HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

MASTER'S THESIS

B. Sc.
Julia Abel

Investigation of mirror/native structures in biomolecules

2020

Faculty of **Applied Computer and Life Sciences**

MASTER'S THESIS

Investigation of mirror/native structures in biomolecules

Author:
Julia Abel

Study Programme:
Molecular Biology/Bioinformatics (M. Sc.)

Seminar Group:
MO18w1-M

First Referee:
Prof. Thomas Villmann

Second Referee:
Dr. Marika Kaden

Mittweida, August 2020

Bibliographic Information

Abel, Julia: Investigation of mirror/native structures in biomolecules, 35 pages, 8 figures, Hochschule Mittweida, University of Applied Sciences, Faculty of Applied Computer and Life Sciences

German Title: *Native oder gespiegelte Strukturen in Biomolekülen*

Master's Thesis, 2020

Typesetting: LaTeX

Abstract

In bioinformatics one important task is to distinguish between native and mirror protein models based on the structural information. This information can be obtained from the atomic coordinates of the protein backbone. This thesis tackles the problem of distinction of these conformations, looking at the statistics of the dihedral angles' distribution regarding the protein backbone. This distribution is visualized in Ramachandran plots. By means of an interpretable machine learning classification method – Generalized Matrix Learning Vector Quantization – we are able to distinguish between native and mirror protein models with high accuracy. Further, the classifier model supplies supplementary information on the important distributional regions for distinction, like α-helices and β-strands.

German Abstract

Eine wichtige Aufgabenstellung der Bioinformatik ist es, zwischen nativen und gespiegelten Proteinmodellen (Konformationen) aufgrund von struktureller Information zu unterscheiden. Diese Information kann von den Atomkoordinaten des Proteinrückgrates erhalten werden. Diese Arbeit befasst sich mit der Problematik der Konformationsunterscheidung, indem die Verteilung der dihedralen Winkel des Proteinrückgrates statistisch betrachtet wird. Diese Verteilung wird in Ramachandran-Plots dargestellt. Mithilfe einer interpretierbaren Klassifizierungsmethode des maschinellen Lernens – *Generalized Matrix Learning Vector Quantization* – kann zwischen nativen und gespiegelten Proteinmodellen mit hoher Genauigkeit unterschieden werden. Des Weiteren liefert das Klassifikationsmodell zusätzliche Informationen zu den für die Unterscheidung relevanten Winkelverteilungen, wie α-Helices und β-Faltblättern.

I. Contents

II. List of Figures

III. List of Tables

IV. Nomenclature

Biological Nomenclature

D	Dextrorotatory
L	Levorotatory
PDB	Protein Data Bank
SCOP	Structural Classification of Proteins
SCOPe	Structural Classification of Proteins – extended

Mathematical Nomenclature

AI	Artificial Intelligence
ARS	Attraction Repulsion Scheme
FN	False Negatives
FP	False Positives
GLVQ	Generalized Learning Vector Quantization
GMLVQ	Generalized Matrix Learning Vector Quantization
LVQ	Learning Vector Quantization
TN	True Negatives
TP	True Positives
WTA	Winner-Takes-All

Mathematical Symbols

$\hat{c} \in \mathscr{C}$	Predicted data class
$\mathbf{w}_k \in W$	Prototype
$\mathbf{x} \in X$	Data point
$\mathscr{C}$	Set of data classes
$c \in \mathscr{C}$	Data class
$d(\mathbf{x}, \mathbf{y})$	Dissimilarity between the vectors $\mathbf{x}$ and $\mathbf{y}$
W	Set of prototypes
X	Set of data points

V. Acknowledgement

I would like to express my gratitude to my thesis advisors for their patience, guidance and help during the process of this thesis. Special thanks to Thomas for providing insight into research itself by declaring: “We do not have any problems, we only have challenges.” A big thanks to Marika for always taking the time to answer any of my questions.

Further, I would like to thank the members of the MaLeKITA Research Group for their valuable input, helpful discussions and the never ending coffee supply.

1 Introduction

The topic of this thesis is in the field of structural bioinformatics. Particularly, the task is to distinguish between mirror and native protein structures based on predicted protein structure data samples by means of machine learning. In detail, interpretable machine learning models will be applied on a theoretical question of the bioinformatic subsection of protein structure prediction. The fundamental question in the process is whether or not native and mirror conformations of proteins can be distinguished by means of the so-called Ramachandran Plot (Ramachandran et al., 1963). This chapter on the one hand serves as an eye-opener for why such a distinction is from crucial importance and on the other to inform the reader on the topic itself.

1.1 Motivation

Since proteins constitute some of the most widespread macromolecules in life organisms (Deng et al., 2018), it is only plausible they propose a wide range of possible research areas such as protein design and structure prediction (Floudas et al., 2006). In the latter field a major role is awarded to the prediction from only the amino acid sequence. This kind of structure prediction holds a few difficulties, for it is hard to find a sufficiently accurate force field for the calculation of the potential energy of molecules needed in *in silico* protein folding. And also the computational expenditure can exceed computational capacities (Deng et al., 2018; Margara et al., 2008; Vassura et al., 2008). Furthermore, the computed protein model, especially when predicted using a protein contact map, has an uncertainty whether it displays the native or mirror conformation of the protein (Noel et al., 2012; Kurczynska and Kotulska, 2018). However, in the fields of drug discovery/development (Wang et al., 2016; Zhao and Lu, 2014), chemical (*de novo*) synthesis of proteins (Kent, 2019), synthetic biology (Weinstock et al., 2014) and orthogonal systems/creation of mirror life (Ling et al., 2019) it is crucial to differentiate exactly between these two conformations.

1.2 Biological Background

The following chapters provide the reader of a non-biological background with the necessary basic preliminaries of proteins in a biological and chemical context.

1.2.1 Proteins

Proteins are molecules, which are essential for many things inside and outside of the cells of living organisms. As *structural proteins* they regulate a cell's shape and act as rails for movement of intracellular organelles and molecules by *motor proteins*. Proteins can also function as an assembler of other proteins for specific functions and then are called *scaffold proteins*. Some proteins can also serve a regulatory function (*regulatory proteins*), authorize the flow of molecules and ions through cell membranes (*membrane transport proteins*) and catalyze chemical reactions as *enzymes* (Lodish et al., 2013).
But what exactly are proteins in detail? To answer this question a short insight into the chemical structure of a protein has to be given.

1.2.1.1 Chemical Structure of a Protein

Proteins are made up of a linear chain of amino acids (Floudas et al., 2006), which are connected via a peptide bond between the carboxyl group ($-\mathrm{COOH}$) of one amino acid and the amino group ($-\mathrm{NH_2}$) of another (Dixon, 1984), see the yellow lines in Figure 1.1 for clarification. This chain of amino acids is called a polypeptide chain or the protein's backbone and is defined by the peptide bonds.
Due to hydrogen-bonding properties in the backbone (Schaeffer and Daggett, 2011), this polypeptide chain or primary structure of the protein then is able to form patterns of local bonding (Floudas et al., 2006) or in other words the secondary structure. There are two secondary structures that are most common: α-helices and β-sheets, which are connected with another secondary structure called loops (Floudas et al., 2006). When the protein folding is done, a particular three-dimensional tertiary structure with a certain biological function is the result (Margara et al., 2008; Li et al., 2018). This specific function is determined by the distinctive chemical properties of the side chains of the amino acids and the protein's structure and folding (Lodish et al., 2013; Kent, 2019).
A closer look at the chemical composition of the protein's backbone gives further insight. In Figure 1.1 a schematic depiction of a few amino acids and the aforementioned side chains (here as a simple R for residue), bound to the α carbon atom (first carbon atom attached to a functional group), are shown. These residues can be manifold, ranging from a single hydrogen atom to more complex groups like hydroxyl or amino groups (Lodish et al., 2013).

The so-called planar peptide bonds of different atoms in a protein's backbone further limit the final folded protein. These planes correspond to the three dihedral or torsion angles of the protein backbone: Φ, Ψ and Ω (Lodish et al., 2013; Hollingsworth and Karplus, 2010). The latter is quite restricted to either 0° or 180° (Neal et al., 2006) and, because of this restriction, will not be further investigated in this thesis. The other two planes can only rotate up to a certain degree, for some combinations of them would produce a steric hindrance in the folded protein (Lodish et al., 2013). The folding of a secondary structure like a helix is also dependent on this planes and their angles (Schaeffer and Daggett, 2011). Figure 1.1 depicts this subject by a schematic section of a protein backbone with the planes matching the angles Φ (blue) and Ψ (red).

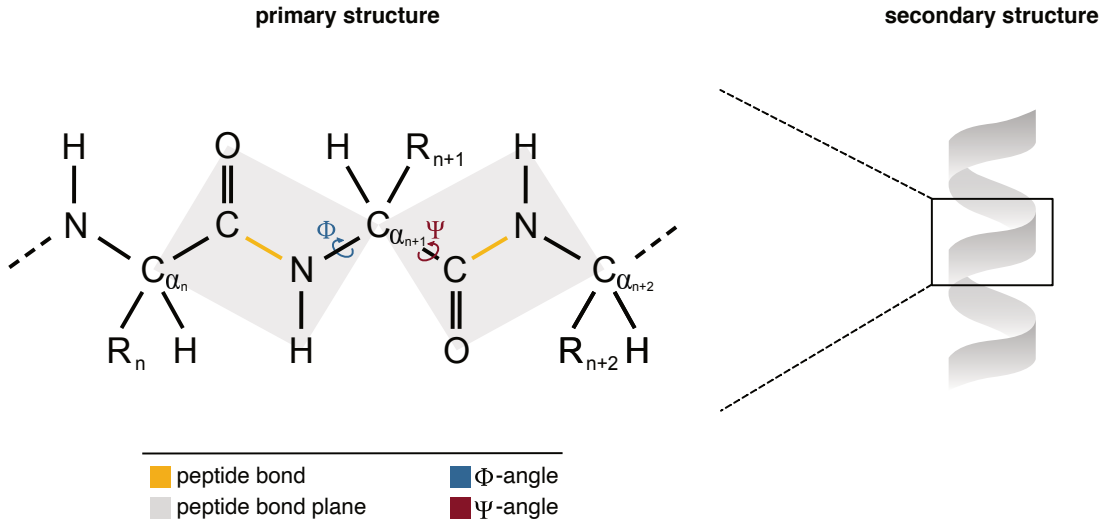

Figure 1.1: Schematic depiction of a protein backbone section with peptide bond planes, their respective dihedral angles and the resulting secondary structure

1.2.1.2 Protein Domains, Families and Classes

If information on a protein regarding its structural and evolutionary relationships is needed and the 3D structure of the protein is known, using the Structural Classification of Proteins (SCOP) database (Murzin et al., 1995) is a good start. Since 2014 only an extended version of SCOP is accessible – SCOPe (Fox et al., 2014) and this thesis will solely refer to SCOPe.

Entries on SCOPe are ordered in a hierarchical fashion with the protein *domain*, derived from the experimentally determined protein structure as the base (Fox et al., 2014) (see Figure 1.2). A *domain* is a structural subunit composed of small repeating patterns of secondary structures like α-helices and β-sheets. These patterns are also known as protein motifs (Schaeffer and Daggett, 2011).

The next level in the hierarchy is *species* in which a specific protein sequence and the existing natural as well as their artificially created variations are put into. Above that is the level *protein* that groups together sequences that are alike and basically have the

same function (Fox et al., 2014; Murzin et al., 1995).
A protein *family* collects proteins which are closely related and evidently have the same evolutionary origin. In contrast to that a *superfamily* groups proteins with domains that are more distantly related, but most likely have a common evolutionary ancestor. The respective similarity usually is limited to mutual structural features (Murzin et al., 1995; Hubbard et al., 1998; Fox et al., 2014).
If the majority of a *superfamily* shares the same global structural features it is merged into a *fold*. These folds are then finally grouped into one of twelve different protein *classes* (Fox et al., 2014; Murzin et al., 1995).
Only seven of these protein classes of $\mathbf{A} - \mathbf{G}$ will be in the scope of this work, based on the data available for this consideration (see Section 3.3).

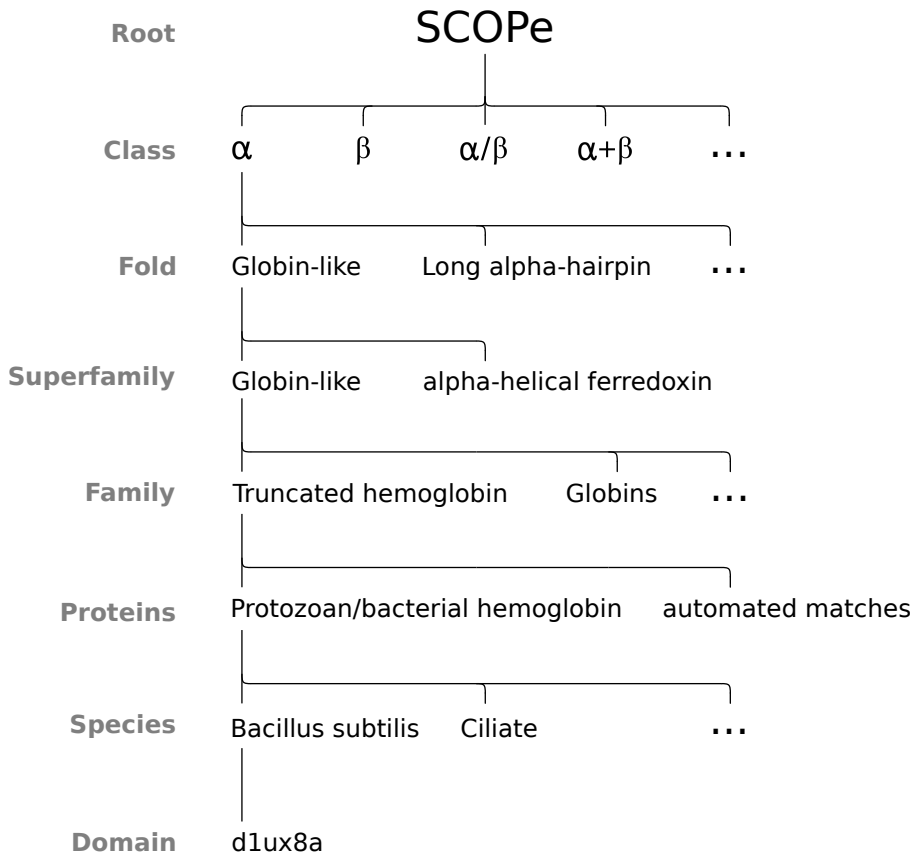

Figure 1.2: Schematic depiction of SCOPe hierarchy, exemplary for the domain of *Bacillus subtilis*

All-alpha proteins (class $\mathbf{A}$) are proteins consisting predominantly of α-helices. One example (Protein Data Bank (PDB)-ID: 1qsa) of this class $\mathbf{A}$ can be seen in Figure 1.3(A). Only loops are to be found in addition to the many α-helices. The class all-beta (class $\mathbf{B}$) however mainly consists of β-strands, see Figure 1.3(B).

Alpha and beta proteins (a/b) (class **C**) show alternations of α-helices and β-strands. The latter are mainly found in the parallel orientation, as can be seen in Figure 1.3(C). This class is not to be confused with alpha and beta (a+b) (class **D**), which contains proteins that show segregated sections of α-helices and β-strands, whereby the strands are mainly present in the antiparallel orientation as can be seen in Figure 1.3(D).

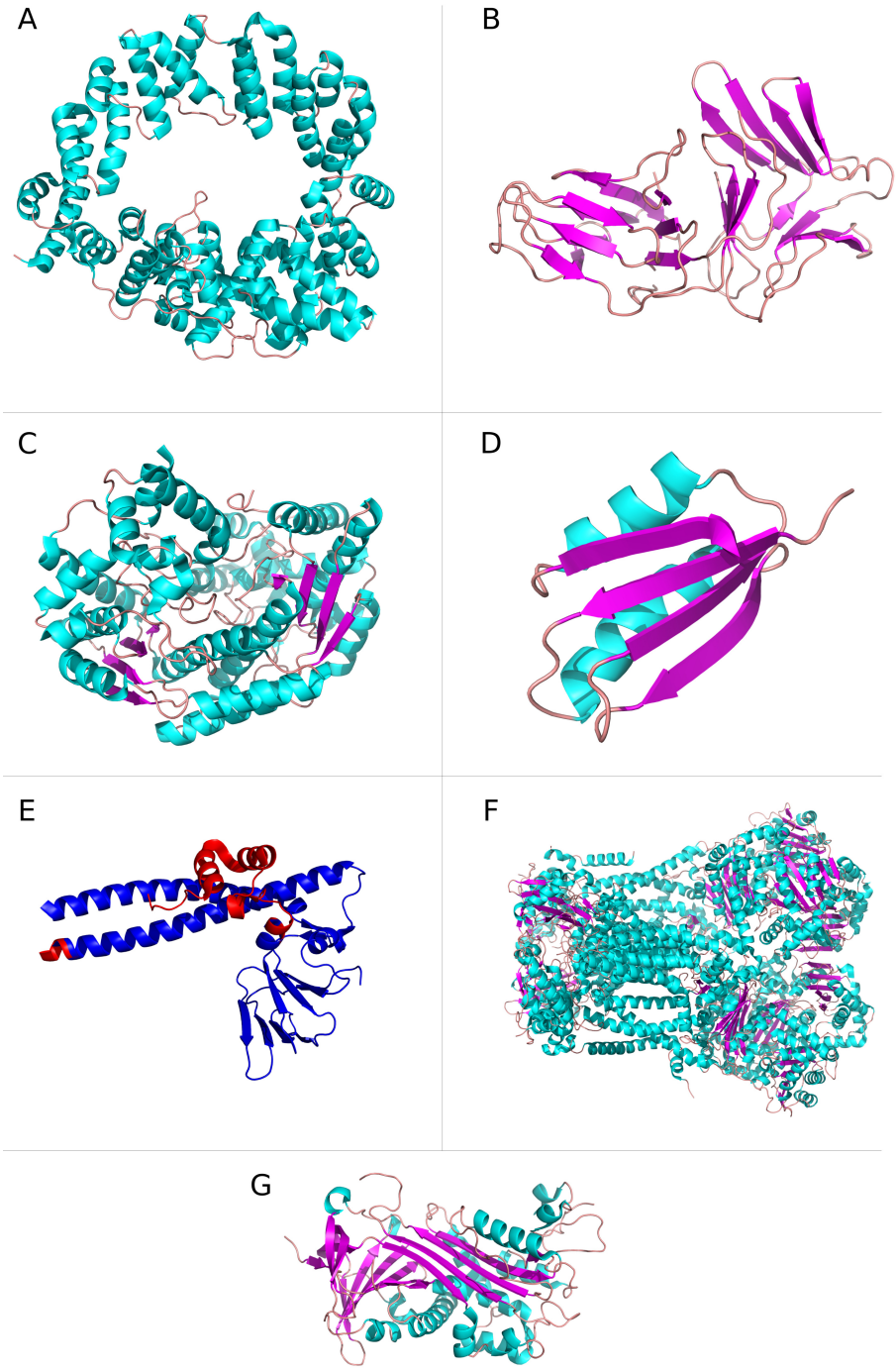

Figure 1.3: Examples for protein classes **A** – **G** with their respective PDB-IDs: A - 1qsa, B - 1b2p, C - 1kqp, D - 1cc8, E - 2aze, F - 1ppj, G - 1oc0

The so-called multi-domain proteins are grouped into class **E** and consist of at least two different domains, see Figure 1.3(E). Proteins and peptides located on surfaces or in membranes are grouped into class **F**. One transmembrane protein is shown in Figure 1.3(F) as an example of this class. Protein class **G** groups small proteins, that have little secondary structure, see Figure 1.3(G) (Schaeffer and Daggett, 2011; Naveenkumar et al., 2019; Murzin et al., 1995; Hubbard et al., 1998).
In Figure 1.3 the light blue spirals show α-helices, the magenta arrows depict β-strands, whereas the salmon colored areas are loops in the protein. To visualize the two domains of class **E** not the prior coloring but blue and red were used, because domains represent patterns of secondary structures rather than single helices or sheets.

1.2.2 Native and Mirror Structure of a Protein

A further concept necessary for understanding this thesis is the difference between the native and mirror conformation of proteins. So far we have already heard about the protein structure and the base for that: the amino acids.
Amino acids as chiral molecules (except for the achiral amino acid Glycin) are able to occur in one of two forms: in a L-configuration (L for levorotatory or left-handed) or D-configuration (D for dextrorotatory or right-handed). The amino acids in a natural or native protein are all of the L-configuration. In contrast to that stand the mirror images or mirror proteins, for they are made of D-amino acids and, therefore, form a mirror structure of the natural protein (Kent, 2019; Zhao and Lu, 2014). In Figure 1.4 the idea of the two forms being mirror images of each other is depicted and it becomes clear that they are not superimposable (Zhao and Lu, 2014; Yeates and Kent, 2012).

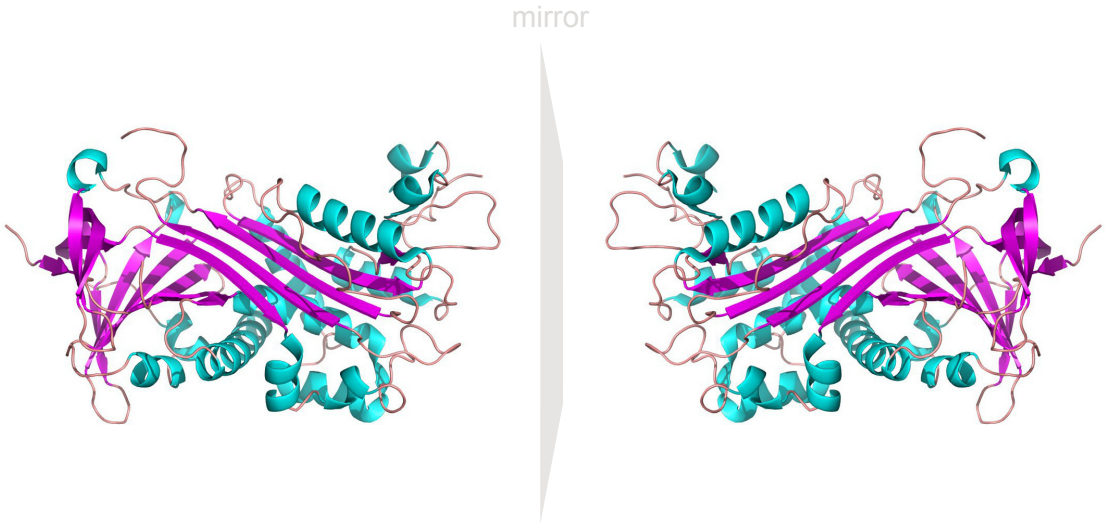

Figure 1.4: Schematic illustration of mirrored forms of a protein (PDB-ID:1oc0)

These mirror or D-amino acids harbor a multitude for applications, especially in synthetic biology and drug development, because the peptides made of D-amino acids can not be recognized or degraded by cellular enzymes (Pentelute et al., 2008; Weinstock et al., 2014). However, there is proof that synthetically produced D-amino acids can be folded into a functioning protein by the help of a native chaperone (another protein aiding in folding of other proteins) (Weinstock et al., 2014).
Even though mirror proteins are resistant to a degradation by native enzymes it is possible for mirror proteins to actually have similar or even the same functions as their natural pendants (Pentelute et al., 2008). In simulations mirror proteins have suggested they are even entropically more stable than native ones and the enantiomers might be competing folds, depending on the surrounding environmental conditions (Noel et al., 2012). Mirror proteins already find use in the so-called mirror image phage display: a technique for identification of new therapeutic target proteins (Kent, 2019).

For full disclosure, the definition of chirality and, therefore, the interpretation of native and mirror proteins is a bit intricate and also converse (Mislow, 2002).
Efimov (2018) first of all states that chirality and handedness are two different things. He declares that chirality indeed is the property of a molecule to have two non superimposable mirror images of itself. In that point, Yeates and Kent (2012), Hoffmann-Ostenhof (1970), Moss (1996), and Zhao and Lu (2014) agree. However, Efimov, and also Pastore et al. (1991), further argue that left-handed and right-handed helices in proteins are not two chiral forms of one structure since the L-amino acids are not converted into D-amino acids, although they are not superimposable. Other researchers claim that mirror image conformations of a protein simply constitute misfoldings (Kachlishvili et al., 2014).

Having said that, this thesis simply and only uses the International Union of Pure and Applied Chemistry's (IUPAC's) definition of handedness and chirality of one structure or molecule being the same (Hoffmann-Ostenhof, 1970; Moss, 1996).

2 Machine Learning Algorithms

This chapter aims to enlighten the reader on machine learning algorithms with the focus on classification, one of which then will be used for solving the task of this thesis: the distinction of native and mirror protein models by using the properties of them.

The automation of intellectual tasks, normally done by humans, is the main core of artificial intelligence (AI). Machine learning constitutes one of the approaches to achieve this goal, by concentrating on AI's learning facet. This is achieved with the development of algorithms that are able to extract knowledge from data. Data sets usually are divided into training, testing and validation (data) sets. Typically four different learning methods can be found in machine learning: supervised, unsupervised, reinforcement learning and semi-supervised (Choi et al., 2020).
In *supervised learning* the training set contains pairings of input as well as desired output (label) information. Mapping features for an output prediction is the key component in this learning method and is accomplished via patterns of the training set (Simeone, 2018). Included in supervised learning are the learning tasks of classification and regression, which predict the category of a certain example and numeric data, respectively (Choi et al., 2020).
Contrary to supervised learning, there are no label information in *unsupervised learning* (Choi et al., 2020). Usually unsupervised learning intends to find patterns and structures of the data generation mechanism (Simeone, 2018).
Reinforcement learning constitutes an intersection of the two aforementioned learning methods. As opposed to unsupervised learning, reinforcement learning does have a kind of supervision, which however is not the label information, as in supervised learning (Simeone, 2018; Choi et al., 2020). Reinforcement learning receives a delayed feedback.
Data analysis tasks that include data with and without label information are calling for a *semi-supervised learning* method (Kaden, 2016).

The task of this thesis is interpreted as a classification problem, for there is label information. In this work the labels are *native* and *mirror* for the given protein models (data samples). Generally speaking the objective of classification is to create a classification model (or classifier) which predicts the labels for given data as best as possible in the context of training data. There are many classifiers available for tackling a classification problem, e.g. Support Vector Machine, k-Nearest-Neighbor and Multi-Layer Perceptron (Kubat, 2017), but this thesis focuses on Learning Vector Quantization which constitutes an interpretable machine learning model (Biehl et al., 2017; Villmann, Saralajew, et al., 2018).

2.1 Variants of Learning Vector Quantization

This section introduces three variants of the prototype-based classification method Learning Vector Quantization (LVQ). In the case of this three variants the learning process differs from variant to variant, whereas the classification procedure stays the same. Generally, LVQ belongs to the class of prototype-based classifiers, i.e. class-dependent prototypes in the data space are taken as reference models for local data. The classification procedure follows the paradigm of nearest prototype classification. According to this paradigm LVQ is easy to interpret (Hofmann et al., 2014).

A training data set $T = \left\{ \left(\mathbf{x}_j, c(\mathbf{x}_j)\right) \in X \times \mathscr{C}, j = 1, \ldots, N \right\}$, where $X \subset \mathbb{R}^n$ is the set of training vectors with class labels $c\left(\mathbf{x}_j\right) \in \mathscr{C}$ and $\mathscr{C}$ is the set of classes. Moreover, a set of prototypes $W = \{\mathbf{w}_k \in \mathbb{R}^n\}$ with the class labels $c(\mathbf{w}_k) \in \mathscr{C}$ is necessary, such that at least one prototype per class is available (Geweniger, 2012).

Following the nearest prototype principle, a data point $\mathbf{x} \in X$ with unknown class label is classified according to the Winner-Takes-All (WTA) rule

$$s(\mathbf{x}) = \arg\min_k \left(d(\mathbf{x}, \mathbf{w}_k)\right) \tag{2.1}$$

with the assigned class

$$\hat{c}(\mathbf{x}) = c(\mathbf{w}_{s(\mathbf{x})}) \tag{2.2}$$

constituting the classification procedure. Here $\mathbf{w}_{s(\mathbf{x})}$ is denoted as winner prototype and $d(\mathbf{x}, \mathbf{w}_k)$ is a dissimilarity measure in the data space (Nebel et al., 2017), often the squared Euclidean distance.

2.1.1 Basic Learning Vector Quantization according to Kohonen

LVQ was introduced by Kohonen (1986) as a protoype-based heuristic classification model. The overall goal of LVQ is to distribute the prototypes in the data space such that a minimization of wrongly classified data points is achieved for the available training data. In the respective learning scheme the first step is a random initialization of labeled prototypes. During the iterations of the algorithm a data point $\mathbf{x}$ is randomly chosen. According to the WTA-rule Equation (2.1) the winning prototype for this data point is determined, where $d(\mathbf{x}, \mathbf{w}_k)$ is the squared Euclidean distance (Geweniger, 2012). The following step of the algorithm updates the prototypes by means of an Attraction Repulsion Scheme (ARS), whereby the winning prototype is either moved toward or away from the data point, depending on the class agreement between the training data label and the winning prototype label.

$$\Delta \mathbf{w}_{s(\mathbf{x})} = \varepsilon \cdot \tau\big(\mathbf{x}, \mathbf{w}_{s(\mathbf{x})}\big) \cdot (\mathbf{x} - \mathbf{w}_{s(\mathbf{x})}) \tag{2.3}$$

with $0 < \varepsilon \ll 1$ as the learning rate. The ARS is realized via the term $\tau(\mathbf{x}, \mathbf{w}_{s(\mathbf{x})})$ according to (Kohonen, 1997)

$$\tau(\mathbf{x}, \mathbf{w}_{s(\mathbf{x})}) = \begin{cases} 1 & if\ c(\mathbf{x}) = c(\mathbf{w}_{s(\mathbf{x})}) \\ -1 & if\ c(\mathbf{x}) \neq c(\mathbf{w}_{s(\mathbf{x})}) \end{cases} . \tag{2.4}$$

The update of the winning prototype is obtained by

$$\mathbf{w}_{s(\mathbf{x})} \rightarrow \mathbf{w}_{s(\mathbf{x})} - \Delta \mathbf{w}_{s(\mathbf{x})} . \tag{2.5}$$

It should be emphasized that the ARS is the basic common principle of all LVQ variants.

2.1.2 Generalized Learning Vector Quantization according to Sato & Yamada

Unfortunately, the original LVQ is a pure heuristic with the motivation to minimize the overall classification error for the training data. To overcome this limitation, Sato and Yamada (1996) introduced a new variant of the LVQ, the Generalized Learning Vector Quantization (GLVQ), optimizing an approximation of this error. This approximation is a differentiable cost function, based on local errors such that Stochastic Gradient Descent Learning (SGDL) (Graf and Lushgy, 2000) can be applied for model adaptation.

In particular, Sato and Yamada (1996) considered the cost function

$$E_{GLVQ}(X, W) = \sum_{j=1}^{N} E(\mathbf{x}_j, W) \tag{2.6}$$

with local errors

$$E(\mathbf{x}, W) = f\big(\mu(\mathbf{x}, W)\big) . \tag{2.7}$$

Here

$$\mu(\mathbf{x}, W) = \frac{d(\mathbf{x}, \mathbf{w}^+) - d(\mathbf{x}, \mathbf{w}^-)}{d(\mathbf{x}, \mathbf{w}^+) + d(\mathbf{x}, \mathbf{w}^-)} \tag{2.8}$$

is the so-called classifier function. In this function $\mathbf{w}^+ = \underset{k}{\operatorname{argmin}}\big(d(\mathbf{x}, \mathbf{w}_k) \mid c(\mathbf{x}) = c(\mathbf{w}_k)\big)$ is the best matching prototype with the same class label as the data point and $\mathbf{w}^- = \underset{k}{\operatorname{argmin}}\big(d(\mathbf{x}, \mathbf{w}_k) \mid c(\mathbf{x}) \neq c(\mathbf{w}_k)\big)$ is the best matching prototype with another class label. Hence, the classifier function gives values in the interval $[-1, 1]$. It becomes

negative when the data point is classified correctly. The squashing function f has to be a monotonically increasing and differentiable function, like the sigmoid function or the identity function. The distance measure itself has to be differentiable to ensure SGDL. The squared Euclidean distance, for example, is such a distance measure and, therefore, can be used for GLVQ (Geweniger, 2012).

Following the SGDL approach the ARS is realized by the update rule

$$\Delta \mathbf{w}^{\pm} \propto \varepsilon \cdot \frac{\partial E(\mathbf{x}_j, W)}{\partial \mathbf{w}^{\pm}} \tag{2.9}$$

for the prototypes for a randomly chosen training data point $\mathbf{x}_j$.

2.1.3 Generalized Matrix Learning Vector Quantization – GMLVQ

GMLVQ is a variant of GLVQ using

$$d_{\Lambda}(\mathbf{x}, \mathbf{w}) = (\mathbf{x} - \mathbf{w})^T \Lambda (\mathbf{x} - \mathbf{w}) \tag{2.10}$$

as differentiable dissimilarity measure (Schneider et al., 2009). The matrix Λ of full size $n \times n$ is assumed to be decomposed into

$$\Lambda = \Omega^T \Omega \tag{2.11}$$

ensuring positive (semi-) definiteness. The matrix $\Omega \in \mathbb{R}^{m \times n}$ is interpreted as a linear mapping of both, data and prototypes, before applying the squared Euclidean metric in the mapping space $\mathbb{R}^m$. In GMLVQ, additionally to the prototype adaptation, also the mapping matrix is adjusted by the SGDL. After training, the Λ-matrix can be interpreted as a classification correlation matrix indicating data feature correlations, which support the classification decision. The diagonal entries Λ_{ii} are the relevances r_i delivering the importance of the ith data feature for class separation. All relevance values $r_i = \Lambda_{ii}$ are collected in the relevance profile vector $\mathbf{r} = (v_1, \ldots, v_n)^T$. This relevance detection is also denoted as *relevance learning*. Hence, it contributes to a better model interpretability (Villmann, Bohnsack, et al., 2017).

2.2 Classification Validation

In binary classification problems there are a few common measures for evaluating the classification, for example the classification accuracy (acc). This measure describes the relative number of data points correctly classified by the model. In a given data set $X, \mathbf{x} \in X$ with known class label $c(\mathbf{x})$ and $\hat{c}(\mathbf{x})$ being the predicted class label by the model, the classification accuracy is

$$acc = \frac{1}{|X|} \sum_{\mathbf{x} \in X} \delta_{c(\mathbf{x}),\hat{c}(\mathbf{x})} \tag{2.12}$$

with $|X|$ being the cardinality of X and δ being the Kronecker delta

$$\delta_{i,j} = \begin{cases} 1 & if\ i = j \\ 0 & else \end{cases} . \tag{2.13}$$

For further classification validation measures like sensitivity and specificity (Choi et al., 2020) the so-called confusion matrix (see Table 2.1) comes in handy. In our work we used 0 as the label for positive (which corresponds to mirror protein models) and 1 for the label for negative (coincides to native protein models). Concluding from these definitions

Table 2.1: Confusion matrix of binary classification problem

	real Positive	**real Negative**
predicted Positive	True Positive (TP)	False Positive (FP)
predicted Negative	False Negative (FN)	True Negative (TN)

True Positives/Negatives and False Positives/Negatives are calculated according to

$$\begin{aligned} \text{TP} &= \sum_{\mathbf{x} \in X} \delta_{0,c(\mathbf{x})} \cdot \delta_{0,\hat{c}(\mathbf{x})} \\ \text{TN} &= \sum_{\mathbf{x} \in X} \delta_{1,c(\mathbf{x})} \cdot \delta_{1,\hat{c}(\mathbf{x})} \\ \text{FP} &= \sum_{\mathbf{x} \in X} \delta_{1,c(\mathbf{x})} \cdot \delta_{0,\hat{c}(\mathbf{x})} \\ \text{FN} &= \sum_{\mathbf{x} \in X} \delta_{0,c(\mathbf{x})} \cdot \delta_{1,\hat{c}(\mathbf{x})} \ . \end{aligned} \tag{2.14}$$

By use of these equations

$$\text{sensitivity} = \frac{TP}{(TP+FN)} \tag{2.15}$$

and

$$\text{specificity} = \frac{TN}{(TN+FP)} \tag{2.16}$$

can be calculated. However, both of these classification validation measures are more useful to imbalanced data (Banerjee et al., 2018), which we do not have.

3 Analysis of Mirror and Native Structures

In this chapter we describe the data set in use as well as how it is generated from protein models. Further, we explain how the machine learning tool is applied to distinguish between native and mirror protein models.

3.1 Related Work for Distinction of Mirror/Native Structures

Giving a short overview of related work in the distinction of native and mirror protein structures is the intention of this section.
In the course of the last years there have been quite a few attempts of solving this differentiation problem. Kurczynska, Kania, et al. (2016) have used energy terms from PyRosetta for a distinction between these predicted models. The application of a simple minimization algorithm with a cost function taking chirality terms into account was also done (Lund et al., 1996; Vendruscolo et al., 1997; Havel and Snow, 1991). However introducing chirality terms commonly only helps in native/mirror distinction in proteins of protein class $\mathbf{A}$, for models from protein class $\mathbf{B}$ do not necessarily reveal a dominant handedness for either one of the conformations. Further, adjustment of a torsion angle in a further processing step was undertaken by Aszoódi et al. (1995) to distinguish between native and mirror protein models.
Summarizing, several research teams have tried to differentiate between the two protein conformations, but were semi-successful or had not as good accuracies doing so. Moreover, Kurczynska and Kotulska (2018) proclaimed, that a distinction by using the Ramachandran Plot, regardless of the analyzed protein class, is not possible.
The aim of this work is to investigate this statement in the light of newly available machine learning methods.

3.2 Ramachandran Plot

Before we take a closer look at the workflow and data generation we have to understand the basic concept of a Ramachandran or R-plot. R-plots visualize the distribution of the torsion angles Φ and Ψ of a protein's backbone by depicting them in a toroidal plot (Ramachandran et al., 1963). Although there are three dehidral angles, mainly Φ and Ψ have an influence on the outcome of the protein folding, since Ω is usually either $0\,^\circ$ or $180\,^\circ$ (Neal et al., 2006). That, however, does not mean that any angle is possible for Φ and Ψ, because these angles are sterically restrained (Lodish et al., 2013).

Depending on the angles, a certain secondary structure like α-helix or β-strand, is folded (Schaeffer and Daggett, 2011) and, therefore, R-plots give information about the secondary structure elements of a protein (Hollingsworth and Karplus, 2010). Further, R-plots show areas of *favored*, *allowed* and *slightly unfavored* regions of the dihedral angles. In Figure 3.1 a generic R-plot with the mentioned regions for right-handed α-helices (α) and left-handed α-helices (L_α) and β-strands (β) and an overlayed $6 \cdot 6$ grid is depicted. Single cells within this R-plot are referred to in coordinate notation.

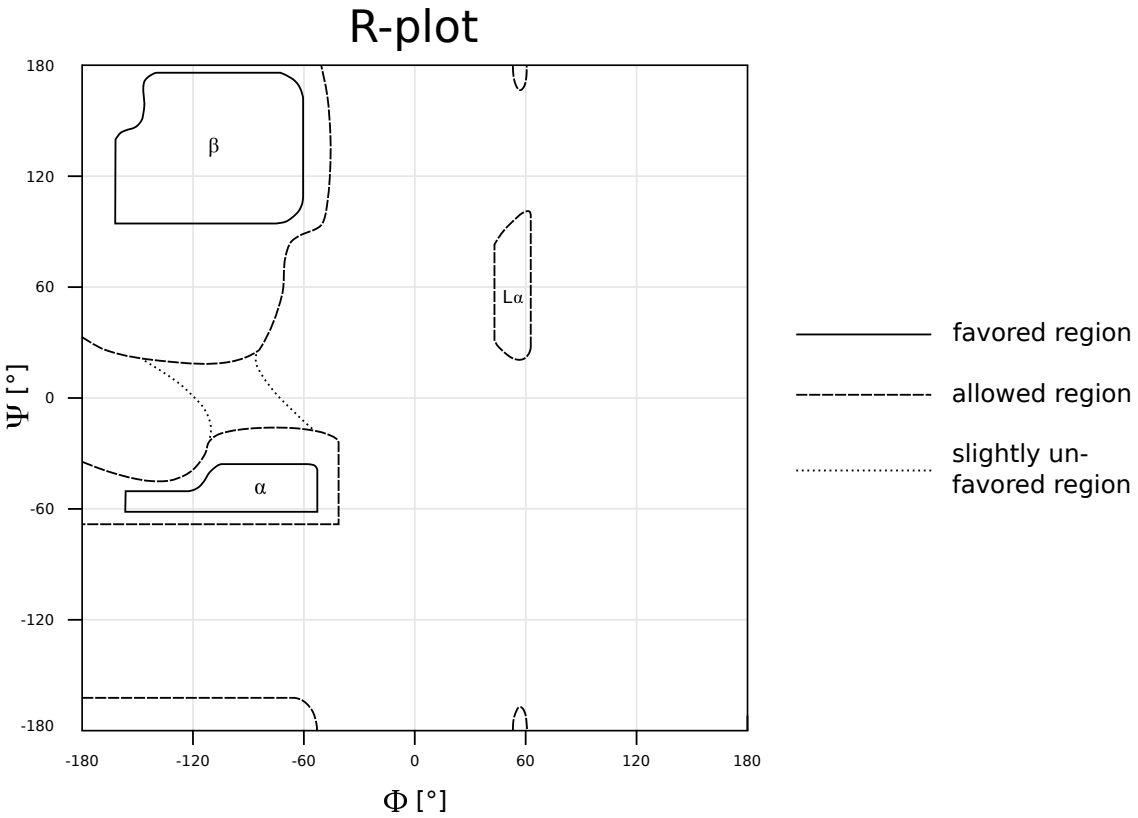

Figure 3.1: Schematic R-plot with the regions for right-handed α-helices (α), left-handed α-helices (L_α) and β-strands (β)

3.3 Mirror/Native Data Set

In this subsection the generation of the (dihedral angle) data set is described. The data set was derived from a protein data set containing the atomic coordinates of each protein (atomic coordinate data set).

3.3.1 Atom Coordinate Data Set

Kurczynska and Kotulska (2018) generated a data set containing atom coordinates for protein domains with the aim to differentiate between mirror and native models. Under the link: http://comprec-lin.iiar.pwr.edu.pl/mirrorModels/ the data is available.
At first 1,961 domains from the SCOP online server were taken. Those domains are representatives of the SCOP superfamilies. A pre-processing of these domains was done in the way that all domains were eliminated, which had one of the following properties:

- special amino acids (like selenocystein)
- heavy atoms in the middle of the chain
- missing residues.

Whenever there was a special amino acid at either the beginning or the end of the chain, said domain was reduced. A reduction also was done, when heavy atoms were not found at the beginning or the end of the chain.

Finally, the data set consisted of 1,305 protein domains from seven different protein classes $\mathbf{A} - \mathbf{G}$ (Kurczynska and Kotulska, 2018).
For creation of protein models with two orientations (native and mirror) the method for structure modeling from contact maps was applied. The contact map for each domain was generated by PConPy and then used as input for C2Sv2.0, which reconstructed 100 structural models for each domain (approximately 50 native and 50 mirror models) (Kurczynska and Kotulska, 2018).
However, the data availability only covers atom coordinates and not the dihedral angles.

3.3.2 Dihedral Angle Data Set

Of all the created protein models a calculation for all dihedral angles (Φ, Ψ and Ω) was done. All in all, the data set of this thesis contained the 100 models from each of the 1,305 protein domains, their respective class label and all dihedral angles of each model. See Table 3.1 for the size of the examined protein classes, which in turn do represent seven distinctive data sets. Moreover, a summary of all models was done into

the class $\mathbf{ALL}$.
However the data set generation is not scope of this work. For detailed information on that, see Kurczynska and Kotulska (2018).

Table 3.1: Numbers of samples for each of the described protein classes $\mathbf{A} - \mathbf{G}$

protein class	protein models
A	343
B	233
C	149
D	368
E	21
F	78
G	113

3.4 Data Analysis Workflow

In this section the workflow of this thesis' data analysis is introduced.
Each of the data sets described in Subsection 3.3.2 defines an unique learning task to differentiate between native and mirror protein models. The differentiation is done by extraction of a relative R-plot histogram vector $\mathbf{x} \in \mathbb{R}^n$ with $n = N \cdot N$ with $N = 6$ for each protein model. These data vectors served as training data for GMLVQ according to the considered tasks.

A separate GMLVQ model with three prototypes per class (native/mirror) was trained for each learning task $\mathbf{A} - \mathbf{G}$ and $\mathbf{ALL}$. In addition to the prototypes also the mapping matrix Ω was adapted with the mapping dimension $m = n = 36$. The obtained averaged test performances of 50 independent runs constitute the classification results. Every run was done as a five-fold cross-validation procedure.

A *summarized R-plot* was generated to enable a visual inspection and evaluation. This R-plot is a collection of all dihedral angle pairs (Φ, Ψ) for all samples of a certain learning task and one was done for each class (native/mirror) separately. In other words, these R-plots can be considered as estimated dihedral angles densities of native and mirror protein models in the (Φ, Ψ)-plane. And by this, a visual inspection of the dihedral angle distributions can be done, see Figures 4.1, 4.2 and 4.3.

4 Results and Discussion

This chapter is essentially based on the subsection *Results and Discussion* of the paper *Detection of native and mirror protein structures based on Ramachandran plot analysis by interpretable machine learning models*, submitted to PLOS ONE in parallel.

By taking advantage of the inherent interpretable nature of GMLVQ, particularly by considering the relevances of the data features, we can extract insightful knowledge regarding the classification decision. Especially the connection of the relevant features with the favored and allowed regions in the R-plots for each protein class offers a biological interpretation. Table 4.1 summarizes the areas of the R-Plot of each protein class which are considered most relevant by the model.

Table 4.1: Most relevant cells (by visual decision) for differentiation of mirror and native models obtained by the classification model

protein class	cells for α-helices	cells for β-strands
ALL	(1,4), (2,4), (4,1)	(1,2), (2,1)
class **A**	(1,3),(1,4), (2,3), (2,4), (4,1), (4,2)	—
class **B**	—	(1,2), (2,1)
class **C**	(1,4), (2,4), (4,1), (4,2)	(2,1)
class **D**	(1,4), (2,4), (4,1), (4,2)	(1,2), (2,1)
class **E**	(1,4), (2,4), (4,1), (4,2)	(2,1)
class **F**	(1,4), (2,4), (4,1), (4,2)	—
class **G**	(2,4)	(1,2), (2,1)

As previous publications suggest (Kurczynska and Kotulska, 2018; Vendruscolo et al., 1997), native and mirror conformations of proteins rich in helices (belonging to class $\mathbf{A}$), can be distinguished based on chirality derived information, or in more detail: right-handed α-helices are favoured in native conformations (Novotny and Kleywegt, 2005). We can confirm this finding, since the features corresponding to the left- and right-handed α-helices predominantly contribute to the class discrimination of mirror and native (see Figure 4.1(A) and Table 4.1). The accuracy for this class is 86.57% (see Table 4.2 for all accuracies). Interestingly, our model achieves an accuracy of 92.56% for class $\mathbf{B}$, which obviously collides with the statement of those native and mirror models being indistinguishable (Kurczynska and Kotulska, 2018; Vendruscolo et al., 1997) and furthermore even exceeds the accuracy of class $\mathbf{A}$. The relevant features for class discrimination are those corresponding to β-strands in the R-plot, see Figure 4.1(B). In detail the important underlying secondary structures in this case might be the right-handed triple helices (collagen) and parallel β-strands (Bella, 2016; Hollingsworth and Karplus, 2010). However, the confirmation of the actual underlying secondary structure as well as their relation to chirality are still pending.

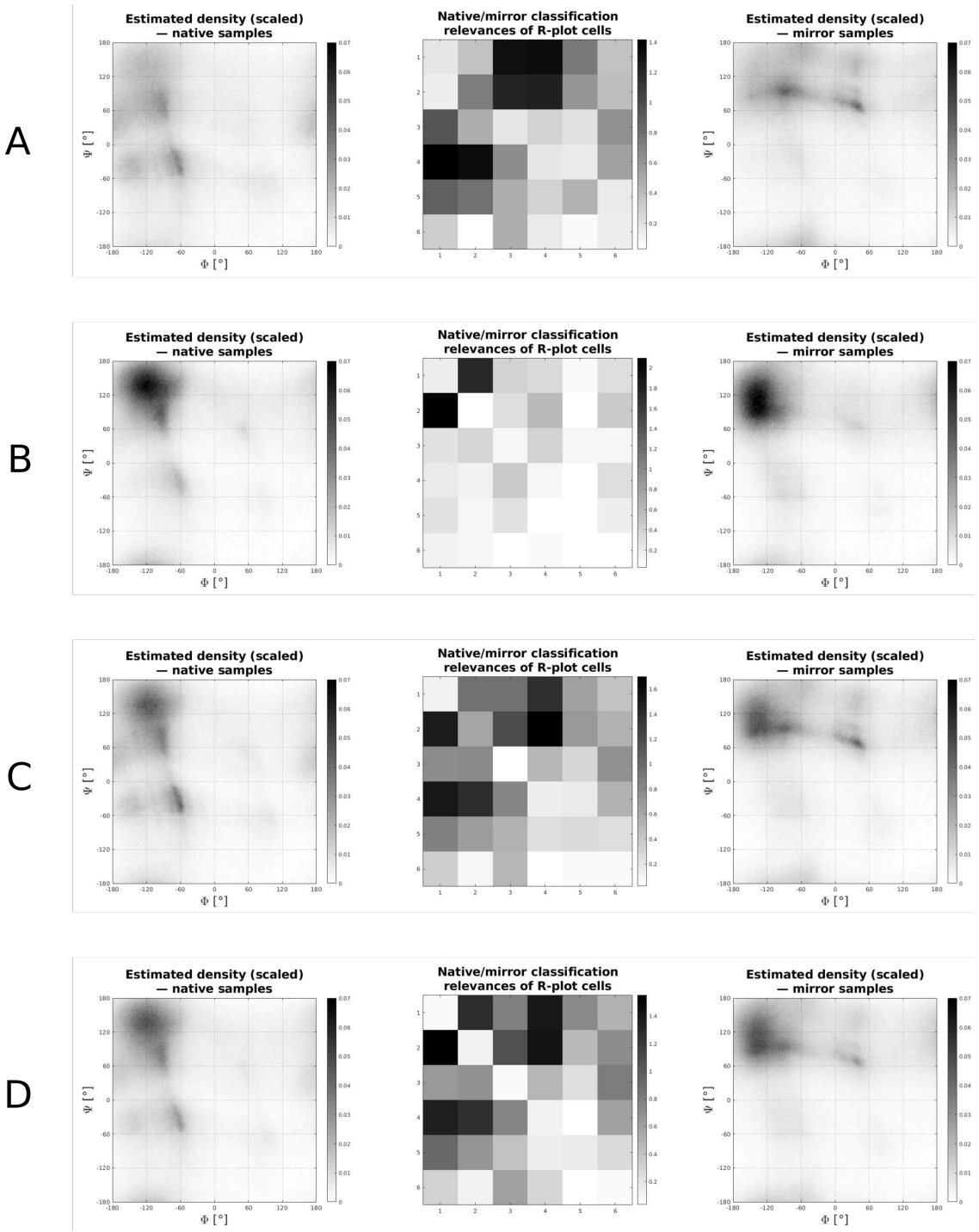

Figure 4.1: The summarized R-plots for native (left) and mirror (right) samples. The plots are estimators for the dihedral angles densities in the (Φ, Ψ)-plane for samples of classes $\mathbf{A} - \mathbf{D}$. The obtained relevance profile vector $\mathbf{r} = (v_1, \ldots, v_{36})^T$ from relevance learning in GMLVQ is depicted in the middle, arranged accordingly to the cells of the R-plot.

As protein class **C** and **D**, see Figure 4.1(C) and 4.1(D) structurally show a combination of the aforementioned two classes *all-alpha* and *all-beta*, the relevant features for class discrimination also do. Class **C** shows the best of all investigated accuracies and this result concurs with the findings in (Kurczynska and Kotulska, 2018). Among the protein classes which were not categorized due to their secondary structure, the multi-domain class **E** by far shows the best accuracy with 91.8%, whereas classes **F** and **G** do not exceed 80%. Even though class **E** has got such a high accuracy it has to be treated with care, since there has not been enough data for this protein class. The relatively poor accuracy for class **F** is most likely due to the fact that membrane proteins propose some difficulties in structure elucidation (Zhou et al., 2004; Postic et al., 2015; Martin and Sawyer, 2019) and this results in low resolutions (Moraes et al., 2014). A poor resolution in turn may lead to inaccurate atom coordinates and that means the calculations of the dihedral angles cannot be correct either and therefore complicate the classification. As for class **G** the obtained low accuracy is probably due to the fact that small proteins do not have that many amino acids, less than 100 (Su et al., 2013; Miravet-Verde et al., 2019), and therefore show less α-helices or β-strands than other proteins.

In order to assess the models suitability for a more general problem we considered all protein classes for training as well as for testing and achieved an overall accuracy of 88.09%. Pursuing this approach, which is more general and more considerable for application, we investigated the behaviour of our model by training with all protein classes but testing with only one protein class at a time. The achieved accuracies are in good agreement with those of the single classes as Table 4.2 shows.

Table 4.2: Obtained averaged (test) accuracies (A), obtained accuracies using the general model for specific classes together with their respective standard deviations for the protein classes (results refer to 50 runs of cross-validated GMLVQ). For comparison, additionally the best classification results A_K from Kurczynska and Kotulska (2018) are given (for all classes as average of the others)

protein class/results	A	σ_A	$A_{overall}$	$\sigma_{A_{overall}}$	A_K
ALL	88.09%	0.03	88.09%	0.03	69%
class **A**	86.57%	1.48	85.34%	0.03	71%
class **B**	92.56%	1.23	90.78%	0.03	76%
class **C**	98.50%	0.70	98.46%	0.01	78%
class **D**	93.04%	0.39	92.76%	0.01	70%
class **E**	91.80%	8.92	92.42%	0.03	70%
class **F**	65.84%	3.03	67.38%	0.03	60%
class **G**	75.45%	2.48	75.50%	0.01	61%

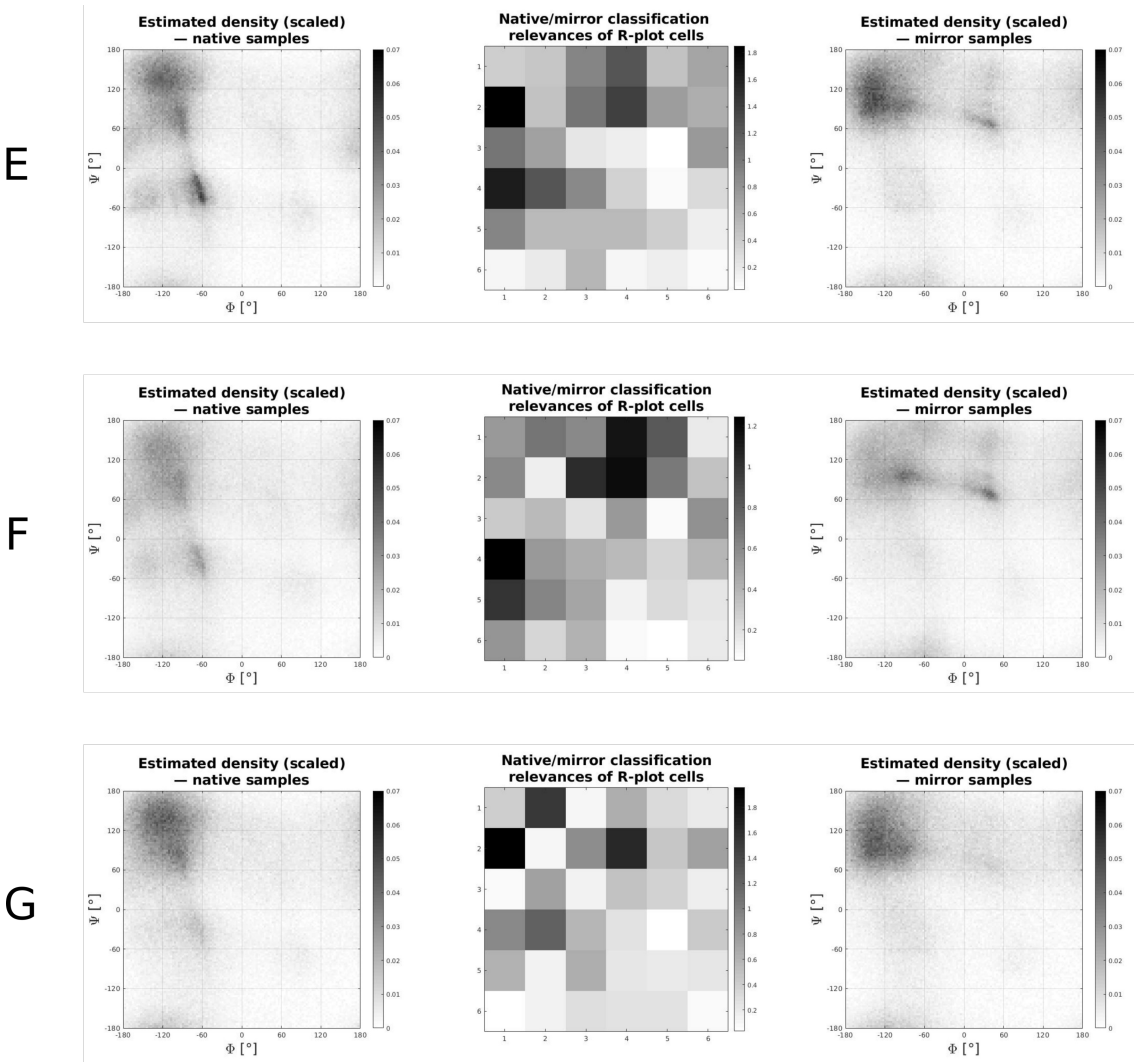

Figure 4.2: The summarized R-plots for native (left) and mirror (right) samples. The plots are estimators for the dihedral angles densities in the (Φ, Ψ)-plane for samples of classes $\mathbf{E} - \mathbf{G}$. The obtained relevance profile vector $\mathbf{r} = (v_1, \ldots, v_{36})^T$ from relevance learning in GMLVQ is depicted in the middle, arranged accordingly to the cells of the R-plot.

All

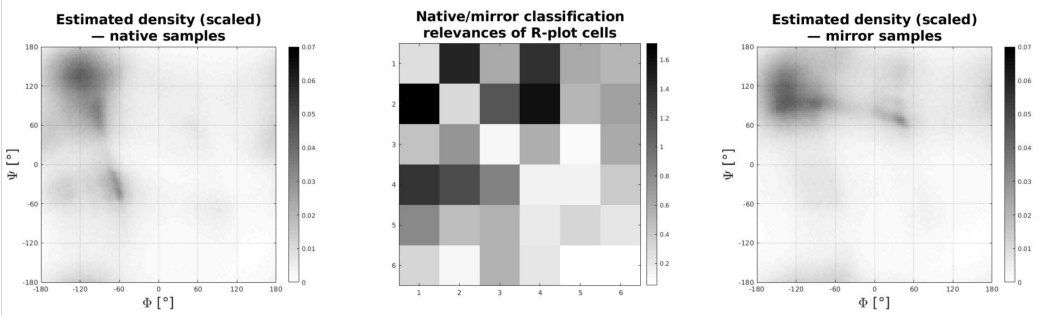

Figure 4.3: The summarized R-plots for native (left) and mirror (right) samples. The plots are estimators for the dihedral angles densities in the (Φ, Ψ)-plane for samples of **ALL** classes. The obtained relevance profile vector $\mathbf{r} = (v_1, \ldots, v_{36})^T$ from relevance learning in GMLVQ is depicted in the middle, arranged accordingly to the cells of the R-plot.

Further, we have also looked at the third dihedral angle Ω and included it into our calculations (results not shown here). However, the accuracies did not change in a drastic matter. This is most likely due to the fact, that Ω takes angle values of either 0° or 180°.

Now we can extract knowledge from the interpretable machine learning method GMLVQ, in particular the obtained relevance profiles (as the diagonals of the Λ-matrix) provide information regarding the secondary structures. As we can see in Figures 4.1 – 4.3 these relevance profiles are in nice agreement with the structural knowledge of proteins as depicted in the generic R-plot (Figure 3.1). The high relevance values of the cells correspond to the biologically predicted allowed regions of secondary structures like α-helices or β-strands. This confirms that the used machine learning method GMLVQ was able to detect the relevant biological structures adequately and make use of that for mirror/native differentiation.

Also, we have calculated the classification validation measures sensitivity and specificity (see Table A.1 in Appendix). But since they do not constitute striking results and rather support our good accuracies, they have not been further discussed in this chapter.

5 Conclusion and Outlook

In this section the previous chapters will be concluded and it will be based on the subsection *Conclusion* of the paper *Detection of native and mirror protein structures based on Ramachandran plot analysis by interpretable machine learning models*, submitted to PLOS ONE in parallel.

In the present contribution we offer a valid approach for distinguishing mirror and native conformations of proteins based on structure information. The approach is based on the evaluation of the respective R-plots by means of an interpretable machine learning model. This model, the Generalized Learning Matrix Vector Quantizer, is known to be robust and highly interpretable according to the underlying reference principle. Moreover, according to the integrated relevance learning metric adaptation, the approach provides beside the classification ability additional knowledge regarding the classification decision. In the context of R-plot analysis this information consists in a weighting of importance of R-plot regions regarding best mirror-native-separability.
Kurczynska and Kotulska (2018) state that structural features are no discriminatory property of native and mirror models. Although information regarding chirality can be used to differentiate models rich in α-helices, according to Vendruscolo et al. (1997) this does not hold for all-β structures. However, we were able to show that a discrimination of native and mirror models using structural features is indeed possible.
The GMLVQ classifier achieves high separation accuracies for all protein classes except class $\mathbf{F}$ and $\mathbf{G}$. At least for the latter one, acceptable results are obtained. In fact, the resulted accuracies for protein classes $\mathbf{F}$ and $\mathbf{G}$ show that a distinction of mirror and native structures by means of R-plots is possible with high specificity and sensitivity. The interpretable model offers additional insights: In particular, the relevance profiles, weighting the regions like α-helices and β-strands of R-plots for mirror-native-discrimination, differ for the considered protein classes. The obtained relevance profiles are in good agreement with respective biological knowledge about protein structure chirality, at least for the considered data set.
Thus, the interpretable GMLVQ method is able to extract biological structure information, which contributes to a good separation of the two cases native and mirror. Of course, other machine learning methods for classification like deep Multi-Layer Perceptrons (Goodfellow et al., 2016) or Support Vector Machines (Schölkopf et al., 2002) probably would achieve similar performance. However, as it is known for Multi-Layer Perceptrons it is not generally obvious whether the distinguishing features detected by them are in compliance with the biological knowledge. Regarding attempts of explanations by heat maps frequently fail because of the problem of vanishing gradients (Hochreiter et al., 2001).

Thus, the presented approach offers a successful alternative to the statistical approach based on energy levels as proposed in Kurczynska and Kotulska (2018) and emphasizes the importance of R-plots for structural analysis of proteins as already mentioned in Ayoub and Lee (2019). Along this line, also data processing is easier in the present approach compared to the complex calculations of the energy levels (Alford et al., 2017).

Further investigation should include to improve the classification performance by finer cell resolutions of R-plots. However, this requires more training data for sufficient learning stability. Further, reject option strategies should be included to detect outliers. Also, the numbers of samples regarding the protein classes **A–G** differ heavily, thus the statistical validation of the results might be on different levels. This in turn also influences the overall result of the learning task **ALL**.

Appendix

Table A.1: Overall model (test) sensitivities (Se) and specificities (Sp) together with their respective standard deviations for the protein classes (results refer to 50 runs of cross-validated GMLVQ)

protein class / results	Se	σ_{Se}	Sp	σ_{Sp}
ALL	87.38%	8.79	88.77%	8.95
class **A**	83.31%	0.12	87.48%	0.06
class **B**	91.70%	0.21	89.93%	0.13
class **C**	98.59%	0.01	98.32%	0.03
class **D**	92.01%	0.06	93.43%	0.03
class **E**	88.51%	0.08	96.33%	0.04
class **F**	65.74%	0.14	69.00%	0.08
class **G**	75.66%	0.15	75.35%	0.13

Bibliography

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., and Gray, J. J. (2017) The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* 13(6):pp.3031–3048.

Aszoódi, A., Gradwell, M. J., and Taylor, W. R. (1995) Global fold determination from a small number of distance restraints. *Journal of Molecular Biology* 251(2):pp.308–326.

Ayoub, R. and Lee, Y. (2019) Rupee: A fast and accurate purely geometric protein structure search. *PLoS ONE* 14(3):pp.1–17.

Banerjee, P., Dehnbostel, F. O., and Preissner, R. (2018) Prediction Is a Balancing Act: Importance of Sampling Methods to Balance Sensitivity and Specificity of Predictive Models Based on Imbalanced Chemical Data Sets. *Frontiers in Chemistry* 6:pp.1–11.

Bella, J. (2016) Collagen structure: New tricks from a very old dog. *Biochemical Journal* 473(8):pp.1001–1025.

Biehl, M., Hammer, B., and Villmann, T. (2017) Prototype-based Models for the Supervised Learning of Classification Schemes. *Proceedings of the International Astronomical Union* (325):pp.129–138.

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., and Peter Campbell, J. (2020) Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science and Technology* 9(2):pp.1–12.

Deng, H. Y., Jia, Y., and Zhang, Y. (2018) Protein structure prediction. *International Journal of Modern Physics B* 32:p.17.

Dixon, H. B. F. (1984) Nomenclature and Symbolism for Amino Acids and Peptides: Recommendations 1983. *European Journal of Biochemistry* 138(1):pp.9–37.

Efimov, A. V. (2018) Chirality and Handedness of Protein Structures. *Biochemistry (Moscow)* 83:pp.103–110.

Floudas, C. A., Fung, H. K., McAllister, S. R., Mönnigmann, M., and Rajgaria, R. (2006) Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science* 61(3):pp.966–988.

Fox, N. K., Brenner, S. E., and Chandonia, J.-M. (2014) SCOPe: Structural Classification of Proteins - extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research* 42(D1):pp.304–309.

Geweniger, T. (2012). Fuzzy Variants of Prototype Based Clustering and Classification Algorithms. PhD thesis. Rijksuniversiteit Groningen.

Goodfellow, I., Bengio, Y., and Courville, A. (2016) *Deep Learning*. MIT Press. Cambridge.

Graf, S. and Lushgy, H. (2000) *Foundations of Quantization for Probability Distributions*. Springer. Berlin.

Havel, T. F. and Snow, M. E. (1991) A new method for building protein conformations from sequence alignments with homologues of known structure. *Journal of Molecular Biology* 217(1):pp.1–7.

Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J. (2001). Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies. In: *A Field Guide to Dynamical Recurrent Networks*. New York: Wiley-IEEE Press.

Hoffmann-Ostenhof, O. (1970) Commission on Biochemical Nomenclature. *Journal of Molecular Biology* 52(1):pp.1–17.

Hofmann, D., Schleif, F. M., Paaßen, B., and Hammer, B. (2014) Learning interpretable kernelized prototype-based models. *Neurocomputing* 141:pp.84–96.

Hollingsworth, S. A. and Karplus, P. A. (2010) A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomolecular Concepts* 1(3-4):pp.271–283.

Hubbard, T. J., Ailey, B., Brenner, S. E., Murzin, A. G., and Chothia, C. (1998) SCOP, structural classification of proteins database: Applications to evaluation of the effectiveness of sequence alignment methods and statistics of protein structural data. *Acta Crystallographica Section D: Biological Crystallography* 54(6 I):pp.1147–1154.

Kachlishvili, K., Maisuradze, G. G., Martin, O. A., Liwo, A., Vila, J. A., and Scheraga, H. A. (2014) Accounting for a mirror-image conformation as a subtle effect in protein folding. *Proceedings of the National Academy of Sciences of the United States of America* 111(23):pp.8458–8463.

Kaden, M. (2016). Integration of Auxiliary Data Knowledge in Prototype Based Vector Quantization and Classification Models. PhD thesis. University Leipzig, Faculty of Mathematics and Computer Science.

Kent, S. B. (2019) Novel protein science enabled by total chemical synthesis. *Protein Science* 28(2):pp.313–328.

Kohonen, T. (1997) *Self-Organizing Maps*. Springer-Verlag. Berlin, Heidelberg.

Kohonen, T. (1986) Learning Vector Quantization for Pattern Recognition. *Technical Report* TKK-F-A601(Helsinski Universtiy of Technology).

Kubat, M. (2017) *An Introduction to Machine Learning*. Springer. Cham.

Kurczynska, M., Kania, E., Konopka, B. M., and Kotulska, M. (2016) Applying PyRosetta molecular energies to separate properly oriented protein models from mirror models, obtained from contact maps. *Journal of Molecular Modeling* 22(111):pp.1–10.

Kurczynska, M. and Kotulska, M. (2018) Automated method to differentiate between native and mirror protein models obtained from contact maps. *PLoS ONE* 13(5):pp.1–19.

Li, B., Fooksa, M., Heinze, S., and Meiler, J. (2018) Finding the needle in the haystack: towards solving the protein-folding problem computationally. *Critical Reviews in Biochemistry and Molecular Biology* 53(1):pp.1–28.

Ling, J., Fan, C., Qin, H., Wang, M., Chen, J., Wittung-Stafshede, P., and Zhu, T. (2019) Mirror-Image 5S Ribonucleoprotein Complexes. *Angewandte Chemie International Edition* 59:pp.1–22.

Lodish, H., Berk, A., Kaiser, C. A., Krieger, M., Bretscher, A., Ploegh, H., Amon, A., and Scott, M. P. (2013) *Molecular Cell Biology*. Katherine Ahr Parker. New York.

Lund, O., Hansen, J., Brunak, S., and Bohr, J. (1996) Relationship between protein structure and geometrical constraints. *Protein Science* 5(11):pp.2217–2225.

Margara, L., Vassura, M., Di Lena, P., Medri, F., Fariselli, P., and Casadio, R. (2008) Reconstruction of 3D structures from protein contact maps. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 5(3):pp.357–367.

Martin, J. and Sawyer, A. (2019) Elucidating the structure of membrane proteins. *BioTechniques* 66(4):pp.167–170.

Miravet-Verde, S., Ferrar, T., Espadas-García, G., Mazzolini, R., Gharrab, A., Sabido, E., Serrano, L., and Lluch-Senar, M. (2019) Unraveling the hidden universe of small proteins in bacterial genomes. *Molecular Systems Biology* 15(2):pp.1–17.

Mislow, K. (2002) Stereochemical terminology and its discontents. *Chirality* 14(2-3):pp.126–134.

Moraes, I., Evans, G., Sanchez-Weatherby, J., Newstead, S., and Stewart, P. D. (2014) Membrane protein structure determination - The next generation. *Biochimica et Biophysica Acta - Biomembranes* 1838(1 PARTA):pp.78–87.

Moss, G. P. (1996) Basic terminology of stereochemistry (IUPAC Recommendations 1996). *International Union of Pure and Applied Chemistry* 68(12):pp.2193–2222.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *Journal of Molecular Biology* 247:pp.536–540.

Naveenkumar, N., Kumar, G., Srinivasan, N., Sowdhamini, R., and Vishwanath, S. (2019) Fold combinations in multi-domain proteins. *Bioinformation* 15(5):pp.342–350.

Neal, S., Berjanskii, M., Zhang, H., and Wishart, D. S. (2006) Accurate prediction of protein torsion angles using chemical shifts and sequence homology. *Magnetic Resonance in Chemistry* 44(7 SPEC. ISS.):pp.158–167.

Nebel, D., Kaden, M., Villmann, A., and Villmann, T. (2017) Types of (dis-)similarities and adaptive mixtures thereof for improved classification learning. *Neurocomputing* 268:pp.42–54.

Noel, J. K., Schug, A., Verma, A., Wenzel, W., Garcia, A. E., and Onuchic, J. N. (2012) Mirror images as naturally competing conformations in protein folding. *Journal of Physical Chemistry B* 116(23):pp.6880–6888.

Novotny, M. and Kleywegt, G. J. (2005) A survey of left-handed helices in protein structures. *Journal of Molecular Biology* 347(2):pp.231–241.

Pastore, A., Atkinson, R. A., Saudek, V., and Williams, R. J. (1991) Topological mirror images in protein structure computation: An underestimated problem. *Proteins: Structure, Function, and Bioinformatics* 10(1):pp.22–32.

Pentelute, B. L., Gates, Z. P., Dashnau, J. L., Vanderkooi, J. M., and Kent, S. B. (2008) Mirror image forms of snow flea antifreeze protein prepared by total chemical synthesis have identical antifreeze activities. *Journal of the American Chemical Society* 130(30):pp.9702–9707.

Postic, G., Ghouzam, Y., Guiraud, V., and Gelly, J. C. (2015) Membrane positioning for high- and low-resolution protein structures through a binary classification approach. *Protein Engineering, Design and Selection* 29(3):pp.87–91.

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* 7:pp.95–99.

Sato, A. and Yamada, K. (1996) Generalized Learning Vector Quantization. *Advances in neural information processing systems* 8:pp.423–429.

Schaeffer, R. D. and Daggett, V. (2011) Protein folds and protein folding. *Protein Engineering, Design and Selection* 24(1-2):pp.11–19.

Schneider, P., Biehl, M., and Hammer, B. (2009) Adaptive relevance matrices in learning vector quantization. *Neural Computation* 21(12):pp.3532–3561.

Schölkopf, B., Smola, A., Smola, A., and Smola, A. (2002) Support Vector Machines and Kernel Algorithms. *Encyclopedia of Biostatistics, (2005)* pp.5328–5335.

Simeone, O. (2018) A Very Brief Introduction to Machine Learning with Applications to Communication Systems. *IEEE Transactions on Cognitive Communications and Networking* 4(4):pp.648–664.

Su, M., Ling, Y., Yu, J., Wu, J., and Xiao, J. (2013) Small proteins: Untapped area of potential biological importance. *Frontiers in Genetics* 4(DEC):pp.1–9.

Vassura, M., Margara, L., Di lena, P., Medri, F., Fariselli, P., and Casadio, R. (2008) FT-COMAR: Fault tolerant three-dimensional structure reconstruction from protein contact maps. *Bioinformatics* 24(10):pp.1313–1315.

Vendruscolo, M., Kussell, E., and Domany, E. (1997) Recovery of protein structure from contact maps. *Folding and Design* 2(5):pp.295–306.

Villmann, T., Bohnsack, A., and Kaden, M. (2017) Can Learning Vector Quantization be an Alternative to SVM and Deep Learning? *Journal of Artificial Intelligence and Soft Computing Research* 7(1):pp.65–81.

Villmann, T., Saralajew, S., Villmann, A., and Kaden, M. (2018). Learning Vector Quantization Methods for Interpretable Classification Learning and Multilayer Networks. In: *Proceedings of the 10th International Joint Conference on Computational Intelligence (IJCCI), Sevilla*. Lissabon, Portugal: SCITEPRESS - Science and Technology Publications, Lda., pp.15–21.

Wang, Z., Xu, W., Liu, L., and Zhu, T. F. (2016) A synthetic molecular system capable of mirror-image genetic replication and transcription. *Nature Chemistry* 8(7):pp.698–704.

Weinstock, M. T., Jacobsen, M. T., and Kay, M. S. (2014) Synthesis and folding of a mirror-image enzyme reveals ambidextrous chaperone activity. *Proceedings of the National Academy of Sciences of the United States of America* 111(32):pp.11679–11684.

Yeates, T. O. and Kent, S. B. (2012) Racemic Protein Crystallography. *Annual Review of Biophysics* 41:pp.41–61.

Zhao, L. and Lu, W. (2014) Mirror image proteins. *Current opinion in chemical biology* 22:pp.56–61.

Zhou, C., Zheng, Y., and Zhou, Y. (2004) Structure prediction of membrane proteins. *Genomics, proteomics & bioinformatics / Beijing Genomics Institute* 2(1):pp.1–5.

Declaration of authorship

I hereby certify that this thesis has been composed by me and is based on my own work, unless stated otherwise. No other persons work has been used without due acknowledgement in this thesis.

All references and verbatim extracts have been quoted, and all sources of information, including graphs and data sets, have been specifically acknowledged. I further declare that I have not submitted this thesis at any other institution in order to obtain a degree.

Mittweida, August 19, 2020

Julia Abel