MASTER'S THESIS

Expert Tuned Profile Hidden Markov Models for Primary and Secondary Structure Based Homology Prediction in Bioinformatics

submitted to the
Department of Information Technology and Systems Management
at the
Salzburg University of Applied Sciences

by

Christian Winkler, BSc

Salzburg University
of Applied Sciences

Head of Department: FH-Prof. DI Dr. Gerhard Jöchtl

Supervisor: FH-Prof Univ.- Doz. Mag. Dr. Stefan Wegenkittl

Salzburg, September 2018

Declaration on Authorship

I confirm that this Master's thesis is my own work and that I have not used any sources other than those listed in the bibliography and identified as references.

This thesis was not previously presented to another examination board and has not been published.

Salzburg, September 2018

1510581041

Christian Winkler, BSc	Matrikelnumber

Common Information

Name:	Christian Winkler, BSc
Institution:	Salzburg University of Applied Sciences
Degree Programme:	Information Technology & Systems Management
Title of Thesis:	Expert Tuned Profile Hidden Markov Models for Primary and Secondary Structure Based Homology Prediction in Bioinformatics
Keywords:	Homology Detection, Sequence Similarity, Hidden Markov Model, Protein Alignment, Bioinformatics
Supervisor:	FH-Prof Univ.- Doz. Mag. Dr. Stefan Wegenkittl

Abstract

Protein homology classification is an important task to better understand proteins whose three-dimensional structure and function are not obvious. Current methods that rely on the primary structure of a protein do not always find distant homologous relationships between proteins. This thesis examines the usage of secondary structure information in profile Hidden Markov Models to improve the classification accuracy in protein family prediction. This is done by extending the emission frequencies of the primary structure by secondary structure frequencies. A generalized mixed frequency set is generated using optimized weighting techniques. The secondary structure is determined by the three-dimensional structure, if available, or predicted from the primary structure. To assess the effectiveness of the different weighting methods, the implementation has been tested with 69 selected sequence alignments, representing distant related families. These sequences have been scored against the SCOP database to determine the accuracy of finding distant homologous relationships between proteins. Results show that the integration of secondary structure information improves the accuracy of homology prediction.

Contents

List of Abbreviations

Å	Ångstrom
ASA	Accessible Surface Area
AUC	Area Under the Curve
CCD	charge-coupled device
DNA	deoxyribonucleic acid
DSSP	Dictionary of Secondary Structures in Proteins
FPR	false positive rate
GUI	graphical user interface
HMM	Hidden Markov Model
MSA	multiple sequence alignment
NMR	Nuclear Magnetic Resonance
NN	neural network
PCA	principal component analysis
PDB	Protein Data Bank
PHD	Profile Network from HeiDelberg
pHMM	profile Hidden Markov Model
PP	physio-chemical property
PSSM	Position-Specific Scoring Matrix
RNA	ribonucleic acid
ROC	Receiver Operating Characteristic
SCOP	Structural Classification of Proteins
SCOPe	Structural Classification of Proteins–extended
SID	SCOP Identifier

SVM support vector machine

TPR true positive rate

UniProtKB UniProt Knowledgebase

XRC x-ray crystallography

List of Figures

List of Tables

1 Introduction

Proteins are the essential building blocks of all forms of life. They can be found in all living organisms and play a key role in different functions of life. All proteins are made up of one or more chains arranged together in specific amino acid sequences. Although these sequences are built from a set of just 20 different elements, they can arrange in an innumerable number of configurations. With regard to the human body, there are more than 20,000 different proteins.

1.1 Research purpose

Understanding the function of proteins is a crucial task in bioinformatics. Each new discovery of a protein structure and its function provides more knowledge about how the macro-molecular mechanisms of life work. This knowledge is important in many different areas, such as the pharmaceutical industry, where it is necessary to know the shape of viruses or bacteria in order to design drugs to counter them. Another application is biotechnology, which aims to develop new technologies, such as self-assembling organic solar cells.

Although at present determining protein sequences is a simple and straightforward task, it produces an enormous amount of protein sequence data. Gaining information about the three-dimensional shape of the protein is a slow and strenuous task. Several prediction methods based on the sequence data already provide a good indication of the structure of certain parts of a protein. Nevertheless, prediction of the complete structure remains an unsolved problem. One approach to gain further information about the function of a protein is to find evolutionary connections, also known as homology, to other known structures. Proteins with the same functions are likely related to each other, as they might have originated from a common ancestor. These connections can be inferred according to similarities in the sequence and structure and thus according to similar functions. However, the conservation of two homologous sequences is not always clear, as due to mutations several amino acids or even longer sections in the chain may have changed over generations.

This thesis investigates the use of profile Hidden Markov Model (pHMM) in protein family classification. Specifically, the common approach, which uses only the primary structure to build a pHMM and score sequences against it, will be extended to make use of secondary structure information. There are particular cases in which protein homology is not well presented by sequence similarity in certain parts of the protein.

It is expected that stability, with respect to the secondary structure, increases the classification quality.

1.2 Overview

Chapter 2 provides an overview of the areas of bioinformatics relevant to this thesis. These include topics such as protein composition, protein structure and function, and the fundamentals of pHMM.

Chapter 3 discusses protein structure determination on its different levels. A special focus is on methods for secondary structure determination.

In Chapter 4, the implementation of various methods for pHMM using secondary structure information in the software package *HMModeler* is described. Based on the theoretical background, this chapter covers the steps from an multiple sequence alignment (MSA), determining secondary structure, building a pHMM and scoring sequences against it.

Chapter 5 evaluates the implementation based on various datasets. In particular, different methods for including secondary structure information with various parameters are compared to one another.

Finally, Chapter 6 concludes by providing a short summary of the thesis' findings.

2 Selected Background Information

Bioinformatics covers a wide range of scientific fields, such as biology, chemistry and computer science. This chapter gives a short introduction to the areas of bioinformatics relevant to the methods and implementation used in later chapters.

2.1 Proteins

Proteins play a key role in almost all biological activity. Proteins are large biological polymers composed of one or more chains of amino acids. These amino acids are held together by spatial bonds called peptide bonds. Further information about the material described in this section can be found in [1, 2, 3].

2.1.1 Amino Acids

Amino acids are the basic building blocks of proteins. Each amino acid consists of a central α carbon (C_α) bound to:

- a carboxyl group ($COOH$)
- an amino group (NH_2)
- a hydrogen atom (H)
- a distinct side chain (or R-group)

Variation in the side chain defines the 20 common amino acids found in protein molecules. For instance, if the side chain contains just one hydrogen atom, the amino acid is glycine, while the side chain CO_2OH forms the amino acid serine. Table 2.1 lists the 20 common amino acids found in proteins with their assigned three-letter abbreviations and one-letter symbols, as defined in [4]. The abbreviations and one-letter symbols are used to simplify computing with amino acid chains.

Several additional symbols for so-called “ambiguous amino acids” are also defined. These are placeholders for positions in a protein where the exact type of a single amino acid cannot be experimentally determined. For example, the symbol B stands for either aspartic acid or asparagine, while X stands for any one of the 20 common amino acids.

2.1.2 Physio-chemical Properties of Amino Acids

The side chains of amino acids give them specific properties that influence how each amino acid can interact with its surroundings. In [5], Meiler et al. collected seven unique physio-chemical properties (PPs) for each of the 20 amino acid molecules. These are as follows:

Amino acid	Abbreviation	One-letter symbol
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Table 2.1: The 20 common amino acids of proteins with their three-letter abbreviations and one-letter symbols [4].

- *Steric parameter:* The steric parameter is the graph-shape index that characterizes the molecular shape of the molecule through its possible deformations.
- *Volume:* The volume is described as the normalized molecular volume occupied by the atoms of an amino acid.
- *Isoelectric point:* The isoelectric point is the pH-value of a solution in which the amino acid exists in a neutral form, where it is neither positively nor negatively charged.
- *Polarizability:* The polarizability is the dynamic response of the molecule to external magnetic fields in order to form instantaneous dipoles.
- *Hydrophobicity:* Hydrophobicity describes the possibility of interactions between polar solvents like water and the side chain of the amino acid. A high hydrophobicity indicates that the residue has a low ability to bind with the solvent.
- *Helix probability:* This is the probability of an amino acid being identified as α-helix in the secondary structure.

- *Sheet probability:* This is the probability of an amino acid being identified as β-sheet in the secondary structure.

The values for these seven PPs for the 20 amino acids can be found in Appendix A.

2.1.3 Polypeptides

Two amino acids can form a dipeptide through a covalent chemical bond. This bond is formed through the cleavage of a water molecule (H_2O) from the carboxyl group of one amino acid and the amino group of another. The resulting $CO - NH$ bond is called a *peptide bond* (see Figure 2.1).

Peptide bonds can extend into peptide chains with a size of up to several thousand amino acid residues. For example, the longest known and experimentally determined peptide chain is *titin*, with almost 36,000 amino acid residues[1]. However, the average sequence length is 336 amino acid residues[2].

By convention, the reading direction of a polypeptide chain is defined from the N- to the C-terminus. The N-terminus refers to the free amino group at one end and the

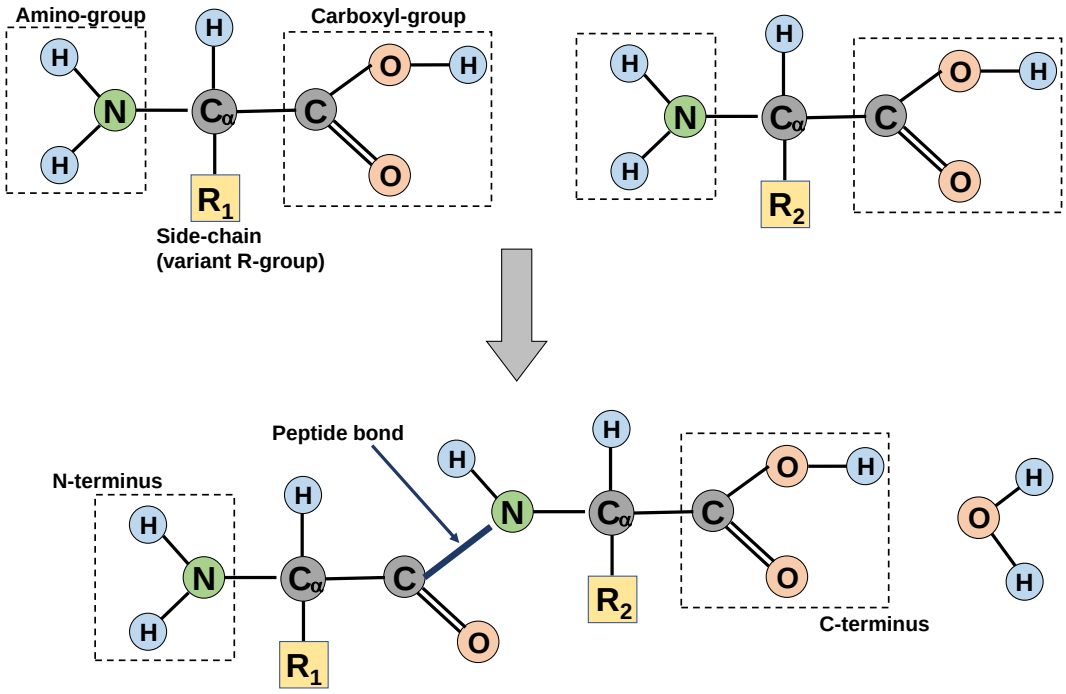

Figure 2.1: Dipeptide forming from the condensation of two amino acids.

[1]Titin UniProt entry: `https://www.uniprot.org/uniparc/UPI000264F4A1` (accessed July 18, 2018).
[2]UniProtKB Statistics: `https://www.ebi.ac.uk/uniprot/TrEMBLstats` (accessed July 18, 2018).

C-terminus to the free carboxyl group at the other end of the chain. Unfolded sequences with up to 50 residues are generally referred to as peptides. For longer sequences, the term "polypeptide" is used. One or more polypeptides that together form a biologically functioning unit when folded into a three-dimensional structure are called a protein.

2.1.4 Protein Backbone

The continuous repeated chain of the central covalent bounded atoms (-NH-C_α-CO-) in a polypeptide forms the main chain and is referred to as the backbone. For each residue in a polypeptide backbone, the three dihedral angles ϕ, ψ, ω define its orientation in space. As shown in Figure 2.2, the dihedral angle is the torsional angle at the intersection of two planes over four consecutive atoms of the backbone, with the angle ϕ between NH and C_α, ψ between C_α and CO, and ω at the peptide bond, from one amino acid residue to the next.

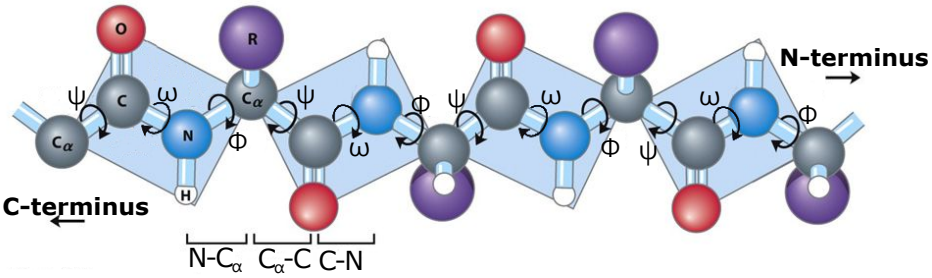

Figure 2.2: The dihedral angles ϕ, ψ, ω in a polypeptide, adapted from [1].

Although the dihedral angles are measured in values from -180° to +180°, the rotation is restricted by several factors. Peptide bonds have a partial double-bonded character because they resonate between a single-bond and double-bond state. This restricts the rotation angle ω to two possible configurations, usually with an angle of 180° or in some rare cases 0°. The other two angles ϕ and ψ are single-bonded and can therefore in principle rotate freely. However, the possible orientations are limited by the variant R-group, as atoms of certain side chains may interfere with the backbone. Ramachandran plots are used to display the statistical distribution of the combination of possible conformations for both dihedral angles of each residue.

The Ramachandran plot in Figure 2.3 shows the freely available conformation space of amino acid residues, based on the atoms' *van der Waals* radii[3]. Green regions mark orientations without interference. The light green regions are less common, especially for residues with larger side chains, but are a possible conformation with reduced van

[3]The van der Waals radius is the spherical region of an atom that represents the distance that another atom approaches before the repulsive force becomes too strong.

der Waals radii. Except for glycine, white regions are not malleable. Glycine, with its small side chain of just one hydrogen atom, has a broader flexibility and therefore can orient itself in all four quadrants of the Ramachandran plot.

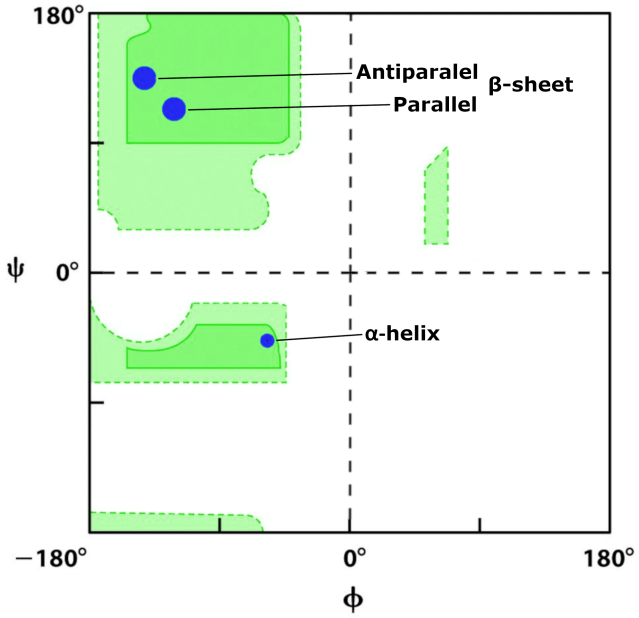

Figure 2.3: Ramachandran plot marking possible conformation of amino acid residues for the dihedral angles ϕ on the horizontal and ψ on the vertical axis, adapted from [1].

2.1.5 Protein Structure

The shape of a protein is critical to its function. When stretched out, polypeptide chains have no functional activity. They become active when arranged in their stable three-dimensional structure, which is dictated by the chain's amino acid sequence. The protein structure can be divided into four hierarchical levels: primary, secondary, tertiary, and quaternary. These four levels, described in more detail below, are shown in Figure 2.4.

The **primary structure** is represented by the linear sequence of amino acids within a polypeptide connected by peptide bonds, where each element corresponds to one of the 20 amino acids described in Section 2.1.1. The primary structure of a polypeptide can be determined from protein sequencing methods such as spectrometry.

The **secondary structure** is defined by the local folding patterns of a region of the protein over a short range of amino acids in a polypeptide. The secondary structure is mainly dependent on the primary structure, which is stabilized by hydrogen bonding due to interaction between the atoms of the backbone. The two most common spatial formations are the α-helix and the β-sheet, connected by irregular segments referred to as loops or coils.

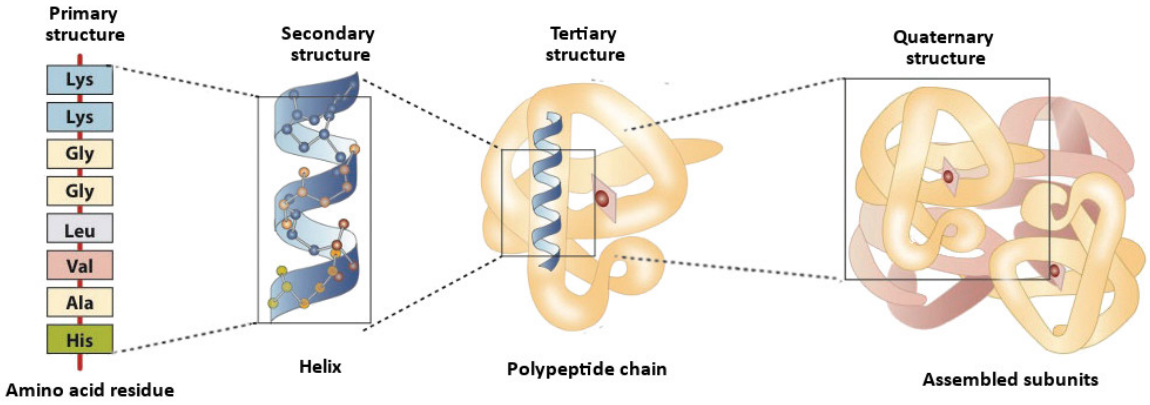

Figure 2.4: The four levels of structure in proteins [1].

The α-helix folds by twisting the polypeptide chain into a right-handed coiled structure, with the side chains positioned on the outside of the helix where they are free to interact with the surroundings. The helix is stabilized by hydrogen bonds between the oxygen of the carboxyl group from one residue and the hydrogen from the amino group of the fourth-next amino acid residue in the polypeptide chain. Each full turn of the helix contains 3.6 amino acid residues. For all amino acids in an α-helix formation, both dihedral angles ϕ and ψ are negative, with regions around of -57° and -47°, respectively, as marked in the third quadrant of the Ramachandran plot in Figure 2.3.

The β-sheet is composed of two or more stretched polypeptide segments, also called β-strands, which are positioned next to each other and held together by hydrogen bonds. The side chains in β-sheets alternate above and below the plane of the strands. Due to the direction of polypeptides, β-sheets can be differentiated into anti-parallel β-sheets when the strands run in opposite directions and parallel β-sheets when they run in the same direction. In the Ramachandran plot, the β-sheet is located in the second quadrant with a positive ϕ and negative ψ of around -139° and +135°, respectively, for parallel β-sheets, and with -119° and 113°, respectively, for anti-parallel β-sheets.

Loops or coils are irregular regions of a polypeptide not recognized as one of the two above, and mainly connect these other structures. However, some patterns are distinguishable, for example hairpin loops between two anti-parallel β-sheets that can be as short as two residues. Turns are narrow 180° loops.

If the tertiary structure is available, the secondary structure can (almost trivially) be computed from the former. Otherwise, it has to be predicted from the primary structure. Methods for secondary structure determination are explained in Section 3.4.

The **tertiary structure** is defined by the coordinates of all the atoms in the protein relative to one another and represents the complete three-dimensional structure of the entire folded polypeptide chain. Beside hydrogen bonds, the tertiary structure is also stabilized by ionic bonds, van der Waals interactions and disulfide bridges. Ionic bonds form between oppositely charged amino acid side chains, such as lysine with aspartic acid. A van der Waals interaction is a weak force of attraction between adjacent atoms that come close to their outer electron cloud, which induces charge fluctuations. Disulfide bonds are like peptide bonds. They are covalent bonds, but they form between the side chains of two *cysteine* amino acids. The tertiary structure of a protein is determined by experimental methods such as x-ray crystallography (XRC) and Nuclear Magnetic Resonance (NMR) spectroscopy.

The **quaternary structure** represents complex protein structures made up of multiple polypeptide chains and shows how they interact with one another. For example, hemoglobin is made up of four polypeptide chains. Linked together, they serve functions such as oxygen and carbon dioxide transport in red blood cells.

2.1.6 Protein Folding

Peptide chains are assembled piece by piece from the synthesis of ribonucleic acid (RNA) molecules. These RNA molecules contain the blueprint of the amino acid sequence given from the genetic code in the deoxyribonucleic acid (DNA). These sequences are built without a specific shape as unfolded chains or random coils. Protein folding is the process in which the unstable chain is translated into its native three-dimensional structure in order to function correctly. The three-dimensional conformation is influenced by several factors, in particular forces from the interaction of the atoms in the side chains and the thermodynamics of the structure. During the synthesis, short local conformations such as helices and strands begin to form. Subsequently, the whole sequence begins to move until it reaches its native shape, which requires the lowest energy conformation to stay in shape. However, folded proteins are still flexible to a certain degree, as their functional properties may require a dynamic structure. One of the biggest challenges in bioinformatics in recent decades has been the protein-folding problem, which tries to fully understand the dynamics and mechanics of the folding process in order to predict the native structure of a protein from its amino acid sequence. Although the folding process of some peptides can already be predicted with reasonable accuracy, understanding the complete process remains an unresolved problem. Improving our knowledge of the folding process is important to treat diseases caused by misfolding [6].

Misfolding of proteins is one of the main causes of many different types of diseases, such as cancer, Alzheimer's and Parkinson's. Mutations in the DNA cause changes in the amino acid composition during the synthesis, which changes the three-dimensional shape of a protein and thus influences the function of the protein. An example in which a small change in amino acid composition has a significant impact on the fold is sickle cell disease, where in one of the peptide chains of hemoglobin, the amino acid valine is placed at the sixth position of the chain instead of glutamic acid. This mutation causes a deformation of the red blood cells from a disc shape to a sickle-shaped structure. This reduces the cells' ability to transport oxygen through the body. The lack of oxygen transport can lead to symptoms such as anemia and bacterial infections and in the long term can cause death.

2.1.7 Protein Functions

Proteins are central elements in all living organisms. They serve many different functions in the body and can be described mainly in terms of the following functional tasks:

- *Enzyme*: Enzymes are catalysts that are responsible for carrying out all biochemical reactions that take place in body cells. For example, *pepsin* is an enzyme protein in the stomach that is responsible for digesting other proteins in food.
- *Messenger*: Messenger proteins, such as most hormones, are responsible for the communication of cells in one part of the body with cells in another part of the body to coordinate biological processes between different parts of the body. Messenger proteins are generally relatively small peptides. For example, *insulin*, with 51 amino acid residues, regulates the metabolism of carbohydrates and fats.
- *Structural Protein:* Structural proteins maintain the structural integrity of cells, organs and connective tissues and are sometimes involved in cell movement. An example is *keratin*, which serves as a protective cover of many different body parts, such as skin, hair and nails.
- *Transport:* Carrier molecules or transport proteins are responsible for carrying and sometimes also storing substances within the body. For example, *hemoglobin* takes oxygen from the lungs and transports it in the blood through the body to the tissues. Other transport proteins include *myoglobin*, which takes the oxygen from the hemoglobin and stores it until needed by the muscle tissue.
- *Antibody:* Antibodies are proteins in the blood that defend the body against

diseases from harmful intruders such as viruses or bacteria. When intruders enter the body, the immune system creates antibodies to identify the intruders in the body and destroy them.

2.1.8 Protein Homology

Two proteins are homologous when they share common evolutionary ancestors. Often, homologous proteins with similar biological functions have similar sequences and structures, at least in some crucial regions. Differences occur due to mutations in the sequence, such as substitutions, insertions, and deletions of single amino acids. The degree of homology is determined by metrics such as the similarity of two sequences. High sequence similarity between two sequences is an indication of a shared ancestor, whereas the probability of these sequences having originated independently of each other increases with decreased sequence similarity. As the structural fold of a protein is crucial to its function, regions that are critical to its function are more conserved than irregular loops. A way to model and describe protein homology is to use MSAs (see Section 2.3.1).

2.2 Protein Databases

Over time, more than 1,600 databases containing bioinformatic data have been created [7]. These databases can be categorized into primary and secondary databases. Primary databases are filled with sequence or structure data derived from researchers' experimental results, whereas secondary databases are composed of data derived from primary databases and organized with additional knowledge, such as family classification.

The following sections detail the sequence database UniProt Knowledgebase (UniProtKB), the structure database the Protein Data Bank (PDB), and the secondary database the Structural Classification of Proteins (SCOP).

2.2.1 UniProtKB

UniProtKB, part of the UniProt database collection, consists of two sections, *TrEMBL* and *Swiss-Prot*. Entries in UniProtKB provide curated information on protein sequences and their function and classification, as well as cross-references to other databases.

TrEMBL currently includes over 115 million sequences[4] derived from high-throughput sequencing methods as described in Section 3.1. These entries are automatically anno-

[4]UniProtKB/TrEMBL statistics: `https://www.ebi.ac.uk/uniprot/TrEMBLstats` (accessed July 18, 2018).

tated by combining identical full-length proteins from one species in single records.

Swiss-Prot provides high-level annotation, where all new entries, taken from *TrEMBL*, are manually annotated and reviewed by experts, using information from the publications dealing with the sequences. Revised entries that have been added to *Swiss-Prot* are removed from *TrEMBL*. This prevents redundancy and allows interoperability of the sections. The manual annotation provides a high-quality data set with minimal redundancy for the cost of the slow processing of new entries. Swiss-Prot currently has less than 600,000 entries[5] [8].

2.2.2 Protein Data Bank (PDB)

In 1971, the Brookhaven National Laboratory established the Protein Data Bank (PDB) as the primary database for the three-dimensional structures of proteins, nucleic acids, and complex structures. Since 2003, the PDB has been managed by multiple organizations around the world under the umbrella of the worldwide PDB (wwPDB) organization[6], whose founding members are RCSB (USA), PDBe (Europe), and PDBj (Japan). Before a new PDB entry is added to one of their mirrored databases, the protein structure information is reviewed and annotated by one of the responsible organizations.

The data in the PDB are typically derived using methods like XRC and NMR spectroscopy. Protein structures are stored in the PDB file format as three-dimensional positions of each atom together with additional data such as temperature factor, associated species and amino acid residues. Each entry published in the PDB has a unique four-character identifier (PDB ID) [9].

Currently, the PDB contains over 142,000 structures[7] and grows in size by approximately 10 % annually [10]. However, compared to more than 115 million sequences in the UniProtKB Protein Database, three-dimensional structures are available for only a fraction of known proteins.

2.2.3 Structural Classification of Proteins (SCOP)

The SCOP database classifies proteins with known structures from the PDB according to their evolutionary, functional, and structural relationships. Each protein in the SCOP is classified in a hierarchical system with the four main levels of *family*,

[5] UniProtKB/Swiss-Prot statistics: `https://www.uniprot.org/statistics/Swiss-Prot` (accessed July 18, 2018).
[6] Worldwide PDB: `https://www.wwpdb.org/` (accessed July 18, 2018).
[7] PDB - Yearly Growth of Total Structures: `https://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total` (accessed July 18, 2018).

superfamily, *fold* and *class* [11].

- **Family** describes proteins with obvious evolutionary relationships due to high sequence similarity across the protein, or with low sequence similarity but high structural and functional similarity.
- **Superfamily** describes families with low sequence similarity, where, based on structural or functional features, a common evolutionary origin is probable.
- **Fold** describes superfamilies with high structural similarities, based on a similar arrangement of secondary structures and their topological connections.
- **Class** describes folds according to the appearance of their secondary structure, such as proteins with α-helices only.

The convention for describing protein classification is `Class.Fold.Superfamily.Family`. In this convention, the class is represented with an alphanumerical letter, while the fold, superfamily, and family are described using numbers. For example, hemoglobin, with the SCOP Identifier (SID) `D1A3NA_` is classified as `a.1.1.2`, belonging to the family *globins* (2), the superfamily and fold *globin-like* (1), and under the class α-helices only (a).

Until version 1.73 of the SCOP, all protein structures were manually classified. With an increasing number of protein structure publications and the subsequent growth of the PDB, manual classification became too slow to classify all new proteins. Therefore, in later versions and with the introduction of the Structural Classification of Proteins–extended (SCOPe), automated processes were introduced [12].

2.3 Profile Hidden Markov Model

This section focuses on the main parts of pHMM, including MSA and scoring methods. More background, especially regarding the fundamentals of Hidden Markov Models (HMMs) such as Markov chains, can be found in [13, 14].

2.3.1 Multiple Sequence Alignment (MSA)

The pHMM used for homology detection is initialized by a set of protein sequences that are aligned to each other for a high degree of structural similarity.

Two sequences can be compared by a pairwise sequence alignment, where the residues of both sequences are directly compared to each other, allowing for the use of gap positions between the residues. Two protein sequences are highly similar when they have many match-states of the same or similar amino acid residues and few gaps. Sequences

with high similarity are also highly likely to be homologous and display structural and evolutionary conservation. An MSA arranges three or more sequences to one another. This method is especially commonly used on a set of sequences that are related to one another in order to identify homologous residues or sequence patterns that may have diverged from common ancestral residues. Functionally important residues in a sequence are assumed to remain stable over generations and are less likely to mutate. Therefore, highly conserved regions, with high similarities over the columns of the MSA, are identified as functionally important. With a higher number of homologous sequences in an MSA, common residues can be identified with a higher probability, while the likelihood of random similarities occurring decreases. Nevertheless, a low sequence similarity does not imply that these sequences are not homologous. Positions of the MSA can be conserved by additional parameters, such as PP. For example, a column or region of the MSA that only contains hydrophobic residues is likely to serve a water-repelling function in the protein. For certain proteins, the homology based on its sequence is not obvious, but a similar three-dimensional structure and function may imply a common evolutionary origin. Figure 2.5 shows an MSA composed of 15 sequences from the same superfamily (see Section 2.2.3).

```
# STOCKHOLM 1.0

d2dlqa4  ..........VECPT...CHKKFLSKYYLKVHNRKHTGEK...........
d2dlqa3  ......SEQVFTCSV...CQETFRRRMELRLHMVSHTGE............
d2dlqa2  .........PFECPK...CGKCYFRKENLLEHEARNCMNR...........
d1wjva1  GSSGSSGMVFFTCNA...CGES.VKKIQVEKHVS.NCRNC...........
d1x5wa1  ........HPEKCSE...CSYSCSSKAALRIHERIHCTD............
d1bboa2  .........PYHCTY...CNFSFKTKGNLTKHMKSKAHSKK..........
d2vy4a1  .......DEVVICPY...DSNHHMPKSSLAKHMASCRLRKMGYTK......
d1m36a_  ...GSRLPKLYLCEF...CLKYMKSRTILQQHMKKCGWF............
d2drpa2  .........VYPCPF...CFKEFTRKDNMTAHVKIIHK.............
d1llmc2  .......EKPFACDI...CGRKFARSDERKRHRDIQHI.............
d1tf3a3  ........KNFTCDSDG.CDLRFTTKANMKKHFNRFHNIK...........
d2ctda2  ........EMFTCHH...CGKQLRSLAGMKYHVMANHNSLP..........
d2glia1  .........ETDCRWDG.CSQEFDSQEQLVHHINSEHIHGER.........
d1wjpa3  .......YKKLTCLE...CMRTFKSSFSIWRHQVEVHNQNNMAPTSGPSSG
d2j7ja3  .........GYPCKKDDSCSFVGKTWTLYLKHVAECH..............
//
```

Figure 2.5: Fifteen sequences from the SCOP superfamily g.37.1 aligned to an MSA.

There are different approaches for generating an MSA, using either manual annotation by expert knowledge or automatic methods based on structural information on its different levels (see [13, Chapter 6]).

2.3.2 Profile Hidden Markov Model

A pHMM is a stochastic model based on Markov chains used especially in bioinformatics to model an MSA and capture its degree of structural conservation. Figure

2.6 shows the structure of a pHMM, which uses three different types of hidden states for each column of the MSA. These are match state M_k, delete state D_k and insert state I_k.

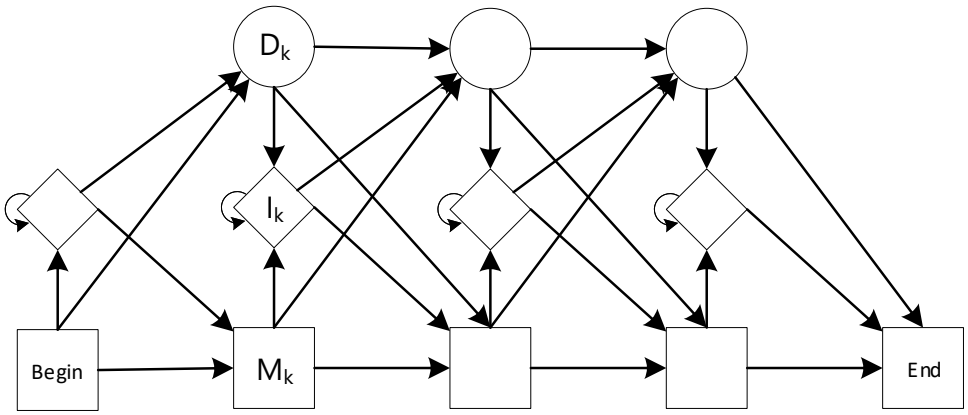

Figure 2.6: Structure of a pHMM with match, insert and delete states, adapted from [13].

Match state M_k is described by a distribution, trained with the symbol frequency in column k of the MSA. Each representative column in the MSA model matches one state. Gaps in the MSA influence the insertion and deletion states.

Insertion states model genetic insertions of additional symbols in a sequence that do not appear in the majority of the sequences of a family of proteins in the MSA. They occur when there are many gaps in a column, typically above 50 %. Insertion states make it possible to insert the symbols in the MSA for those sequences that do not show gaps. Insertion states have a self-transition to cover repeated insertions over multiple columns. Instead of the actual symbol frequency of the MSA column, the emission probabilities can be set to a background probability based on a general frequency of the symbols, as a few residues in the column might not be representative.

Deletions allow a sequence to skip over a match state without emitting a symbol and model situations of genetic deletions, where a certain sequence has one or more positions fewer than the other proteins in a family modeled by the MSA. Deletions are silent states without emitting symbol probabilities. Self-transitions are not allowed for deletions. However, gaps over multiple columns are modeled as a sequence of deletion states, allowing different transition probabilities over multiple gap positions.

2.3.3 Decoding and Scoring

In a fully defined pHMM, any sequence can be represented by a matching score, by multiplying the transition and emission probabilities along the path. Due to the insert and delete states, many different paths can produce the same sequence. To find the

best-suited path for a specific sequence, dynamic programming approaches are used, such as the *forward algorithm* or the *Viterbi algorithm*.

The forward algorithm calculates the overall probability as a sum for each individual state of the pHMM. The Viterbi algorithm calculates the probability for the most likely sequence of states by using the maximized path over the pHMM.

To prevent an underflow caused by the multiplication of many probabilities in the range of 0–1, a logarithmic scale is used to calculate the raw score. However, these raw scores are highly dependent on the overall sequence length, as longer sequences generate smaller scores, and thus these raw scores are difficult to compare. Therefore, the scores can be normalized using correction methods such as null models, which can either be simple null models or reversed-sequence null models. The simple null model uses a simplified one-state HMM with a specific residue composition based on general occurrence frequencies of the amino acids and scores the sequence against it. The reversed-sequence null model uses the original pHMM and scoring algorithm, but reverses the direction of the sequence employed. The final score without length dependency is calculated by subtracting the null model score from the raw score in the logarithmic scale.

3 Protein Structure Determination

This chapter covers the fundamental methods of protein structure determination on its different levels. It begins with protein sequencing for determining the primary structure of a polypeptide. The discussion proceeds with the two major methods for determining the tertiary structure, namely XRC and NMR. Finally, methods for secondary structure determination are described, where the secondary structure is either derived from the tertiary structure or predicted from the primary structure.

3.1 Protein Sequencing

Protein sequencing is the process of determining the primary structure of a polypeptide. Protein sequencing is the first essential step for gathering more information about the structure and function of a protein. Two major methods for protein sequencing are Edman degradation and mass spectrometry.

Edman degradation, developed by Pehr Edman in 1950, determines each amino acid residue through a repetitive approach, where the peptide bond of the first amino acid at the N-terminus of the peptide chain is labeled and chemically separated from the chain while keeping the remaining bonds untouched. The separated residue is identified by procedures such as ion exchange chromatography. Ion exchange chromatography separates ionizable molecules according to differences in their total charge by increasing the ionic force of the sample medium and measuring the strength needed for each element to separate. The Edman degradation method is only reliable for short peptides up to 40–60 residues. Longer proteins must be analyzed by fragmentation, where the polypeptide is split into shorter sections and these sections are analyzed one after another [3].

Another approach is protein sequencing by mass spectrometry. Mass spectrometry uses the spectrum of a peptide to determine its sequence according to the different masses of the 20 amino acid residues. This is done by ionizing the protein molecules in a strong electric field. The ionized particles can then be detected by a mass analyzer, which provides the mass-to-charge ratio for each ion. The mass-to-charge ratio over the induced electrical energy produces a mass spectrum. The amino acid sequence can be determined from this mass spectrum by fragmentation of the molecules and also by computational analysis and matching with databases [15].

With high-throughput variations of mass spectrometry such as tandem mass spectrometry or protein sequencing machines that automatically perform the process of Edman

degradation, the amino acid sequence can be determined within hours [1].

3.2 X-ray crystallography

The first high-resolution three-dimensional protein structure myoglobin was identified in 1958 by John Kendrew using x-ray crystallography (XRC), for which he received the 1962 Nobel Prize in Chemistry [16]. Ninety percent of the determined protein structures in the PDB have been solved using XRC [10].

X-ray crystallography is a form of microscopy that uses electromagnetic radiation beams. These beams have a wavelength of around 1 Ångstrom (Å)[1]. For comparison, the inter-atomic distances of protein crystals are approximately 1–3 Å. The X-ray beams are diffracted by the protein's electrons when passing the crystal. The resulting diffraction pattern is collected by digital charge-coupled device (CCD) image sensors for further processing. Several measurements with different angles and intensities produce multiple diffraction patterns, from which a three-dimensional electron density map can be computed. The density map provides direct information about the mean positions and size of the atoms and the length and types of chemical bonds, among other aspects, which enables its primary structure to be modelled into its three-dimensional structure.

In order to obtain usable results from the protein molecules, the sample has to be purified and set in a stable crystal state. Protein molecules in their natural form are not stagnant, as there is always movement, rotation and bending in the structure. In addition, the x-ray radiation during the measurement influences the sample. A high-quality crystal is composed of a regular repeating arrangement of proteins, with a size in all dimensions of at least 20 µm that can extend up to and beyond 0.1 mm.

Protein crystallization is the crucial part of XRC which determines whether a protein structure can be determined using this approach. There is no straightforward approach for crystallization, as each new protein may respond differently. The path from a protein to a usable crystal often involves an extensive trial and error approach, which can last months. Furthermore, some proteins do not form crystals at all, without any indication why this is the case [17]. There are several different physical, chemical, and biochemical factors affecting the crystallization process. For example, the earth's gravity force can prevent crystals from growing to the required size. Experiments in microgravity have in many cases improved the size and quality of protein crystals. Currently, on a regular basis, protein samples that fail to grow crystals with the required

[1] Ångstrom (Å) is a unit to measure the wavelength of light, 1 Å equals to 10^{-10} meter or 0.1 nanometer.

quality and size are transported to the International Space Station (ISS), where the crew continues to search for usable crystals[2]. However, these experiments presuppose a certain degree of crystallization success on Earth, as the process must be highly automated and practicable for the ISS crew to be able to implement it [18].

3.3 Nuclear Magnetic Resonance Spectroscopy

Nuclear Magnetic Resonance (NMR), which has been used to determine 9 % of protein structures in the PDB, is the second-most-used application for identifying the three-dimensional structures of proteins [10].

This method makes use of the quantum-mechanical property of subatomic particles known as spin, which can be thought of as a rotation around the particle's own axis. Atomic nuclei with an odd number of protons and/or neutrons have spin, while nuclei with even numbers of these particles have no spin. Nuclei without spin cannot absorb or emit electromagnetic radiation, and therefore cannot be used for NMR. However, most nuclei in proteins, such as those in the backbone (^{2}H, ^{13}C, ^{15}N, ^{17}O), do have spin.

The spinning of the nucleus generates a nuclear magnetic moment. When an external magnetic field is applied, the nucleus acts like a magnetic dipole and can orient itself in two different spin states, referred to as α and β. The α state is the preferred orientation in terms of energy, where the magnetic moment matches with the applied magnetic field. The orientation can be reversed to the β state by irradiating the nucleus with an additional electromagnetic radiation frequency. The so-called resonance frequency is equal to the energy difference needed for the nucleus to switch its state and is directly proportional to the magnetic field applied.

By varying the frequency while maintaining a static magnetic field or vice versa, a resonance spectrum for a molecule can be obtained. The resonance spectrum indicates the energy necessary to put various nuclei in resonance. Each nucleus has its own characteristic resonance frequency, for example ^{1}H with approximately 500 MHz and ^{13}C with 126 MHz. However, the resonance frequency of a single nucleus also varies at different locations in the molecule, due to various interactions between other subatomic particles. For example, negatively charged electrons have a so-called shielding effect, which reduces the force from the applied magnetic field that is absorbed by the nearby nuclei. This effect also reduces the electromagnetic radiation intensity

[2]NASA Protein Crystals in Microgravity: `https://www.nasa.gov/mission_pages/station/research/benefits/mab` (accessed July 18, 2018).

needed to spin the nucleus in β state. Furthermore, the interaction from one nucleus with its surrounding nuclei can be measured by emitting frequency pulses that only bring specific nuclei in resonance. The information obtained is used to calculate the three-dimensional structure of a protein through computational analysis and molecular modeling procedures [19].

An advantage of NMR over XRC is that the protein can be analyzed while in solution. Therefore, NMR can be used on proteins that refuse to form crystals. The liquid form also provides information about the protein's stability and its dynamic processes, as the protein has freedom to move. However, the fluid state also impacts the protein's stability, as its structural integrity must be maintained over the entire experiment. The movement of the protein in solution also restricts the size of the protein, in most cases to less than 30 kDa[3] or on average 250 amino acid residues [20]. Furthermore, high-resolution NMR spectrometers are relatively expensive, as the resolution is directly related to the magnetic field strength. To create the magnetic field necessary for protein structure determination, NMR spectrometers need expensive liquid-helium-cooled superconducting magnets.

3.4 Secondary Structure Estimation

Secondary structure prediction methods rely on the annotation of known three-dimensional data, as the annotated structure serves as the training input for prediction methods. The accuracy of secondary structure prediction methods is typically measured in the percentage of correctly assigned elements for each of the three main elements, namely α-helix (Q_H), β-strand (Q_E) and turn (Q_C), and in particular the overall three-state accuracy (Q_3).

Early methods of secondary structure prediction, such as the *Chou-Fasman method* developed in 1978, used an approach which assigns the classes α-helix and β-strand based on statistical properties on the amino acid residues. The later *GOR method* includes statistical principles based on Bayes' theorem to include the conformation probability of the chain. While early implementations of GOR and Chou-Fasman only reached a Q_3 accuracy of 50–60 %, the current implementation of GOR V reaches a Q_3 accuracy of 73.5 % [21].

Modern approaches make use of statistical methods like neural networks (NNs), support vector machines (SVMs) and HMMs. These methods use black-box approaches, where

[3]Dalton (Da) is a unit to measure atomic mass, 1 Da is equal to 1/12 the mass of a single carbon-12 atom.

the path from the sequence to the structure is not obvious, due to the model being trained using machine learning. These methods reach a Q_3 accuracy of $> 80\,\%$. With growing training data from new experimentally determined tertiary structures, this value is also increasing over time. Nevertheless, in theory the average Q_3 accuracy limit for predicting secondary structure is approximately $88\,\%$ [22].

In what follows, methods for determining secondary structure information will be explained. First, the secondary structure annotation method Dictionary of Secondary Structures in Proteins (DSSP), which uses the tertiary structure, will be described. Subsequently, secondary structure prediction approaches based on the primary structure will be discussed. These include the state-of-the-art methods Profile Network from HeiDelberg (PHD), SPINE-X and MetaSSPred.

3.4.1 The Dictionary of Secondary Structures in Proteins (DSSP)

The DSSP, described in [23], is a method developed by Kabsch and Sander to determine the secondary structure of a protein based on the atomic coordinates of the protein. It uses pattern recognition in hydrogen bonding and specific geometric features to assign one of eight secondary structures (see Table 3.1) to each residue in a protein. These eight types can be grouped into three major components: helix (H, G, I), strand (B, E) and loop (T, S, C).

3-state SS	DSSP	Secondary Structure
α (H)	H	α-helix
	G	3_{10} helix
	I	π-helix
β (E)	B	residue in isolated β bridge
	E	extended strand
L (C)	T	hydrogen bounded turn
	S	bend
	C	loop or irregular

Table 3.1: Structure types determined by DSSP [23].

Hydrogen bonds are determined by approximation of the interaction energy between the NH group of one amino acid and the CO group of another spatially proximate amino acid. The total electrostatic energy E, measured in $\frac{kcal}{mol}$, is calculated according to (3.1), where r_{AB} represents the inter-atomic distance of the two atoms A and B measured in Å, the electron charges $q^+ = 0.2\,e$ and $q^- = -0.42\,e$ between NC and OH, respectively, and a dimensional factor $f = 332\,\text{Å}\frac{kcal}{e^2}$.

$$E = fq^{+}q^{-}(\frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}}) \tag{3.1}$$

When $E < -0.5\,\frac{kcal}{mol}$, it is assumed that this position forms a hydrogen bond. After determining all hydrogen bonds, the secondary structure is assigned by identifying the basic hydrogen binding patterns. For example, the α-helix is identified by two consecutive hydrogen bonds between amino acid residues of the backbone that are four positions apart. The 3_{10}-helix and π-helix are variations with hydrogen bond distances of three and five residues, respectively. The two types of beta strands are stretched chains arranged for repeating hydrogen bonds with other strands

There are many other annotation methods to annotate secondary structure. For example, STRIDE uses a similar variation of hydrogen-bond patterns, but also takes into account the dihedral angles ϕ and ψ of the protein backbone [24]. In contrast, DEFINE uses a different approach, by comparing only the coordinates of the central C_α atoms with linear distance masks of the different ideal secondary structures [25]. The secondary structure outputs from these methods are not unique. By comparing the results reduced to the three major components, the two similar approaches DSSP and STRIDE agree on 95 % of all positions. Including DEFINE, all three methods have only 75 % agreement [26]. While there are more modern and complex approaches than DSSP available, this approach has become the standard method for most secondary structure annotation tasks. In [27], Wilman explains the general acceptance of DSSP based on the simplicity of the algorithm and the availability of the code and software under the permissive non-copyleft free software license Boost[4].

3.4.2 Profile Network from HeiDelberg (PHD)

The secondary structure prediction method Profile Network from HeiDelberg (PHD), introduced in [28] by Rost and Sander, was the first method to predict secondary structure with an accuracy greater than 70 % from the primary structure alone.

The PHD method uses three steps to predict the secondary structure. In the first step, sequences with a high sequence similarity are obtained and aligned to an MSA. This includes additional evolutionary information, as sequences with a high sequence similarity are likely to share the same function, and thus a similar structure. In the second step, the probability for each amino acid in the MSA is fed into a two-level feed-forward NN system, previously trained through back-propagation with proteins

[4]Boost Software License: `https://www.boost.org/LICENSE_1_0.txt` (Accessed: July 18, 2018).

of known structures. In the first level, for each position in the sequence, a sliding window of 13 consecutive amino acids is fed. This is done with the frequency of the 20 amino acids and an additional position for padding of both ends of the sequence. The first NN level outputs a likelihood of the central residue being a helix, sheet, or loop. The independently trained second NN level receives the output from the first level and inputs the likelihood of the three states with a sliding window of 17 elements. The second-level output is an optimized likelihood for the same three major secondary structure elements. The highest value for each element of the NN determines the secondary structure. A reliability index for each residue is set with the difference between the two highest values. The final step is a simple filter to remove obvious errors, like helices shorter than three residues long.

Secondary structure prediction with PHD is available online with the PredictProtein web server[5] or locally through the Debian Linux package PROFPhd. The expected accuracy for the three major secondary structure types is 76 % on average [29].

3.4.3 SPINE-X

Another high-accuracy secondary structure prediction method is SPINE-X by Faraggi et al., as described in [30].

In addition to the secondary structure, SPINE-X also predicts the Accessible Surface Area (ASA) and the dihedral angles ϕ and ψ for each amino acid residue. The ASA is the surface area of an amino acid that can interact with a solvent and is measured in square Å. SPINE-X uses an iterative approach of six NNs, each fed with additional data predicted from the previous steps. Like PHD, SPINE-X includes evolutionary information about the sequence by generating a Position-Specific Scoring Matrix (PSSM), which describes column-wise sequences similar to MSAs. The aligned sequences are collected by PSI-BLAST, a tool for collecting and aligning distant relatives' proteins for the query sequence (see [31]). Furthermore, the NNs are also individually trained with back-propagation and make use of sliding windows.

The first NN predicts the secondary structure by using the PSSM generated from the sequence and seven PPs for each amino acid residue as input. These PPs include hydrophobicity, volume, steric parameter, polarizability, isoelectric point, helix probability and sheet probability (see section 3.4). The secondary structure predicted from the first step, together with the PSSM and the PPs, is fed to the second NN and provides the ASA for each residue. The third NN uses the same properties as before

[5]PredictProtein web server: `https://www.predictprotein.org/` (Accessed July 18, 2018).

and includes the ASA in order to predict the two dihedral angles. The stages before the fourth step also include the output from the last step, but without the secondary structure predicted from the first step, in order to predict an optimized version of the secondary structure. Including the secondary structure in the input, the two dihedral angels are re-predicted in the fifth NN. The sixth and final NN predicts the final secondary structure elements, based on the PSSM, PPs and ASA from step two and the dihedral angles from step five. SPINE-X predicts the secondary structure based on the eight-state DSSP assignment. To increase accuracy, the final result is grouped into the three major types helix (H), strand (E) and coil (C) (see section 2.1.4).

SPINE-X reaches a Q_3 accuracy of 82–83.8 % and is available as Linux package or online through a web server[6].

3.4.4 MetaSSPred

MetaSSPred is a secondary structure prediction method, developed in the Bioinformatics and Machine Learning Lab[7] at the University of New Orleans (see [32]).

MetaSSPred uses three binary SVMs, together also referred as cSVM, to predict the secondary structure, then combines its results with SPINE-X to improve the accuracy, especially for β-sheets. While modern secondary structure prediction methods reach an average Q_3 accuracy of more than 80 %, the accuracy of each type varies greatly. In particular, β-sheet scores are likely to be far below the others. For example, while SPINE-X reaches a Q_3 accuracy of 82 %, helices are correctly predicted with a Q_H of 86.6 %, coils with a Q_C of 81.5 %, and sheets with a Q_E of 75.3 % [30].

The SVMs are trained by 33 features for each amino acid listed in Table 3.2. These features are assigned directly by the primary structure or gathered by the tools described below. For each amino acid residue, the seven PPs as described in Section 2.1.2 are used. The terminal indicator marks whether the residue is one of the first or last five residues in the sequence. As in the case of SPINE-X, the PSSM is determined by PSI-BLAST. The disorder probability is the probability that an amino acid residue in the chain does not have a well-defined three-dimensional structure and can thus arrange in several other conformations. The disorder probability is predicted by the application DisPredict (see [33]). The two dihedral angles ϕ and ψ are predicted by SPINE-X (see Section 3.4.3). While ASA could also be taken from the SPINE-X prediction, the application REGAd^{3}p with its higher prediction accuracy is used [34].

[6]SPINE-X web server: `http://sparks-lab.org/SPINE-X/` (Accessed July 18, 2018).
[7]Bioinformatics & Machine Learning Lab UNO: `http://biomall.cs.uno.edu/` (accessed July 18, 2018).

Feature	Count	Method
Amino acid (AA)	1	Primary Structure
physio-chemical property (PP)	7	Primary Structure
Terminal indicator	1	Primary Structure
Position-Specific Scoring Matrix (PSSM)	20	PSI-BLAST
Disorder probability	1	DisPredict
Torsion angles (ψ, ψ) fluctuation	2	SPINE-X
Accessible Surface Area (ASA)	1	REGAd^{3}p

Table 3.2: The 33 features used by MetaSSPred with the methods to determine them from the primary structure [32].

For one of the three secondary structure classes, each SVM predicts whether a residue belongs to the secondary structure class and returns a probability that this is the case. For each residue in the sequence, the secondary structure element with the highest probability is assigned. In the final step, the output from cSVM is combined with the prediction from SPINE-X. For each residue, the predicted secondary structure class is compared. For all positions of cSVM where the class sheet (E) is predicted, the result from cSVM will be accepted if it belongs to the class sheet, otherwise the predicted class from SPINE-X is used. MetaSSPred is available as a Linux-based standalone software package[8].

[8]MetaSSPred package: `http://biomall.cs.uno.edu/software/` (accessed July 18, 2018).

4 Implementation

This chapter covers the implementation based on the HMModelers workflow. First, the processing of input data by including secondary structure data is described. This is followed by a description of the adaptation of the pHMM and of the scoring algorithms. This section focuses on the modifications made in HMModeler. More information on HMModeler and its other implementations can be found in [35, 36, 37].

4.1 HMModeler

HMModeler is a software package for protein classification jointly developed by researchers at the Salzburg University of Applied Sciences and the University of Salzburg. Although HMModeler was designed as an extension for UCSF Chimera[1], the current version runs independently with its own graphical user interface (GUI). The core of HMModeler is written in Python, while time-consuming algorithms are also implemented with faster C++ libraries. The web-based GUI communicates with the core system via RESTful[2] web services. HMModeler is platform-independent and runs both on Windows and Linux operating systems. The program uses the Smith-Watermann-style variant of pHMM, shown in Figure 4.1. This variant is used for local alignments, by using flanking states with transitions from the beginning to each match state as start-model, and from each match state to the end state as end-model.

The GUI allows skilled experts to introduce prior information about the targeted protein family into the HMM. In particular, the user can interactively define parts of the protein with increased or decreased insertion and deletion probabilities. The user can also modify the extent to which the emission probabilities in the single-model columns of the pHMM are extracted. This can be done purely through the given MSA; alternatively, the probabilities are determined from a priori distributions. Finally, the user can define so-called expert sets that override other estimation methods and set the possible emissions in certain model states to an explicit set of amino acids. Currently, the software only uses the primary structure for building the pHMM and scores sequences against it with the Viterbi and forward algorithms.

[1] UCSF Chimera `https://www.cgl.ucsf.edu/chimera/` (accessed July 18, 2018).
[2] Representational State Transfer (REST) is an architectural paradigm for communication between distributed machines or applications.

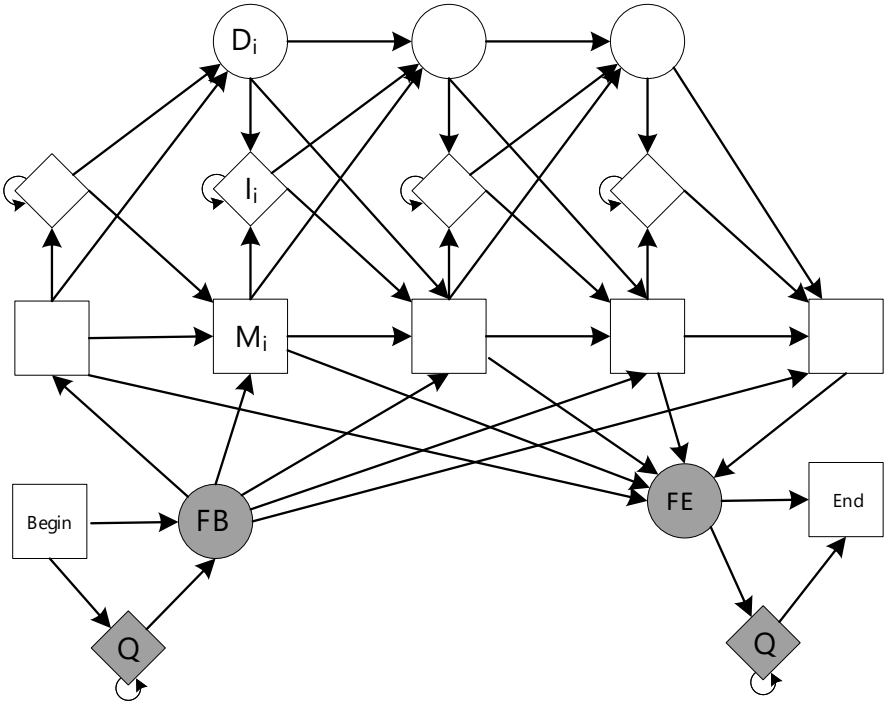

Figure 4.1: Smith-Waterman variant of a pHMM, adapted from [13].

4.2 Input Data

First, HMModeler takes an MSA as input to train its pHMM. HMModeler can process files in the *Stockholm* file format, uploaded by the user or referenced by a project ID from the multiple structure alignment Server PIRATES[3].

The Stockholm format is a markup format for MSA, where each sequence can be annotated with additional features, such as the corresponding secondary structure for each amino acid residue. The definition of the Stockholm format can be found in [38].

Figure 4.2 shows the workflow implemented for determining the secondary structure. First, all sequences in the MSA are read. If no structure information is provided for one or more sequences, the secondary structure is retrieved from the PDB or predicted using the MetaSSPred method (see Section 3.4.4).

The secondary structure from sequences corresponding to three-dimensional data in the PDB is calculated using the DSSP program (see Section 3.4.1). The extracted structural information must be aligned with the original sequence, considering the exact gap positions and possible differences in the overall sequence length.

[3]Multiple structure alignment server PIRATES: `https://biwww.che.sbg.ac.at/pirates` (accessed July 18, 2018).

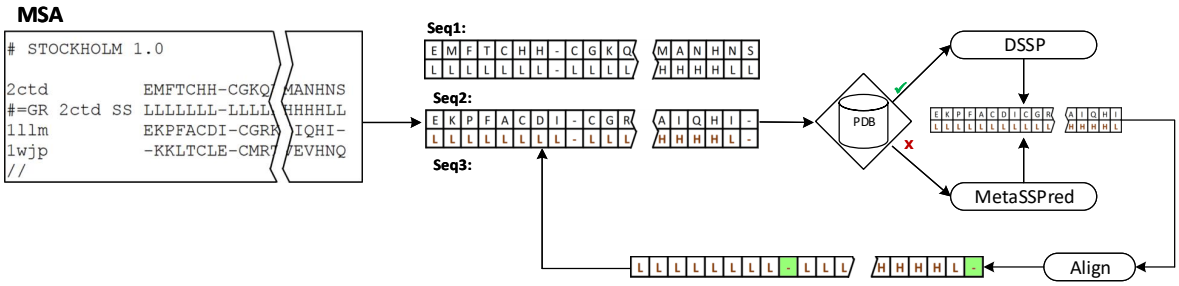

Figure 4.2: Workflow for processing input data and determination of secondary structure.

The Needleman–Wunsch algorithm, described in [13, p. 19–21], is used for this task. Needleman–Wunsch is a dynamic programming approach for global sequence alignment that calculates the optimal path through a scoring matrix using match, mismatch, and gap penalties. The derived sequences, linked to their corresponding secondary structure, are aligned with the original sequence. Provided the original sequence from the MSA matches the aligned sequence from the PDB, the determined secondary structure is used.

If there is no match in the PDB or the alignment fails, the secondary structure is predicted with MetaSSPred. For each sequence that must be predicted, a Fasta-formatted file with all gaps removed in the sequence is generated in the input directory of MetaSSPred and processed by the software. For the prediction process, ambiguous amino acids have to be treated separately, as most prediction methods including MetaSSPred only support the 20 common amino acids. Therefore, those symbols are also removed for the prediction process. Subsequently, the symbol X is added at the removed position in the secondary structure data, which will be handled by HMModeler in the same way as all three secondary structure types. Finally, the predicted secondary structure is aligned to the sequence with the Needleman–Wunsch algorithm.

4.3 Training the HMM

Based on the MSA, which now includes both primary and secondary structures, the transitions and emission distribution are assigned to the pHMM. These are calculated by the frequencies across sequences for each column in the MSA, according to (4.1), where a_{kl} is the transition probability from state k to l, and (4.2) for the emission probability $e_k(a)$ of the symbol a at state k.

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \tag{4.1}$$

$$e_k(a) = \frac{E_k(a)}{\sum_{a'} E_k(a')} \tag{4.2}$$

To prevent overfitting by symbols with a probability of zero, it is important to add some background frequency to each symbol, such as a small non-zero prior probability $pc(a)$ (see (4.3)). The simplest approach is the Laplace smoother, which adds a hypothetical observation in the form of one pseudo-count to each emission and transition.

$$e_k(a) = \frac{E_k(a) + pc(a)}{\sum_{a'} E_k(a') + \sum pc(a')} \tag{4.3}$$

The implementation for the emission probabilities was extended by calculating the primary structure with 20+1 symbols to generate two additional matrices, one with 3+1 symbols for the secondary structure and another with 84 symbols containing mixed probabilities from the other two sets. The additional symbol for the primary and secondary structures was added to cover a wider set of sequences in the test database (see Section 5.1), where single residues or structure elements in sequences may be unknown and therefore marked with the letter X. As the symbol X represents all other symbols in the set, the probability is set to 1. This is equal to the sum of all other probabilities.

The transition probabilities only rely on the gap positions of the MSA. Therefore, no modifications in the current implementation of HMModeler are needed. Moreover, the functional emission probability implementation for the primary structure remains the same.

The emission probabilities for the secondary structure are calculated by using the three major types of helix, strand and loop. Therefore, secondary structure data from methods with higher granularity, such as DSSP, are translated to the three-type annotation (see Table 3.1). To prevent zero probabilities, a prior probability can be set by the user in the HMModeler GUI.

Using both probability sets, the original primary structure implementation and the new secondary structure probabilities, a weighted emission frequency set that covers both primary and secondary probabilities is generated. Mayer discusses the emission probability weighting for use with pHMM in [37]. Following his research, three different implementations are explained below.

4.3.1 Linear Weighting

In [37], Mayer discusses the problem with the different length of the two alphabets, as the primary structures involve 20 amino acid symbols and the secondary structure only involves three symbols. Therefore, he recommends the approach in (4.4), where the mixed emission probabilities $e_k(p, s)$ are weighted by their symbol length with 20 amino acids for the primary probability $e_k(p)$ and the three structure symbols in $e_k(s)$.

$$e_k(p, s) = \frac{20}{20 + 3} \cdot e_k(p) + \frac{3}{20 + 3} \cdot e_k(s) \tag{4.4}$$

However, the most convenient approach for mixing two probabilities, which is by means of a simple multiplication as in (4.5), will also be implemented and tested.

$$e_k(p, s) = e_k(p) \cdot e_k(s) \cdot k \tag{4.5}$$

As mentioned in Section 4.4 below, a threshold can be used to define whether the scoring process uses the mixed or the primary probability. The additional factor k is introduced for scaling the resulting mixed probability. With this factor, the differences in the pHMM resulting from the use of either mixed probability or primary probability can be reduced.

For example, if the three secondary structure elements are equally distributed with a probability of $e_k(s) = \frac{1}{3}$ and a $k = 1$, the value of the mixed probability would also be $\frac{1}{3}$ of the value of the primary probability. By setting the factor $k = 3$, an equally distributed secondary structure would not differ if the primary probability or the mixed probability were used.

4.3.2 Weighting by Shannon

In [37], Mayer introduces weighting the two emission probabilities by the amount of information in each column using the Shannon theorem, defined in [39] as

$$H = -\sum_{i=1}^{N} p_i \log_b p_i \tag{4.6}$$

where the entropy H is the negative sum over the probability of each possible outcome p_i multiplied by the logarithm of the same probability. If a single probabilistic result of an event has a probability of 1, the entropy is 0. In this case, there is no uncertainty. Conversely, if all probabilistic results have equal probability, the uncertainty is max-

imal, with an entropy of $\log_b(N)$, where N is the number of possible outcomes. The logarithmic base b is usually 2, as the common use is digital communication. In other cases, the natural logarithm e is used.

Figure 4.3 shows the entropy as a function for a binary event with probabilities p and $q = 1 - p$ with a natural logarithm and logarithmic base of 2. The maximum entropy occurs when $p = q = 0.5$.

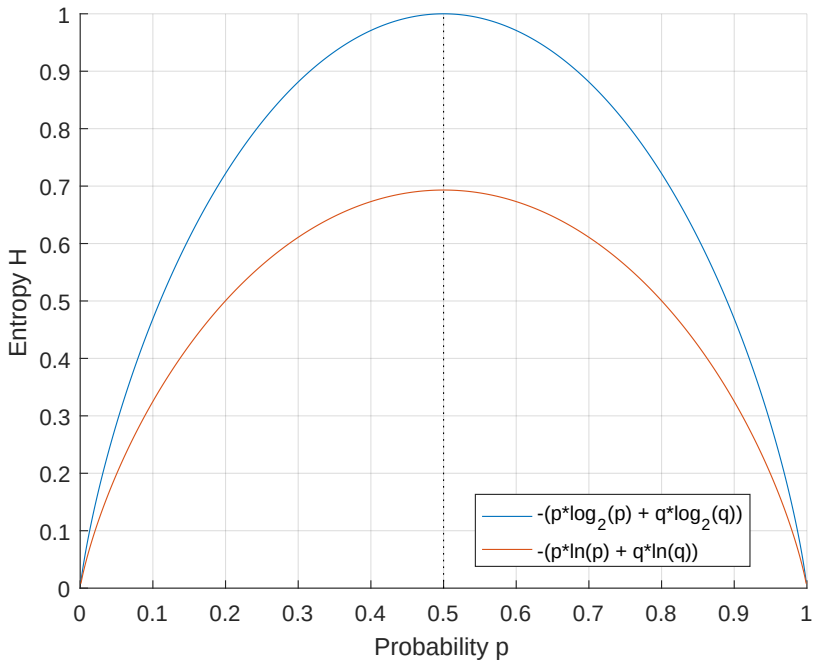

Figure 4.3: Entropy H for an event with two possible outcomes, with natural logarithm (red) and normalized logarithm with base 2 (blue).

The lower the entropy, the higher the degree of information and thus the relevance of the mixed probability. The rate of information I is determined by subtracting the entropy H from the maximal possible entropy $\log_b(N)$:

$$I = \log_b(N) - H \tag{4.7}$$

With the degree of information for both the primary structure $I(p)$ and the secondary structure $I(s)$, the probabilities are weighted as follows:

$$e_k(p, s) = \frac{I(p) \cdot e_k(p) + I(s) \cdot e_k(s)}{I(p) + I(s)} \tag{4.8}$$

Instead of using the natural logarithm for both the primary and secondary structure, as recommended by Mayer, the normalized logarithm with a base equal to the corresponding alphabet size is used in the present thesis. Without the normalization, the probabilities would not be equally weighted, as the maximum entropy for the primary structure would be $ln(20) = 2.9957$ and for the secondary structure $ln(3) = 1.0986$. By contrast, the maximum entropy for both structures using the normalized logarithm is $\log_3(3) = \log_{20}(20) = 1$.

For programming languages that do not support nonstandard logarithmic bases, any logarithmic equation with base a can be converted to any other base b by dividing through the logarithm with the same base a, with the argument set to the new base b:

$$\log_b(x) = \frac{\log_a(x)}{\log_a(b)} \tag{4.9}$$

4.4 Scoring Sequences

HMModeler calculates nine different types of scores, (see Table 4.1) using the Viterbi and forward algorithms. It also computes variations of these for length-corrected scores using the simple null model and the reversed sequence null model (see Section 2.3.3).

Score type	Description
Viterbi score	$\mathrm{V}(s)$
Forward score	$\mathrm{F}(s)$
Simple null model	$\mathrm{N}(s)$
Reverse Viterbi null model	$\mathrm{V}(s^{-1})$
Reverse forward null model	$\mathrm{F}(s^{-1})$
Simple corrected Viterbi score	$\mathrm{V}(s) - \mathrm{N}(s)$
Simple corrected forward score	$\mathrm{F}(s) - \mathrm{N}(s)$
Reverse corrected Viterbi score	$\mathrm{V}(s) - \mathrm{V}(s^{-1})$
Reverse corrected forward score	$\mathrm{F}(s) - \mathrm{F}(s^{-1})$

Table 4.1: Scores calculated by HMModeler.

Only the algorithms for the Viterbi score and the forward score must be adapted for the use of the secondary structure. The reversed null models for both the Viterbi and forward scores use the same underlying algorithm but with the sequence reversed. The simple null model uses a one-state HMM with general background distribution based on the primary structure that will continue to be used.

As described in Section 4.3, the mixed emission probabilities $e_k(p, s)$ are generated based on the primary structure probabilities $e_k(p)$ and secondary structure probabilities

$e_k(s)$. However, with the aim of using only the primary structure instead of the mixed probabilities for regions in the MSA where the primary structure is highly conserved, a threshold can be set. If the highest emission probability in the primary structure is set below the threshold, the mixed probabilities are used. Otherwise, only the probabilities for the primary structure are used:

$$e_k(m) = \begin{cases} e_k(p) & \text{if } \max(e_k(p')) > threshold \\ e_k(p,s) & \text{otherwise} \end{cases} \tag{4.10}$$

The resulting probability $e_k(m)$ will be used in the Viterbi equation for the transition to the match-state V_j^M:

$$V_k^M(k) = \log e_k(m) + \max \begin{cases} V_{k-1}^M(i-1) + \log a_{M_{k-1}M_k} \\ V_{k-1}^I(i-1) + \log a_{I_{k-1}M_k} \\ V_{k-1}^D(i-1) + \log a_{D_{k-1}M_k} \\ V_{k-1}^{FB}(i-1) + \log a_{FBM_k} \end{cases} \tag{4.11}$$

The other Viterbi equations remain unchanged; as for insertion states in (4.12), a background probability $q(p)$ based on the primary structure alphabet will be emitted, and the silent deletion states in (4.13) do not emit any symbol.

$$V_k^I(k) = \log q(p) + \max \begin{cases} V_k^M(i-1) + \log a_{M_kD_k} \\ V_k^I(i-1) + \log a_{I_kD_k} \\ V_k^D(i-1) + \log a_{D_kD_k} \end{cases} \tag{4.12}$$

$$V_k^D(k) = \max \begin{cases} V_{k-1}^M(i-1) + \log a_{M_{k-1}D_k} \\ V_{k-1}^I(i-1) + \log a_{I_{k-1}D_k} \\ V_{k-1}^D(i-1) + \log a_{D_{k-1}D_k} \end{cases} \tag{4.13}$$

The same modification applies to the match state in the forward algorithm in (4.14), which is similar to the Viterbi algorithm. An exception is that instead of using the path with the maximum transition state, it sums up all transition states.

$$
\begin{aligned}
F_k^M(i) = \log e_k(m) + \log[&a_{M_{k-1}M_k} \exp(F_{k-1}^M(i-1)) \\
&+ a_{I_{k-1}M_k} \exp(F_{k-1}^I(i-1)) \\
&+ a_{D_{k-1}M_k} \exp(F_{k-1}^D(i-1)) \\
&+ a_{FBM_k} \exp(F_{k-1}^{FB}(i-1))]
\end{aligned}
\tag{4.14}
$$

5 Tests and Results

In the previous chapter, the implementation of the secondary structure logic in HMModeler was described. In the following chapter, the changes made to HMModeler will be evaluated by testing the different scoring and weighting methods that were used.

All test datasets and the database, as well as the generated scores and plots mentioned in the following chapter, can be found on the attached disk (see Appendix B).

5.1 Test Dataset

The ASTRAL SCOP 1.73 sequence database[1], filtered to retain entries with less than 40 % sequence similarity to each other, was used. This specific version was chosen because it was the latest complete manually curated version available at the time of writing this thesis. The database contains 9,536 sequences with an average sequence length of 178 residues. While most residues in the database are from the 20 common amino acids listed in Table 2.1, 452 are ambiguous residues. Apart from one, these ambiguous residues are all defined as *any of the 20 common amino acid residues* and represented by the letter X. As described in Section 4.2, the logic for sequences with the symbol X is implemented for both the primary and secondary structure. The single additional ambiguous residue, marked by the symbol Z, occurs in the sequence with the SID `d4cpai_`, representing a position with either glutamine (Q) or glutamic acid (E). This sequence was removed from the database, as these rare special cases would lead to an extensive expansion of the implementation, such as an additional emission case for each possible ambiguous amino acid. Furthermore, this would also lead to an increased run-time of the training and scoring algorithms. Another approach to handle ambiguous amino acids is, as described in [14, p. 183], "the 'benefit of the doubt' approach", where ambiguous amino acids are replaced with the most likely candidates for these positions. With the sequence containing the symbol Z removed from the database, the test set finally contained 9,535 sequences. As each entry listed in SCOP is linked to three-dimensional structure information in the PDB, the secondary structure for the test set was determined using DSSP.

A set of 69 MSAs based on the same ASTRAL SCOP 1.73 database was provided by the Department of Molecular Biology at the University of Salzburg[2]. Each MSA is built up from 15 homologous sequences within the same superfamily. For the first

[1] Astral SCOP 1.73: `https://scop.berkeley.edu/astral/ver=1.73` (accessed July 18, 2018).
[2] University of Salzburg: Department of Molecular Biology: `https://biwww.che.sbg.ac.at/` (accessed July 18, 2018).

tests of the scoring and weighting methods of the pHMM, the secondary structure from DSSP was also used for the MSA, as it represents the perfect condition in combination with the used test database assigned the same way.

For the first part of the following section, the MSA representing the SCOP superfamily `c.67.1` will mainly be used in order to show changes in the score types when including secondary structure. The superfamily `c.67.1` is from the class *alpha and beta proteins (a/b)* and represents *pyridoxal phosphate-dependent transferase* proteins that are involved in the biosynthesis of amino acids dependent on *pyridoxal phosphate*. This is the active form of vitamin B6. The MSA with 15 sequences is 501 columns long. The pHMMs trained from the MSA have a length of 384 match-states. More than 50 % of the remaining 116 columns consist of gaps and thus are represented as insertion states in the pHMM. In total, the MSA includes 5,966 residues and 1,549 gap positions. The complete MSA `c.67.1` in its Stockholm format is provided in Appendix C. In the SCOP database that was used, 60 out of the 9,535 sequences are classified as belonging to the superfamily `c.67.1`.

The secondary structure for the 15 sequences of the MSA was assigned by DSSP, using the associated structure files in the PDB. The class *alpha and beta proteins (a/b)*, to which the MSA belongs, is structurally composed of alternating alpha helices and beta sheets in which the beta sheets are mostly parallel to each other. Therefore, the MSA is represented by a balanced number of all three secondary structure elements. Table 5.1 lists the number of occurrences and the percentage distribution of the secondary structure classes assigned by DSSP for the database that was used and the MSA `c.67.1`. The data show that in general, loops are the most common structure with 43.1 %, followed by helices, and finally sheets, which are the least common.

SS-Class		SCOP 1.73		MSA c.67.1
Helix H	35.1 %	594,399	41.9 %	2,497
Sheet E	21.8 %	368,469	16.6 %	991
Loop C	43.1 %	729,494	41.4 %	2,472
Any X	< 0.1 %	451	0.1 %	6
Total	100 %	1,692,815	100 %	5.966

Table 5.1: Distribution of the secondary structure elements H, E, C and X in the SCOP database and the MSA c.67.1.

5.2 Changes in Scores Following the Inclusion of Secondary Structure

Initially, for all nine scores, the original scores obtained using only the primary structure and the new scores produced with secondary structure information were compared. An overview of all 69 scores will be given in Section 5.4. This section focuses on the MSA `c.67.1` to highlight the impact of the secondary structure information on the scores. The method and parameters for the scores calculated using secondary structure is `M2_025_1_3`. The exact configuration will be described in Section 5.4.

The pHMM was trained with the MSA and scored against the database first with primary structure information only and then with both primary and secondary structure information. Table 5.2 lists the average changes between the two methods' scores across all scores from the database. The changes are separated into the superfamily `c.67.1` and the other scores.

Method	Family	Average score-change
Viterbi score	Superfamily:	+72.7664
	Other:	+21.8893
Forward score	Superfamily:	+71.0713
	Other:	+25.1106
Simple null score	Both:	0
Reversed Viterbi	Superfamily:	+53.9577
	Other:	+21.3254
Reversed forward	Superfamily:	+64.7312
	Other:	+24.6887
Simple corrected Viterbi	Superfamily:	+72.7664
	Other:	+21.8893
Simple corrected forward	Superfamily:	+71.0713
	Other:	+25.1106
Reverse corrected Viterbi	Superfamily:	+18.8087
	Other:	+0.5639
Reverse corrected forward	Superfamily:	+6.3401
	Other:	+0.4219

Table 5.2: Average change for MSA c.67.1 scored against the SCOP database using primary structure only to including secondary structure.

The first two scoring methods, the plain Viterbi and forward scores, have an average increase three times higher than that of the other scores. The score for the simple null score is unchanged, as it uses a one-state HMM with a general background distribution. As there is no change in the simple null model, the average change for both simple-corrected scores is the same as the plain scores. The difference between the

superfamily and other scores for the reversed scores is lower than for the plain scores, while the difference for the other scores stays within decimal range. This difference between the plain and reversed scores results in an improvement of the reverse-corrected score. On average, the superfamily scores compared to the others increase by a factor of 33 for the Viterbi method and a factor of 15 for the forward method. Although this increase is an indication of a general improvement, these scores must be investigated in more detail, as there are several uncertain factors, such as the different sample sizes between the 60 superfamily scores compared to the remaining 9,475 other scores. The changes in these scores are investigated in detail in the following figures.

Figure 5.1 shows the change in the scores for the plain and reversed scoring methods generated by HMModeler. A scatterplot is used to visualize the variation for each sequence, with the score using the secondary structure on the horizontal axis. On the vertical axis, the difference between the same score and the original score without secondary structure information is displayed. The scores in the database from sequences in the same superfamily (e.g. the MSA used to build the pHMM) are shown as red circles. All other scores are represented by blue circles. The left plots show the scores related to the Viterbi scoring method and the right plots show the forward method scores.

The upper plots represent the plain Viterbi and forward scores. They show that all scores increase, both those from the target superfamily and all the other scores. However, the rise in the superfamily scores is noticeably higher than the rise in the other scores. Moreover, the higher the original scores from sequences that are not from the superfamily, the smaller the difference that results from using the secondary structure. This implies that using secondary structure information improves the plain scores significantly. However, there are still a considerable number of scores from other superfamilies that are higher than the scores of the tested superfamily. These higher-scored sequences are mainly sequences with a low residue count, as the plain scores are strongly influenced by the sequence length. The reversed scores are used to compensate for the sequence length of the plain scores, by subtracting the reversed scores from the plain scores. It is expected that the difference from the other scores will remain in the range of the plain scores above, whereas the difference for the superfamily scores should be reduced. This behavior can be observed in the two lower plots of Figure 5.1, which show the scores for the reverse Viterbi (left) and the reverse forward (right).

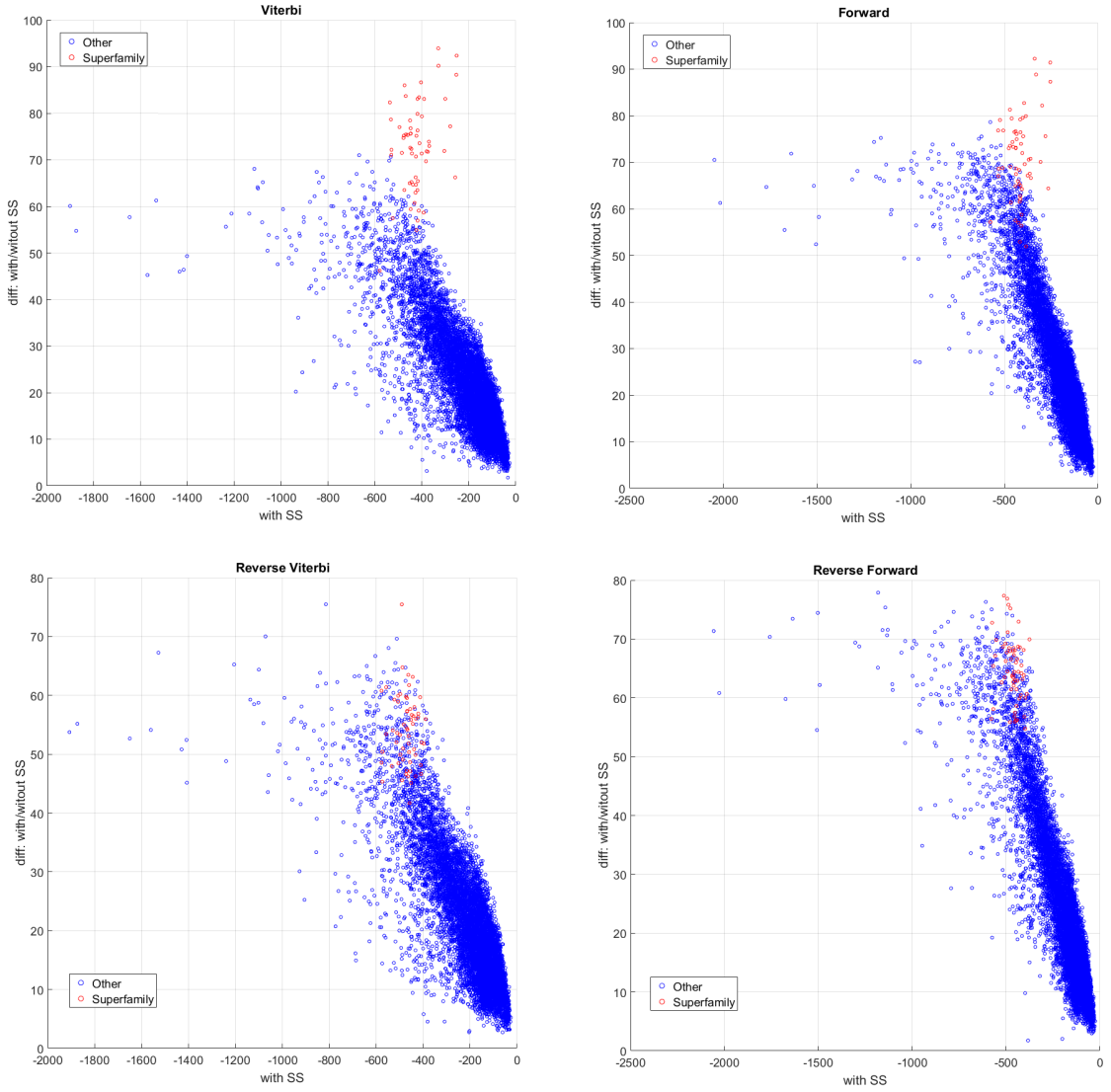

Figure 5.1: Comparison between plain scores with and without secondary structure information as scatterplots, with the score including secondary structure on the horizontal axis and on the vertical axis the difference from the scores with and without secondary structure information.

Figure 5.2 shows the reverse-corrected scores. These are the combinations of the plain Viterbi and forward scores with their respective reversed scores in Figure 5.1, displayed as a scatterplot of the scores with and without secondary structure information. The horizontal axis of the plot shows the original scores obtained using only the primary structure. The vertical axis gives the corresponding scores calculated by including the secondary structure. The red line straight through the origin assists in identifying the change in each score. Scores where the original score equals the new score are positioned on the red line. The farther away a score is from the zero line, the larger the change in the score between the two methods. Scores above the line indicate an

increase; scores below the line mark a decrease in the score when including secondary structure information. A so-called *rug plot* is included in the following figures. The rug plot is a one-dimensional representation of the data points projected along the axis. These rug plots are separated for scores from the superfamily, indicated as red ticks, and the other points, shown as blue ticks.

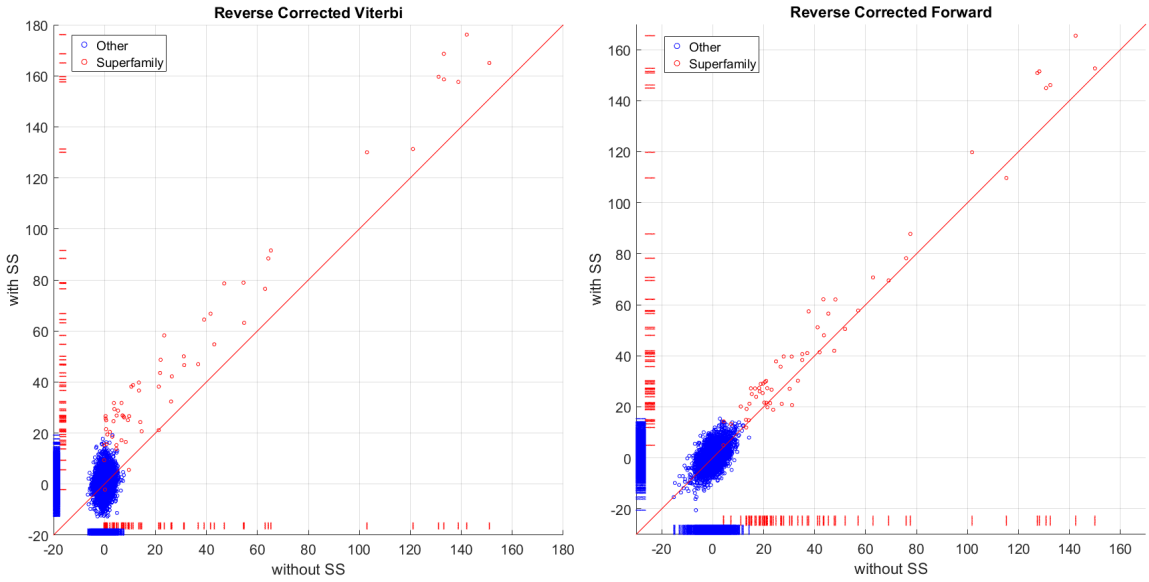

Figure 5.2: Scatterplots for comparing reverse-corrected scores with and without secondary structure information on the horizontal axis and with secondary structure information in the vertical axis.

Noticeable for the reverse-corrected Viterbi score is the dispersal of the other scores over twice the region when using secondary structure information. This difference is interesting due to its average change of below 0.5. However, except for three scores, the superfamily scores increase consistently. In particular, most superfamily scores close to zero and therefore surrounded by other scores increase more than the average and move away from the other scores. In contrast, the dispersal of the other scores for reverse-corrected forward scores remain closer to region of the original score. In addition, the rise of the superfamily scores is lower. While several low scores perform better than the other scores, there are also some decreasing superfamily scores.

Figure 5.3 shows the same plot as Figure 5.2, but with the two simple-corrected scores. These are the simple null model scores subtracted from the plain Viterbi and forward scores. The simple null model uses the same general background distribution; therefore, there is no change when including secondary structure information. As a consequence, the region of all scores for the two simple-corrected scores increases according to the average score change listed in Table 5.2. Moreover, except for a few outliers, the superfamily scores separate clearly from the other scores.

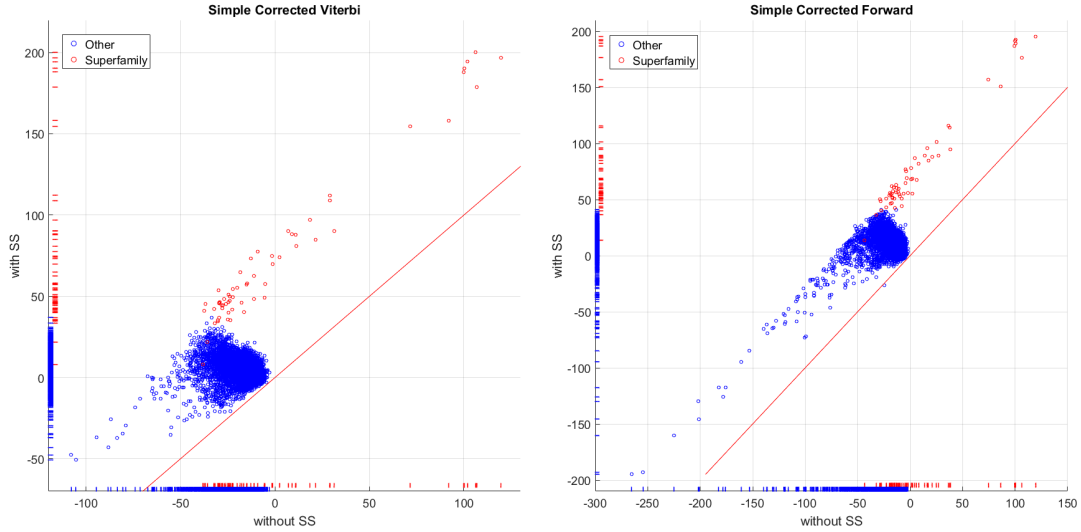

Figure 5.3: Scatterplots for comparing simple-corrected scores with only the primary structure information on the horizontal axis with scores calculated using secondary structure information on the vertical axis.

5.3 Combining Scores

Principal component analysis (PCA) is a statistical approach that uses orthogonal transformation in order to reduce the dimensions of correlated data while preserving as much variability and thus information as possible (see [40]). In consideration of the difference between the score with secondary structure information and those without, a linear combination of both scores might further improve the results. This is done by using PCA to obtain the first principle component of the two-dimensional data containing the score with and without secondary structure information.

Table 5.3 lists the average changes between the scores without secondary structure information and the first principal component of the same score, and the score including secondary structure. The column *SS-Score* lists the average change without applying

Method	Family	SS score	PCA score
Simple corrected Viterbi	Superfamily:	+72.7664	+64.3552
	Other:	+21.8893	+17.2243
Simple corrected forward	Superfamily:	+71.0713	+50.0806
	Other:	+25.1106	+16.0857
Reverse corrected Viterbi	Superfamily:	+18.8087	+27.4886
	Other:	+0.56383	−0.41845
Reverse corrected forward	Superfamily:	+6.34041	+21.3730
	Other:	+0.42186	−0.55353

Table 5.3: Average change for MSA c.67.1 using primary structure only to the PCA optimized score.

PCA (see Table 5.2). Applying PCA on both simple-corrected scores reduces the average change from the scores without secondary structure information by up to 36 %. In contrast, the superfamily scores for both reverse-corrected scores increase significantly, while the other scores decrease.

The impact of using PCA is shown in Figure 5.4 with scatterplots of the PCA-generated scores over the scores with secondary structure. Both simple-corrected scores suffer from high deterioration. The scores in the transition region perform even more poorly, where the superfamily scores should separate from the other scores. On the other hand, the reverse-corrected scores provide a better result. The superfamily scores increase more than the other scores. Some of the reverse-corrected Viterbi scores decline in the transition area, but in the same region there is also a drop in the other scores.

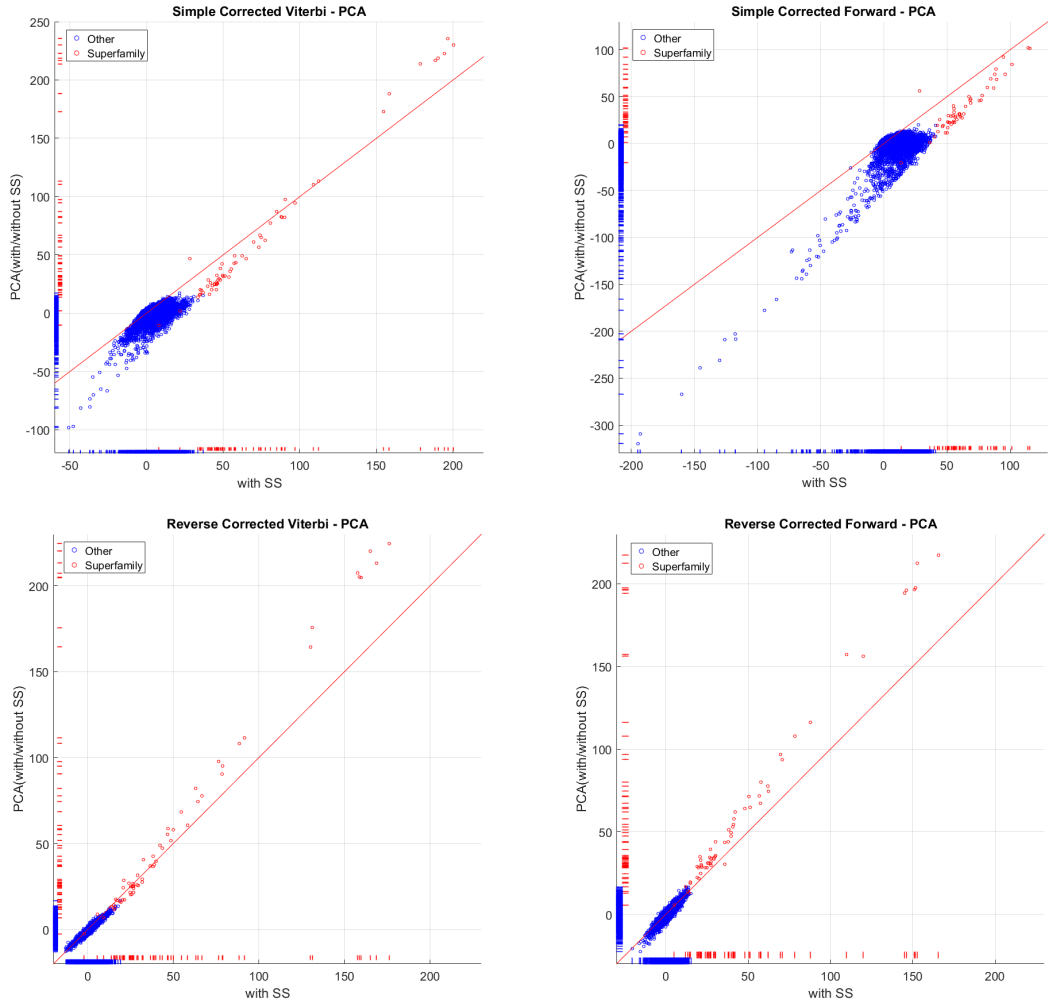

Figure 5.4: Comparison of scores corrected by applying PCA over the secondary structure scores.

From the figures shown previously in this chapter, it is obvious that the scores for both

methods, i.e. the Viterbi algorithm and the forward algorithm, are highly correlated. Therefore, a linear combination of the scores from the same type should reduce the dimensional complexity of the data by uniting the information from each method.

Principal component analysis was applied separately on both simple-corrected and reverse-corrected scores. These new scores are shown in Figure 5.5 on the vertical axis over their associated Viterbi score on the horizontal axis. The mostly linear distribution shows that scores from the same type but different methods mostly share the same information. An exception is the lower section of the scores not related to the `c.67.1` superfamily. The reason for this drop in scores can be found in the distribution of both scores, visualized in the rug plot in Figure 5.3 on the vertical axis. While both superfamily scores are spread around the same region, the other scores from the forward algorithm are spread over four times the area compared to those from the Viterbi algorithm.

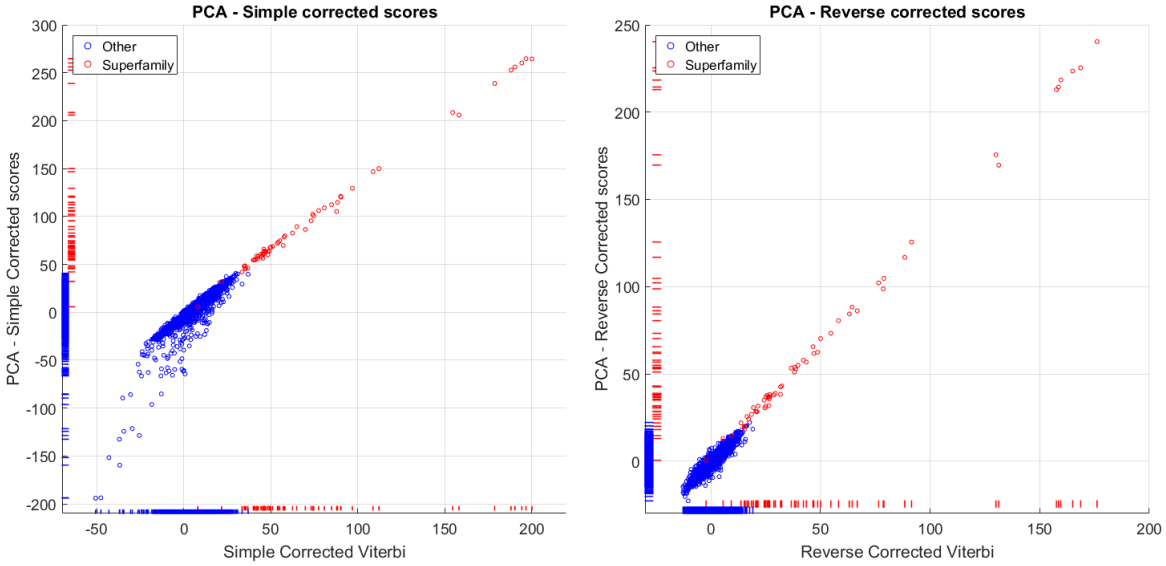

Figure 5.5: PCA applied to the scores over the Viterbi and forward methods.

Up until this point, the focus was mainly on the single scores, mostly from the same type and method, to improve the separation between the scores of the MSA superfamily and the others. The following section analyzes the reverse-corrected scores and the simple-corrected scores from the same method together.

Figure 5.6 shows scatterplots with the simple-corrected scores on the horizontal axis and the reverse-corrected scores on the vertical axis, generated from the original score with the primary structure only on top, and including the secondary structure information

below. For both approaches, Viterbi and forward scoring, the improvements using the secondary structure are obvious, as the separation between the superfamily scores and the others increases. However, this result was already expected based on the figures provided in this section. The same plot for all other variations of the MSA `c.67.1` shown in this section can be found in Appendix D.

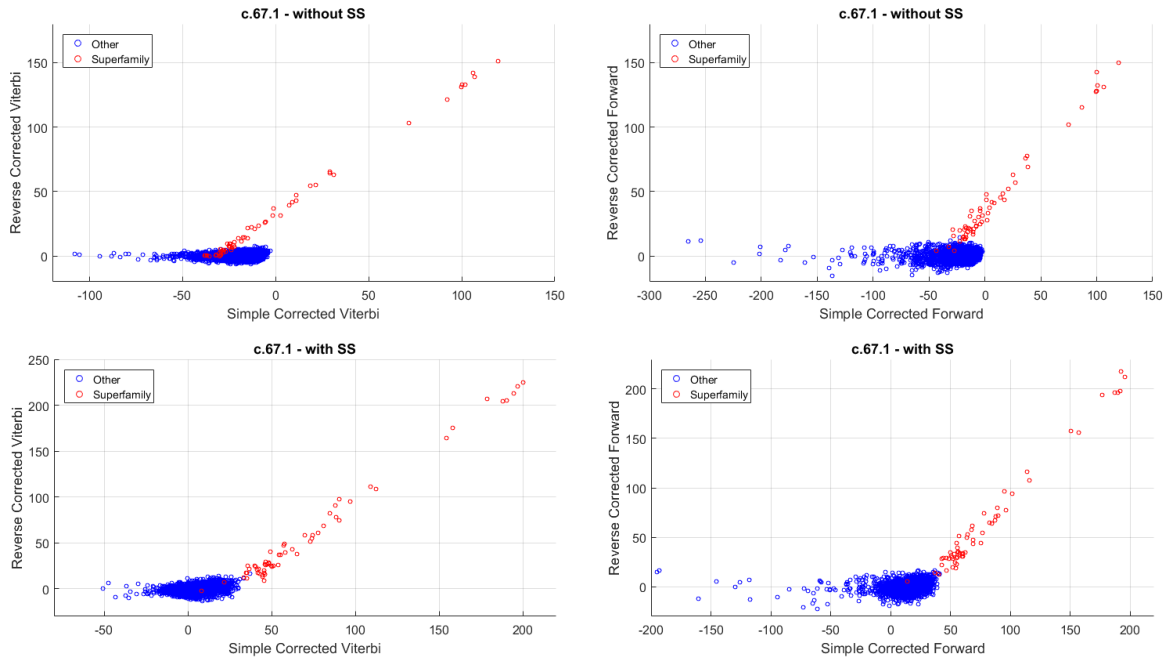

Figure 5.6: Scatterplots of the corrected scores with and without secondary structure information.

This two-dimensional representation with its variety of post-processing methods is used in the following section for the evaluation of all 69 MSAs with varying weighting methods.

5.4 Evaluation of the Different Weighting Methods

The different weighting methods and their parameters described in Section 4.3 will be tested by scoring all 69 MSAs against the SCOP database with different settings. The following five configurations will be explained in more detail.

- `M1_025_1` uses the weighting approach in (4.4), with a pseudo-count of 1 for the secondary structure emission probabilities and a threshold of 0.25 for the highest primary structure emission probability.
- `M1_025_3` uses the same approach as above, except with a pseudo-count of 3.
- `M2_025_1_3` uses the weighting approach in (4.5), also with a threshold of 0.25

and a pseudo-count of 1. The scale factor k is set to 3.

- `M2_025_1_5` uses the same parameters as M2_025_1_3 except with a scaling factor of 5.
- `M3_100_5` uses the Shannon approach in (4.8) with a pseudo-count of 5 for generating the secondary structure emission probabilities. A threshold of 100 % is also employed, resulting in only the mixed probabilities being used for scoring.

For example, with the MSA `c.67.1`, the threshold of 0.25 for the first four approaches results in 272 columns where the highest emission probability is below the threshold; therefore, the mixed probabilities are used, containing both the primary and the secondary structure. For the other 117 columns of the pHMM, only the emission probabilities for the primary structure are used. For the last method, using the Shannon entropy, it was found that the best results are achieved using the mixed probabilities only.

The different methods over the MSAs are compared in MATLAB as listed in 5.1 by generating a Receiver Operating Characteristic (ROC) curve. A ROC curve compares the true positive rate (TPR) against the false positive rate (FPR) for varying thresholds of a classifier and is used to compare the quality of classifiers. The TPR measures the proportion of positives that are correctly classified as such, while the FPR identifies negatives that are wrongly classified as positives (see [41, p. 34–35]).

```
1 mdlSVM = fitcsvm(hmmScores,classes,'Standardize',true);
2 mdlSVM = fitPosterior(mdlSVM);
3 [~,score_svm] = resubPredict(mdlSVM);
4 [X,Y,T,AUC] = perfcurve(hmmScores,score_svm(:,mdlSVM.
     ClassNames),'true');
5 plot(X,Y)
```

Listing 5.1: Matlab implementation for generating the ROC curve.

For each method and MSA, the function *fitcsvm* trains a binary SVM classifier from *hmmScores*, containing the two dimensions for the reverse-corrected and the simple-corrected scores and the *classes*, indicating whether the score relates to the tested superfamily or not. The function *fitPosterior* calculates the posterior probabilities for all scores, allowing *perfcurve* to generate the ROC curve.

Figure 5.7 shows the ROC curve generated for the superfamily `c.67.1` using the methods described above, as well as the original score using only the primary structure listed

as AA. The scores used are the two corrected Viterbi scores, with the reverse-corrected Viterbi as the first principal component from the score with and without secondary structure information. The six scatterplots used for this ROC plot are shown in Appendix E. As an optimal classifier would be a rectangular graph with 100% TPR at 0% FPR, the improvement of the new methods using secondary structure information is obvious.

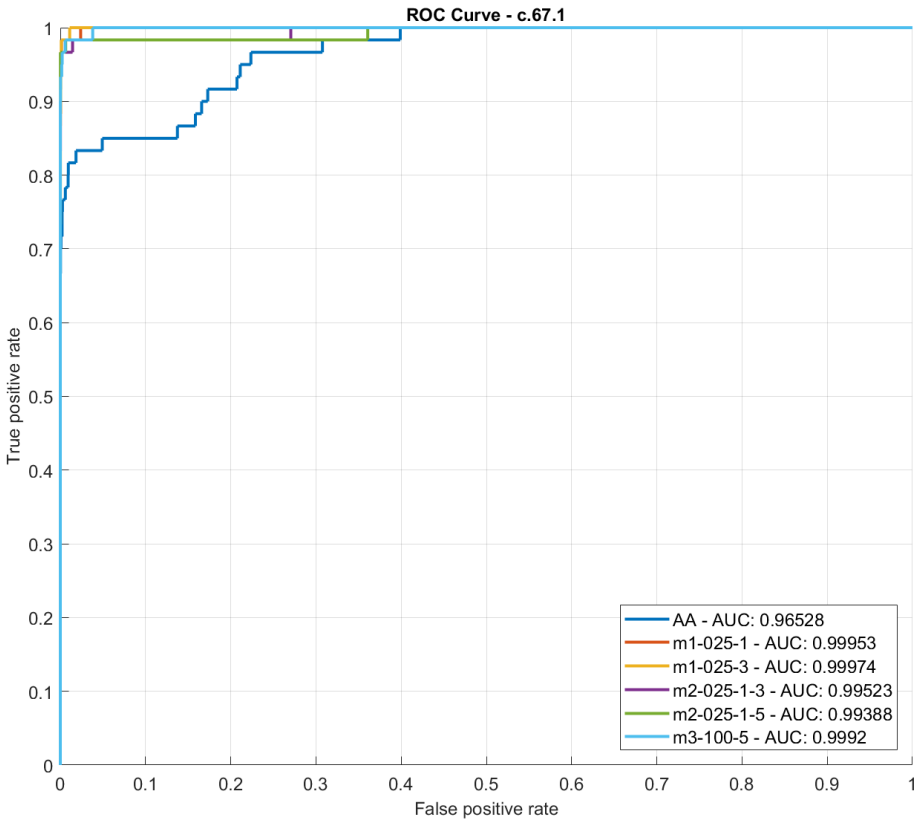

Figure 5.7: ROC curve for the MSA c.67.1.

For comparison of the different methods, the ROC curves can be ranked with the Area Under the Curve (AUC) score. The AUC score combines the whole ROC curve into one number in the range of 0–1 and is calculated by the integral of the curve.

Table 5.5 provides an overview of the selected methods for the MSA. However, the table lists only 56 MSAs, as those with a variation of the AUC score below 1% were filtered out. Column AA lists the AUC score using only the primary structure. The highest AUC score for each MSA is highlighted in green, while all scores below those obtained using the original method are highlighted in red. To summarize the table, it can be said that except for a few scores, the secondary structure information improves the quality of the homology detection. For all MSAs, the original score was improved by

at least one method including the secondary structure. The best results were achieved with the method `M2_025_1_5`, with only two scores below the original score and 29 with the highest rank.

The AUC scores for all MSAs were also compared to the different corrected scores and their combinations discussed in Section 5.2. The results are summarized in Table 5.4, with the first number of each cell providing the number of scores above those of the original method using only the primary structure. The second number counts the MSA with the highest AUC scores compared to all other methods. The table shows that the method `M2_025_1_5` performed best for the most score types, compared to the other methods. The best score type varies across the different methods. For the highest-ranked method, the Viterbi method performs better than the old method for 65 of 69 scores. However, over all methods, both PCA variations for the Viterbi yield a larger improvement than the plain scores. The last column represents the combination of the Viterbi and forward scores. In general, this score type performs worse than the other types, with seven MSAs where the original method has higher scores than all methods that use the secondary structure information.

Score-type / Method	# scores with AUC > AA ; # highest scores for method									
	Viterbi		Forward		PCA-Vit.		PCA-For.		PCA-VitFor	
AA		1		1		2		2		7
M1_025_1	55	13	54	12	60	8	62	12	55	11
M1_025_3	57	16	56	16	59	16	63	11	51	13
M2_025_1_3	49	4	50	4	54	4	58	2	53	5
M2_025_1_5	65	28	64	28	64	33	62	29	56	27
M3_100_5	55	7	46	8	54	6	59	13	55	6

Table 5.4: Performance of different scoring methods for the corrected scores over all 69 MSAs. First column lists number of AUC scores above the score using primary structure only (AA), second column number of highest scores for the method.

The AUC scores for each score type and the method of all 69 MSAs represented in Table 5.4, including their associated scatterplots and ROC curves, can be found on the attached disk (see Appendix B).

MSA	AA	M1_025_1	M1_025_3	M2_025_1_3	M2_025_1_5	M3_100_5
a.1.1	0.981660	0.999138	0.998953	0.992644	0.997789	0.997729
a.118.1	0.658467	0.753259	0.705426	0.746597	0.829309	0.644325
a.118.8	0.955319	0.984794	0.985946	0.986001	0.973027	0.977922
a.121.1	0.949624	0.995763	0.995620	0.996995	0.997438	0.983587
a.25.1	0.841495	0.937005	0.932636	0.941242	0.957592	0.937555
a.26.1	0.874773	0.983614	0.985165	0.969996	0.985939	0.964079
a.39.1	0.952474	0.981174	0.982275	0.959055	0.979119	0.948074
a.4.1	0.943772	0.955631	0.952669	0.969677	0.964489	0.918509
a.4.5	0.855347	0.926701	0.925346	0.908423	0.929703	0.916538
b.1.18	0.792153	0.817158	0.807228	0.815886	0.838580	0.802668
b.1.2	0.969961	0.995894	0.995995	0.981565	0.993793	0.997065
b.121.4	0.773631	0.967558	0.966382	0.866694	0.992417	0.953065
b.122.1	0.848138	0.845888	0.848104	0.845576	0.851568	0.779088
b.18.1	0.761240	0.930075	0.936217	0.817446	0.866923	0.929797
b.29.1	0.696135	0.960836	0.951235	0.849526	0.983435	0.940170
b.40.4	0.562231	0.645208	0.670961	0.569530	0.338704	0.567449
b.55.1	0.911226	0.979285	0.980650	0.929043	0.959400	0.957248
b.6.1	0.908563	0.960190	0.966697	0.937017	0.955984	0.957871
b.60.1	0.930032	0.984104	0.984354	0.970134	0.990348	0.978497
b.82.1	0.825039	0.938847	0.919836	0.866187	0.955784	0.912106
c.1.10	0.873501	0.874392	0.869691	0.872736	0.940628	0.872383
c.1.8	0.771208	0.928533	0.936710	0.773840	0.930788	0.945124
c.1.9	0.776061	0.984452	0.987866	0.849393	0.953804	0.968378
c.14.1	0.922587	0.973523	0.978553	0.895175	0.937373	0.943083
c.2.1	0.832922	0.877780	0.878903	0.830230	0.874660	0.849042
c.23.16	0.953418	0.984812	0.985526	0.946543	0.970777	0.979098
c.23.1	0.982697	0.997496	0.996486	0.996668	0.998236	0.994347
c.26.1	0.948562	0.976231	0.976961	0.960769	0.965366	0.976797
c.26.2	0.762557	0.851986	0.846154	0.772628	0.903711	0.878560
c.3.1	0.873175	0.889874	0.884938	0.874126	0.900207	0.855220
c.37.1	0.566633	0.757238	0.705066	0.633610	0.795116	0.721068
c.47.1	0.758197	0.851974	0.840546	0.832753	0.889316	0.819158
c.52.1	0.715274	0.687236	0.689456	0.728182	0.729141	0.689634
c.55.1	0.841824	0.890524	0.888006	0.872482	0.879920	0.896231
c.55.3	0.661511	0.724718	0.730029	0.745105	0.827794	0.708841
c.56.5	0.932965	0.975813	0.976139	0.931557	0.988531	0.970057
c.66.1	0.691936	0.792645	0.805687	0.762319	0.828899	0.826029
c.67.1	0.965279	0.999527	0.999741	0.995230	0.993875	0.999198
c.68.1	0.840345	0.954689	0.955924	0.896832	0.953844	0.942945
c.69.1	0.769510	0.951119	0.954956	0.886774	0.945654	0.938550
c.94.1	0.786186	0.885272	0.891178	0.822143	0.910847	0.838724
d.108.1	0.842174	0.985350	0.985037	0.890825	0.959523	0.956787
d.129.3	0.770835	0.906587	0.886342	0.779503	0.887589	0.929245
d.14.1	0.800006	0.820208	0.828362	0.813035	0.884108	0.847360
d.144.1	0.970321	0.980958	0.968978	0.968810	0.985398	0.979926
d.15.1	0.820850	0.841835	0.836335	0.938895	0.945733	0.819833
d.153.1	0.904484	0.963783	0.964523	0.911143	0.941617	0.949000
d.169.1	0.914845	0.984051	0.986197	0.949123	0.987850	0.936318
d.17.4	0.917448	0.997716	0.997654	0.949466	0.988231	0.996560
d.3.1	0.781971	0.922795	0.913319	0.795755	0.889231	0.920976
d.32.1	0.943212	0.988796	0.991518	0.974388	0.983423	0.983284
d.38.1	0.927538	0.975661	0.971830	0.955195	0.986411	0.971640
d.58.4	0.923971	0.939588	0.928788	0.936421	0.951176	0.935353
d.81.1	0.803026	0.821645	0.809742	0.798986	0.823577	0.824889
d.92.1	0.747920	0.791525	0.794027	0.803735	0.833569	0.819230
g.39.1	0.984596	0.982231	0.982742	0.988259	0.983898	0.976715

Table 5.5: AUC scores for the different scoring methods from the corrected Viterbi scores postprocessed with PCA compared to the original method in column AA using primary structure only. The highest score for each MSA is marked green. Red scores are below the original method.

5.5 Scoring with Predicted Secondary Structure

The previous tests were all done with secondary structure information gathered via DSSP from the three-dimensional structure information. Homology prediction is a procedure commonly used for protein sequences for which the three-dimensional structure is unknown. Furthermore, gaining knowledge about the homology is used to improve methods for determining the proteins tertiary structure.

The following scatterplots in Figure 5.8 compare the results from the different methods of secondary structure determination on the MSA `c.67.1`. All plots show the simple-corrected scores on the horizontal axis and the reverse-corrected scores on the

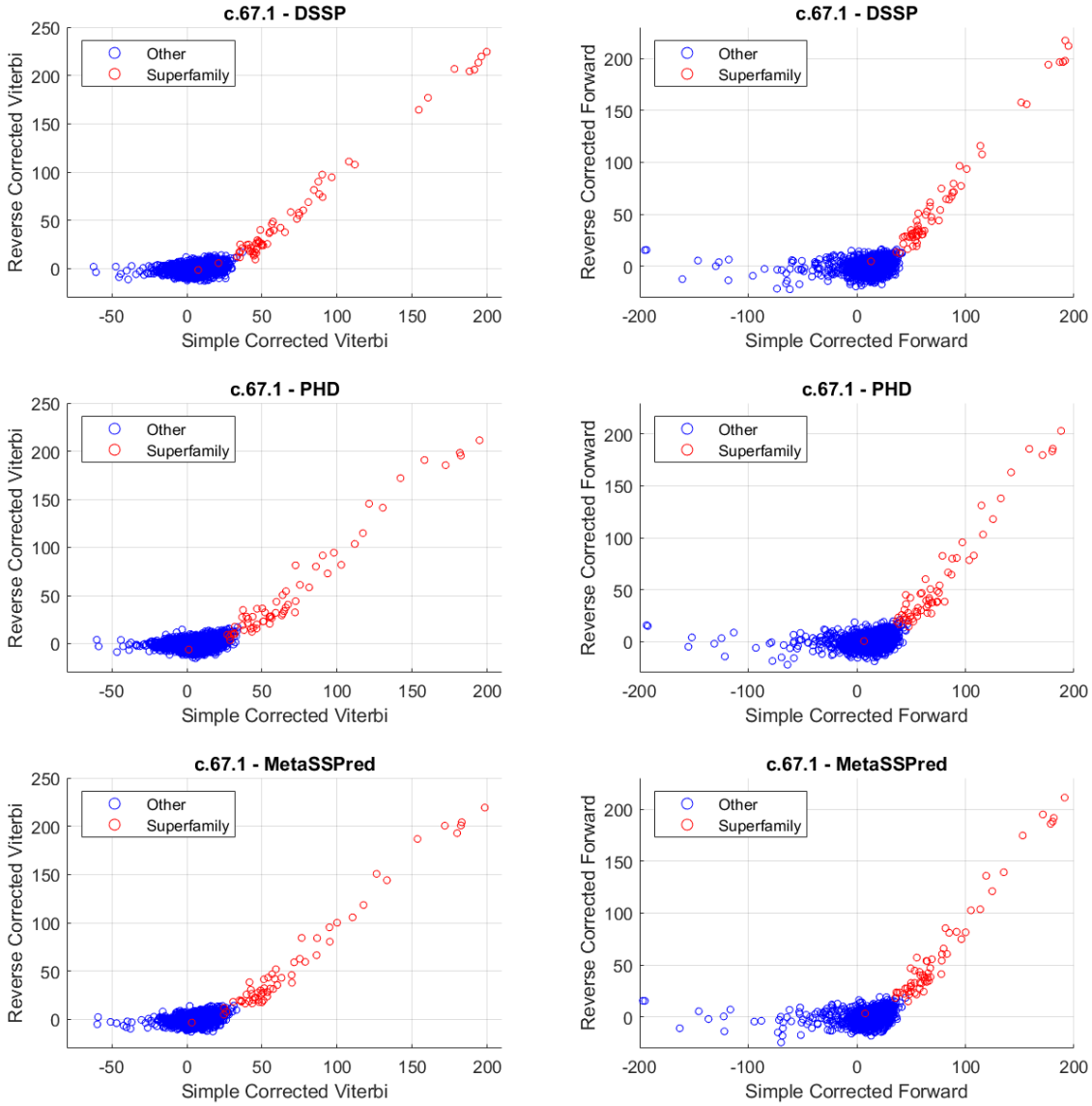

Figure 5.8: Comparison of the different secondary structure estimation methods on MSA c.67.1.

vertical axis. The first row is the reference plot with the secondary structure obtained from DSSP. The structures for the second and third rows are predicted with PHD and MetaSSPred, respectively. As in the previous sections, the method `M2_025_1_3` is used for training the pHMM and the test database is scored against it. For the reverse-corrected scores, the PCA between the scores with and without secondary structure is used. Both the Viterbi scores on the left and the forward scores on the right show a slight decline. Even though the scores in the transition region between the superfamily and the other classes move closer to one another, there is still an outstanding improvement over the scores without secondary structure. This is also apparent in the ROC curve in Figure 5.9 based on the forward scores from Figure 5.8 and their associated primary-structure-only scores.

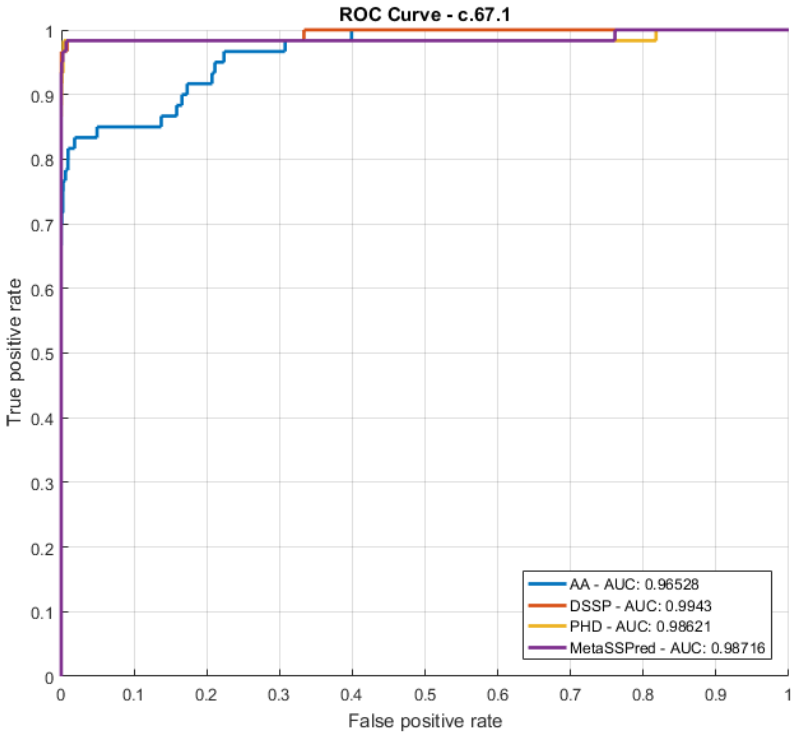

Figure 5.9: ROC curves comparing different secondary structure determination methods.

6 Conclusion

The following chapter presents a brief summary of the thesis' findings and subsequently provides an outlook of potential future work based on these results.

6.1 Summary

This thesis has presented methods for protein homology detection with pHMM using secondary structure information. The key tasks were to determine secondary structure information from primary or tertiary structure, build a pHMM using both primary and secondary structure information, and extend the scoring methods to make use of secondary structure information. These steps were implemented in the software package HMModeler.

Chapter 2 introduced the biological background of proteins. The discussion included the fundamentals of proteins, their composition by chains of amino acid residues, and their structural levels. Furthermore, relevant databases containing different levels of protein information were introduced. Finally, the statistical model pHMM for analyzing proteins was presented.

Chapter 3 focused on methods for determining the different structural levels of proteins. This covered protein sequencing and experimental methods for determining the three-dimensional structure on an atomic level. Subsequently, the secondary structure annotation method DSSP, which relies on the tertiary structure, was explained. A special focus was placed on secondary structure prediction methods. These include the two neural-network-based methods PHD and SPINE-X, and MetaSSPred, which improves the accuracy of SPINE-X using SVMs.

The fourth chapter explained the implementation in the software package HMModeler. In particular, the changes made to the workflow to make use of secondary structure information were discussed. This procedure starts with processing the input data by gathering the secondary structure from the three-dimensional structure using DSSP, if available, or otherwise through prediction using the primary structure. Subsequently, the methods for training the pHMM were covered. This discussion presented the additional emission frequencies for the secondary structure and methods for mixing both the primary and secondary structure frequencies. Finally, the Viterbi and forward algorithms used for aligning sequences against the pHMM were extended by an additional threshold, which determines whether the mixed probabilities or just the primary structure are used.

Finally, in Chapter 5, the implementation was tested using the ASTRAL SCOP database and a set of 69 MSAs representing different superfamilies. The MSAs were trained against the database using the different methods with varying parameters for generating the emission frequencies. The improvements of both scoring algorithms for the different score types were analyzed and compared with the scores from the old implementation. In addition, methods for further processing the scores using PCA were tested. This showed that a linear combination of the reverse-corrected scores from the scores with and without secondary structure information can further increase the distinguishability of the target family against the remaining database. Finally, the weighting methods and parameters that were discussed were compared. For this task, MATLAB was used to train an SVM classifier with varying thresholds for generating ROC curves for each MSA score. Compared to the old implementation using only the primary structure, most scores were improved by including secondary structure information. The best results were achieved with the weighting method in (4.5) with a scale factor of 5.

6.2 Outlook

This thesis has presented improvements for the implementation of using secondary structure information for superfamily classification. For current tests, the manually curated SCOP database has been used. Future developments should include more recent databases, such as the automatically curated SCOPe.

Certain parts of the code, such as the scoring algorithm, are already implemented in the fast low-level language C++. Other time-consuming calculations, like the processing of the database or the handling of secondary structure information, should also be moved from Python to C++. Furthermore, the sequential processing of each sequence in the database can be improved by parallel computing over multiple CPU or GPU cores.

Bibliography

[1] D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, 6th ed. W.H. Freeman and Company, 2013.

[2] M. M. Gromiha, *Protein Bioinformatics: From Sequence to Function.* Academic Press, 2010.

[3] J. M. Berg, J. L. Tymoczko, L. Stryer, and N. D. Clarke, *Biochemistry*, 5th ed. W.H. Freeman, 2002.

[4] IUPAC-IUB Joint Commission on Biochemical Nomenclature, "Nomenclature and symbolism for amino acids and peptides. recommendations 1983," *European Journal of Biochemistry*, vol. 138, no. 1, pp. 9–37, 1984.

[5] J. Meiler, A. Zeidler, F. Schmäschke, and M. Müller, "Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks," *Journal of Molecular Modeling*, vol. 7, no. 9, pp. 360–369, 2001.

[6] K. A. Dill and J. L. MacCallum, "The protein-folding problem, 50 years on," *Science (New York, N.Y.)*, vol. 338, no. 6110, pp. 1042–1046, 2012.

[7] M. Y. Galperin, X. M. Fernández-Suárez, and D. J. Rigden, "The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes," *Nucleic acids research*, vol. 45, no. D1, pp. D1–D11, 2017.

[8] The UniProt Consortium, "UniProt: The universal protein knowledgebase," *Nucleic acids research*, vol. 45, no. D1, pp. D158–D169, 2017.

[9] F. C. Bernstein *et al.*, "The protein data bank. a computer-based archival file for macromolecular structures," *European Journal of Biochemistry*, vol. 80, no. 2, pp. 319–324, 1977.

[10] S. K. Burley *et al.*, "RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education," *Protein science : a publication of the Protein Society*, vol. 27, no. 1, pp. 316–330, 2018.

[11] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of molecular biology*, vol. 247, no. 4, pp. 536–540, 1995.

[12] N. K. Fox, S. E. Brenner, and J.-M. Chandonia, "SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures," *Nucleic acids research*, vol. 42, no. Database issue, pp. D304–D313, 2014.

[13] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[14] P. Baldi and S. Brunak, *Bioinformatics: The machine learning approach*, 2nd ed. MIT Press, 2001.

[15] V. H. Wysocki, K. A. Resing, Q. Zhang, and G. Cheng, "Mass spectrometry of peptides and proteins," *Methods (San Diego, Calif.)*, vol. 35, no. 3, pp. 211–222, 2005.

[16] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips, "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis," *Nature*, vol. 181, no. 4610, pp. 662–666, 1958.

[17] A. McPherson and J. A. Gavira, "Introduction to protein crystallization," *Acta crystallographica. Section F, Structural biology communications*, vol. 70, no. 1, pp. 2–20, 2014.

[18] A. McPherson and L. J. DeLucas, "Microgravity protein crystallization," *Npj Microgravity*, vol. 1, no. 15010, 2015.

[19] J. C. Edwards, "Principles of NMR," *Process NMR Associates LLC, 87A Sand Pit Rd, Danbury CT*, vol. 6810, 2009.

[20] R. Milo and R. Phillips, *Cell biology by the numbers*. Garland Science, 2016.

[21] T. Z. Sen, R. L. Jernigan, J. Garnier, and A. Kloczkowski, "GOR V server for protein secondary structure prediction," *Bioinformatics*, vol. 21, no. 11, pp. 2787–2788, 2005.

[22] B. Rost, "Rising Accuracy of Protein Secondary Structure Prediction," *Protein structure determination, analysis, and modeling for drug discovery*, pp. 207–249, 2003.

[23] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[24] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment," *Proteins*, vol. 23, no. 4, pp. 566–579, 1995.

[25] F. M. Richards and C. E. Kundrot, "Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure," *Proteins*, vol. 3, no. 2, pp. 71–84, 1988.

[26] J. A. Cuff and G. J. Barton, "Evaluation and improvement of multiple sequence methods for protein secondary structure prediction," *Proteins*, vol. 34, no. 4, pp. 508–519, 1999.

[27] H. Wilman, "DSSP | Oxford Protein Informatics Group," 2014. [Online]. Available: https://www.blopig.com/blog/2014/08/dssp/ [Accessed: 15- Aug- 2018].

[28] B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70% accuracy," *Journal of molecular biology*, vol. 232, no. 2, pp. 584–599, 1993.

[29] B. Rost, G. Yachdav, and J. Liu, "The PredictProtein server," *Nucleic acids research*, vol. 32, no. Web Server issue, pp. W321–W326, 2004.

[30] E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, and Y. Zhou, "SPINE X: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles," *Journal of computational chemistry*, vol. 33, no. 3, pp. 259–267, 2012.

[31] S. F. Altschul and E. V. Koonin, "Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases," *Trends in Biochemical Sciences*, vol. 23, no. 11, pp. 444–447, 1998.

[32] Md N. Islam, S. Iqbal, A. R. Katebi, and Md T. Hoque, "A balanced secondary structure predictor," *Journal of theoretical biology*, vol. 389, pp. 60–71, 2016.

[33] S. Iqbal and M. T. Hoque, "Dispredict: A predictor of disordered protein using optimized rbf kernel," *PloS one*, vol. 10, no. 10, p. e0141551, 2015.

[34] S. Iqbal, A. Mishra, and M. T. Hoque, "Improved prediction of accessible surface area results in efficient energy function application," *Journal of theoretical biology*, vol. 380, pp. 380–391, 2015.

[35] R. Graf, M. Aigner, M. Lechner, D. Schroffner, P. Lackner, and S. Wegenkittl, Eds., *HMModeler, a new approach for designing profile HMMs for protein families.* Proceedings of the 6th European Conference on Computer Systems, 2011.

[36] M. Oberkirchner, "Softwareentwicklung einer Client-Server-Architektur für Scientific Computing am Beispiel des Bioinformatiktools HMModeler," Master's thesis, Salzburg University of Applied Sciences, Salzburg, 2014.

[37] U. Mayer, "Secondary Structure Profile HMMs," Master's thesis, Salzburg University of Applied Sciences, Salzburg, 2014.

[38] E. Sonnhammer, "Stockholm format." [Online]. Available: http://sonnhammer.sbc.su.se/Stockholm.html [Accessed: 15- Aug- 2018].

[39] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[40] G. H. Dunteman, *Principal components analysis*, ser. Quantitative Applications in the Social Sciences. Newbury Park, Calif.: Sage Publ, 1989, vol. 69.

[41] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition).* John Wiley & Sons, 2000.

A Physical Properties of Amino Acids

Name	Ξ^a	α^b	$\upsilon_v{}^c$	π^d	I^e	α^f	β^g
ALA	1.28	0.05	1.00	0.31	6.11	0.42	0.23
GLY	0.00	0.00	0.00	0.00	6.07	0.13	0.15
VAL	3.67	0.14	3.00	1.22	6.02	0.27	0.49
LEU	2.59	0.19	4.00	1.70	6.04	0.39	0.31
ILE	4.19	0.19	4.00	1.80	6.04	0.30	0.45
PHE	2.94	0.29	5.89	1.79	5.67	0.30	0.38
TYR	2.94	0.30	6.47	0.96	5.66	0.25	0.41
TRP	3.21	0.41	8.08	2.25	5.94	0.32	0.42
THR	3.03	0.11	2.60	0.26	5.60	0.21	0.36
SER	1.31	0.06	1.60	-0.04	5.70	0.20	0.28
ARG	2.34	0.29	6.13	-1.01	10.74	0.36	0.25
LYS	1.89	0.22	4.77	-0.99	9.99	0.32	0.27
HIS	2.99	0.23	4.66	0.13	7.69	0.27	0.30
ASP	1.60	0.11	2.78	-0.77	2.95	0.25	0.20
GLU	1.56	0.15	3.78	-0.64	3.09	0.42	0.21
ASN	1.60	0.13	2.95	-0.60	6.52	0.21	0.22
GLN	1.56	0.18	3.95	-0.22	5.65	0.36	0.25
MET	2.35	0.22	4.43	1.23	5.71	0.38	0.32
PRO	2.67	0.00	2.72	0.72	6.80	0.13	0.34
CYS	1.77	0.13	2.43	1.54	6.35	0.17	0.41

[a] Steric parameter (graph shape index)
[b] Polarizability
[c] Volume (normalized van der Waals volume)
[d] Hydrophobicity
[e] Isoelectric point
[f] Helix probability
[g] Sheet probability

Figure A.1: Physical properties of amino acids [5].

B Disk

Content of the Disk:

```
|---Data
    |---input
    |   |---database
    |   |---MSA
    |---roc
    |   |---ROC_Forward
    |   |   |---jpg
    |   |   |---ROC
    |   |---ROC_FOR_PCA
    |   |   |---jpg
    |   |   |---ROC
    |   |---ROC_Viterbi
    |   |   |---jpg
    |   |   |---ROC
    |   |---ROC_VitForPCA
    |   |   |---jpg
    |   |   |---ROC
    |   |---ROC_VIT_PCA
    |   |   |---jpg
    |   |   |---ROC
    |   |---RocScores
    |---scores
        |---AA
        |---DSSP_m1_025_1
        |---DSSP_m1_025_3
        |---DSSP_m2_025_1_3
        |---DSSP_m2_025_1_5
        |---DSSP_m3_025_1
        |---DSSP_m3_025_3
        |---DSSP_m3_100_1
        |---DSSP_m3_100_5
        |---SSP_msa_c.67.1
```

Listing B.1: Contents of the Disc

C MSA c.67.1

```
# STOCKHOLM 1.0

d1vefa1 ...WRALLEAEKTLDSG........VYNKHDLLIVRGQGARVWDAEGNEYIDCVGGYGVANLGHGNPEVVEAVKRQAET.
d2epja_ GEKSRMLFERTKELFPGGVNSPVRAAVKPYPFYVKRGEGAYLYTVDGARIVDLVLAYGPLILGHKHPRVLEAVEEALARG
d1zoda_ .......LNDDATFWRNARHHLVRYGGTFEPMIIERAKGSFVYDADGRAILDFTSGQMSAVLGHCHPEIVSVIGEYAGK.
d3doda_ ...HDLIEKSKKHLWLP...FTQMKDYDENPLIIESGTGIKVKDINGKEYYDGFSSVWLNVHGHRKKELDDAIKKQLGK.
d1fg7a_ ..............................TVTITDLARENVRNLTPYQSARRLGGNGDVWLNAN..EYPTAVEFQLTQQ
d2f8ja_ ..............................HMNPLDLIAKRAYPYETEKRDKTYLALNENPFPFP.EDLVDEVFRRLNSD
d1lc5a_ ..............................LFNTAHGGNIREPATVLGISPDQLLDFSANINPLG.MPVSVKRALIDNLD
d1bw0a_ ........WDVSMSNHAGLVFNPIRTVSDNAKPSPSPKPIIKLSVGDPTLDKNLLTSAAQIKKLKEAIDSQECNGYFPTV
d2gb3a1 ......................FSDRVLLTEESPIRKLVPFAEMAKKRGVRIHHLNIGQPDLKTPEVFFERIYENKPE
d2dkja_ ...................KRDEALFELIALEEKRQREGLELIASENFVSKQVREAVGSVLTNKYAEGYPGARYYGGCE
d2ez1a_ ................NYPAEPFRIKSVETVSMIPRDERLKKMQEAGYNTFLLNSKDIYIDLLTDSGTNAMSDKQWAGMMM
d2rfva_ ..............................SDXRTYGFNTQIVHAGQQPDPSTGALSTPIFQTSTFVFDSAEQGAARFG
d2c81a_ ............................WPEWPQHSDRTRRKIEEVFQSNRWAISGYWTGEESMERKFAKAFADFNGVPY
d3k40a_ ............MEAPEFKDFAKTMVDFIAEYLENIRERRVLPEVKPGYLKPLIPDAAPEKPEKWQDVMQDIERVIMPG
d1m32a_ ............................YLLLTPGPLTTSRTVKEAMLFDSCTWDDDYNIGVVEQIRQQLTALATASEG

d1vefa1 .LMAMPQ..TLPTPMRGEFYRTLT..AILPPELNRVFPVNSGTEANEAALKFARAHTGRKKFVAAMRGFSGRTMGSLSVT
d2epja_ WLYGAPG..EAEVLLAEKILG.......YVKRGGMIRFVNSGTEATMTAIRLARGYTGRDLILKFDGCYHGSHDAVLVAA
d1zoda_ .LDHLFS..EMLSRPVVDLATRLA..NITPPGLDRALLLSTGAESNEAAIRMAKLVTGKYEIVGFAQSWHGMTGAAASAT
d3doda_ IAHSTLL..GMTNVPATQLAETLI..DISPKKLTRVFYSDSGAEAMEIALKMAFQYWKNIGKPEKQKFIAMKSYKAPIPY
d1fg7a_ TLNRYPE..CQPKAVIENYAQ.......YAGVKPEQVLVSRGADEGIELLIRAFCEPGKDAILYCPPTYGMYSVSAETIG
d2f8ja_ ALRIYYD..SPDEELIEKILSYLD..TDFLSKNN..VSVGNGADEIIYVMMLMFDRS.....VFFPPTYSCYRIFAKAVG
d1lc5a_ CIERYPD..ADYFHLHQALAR.......HHQVPASWILAGNGETESIFTVASGLKPR...RAMIVTPGFAEYGRALAQSG
d1bw0a_ GSPEARE..AVATWWRNSFVHKEE..LKSTIVKDNVVLCSGGSHGILMAITAICDAG..DYALVPQPGFPHYETVCKAYG
d2gb3a1 VVYYSHS..AGIWELREAFASYYKRRQRVDVKPENVLVTNGGSEAILFSFAVIANPG..DEILVLEPFYANYNAFAKIAG
d2dkja_ VIDRVES..LAIERAKALFGAAWAN...........VQPHSGSQANMAVYMALMEPG..DTLMGMDLAAGGHLTHGSRVN
d2ez1a_ GDEAYAG..SENFYHLERTVQELFG.....FKHIVPTHQGRGAENLLSQLAIKPGQYVAGNMYFTTTRYHQEKNGAVFVD
d2rfva_ YIYTRLG..NPTTDALEKKLAVLE.......RGEAGLATASGISAITTTLLTLCQQG..DHIVSASAIYGCTHAFLSHSM
d2c81a_ CVPTTSG..STALMLALEALGIGEG....DEVIVPSLTWIATATAVLNVNALPVFVDVEADTYCIDPQLIKSAITDKTKA
d3k40a_ VTHWHSPKFHAYFPTANSYPAIVADMLSGAIACIGFTWIASPACTELEVVMMDWLGKMLELPAEFLACSGGKGGGVIQGT
d1m32a_ YTSVLLQ..GSGSYAVEAVLGSALG......PQDKVLIVSNGAYGARMVEMAG........LMGIAHHAYDCGEVARPDV

d1vefa1 .WEPKYREPFLPLVEPVEFIPYNDVEALKR.................AVDEE...TAAVILEPVQGEGGVRPATPEFLRAA
d2epja_ GGVPTSAGVPEAVARLTLVTPYNDVEALER.................VFAEYGDRIAGVIVEPVIANAGVIPPRREFLAAL
d1zoda_ YSAGRKGVGPAAVGSFAIPAPFTYRPRFERNGAYDYLAELDYAFDLIDRQSSGNLAAFIAEPILSSGGIIELPDGYMAAL
d3doda_ VYRSESGDPDECRDQCLRELAQLLEEHHEE.......................IAALSIESMVQGASGMIVMPEGYLAGV
d1fg7a_ .....VECRTVPTLDNWQLDLQGISDKLDG.....................VKVVYVCSPNNPTGQLINPQD..FRTL
d2f8ja_ .....AKFLEVPLTKDLRIPEVNVGE.........................GDVVFIPNPNNPTGHVFEREE......
d1lc5a_ ...CEIRRWSLREADGWQLTDAILEALTPD.....................LDCLFLCTPNNPTG..LLPERPLLQAI
d1bw0a_ ...IGMHFYNCRPENDWEADLDEIRRLKDD.....................KTKLLIVTNPSNPCGSNFSRKH..VEDI
d2gb3a1 ...VKLIPVTRRMEEGFAIPQNLESFINER.....................TKGIVLSNPCNPTGVVYGKDE..MRYL
d2dkja_ ....FSGKLYKVVSYGVRPDTELIDLEEVR.....................RLALEHRPKVIVAGASAYPRFWDFKAF
d2ez1a_ IVRDEAHDAGLNIAFKGDIDLKKLQKLIDE...................KGAENIAYICLAVTVNLAGGQPVSMANMRAV
d2rfva_ P....KFGINVRFVDAAKPEEIRAAMRPET.....................KVVYIETPANPTLSLVDIET.....V
d2c81a_ IIPVHLFGSMANMDEINEIAQEHNLFVIED.....................CAQSHGSVWNNQRAGTIGDIGAFSCQQG
d3k40a_ ASESTLVALLGAKAKKLKEVKELHPEWDEHTILG.................KLVGYCSDQAHSSVERAGLLGGVKLRSV
d1m32a_ QAIDAILNADPTISHIAMVHSETTTGMLNP.....................IDEVGALAHRYGKTYIVDAMSSFGGIP

d1vefa1 REITQEKGALLILDEIQTGMG......RTGKRFAFEHFGIVPDILTLAKA..LGGG.VPLGVAVMREEVARSMPKGG...
d2epja_ QRLSRESGALLILDEVVTGF.......RLGLEGAQGYFNIEGDIIVLGKI..IGGG.FPVGAVAGSREVMSLLTPQGK..
d1zoda_ KRKCEARGMLLILDEAQTGVG......RTGTMFACQRDGVTPDILTLSKT..LGAG.LPLAAIVTSAAIEERAHELG...
d3doda_ RELCTTYDVLMIVDEVATGFG......RTGKMFACEHENVQPDLMAAGKG..ITGGYLPIAVTFATEDIYKAFYDDYENL
d1fg7a_ LELTRGKAIVVADEAYIEFCP......QASLAGWLAEYPHLAILRTLSKA..FALAGLRCGFTLANEEVINLL.......
d2f8ja_ IERILKTGAFVALDEAYYEFH......GESYVDFLKKYENLAVIRTFSKA..FSLAAQRVGYVVASEKFIDAYN......
d1lc5a_ ADRCKSLNINLILDEAFIDFIPH....ETGFIPALKDNPHIWVLRSLTKF..YAIPGLRLGYLVNSDDAAMARMR.....
d1bw0a_ VRLAEELRLPLFSDEIYAGMVFKGKDPNATFTSVADFETTVPRVILGGTAXNLVVPGWRLGWLLYVDPHGNGPSFLEG..
d2gb3a1 VEIAERHGLFLIVDEVYSEIVFR.....GEFASALSIESDKVVVIDSVSX.KFSACGARVGCLITRNEELISHAMKLA..
d2dkja_ REIADEVGAYLVVDMAHFAGLVAAGLHPNPLPYAHVVTSTTHKTLRGPRGGLILSNDPELGKRIDKLIFPGIQGGP....
d2ez1a_ RELTEAHGIKVFYDATRCVENAYFIKEQEQGFENKSIAEIVHEMFSYADG...CTMSGKXDCLVNIGGFLCMNDDEMFSS
d2rfva_ AGIAHQQGALLVVDNTFMSPY.....CQQPLQLGADIVVHSVTXYINGHGDVIGGIIVGKQEFIDQARFVGLKDITGGCM
d2c81a_ KVLTAGEGGIIVTKNPRLFELIQQLRADSRVYCDDSSELMHGDMQLVKKG.DIQGSNYCLSEFQSAILLDQLQELDDK..
d3k40a_ QSENHRMRGAALEKAIEQDVAEGLIPFYAVVTLGTTNSCAFDYLDECGPVGNKHNLWIHVDAAYAGSAFICPEYRHLMKG
d1m32a_ MDIAALHIDYLISSANKCIQG......VPGFAFVIAREQKLAACKGHSRS.......LSLDLYAQWRCMEDNHG......
```

```
66
67  d1vefa1 .....HGTTFGGNPLAMAAGVAAIRYLERTRLWERAAELGPWFMEKLRAIPSPK......IREVRG.MGLMVGLELKEK.
68  d2epja_ ...VFNAGTFNAHPITMAAGLATLKALEEEPVYSVSREAAKALEEAASEVLDRTGLPYTINRVESM.MQLFIGVEEVSN.
69  d1zoda_ ...YLFYTTHVSDPLPAAVGLRVLDVVQRDGLVARANVMGDRLRRGLLDLMERFD....CIGDVRG.RGLLLGVEIVKDR
70  d3doda_ K.TFFHGHSYTGNQLGCAVALENLALFESENIVEQVAEKSKKLHFLLQDLHALPH.....VGDIRQ.LGFMCGAELVRSK
71  d1fg7a_ ......MKVIAPYPLSTPVADIAAQALSPQGIVAMRERVAQIIAEREYLIAALKEIPCVEQVFDSE.TNYILARFKASS.
72  d2f8ja_ .......RVRLPFNVSYVSQMFAKVALDHREIFEERTKF..IVEERERMKSALREMG..YRITDSR.GNFVFVFMEKEE.
73  d1lc5a_ .......RQQMPWSVNALAALAGEVALQDS..AWQQATWHWLREEGARFYQALCQLP.LLTVYPGR.ANYLLLRCERED.
74  d1bw0a_ ....LKRVGMLVCGPCTVVQAALGEALLNTPQEHLDQIVAKIEESAMYLYNHIGECIGLAPTMPRG.AMYLMSRIDLEKY
75  d2gb3a1 .......QGRLAPPLLEQIGSVGLLNLDDSFFDFVRETYRERVETVLKKLEEHGLKR...FTKPSG.AFYITAELPVEDA
76  d2dkja_ .....LEHVIAGKAVAFFEALQPEFKEYSRLVVENAKRLAEELARRGYRIVTGGTDNHLFLVDLRP.KGLTGKEAEERLD
77  d2ez1a_ A.KELVVVYEGMPSYGGLAGRDMEAMAIGLREAMQYEYIEHRVKQVRYLGDKLKAAGVPIVEPVGG.HAVFLDARRFCEH
78  d2rfva_ SPFNAWLTLRGVKTLGIRMERHCENALKIARFLEGHPSITRVYYPGLSSHPQYELGQRQMSLPGGI.ISFEIAGGLEAGR
79  d2c81a_ ...NAIREKNAMFLNDALSKIDGIKVMKRPPQVSRQTYYGYVFRFDPVKFGGLNADQFCEILREKLNMGTFYLHPPYLPV
80  d3k40a_ .IESADSFNFNPHXWMLVNFDCSAMWLKDPSWVPLGRRFRALKLWFVLRLYGVENLQAHIRRHCNFAKQFGDLCVADSRF
81  d1m32a_ .....KWRFTSPTHTVLAFAQALKELAKEGGVAARHQRYQQNQRSLVAGMRALGFNTLLDDELHSPIITAFYSPEDPQYR
82
83  d1vefa1 ....AAP........YIARLEKEHRVLALQAGPT............VIR.FLPPLVIEKEDLERVVEAVRAVLA......
84  d2epja_ ....AAQARKADKKFYVKLHEEMLRRGVFIAPSN............LEA.VFTGLPHQGEALEIAVEGLRSSLKTVLGS.
85  d1zoda_ RTKEPADGLGAKITRECMNLGLSMNIVQLPGMGG............VFR.IAPPLTVSEDEIDLGLSLLGQAIERAL...
86  d3doda_ ..ETKEPYPADRRIGYKVSLKMRELGMLTRPLGD............VIA.FLPPLASTAEELSEMVAIMKQAIHEVTSLE
87  d1fg7a_ ............AVFKSLWDQGIILRDQNKQPS............LSG.CLRITVGTREESQRVIDALRAEQV......
88  d2f8ja_ ................KERLLEHLRTKNVAVRS............FRE.GVRITIGKREENDMILRELEVF........
89  d1lc5a_ ............IDLQRRLLTQRILIRSCANYPG............LDSRYYRVAIRSAAQNERLLAALRNVL.......
90  d1bw0a_ R......DIKTDVEFFEKLLEEENVQVLPGTIFH............APGFTRLTTTRPVEVYREAVERIKAFCQRHAA..
91  d2gb3a1 EEFAR..WMLTDFNMDGETTMVAPLRGFYLTPGL............GKKEIRIACVLEKDLLSRAIDVLMEGLKMFCS..
92  d2dkja_ AVGITVNKNAIPFDPKPPRVTSGIRIGTPAITTRGFT.........PEEMPLVAELIDRALLEGPSEALREEVRRLALAH
93  d2ez1a_ LTQDEFPAQSLAASIYVETGVRSMERGIISAGRNNVTGEHHRPKLETVRLTIPRRVYTYAHMDVVADGIIKLYQHKEDIR
94  d2rfva_ RMINSVELCLLAVSLGDTETLIQHPASMTHSPVAPEER.....LKAGITDGLIRLSVGLEDPEDIINDLEHAIRKAT...
95  d2c81a_ HKNPLFCPWTKNRYLKSVRKTEAYWRGLHYPVSERASG.......QSIVIHHAILLAEPSHLSLLVDAVAELARKFCV..
96  d3k40a_ ELAAEINMGLVCFRLKGSNERNEALLKRINGRGHIHLVPAKIKDVYFLRMAICSRFTQSEDMEYSWKEVSAAADEMEQEQ
97  d1m32a_ ...........FSEFYRRLKEQGFVIYPGKVSQS............DCFRIGNIGEVYAADITALLTAIRTAMYWT....
98
99  d1vefa1 ..................
100 d2epja_ ..................
101 d1zoda_ ..................
102 d3doda_ D.................
103 d1fg7a_ ..................
104 d2f8ja_ ..................
105 d1lc5a_ ..................
106 d1bw0a_ ..................
107 d2gb3a1 ..................
108 d2dkja_ PMP...............
109 d2ez1a_ GLKFIYEPKQLRFFTARFDYI
110 d2rfva_ ..................
111 d2c81a_ ..................
112 d3k40a_ ..................
113 d1m32a_ ..................
114 //
```

Listing C.1: MSA c.67.1

D Comparision of Scatterplots for MSA c.67.1

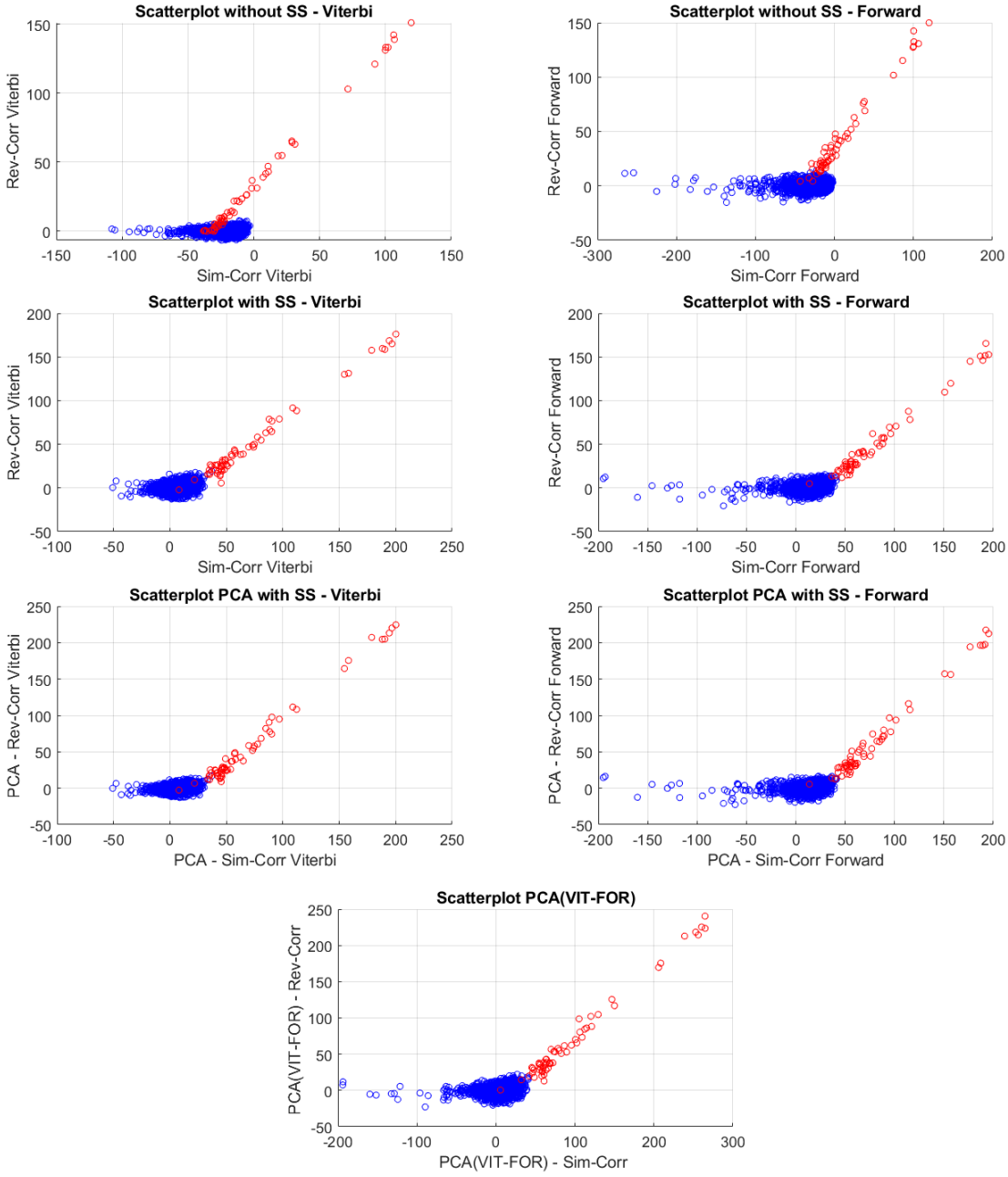

Figure D.1: Comparision of scatterplots for optimized scores of MSA c.67.1.

E Scatterplots for MSA c.67.1 used for ROC

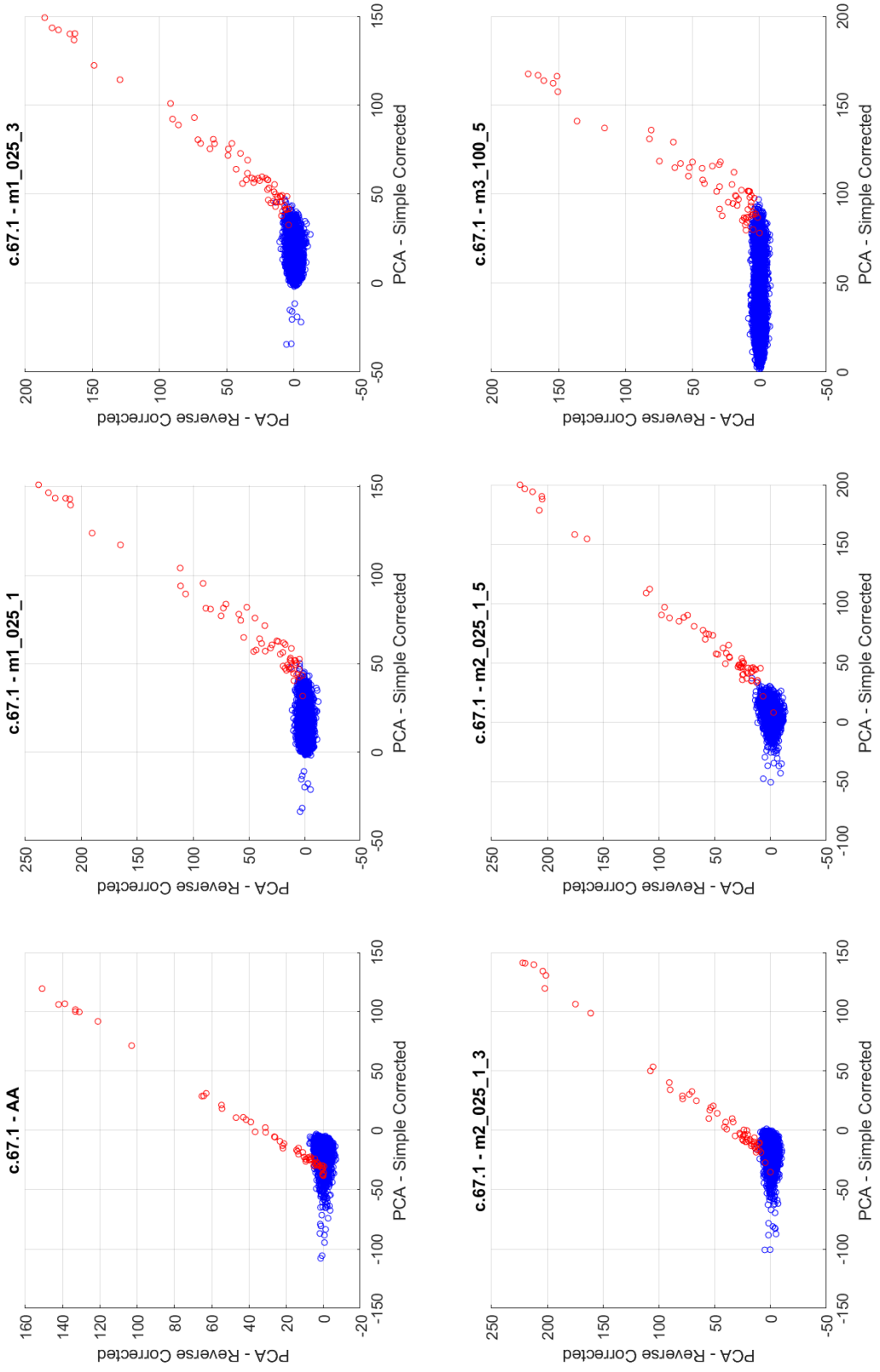

Figure E.1: Scatterplots of MSA c.67.1 using different weighting methods.