Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Technische Universität München

Transmembrane protein contacts and signal peptide evolution

Peter Hönigschmid

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Prof. Dr. Jan Baumbach

Prüfende der Dissertation:

1. Prof. Dr. Dmitrij Frischmann
2. Prof. Dr. Burkhard Rost

Die Dissertation wurde am 06.09.2018 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 06.12.2018 angenommen.

Abstract

Most transmembrane and exported proteins follow a similar pathway to reach their destination. Bound to the signal recognition particle and guided to the membrane, the N-termini of both protein types share similar characteristics, *i.e.* an α-helical hydrophobic stretch. In case of an exported protein, this signal sequence gets cut off at the cleavage site, while it remains a part of the mature protein as transmembrane segment if it is inserted into the membrane resulting in a transmembrane protein.

Given the similarity of the N-terminus, separating exported from transmembrane proteins only by their sequence is a difficult task. In order to facilitate the differentiation, we hypothesized that spatial residue-residue contacts as a predicted structural features could be beneficial for the separation. While we could validate our hypothesis, the discriminative strength of this feature does not reach the performance of already established methods for the prediction of signal peptides and their discrimination from transmembrane protein domains.

In the process of testing our initial hypothesis we were able to develop a new contact prediction method for α-helical transmembrane proteins. The method called MemConP, combines the latest developments in co-evolution analysis with a machine learning approach. These co-evolution methods experienced remarkable progress over the recent years through the application of novel co-variation algorithms which eliminate transitive evolutionary connections between residues.

The downloadable standalone tool MemConP achieves a substantially improved accuracy (precision: 56.0%, recall: 17.5%, MCC: 0.288) compared to the use of either machine learning or co-evolution methods alone. The method also achieves 91.4% precision, 42.1% recall and a MCC of 0.490 in predicting helix-helix interactions based on predicted contacts. The approach was trained and rigorously benchmarked by cross-validation and independent testing on up-to-date non-redundant datasets of 90 and 30 experimental three dimensional structures, respectively.

A further investigation focused on the secretion of orthologous proteins. Related sequences typically perform similar functions and should therefore also appear at the same location. To assess this conclusion, we examined the distribution of signal peptides within the orthologous groups of Enterobacterales.

Parsimony analysis and sequence comparisons revealed a large number of signal peptide gain and loss events, in which signal peptides emerge or disappear in the course of evolution. Signal peptide losses prevail over gains, an effect which is especially pronounced in the transition from the free-living or commensal to the endosymbiotic lifestyle. The disproportionate decline in the number of signal peptide-containing proteins in endosymbionts cannot be explained by the overall reduction of their genomes.

Signal peptides can be gained and lost either by acquisition/elimination of the corresponding N-terminal regions or by gradual accumulation of mutations. The gained results lead inevitably to the conclusion that the evolutionary dynamics of signal peptides in bacterial proteins represents a powerful mechanism of functional diversification.

Zusammenfassung

Die meisten Transmembran- wie auch exportierten Proteine nutzen einen ähnlichen Mechanismus um zu ihrem Zielort zu gelangen. Gebunden an das Signalerkennungspartikel und damit zur Membran geführt, haben die N-Termini beider Arten von Proteinen ähnliche Charakteristika, einen α-helikalen hydrophoben Abschnitt. Im Falle eines exportierten Proteins, wird jene Signalsequenz an einer bestimmten Schnittstelle abgetrennt. Sofern dieses Protein in die Membran eingebettet wird und somit als Transmembranprotein fungiert bleibt der N-Terminus jedoch als ein Teil des finalen Proteins als erstens Transmembransegment erhalten..

In Anbetracht der Ähnlichkeit der N-Termini ist die Unterscheidung von exportierten und Transmembranproteinen nur anhand ihrer Sequenz eine diffizile Aufgabe. Mit dem Ziel die Differenzierung zu verbessern, stellten wir die Hypothese auf, dass räumliche Kontakte zwischen Aminosäuren als ein vorhergesagtes strukturelles Merkmal bei der Unterscheidung hilfreich sein könnten. Auch wenn sich unsere Theorie als richtig erwiesen hat, ist das Signal der Methode nicht ausgeprägt genug um bessere Ergebnisse zu liefern als bereits etablierte Methoden zur Vorhersage von Signalsequenzen und ihre Abgrenzung von Transmembranproteindomänen.

Bei der Auswertung unserer Hypothese ist es uns gelungen eine neue Kontaktvorhersagemethode für α-helikale Transmembranproteine zu entwickeln. Die Methode mit dem Namen MemConp vereint die neuesten Entwicklungen der Koevolutionsanalyse mit einem Ansatz des maschinellen Lernens. Die Methodik der Koevolutionsanalyse hat in den letzten Jahren erhebliche Fortschritte erzielt. So wurden neue Algorithmen entwickelt welche transitive evolutionäre Korrelationen zwischen Aminosäuren eliminieren.

Das eigenständig lauffähige und öffentlich verfügbare Programm MemConP erreicht eine erheblich höhere Genauigkeit (precision: 56.0%, recall: 17.5%, MCC: 0.288) verglichen mit der Vorhersage alleinig durch maschinelles Lernen oder durch Koevolutionsmethoden. Darüber hinaus erreicht die Methode 91.4% precision, 42.1% recall und einen MCC von 0.490 bei der Vorhersage von Helix-Helix Interaktionen basierend auf vorhergesagten Kontakten. Die Methode wurde auf einem Datensatz von 90 und 30 aktuellen, experimentell bestimmten dreidimensionalen Strukturen trainiert und mittels Kreuzvalidierung bzw. unabhängiger Testung streng evaluiert.

Anschließend eruierten wir mögliche Unterschiede im Exportverhalten von orthologen Proteinen. Verwandte Sequenzen haben typischerweise ähnliche Funktionen und sollten somit auch dieselbe Lokalisierung aufweisen. Um diese Aussage zu überprüfen untersuchten wir die Verteilung von Signalsequenzen innerhalb orthologer Gruppen innerhalb der Ordnung der Enterobacteriales.

Die Analyse der maximalen Sparsamkeit und Sequenzvergleiche brachten eine große Anzahl von Ereignissen zum Vorschein, bei denen Signalsequenzen im Laufe der Evolution hinzugewonnen werden oder verloren gehen. Signalsequenzverluste sind häufiger als Zugewinne, ein Effekt der besonders ausgeprägt beim Wandel von einer freilebenden oder kommensalen zu einer endosymbiontischen Lebensweise zu beobachten ist. Die überproportionale Abnahme der Anzahl von Proteinen mit Signalsequenzen in Endosymbionten kann nicht mit der generellen Verkleinerung ihrer Genome begründet werden.

Signalsequenzen können entweder durch den Erwerb bzw. Verlust der entsprechenden N-terminalen Regionen oder durch die allmähliche Akkumulation von Mutationen hinzugewonnen werden bzw. verloren gehen. Die gewonnenen Erkenntnisse führen unweigerlich zu der Annahme, dass die evolutionäre Dynamik von Signalsequenzen in bakteriellen Proteinen einen bedeutenden Mechanismus zur funktionellen Diversifikation darstellt.

Acknowledgements

In my years at our department obtaining the doctoral degree was not the only task I had to conquer. Being involved in side-projects, collaborations, travel, teaching, conferences, seminars, supervising thesis and administrative tasks, it is clearly visible that my studies provided me with versatile challenges. Everybody who has been involved in such a project, or dedicated a similar amount of his time and effort into a task knows, that obstacles on the way are easier to overcome if you are not on your own. Therefore I do not only want to give acknowledgements but sincerely thank several people for accompanying me on my way.

First of all, I want to thank my supervisor Prof. Dr. Dmitrij Frishman for spending his valuable time teaching me to become a researcher. Of course I appreciate his scientific knowledge, his valuable comments, fruitful discussions and his habit to be always available to his students. I even more value his patience to let me grow into my projects and to trust in my decisions.

Special thanks go to Prof. Dr. Burkhard Rost for helping me to find my passion for machine learning, playing a major role in my bioinformatics education and for allocating some of his precious time to be my second corrector.

I would like to express my gratitude to my former and present colleagues at our department not only for fruitful discussions and debates about our projects, but also for all the other conversations whether they were serious, educating, or just hilarious. Thanks Léonie Corry, Roswitha Weinbrunn and Erna Look for keeping the department up and running, a task which a group of scientists would never be able to solve as flawlessly and smiling as they do. Thanks to Drazen Jalsovec for managing hard- and software, air conditioning and power outages, version conflicts and grid queue jams. Thanks to Michael Kiening for always being a good friend and office room mate, and for always having the time for discussions. Thanks to Martina Weigl for accompanying, motivating and helping me during the finalization period of this thesis and for always asking the right questions. Thanks to Usman Saeed and Pinar Karabulut for being good friends during our greenhorn stage. Special thanks go to Evans Kataka, Xeynub Asrar, Bo Zeng, Jan Zaucha, Anja Mösch, Marina Parr, Fei Qi, Alec Steep, Jinlong

Ru, Jianfeng Sun, Hengyuan Liu, Hongen Xu, Florian Goebels, Kerstin Haase, Stefanie Kaufmann for being great colleagues and people.

I am very grateful to Prof. Dr. Jan Baumbach who kindly agreed to be the chair of my examination committee.

Last but not least want to express my deepest gratitude to my family: To my parents Sigrid and Peter for their continuous love and support, and for providing me with the tools to become the person I am today. To my siblings Ulrike for her always honest opinion, Marion for always being willing to listen and Florian for being my inspiration to get into computer science. And to my grandfather Gerhard for probably being more proud of this thesis than I am.

Publications

The two entries written in bold are part of this thesis.

- Hamp, Tobias, Kassner, Rebecca, Seemayer, Stefan, Vicedo, Esmeralda, Schaefer, Christian, Achten, Dominik, Auer, Florian, Boehm, Ariane, Braun, Tatjana, Hecht, Maximilian & others (2013). Homology-based inference sets the bar high for protein function prediction. BMC bioinformatics, 14, S7.

- **Hönigschmid, Peter & Frishman, Dmitrij (2016). Accurate prediction of helix interactions and residue contacts in membrane proteins. Journal of structural biology, 194, 112-123.**

- **Hönigschmid, Peter, Bykova, Nadya, Schneider, René, Ivankov, Dmitry & Frishman, Dmitrij (2018). Evolutionary interplay between symbiotic relationships and patterns of signal peptide gain and loss. Genome biology and evolution, 10, 928-938.**

- Ivankov, Dmitry N, Bogatyreva, Natalya S, Hönigschmid, Peter, Dislich, Bastian, Hogl, Sebastian, Kuhn, Peer-Hendrik, Frishman, Dmitrij & Lichtenthaler, Stefan F (2013). QARIP: a web server for quantitative proteomic analysis of regulated intramembrane proteolysis. Nucleic acids research, 41, W459-W464.

- Kloppmann, Edda, Hoenigschmid, Peter & Rost, Burkhard (2013). Protein Secondary Structure Prediction in 2012. Encyclopedia of Biophysics, 2040-2047.

- Kulandaisamy, A, Priya, S Binny, Sakthivel, R, Tarnovskaya, Svetlana, Bizin, Ilya, Hönigschmid, Peter, Frishman, Dmitrij & Gromiha, M Michael (2018). MutHTP: Mutations in Human Transmembrane Proteins. Bioinformatics, 1, 2.

- Radivojac, Predrag, Clark, Wyatt T, Oron, Tal Ronnen, Schnoes, Alexandra M, Wittkop, Tobias, Sokolov, Artem, Graim, Kiley, Funk, Christopher, Verspoor, Karin, Ben-Hur, Asa & others (2013). A large-scale evaluation of computational protein function prediction. Nature methods, 10, 221.

- Yachdav, Guy, Kloppmann, Edda, Kajan, Laszlo, Hecht, Maximilian, Goldberg, Tatyana, Hamp, Tobias, Hönigschmid, Peter, Schafferhans, Andrea, Roos, Manfred, Bernhofer, Michael & others (2014). PredictProtein—an open resource for online prediction of protein structural and functional features. Nucleic acids research, 42, W337-W343.

Contents

List of Figures

List of Tables

Acronyms

AP	average precision.
BP	biological process.
CC	cellular component.
COG	cluster of orthologous groups.
com	commensal.
EC	evolutionary coupling.
ENA	European Nucleotide Archive.
endo	endosymbiont.
F_β	F-score.
fl	free-living bacterium.
FN	false negative.
FP	false positive.
GO	Gene Ontology.
HMM	hidden Markov model.
LIPS	LIPid-facing Surface.
MCC	Matthews correlation coefficient.
MF	molecular function.
mfDCA	mean-field direct coupling analysis.
MI	mutual information.
MSA	multiple sequence alignment.

NCBI	National Center for Biotechnology Information.
OMA	Orthologous MAtrix.
OPM	Orientations of Proteins in Membranes.
P	precision.
PDB	Protein databank.
PDBTM	Protein databank of transmembrane proteins.
PFAM	Protein Families.
R	recall.
SA	signal anchor.
Sec	Secretion.
SP	signal peptide.
SRP	signal recognition particle.
SSEP	Segment specific evolutionary profile.
Tat	Twin-arginine translocation.
TM	transmembrane.
TMP	transmembrane protein.
TMS	transmembrane segment.
TN	true negative.
TOPDB	Topology Data Bank of Transmembrane Proteins.
TP	true positive.
UPS	unconventional protein secretion.

1 Introduction

The following introduction is designed to provide information about the biological and methodological background as well as its implications for the approaches used in the published articles. Proteins are versatile macromolecules, playing a crucial role in almost every process in the cell. Being synthesized in the cytoplasm based on their encoding in the genomic DNA sequence, their final sphere of action does not have to be at this same place. Two types of polypeptides which operate elsewhere, exported and transmembrane proteins, are the focus of this thesis.

Both types of proteins contain a biological postal code encoded in their N-terminal sequence determining their target localization. Certain similarities in their N-termini lead to difficulties distinguishing them from each other just by their sequences.

The N-terminal sequence for exported proteins is called signal peptide. This introduction will give an overview starting from the most important export pathway. It continues with the current estimate of signal peptide annotations and gives insights into their experimental and computational annotation.

The second part is dedicated to transmembrane proteins with focus on their spatial structure. The experimental determination and the difficulties arising especially for this class of proteins will be examined. A further section outlines structure prediction in general, and the history of contact prediction approaches. These methods were first employed on globular proteins and later applied to transmembrane proteins as well.

Finally, the challenging task of differentiating exported proteins containing a signal peptide and transmembrane proteins having similar hydrophobic N-termini will be covered, including our unpublished results to solve the problem using structural contacts as a separation criterion.

1.1 Signal peptides

1.1.1 The Sec pathway

The secretome denotes the entirety of molecules of any kind which are exported from a cell via varying mechanisms. While this per definition also includes for example inorganic elements, in the field of bioinformatics the term in most cases refers to the export of proteins.

Several pathways exist which enable proteins to be exported through the cytoplasmic membrane. The most prominent one, which accounts for 96% of all secreted proteins in *Escherichia coli* [1], is the Secretion (Sec)-pathway. In order to be recognized by this secretion machinery, the protein is required to contain a N-terminal signal sequence, the SP, which has an approximate length of 20 to 30 amino acids. These N-terminal peptides have a tripartite structure consisting of a positively charged N-terminus for their correct orientation during the process, a hydrophobic stretch to enter the membrane, and a cleavage site which gets recognized by the signal peptidase I [2, 3].

Cytoplasmic proteins, *i.e.* proteins which are not secreted, fold during or immediately after their synthesis. Non-cytoplasmic proteins are either embedded into the membrane or secreted to the other side of the plasma membrane, to the periplasm, or even beyond to the outer membrane. Most of the secreted proteins evade immediate folding in the cytoplasm and are delivered to their final destination first where they then become folded. However there is a minority of proteins that are exported in a folded state being the exception of that rule [4].

Exported proteins face several challenges: (i) they must remain unfolded and soluble until they reach their final destination, (ii) they must be distinguished from cytoplasmic proteins, (iii) they need to be correctly guided to transmembrane channels, (iv) activate their opening, (v) they need energy for the translocation process, (vi) they have to detach from the transportation machinery to become released, (vii) and they have to fold in the cell envelope.

The co-translational process of Sec dependent protein export starts with the recognition of unfolded pre-proteins that contain a SP during their translation. First targeted by the signal recognition particle (SRP) and then bound to its membrane receptor FtsY, the complex is transported to the transmembrane SecYEG channel [5]. An alternative, post-translational export is aided by chaperones: the trigger factor [6], and SecB [7] bind to the translated pre-protein on the ribosome in the cytoplasm and maintain its unfolded state [8]. Pre-proteins are then targeted to the SecYEG-SecA translocase residing in the membrane and translocated through SecYEG to the periplasm or into the plasma membrane [8]. This process is powered by repeated cycles of ATP binding and hydrolysis by SecA and the proton motive force [9]. The auxiliary components SecDF–YajC [10] and YidC [5] enhance translocation efficiency. The signal peptidases finally cleaves the signal peptide and the mature domain is released into the periplasm [11].

While the Sec dependent secretion is the most common export mechanism, other pathways for proteins to traverse the membrane exist. The Twin-arginine translocation (Tat)-pathway for example is meant to extend the Sec-pathway because of its ability to export already folded proteins [12]. Tat SPs have a similar tripartite structure compared to Sec SPs with a charged N-terminus, a hydrophobic stretch and a cleavage site. However they contain an additional motif consisting of two Arginine residues preceded by a polar residue and followed by two hydrophobic residues with a distance of one amino acid.

Being known for several decades, the Sec and Tat pathways, in their co- as well as post-translational version are thoroughly studied and cover most of the secretion machinery. Nevertheless, so-called unconventional protein secretion (UPS) pathways gained attention as they represent yet unknown export mechanisms which are complex and do not rely on a SPs sequence [13].

1.1.2 Signal peptide content

Providing the facility to interact with the environment, for instance to respond on external stimuli, it is no surprise that protein secretion is a common phenomenon. The first assessments in 2004 for *Escherichia coli* and in 1997 for *Haemophilus influenzae* suggested exported proteins to represent 20% of the proteome [14, 15]. This number underwent multiple corrections towards more conservative estimates. The most recent studies still hypothesize that 10% of the proteins in *Escherichia coli* are secreted [16]. This decrease of approximately 50% is the result of improved prediction methods [17, 18], such as SignalP (currently in the version 4.1). Improved models and an increased amount of training data besides an enhanced ability to separate TMSs from SPs decreased the number of false positive predictions. Additionally, the raise of more powerful experimental methods, especially the use of proteomics data, allowed for more precise protein annotations on a large scale, increasing the accuracy of the estimates even further [19, 20].

1.1.3 Experimental determination of signal peptides

The first SPs were identified by investigating each sequence separately. One of the employed methods was Edman degradation, an approach which identifies the amino acid sequence of a protein one residue at a time, starting from its N-terminus. By using the fact that SPs are cleaved off and are, therefore, no longer present in the mature protein, SPs can be identified by a comparison with the respective translated genomic sequence.

The basic knowledge about SPs like their tripartite structure, their typical length or the composition of their cleavage site was accumulated by the first studies employing this approach [2, 3].

While the concept of comparing the N-terminal sequence of a protein to the predicted gene did essentially not change, follow up methods using mass-spectrometry allow the high-throughput identification of SPs. In mass-spectrometry for proteomics, the protein is digested into several peptides by the use of the peptidase Trypsin which cuts after Arginine and Lysine. The first peptide of a given protein sequence which has a non-tryptic N-terminus, *i.e.* a peptide whose preceding amino acid is neither Arginine nor Lysine, is used to determine candidates for potentially secreted proteins. Similar to the Edman degradation approach, a SP is present if this first non-tryptic peptide is located within 15-50 amino acids of the potential protein start site which is either predicted or experimentally determined. Therefore the region between the initial start codon or Methionine and the first observed non-tryptic peptide is a potential SP.

In order to increase the reliability of the experimental approach, additional filtering criteria are applied to ensure that certain SP characteristics are met. This includes the potential SP's length and other characteristics like the tripartite structure. This high-throughput method was employed in several studies about different organisms which include *Human* [21], *Halobacterium salinarum* and *Natronomonas pharaonis* [22], *Shewanella oneidensis* [20], *Yersinia* [23], *Aspergillus niger* [24] and even whole communities [25, 19], leading to a significant increase in SP annotations.

1.1.4 Prediction of signal peptides

Like in many fields of computational biology, the more experimental determined and validated data is available, the more sophisticated prediction models can be implemented. The first method for predicting SPs was based on a weight matrix and was able to determine the existence of a SP and the position of the cleavage site [26]. This method specifically used 161 eukaryotic and 36 prokaryotic non-homologous SPs with known cleavage sites for the calculation of the weight matrix. This weight matrix represents a profile and is built by comparing the actual frequency of a given amino acid at a specific position in the SP to the background probability, *i.e.* the amino acid composition of the whole protein sequence. The authors of this first SP prediction method claim a prediction accuracy of 75-80% for both prokaryotic and eukaryotic proteins.

The next generation of SP prediction methods were already based on machine learning algorithms leading to more powerful methods such as SignalP or Phobius, which are still widely used [15, 14].

Especially the SignalP series underwent ongoing improvements to increase sensitivity and overall performance. The first version of SignalP (1.1) [15] employed six neural networks, two for Eukaryotes, two for Gram-positive and two for Gram-negative organisms. For each type of organism, one neural network was used to predict the cleavage site, and one classified the residues belonging to either the SP or the mature protein. A final decision was made by averaging the output of the first neural network with the slope of the second. The reported Matthews correlation coefficient (MCC) for the discrimination of a protein containing a SP or not ranged from 0.88 for Gram-negative to 0.97 for Eukaryotes.

Version 2.0 of SignalP [27] switched from neural networks to hidden Markov models (HMMs) directly modelling the n-, h- and c-region of a potential SP. While this approach achieved similar performance to version 1.1, improvement of the model for the separation between SPs and signal anchors (SAs) added further benefits to the method.

The follow up method, SignalP 3.0 [17] basically combined the HMM and neural network approach, introduced new features and refined the model further, e.g. selecting more appropriate window sizes for the neural network. This version improved the discrimination between proteins with and without SP further, while the main benefit consisted of the more precise determination of the cleavage site position.

The latest and up-to date version, SignalP 4.1 [18] included TMPs in the dataset. As TMSs at the beginning of a TMP show similar characteristics like the hydrophobic α-helical stretch (see section 1.3), many false positives could be corrected.

The other most widely used method Phobius and its follow-up PolyPhobius [14, 28] was primarily designed to predict TMP topology. Phobius employs a HMM which tries to model the architecture of a TMP as accurate as possible. The N-terminus of a protein can either be a short loop, a globular domain or a SP, successively followed by a number of TMSs, short loops or additional globular domains. Not only the topology prediction of this approach can be still classified as state of the art, but also the discrimination of proteins with and without SP shows high precision. PolyPhobius which was only released one year after the first version, included information from MSAs to increase the prediction accuracy.

During the following years, several other software tools have been developed for prediction of signal peptides [29, 30, 31, 32] and protein localization [33, 34, 35, 36, 37, 38].

1.1.5 Signal peptides and evolution

Besides their functional diversity, proteins can evolve aberrant functionality due to evolutionary changes in the genomic sequence. Duplications, mutations, insertions or deletions are events that can alter the proteins behaviours to either a gain, loss or change in their functionality.

More specifically gain-of-function mutations can lead either to an increased activity of a protein or even to a complete new function. The development of a neomorphic allele, *i.e.* an allele introducing a new function, in a tumor's genes for example can lead to an unanticipated phenotypic outcome. This raises the possibility that tumors with this mutation may not respond to therapies designed to target the wildtype of the protein [39].

The more common loss-of-function mutations either reduce the activity of a gene or completely prohibit functionality. Patients with Sickel cell disease for example have hemoglobin which clumps up leading to a reduced capability of transporting oxygen [40], *i.e.* the protein loses the functionality to transport oxygen properly.

Protein studies have revealed results which stay in contrast to the central dogma of molecular biology, postulating the determination of protein function by its structure. That means two proteins which have the same structural fold, can exhibit different functions, exemplary shown in [41]. Their analysis showed that a sequence identity of 45% almost guarantees structural similarity, while 85% sequence identity are necessary to transfer the proper function. Even orthologous proteins which have highly similar amino acid sequences, do not always own the same function [42, 43] and they can differ for example in their binding specificity [44] or their interaction sites [45]. In addition to changes in the amino acid sequence itself, other protein characteristics which influence the protein's function, like phosphorylation sites or protein domains, can be altered during the course of evolution.

Literature research revealed further findings of changes in protein function, focusing on binding dynamics and protein-protein or protein-ligand interactions. Functional changes induced by alterations of cellular targeting signals, e.g. SPs, on the other hand were only examined by a few studies mainly limited to sequence diversity of these peptides [46]. Two examples are the amino acid composition of Mitochondrial matrix targeting signals which has constraints inflicted by the N-terminal sequence of the mature protein [47] and the prediction of localization signals using the sequence divergence as a significant feature [48].

Functional change is often necessary for the adaption to new environmental conditions linked to the lifestyle of an organism. Hence, a different secretion profile would be expected between pathogenic and non-pathogenic species. This assumption was confirmed by showing that the major difference between pathogenic and non-pathogenic *Listeria* species, manifests itself strongest in the secretome, as the pathogenic species showed more exported virulence factors [49]. A more recent study found that within Gram-negative bacteria, intracellular pathogens had the smallest secretomes. Trying to generalize a universal connection between pathogenicity and the secretome size however failed as the secretomes of certain bacteria did not fit into this pattern [50]. The same study reported a positive correlation between the percentage of secreted proteins and the number of genes in the gram-negative, but not in the gram-positive organisms.

1.1.6 Gain and loss of signal peptides

As described in chapter 3, we further investigated the evolutionary events connected to SPs on the basis of bacteria belonging to the *Enterobacterales* order. We show that homologous proteins which are assumed to carry out the same function are not always exported to the same location. In addition, we could prove that SPs can be lost or gained during the course of evolution and give insights about the mechanism leading to these events. Finally, we reveal a connection between these events and the lifestyle of *Enterobacterales* bacteria [51].

1.2 Transmembrane proteins

Every cell is surrounded by at least one membrane, a lipid bilayer, which separates the cell's interior from the exterior environment. To enable the cell to communicate with its exterior, e.g. reacting to external signals, export its content, or import necessary molecules an interface through the membrane is needed. Therefore, a type of protein called TMPs exists, serving to connect the inside (intracellular space) to the outside (extracellular space) of the cell. TMPs possess TMSs which contain mostly hydrophobic amino acids, span the the entire membrane and often form an α-helical structure. The hydrophobic nature of the TMSs stabilizes the position as well as the structure of the TMP by hydrophobic interactions with the lipid bilayer.

TMPs can be classified according to the localization of their N- and C-terminal domains and the number of TMSs. Bitopic or single-pass TMPs cross the membrane only once while polytopic or multi-pass TMPs cross the membrane multiple times. Depending on the mechanism which inserts the protein into the membrane, these two types can be further separated into proteins with either their N- or their C-terminal at the outside of the membrane.

Being integrated in the membrane, TMPs play several roles for the cells' functions and their interaction or communication with their surrounding. Transporters for example facilitate the exchange of molecules between the two sides of the membrane. Carrier proteins actively support the process of translocation with glucose transporters being one example for this class [52]. TMPs are able to form channels which serve as control guards for molecule streams at the membrane. Molecules are allowed to cross the membrane by the channels conformational change, e.g. voltage-gated sodium channels [53]. Another form of transporter is a pore which is an open channel without control function. A prominent example for a pore are the aquaporins, enabling water to flow into and out of the cell [54].

Receptors like G protein-coupled receptors (GPCRs) [55], with an example shown in Figure 1.1, as another class of TMPs can interact with substrates like hormones, neurotransmitters, cytokines, growth factors, cell adhesion molecules or nutrients at the extracellular side of the membrane. This leads to conformational changes of the TMP including its intracellular domain. This conformational change activates a cascade of further processes at the cell's interior.

Another important class consists of those TMPs which provide enzymatic activity, e.g. the methane monooxygenase [56].

1.2.1 Insertion of membrane proteins into the membrane

Similar to exported proteins, the SRP plays a key role to recognize and guide the TMP to the membrane. Exemplary in *E. coli*, the SRP attaches to the site where the polypeptide exits the ribosome and binds to the N-terminal hydrophobic segment of the newly synthesized protein [57]. This complex interacts with the SRP receptor FtsY at the membrane and gets attached to the SecYEG translocon. After proper attachment, two GTPs get hydrolyzed, one from the SRP and one from the FtsY GTPase domain, which leads to a release of the synthesized protein into the translocon. SecYEG contains a lateral gate from which the transmembrane segments of the protein get inserted into the membrane consecutively [58].

1.2.2 The sequence structure gap

The protein sequence-structure gap [59] describes the discrepancy between the number of available protein sequences and the number of available experimentally solved three-dimensional protein structures. UniProt contained about 10 thousand sequences back in 1990, a number which rose to 116 million sequences today in 2018 which is a 11600-fold increase [60]. The number of structures in the PDB increased only 280 fold from 500 to 140 thousand at the same time. This effect is even more pronounced for TMPs although they represent almost 30% of the human proteome. While 0.121% (140 thousand out

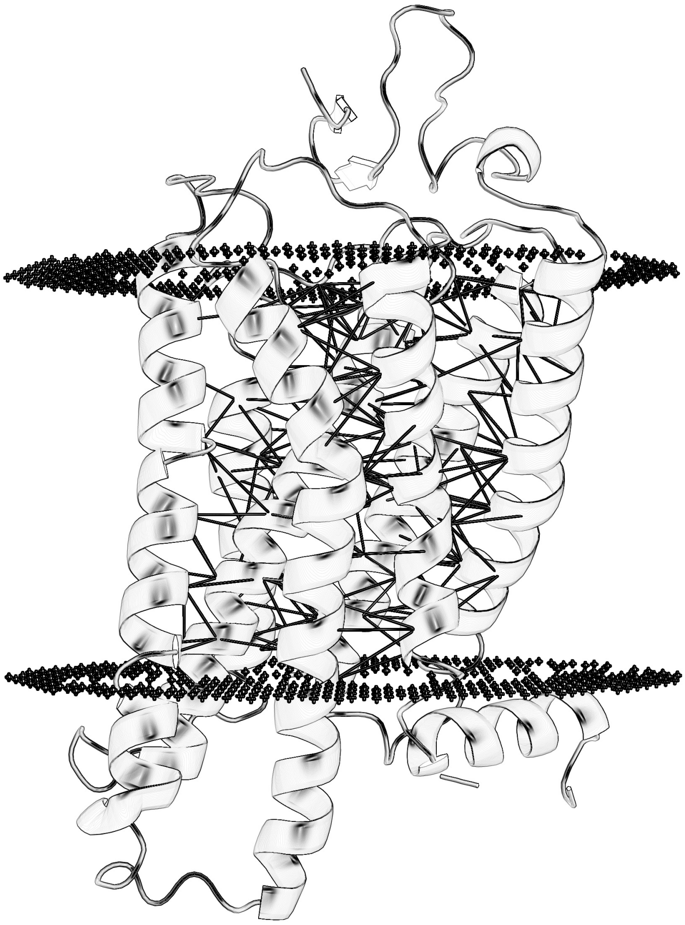

Figure 1.1: Visualization of the PDB entry *1gzm*, the structure of bovine rhodopsin, a G-protein coupled receptor. The dotted planes indicate the membrane boundaries, while black lines show amino acid pairs which C_α atoms have a distance less than 8Å.

of 116 million) of the known UniProt sequences have a corresponding structure in the PDB, this holds true for only 0.01% (2200 [61, 62] out of 23 million) of the known TMP sequences.

Figure 1.2 shows the yearly increase in the number of structures and sequences since 1985. An additional tendency is clearly visible from the same figure: The number of added Swiss-Prot annotated entries went lower since 2007. This saturation could indicate that only redundant sequences are added to the UniProt while the reviewed entries (Swiss-Prot) converge more and more with the real world sequence space.

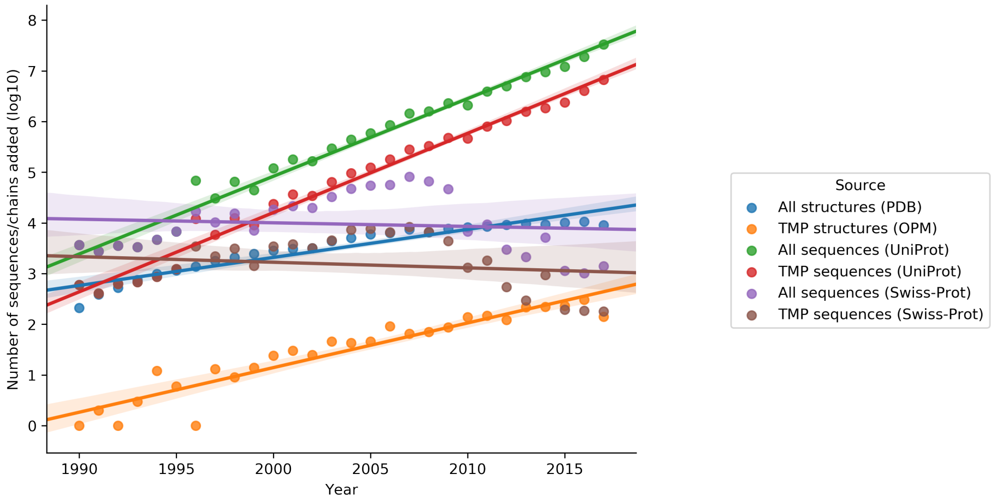

Figure 1.2: Number of sequences and structures added pear year. The sequences were derived from UniProt and Swiss-Prot, whereas proteins were annotated as TMPs if they contained a TM region in the subcellular localization section. The structures were derived from the PDB and the OPM. While the points represent the actual datapoints, the lines show the linear fit in order to make the tendencies visible more clearly.

1.2.3 Experimental determination of transmembrane protein structures

Knowing that TMPs account for about 30% of the human proteome the reason for the discrepancy between the number of solved structures for TMPs and globular proteins cannot be explained by lacking interest. Instead the experimental determination of TMP structures is the factor influencing these numbers. Two commonly used methods for the determination of protein structures are NMR spectroscopy and X-ray crystallography. NMR structures are derived using constraints resulting from NMR spectra. These constraints aim primarily on inter-atomic distances [63]. The result of this approach is an ensemble of multiple possible structures, also called models. Multiple models have the benefit of showing potential disordered regions or at least regions of high mobility. Another benefit of NMR spectroscopy is that the molecule to be studied can be left in its natural solution. Disadvantages of the method include a limited molecule size of the subject and, on average, less precise structures than the most common method, the X-ray crystallography [64].

X-ray crystallography, is the favoured method, at least according to the number of structures deposited in the PDB, where 86% of the entries are experimentally determined by this approach. To solve the structure, the molecule has to be crystallized. This prerequisite is also one of the biggest drawbacks of the method especially for membrane proteins, proteins with natively disordered regions or for transient complexes as it is difficult to crystallize them in their native structure. In order to overcome these issues, the molecules are often altered by introducing disulfide bonds to stabilize the structure. The electron-density map resulting from the X-ray assay is used as a template for the atomic coordinates derived from the protein sequence resulting in the final solved structure. Another drawback aside from the need of crystallization is that the crystallized structure is held in a rigid state, therefore removing most of the mobility information from the molecule [64].

In their work about the purification of TMPs from *Saccharomyces cerevisiae* for X-ray crystallography [65] the difficulties of crystallizing TMPs are discussed:

First, it is difficult to overexpress TMPs compared to soluble proteins and they generally show lower abundance. The second reason is the use of detergents which are necessary to remove the TMP from the membrane. These detergents can denature the proteins just as much as the removal of the protein from its natural environment, the membrane. Therefore, it is often necessary to create stabilized versions of these proteins by introducing mutations, e.g. to force the formation of disulfide bonds. In addition, the association of the TMPs with the detergent creates large complexes, which are difficult to crystallize and to be of use for x-ray crystallography.

1.2.4 Protein structure prediction

The lack of solved structures can be partly compensated by predicting them. The most successful approaches for protein structure prediction belong to the class of homology modelling methods. For a query sequence of unknown structure to be predicted by homology modelling, there is one prerequisite: there has to be a homologous protein of known structure high sequence similarity to the query. The homolog's structure serves as a template for the modelling process. These methods typically involve several steps beginning with the search of a suitable template structure in a database, and mapping the query sequence on to that template sequence. Afterwards, the backbone is generated by transferring the query onto the template coordinates. The next steps include the modelling of loops which are often more flexible regions of the protein and side chains of the amino acids. These two can be either facilitated in a knowledge-based way, e.g. rotamere libraries in the case of side chain orientations, or by energy functions. The model is finally optimized based on energetic measures and the validation of the model according to certain criteria. A well known and widely used method conducting homology modelling is SWISS-MODEL [66].

In more difficult cases, where no well-defined template covering the whole query sequence can be found, fragment based methods serve as a more versatile approach for structure prediction. Fragment based methods involving similar processes as homology modelling and can be seen as a subclass thereof. The difference is, that these methods do not rely on a single template. I-TASSER, for example, threads the query through a representative PDB structure library. The found fragments are extracted from regions aligned before and reassembled to full-length models. Similar to homology modelling, the unaligned regions are built by *ab initio* modelling [67].

Ab initio protein structure prediction methods are applied if no suitable template can be found at all. This approach can already be a part of homology modelling or fragment based methods at the stage of loop modelling. *Ab initio* methods rely exclusively on the amino acid sequence of the query and do at most use very small fragments of known structures. Although the mechanics of protein folding are by far not fully understood, *ab initio* methods like ROSETTA [68] try to simulate the physical forces acting on the protein chain to find a conformation with the lowest free energy possible. Taking into account all degrees of freedom of the folding process, the possible number of conformations is virtually uncountable, at least for a protein with a significant length. This limitation makes the computation unfeasible at least for routine use [69].

Correlating with the significant rise in performance of residue-residue contact prediction methods which is introduced in the next sections, another type of structure prediction approaches became more successful: the contact-guided structure prediction. By using predicted contacts as restraints during the *ab initio* modelling process, the number of possible conformations can be reduced significantly. The recently published method CONFOLD2 [70] uses various subsets of input contacts and employs a soft square energy function which takes into account these restraints to explore the reduced number of conformations.

1.2.5 Early residue-residue contact prediction methods

One of the most useful restraints for protein structure prediction are residue-residue contacts as they reduce the search space for the lowest energy structure. Especially their feature to infer contacts from sequence makes them useful for *ab initio* structure prediction methods where no template is available.

The foundation for residue-residue contacts prediction was laid by investigating the known structure of the *Tobamovirus* coat protein and the corresponding MSA which consisted of seven family members [71]. By searching for residue pairs with identical conservation patterns, the authors found those pairs to be amino acids which are spatially close in the structure. Shortly after, the hypothesis of such correlated mutations was verified by the same authors in three other families: serine proteases, cysteine proteases and haemoglobins [72]. In this follow-up study, the determination of structural contacts from sequence was hindered by larger alignments where the strict criterion of identical conservation patterns could no longer be applied.

Not surprisingly, these findings were used in the opposite direction by using MSAs to find correlated mutations, and infer amino acids which are spatially close [73]. In this first approach, substitution matrices were derived for each MSA column and a residue-residue contact was predicted if the correlation criteria were fulfilled. Tested on 11 protein families, the prediction accuracy of the highest scoring pairs ranged from 37% to 68%. Although still in is infancy and a low accuracy for the restriction to only the most confident predictions, the method's performance was already from 1.4 to 5-fold higher than a random prediction.

When MSAs with more sequences were available, further improvements could be achieved by changing the type of MSA representation or correlation calculation. Statistical coupling analysis (SCA), introduced in 1999 [74], compared the co-evolution of two residues not only in the whole MSA, but also in several subsets. A follow-up method refined this approach [75]. Another co-evolution measure, mutual information, did not change the MSA but used an approach known from information theory to search for co-evolving residues in proteins [76].

Further significant improvements in terms of accuracy could not be achieved by varying either the MSA representation nor the co-evolution measure. Especially the number of false positive predictions limited the usefulness of these methods, as residue contacts are not the only reason for the occurrence of correlated mutations.

Therefore, subsequent approaches tried to incorporate other features derived from sequences and MSAs, including sequence conservation, sequence separation along the chain, alignment stability, family size, residue-specific contact occupancy, the formation of contact networks and phylogenetic information by simple linear combination [77].

With the rise of computational power, a multitude of machine learning methods arrived in the field of sequence based predictions. The approaches started to substitute the easy linear combination with machine learning methods such as neural networks. In addition, other features such as the predicted secondary structure or solvent accessibility were paired with correlated mutations and other values derived from the MSA by using machine learning to increase the prediction accuracy of residue contacts [78].

1.2.6 Residue contacts in transmembrane proteins

Invariably all these methods were designed to predict residue contacts in soluble proteins. Because TMPs account for approximately 20-30% of all proteins in the human genome [79], the reason for this circumstance cannot be found in this class of macromolecules being of too little interest. With the combination of general co-evolution methods and machine learning trained on specific sequences resulting in the highest performance, the first approach would have been to apply the same methods to TMPs.

As the machine learning models were trained on globular proteins, they also learned the characteristics specific for this protein class. TMPs, however, have unique characteristics which distinguish them from globular proteins. First and foremost, the α-helical stretches which span across the membrane and the lipids surrounding them. These helices possess a significantly more hydrophobic amino acid composition to anchor the TMP in the lipid bilayer by residue-lipid interactions. For that reason, it is not surprising that residue contact prediction methods developed for globular proteins perform poorly on TMPs [80].

The insufficient amount of solved TMP structures these days made it difficult to properly develop and evaluate a specialized method for predicting residue contacts in TMPs. As explained in section 1.2.3, these proteins are embedded in the membrane and the issues solving their structures experimentally emerge from the fact that they denature if removed from their natural environment, *i.e.* the membrane.

The study from 2007 [80] was the first large-scale analysis of co-evolving residues in membrane proteins. In addition to benchmarking different prediction methods, an application of residue contacts for the prediction of interacting helices was analysed. While the prediction accuracies did not exceed 10% with any of the tested methods, almost 49% of all predicted contacts were found to be within one helical turn of an actual contact. By employing a meta approach, *i.e.* combining the tested prediction algorithms, the authors were able to increase the accuracy to 53%. Finally, interacting helices could be predicted at a specificity of 83% and sensitivity of 42%.

TMHcon, a method specifically for TM residue contacts was published two years later [81]. TMHcon is a neural network based method combining several co-evolution measures with features derived from sequence. These features include the evolutionary profile calculated from the MSA, the position of the residues in the TMS, the orientations of the sidechains, the protein length and its number of TMSs. It was trained and evaluated on 62 non-redundant protein chains and achieved a more than 2-fold increase in accuracy compared to co-evolution methods alone as well as approaches using information derived from globular proteins.

1.2.7 Latest developments in co-evolution analysis

In the meantime, not only the number of TMP structures increased from 800 in 2009 to 2200 in 2017 (see section 1.2.2) but also a huge improvement could be made in calculating residue co-evolution. Because the earlier approaches described in section 1.2.5 only considered two positions in a MSA, they all were facing the problem of transitive connections. In detail, two residues A and B look like they are co-evolving, but in reality residue A as well as residue B are co-evolving with a third residue C, but not with each other. Recent methods solve this problem including all information available in the MSA by building a statistical model which tries to explain the measured co-evolution between all residue pairs at the same time.

Two pioneer approaches based on these novel ideas were developed: Firstly mean-field direct coupling analysis (mfDCA), implemented as EVFold [82] which was directly applied to predicting structures using the derived contacts as structural constraints. The second method is based on the estimation of a sparse inverse covariance matrix, as used in PSICOV [83]. Both methods were reimplemented reducing their runtime while maintaining their predictive performance in FreeContact [84].

Along the discovery of the first co-evolution methods, improved approaches to predict residue contacts in soluble proteins building on these developments have been released. They employ either enhanced algorithms (CCMpred, [85]), or combine several co-evolution methods (PconsC2 [86], MetaPSICOV [87]).

1.2.8 MemConP

As will be described in chapter 2, we created a follow-up method to TMHcon incorporating the latest developments in co-evolutionary analysis. Paired with random forest as a machine learning method which has proven to be successfully applicable to many recent problems, we were able to create a method to predict residue-residue contacts and helix interactions in TMPs with state of the art prediction performance. While we applied our method only to the problem of distinguishing SPs from TMSs, we believe that our tool can be supply contacts as restraints for structure prediction increasing their accuracy and decreasing their runtime by reducing the number of conformations which have to be explored [88].

1.3 Distinguishing signal peptides from transmembrane segments using residue contacts

One of the difficulties SP predictors have to overcome is the differentiation between SPs or SAs and TMSs at the N-terminus of the sequence. This is due to the fact that both, the h-region of SPs as well as TMSs, are composed of hydrophobic residues and form a α-helical structure. Of course, the reverse is true: tools developed to predict TMSs, have to ensure not to confound SPs, SAs and TMSs. While for the SP prediction only N-terminal TMSs can lead to mistakes, TMSs can be similar to numerous other structures crossing or penetrating the membrane: amphipathic helices, and re-entrant helices, *i.e.* helices which enter and exit the membrane on the same side which is common in many ion channel families, or, if the tool is specific to α-helical proteins, β-sheets. Depending on the aim of the prediction, one possibility is to use both kind of tools, e.g. SignalP to prefilter for proteins containing a SP, followed by a topology predictor on the remaining sequence. Furthermore, approaches incorporating both prediction targets exist. Phobius, an example mentioned before [14], predicts SPs as well as the TM topology using a HMM-based approach. The successor to Phobius, named Polyphobius, even includes evolutionary information using a MSA [28]. There are more specialized tools, e.g. methods such as TMLOOP [89], TOP-MOD [90] and OCTOPUS [91] which have attempted to identify re-entrant regions.

Although TMSs and SPs have a hydrophobic stretch, TMSs are generally longer. In addition, TMSs do not have cleavage sites, but as the cleavage-site pattern itself is somehow variable it is not always a sufficient feature to separate SPs from TMSs. In consequence, trying to annotate all SP in a genome computationally results in a lot of false positive predictions.

This issue is where the improvement of SP prediction methods took place, e.g. SignalP 3.0 included submodels for different types of sequences: one for SPs, one for SAs and one for other proteins. Another approach to overcome these issues is the joint prediction of SPs and TMSs as done by Phobius or PolyPhobius mentioned earlier.

1.3.1 Residue-residue contacts as a criterion

Before, we introduced the difficulties when separating SPs from TMSs and the developments in the field of co-evolutionary measures in section 1.2.7. With these two aspects in mind, we hypothesized that there are less contacts between the hydrophobic stretch found in SPs and remaining TMSs than between any two TMSs. If that hypothesis holds, the prediction of residue-residue contacts in TMPs would be beneficial for the process of protein annotation. Because the latest contact predictor aimed towards TMPs, TMHcon, was released in 2009, we decided to implement an updated method MemConP as described in chapter 2.

While we were not able to validate the hypothesis to publish it, the following approach and results still give valuable insights.

1.3.2 Data acquisition

The initial dataset consisted of 5209 human protein entries downloaded from the UniProt database [60] including their TMS as well as their SP annotations. These entries were reviewed, *i.e.* part of the Swiss-Prot database, and had at least one TMS. The initial dataset was filtered so that every TMS had to have valid start/end positions and each protein had to contain either at least two TMSs or one TMS and a SP to be able to calculate contacts between two helices. From the 4109 sequences being valid according to these criteria, 2430 formed the final dataset after redundancy reduction to 40% sequence identity using CD-HIT [92].

1.3.3 Co-evolution measures and residue contact prediction methods

We calculated 14 scores from four sources to distinguish between TMSs and SPs. Two SP prediction methods, SignalP and Phobius, served as state of the art methods for the task at hand. SignalP outputs a Dscore between zero and one, where a higher Dscore represents a higher probability for the protein to contain a SP. Phobius, on the other hand, only predicts the presence or the absence of a SPs, which was converted into a binary variable being either one or zero, respectively.

The other two sources were EVfold [82], or rather the Freecontact implementation [84], and our developed tool MemConP [88]. Freecontact calculates two values for each pair of residues, the mutual information (MI) [76] representing one of the early co-evolution measures, and a evolutionary coupling (EC)-value being a state of the art co-evolution measure. Our tool, MemConP, predicts a contact score for each pair of TM residues and a helix-helix interaction score for each pair of TMSs. To predict contacts between a SP and the remaining TMSs, the SP was annotated as TMS to be processed by MemConP.

Three methods were applied to convert these scores into a prediction, whether the first helix is a SP or a TMS. To express the strength s of a potential interaction between two helices i and j using measure or method m, the scores of each possible residue pair between these two helices were accumulated:

$$s_{i,j} = \sum_{a=1}^{k_i} \sum_{b=1}^{l_j} m(a,b)$$

where a and b are residues in the TMSs i and j, respectively and k_i and j_i the respective lengths of these TMSs.

For the remaining scores, MI, EC-value, the MemConP contact scores and the MemConP helix-helix interaction score, the strength between two TMSs was used in three ways. The strength between the first and the second TMSs, where the first TMS can be a SP. The second is the average strength between the first and all other TMSs. The third is the maximum strength between the first and any other TMS.

1.3.4 Distributions of interaction strengths

Figure 1.3 shows the distribution of the 14 different scores separated by proteins with and without SP. The visually clearest separation takes place with the two reference tools SignalP and Phobius where a low score indicates the absence of a SP. This result is expected as these two methods were trained to separate these two types of proteins. Independent of the score calculation, MI is able to separate proteins with SP from proteins without SP. The same is true for the EC-values, although not as significantly as with the MI scores. Our tool, which is tailored towards the prediction of residue-residue contacts and helix interactions in TMPs and has proven by rigorous benchmarking to be successful at that task, does not perform well when predicting with the aim to separate proteins with and without SP. While the helix interaction scores perform well in many cases, *i.e.* predicting a lower interaction score between SPs and TMS, the contact scores are quite similar between the two classes, or even favour the opposite from our expectation. An explanation is that some of the features, such as the number of TMSs, the position of the residue or the index of the TMS create false signals. The score distributions, for example, indicate that the predictor learned that the first and the second TMS in a TMP interact in most cases. As it has not been presented with SPs during training, the predictor transfers this observation to proteins with SP as well. The visually best separation is achieved by using the average contact strength of the first TMS with the other TMSs, although SPs are predicted to have higher contact scores which is not expected.

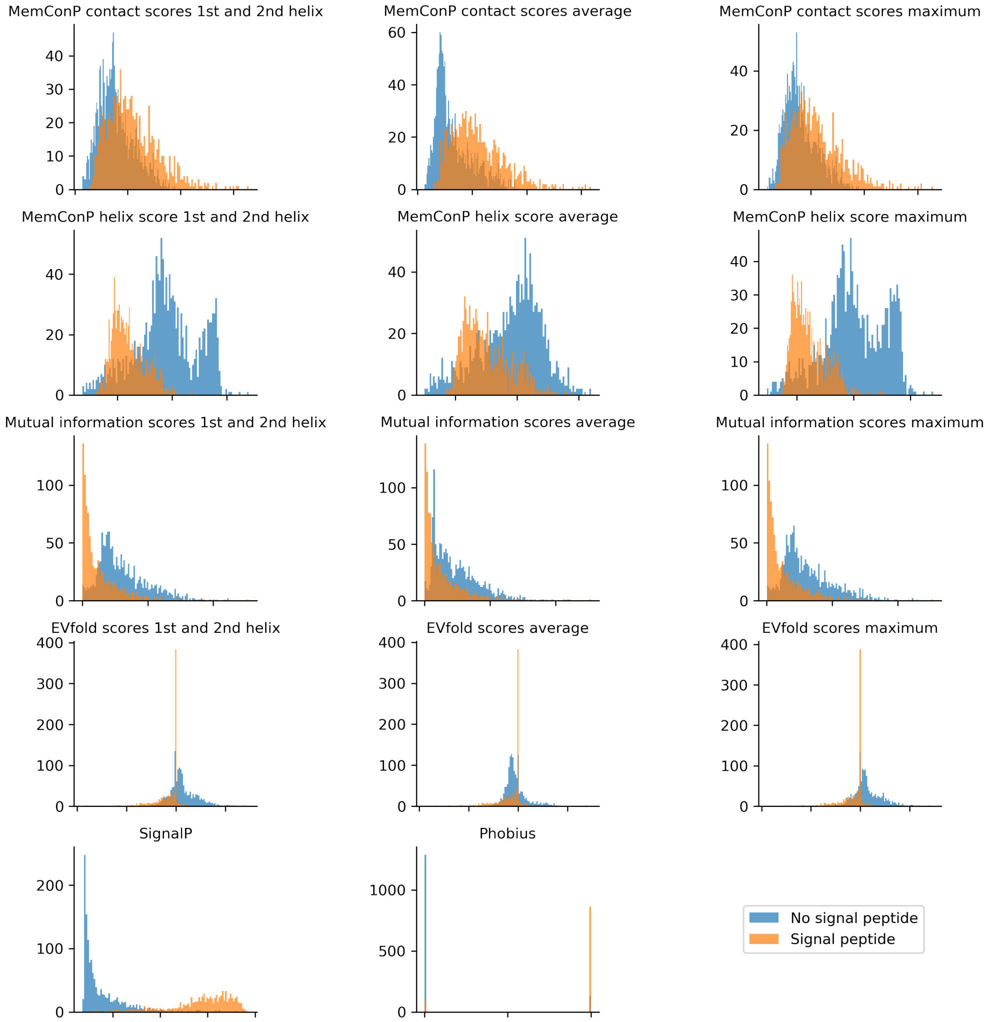

Figure 1.3: Score distributions for the first helix which is either a TMS or a SP using different co-evolution measures and contact prediction method scores.

1.3.5 Performance using interaction strengths as a signal peptide predictor

In order to measure the predictive performance of a score or method and quantify the visual inspection, the average precision (AP) was used. The AP summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight:

$$\mathrm{AP} = \sum_n (R_n - R_{n-1}) P_n$$

where P_n and R_n are the precision and recall at the nth threshold

Table 1.1 shows the AP for the different scores and methods. Because we expect to have less contacts between a SP and a TMS, the prediction scores of MemConP, the MI and the EC-values had to be inverted such that a high value indicates a SP. These inverted scores are indicated with "(rev.)" in the table. The quantified results confirm the visual inspection as the best separation between SP and TMS is achieved by SignalP with an AP of 0.95 followed by Phobius with an AP of 0.82. While not clearly visible in Figure 1.3, MI excels the EC-values by 10%, with the scores achieving APs of 0.78 and 0.71, respectively. Regarding MemConP, only the helix interaction scores give a meaningful result, situated between MI and the EC-values with an AP of 0.73. The contact scores predicted by MemConP on the other hand as well achieve a AP of 0.70, but only when not using the scores in their reversed form as we had expected. This means that the contact scores of MemConP are able to separate SPs from TMSs with an acceptable precision, but only under the assumption that SPs have more contacts or a higher contact strength to the remaining TMSs.

In sum, it is obvious that the already available tools Phobius and, especially, SignalP do very well on the task. We were surprised, however, that MI performs way better than the EC-values and that only the MemConP helix interaction but not the contact scores follow our assumption of lower scores between SPs and TMSs than the other way around.

Table 1.1: Sorted threshold independent average precision using different co-evolution measures and contact prediction method scores to predict whether the first helix of a TMP is a TMS or a SP. (rev.) signifies that the score has been multiplied with -1 and thus been reversed, *i.e.* a low score predicts a SP.

Score	**Average precision**
SignalP	0.95
Phobius	0.82
(rev.) Mutual information scores maximum	0.78
(rev.) Mutual information scores 1st and 2nd h...	0.77
(rev.) MemConP helix score maximum	0.73
(rev.) EVfold scores maximum	0.71
(rev.) Mutual information scores average	0.70
MemConP contact scores average	0.70
(rev.) EVfold scores 1st and 2nd helix	0.70
(rev.) MemConP helix score 1st and 2nd helix	0.67
MemConP contact scores 1st and 2nd helix	0.64
MemConP contact scores maximum	0.64
(rev.) MemConP helix score average	0.56
(rev.) EVfold scores average	0.53
EVfold scores average	0.42
(rev.) Phobius	0.37
EVfold scores 1st and 2nd helix	0.30
EVfold scores maximum	0.30
(rev.) MemConP contact scores maximum	0.30
(rev.) MemConP contact scores 1st and 2nd helix	0.29
Mutual information scores average	0.29
MemConP helix score average	0.28
Mutual information scores maximum	0.27
Mutual information scores 1st and 2nd helix	0.27
(rev.) MemConP contact scores average	0.27
MemConP helix score 1st and 2nd helix	0.26
MemConP helix score maximum	0.25
(rev.) SignalP	0.24

2 Accurate prediction of helix interactions and residue contacts in membrane proteins

The aim of this work was to create a sequence based classifier which is able to predict intramolecular residue-residue contacts between amino acids residing in the transmembrane segments of membrane proteins as well as inter-helical contacts.

The data, *i.e.* the protein chains and the TMS definitions, were downloaded from the PDBTM database. To show the impact of a larger dataset size, we first reused one dataset for cross-validation and one for independent testing derived from earlier publications, while the final cross-validation and independent testing was done on our own recent rigorously redundancy reduced dataset. The mentioned redundancy reduction procedure not only involved the number of identical residues between two proteins, but also the TM-score, a measure of structural similarity, and Protein Families (PFAM) families/clans.

To separate the residue pairs into two classes, contacting and non-contacting residues, two definitions were used to ensure future comparability. In the first definition, two residues are in contact if any of their atoms are closer than 5.5Å, while the second definition asks for the C_α atoms to be closer than 8Å.

The random forest algorithm was used as the machine learning method of choice, as it is easy to train and not prone to overfitting. The protein sequence was converted to several features, which are either global, such as the amino acid composition, local e.g. the evolutionary profile derived from the MSAs and several physico-chemical amino acid properties, or pairwise like the co-evolutionary measure calculated by EVFold/Freecontact.

As a result, we were able to improve the accuracy significantly compared to other contact prediction methods. Including we were able to quantify the impact of each of our included improvements, *i.e.* the use of the most recent co-evolutionary methods, the increased amount of training data and the application of the newest database search methods.

By converting the residue contacts into helix interaction, we provide a tool which gives important insights into the topology for TMPs where no structure is known yet.

Peter Hönigschmid planned and conducted the experiment and wrote the manuscript. Dmitrij Frishman supervised and planned the project and wrote the manuscript.

2.1 Introduction

Protein sequence-structure gap [59], already quite dramatic for globular proteins, is even more pronounced for membrane proteins, with merely two thousand atomic structures available [61, 62] for over one million amino acid sequences containing at least one predicted TM region [60]. The bulk of this huge discrepancy stems from the challenge to crystallize membrane proteins, as they are likely to lose their original structure when removed from their natural lipid environment due to their strongly hydrophobic surfaces, flexibility, and lack of stability [93]. The low number of known 3D structures also limits our ability to increase the structural coverage of membrane proteins by template-based structure prediction methods. On the other hand, sequence-based methods to predict the topology of TMPs, while highly accurate and useful, are unable to shed light on their spatial architecture.

Perhaps the only sequence-based approach able to provide information about the spatial arrangement of polypeptide chains and, in particular, useful constraints for 3D structure modeling, involves predicting contacts between amino acid residues. Prediction methods of the first generation exploited the idea of compensatory residue substitutions as an indication of a residue contact and utilized statistical methods of varying degree of sophistication to identify correlated mutations between pairs of positions in a multiple alignment (reviewed in [80]). More recent methods additionally applied machine learning algorithms to extract information about potential contacts form multidimensional data, such as evolutionary profiles, physico-chemical properties of amino acids, and other sequence specific features [78]. However, all these methods, without exception, were designed to predict residue contacts in soluble proteins.

For a very long time sparseness of structural data precluded the application of contact prediction techniques to TMPs. Not surprisingly, methods trained on globular proteins produce extremely poor results when applied to membrane protein sequences due to their very specific biophysical properties, most notably the fact that their exterior is much more hydrophobic than the interior due to the interaction with the lipid environment. In 2009 we developed the first contact predictor (TMHcon) specifically geared towards α-helical membrane proteins, which employed a neural network trained on sequence features and correlation measures, which dramatically outperformed earlier methods used for globular proteins in terms of precision and recall [81].

Since the release of TMHcon the number of experimentally determined three-dimensional structures of TMPs that can be used for training prediction algorithms increased significantly, from a mere 160 high resolution structures (non-redundant at 40% sequence identity) in 2009 to over 330 today. Concomitantly, the recent availability of more sensitive database search methods, such as HHblits [94], allows to create better evolutionary sequence profiles by detecting more homologous sequences to be included in the MSA. Finally, and most importantly, there has been a quantum leap in our ability to detect compensatory mutations, which are indicative of structural contacts. While earlier methods assessed residue co-variation between each pair of positions in a MSA individually using simple correlation measures, such as mutual information, recent methods rely on global statistical models. These models attempt to infer causative correlations from the entire alignment and are thus able to distinguish between direct structural contacts and transitive connections between residues. The two pioneer approaches based on these novel ideas are mean-field direct coupling analysis (mfDCA), implemented as EVFold [82], and the estimation of a sparse inverse covariance matrix, as used in PSICOV [83]. For both methods an accelerated implementation called Freecontact [84] is available. Recently improved methods to predict residue contacts in soluble proteins have been released, which either employ enhanced algorithms (CCMpred, [95]), or combine several co-evolution methods (PconsC2 [86]), MetaPSICOV [87]).

Here we introduce a novel computational method, MemConP (**Mem**brane **Con**tact **P**rediction), which is specifically geared towards predicting residue contacts and helix interactions in TMPs. The tool takes advantage of the recent surge in the number of 3D structures, more sensitive sequence analysis techniques, and vastly improved approaches to residue co-variation. It employs the random forest classification algorithm, which utilizes a large number of decision trees, each trained on a randomly chosen subset of training data and features. The resulting ensemble of classifiers determines the outcome by a majority voting. The random forest approach is used to combine several sequence-derived (evolutionary profiles, amino acid properties) and structure-derived (predicted TM topology) features with the *mfdca* approach offered by Freecontact. We also introduce a new highly non-redundant dataset for training machine learning methods on TMPs, as well as a new independent test dataset, which can serve for performance comparison with future methods. We compare the performance of MemConP with several recent predictive techniques, which employ residue co-evolution.

2.2 Materials and Methods

2.2.1 Definition of transmembrane segments, residue contacts, and helix interactions

For comparison of our method with other techniques we used the definition of TM regions obtained from the PDBTM database [61]. PDBTM definitions were also utilized for benchmarking of contact predictions. For benchmarking the quality of helix interaction predictions we rely both on PDBTM as well as on TM topology predictions produced by PolyPhobius [28]. To make our method comparable to the already existing and future ones (including our own previous work [81], we used the definition of residue contacts based on the Euclidean distance between any two atoms of less than 5.5Å. A pair of helices was defined to be interacting if there was at least one residue contact between them. Another common contact definition is the distance between the C_β atoms of two residues of less than 8Å. Performance measures for this alternative definition are reported the *Appendix* (Tables A.1 and A.2).

2.2.2 Datasets

We used four datasets to train and benchmark the predictor: *OldTrain*, *OldTest*, *NewTrain* and *NewTest*. The first two datasets, *OldTrain* and *OldTest*, were used by all recent TM helix contact prediction methods and thus served as comparison datasets. *OldTrain* (introduced by [81]) originally consisted of 62 redundancy reduced X-ray structures of TM proteins extracted from PDBTM, Topology Data Bank of Transmembrane Proteins (TOPDB) [61], and OPM [96], with a resolution better than 3.5Åand possessing at least three TMSs. We omitted the entry *2a79* from this dataset, as the corresponding topology data was deleted from PDBTM. *OldTest* was introduced by [97] and contains 21 TMPs, of which none has a sequence identity above 40% to any other protein in this dataset, nor to those in *OldTrain*. To create the *NewTrain* and *NewTest* datasets, used to train and test our final predictor, atomic coordinates and the annotation of membrane-spanning regions were extracted from the PDBTM database. PDBTM contains 3D structure information of experimentally solved TMP structures, including atomic coordinates and the annotation of TM regions generated by TMDET [98]. We used the "Redundant Alpha" dataset of June 2015 from PDBTM containing 7374 protein chains as the initial dataset of TMPs. In order to produce a training dataset which is not biased towards an overrepresented family of proteins, and a test dataset which is totally independent from the training data, the initial dataset had to be redundancy reduced. Unfortunately, all existing approaches are aimed towards redundancy reduction of globular proteins. These methods take into account global or local sequence similarity using a substitution matrix which is designed for globular proteins and not optimized for highly hydrophobic TMSs. We therefore applied a very rigorous procedure to reduce redundancy both within and between our training and test datasets, incorporating structural similarity and PFAM family/clan membership [99] in addition to sequence similarity. Specifically, we calculated the length-independent measure of structural similarity, the so called TM-score, using the TMalign method [100]. The PFAM family/clan membership was added as an additional criterion to eliminate similarity between multi-domain proteins as well as to address those cases where even structural similarity comparison fails. Two proteins were declared similar if they i) either shared a sequence identity of more than 35%, ii) or displayed a TM-score above 0.5, which, according to the authors, implies that they share the same fold, iii) or belonged to the same PFAM family or clan. To minimize the bias towards a specific type of TMPs we grouped all proteins in this initial dataset according to the number of TMSs they possess. Subsequently protein chains were drawn from each of these groups, one at a time, and added to the *NewTest* dataset. At the same time, the sequences in the same group, which were similar to the drawn protein, were removed from the initial dataset. Upon achieving a certain pre-defined size of the *NewTest* dataset, N_{test}, the procedure was continued and the drawn proteins added to the *NewTrain* dataset until the initial dataset was depleted, automatically yielding a certain size of the *NewTrain* dataset, N_{train}. By applying the described procedure we sought to ensure the lack of similarity both within or between the newly created datasets. To achieve a uniform increase in the size of the newly created datasets (*NewTrain*/*NewTest*) relative to the old ones (*OldTrain*/*OldTest*), we chose N_{test} to be 30, resulting in N_{train} of 90, which is a 1.5-fold increase. We additionally required i) structure resolution to be better or equal to 3.5Å(with preference given to higher resolution structures for proteins identical in sequence), ii) protein chains to contain at least three TM helices (as we are interested in inter-helical interactions), and iii) sequences to not contain any unknown residues. The increase of performance due to using *NewTrain* over *OldTrain* for training was assessed on the newly derived independent dataset *NewTest*. We tried to keep the redundancy between the *OldTrain* and *NewTest* datasets to a minimum. This was achieved by first sorting the proteins in the initial dataset in ascending order according to the highest TM-score they display to any protein contained in the *OldTrain* dataset. Proteins most dissimilar to any protein in *OldTrain* were drawn first during the redundancy reduction procedure and added to *NewTest* first. As a result, 21 of the 30 proteins in *NewTest* have a TM-score smaller than 0.5 to any protein in the *OldTrain* dataset. We identified residues making contacts with each other (according to the definition described above) and situated on different TMSs, which resulted in an approximate contacting to non-contacting residue pair ratio of 1:50 and interacting/non-interacting helix pair ratio of 1:1.

2.2.3 Database search and multiple sequence alignments

The evolutionary background of protein sequences was retrieved using HHblits [94]. The search results of HHblits are represented as multiple alignments, which served as input for assessing residue co-evolution (see below) and for calculating evolutionary profiles for each protein sequence. HHblits was used with the latest Uniprot20 database from June 2015 and the parameters ”-Z 999999999 -B 999999999 -maxfilt 999999999 -id 99 -diff inf” to maximize alignment size.

2.2.4 Evaluation measures

To benchmark the predictive performance of our method and make it comparable to other techniques, several evaluation measures were used:

- precision (P), also called accuracy or positive predictive value

$$P = \frac{TP}{TP + FP}$$

- recall (R), often referred to as sensitivity or coverage

$$R = \frac{TP}{TP + FN}$$

- F-score (F_β), a weighted average of precision and recall

$$F_\beta = (1 + \beta^2)\frac{Precision * Recall}{(\beta^2 * Precision) + Recall}$$

  We chose the weight factor βin order to favor precision over recall, as we are interested in a rather small amount of high quality contacts. Throughout this text we refer to this measure as $F_{0.25}$ score.

- MCC

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

In these formulas true positives (TPs) are the residue pairs which are correctly predicted to be in contact, true negatives (TNs) are the correctly predicted non-contacting residue pairs, false positives (FPs) are falsely predicted contacting residues, and false negatives (FNs) are residues pairs that are falsely predicted not to be in contact. For the residue-residue contact prediction, the performance values are the average over these measures for each protein, while the performance measures for helix-helix interactions are given for the total number of interacting helices. For a threshold independent assessment of the performance precision-recall plots were generated by sorting all the prediction results according to either their random forest output value (RFscore: between 0 and 1) or their L/x best predictions and plotting the precision against the recall using the predicted value as threshold at each of the points shown in the plot. While the threshold independent evaluation provides insight into the method's overall performance, a fixed threshold for the random forest output has to be applied while using the final predictor. Therefore, we probed the entire RFscore range in order to find the threshold resulting in the highest $F_{0.25}$ score averaged over all proteins. Because some of the sequences in the dataset contain only residue pairs with a low RFscore, we required only 90% of the proteins to have a RFscore at that threshold to account for these exceptions. The cutoff derived from these two criteria is denoted RFscore$^{F_{0.25}}$ in the text. We also determined the number of top predictions according to the $F_{0.25}$ measure, *i.e.* we optimized the fraction x of the L/x top predictions, and the corresponding threshold is denoted L/x$^{F_{0.25}}$.

2.2.5 Random forest parameters, training and output

The classifier of choice for the predictor is a random forest, as implemented in the R [101] package randomForest [102]. This supervised machine learning method is an ensemble classifier that employs *ntree* decision trees. For each tree, a subset of training samples is created by drawing as many samples out of the complete dataset as it contains data points. As this sampling is done with replacement, it results in an approximate usage of two thirds of the complete training data for each tree. Each tree is trained using $\sqrt{M}$ input features, where M is the total number of input features available. The output of the random forest is a value between 0 and 1, representing the fraction of decision trees voting for the residue pair to be in contact. Thus, two residues are predicted as contacting if more than 50% of the votes indicate a contact. A significant benefit of the random forest approach is that due to its mechanism of decision finding it is not necessary to normalize the data. To get a realistic assessment of the trained predictor, we performed a 10-fold cross-validation on the *OldTrain*/*NewTrain* datasets (containing

61 and 90 proteins, respectively) by randomly splitting the data into 10 bins and using 9 bins for training and the remaining one for testing. This process was repeated 10 times, such that each protein was used for testing the performance only once. Applying the trained random forest model to a data sample results in an output value (RFscore) between 0 and 1, representing the fraction of trees voting for a contacting residue pair. Output values closer to 0 and 1 correspond to more confident predictions. As the residue contact data shows a huge class imbalance of 1:50, with non-contacting residue pairs being far more abundant than contacting pairs, the random forest, like any other machine learning method, tends to favor the overrepresented class. To deal with this problem we tried out different balancing methods, but the best performance was achieved by using i) only residue pairs between interacting helices ii) performing under sampling for the majority class with the ratio of 1:5 separately for each tree and iii) choosing the threshold based either on the RFscore or a certain number of top predictions, as explained in the *Evaluation measures* section. We determined the thresholds separately for the residue-residue contact prediction and the helix interactions on the cross-validation data, which might lead to an optimistic assessment on the training data, while the evaluation on the independent test data is not biased.

2.2.6 Input features

Below we list the input features used to train the method described in this work, MemConP.

Amino acid composition was represented as a vector of length 20 containing the fractions of each amino acid in the sequence alignment of interest.

Number of α-helical TMSs.

Evolutionary profile. Evolutionary profiles are represented by vectors of the length 20, each position referring to one of the 20 amino acids. The log likelihood $m_{i,j}$ to find amino acid i at position j in the multiple alignment is calculated for each amino acid as $m_{i,j} = log(p_{i,j})/p_i)$, where $p_{i,j}$ is the relative frequency of amino acid i at column j in the MSA retrieved by HHblits, and p_i is the relative frequency of amino acid i in the whole alignment. $m_{i,j}$ was normalized using the sigmoid function $1/(1 + exp(-m_{i,j})))$.

Segment specific evolutionary profile (SSEP). This feature describes the evolutionary profile of a TMS as a vector of length 20. It is calculated similar to the evolutionary profile, with the difference that instead of pi,j being calculated on a per-residue basis, here pi,j reflects the relative amino acid frequency of all columns within a TMS.

Relative position. Three values were used to describe the position of the contacting residues in the protein: the position in the protein as such, normalized by the length of the protein, the distance from the N-boundary of the TMS the residues are located in divided by its length, and the sequential number of the TMS divided by the total number of TMSs in the protein.

Transmembrane segment length. The absolute length of the TMS harboring a given amino acid residue.

Amino acid properties. Vectors of length 13 containing five numeric (charge, hydrophobicity, pI, volume and mass) and eight discrete (hydrophobic, aliphatic, aromatic, polar, negative, positive, small, C_β-branched) characteristics of amino acids. The data was obtained from the AAindex database [103]. For each property the log likelihood was calculated by dividing the average value in each alignment column by the average in the whole alignment. The result was normalized using the sigmoid function. To obtain a log likelihood for the discrete characteristics, they were transformed into a numeric representation, e.g. the "small" property was assigned the values of 1, 2, and 3 for "normal sized", "small", and "tiny", respectively.

Segment specific properties. A vector of length 13 containing the same amino acid characteristics as the "Amino acid properties" feature. The log likelihood, however, is calculated by taking the average value over all alignment columns in the TMS instead of a per-residue basis, and dividing it by the average in the entire alignment.

Predicted helix orientation. Amino acid residues located in a TMS are partitioned into groups representing seven possible helical surface patches. Each of these faces ($f \in [1, 7]$) contains the residues $f + i$, $f + i + 3$ and $f + i + 4$, with i being every seventh residue of the TMS. Each face is assigned a LIPid-facing Surface (LIPS) score [104] calculated by the alignment entropy and the average lipophilicity of the residues it contains. For each residue the feature consists of four values: the rank of the surface the residue is located in, the LIPS score of this face, residue lipophilicity and its entropy.

Evolutionary coupling values. EVFold [82] is a method to assess evolutionary co-variation between two contacting residues in a protein, which outputs a so-called EC-value. This number provides a quantitative measure of co-evolution between two residue positions in a multiple alignment. While simple approaches such as mutual information only consider the two alignment columns in question, the advantage of EVFold is that it considers the entire alignment by calculating a statistical model that fits best the properties of the alignment. Freecontact [84] is a reimplementation of the EVFold algorithm in C (which was originally written in Matlab), optimized for computational speed. In this study we used the raw EC-values from Freecontact in

EVFold mode with all default parameters. MSAs produced by HHblits were used as input for the Freecontact method.

We utilized three different variations of EC-values, dependent on which residue pairs were considered:

Local EC-values. The EC-values between the current residue and its 8 neighboring residues (4 at each side).

Pair EC-values. The 25 EC-values between residue pairs, which are located on different TM helices and are presumed to face each other. These are the two central residues at positions i and j, and possible combinations of residue positions $i \pm a$ and $j \pm b$, with $(a, b) \in (0, 0), (0, 1), (0, 3), (0, 4), (1, 0), (3, 0), (3, 4), (4, 0), (4, 3), (4, 4)$.

Residue separation in the primary structure. Five features were used to represent the sequence separation between two contacting residues. The first two are the absolute and relative distances between the two residues. The other three are the difference of the TMSs index numbers (absolute and relative, normalized by the number of TMSs), and whether this difference is an odd or even number. The latter feature indicates how the two TMSs are oriented with respect each other (assuming that the topology is correct), *i.e.* an odd number means that the N-terminus of the first TMS is located on the same side of the membrane as the C-terminus of the other one.

2.2.7 Global, context-dependent, segment-specific and position-specific features

Protein features listed above can be subdivided into four categories: i) global features, which includes amino acid composition and the number of TM helices, ii) context-dependent features, which includes the relative position of the residue, the TMS length, and the residue separation in the primary structure, iii) segment-specific features, namely SSEP and segment specific properties, and iv) position-specific features - all other features, namely the evolutionary profile, amino acid properties, predicted helix orientation, and EC-values, where we used a window approach to incorporate relevant information about the adjacent residues. To achieve this, the input vectors from each of the residues inside this window were concatenated.

2.2.8 From residue-residue contacts to helix interactions

The performance of our method is given using the RFscore$^{\mathrm{F}_{0.25}}$ and L/x$^{\mathrm{F}_{0.25}}$ random forest threshold and the well-established benchmark of top L/x predictions from a ranked list, where L is the total number of residues residing in TMSs and x is a number used to obtain a fraction of this amount. We used the L/5 and L best predictions to make our results comparable with other methods, while the final method's predictions are solely based on the RFscore$^{\mathrm{F}_{0.25}}$ random forest output threshold, which gives information about the certainty of the prediction.

We classified two helices as interacting if there exists at least one predicted residue contact between them. Here, too, we used the L/5 and L best predictions of residue pairs for comparison of MemConP with other methods. The set of helix interactions observed in known 3D structures was generated using the same approach, by classifying two helices as interacting if at least one residue pair was in contact according to the contact definition.

2.2.9 Feature importance

The randomForest package offers two possibilities to evaluate the importance of different features. The first measure reflects the average decrease of accuracy (or increase of error rate) when permuting the feature values of the predictor variable in question. The second measure used is the total decrease in node impurity measured by the Gini index, which is a value describing how well a variable serves for separating the two classes. It is defined by $1 - \sum_{i=1}^{m} g_i$, with g_i being the fraction of samples of the two classes (contact and non-contact) in the data at this node of the tree. We used the Gini index and the decrease of accuracy for the evaluation of feature importance on the final classifier trained on *NewTrain*. To assess the specific impact of the Freecontact method on the prediction performance, we also trained a random forest with all features except for the EC-values.

2.2.10 Performance values for other methods

We compared the performance of MemConP with several previously published techniques. We were not able to test MemBrain [105], reportedly the best performing method based on PSICOV [83] predictions, as it uses its own TMS definition and does not allow to submit user specified TM topology of a protein. We therefore use performance values provided in the respective publication. The precision and recall for PSICOV, Freecontact/EVFold, CCMpred [95], PconsC2 [86], MetaPSICOV [87], and MemConP were calculated using the topology given by PDBTM, which also made it possible to assess the predictive performance for the *NewTrain* and *NewTest* datasets. Except for Freecontact, all other methods failed to return a prediction for every target within 24 hours. Missing predictions were left out for the benchmark results of each specific method, probably leading to a performance overestimation. PconsC2 was completely omitted from the precision-recall plots as it was not able to return predictions for about 30% of the proteins, involving many cases for which the other methods performed below their average and thus can be considered as difficult targets for prediction.

2.3 Results and discussion

2.3.1 Model training

The random forest classifier makes use of several parameters for the training process, which are the number of sampled features, the number of training examples used for each tree, and the total number of trees in the ensemble *ntree*. In addition, it is possible to change the window size for each feature. As an exhaustive grid search over all these parameters would have been computationally prohibitive, we tried out a range of parameter combinations empirically. Except for *ntree*, the optimal values of the parameters controlling the random forest itself turned out to be their default values, as described in the *Materials and Methods* section. While the number of trees set to the default value of 500 already gave good results in terms of the error rate of the classifier, an increase of *ntree* to 2000, our final choice for this parameter, further improved the performance of the decision ensemble and stabilized it by introducing additional training sample/feature combinations, as every tree is built only by using a subset of them. The optimal window size for position-specific features (see *Materials and Methods*) was determined to be 9.

2.3.2 Variable importance

As described in the *Methods* section variable importance gives information about how much a feature contributes to the discrimination of the two classes by estimating the loss of accuracy when randomly permuting the feature values of the variable in question or the total increase in node impurity measured by the Gini index. Although these measures provide only a rough estimate, they give a useful overview over the ranking of individual features (Table 2.1).

Based on these criteria we found the absolute residue distance to be the most important MemConP feature apart from co-evolution, as evidenced by an average decrease of accuracy (Acc_{dec}) of 5.4% and a decrease in node impurity (Gini_{dec}) of 88.8. This can be explained by the fact that residues located in different TMSs, which are connected only by a short loop, have a high probability to be in contact. As Freecontact itself performs very well (see below), it is not surprising that its output value (reflecting contact propensity between a pair of residues) constitutes the most important feature, (Acc_{dec}: 2.4%, Gini_{dec}: 149.3). Interestingly, EC-values calculated by Freecontact for residue pairs situated four positions down and upstream from the central residue pair (*i.e.* in the positions which form hydrogen bonds with the central residue and are located on the same face of the helix, resulting in four possible pairs), are assigned a high importance (Gini_{dec} of 47.7 to 86.5, Acc_{dec} of 2.7 to 3.3). For comparison, the averages over all features regarding Acc_{dec} and Gini_{dec} values are 2.2% and 33.1 respectively.

Table 2.1: Variable importance for features or groups of features (bold face). For the features using a window or feature groups, the values are averaged.

Feature	Acc_{dec}	Gini_{dec}	**Feature**	Acc_{dec}	Gini_{dec}
Amino acid composition	2.2	17.3	**Evolutionary profil**e	1.8	20.5
Local EC-values	2.5	24.1	**Segment specific properties**	2.1	21.2
Pair EC-values	2.6	59.8	Aliphatic	2.4	21.8
4, 4	3.1	149.3	Aromatic	2.0	19.2
0, 0	2.4	149.3	C -branched	2.2	21.9
4, 0	3.3	86.5	Charge	1.7	20.9
0, 4	3.1	86.3	Hydrophobicdiscrete	2.0	21.9
0, -4	2.9	51.7	Hydrophobicitycontinuous	2.6	21.3
4, -4	3.3	48.3	Mass	2.4	21.9
-4, 0	2.7	47.7	Negative	1.9	20.9
-4, 4	3.2	45.0	pI	2.0	20.6
-4, -4	2.4	34.8	Polar	2.2	22.6
Predicted helix orientation	1.6	19.3	Positive	2.2	19.9
Entropy	1.8	22.6	Small	1.7	21.3
LIPS score of helix face	1.8	22.9	Volume	2.3	21.1
Rank of helix face	1.0	10.6	**SSEP**	2.2	20.6
LIPS score of residue	1.7	21.2	**Number of TMSs**	1.7	9.4
Amino acid properties	2.1	24.1	**Relative position**	1.86	24.17
Aliphatic	2.1	23.2	Relative residue position	2.5	29.9
Aromatic	2.3	22.6	Rel. N-boundary distance	1.3	25.9
C -branched	2.1	24.8	Rel. TMS number	1.8	16.8
Charge	2.3	23.6	**TMS length**	1.1	16.4
Hydrophobicdiscrete	2.0	24.5	**Residue separation**	3.4	46.5
Hydrophobicitycontinuous	2.0	23.5	Abs. residue distance	5.4	88.8
Mass	2.0	26.4	Abs. TMS distance	3.0	34.7
Negative	1.9	23.8	(Anti)parallel TMSs	1.9	11.4
pI	2.2	24.2	Rel. residue distance	3.7	63.56
Polar	1.8	23.6	Rel. TMS distance	3.12	34.16
Positive	2.2	22.9			
Small	2.0	25.1			
Volume	2.1	25.0			

2.3.3 Evaluation of residue contact prediction and comparison to existing methods on established datasets

To compare MemConP to other recent methods we used the same training and test datasets *OldTrain* and *OldTest* for evaluating PSICOV, Freecontact/EVFold, CCMpred, PconsC2, MetaPSICOV and MemBrain. According to the L/5 criterion the precision and recall values of MemConP are 72.5% and 12.3%, respectively, on the *OldTrain* dataset (Table 2.2), a 10.5%/2.1% increase compared to the claimed values of MemBrain, the reportedly best performing method for TMPs. Using the $\text{RFscore}^{\text{F}_{0.25}}$ threshold instead of the L/5 criterion, MemConP performs with a more than 2.5-fold increased recall value (31.1%) compared to MemBrain, at the cost of only a slight decrease in precision (72.4%). MCC for the RFscore$^{F_{0.25}}$ threshold is also much better (0.435) compared to 0.284 for the L/5 threshold. Similar results were obtained for the *OldTest* dataset (Table 2.2), for which the increase in precision compared to MemBrain is even higher (81.2% to 64.1%), while the gain in recall is relatively small (10.4% to 8.3%). Using the RFscore$^{F_{0.25}}$ threshold, the MCC increases to 0.446 on this dataset, significantly outperforming all other methods. The difference in performance between Freecontact (EVFold) and PSICOV is clearly in favor of Freecontact, which is conceivably one of the reasons for MemConP performing better than MemBrain on *OldTrain* and *OldTest*. The other methods designed for soluble proteins, namely CCMpred, PconsC2 and MetaPSICOV, perform remarkably well on the TM specific data, outperforming PSICOV and Freecontact and even MemBrain in most dataset/measure combinations. Unfortunately, some of the predictions could not be made by these methods, e.g. PconsC2 gave no result for 10 proteins from the *OldTrain* dataset, even after 24 hours of runtime.

All recent methods perform better using *OldTest* over *OldTrain*, which indicates that the contacts in these proteins are easier to predict. We tried to correlate Freecontact's predictive performance with several characteristics of the query proteins such as the relative/absolute number of sequences in the input alignment, number of TMSs and their lengths or the alignment entropy, but were not able to find the reason why these proteins seem to be so much easier to predict.

Table 2.2: Residue-residue contact prediction performance comparison. The '-'-column indicates the number of proteins for which no prediction was returned within 24 hours of runtime. These proteins were not considered when measuring the performance of the respective methods. Bold values indicate the highest performance for a given measure/dataset combination. Despite the large amount of missing predictions PconsC2's performance is shown for completeness, but is greyed out.

Method	**Threshold**	**P**	**R**	**$F_{0.25}$**	**MCC**	**-**	**P**	**R**	**$F_{0.25}$**	**MCC**	**-**
		OldTrain					*OldTest*				
PSICOV	L/5	42.8	6.6	31.7	0.157	60	58.0	7.9	41.6	0.206	21
Freecontact	L/5	58.1	9.7	43.4	0.222	61	70.4	9.3	50.0	0.247	21
CCMpred	L/5	64.5	10.8	48.1	0.248	60	77.8	10.3	55.3	0.275	21
PconsC2	L/5	56.8	9.2	42.5	0.216	51	80.7	9.8	56.2	0.275	15
MetaPSICOV	L/5	64.8	11.5	48.7	0.254	58	76.2	9.9	54.0	0.267	20
MemBrain	L/5	62.0	10.2	-	-	-	64.1	8.3	-	-	-
MemConP	L/5	72.5	12.3	54.5	0.284	61	81.2	10.4	57.3	0.284	21
PSICOV	L	22.7	17.5	22.1	0.174	60	30.1	19.8	29.1	0.227	21
Freecontact	L	34.0	26.4	33.2	0.276	61	41.1	26.6	39.6	0.314	21
CCMpred	L	35.8	28.0	34.9	0.291	60	45.5	29.5	43.9	0.350	21
PconsC2	L	36.3	27.9	35.5	0.295	51	50.4	30.1	48.3	0.375	15
MetaPSICOV	L	41.2	32.7	40.2	0.341	58	51.0	32.4	49.1	0.392	20
MemConP	L	52.5	42.0	51.3	0.444	61	59.8	37.6	57.5	0.460	21
MemConP	L/x$^{F_{0.25}}$	65.2	25.0	58.5	0.384	61	72.7	21.4	63.2	0.384	21
MemConP	RFscore$^{F_{0.25}}$	72.4	31.1	61.2	0.435	55	74.5	33.2	61.4	0.446	21
		NewTrain					*NewTest*				
PSICOV	L/5	46.0	7.0	33.9	0.169	87	42.3	5.9	30.0	0.147	30
Freecontact	L/5	57.3	8.7	42.2	0.213	90	54.1	8.1	38.9	0.196	30
CCMpred	L/5	65.4	11.0	48.2	0.247	86	61.3	9.4	44.2	0.226	29

PconsC2	L/5	60.6	9.1	44.6	0.224	73	64.7	9.7	47.1	0.238	20
MetaPSICOV	L/5	62.4	9.5	46.1	0.233	87	58.7	8.1	41.7	0.207	30
MemConP	L/5	70.9	10.9	52.4	0.268	90	65.6	9.7	47.0	0.239	30
PSICOV	L	24.4	17.9	23.8	0.188	87	21.3	14.6	20.5	0.155	30
Freecontact	L	32.1	24.8	31.2	0.257	90	29.2	20.9	28.1	0.224	30
CCMpred	L	35.8	27.1	34.8	0.286	86	31.9	22.7	30.8	0.246	29
PconsC2	L	36.8	26.7	35.8	0.293	73	37.4	26.3	36.2	0.292	20
MetaPSICOV	L	39.5	28.9	38.4	0.318	87	36.8	24.6	35.4	0.280	30
MemConP	L	48.3	37.2	47.0	0.400	90	41.0	28.7	39.5	0.320	30
MemConP	L/x$^{F_{0.25}}$	64.7	18.5	55.6	0.333	90	57.8	15.6	48.5	0.284	30
MemConP	RFscore$^{F_{0.25}}$	63.6	25.7	53.8	0.369	89	56.0	17.5	45.7	0.288	29

2.3.4 Benchmarking the residue contact predictor trained on latest data

The use of improved methods, such as HHblits and Freecontact, combined with a newer dataset containing 30% more TMP chains resulted in an improved generalization power of the model. On the *NewTrain* dataset precision and recall of MemConP reach 63.6% and 25.7%, respectively, using the RFscore$^{F_{0.25}}$ threshold, and 70.9%/10.9% using the L/5 criterion (Table 2.2).

The random forest trained on the *NewTrain* dataset with 90 proteins, which is the final released predictor, achieves results on *NewTest* (P/R/MCC: 65.6%/9.7%/0.239), using the RFscore$^{F_{0.25}}$ threshold, the performance reaches 56.0% precision, 17.5% recall and a MCC of 0.317. In general, a good indicator to check whether the model is overtrained or generalizes well is to compare the performance between the cross-validation and the independent test data (*NewTrain* and *NewTest*). An overtrained model would perform worse on independent data, provided that these data are non-redundant. Whilst there is indeed a drop in MCC values between the *NewTrain* and *NewTest* datasets (0.268 vs. 0.239), this is also true for the methods not employing any training. We speculate that this drop in performance is due to the rigorous redundancy reduction procedure employed to create the *NewTest* dataset (see *Methods*), which resulted in a higher fraction of difficult prediction targets. It should also be noted that Freecontact substantially outperforms PSICOV, which makes it a better choice as input for our method (Table 2.2). Again, among the more recent contact prediction tools, CCMpred and MetaPSICOV perform fairly well on the latest datasets. Based on the list of predictions sorted by the random forest output values the trained classifier was capable of sustaining higher precision values than the other methods over the whole recall range for the L/x best predictions (Figure 2.1). The problem of missing predictions is still evident especially for PconsC2, which was not able to process one third of the *NewTest* dataset. These 10 missing cases were some of the most difficult prediction targets for the other methods as well, which is also discussed briefly below. Therefore PconsC2 was not included in Figure 2.1, as it would give a distorted view of its real life performance.

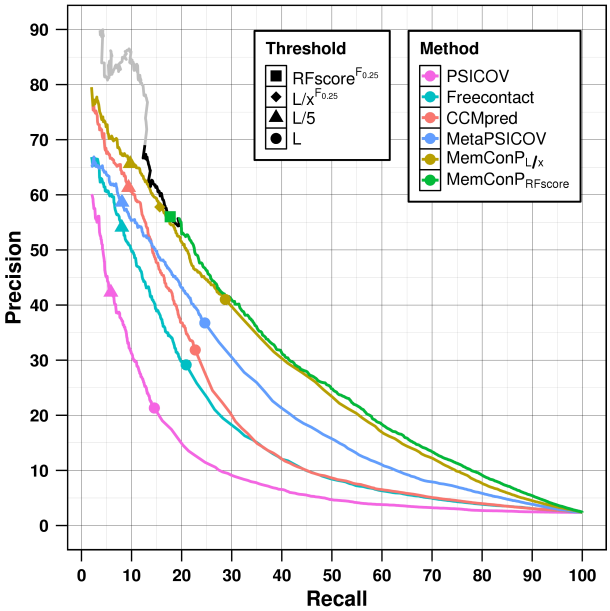

Figure 2.1: Precision-recall curve created by sorting all predictions of residue contacts according to their score, which is either the output value of the random forest (RFscore) for MemConP or the top L/x predictions for MemConP and the other methods for each protein, and averaging over them. The shapes show the performance at different thresholds (RFscore$^{F_{0.25}}$, L/x$^{F_{0.25}}$, L/5 and L). The black line indicates the range of positions, where at least one positive MemConP prediction per protein is present for more than 90% but not for all of the proteins in the dataset, while data points on the grey line do not meet the 90% criterion. Protein/method combinations where no prediction could be made were not included in performance calculation.

A particular combination of the precision and recall values should be selected by the users depending on the task in hand. For example, a user who intends to perform costly mutagenesis experiments on contacting residues, might want to use only a few of the most reliable predictions while accepting to miss most of the true positives. On the other hand, for those users interested in investigating the abundance of possible contacts between different pairs of helices obtaining as many true positives as possible will be the top priority, even at the expense of higher false positive rate.

2.3.5 Impact of updated training data

To assess the benefit of the updated datasets and employing the random forest, we trained the random forest as a stand-alone method without using any co-evolution information (*i.e.* without the EC-values, denoted *NoEC*) on *OldTrain* or *NewTrain* and then tested it on *NewTest*. Although we tried to keep the *NewTest* dataset independent from *OldTrain*, only 21 out of the 30 *NewTest* proteins have TM-score below 0.5 when compared to *OldTrain*. Only these 21 proteins were used for this analysis.

Interestingly, MemConP$_{\text{OldTrain/NoEC}}$, the basic predictor trained on *OldTrain* without the Freecontact feature, still performs remarkably well with 27.9% precision, 5.5% recall and a MCC of 0.111 (Figure 2.2), especially when compared to earlier methods, which only incorporated basic co-evolution measures such as mutual information. TMHcon, for example, has the reported precision/recall values of 14.8%/3.9%. Upon switching from *OldTrain* to *NewTrain* while still not using any co-evolution information, there is an increase in performance to a precision and recall of 32.0% and 5.1% respectively, which implies an improved generalization power of the model trained on the *NewTrain* dataset containing 30% more proteins. For comparison, PSICOV, which is a second generation co-evolution method not employing any training, achieves 36.3% precision and 5.9% recall on these 21 proteins. As expected, incorporation of co-evolution results in the most significant improvement, raising precision/recall using *OldTrain* to 57.7%/10.2% and using *NewTrain* to 60.3% and 10.4% respectively.

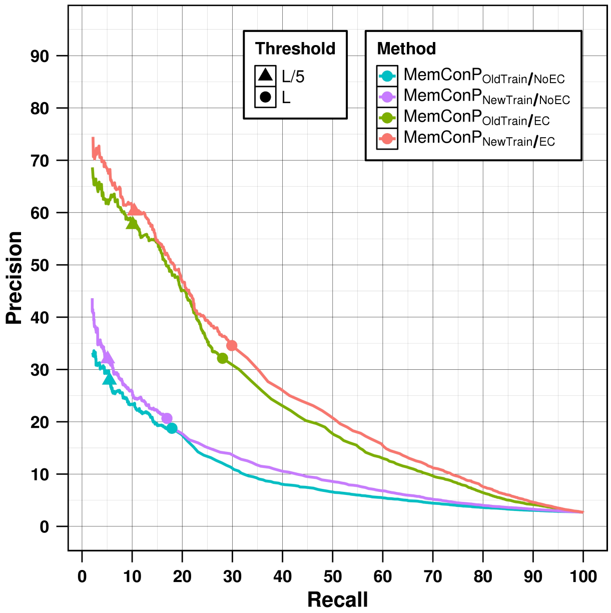

Figure 2.2: Precision-recall curve for the 15 protein chains in the *NewTest* dataset, which exhibit no similarity to any proteins in *OldTrain* and *NewTrain*. The plot shows the performance of the random forest models using each of the two training datasets (*OldTrain*/*NewTrain*) with or without employing co-evolution as an input feature (EC/NoEC).

2.3.6 Dependence of the performance on the number of transmembrane segments and alignment size

We assessed the performance of all methods depending on two properties: the size of the alignment and the number of TMSs. As seen in Figure 2.3, there is no clear correlation between the alignment size and the prediction performance. All methods perform above their average between 7 and 11 or more than 61 thousand sequences and suffer a performance drop between 13 and 18 thousand sequences. MemConP, for example, has an average L/5 precision of 65.6%, but achieves 100% in the alignment size range of 10-11 thousand proteins, while for 17-18 thousand aligned proteins its accuracy is only slightly above 45%. However, both of these alignment size ranges contain only one prediction target, and are thus not statistically representative. In general, all alignment size intervals up to 7-8 thousand proteins contain more than two proteins, while larger alignments correspond to one or two prediction targets. Notably, while most methods follow a similar trend for each alignment size, MemConP sets itself apart for alignment sizes of 4 to 6 thousand aligned proteins. A more detailed view of the relation between alignment size and the prediction performance without the aggregation of individual prediction results can be found in Figure 2.4.

Similar to the alignment size, there is no obvious general trend in the dependence of performance values on the number of TMSs (Figure 2.5). Precision does increase slightly with the increasing number of TMSs, but this is presumably due to the use of the L/5 measure: while the number of possible interactions increases exponentially upon adding each TMS, the length L only increases linearly. Especially for proteins containing 3 to 6 TMSs, all methods perform worse than for proteins with 7 TMSs or more. MemConP follows this trend in that it performs above its average L/5 precision (65.6%) in all cases with more than 4 TMSs, except for the proteins with 16 TMSs. The figure also exposes PconsC2 problems with more difficult prediction targets, as in most cases where it seems to beat the other methods (e.g. for 5, 9, 12 TMSs), it is missing some predictions (smaller sizes of points in Figure 2.5 corresponding to fewer prediction targets). Another interesting feature of the *NewTest* dataset made apparent by Figure 2.5 is that prediction targets are relatively evenly distributed over TMS bins, with most of the bins containing two test examples.

2.3.7 Evaluation of helix-helix interaction prediction

MemConP identifies interacting helices in TMPs by considering the highest scoring pairs of contacting residues. Unlike the residue contacts, the performance for helix interac-

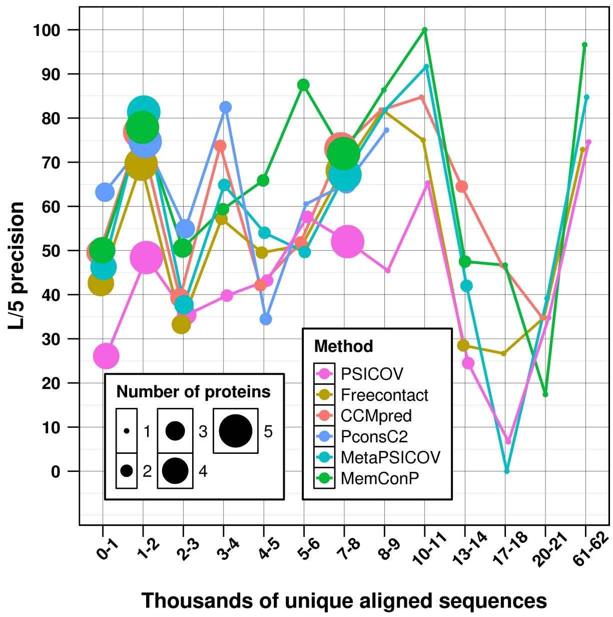

Figure 2.3: Dependence of precision on the number of sequences in the alignment for all prediction methods. The diameter of the points indicates the number of proteins with alignments in the respective alignment size range. Small displacements of points were introduced in order to better distinguish them from each other.

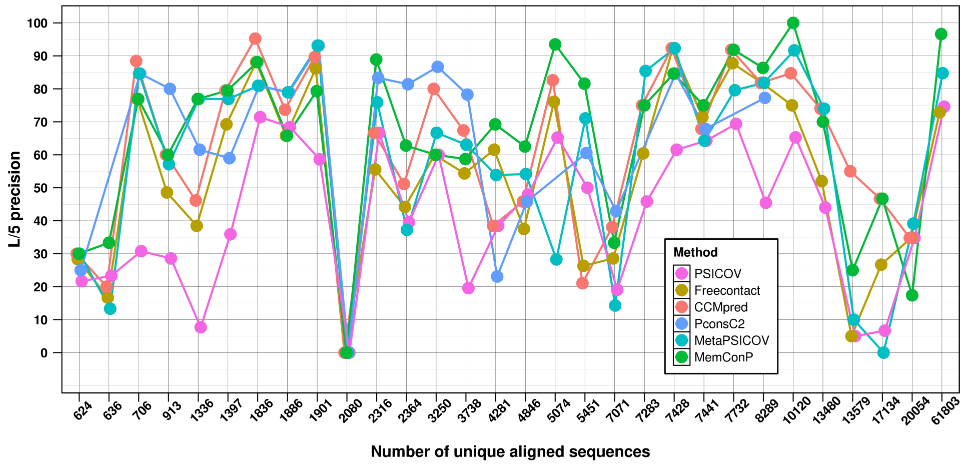

Figure 2.4: Dependence of precision on the number of sequences in the alignment for all prediction methods. Small displacements of points were introduced in order to better distinguish them from each other.

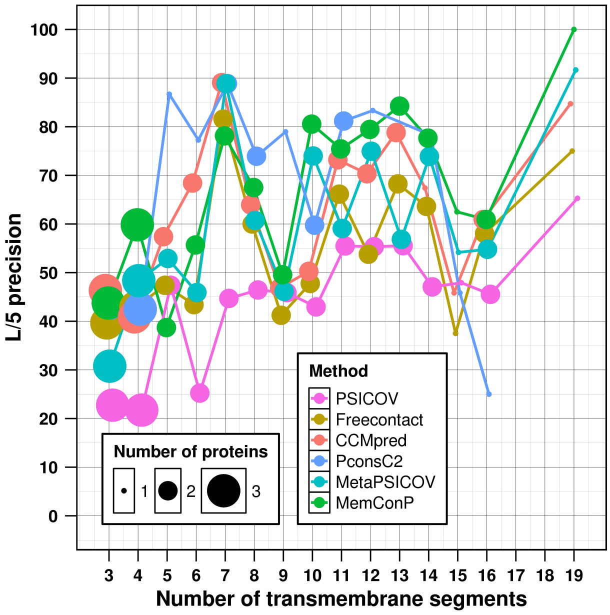

Figure 2.5: Dependence of precision on the number of TMSs in proteins for all prediction methods. The diameter of the points indicates the number of proteins with the respective number of TMSs. Small displacements of points were introduced in order to better distinguish them from each other.

tions is highly dependent on the chosen threshold and benchmark measure (Table 2.3). MemConP achieves 98.8% precision and 49.4% recall on *OldTest* using the L/5 best predictions. PSICOV, on the other hand, also performs well using the same threshold (precision: 81.2%, recall: 64.7%), but achieving only a precision of 58.4% when using the L best predictions, a value that would be expected from a random predictor. All other benchmarked methods show comparable performance, mostly gaining a higher recall compared to MemConP at the cost of slightly decreased precision. Using the RFscore$^{F_{0.25}}$ threshold, MemConP is even able to achieve 100% precision at 43.3% recall. MemConP is clearly superior to all other contact prediction methods on the *NewTest* dataset. In terms of helix interaction prediction MemConP shows a comparable performance with CCMpred and Freecontact. Its ranks first in terms of L and L/5 precision (92.2% and 72.6%, respectively) at the cost of a decreased recall (43.3% and 75.3%, respectively). In particular, CCMpred achieves a higher recall (L/5: 52.8%, L: 80.6%) and slightly lower but comparable precision (L/5: 91.7%, L: 65.6%) when predicting helix interactions. But since the method stays behind MemConP for predicting residue contacts, it can be assumed that the high ranking predicted residue contacts are more evenly distributed among the interacting helix pairs. Figure 2.6 confirms this assumption, with the precision-recall curve of CCMpred being slightly above MemConP. PSICOV has the highest recall (83.9%) accompanied by a rather low precision (55.6%) compared to the other methods. Aside from MemConP and CCMpred, Freecontact achieves comparable performance predicting helix interactions, while MetaPSICOV lags slightly behind these three tools. Just like in the case of residue contact prediction, PconsC2 could not be benchmarked properly on helix interactions because of missing predictions.

Table 2.3: Comparison of helix-helix interaction prediction methods. Bold values indicate the highest performance for a given measure/dataset combination. Despite the large amount of missing predictions PconsC2's performance is shown for completeness, but is greyed out.

Method	**Threshold**	**P**	**R**	**$F_{0.25}$**	**MCC**	**P**	**R**	**$F_{0.25}$**	**MCC**
		OldTrain				*OldTest*			
PSICOV	L/5	81.4	66.4	80.4	0.532	81.2	64.7	80.0	0.515
Freecontact	L/5	87.5	63.5	85.6	0.577	95.3	60.2	92.1	0.619
CCMpred	L/5	90.0	60.4	87.5	0.579	94.0	61.4	91.2	0.618
PconsC2	L/5	89.4	59.0	86.8	0.555	96.2	57.3	92.5	0.608
MetaPSICOV	L/5	88.7	50.6	84.9	0.493	91.8	50.3	87.6	0.519
MemBrain	L/5	90.1	56.2	-	0.555	87.9	56.3	-	0.526
MemConP	L/5	94.9	49.5	90.0	0.542	98.8	49.4	93.3	0.567
PSICOV	L	57.0	91.9	58.3	0.320	58.4	93.7	59.7	0.360
Freecontact	L	66.0	88.4	67.0	0.480	71.0	84.4	71.7	0.520
CCMpred	L	67.7	88.2	68.6	0.508	70.4	86.2	71.2	0.524
PconsC2	L	73.0	81.2	73.4	0.517	75.4	83.2	75.9	0.579
MetaPSICOV	L	74.0	82.1	74.4	0.544	71.6	80.8	72.1	0.502
MemConP	L	77.5	81.6	77.7	0.596	79.8	80.5	79.9	0.609
MemConP	$L/x^{F_{0.25}}$	94.5	53.0	90.3	0.564	98.9	52.4	94.0	0.591
MemConP	$RFscore^{F_{0.25}}$	98.4	44.3	91.8	0.529	100.0	43.4	92.9	0.530
		NewTrain				*NewTest*			
PSICOV	L/5	81.2	62.7	79.9	0.524	81.5	56.1	79.3	0.509
Freecontact	L/5	89.8	54.8	86.6	0.551	90.1	52.6	86.4	0.555
CCMpred	L/5	89.6	55.9	86.6	0.556	91.7	52.8	87.9	0.568
PconsC2	L/5	92.3	51.1	88.1	0.536	95.0	47.8	89.8	0.536
MetaPSICOV	L/5	88.2	48.7	84.2	0.496	88.2	40.8	82.5	0.459
MemConP	L/5	91.4	46.5	86.5	0.505	92.2	43.3	86.5	0.504
PSICOV	L	56.1	90.0	57.4	0.338	55.6	83.9	56.7	0.361
Freecontact	L	67.4	78.6	67.9	0.465	67.8	77.0	68.3	0.499
CCMpred	L	65.8	81.0	66.6	0.455	65.6	80.6	66.3	0.491
PconsC2	L	76.9	76.7	76.9	0.559	80.9	72.1	80.3	0.578
MetaPSICOV	L	70.1	76.5	70.5	0.490	72.6	72.5	72.6	0.526
MemConP	L	71.6	78.7	71.9	0.522	72.6	75.3	72.8	0.544
MemConP	$L/x^{F_{0.25}}$	91.0	49.9	86.8	0.526	90.6	46.3	85.8	0.514
MemConP	$RFscore^{F_{0.25}}$	94.6	44.5	88.7	0.514	91.4	42.1	85.5	0.490

An example of a helix interaction prediction and its dependence on the chosen helix interaction threshold is given in Figure 2.7A. In total, 10 out of 14 contacting helices are predicted correctly using the $RFscore^{F_{0.25}}$ threshold, despite the low number of true positive contact predictions for some of the helix pairs. Notably, there is neither a predicted residue contact, nor a predicted helix interaction between helix 2 (light green) and helices 5, 6 and 7 (cyan, yellow and dark green), which appear to actually interact with helix 2 based on the structure *4pgr* chain A. In order to investigate possible reasons for this

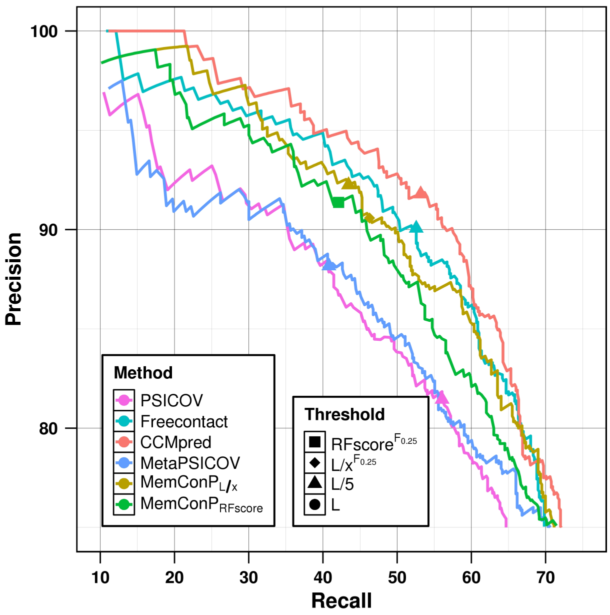

Figure 2.6: Comparison of precision-recall curves created by sorting all predictions of helix interactions according to their score, which is either the output value of the random forest (RFscore) for MemConP or the top L/x predictions for MemConP and the other methods for each protein, and averaging over them. The shapes show the performance at different thresholds (RFscore$^{F_{0.25}}$, L/x$^{F_{0.25}}$, L/5 and L).

false negative prediction we superimposed *4pgr* with another PDB entry for the same sequence - *4pgs* chain A (Figure 2.8). Both structures match perfectly except for the second helix, which does not make any contacts to helices 5, 6 and 7 according to *4pgs*. The position of helix 2 thus appears to be ambiguous, possibly due to either crystallization artifacts or to or to functionally relevant conformational mobility. Comparison of helix interaction graphs shown in Figure 2.7A makes apparent the advantage of using the RFscore$^{F_{0.25}}$ threshold. Compared to the RFscore$^{F_{0.25}}$ threshold, with its 10 true positive and one false positive prediction, using the L best predictions as the threshold generates four false positive interactions, while using the more stringent L/5 threshold misses four true positives interactions. The second example (Figure 2.7B) highlights the advantages and drawbacks of different thresholds as well. Even using the RFscore$^{F_{0.25}}$ threshold, which is geared towards high precision, two false positive interactions between helices 2/7 and 2/9 are produced whereas using the L/5 highest scoring predictions only results in one of these false positives (2/7). In this specific case, however, this interaction would in fact be a true inter-molecular interaction if contacts between all subunits of the structure, which is a homotrimer, were investigated.

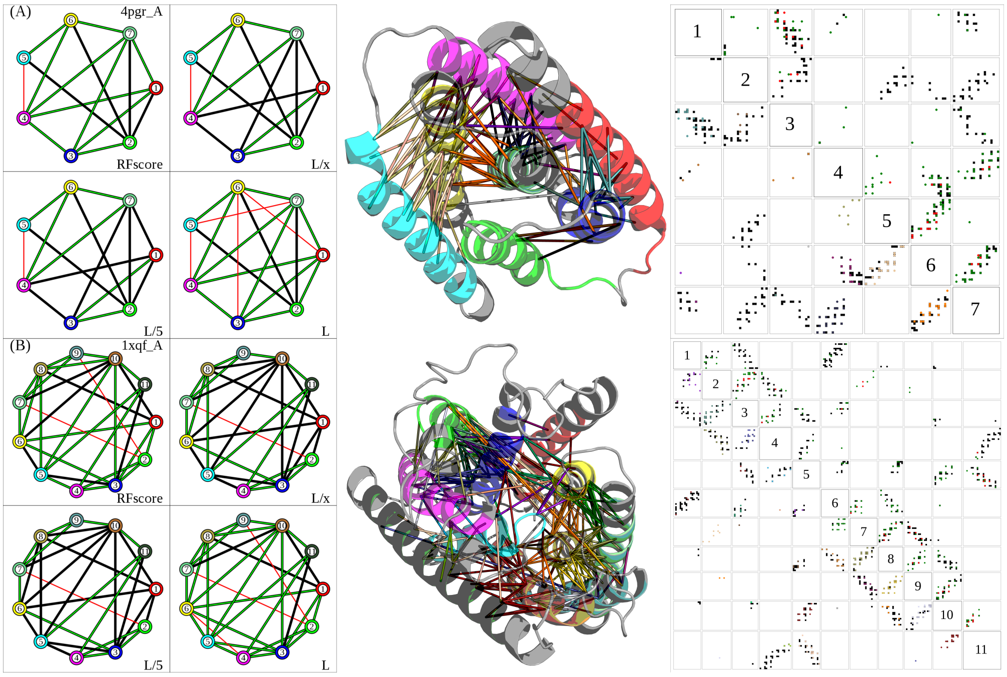

Figure 2.7: Visualization of contacts and helix-helix interactions predicted by MemConP at different thresholds by helix interaction graphs (left), ribbon models of structures (center), and contact maps (right) for the *Bacillus subtilis* pH-sensitive calcium leak channel YetJ (*4pgr*, chain A) (A) and for the ammonium transport protein AmtB from *Escherichia coli* (*1xqf*, chain A) (B). In the helix interaction graphs, black edges depict the observed contacts as extracted from the PDBTM database. Green edges correspond to true interactions which were also predicted by our method, while red lines indicate false positives, *i.e.* interactions predicted by MemConP, which are not observed in the experimental structure. α-helices, represented as circles on the helix interaction graphs, are encoded by the same color as in the structure image. On the contact maps, black dots are the observed contacts as extracted from the structure. Colored dots are contacts predicted by MemConP according to different thresholds: top L predictions (green, upper right triangle), top L/5 predictions (red, upper right triangle), RFscore$^{F_{0.25}}$ threshold (various colors, lower left triangle). The dots representing the predicted contacts at the adjusted threshold are drawn in the same color as lines in the structure image connecting residues predicted to interact. For example. the orange dots in the lower left triangle (A) between helices 6 and 7 are clearly visible between the corresponding helices (colored in yellow and dark green) in the structure image.

Similar to the residue contact prediction, threshold plots can be used to estimate the trustworthiness of a particular prediction by assessing performance measures at different thresholds (Figure 2.6).

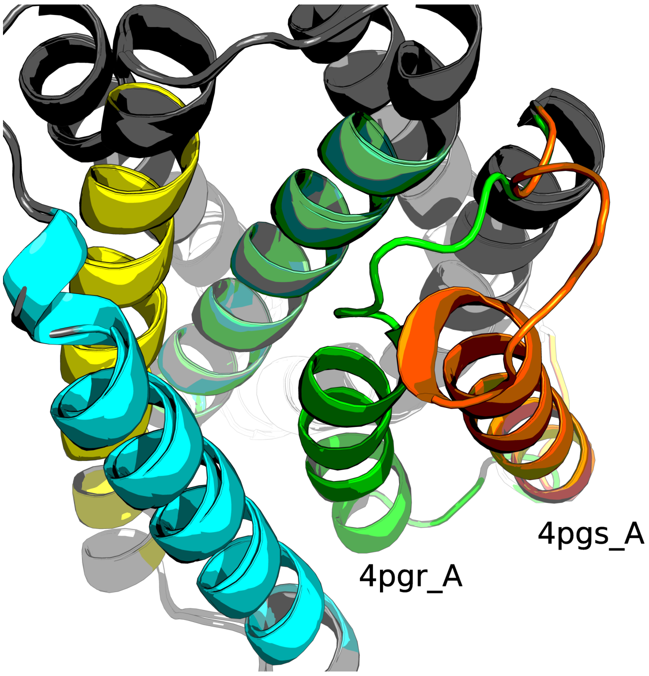

Figure 2.8: Superposition of two alternative structures for the *Bacillus subtilis* pH-sensitive calcium leak channel YetJ. The structures match perfectly except for the localization of the second TMS colored in light green (*4pgr*, chain A) and orange (*4pgs*, chain A), which explains the missing contact predictions between the second and the other three helices colored in cyan, yellow and dark green.

2.3.8 Prediction of helix interaction patterns using sequence derived 2D-topology

The positions of TMSs used for benchmarking of our methods were derived from known 3D structures. In real use cases these structures are not known and thus the TMSs have to be predicted from sequence. We used for this purpose PolyPhobius, a highly rated [106] hidden Markov model-based topology prediction method using multiply aligned sequences as the sole input. Because the true TM topology is dependent on its definition and may therefore vary across different structure databases by several residues, we mapped each predicted TMS to the one derived from structure using the segment overlap (SOV) approach [107]. SOV between a predicted and a true TMS was calculated as

$$SOV(predicted, true) = \frac{overlap(predicted, true)}{max(predicted_{last}, true_{last}) - min(predicted_{first}, true_{first}) + 1}$$

where *overlap* is the number of shared, *i.e.* correctly predicted TM residues, and *first/last* are the start and end positions of the *true/predicted* TMS. For a successful mapping, we required the SOV to be at least 0.5. Out of the 30 proteins from *NewTest* only 6 could be mapped completely, *i.e.* each of the predicted TMSs had a corresponding structure-derived segment with a SOV greater or equal than 0.5, which means that there were no additional or missing segments. Taking only these 6 proteins into account, the performance using the predicted topology is close to the performance using the structure derived segments with a precision of 95.3%, a recall of 48.2% and a MCC of 0.570. For the 24 proteins having wrong predicted topology, unmapped TMSs can either result in false positive helix interactions, in case an unmapped segment is predicted to interact with another helix, or in a true negative interaction. Conversely, a 3D-structure derived TMS involved in an interaction, but missed by prediction, would result in a false negative interaction. Although 80% of the topologies were not predicted correctly, the performance drop is very low (P: 75.8%, R: 32.2%, MCC: 0.359), which demonstrates that MemConP is capable of producing accurate predictions based on amino acid sequence information.

2.4 Conclusions

We have developed a novel method for predicting intramolecular interactions in membrane proteins. MemConP predicts individual residue-residue contacts and helix interactions using correlated mutations and sequence features. The method heavily relies on the recent co-evolution algorithms, which eliminate transitive evolutionary connections between residues. Figure 2.9 summarizes the key methodological developments in predicting intramolecular interactions in membrane proteins, by showing how the improvements in three areas - training data, machine learning and co-evolution methods - impact prediction performance. While contact prediction methods for soluble proteins were established in mid-90s and have matured by now, the first approaches achieving a comparable performance for membrane proteins appeared more than a decade later. HelixCorr [80] was a predictor based on the combination of several first generation co-evolution methods. The follow-up method TMHcon [81] additionally exploited a neural network-based machine learning approach. In this work we demonstrate that a baseline machine-learning predictor, which utilizes a highly sensitive database search method (HHBlits), surpasses these early methods on the same training dataset, even though it does not consider residue co-evolution. However, it still lags behind the modern co-evolution methods, such as Freecontact [84], which employ global statistical models to infer residue coupling in the entire sequence alignment. On a much more diverse training dataset the random forest classifier edges closer to the more recent methods. Our final method, which combines machine learning with the co-evolution signal as an additional input feature, outperforms all currently available methods by a wide margin.

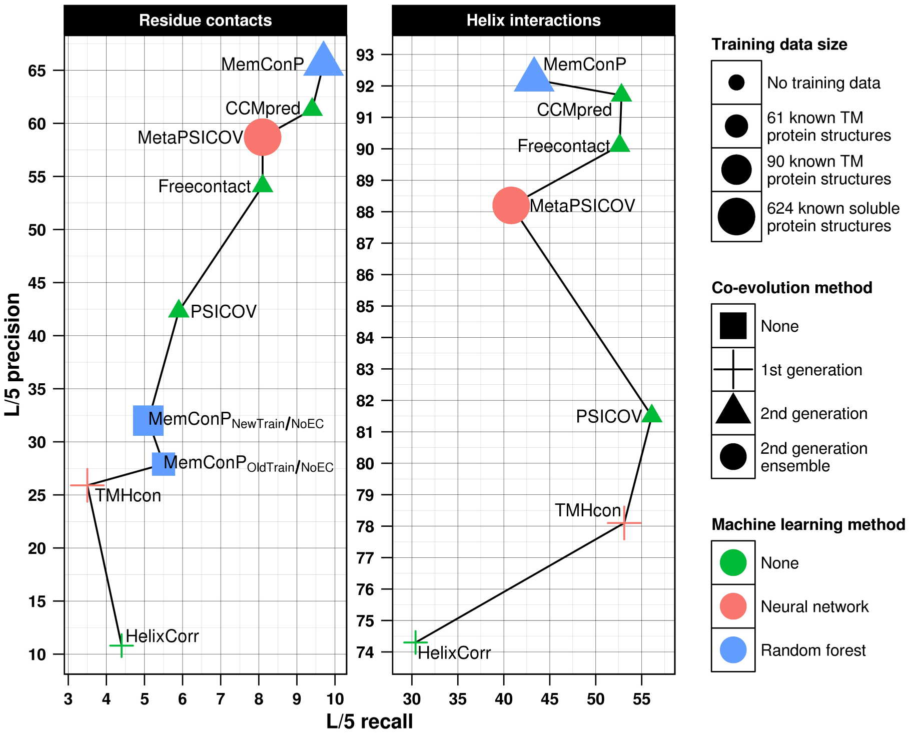

Figure 2.9: Illustration showing the progress of co-evolution methods, training data, machine learning and their combinations in relation to the corresponding benefits in residue contact and helix interaction prediction performance. For each combination an example method is given. Shapes (crosses, triangles, squares, circles) indicate the particular type of the co-evolution method applied. Colors depict machine learning approaches utilized, from earlier methods based on neural networks to more recent random forest models. The size of the symbol indicates the dataset used for training the classifier.

Building on this unmatched performance in predicting individual residue contacts, we predicted helix interactions as well. In contrast to all previous studies we conducted benchmarking of MemConP not only on TM helix positions obtained from crystal structures, but also on TM topologies predicted from sequence, which would be the normal use case for this sequence-based method. Although less than a third of our test cases had the correctly predicted topology, the quality of helix interaction predictions did not noticeably deteriorate compared with using 3D structure-derived helix positions. Our approach can thus be used to obtain valuable insights into TMP folds from sequence alone.

3 Evolutionary interplay between symbiotic relationships and patterns of signal peptide gain and loss

In this publication, we investigated orthologous groups of proteins from the *Enterobacterales* order to find cases of SP gain and loss events, *i.e.* if orthologous proteins with different SP assignments exist, and how they evolved.

Therefore, we prepared a dataset deriving COGs from the OMA and scanned the proteins for SPs using a combination of three prediction methods, SignalP, Phobius and TatP in order to receive a high quality dataset in terms of reliability of the SP assignments.

To be able to distinguish between SP gain and loss events and to bring these events into the evolutionary context, we calculated a phylogenetic tree for each orthologous group and did a standard parsimony analysis, *i.e.* reconstructed the ancestral states, using the Fitch method. An additional parsimony analysis was conducted after a gene start correction procedure, applied to reduce the number of false positive events.

As a result, we found out that 1.9% of all investigated COGs contain both, proteins with and without SPs, and, therefore, must have undergone either a SP gain or loss event. The two subsequent conducted parsimony analyses showed that SP loss events happen more often than gain events. In addition, we found the tendency of gains to be more ancient while loss events seemed to have happened more recently.

Looking into the mechanics behind the events showed mainly two possible patterns. The first is the complete deletion or insertion of the SP, the second is the mutation of amino acids while retaining the full length of the N-terminal sequence.

Finally, we tried to relate the lifestyle of the bacteria from the *Enterobacterales* order to their SP content. As a result, we were able to show that in some COGs the SP assignments were enough to discriminate between endosymbionts and free-living or commensal bacteria. Additionally, we found out that SP loss events correlate with a change of lifestyle as these events are often accompanied by a transition towards an endosymbiotic lifestyle.

Peter Hönigschmid conducted the whole experiment and wrote the manuscript. The original project was initiated by Nadya Bykova, Dmitry Ivankov and Dmitrij Frishman. René Schneider did some preliminary analyses during his Bachelor thesis. Dmitrij Frishman supervised and planned the project and wrote the manuscript.

3.1 Introduction

Protein function is not set in stone – it can undergo both gradual and drastic changes due to a variety of evolutionary events, including mutations, insertions, deletions, and duplications. Early on it was noted that proteins sharing the same structural fold can have vastly divergent functional roles [41]. Although functional equivalence of orthologs is often assumed, recent assessments indicate a rather low degree of functional similarity between pairs of orthologous genes [42], even when they share very high overall sequence identity [43]. Specific aspects of proteins function may vary between orthologs significantly, including enzymatic specificity [44] and protein interaction sites [45]. Local molecular determinants of protein function, such as phosphorylation sites, as well as entire protein domains, can be gained and lost in the course of evolution.

While the evolutionary dynamics of enzymatic and binding activities has been extensively studied, functional shifts associated with the evolution of cellular targeting signals have received much less attention, and most of the work done so far focused on the sequence diversity of eukaryotic SPs, mitochondrial targeting signals, and chloroplast transit peptides [47, 48, 46]. In particular, differences in the evolutionary rates between intra- and extracellular proteins have been reported for mammals and yeast [108, 109], and shown to depend on tissue-specific gene expression [110]. In bacteria, the majority of the secreted proteins (96% in *Escherichia coli* [1]) are translocated across the cytoplasmic membrane in a Sec-pathway-dependent manner and possess cleavable SPs – short sequence segments of 20-30 amino acids in length, which act as targeting signals [3]. SPs exhibit a tripartite structure, consisting of a positively charged N-terminal region, a central hydrophobic region, and a polar C-terminal region, which contains a three-residue cleavage motif recognized by the signal peptidase I [111]. The limits of sequence variation within SPs have been extensively studied [112] and a large number of non-conventional taxon-specific sequences have been identified by proteogenomic experiments [113]. However, all these studies were primarily aimed at understanding the minimal sequence requirements of the signal recognition machinery and did not consider evolutionary effects associated with elimination or acquisition of SPs.

Given the importance of SPs for protein sorting and localization it is no wonder that they constitute an important element of protein and genome annotation. Early analyses of completely sequenced genomes suggested that around 20% of proteins are secreted in a typical bacterium, such as *Haemophilus influenzae* [15] or *Escherichia coli* [14]. More recently these estimates have been revised due to both improved accuracy of bioinformatics predictions [18] and the availability of proteogenomics data [20, 19], and for the best studied bacterium *Escherichia coli* they currently converge to 10% of proteins possessing a SP [16]. The size and the composition of the secretome are highly informative for understanding organism's physiology. An important driving force for functional divergence in bacteria is constituted by environmental variation and the ensuing changes of lifestyle. In general, pathogenic and non-pathogenic species would be expected to secrete different proteins [49], but a recent study [50] failed to establish any connection between pathogenicity and the secretome size. A positive correlation between the percentage of secreted proteins and the number of genes in the gram-negative, but not in the gram-positive organisms, was reported.

Here we present a comparative secretome analysis of *Enterobacterales*, focusing not only on the relative number of secreted proteins, but also on the conservation of their ability to be secreted in relation to the bacterial lifestyle. In order to conduct this analysis, we integrated evolutionary trees of orthologous protein groups with SP predictions and functional annotation. Parsimony analysis and sequence comparisons revealed a large number of SP gain and loss events, in which SPs emerge or disappear amongst orthologous proteins in the course of evolution. We also attempted to shed light on the molecular mechanism leading to these events and their relationship to the symbiotic lifestyle of an organism. Our results indicate that SP losses prevail over gains, an effect which is especially pronounced in the transition from the free-living or commensal to the endosymbiotic lifestyle. The disproportionate decline in the number of SP-containing proteins in endosymbionts cannot be explained by the overall reduction of their genomes [114]. SPs can be gained and lost either by acquisition/elimination of the corresponding N-terminal regions or by gradual accumulation of mutations.

3.2 Materials and Methods

3.2.1 Genomes, orthologous clusters, and Gene Ontology terms

The *Enterobacterales* order is a large and diverse group of gram-negative bacteria within the class *Gammaproteobacteria*. Its taxonomic tree has been recently updated and refined [115]. This group, to which the best studied model organism *Escherichia coli* also belongs, contains bacteria occupying a variety of habitats and involved in diverse kinds of symbiotic relationships. The taxonomic identifiers of these organisms were extracted from the National Center for Biotechnology Information (NCBI) taxonomy database [116] in November 2016. The corresponding genomes were downloaded either from the European Nucleotide Archive (ENA) [117] or the EnsemblGenome database [118]. *Enterobacterales* COGs with associated GO-terms [119, 120] were retrieved from the OMA orthology database in June 2016 [121]. The resulting dataset contains 626680 proteins from 153 distinct species, of which 557556 proteins are mapped onto 24837 orthologous clusters.

3.2.2 Evolutionary trees

Evolutionary trees for all clusters were built with PhyML 3.0 [122] using MSA of cluster members as input. MSAs were computed by Clustal Omega [123] with the default parameters. As PhyML only produces unrooted trees, which do not provide any information about the direction of evolution, we rooted the tree using the midpoint rooting method, which takes the longest distance between two leafs in the tree, and inserts the root at the exact midpoint between them. Since at least three proteins are required to calculate an evolutionary tree, clusters with one or two members were not considered.

3.2.3 Signal peptide data

SPs were identified in the *Enterobacterales* gene products based on three data sources with a different degree of confidence. First, SPs were predicted by the latest and most accurate version of the SignalP (SignalP 4.1 [18]) software with all default parameters using the gram-negative model. In addition, SPs were predicted by Phobius [14, 28], which, in contrast to SignalP, returns discrete predictions rather than scores.

As we focus on Sec-mediated protein secretion, we used TatP [124] to remove COGs containing proteins utilizing the Tat pathway.

Results of these three methods were combined to derive a consensus prediction with four possible outcomes: i) Tat SP predicted by TatP (leads to rejection of the entire COG), ii) Sec SP reliably predicted (positive SignalP and Phobius predictions) iii) absence of a Sec SP reliably predicted (negative predictions by both SignalP and Phobius), iv) discordant Sec SP assignments by SignalP and Phobius (protein gets discarded).

In order to find COGs with contradicting SP assignments, *i.e.* those clusters where SP gain and loss events happened, they were subdivided into positive, negative, or mixed clusters containing only positive, only negative, or both positive and negative predictions.

3.2.4 Assignment of symbiont status to bacteria

We manually annotated organisms according to their lifestyle as either symbiotic or free-living bacteria. The symbionts were further sub-divided into either endosymbionts or commensals. In the former relationship both partners benefit from the interactions, while in the latter relationship only one partner gains benefits, while the other is affected neither in a positive nor in a negative way. Out of the 153 genomes, 33 (21.6%) were classified as symbionts - 12 of them as commensals and 21 as endosymbionts.

3.2.5 Evolutionary model and parsimony analysis

We seek to identify SP gain and loss events in the evolutionary history of *Enterobacterales* orthologous families. The input data for this analysis are constituted by the evolutionary tree of the extant protein sequences in each family and the predicted SP states of the exterior nodes (leafs). The latter can be expressed as a presence/absence dichotomy. SP states for the internal nodes are reconstructed using the parsimony method by Fitch [125], which essentially assigns the SP states such that the number of state transitions in the tree is minimal. Given the tree, the inferred states at the internal nodes and the states at the leaf nodes, where a negative state (0) and a positive state (1) indicate the absence and the presence of a SP, respectively, a gain event corresponds to the transition from a negative state to a positive state at some branch of the tree, while the loss event corresponds to the opposite transition.

We conducted this standard parsimony analysis for all protein families with contradicting SP assignments between individual family members ('mixed' families). Only discrete SP data (*i.e.* presence or absence) were considered to infer ancestral states. Tentative SP loss events resulting from the first round of parsimonious reconstruction were verified by comparative genomics and used to conduct a gene start correction procedure, as described in the next section. Subsequently a second parsimony analysis was conducted to infer the final SP states for all internal nodes of the trees and to estimate the effect of the start correction procedure.

Along with the second parsimony analysis for SPs, the Fitch algorithm was also applied to the symbiont states. The leaf nodes (organisms) were assigned either state 2 if the organism lives in a commensal relationship, state 1 if it lives in an endosymbiotic relationship, or state 0 if it is a free-living bacterium. After inferring the ancestral states using the Fitch algorithm, transition events between all three states along the evolutionary tree were derived.

3.2.6 Gene start correction

Based on the results of the initial parsimony analysis we investigated the possibility of spurious gain or loss events caused by incorrect gene starts. All trees containing leaves (extant proteins) with contradicting SP assignments, *i.e.* the mixed clusters, were traversed. In case a leaf was predicted not contain a SP both by SignalP and Phobius, a set of proteins with alternative start positions (considering the ATG, GTG, and TTG start codons) was constructed for this specific protein. The size of the sequence neighborhood scanned up- and downstream for an alternative gene start was determined based on the MSAs calculated in the first round of the parsimony analysis as follows. The position of the first residue in the MSA of each protein without a SP prediction was compared against all first residue positions of proteins with SPs. The maxima of these distances in both directions, up- as well as downstream (plus another 30 residues in each direction) were used as search space. Subsequently SignalP, Phobius and TatP predictions were made for the N-termini of these new proteins. A start position was chosen dependent on the prediction outcomes in the following order of priority: i) positive TatP prediction, resulting in the deletion of the entire COG, ii) reliable positive or negative prediction (agreement between SignalP and Phobius), iii) disagreement between SignalP or Phobius, resulting in the deletion of the protein, or iv) gene start left unchanged, *i.e.* the reliable negative prediction remains valid. In cases where multiple gene starts lead to a reliable positive prediction, the one with the highest SignalP prediction score was chosen.

3.2.7 Functional annotation of protein groups

To calculate the enrichment of GO terms in the positive, negative, and mixed groups, GO annotations assigned to each individual protein were supplemented with their parent terms according to the GO tree. Searching for enriched terms was then achieved by applying a one-sided Fisher's exact test to each term in each group using the occurrence frequency of the term in all groups as a background model. A similar analysis was performed solely on the proteins in the mixed groups in order to understand the functional implications associated with the gain and loss of SPs.

3.2.8 Assignment of taxonomic positions to signal peptide gain and loss events

For each event reconstructed on the evolutionary tree by the method described above we first determined all children leafs of the node where the event happened, and the species, genus, family and order of each of the corresponding organisms. We then identified the minimal common taxonomy rank for this resulting group of genomes using the NCBI taxonomy tree. As a result, the taxonomic rank of that event could be determined.

3.2.9 Discrimination score

For each COG g a discrimination score $d(a, b, g)$ was calculated as

$$d(a, b, g) = \frac{a_{sp} - a_{\overline{sp}}}{a_{sp} + a_{\overline{sp}}} - \frac{b_{sp} - b_{\overline{sp}}}{b_{sp} + b_{\overline{sp}}}$$

where a and b are two lifestyles to be compared, *i.e.* free-living bacteria, commensals or endosymbionts, while a_{sp} and $a_{\overline{sp}}$ are the numbers of proteins associated with the lifestyle a and b_{sp} and $b_{\overline{sp}}$ are the numbers of proteins associated with the lifestyle b with and without SP in COG g. The result ranges from -2 to 2, where more negative values mean that in this group bacteria of type a tend to have fewer SPs than bacteria of type b, while a more positive value means the opposite. In addition, the closer the result is to the two extrema -2 and 2, the more discriminating the possession of a SP is for separating lifestyles a and b in a particular group g, while values close to zero can be considered as indecisive.

3.3 Results and discussion

3.3.1 Signal peptides in the Enterobacterales order

We conducted a comprehensive analysis of *Enterobacterales* secretomes based on bioinformatics predictions. Out of 626680 gene products encoded in 153 *Enterobacterales* genomes derived from the OMA orthology database, 52902 (8.4%) were identified as containing reliable SPs based on the intersection of positive SignalP, positive Phobius and negative TatP predictions, respectively, while 518174 (82.7%) proteins were determined to be reliable negatives. The remaining 55604 (8.9%) cases consist of 52050 (8.3%) discordant predictions (51787 predicted positive only by Phobius, 263 only by SignalP), and 3554 (0.6%) Tat SPs predicted by TatP. The average percentage of proteins with SPs per genome in our data is 7.7±2.6%; the percentage scales roughly linearly with the genome size, increasing from 0.2% in *Riesia pediculicola USDA* over 10.1% in the *Escherichia coli K12/MC4100/BW2952* to 10.7% in a yet unclassified *Enterobacteriaceae* bacterium (Figure 3.1A). The *Escherichia coli* annotation is thus in line with our previous estimate (10%) of the secretome size for this genome [16].

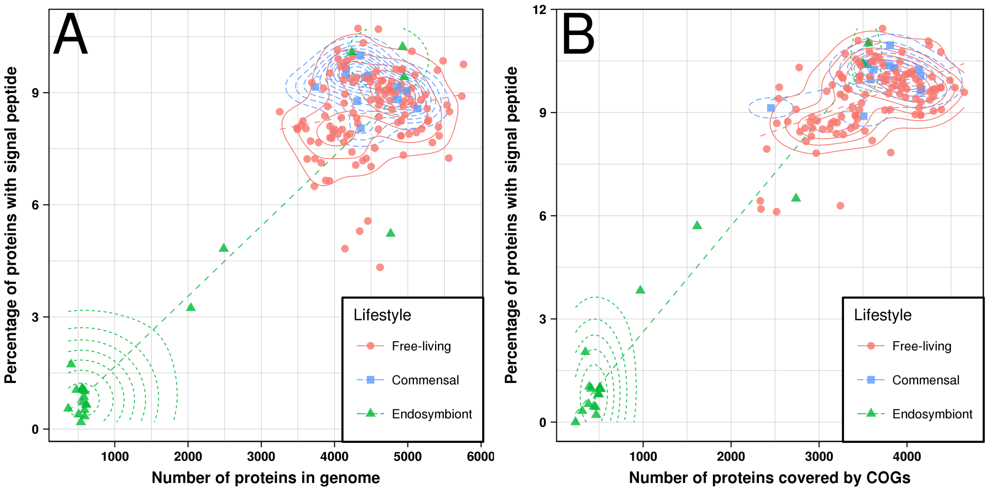

Figure 3.1: Number of proteins in a genome vs. the percentage of proteins that possess a SP (A) using the full dataset, and (B) after mapping of the proteins to COGs. In addition to the raw values, the two-dimensional density and a linear fit (dashed lines) for each lifestyle is shown.

3.3.2 Occurrence of signal peptides in Enterobacterales clusters of orthologous groups

In total, 557556 of the 626680 proteins (89.0%) belong to 24837 COGs with at least three members. On average 88.6±8.7% of proteins in the species considered are covered by COGs - from 52.9% in *Hamiltonella defensa subsp. Acyrthosiphon pisum 5AT* to 99.5% in *Buchnera aphidicola subsp. Acyrthosiphon pisum Tuc7*. The average COG coverage of small genomes, consisting of less than 1000 genes, tends to be similar (86.3±11.2%) to that of large genomes with more than 3000 genes (89.2±7.8%) (Figure 3.2) (p = 0.5, Kolmogorov-Smirnov test). The former correspond to endosymbiotic genomes that are thought to retain only the most functionally important and evolutionary conserved genes. The size of the clusters is 22.4 on average and ranges from three (4767 clusters or 19.2%), which is the smallest possible size, to 153 (7 clusters or 0.03%), which is a cluster containing a protein from every organism (Figure 3.3).

After removal of 1893 COGs which either contained a positive TatP prediction or did not satisfy the minimum number of three members after the removal of discordant SP predictions, 498690 of the initial 626680 proteins (79.6%) were left in the dataset and mapped to a COG. The percentage of these COG proteins possessing a SP does not significantly differ from the percentage of SP containing proteins in the entire proteomes. The total amount of proteins assigned as having a SP is 47139 (9.5%), with 8.6±2.8%

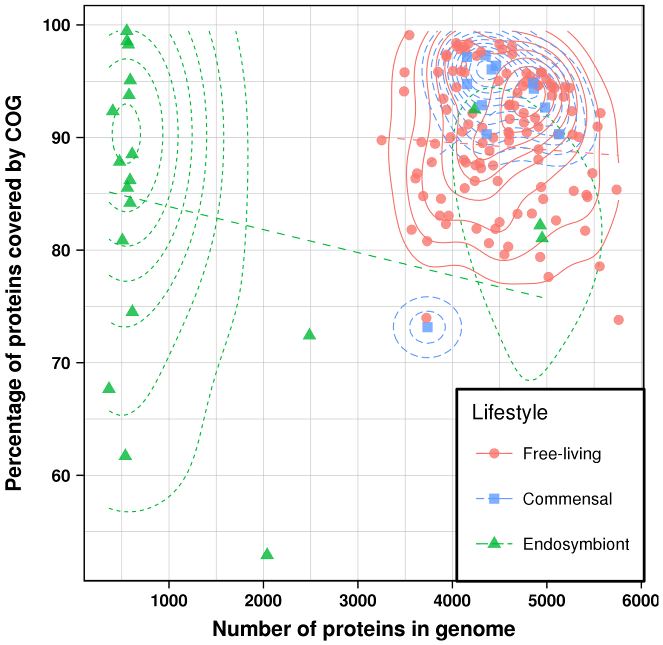

Figure 3.2: Number of proteins in a genome vs. the percentage of proteins that are members of a COG. In addition to the raw values, the two-dimensional density and a linear fit (dashed lines) for each lifestyle is shown.

on average per genome. Also, the dependence on the genome size is essentially the same (Figure 3.1B).

We subdivided the remaining 22944 COGs according to the SP assignments present in a cluster as described in the *Materials and Methods* section. This resulted in 20363 negative clusters (88.8%), containing only proteins without SPs, 1507 positive clusters (6.6%), containing only proteins with SPs, and 1074 mixed clusters (4.7%), containing proteins both with and without SPs (see Table 3.1). The mixed clusters can be assumed to contain those proteins that changed their cellular localization at least once in their evolutionary history, but could also result from wrong gene start annotation or wrong SP assignments.

Since we are primarily interested in gain and loss of SPs, mixed clusters were further examined in order to estimate the scale of annotation errors and determine the biological significance of evolutionary events.

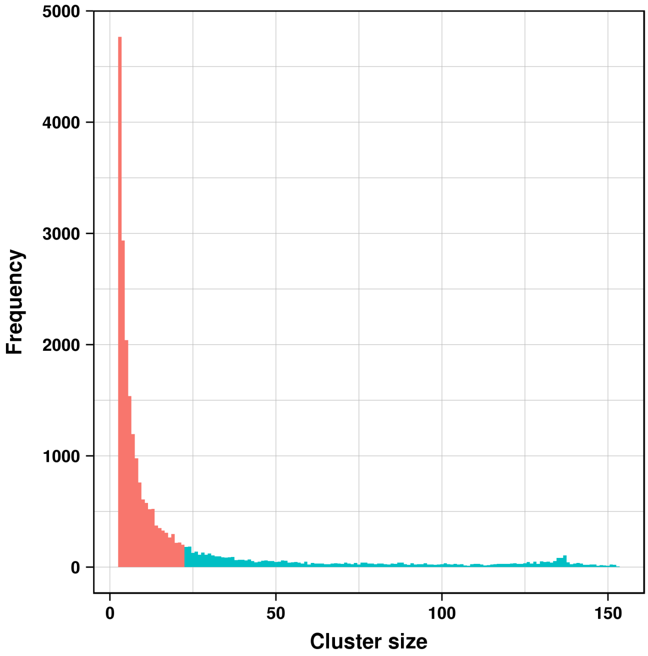

Figure 3.3: Distribution of cluster sizes. Histogram bins containing clusters that are smaller and larger than the average cluster size are colored red and green, respectively.

Table 3.1: Statistics on clusters and events for the two rounds of parsimony analysis before and after the gene start correction procedure.

Parsimony	**Clusters**			
round	**Negative**	**Positive**	**Mixed**	**Total**
1	20363 (88.8%)	1507 (6.6%)	1074 (4.7%)	22944
2	20363 (89.0%)	2087 (9.1%)	440 (1.9%)	22890
Parsimony	**Events**			
round	**Gain**	**Loss**	**Uncertain**	**Total**
1	325 (13.5%)	1235 (51.2%)	852 (35.3%)	2412
2	83 (11.6%)	288 (40.2%)	346 (48.3%)	717

3.3.3 Parsimony analysis and gene start correction

We conducted a first round of the parsimony analysis of the SP assignments for the 'mixed' COG clusters as described in the *Materials and Methods* section, *i.e.* using the Fitch algorithm. In total 2412 events were revealed, including 325 gains (13.5%), 1235 losses (51.2%), and 852 uncertain events (35.3%) where the state could not be resolved by parsimony (Table 3.1). SP losses thus prevailed over gains significantly (almost 4-fold).

Following the first round of the parsimony analysis we attempted to improve gene start annotation in order to minimize the number of false SP events. Each protein without an assigned SP was tested for a potential false negative prediction by shifting its gene start over a certain range determined by the SP containing proteins in the same group (see *Materials and Methods*). After the gene start correction, the MSAs and the trees were recalculated using the updated sequences. Altogether, the correction procedure affected 3005 proteins from 147 species, with the most affected genomes being *Cronobacter turicensis DSM 18703/LMG 23827/z3032* (127 corrections) and *Klebsiella pneumoniae subsp. pneumoniae ATCC 700721/MGH 78578* (54 corrections). In most cases gene starts underwent relatively small changes of their positions (Figure 3.4), with the average value of the absolute shift of +1.2 amino acids and the median value of +9; there were fewer corrections towards upstream gene start positions (1450) then towards downstream positions (1555).

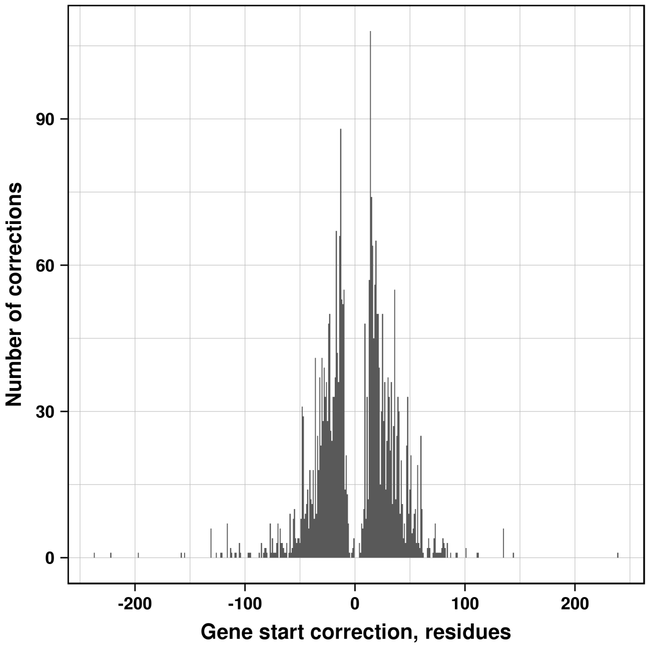

Figure 3.4: Distribution of gene start corrections, *i.e.* the number of residues by which the protein sequence was extended (negative values) or truncated (positive values).

The gene start correction procedure led to changed SP assignments for a number of proteins from 'negative' to 'positive', the removal of proteins in which the correction revealed discordant predictions, and the deletion of certain mixed clusters due to either positive TatP predictions or fewer than three remaining proteins in the COG. Overall, only 29.7% of the events were kept compared to the first round of parsimony analysis, while 41.0% of mixed clusters remained (Table 3.1). Based on these new assignments we conducted a second round of parsimony analysis on the remaining 440 mixed clusters, which revealed 83 gain (11.6%), 288 loss (40.2%), and 346 uncertain events (48.3%) out of 717 events in total (Table 3.1). Therefore, out of the 1235 loss events from the first round of parsimony analysis, 947 events were recognized as false positives and 242 gain events were also eliminated. The ratio between gains and losses decreased only slightly, still being almost 4-fold. The percentage of SPs in our final data after mapping to COGs, removal of Tat signal containing groups and gene start correction is 48817 out of 497338 proteins (9.8%), with an average of 8.9±2.9% per genome (Figure 3.5).

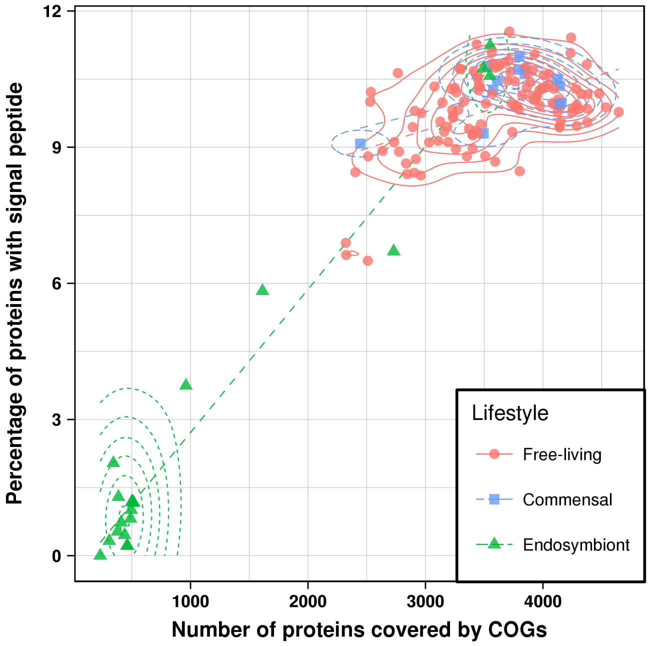

Figure 3.5: Number of proteins in a genome vs. the percentage of proteins that possess a SP after mapping of proteins to COGs, and after the gene start correction procedure. In addition to the raw values, the two-dimensional density and a linear fit (dashed lines) for each lifestyle are shown.

3.3.4 Sequence similarity of secreted and non-secreted proteins

In order to find out whether the gain and loss patterns of SPs correlate with the evolutionary distance we compared amino acid sequences of the proteins in the mixed groups. All possible pairwise sequence alignments were extracted from the MSA of each group and the pairwise sequence identity was calculated by dividing the number of identical residues by the length of the shorter sequence. We plotted the distributions of sequence identities for sequence pairs in which both, none or only one sequence had a SP (Figure 3.6). As expected, the mean of sequence identities for the pairs in which either no or both proteins possess a SP (80.9%, 80.6%) is higher than for the pairs where only one protein gets secreted (64.8%), because in the latter case a smaller number of almost identical sequences occurs. If only protein pairs with a sequence identity below 95% are considered, the three groups have much closer means (both have SPs: 71.5%, none has SP: 73.4%, one has SP: 59.9%).

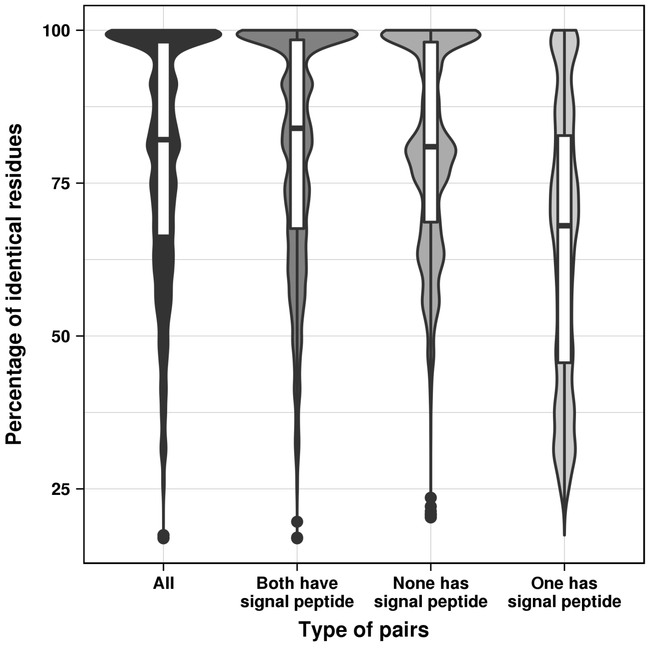

Figure 3.6: Comparison of sequence identity distributions between pairs of proteins where either both proteins have a SP, or both have none, or only one protein has a SP.

3.3.5 Evolutionary mechanisms leading to gain and loss of signal peptides

How are SPs gained and lost, at the molecular level? To answer this question, we analyzed the alignments of extant proteins that descended from their last common ancestor before the gain or loss event, such that some of them contain SPs while others do not. Note that only the latest events in the evolutionary sense were taken into account, e.g. if a gain event was later on reversed by a loss event, only the loss event was considered. For each alignment associated with a gain or loss event we calculated the length ratio *lr* between SPs and the N-termini devoid of SPs, as shown in Figure 3.7A. The distribution of *lr* values (Figure 3.8A) points to the existence of two categories of events. The first category, covering 145 loss and 34 gain events, is characterized by *lr* values close to zero, reflecting a full deletion or insertion of an entire SP. An example of such a loss event can be found in the ”Pectinesterase” OMA-group 189619. Pectin methylesterases, found in plant pathogens, play a major role in the first step of soft rot infections. They help to degrade pectin in the plant cell wall, destabilizing it and leading to cell necrosis and tissue maceration. Different plant pathogens have a different inventory of these secreted proteins [126]. Figure 3.7B shows the alignment of the SP-containing pectin methylesterases (*pemB*) from two Dickeya (former Erwinia) species and four *pemB* orthologs from other Pectobacteria, which lack a SP. Beyond the N-terminal part of the alignment the proteins are highly similar. It should be noted that *pemA*, another pectin methylesterase, does contain a SP in all of these six organisms. The observation that *pemB* is not exported in all pectin degrading bacteria is in line with an earlier experimental study, which showed that *pemB* is exported in some but not all Dickeya strains [127]. We therefore speculate that, although *pemB* is encoded in all of the Dickeya genomes, its activity may vary dependent on whether or not a SP is present.

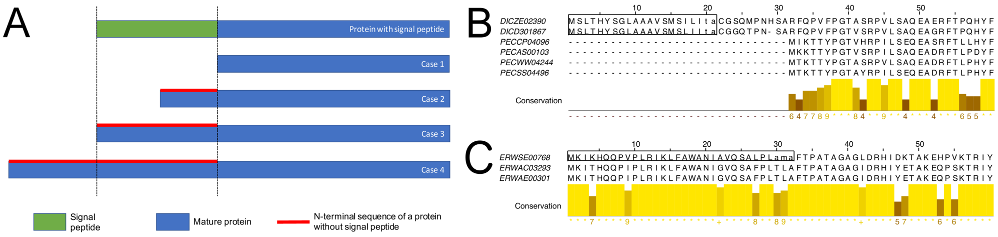
A
Protein with signal peptide
Case 1
Case 2
Case 3
Case 4
Signal peptide
Mature protein
N-terminal sequence of a protein without signal peptide
B
10
20
30
40
50
DICZE02390 MSLTHYSGLAAAVSMSILIta CGSQMPNHSARFQPVFPGTASRPVLSAQEAERFTPQHYF
DICD301867 MSLTHYSGLAAAVSMSLIIta CGGQTPN- SARFQPVFPGTVSRPVLSAQEAGRFTPQHYF
PECCP04096 - MIKTTYPGTVHRPILSEQEASRFTLLHYF
PECAS00103 - MKKTTYPGTVSRPILSAQEADRFTLPDYF
PECWW04244 - MTKTTYPGTASRPVLSAQEADRFTLPHYF
PECSS04496 - MTKTTYPGTAYRPILSEQEADRFTLPHYF
Conservation
- 647789***84**9**4***4***655**
C
10
20
30
40
50
ERWSE00768 MKIKHQQPVPLRIKLFAWANIAVQSALPLama FTPATAGAGLDRHIDKTAKEHPVKTRIY
ERWAC03293 MKITHQQPIPLRIKLFAWANIGVQSAFPLTLAFTPATAGAGIDRHIYETAKEQPSKTRIY
ERWAE00301 MKITHQQPIPLRIKLFAWANIGVQSAFPLTLAFTPATAGAGIDRHIYETAKEQPSKTRIY
Conservation
79**************+****8**89**********+****57****6*6*****

Figure 3.7: (A) The four possible cases for SP gain and loss events. In proteins devoid of SPs the N-teminal sequence can be completely eliminated (case 1), shortened (case 2), have the same length (case 3), or be extended (case 4). Cases one and three are by far the most prevalent ones. (B/C) The first 60 positions in the MSAs of the proteins involved in a SP gain event in ”Pectinesterase” OMA group containing two Dickeya and four Pectobacteria (UniProt identifiers: C6CL61, Q47474 (reviewed), C6DIG6, Q6DAZ5, D0KDA3, P55743 (reviewed)) (B) and the gain event in the ”putative Invasin” group containing three Erwinia species (UniProt identifiers: E3DHH7, D4I2A7, unknown) (C). Rectangles indicate SPs, with cleavage sites in lowercase letters.

We tested the hypothesis that complete deletions and insertions could be caused by transposable elements, but no such elements in proximity to the N-termini of the proteins in the mixed clusters were found by ISEScan [128].

In the second category, covering 25 loss and 20 gain events, proteins with and without SPs possess N-terminal amino acid sequences of comparable length. The events are therefore caused by amino acid substitutions, with *lr* values close to one. In most of the

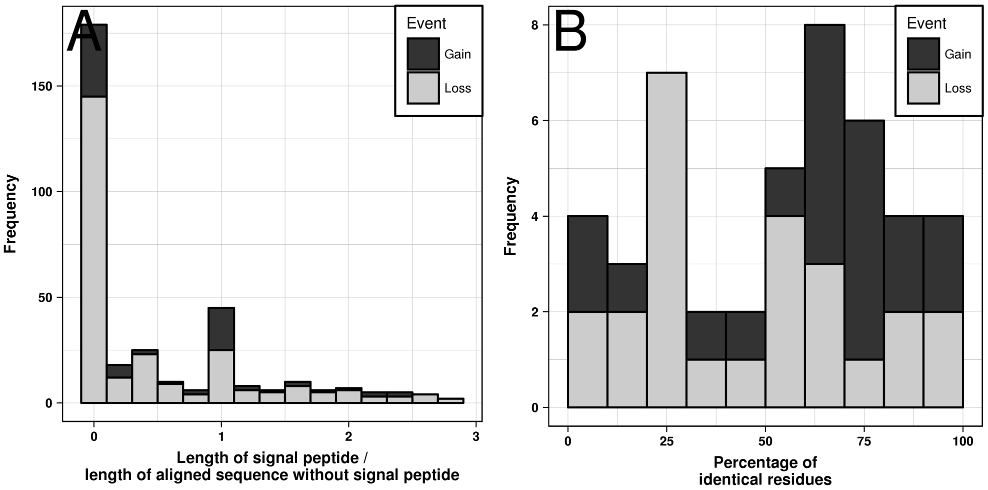

Figure 3.8: Comparison of SP sequences and the aligned N-terminal sequences without a SP. (A) Sequence length ratio. (B) Percentage of identical residues for those cases where the length ratio is between 0.9 and 1.1, *i.e.* where both sequences have a comparable length.

cases the N-terminal regions maintain an even higher sequence identity than the average of 52.7% (Figure 3.8B). For example, the gain event alignment of the ”putative Invasin” (OMA-group: 83250) (Figure 3.7C) contains three similar N-terminal sequences, but only one of them possesses a SP. From the six mutations contributing to the difference between the N-termini with and without SPs, four mutations strengthen the tripartite structure of a common SP: i) replacement of threonine by lysine at position four introduces an additional positively charged amino acid, ii) replacement of glycine by alanine at position 22 extends the hydrophobic stretch, and iii) two further mutations affect the cleavage site by changing its sequence from TLA to AMA and thus make it more similar to the canonical AxA motif.

While the conducted analysis of the mechanism included only the latest events, we were also able to identify 11 mixed clusters where preceding events were reversed. In seven cases, earlier loss events were inverted by a later gain event (”putative lipoprotein”, ”hemolysin activator protein”, ”RND efflux system outer membrane lipoprotein NodT”, ”Fimbrial biogenesis outer membrane usher protein”, ”Biofilm PGA synthesis protein pgaA” and two ”Putative uncharacterized proteins”), while in two groups a reversal in the opposite direction occurred (”acetyl-coA acetyltransferase” and ”secretion monitor”). In the remaining two COGs, the SP was lost, regained and lost again (”cytochrome b562” and ”Soluble lytic murein transglycosylase and related regulatory proteins some contain LysM/invasin domain”).

Our findings indicate that loss events are due to insertions/deletions almost seven times more often than due to mutations. For gain events, this ratio is only 1.5-fold. Indeed, a shift of the gene start will likely delete a SP, while a functional SP is not likely to be gained by randomly prepending amino acids to the protein N-terminus. Intuitively, the deletion or mutation of the cleavage site would be the most economical way to disable a SP, but our data do not support this assumption. We calculated sequence identities between the cleavage sites and the remaining N-terminal sequences for protein pairs with and without SP having lr values close to one. The Spearman's rank correlation coefficient between these two sequence identity values is 0.39 for gain and loss events together (p = 0.008), 0.49 for loss events (p = 0.013), but only 0.22 for gain events (p = 0.346), which indicates that the mutation rate in the cleavage sites does not differ from other positions within the SP sequence (see also Figure 3.9).

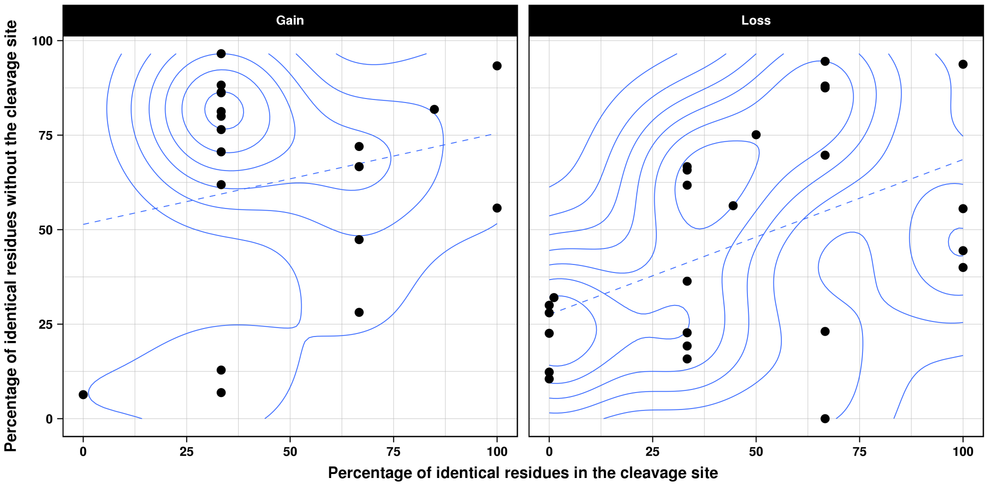

Figure 3.9: Identity between SP sequences and the aligned N-terminal sequences without a SP according to the sequence identity at the cleavage site and the remaining positions separated by gain and loss events.

3.3.6 Functional classification

We investigated the functional distribution and the localization of the positive/negative and mixed groups based on GO annotations (GO-terms) from three domains: biological process (BP), molecular function (MF), and cellular component (CC). In general, the distribution of GO terms in the mixed clusters is clearly more similar to the one of the positive than in the negative clusters (Figure 3.10). COG functions tend to reflect their SP content, with positive and mixed clusters containing significantly more GO terms associated with exported proteins, while the negative clusters are mostly associated with intracellular processes, functions and components. For example, processes involving DNA or RNA, which are localized within the cell, such as ”nucleobase-containing compound metabolic process” (GO:0006139) in the BP category and ”nucleotide binding” in the MF category, are prevalent in the negative group. On the other hand, ”Cell adhesion” (GO:0007155), a process which occurs outside of the cell, is almost exclusively found in the positive and mixed groups. The CC categories ”outer membrane” (GO:0019867) and ”pilus” (GO:0009289) are over-represented in the positive and mixed groups, while ”intracellular” (GO:0005622) and ”cytoplasm” (GO:0006737) are more often found in the negative groups. While the terms in the mixed groups are often similar to those in the positive groups, there are some exceptions, e.g. the ”aminoglycan metabolic process” (GO:0006022) from the BP category is prevalent in the mixed groups (in about 7.6% of its proteins), while almost absent in the other two groups (0.8% of the proteins in the negative groups, and 2.1% of the proteins in the positive groups).

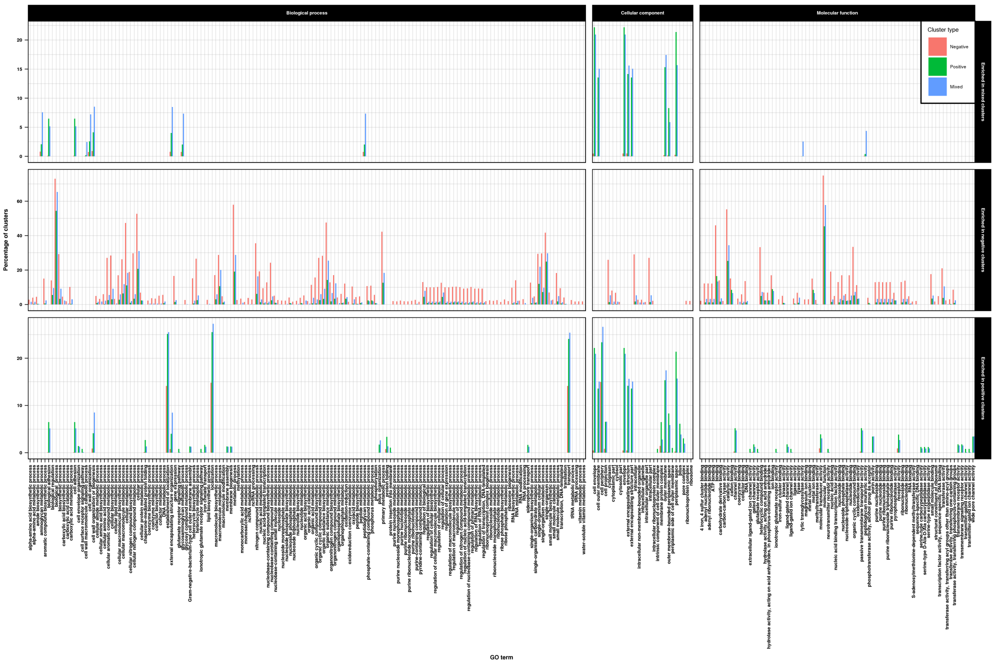

Figure 3.10: Percentage of proteins having a specific enriched GO-term for the mixed (first row), negative (second row) and positive (third row) groups. The columns match the three ontologies, biological process, cellular component and molecular function.

3.3.7 Taxonomy distribution of events

For each event we identified the minimal common taxonomic rank of the descendants of the node where it happened. Gain events preferentially occurred at the order level (32.5%), and somewhat less frequently at the family (28.9%), genus (22.9%) and order level (15.7%), while loss events occurred mostly at the species level (33.7%) (Table 3.2). The number of uncertain events increases with the level of the taxonomic rank, from 10.4% at the species level to 60.1% at the order level, mainly because the assignment of a definite SP state gets more difficult towards the root of the tree.

Table 3.2: Taxonomic rank of SP gain, loss and uncertain events.

Event	Species	Genus	Family	Order	Total
Gain	13 (15.7%)	19 (22.9%)	24 (28.9%)	27 (32.5%)	83
Loss	97 (33.7%)	45 (15.6%)	93 (32.3%)	53 (18.4%)	288
Uncertain	36 (10.4%)	18 (5.2%)	84 (24.3%)	208 (60.1%)	346

3.3.8 Symbiotic relationships and the loss of signal peptides

We investigated the inter-relationships between SPs, genome sizes, and bacterial lifestyle at two levels: the fraction of SP containing proteins as a function of genome size (Figure 3.5), and the correlation of SP gain/loss events with the transition from a free-living organism to a commensal organism or an endosymbiont and vice versa. It should be noted that these analyses were conducted on our final dataset, *i.e.* only with proteins which could be mapped to a COG and have a reliably assigned SP status after the gene start correction, which led to a slightly reduced number of proteins per genome.

In our dataset, the 120 free-living bacteria contain on average 3596 proteins, compared to 3730 proteins in the 12 commensals and 1066 proteins in the 21 endosymbionts. For reference, the average numbers of proteins in the complete genomes of free-living bacteria, commensals and endosymbionts were 4511, 4481 and 1500, respectively. The Kolmogorov-Smirnov test shows that the protein size distributions between free-living bacteria and commensals are similar (p = 0.12), while both of them differ significantly from the endosymbiont distribution (p = 1.3e-10 and p = 1.2e-5). The same is true for the percentage of proteins containing SPs, with the average numbers being 9.5% for the free-living bacteria, 10.0% for the commensals and 2.8% for the endosymbionts. Again, the distributions are significantly different when comparing free-living bacteria or commensals against endosymbionts (p = 5.8e-11 and p = 3.8e-6), while being similar between the latter two (p = 0.18). The same holds true according to the two sample Cramér-von

Mises test calculated for the multivariate distributions of protein sizes and fractions of SPs between the three classes (p-values close to zero between free-living/commensal and endosymbionts; p = 0.26 for free-living and commensals).

Symbionts tend to have reduced genomes as a consequence of losing genes whose functions are delegated to the host organism. As a result of genomic shrinkage, a larger proportion of the remaining genes is involved the basic cellular functions, such as replication, transcription, and translation, while many less essential functions, including those associated with amino acid synthesis or other metabolic processes, which can be provided by the partner or host may be lost [114]. We calculated a discrimination score $d(a, b, g)$ for each COG g (see *Materials and Methods*) to evaluate whether or not the possession of a SP is a sufficiently discriminative characteristic for telling apart endosymbionts (endos), commensals (coms), and free-living bacteria (fls). Out of the 440 mixed groups, 182 contained at least one free-living bacterium and at least one endosymbiont, 104 at least one commensal and at least one endosymbiont, and 221 contained at least one free-living bacterium and at least one commensal. According to the two-tailed Fisher's exact test discrimination between endosymbionts and free-living bacterial was significant in seven groups, of which the following six yielded $d(fl, endo, g)$ scores above zero (Figure 3.11), indicating an association of the SP-less proteins with endosymbionts: "flagellar biosynthetic protein flip", "endonuclease I", "mechanosensitive ion channel", "D-alanyl-D-alanine carboxypeptidase", "ErfK/YbiS/YcfS/YnhG family protein", and "N-acetylmuramoyl-l-alanine amidase". We found only one COG ("Spore coat U domain protein") with a significant discrimination and a $d(fl, endo, g)$ below zero, indicating that SPs preferentially occur in the proteins from symbiotic bacteria rather than in free-living organisms. In three out of the 104 COGs containing both endosymbionts and commensals the SP state was significantly associated with the lifestyle. We found two groups with $d(com, endo, g)$ above zero ("putative transferase" and "mechanosensitive ion channel"), as well as one below zero ("tonB-system energizer ExbB"). Comparing the groups containing free-living and commensals, there were also three significant groups, two with a $d(fl, com, g)$ above zero ("Putative uncharacterized protein", "peptidase M15D vanX D-ala-D-ala dipeptidase") and one below zero ("putative transferase"). The Spearman's rank correlation coefficient of 0.74 between all $d(fl, endo, g)$ and $d(com, endo, g)$ scores is highly significant (p = 2.2e-16), reflecting resemblance in genome size and SP content of free-living bacteria and commensals. The overall distribution of significant $d(a, b, g)$ scores (Figure 3.11) indicated that SPs are a discriminating feature between endosymbionts and free-living bacteria or commensals.

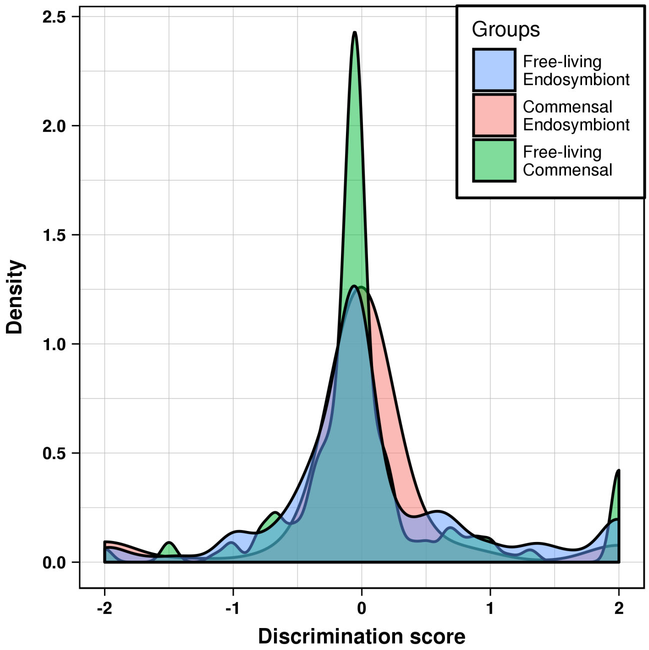

Figure 3.11: Density plot of discrimination scores between different lifestyles of bacteria.

We analyzed the GO annotations of the individual proteins with or without SPs in the mixed clusters (Figure 3.12). With regard to CC non-secreted proteins are preferably tagged as "cytoplasm" (GO:0005737), while the secreted ones are annotated with "membrane" (GO:0016020) which includes "outer membrane" (GO:0019867), "periplasmic space" (GO:0042597) and similar terms. In the MF and BP categories proteins containing a SP are involved in "channel activity" (GO:0015267) and "transport" (GO:0006810), while those without a SP take part in "nucleotide binding" (GO:0000166) and "carboxylic acid biosynthetic process" (GO:0046394).

While the previous analysis was conducted for all bacteria in our dataset, we additionally compared GO-term annotations of proteins with and without a SP for each lifestyle separately and found that functional assignments generally do not correlate with the lifestyle, with few exceptions. Some GO-terms are more (MF: "nucleotide binding") or less (CC: "membrane") frequently associated with endosymbionts compared to free-living bacteria and commensals (Figure 3.12).

Assuming that some species may have changed their lifestyle in the course of evolution, we conducted an additional parsimony analysis using the endosymbiont/commensal/free-living annotations together with the SP events (Table 3.3). The proportions of gain/loss events are similar for all transitions to any lifestyles, e.g. 1.1% of the transitions to

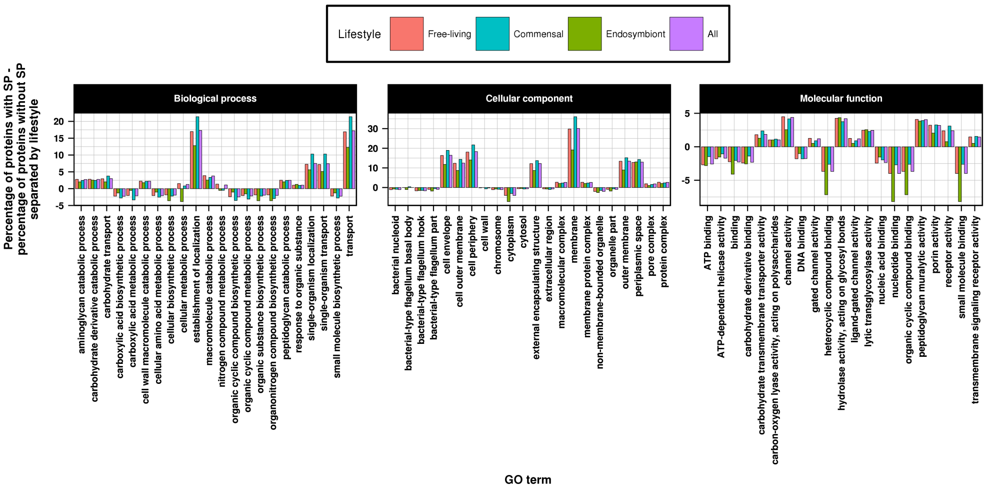

Figure 3.12: Percentage of proteins without SP subtracted from the percentage of proteins with SP in the mixed clusters having a specific enriched GO-term and belonging to organisms with a certain lifestyle.

endosymbionts are accompanied by a loss event but only 0.4% by gain events. However, dependent on the nature of a transition there is a noticeable difference in the number cases where SP assignments remain negative: this applies to 28.7% of the transitions to endosymbionts, but only to 19.8% and 15.6% of the transitions to free-living bacteria and commensals, respectively. We speculate that in many such cases the loss of the SP might not have happened simultaneously with the transition to a specific lifestyle, but rather before or after it. Qualitatively, this apparent difference seems to strengthen our conjecture, but it fails to reach statistical significance as the number of such events is quite low compared to the total number of events in our analysis.

Table 3.3: Contingency table of SP gain and loss events and their correlation with changes of bacterial lifestyles.

Event	**Transition to free-living bacterium**	**Transition to endosymbiont**	**Transition to commensal**	**Uncertain transition**	**Total number of SP events**
Gain	76 (0.4%)	2 (0.4%)	2 (0.1%)	3 (0.9%)	83
Loss	263 (1.3%)	5 (1.1%)	14 (0.9%)	6 (1.7%)	288
Uncertain	268 (1.3%)	15 (3.3%)	5 (0.3%)	58 (16.9%)	346
Keep SP	15581 (77.2%)	298 (66.4%)	1312 (83.0%)	227 (66.2%)	17418
Stay without SP	4006 (19.8%)	129 (28.7%)	247 (15.6%)	49 (14.3%)	4431
Total number of transition events	20194 (100%)	449 (100%)	1580 (100%)	343 (100%)	

3.4 Conclusions

Computational prediction of SPs is an indispensable step in bacterial genome annotation, but their evolutionary dynamics has not been comprehensively studied. We investigated the gain and loss patterns of SPs between orthologous proteins from *Enterobacterales* and found that 1.9% of COGs contain proteins both with and without SPs. Reconstruction of ancestral SP states by parsimony analysis in such mixed groups clearly indicates that SPs get lost more often in the course of evolution than they are gained. We also show that SP gains tend to be more ancient events, predominantly occurring at the family and probably at the order level, although a high number of uncertain events at this latter level makes it impossible to draw definitive conclusions. At the same time, SP losses might be more recent events as we found most of them at the species level. Gain and loss events occur by either a complete insertion or deletion of the entire SP sequence or by retaining the N-terminal sequence and mutating residues to enable or disable the SPs functionality. The prevalent loss of SPs is accompanied by genome reduction, with smaller genomes of endosymbiotic bacteria containing a lower percentage of SPs than free-living and commensal bacteria. In some enterobacterial COGs the presence or absence of a SP alone is sufficient to discriminate between endosymbionts, on the one hand, and free-living bacteria or commensals, on the other hand. Finally, we demonstrate that SP loss events preferentially occur in the course of transition from free-living bacteria/commensals to endosymbionts.

Bibliography

[1] A. Tsirigotaki, J. De Geyter, N. Šoštaric, A. Economou, and S. Karamanou. Protein export through the bacterial Sec pathway. *Nature Reviews Microbiology*, 15(1):21–36, November 2016.

[2] G. von Heijne. Signal sequences. *J Mol Biol*, 184(1):99–105, July 1985.

[3] G. von Heijne The Journal of membrane biology. The signal peptide. *Springer*, 1990.

[4] J. De Geyter, A. Tsirigotaki, G. Orfanoudaki, V. Zorzini, A. Economou, and S. Karamanou. Protein folding in the cell envelope of Escherichia coli. *Nature microbiology*, 1(8):16107, July 2016.

[5] D. Akopian, K. Shen, X. Zhang, and S.-o. Shan. Signal recognition particle: an essential protein-targeting machine. *Annual Review of Biochemistry*, 82(1):693–721, 2013.

[6] T. Saio, X. Guan, P. Rossi, A. Economou, and C. G. Kalodimos. Structural basis for protein antiaggregation activity of the trigger factor chaperone. *Science*, 344(6184):1250494–1250494, May 2014.

[7] K. E. Chatzi, M. F. Sardis, A. Economou, and S. Karamanou. SecA-mediated targeting and translocation of secretory proteins. *Biochimica et biophysica acta*, 1843(8):1466–1474, August 2014.

[8] K. E. Chatzi, M. F. Sardis, S. Karamanou, and A. Economou. Breaking on through to the other side: protein export through the bacterial Sec system. *Biochemical Journal*, 449(1):25–37, January 2013.

[9] E. Schiebel, A. J. Driessen, F. U. Hartl, and W. Wickner. Delta mu H+ and ATP function at different steps of the catalytic cycle of preprotein translocase. *Cell*, 64(5):927–939, March 1991.

[10] R. J. Schulze, J. Komar, M. Botte, W. J. Allen, S. Whitehouse, V. A. M. Gold, J. A. Lycklama A Nijeholt, K. Huard, I. Berger, C. Schaffitzel, and I. Collinson. Membrane protein insertion and proton-motive-force-dependent secretion through the bacterial holo-translocon SecYEG-SecDF-YajC-YidC. *Proc Natl Acad Sci U S A*, 111(13):4844–4849, April 2014.

[11] S. M. Auclair, M. K. Bhanu, and D. A. Kendall. Signal peptidase I: cleaving the way to mature proteins. *Protein Science*, 21(1):13–25, January 2012.

[12] P. Natale, T. Brüser, and A. J. M. Driessen. Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane–distinct translocases and mechanisms. *Biochimica et biophysica acta*, 1778(9):1735–1756, September 2008.

[13] C. Rabouille. Pathways of Unconventional Protein Secretion. *Trends in Cell Biology*, 27(3):230–240, March 2017.

[14] L. Käll, A. Krogh, and E. L. L. Sonnhammer. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *J Mol Biol*, 338(5):1027–1036, May 2004.

[15] H. Nielsen, J. Engelbrecht, S. Brunak, and G. Von Heijne. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein engineering*, 10(1):1–6, January 1997.

[16] D. N. Ivankov, S. H. Payne, M. Y. Galperin, S. Bonissone, P. A. Pevzner, and D. Frishman. How many signal peptides are there in bacteria? *Environmental Microbiology*, 15(4):983–990, April 2013.

[17] J. Dyrløv Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak. Improved Prediction of Signal Peptides: SignalP 3.0. *J Mol Biol*, 340(4):783–795, July 2004.

[18] T. N. Petersen, S. Brunak, G. von Heijne, and H. Nielsen. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*, 8(10):785–786, September 2011.

[19] E. Venter, R. D. Smith, and S. H. Payne. Proteogenomic Analysis of Bacteria and Archaea: A 46 Organism Case Study. *PLoS One*, 6(11):e27587, November 2011.

[20] N. Gupta, S. Tanner, N. Jaitly, J. N. Adkins, M. Lipton, R. Edwards, M. Romine, A. Osterman, V. Bafna, R. D. Smith, and P. A. Pevzner. Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. *Genome Research*, 17(9):1362–1377, July 2007.

[21] K. Gevaert, M. Goethals, L. Martens, J. Van Damme, A. Staes, G. R. Thomas, and J. Vandekerckhove. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nature Biotechnology*, 21(5):566–569, March 2003.

[22] M. Aivaliotis, K. Gevaert, M. Falb, A. Tebbe, K. Konstantinidis, B. Bisle, C. Klein, L. Martens, A. Staes, E. Timmerman, J. Van Damme, F. Siedler, F. Pfeiffer, J. Vandekerckhove, and D. Oesterhelt. Large-Scale Identification of N-Terminal Peptides in the Halophilic Archaea Halobacteriumsalinarumand Natronomonas-pharaonis. *Journal of Proteome Research*, 6(6):2195–2204, June 2007.

[23] S. H. Payne, S.-T. Huang, and R. Pieper. A proteogenomic update to Yersinia: enhancing genome annotation. *BMC Genomics*, 11:460, August 2010.

[24] M. Braaksma, E. S. Martens-Uzunova, P. J. Punt, and P. J. Schaap. An inventory of the Aspergillus niger secretome by combining in silico predictions with shotgun proteomics data. *BMC Genomics*, 11(1):584, October 2010.

[25] B. K. Erickson, R. S. Mueller, N. C. VerBerkmoes, M. Shah, S. W. Singer, M. P. Thelen, J. F. Banfield, and R. L. Hettich. Computational Prediction and Experimental Validation of Signal Peptide Cleavages in the Extracellular Proteome of a Natural Microbial Community. *Journal of Proteome Research*, 9(5):2148–2159, May 2010.

[26] G. Von Heijne. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res*, 14(11):4683–4690, June 1986.

[27] H. Nielsen and A. Krogh. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, 6:122–130, 1998.

[28] L. Kall, A. Krogh, and E. L. L. Sonnhammer. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, 21(Suppl 1):i251–i257, June 2005.

[29] K. Hiller, A. Grote, M. Scheer, R. Munch, and D. Jahn. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res*, 32(Web Server):W375–W379, July 2004.

[30] K.-C. Chou and H.-B. Shen. Signal-CF: A subsite-coupled and window-fusing approach for predicting signal peptides. *Biochemical and Biophysical Research Communications*, 357(3):633–640, June 2007.

[31] K. Frank and M. J. Sippl. High-performance signal peptide prediction based on sequence alignment techniques. *Bioinformatics*, 24(19):2172–2176, September 2008.

[32] S. M. Reynolds, L. Käll, M. E. Riffle, J. A. Bilmes, and W. S. Noble. Transmembrane Topology and Signal Peptide Prediction Using Dynamic Bayesian Networks. *PLoS computational biology*, 4(11):e1000213, November 2008.

[33] K. Nakai and P. Horton. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in Biochemical Sciences*, 24(1):34–36, January 1999.

[34] J. L. Gardy. PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res*, 31(13):3613–3617, July 2003.

[35] N. Y. Yu, J. R. Wagner, M. R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S. C. Sahinalp, M. Ester, L. J. Foster, and F. S. L. Brinkman. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26(13):1608–1615, May 2010.

[36] R. Nair and B. Rost. Mimicking Cellular Sorting Improves Prediction of Subcellular Localization. *J Mol Biol*, 348(1):85–100, April 2005.

[37] T. Goldberg, T. Hamp, and B. Rost. LocTree2 predicts localization for all domains of life. *Bioinformatics*, 28(18):i458–i465, September 2012.

[38] T. Goldberg, M. Hecht, T. Hamp, T. Karl, G. Yachdav, N. Ahmed, U. Altermann, P. Angerer, S. Ansorge, K. Balasz, M. Bernhofer, A. Betz, L. Cizmadija, K. T. Do, J. Gerke, R. Greil, V. Joerdens, M. Hastreiter, K. Hembach, M. Herzog, M. Kalemanov, M. Kluge, A. Meier, H. Nasir, U. Neumaier, V. Prade, J. Reeb, A. Sorokoumov, I. Troshani, S. Vorberg, S. Waldraff, J. Zierer, H. Nielsen, and B. Rost. LocTree3 prediction of localization. *Nucleic Acids Res*, 42(Web Server issue):W350–5, July 2014.

[39] V. Takiar, C. K. M. Ip, M. Gao, G. B. Mills, and L. W. T. Cheung. Neomorphic mutations create therapeutic challenges in cancer. *Oncogene*, 36(12):1607–1618, March 2017.

[40] GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet (London, England)*, 388(10053):1545–1602, October 2016.

[41] D. Devos and A. Valencia. Practical limits of function prediction. *Proteins*, 41(1):98–107, October 2000.

[42] Quest for Orthologs consortium, A. M. Altenhoff, B. Boeckmann, S. Capella-Gutierrez, D. A. Dalquen, T. DeLuca, K. Forslund, J. Huerta-Cepas, B. Linard, C. Pereira, L. P. Pryszcz, F. Schreiber, A. S. da Silva, D. Szklarczyk, C.-M. Train, P. Bork, O. Lecompte, C. von Mering, I. Xenarios, K. Sjölander, L. J. Jensen, M. J. Martin, M. Muffato, T. Gabaldón, S. E. Lewis, P. D. Thomas, E. Sonnhammer, and C. Dessimoz. Standardized benchmarking in the quest for orthologs. *Nat Methods*, 13(5):425–430, April 2016.

[43] N. L. Nehrt, W. T. Clark, P. Radivojac, and M. W. Hahn. Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. *PLoS computational biology*, 7(6):e1002073, June 2011.

[44] B. Rost. Enzyme Function Less Conserved than Anticipated. *J Mol Biol*, 318(2):595–608, April 2002.

[45] P. Aloy, H. Ceulemans, A. Stark, and R. B. Russell. The relationship between sequence and interaction divergence in proteins. *J Mol Biol*, 332(5):989–998, October 2003.

[46] E. J. Williams, C. Pal, and L. D. Hurst. The molecular evolution of signal peptides. *Gene*, 253(2):313–322, August 2000.

[47] S. R. Doyle, N. R. P. Kasinadhuni, C. K. Chan, and W. N. Grant. Evidence of Evolutionary Constraints That Influences the Sequence Composition and Diversity of Mitochondrial Matrix Targeting Signals. *PLoS One*, 8(6):e67938, June 2013.

[48] Y. Fukasawa, R. K. Leung, S. K. Tsui, and P. Horton. Plus ça change – evolutionary sequence divergence predicts protein subcellular localization signals. *BMC Genomics*, 15(1):46, 2014.

[49] M. Trost, D. Wehmhöner, U. Kärst, G. Dieterich, J. Wehland, and L. Jänsch. Comparative proteome analysis of secretory proteins from pathogenic and non-pathogenic Listeriaspecies. *PROTEOMICS*, 5(6):1544–1557, April 2005.

[50] C. Song, A. Kumar, and M. Saleh. Bioinformatic Comparison of Bacterial Secretomes. *Genomics, Proteomics & Bioinformatics*, 7(1-2):37–46, June 2009.

[51] P. Hönigschmid, N. Bykova, R. Schneider, D. Ivankov, and D. Frishman. Evolutionary Interplay between Symbiotic Relationships and Patterns of Signal Peptide Gain and Loss. *Genome biology and evolution*, 10(3):928–938, March 2018.

[52] Y. Oka, T. Asano, Y. Shibasaki, J. L. Lin, K. Tsukuda, H. Katagiri, Y. Akanuma, and F. Takaku. C-terminal truncated glucose transporter is locked into an inward-facing form without transport activity. *Nature*, 345(6275):550–553, June 1990.

[53] W. A. Catterall. From ionic currents to molecular mechanisms: the structure and function of voltage-gated sodium channels. *Neuron*, 26(1):13–25, April 2000.

[54] P. Agre. Aquaporin Water Channels. *Bioscience Reports*, 24(3):127–163, July 2005.

[55] B. Trzaskowski, D. Latek, S. Yuan, U. Ghoshdastider, A. Debinski, and S. Filipek. Action of molecular switches in GPCRs–theoretical and experimental studies. *Current medicinal chemistry*, 19(8):1090–1109, 2012.

[56] A. J. Holmes, A. Costello, M. E. Lidstrom, and J. C. Murrell. Evidence that participate methane monooxygenase and ammonia monooxygenase may be evolutionarily related. *FEMS Microbiology Letters*, 132(3):203–208, 1995.

[57] J. J. Flanagan, J.-C. Chen, Y. Miao, Y. Shao, J. Lin, P. E. Bock, and A. E. Johnson. Signal recognition particle binds to ribosome-bound signal sequences with fluorescence-detected subnanomolar affinity that does not diminish as the nascent chain lengthens. *Journal of Biological Chemistry*, 278(20):18628–18637, May 2003.

[58] K. Ojemalm, S. C. Botelho, C. Stüdle, and G. von Heijne. Quantitative analysis of SecYEG-mediated insertion of transmembrane α-helices into the bacterial inner membrane. *J Mol Biol*, 425(15):2813–2822, August 2013.

[59] B. Rost and C. Sander. Bridging the Protein Sequence-Structure Gap by Structure Predictions. *Annual Review of Biophysics and Biomolecular Structure*, 25(1):113–136, June 1996.

[60] C. The UniProt. UniProt: a hub for protein information. *Nucleic Acids Res*, 43(D1):D204–D212, October 2014.

[61] D. Kozma, I. Simon, and G. E. Tusnády. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res*, 41(D1):D524–D529, November 2012.

[62] G. E. Tusnady. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*, 33(Database issue):D275–D278, December 2004.

[63] P. R. L. Markwick, T. Malliavin, and M. Nilges. Structural Biology by NMR: Structure, Dynamics, and Interactions. *PLoS computational biology*, 4(9):e1000168, September 2008.

[64] S. I. O'Donoghue, D. S. Goodsell, A. S. Frangakis, F. Jossinet, R. A. Laskowski, M. Nilges, H. R. Saibil, A. Schafferhans, R. C. Wade, E. Westhof, and A. J. Olson. Visualization of macromolecular structures. *Nat Methods*, 7(3):S42–S55, March 2010.

[65] K. M. Clark, N. Fedoriw, K. Robinson, S. M. Connelly, J. Randles, M. G. Malkowski, G. T. DeTitta, and M. E. Dumont. Purification of transmembrane proteins from Saccharomyces cerevisiae for X-ray crystallography. *Protein expression and purification*, 71(2):207–223, June 2010.

[66] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, R. Lepore, and T. Schwede. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*, 14(Suppl. 1):155, May 2018.

[67] J. Yang and Y. Zhang. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res*, 43(W1):W174–81, July 2015.

[68] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*, 268(1):209–225, April 1997.

[69] Y. Zhang. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3):342–348, June 2008.

[70] B. Adhikari and J. Cheng. CONFOLD2: improved contact-driven ab initio protein structure modeling. *BMC Bioinformatics*, 19(1):22, January 2018.

[71] D. Altschuh, A. M. Lesk, A. C. Bloomer, A. K. J. o. m. biology, and 1987. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Elsevier*, 193(4):693–707, February 1987.

[72] D. Altschuh, T. Vernet, P. Berti, D. Moras, D. Selection, , and 1988. Coordinated amino acid changes in homologous protein families. *academic.oup.com*, 1988.

[73] U. Göbel, C. Sander, R. Schneider, and A. Valencia. Correlated mutations and residue contacts in proteins. *Proteins*, 18(4):309–317, April 1994.

[74] S. W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, October 1999.

[75] J. P. Dekker, A. Fodor, R. W. Aldrich, and G. Yellen. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, 20(10):1565–1572, July 2004.

[76] L. C. Martin, G. B. Gloor, S. D. Dunn, and L. M. Wahl. Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21(22):4116–4124, November 2005.

[77] O. Olmea and A. Valencia. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding & design*, 2(3):S25–32, 1997.

[78] M. Punta and B. Rost. PROFcon: novel prediction of long-range contacts. *Bioinformatics*, 21(13):2960–2968, June 2005.

[79] E. Wallin and G. Von Heijne. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Science*, 7(4):1029–1038, April 1998.

[80] A. Fuchs, A. J. Martin-Galiano, M. Kalman, S. Fleishman, N. Ben-Tal, and D. Frishman. Co-evolving residues in membrane proteins. *Bioinformatics*, 23(24):3312–3319, December 2007.

[81] A. Fuchs, A. Kirschner, and D. Frishman. Prediction of helix-helix contacts and interacting helices in polytopic membrane proteins using neural networks. *Proteins*, 74(4):857–871, March 2009.

[82] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS One*, 6(12):e28766, December 2011.

[83] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190, November 2011.

[84] L. Kaján, T. A. Hopf, M. Kalaš, D. S. Marks, and B. Rost. FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics*, 15(1):85, 2014.

[85] S. Seemayer, M. Gruber, and J. Söding. CCMpred–fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, 30(21):3128–3130, November 2014.

[86] M. J. Skwark, D. Raimondi, M. Michel, and A. Elofsson. Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. *PLoS computational biology*, 10(11):e1003889, November 2014.

[87] D. T. Jones, T. Singh, T. Kosciolek, and S. Tetchner. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31(7):999–1006, November 2014.

[88] P. Hönigschmid and D. Frishman. Accurate prediction of helix interactions and residue contacts in membrane proteins. *JOURNAL OF STRUCTURAL BIOLOGY*, 194(1):112–123, April 2016.

[89] T. Nugent and D. T. Jones. Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, 10(1):159, 2009.

[90] H. Viklund, E. Granseth, and A. Elofsson. Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins: application to complete genomes. *J Mol Biol*, 361(3):591–603, August 2006.

[91] H. Viklund and A. Elofsson. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics*, 24(15):1662–1668, August 2008.

[92] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, June 2006.

[93] E. P. Carpenter, K. Beis, A. D. Cameron, and S. Iwata. Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology*, 18(5):581–586, October 2008.

[94] M. Remmert, A. Biegert, A. Hauser, and J. Söding. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods*, 9(2):173–175, December 2011.

[95] S. Seemayer, M. Gruber, and J. Söding. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*, 30(21):3128–3130, July 2014.

[96] M. A. Lomize, A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg. OPM: Orientations of Proteins in Membranes database. *Bioinformatics*, 22(5):623–625, February 2006.

[97] X.-F. Wang, Z. Chen, C. Wang, R.-X. Yan, Z. Zhang, and J. SONG. Predicting Residue-Residue Contacts and Helix-Helix Interactions in Transmembrane Proteins Using an Integrative Feature-Based Random Forest Approach. *PLoS One*, 6(10):e26767, October 2011.

[98] G. E. Tusnady, Z. Dosztanyi, and I. Simon. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*, 21(7):1276–1277, March 2005.

[99] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta. Pfam: the protein families database. *Nucleic Acids Res*, 42(D1):D222–D230, December 2013.

[100] Y. Zhang. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*, 33(7):2302–2309, April 2005.

[101] R Core Team. R: A Language and Environment for Statistical Computing, 2013.

[102] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, November 2003.

[103] S. Kawashima and M. Kanehisa. AAindex: amino acid index database. *Nucleic Acids Res*, 28(1):374, January 2000.

[104] L. Adamian and J. Liang. Prediction of transmembrane helix orientation in polytopic membrane proteins. *BMC Structural Biology*, 6:13, June 2006.

[105] J. Yang, R. Jang, Y. Zhang, and H.-B. Shen. High-accuracy prediction of transmembrane inter-helix contacts and application to GPCR 3D structure modeling. *Bioinformatics*, 29(20):2579–2587, August 2013.

[106] J. Reeb, E. Kloppmann, M. Bernhofer, and B. Rost. Evaluation of transmembrane helix predictions in 2014. *Proteins*, 83(3):473–484, January 2015.

[107] B. Rost, C. Sander, and R. Schneider. Redefining the goals of protein secondary structure prediction. *J Mol Biol*, 235(1):13–26, January 1994.

[108] K. Julenius. Protein Evolution Is Faster Outside the Cell. *Molecular Biology and Evolution*, 23(11):2039–2048, August 2006.

[109] B. Y. Liao, M. P. Weng, and J. Zhang. Impact of Extracellularity on the Evolutionary Rate of Mammalian Proteins. *Genome biology and evolution*, 2(0):39–43, May 2010.

[110] E. E. Winter. Elevated Rates of Protein Secretion, Evolution, and Disease Among Tissue-Specific Genes. *Genome Research*, 14(1):54–61, December 2003.

[111] G. Von Heijne. Signal sequences. The limits of variation. *J Mol Biol*, 184(1):99–105, July 1985.

[112] R. S. Hegde and H. D. Bernstein. The surprising complexity of signal sequences. *Trends in Biochemical Sciences*, 31(10):563–571, October 2006.

[113] S. H. Payne, S. Bonissone, S. Wu, R. N. Brown, D. N. Ivankov, D. Frishman, L. Pasa-Tolic, R. D. Smith, and P. A. Pevzner. Unexpected Diversity of Signal Peptides in Prokaryotes. *mBio*, 3(6):e00339–12–e00339–12, October 2012.

[114] S. G. E. Andersson and C. G. Kurland. Reductive evolution of resident genomes. *Trends in Microbiology*, 6(7):263–268, July 1998.

[115] M. Adeolu, S. Alnajar, S. Naushad, and R. S Gupta. Genome-based phylogeny and taxonomy of the 'Enterobacteriales': proposal for Enterobacterales ord. nov. divided into the families Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morganellaceae fam. nov., and Budviciaceae fam. nov. *International journal of systematic and evolutionary microbiology*, 66(12):5575–5599, December 2016.

[116] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, V. Miller, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 35(Database):D5–D12, January 2007.

[117] N. Pakseresht, B. Alako, C. Amid, A. Cerdeño-Tárraga, I. Cleland, R. Gibson, N. Goodgame, T. Gur, M. Jang, S. Kay, R. Leinonen, W. Li, X. Liu, R. Lopez, H. McWilliam, A. Oisel, S. Pallreddy, S. Plaister, R. Radhakrishnan, S. Rivière, M. Rossello, A. Senf, N. Silvester, D. Smirnov, S. Squizzato, P. t. Hoopen, A. L. Toribio, D. Vaughan, V. Zalunin, and G. Cochrane. Assembly information services in the European Nucleotide Archive. *Nucleic Acids Res*, 42(D1):D38–D43, December 2013.

[118] P. J. Kersey, J. E. Allen, I. Armean, S. Boddu, B. J. Bolt, D. Carvalho-Silva, M. Christensen, P. Davis, L. J. Falin, C. Grabmueller, J. Humphrey, A. Kerhornou, J. Khobova, N. K. Aranganathan, N. Langridge, E. Lowy, M. D. McDowall, U. Maheswari, M. Nuhn, C. K. Ong, B. Overduin, M. Paulini, H. Pedro, E. Perry, G. Spudich, E. Tapanari, B. Walts, G. Williams, M. Tello-Ruiz, J. Stein, S. Wei, D. Ware, D. M. Bolser, K. L. Howe, E. Kulesha, D. Lawson, G. Maslen, and D. M. Staines. Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res*, 44(D1):D574–D580, January 2016.

[119] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[120] The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res*, 43(D1):D1049–D1056, January 2015.

[121] A. M. Altenhoff, N. Škunca, N. Glover, C.-M. Train, A. Sueki, I. Piližota, K. Gori, B. Tomiczek, S. Müller, H. Redestig, G. H. Gonnet, and C. Dessimoz. The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res*, 43(D1):D240–D249, November 2014.

[122] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic biology*, 59(3):307–321, May 2010.

[123] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1):539–539, January 2011.

[124] J. D. Bendtsen, H. Nielsen, D. Widdick, T. Palmer, and S. Brunak. Prediction of twin-arginine signal peptides. *BMC Bioinformatics*, 6:167, July 2005.

[125] W. M. Fitch. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic biology*, 20(4):406–416, December 1971.

[126] D. W. Abbott and A. B. Boraston. Structural Biology of Pectin Degradation by Enterobacteriaceae. *Microbiology and Molecular Biology Reviews*, 72(2):301–316, June 2008.

[127] V. E. Shevchik, G. Condemine, N. Hugouvieux-Cotte-Pattat, and J. Robert-Baudouy. Characterization of pectin methylesterase B, an outer membrane lipoprotein of Erwinia chrysanthemi 3937. *Molecular Microbiology*, 19(3):455–466, February 1996.

[128] Z. Xie and H. Tang. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics*, 33(21):3340–3347, July 2017.

A Appendix

A.1 Additional tables

Table A.1: Residue-residue contact prediction performance comparison. The '-'-column indicates the number of proteins for which no prediction was returned within 24 hours of runtime. These proteins were not considered when measuring the performance of the respective methods. Bold values indicate the highest performance for a given measure/dataset combination. Despite the large amount of missing predictions PconsC2's performance is shown for completeness, but is greyed out.

Method	Threshold	P	R	$F_{0.25}$	MCC	-	P	R	$F_{0.25}$	MCC	-
	NewTrain					*NewTest*					
PSICOV	L/5	44.07	7.08	32.7	0.165	87	40.29	6.08	28.96	0.144	30
Freecontact	L/5	54.94	8.69	40.72	0.207	90	50.87	8	36.81	0.188	30
CCMpred	L/5	62.89	11	46.56	0.241	86	56.89	8.89	41.13	0.211	29
PconsC2	L/5	58.19	9.15	43.07	0.219	73	61.88	9.97	45.48	0.235	20
MetaPSICOV	L/5	64.07	10.01	47.46	0.243	87	60.42	8.93	43.48	0.22	30
MemConP	L/5	71.61	11.5	53.38	0.276	90	66.26	10.49	47.85	0.247	30
PSICOV	L/1	23.35	17.66	22.73	0.181	87	20.5	14.86	19.8	0.153	30
Freecontact	L/1	31.19	24.92	30.38	0.253	90	28.12	21.39	27.17	0.22	30
CCMpred	L/1	34.85	27.32	33.92	0.283	86	30.31	22.66	29.31	0.238	29
PconsC2	L/1	36.36	27.07	35.38	0.292	73	36.66	27.9	35.57	0.297	20
MetaPSICOV	L/1	40.02	29.91	38.94	0.326	87	36.64	25.4	35.31	0.284	30
MemConP	L/1	47.05	37.48	45.87	0.395	90	40.81	29.83	39.37	0.324	30
MemConP	L/x$^{F_{0.25}}$	65.85	16.03	54.5	0.312	90	59.65	14.11	48.14	0.272	30
MemConP	RFscore$^{F_{0.25}}$	65.85	23.41	53.83	0.357	87	63.53	15.96	48.12	0.289	29

Table A.2: Comparison of helix-helix interaction prediction methods. Bold values indicate the highest performance for a given measure/dataset combination.

Method	Threshold	P	R	$F_{0.25}$	MCC	P	R	$F_{0.25}$	MCC
		NewTrain				*NewTest*			
PSICOV	L/5	78.36	64.64	77.39	0.528	77.83	57.78	76.27	0.508
Freecontact	L/5	88.44	57.75	85.76	0.575	87.5	55.1	84.57	0.567
CCMpred	L/5	88.04	58.67	85.52	0.577	88.69	54.92	85.59	0.573
PconsC2	L/5	90.98	53.42	87.36	0.555	92.22	50	87.86	0.546
MetaPSICOV	L/5	86.28	50.91	82.9	0.511	86.02	42.93	81.23	0.474
MemConP	L/5	88.95	49.55	84.98	0.522	89.21	44.36	84.2	0.505
PSICOV	L/1	53.13	91.12	54.46	0.339	51.43	83.72	52.62	0.336
Freecontact	L/1	64.91	81.05	65.68	0.48	64.62	79.07	65.32	0.502
CCMpred	L/1	63.09	82.92	63.99	0.463	61.46	81.25	62.35	0.471
PconsC2	L/1	74.11	78.37	74.35	0.559	76.49	73.49	76.31	0.565
MetaPSICOV	L/1	67.51	78.68	68.08	0.499	69.6	74.96	69.9	0.534
MemConP	L/1	67.48	81.71	68.18	0.518	68.3	77.46	68.78	0.535
MemConP	L/x$^{F_{0.25}}$	91.45	43.89	85.97	0.501	91.88	38.46	84.94	0.479
MemConP	RFscore$^{F_{0.25}}$	91.79	44.05	86.29	0.504	90.15	42.58	84.59	0.498