism. The membrane associated are crucial proteins here. These can form channels into membrane and change the permeability for ions or polar molecule specifically with that. The function of a transporter protein is to transporters they channel metabolic products, ions and even proteins about the cell membrane. They contribute to the energy conversion and use proteins with enzyme character. The last protein characteristic is to work as receptors. In this function they can form antennae which take signals of the outer surroundings and pass information on into the cell inside. Membrane proteins are usually of mainly ranges of the proteins the complete membrane, this described as integral membrane protein. These are subdivided into areas. The transmembrane region consists of one or several right-handed $\alpha$-helix and nontransmembrane region is divided up into extracellular space and intracellular space. Usually consists the transmembrane region of a hydrophobic core and the nontransmembrane region consists of hydrophilic segments.

Proline is able influence the functions of the membrane proteins. Primarily at transporters and ion channels, many became proline in the transmembrane regions often found here. This fact says that the isomerization is a switch function in ion channels [Lu et al., 2007]. Without a catalyst is the change between *cis* and *trans* rather slowly. However it is unclearly triggers the *trans*-to-*cis* isomerization like neurotransmitter and which rolls has the PPIases in the conformational switch [Lu et al., 2007]. The *cis-trans* isomerization is able to open the pore of a neurotransmitter-gated ion channel. This phenomenon was investigated at the 5-Hydroxytryptamine type 3 (5-$HT_3$) receptor. This receptor is a member of the Cys-loop receptor superfamily. The ion channel opening (gating) if the neurotransmitter binding in these protein trigger. As yet this process is based on a not characterized conformation change. The isomerism change takes place at the specific proline (Pro 8$^\star$) which is between two transmembrane helices at the top of the loop. Therefore between the second helix referred to as M2 and third helix referred to as M3. This proline can move the backbone of the protein by the change of the *cis-trans* isomerization. In detail, Pro 8$^\star$ is in the *trans* conformation, ion channel is in the closed state. Once a ligand binds, a structural change is per initialized and culminates to the *cis*. The transmembrane helix M2 lines up newly through this, the channel is open and the ion exchange can happen. In addition, was Pro 8$^\star$ with the amino acids Cys, Val, Gly, Ala, Asn or Lys replaces, nevertheless a correct folded receptors with antagonist binding functionality. However, the functionality was not given, i. e. only the Pro 8$^\star$ is able to regulate the ion channel gating [Lummis et al., 2005].

## 1.3 Prolyl Cis-Trans Isomerization

The peptide backbone is a repeating chain defined as three dihedral angels per amino acid residue phi $\phi$, psi $\psi$ and omega $\omega$. This angle is defined as the angle between two planes, spanned by three bond vectors between four atoms. In detail the torsion angel phi describe rotation about the $N - C^{\alpha}$ bond and contains the four atoms $C'_{(i-1)} - N_{(i-1)} - C^{\alpha}_{(i-1)} - C'_{(i-1)}$ and the angle psi describe rotation about the $C^{\alpha} - C'$ bond with atoms $N_{(i-1)} - C^{\alpha}_{(i-1)} - C'_{(i-1)} - N_{(i)}$ involved. Omega is the conformation determining torsion angle. The peptide joining two adjacent atoms $C' - N$ in the peptide backbone. The torsion angle omega can occur in *cis* ($\omega = 0°$) or in *trans* ($\omega = 180°$) conformation. Furthermore it is involved the four atoms $C^{\alpha}_{(i-1)} - C'_{(i-1)} - N_{(i)} - C^{\alpha}_{(i)}$. Also rotates the chemical bond as an interval from $-180°$ to $180°$. *Cis-trans*-isomerism appears when the free rotatability is not given around a bond axis between two atoms and these atoms always carry substituent. The angles phi, psi and omega are represented in the illustration in figure 1.1.

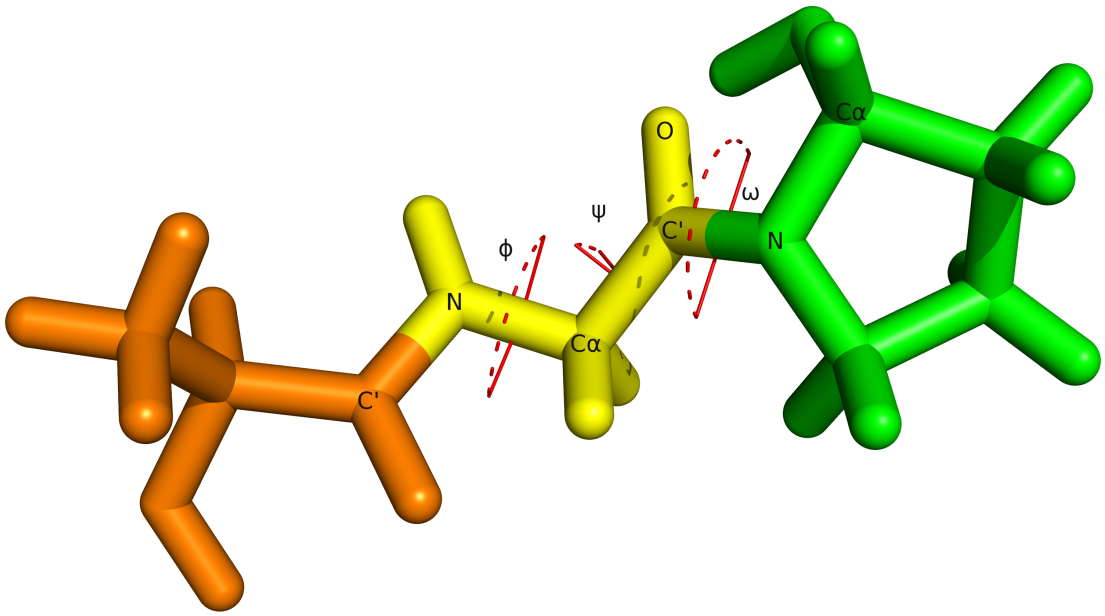


Figure 1.1: Tripeptide alanine (orange), glycine (yellow) and proline (green) with $\phi$,$\psi$ and $\omega$ dihedral angles

For all peptide bond is the *trans* conformation the favorable conformation with the least steric interactions. The peptide bond $C'$ and $N$ educates a partial double bond character. This is the resulting from the plane arrangement of the six backbone atom $C^{\alpha}$,$C'$,

$O$, $N$, $C^{\alpha}$ and the nitrogen lon electron pair. The hydrogen atom is included at the nitrogen, but not present in proline [Pahlke et al., 2005]. This double bond character leads to the abolition of the free rotation around the $CO-NH$ bond. The peptide bond is therefore planar and rigid. Furthermore, experimental and theoretical observations indicate a high energy barrier to about $20\ kcal{\cdot}mol^{-1}$ between the *cis* and *trans* form of the $C'-N$ bond. The resonance are represented in the illustration in figure 1.2. The partial double bond character accrues from delocalisation of the electrons over $\pi$ orbitals of the $O-C'-N$-bond. The length of the peptide bond is $1.32$ Å this is between the length of a $C'-N$ single bond ($1.49$ Å) and $C'=N$ double bond ($1.27$ Å).

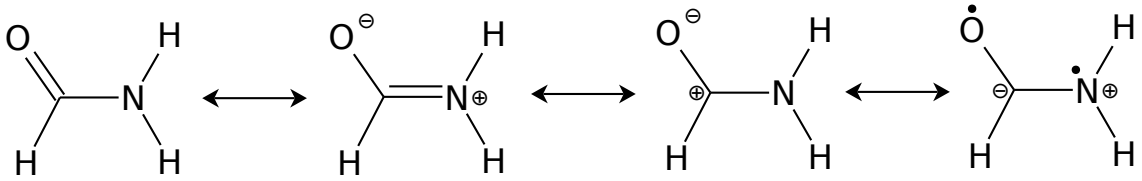

Figure 1.2: Resonance of the peptide bond

At Xaa-Pro-bonds (Xaa = any amino acid) are capable to both *cis* and *trans* isomers. Non-prolyl *cis* peptide bonds in folded protein to be extremely rare, because the *trans* form is higher stability and is far more energetically favorable than the *cis* form. This is caused by the interaction between the two $C^{\alpha}$ atoms of the adjacent residues and the hydrogen atom bound to it. An example of the two conformations are depicted in Figure 1.3 and 1.4 on page 5. Whereas peptide bonds to proline, the peptidyl-prolyl-peptide-bond, in *cis* conformation is more frequent, because the smaller energy difference between the *cis* and *trans* isomeric. The *cis* and the *trans* form are almost isoenergetic in Xaa-Pro fragments. In the two forms of Xaa-Pro the energy barrier is reduced for the unfavorable interactions. Through this it comes to a higher *cis* conformation propensity in the Xaa-Pro fragments in relation to the non-proline fragments. In fact, the *cis* conformation occurs with a frequency of $5-6\%$ in protein structures and a large majority of the *cis* Xaa-Pro occur in bend and turn regions and is typically solvent exposed on the surface of proteins [Pahlke et al., 2005].The rotation of the $C'-N$ bond is highly energetically restricted. This barrier limits the rate of interconversion between these two structural isomerism. Between *cis* and *trans* there is also the transition state. This is described as a *syn* transition state ($\omega = 90^{\circ}$) and has an energy that is significant higher than that of the *trans* state, which indicates that conversion from *cis* to *trans* state occurs very rarely [Lu et al., 2007]. Furthermore the interconversion rate is also solvent dependent and generally higher in polar solvents [Wedemeyer et al., 2002]. Linus Pauling purposed the following equation to approximately decide the energy associated with the rotation about the $C'-N$ peptide bond:

$$E = 30\sin^2\delta[kcal{\cdot}mol^{-1}] \tag{1.1}$$

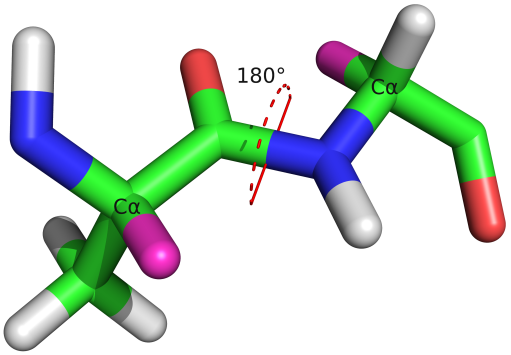


Figure 1.3: Dipeptide Alanine-Glycine in *trans* conformation

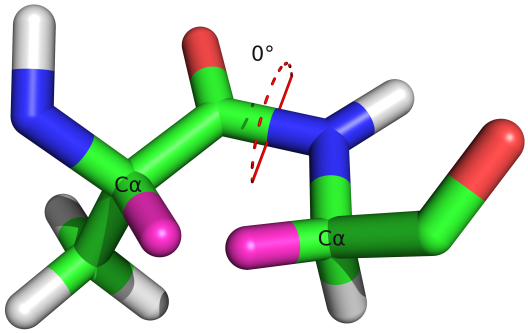


Figure 1.4: Dipeptide Alanine-Glycine in *cis* conformation

whereas $\delta$ is the torsion angle measuring deviations from the *trans* conformation. New analyses of high resolution proteins confirmed the remarkable precision of this early approximation [Lu et al., 2007].

Furthermore showed NMR experiments that the *cis-trans*-isomerism is influenced by the adjacent amino acids sequence by proline. In addition, one has seen through the correlations between the bulkiness of the side chain of a preceding residue and the isomerization rate. If the bulkiness of the side chain increases, the isomerization rate becomes smaller. For example the preceding aromatic residues can cause approximately a tenfold reduction of the isomerization rate in comparison to alanine. NMR studies proved a coherence between the *cis-trans* isomerization rate and the nature of the proline succeeding residue. Positive charged sidahins, destabilize presumably the *cis* conformation relative to *trans* and in contrast stabilize the *cis* form the amino acids aspartate, asparagine and glycine [Pahlke et al., 2005].

## 1.4 Relationship of Structural Information

As yet there is now the information on the primary structure based on the sequence plain and the secondary structure based on models and algorithms. For example the one-dimensional informations based on the FASTA sequences. This could be solvent accessibility, physiochemical properties, secondary structure of the adjacent amino acids and further [Exarchos et al., 2009].It is for certain that information content is much contained in the FASTA sequence. Novel is the three-dimensional plane information analysis for the peptidyl-prolyl-*cis-trans* isomerism. It is the idea that the three-dimensional plane on also contains a high information content. Perhaps with the same features which were already found and used in publications.

A protein in the space offers a large spectrum to information content. Restrictions and filtering is also importantly, otherwise one does not get any exact data. A protein exists from several hundred amino acids with different features but not everybody these features have an influence on the *cis-trans* isomerism. The amino acids in the direct surroundings of the peptidyl-prolyl-peptide-bond ($\leq 10$Å) could have an influence on it.

The omega angle is the main and first information. This is extracted from the protein chain. This determines the *cis-trans* isomerism in a defined interval. The definition of the angular position is very important for the angle recognition in protein chains. In detail the interval for *cis* $-30^\circ \leq \omega \leq 30^\circ$ and for *trans* $-180^\circ \leq \omega \leq -150^\circ$ and $150^\circ \leq \omega \leq 180^\circ$. If the angle is outside this intervals, this is classified as intermediate state *cis/trans*.

The next information is the real secondary structure from the protein chain scaled in $\alpha$-helices, $\beta$-strand and coil. *Cis* and *trans* isomers occur with a different affinity in the certain secondary structure elements. Whereas the *trans* conformations are found in all secondary structure classes [Pahlke et al., 2005]. Secondary structure describes the convolution of the peptide chain as a secondary structure himself without the individual residues. Therefore only this is backbone of the peptide chain involved. Hydrogen bridges connect $-CO$ and $-NH$ groups peptide bonds more differently with each other. The hydrogen bonds between different parallel or anti-parallel develop bind for the helices. All amino acids can theoretically form a helix with the exception of proline. In detail proline has a secondary amino group whose hydrogen atom at the peptide bond is lost. Therefore no hydrogen bonds can emerge and proline occurs at the beginning or end. If a proline is in the helix centrally, then will that cause a sharp kink of $\geq 20^\circ$. The *cis* conformation is usually confined to bend and turn regions [Pahlke et al., 2005].

The next an the third structural information is the inside/outside classification. This classification is an altenative method for the SASA (solvent accessible surface area). A inside/outside criterion for globular proteins which is based on the calculation of the SASA in dependence of the physikochemischen composition of the observed amino acid residue was defined. For $\alpha$-helical membrane proteins must the approach be modified. The residue of the helices are reason for it in the transmembrane regions which are exposed in the hydrophobic environment of the cell membrane. However, external residues in the extracellular space are exposed to hydrophilic environment. Therefore the auxiliary condition of the statistics is extended. The position of the observed atom in the protein structure is indicated now. An amino acid must be determined as the outside and inside. In addition, it must be explored whether an amino acid located in the transmembrane or in extracellular space. The inside/outside criterion of SASA is not sensitive enough to grasp the inside/outside conditions of the residues in transmembrane segments correctly. An alternative to the SASA approximation represents the following inside/outside classification. Proline is the interesting amino acid at the *cis-trans* isomerism and therefore the calculation is applied to proline:

**Definition 1.1** Proline is classified as $inside$ if the position of $C^{\beta}$ atom of proline and $C^{\alpha}$ of proline and the centroid $c$ of all $C^{\alpha}$ atoms within a sphere of $10$ Å around proline fulfills:

$$|C^{\alpha} - c| < 5 \cup (C^{\beta} - C^{\alpha})(c - C^{\alpha}) < 0 \tag{1.2}$$

if the above condition is not valid, proline is determined as $outside$.

The first logical statement corresponds to the case differentiation of the distance of the $C^{\alpha}$-atom to the emphasis $c$. An assignment of proline as $inside$ is carried out only at compliance of the second case differentiation. This describes the orientation of the side chain of proline. If the side chain depicts in the direction of the emphasis, the condition is regarded as filled. The scalar product of the two vectors is $(C^{\beta} - C^{\alpha})$ and $(c - C^{\alpha}) < 0$ [Heinke and Labudde, 2012].

The information based on the surrounding amino acids in a defined space and their effect on the *cis-trans* isomerism is the next information. These include the amino acid informations hydrophobicity, polarity mutability, bulkiness and flexibility. This is all to scaled values. Each amino acid is a get a numeric value that corresponds to their characterization. Each of these properties are developed by different researchers. A summary can be found on the ProtScale (`http://web.expasy.org/protscale/`), there are all values collected. With ProtScale you can determine course of properties along the amino acid sequence, this different procedure.

The last structure information is the proline energy. By means of a coarsegrained energy model based on $C^{\alpha}$ and $C^{\beta}$ coordinate information extracted from known protein structures can determine approximately the energy of an amino acid. It examines the interaction tendency of amino acid residue with other residue. In addition the information about the localization of the residue of membrane lipid bilayer in membrane proteins [D. Labudde, 2012].

## 1.5 Support Vector Machine

The idea of using support vector machine (SVM) is first to find an optimal parting plane to the separation of data tuple of two classes. These two objectives are as follows: the parting plane should have a largest possible distance (safe distance) to the nearest data tuple and thereby cause no or at least minimize classification errors. The next remote data tuple and perhaps classification errors are called support vector and determine the location of the parting plane. In the mathematical formulation of two classes are always encoded by the values $y = 1$ and $x = -1$.

This important learning method generated label input output mapping functions for the training data. The mapping function can be either a classifier function (category of the input data, or a regression function). To be considered to be non linear kernel functions

in are often used to transform input-data in a high-dimensional transform feature space. Which become input-data in multiple-separable planes in the comparison to the original input space transforms [Kecman, 2005].

# 2 Methods

All won data were caused and processed with Xaa-PIPT. This software is basing Java and them is able to process large text data. The programming language Java is an object-oriented programming language and is platform independent. Florian Kaiser and I have used Java due to the high complexity of the task to develop the software. We have developed the algorithms and source code together. BioJava (version 3.0.4) has helped us here to extract the information from the PDB files. BioJava is an open source project. A Java framework was created, to work with biological data possible to make. BioJava can manipulate biological sequences, files parse, access to BioSQL to provide and it provides tools for the statistical analysis of sequences. The software works with multithreading capability to have faster data processing. Our data collection is based on the same *cis-trans* isomerism background, a cooperation was therefore possible.

## 2.1 Basic Dataset

A dataset which contains only membrane proteins structures is necessary for a statistical returns. In this work I analyzed all 389 protein chain entries found on the PDBTM server (`http://pdbtm.enzim.hu/`), which are nonredundant. The protein chains entries are listed among each other. The individual entries consist of the PDB ID (all in lower case letters) and chain ID (all in upper case letters) separated by one underlined (e.g. $2a06\_E$). A PDB ID is for a unique $4$ character code, this one an entry is in the PDB. The first character must be a number between $1$ and $9$ and the other three characters can consist of letters and numbers. The TMDET algorithm is available for separation. Membrane planes are relatively determined with the position of atomic coordinates. With the help of a discrimination function the algorithm is able to separate transmembrane and globular proteins even in cases of low resolution or incomplete structures such as fragments or parts of large multi chain complexes. The complete PDB entires was searched and processed with the TMDET algorithm and the results were stored in PTBTM database. The success rate of the TMDET algorithm is $> 95\%$ [Frishman, 2010]. This server exclusively provides membrane proteins chains. These are not many proteins but I will cause statistics. I have analyzed the *cis-trans* isomerism in membrane proteins because to this no statistics exist it yet. Such statistics can be very important in the future. Connections between the *cis-trans* isomerism and the surrounding amino acids and the features of theses can be discovered. This analysis examines the Xaa-Pro *cis-trans* isomerization structure information. The PDBTM provides a list which contains the 389 protein chain entries. I can download the PDB files from the RCSB PDB (`http://www.rcsb.org/pdb/home/home.do`) using this list. Experimentally determined 3D coordinates of protein molecules are retrievable in the PDB. These data are derived from X-ray crystallography and NMR spectroscopy mostly.

Some 3D structures are also theoretical models. At present, the PDB is operating by RCSB and is the longest established information source for 3D structure data of proteins. However, must have knowledge about the construction so that can the PDB files analyze. The files have a standardized format. The information reside column oriented, i. e. in every line is always stored in columns defined before the same information. The format distinguishes between 51 line types which are indicated by keywords at the beginning of the line. The order of the lines is set predefined in which some lines are optional and are not published in every file. The description of the individual atoms takes the greatest amount of a file [wwPDB, 2012]. The PDB format is suited for regular expressions well and therefore is responsive well. With the help of one download function in Xaa-PIPT, be the PDB files local stored on a select path. The PDBTM list is to the submit Java class, the PDB and chain ID is extracted about a regular expression

```
String regex = "(\\w{4})_(\\w)";
```

and is submitted to the download method. The variable $s$ save the selected PDB ID and chain ID and integrated into the URL

```
ftp://ftp.wwpdb.org//pub//pdb//data//structures//all//pdb//pdb" +
s.toLowerCase() + ".ent.gz
```

and the download can start. The instruction `toLowerCase()` guarantees which submit all ID's in lower case letters. The two attitude possibilities can be recognized for the input file in the figure 2.1 on page 11. Xaa-PIPT supports two input formats, the PDB list and an PISCES input file (attainable about `http://dunbrack.fccc.edu/PISCES.php`). It is important that the files are saved locally for the analysis.

Example of the PDB list format:

```
2a06_E
2a06_G
...
```

Example of the PISCES input format:

```
IDs          length   Exptl.  resolution  R-factor FreeRvalue
7ODCA        424      XRAY        1.600     0.20     0.23
1GH9A         71      NMR            NA     1.00     1.00
...
```

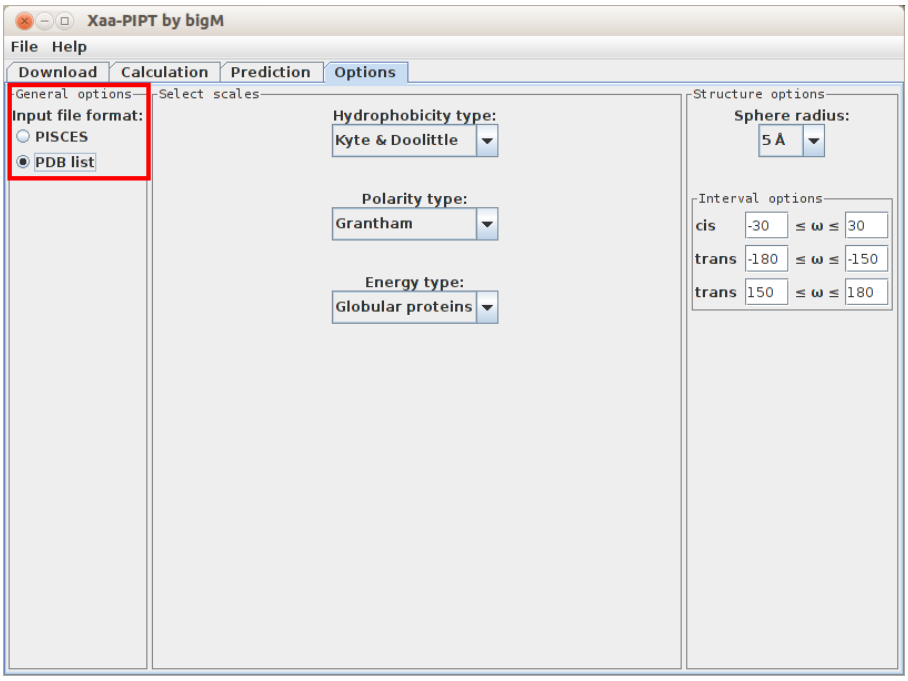


Figure 2.1: Options panel of Xaa-PIPT with red highlighted input file format options (PISCES and PDB list)

The download surface is represented in figure 2.2 on page 12. The choice of the input file path and download directory can be appointed here. The download of the PDB files then can be started by the download button.

## 2.2 Extraction of Information

Attitudes still can be carried out under the option panel in front of the real calculations. Under the part *select scales* can the user select under the categories *hydrophbicity type*, *polarity type* and *energy type*. These data are scaled respectively differently. Furthermore the user can select the *sphere radius* of $1$ to $10$ Å under *structure options*. The *sphere radius* indicates how large the spheric are formed around the proline. However, this says nothing about the influence of the surrounding amino acids. Amino acids are into sphere around proline with different features. These features also have influence of the *cis-trans* isomerism and the *sphere radius* predicates the amino acids are far distant from the proline. The omega angle *interval options* is the last attitude possibility. The user can select the interval independently. The interval therefore says omega angle as *cis* and *trans* where defines until. The interval for *cis* $-30° \leq \omega \leq 30°$ and for *trans* $-180° \leq \omega \leq -150°$ and $150° \leq \omega \leq 180°$ by default. For the calculation I have selected the *hydrophbicity type* Kyte & Doolittle [Kyte and Doolittle, 1982], the *polarity type* Grantham [Grantham, 1974] and the *energy type* transmembrane proteins [D. Labudde, 2012], for the *sphere radius* I chose $5$ Å. I defined my *interval options* on *cis* $-30° \leq \omega \leq 30°$ and for *trans* $-180° \leq \omega \leq -150°$ and $150° \leq \omega \leq 180°$. The

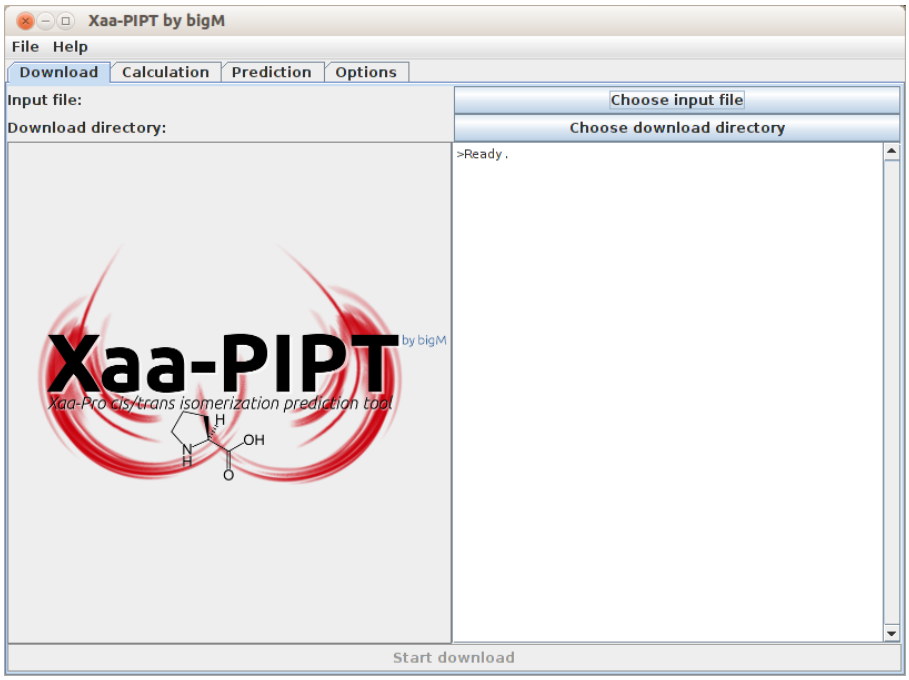


Figure 2.2: Download panel of Xaa-PIPT, input file and download directory can be chosen here and the download button is activated

option surface is represented in figure 2.3 on page 13. The next step is calculation of the features. This is carried out in the calculation surface. The essentially three-dimensional structure information is here possible. These include: secondary structure, inside/outside classification, proline energy and the sphere informations hydrophobcity, polarity, mutability, bulkiness and flexibility. A one-dimensional structure information is also still present, this is the sequence fragments. The position of the proline, the omega angle and the *cis-trans* conformation resulting from it is the base of all following calculations. All features listed in Figure 2.4 on page 13 illustrates in the calculation panel of Xaa-PIPT.

The calculations are used for the production of *cis-trans* statistics. Before the calculations can be started, the paths still must be chosen. The PDB list must be submitted to Xaa-PIPT again, a output directory for the PIPT files must be chosen and the list of the PDB files must be indicated. After that the activation of the calculation button is carried out. All calculated informations is saved in the PIPT file (the files end on .pipt). All the PIPT files have a fixed defined header. The informations is saved as a ASCII-text and separated with a semicolon, the typical CSV format.

Section of a example PIPT file (3zux_A.pipt):

```
res;omega;cis/trans;xPx;xxPxx;xxxPxxx;ss_i-2;ss_i-1;ss;ss_i+1;...
30;-178.0532;trans;APD;FAPDT;FFAPDTF;h;h;h;h;h;outside;-0.0031;...
38;178.0693;trans;GPY;AGPYI;WAGPYIP;h;h;h;h;h;outside;-0.0087;...
```

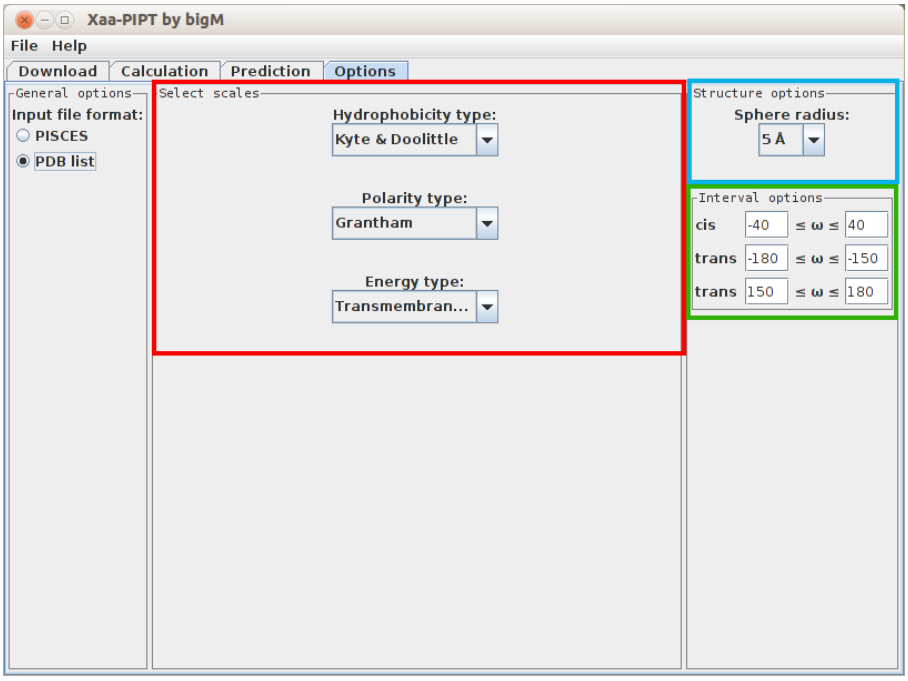


Figure 2.3: Options panel of Xaa-PIPT, red highlighted area shows *select scales* with the attitude possibilities, blue highlighted area shows *structure options* with the attitude possibilities and green highlighted area shows *interval options* with the attitude possibilities

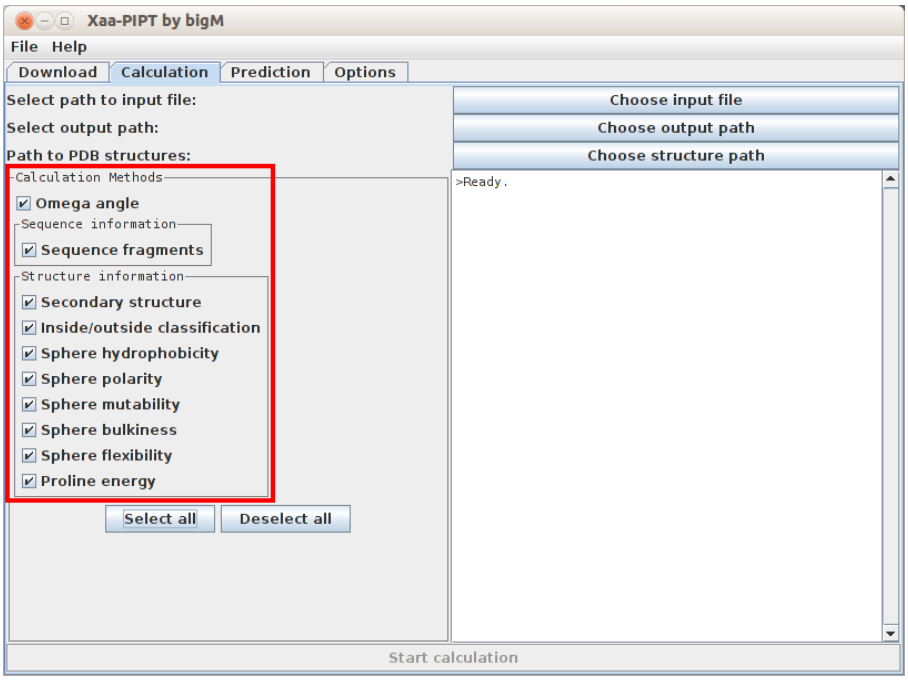


Figure 2.4: Calculation panel of Xaa-PIPT, red highlighted area shows all enabled structure informations

### 2.2.1 Calculation The Structure Information

Following calculations and information is the surrounding information to Xaa-Pro. Each individual feature is expressed as a numeric value and is used for the SVM input. The SVM supports only numeric values and multi categorical represented as vectors in binary format. Only complete dataset be used for the calculation. If a feature is missing, the complete incomplete dataset is removed. So no omissions arise and each property is equally often represented.

First is the angle calculated omega and the corresponding conformation. Residue number, which angle and isomerism, *cis*, *trans* and *cis/trans*, are stored in the PIPT-file. The following is required for the calculation of omega angle. First three vectors are needed. The vector $b_1$ represents the bond between the atomic bond $C^{\alpha}_{(i-1)}$ and $C'_{(i-1)}$, $b_2$ represents the bond between $C'_{(i-1)}$ and $N_{(i)}$ and $b_3$ represents the bond between $N_{(i)}$ and $C^{\alpha}_{(i)}$.

$$b_1 = \begin{pmatrix} C'_{(i-1)_x} - C^{\alpha}_{(i-1)_x} \\ C'_{(i-1)_y} - C^{\alpha}_{(i-1)_y} \\ C'_{(i-1)_z} - C^{\alpha}_{(i-1)_z} \end{pmatrix}, b_2 = \begin{pmatrix} N_{(i)_x} - C'_{(i-1)_x} \\ N_{(i)_y} - C'_{(i-1)_y} \\ N_{(i)_z} - C'_{(i-1)_z} \end{pmatrix}, b_3 = \begin{pmatrix} C^{\alpha}_{(i)_x} - N_{(i)_x} \\ C^{\alpha}_{(i)_y} - N_{(i)_y} \\ C^{\alpha}_{(i)_z} - N_{(i)_z} \end{pmatrix}$$

The vectors are used in the $atan2(y,x)$ function (equation 2.1). The $atan2(y,x)$ function can observe the sign of the parameter $y$ and $x$. The function returns the arctangent, or the inverse tangent of y and x. Returns the result in radians or degrees, which is dependent on angle settings. The interval returned then by $-\pi$ to $\pi$ (radians) or from $-180^{\circ}$ to $180^{\circ}$ (degrees).

$$\omega = atan2(|b_2|b_1 \cdot [b_2 \times b_3], [b_1 \times b_2] \cdot [b_2 \times b_3]) \tag{2.1}$$

The calculation on several individual functions distributed in the program code. The memory in the PIPT-file, the residue number, the calculated angle and the accompanying conformation is traded in the program with the following three methods:

1. the result stores for *cis*:

```
result.add(currentPro.getResidueNumber() + ";" + currentOmega +
";" + "cis");
```

2. the result stores for *trans*:

```
result.add(currentPro.getResidueNumber() + ";" + currentOmega +
";" + "trans");
```

3. the result stores for *cis/trans*:

```
result.add(currentPro.getResidueNumber() + ";" + currentOmega +
";" + "cis/trans");
```

The next structure information is the secondary structure. The secondary structure elements $\alpha$-helix, $\beta$-strand and coil are treated here. For the SVM input the secondary structures get a numeric value, the $\alpha$-helix receives the value $-1$, $\beta$-strand receives the value $0$ and coil receives the value $1$. In addition, the secondary structure element of the amino acids before and behind proline are investigated. Proline (position $i$) stands in the middle and the secondary structure element of the positions $i-2$, $i-1$, $i+1$ and $i+2$ are specified. BioJava has a simple solution for the extract of the secondary structure provided (`getSecondaryStructure()`).

The inside/outside classification is done in equation1.2 and definition1.1 on page 7 in section1.4. For the implementation of the equation are the coordinates of the $C^{\alpha}$ $(P_{C^{\alpha}} = (x,y,z))$ and $C^{\beta}$ $\left(P_{C^{\beta}} = (x,y,z)\right)$ atoms of proline required and the centroid $c$ must be determined. Depending on the where the proline is located in the structure, it is classified as inside or outside. For the SVM classification is value $-1$ $inside$ and $outside$ of the value $1$ reserved.

The sphere information includes the surrounding amino acids within a defined radius. This amino acids have different physicochemical and other properties, these interact differently with the proline and the Xaa-Pro conformation. Therefore they are surrounding amino acids on this investigated. The radius can be chosen between $1$ Å and $10$ Å. I have chosen a radius of $5$ Å. This was recognized by many runs as best (see section 3.3 on page 28). Described in section 1.4 on page 5 is already a scaling of the amino acids in dependence their properties on ProtScale (a category of the Expert Protein Analysis System(ExPASy) server). The scaling is not sufficient alone to normalize the amino acid properties in the sphere. The density ($\rho$) is used for the normalization. Instead of the mass is the quantity of occurring amino acids ($q_{aa}$) used in the sphere with a defined radius ($r$). The volume ($V$) of the sphere is used for the volume. The following equation 2.2 shows the connectedness:

$$\rho = \frac{q_{aa}}{\frac{4}{3}\pi r^3} \tag{2.2}$$

Therefore, all properties of the amino acids with this equation2.2 have been normalized. The unit calculated for this equation is $\left[\frac{aa}{\mathring{A}^3}\right]$.

I have used the scale values of by Kyte and Doolittle [Kyte and Doolittle, 1982] for the property of hydropathicity. These values are listed in the following table 2.1 on page 16. The table contains positive and negative scale values. These provide an overview of the type and starch of the hydropathicity of the different amino acids. The positive values represent thereby the hydrophobic amino acids and the negatives represent the hydrophilic amino acids.

Table 2.1: Hydropathicity scale by Kyte and Doolittle [Kyte and Doolittle, 1982]

| **Amino acid** | Ala | Arg | Asn | Asp | Cys | Glu | Gln | Gly | His | Ile |
|---|---|---|---|---|---|---|---|---|---|---|
| **Hydropathicity** | 1.8 | -4.5 | -3.5 | -3.5 | 2.5 | -3.5 | -3.5 | -0.4 | -3.2 | 4.5 |
| **Amino acid** | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
| **Hydropathicity** | 3.8 | -3.9 | 1.9 | 2.8 | -1.6 | -0.8 | -0.7 | -0.9 | -1.3 | 4.2 |

The hydrophobicity value is calculated as follows. The sum of the hydrophobicity values of amino acids occurring in the sphere about proline is educated and will multiply with the dense spheres. This was implemented about the following program code in Xaa-PIPT:

```
sphereHphob = sphereHphob + properties.getHphob(aa);

sphereHphob = sphereHphob * sphere.size()/
(4/3*Math.PI*Math.pow(sphereRadius,3));
```

The polarity scale value was built by Grantham in 1974 and table 2.2 on page 16 shows these for the different amino acids [Grantham, 1974]. The more highly this value of an amino acid, the high polar for it and the smaller the value of amino acid, the polarity of this is less. Also the polarity is the same calculation. The sum of the polarity values of amino acids occurring in the sphere about proline is educated and will multiply with the density of these sphere.

Table 2.2: Polarity (p) scale by Grantham [Grantham, 1974]

| **Amino acid** | Ala | Arg | Asn | Asp | Cys | Glu | Gln | Gly | His | Ile |
|---|---|---|---|---|---|---|---|---|---|---|
| **Hydropathicity** | 8.1 | 10.5 | 11.6 | 13.0 | 5.5 | 10.5 | 12.3 | 9.0 | 10.4 | 5.2 |
| **Amino acid** | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
| **Hydropathicity** | 4.9 | 11.3 | 5.7 | 5.2 | 8.0 | 9.2 | 8.6 | 5.4 | 6.2 | 5.9 |

This was implemented about the following program code in Xaa-PIPT:

```
spherePolar = spherePolar + properties.getPolar(aa);

spherePolar = spherePolar * sphere.size()/
(4/3*Math.PI*Math.pow(sphereRadius,3));
```

The relative mutability (table 2.3) are also scaled values, defined by Dayhoff M.O., Schwartz R.M. and Orcutt B.C. in 1978. The amino acid alanine assigned the value $100$, all other relative mutability refer to this value. The researchers have determined the probability of a mutation for each amino acid in a period of evolutionary time. So a high indicates value a more frequent observed mutation of the considered amino acid relative to alanin. The surrounding property is also calculated as the hydrophobicity and polarity. The sum of the relative mutability values of amino acids occurring in the sphere about proline is educated and will multiply with the density of these sphere.

Table 2.3: Relative mutability of amino acids (Ala=100) scale by Dayhoff M.O., Schwartz R.M., Orcutt B.C. [Dayhoff M.O., 1978]

| **Amino acid** | Ala | Arg | Asn | Asp | Cys | Glu | Gln | Gly | His | Ile |
|---|---|---|---|---|---|---|---|---|---|---|
| **Hydropathicity** | 100 | 65 | 134 | 106 | 20 | 93 | 102 | 49 | 66 | 96 |
| **Amino acid** | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
| **Hydropathicity** | 40 | 56 | 94 | 56 | 120 | 97 | 18 | 41 | 72 | 5.9 |

This was implemented about the following program code in Xaa-PIPT:

```
sphereMutability = sphereMutability + properties.getMutability(aa);

sphereMutability = sphereMutability * sphere.size()/
(4/3*Math.PI*Math.pow(sphereRadius,3));
```

The property bulkiness (table 2.4 page 18) has been scaled 1968 from Zimmerman J.M., Eliezer N. and Simha R. and this property describes how bulky an amino acid residue is [Zimmerman et al., 1968]. The smaller the scale value, the less is the bulk of the amino acid residue. The calculation is also the same. The sum of the bulkiness values of amino acids occurring in the sphere about proline is educated and will multiply with the density of these sphere.

Table 2.4: Bulkiness scale by Zimmerman J.M., Eliezer N., Simha R. [Zimmerman et al., 1968]

| **Amino acid** | Ala | Arg | Asn | Asp | Cys | Glu | Gln | Gly | His | Ile |
|---|---|---|---|---|---|---|---|---|---|---|
| **Hydropathicity** | 11.50 | 14.28 | 12.82 | 11.68 | 13.46 | 14.45 | 13.57 | 3.40 | 13.69 | 21.40 |
| **Amino acid** | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
| **Hydropathicity** | 21.40 | 15.71 | 16.25 | 19.80 | 17.43 | 9.47 | 15.77 | 21.57 | 18.03 | 21.57 |

This was implemented about the following program code in Xaa-PIPT:

```
sphereBulkiness = sphereBulkiness + properties.getBulkiness(aa);

sphereBulkiness = sphereBulkiness * sphere.size()/
(4/3*Math.PI*Math.pow(sphereRadius,3));
```

The average flexibility scale value was built by Bhaskaran R. and Ponnuswamy P.K. in 1974 and table 2.5 shows these for the different amino acids [Bhaskaran R., 1988].The amino acid flexibility describes the conformation the respective amino acid residue. The value is high, there is a high flexibility. This property is also calculated as all the other sphere features. The sum of the average flexibility values of amino acids occurring in the sphere about proline is educated and will multiply with the density of these sphere. But for the energy calculation in the sphere, is the sum of all occurring energy educated in the sphere.

Table 2.5: Average flexibility index scale by Bhaskaran R. and Ponnuswamy P.K. [Bhaskaran R., 1988]

| **Amino acid** | Ala | Arg | Asn | Asp | Cys | Glu | Gln | Gly | His | Ile |
|---|---|---|---|---|---|---|---|---|---|---|
| **Hydropathicity** | 0.36 | 0.53 | 0.46 | 0.51 | 0.35 | 0.49 | 0.50 | 0.54 | 0.32 | 0.46 |
| **Amino acid** | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
| **Hydropathicity** | 0.37 | 0.47 | 0.30 | 0.31 | 0.51 | 0.51 | 0.44 | 0.31 | 0.42 | 0.39 |

This was implemented about the following program code in Xaa-PIPT:

```
sphereFlexibility = sphereFlexibility + properties.getFlexibility(aa);

sphereFlexibility = sphereFlexibility * sphere.size()/
(4/3*Math.PI*Math.pow(sphereRadius,3));
```

The inside/outside classification was defined in section 1.4 on page 5. This is the basis for the amino acids energy approximation. The equation 2.3 for the energy approximation can be found at the ePros (`http://bioservices.hs-mittweida.de/Epros/Index?page=intro`).

$$e_{aa} \propto \ln\left(\frac{n_{aa}^{in}}{n_{aa}^{out}}\right) \rightarrow e_{aa_{ij}} = \left(e_{aa_i} + e_{aa_j}\right) \rightarrow E_{aa_i} = \frac{1}{2}\sum_{<i,j>} e_{aa_{i,j}} \tag{2.3}$$

The energy of any protein is generally through the first part of the equation given. The two parameters $\left(n_{aa}^{in}\right)$ and $\left(n_{aa}^{out}\right)$ are equal of observed amino acids that are buried in the structure or that are exposed on the solvent at the protein surface. The two parameters are derived from already known membrane and globular proteins. The interaction energy $e_{ij}$ between the two amino acids $a_i$ and $a_j$ is equal to the sum of $e_i$ and $e_j$. In the second part of the equation, the $E_i$ of the residue of the amino acid $A_i$ is then calculated. The total energy in a protein structure corresponds to the interaction over all amino acids. Therefore, the sequence of all occurring energy corresponds to the energy profile of a protein structure.

Table 2.6: Amino acid-specific average coarse-grained energies, 146 non-redundant membrane proteins [D. Labudde, 2012]

| **Amino acid** | Ala | Arg | Asn | Asp | Cys | Glu | Gln | Gly | His | Ile |
|---|---|---|---|---|---|---|---|---|---|---|
| **Hydropathicity** | 7.88 | 24.29 | 17.84 | 18.34 | 8.43 | 20.45 | 21.16 | 6.26 | 16.68 | 14.58 |
| **Amino acid** | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
| **Hydropathicity** | 16.93 | 27.62 | 14.51 | 11.97 | 14.83 | 10.02 | 13.33 | 15.17 | 14.03 | 13.11 |

This was implemented about the following program code in Xaa-PIPT:

```
ePro = ePro + properties.getEnergy(aa);
```

## 2.3 Training Dataset for SVM

Each property is a feature and each of these has adopted a discrete value. All proline have the same number of features entries. The proline entries are missing where take place does not calculate the different properties found. Therefore a less proline features in the training, as found in total. There are different reasons. the omega angle could not get calculated because no amino acid stands in front of the proline. If a selected feature cannot be calculated, the whole proline is removed from the training dataset. It is also possible that no proline in a chain occurs. In the following table 2.7 on page 20 are the values collected.

Table 2.7: Proline found in all chains; proline transmitted in the training dataset

| | **total** | **relative** | ***cis*** | ***trans*** | ***cis/trans*** | **undefined** |
|---|---|---|---|---|---|---|
| found prolines in dataset | 4026 | 100 % | 81 | 3813 | 35 | 97 |
| received prolines in training | 3859 | 95.85 % | 81 | 3783 | – | – |

The table shows that of the $4026$ (*cis*$= 81$ and *trans*$= 3813$) found prolines a number of $3866$ (*cis*$= 81$ and *trans*$= 3783$) prolines in the training dataset are imported.

The two conformations are classified with $-1$ for *cis* and $1$ for *trans* in the SVM training dataset. The *cis/trans* conformation is not marketable in the classification, because these SVM supports only two classes. *Cis* and *trans* are the class label and are available first, followed by the features as a discrete value.

Section of a example SVM input file:

```
...
-1.0 6:1.0000 7:0.3963 8:-0.8241 9:-0.8339 10:-0.8784 11:-0.8107
-1.0 1:1.0000 2:1.0000 3:1.0000 4:1.0000 5:1.0000 6:1.0000
-1.0 1:1.0000 2:1.0000 3:1.0000 4:1.0000 5:1.0000 6:1.0000
...
```

The creation of the training files is performed also with Xaa-PIPT. Here choose the features that are used for training and future prediction. You can choose the individual features on the generate SVM input surface in the task of prediction. The directories for the created PIPT files and the output are defined here and selected. The generate SVM input surface is represented in figure 2.5 on page 21.

The conformation *trans* is very much more frequently represented as the *cis*. Thus, there is an overrepresented class *trans* and an underrepresented class *cis*. The conformation occur unevenly distributed. The SVM training input requires a uniform distribution. Otherwise, it comes to unbalanced train of the dataset. This would lead to a distorted and incorrect prediction. A regulation via the selection of randomly 70% of the underrepresented *cis* class. The same number of overrepresented *trans* class are chosen randomly. This process is repeated five times. This is to ensure that occur all classes vectors at the same rate and that all classes vectors are represented. From this, I have created five files. These are necessary for the exercise of the classifier and the testing data. In Xaa-PIPT you can select the number of randomly generated and balanced SVM datasets (figure 2.6, page 22). Furthermore the percentage of elements for underrepresented class per dataset (figure 2.6, page 22).

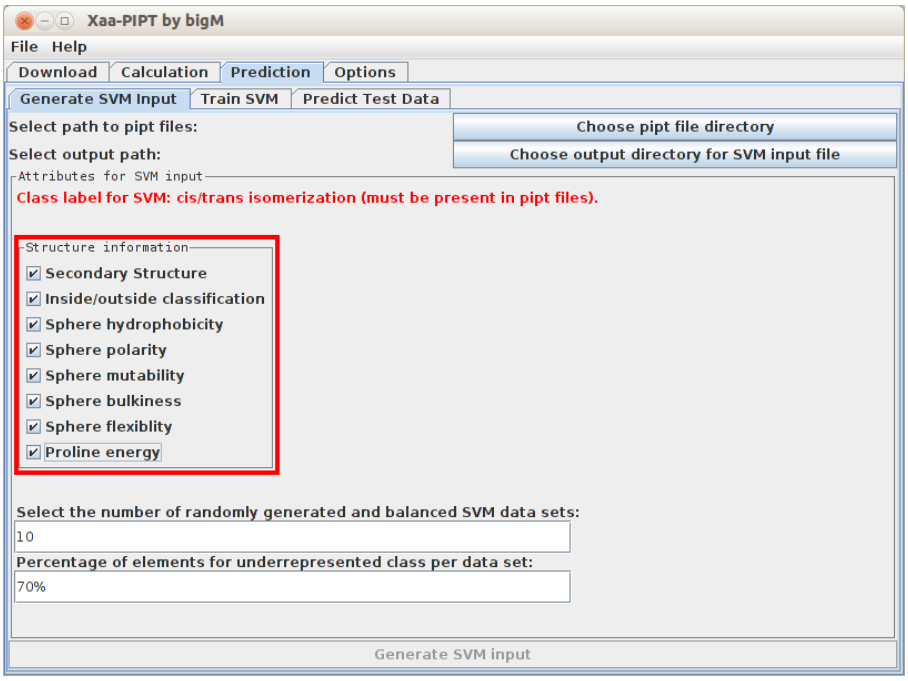


Figure 2.5: Generate SVM input surface of Xaa-PIPT, red highlighted area shows *structure information* with the attitude possibilities

I select five for the number of randomly generated and balanced SVM datasets and $70\%$ for the percentage of elements for underrepresented class per dataset. These settings are sufficient for a small set of data. In addition, I have selected all of the structure information for the training. The following five files have been created based on these settings.

```
svm_balanced_1.pipt
svm_balanced_2.pipt
svm_balanced_3.pipt
svm_balanced_4.pipt
svm_balanced_5.pipt
```

In these files are $70\%$ of the underrepresented *cis* class, with $58$ entries and the same number of randomly of the overrepresented *trans* class, with $57$ entries. So a file contains $115$ entries.

The number of containing entries is too low. A training for the SVM with these low information leads to no useful training results.

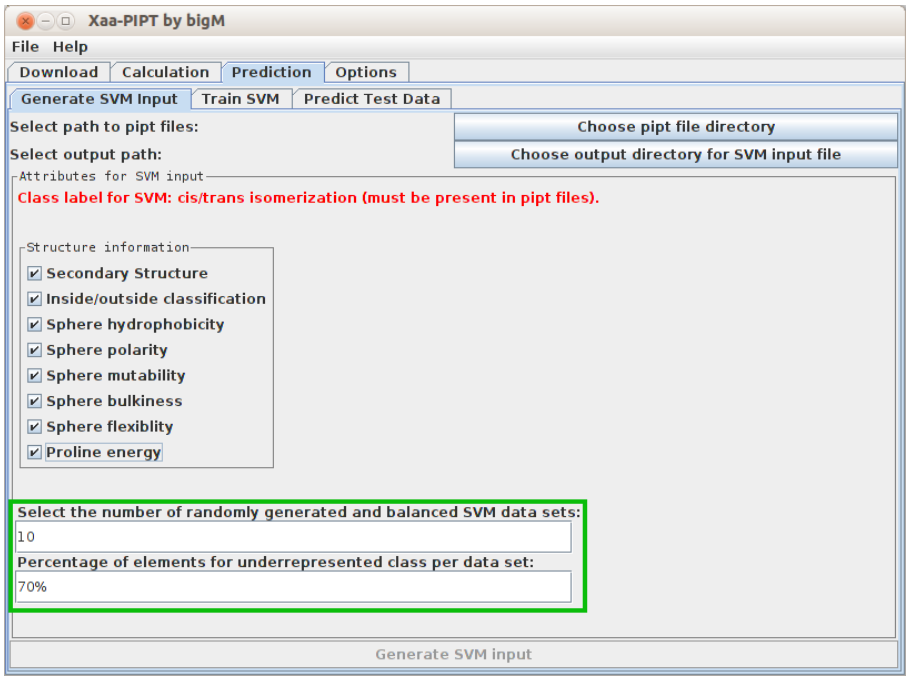


Figure 2.6: Generate SVM input surface of Xaa-PIPT, green highlighted area shows settings option for the dataset

## 2.4 Scaling for Consistency

Scaling is an important step for the apply of the SVM. This prevents the large numbers suppress small numbers. The $[-1;+1]$ rank is made for linear scaling. The scale for the training and testing must be equivalent. The advantage to avoid numeric difficulties in the calculation. This is done through a maximum and minimum value. It is also important that the same scaling for training dataset and test dataset is made [Chih-Wei and Chih-Chung, 2010].

## 2.5 SVM Feature Selection

There are many different ways to do a feature selection. The SVM feature selection was not performed with Xaa-PIPT. For this, a tool was provided by the libSVM that brings a quick prediction for the best features. The tool is a python script named **fselection.py**. The absolute advantage of the tool is the speed. But the disadvantage is that it shows no relationship between the individual features and it is also not obvious why the features were so selected. Nevertheless is a very good tool that provides a quick overview.

## 2.6 SVM Parameter Grid-Search and Cross-validation

The two parameters $C$ and $\gamma$ are for RBF kernel. The parameter values of a problem are not known previously. Therefore, a parameter search (model selection) must be made. The goal of this search is good to identify the parameters $C$ and $\gamma$. So that a classification of unknown data is possible and thus can predict test records. It should be noted that it is may not be useful to achieve a high training accuracy. In detail, a classification that exactly predicts training data to their class labels are known [Chih-Wei and Chih-Chung, 2010].

The goal is to divide the dataset into two datasets. The predictive accuracy that is obtained from a unknown set, affects the more precisely classification of an independent dataset. An improved process can be very useful. This process is known as cross-validation.

In the simple cross-validation, a record is divided into $K$ subsets of data. A dataset is selected as a test dataset. A regression model is adapted to the remaining $K-1$ datasets. These records shall be selected at random. This method is used to the parameters for the SVM to finding. Several pairs of parameters are tested and the pair with the best accuracy of cross-validation is selected. In addition, they are exponential growing sequences of the $C$ and $\gamma$ parameters current and suitable methods at find best parameters (for example, $C = 2^{-5}, 2^{-3}, ..., 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, ..., 2^{3}$).

Also, the grid search is not performed with Xaa-PIPT. Of the libSVM supplied another tool named **grid.py**, that supports the grid search. The tool is also a python script and it used the five-fold cross-validation.

## 2.7 SVM Training

After the best parameters were found that begins the training which with the five datasets. The SVM is trained with the three kernels. The best parameters that were used are in section 3.5 to find.

## 2.8 SVM Prediction

After that the best parameters were found, is a further cross-validation to find the best model file. From the five generated svm balanced files (section 2.3) a cross-validation executed, to perform representative predictions. As already described, all five models against the other four dataset (which were not used for the training of model) be predicted. The model file that brings the best prediction performance is implemented in $\mu$Xaa-PIPT to used for the subsequent Xaa-Pro predictions in protein structures.

# 3 Results

## 3.1 Quantities of *cis-trans* Conformations

An overview of the distribution of the 4010 available prolines is given in the piechard 3.1. Of the available prolines are 3891 in *trans*, 83 in *cis* and 36 in *cis/trans* conformation. A detailed overview is in table 3.1.

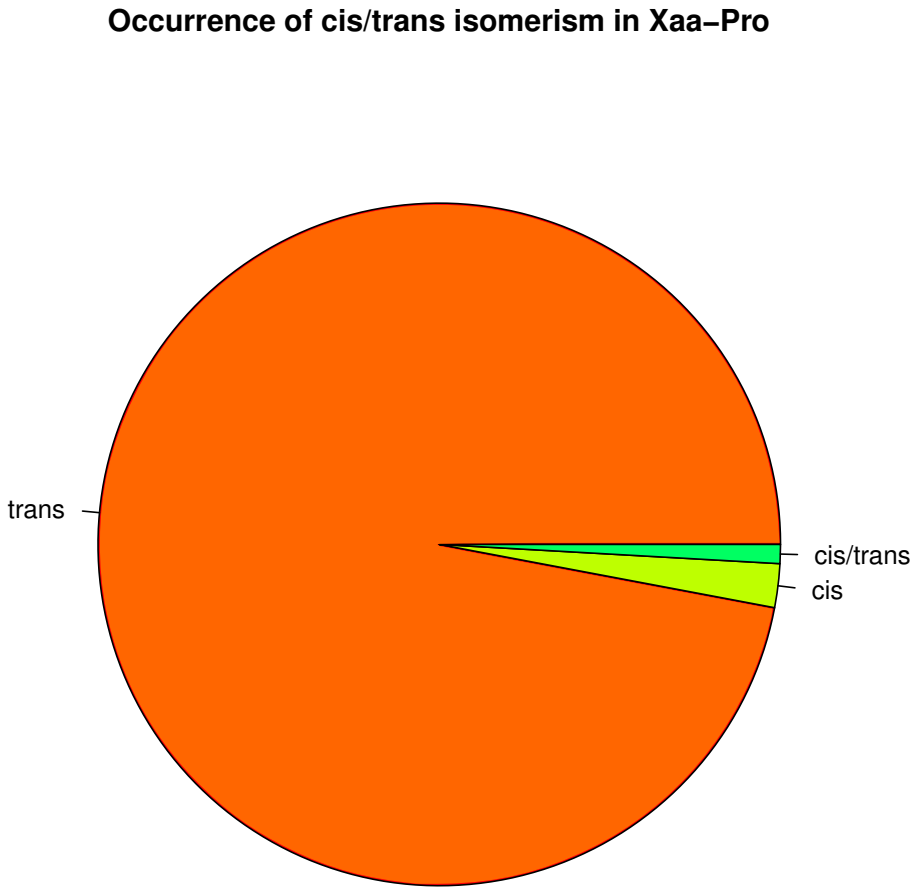


Figure 3.1: Overview of the distribution of the 4010 available prolines

Table 3.1: Overview the found conformation in 4010 prolines.

| conformation | total | relative |
|---|---|---|
| *trans* | 3891 | 97.0324% |
| *cis* | 83 | 2.0698% |
| *cis/trans* | 36 | 0.8949% |

## 3.2 Attribution of the Omega Angle

The boxplot is a graph that is used to plot the cardinal scaled data distribution. It summarizes these different robust spread and location mass in a presentation. A box plot gives an overview of the distribution of data in the different areas. That is why are all values of the so-called five point summary, so the median, represented the two quartiles and the two extremes.

The boxplot is used to represent the distribution of the omega angle, based on the state of the Xaa-Pro isomerization. The distribution of $83$ *cis* conformations is shown in figure 3.2. In the table 3.2 is an overview of the characteristic values.

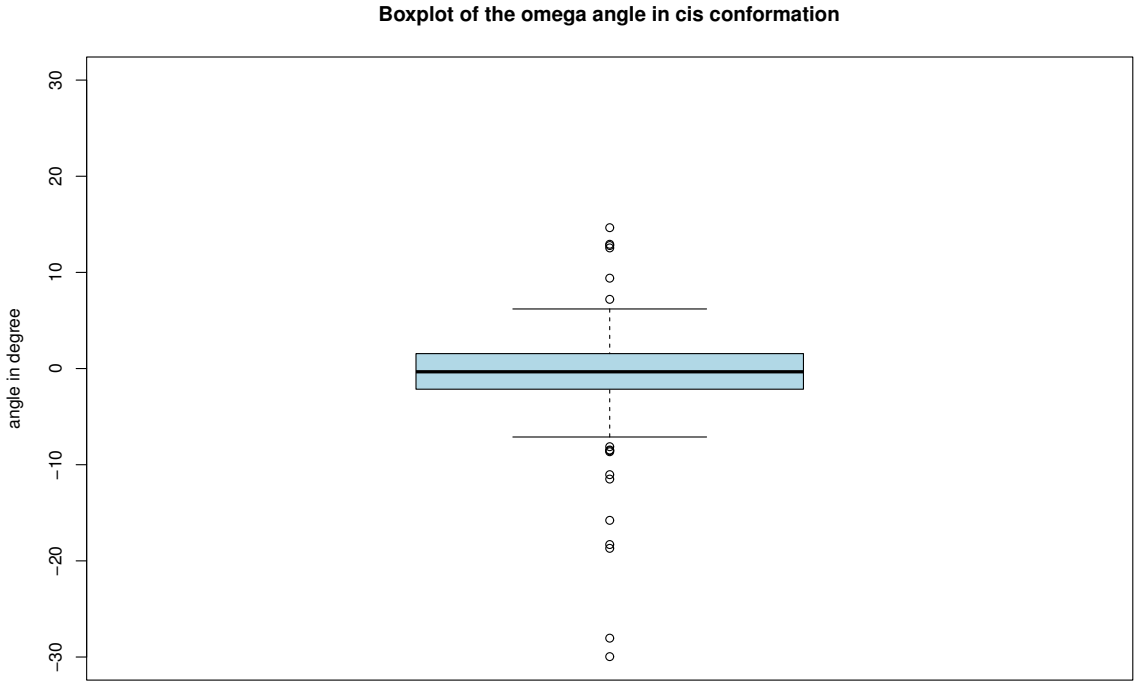


Figure 3.2: Boxplot for the distribution of the omega angle in *cis* conformation

Table 3.2: Distribution of the omega angle in *cis* conformation

| **boxplot parameter** | **characteristic value** |
|---|---|
| minimum | -29.9584 |
| 1st quantile | -2.1476 |
| median | 0.3348 |
| mean | -1.1558 |
| 3st quantile | 1.5539 |
| maximum | 14.6509 |

For the *trans* conformation two illustrations are prepared to the better overview. The

omega angle of *trans* in the range of $150^\circ$ to $180^\circ$ is shown in figure 3.3. In the table 3.3 is an overview of the characteristic values of these range.

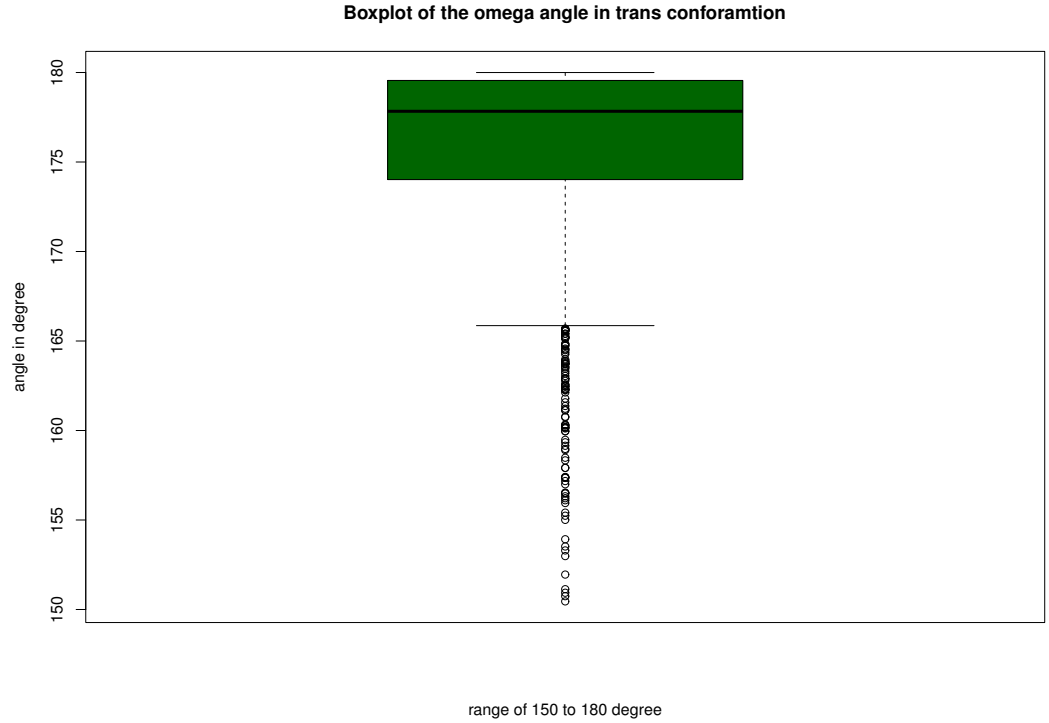


Figure 3.3: Boxplot for the distribution of the omega angle in *trans* conformation; $150^\circ$ to $180^\circ$

Table 3.3: Distribution of the omega angle in *cis* conformation

| **boxplot parameter** | **characteristic value** |
|---|---|
| minimum | 150.5 |
| 1st quantile | 174.0 |
| median | 177.8 |
| mean | 176.0 |
| 3st quantile | 179.6 |
| maximum | 180.0 |

The omega angle of *trans* in the range of $-180^\circ$ to $-150^\circ$ is shown in figure 3.4 on page 28. In the table 3.4 on page 28 is an overview of the characteristic values of these range.

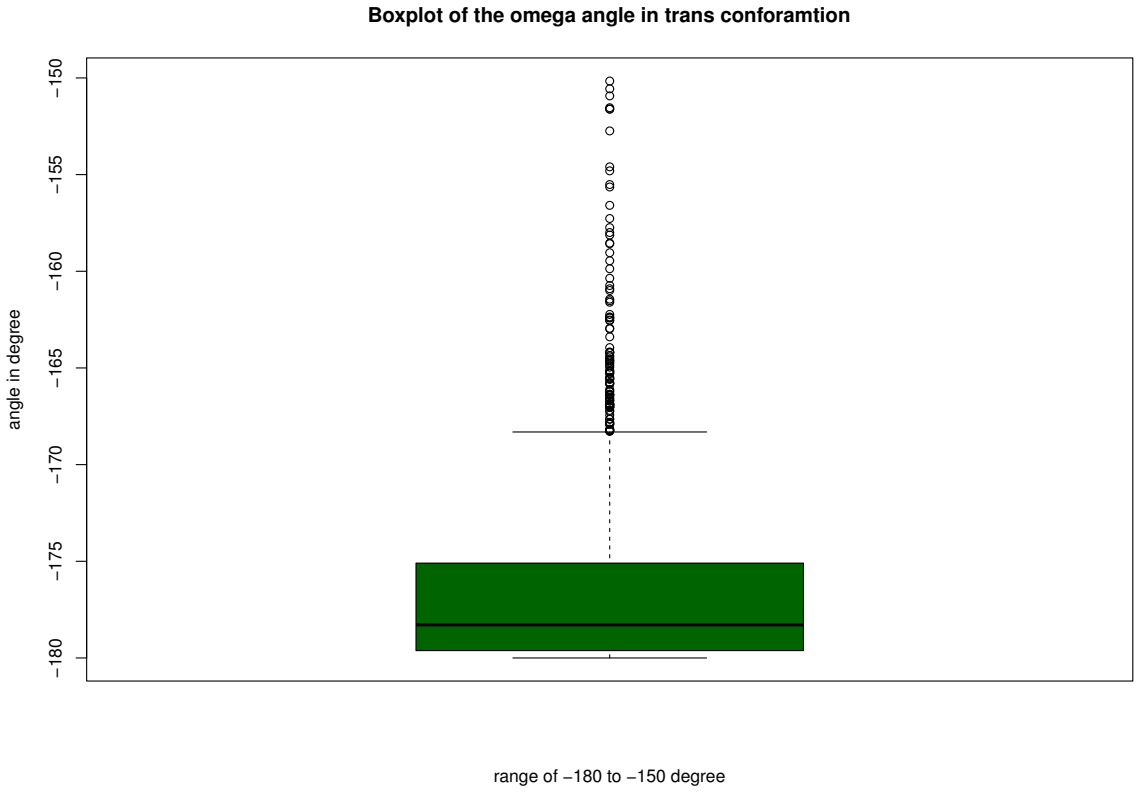


Figure 3.4: Boxplot for the distribution of the omega angle in *cis* conformation

Table 3.4: Distribution of the omega angle in *trans* conformation; range of $-180°$ to $-150°$

| boxplot parameter | characteristic value |
|---|---|
| minimum | -180 |
| 1st quantile | -179.6 |
| median | -178.3 |
| mean | -176.6 |
| 3st quantile | -175.1 |
| maximum | -150.2 |

## 3.3 Sphere Radius Specification

The different features were tested with the different radii for the determination of optimal sphere radius. A five-fold cross-validation was performed about the $322$ membrane protein chains. The loose grid search was carried out with RBF kernel and with the parameters $C : [2^{-5}; 2^{15}]$ and $\gamma : [2^{-15}; 2^{3}]$ with gradual increase of $1$. The average grid search accuracy were identified and considered. From the respective individual grid search results, the subset was formed and provided with the other compared to and associated with the radial in context. Table 3.5 are these values to find. It should be

noted that the consideration of the radii of $1$ Å and $2$ Å have not been tested. There are not enough information in this small radius.

Table 3.5: Average prediction compared of sphere radius (used RBF kernel)

| radius in Å | average prediction in percent |
|---|---|
| 3 | 65.4546 |
| 4 | 68.5217 |
| 5 | **71.1304** |
| 6 | 69.5652 |
| 7 | 70.9565 |
| 8 | 69.3956 |
| 9 | 70.9565 |
| 10 | 69.5652 |

The best accuracy was calculated in a radius of $5$ Å (highlighted in table 3.5). The accuracy is $71.1304\%$. The radius has the lowest accuracy of all was $3$ Å this is $65.4546\%$. The radius of $10$ Å has still an accuracy of $69.5652\%$. The following figure 3.5 on page 29 shows the facts.

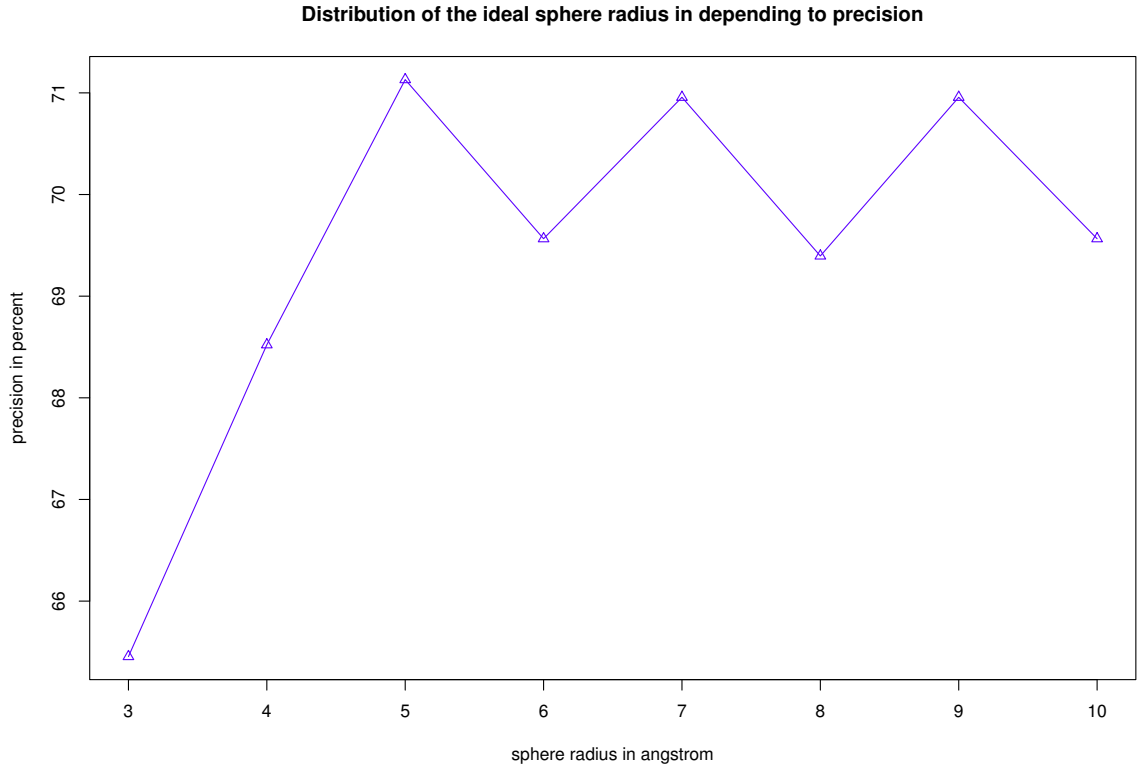


Figure 3.5: Accuracy depending on sphere radius

## 3.4 F-score for Features

The Python tool (fselect. py) helps to find the best feature. The F-score is used to describe the feature. In table 3.6 is an overview of the numeric values of individual datasets. The average was calculated from the feature F-scores of the different features.

Table 3.6: F-score compared to the datasets

| features | dataset1 | dataset2 | dataset3 | dataset4 | dataset5 | average |
|---|---|---|---|---|---|---|
| $ss_{i-2}$ | 0.209682 | 0.060137 | 0.016156 | 0.090547 | 0.077662 | 0.090837 |
| $ss_{i-1}$ | 0.324751 | 0.166775 | 0.058465 | 0.125981 | 0.077468 | 0.150688 |
| $ss_{i}$ | 0.327682 | 0.152831 | 0.115528 | 0.178431 | 0.131742 | **0.181243** |
| $ss_{i+1}$ | 0.222466 | 0.045594 | 0.071047 | 0.061728 | 0.052816 | 0.090730 |
| $ss_{i+2}$ | 0.247381 | 0.060607 | 0.045611 | 0.075506 | 0.073106 | 0.100442 |
| inside/outside | 0.000032 | 0.001471 | 0.008668 | 0.014603 | 0.004009 | 0.005757 |
| $sphere_{hydrophobicity}$ | 0.129166 | 0.010148 | 0.026313 | 0.041931 | 0.033841 | 0.048280 |
| $sphere_{polarity}$ | 0.023890 | 0.024697 | 0.034741 | 0.026017 | 0.001158 | 0.022101 |
| $sphere_{mutability}$ | 0.016962 | 0.039019 | 0.030346 | 0.032392 | 0.005318 | 0.024807 |
| $sphere_{bulkiness}$ | 0.019015 | 0.043641 | 0.025581 | 0.032894 | 0.016360 | 0.027498 |
| $sphere_{flexibility}$ | 0.017214 | 0.021269 | 0.026313 | 0.014594 | 0.001770 | 0.016232 |
| $sphere_{energy\ Pro}$ | 0.040308 | 0.001288 | 0.000003 | 0.014958 | 0.002914 | 0.011894 |

The secondary structure reached the best average with $0.181243$, followed by the secondary structure element before prolin (position $i-1$) with $0.150688$. The inside/outside classification has the smallest average with $0.005757$

## 3.5 Results of Grid Search

After a loose grid search, a finer grid search is carried out to improve the prediction, it increases the accuracy of prediction. An improvement has not occurred and the values have not changed or improved. Therefore remain the parameters $C : [2^{-5}; 2^{15}]$ and $\gamma : [2^{-15}; 2^{3}]$ for the search.

For the search of the appropriate kernel, was a direct comparison made. For this comparison the same parameters were used for all kernel. So linear and polynomial were assigned to the parameters $C : [2^{-5}; 2^{15}]$ and the RBF kernel in addition still the $\gamma : [2^{-15}; 2^{3}]$. The five-fold cross-validation is done with 1 step-wise for all. The single results for the datasets and the average value is also included are listed in table 3.7.

Table 3.7: Kernel compared to the datasets

| | RBF | polynomyal | linear |
|---|---|---|---|
| dataset 1 | 75.6522% | **73.0435%** | **71.3043%** |
| dataset 2 | 69.5652% | 68.6957% | 68.087% |
| dataset 3 | 64.3478% | 61.7391% | 60.0% |
| dataset 4 | **76.5217%** | 72.1739% | 64.3478% |
| dataset 5 | 69.5652% | 68.6957% | 65.2174% |
| **average** | **71.1304%** | 68.8696% | 65.3913% |

The RBF kernel has brought the best results (highlighted in table 3.7). This bring the highest average prediction accuracy of $71.13\%$. The linear kernel has in the comparison to the other two the lowest prediction accuracy $65.39\%$. In the figure 3.6 on page 31 are the kernel in the direct comparison to the individual datasets. The datasets are under each other on the left side of the Dotchart.

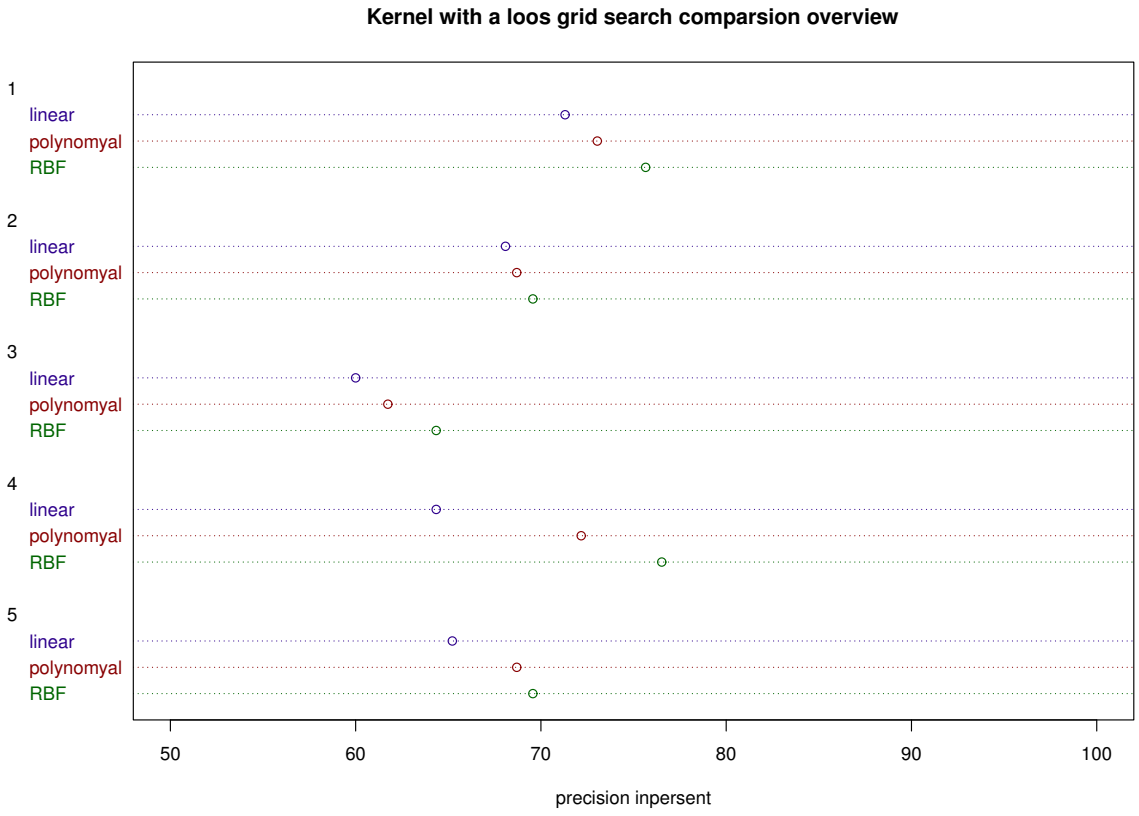


Figure 3.6: Accuracy depending on kernels

The figure 3.7 on page 32 shows the exact output and results of best record with the best prediction accuracy. The record 4 has the best prediction accuracy $76.52\%$. So were the best parameters $C = 2^{10}$ and $\gamma = 2^{-6}$ for the RBF kernel. For the polynomial

kernel were the best parameter $C = 2^{10}$ with a prediction accuracy $73.04\%$ and for the linear kernel were the best parameter $C = 2^{-3}$ with a prediction accuracy $71.30\%$. The RBF kernel is therefore the best of the three kernels.

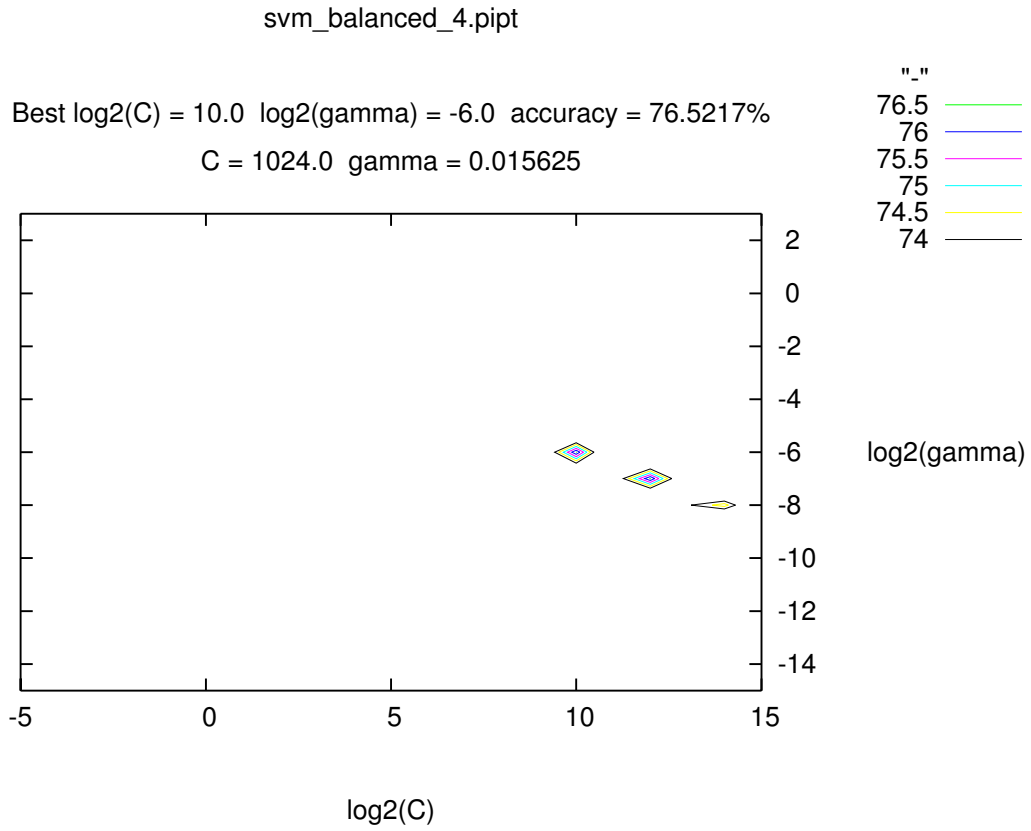


Figure 3.7: Overview of the loose grid search result with RBF kernel for dataset 4

## 3.6 Prediction Results

Five model files have been created with the RBF kernel and the parameters $C : [2^{-5}; 2^{15}]$ and $\gamma : [2^{-15}; 2^{3}]$. The five-fold cross-validation has determined the accuracy has failed. The five-fold cross-validation brings not a single file for the prediction. Therefore, the prediction is failure. The production prediction accuracy can be seen in table 3.8.

Table 3.8: Prediction performance of RBF kernel training with five-fold cross-validation

| dataset | accuracy in percent |
|---|---|
| 1 | 50.4348 |
| 2 | 50.4348 |
| 3 | 50.4348 |
| 4 | 50.4348 |
| 5 | 50.4348 |

Would made no prediction for the polynomial and linear kernel. They have brought in the grid search of poor results as the RBF kernel.

# 4 Discussion

## 4.1 Quantity of the Omega Angel

The occurrence of the *cis* conformation is $2.0698\%$, this value corresponds not to the value of $4-5\%$ recorded in the literature [Pahlke et al., 2005]. The less occurrence of *cis* can be dependent on the small dataset. There were $389$ membrane proteins with $4026$ prolines of the $4010$ evaluated were available. To membrane protein and their statistical study on the isomerization of Xaa-Pro, there is still no comparable study.

An extension of the omega angle for *cis* would result in the emergence of *cis* increases. It would not contribute to a exactly prediction determine. A few statistical outlier is located only in the range of $-20^\circ$ to $-30^\circ$, which can analyze neglected. The maximum value of the *cis* omega angle is located at $14.6509^\circ$. Therefore a restriction of the angle to $-20 \leq \omega \leq 20$ could be made. The interval of omega angle for the *trans* state is good selected for the feature extraction. The range are close off on the $180^\circ$ and $-180^\circ$. There are a few statistical outlier is located in $-150.2^\circ$ and $150.5^\circ$.

## 4.2 Sphere Radius

The best sphere radius has been determined at 5 Å . At this radius, the features were extracted the best. The average forecast has achieved a rate of $71.13\%$. I. e. in the environment, most of the information from the sphere could be determined. If the radius is reduced, then also the included information content reduced. In the environment, therefore, no information can be reported about the influence on the isomerism. Therefore an analysis of the radius under $3$ Å is not useful. The average predicts decreases. A similar result is illustrated in a radius of $10$ Å with an average prediction of $69.57\%$. Therefore become the surroundings information of increasing radius always has less influence on the isomerism. A radius of $5$ Å to choose I think useful. When a radius is selected, you can choose between the radii $4$ Å , $5$ Å and $6$ Å. This environment contained enough information, which guarantee a good average prediction.

If you have a large dataset available, you can better determine the exact radius for sphere. The small dataset are only an indication, no defiled statement made but it can be. Sure this too small and too large radii is the information content in the space decreases.

## 4.3 Features Valuation

The best feature of all is the secondary structure element, with a average value of $0.181243$. The F-score has rated the secondary structure as the best feature over the others. This feature was already by other researchers check and also evaluated a useful feature [Pahlke et al., 2005]. It also follow a secondary structure element, with a average value of $0.150688$. This is position in front of the proline $(i-1)$, thus the position that is instrumental for the omega angle to them. In biology, proline is a helix breaker, therefore it is very rarely found in a helix. Proline usually occur before and after a helix [Pahlke et al., 2005]. The top five features is all to secondary structure elements. The other studied environment properties are not very good. Sixth is hydrophilicity with a value of $0.048280$. This feature is well chosen, as helix regions mostly in the membrane are, which are rather hydrophobical. The bulkiness could also have influence on the fold, but this feature has not very often. Physical quantities have in general no great influence on the environment and of the Xaa-Pro *cis-trans* isomerism. Also have the features proline energy $(0.011894)$ and inside/outside $(0.005757)$ insufficient rated. The inside/outside classification is already designed for membrane proteins, is probably still not particularly well suitable. Proline energy it is also so here is the assumption that the energy calculation to coarse is. An own energy its for the sphere should be made here.

Summary can be assumed that the environment information to Xaa-Pro has an influence on the isomerism. This acceptance was again clarified in this work. Necessary, the other methods is further development for better predictions. At the moment they are still immature and roughly indicate the influence of the features the isomerism. In addition the one-sided view of the features. There could be still more environment information that has not yet been observed. The view of the amino acid property is still and is a good starting point for further research. Despite the relative dataset information content of the record, a feature look is quite possible. The normalization could be the reason for the bad room information. The density as a normalization factor is probably not appropriate. Better during defined space coordinates, that better describe the position. This method could be complicated because no individual discrete values could be used.

## 4.4 Assessment of Grid Search

The cross-validation is suitable for large amounts of data. However, this is not suitable for small data quantities. Therefore, the predictions of the trained data sets are not good. The RBF kernel has prevailed of all kernel. He brought the best prediction. However, these values not binding, are there which data set is very small and has therefore no great credibility. Have produced worse results the polynomial and the linear. The three kernel differ in the time. The polynomial spent the longest time in the calculation. A finer grid search is not specify eight in all three kernels. The small number of data would

have resulted in no improvement of the result and consume unnecessary time. For a larger dataset it is that through possible to win a good prediction.

## 4.5 Impact of Prediction

Five model files have emerged from the SVM training. The five-fold cross-validation could not be applied. It has emerged no model file, thus no prediction can be made. The model file with the best prediction accuracy can be used in $\mu$Xaa-PIPT, however, would lead to no significant prediction. Therefore $\mu$Xaa-PIPT not joined in this work used. $\mu$Xaa-PIPT cannot be used without a significant model file. If a model file has been created, it can be implemented in $\mu$Xaa-PIPT and a prediction can be made. No prediction for other membrane protein structures can be performed with the current record of membrane proteins. The main reason is that no prediction can take place the low number of available data. The methods for a prediction are created and can each time used.

## 4.6 Summary

The context of this work is the development of a method for the prediction of a *cis-trans* isomerism based on structural characteristics. The prediction should be made possible with an SVM. A prediction was not reached with the SVM. The prediction with the SVM method is suitable for predictions with a variety of datasets, however the datasets of the membrane proteins are not sufficient. The developed software Xaa-PIPT support the extraction of structure data for the compilation of the used features. $\mu$Xaa-PIPT not in the context of this work was used. A model file needed for $\mu$Xaa-PIPT is not created. No prediction for membrane proteins structures can be done without a model file. If a suitable model file exists it can be implemented in $\mu$Xaa-PIPT and be used for prediction of *cis-trans* isomerism in Xaa-Pro. The amino acids within a defined radius of proline have a direct impact on the isomerism. Extraction of the structure features are too rough, but are a good approach. The number of the membrane protein chains and the achievable proline are too low to make a statement.

## 4.7 Perspective

The basis for the extraction of structural characteristics is given and can be developed further. In the further development, the improvement of the methods and evaluation of the features in focus should be. It can be aimed at a uniform file format for further processing of the information. The SVM is a very good way for a prediction of two class method *cis-trans*. It is suitable for the search of suitable features, training of the data and the prediction of unknown protein structures. The low record is still a barrier to the

study. However the number of dissolved proteins is increasing, so that this problem its borders has. The datasets of the PDB grow constantly rise to. The PDF format is not very good for extracting information. Also here an improvement of consistency should in be made. Membrane proteins in regard of the *cis-trans* isomerism are still unexplored. The research in this direction should be extended. The exact biological research of the Xaa-Pro *cis-trans* isomerism should always takes into consideration are. There are a few studies that are extensive and give a little insight on the function and the benefits.

# Appendix A: Appendix

A DVD is attached in the appendix and contains the following content:

1. in directory: `aeisoldBachelorThesis/`
    - aeisoldBachelorThesis.pdf
2. in directory: `PDBTM_FullList/`
    - PDBTM_list.txt
3. in directory: `software/`
    - Xaa-PIPT.jar
4. in directory: `SVM_Data/models`
    - svm_balance_1.pipt.model
    - svm_balance_2.pipt.model
    - svm_balance_3.pipt.model
    - svm_balance_4.pipt.model
    - svm_balance_5.pipt.model
5. in directory: `SVM_Data/testing`
    - svm_balance_1.txt
    - svm_balance_2.txt
    - svm_balance_3.txt
    - svm_balance_4.txt
    - svm_balance_5.txt
6. in directory: `SVM_Data/testing`
    - svm_balance_1.pipt
    - svm_balance_2.pipt
    - svm_balance_3.pipt
    - svm_balance_4.pipt
    - svm_balance_5.pipt
    - svm_raw.pipt

- svm_scaled.pipt

The launch of the software requires the Java Runtime Environment (JRE).
*Launch Xaa-PIPT on Unix*:
Use to execute the command: `java -jar XaaPIPT.jar`

*Launch Xaa-PIPT on Windows*:
Double click the Xaa-PIPT.jar

# Bibliography

[Bhaskaran R., 1988] Bhaskaran R., P. P. (1988). Average flexibility index. Website. Available online at http://web.expasy.org/protscale/pscale/Averageflexibility.html; visited on May 15th 2012. [Reference: Int. J. Pept. Protein. Res. 32:242-255(1988)].

[Chih-Wei and Chih-Chung, 2010] Chih-Wei, H. and Chih-Chung, C.-J. L. (2010). A practical guide to support vector machine classification. Department of Computer Science,National University Taiwan, Taipei 106.

[D. Labudde, 2012] D. Labudde, F. Heinke, S. S. D. S. (2012). 2.1 amino acid-specific average coarse-grained energies. Website. Available online at http://bioservices.hs-mittweida.de/Epros/; visited on July 25th 2012 University of Applied Sciensces.

[Dayhoff M.O., 1978] Dayhoff M.O., Schwartz R.M., O. B. (1978). Relative mutability of amino acids (ala=100). Website. Available online at http://web.expasy.org/protscale/pscale/Relativemutability.html; visited on May 8th 2012. [Reference: Atlas of Protein Sequence and Structure, Vol.5, Suppl.3 (1978)].

[Exarchos et al., 2009] Exarchos, K. P., Papaloukas, C., Exarchos, T. P., Troganis, A. N., and Fotiadis, D. I. (2009). Prediction of cis/trans isomerization using feature selection and support vector machines. *J Biomed Inform*, 42(1):140–149. [DOI:10.1016/j.jbi.2008.05.006] [PubMed:18586558].

[Frishman, 2010] Frishman, D., editor (2010). *Structural Bioinformatics of Membrane Proteins*. Number 51. Springer, Holzhausen Druck und Neue Medien GmbH, 1140 Wien, Austria, 1st edition. ISBN-10: 3709100445; ISBN-13: 978-3709100448.

[Grantham, 1974] Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science*, 185(4154):862–864. [PubMed:4843792].

[Heinke and Labudde, 2012] Heinke, F. and Labudde, D. (2012). Membrane protein stability analyses by means of protein energy profiles in case of nephrogenic diabetes insipidus. *Comput Math Methods Med*, 2012:790281. [PubMed Central:PMC3312259] [DOI:10.1155/2012/790281] [PubMed:22474537].

[Kecman, 2005] Kecman, V. (2005). Support vector machines: Theory and applications. Springer Berlin.

[Kyte and Doolittle, 1982] Kyte, J. and Doolittle, R. F. (1982). A simple method for

displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157(1):105–132. [PubMed:7108955].

[Lu et al., 2007] Lu, K. P., Finn, G., Lee, T. H., and Nicholson, L. K. (2007). Prolyl cis-trans isomerization as a molecular timer. *Nat. Chem. Biol.*, 3(10):619–629. [DOI:10.1038/nchembio.2007.35] [PubMed:17876319].

[Lummis et al., 2005] Lummis, S. C., Beene, D. L., Lee, L. W., Lester, H. A., Broadhurst, R. W., and Dougherty, D. A. (2005). Cis-trans isomerization at a proline opens the pore of a neurotransmitter-gated ion channel. *Nature*, 438(7065):248–252. [DOI:10.1038/nature04130] [PubMed:16281040].

[Pahlke et al., 2005] Pahlke, D., Freund, C., Leitner, D., and Labudde, D. (2005). Statistically significant dependence of the Xaa-Pro peptide bond conformation on secondary structure and amino acid sequence. *BMC Struct. Biol.*, 5:8. [PubMed Central:PMC1087856] [DOI:10.1186/1472-6807-5-8] [PubMed:15804350].

[Wedemeyer et al., 2002] Wedemeyer, W. J., Welker, E., and Scheraga, H. A. (2002). Proline cis-trans isomerization and protein folding. *Biochemistry*, 41(50):14637–14644. [PubMed:12475212].

[wwPDB, 2012] wwPDB (2012). Pdb entry requirements. Website. Available online at `http://www.wwpdb.org/policy.html#toc_requirements`; visited on July 23th 2012.

[Zimmerman et al., 1968] Zimmerman, J. M., Eliezer, N., and Simha, R. (1968). The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.*, 21(2):170–201. [PubMed:5700434].

# Glossary

**amino acid** Are a class of small organic molecules with at least a carboxyl group ($-COOH$) and at least a amino group ($-NH_2$).

**ASCII-text** General interface for text files to the transfer in arbitrary word processing programs.

**CSV** Save values and text of the active table if data columns are separated by separator.

**Expert Protein Analysis System** EyPASy is the SIB (Swiss Institute of Bioinformatics) bioinformatics resource portal which access provides databases and tools in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. Also It is a portal for many resources of different SIB groups and of external institutions.

**F-score** Is a value indicating the significance of the feature.

**FASTA** The FASTA format is the simplest sequence format. It starts with a arrow bracket (>), followed by it sequence name. After that a description of the sequence still can attach. The sequence then is in the second line in the one-letter-code.

**feature** Classes variable which are used for the input of the SVM classification. Each feature represents a property and is described in a numerical value..

**libSVM** Is a software library for the support of the SVM. It contains individual libraries for different programming languages.

**PPIases** Enzym type, catalyzes the change from *cis* in the *trans* position (and reversed) and therefore is important at the protein folding.

**SASA** Is a very simple solvent model represents the solvent accessible surface area approximation. The free solvent energy ($\Delta G_{\text{solvation}}$) is calculated the sum of the empirical parameters ($\sigma_{\text{i}}$)(atomic surface tension) and the surface ($A_{\text{i}}$) of the atom $i$ which is accessible to the solvent.

**SVM** A machine learning technique and is used for prediction of unknown data based on known data..

# Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Mittweida, im August 2012