Computational Systems Biology Methods

for Functional Classification of Membrane Proteins and Modeling of Quorum Sensing in *Pseudomonas aeruginosa*

Dissertation

zur Erlangung des Grades

des Doktors der Naturwissenschaften

der Naturwissenschaftlich–Technischen Fakultät III

Chemie, Pharmazie, Bio- und Werkstoffwissenschaften

der Universität des Saarlandes

von

Nadine S. Schaadt

Saarbrücken

April 2013

Tag des Kolloquiums:	30.08.2013
Dekan:	Prof. Dr. Volkhard Helms
Berichterstatter:	Prof. Dr. Volkhard Helms
	Prof. Dr. Hans-Peter Lenhof
Vorsitz:	Prof. Dr. Rolf W. Hartmann
Akad. Mitarbeiter:	PD Dr. Matthias Bureik

Parts of this work have already been published in or submitted for publication:

- Schaadt, N.S., Christoph, J., and Helms, V., "Classifying substrate specificities of membrane transporters from *Arabidopsis thaliana*", Vol. 50, p. 1899–1905, *Journal of Chemical Information and Modeling* (2010).
- Schaadt, N.S. and Helms, V., "Functional classification of membrane transporters and channels based on filtered TM/non–TM amino acid composition", Vol. 97, p. 558–567, *Biopolymers* (2012).
- Schaadt, N.S., Steinbach A., Hartmann R.W., and Helms, V., "Rule–based regulatory and metabolic model for Quorum sensing in *P. aeruginosa*", submitted to *BMC Systems Biology* (2013).

In addition, current aspects on functional classification of membrane proteins have been summarized in:

- Metzger, J., Schaadt, N.S., Hayat, S., and Helms, "Predicting structural and functional properties of membrane proteins from protein sequence", V., Vol. 7, p. 39–64. Ed. Wheeler, R.A., *Annual Reports in Computational Chemistry* (2011).

Acknowledgments

I am very grateful to those who helped me in my research work and in preparing this thesis. I thank especially:

- my supervisor Prof. Dr. Volkhard Helms for giving me the opportunity to work on two interesting projects and for his competent guidance. He was considerably helpful in successful completion of this thesis. Further, he offered me the possibility to participate at national and international conferences.
- Dr. Anke Steinbach of the research group "Drug Design and Optimization" in the Helmholtz Center for many helpful discussions and suggestions about the biological and pharmaceutical area in the Quorum sensing project.
- PD Dr. Michael Hutter for careful readings of my manuscripts and several scientific and non–scientific discussions.
- Ahmad Barghash for reading the functional classification part of my thesis and for fruitful comments.
- Dr. Tihamér Geyer for scientific discussions and suggestions in the area of Computational Systems Biology; Dr. Sikander Hayat who helped me in processing TMX; and Jan Christoph who was involved in the initial steps of the functional classification project.
- all former and current members of the whole Computational Biology group for providing a nice working atmosphere.
- Sebastian Schwarzbach and my family for their moral support and always encouraging me.

Abstract

Due to the function of membrane proteins and the effort required for experimental annotations, bioinformatical approaches to functionally classify uncharacterized sequences are desirable. For this, the similarity between sequences of different membrane proteins was statistically analyzed based on several amino acid compositions. To discriminate between functional classes, a ranking method was developed.

We showed that including further information in the amino acid composition and filtering into different sequence regions improved the classification quality. Subsets based on function achieved sensitivities of about 80%, whereas those of random subsets are in the range of 30–35%.

The pathogen *Pseudomonas aeruginosa* produces many virulence factors that are regulated by Quorum sensing. The number of infecting strains with antibiotic resistance is growing. Thus, new strategies focus on Quorum sensing inhibitors that target the regulatory pathways of virulence factors.

Pseudomonas aeruginosa contains three Quorum sensing systems that were simulated with an extended multi–level logical formalism to study the influence of Quorum sensing inhibitors on the autoinducer and virulence factor formation.

A topology analysis suggested that the proteins PqsR and PqsE act as receptors. Both are required together with an autoinducer to form pyocyanin. Enzyme inhibitors were more useful to block the autoinducer formation, whereas PqsR antagonists inhibited the pyocyanin biosynthesis stronger.

Kurzfassung

Aufgrund der Funktionen von Membranproteinen und dem Aufwand experimenteller Charakterisierungen sind bioinformatische Ansätze zur Klassifizierung unbekannter Sequenzen sinnvoll. Daher wurde deren Ähnlichkeit basierend auf verschiedenen Aminosäurenkompositionen bestimmt und statistisch analysiert. Eine Ranking–Methode wurde zur Einteilung in funktionelle Klassen entwickelt.

Wir konnten zeigen, dass die Vorhersagegenauigkeit durch Hinzunahme weiterer Informationen und durch Unterscheidung verschiedener Sequenzregionen verbessert werden kann. Proteingruppen mit derselben Funktion erzielten Sensitivitäten von etwa 80%, während zufällig zusammengestellte Gruppen nur 30–35% erreichten.

Der Krankheitserreger *Pseudomonas aeruginosa* produziert viele durch Quorum Sensing regulierte Virulenzfaktoren. Wegen der wachsenden Anzahl Antibiotika–resistenter Stämme greifen neue antibakterielle Strategien gezielt diese Regulationsmechanismen an.

Die drei Quorum Sensing–Systeme von *Pseudomonas aeruginosa* wurden mit einem erweiterten logischen Formalismus modelliert um den Einfluss von Quorum Sensing–Inhibitoren auf die Bildung von Autoinducern und Virulenzfaktoren zu untersuchen.

Eine Topologie–Analyse zeigte, dass die Proteine PqsR und PqsE anscheinend als Rezeptoren zusammen mit einem Autoinducer Pyocyanin regulieren. Enzym–Hemmstoffe waren besser geeignet, die Bildung von Autoinducern zu blockieren, während PqsR–Antagonisten die Pyocyanin–Biosynthese besser hemmten.

Zusammenfassung

Moderne Systembiologie umfasst experimentelle und computerbasierte Methoden zur Charakterisierung biologischer Zellen sowie theoretische Verfahren zur Untersuchung des Zusammenhangs zwischen einzelnen Komponenten. Diese Arbeit präsentiert Studien aus beiden Bereichen.

Funktionelle Klassifizierung von Membranproteinen

Entweder passive Kanäle oder aktive Transporter werden gebraucht um große und polare Substanzen durch Membranen zu transportieren. Die Aminosäurensequenz vieler Membranproteine ist heutzutage verfügbar. Wegen des enormen experimentellen Aufwandes ist deren spezifische Funktion dennoch oft unbekannt. Daher sind bioinformatische Ansätze zur funktionellen Klassifizierung und zur Vorhersage möglicher Substrate von Membrantransportern sinnvoll.

Wir betrachteten hauptsächlich Membranproteine der Subfamilien 1.A.1, 2.A.1 und 3.A.1 aus der "Transporter Classification Database" sowie Aminosäure–, Oligopeptid–, Phosphat– und Hexose–Transporter von *Arabidopsis thaliana*. Zur statistischen Analyse und funktionellen Klassifizierung dieser Datensätze wurden hypothetische Tests wie "Analysis of Variance" oder Wilcoxon–Mann–Whitney, und klassische Techniken des maschinellen Lernens wie "Support Vector Machines", "Principal Component Analysis" und hierarchisches Clustern ausgeführt. Hierfür wurden mehrere Features basierend auf der Aminosäurenkomposition der Proteine verwendet. Diese beinhalten unter anderen Nachbarschaftskorrelationen, physikochemische Eigenschaften und Konservierungsprofile. Außerdem wurde zwischen verschiedenen Sequenzregionen etwa abhängig von der transmembranen Lokalisierung oder abhängig vom Konservierungsgrad unterschieden. Zusätzlich wurde zum sequenzbasierten funktionellen Klassifizieren eine Rankingmethode entwickelt, die auf kleine Datensätze angewendet werden kann. Dabei wird ein kombiniertes Ranking mit der Euklidischen Distanz als Ähnlichkeitsmaß zwischen dem zu untersuchenden Protein und den Vertretern einer bestimmten funktionellen Klasse erstellt. Je nach betrachtetem Datensatz ist die funktionelle Klasse dabei entweder durch ein Durchschnittssuchprofil oder durch einzelne Kompositionen dargestellt. Die Vorhersagequalität wurde mit Standardverfahren gemessen und die Ergebnisse wurden mit zufällig generierten Resultaten verglichen. Desweiteren wurde der Zusammenhang zwischen Loops in den nichttransmembranen Regionen und der Funktion von Membranproteinen analysiert.

Zuerst haben wir gezeigt, dass sich die Aminosäurenkomposition zur Klassifizierung der untersuchten Datensätze eignet. Die Unterschiede zwischen Proteinen unterschiedlicher Funktion

waren in den transmembranen Regionen deutlich stärker ausgeprägt als in der gesamte Sequenz. Eine Beschränkung auf transmembrane Segmente verbesserte folglich die Vorhersagegenauigkeit. Die Rankingmethode konnte generell Proteingruppen derselben Funktion mit etwa 80%–iger Sensitivitäten klassifizieren, während zufällig zusammengestellte Proteingruppen zu Sensitivitäten im Bereich von 30–35% führten. Am Erfolgreichsten war die Komposition, die mit Konservierungsdaten aus Multiplen Sequenzalignments angereichert wurde.

Die Methode kann mögliche Substratklassen für unbekannte Transporter vorschlagen und somit den Suchraum für Experimente drastisch einschränken.

Quorum Sensing in *Pseudomonas aeruginosa*

Infektionen durch Krankheitserreger wie *Pseudomonas aeruginosa* betreffen insbesondere Patienten mit schwachen Immunsystem. *Pseudomonas aeruginosa* bildet Biofilme und produziert eine Reihe von Virulenzfaktoren, die unter anderem das Gewebe des Wirtes schädigen. Aufgrund der steigenden Anzahl resistenter Stämme sind neuartige Strategien als Alternative zu Antibiotika notwendig um die Bakterien zu bekämpfen. Die Pathogenität von *Pseudomonas aeruginosa* wird durch Quorum Sensing reguliert. Daher sind Quorum Sensing–Inhibitoren, die direkt diesen Signalweg blockieren, effektiv um Selektionsdruck zu vermeiden.

Wegen des Mangels an experimentellen Daten haben wir einen diskreten, regelbasierten Ansatz mit möglichst wenigen Parametern verwendet, um das Quorum Sensing von *Pseudomonas aeruginosa* zu modellieren. Boolsche Netze sind eine zu starke Vereinfachung für die erfolgreiche Simulation der drei Quorum Sensing–Systeme und ihrer hierarchischen Verbindung. Daher wurde ein logischer Mehrfachzustandsformalismus erweitert und auf den zugrundeliegenden Signalweg angepasst. Die Konzentrationsabhängigkeiten der Signalprozesse wurden dabei mit Hilfe von Schwellenwerten umgesetzt. Diese Durchführung der enzymatischen und regulatorischen Reaktionen innerhalb einer Zelle wurde in ein System, das eine vollständige Bakterienkultur darstellt, eingebettet. Ein wichtiger Bestandteil des Quorum Sensings ist die bakterielle Kommunikation, die durch Transportvorgänge zwischen einzelnen Zellen und der gemeinsamen Umgebung sichergestellt ist. Außerdem wurden Wachstumsprozesse und zufällig auftretende Mutationen hinzugefügt. Neben dem Wildtyp wurden auch Knock–out– und Gain–of–function–Mutanten sowie Quorum Sensing–Inhibitoren modelliert.

Die Durchschnittslevel von HHQ, PQS und Pyocyanin, die in großen Mengen vom Wildtyp gebildet werden, konnten von PqsBCD–Hemmstoffe verringert werden. Dahingegen wurde das Pyocyaninlevel von PqsR–Antagonisten nur bei sehr großen Inhibitionsleveln reduziert. Da dies ein Widerspruch zur Literatur ist und da auch $PqsE^-$ Mutanten zu viel Pyocyanin produzierten, scheint der Biosynthese–Pathway von Pyocyanin unvollständig zu sein. Daher wurden viele verschiedene Topologien des Netzwerkes analysiert. Danach scheinen PqsR and PqsE als

Rezeptoren die Pyocyanin–Produktion zu aktivieren und ein Autoinducer als Ligand zu fungieren. Außerdem stellte sich heraus, dass PqsBCD–Hemmstoffe besser geeignet sind um die Bildung von HHQ und PQS zu inhibieren, während PqsR–Antagonisten besser gegen Pyocyanin–Synthese wirken. Zusätzlich wurde eine gemischte Zellkultur simuliert, in der Wildtyp–Zellen mit Mutanten, die kein HHQ, PQS und Pyocyanin produzieren, sind. In dieser Kultur wurde mehr Pyocyanin gebildet als von den Wildtyp–Zellen möglich gewesen wäre, was auf die Transportprozesse zurückzuführen ist.

Contents

Part I.

Introduction

The first part of this work provides a detailed, sequence–based study of membrane transporters and their functional classification, given in Part II. For this approach, we used the sequences of membrane proteins from the database Orientations of Proteins in Membranes (OPM) [104], the Transporter Classification Database (TCDB) [156], and from *Arabidopsis thaliana* (*A. thaliana*) membrane transporters from the database Aramemnon [167]. Secondly, the Quorum sensing in *Pseudomonas aeruginosa* (*P. aeruginosa*) has been computationally analyzed in order to find targets considered in efficient drug development, presented in Part III. This section starts with a motivation. Then, the biological background of membrane proteins and of Quorum sensing regulatory pathways in *Pseudomonas aeruginosa* is explained.

Modern systems biology encompasses experimental and computational technologies for characterizing the inventory of biological cells as well as integrative, mostly computational approaches for studying the interplay of these components. This thesis presents work that falls into both of these areas.

Either highly selective passive channels or active carriers are needed to transport large and polar substances through membranes [17, 61, 153]. Membrane transporters also play a role in drug resistance. Nowadays, for a large number of membrane proteins, the amino acid sequence is resolved. The specific function of many putative transporters, however, is still not understood because of the huge amount of available data and the large experimental effort for substrate annotation. Consequently, computational methods are very useful to classify putative membrane transporters according their functional class and to predict their possible substrates. Computational sequence analysis can help experimentalists to decrease the search space while determining the function of new sequences or finding additional transporters. Unknown proteins are usually attributed to specific functional families mainly based on sequence similarity, which is obviously not successful for proteins sharing a similar function and having too different sequences. To distinguish between different proteins, features in which proteins are highly diverse are required. The continuous development and improvement of classification tools is further necessary. Nevertheless, an organization into family or sub–cellular localization is often not satisfying, but the transported substrates themselves are needed — especially for

drug design. Thus, this thesis focuses on substrate prediction of membrane transporters using the amino acid frequencies, physicochemical properties, and topological information.

Particularly, people with weak immune system are likely to become infected by pathogens such as *Pseudomonas aeruginosa*. This bacteria forms biofilms and produces several virulence factors that disrupt tight junctions, kill epithelial cells, and damages tissues. Due to an increasing number of resistant strains, novel strategies have to be developed instead of using current antibiotics. Since the pathogenicity of *Pseudomonas aeruginosa* is regulated by a population density dependent signaling mechanism (Quorum sensing), the direct inhibition of this signaling pathway may be useful to avoid the selection of resistant strains. In this thesis, the topology and the dynamical behavior of the *las*, *rhl*, and *pqs* Quorum sensing systems was analyzed with a multi–level logical formalism. The influence of inhibitors of PQS biosynthesis and antagonists of receptor PqsR, which up–regulates the biosynthesis, on formation of the signaling molecules HHQ and PQS, as well as the virulence factor pyocyanin were compared. To overcome resistance, it is important to know how resistance is developed. Then, the medication can be adapted to avoid selection pressure. For this, a study was started with randomly occurring mutations to analyze their effect on bacterial growth and virulence factor formation.

1. Membrane Proteins

Membrane proteins fulfill a variety of essential cellular functions and are so frequently used as drug targets. About 60% of the available pharmaceutical compounds interact with membrane proteins [129]. Despite of their physiological importance, membrane proteins are not well understood because of immense effort of biological experiments, such as the difficulty of X–ray diffraction for membranes. Often, their structure and function are not annotated while their sequence is known due to large number of genome projects in the recent years. Accordingly, it is highly desirable to analyze the sequences of membrane proteins with computational tools and to develop new sequence–based methods for predicting their function.

A biological membrane is a lipid bilayer structure that forms a semi–permeable barrier either between different cell compartments or between cells and their environments. Membranes consist of amphiphilic phospholipids with a hydrophilic head group with a negatively charged phosphate group pointing into the polar surrounding of the membrane and a hydrophobic tail. These aliphatic tails of the phospholipids, long fatty acid hydrocarbon chains, are in the interior of the bilayer. Membrane proteins are attached to such a bilayer either in a peripheral, non–covalent bonded, or an integral way, illustrated in Figure 1.1. Integral membrane proteins are

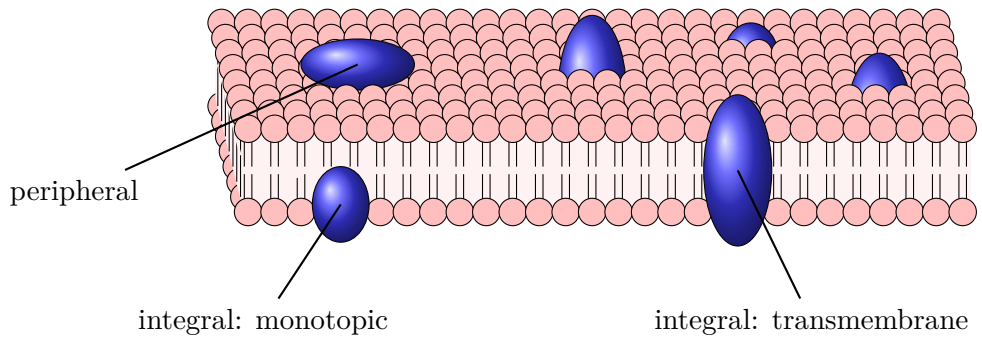


Figure 1.1.: **Fluid mosaic model** introduced by Singer–Nicholson [175] containing phospholipids and membrane proteins illustrated as blue ellipsoids.

permanently embedded into the membrane either as bitopic and polytopic transmembrane or as monotopic proteins. While integral monotopic proteins are only located on one side of the lipid bilayer, transmembrane proteins span through the whole hydrophobic interior of the membrane consisting of an intracellular domain, a transmembrane region, and an extracellular domain.

Since the physicochemical characteristics of membranes vary clearly from those of water, there are likewise strong differences between transmembrane and water–soluble parts of proteins in their amino acid frequency and structural properties. The globular integral membrane proteins must also be amphiphatic as the phospholipids. Polar amino acids in the transmembrane segment (TMS) of the protein must be isolated from the hydrophobic environment. This mixed structure between lipids and some proteins was introduced by Singer and Nicholson in 1972 as the so–called fluid mosaic model [175]. However, the thickness of a membrane is highly diverse due to adaption of lipids to the large number of proteins [46].

1.1. Structural Features

So far, known X–ray structures show that transmembrane proteins comprise either pure α–helical bundles or β–barrels, see Figure 1.2. Since there is no way to form hydrogen bonds in the non–polar inside of the membrane, no other secondary structure motifs have been observed so far than the commonly found pure α–helical bundles and β–barrels are possible for transmembrane proteins. The central helix running through the pore of a β–barrel is the only exception, for example, the human voltage–dependent anion channel (PDB entry 2JK4, see Figure 1.2). While β–barrel membrane proteins are mostly channels used for mediating the

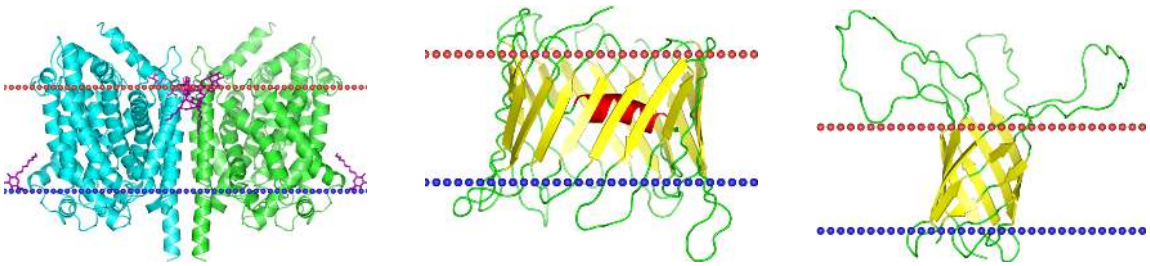


Figure 1.2.: **Crystal structures with their orientation in the membrane:** Left: α–helical leucine transporter LeuTAa (PDB–ID: 2A65). The red dotted line denotes the boundary to the periplasmic side and the blue dotted line that to the cytoplasmic side. Middle: Human voltage–dependent anion channel (PBD–ID: 2JK4). The red dotted line denotes the boundary to the intermembrane space side and the blue dotted line that to the cytoplasmic side. Right: β–barrel OprH from *Pseudomonas aeruginosa* (PDB–ID: 2LHF). The red dotted line denotes the boundary to the extracellular side and the blue dotted line that to the periplasmic side. Graphics obtained from the OPM database[1][104].

passive passage of substrates across the outer membrane of Gram–negative bacteria, chloroplasts, or mitochondria, most membrane transporters in the plasma membrane that catalyze active transport of substances are usually α–helical [164]. In general, 20–30% of all proteins in a genome are α–helical integral membrane proteins [198].

[1] `http://opm.phar.umich.edu/`

The topology of membrane proteins describes the number of membrane–spanning segments and their orientation in the membrane. In general, α–helical transmembrane proteins have a relatively simple topology resembling a helical bundle of several transmembrane helices (TMHs) connected by loops or other secondary structure elements outside the membrane. Often, membrane proteins with many TMSs have fewer residues outside the membrane than membrane proteins with few TMSs [198]. Since the TMS passes the membrane several times, in polytopic proteins, pores of ion channels can be switched between open and closed states. However, some helical membrane proteins also consist of more complicated structural domains, such as re–entrant regions where the TMS enters and exits the interior of the membrane on the same side, short helices followed by a transmembrane loop, connections between two half–helices, or interfacial helices that are oriented parallel to the membrane surface [193, 209]. In water–soluble β–barrels, such as the green fluorescent protein, the residues are oriented such that the interior of the barrel forms a hydrophobic core and the outside is polar. In membrane β–barrels, this pattern is actually inverted, whereby their membrane spanning regions, the transmembrane β–strands (TMBs), are formed by anti–parallel β–strands creating a barrel [165]. They are more polar in the pore than outside. Therefore, they are also called *inside–out* proteins [183].

Polytopic membrane proteins containing more than one TMS need to be inserted into the membrane via the Sec translocon [212]. To ensure that the first N–terminal TMH is correctly targeted to the membrane, it is normally more hydrophobic than the following TMHs [14, 134]. In comparison to helices in water, the interior of TMHs is more tightly packed with some more loosely packed interhelical regions with polar residues that line the substrate binding sites or pores [44]. In water–soluble proteins, the amino acids glutamic acid, methionine, alanine, and leucine are helix formers, while proline and glycine do mainly not occur in helices due to their size and chemical structure [27]. Valine, isoleucine, and tyrosine are often contained in β–sheets, while aspartic acid, glutamic acid, and again proline are rarely included. Glycine, proline, aspartic acid, asparagine, and serine, which mostly occur in loops, are generally known as helix and sheet breakers in water–soluble proteins [19]. However, in membrane proteins, glycine plays a different structural role. Instead of disrupting the secondary structure, it stabilizes the structure of transmembrane proteins. Due to its short side chain, glycine is important for facilitating closer packing of helices, helix–helix interactions, and crossing points [74]. At helix crossing points, it is the most frequent amino acid. Furthermore, glycine is often found as GxxxG motif in membrane α–helices [170]. In transmembrane regions, the amino acids alanine, glycine, isoleucine, leucine, phenylalanine, serine, threonine, and valine generally occur repeatedly mainly because of the more non–polar properties of most of these amino acids. Taken together, these amino acids constitute for 75% of all amino acids in transmembrane regions [79]. The positively charged amino acids arginine and lysine occur more frequently in the loops exposed to the cytoplasmic region which is described by the so–called positive inside rule [195, 196]. The high diversity between transmembrane, cytoplasmic, and peripheral regions has been exploited by several topology predictors (see Section 1.2).

1.2. Topology Prediction

Nowadays, the crystal structures of about 400 unique membrane proteins are available[2]. Due to this still quite small number of known three–dimensional (3D) structures for membrane proteins, we used for our analysis available computational tools that assign putative membrane–spanning segments from protein sequence. Today, predicting the α–helical and β–strand TMSs from protein sequence is a well–established area in bioinformatics.

As already mentioned, the hydrophobic interior of the membrane results in a higher frequency of hydrophobic and apolar amino acids in TMHs than in other sequence regions. In 1986, von Heijne made a seminal discovery that the loops enriched with positive amino acids, such as arginine and lysine, tend to be located on the cytoplasmic side [195]. Topology predictors often combine a hydrophobicity analysis [8] and the positive inside rule together in a machine learning technique, such as a hidden Markov model or a neural network [88, 111]. For example, SPLIT 4.0 searches the preferred structural region of an amino acid using charged motifs [80].

Most modern prediction methods to identify TMHs are also based on sequence profiles integrating sequence information of many homologs. For this, a position–specific scoring matrix (PSSM), see Figure 1.3, is usually generated by the position–specific iterative basic local alignment search tool (PSI–BLAST) [4, 161] including a row for each amino acid and a column for each position in the corresponding amino acid sequence. Derived from a multiple sequence alignment (MSA) with a certain number of different sequences, the PSSM entry $s_{i,j}$ contains the score for the expected a priori probability and the likelihood that an amino acid at position j in the protein sequence is replaced by amino acid i, whereby a more frequently than expected occurrence of a certain substitution is represented by a positive value.

In 2007, Nugent and Jones presented the prediction tool Memsat3, which predicts the localization of TMHs in a protein sequence based on a hidden Markov model [78]. The improved version Memsat–SVM of 2009 uses a support vector machine (SVM) that discriminates between TMHs and non–TMHs, as well as inside and outside loops, re–entrant helices, or signal peptides. The method is based on sequence profiles that integrate the information of many homologous protein sequences in the form of a PSSM. Its accuracy has been reported as 95% for detecting TMHs and 89% for individual residues [122]. However, Tsirigos *et al.* evaluated the performance of several topology predictors on experimental data sets and reported that Memsat–SVM performed not very well for their data sets [187]. Similarly, the consensus tool TOPCONS [15] combining the profile–based hidden Markov models PRO–TMHMM and PROVID–TMHMM [191], the model based predictors SCAMPI–single and SCAMPI–multi

[2] In January 2013, the membrane protein data bank (MPDB) [144] contained 407 membrane proteins and the database Membrane Proteins of Known 3D Structure (`http://blanco.biomol.uci.edu/mpstruc/listAll/list`) 377 membrane proteins.

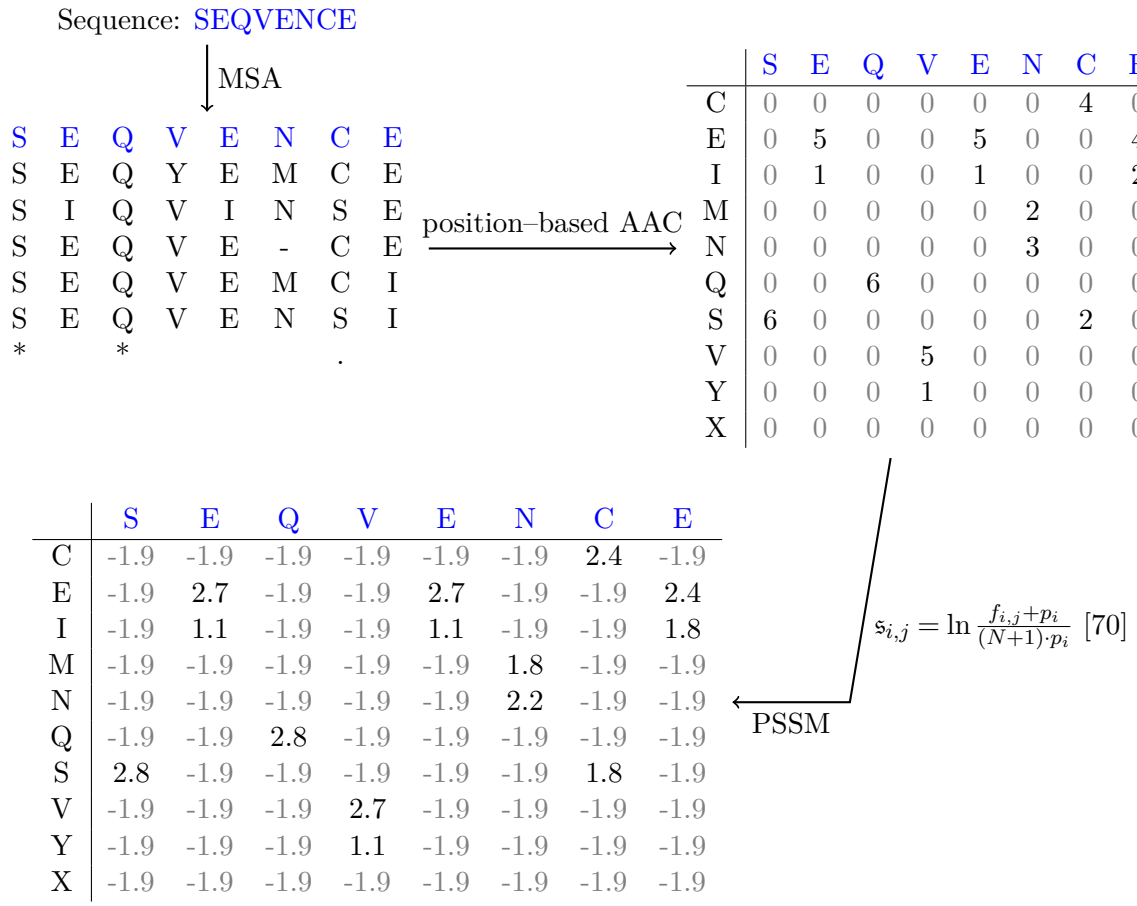


Figure 1.3.: **Position–specific scoring matrix (PSSM)** $\mathfrak{s}$ for an example sequence with 20 rows (number of amino acids) and eight columns (sequence length). The occurrence $f_{i,j}$ of each amino acid type i at every sequence position j (position–based AAC) in an multiple sequence alignment (MSA) with N sequences is required as intermediate calculation step. Here, X denotes any not listed amino acid. In this example, the expected a priori probability p is 0.05 for all amino acids.

[14], and OCTOPUS [192] detects the number and localization of TMHs with a reported accuracy of 83%. OCTOPUS uses a sliding window with PSSM profiles as input of a neural network together with a hidden Markov model. We compared the results of the current most accurate TMH predictors Memsat–SVM and TOPCONS for some membrane proteins with known 3D structure with each other. Finally, we decided to use Memsat–SVM for our classifications. To set up the data set, we also considered the secondary structure elements predicted by SPLIT 4.0.

Topology predictors of β–barrel membrane proteins often implement such general rules as an even number of TMBs or the anti–parallel connection between direct neighboring TMBs [165]. TMBETAPRED–RBF detects TMBs with an accuracy of 96% for detecting transmembrane strands and 87% for individual residues [128]. It uses a PSSM and the composition of the

amino acids alanine, aspartic acid, histidine, tyrosine, and valine as input features for a radial basis network.

1.3. Cellular Function

Membrane proteins are potential targets for the design of pharmaceutical compounds and are also associated with bacterial virulence and multi–drug resistance due to their important soles in many crucial processes of cells [30, 101, 120]. First of all, structural membrane proteins stabilize the membrane itself as well as the whole cell by attaching it to the cytoskeleton and the extracellular matrix. Membrane proteins functioning as cell adhesion molecules connect neighboring cells with each other and enable interactions. In general, cell surface receptors activated by an extracellular signal initiate a response via intracellular signal transduction [179]. For example, G–protein coupled receptors (GPCRs) recognize a ligand, such as light–sensitive compounds, odors, pheromones, hormones, and neurotransmitters, change their conformation, and thus activate the G protein [151]. Further, membrane proteins may work as enzymes that, e.g., generate chemical energy. In photosynthesis, helical membrane proteins are very important for catalyzing oxidations and reductions. Moreover, cell membrane adhesion molecules are also involved in signaling of immune regulation, activating pathways, and in inflammations [184]. Transduction receptors cause physiological cellular responses when interacting with a signaling ligand. Therefore, they are often related to cancer [184].

1.3.1. Transport of Substrates

The first part of this thesis is focused on another major function of membrane proteins, namely the transport of substrates. In all organisms, substances required for biosynthesis must be transported to the compartments where synthesis takes place and the synthesized products to their consumption or storage locations. On the other hand, organisms have to release biowaste and toxic compounds to their environment. In both cases, different substances have to cross membranes. Small, hydrophobic molecules are able to permeate the membrane by osmosis without involvement of any transporting protein. The rate of diffusion depends on the state, the size, and the polarity of the substance. Such molecules as oxygen, benzene, glycerol, or hemoglobin, for example, can simply diffuse across membranes [89]. However, phospholipid bilayers are impermeable for polar and charged substances, such as ions, sugars, or amino acids. The passage of these and other larger molecules needs to be catalyzed by passive and active transport processes.

Often, transmembrane proteins with a 3D structure that enables conformational transitions are required for transport across membranes. In passive transport, the molecules go through the membrane with the concentration gradient faster than using standard osmosis. Ordinarily,

channels and carriers mediate passive transport. Channels, such as membrane–spanning β–barrel porins or transmembrane α–helical ion channels, are highly selective pore proteins which are gated either by the membrane potential (voltage) or by binding of ligands. They can transport metabolites, such as inorganic ions, sugars, amino acids, or nucleotides [66]. Porins usually transport water or glycerol. Therefore, one distinguishes between aquaporins and glycerolporins. Carrier proteins acting as uniporters, symporters, or antiporters are specific for a single molecule or a group with very similar characteristics as substrates. The transported substances attach to the binding site of the carrier. The transport cycle of uniporters involves a conformational change and transports only a single molecule at a time. In contrast, symporters transport two different molecules in the same direction; and antiporters transport one molecule in one direction and another molecule in the opposite direction. Carriers often transport such metabolites as sugars, amino acids, and nucleosides [148] and are selective for special substrates [17, 61, 153].

In active transport, where one generally distinguishes between primary, secondary, tertiary active transport, and group translocation, substrates are transported against concentration gradients between the two compartments separated by the lipid bilayer [89]. Thus, for those processes energy is required. Primary active transporters, such as ATP binding cassette (ABC) transporters or large ATPases, use the energy delivered by ATP hydrolysis. ABC transporters have a wide range of substrates, e.g., ions, amino acids, peptides, sugars, small toxins, or lipids. ABC transporters, such as multi–drug resistance–related proteins, transport as efflux pumps antibiotics and other medications out of the cell. Therefore, bacteria have multi–drug resistances [95]. In addition, ABC transporters play a considerable role in virulence and pathogenicity of bacteria [31, 34]. Consequently, these transporters are interesting for antimicrobial drug development. On the other hand, proteins belonging to the secondary active transporters, such as proteins of the major facilitator superfamily (MFS), utilize energy from electrochemical potential differences by transporting a further type of substrates [130]. Today, several crystal structures of active membrane transporters are available revealing the molecular basis for active transport across the membrane.

It is known that transporters are often able to carry out the transport of various substrates [94], and that often multiple transporters exist for one particular substrate. It is therefore quite likely that only computational methods may be able to provide an integrated picture of the transportome. Moreover, transporters selective for the same substrate may have very different amino acid sequences. For example, the phosphate transporters AtPHT2 and AtPHT4 have a BLAST [3] E–value of 4.1. Furthermore, proteins with very similar sequence have possibly different functions, such as the phosphate transporter Pho84 in *Saccharomyces cerevisiae* and the sugar symporter AtSTP14 that have an E-value of $8 \cdot 10^{-14}$.

1.4. Heterogeneity

Obviously, a high heterogeneity is important in the context of classifications. As already mentioned, membrane proteins can be distinguished into α–helical or β–strand and into single–pass or multi–pass. Additionally, transmembrane proteins have a high diversity in sequence and structural aspects as a different ratio of TMSs and loops. In uni–cellular organisms, membrane proteins with six TMHs are mainly transporters for small solutes and membrane proteins with twelve TMHs are amino acid, sugar, or ABC transporters [198]. In contrast, GPCRs have seven TMHs [151]. Due to a possible relation to function, those differences can be used for classifications.

TMSs vary strongly from water–soluble parts due to different environmental conditions. Further, buried parts of TMSs are generally strongly conserved [78] and TMSs are more conserved than external loops [194]. Due to the direct interaction with the substrate, it can be assumed that the TMSs are related to function. Therefore, a detailed analysis of TMSs in the context of functional classification could be useful, see Section 6.5.

2. Functional Classification of Membrane Proteins

Although the amino acid sequences of many putative membrane proteins are known, often the roles they play in an organism has not been recognized by experiments due to the huge effort involved in characterizing an unknown function of a protein. Hence, computational strategies are required for searching the function of membrane proteins and in the case of membrane transporters to identify putative transported substrates or substrate groups.

2.1. Homology, Phylogeny, and Motifs

Since a large sequence identity usually implies similar tasks of the proteins [143], the commonly used way of classification investigates significant sequence similarities to functionally annotated proteins by pairwise BLAST [3] alignments. Then, the unknown sequences are often assigned to the protein family of a highly similar sequence [149]. Those approaches rely on the idea that a sequence identity may be correlated to homology which is given when two proteins share many features inherited from a common ancestor. The tool TransportTP incorporates a SVM including number of TMSs, Pfam domains, and Gene Ontology terms into a sequence similarity search to avoid false positives determined by purely homology analysis [98]. Homology represents either paralogy, duplications in a genome, or orthology, duplications descended from a common ancestor. The role of a protein is generally conserved in orthologs. Consequently, sequence similarity is not necessarily associated to structure or function such that it is more reliable to detect functional connections by phylogenetic analysis.

In general, it is assumed that a similar structure or evolution of proteins indicates a similar function. Hence, some recognition methods for membrane proteins incorporate such phylogenetic information or multiple sequence alignments [36, 154, 155]. Here, the evolutionary relationship between certain features of different organisms is studied. As evolution is a branching process in which each species derived from a single organism, a rooted and directed graph is usually generated on the basis of molecular sequencing as well as morphological data. In this phylogenetic tree, a node represents a certain species and an edge represents the connection to the direct ancestor or descendants of the species, whereby the number of observed mutations between related sequences is indicated by the length of the edges or the edge weights. Thus, the path from the root to the leaves describes the evolution of the organisms. Besides evolution

between organisms, a modified phylogenetic tree is also useful to consider individual genes. Since similarity in sequence is just a prerequisite for homology, additional characteristics, such as sequence length, topology, or 3D structures, are required to indicate homology and build a phylogenetic tree [109].

Some predictors [110] use the fact that protein areas important for the function may be represented by sequence motifs shared by most or all members of a protein family or by other conserved residues. Therefore, membrane transporters have been clustered with a reported positive classification rate of 72.3% by a combination of motifs, topology, and homology [99]. A motif is typically linked to the structure of a protein, but not necessarily to its purpose. Further, it could be the case that some family members do not contain all motifs. Here, ignored sequence areas may include important information.

2.2. Amino Acid Composition

The so–called amino acid composition (AAC), a vector containing the frequency of each single amino acids over the full protein sequence, is a very robust and simple method to assign proteins according to their structure or function, see Figure 2.1. In 1983, Nishikawa *et al.* introduced

Sequence: SEQVENCE $\xrightarrow{\text{AAC}}$

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	1	0	3	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0

Figure 2.1.: **Amino acid composition (AAC)** for an example sequence.

the AAC to discriminate enzymes from other proteins and located out of or into a biological cell [121]. Besides many other applications the AAC has since then been used for secondary or tertiary structure prediction [10, 42, 56, 152] or to functionally characterize GPCRs [72].

Gromiha and Yabuki used the AAC in machine learning techniques to sort membrane proteins from TCDB into channels/pores, electrochemical potential–driven transporters and primary active transporters [62] with a reported accuracy of about 68% for neural networks. Additionally, the k–nearest neighbor clustering algorithm was utilized to distinguish between α–helical or β–barrel proteins and reached an accuracy of 85%. Accordingly, asparagine, which is related to the stability and the role of β–barrel membrane proteins, and tyrosine are more frequently contained in pores or channels than in carriers or transporters, whereas the amino acids glutamine, glycine, isoleucine, and valine are prevalent in electrochemical potential–driven transporters.

2.2.1. Pair Amino Acid Composition

Moreover, Park and co–workers predicted sub–cellular localizations using the improved pair amino acid composition (PAAC) [131]. Here, a vector with 400 components is calculated, where a frequency for each possible pair of amino acids represents an entry, see Figure 2.2. Thus, it is counted how frequently a certain amino acid, e.g., asparagine, follows another

$$\text{Sequence: SEQVENCE} \xrightarrow{\text{PAAC}} \begin{pmatrix} \text{SE} & \text{SQ} & \text{SV} & \text{SN} & \text{SC} & \text{SS} & \text{EE} & \text{EQ} & \text{EV} & \text{EN} & \text{EC} & \text{ES} & \cdots \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & \cdots \end{pmatrix}$$

Figure 2.2.: **Pair amino acid composition (PAAC)** for an example sequence.

amino acid, e.g., glutamic acid, in the protein sequence. Whereas the location in a sub–cellular compartment and the purpose of a protein are directly associated, while information about the specific substrate of a transporter is not contributed.

Membrane proteins have been also classified by a radial basis function network with a reported accuracy of 75.4% regarding AAC, PAAC, PSSM, and PSSM including biochemical features [126, 127].

2.2.2. Pseudo Amino Acid Composition

Chou *et al.* developed the pseudo amino acid composition (PseAAC) to prevent ignoring essential information of biochemical properties, such as hydrophobicity and hydrophilicity values, mass, pK values, or isoelectric point of the involved amino acids [23]. An entry v_i of the $20+\lambda$ entries containing PseAAC is given in equation (2.1),

$$v_i = \begin{cases} \dfrac{f_i}{\sum_{j=1}^{20} f_j + \omega \sum_{j=1}^{\lambda} \tau_j}, & \text{if } 1 \leq i \leq 20 \\ \dfrac{\omega \tau_{i-20}}{\sum_{j=1}^{20} f_j + \omega \sum_{j=1}^{\lambda} \tau_j}, & \text{if } 20 \leq i \leq \lambda \end{cases} \tag{2.1}$$

whereby f denotes the amino acid frequencies and ω a weighting factor. The normalization is based on the sequence length of the corresponding protein. Correlations of physicochemical characteristics between residue pairs are indicated in τ_j, defined as in equation (2.2),

$$\tau_j = \frac{1}{n+1-j} \sum_{k=1}^{n+1-j} T(\mathcal{R}_k, \mathcal{R}_{k+j+1}). \tag{2.2}$$

whereby n here represents the corresponding sequence length. The sequence order correlation between all the residues of a protein chain is schematically drawn in Figure 2.3. The set *cha*

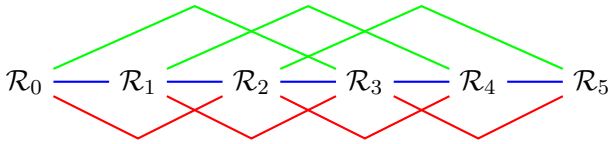


Figure 2.3.: **Illustration of the sequence order correlation** of the first five residues ($\mathcal{R}_1$–$\mathcal{R}_5$) whereby the blue lines denote τ_0, the red lines τ_1, and the green lines τ_2 [23].

contains all considered biochemical characteristics. The correlation function T, as in equation (2.3),

$$T(\mathcal{R}_a, \mathcal{R}_b) = \frac{1}{|cha|} \sum_{l=1}^{|cha|} (cha_l(\mathcal{R}_a) - cha_l(\mathcal{R}_b))^2 \tag{2.3}$$

represents the average distance of the amino acid properties between the residues $\mathcal{R}_a$ and $\mathcal{R}_b$. Therefore, parts of the long–range sequence–order information are included in the PseAAC. It has been applied to a large variety of properties, such as to predict sub–cellular localizations [23], protein cellular attributes [24], enzyme subfamilies [25], structures [100], and function of GPCRs [206]. Further, the PsePSSM including these property correlations into PSSMs was applied in an evidence–theoretic k–nearest neighbor operator to identify the type of membrane proteins [26].

In 2012, several tools combine different AAC variations and further characteristics in the context of functional prediction of membrane proteins, e.g., Proclusensem [200]. Yu *et al.* used a SVM containing amino acid substitution matrices and auto covariance transformation to specify sub–cellular localizations of apoptosis proteins [211]. Another program includes AAC, PAAC, PsePSSM, Pseudo–average chemical shift, and physicochemical properties into a SVM [47]. Hayat *et al.* presented Mem–EnsSAAC, an ensemble calculator for membrane protein integrating SVM, probabilistic neural network, k–nearest neighbor, Adaboost, and random forest. It is based on PseAAC, discrete wavelet analysis, and the so–called split amino acid composition (SAAC) for which the protein sequence is divided into three parts: a) 25 amino acids of N termini, b) 25 amino acids of C termini, and c) region in–between [64].

2.3. Physicochemical Properties

Beside sequence similarity, sequence motifs, and AACs, physicochemical properties of amino acids were used to functionally classify membrane proteins [32]. Therefore, Davies *et al.* set up five z–values incorporating lipophilicity ($z1$), steric properties ($z2$), polarity ($z3$), and electronic effects ($z4$ and $z5$) with a principal component analysis (PCA). Then, a protein is defined by a matrix consisting of these z–values for each residue. GPCRTree discriminates between family levels hierarchically at each node level. At these different levels, several classification ways were applied based on the z–values [33, 168].

3. Quorum Sensing in *Pseudomonas aeruginosa*

The communication of bacteria and the following regulation of certain genes depending on the population density of the cell culture is called Quorum sensing. A famous bacteria controlling processes via Quorum sensing is the human pathogen *Pseudomonas aeruginosa*. This bacterium is multi–resistant against currently available antibiotics that block bacterial growth. Since the production of virulence factors and the formation of biofilms are regulated by Quorum sensing systems, it has been suggested that developing new strategies to combat Quorum sensing may lead to alternative treatment options [69, 146].

3.1. Quorum Sensing

The competence of microorganisms to measure the density of their population by communication and to react has been termed Quorum sensing. It describes a regulation of specific genes depending on the density of small chemical signaling molecules, so–called autoinducers [7]. Inefficient processes for a single cell, such as biofilm formation or bioluminescence, are often applied by Quorum sensing [202]. In the case of a small cell density, the transcription of the corresponding genes is at a basal rate [38].

Several Gram–negative bacteria use Quorum sensing mechanisms, e.g., *Agrobacterium tumefaciens* to initiate tumor growth in its plant host cell. Quorum sensing is also widespread in Gram–positive bacteria, such as in *Streptococcus pneumoniae* to reach the "competent state" for genetic transformation, in *Bacillus subtilis* for sporulation, or in *Staphylococcus aureus* for virulence systems [118]. Many of those bacteria are human pathogens with multi–resistance against most current antibiotics such that it is important to understand the Quorum sensing mechanisms in detail.

In general, autoinducers can freely diffuse through cell membranes due to their small size and amphiphilicity. Alternatively, there are other transport processes for the autoinducers. In this way, autoinducers can be indirectly used to measure the population density. In the case of Gram–negative bacteria, most autoinducers are acyl homoserine lactones. Figure 3.1 gives the chemical structure of acyl homoserine lactones. In the case of Gram–positive bacteria, the

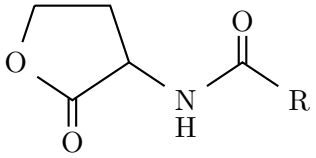


Figure 3.1.: **Chemical structure of acyl homoserine lactones:** Most autoinducer in Quorum sensing are acyl homoserine lactones in the case of Gram–negative bacteria.

autoinducers are oligopeptides that are typically transported by ABC transporters through the membrane [118].

Figure 3.2 shows a schematic illustration of Quorum sensing. Autoinducers form together with certain receptors complexes. Such complexes build dimers or higher multimers. The monomer or dimer binds specifically to an operon to up–regulate the gene expression of autoinducer synthases as well as biofilm or light producing proteins. Thus, a positive feedback loop is generated because these autoinducers initiate their own production. However, the up–regulation requires a certain threshold concentration of the autoinducer since the regulated process should only be produced if a certain cell density is exceeded [118, 119]. Often, there are multiple positive feedback loops each with their own autoinducers that are linked to each other in some way.

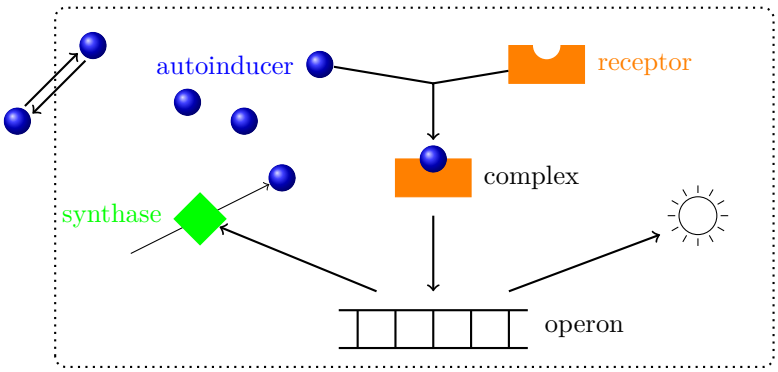


Figure 3.2.: **Schematic illustration of Quorum sensing:** Autoinducers (blue balls) and receptors (orange rectangles) form a complex which up–regulates the gene expression of an operon. Squares in green represent the autoinducer synthase. The sun denotes the controlled process, such as production of virulence factors, biofilm formation, or bioluminescence.

In the earlier 1970s, the *lux* system [53] of the marine Gram–negative bacterium *Vibrio fischeri*, which colonizes the *Euprymna scolopes*, has been described as first Quorum sensing system. The *lux* system consists of an acyl homoserine lactone molecule as autoinducer, the receptor LuxR, and the autoinducer synthase LuxI. Together with a second Quorum sensing system, the *ain* system (with another acyl homoserine lactone and the synthase AinS), it is responsible for the up–regulation of the *lux* operon, which encodes the bioluminescent luciferase [119].

3.2. *Pseudomonas aeruginosa*

Pseudomonas aeruginosa (*P. aeruginosa*) is a Gram–negative aerobic and rarely also anaerobic bacterium that is able to live in a large variety of often times unaspiring environments, shown in Figure 3.3. It colonizes in soil, in water, or in other organisms, in which it is independent

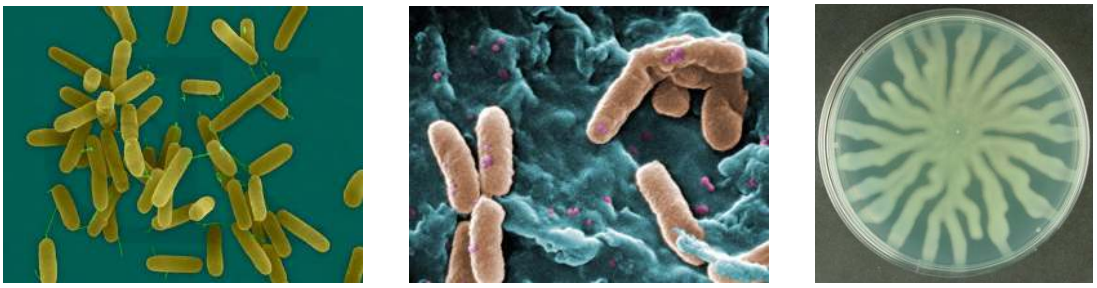


Figure 3.3.: ***Pseudomonas aeruginosa:*** Shown is the bacteria under a microscope. On the right side, swarming is illustrated. The picture on the left side was obtained from the *Pseudomonas* genome database[1], the photo in the middle from the *Pseudomonas aeruginosa* website[2], and the picture on the right from Verstraeten *et al.* (2008) [190].

of the tissue. Due to its wide range of food sources and its high adaptability, it can appear in environments in which other microorganisms miss nutrients or are killed by toxic compounds.

3.2.1. Pathogenicity

The generally harmless *P. aeruginosa* is an opportunistic pathogen for plants and animals [45]. It is typically dangerous for patients with immune system deficiencies, such as cancer or burn patients. Then, it causes a large number of diseases, such as pneumonia, sepsis, or infected wounds.

Often, the respiratory tract of cystic fibrosis patients is affected by *P. aeruginosa*. Then, the bacteria causes inflammations and breathing difficulties. Cystic fibrosis is an autosomal recessive inherited metabolic disorder with a mutation in the cystic fibrosis transmembrane conductance regulator gene encoding an ABC transporter in epithelial cells. Mainly affected tissues are lungs, pancreas, and respiratory tract.

In 2011, about 10% of hospital–acquired infections were related to this microorganism [35]. Besides producing a large number of toxins, *P. aeruginosa* forms biofilms.

[1] `http://www.pseudomonas.com/images/paeruginosa.jpg` in 2012
[2] `http://www.pseudomonas-aeruginosa.de/bilder/pseudomonas-aeruginosa.jpg` in 2013

Virulence Factors

P. aeruginosa produces several virulence factors, such as the enzymes elastase or alkaline protease, rhamnolipids, phenazines, cyanide, and pyocyanin. Figure 3.4 shows some 3D and chemical structures of virulence factors that were used in our simulations. The toxic compounds

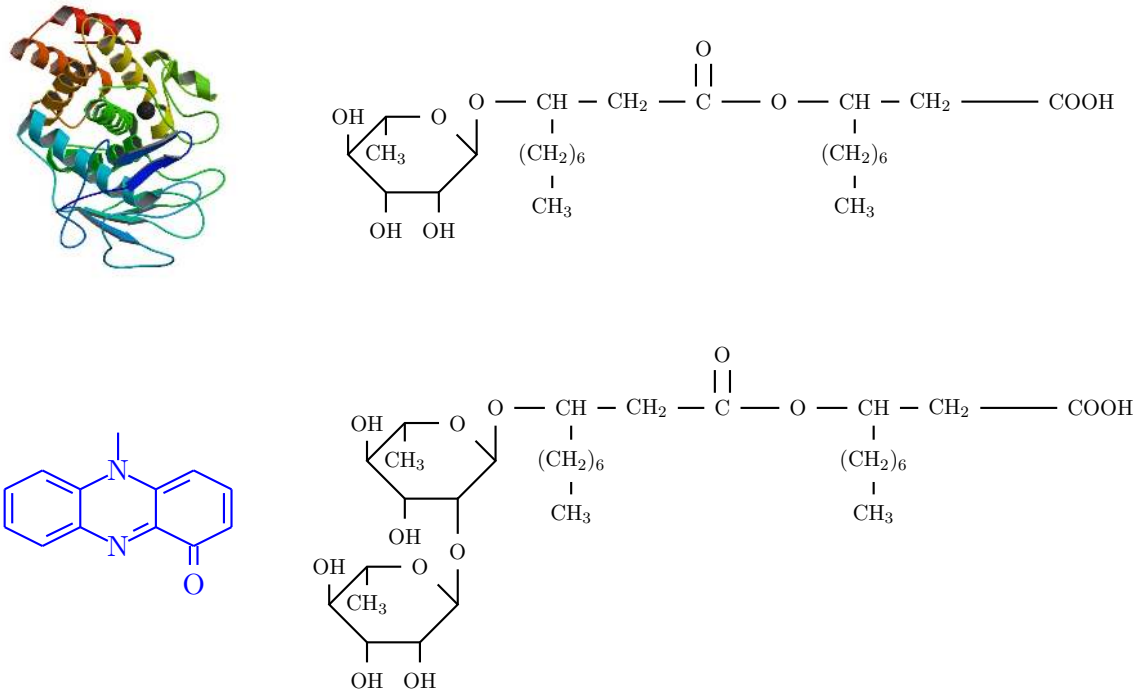


Figure 3.4.: **Virulence factors:** Shown is the three–dimensional structure of elastase (PBD–ID: 1EZM) obtained from the PDB[3][12] on the left side in the first row, the chemical structure of pyocyanin on the left side in the second row, and the chemical structures of the rhamnolipids on the right side.

kill host cells, disrupt membranes or whole tissues, and are able to interact with the host defense mechanisms. Elastase is a protease that breaks peptide bonds. In infections caused by *P. aeruginosa*, elastase destroys tight junctions such that tissues are proteolytically damaged. In human lung tissues, it further disrupts elastin, which is required for lung expansion and contraction [189]. Additionally, elastase interrupts the human immune system.

Amphiphilic glycolipids, so–called rhamnolipids, together with their precursors are required for swarming and biofilm formation [190]. They boost the formation of microcolonies and bacterial migration. Rhamnolipids build also water channels in a biofilm such that water and nutrients can penetrate full biofilm depth. Furthermore, rhamnolipids kill epithelial cells and disrupt tight junctions.

[3] `http://www.pdb.org/pdb/explore/explore.do?structureId=1ezm` in 2012

Pyocyanin is a small blue–green colored substance that can easily cross cell membranes. It oxidizes or reduces other compounds. Pyocyanin affects the electron transport chain and the vesicular transport, inhibits the growth of cells and cell respiration, and causes apoptosis in neutrophils [93]. Moreover, it is able to kill competing microorganisms.

Biofilm Formation

A biofilm is an adhesion of bacteria to a surface or an interface consisting of an extracellular polymeric substance. A first, some cells attach reversibly to the surface. Then, Quorum sensing mechanisms up–regulate certain genes that are, e.g., responsible for formation of the extracellular polymeric substance. The gene expression profiles of cells in a biofilm are totally different from individual cells. Afterwards, the biofilm changes its shape and size due to cell divisions.

The extracellular polymeric substance provides a common environment to communicate via autoinducers, protects the bacterial cells from the immune system of the host, and is a barrier for antimicrobial agents [28]. Thus, antibiotics are not able to reach closed areas of the biofilm.

3.2.2. Resistance

A rapidly increasing number of infecting *P. aeruginosa* strains is resistant against current antibiotics. Often, classical antimicrobial agents are connected to the central metabolism or DNA replication and therefore affect the bacterial growth. Consequently, the evolutionary pressure is increased and resistances are developed. Due to mobile DNA elements, Gram–negative bacteria consist of several resistance genes coding for proteins that degrade antibiotics [35]. In mutants, efflux pumps, such as MexAB–OprM or MexXY–OprN, that also lead to resistance are over–expressed. Further, mutations affecting the drug targets are frequently found.

3.3. Regulatory Pathways in *Pseudomonas aeruginosa*

In *P. aeruginosa*, Quorum sensing systems are involved in the gene expression required for the formation of virulence factors, such as elastase, rhamnolipids, and pyocyanin. The virulence factors are produced at the beginning of the stationary growth phase [40]. To regulate their formation, there are three individual Quorum sensing systems, namely *las*, *rhl*, and *pqs*, that are hierarchically organized. The *las* system initiates both other Quorum sensing systems, whereas it has been described that *rhl* negatively regulates the *pqs* system and the *pqs* system induces the *rhl* system [113]. An overview over the important metabolites and their relationship is given in Figure 3.5. The references of the individual reactions that are here taken into account are listed in Table B.1 in appendix B.1.

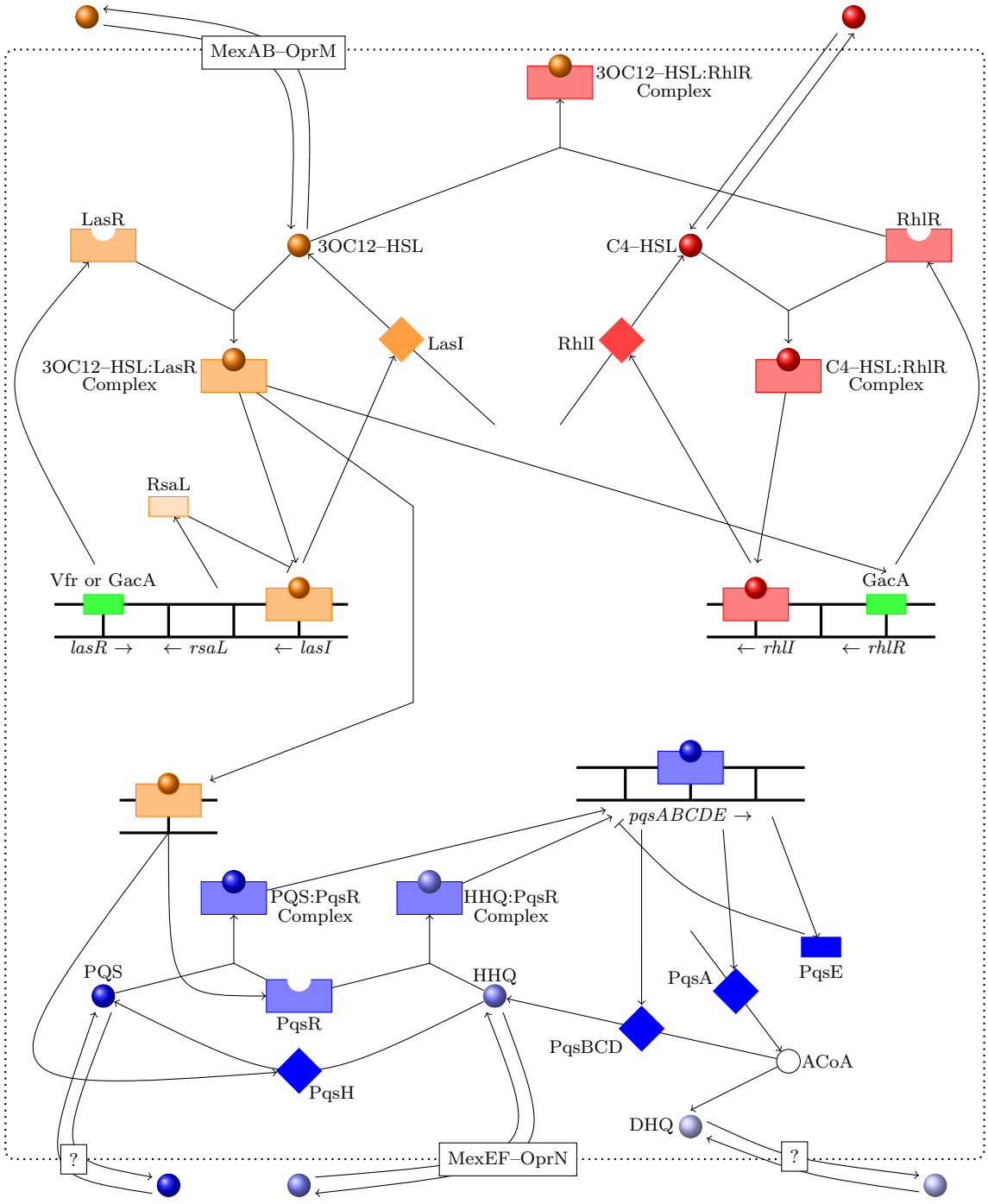


Figure 3.5.: **Quorum sensing network of *Pseudomonas aeruginosa*:** The three systems *las* (in orange), *rhl* (in red), and *pqs* (in blue) and their hierarchical organization are illustrated. Colored balls denote the corresponding autoinducers, squares enzymes, and rectangles receptors or other proteins. MexAB–OprM and MexEF–OprN are efflux pumps coded by the genes *mexA* and *mexB*, as well as *mexE* and *mexF*.

3.3.1. The *las* system

Analogous to the *lux* system in *Vibrio fischeri*, there is the *las* system (colored in orange in Figure 3.5), the first Quorum sensing system in *P. aeruginosa*. Here, the biosynthesis of the first autoinducer

N–3–oxododecanoyl–homoserine lactone (3–oxo–C_{12}–HSL, AI–1)

is catalyzed by the synthase LasI (LuxI homolog). 3–oxo–C_{12}–HSL forms a complex (3–oxo–C_{12}–HSL:LasR, C1) with the LuxR homologous receptor LasR that initiates the transcription of LasI.

Further, 3–oxo–C_{12}–HSL:LasR up–regulates the formation of RsaL, which blocks the expression of *lasI* [37, 136, 169]. Since 3–oxo–C_{12}–HSL is only able to diffuse very slowly through membranes, an over–expressed efflux pump MexAB–OprM is useful to transport 3–oxo–C_{12}–HSL across the membrane [138]. The expression of gene *lasR* is up–regulated by the global activator GacA and the activator Vfr [2, 147].

3.3.2. The *rhl* system

Similarly, the *rhl* system (colored in red in Figure 3.5) uses a positive feedback loop for a fast autoinducer formation. A second autoinducer [137]

N–butyryl–homoserine lactone (C_4–HSL, AI–2)

binds to the receptor RhlR, which is structurally similar to LasR. This complex (C_4–HSL:RhlR, C2) activates the transcription of the synthase RhlI [124].

C_4–HSL is able to diffuse rapidly through membranes. Despite the high similarity between receptors and autoinducers of the *las* and *rhl* system, they are very specific. C_4–HSL does not bind to LasR. In contrast, 3–oxo–C_{12}–HSL is able to build a complex (3–oxo–C_{12}–HSL:RhlR, C4) with RhlR, but this 3–oxo–C_{12}–HSL:RhlR does not up–regulate the transcription of the *rhl* operon [141]. RhlR is positively regulated by the activator GacA [147] and by complex 3–oxo–C_{12}–HSL:LasR of the *las* system.

3.3.3. The *pqs* system

In contrast to classical Quorum sensing systems in Gram–negative bacteria, the *pqs* system (colored in blue in Figure 3.5) of *P. aeruginosa* consists of more than 50 different small signaling molecules that are structurally different from acyl homoserine lactones [67]. Those molecules are alkyl–quinolones and some of them able to weakly bind to the receptor PqsR. Figure 3.6 shows the chemical structures of two such alkyl–quinolones that activate as agonists the response of

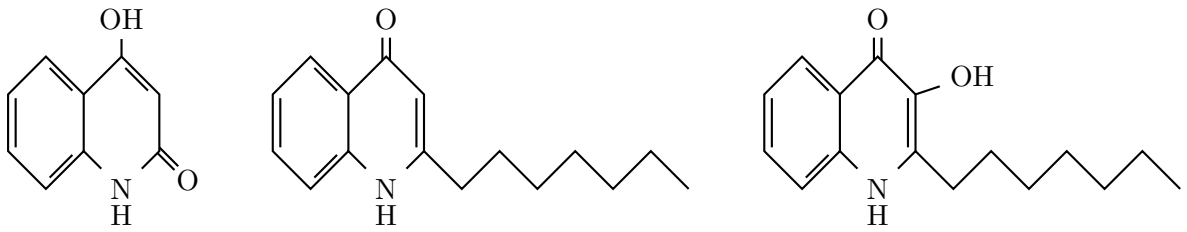


Figure 3.6.: **Chemical structures** of DHQ (on the left), HHQ (in the middle), and PQS (on the right).

PqsR as well as the chemical structure of the molecule 2,4–dihydroxy–quinoline (DHQ) that is structurally related to alkyl–quinolones. The specific function of DHQ and its relation to the production of virulence factors in *P. aeruginosa* or the formation of biofilms is so far not understood.

The main autoinducer of the *pqs* system

2–heptyl–3–hydroxy–4–quinolone (*Pseudomonas* quinolone signal, PQS)

is the heterobicyclic aromatic compound [140]. The enzyme PqsH, which is activated by LasR, is required to convert the precursor

4–hydroxy–2–heptyl–quinoline (HHQ)

into PQS. The exact transport mechanisms of HHQ and PQS are so far not completely understood. The efflux pump MexEF–OprN is able to transport HHQ across membranes in the case of an over–expression [92].

Both, HHQ and PQS, bind strongly to PqsR. The complexes PQS:PqsR (C3) and HHQ:PqsR (C5) control the expression of several genes, such as the biosynthesis operon *pqsA–E* and the *phnAB* operon [39, 205]. The proteins PhnA and PhnB are responsible to form anthranilic acid. The benzoate coenyzme A ligase PqsA catalyzes the formation of anthraniloyl–coenzyme A (ACoA) using anthranilic acid and coenzyme A. Afterwards, ACoA together with β–ketodecanoic acid or its bioactivated thioesters is converted into HHQ or DHQ by PqsD [142]. ACoA is also used to form 2–heptyl–4–hydroxyquinoline *N*-oxide (HQNO) catalyzed by PqsL. The enzymes PqsB and PqsC are related to a fatty acid pathway and their specific role remains still unknown, but they are involved in the HHQ biosynthesis [67, 145]. The formation of HHQ and the conversion of HHQ to PQS happen in two different cells [39].

The so–called PQS signal response protein (PqsE) negatively regulates the *pqsA–E* operon [145]. PqsE is additionally linked to the *las* and *rhl* systems [48, 145].

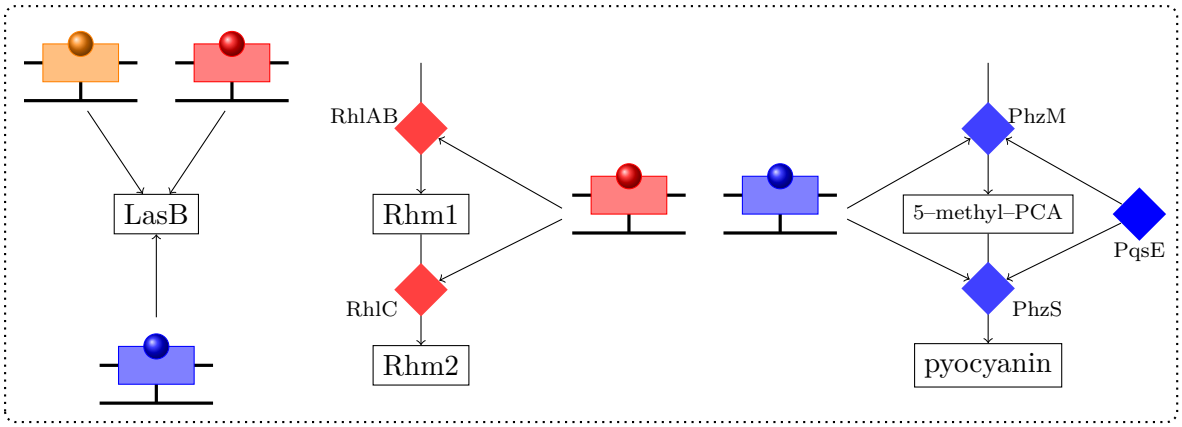


Figure 3.7.: **Pathways of virulence factors:** The regulation of LasB, rhamnolipids (Rhm1 and Rhm2), and pyocyanin is illustrated. Rectangles with balls denote the complexes between receptor and autoinducer, where the *las* system is colored in orange, the *rhl* system in red, and the *pqs* system in blue. Colored squares represent enzymes of the corresponding system.

3.3.4. Formation of Virulence Factors

Figure 3.7 illustrates the biosynthesis of the virulence factors LasB, rhamnolipids, and pyocyanin. The used references of virulence factor formation are given in Table B.1 in appendix B.1. The gene *lasB* encodes the elastase. The complexes 3–oxo–C_{12}–HSL:LasR, C_4–HSL:RhlR, and PQS:PqsR all trigger the production of elastase [55, 204].

The transcription of the operon *rhlAB* and the gene *rhlC* is up–regulated by the complex C_4–HSL:RhlR of the *rhl* system. The gene *rhlB* encodes the rhamnosyltransferase required to form monorhamnolipids (Rhm1). RhlA is also involved in synthesis of monorhamnolipids. RhlC catalyzes the conversion of monorhamnolipids into dirhamnolipids (Rhm2) [106, 114, 123].

Complex PQS:PqsR and the protein PqsE activate the gene expression of two homologous *phzA–G* operons and the genes *phzH*, *phzM*, and *phzS*. 5–methylphenazine–1–carboxylic acid betaine (5–methyl–PCA) is formed by the *S*–adenosyl methionine–dependent methyltransferase PhzM using phenazine–1–carboxylic acid that is controlled by the *phzA–G* operon. Then, the NADH–dependent flavoprotein monooxygenase PhzS is responsible for converting 5–methyl–PCA into pyocyanin [112].

3.4. Quorum Sensing Inhibitors

As already mentioned, *P. aeruginosa* is multi–resistant against most antibiotics available today except for azithromycin, ceftazidime, and ciprofloxacin. Those antibiotics decrease not only the growth of cells but influence also the permeability of the membranes [176]. Therefore,

the diffusion of 3–oxo–C_{12}–HSL is changed. Nevertheless, the evolutionary pressure leads to development of resistance. To reduce the probability of resistance development, Quorum sensing inhibitors are used alternatively to antibiotics. In the case of Quorum sensing inhibitors, the Quorum sensing systems are directly targeted, i.e., the biosynthesis, the autoinducer itself, or the activating complex is blocked. Quorum sensing inhibitors are either receptor antagonists or enzyme inhibitors of a protein in the Quorum sensing system.

Receptor antagonist Receptors, such as LasR, RhlR, and PqsR, are proteins that bind highly selective to their specific ligands with high affinity that often causes a conformational change. A ligand that induces some response with high efficiency is called agonist, whereas a ligand with no efficiency that blocks the active side of a receptor without activation is called competitive antagonist. The complex between receptor and antagonist may be reversible or irreversible.

Enzyme inhibitor A metabolite that binds an enzyme either reversibly or irreversibly and therefore blocks the substrate–binding position of the enzyme is called enzyme inhibitor.

Dose–response curve A dose–response curve describes the connection between the amount of a antagonist and the response of a receptor and its agonist. For example, the half maximal inhibitory concentration (IC_{50}) defines the required concentration of an antagonist or inhibitor to block half of the response.

Alternatively, it is possible to chemically degrade the autoinducer molecules. Depending on the temperature and the length of the acyl side chain, pH levels above seven can cause a ring opening in acyl homoserine lactones and therefore a decreased specificity to the receptor [146]. Further, there are enzymes that degrade acyl homoserine lactone molecules, e.g., PvdQ in *P. aeruginosa*. In human epithelial cells, it is possible to inactivate 3–oxo–C_{12}–HSL of the *las* system, but not C_4–HSL of the *rhl* system.

Several small molecules structurally similar to acyl homoserine lactones have been considered as receptor ligands. It seemed that the ring structure decides whether a ligand acts as agonist or antagonist [146]. Based on this, the first LasR antagonists have been described in 2003 [177, 181]. Further, Quorum sensing and biofilm formation inhibiting molecules have been presented [58]. A structure–based virtual screening approach ranked candidates for LasR antagonists found in the SuperDrug and SuperNatural databases [207]. In 1999, an analog of an acyl homoserine lactone precursor was shown to block RhlI in vitro [135].

Due to the central role of the *pqs* system in the pathogenicity of *P. aeruginosa*, it makes sense to block directly the *pqs* system. Quinazolinones and derivates of the natural effector HHQ are potent antagonists for PqsR [35, 105]. The production of pyocyanin is strongly decreased by PqsR antagonists in the low micromolar range [85]. In 2012, PqsD inhibitors have been presented by our experimental collaborators that detectably decrease the concentration of HHQ and PQS [180].

4. Computational Analysis of Biological Systems

Studying functional interactions between different species of a gene regulatory network or of a protein pathway, analyzing dependencies on environmental conditions, as well as interpreting the relationship between different networks is a challenge, e.g., to understand and treat diseases. Looking at the large complexity of biological systems in which a huge number of genes and proteins, several cells, and whole tissues may be involved containing very fast interactions, such as enzymatic reactions, and very slow processes, such as the whole life of an organism, it nowadays appears almost impossible to investigate experimental data without computational tools. These computational methods play an important role in the continuously ongoing systems biology cycle including data generation, mathematical modeling, simulations with predictions, validation, and following experiments.

4.1. Modeling and Simulation in Systems Biology

Single biological pathways can be modeled by several theoretical approaches. For this, the interactions between the considered species are organized in a directed acyclic graph called network.

Boolean network Dynamic systems with discrete states and time steps are called Boolean networks. The interactions are described by logical rules. In classical Boolean networks, the species are Boolean variables. Unfortunately, it is not possible to achieve quantitative results or to apply Boolean networks to inconsistent and incomplete networks [178]. Further, the runtime increases exponentially with an increasing number of nodes.

Kauffman network In 1969, Kauffman introduced random Boolean networks in which the genes are randomly assigned [9, 82].

Probabilistic Boolean network Probabilistic Boolean networks combine the rule–based dependencies of the Boolean variables with stochastic processes including random variables [91, 173].

Bayesian network In dynamic Bayesian networks, the nodes represent random variables with conditional probabilities and the edges conditional dependencies [91]. The model has been used, e.g., to study gene interactions in the cell–cycle of *Saccharomyces cerevisiae* considering mircoarray data [52].

Petri net A Petri net consists of places that denote resources, transitions that denote actions, arcs, and tokens [178]. For example, the glycolysis and citric acid cycle has been modeled as a Petri net [57].

Continuous model In contrast to the other listed models, ordinary differential equations or partial differential equations allow for a quantitative simulation. Therefore, precise rate constants of all considered biochemical reactions are required, but those remain often unknown. Continuous models have been frequently used, e.g., the leaf carbohydrate metabolism in *Arabidopsis thaliana* metabolism has been modeled using ordinary differential equations [68].

Combining several single pathways or incorporating them into cell cultures lead to extended multi–scale models. For example, the multi–scale approach by Jiang *et al.* divided tumor growth into three levels where the cell growth is formulated as a Monte Carlo model, the cell–cycle as a Boolean network, and external diffusion processes as differential equations [77].

4.1.1. Discrete Rule–based Models

Several approaches to simulate regulatory networks are based on Boolean networks. For example, the FA/BRCA pathway has been considered [150] or the cell–cycle of *Saccharomyces cerevisiae* [97]. The flower morphogenesis of *Arabidopsis thaliana* has been studied by a weighted interactions in a Boolean model with a semi–synchronous updating scheme [116]. In the software gene interaction network simulation (GINsim), different logical algorithms based on either synchronous or asynchronous updating schemes are implemented to analyze regulatory networks [59]. To avoid runtime problems, a state transition graph is generated to focus on important parts of the network.

Besides probabilistic Boolean networks, several extensions of the logical formalisms have been developed to adopt the concept to the biological question. Handorf and Klipp considered a signaling network containing the interactions between MAP kinase cascade and Wnt pathway. For this, they introduced a second Boolean variable for each node to describe depletions [63]. In 2001, a temporal Boolean network was presented in which the state at time step $t+1$ of a node depends not only on the system state of time t but also on early system states [174]. A multi–level discrete approach with the same number of possible levels of a certain species was applied to analyze the cell–cycle of budding yeast [49]. Another multi–level logical model considered for each node a different number of possible levels based on the biological relation. This formalism was used to simulate the Gap–gene system of *Drosophila* [157].

4.2. Quorum Sensing

Different methodologies have been used before this work to simulate the dynamics of Quorum sensing pathways. For this, often a theoretical simple positive feedback loop is considered. For

example, the *las* system of *Pseudomonas aeruginosa* was analyzed by a continuous model to understand the influence of autoinducer degradation on burn infections [86]. The missing parameters were estimated. Another mathematical model solved ordinary and partial differential equations of a theoretical Quorum sensing system or of the actual *las* system of *Pseudomonas aeruginosa* [41, 201]. A simplified *lux* system with autoinducer diffusion described by ordinary differential equations was included into a cell growth model in which bacteria were represented by two half–spheres [115]. The *lux* system of *Vibrio fischeri* was amongst others considered by so–called P systems [13]. A combined regulatory network regarding both Quorum sensing system *lux* and *ain* of *Vibrio fischeri* was modeled using ordinary differential equations [90].

Due to the responsibility of Quorum sensing for virulence factor formation, the *las* system of *Pseudomonas aeruginosa* in a growing batch culture was modeled together with a LasR or autoinducer degrading substance using differential equations [6].

4.3. Biofilms

To understand the impact of extracellular production of polymeric substances, which is regulated by Quorum sensing systems, on growing biofilms, a differential mass balance of extracellular polymeric substance, autoinducers, and nutrients with a discrimination between down– and up–regulated cells was published [51]. Moreover, the formation of an one–dimensional biofilm in *Pseudomonas aeruginosa* was considered by Chopp and co–workers using the *las* system as autoinducer production [22].

Similarly, Anguige *et al.* modeled biofilms and the corresponding extracellular polymeric substance production that is regulated by *las* system of *Pseudomonas aeruginosa* in dependency of Quorum sensing inhibitors [5]. Biofilms were also studied as 3D system with a bulk liquid compartment and a biofilm compartment that contains cells [50]. Each cell was able to consume nutrients, produce extracellular polymeric substance and autoinducers, and could divide. The regulation of autoinducers was realized as a self–activation. In this study, Forzard *et al.* analyzed the effect of Quorum sensing inhibitors added to the system at varying time points and different concentrations.

Part II.

Functional Prediction and Classification

So far, for many membrane proteins that are involved in transport processes, their transported substrates have still not been characterized by time–consuming experiments. For this reason, we are searching for highly efficient computational methods to predict the function of membrane proteins. First, we will introduce the developed and used methods to analyze and classify some data sets with membrane proteins. Afterwards, the reliability and successfulness of the considered features in the context of functional classification are discussed. A comparison between different amino acid compositions containing higher sequence order information, amino acid properties, or multiple sequence alignments based on *Arabidopsis thaliana* transporters was published in *J. Chem. Inf. Model.* in 2010 [158]. In 2012, we published a study using transmembrane or non–transmembrane regions separately in the amino acid composition [159].

5. Materials and Methods

To download all required data, to calculate the used features, and to statistically analyze the considered membrane proteins, I wrote a Java program. The developed ranking method was also implemented in Java.

5.1. Data Resources

5.1.1. Orientations of Proteins in Membranes Database

The database Orientations of Proteins in Membranes (OPM)[1] [104] contains peripheral, monotopic, and transmembrane proteins with known 3D structures listed in the protein data bank (PDB)[2] [12]. The transmembrane proteins are clustered in three classes, namely α–helical polytopic (1.1), α–helical bitopic (1.2), and β–barrel (1.3) transmembrane proteins. The OPM database further provides the position of TMS in the lipid bilayer with a theoretical computation which is evaluated by experimental data.

5.1.2. Transport Classification Database

The Transporter Classification Database (TCDB)[3] introduced by the Saier Lab Bioinformatics Group provides an accepted annotation of about 5,600 membrane transporters from various organisms that are classified into around 600 transporter families on the basis of the so–called TC–system [156]: are explained $N_1.A.N_2.N_3.N_4$ whereby $N_i \in \mathbb{N}$ is a number and A is a character. N_1 represents the transport class, e.g., channels or primary active transporters. A refers to the transporter subclass, for example, 3.D oxidoreduction–driven transporters where the character stands for the energy source. N_2 corresponds to the (super)family, such as the ABC superfamily (3.A.1), N_3 to the transporter subfamily and N_4 to a transporter and its particular substrates.

[1] `http://opm.phar.umich.edu/`
[2] `http://www.pdb.org/`
[3] `http://www.tcdb.org/`

5.1.3. Aramemnon

The plant membrane protein database Aramemnon[4] from the Flügge Lab [167] contains amino acid sequences, substrate annotations, TC information, topology predictions, and homologous relations about membrane proteins with at least one TMS from, e.g., *Arabidopsis thaliana* (*A. thaliana*) or *Oryza sativa*. The functional descriptions are manually created and directly linked to the original bibliographical data such that the substrate annotations are quite reliable.

5.2. Data Sets

In this thesis, three independent data sets were analyzed and used in the context of classification, namely a OPM data set containing membrane proteins with known structure, a TCDB data set, and an *A. thaliana* data set retrieved from Aramemnon. For the TCDB and *A. thaliana* data sets, subsets representing either a certain family or substrate group, so–called positive sets, were considered. For those positive sets, complementary negative sets with the same size as the corresponding positive set were additionally generated for validation purpose. Since there does not exist evidence (at least not on a large scale) that certain transporters do not transport certain substrates, those negative sets were randomly created using members of non–related positive sets that are expected to not transport the respective substrate.

5.2.1. OPM Data Set

The OPM data set downloaded in 2011 from the OPM database [104], consists of 102 α–helical (H.OPM) and 35 β–barrel (B.OPM) transmembrane proteins with 3D structures determined at high resolution. To reach a non–redundant set, BLAST [3] was used to exclude homologous sequences with a sequence identity higher than 20% and an E–value threshold of less than 10^{-50}. In order to test the reliability of secondary structure predicting tools, TMSs were predicted by Memsat–SVM [122], TOPCONS [15], and TMBETAPRED [128].

Considering the α–helical polytopic membrane proteins, the largest protein families in this data set are 1.1.01 (Rhodopsin–like proteins), 1.1.02 (photosynthetic reaction centers and photosytems containing the TC subfamily 3.E.2), 1.1.04 (transmembrane cytochrome b like), 1.1.05 (cytochrome c oxidases belonging to 3.D.4), and 1.1.27 (major intrinsic protein superfamily). Among these, 1.1.01 and 1.1.02 were considered as positive sets in the approach mapping non–transmembrane regions, see Section 5.9.

[4] `http://aramemnon.botanik.uni-koeln.de/`

5.2.2. TCDB Data Set

In May 2011, membrane proteins and their families were downloaded from TCDB [156]. We did not consider *uncharacterized*, *putative*, or *probable* labeled sequences or proteins grouped to more than one family. Only proteins with at least 100 amino acids were used. All homologous sequences were again deleted applying a threshold of 20% for sequence identity and of 10^{-50} for E–value. The TMHs were annotated by Memsat–SVM and TMBETAPRED. Then, 73 further sequences were removed for which the predicted TMHs were located by Memsat–SVM after the end of the protein as well as all proteins without at least one assigned TMH or two TMBs. This gave a final data set of 2126 proteins.

The subfamilies 1.A.1 containing voltage–gated ion channels, 2.A.1 representing the MFS, and 3.A.1 including ABC transporters were used. For those, we discriminated between sequences showing either an α–helical or a β–barrel character. The corresponding α–helical positive sets were labeled as H.TCDB, H.1.A.1, H.2.A.1, H.3.A.1 and the β–barrel positive sets as B.TCDB, B.1.A.1, B.2.A.1, B.3.A.1, respectively. Figure 5.1 shows an overview over all positive sets with their size. For each positive set, five different negative sets were generated out of sequences

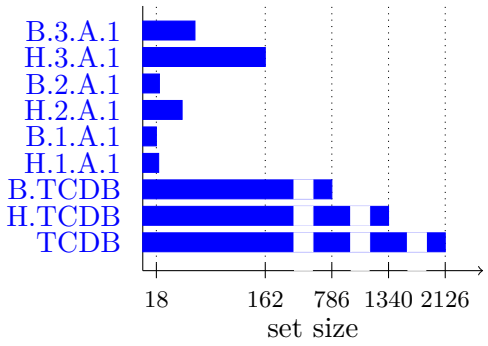


Figure 5.1.: **TCDB data set** with the size of all subsets.

from positive sets containing proteins of not considered TC families.

5.2.3. *Arabidopsis thaliana* Data Set

A. thaliana transporters and carriers with their substrate annotations were downloaded from the Aramemnon database [167] in 2009. Due to the relatively small size of subsets sharing the same substrate, a less restrictive sequence identity of 90% was used to ignore redundant proteins. TMHs were predicted by Split 4.0 [80] or MEMSAT–SVM [122] and sequences without TMHs were again removed. This yielded a final data set of 793 membrane proteins.

Four individual substrate groups, namely amino acid, oligopeptide, phosphate, and hexose were considered as positive sets without including transporters labeled as *putative*. The members of

those sets belong to several TC families. Most proteins of the amino acid transporter set are amino acid/auxin permeases (2.A.18). The oligopeptide transporter set consists of oligopeptide transporters (2.A.67) and proton–dependent oligopeptide transporters (2.A.17). The phosphate transporter set is a mixture of multiple TC families, namely 2.A.1, 2.A.20, 2.A.29. The hexose transporters include the families 2.A.1 and 2.A.7. Figure 5.2 gives an overview of the four positive sets with their size. For each positive set, a certain number of corresponding negative

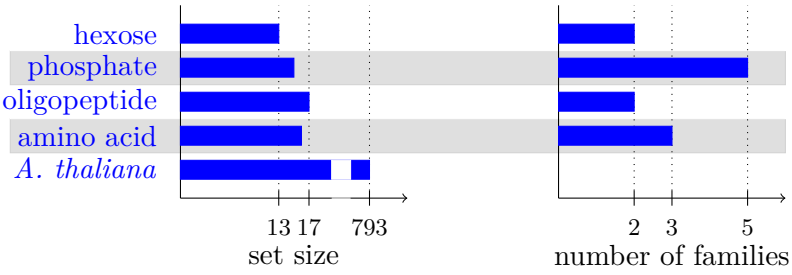


Figure 5.2.: ***Arabidopsis thaliana* data set** with the size of all subsets.

sets were randomly constructed assuming, e.g., that sugar molecules are not transported by amino acid or oligopeptide transporters.

5.3. Features

5.3.1. Amino Acid and Extended Compositions

Several different variations of the amino acid composition (AAC) were used in the later analysis and classification. For all considered AACs, the entries were normalized by the total number of included residues. In the following, the original AAC containing the frequencies of the 20 amino acids in the full protein sequence is called oAAC. The extended PseAAC developed by Chou with λ further entries (see Section 2.2.2) includes neighborhood correlations based on amino acid properties [23]. The PAAC introduced by Park *et al.* incorporates the 400 frequencies of all possible amino acid pairs [131].

Furthermore, a combination between both extensions PseAAC and PAAC, termed PsePAAC, containing $400 + \lambda$ entries was considered. The entries representing the pair frequencies and neighborhood correlations are given in equation (5.1),

$$v_i = \begin{cases} \frac{f_i}{\mathfrak{n}} & \text{if } 1 \leq i \leq 400 \\ \frac{\omega \tau_{i-400}}{\mathfrak{n}} & \text{if } 400 \leq i \leq \lambda \end{cases} \tag{5.1}$$

where ω is a weighting factor and $\mathfrak{n}$ is a sequence length based normalization. Hydrophobicity, hydrophilicity, side–chain mass, pK values, and isoelectric point of the residues as described by Shen *et al.* [172] were included in the correlation factors (τ). Table A.1 in appendix in Section A.1 lists the used characteristic values.

5.3.2. Profile–based Composition

Additionally, a profile–based version termed MSA–AAC was constructed including the frequencies of all amino acids in a ClustalW [186] MSA of the considered sequence. An example MSA–AAC is given in Figure 5.3. The MSA was generated applying the procedure presented in

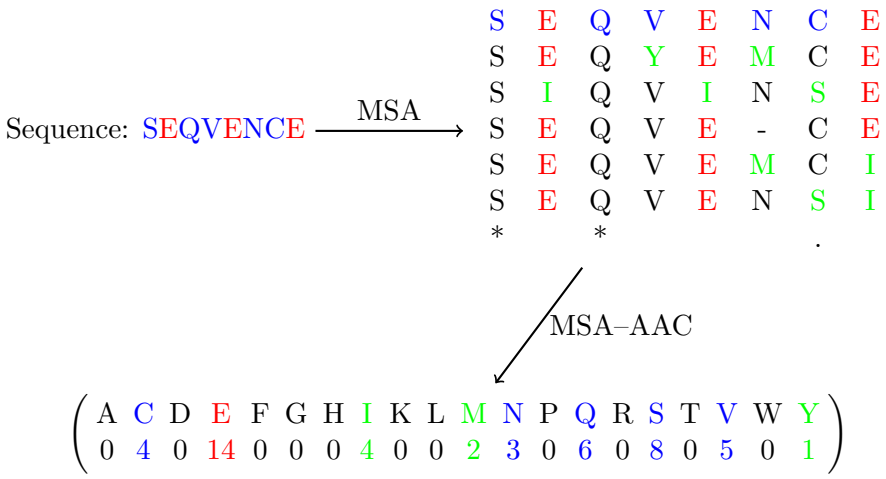


Figure 5.3.: **Profile–based amino acid composition (MSA–AAC)** for an example sequence and a multiple sequence alignment containing similar sequences.

[132]; up to 1000 homologous sequences (the top hits) of the non–redundant database searched with BLAST were used to set up an initial MSA. Then, proteins with an identity less than 25% were excluded and the MSA was realigned.

5.3.3. Filtered Compositions

For these different AACs, we also constructed various filtering options covering only the amino acid frequencies of selected sequence regions. A first variant only considered positions in transmembrane segments (either purely TMHs or TMBs) instead of the full sequence. Figure 5.4 illustrates the calculation of this AAC over TMSs. In contrast, another AAC contains only

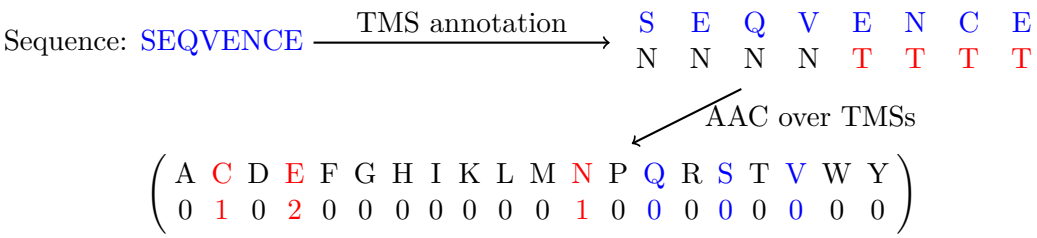


Figure 5.4.: **Amino acid composition over transmembrane segments** for an example sequence and its topology annotation.

residues of non–TMSs. A joined version of both defined as a 40 entries long vector, denoted as TMS–nonTMS, retains the filtering advantages and all necessary information of the full sequence. Moreover, this combination was further concatenated with the AAC over the full sequence to a vector with 60 components, termed TMS–nonTMS–AAC. In addition, buried or exposed positions assigned by the transmembrane exposure prediction [132] were solely considered as well as regions marked as conserved in the MSA described above. Since MSAs distinguish between total, strong, weak, and no conservations at each position, three different filters were constructed, namely conserved, strong, and total. Here, conserved represents all positions indicated as conserved independent of the conservation level (total + strong + weak). The vector strong includes all as strong or total conserved assigned positions (total + strong) and the vector total only contains total conserved positions (total).

5.4. Similarity Measurement

The similarity $d(\mathfrak{p})$ within a positive set[5] $\mathfrak{p}$ with q members was calculated as given in equation (5.2),

$$d(\mathfrak{p}) = \frac{2}{q(q-1)} \sum_{i,j} \sum_{k=1}^{20} (v_{i_k} - v_{j_k})^2 \tag{5.2}$$

whereby v_i and v_j indicate the AACs of two transporters i and j of the positive set $\mathfrak{p}$. This squared Euclidean distance between every transporter pair in the positive set is normalized by the number of pairs. A small value of this similarity $d(\mathfrak{p})$ denotes a homogeneous positive set.

In an analogous fashion, we also estimated the pairwise similarity or dissimilarity $d(\mathfrak{p}_1, \mathfrak{p}_2)$ between two different positive sets $\mathfrak{p}_1$ and $\mathfrak{p}_2$ by the squared Euclidean distance as shown in equation (5.3),

$$d(\mathfrak{p}_1, \mathfrak{p}_2) = \sum_{k=1}^{20} (\mu_{\mathfrak{p}_{1_k}} - \mu_{\mathfrak{p}_{2_k}})^2 \tag{5.3}$$

whereby each positive set is represented by its mean AAC (μ). Then, a distance less than a certain threshold characterizes a similarity between the considered positive sets.

Additionally, the residues mainly responsible for dissimilarities between two sets $\mathfrak{p}_1$ and $\mathfrak{p}_2$ were computed. For this, the standard deviation (σ) of the AAC in each positive set was measured. In the case of a Gaussian distribution, 95.4% of all elements are in the range $[\mu - 2\sigma, \mu + 2\sigma]$. Therefore, the following condition was checked for every amino acid type k:

$$\mu_{\mathfrak{p}_{1_k}} + 2\sigma_{\mathfrak{p}_{1_k}} \leq \mu_{\mathfrak{p}_{2_k}} - 2\sigma_{\mathfrak{p}_{2_k}} \ \vee \ \mu_{\mathfrak{p}_{2_k}} + 2\sigma_{\mathfrak{p}_{2_k}} \leq \mu_{\mathfrak{p}_{1_k}} - 2\sigma_{\mathfrak{p}_{1_k}}, \tag{5.4}$$

under the assumption that the positive sets show a similar behavior as a Gaussian distribution. Thus, amino acid k causes a dissimilarity if this condition is true.

[5] positive set: a subset of the full data set in which all proteins share a certain common ground

5.5. Significance of Dissimilarities

In order to identify the significant differences between a pair of positive sets based on ten non–disjoint categories of physicochemical amino acid characteristics, p–values were computed by the statistical hypothesis tests analysis of variance [73, 162] and Wilcoxon–Mann–Whitney [107, 203].

5.5.1. Analysis of Variance

Since an analysis of variance (ANOVA) [73, 162] requires normal distributed observations, all frequencies of each category in every positive sets were tested whether they could be assumed as Gaussian distributions. For this, the Shapiro–Wilk–Test [171] was used with a significance level $\alpha = 0.05$. Under the null hypothesis H_0 that the mean values are the same, ANOVA was performed to compute the F–distribution and the corresponding p–values based on the variances within and between positive sets. For p–values less than 0.001, the null hypothesis was rejected assuming that the given differences between both considered distributions are not only caused by random fluctuations.

5.5.2. Wilcoxon–Mann–Whitney Test

The Wilcoxon–Mann–Whitney test [107, 203], also known as Wilcoxon rank–sum test or Mann–Whitney U test, is a non–parametric significance test that compares medians and is therefore highly robust against outliers. In comparison to ANOVA, the main advantage of the Wilcoxon–Mann–Whitney test is that a normal distribution is not necessary. Again, we tested the null hypothesis H_0 that the mean values are the same. Here, the observations in two positive sets of size q_1 and q_2 are sorted and represented by their rank ρ. Then, the value U is calculated as given in equation (5.5),

$$U = \min\left(q_1 q_2 + \frac{q_1(q_1+1)}{2} - \sum_{i=1}^{q_1} \rho_{1_i} \ , \ q_1 q_2 + \frac{q_2(q_2+1)}{2} - \sum_{i=1}^{q_2} \rho_{2_i} \right), \tag{5.5}$$

which is used to compute the p–values. The null hypothesis was again rejected if the p–value was less than 0.001.

5.6. Ranking Procedure for Classification

Figure 5.5 demonstrates a general overview of the work flow. At first, a certain AAC was calculated for each protein in the full data set. Then, the Euclidean distance d between all individual AACs and the considered positive set was measured and used to construct several individual rankings r, where all proteins were sorted according to the similarity to the positive set starting with the sequence having the highest similarity. For this, there are two different

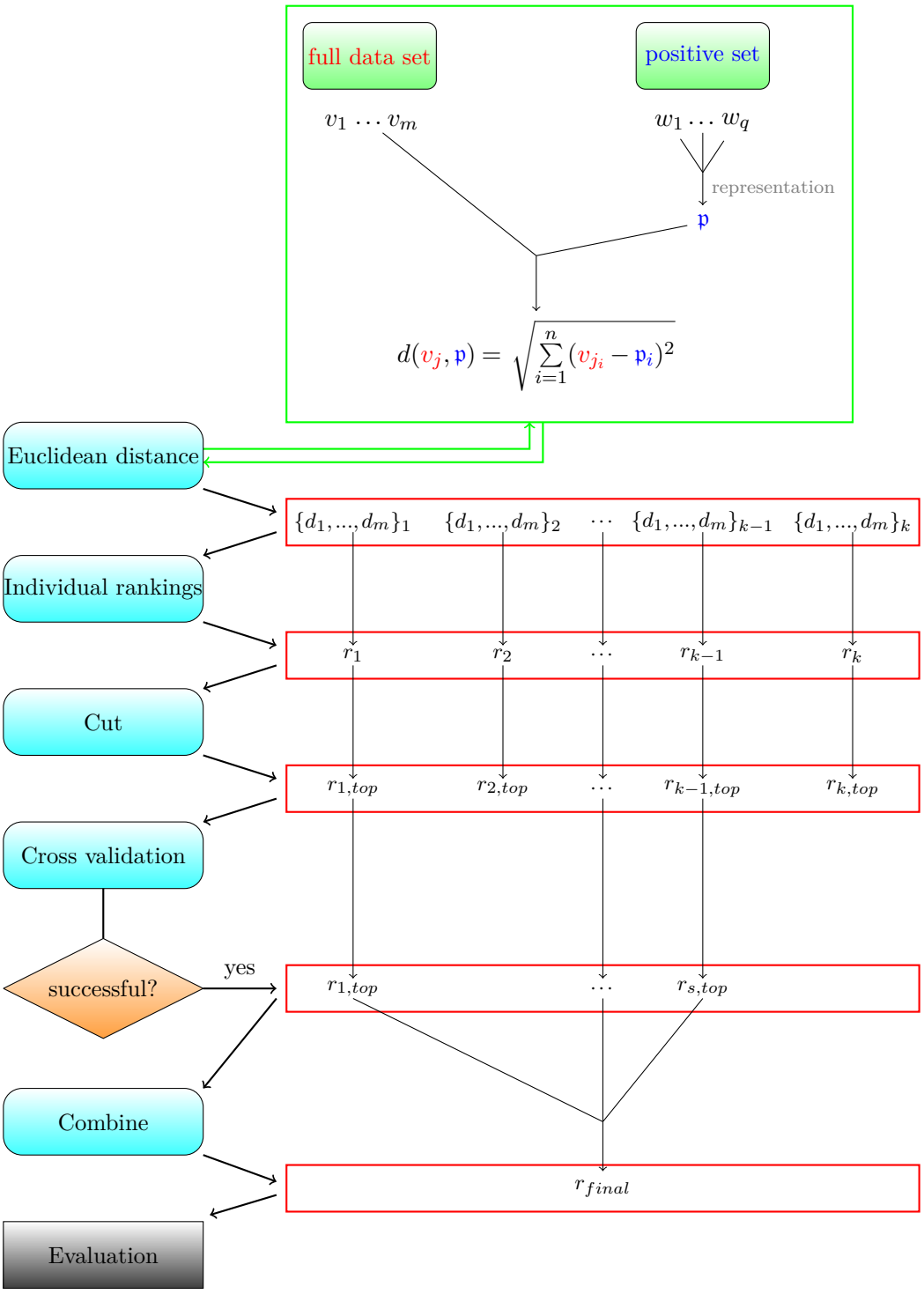


Figure 5.5.: **Work flow of ranking procedure:** Shown are the main calculation steps, namely the AAC v of each protein, the distance measure $d = (v, \mathfrak{p})$ between v and the considered positive set $\mathfrak{p}$, the evaluation of individual rankings r_i, and the final ranking r_{final}. Here, s gives the number of rankings that are classified as successful. In the case of the search profile based method, $\mathfrak{p}$ denotes the average AAC μ of the positive set and k the number of individual rankings as in equation (5.7). In the case of the individual composition based method, $\mathfrak{p}$ indicates the AAC of any single sequence in the positive set and $k = q$.

representation ways of the positive set, namely based on an average search profile and based on individual compositions as explained in the following.

The individual rankings were cut after the top ten positions. For positive sets with less than 20 members, a leave one out cross validation (LOOCV) quantified the quality of every ranking. Here, the procedure was applied for a positive set of size $q-1$ in which a single protein was removed. Then, it was tested whether this protein was contained among the considered top ten entries. This was iterated for all proteins in the positive set. When at least one member of the positive set was found in the LOOCV, a ranking was defined as successful. For large positive sets, a five–fold cross validation quantified the ranking quality. For this, five equally sized folds were generated randomly out of the considered positive set. Then, the number of elements in the omitted fold found in the ranking built by the other four folds was measured. This was repeated for all possible omitted folds. For statistical reasons, the average value over 20 procedures with different folds were used to decide whether a ranking is successful.

Finally, all successfully marked individual rankings containing ten proteins were merged to a single final ranking using the cross entropy Monte Carlo method by Lin and Ding [102]. This method integrates sorted lists by an iterative procedure so that the sum of distances between the positions of the entries in the final and the original lists gets minimal. Section 5.6.3 describes the evaluation of the final ranking.

5.6.1. Ranking based on an Average Search Profile

In this case, the average AAC of the positive set was calculated as search profile. Then, the Euclidean distance $d(v_j, \mu)$ between a sequence and the positive set was determined as:

$$d(v_j, \mu) = \sqrt{\sum_{i=1}^{n} (v_{j_i} - \mu_i)^2}, \tag{5.6}$$

whereby v_j represents the AAC of any protein in the full data set, μ the average AAC of the positive set, and n the number of entries in the corresponding AAC.

Since outliers in the positive set could effect systematic shifting, different search profiles were generated. For this, all possible subsets of the positive set with $(q-1)$ or $(q-2)$ members each were used. Then, for each resulting search profile an individual ranking was constructed based on the distance $d(v_j, \mu)$, whereby equation (5.7)

$$|\text{search profiles}| = 1 + \sum_{m=1}^{q} m \tag{5.7}$$

gives the number of considered search profiles and corresponding rankings.

5.6.2. Ranking based on Individual Compositions

For each element in the positive set, an individual ranking was set up by sorting the sequences of the full data set by comparing their Euclidean distance d to this positive set member. For this, the Euclidean distance $d(v_j, y_l)$ was computed between each sequence of the full data set and every single sequence of the positive data set, as defined in equation (5.8),

$$d(v_j, y_l) = \sqrt{\sum_{i=1}^{n} (v_{j_i} - y_{l_i})^2}, \tag{5.8}$$

whereby v and y indicate the AAC of the two proteins and n again denotes the number of entries in the corresponding AAC.

5.6.3. Evaluation of Ranking Procedure

For validation, we determined the number of proteins that were correctly and wrongly classified by the ranking method among the $2q$ top positions of the final ranking. The number of detected elements in the positive set with size q represents the true positives (TP). The number of detected elements in the negative set are the false positives (FP). Based on the number of the positives, and the size of the sets, the number of false negatives (FN) and true negatives (TN) were counted to estimate the commonly used quality measurements sensitivity, specificity, and accuracy:

$$sensitivity = \frac{TP}{TP+FN} \tag{5.9}$$

$$specificity = \frac{TN}{TN+FP} \tag{5.10}$$

$$accuracy = \frac{TP+TN}{TP+TN+FN+FP}. \tag{5.11}$$

The sensitivity determines the likelihood of finding positive set elements, whereas the specificity is influenced by the negative set. However, for positive and negative sets that are clearly smaller than the full data set, the size may affect those performance measures. Further, the selection of the negative sets is not unique. Therefore, random sets with the same or a similar size as the respective positive set were used to evaluate the behavior of this actual positive set. Such a random set consists of randomly chosen membrane proteins from the full data set irrespective of their family or substrate annotation. The quality measurements of the ranking method for several different random sets with same size were averaged (μ) and compared with the evaluation measure of the corresponding positive set.

Moreover, the standard deviation (σ) of the random sensitivities and the integer value ξ with the characteristics given in equation (5.12) were calculated,

$$\mu_r + \xi\sigma_r < sensitivity_p < \mu_r + (\xi+1)\sigma_r \tag{5.12}$$

whereby $\mathfrak{r}$ denotes the random sets and $\mathfrak{p}$ the actual positive set.

The significance test named one–sample t–test was used to check the null hypothesis H_0 that the averaged random sensitivity (μ) is equal to the actual sensitivity [87]. For this, two significance levels $\alpha_1 = 0.001$ and $\alpha_2 = 0.1$ were considered. This test is based on the mean of the sample:

$$t = \frac{(\mu - \mu_0)\sqrt{n}}{\sigma}, \tag{5.13}$$

whereby μ_0 here denotes the sensitivity of the actual positive set and n the number of random sets. Then, there are $n-1$ degrees of freedom. The critical t–values are taken from a t–distribution table according to the significance levels and the degrees of freedom. Thus, the null hypothesis was rejected if the computed t value was higher than $t_1 = 3.883$ or $t_2 = 1.328$, respectively.

5.7. Support Vector Machine

Alternatively to the ranking procedure, a support vector machine (SVM) was applied for functional classification. In general, SVMs are used to find a separation boundary for overlapping classes. For this, a linear boundary is calculated between the data points which are represented by a vector in a higher version of the feature space. In that way, a distance between a separating hyperplane and the closest vector to this hyperplane is maximized. Afterwards, the hyperplane is transformed back to the original space [16].

The distance between the features can be measured with linear, polynomial, sigmoid, or radial kernel functions. We usually used a linear kernel.

For each positive set, five different training and test sets were randomly constructed. A test set contains 25% of the proteins in the positive set (class: 1) and in the corresponding negative set (class -1) as well as the same number of other proteins in the full data set (class: 0). The remaining proteins were included in the training set.

The features used here are described in Section 5.3. The formation, training, and testing of the SVM was performed by R using the library e1071.

To evaluate the predictions again the commonly used quality measurements, given in equations (5.9), (5.10), and (5.11), were considered. Positive set members assigned to class 1 were counted as true positives, those assigned to class -1 as false negatives, and members of the negative set either as false positives or true negatives, respectively.

5.8. Principal Component Analysis and Hierarchical Clustering

Principal component analysis (PCA) [71, 139] is a statistical method to structure a Gaussian distributed but quite complex data set. For this, the orthogonal transformation reconstructs the correlated data points into the so–called principal components that are linear and uncorrelated. In our case, the input is given as a data matrix containing a certain number of proteins represented by different features. Then, the orthogonal transformation matrix is calculated based on the eigenvectors of the covariance matrix from the data matrix.

In hierarchical clustering, similar proteins are grouped into the same cluster, whereas proteins with a larger Euclidean distance are grouped to another cluster. At first, each protein gets its own cluster. In the next step, every two clusters with the smallest distance in–between are combined to a larger cluster. This is iteratively done until all proteins belong to a single cluster.

Both, the PCA and the hierarchical clustering were realized by R using the functions `prcomp()` and `hclust()`.

5.9. Mapping non–Transmembrane Regions

To analyze the conservation level of loop lengths, non–transmembrane and transmembrane regions were mapped to each other for sequences belonging to the same or to different protein families. For this, multiple binary string alignments (MBA) were used. For this procedure, I wrote a Java program.

5.9.1. Multiple Binary String Alignment

At first, the amino acid sequence of every protein was translated into a binary string containing a W for residues of the non–transmembrane region and a P for transmembrane residues. Thus, W and P here did not represent the amino acids tryptophan and proline, but were chosen due to strong physicochemical differences between both amino acids. Then, the construction of the MBA starts with the last residue of a non–TMS, i.e., the last position of a W box, and the first residue of the TMS that follows in the protein sequence. Both residues represent two directly neighbored columns in the alignment. The alignment grows in both directions filling columns either with W or P until the length of the longest non–TMS or TMS is reached. The boundary columns of the sequences with shorter non–transmembrane and transmembrane regions contain a gap. Afterwards, the same procedure is repeated for the next non–transmembrane and transmembrane regions until the full sequence is included in the MBA. Since the number of TMSs between different membrane proteins varies strongly, the main problem is to figure out

which segment of the first protein should be mapped to which segment of another protein. For this, the following strategy was chosen.

For all sequences in a certain protein family, a MSA was calculated by the very efficient tool Multiple Alignment using Fast Fourier Transform (MAFFT) [81] as a so–called reference MBA. MAFFT applies the fast Fourier transform regarding the correlation between amino acid sequences based on their volume and polarity. Here, MAFFT version 6 was used. A second MBA was built in the same way for all members of the corresponding protein family and the protein of unknown function denoted as actual MBA. Both MBAs were realigned such that no mismatches occur and all gaps are grouped together, see Figure 5.6. For this, a column

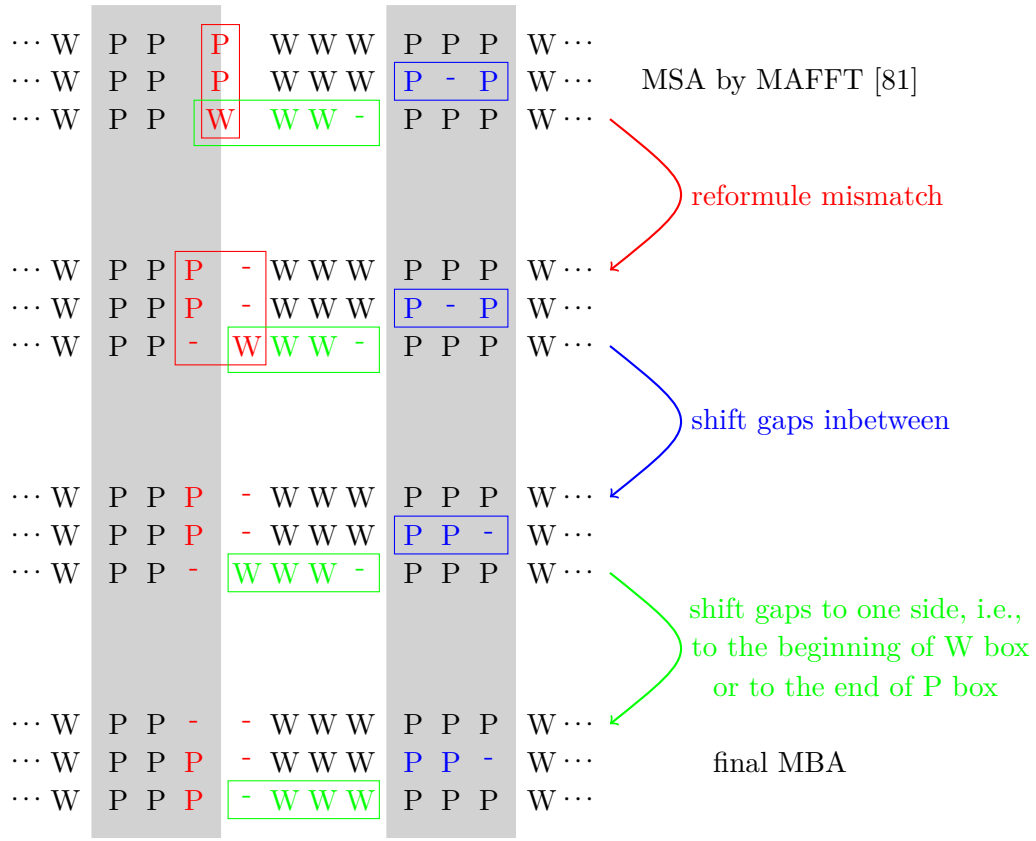


Figure 5.6.: **Realigning MBA:** Starting with a MSA from a binary sequence of several proteins generated by MAFFT [81], a final MBA is constructed by removing mismatches (shown in red) and shifting wrongly located gaps (gaps in–between are shown in blue and gaps at the wrong side are shown in green).

with mismatches was divided into two columns with only one character and gaps each. In the following step, gaps that were placed in the middle of a W or P box were shifted to a boundary. Then, there are possibly gaps at both ends of a box. In the case of a W box, all gaps were arranged at the beginning of the box, whereas all gaps of a P box were located at its end. It may be that during this realignment process columns arise which only contain gaps. Those useless columns were deleted.

Afterwards, a score s for both MBAs was computed, as described below. Finally, the unknown protein can be classified to the considered protein family if the score of the actual MBA is only slightly different from the score of the reference MBA.

5.9.2. Score Measurement

There are several different possibilities to measure the similarity between sequences in an alignment. For simplicity, the number of gaps $|gaps|$ normalized by the alignment size was calculated as given in equation (5.14),

$$s_{gap}(MBA) = \frac{|gaps|}{|gaps| + |W| + |P|} \tag{5.14}$$

whereby $|W|$ and $|P|$ represent the total number of non–transmembrane and transmembrane residues in all considered proteins. As a second measurement, the number of well matched boxes was computed as defined in equation (5.15),

$$s_{box}(MBA, k) = 1 - \frac{|mbox_k|}{|box_k|} \tag{5.15}$$

whereby $k \in \{W, P\}$, $|box|$ denotes the maximum number of segments, and $mbox$ represents a box with $\leq 2q$ gaps for q proteins in the alignment. The smaller the scores are, the more similar are the proteins in the alignment according to their loop lengths.

5.9.3. Evaluation of Mapping Procedure

Again, a LOOCV was done to evaluate the procedure. Here, the MBA from the full protein family of size q represents the actual MBA and a MBA from a subset of size $q-1$ the reference MBA. This was repeated for all possible subsets. Further, a MBA was built for a protein family together with a protein belonging to another family. Its score was compared to that of the reference MBA built by the considered family.

6. Results and Discussion

To predict putative substrates of *A. thaliana* membrane transporters and TC protein families, different AACs were used as simple mathematical representation. For this, the AAC was at first tested in the context of a suitable feature and a detailed analysis how TMSs are connected to the function of membrane proteins was then provided.

6.1. Relation between Amino Acid Composition and three–dimensional Structure

Due to their similar 3D structures, we considered the four membrane transporters LeuT, BetP, *Pm*CaiT, and *Ec*CaiT to test whether the amino acid frequencies are mainly responsible for the structure of a protein or for its specific function. LeuT is a leucine transporter from *Aquifex aeolicus* and BetP a sodium–coupled glycine betaine symporter from *Corynebacterium glutamicum*. The two CaiT proteins are sodium–independent antiporters for carnitine and butyrobetaine in *Proteus mirabilis* (*Pm*) and *Escherichia coli* (*Ec*). Both CaiT proteins have a sequence identity of 87% and a RMSD[1] of less than 1 Å to each other, a sequence identity of 25% and a RMSD of 2.2 Å to BetP as well as 10% and 4.3 Å to LeuT [166]. The similarity based on amino acid frequencies is sketched in Figure 6.1. Obviously, the two CaiT proteins

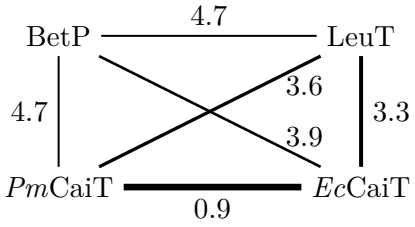


Figure 6.1.: **Similarity graph based on original amino acid composition considering membrane proteins with similar three–dimensional structure:** The thicker the line, the smaller is the distance, which is given as edge labels, according to the original amino acid composition.

showed the smallest AAC distance. In contrast to sequence and structure similarity, the AAC similarity of CaiT to LeuT was higher than to BetP. This suggests that the AAC is related more closely to function than to structure.

[1] Root–mean–square deviation (RMSD): measures the average distance between equivalent atoms of the two proteins

6.2. Amino Acid Composition for Substrate Annotation

Suitable features for substrate prediction must show a clearer similarity between proteins with similar function than without. Since substrates have to interact with their transporters, physicochemical characteristics and accordingly amino acid frequencies are supposed to show a more pronounced characteristics in a certain transporter class than between miscellaneous groups.

6.2.1. Similarities and Dissimilarities

For this, the Euclidean distances of the oAACs were calculated between all *A. thaliana* transporters of a particular substrate, namely amino acid, oligopeptide, phosphate, and hexose denoted as positive sets as well as between two of those groups. Figure 6.2 illustrates the found similarities and dissimilarities. Compared to the full *A. thaliana* data set with an averaged similarity score of 0.374, the members of the positive sets were closer to each other according to the AAC. Since they are composed of more different protein families, the phosphate and hexose transporter sets were more heterogeneous than the oligopeptide and amino acid sets. Further, the phosphate transporters differed strongly in their sequence length, see Figure A.1 in appendix in Section A.2. In general, the AACs differed in the same way from each other as between proteins with LeuT fold, see Section 6.1. Thus, distances of amino acid frequencies generated by a specific 3D structure were not smaller than those determined by certain substrates.

As expected, amino acid transporters were directly linked to oligopeptide transporters whose substrates consists of several amino acids and hence have similar properties as individual amino acids. Furthermore, the phosphate transporter set was connected to the sets of oligopeptide and hexose transporters. Subsequently, we determined the amino acid types having the largest

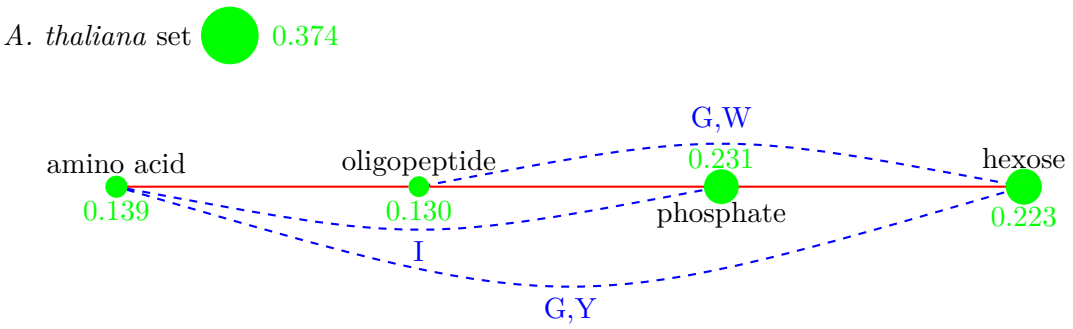


Figure 6.2.: **Similarity graph based on original amino acid composition considering *Arabidopsis thaliana* data sets:** The larger the green colored dot, the more different is the positive set according to the averaged similarity scores given as green numbers. Straight red line denotes a significant similarity between the two connected positive sets and blue dashed lines indicate dissimilarities. The edge labels denote which amino acids are mainly responsible for the differences.

differences between transporter classes according to equation (5.4), see Section 5.4. The frequencies of glycine and tyrosine cause mostly the dissimilarities between amino acid and hexose transporters, of glycine and tryptophan between oligopeptide and hexose transporters, and of isoleucine between amino acid and phosphate transporters, also given in Figure 6.2. In summary, we found that proteins with similar function were also related in their AAC and there was a clear border between different transporter groups. Hence, it seems that the AAC is a suitable feature for substrate prediction.

6.2.2. Significance of Dissimilarities

Figure 6.3 shows significant differences according to ANOVA p–values between the individual transporter sets based on non–disjoint categories (listed in Table 6.1) of amino acid properties.

category	amino acids
hydrophobic	A, C, F, I, L, M, P, T, V, W, Y
aliphatic	I, L, V
aromatic	F, H, W, Y
polar	C, D, E, H, K, N, R, Q, S, T, W, Y
charged	D, E, H, K, R
positive	H, K, R
negative	D, E
large	F, K, R, W, Y
small	A, C, D, G, N, P, S, T, V
tiny	A, C, G, S

Table 6.1.: **Categories based on physicochemical properties** of amino acids given in one–letter code.

Considering a p–value less than 0.001 to reject the null hypothesis that the content was similar in both sets, oligopeptide transporters diverged from amino acid and phosphate transporters in the frequency of small amino acids. Additionally, the hexose transporter set showed a significant difference to amino acid transporters regarding aromatic and hydrophobic amino acids, to oligopeptide transporters in the content of polar, large, and aliphatic amino acids as well as to phosphate transporters with respect to aliphatic amino acids. For a less strict p–value threshold, aliphatic amino acids facilitated, for example, to distinguish between amino acid and phosphate transporter sets. The remaining p–values are given in appendix in Table A.2 in Section A.2.

As shown in appendix in Table A.3 in Section A.2, the positive sets were quite similar with respect to these categories based on physicochemical properties. In comparison to the full *A. thaliana* set, the oligopeptide transporter set contained 25% more aromatic amino acids and 10% more large amino acids. The hexose transporter set had more than 15% fewer positive amino acids. Additionally, amino acid transporters consisted clearly less of negative amino acids.

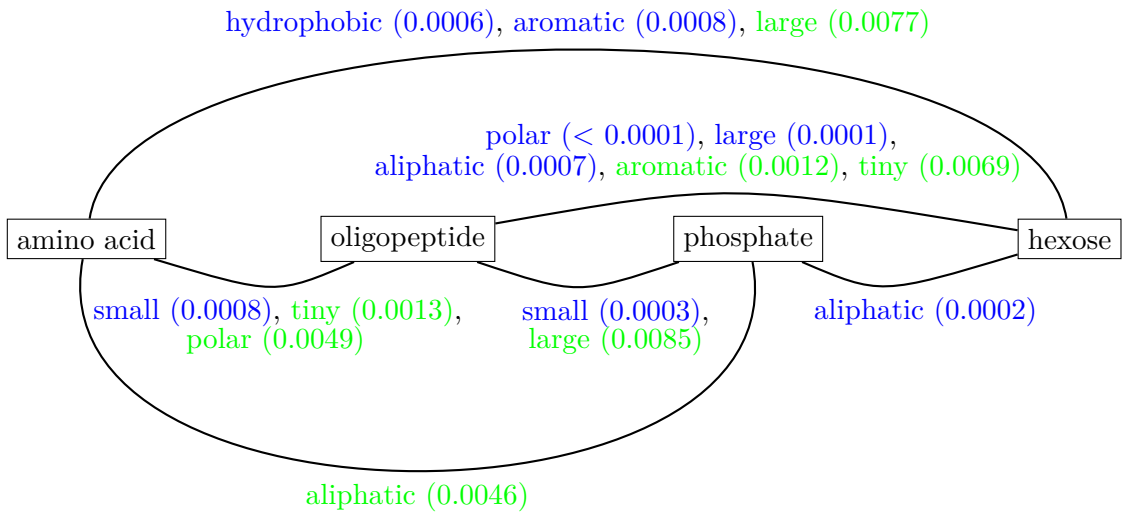


Figure 6.3.: **Significant differences** according to p–values obtained by an analysis of variance. Blue colored physicochemical categories (see Table 6.1) represent a strong significance with $p < 0.001$, green colored amino acid types denote $p < 0.01$.

6.2.3. Comparison with Sequence Identity

As already mentioned, BLAST is often used to functionally identify membrane proteins [98]. However, the membrane proteins are usually matched to established protein families, such as TC families, and not to their specific substrates. In order to test whether sequence homology could be also useful to distinguish between transported substrates, the average sequence identity between members of the same and different positive sets was analyzed using BLAST [3], see Table 6.2. In general, the E–values were quite high and there was no significant difference

set	amino acid	oligopeptide	phosphate	hexose
amino acid	1.90	1.78	1.99	1.23
oligopeptide		1.47	1.37	1.26
phosphate			1.67	0.84
hexose				1.43

Table 6.2.: **Sequence identity:** Shown is the averaged BLAST [3] E–value between members of the same positive set having different TC families (red diagonal entries) and between members of two different positive sets.

between members of the same positive set with different TC family or of different positive sets. As seen in the similarity graph (Figure 6.2) according to AAC, phosphate and hexose transporters were quite similar to each other. In contrast to AAC, the sequence similarity was high between hexose and oligopeptide or amino acid transporters. Therefore, sequence homology based methods seem to be promising for prediction of substrate specificities beyond the boundaries of single TC families. This conclusion is in perfect agreement with a forthcoming

study by Barghash and Helms in 2013 [11]. They found that sequence homology based tools, such as BLAST or HMMER3 [43], are able to classify membrane proteins according to their TC families, but clearly less satisfying according to their substrate groups. They further suggested the usage of E–value thresholds lower than 10^{-4} for a reliable result. In our *A. thaliana* positive sets, only two transporter pairs having different TC families and belonging to the same positive set fulfilled this requirement.

6.3. Reliability of Transmembrane Segment Annotation

Due to the lack of available 3D structures, the TMHs and TMBs used for classification were generated by the tools Memsat–SVM [122] and TMBETAPRED [128]. Since the accuracy of those methods strongly influences the prediction quality, the results of Memsat–SVM, TOPCONS [15], and TMBETAPRED were compared with TMS positions assigned by OPM for membrane proteins with known 3D structure. At first, this shows that the tools were quite reliable to detect transmembrane regions. Although, most wrongly identified residues appear at both borders of TMSs. Figure 6.4 demonstrates that for α–helical membrane proteins Memsat–

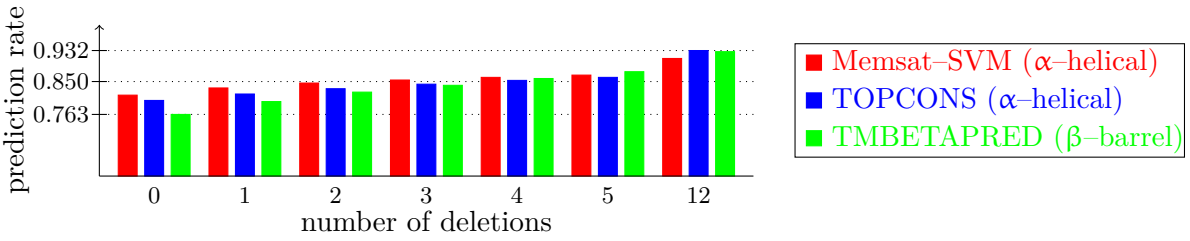


Figure 6.4.: **Connection between correctly predicted residues and transmembrane boundaries:** Shown is the rate of correctly placed transmembrane segment positions when removing amino acids at the beginning and end of a transmembrane region.

SVM correctly recognized 81.4% of the residues considering the full transmembrane region and 91.1% when twelve residues were removed from right and left transmembrane boundary and for β–barrel membrane proteins TMBETAPRED 76.3% and 92.9%, respectively. In the case of twelve removed amino acids, the wrongly assigned residues represented missed TMSs. Additionally, those observations indicated that Memsat–SVM behaves similar to TOPCONS, which correctly annotated about 80% of the amino acids. In contrast, Tsirigos *et al.* reported in 2012 that TOPCONS behaves clearly better than Memsat–SVM considering the N– and C–terminal location, the number of TMHs, and the location of TMHs regarding the overlap of at least five amino acids [187]. Nevertheless, in the following TOPCONS was not used because the overlap of most residues was important for the later functionally classification.

Paradoxically, the tools predicted some sequence parts as both TMHs and TMBs for about 50% of the proteins in the TCDB data set. For helical membrane proteins, the overlap between Memsat–SVM and TMBETAPRED assignments is 74.1% and for β–barrel proteins 57.8%. Although, most considered proteins were known to be α–helical. Nevertheless, about 25% of the proteins in the α–helical family 3.A.1 were annotated as β–barrel. Therefore, the overlap between Memsat–SVM and TMBETAPRED predictions was analyzed in detail on the basis of the OPM data set. While TMBs usually contained more than twelve amino acids TMHs were often longer with at least 20 residues. Since short TMBs were therefore not annotated as TMHs, the overlap for α–helical proteins was with 63.8% remarkably higher than for β–barrel proteins with 49.2%. However, the predictions still covered 64.2% in the case of Memsat–SVM for β–barrel proteins and 72.6% in the case of TMBETAPRED for α–helical proteins of the TMSs given by the OPM database.

6.4. Discrimination between α–helical and β–barrel Proteins

Due to the overlaps between TMHs and TMBs, a discrimination between α–helical and β–barrel proteins was required. For simplicity, the oAAC without any further information was used. The Euclidean distance was computed between the AAC of the protein with unknown character and two average AACs, one representing all α–helical membrane proteins of the OPM data set and the other all β–sheet proteins. Then, the protein was matched to the character with the smaller distance. This method was applied to 44 proteins of the OPM data set for which Memsat–SVM and TMBETAPRED assigned overlapping TMSs. Here, 14 out of 20 α–helical and all 24 β–barrel membrane proteins were correctly identified. Considering the full OPM data set, only six proteins were wrongly classified to have β–barrel character. In general, wrong annotations have the same effect as wrongly predicted boundaries of TMSs.

6.5. Number and Length of Transmembrane Segments

Splitting the AAC into TMSs and non–TMSs for functional classification was assumed to be rather successful when the secondary structure is conserved. In Figure 6.5, TMSs were considered in detail for the positive sets. In general, the *A. thaliana* sets were more homogeneous than the TCDB families regarding the number of TMS except for phosphate transporters with a standard deviation of 3.1 and for H.2.A.1 with a standard deviation of 2.9. In contrast, the most heterogeneous sets H.1.A.1 and H.3.A.1 were characterized by a quite large standard deviation and a relatively low number of TMHs on average.

For some members of the helical TCDB sets which are actually specified in TCDB as transport proteins, Memsat–SVM found only a single TMH. Further, whereas the galactose transporter

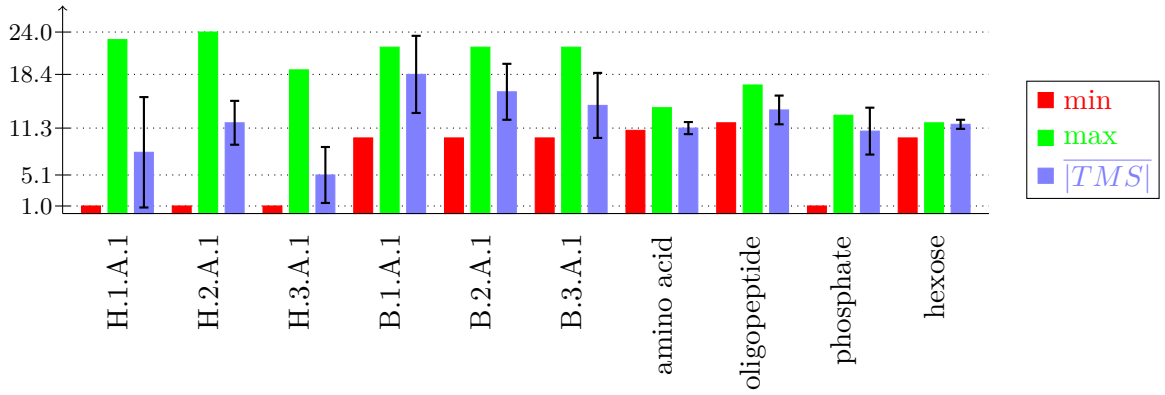


Figure 6.5.: **Occurrence of transmembrane segments in positive sets:** The red and green bars symbolize the minimal and maximal numbers of transmembrane segments. Blue bars the average number of transmembrane segments with their standard deviations.

vSGLT contains 14 TMHs [1], a structure of twelve TMHs is strongly conserved in sugar transporters [108, 60] as also fulfilled for a majority of membrane proteins in the hexose transporter set.

Incorrectly annotated TMS boundaries have a stronger impact on the classification quality for short transmembrane regions. Figure 6.6 demonstrates that the ratio of the cumulative TMS length and the total sequence length was higher for *A. thaliana* sets than for TCDB sets, especially for the the positive set B.2.A.1.

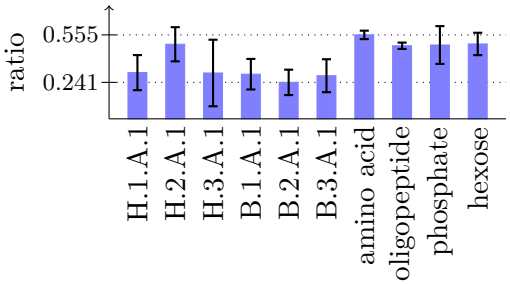


Figure 6.6.: **Ratio between the cumulative length of transmembrane segment and the total sequence length:** Shown are average values in the positive sets with their standard deviations.

Despite the large standard deviation the AAC splitting into TMSs and non–TMSs may be helpful for classification, when the mean of the set is placed between two or more independent clusters that each have a small standard deviation. Otherwise, the large standard deviation results from a large variation in the set. Therefore, Figure 6.7 clarifies the distribution of TMHs for H.1.A.1 and H.3.A.1, the sets with the largest standard deviation. It also displays

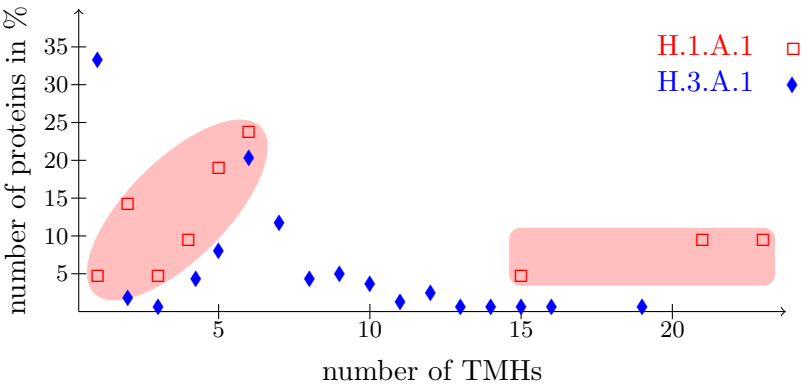


Figure 6.7.: **Distribution of Transmembrane helix number** for the TC subsets H.1.A.1 and H.3.A.1. Areas in light red exemplify clusters.

two homogeneous clusters into which H.1.A.1 could be possibly divided up. When ignoring the large number of proteins with only one TMH, the distribution of H.3.A.1 could be approximated by a Gaussian distribution, but still with a rather wide dispersion.

6.6. Amino Acid Frequencies based on Physicochemical Properties

Now, sequence–based diversity between the studied sets of membrane proteins was analyzed with respect to sequence regions in order to identify that part of the sequence having the highest discrimination. For this, Figure 6.8 illustrates the over- and underrepresentation of non–disjoint physicochemical categories (see Table 6.1) comparing the full sequence, TMS, and non–TMSs. As generally known for membrane transporters, the categories hydrophobic and polar occurred most frequently in the proteins of the considered positive sets. Polar amino acids, such as arginine, asparagine, aspartic acid, glutamine, glutamic acid, histidine, and lysine, form hydrogen–bonds to recognize the transported substrate [61]. In the full sequence, the occurrence of aliphatic amino acids was detectably reduced for B.OPM.

Compared to non–TMSs, positive, negative, and polar residues were remarkably overrepresented in TMHs, but not that strongly in TMBs due to the apolar external surface of β–barrels [65, 197, 75]. However, the frequency of negatively charged amino acids was noticeably higher in the full sequence for β–barrel proteins than for α–helical ones. Concerning the low polarity in membranes, the fraction of the aliphatic amino acids isoleucine, leucine, and valine as well as of the hydrophobic amino acids alanine, glycine, and phenylalanine was clearly higher in TMSs than in non–TMSs as in common for integral membrane proteins [188]. Further, aromatic amino acids were enriched in TMSs in agreement with data about the interactions between aromatic rings and phosphate groups of phospholipids in the interface region [84]. Connected to the packing ability [199], the tendency of amino acids to be in TMSs depends on their partial

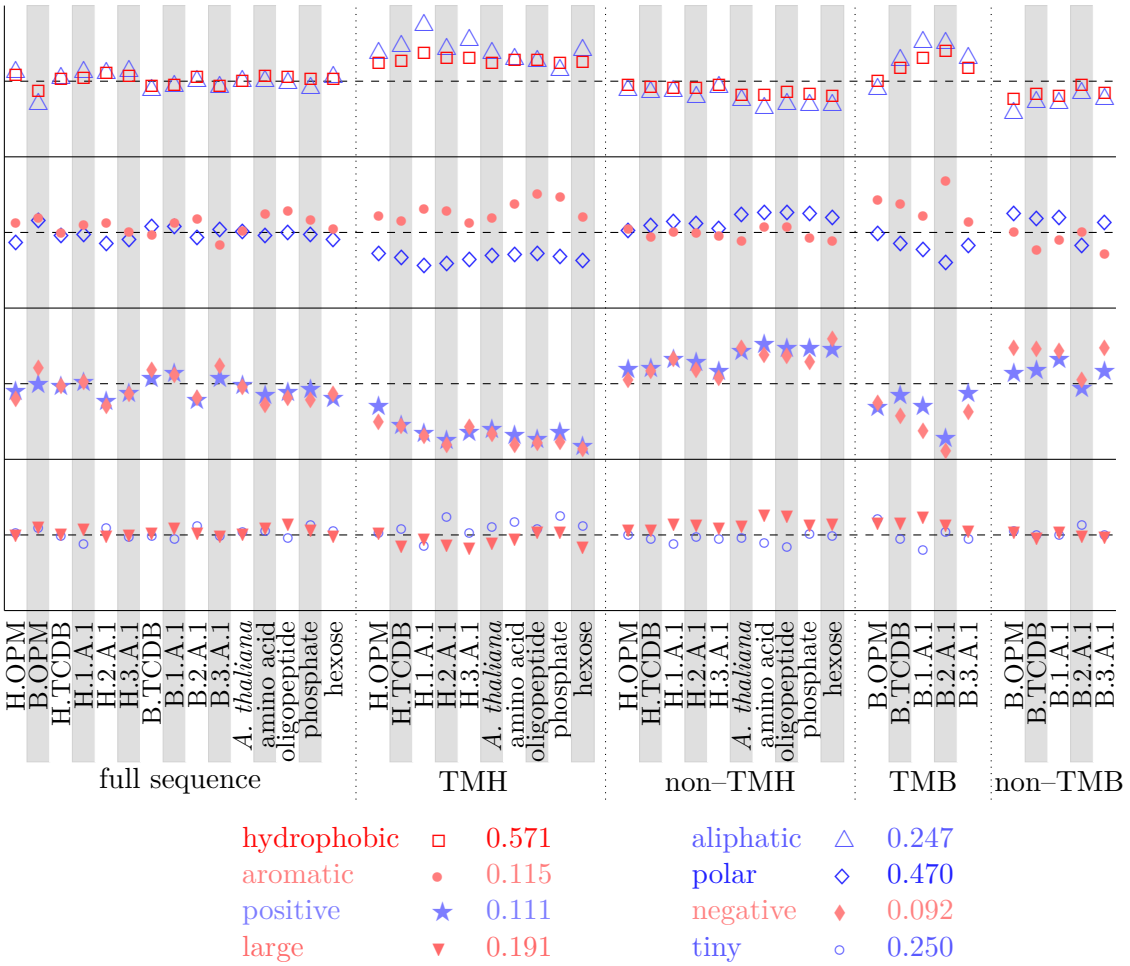


Figure 6.8.: **Average frequencies based on physicochemical properties** for the different sequence regions. Shown are deviations from the average value that is denoted as dashed line considering different categories (see Table 6.1). The frequency values are given in appendix Section A.2 in Tables A.3, A.4, and A.5.

specific volume [133]. Thus, it is not surprising that TMSs reveal more tiny and fewer large amino acids than non–TMSs. However, instead of substrate properties, structural reasons may be responsible for the higher content of tiny residues in TMHs. In this way, alanine is usually preferred to form helices [19]. Glycine, the breaker of soluble helices [20], is in membrane proteins related to facilitating a closer packing of TMHs, to crossing points and helix–helix interactions [74]. Also, glycine is involved in the GxxxG motif that is quite common in TMHs [170]. The frequency of alanine in TMHs was on average 0.022 higher than of the full sequence and the difference in those regions for glycine was less than 0.004.

With respect to the task of classification, of primary importance are the obvious differences based on the frequencies of these categories between the separate positive sets. Whereas the physicochemical frequencies of the individual sets were rather homogeneous, most reliable variations were detected for hydrophobic, aromatic, polar, positive, and negative amino acids.

Since splitting the sequence into TMSs and non–TMSs allows for a better classification for intensification of those differences in these areas compared to the full sequence, we searched for regions with the strongest variations between the positive sets. For this, pairwise dissimilarities between positive sets according to the frequencies of categories with the most variations are considered in the following. Further, we tried to connect the enrichment to the substrates of the proteins.

6.6.1. Frequency Differences in TCDB Sets

Positively and negatively charged residues were dominant in H.1.A.1 and B.1.A.1, which include voltage dependent ion channels, (see Figure 6.8) as expected because the substrates, the small and charged ions, may directly interact with the oppositely charged amino acids of the channel. Moreover, the conformational transition of some membrane–voltage gated ion channels is mediated by charged residues, e.g., in the S4 segment [18, 76]. In contrast, the MFS (2.A.1) and ABC (3.A.1) superfamilies have a wide range of substrates, such as amino acids or sugars. In B.2.A.1, hydrophobic, aliphatic, and aromatic residues were strongly overrepresented and polar, positive, and negative underrepresented compared to the other β–barrel TCDB proteins. This confirmed the previous observations by Gromiha and Yabuki that the polar amino acids asparagine, aspartic acid, and tyrosine are overrepresented in channels/pores (TCDB class 1), and hydrophobic amino acids, such as phenylalanine, isoleucine, leucine, and valine, in electrochemical potential–driven transporters (TCDB class 2) [62].

Figure 6.9 shows that H.1.A.1 can be distinguished from H.2.A.1 on the basis of the categories positive and negative, independent of the sequence region. In comparison, H.1.A.1 and H.3.A.1 were more similar to each other and differ mainly in the frequency of negative residues. Here, TMSs were generally more suitable for a discrimination. The differences between H.2.A.1 and H.3.A.1, which were especially given by positive and negative amino acids, were also clearly more pronounced in transmembrane parts. As already mentioned, B.2.A.1 varied remarkably from the other sets. For the β–barrel TCDB proteins, the differences were typically larger in transmembrane regions than in the full sequence.

6.6.2. Frequency Differences in *Arabidopsis thaliana* Sets

According to Figure 6.8, the frequency of positive amino acids in the phosphate transporter set was higher than that of negative amino acids and higher in comparison to the other *A. thaliana* sets possibly due to interactions with the negatively charged phosphate ions. In contrast, hexose transporters contained less positive and large as well as more aliphatic residues than the other *A. thaliana* sets. To recognize and bind hexose molecules, hydrophobic amino acids are quite important; valine and tryptophan, for example, form interactions for structure stabilization and allow a compact protein fold [108]. Whereas the content of valine was indeed high, the tryptophan and tyrosine frequencies were low resulting in a corresponding low frequency of

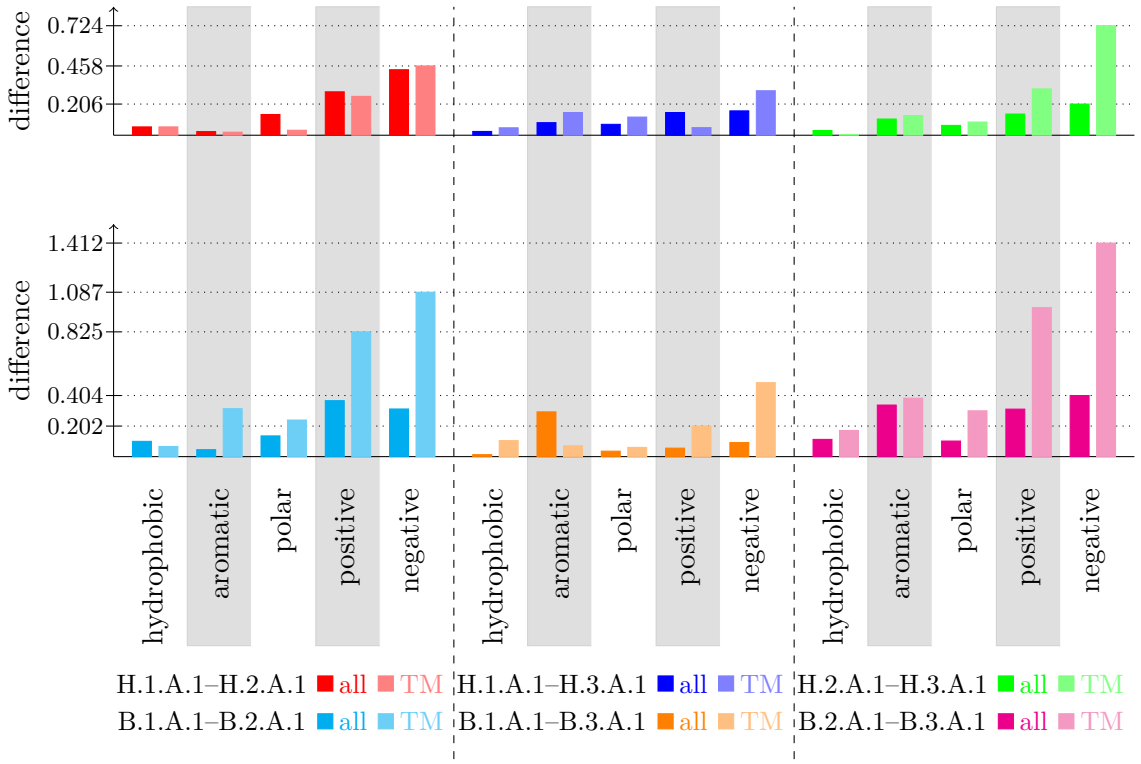


Figure 6.9.: **Frequency differences for TCDB sets:** Shown are the absolute differences of two positive sets normalized by their average frequency either for the full sequence (all) or for the transmembrane regions alone (TM).

aromatic residues for hexose transporters. In general, tryptophan and tyrosine residues are involved in stabilization of membrane proteins due to the formation hydrogen bonds [163, 208].

Figure 6.10 shows that the *A. thaliana* sets differed primarily from each other in the content of negative and positive residues of the transmembrane region. Here, the differences were typically smaller in comparison to the TCDB sets which include a clearly wider range of different membrane proteins. Further, hexose transporters can be distinguished from the other sets considering polar amino acids again in TMSs. The frequency of aromatic residues also varied between the different sets. There was a detectable difference between hexose and amino acid or oligopeptide transporters independent of the sequence region, between oligopeptide and phosphate transporters considering the full sequence, as well as between amino acid and oligopeptide transporters or phosphate and hexose transporters in the transmembrane parts.

Next, we determined the significance of the diversity in these physicochemical characteristics between the positive sets with the Wilcoxon–Mann–Whitney test. Figure 6.11 illustrates the significant dissimilarities when comparing the full sequence with transmembrane regions. The null hypothesis that the values are the same for both positive sets was rejected if the p–value was less than 0.001. In the TMSs, the variations between the positive sets were typically larger than over the full sequence. The p–values according to TMSs were detectably smaller than

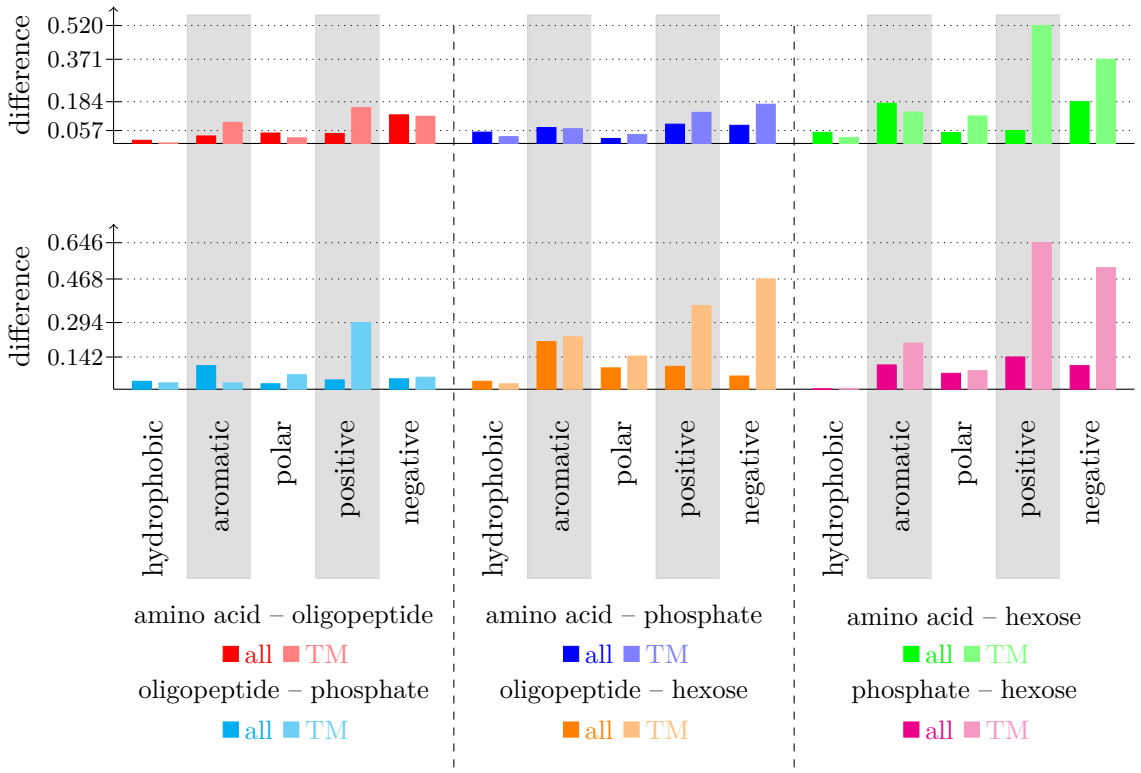


Figure 6.10.: **Frequency differences for *Arabidopsis thaliana* sets:** Shown are the absolute differences of two positive sets normalized by their average frequency either for the full sequence (all) or for transmembrane regions alone (TM).

p–values considering the full sequence. For example, the p–values corresponding to polar and large between oligopeptide and hexose transporters as well as for aliphatic between phosphate and hexose transporters in transmembrane regions were about half of that over the full sequence. Consequently, more dissimilarities became significant. Whereas the amino acid and oligopeptide transporters were again quite similar in the full sequence, at least the frequency of tiny residues was clearly different in transmembrane sequence parts. The phosphate and hexose transporter sets diverged in the frequencies of aliphatic, large, positive, and tiny amino acids. In general, the Wilcoxon–Mann–Whitney p–values for the full sequence were rather similar to the ANOVA p–values (see Figure 6.3 in Section 6.2), whereby the Wilcoxon–Mann–Whitney test is slightly more robust resulting in slightly less pronounced significances. The remaining p–values are given in appendix Section A.2 in Table A.6.

In summary, the AAC or even more rough frequencies based on physicochemical characteristics can be used to group membrane proteins into their corresponding family or substrate classes. In general, the variations were stronger in TMSs than in the full sequence such that a filtering between transmembrane and non–transmembrane parts for classification purpose should be advantageous.

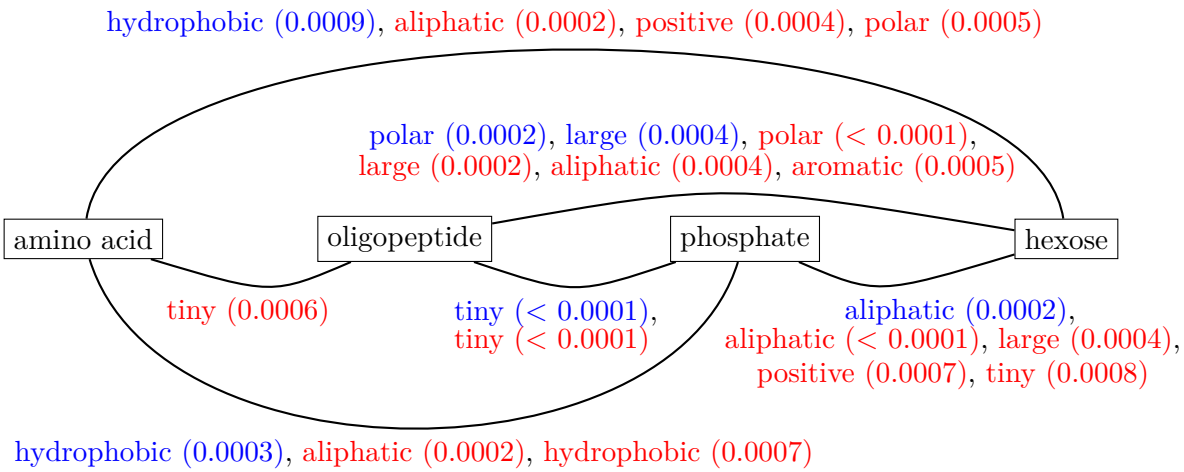


Figure 6.11.: **Significant differences** according to Wilcoxon–Mann–Whitney p–values with a significance level of $p < 0.001$. Blue colored physicochemical categories (see Table 6.1) denote a significant diversity over the full sequence, red colored amino acids types denote a significant diversity in transmembrane regions.

6.7. Classifications based on Families and Substrates

As discussed in detail in the previous sections, the AAC was expected to successfully discriminate different membrane proteins and a splitting into TMSs and non–TMSs may improve the prediction quality. The idea behind this was to assume that the properties of a substrate are closely related to the transmembrane characteristics due to the necessary contact between substrate and transporter when the substrate passes the central pore in the transmembrane region. Now, the complete TCDB and *A. thaliana* sets were classified regarding their family or substrate groups. Since several score measures behaved similarly, the Euclidean distance was used for simplicity in all predictions.

6.7.1. Comparison between different Amino Acid Composition based Features

A SVM with linear kernel based on the AAC was used to classify *A. thaliana* membrane transporters according to their substrate classes, see Table 6.3. With about 85% on average, the sensitivity was quite acceptable. The specificity was slightly lower than the sensitivity. The difference between the performance of training and test sets was relatively small such that there is no indication for overtraining. Since SVMs are generally most reliable for very large data sets with equally sized subsets, the ranking method was preferred for the comparably small data sets considered in this thesis.

Further, we used the ranking method based on an average search profile, see Section 5.6, for the *A. thaliana* data sets. Figure 6.12 sketches a ranking for the amino acid transporter

positive set	sensitivity	specificity	accuracy
amino acid	0.900	0.867	0.895
oligopeptide	0.827	0.583	0.716
phosphate	0.703	0.617	0.673
hexose	0.950	0.950	0.950
average	0.845	0.754	0.809

Table 6.3.: **Quality of classification** for *Arabidopsis thaliana* sets using a support vector machine with linear kernel based on the amino acid composition. Given are the quality measurements averaged over five different test sets.

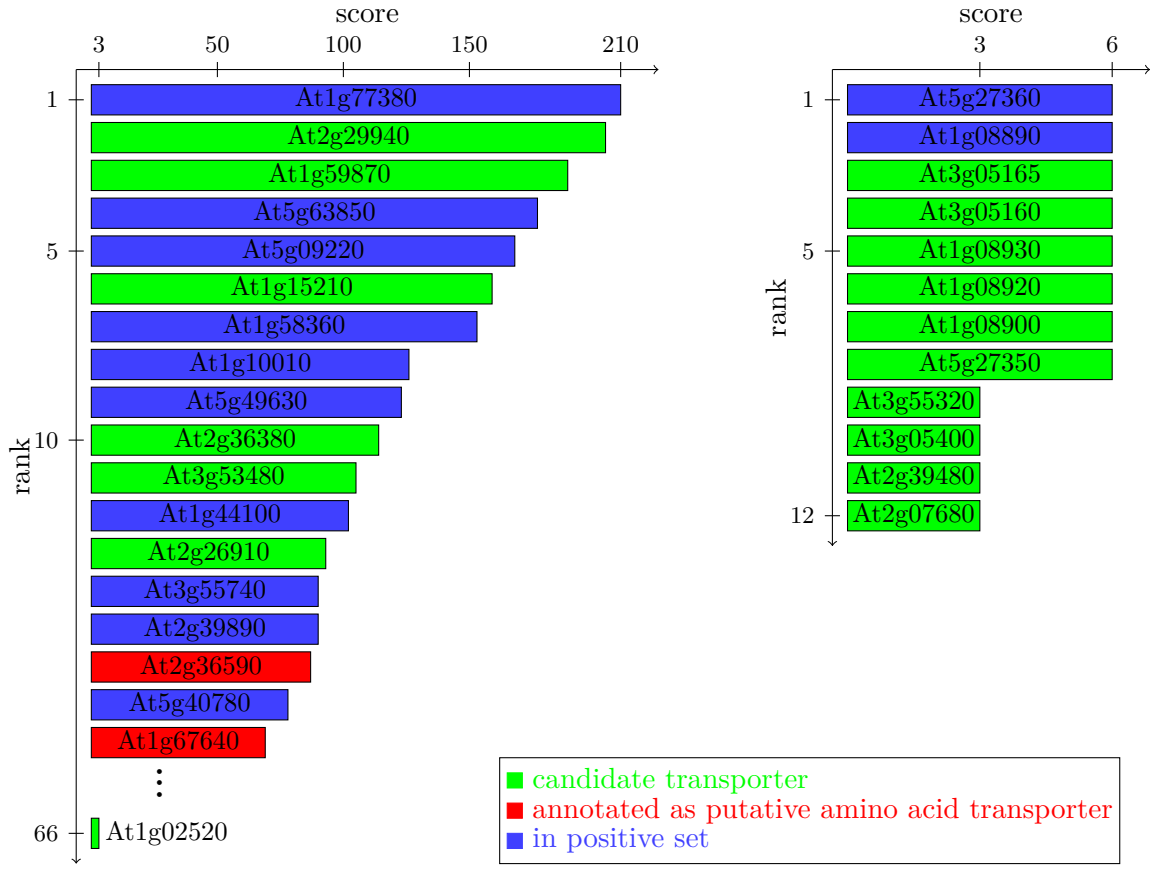


Figure 6.12.: **Final joint ranking** of the amino acid transporter set (left) and a random set (right). The score equals the number of methods that ranked the transporter among the ten most similar proteins to the corresponding positive set. Here, the ranking method based on average search profile was used.

set and a randomly generated positive set. This random set contains, e.g., different ion and putative sugar transporters. All members of the amino acid transporter set were found by an individual method and most of them were listed at the top of the ranking. Interestingly, proteins annotated as putative amino acids transporters in Aramemnon [167] and therefore not included in the positive set, were also found among the top positions. In contrast, only two members of the random set were listed in the ranking. In comparison to the random set, the scores of true positives and possible candidates were clearly higher.

As listed in Table 6.4, the sensitivity and accuracy are generally over 70%. In this context,

positive set	feature	sensitivity	specificity	accuracy
amino acid	oAAC	0.875	0.875	0.875
	PAAC	0.938	0.867	0.903
	PseAAC	0.875	0.875	0.875
	MSA–AAC	0.875	1.000	0.938
	PsePAAC	0.938	0.875	0.906
oligopeptide	oAAC	0.941	1.000	0.970
	PAAC	0.933	1.000	0.968
	PseAAC	0.882	1.000	0.939
	MSA–AAC	0.882	1.000	0.939
	PsePAAC	1.000	1.000	1.000
phosphate	oAAC	0.800	0.667	0.733
	PAAC	0.933	1.000	0.968
	PseAAC	0.800	0.667	0.733
	MSA–AAC	0.933	1.000	0.968
	PsePAAC	0.933	1.000	0.968
hexose	oAAC	0.769	0.909	0.833
	PAAC	0.769	1.000	0.875
	PseAAC	0.769	0.909	0.833
	MSA–AAC	0.769	1.000	0.875
	PsePAAC	0.769	1.000	0.875
average	oAAC	0.846	0.863	0.853
	PAAC	0.893	0.967	0.929
	PseAAC	0.832	0.863	0.845
	MSA–AAC	0.865	1.000	0.930
	PsePAAC	0.910	0.969	0.937

Table 6.4.: **Quality of classification:** Sensitivity, specificity, and accuracy are given regarding amino acid composition based features for the *Arabidopsis thaliana* positive sets. Red colored entries denote the highest measures for a data set. Here, the ranking method based on average search profile was used.

the sensitivity is the more reliable measurement due to the influence of the randomly compiled negative sets on the specificity. Using the original AAC (oAAC) and the physicochemical properties including PseAAC led to a similar classification quality, whereas the amino acid pairs considering PAAC and the profile–based MSA–AAC clearly improved the results. The combined PsePAAC further enhanced the sensitivity such that it achieved a very high accuracy of 93.7% on average. For comparison, Park and Kanehisa reported that the PAAC improved the total accuracy from 72.4% for the oAAC to 75.9% to predict subcellular locations [131].

Interestingly, the accuracy was only marginally higher for the homogeneous amino acid and oligopeptide transporter sets than for phosphate and hexose transporters belonging to multiple different protein families. In the case of the phosphate set, the sensitivity was strongly increased

by the more accurate AACs. Surprisingly, the specificity was unexpectedly high independent of the applied feature and the oAAC provides the highest sensitivity for oligopeptide transporters.

To validate these results, the behavior of the different features for a certain transporter set was compared with several randomly generated positive sets having the same size as the corresponding transporter set. For each positive set, 20 different random sets were considered. The positive and negative sets are much smaller than the full data set such that the chance to find a negative was typically low. As expected, the specificity was only slightly decreased on average, see Table 6.5. Nevertheless, the specificity was reduced to about 76% of the actual

corresponding positive set	feature	sensitivity		specificity	accuracy
		μ	σ	μ	μ
amino acid	oAAC	0.325	0.097	0.837	0.580
	PAAC	0.263	0.214	0.753	0.505
	MSA-AAC	0.257	**0.064**	0.894	0.576
oligopeptide	oAAC	0.324	0.092	0.916	0.613
	PAAC	0.308	0.207	0.809	0.555
	MSA-AAC	0.294	**0.072**	0.916	0.595
phosphate	oAAC	0.373	0.095	0.880	0.627
	PAAC	0.370	0.229	0.787	0.578
	MSA-AAC	0.273	0.094	0.930	0.602
hexose	oAAC	0.358	0.122	0.918	0.611
	PAAC	0.319	0.297	0.586	0.442
	MSA-AAC	0.273	**0.089**	0.927	0.573
average	oAAC	0.345	0.102	0.888	0.608
	PAAC	0.315	0.237	0.734	0.520
	MSA-AAC	0.274	**0.080**	0.917	0.587

Table 6.5.: **Quality of random classification:** Mean (μ) and standard deviation (σ) of the sensitivity as well as mean (μ) of the specificity and the accuracy are given regarding amino acid composition based features averaged over 20 computations with different random sets having the same size as their corresponding positive sets. Entries illustrated in bold denote the smallest standard deviations. Entries in red represent a specificity $> 95\%$ of the corresponding positive set specificity, magenta $> 90\%$, and orange $> 85\%$. Entries in green represent a sensitivity $< 40\%$ of the corresponding sensitivity, cyan $< 35\%$, and blue $< 30\%$. For example, the MSA–AAC on average reached $\frac{0.274}{0.865} = 0.317 < 35\%$ of the sensitivity for the corresponding positive set. Here, the ranking method based on average search profile was used.

specificity on average when using the PAAC because of a probably higher similarity between positive and negative set in the case of randomly constructed sets.

Independent of the feature and the size of the random set, the sensitivity was clearly lower (less than 40%) than the sensitivity for the real data. Since the size of the different sets was rather similar, the sensitivity of the several random sets differs only slightly when considering

the same feature. In general, the sensitivity was stronger reduced for the PAAC than for the oAAC and was strongest reduced for the MSA–AAC. Moreover, the MSA–AAC provided the lowest standard deviation of the sensitivity between different random sets with the same size. Further, a t–test with the null hypothesis that both are in the same range was applied to compare actual and random sensitivities. The null hypothesis was rejected for all positive sets and all features. Therefore, the random sensitivities were significantly smaller than the actual ones.

Additionally, Figure 6.13 illustrates the comparison measurement ξ defined in equation (5.12) in Section 5.6.3 between actual and random predictions. The higher the value is, the more

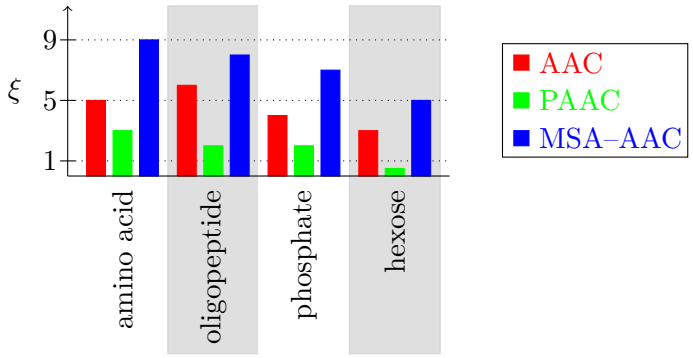


Figure 6.13.: **Comparing measurement** ξ as defined in equation (5.12) between actual and random sets for different features. High values ($\geq$ 3) represent a convincing separation.

reliable is the considered classification. Whereas the relatively low ξ–values of PAAC may be a hint for over–training, the MSA–AAC here performed again very well.

In conclusion, all tested AAC features worked quite well for functionally classifying *A. thaliana* transporters based on substrate groups. Whereas, the PsePAAC achieved the best sensitivities and accuracy for the single positive sets, the profile–based MSA–AAC seems to perform generally best according to the analysis with random classifications.

6.7.2. Amino Acid Composition in different Sequence Regions

As known from literature [21] and confirmed in the previous analysis, TMSs clearly differ in the frequency of physicochemical categories from non–TMSs such that we now try to find out whether distinguishing between TMSs and non–TMSs in the AAC can improve the prediction quality of functional classifications for membrane transporters.

At first, we tried to use PCA to separate the transporters according to their function considering the AAC in TMSs as transporter representation, see Figure 6.14 (left). Indeed, most hexose

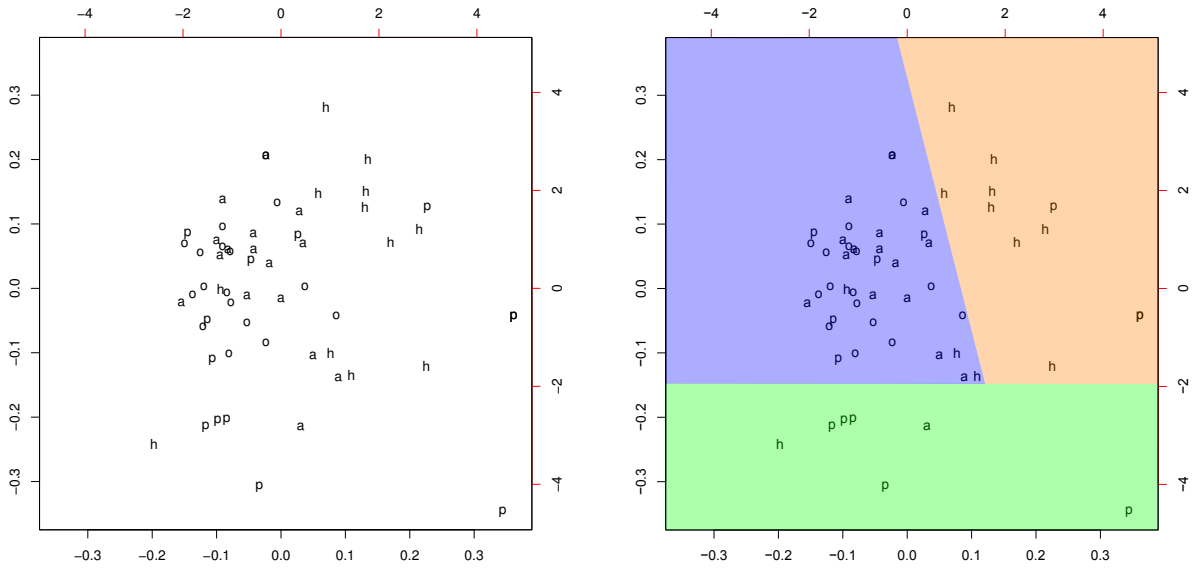


Figure 6.14.: **Principal component analysis based on the amino acid composition in transmembrane positions** for the *A. thaliana* positive sets. Here, "a" denotes amino acid transporters, "o" oligopeptide transporters, "p" phosphate transporters, and "h" hexose transporters. The illustration on the right side tries to cluster the proteins based on the principal component analysis result given on the left side. The blue colored area denotes amino acid and oligopeptide transporters, the green area phosphate transporters, and the orange area hexose transporters.

transporters were located in a certain area. However, the phosphate transporters were spread over the entire area. Further, distinguishing between amino acid and oligopeptide transporters was also not possible since they are closely related according to this PCA. The right graph in Figure 6.14 demonstrates a way to group the proteins into substrate classes. As mentioned, the amino acid and oligopeptide transporter sets belong to a common substrate group (colored in blue) and the number of false negatives is relatively high in the case of the hexose transporter set (colored in orange). Due to the widespread location of the phosphate transporters (colored in green), there are many false positives in both areas. This procedure resulted in a sensitivity of 68.7%, a specificity of 85.1%, and an accuracy of 84.2% on average. The sensitivity of the blue colored amino acid and oligopeptide transporter area was quite high with 97.0%. Further, the specificities of the orange colored hexose and green colored phosphate transporter areas were also relatively high with 91.7% and 95.7%, whereas the sensitivity was very small in the green colored phosphate transporter area with 40.0%. For this reason and since this procedure did not manage to discriminate between amino acid and oligopeptide transporters, we did not consider PCA further.

Figure 6.15 shows the hierarchically clustered *A. thaliana* positive sets based on the AAC over the full sequence. About half of the amino acid transporters (shown in red) were grouped together as a cluster, whereas the remaining proteins are either pairwise clustered or directly to a member of another positive set. The same behavior was found for the oligopeptide (shown

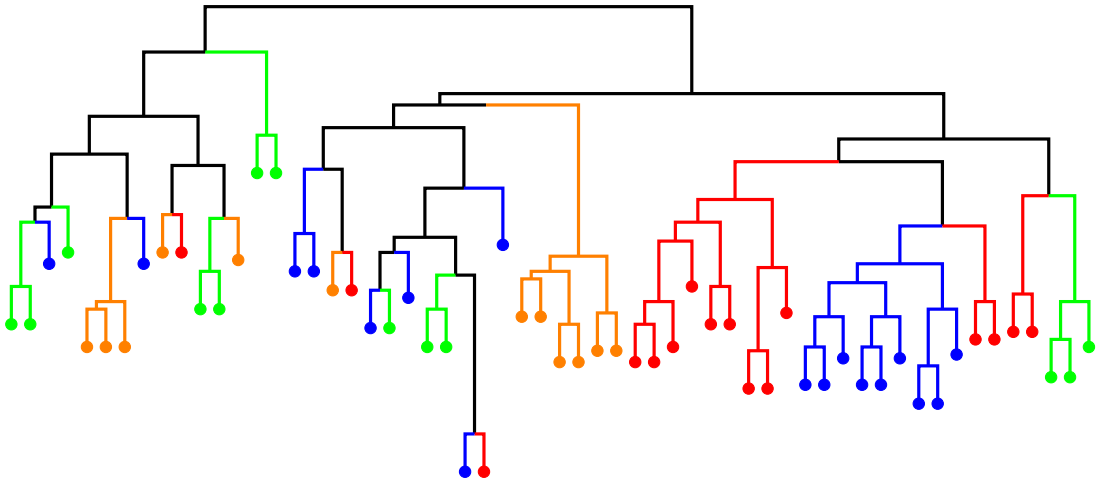


Figure 6.15.: **Hierarchical Clustering based on the amino acid composition over the full sequence** for the four *A. thaliana* positive sets. Red nodes indicate amino acid transporters, blue nodes oligopeptide transporters, green nodes phosphate transporters, and orange nodes hexose transporters. The longer an edge is in vertical direction, the higher is the distance between the connected clusters.

in blue) and hexose (shown in green) transporters. The phosphate transporter set was too heterogeneous to belong to same clusters. Again, the amino acid cluster is directly connected to the oligopeptide cluster. In comparison, Figure 6.16 shows the hierarchically clusters according to transmembrane regions. Here, the proteins of the oligopeptide and hexose transporter sets

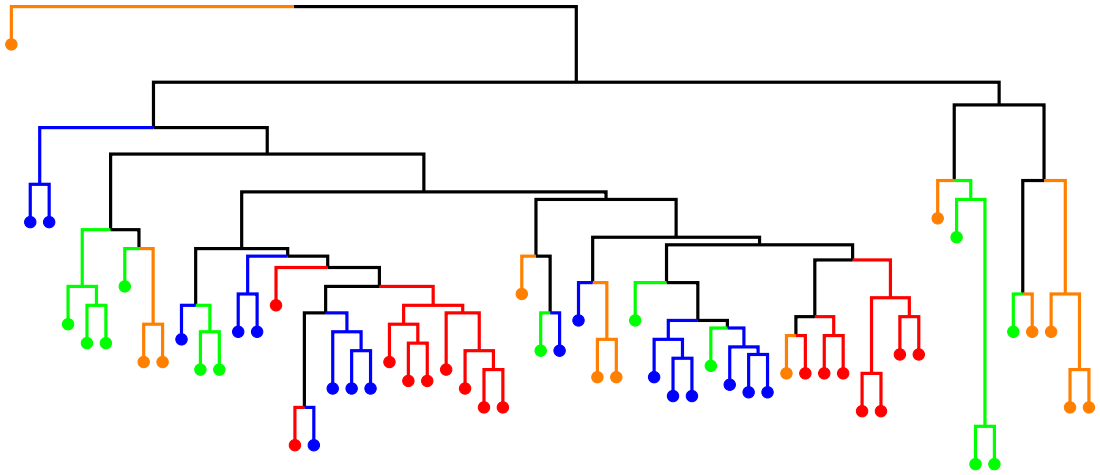


Figure 6.16.: **Hierarchical Clustering based on the amino acid composition in transmembrane positions** for the four *A. thaliana* positive sets. Red nodes indicate amino acid transporters, blue nodes oligopeptide transporters, green nodes phosphate transporters, and orange nodes hexose transporters. The longer an edge is in vertical direction, the higher is the distance between the connected clusters.

were more widespread. However, the amino acid transporter set was grouped into two clusters. One of those contains also four oligopeptide transporters and is directly linked to another two

oligopeptide transporters. The other cluster is related to an oligopeptide cluster. In general, the distances between proteins in a certain cluster or between two clusters are clearly smaller in the case of the transmembrane regions than over the full sequence. Clusters including different transporter types show a larger distance considering TMSs, whereas this effect is not that strongly pronounced considering the full sequence. Therefore, it seems that filtering into different sequence regions may be helpful to functionally cluster membrane proteins.

Now, the ranking method based on individual compositions, see Section 5.6, was applied for the TCDB and *A. thaliana* data sets. As shown in Tables 6.6 and 6.7, concentrating on the TMH in oAAC instead of considering the full sequence improved the prediction quality slightly. Indeed, the sensitivity was reduced using non–TMSs, but

1. the difference to the full sequence was only 0.052 on average and
2. on average the specificity was surprisingly the highest here of all analyzed features.

In general, a combining the AACs of the pure TMS and the purely non–TMS into a single vector reached the best sensitivity (69.4% on average) and accuracy (80.3% on average). A further inclusion of the oAAC over the full sequence could not increase the quality any further, although an accuracy of 79.1% on average was still a good result in comparison to 75.5% for the oAAC only over the full sequence. It seems that a TMS–nonTMS–all combination is able to neglect possible wrongly annotated TMSs.

As mentioned before, the specificity was expected to depend on the size of the positive set. Actually, the large sets H.3.A.1 and B.3.A.1 achieved lower specificity than the small sets. Since the TCDB sets were more heterogeneous than the *A. thaliana* sets in their function, their TMS structure, and their physicochemical properties, their average sensitivity was remarkably lower with 62.2% in comparison to 80.3%. According to the standard deviation of the TMS number in a TCDB set (see Figure 6.5), a stronger focusing on TMSs obviously improved the classification quality for H.2.A.1. Since the ratio between cumulative TMS length and total sequence length was quite small for H.1.A.1, H.3.A.1, and the β–barrel sets (see Figure 6.6), it was expected that considering only TMSs behaves rather poorly. Indeed, the sensitivity of H.3.A.1, B.2.A.1, and B.3.A.1 was lower in TMSs than over the full sequence and for B.1.A.1 the sensitivity was somehow increased whereas the specificity was strongly decreased. In contrast, for H.1.A.1 as well as for oligopeptide and phosphate transporters, considering only TMHs gave the highest sensitivity of all features. However, this was nevertheless expected since the relatively high standard deviation for H.1.A.1 resulted from two clusters as illustrated in Figure 6.7.

Further, as shown in Figure 6.10, the differences between different positive sets based on physicochemical properties was strongly pronounced in the transmembrane region for oligopeptide and phosphate transporters.

positive set	feature	sensitivity	specificity	accuracy
H.1.A.1	oAAC	0.476	0.914	0.695
	TMH	0.524	0.952	0.738
	nonTMH	0.524	0.962	0.743
	TMH–nonTMH	0.429	0.924	0.676
	TMH–nonTMH–oAAC	0.476	0.943	0.709
B.1.A.1	oAAC	0.444	0.978	0.711
	TMB	0.500	0.911	0.705
	nonTMB	0.444	0.967	0.705
	TMB–nonTMB	0.556	0.944	0.750
	TMB–nonTMB–oAAC	0.500	0.978	0.739
H.2.A.1	oAAC	0.614	0.914	0.765
	TMH	0.630	0.897	0.763
	nonTMH	0.390	0.911	0.644
	TMH–nonTMH	0.725	0.926	0.817
	TMH–nonTMH–oAAC	0.679	0.930	0.805
B.2.A.1	oAAC	0.818	0.900	0.859
	TMB	0.636	0.955	0.795
	nonTMB	0.636	0.900	0.768
	TMB–nonTMB	0.864	0.918	0.891
	TMB–nonTMB–oAAC	0.773	0.927	0.850
H.3.A.1	oAAC	0.471	0.688	0.583
	TMH	0.418	0.684	0.551
	nonTMH	0.556	0.736	0.646
	TMH–nonTMH	0.552	0.729	0.640
	TMH–nonTMH–oAAC	0.518	0.710	0.614
B.3.A.1	oAAC	0.523	0.807	0.664
	TMB	0.512	0.774	0.643
	nonTMB	0.514	0.804	0.658
	TMB–nonTMB	0.605	0.808	0.692
	TMB–nonTMB–oAAC	0.621	0.815	0.706
average – H.TCDB	oAAC	0.520	0.839	0.681
	TMH	0.524	0.844	0.684
	nonTMH	0.490	0.870	0.678
	TMH–nonTMH	0.569	0.860	0.711
	TMH–nonTMH–oAAC	0.558	0.861	0.709
average – B.TCDB	oAAC	0.595	0.895	0.745
	TMB	0.549	0.880	0.714
	nonTMB	0.531	0.890	0.710
	TMB–nonTMB	0.675	0.890	0.778
	TMB–nonTMB–oAAC	0.631	0.907	0.765
average – TCDB	oAAC	0.558	0.867	0.713
	TMS	0.537	0.862	0.699
	nonTMS	0.511	0.880	0.694
	TMS–nonTMS	0.622	0.875	0.744
	TMS–nonTMS–oAAC	0.595	0.884	0.737

Table 6.6.: **Quality of classification:** Sensitivity, specificities, and accuracy are given based on the amino acid compositions over different sequence regions for the TCDB positive sets. Red colored entries denote the highest measures for a data set. Here, the ranking method based on individual compositions was used.

positive set	feature	sensitivity	specificity	accuracy
amino acid	oAAC	0.813	0.938	0.875
	TMH	0.813	0.950	0.881
	nonTMH	0.750	1.000	0.875
	TMH–nonTMH	0.875	1.000	0.938
	TMH–nonTMH–oAAC	0.875	1.000	0.938
oligopeptide	oAAC	0.588	1.000	0.794
	TMH	0.824	0.894	0.859
	nonTMH	0.706	1.000	0.853
	TMH–nonTMH	0.824	1.000	0.912
	TMH–nonTMH–oAAC	0.765	1.000	0.882
phosphate	oAAC	0.667	0.836	0.750
	TMH	0.800	0.890	0.843
	nonTMH	0.467	0.917	0.689
	TMH–nonTMH	0.667	0.915	0.790
	TMH–nonTMH–oAAC	0.733	0.929	0.831
hexose	oAAC	0.769	0.938	0.854
	TMH	0.769	1.000	0.885
	nonTMH	0.769	1.000	0.885
	TMH–nonTMH	0.846	1.000	0.923
	TMH–nonTMH–oAAC	0.769	0.905	0.836
average – *A. thaliana*	oAAC	0.709	0.928	0.818
	TMH	0.802	0.934	0.867
	nonTMH	0.673	0.979	0.826
	TMH–nonTMH	0.803	0.979	0.891
	TMH–nonTMH–oAAC	0.786	0.959	0.872
average	oAAC	0.618	0.891	0.755
	TMS	0.643	0.891	0.766
	nonTMS	0.586	0.920	0.747
	TMS–nonTMS	0.694	0.916	0.803
	TMS–nonTMS–oAAC	0.671	0.914	0.791

Table 6.7.: **Quality of classification:** Sensitivity, specificities, and accuracy are given based on the amino acid compositions over different sequence regions for the *A. thaliana* positive sets. Red colored entries denote the highest measures for a data set. Additionally, average values of all TCDB sets (see Table 6.6) and all *A. thaliana* sets are shown. Here, the ranking method based on individual compositions was used.

size	sets	oAAC	TMS	nonTMS	TMS–nonTMS	TMS–nonTMS–oAAC
tiny	*A. thaliana* sets	0.221	0.186	0.216	0.270	0.201
small	H.1.A.1	0.185	0.184	0.190	0.213	0.198
	B.1.A.1, B.2.A.1	0.295	0.287	0.279	0.297	0.298
large	H.2.A.1	0.184	0.191	0.212	0.215	0.201
	B.3.A.1	0.352	0.375	0.354	0.400	0.408
huge	H.3.A.1	0.374	0.339	0.412	0.404	0.407

Table 6.8.: **Quality of random classification:** Mean sensitivity is given based on different filtered amino acid compositions averaged over 50 computations with different random sets. Entries in magenta represent a sensitivity $< 55\%$ of the corresponding positive set sensitivity, red $< 50\%$, orange $< 45\%$, green $< 40\%$, cyan $< 35\%$, and blue $< 30\%$. For example, the amino acid composition over TMSs of *Arabidopsis thaliana* sets reached $\frac{0.221}{0.802} = 0.276 < 30\%$ of the sensitivity for the corresponding positive set. Here, the ranking method based on individual compositions was used.

For validation purposes, the performance of randomly generated sets of different sizes compared to the actual positive sets were again analyzed and averaged over 50 different computations, see Table 6.8. For this, tiny random sets consisting of *A. thaliana* transporters represent the *A. thaliana* positive sets. Small random sets represent H.1.A.1, B.1.A.1, or B.2.A.1, large sets represent H.2.A.1 or B.3.A.1, and huge sets represent H.3.A.1. In general, the sensitivity is clearly lower for random than for actual positive sets. One can recognize a relation between the size of the random set and its sensitivity. The sensitivity values itself increase with increasing number of members in the random set, whereas the sensitivity of actual positive sets seems in contrast to be independent of their size. Therefore, the random sensitivity was less than 55% of the actual sensitivity for tiny and small sets and higher than 65% for large and huge sets except for H.2.A.1. The very successful behavior of H.2.A.1 is based on the previously discussed homogeneity in this set. Therefore, the ranking method applied here that is based on AACs focusing on different sequence regions was mainly advantageous for small positive sets. This is rather unusual comparing to standard techniques, such as SVMs or motif–based approaches that require really large data sets to achieve a satisfactory quality.

Whereas the AAC only over TMSs did not increase the sensitivity in comparison to the AAC over the full sequence for random sets, the joined vector TMS–nonTMS reached an improved sensitivity. Consequently, all features gave on average a random sensitivity of about 49% of the actual sensitivity except for nonTMS with a fraction of 53%. Considering only small sets and the homogeneous H.2.A.1, a filtering of TMSs performed slightly better than taking into account the full sequence and the joined TMS–nonTMS performed even slightly worse according to this fraction. Hence, the simple TMS based AAC seems to be a useful feature for functional classification.

Since the predictions can be strongly improved by PsePAAC and MSA–AAC, Table 6.9 shows a combination between this profile–based MSA–AAC and the concentration on a certain sequence

region for the *A. thaliana* positive sets. Filtering did not further improve the good performance

	full sequence	TMH	nonTMH	TMH–nonTMH	TMH–nonTMH–AAC
oAAC	0.709	0.802	0.673	0.803	0.786
MSA–AAC	0.861	0.832	0.817	0.811	0.846

Table 6.9.: **Quality of classification:** Sensitivity is given based on the original amino acid composition and the profile–based amino acid composition (MSA–AAC) in different sequence regions averaged over all *Arabidopsis thaliana* positive sets. Here, the ranking method based on individual compositions was used.

of the MSA–AAC over the full sequence. However, the MSA–AAC increased the sensitivity in all sequence regions compared to the oAAC. Here, the largest differences between oAAC and MSA–AAC were surprisingly found when considering the full sequence and non–TMSs. Generally, TMSs are clearly more conserved than non–TMSs such that the MSA–AAC differs strongly in non–transmembrane regions. This deviations may better represent all proteins in a positive set. In contrast, an enrichment of the oAAC in the transmembrane regions changes the vector only marginally such that the sensitivity was not that remarkably enhanced.

Since the ranking method based on an average search profile, see Section 5.6, performed better for homogeneous positive sets, it was used for the *A. thaliana* amino acid, oligopeptide, and hexose transporter sets to analyze further sequence regions. Table 6.10 shows the achieved sensitivities using MSA–AAC. In general, considering conserved regions gave worse results

	full sequence	buried	exposed	conserved	strong	total
amino acid	0.875	0.875	0.813	0.800	0.733	0.786
oligopeptide	0.882	0.941	0.706	0.706	0.800	0.556
hexose	0.769	0.846	0.692	0.385	0.375	0.583
average	0.842	0.887	0.737	0.630	0.636	0.642

Table 6.10.: **Quality of classification:** Sensitivity is given based on the profile-based amino acid composition (MSA–AAC) over different sequence regions for *A. thaliana* data sets. Red colored entries denote the highest sensitivity for a data set. Here, the ranking method based on an average search profile was used.

than considering the full sequence. Possibly, too many residues were ignored by these three filters. As expected, the concentration on buried positions reached a higher sensitivity than a concentration on exposed positions or considering the full sequence.

Filtering into varying sequence regions is closely related to Mem–EnsSAAC by Hayat *et al.* [64]. This approach is based on the split amino acid composition (SAAC), a 60 entries long vector which considers the AAC in three different sequence parts, namely 25 amino acids

of N termini, 25 amino acids of C termini, and residues in–between. They compared the SAAC with the PseAAC and a discrete wavelet analysis using an ensemble classifier with SVM, probabilistic neural network, k–nearest neighbor, Adaboost, and random forest. Whereas the discrete wavelet analysis reached a sensitivity of 67% and an accuracy of about 75% for a jackknife test, the PseAAC and SAAC performed clearly better with a sensitivity of 90% and 91% and an accuracy of about 90% and 92%, respectively. In comparison to the SAAC, it seems that the AAC in TMSs or TMS–nonTMS improve the sensitivity to a stronger extent.

6.8. Mapping non–Transmembrane Regions

Figure 6.17 shows the individual lengths of non–TMSs in the OPM positive sets 1.1.01 and 1.1.02. Whereas most TMSs of family 1.1.01 had a size less than 40 residues, the TMSs of family 1.1.02 differed strongly in their length. In the case of a classification based on the length of non–TMSs, this may be quite problematic. Similarly, a relatively high homogeneity

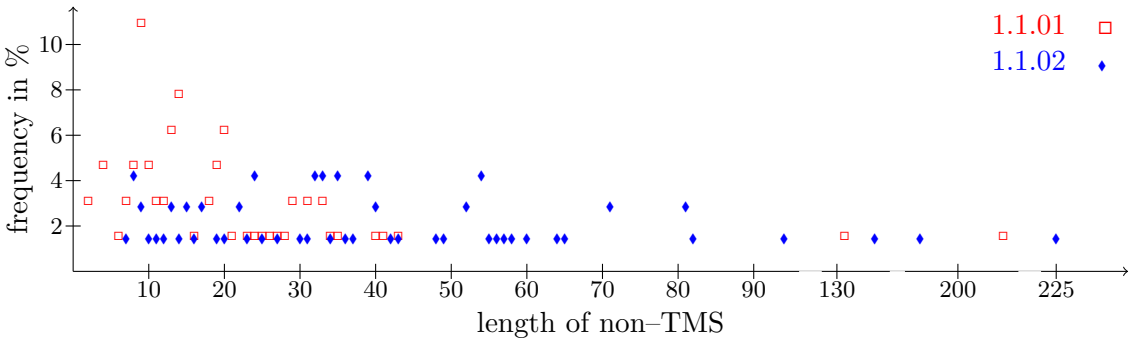


Figure 6.17.: **Lengths of individual non–transmembrane segments:** Shown is their distribution in the OPM positive sets 1.1.01 and 1.1.02.

in the number of TMSs is required. As shown in Figure 6.18, all members of 1.1.01 had seven TMHs. In contrast, the TMH number varies in the range from one to eleven in the positive set 1.1.02. A MBA is only well–defined and unique for the same number of TMSs in all considered proteins. To solve this problem of a different TMS number, MAFFT [81] was used to set up an initial alignment. Since MAFFT tries to match each character and not a certain box of the same characters in an optimal way, the initial alignment may be suboptimal in this case. MSAs often prefer mutations (mismatches) instead of insertions or deletions (gaps) depending on the alignment score. During the realigning step, those mismatches were replaced by gaps. However, the mapping of the TMS between all proteins remained unchanged.

Figure 6.19 shows the score values s_{gap} and s_{box} for the positive sets 1.1.01 and 1.1.02. Allowing less than $2q$ gaps for q aligned proteins in a certain box, the number of correctly mapped boxed was quite small. In the case of the more homogeneous set 1.1.01, one out of seven TMS boxes

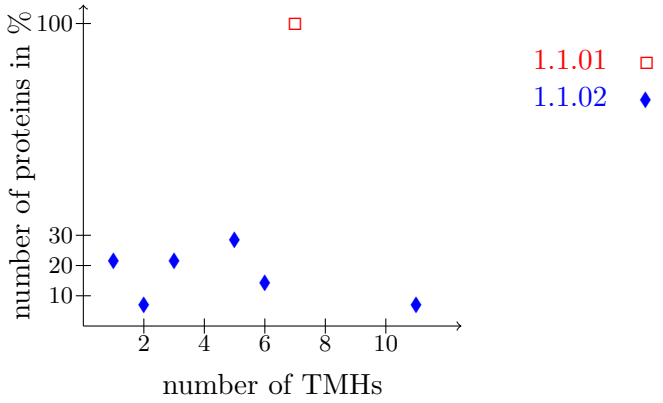


Figure 6.18.: **Transmembrane helix number distribution** of 1.1.01 and 1.1.02.

(P) was generally well matched. Boxes of non–TMSs (W) differed usually stronger in their length. The gap counting score is clearly lower for family 1.1.01 than 1.1.02 because of its dependence of sequence length and number of TMSs. To classify an unknown protein, it is assumed that the score is higher for an alignment with $q+1$ proteins than with q proteins. The LOOCV (denoted as 1.1.01- in Figure 6.19) showed an only slightly smaller score s_{gap} on

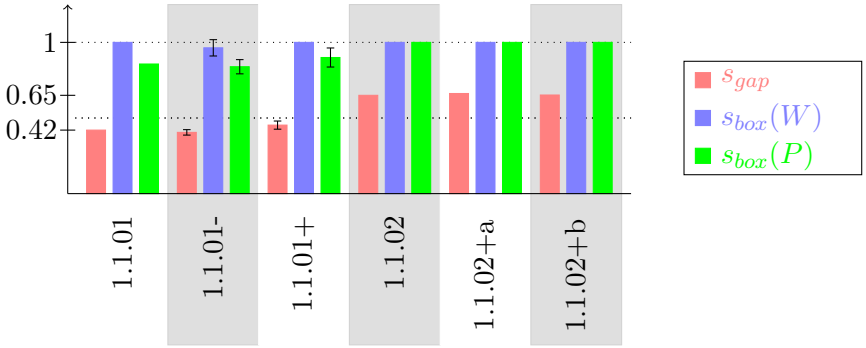


Figure 6.19.: **Values of different scores** s_{gap}, $s_{box}(W)$, and $s_{box}(P)$ for different subsets or supersets of the OPM positive sets 1.1.01 and 1.1.02. 1.1.01- represents the average over all possible sets without a member, 1.1.01+ the average over all possible sets that further include a member of 1.1.02, as well as 1.1.02+a and 1.1.02+b sets with a certain member of 1.1.01. In the case of average scores, the standard deviations are given.

average. Therefore, it could be possible to group a protein to a certain family when the score of the alignment contains the whole family and the considered protein differed from the score of the family alignment in the same range as the LOOCV. Obviously, the average score for joined alignments of family 1.1.01 with a single protein of family 1.1.02 (denoted as 1.1.01+ in Figure 6.19) was higher than that for the family alignment. The distance between those scores was smaller for one protein and equal for two proteins of 1.1.02 compared to the score difference of the LOOCV. It seems that the length of non–TMSs is not related to the protein family. However, those proteins also differed clearly in their number of TMSs and sequence

length. Figure 6.19 further presents the scores of two alignments considering the whole family 1.1.02 together with a single protein of family 1.1.01 (denoted as 1.1.02+). For this, the two proteins with the largest deviation of score from family 1.1.01 were chosen. Both proteins increased the score of family 1.1.02 only marginally. Therefore, it is not possible to classify a protein to that family with the smallest deviation.

In 2001, Otaki and Firestein reported the functional importance of loop lengths [125]. They statistically analyzed and classified GPCRs into three functional subclasses based on the loop length. Advantageously, all GPCRs consists of seven TMHs. In 2003, Sugiyama *et al.* used the number of TMSs and the loop length to group transmembrane proteins into functional classes [182]. For this, they introduced a binary topology pattern that contains a zero for a loop which is shorter than a certain threshold and a one otherwise. However, they only compared proteins with the same number of TMSs.

7. Summary

Due to the great importance of membrane proteins and the often missing experimental annotation of their sequences, proteins from the Transporter Classification Database (TCDB) or *Arabidopsis thaliana* transporters and carriers were analyzed in detail and functionally classified with several computational, sequence–based procedures. The four substrate classes amino acids, oligopeptides, phosphates, and hexoses were studied for the *Arabidopsis thaliana* data set and the three subfamilies 1.A.1, 2.A.1, and 3.A.1 for the TCDB set.

The studies showed that amino acid compositions (AACs) between functionally related proteins belonging to the same substrate group were more similar to each other than between proteins sharing the same three–dimensional structure. In this way, a clear boundary between different transporter types was found based on the AAC and on frequencies of physicochemical categories, e.g., an analysis of variance detected several categories as significantly different. In contrast, a homology search by BLAST did not yield satisfactory results for our data sets due to too large E–values between proteins of the same substrate class and different TC family. Therefore, the AAC was chosen as suitable feature for functional annotation of membrane transporters.

Assuming that a substrate has similar characteristics as transmembrane segments (TMSs) due to required interactions in the transport process, filtering between transmembrane and non–transmembrane regions can be expected to improve functional classification. Indeed, the Wilcoxon–Mann–Whitney test detected more and stronger significant differences between different transporter types based on physicochemical categories when including only TMSs than when including the full sequence.

Since the data sets were too small to be used as input for a support vector machine, a ranking method either based on an average search profile or based on individual compositions was used to classify the proteins. For this, several AAC based features integrating physicochemical properties, conservation levels, and sequence order information were considered to quantify similarities in the data set and to generate ranked similarity lists using the Euclidean distance. Those AACs led all to relatively high quality measures, e.g., the original AAC gave a sensitivity and an accuracy of at least 75% for the *Arabidopsis thaliana* positive sets, respectively. The performance of the substrate prediction was enhanced up to sensitivities and accuracies higher than 80% by including further characteristics into the AAC. Randomly constructed positive sets

showed accuracies of less than 60% and sensitivities of 30–35%. Considering the sensitivities and accuracies of individual sets, the PsePAAC proved best. However, a comparison with random classifications indicated that overall the profile–based MSA–AAC yielded the best results.

Hierarchical clustering was applied to compare between AACs over the full sequence and over TMSs. In general, the clusters were not that strongly pronounced. The phosphate transporters were relatively widespread. Nevertheless, functional clustering worked better in the case of regarding only TMSs. Further, a principal component analysis based on the AAC in transmembrane positions was not able to discriminate between amino acid and oligopeptide transporters. Therefore, the ranking method was again used for classification according to filtering into different sequence regions. The prediction quality was clearly dependent on the used sequence regions. The AAC over the full sequence gave an accuracy of 76% on average, whereas the joined AAC that separately includes TMSs and non–TMSs achieved a slightly higher average accuracy of 80%. Obviously, the improvements were larger for sets that are homogeneous in their TMS distribution. Comparing the quality measures of the actual positive sets with those of randomly generated sets, the AAC based solely on TMSs performed best. In general, the sensitivities of very small actual sets were about three times higher than that of very small random sets.

Due to the high sensitivities, the ranking procedure combining several AACs over different sequence regions could help to identify unknown proteins of a certain substrate class and to narrow down the huge search space of experiments that determine the substrate specificities of a certain transporter.

It seems that loop length patterns improve the prediction quality only for data sets that are homogeneous in their TMS numbers.

Part III.

Quorum Sensing in *Pseudomonas aeruginosa*

Unfortunately, there are currently many questions open regarding the Quorum sensing in *Pseudomonas aeruginosa* and in particular regarding the *pqs* system. Since Quorum sensing is responsible for producing virulence factors in the multi–resistant human pathogen *Pseudomonas aeruginosa*, the signaling pathways were modeled. Highly selective Quorum sensing inhibitors without impact on bacterial viability delay are assumed to avoid resistance in comparison to targeting central metabolism or DNA replication [69, 146]. Therefore, we set up a rule–based regulatory and metabolic model to study the effect of added enzyme inhibitors and receptor antagonists. The results of this chapter were recently submitted to *BMC Systems Biology* [160]. Further, the influence of randomly occurring mutations on virulence factor formation was analyzed.

8. Computational Model

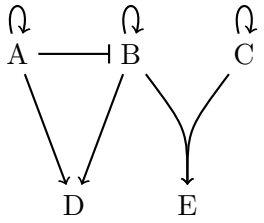
To analyze the Quorum sensing in *Pseudomonas aeruginosa*, we wanted to apply a very robust method that generates easily interpretable results. Due to the lack of rate constants and other experimental data for the involved association and dissociation processes as well as enzymatic reactions, the approach is required to be as independent of parameters as possible. Therefore, we decided to use a qualitative logical formalism.

8.1. Logical Formalism

Logical formalisms, such as Boolean networks, are commonly applied to purely gene regulatory networks, but they are also useful for combined regulatory and metabolic systems including the three biological layers genes, proteins, and metabolites. Complex chemical reactions with often unknown rate constants are described by simple mathematical rules, as, e.g., "If A is present, then D will be produced". Then, a considered network consists of those logical functions (the edges in the graph) and a given number of species (the nodes in the graph) representing the genes, proteins, or metabolites. Instead of a continuous concentration, a finite number of discrete states x_i determines the level of a species i. The progressing time t during the simulation is realized as discrete propagation steps. The dependencies containing logical functions may be either converted into condition tables or a weighting matrix $\mathbf{W} = (w_{ij})$ [82, 116]. Thus, the next propagation state $\mathbf{S}(t+1)$ at the next time step $t+1$ is determined by those functions and the current system state $\mathbf{S}(t)$ at time t.

In a synchronous updating scheme, the states of all species are simultaneously changed [82]. However, not all reactions in the considered network happen naturally at exactly the same time. In contrast, the nodes are updated one after the other in a chosen order in the case of the so–called asynchronous updating scheme [185]. Here, the behavior of the system is highly dependent on the chosen order. In a semi–synchronous scheme, groups of nodes are updated one after the other in a chosen order, where all nodes in a certain group are updated simultaneously [116]. To cluster the nodes into different groups and to define the order of these groups, still further experimental data are required.

For example, Figure 8.1 shows a simple network with the five Boolean variables A–E and the corresponding condition tables. Here, species E will only be produced when B and C are both



Logical rules:
B: $\neg A \wedge B$
D: $A \vee B$
E: $B \wedge C$

$A(t+1)$	$A(t)$
1	1
0	0

$B(t+1)$	$A(t)$	$B(t)$
0	0	0
1	0	1
0	1	0
0	1	1

$C(t+1)$	$C(t)$
1	1
0	0

$D(t+1)$	$A(t)$	$B(t)$
0	0	0
1	0	1
1	1	0
1	1	1

$E(t+1)$	$B(t)$	$C(t)$
0	0	0
0	0	1
0	1	0
1	1	1

Figure 8.1.: **Example for a Boolean network** with five Boolean variables A–E and the corresponding condition tables.

present together. In the case of a synchronous updating scheme, state

$$\mathbf{S}(t) = \{A(t), B(t), C(t), D(t), E(t)\} = \{1,1,1,0,0\}$$

yields the next state

$$\mathbf{S}(t+1) = \{A(t+1), B(t+1), C(t+1), D(t+1), E(t+1)\} = \{1,0,1,1,1\}.$$

However, when being considered by an asynchronous updating scheme with the order A, B, C, D, and E, the next state is defined as:

$$\mathbf{S}(t+1) = \{A(t+1), B(t+1), C(t+1), D(t+1), E(t+1)\} = \{1,0,1,1,0\},$$

whereas the order E, D, C, B, and A ends up in the same state $\mathbf{S}(t+1)$ as for the synchronous updating scheme.

For each system state $\mathbf{S}(t)$, the following state $\mathbf{S}(t+1)$ is well–defined and unique such that Boolean models show a deterministic dynamics. Due to the finite number of system states 2^n for a network with n species with two states (0, 1) each and the deterministic dynamics, a simulation results in periodic trajectories reaching an already visited state. A periodic sequence of system states (i.e. an ordered list of consecutive states where the last state runs again into the first state of this sequence) is called attractor or orbit. All states ending in a certain at-

tractor are named as basins of attraction. The length of an attractor is denoted as its period. Usually, one studies the oscillations in the attractors and how likely the system runs into a certain orbit.

8.1.1. Boolean Network with Weighted Interactions

In a classical Boolean network, the possible states are zero (off, inactive) and one (on, active at maximum rate). Then, the state $\mathbf{S}(t)$ of the network at time t is defined as given in equation (8.1),

$$\mathbf{S}(t) = \{x_1(t), x_2(t), ..., x_n(t)\} \qquad \forall x_i(t) \in \{0,1\} \tag{8.1}$$

where n is the number of nodes in the network and x the state of a certain species. To combine the discrete states of the logical formalism with relative interaction weights, Mendoza *et al.* implemented the following equation [116].

$$x_i(t+1) = L\left(\sum_{j=1}^{n} w_{ij} x_j(t) - \gamma_i\right) \tag{8.2}$$

The entries w_{ij} of the weighting matrix are positive integers for activations and negative integers for inhibitions. For simplicity, the threshold $\gamma \in \mathbb{Z}$ is also an integer. To keep the states as Boolean, the heavyside step function L is used as defined in equation (8.3),

$$L(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{else} \end{cases}. \tag{8.3}$$

Whereas the weights and thresholds are further required parameters, they allow to include relative strengths of connections found in experimental data.

8.2. Quorum Sensing as Boolean Network

In order to better understand the behavior of the positive feedback loop in the Quorum sensing of *P. aeruginosa*, a Boolean formalism with a synchronous updating scheme was used to analyze a simplified network of the internal pathways in a single cell. Based on the main important nodes in the pathway illustration of Figure 3.5 in Section 3.3, we set up the small network that is shown in Figure 8.2 as a Boolean diagram.

At first, the receptor proteins LasR, RhlR, and PqsR are assumed to be always active. Since the complex between LasR and AI–2 (C4) has only a weak effect on the concentration of AI–1, it was not considered here. Due to the preference of PqsR to bind PQS instead of HHQ, a complex between PqsR and HHQ (C5) was not added, here. Instead of distinguishing between different virulence factors and their formation pathways, a general species called virulence was

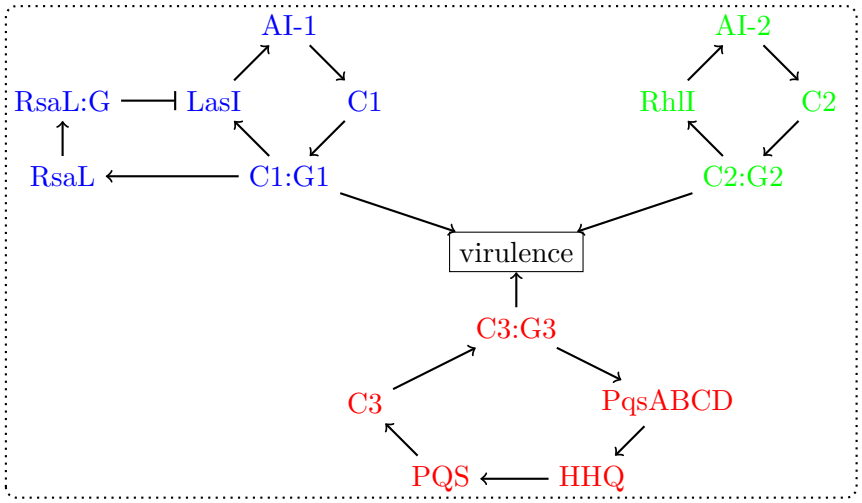


Figure 8.2.: **Simplified network** of the three Quorum sensing systems *las* (given in blue), *rhl* (in green), and *pqs* (in red) in Boolean topology. The receptors LasR, RhlR, and PqsR are assumed as active the whole time.

introduced. In the *pqs* system (shown in red), the enzymes PqsA, PqsB, PqsC, and PqsD were combined to a single species. Then, only the autoinducers AI–1, AI–2, HHQ, and PQS (AI–3), the synthases LasI, RhlI, and PqsABCD, which trigger the autoinducer production, the complexes C1, C2, and C3, which are activated by the autoinducers, as well as their binding to the operons C1:G1, C2:G2, and C3:G3 were considered. For the *las* system (shown in blue), the LasI inhibitor RsaL activated by C1:G1 was additionally included. Thus, the simplified network contains three independent positive feedback loops for every Quorum sensing system.

	AI–g	HHQ	Cg	C1:G1	Cm:Gm	Yh	Y3	RsaL	RsaL:G	V
AI–h	0	0	0	0	0	1	0	0	0	0
HHQ	0	0	0	0	0	0	1	0	0	0
AI–3	0	1	0	0	0	0	0	0	0	0
Cg	1	0	0	0	0	0	0	0	0	0
Cg:Gg	0	0	1	0	0	0	0	0	0	0
Y1	0	0	0	1	0	0	0	0	-2	0
Ym	0	0	0	0	1	0	0	0	0	0
RsaL	0	0	0	1	0	0	0	0	0	0
RsaL:G	0	0	0	0	0	0	0	1	0	0
V	0	0	0	1	1	0	0	0	0	0

Table 8.1.: **Weighting matrix of Boolean model** where Y represents either the autoinducer synthases LasI, RhlI, or PqsABCD and V the virulence factors. g is one (*las* system), two (*rhl* system), or three (*pqs* system) with AI–3 = PQS. h is either one or two and m either two or three. The species belonging to the *las* system are shown in blue, *rhl* system in green, and *pqs* system in red.

The Boolean representation was realized as given in equation (8.2) [116] without thresholds ($\forall i : \gamma_i = 0$) and a corresponding weighting matrix as shown in Table 8.1. All activations were equally weighted and an inhibition was two times stronger than an activation. Therefore, inhibiting relations typically compensate all activating relations of a certain node. In this model, nodes that are not explicitly activated are inactive in the next time step in contrast to some classical Boolean approaches [97].

8.3. Extended multi–level Logical Formalism

In this work, we wanted to explicitly analyze the three hierarchically layered Quorum sensing systems that are organized in a combined regulatory and metabolic network. Different types of species belong to this network, namely genes, proteins, and cells. Due to the critical simplifications that had to be made when setting up the model, a classical Boolean approach was not applicable. Therefore, an extended logical formalism with discrete multi–level variables was applied, where every species may take on another fixed number of possible states.

Still, we restricted the level of detail to a mostly parameter–free minimal system to maintain the robustness and interpretability. For this, many species were still designed as Boolean nodes with the two states “on” or “off”. Only a few species were treated as multi–level nodes with more than two possible states when this was essential for the correct functioning of the network. Thus, our extension differs from other multi–level logical approaches that consider for each species the same number of possible states [49].

In our model, the propagation of states was based on logical functions and a synchronous updating scheme with discrete time steps. Due to the missing time scale, all reactions are implemented at the same rate. Here, a certain time step contains either a relatively fast enzymatic catalysis or a much slower, complete protein biosynthesis including transcription and translation. Enzymatic reactions often take place in the microsecond to millisecond range, whereas gene regulations are in the range of seconds to minutes [29]. To entangle possibly occurring critical behaviors in the case of reactions with strongly different reaction time scales or with further conditions, multi–level nodes are required at some points of the network.

Further, the model was expanded to treat multiple cells that work independently from each other with either the same or a different corresponding network, see Figure 8.3. Those cells were realized as a growing culture in which single cells are able to mutate. An exchange between different cells was included by transport processes of the autoinducers from a single cell to their environment or from the environment to a certain cell. For this, all cells were located in the same environment, i.e., they all share the same external autoinducer concentrations.

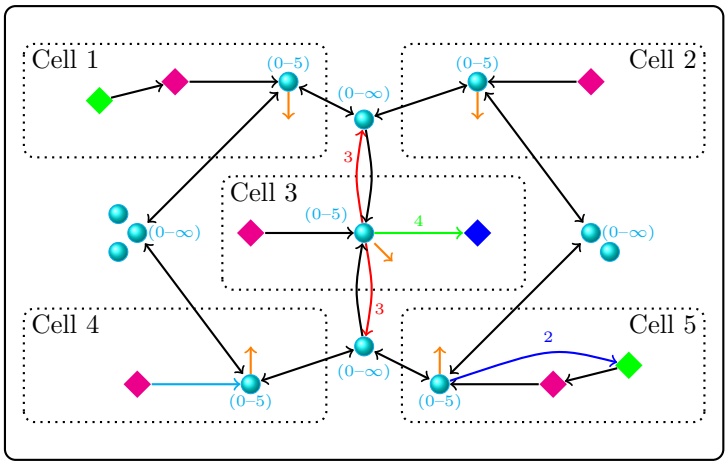


Figure 8.3.: **Multiple cells in a common environment** with an exchange via autoinducer transport. The colored squares denote Boolean logical variables and the cyan colored balls are autoinducers modeled as multi–level logical variables with an infinite (∞) number of levels in the environment and six levels inside of a cell. Black, blue, red, or green colored arrows characterize activations for which the autoinducer level must reach a certain threshold. Cyan colored edges denote randomly occurring activations and orange colored edges infrequently occurring degradations.

Figure 8.4 illustrates the important steps in the work flow. Starting with a fixed number of cells, each with a given network, the system state is updated iteratively according to the mathematical rules in the generalized logical approach and the transport processes. At each time step, growth and random mutation processes are possible. The simulation is stopped after a certain number t_{max} of time steps.

8.3.1. Propagation in a Single Cell

As already mentioned, all states independent of their maximal possible level $\mathcal{M}$ are modified together at the beginning of each time step. Based on the states of all connected species and the corresponding interaction type, equation (8.4) defines the next state $x_i(t+1)$ of node i.

$$x_i(t+1) = M\left(B\left(\left(\sum_{j=1}^{n} w_{ij} K(x_j, \epsilon_{ij})(t)\right)\right) + G(x_i(t), i), i\right) \tag{8.4}$$

In this way, it is measured how many levels are produced and added to the previous level considering the number of possible reactions. For this, the weighting matrix $\mathbf{W} = (w_{ij})$ includes the connections between all species. According to our model all activations are weighted equally and any deactivation should have a stronger effect than the sum of all possible activations. This was realized by setting each activating relation from node j to node i as $\frac{1}{a}$ in entry w_{ij} and all inhibiting edges as $-k$. Here, $a-1$ denotes the number of nodes that are in common with node j necessary to produce species i (logical "and"–relation) and k is the number of activations for node i (logical "or"–relations). When there is no connection between node i and j, w_{ij} is zero.

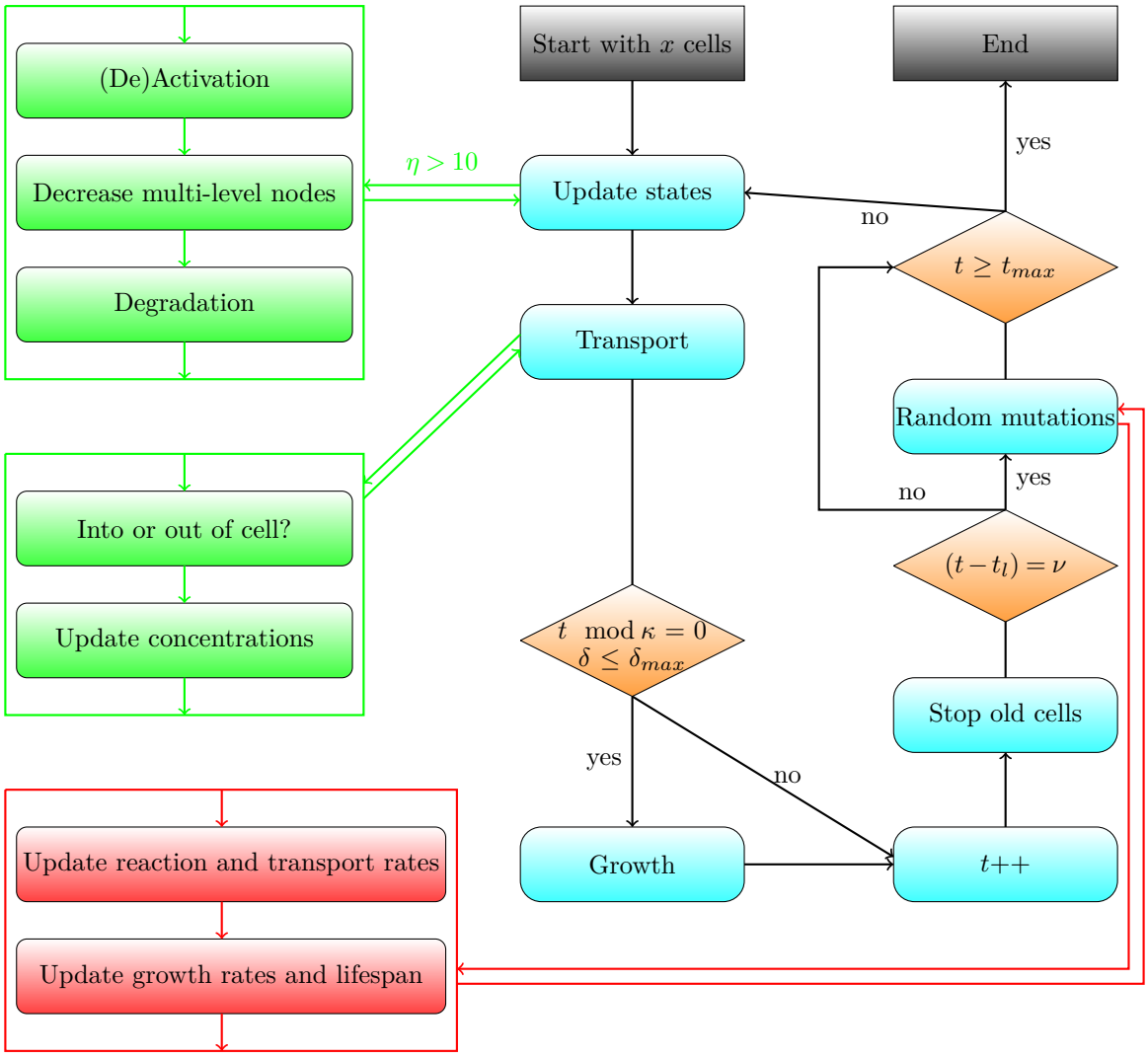


Figure 8.4.: **Work flow:** t_{max} iterations are performed that update the system state in each cell with a life time $\eta > 10$, transport of autoinducers, growth of cell culture, and random mutations. t represents the current time step, t_l the last time when a mutation occurred, κ the growth rate, δ the number of cell divisions, and ν the mutation rate.

A reaction can only take place when the states of all involved activating or inhibiting nodes exceed a certain threshold ϵ. This threshold is different for every reaction, e.g., two of species A are required to produce D, but three of A to form C, and one of D is enough to activate C. To fulfill this requirement, the heavyside step function K,

$$K(x, \epsilon) = \begin{cases} 1 & \text{if } x \geq \epsilon \\ 0 & \text{else} \end{cases}, \tag{8.5}$$

ensures for all nodes j that are connected to node i that their respective state $x_j(t)$ is equal or larger than this threshold ϵ_{ij} at the current time point t. Usually, ϵ_{ij} equals one. These cases are later denoted by black edges between nodes i and j. Often, those species are Boolean nodes.

In the case of multi–level nodes, larger thresholds are necessary. In all network representations below, an ϵ of two is symbolized by blue colored edges, an ϵ of three by red colored edges, and an ϵ of four by green colored edges.

Due to allowing for rational numbers in the weighting matrix, it is required to round the resulting sum down to an integer, included in function B,

$$B(x) = \begin{cases} \lfloor x \rfloor & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}. \tag{8.6}$$

This ensures that a reaction with several reactants ("and"–relation) takes only place when the level of all reactants is high enough. For example, let us assume that species A_1 is activated by species A_2 "and" A_3 and the sum over the weights and the current states of A_2 and A_3 is 0.5. Then, either A_2 and A_3 are deactivated such that this reaction is not possible. Here, function B rounds the value to zero. At this point, the equation is simplified because no species were considered that are formed by two "and"–relations that are linked with an "or"–relation $((A \wedge C) \vee (D \wedge E))$.

Since function B detects only the number of newly formed metabolites, the previous level $x_i(t)$ must be added for a multi–level node i. In contrast, Boolean nodes should be only active when they were explicitly activated. Therefore, function G,

$$G(x,i) = \begin{cases} x & \text{if } \mathcal{M}_i > 1 \\ 0 & \text{else} \end{cases}, \tag{8.7}$$

distinguishes between multi–level and Boolean variables.

To avoid levels higher than the maximal possible level $\mathcal{M}_i$ of a node i, function M is used:

$$M(x,i) = \begin{cases} x & \text{if } \mathcal{M}_i \geq x \\ \mathcal{M}_i & \text{else} \end{cases}. \tag{8.8}$$

In the case of Boolean nodes, the state x_i of node i at time $t+1$ is independent from the number of performed reactions at time t that used this species. In contrast, the state of a multi–level variable must be decreased by the count of reactions that take place during this time step and consumed this metabolite. For example, let us say that a multi–level node A forms species C and in another reaction species D. Then, it is necessary to reduce the level of A by two.

Some reactions may be much slower than the other ones or may occur infrequently. In those cases, multi–level variables may not able to avoid the critical behavior of a synchronous updating scheme. Therefore, the formalism was extended by random reactions that occur with a

certain probability. This probability is called conversion frequency ζ. Such reactions are later colored in cyan in network representations.

Since an explicit activation to a level higher than zero is not required for multi–level nodes, all molecules of this species exist after their formation until the end of the simulation t_{max}. Thus, all nodes without any out–going edge remain at their maximum levels after they have reached them. However, in a natural cell, molecules are typically degraded after a particular time period. Further, those species may be involved in reactions that are not included in the analyzed regulatory network. Hence, we also included degradation reactions in this logical formalism. Here, the state of multi–level nodes is decreased by one after a certain number of time steps, whereby this number is chosen as the same for each multi–level variable independent of its maximal possible level. The inversed value of this number of time steps is called degradation frequency ϖ. A degradation step takes place after updating the states based on activations and deactivations such that at this time step no multi–level node reaches its maximum level.

To interpret the predicted level of a certain species i averaged over a time interval from t_1 to t_2, the average theoretical maximum level Υ is calculated as given in equation (8.9),

$$\Upsilon = \frac{\mathcal{M}_i\Big(t_2 - t_1 - \mathfrak{d}\Big) + \Big(\mathcal{M}_i - 1\Big)\mathfrak{d}}{t_2 - t_1}, \tag{8.9}$$

whereby $\mathfrak{d}$ denotes the number of degradations in the considered time interval. It is estimated in the following way:

$$\mathfrak{d} = \varpi\Big(t_2 - t_1\Big) \tag{8.10}$$

8.3.2. Transport Processes

In Quorum sensing, bacteria communicate with each other via the exchange of small signaling molecules, the autoinducers. For this reason, a simple transport process was implemented where we do not distinguish between the different forms of transport mechanisms, such as diffusion, usage of passive carriers (e.g. efflux pumps), or active transport. After updating the system state in a single cell, a transport process takes place at each time step. Transport reactions, later drawn as dotted arrows, are not considered during the propagation steps in a single cell. At first, it is determined which autoinducer should be transported into or out of which cell and then, the internal and external levels are updated. This decision is synchronously performed for all cells. According to the level of autoinducer molecules inside and outside of a cell it is decided whether the autoinducer is transported into or out of the cell, see Figure 8.5. If the internal level (x_{in}) of an autoinducer in a considered cell reaches at least the required transport threshold χ and is not larger than the corresponding external level relative to the number of cells in the environment, then the autoinducer is transported out of the cell. In the remaining cases, the autoinducer is transported into the cell if there is no autoinducer in the environment and there is no transport if the external concentration is not zero.

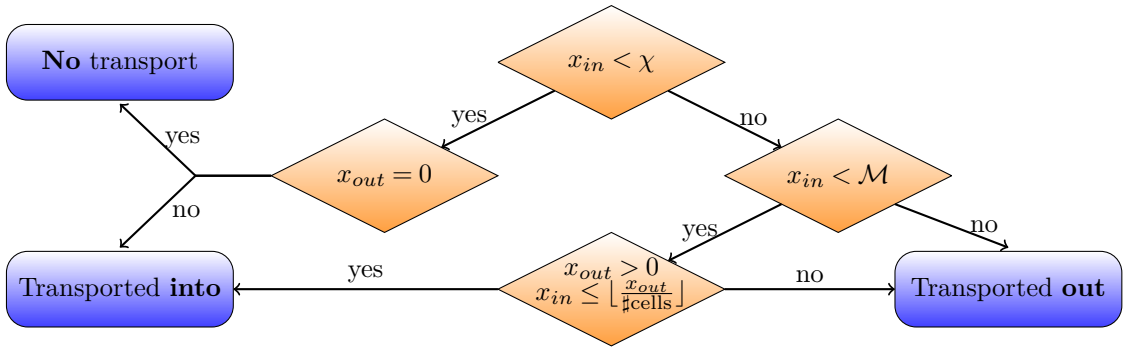


Figure 8.5.: **Transport decision:** Depending on the external level x_{out} and the internal level x_{in} of a certain autoinducer, it is decided whether the autoinducer is transported into or out of a cell. Hereby, χ denotes the transport threshold, $\mathcal{M}_i$ the maximal possible level, and $\sharp$ cells the number of cells in the environment.

8.3.3. Growth Processes

The starting cells characterize a biological cell culture in the last stages of the exponential growth phase. Here, the exponential growth is modeled as a certain number of cell divisions δ_{max}. In a growth process as illustrated in Figure 8.6, after a certain number of time steps, a mother cell c_m produces a new cell c_c. Hereby, the daughter cell c_c inherits its network, its

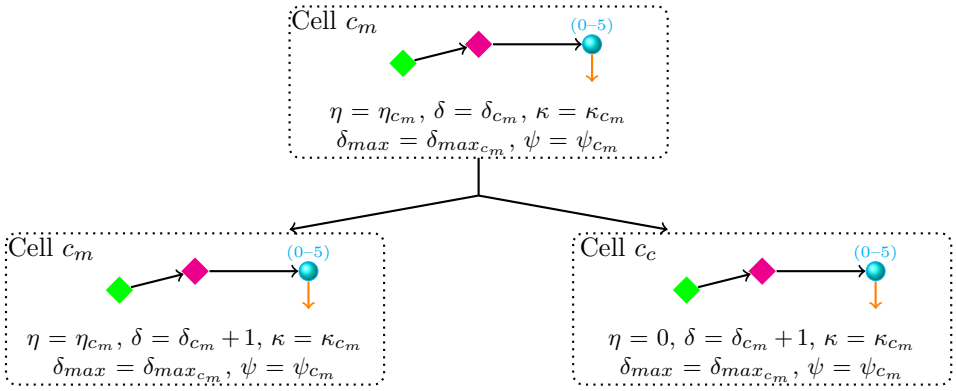


Figure 8.6.: **Growth:** A mother cell c_m produces a new daughter cell c_c with the same network, same growth rate κ, same number of maximum cell divisions δ_{max}, and same mortality rate ψ. The life rate η is unchanged for cell c_m and set to zero for cell c_c. The number of already happened cell divisions δ is increased by one for both cells.

growth rate κ_{c_m}, the number of maximal possible cell divisions $\delta_{max_{c_m}}$, and its mortality rate ψ_{c_m} from the mother cell. The growth rate κ is the inversed number of time steps between two growth processes. The first system state of the new cell is initialized as in the beginning of the mother cell. The system state and the life time η_{c_m} of the mother cell are unchanged. The life time η_{c_c} of the daughter cell is set to zero. For both cells, the number of already

occurred divisions δ is one higher than the previous number δ_{c_m} of the mother cell. The new cell starts with ten steps during which the systems states are not updated based on the logical rules. During this delay, only transport processes are allowed.

The life time η of each cell is increased by one when the next time step starts. If the life time reaches the mortality rate ψ of a cell, the propagation of this cell is stopped. After the exponential growth phase ($\delta > \delta_{max}$), the growth rate of the whole cell culture is equal to the mortality rate of the whole cell culture (stationary phase). In the case of a mortality rate $\psi = t_{max} + 1$, no new cell is produced. Otherwise, during the growth process it is counted how many cells are stopped after the last cell division until the current time step. The same number of cells are randomly selected and divided afterwards. When the number of new cells is smaller than the number of stopped cells the death phase of the culture starts.

Hereby, each cell may have its own parameters δ_{max}, κ, and ψ. Therefore, a growth phase refers to a subculture with the same parameters.

8.3.4. Random Mutations

In a further variant of our model, random mutations in the network and in the parameters of a cell are possible at the end of each time step. Then, it is tested for every cell whether a mutation should take place depending on the time point of the last mutation t_l and on the mutation rate ν:

$$\textbf{if } (t - t_l) \mod \nu = 0 \textbf{ then} \text{ there is a mutation.}$$

This mutation rate is the same for all cells. Figure 8.7 demonstrates the possible consequences of a mutation. At first, an edge of the network is chosen randomly for which the frequency

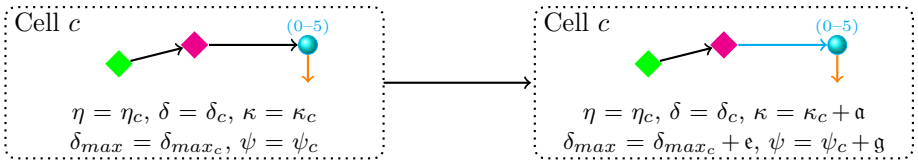


Figure 8.7.: **Random mutations:** As a consequence of a mutation, the growth rate κ, the number of cell divisions δ, the mortality rate ψ, and the probability how often a certain reaction happens may be changed. Here, $\mathfrak{a}$, $\mathfrak{e}$, and $\mathfrak{g}$ indicate some constants.

of occurrence is changed. For this, a randomly selected value in the range $\{-10, -5, 0, 5, 10\}$ is added to this probability. Since naturally occurring mutants often show a decreased fitness compared to the wild type, the parameters δ_{max}, κ, and ψ may be also influenced, in the

following way:

$$\kappa_m = \begin{cases} \kappa_o + \mathfrak{a} & \text{if } \kappa_o + \mathfrak{a} > 0 \\ 0 & \text{else} \end{cases} \qquad \mathfrak{a} \in \{-10, -9, \cdots, 9, 10\} \tag{8.11}$$

$$\delta_{max_m} = \begin{cases} \delta_{max_o} + \mathfrak{e} & \text{if } \delta_{max_o} + \mathfrak{e} > 0 \\ 0 & \text{else} \end{cases} \qquad \mathfrak{e} \in \{-1, 0, 1\} \tag{8.12}$$

$$\psi_m = \begin{cases} \psi_o + \mathfrak{g} & \text{if } \psi_o + \mathfrak{g} > 0 \\ 0 & \text{else} \end{cases} \qquad \mathfrak{g} \in \{-10, -9, \cdots, 9, 10\} \tag{8.13}$$

whereby κ_m, δ_{max_m}, and ψ_m are the mutated parameters and κ_o, δ_{max_o}, and ψ_o the previous ones. The constants $\mathfrak{a}$, $\mathfrak{e}$, and $\mathfrak{g}$ are Gaussian distributed random values. Hereby, the distribution is defined by the mean μ and the standard deviation σ over the domain of the parameters. The Gaussian distribution ensures that parameter values around zero are more frequently taken than values at the boundaries of the domain. As the simulation time t_{max} remains the same, the exponential phase is very short for small κ and δ_{max} and long for large values. Obviously, the life time η and the number of cell divisions δ are kept the same in a mutation.

8.3.5. Implementation

I implemented the extended multi–level logical formalism in Java. The program requires the network as a ".txt" file. It is further possible to change some default parameters. The output is a ".txt" file containing the average levels of each species in the considered network. More information is given in appendix Section B.3.

Random Numbers

As mentioned before, the formation of PQS was stochastically modeled, died cells are randomly chosen, and the changed edge is randomly selected during a mutation process. Therefore, random numbers are required during the whole propagation. For this, the function `Math.random()` in Java is used. Further, the conversion frequency, the growth rate, the number of maximal possible cell divisions, and the mortality rate are updated based on Gaussian distributed random numbers in a mutation process. Here, the function `Random.nextGaussian()` in Java is applied. For statistical reasons, random processes demand multiple repeats of the simulation with different random values. Thus, all levels are averaged over ten different runs that use different random numbers.

In order to reach comparable and well interpretable results, ten independent sets (for each function) with a list of different random numbers were constructed. For each list, a full computation is done and the random numbers that are included in this list are successively used during the propagation.

8.4. Quorum Sensing as Extended multi–level Logical Formalism

The extended multi–level logical formalism explained in Section 8.3 including growth, transport, and mutation processes was applied to a cell culture of *P. aeruginosa* in which the bacteria communicate and regulate the formation of virulence factors via Quorum sensing systems.

8.4.1. Modeling the Wild type Network

According to the pathway diagram given in Figure 3.5 in Section 3.3 of the three Quorum sensing systems, a topology containing the same nodes and interactions was adopted to the logical formalism, see Figure 8.8. In the model, the enzymes PqsB, PqsC, and PqsD are combined to one single species and ACoA is assumed to be active over the whole time such that it is explicitly modeled as a species.

Due to the transport processes and accumulations, the autoinducers and virulence factors must be modeled as multi–level nodes. Whereas, the number of external autoinducer levels may be arbitrarily large. Therefore, each of the internal signaling molecules AI–1, AI–2, HHQ, and PQS as well as the virulence factors LasB, rhamnolipids (Rhm2), and pyocyanin have six different levels (0–5).

Since receptor proteins are described to prefer a certain metabolite in the literature, we included an inhibition of the weaker complex by the strong binding species. HHQ has a 100–fold lower affinity to PqsR than PQS [205] such that PQS inhibits the complex C5 between PqsR and HHQ when the concentration of PQS is high ($\epsilon = 2$). In a similar way, AI–1 of the *las* system has a weak affinity to the receptor RhlR. To form this complex C4, an ϵ of four is required. When AI–2 of the *rhl* system reaches very high concentrations ($\epsilon = 4$), C4 is inhibited.

The up–regulation of gene expressions is a concentration–dependent process. Instead of including the concentration directly in the activation of Ci:Gi by Ci, the complex Ci between a receptor and an autoinducer is only formed when the autoinducer is at least two ($\epsilon = 2$). Hence, the positive feedback loop is initiated by an autoinducer concentration higher than a critical threshold. The transcription of the autoinducer synthase LasI in the *las* system is blocked by a regulator RsaL, which is again a concentration–dependent process. Therefore, RsaL is also a multi–level variable with three different states, namely “not available” (0), “available at low concentrations” (1), and “sufficient to block LasI” (2).

Dissociations of a complex bound to an operon (Ci:Gj) into the corresponding receptor and autoinducer were included through a further activation edge of the autoinducer.

In reality, the formation of PQS requires two different cells. In the first cell, PqsA uses anthranilic acid and coenzyme A to form AcoA which is then converted in HHQ by the enzyme PqsD. Then, HHQ is transported into the other cell, where it is finally converted into PQS [39]. This may result in a time delay. However, the whole PQS formation starting with anthranilic

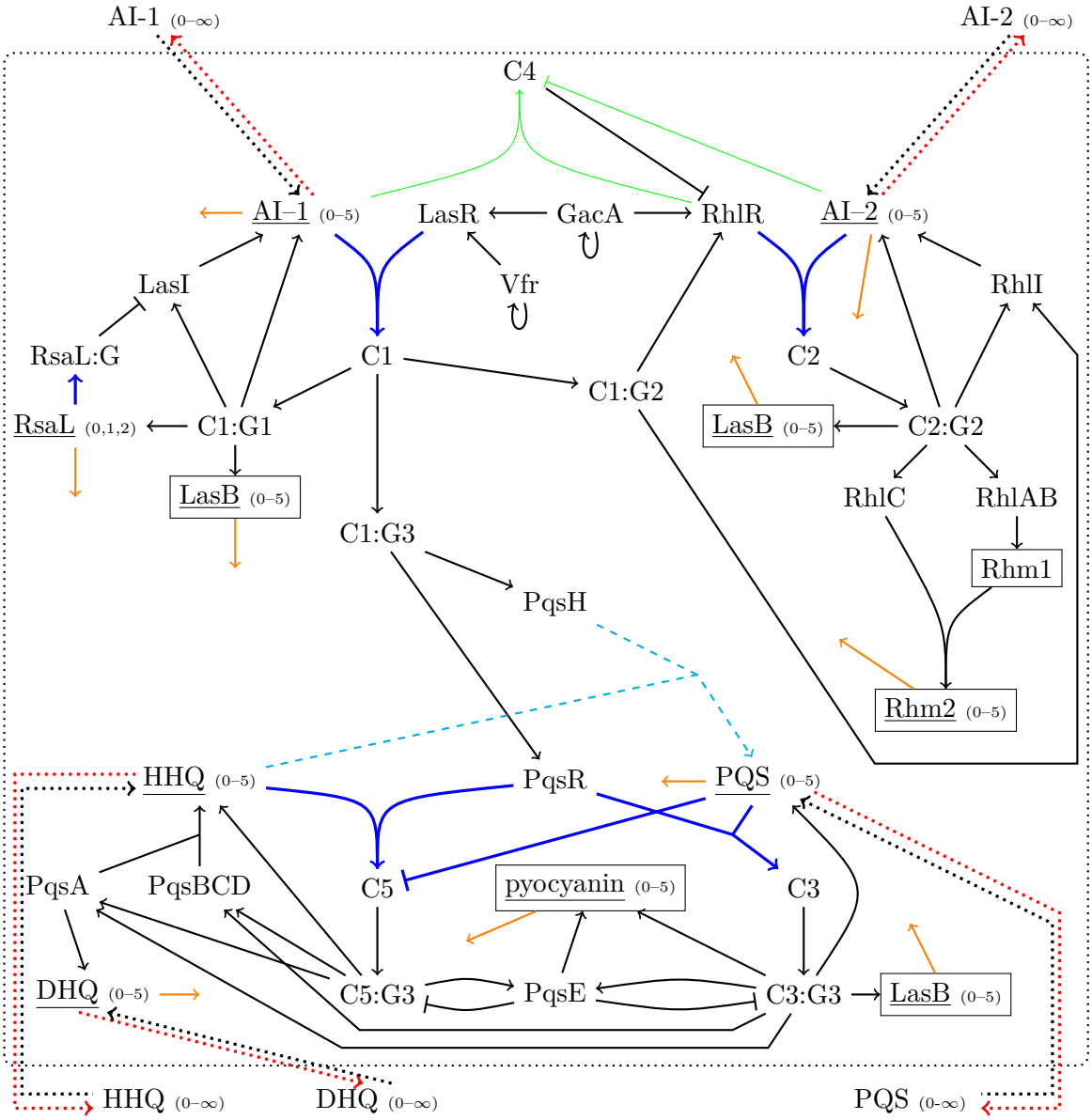


Figure 8.8.: **The three Quorum sensing systems as logical formalism network:** Complexes between a receptor and an autoinducer are labeled as C and complexes bound to an operon as C:G. Multi–level nodes are underlined and their possible levels given in brackets. Black edges denote reactions with a threshold ϵ of one. Blue and thick edges represent reactions with $\epsilon = 2$ for multi–level nodes. In the case of red edges, $\epsilon = 3$ and $\epsilon = 4$ in the case of green and thin edges. Edges colored in orange (degradations) take place after a fixed number of time steps. Cyan and dashed arrows indicate reactions that happen randomly. Dotted arrows characterize transport reactions.

acid was modeled in a single cell in this system. Accordingly, the transformation of HHQ to PQS (in Figure 8.8 illustrated as cyan and dashed arrow) was realized as a reaction that takes place randomly with a certain probability.

8.4.2. Modeling of Knock–out and Gain–of–function Mutants

To validate the network, we analyzed some knock–out and gain–of–function mutants which are described in literature. All considered mutants are listed in Table 8.2, whereby Protein1$^-$ denotes a knock–out mutant in which Protein1 is not formed and Protein2$^+$ produces more Protein2 than the wild type. However, the behavior of most of them was already used to set up the model. For the knock–out mutants, all edges that are related to the formation of the corresponding protein were deleted. In the case of a gain–of–function mutant, a self–activating loop of the respective protein is included.

mutant	changed edges
LasI$^-$	C1:G1 → LasI
LasR$^-$	GacA → LasR, Vfr → LasR
RhlI$^-$	C2:G2 → RhlI, C1:G2 → RhlI
RhlR$^-$	GacA → RhlR, C1:G2 → RhlR
PqsA$^-$	C3:G3 → PqsA, C5:G3 → PqsA
PqsBCD$^-$	C3:G3 → PqsBCD, C5:G3 → PqsBCD
PqsR$^-$	C1:G3 → PqsR
PqsE$^-$	C3:G3 → PqsE, C5:G3 → PqsE
PqsE$^+$	PqsE → PqsE

double mutants
PqsA$^-$–PqsBCD$^-$
PqsE$^-$–PqsR$^-$
PqsE$^-$–PqsA$^-$–PqsBCD$^-$

Table 8.2.: **Knock–out and gain–of–function mutants** with their changed formation edges. Red colored arrows denote removed reactions and blue colored arrows newly added edges.

8.4.3. Modeling of Quorum Sensing Inhibitors

Further, the behavior of the wild type itself was compared with enzyme (PqsBCD) inhibitors and receptor (PqsR) antagonists added to the wild type or randomly occurring mutants. The

Figure 8.9.: **Modeling of Quorum sensing inhibitors** as Boolean nodes with self–activating loop that inhibit the target with a certain inhibition level. This by chance occurring reaction is drawn in dashed and cyan.

inhibitors and antagonists were also combined with some mutants of Table 8.2. Since inhibitors

and antagonists naturally block their target incompletely, such inhibitions of a target were realized with a certain probability as illustrated in Figure 8.9. This inhibition level defines how frequently the target is blocked. The time point of the inhibition is selected randomly (random number $\leq$ inhibition level). For this, the inhibitors are included as Boolean nodes with a self–activating loop that are initialized as active.

9. Results and Discussion

The simple Boolean approach and the extended multi–level logical formalism introduced in the previous chapter were now applied to the three hierarchically organized Quorum sensing systems in *Pseudomonas aeruginosa*. In this chapter, the results are presented and detailed discussed.

9.1. Quorum Sensing as Boolean Network

The positive feedback loop in Quorum sensing was analyzed using a Boolean formalism for the simplified network given in Figure 8.2. Here, the three Quorum sensing systems *las*, *rhl*, and *pqs* were considered separately instead of including their hierarchical linkage.

The propagation of the *las* system resulted in three orbits, namely two cyclic attractors given in Figure 9.1 and Figure 9.2 that oscillate in the positive feedback loop as well as the point attractor that is stuck in a disfunctional state in which all nodes are deactivated. During the

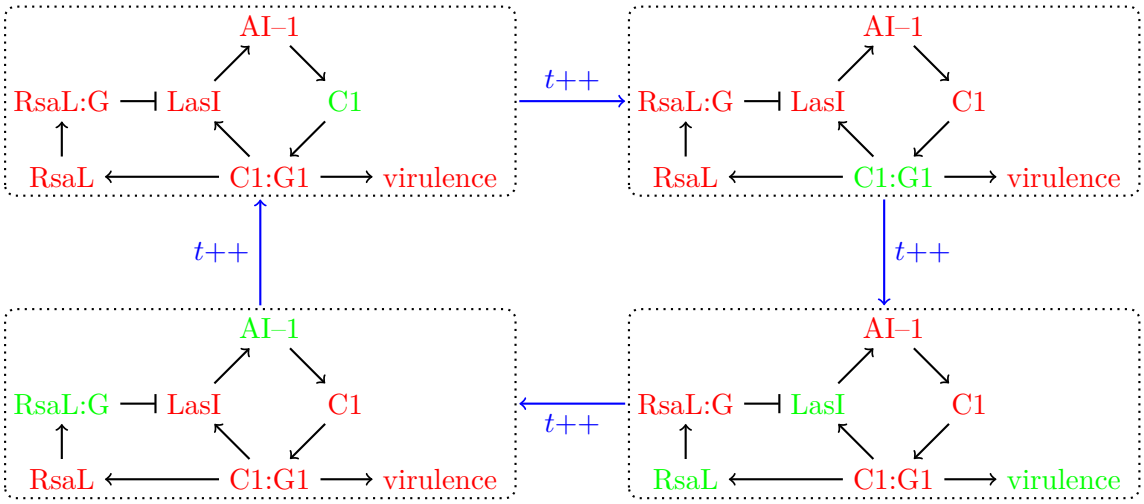


Figure 9.1.: **Cycle attractor 1 resulting from *las* system:** Red colored nodes indicate a deactivated state and green colored nodes an activated state. Blue colored edges denote the update of the system state to the next time step.

oscillation, the formation of virulence factors was activated for both orbits. The first cycle

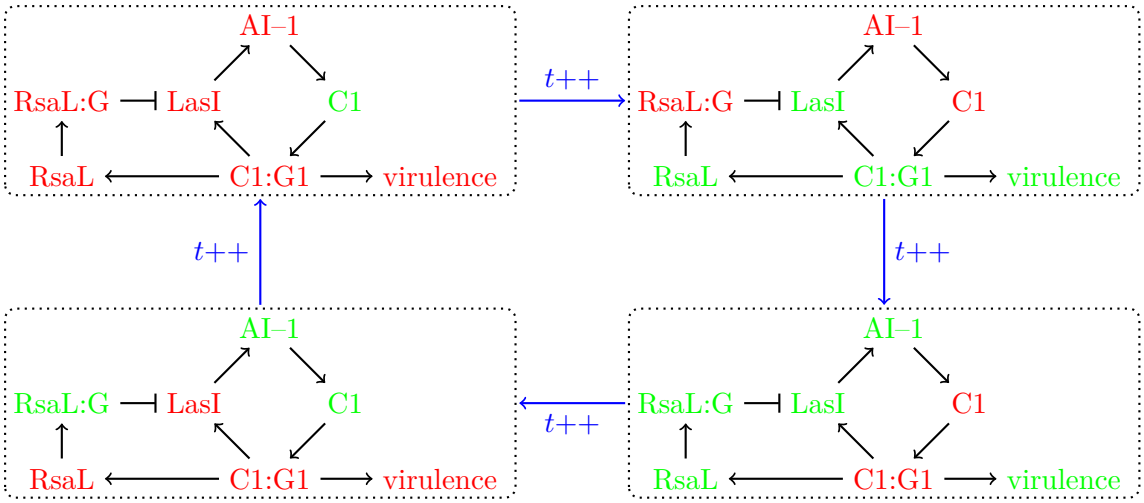


Figure 9.2.: **Cycle attractor 2 resulting from *las* system:** Red colored nodes indicate a deactivated state and green colored nodes an activated state. Blue colored edges denote the update of the system state to the next time step.

attractor has a coverage of about 47% (meaning that 47% of all possible initial conditions converge to this attractor), whereas the second attractor clearly less frequently is reached with a coverage of about 14%. Initial system states that have exactly two directly connected nodes in the positive feedback loop active and the two others inactive end up in the second orbit. In addition, either RsaL, RsaL:G, or both may be initially activated depending on the localization of both activated nodes in the loop. Obviously, the node representing the virulence factors has no influence on the orbit behavior. The negative regulator RsaL prevents the emergence of further attractors.

In the case of the *rhl* and *pqs* systems, an inhibitor is missing. Therefore, there are several cycle attractors and two point attractors, namely the totally deactivated system state and a totally activated system state in which all nodes are active. For the cycle attractors, the activation and deactivation points run step–wise in the positive feedback loop from node to node. Except for the totally deactivated orbit, virulence factors were built at some time step in all cases.

9.2. Quorum Sensing as Extended multi–level Logical Formalism

The three hierarchically layered Quorum sensing systems of *P. aeruginosa* were considered as a combined regulatory and metabolic network in a computational model with multi–level nodes and logical rules. Besides the behavior of the wild type, the consequence of several individual gene knock–outs as well as the influence of Quorum sensing inhibitors on the formation of autoinducers and virulence factors was studied. Simultaneously, we analyzed the implemented topology of the wild type network and adapted it to take into account further literature about

the effects of drug candidates. Additionally, some preliminary results achieved by randomly occurring mutations in the network or simulation parameters were discussed based on the growth of the bacteria population.

9.2.1. Initialization of System State

Since the computation time of logical approaches increases strongly with the number of considered species, not every possible system state was used to find all orbits. Further, the system will not run into a classical attractor due to effects of the stochastically modeled reactions. A certain system state $\mathbf{S}$ may come to state $\mathbf{S}'$ or to state $\mathbf{S}''$ in the next time step. For this reason, it is necessary to define an end point for the simulation t_{max}. Nevertheless, only "attractors" in which all three Quorum sensing systems are active give helpful information. In contrast, the level of every species stays at zero, i.e., deactivated or not available, during the whole propagation, when each initial state is zero. Then, no transport can happen and random mutations only influence the growth process.

In this logical model, initially active complexes or complexes bound to the corresponding operon are required due to missing basal transcription levels that are often found in biological systems. In the case of including basal transcriptions as standard formation reactions, the basal transcription rates are in the same range as activated gene expressions. Then, the autoinducers are not required for the positive feedback loop and for the formation of virulence factors. Thus, the dynamical behavior of the system is changed. Alternatively, basal transcriptions could be realized as stochastic reactions such that a further parameter becomes necessary. Typically, this rate is assumed to be in the same range as the degradation frequency. However, this does not initiate the system developed here. Therefore, instead of including basal transcription rates the system is "manually" started with an adequate initialization.

To activate the *las* system, at least one of the two global activators GacA or Vfr must initially be active. Alternatively, it is possible to set C1:G1 to one, i.e., complex C1 between LasR and AI–1 activates the *las* operon G1. C1 binds also to the operon G2. Consequently, the *rhl* system will be initiated. In the case of purely starting with Vfr, there is a short time delay. If C2:G2 is also initially present, i.e., complex C2 between RhlR and AI–2 activates the *rhl* operon, the activation of the *rhl* system happens directly. To switch–on the *pqs* system, initially active C3:G3 (complex between PqsR and PQS binds to the *pqs* operon) and C5:G3 (complex between PqsR and HHQ binds to the *pqs* operon) are further required. An activation of PqsR avoids a certain time delay since the formation of PqsR causes this delay. Another possibility is to start with elevated initial levels of the internal autoinducers or to use an adequate initial level of external autoinducers that are transported into the cells during the first time steps of the simulation.

An initialization with certain active species is called a minimal system, if any deactivation of every individual species switches the whole system off. Obviously, it is possible to enlarge the

initialization with further active species. Here, only negative regulators, such as RsaL or PqsE, may turn off the system. Figure 9.3 compares the influences of a minimal system in which Vfr, C1:G1, C3:G3, and C5:G3 are one and of a maximal system in which all species except for the virulence factors and the external autoinducers are one on the average levels depending on different time intervals. In the beginning, the system states differed clearly from each other

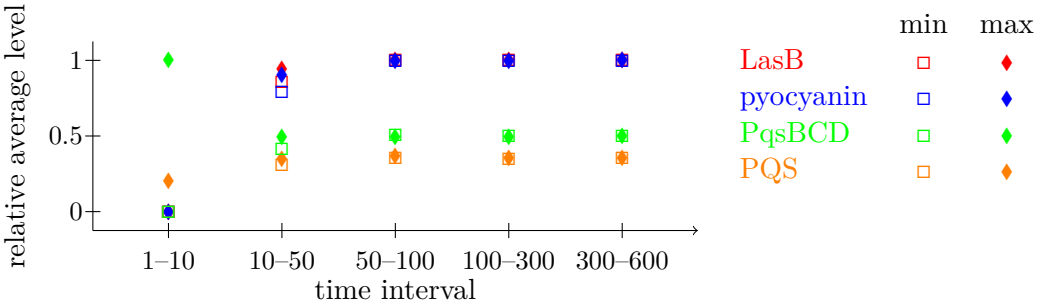


Figure 9.3.: **Comparison between a minimal and a maximal initial system:** Shown is the effect on the average levels of LasB, pyocyanin, PqsBCD, and PQS in different time intervals. The average levels averaged over ten runs are divided by their average theoretical maximum levels Υ. A minimum set denoted as min consists of only Vfr, C1:G1, C3:G3, and C5:G3 initially set to one and a maximum set labeled as max with all species except for LasB, Rhm2, pyocyanin, and external autoinducers initially set to one.

since the whole system has to be turned–on at first in the case of the minimal system and is already active in the case of the maximal initialized system. After around 100 time steps, the averaged levels and the dynamic behavior were very similar independent of the initialization choice. Also, intermediate initialization systems ended up in the same average levels.

In the following, a minimal system with Vfr, C1:G1, C3:G3, and C5:G3 initially activated and all other species levels at zero was used for all simulations.

9.2.2. Parameter Fitting

Due to missing experimental data, the Quorum sensing was modeled as simply as possible. Nevertheless, we had to introduce some adjustable parameters in order to fulfill necessary requirements and to yield reliable and meaningful results. The parameters mainly affect growth, transport, and mutation processes as well as reaction types that were additionally added to the logical formalism. So far, the mortality rate ψ was not included in the approach such that $\psi \geq t_{max}$ for all simulations without testing the effect of this parameter.

Simulation Time (t_{max})

In general, the system achieved a stable regime after 100 time steps as discussed in Section 9.2.1. For this, an example trajectory of a wild type cell is also listed in Table B.2 in appendix Section B.2. Due to randomly modeled reactions, e.g., the conversion of HHQ to PQS, the calculation of equilibrated average levels over multiple time steps after time 100 is required, see Figure 9.3. Therefore, the total number of time steps was chosen as t_{max} equals 600 for statistical reasons. This simulation time was always kept the same except for fitting of growth parameters.

Growth Conditions (δ_{max} and κ)

For simplicity, the growth conditions were analyzed without considering any random mutations ($\nu \geq t_{max}$). Since a growth process depends on t_{max}, κ, and δ_{max}, it is enough to modify t_{max} and κ. The maximal possible number of cell divisions δ_{max} was set to six for all following propagations. The growth rate κ was switched between $\frac{1}{60}$, $\frac{1}{120}$, or $\frac{1}{240}$ time steps together with a suitable t_{max} such that around 200 time steps are simulated after the last cell division. When comparing these three growth processes, the average levels of a certain species typically differed only marginally from each other. For internal nodes that are not influenced by the stochastically realized PQS formation, the simulations yield equal levels. In the *pqs* system, the level deviation is as high as the standard deviation of the ten runs with different random numbers that all share the same growth conditions. Since the degradations of multi–level nodes occur at different system states when varying the growth parameters, the levels of external autoinducers were also slightly changed. Considering levels averaged over the last 100 time steps of the stationary phase, the level of external AI–1 relative to the total number of cells is 3.1 when κ equals $\frac{1}{60}$ and t_{max} equals 600, 3.3 when κ equals $\frac{1}{120}$ and t_{max} equals 1000, and 2.9 when κ equals $\frac{1}{240}$ and t_{max} equals 1700. Here, 64 cells were simulated in the stationary phase. For a propagation without any cell divisions ($\kappa < \frac{1}{t_{max}}$ and $t_{max} = 600$), the average levels of internal and external species were somehow different. Nevertheless, the qualitative behavior remained unchanged. For such a single cell, the level of external AI–1 averaged in the time interval 500 to 600 is 2.6. Thus, the formation of autoinducers and virulence factors as well as the dynamic behavior of the system in general is practically independent of the frequency and the specific time step of cell divisions.

In all following simulations, the growth rate κ equals $\frac{1}{60}$ which may be changed during a random mutation in the case of a mutation rate ν less than t_{max}.

Probability of PQS Formation ($\zeta_{HHQ \to PQS}$)

As the conversion of HHQ to PQS [205] was designed as random reaction, Figure 9.4 demonstrates the dependency of pyocyanin and autoinducer levels on the probability that the reaction

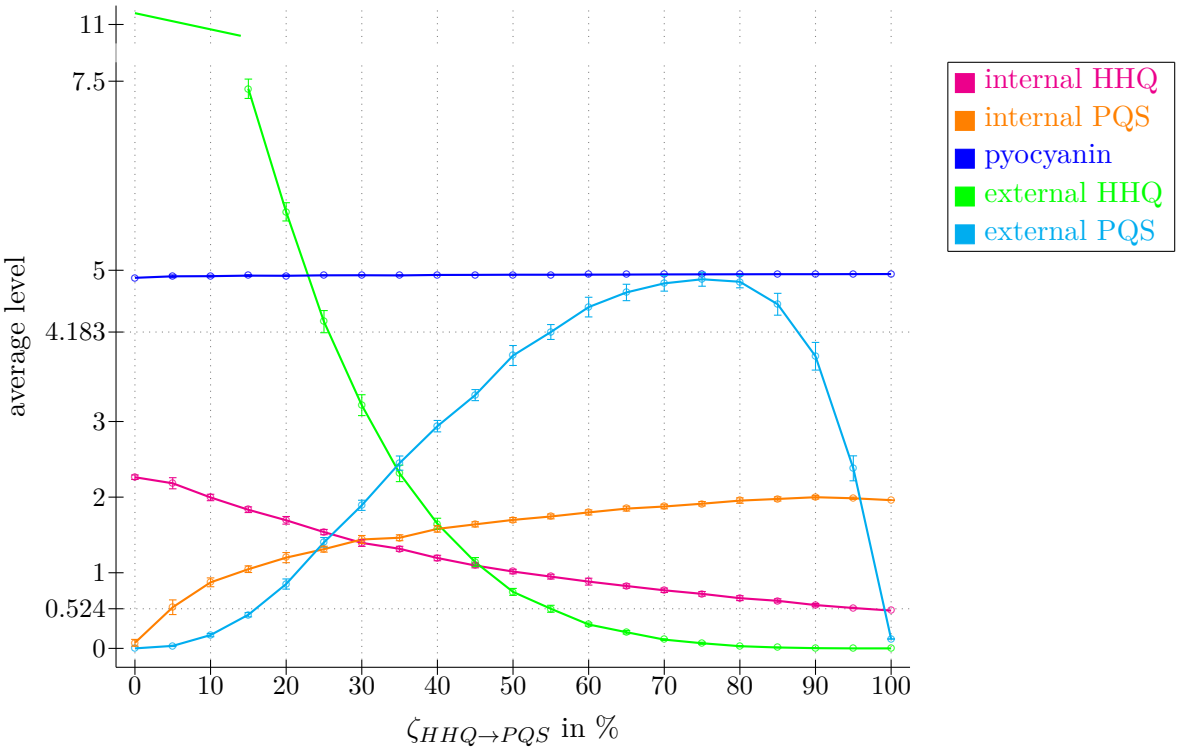


Figure 9.4.: **Influence of PQS formation frequency** $\zeta_{HHQ\to PQS}$ on average levels of HHQ, PQS, and pyocyanin. Shown are average levels in the time interval 100 to 600 averaged over ten runs and their standard deviations. Here, $\varpi = \frac{1}{20}$ and $\chi = 3$.

takes place. The level of pyocyanin was mainly not influenced by this conversion frequency $\zeta_{HHQ\to PQS}$ since PqsE is also formed via C5:G3 using HHQ without any PQS and PqsE is an alternative to activate the biosynthesis of pyocyanin. As expected, the internal and external HHQ levels decrease and the internal and external PQS levels increase with an increasing conversion frequency. As published by Kesarwani and coworker, the concentrations of HHQ and PQS depend strongly on the density of the bacterial culture and the concentration of external HHQ is around 12% of the external PQS concentration after about 11 h [83]. This represents a time step in the beginning of the stationary growth phase. For further propagations, a conversion frequency of 55% was chosen in order to fulfill the experimental findings.

Degradation Frequency (ϖ) of multi–level Species

As mentioned before, the level of multi–level species is reduced by one after a certain number of time steps. The influence of the degradation frequency on the average levels of the virulence factors is illustrated in Figure 9.5. For this, the actual average levels were compared to the average theoretical maximum levels Υ [1]. Obviously, very frequently occurring degradations have a strong effect on the average levels. Due to the required autoinducers in the positive feedback loop, the complete network becomes inactive except for Vfr and LasR in the case of

[1] average theoretical maximum level Υ is defined in equation (8.9) in Section 8.3.1

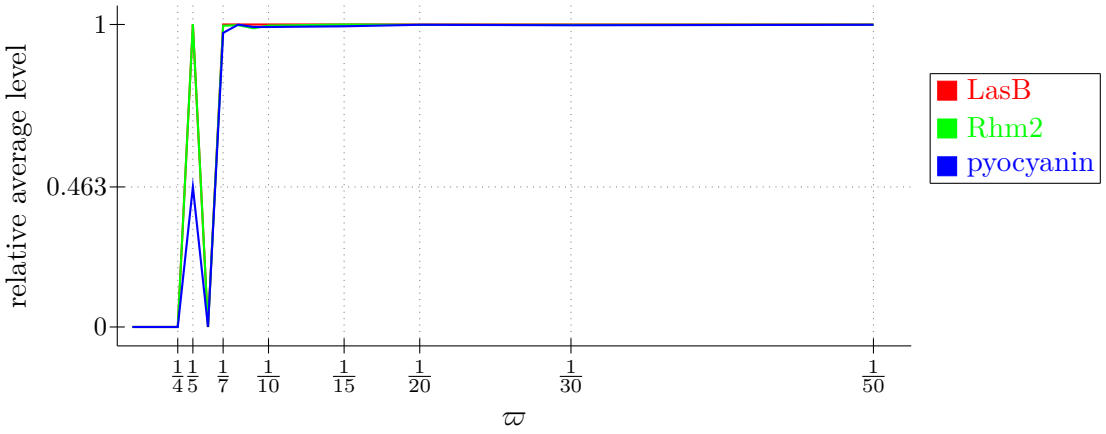


Figure 9.5.: **Influence of degradation frequency** on average levels of the the virulence factors LasB, Rhm2, and pyocyanin in the time interval 100 to 600 relative to the average theoretical maximum levels Υ averaged over ten runs. Here, $\zeta_{HHQ \to PQS} = 55\%$ and $\chi = 3$.

frequently occurring degradations with less than five time steps. Vfr is initially set to one, activated by a self–loop at each time step, never inhibited, and not influenced by degradations due to its Boolean character. Further, it activates LasR at each time step.

In contrast, the influence on the average levels of autoinducers and virulence factors decreased with decreasing degradation frequency. This could be easily explained when considering a positive feedback, e.g., the *rhl* system as illustrated in Figure 9.6 for large frequencies. The plot for smaller frequencies is shown in Figure B.1 in appendix in Section B.2. Starting with an internal AI–2 level of two and an activated RhlI at time t complex C2 is activated in the next time step $t+1$ and C2:G2 afterwards. At time $t+3$, this complex bound to an operon dissociates such that the level of AI–2 is increased by one. In the next time step, AI–2 is built by RhlI. When a degradation of AI–2 happens in this circle, the level decreases under the required threshold to form the complex. Figure 9.6 shows that the system is switched off independent of the time point of a degradation in the cases where ϖ equals one and ϖ equals one half. In contrast, the behavior of *rhl* system depends on the time point of the degradation (see Figure B.1) when ϖ equals one third. For this plot, we assumed a constantly activated RhlR. However, due to frequently occurring degradations, this must not be necessarily the case such that the probability of a switching–off may be higher than illustrated. Nevertheless, the average theoretical maximum level Υ was approximated by the averaged pyocyanin level for degradation frequencies less than $\frac{1}{6}$ and reached for frequencies less than $\frac{1}{19}$.

Thus, a degradation frequency of $\frac{1}{20}$ was used for all further simulations. Since the whole protein biosynthesis with transcription and translation is done in a single time step in our model, degradation processes occur rarely for a frequency of $\frac{1}{20}$. In contrast, a single degradation reduces the concentration of a certain species strongly due to the sparse number of possible levels. A degradation process consumes one third of the RsaL concentration and a sixth of the other multi–level nodes.

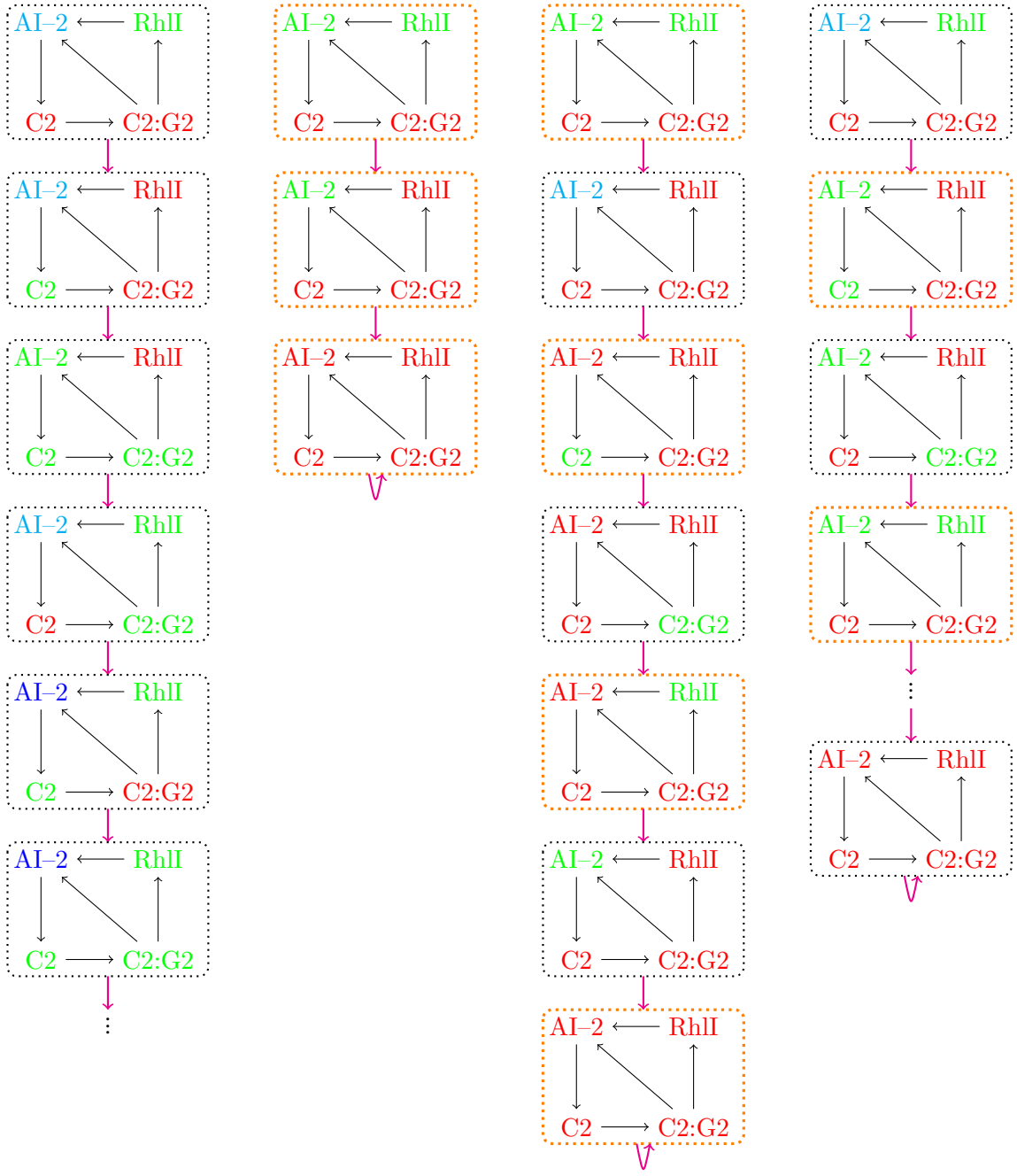


Figure 9.6.: **Influence of degradation frequency on *rhl* system:** Shown is a simplified, positive feedback loop of the *rhl* system. Species colored in red represent a level of zero, in green a level of one, in cyan a level of two, and in blue a level of three. Arrows colored in magenta denote the update of the system state to the next time step and cells labeled in orange mark the time point at which a degradation happens. The left column demonstrates a simulation without degradations, in the second column degradations take place at each time step ($\varpi = 1$), and in the two columns on the right side a degradation occurs in every second step ($\varpi = \frac{1}{2}$).

Transport Threshold (χ)

The decision in which direction an autoinducer is transported is based on the idea of concentration gradients. In the case of an autoinducer transport out of a cell, the level of the corresponding internal autoinducer has further to reach a certain threshold χ. For all considered autoinducers, the same threshold is applied.

In a time interval from 100 to 600, the influence of different transport thresholds is typically marginal. Obviously, the smaller the transport threshold (i.e. the more frequently autoinducers are transported out of a cell), the smaller are the internal and the higher are the external autoinducer levels. Since the average levels of HHQ and PQS are usually smaller than their maximum values, the external levels of both were zero in the case when χ is less than five. In contrast, the levels of internal HHQ and PQS were quite small in the case when χ equals one such that the level of pyocyanin was also reduced a little.

The effect of χ on the LasB level during the starting phase of the single starting cell is demonstrated in Figure 9.7. LasB reaches its maximal possible level only three time steps earlier for

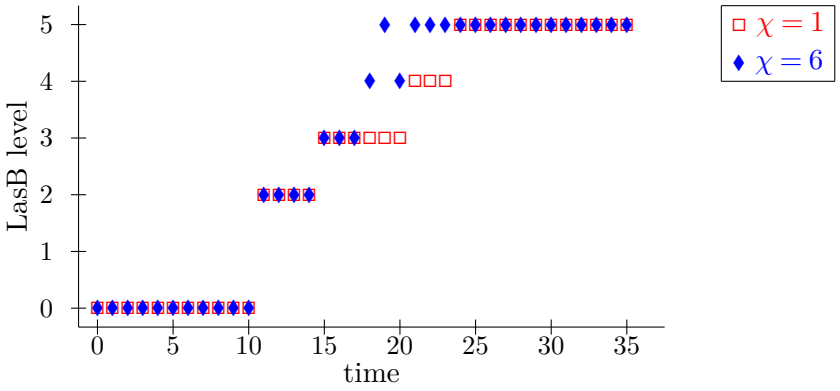


Figure 9.7.: **Influence of transport threshold χ on LasB formation**: Illustrated are the levels of LasB using $\chi = 1$ with $\chi = 6$ during the switching–on. Here, $\varpi = \frac{1}{20}$ and $\zeta_{HHQ \to PQS} = 55\%$.

χ equal to six than for χ equal to one. Therefore, the influence of χ on the starting behavior is smaller than the effect of other initializations.

For all further propagations, a transport threshold χ equal to three is used, what should represent a quite high concentration of autoinducers in reality.

Mutation Rate (ν)

The mutation rate ν, which is equal for all cells and unchanged during the whole simulation, controls how frequently mutation processes happen, where the number of mutation processes equals $\frac{t_{max}}{\nu}$. In a single mutation process, a certain action occurs with a given probability P. The used random variables are Gaussian distributed and rounded to an integer. Therefore, for random variables b in the interval $[\mu-0.5, \mu+0.5]$ no action takes place. Then, the probability is given as in equation (9.1),

$$\begin{aligned} P(\text{action}) &= 1 - P\left(\mu - \frac{1}{2} \leq b \leq \mu + \frac{1}{2}\right) = 1 - \int_{\mu-\frac{1}{2}}^{\mu+\frac{1}{2}} P(b)\ \mathrm{d}b \\ &= 1 - \left(\Phi\left(\frac{1}{2\sigma}\right) - \Phi\left(\frac{-1}{2\sigma}\right)\right) = 2 - 2\Phi\left(\frac{1}{2\sigma}\right) \end{aligned} \tag{9.1}$$

where μ is the mean and σ the standard deviation of the corresponding Gaussian distribution and Φ is the probability density function of the standard normal distribution. The values were taken from a density function table. Accordingly, the probability of a growth rate modification is 93.6%, the probability that the maximum number of cell divisions δ_{max} is changed equals 54.2%, and the probability that the occurring frequency ζ of a certain reaction in the network is varied equals 72.6%. Thus, it can be expected that an action appears about $\frac{P(\text{action})t_{max}}{\nu}$ times during the full simulation.

If the network is modified, the corresponding edges are chosen randomly. The network of the wild type and therefore all considered randomly occurring mutants contains 60 reactions. Since most reactions of the original network have a probability of 100%, the actual probability that the frequency of those reactions is modified is about 36.3%. For a mutation rate ν less than 33, a reaction is affected twice by mutation processes with a probability higher than 95%. Even a mutation rate ν less than eleven can not guarantee that each reaction is affected at least once in a mutation process.

Figure 9.8 illustrates the effect of the mutation rate on the number of modifications and on the total number of cells. Obviously, the number of modifications of the parameters κ and δ_{max} increased detectably with a decreasing mutation rate. Hereby, κ was more frequently varied. Modifications in the growth rate or in the maximum number of cell divisions changed the total number of cells and the start time of individual cells. The number of cells and the standard deviation clearly were increased with a decreasing mutation rate.

To significantly vary the average level of some species, it is required to update the network. Whereas the possibility of a network change increases with an decreasing mutation rate, the conversion frequencies was on average only modified by a factor of four in the case of a mutation

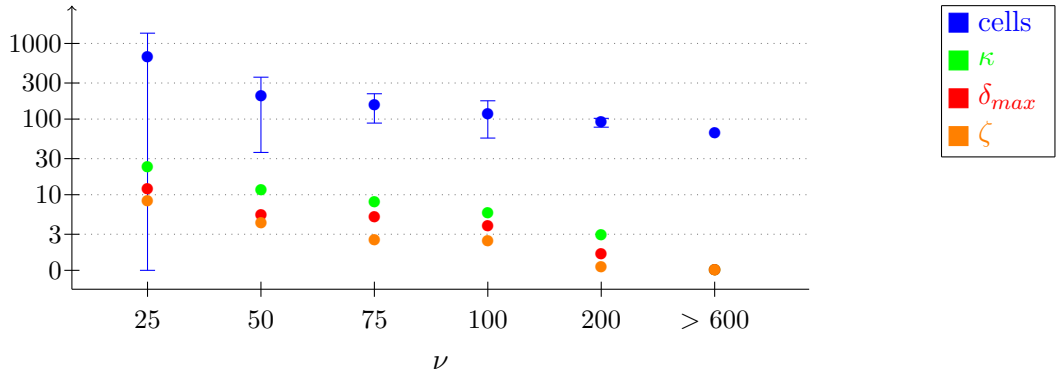


Figure 9.8.: **Influence of mutation rate ν on growth process:** Shown are the average numbers of cells in a culture at the end of the simulation (t_{max}) and the corresponding standard deviations. The number how often the growth rate κ is changed is given in green, the number how often the maximal possible number of cell divisions δ_{max} is changed in red, and the number how often any conversion frequency ζ in the network is changed is given in orange. The values are averaged over ten different runs.

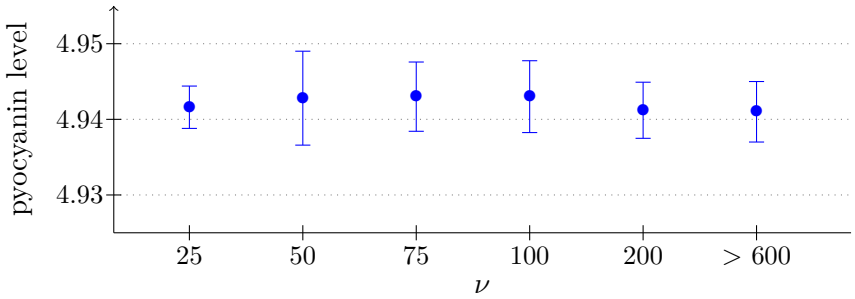


Figure 9.9.: **Influence of mutation rate on pyocyanin formation:** Shown are average levels in the time interval 100 to 600 averaged over ten different runs and their standard deviations. Here, $\varpi = \frac{1}{20}$, $\chi = 3$, and $\zeta_{HHQ \to PQS} = 55\%$.

frequency of 50. Figure 9.9 shows the behavior of the pyocyanin level depending on different mutation rates. As expected, the standard deviation of the pyocyanin level averaged over ten different simulations was generally strongly increased with a decreasing mutation rate — interestingly, except for a mutation rate of 25. Due to the infrequently occurring modification of conversion frequencies, the pyocyanin formation may be directly and also indirectly affected.

In the following, $\nu \geq t_{max}$ was used to analyze Quorum sensing inhibitors and to study the network topology. Due to the infrequent occurrence of modifications in the network for large mutation rates, ν equal to 25 is used when considering mutation processes in detail, see Section 9.2.9.

Overview

To summarize the observations of the parameter fitting, Table 9.1 gives an overview over all parameters required in the model and their optimized values. In all the following simulations,

Parameter	function	value
t_{max}	total number of time steps	600
δ_{max}	maximal possible number of cell divisions	6
κ	growth rate	$\frac{1}{60}$
ψ	mortality rate	$\geq t_{max}$
ν	mutation rate	25 or $\geq t_{max}$
χ	transport threshold	3
$\zeta_{HHQ \rightarrow PQS}$	conversion frequency for reaction HHQ $\rightarrow$ PQS	55%
ϖ	degradation frequency for multi–level nodes	$\frac{1}{20}$

Table 9.1.: **Overview over the parameters** required in the Quorum sensing model based on an extended multi–level logical formalism and their commonly used values. Parameters colored in red may be influenced during random mutations. Then, the given value is the initial value of this parameter.

these values were used when no parameters are explicitly listed. When the mutation rate ν was less than t_{max}, δ_{max} and κ were initialized with the corresponding value, whereas they may be changed during a random mutation. For simplicity, $\nu \geq t_{max}$ in Sections 9.2.3 – 9.2.7. In Section 9.2.9, $\nu = 25$. In general, only the average values of the starting cells are discussed. Often, a simulation starts with only a single cell representing a cell culture.

9.2.3. Behavior of Switching–on

After a delay time of ten steps during which the initial system state of a new cell may only be changed by transport processes, the system state is updated based on the logical rules. Incipiently, the *las* system is activated and then, the *rhl* and *pqs* systems become active starting by forming complex C1 between LasR and AI–1. C1 that is required to form C1:G2 and C1:G3 causes the delay for turning–on the *rhl* and *pqs* systems. The switching–on behavior of the complexes is demonstrated considering the single start cell in Figure 9.10. Since starting from the initially active C1:G1 the three reactions, comprising the whole biosynthesis,

1. C1:G1 $\rightarrow$ LasI
2. LasI $\rightarrow$ AI–1
3. AI–1 + LasR $\rightarrow$ C1

are necessary to form complex C1. It is first formed at time step 13. LasR is active the whole time after time step eleven and C1 after time 19.

Besides the formation of C1, the reactions

1. C1:G2 $\rightarrow$ RhlI
2. RhlI $\rightarrow$ AI–2
3. AI–2 + RhlR $\rightarrow$ C2

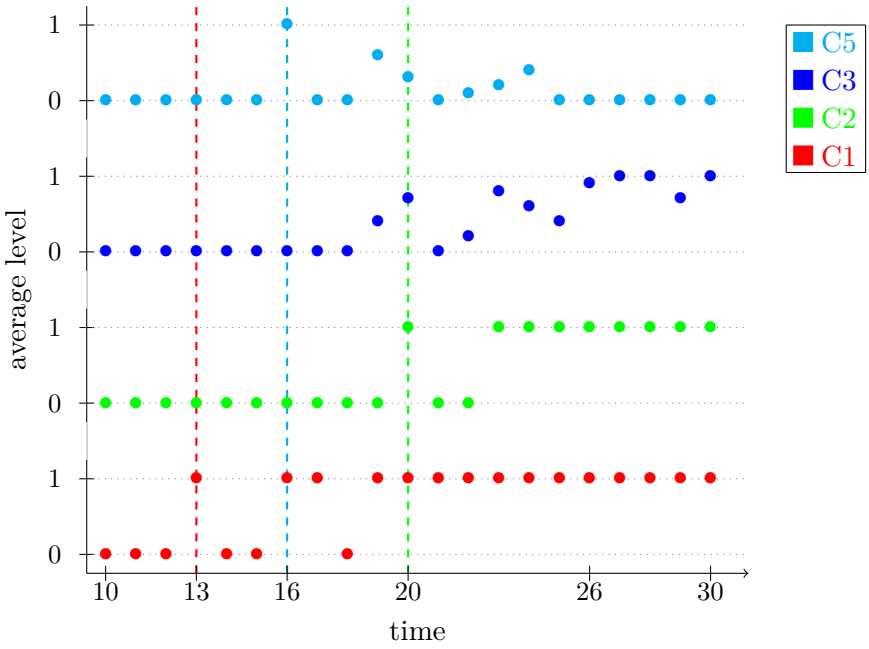


Figure 9.10.: **Switching–on behavior illustrated the formation of the complexes:** Shown are the levels averaged over ten different runs. Dashed lines represent the time step of the first activation for the complex with the same color.

are further necessary to form the complex C2 of the *rhl* system. Yet, the third reaction can not take place due to the threshold ϵ of two for AI–2. Hence, the cycle must be completed once more such that C2 is activated for the first time at time 20.

Due to the initial activations of C3:G3 and C5:G3, the state of internal HHQ is two and the state of internal PQS is one at time twelve. PqsR and PqsH become active for the first time at time step 15 after C1 bound to genome G3. Then, complex C5 between PqsR and HHQ is built in the next time step ($t = 16$). The occurrence of C3 and C5 is based on the stochastically realized conversion of HHQ to PQS by PqsH. The first peak of complex C3 formed between PqsR and PQS is observed between time 19 and time 23 depending on the used random numbers. Since PQS inhibits C5 with the same threshold ($\epsilon = 2$) that is required to form C3, C5 is inactive if C3 is active in later time steps.

9.2.4. Simulation of Wild type

Figure 9.11 shows the average levels of some species simulating the wild type network with the standard parameter set–up. The virulence factors LasB and Rhm2 were produced at their average theoretical maximum levels Υ. The level of the virulence factor pyocyanin was on average only slightly less than the maximum. Since there is no negative regulator in the *rhl* system, RhlI was active the whole time and the levels of AI–2 were quite high with almost four in the interior of the starting cell and about 140 in the external environment. Since AI–2

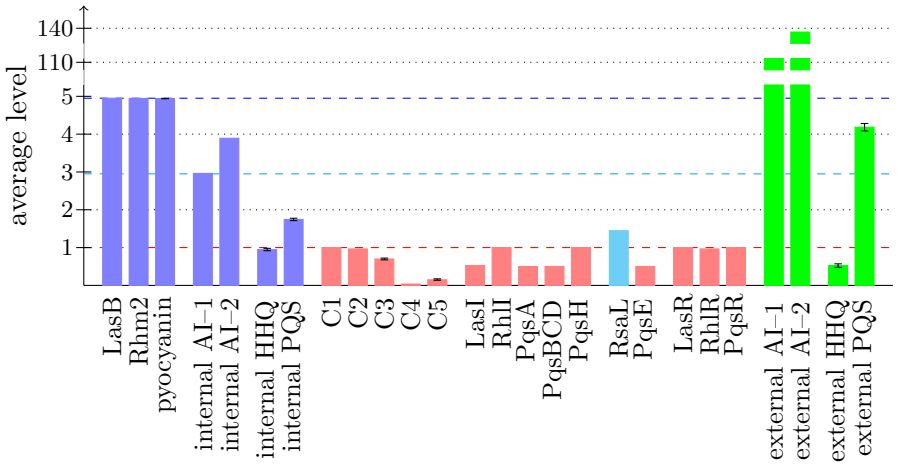


Figure 9.11.: **Average levels of wild type species** in the time interval 100 to 600 averaged over ten runs together with their standard deviations. The dashed line in blue represents the average theoretical maximum level Υ of species (given as blue bars) with six possible states, the dashed line in cyan represents Υ of species with three possible states, and the dashed line in red represents the maximum level of Boolean nodes. The level of species with an unfixed number of possible states are given in green.

inhibits the formation of C4 (complex between AI–1 and RhlR) at high concentrations, the average level of C4 is close to zero. RhlR and therefore C2 were indirectly influenced in earlier time steps by RsaL via C1:G2. Nevertheless, their average levels were around one.

Due to the self–regulating edge of Vfr, LasR was always active. In the *las* system, RsaL inhibits the formation of LasI such that LasI was deactivated half of the time. Therefore, the levels of internal and external AI–1 were less than those of AI–2. The level of RsaL was about half of Υ. However, the internal AI–1 level was high enough to permanently build the complex C1. Consequently, PqsR and PqsH were active the whole time. The internal and external levels of HHQ and PQS were clearly less than those of AI–1 and AI–2 because PqsE is activated together with the biosynthesis enzymes PqsA and PqsBCD and works as negative regulator in the *pqs* system. The average internal levels of HHQ and PQS were so small that PqsA, PqsBCD, and PqsE were active only about every other step. However, the average level of internal PQS is almost two. Therefore, the average level of complex C3 (between PQS and PqsR) is about three times higher than that of complex C5 (between HHQ and PqsR).

9.2.5. Simulation of Knock–out Mutants and Gain–of–function Mutants

Figure 9.12 compares the behavior of the mutants listed in Table 8.2 in Section 8.4.2 with the wild type based on the average levels of pyocyanin and internal PQS. Here, the double mutant PqsA$^-$–PqsBCD$^-$ is denoted as PqsABCD$^-$ and PqsE$^-$–PqsA$^-$–PqsBCD$^-$ as PqsABCDE$^-$. Complex C1 is never built due to the lack of AI–1 in the case of LasI$^-$ and due to absent

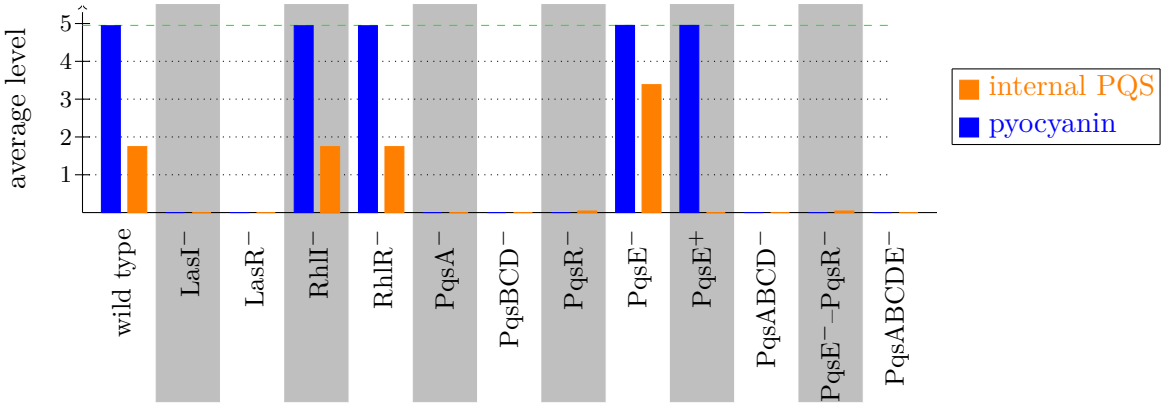


Figure 9.12.: **Average levels of internal PQS and pyocyanin levels for wild type and mutants** in the time interval 100 to 600 averaged over ten runs. The green colored dashed line represents the average theoretical maximum levels Υ.

LasR in the case of LasR$^-$. This resulted in a lack of PqsR, which is required to form PQS and pyocyanin. Since RhlI and RhlR have no influence on the activation of PqsR or on the formation of the autoinducers in the *pqs* system, the average levels of PQS and pyocyanin for RhlI$^-$ and RhlR$^-$ knock–out mutants were as high as in wild type cells.

In the literature, it has been reported that knock–out mutants related to *pqsR* or the *pqsA–E* operon are deficient in pyocyanin in comparison to the wild type [39, 54, 145]. Consistently, the PQS and pyocyanin production was drastically decreased for the modeled and simulated knock–out mutants of the corresponding genes. In our model, an increased level of internal PQS was achieved when deleting PqsE. Since C3:G3 and C5:G3 were not longer blocked, this mutant was able to form pyocyanin at the average theoretical maximum levels Υ. Further, the level of external PQS was clearly increased. In contrast, PqsE$^-$ knock–out mutants are described to produce strongly reduced amounts of pyocyanin compared to wild type cells and a cell with an increased concentration of PqsE reaches a higher pyocyanin concentration [48, 145]. Further, it is known that PqsE regulates the pyocyanin formation independent of the *pqs* system (without PQS or PqsR), but with a dependence on the *las* and *rhl* systems [48]. For the autoinducers, it has been described that the external concentration of a PqsE$^-$ knock–out mutant is in the same range as for a wild type cell [48], whereas the HHQ and PQS concentrations are decreased in PqsE$^+$ mutants [145]. The modeled gain–of–function mutant PqsE$^+$ produces more pyocyanin than the wild type, but not PQS. The discrepancy between experimental findings and the model for pyocyanin is discussed in detail in Section 9.2.7.

9.2.6. Simulation of Quorum Sensing Inhibitors

In the following, we studied the effects on the *pqs* system of antagonists for receptor PqsR and inhibitors for PQS biosynthesis by blocking the enzyme PqsBCD. For this, Figure 9.13 again

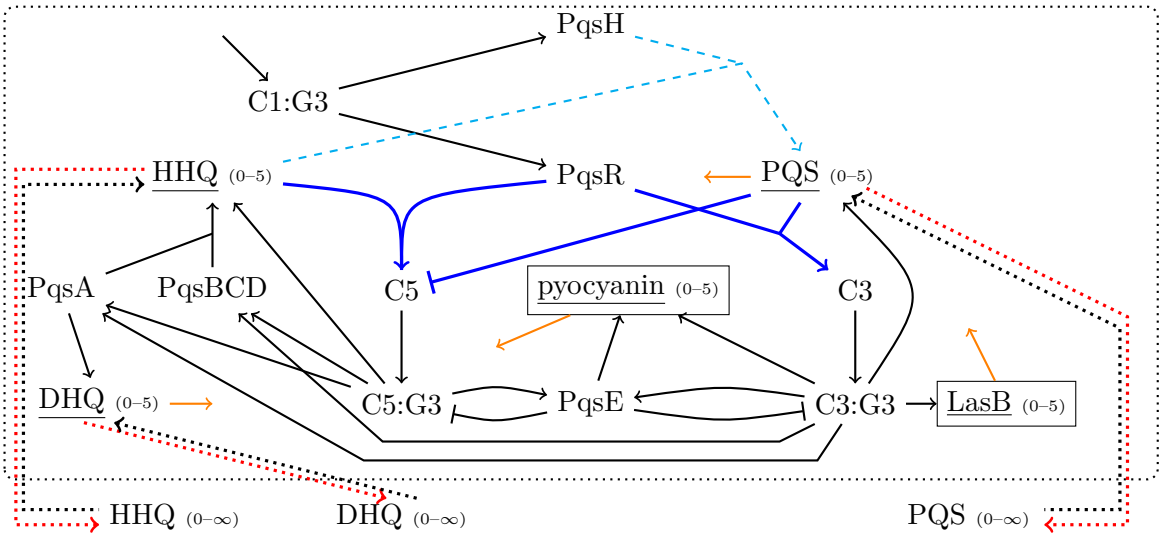


Figure 9.13.: **The *pqs* system as logical formalism network:** Complexes between a receptor and an autoinducer are labeled as C and complexes bound to an operon as C:G. Multi–level nodes are underlined and their possible levels given in brackets. Black edges denote reactions with a threshold ϵ of one. Blue and thick edges represent reactions with $\epsilon = 2$ for multi–level nodes. Edges colored in orange (degradations) take place after 20 time steps. The cyan and dashed arrow indicates the randomly happening formation of PQS with a conversion frequency of 55%. Dotted arrows characterize transport reactions with a threshold $\chi = 3$ for red colored edges.

shows the Quorum sensing system that was simulated together with the *las* and *rhl* systems using the previously listed initial conditions and parameters. The influence of the antagonists and enzyme inhibitors on the average levels of internal and external HHQ and PQS as well as of pyocyanin was compared with each other and the impact of varying inhibition levels was analyzed. Here, inhibition level denotes the number of blocked receptors or enzymes. Figure 9.14 compares the behavior of external HHQ and PQS in a wild type cell (no inhibition) with different blocked networks.

PqsR antagonists avoid the activation of complex C3 (between PqsR and PQS) such that more PQS can be transported out of a cell. Therefore, the average level of external PQS was noticeably increased. For inhibition levels higher than 90%, the fact that neither C3 nor C5 can induce the transcription of PqsBCD, which is required to form PQS, is stronger than the transport effect. PqsBCD inhibitors remarkably reduced the average levels of external HHQ and PQS with a clear dependence of the inhibition level. The external HHQ level was decreased by about 80% and the external PQS level by about 55% in the case of a PqsBCD inhibitor with inhibition level of 30%. In fact, Storz *et al.* reported that an actual PqsD inhibitor with around 42% less external PQS produces also around 77% less external HHQ [180].

The influence of PqsR receptor antagonists with different inhibition levels on internal HHQ, PQS, and pyocyanin levels is shown in Figure 9.15. The internal autoinducer levels are in-

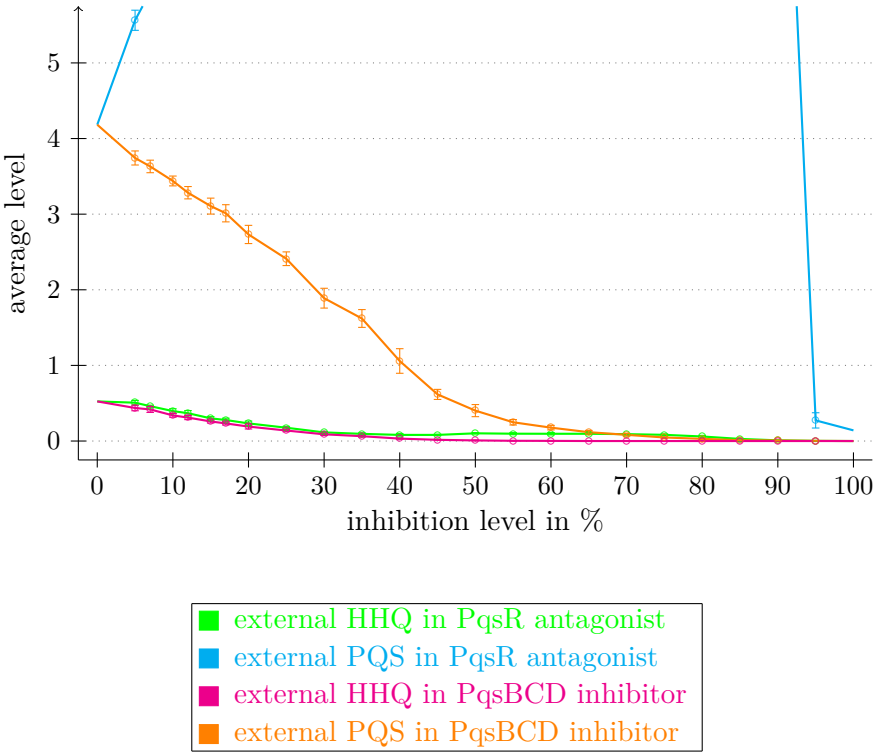


Figure 9.14.: **Average levels of external HHQ and PQS for receptor antagonists and enzyme inhibitors** in the time interval 100 to 600 averaged over ten runs together with their standard deviations. Considered are receptor antagonists blocking PqsR and enzyme inhibitors blocking PqsBCD with different inhibition levels. An inhibition level of 0% represents the wild type.

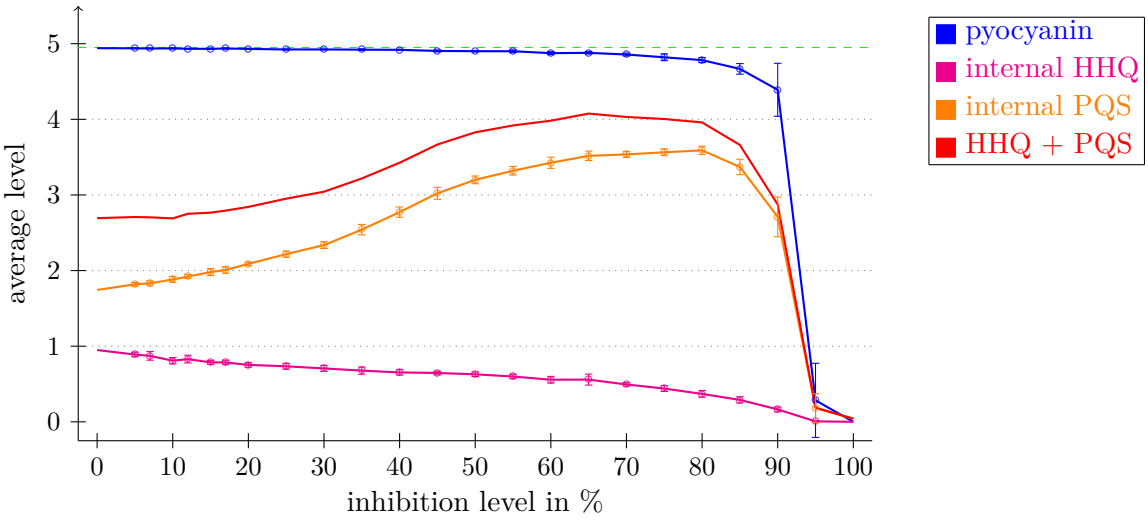


Figure 9.15.: **Average levels of internal HHQ, PQS, and pyocyanin for receptor antagonists** in the time interval from 100 to 600 averaged over ten runs together with their standard deviations. Considered are receptor antagonists blocking PqsR with different inhibition levels. The dashed green colored line denotes the average theoretical maximum levels Υ.

creased as long as less than 60% of PqsR is blocked due to a reduced usage and are clearly decreased for larger inhibition levels due to their missing biosynthesis. The pyocyanin level was also only marginally reduced for inhibition levels until 60% for PqsR antagonists taken from the literature. For HHQ analogs with K_D values in a low nanomolar range acting as PqsR antagonists, the pyocyanin production was reduced by around 75% at 3 µM [105]. Other antagonists with affinity in the low micromolar range also strongly decreased the pyocyanin concentration with an IC_{50} of 87 µM [85]. Nevertheless, the predicted PqsR antagonists showed clearly less pyocyanin formation for inhibition levels higher than 90%.

The effect of weak and strong PqsBCD inhibitors on the internal autoinducer and pyocyanin levels is given in Figure 9.16. As expected, the internal levels of HHQ and PQS are also

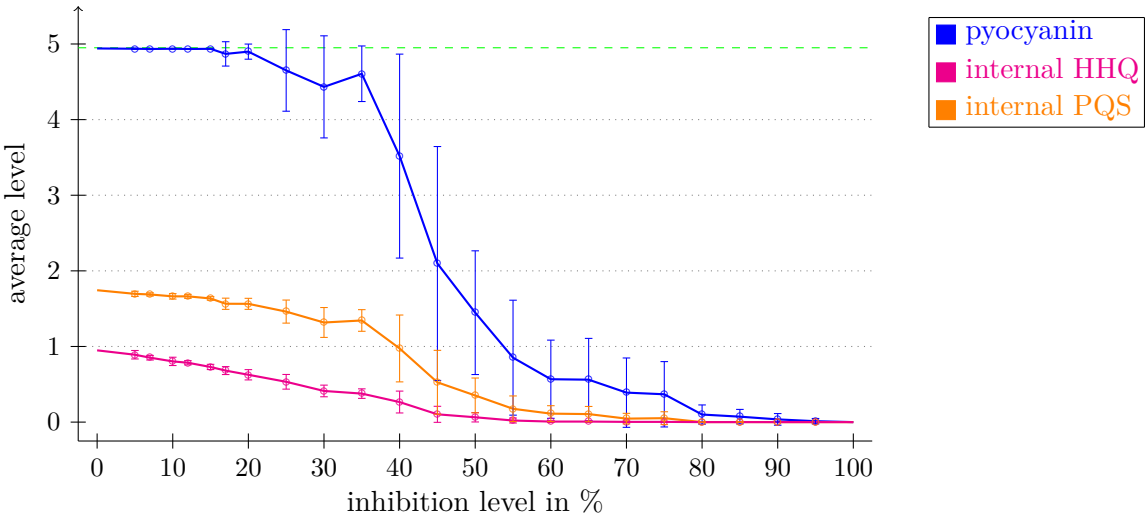


Figure 9.16.: **Average levels of internal HHQ, PQS, and pyocyanin for enzyme inhibitors** in the time interval from 100 to 600 averaged over ten runs together with their standard deviations. Considered are enzyme inhibitors blocking PqsBCD with different inhibition levels. The dashed green colored line denotes the average theoretical maximum levels Υ.

decreased. Interestingly, the pyocyanin level is strongly reduced with a relatively high standard deviation for inhibition levels in the range from 25% to 75%. For example, the pyocyanin level is decreased by around 10% on average in the case of inhibiting PqsBCD by 30%.

9.2.7. Systematically Varying Topology

The levels of internal and external HHQ and PQS as well as of other Quorum sensing metabolites agree with experimental findings on wild type bacteria, several mutants, and with the behavior of specific receptor antagonists or enzyme inhibitors added to the bacterial culture. In contrast, the pyocyanin level calculated in our simulations and the actually observed behavior of pyocyanin concentrations are in some points different (considering PqsE mutants and

Quorum sensing inhibitors) different. Therefore, it seems that additional regulatory mechanisms of the pyocyanin formation are missing. It is possible, but not necessary, that functionally unknown proteins of those mechanisms are related to the *pqs* system. Since none of the metabolic pathways outside the Quorum sensing systems was included in the simulations, only proteins connected to the system are interesting for follow–up studies. Hence, we analyzed several networks with altered relations in the *pqs* system to better understand the network topology. Due to the large number of possible modification, only additional activations between already included proteins were considered. Modified parameters, such as reaction rates or different number of maximal possible states for some species or negative regulations, were ignored. Further, we replaced the feed–forward link

$$\textit{pqs}\text{–member} \rightarrow X \rightarrow \text{pyocyanin},$$

where *pqs*–member represents a known protein or metabolite in the *pqs* system and X denotes a hypothetical unknown protein, by the direct connection

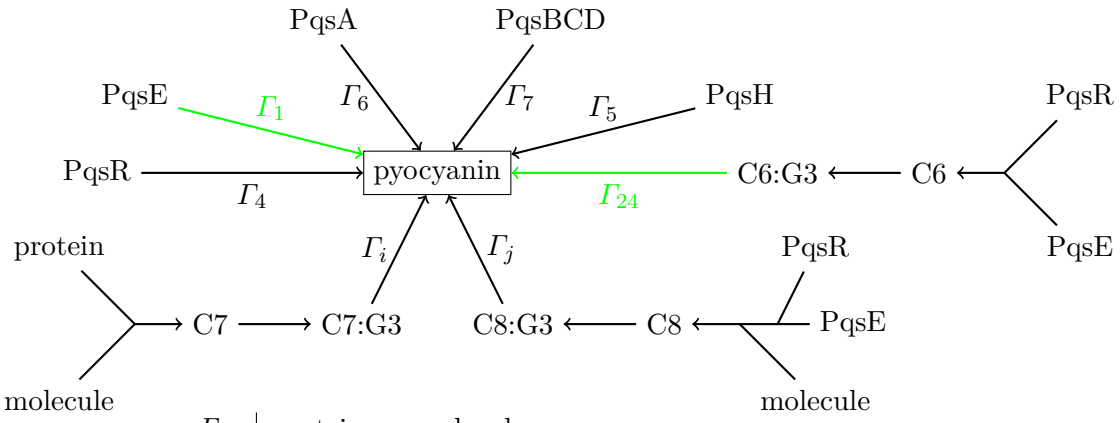
$$\textit{pqs}\text{–member} \rightarrow \text{pyocyanin}.$$

Therefore, it is not necessary to insert any new protein. Due to the usage of a synchronous updating scheme instead of precise rate constants in the logical formalism, this simplification does not change the dynamic behavior of the system. Moreover, linkages are typically assumed to be between small molecules and proteins that act as receptors.

Modified Networks

Based on these restrictions, all theoretically imaginable connections Γ to build pyocyanin are illustrated in Figure 9.17. Due to experimental findings that demonstrate the need for PqsR, PqsA, PqsBCD, and PqsE, a meaningful network has to contain either a reaction that is colored in blue or a combination of a red and a green colored reaction. As an alternative to red colored reactions, it is possible to include one of the reaction pairs Γ_6 and Γ_7, Γ_{14} and Γ_7, Γ_{11} and Γ_7, or Γ_8 and Γ_7. Additionally, a combination of Γ_7 and Γ_{20} fulfills the four conditions. All discussed mutants and Quorum sensing inhibitors have no impact on PqsH. Hence, reactions Γ_5 and Γ_9–Γ_{11} that include PqsH were not used to set up new networks. Since PqsA and PqsD play an enzymatic role in the biosynthesis of HHQ, DHQ, and PQS, reactions Γ_{12}–Γ_{17} were not considered. PqsR antagonists described in the literature showed a strongly reduced pyocyanin level, whereas our modeled PqsR antagonists decreased the pyocyanin level only at high inhibition levels. This suggests that networks with either Γ_4 or Γ_{21}–Γ_{24} perform most realistic.

The considered networks are given in Table 9.2. The original network (see Figure 8.8 in Section 8.4.1) that was discussed in detail in previous sections is now labeled as $\mathcal{N}_1$. In network



Γ_i	protein	molecule
Γ_2	PqsR	PQS
Γ_3	PqsR	HHQ
Γ_8	PqsR	DHQ
Γ_{18}	PqsE	HHQ
Γ_{19}	PqsE	PQS
Γ_{20}	PqsE	DHQ
Γ_{12}	PqsA	HHQ
Γ_{13}	PqsA	PQS
Γ_{14}	PqsA	DHQ
Γ_{15}	PqsBCD	HHQ
Γ_{16}	PqsBCD	PQS
Γ_{17}	PqsBCD	DHQ
Γ_9	PqsH	HHQ
Γ_{10}	PqsH	PQS
Γ_{11}	PqsH	DHQ

Γ_j	molecule
Γ_{21}	HHQ
Γ_{22}	PQS
Γ_{23}	DHQ

Figure 9.17.: **Possibilities for pyocyanin formation** that may either act alone or in combination as part of the Quorum sensing network. Reactions colored in green denote a positive regulation of pyocyanin by PqsE and consequently also by PqsR, reactions colored in red by PqsA, PqsBCD and consequently also by PqsR, as well as reactions colored in blue by PqsE, PqsA, PqsBCD, and PqsR. The table on the left lists all reactions Γ to form complex C7 and the table on the right to form complex C8.

$\mathcal{N}_3$, a complex between PQS and PqsE activates the formation of pyocyanin represented by reaction Γ_{19}. However, it has been reported that PqsE is not able to bind PQS [210]. Nevertheless, the actual ligands of PqsE remain uncharacterized. Network N2 uses HHQ instead of PQS to form the complex. The corresponding biosynthesis pathways of pyocyanin are shown in Figure 9.18. Here, low concentrations of HHQ or PQS ($\epsilon = 1$) are enough to form complex C6 between PqsE and the autoinducer and the degradation of C6:G4 is modeled as activation edge to HHQ or PQS. Figure 9.19 illustrates the pyocyanin biosynthesis in networks N9 and N10 using reactions Γ_{21} and Γ_{22}, respectively. In this case, complex C6 is built from PqsR, PqsE, and an autoinducer, whereby again ϵ equals one.

Network	used reactions	$\zeta_{HHQ \to PQS}$
$\mathcal{N}_1$	Γ_1, Γ_2	55%
$\mathcal{N}_2$	Γ_{18}	30%
$\mathcal{N}_3$	Γ_{19}	50%
$\mathcal{N}_4$	Γ_{20}, Γ_2	55%
$\mathcal{N}_5$	Γ_1, Γ_2, Γ_4	55%
$\mathcal{N}_6$	Γ_{18}, Γ_4	30%
$\mathcal{N}_7$	Γ_{19}, Γ_4	50%
$\mathcal{N}_8$	Γ_{20}, Γ_2, Γ_4	55%
$\mathcal{N}_9$	Γ_{21}	30%
$\mathcal{N}_{10}$	Γ_{22}	50%
$\mathcal{N}_{11}$	Γ_{23}, Γ_2	55%
$\mathcal{N}_{12}$	Γ_{24}, Γ_2	55%

Table 9.2.: **Different network topologies** with the used pyocyanin formation reactions of Figure 9.17 and the corresponding conversion frequency $\zeta_{HHQ \to PQS}$.

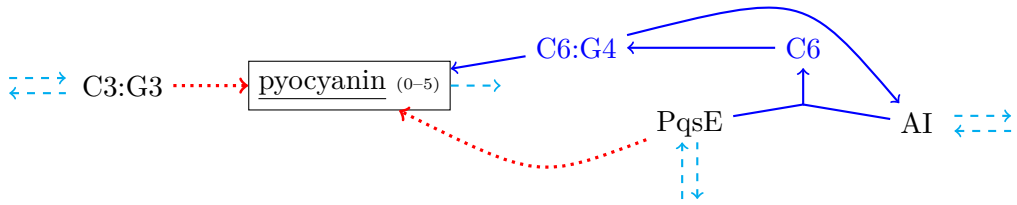


Figure 9.18.: **Networks $\mathcal{N}_3$ and $\mathcal{N}_4$:** Edges colored in blue represent newly included relations, dashed cyan edges unchanged connections, and dotted red edges removed reactions in comparison to network $\mathcal{N}_1$ which is given in Figure 9.13. Autoinducer denotes HHQ in the case of $\mathcal{N}_2$ and PQS in the case of $\mathcal{N}_3$.

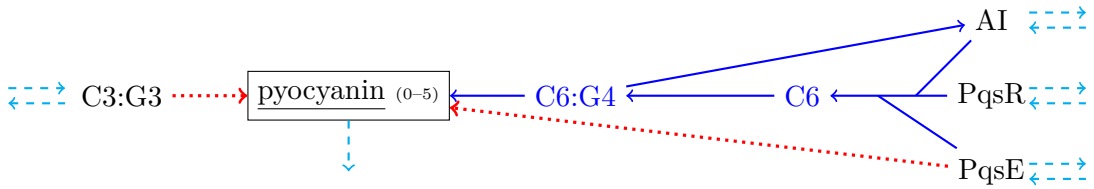


Figure 9.19.: **Networks $\mathcal{N}_9$ and $\mathcal{N}_{10}$:** Edges colored in blue represent newly included relations, dashed cyan edges unchanged connections, and dotted red edges removed reactions in comparison to network $\mathcal{N}_1$ which is given in Figure 9.13. Autoinducer denotes HHQ in the case of $\mathcal{N}_9$ and PQS in the case of $\mathcal{N}_{10}$.

When HHQ or PQS are used in further reactions, the conversion frequency $\zeta_{HHQ \to PQS}$ must be again adapted to the experimental observations. By testing different probabilities, the frequency of 55% was either confirmed or changed for each network. The resulting values that were used in the following are also listed in Table 9.2.

Behavior of Modified Networks

The behavior of the considered networks compared to existing knowledge is shown in Table 9.3. Due to the "or"–linkage with reaction Γ_2 that uses PqsR and PQS, PqsE$^-$ mutants form

Network	PqsABCD$^-$	PqsE$^-$	PqsR antagonists	PqsBCD inhibitors
$\mathcal{N}_1$	- - - - -	pyocyanin	pyocyanin	- - - - -
$\mathcal{N}_2$	- - - - -	- - - - -	pyocyanin	- - - - -
$\mathcal{N}_3$	- - - - -	- - - - -	pyocyanin	PQS
$\mathcal{N}_4$	- - - - -	pyocyanin	pyocyanin	- - - - -
$\mathcal{N}_5$	pyocyanin	pyocyanin	pyocyanin	- - - - -
$\mathcal{N}_6$	pyocyanin	pyocyanin	pyocyanin	- - - - -
$\mathcal{N}_7$	pyocyanin	pyocyanin	pyocyanin	PQS
$\mathcal{N}_8$	pyocyanin	pyocyanin	pyocyanin	- - - - -
$\mathcal{N}_9$	- - - - -	- - - - -	- - - - -	- - - - -
$\mathcal{N}_{10}$	- - - - -	- - - - -	- - - - -	PQS
$\mathcal{N}_{11}$	- - - - -	pyocyanin	pyocyanin	- - - - -
$\mathcal{N}_{12}$	- - - - -	pyocyanin	pyocyanin	- - - - -

Table 9.3.: **Behavior of modified networks:** Listed are species for which the predicted level is detectably higher than published observations. Entries colored in red denote average levels that reach their maximal possible levels, i.e., at least 100% of the corresponding wild type level. Entries colored in blue denote average levels that reach more than 95% of the corresponding wild type level and entries in green represent average levels that reach more than 85% of the corresponding wild type level. An inhibition level of 70% is used in the case of antagonists and an inhibition level of 30% in the case of enzyme inhibitors to compute the color scale.

pyocyanin at the average theoretical maximum level Υ in networks $\mathcal{N}_1$, $\mathcal{N}_4$, $\mathcal{N}_5$, $\mathcal{N}_8$, $\mathcal{N}_{11}$, and $\mathcal{N}_{12}$. Networks $\mathcal{N}_5$–$\mathcal{N}_8$ are based on networks $\mathcal{N}_1$–$\mathcal{N}_4$ and additionally include reaction Γ_4 as "or"–connection. Here, an activated PqsR is sufficient to form pyocyanin independent of the HHQ, PQS, or PqsE levels. Since PqsR is active for the whole simulation time after the starting phase, the pyocyanin production of knock–out mutants in the *pqsA–E* operon achieves Υ as a contradiction to the reported observations. Adding reaction Γ_4 does not change the level of the other species such that there is no difference to the corresponding network. For example, network $\mathcal{N}_5$ forms the same levels as network $\mathcal{N}_1$. In contrast, pyocyanin is never activated in the knock–out mutants of the *pqsA–E* operon for networks $\mathcal{N}_2$, $\mathcal{N}_3$, $\mathcal{N}_9$, and $\mathcal{N}_{10}$. In those four networks, the required components PqsR, PqsE, PqsA, and PqsBCD are combined in an "and"–relation (in Figure 9.17 colored in blue). PqsR antagonists with low inhibition levels still were not able to reduce the pyocyanin level for networks $\mathcal{N}_2$ and $\mathcal{N}_3$.

A complex between PqsR, PqsE, and an autoinducer up–regulates the formation of pyocyanin in networks $\mathcal{N}_9$ and $\mathcal{N}_{10}$. Therefore, they perform as described in the literature. The levels of pyocyanin and of internal and external HHQ and PQS were clearly reduced with a clear

dependence on the inhibition level. However, the PQS reducing rate of PqsBCD inhibitors in network $\mathcal{N}_{10}$ is less pronounced than of corresponding inhibitors in the other networks. As reported by Storz *et al.* [180], the predicted levels of HHQ were reduced more strongly than that of PQS considering the same inhibition level. Further, it seems that PqsBCD inhibitors are typically more reliable to combat HHQ and PQS, whereas PqsR antagonists are better to decrease the pyocyanin level.

Biological Interpretation

The most plausible networks $\mathcal{N}_9$ and $\mathcal{N}_{10}$ contain the reactions Γ_{21} and Γ_{22}, respectively. Since we assume that a complex between two receptor proteins and a ligand is quite unlikely, there are three possibilities to understand the relationship between PqsR, PqsE, and the autoinducer.

1. There are two different complexes that activate the transcription of two independent enzymes. Then, the first enzyme is responsible for building a certain metabolite. Afterwards, this metabolite is converted into pyocyanin or a precursor by the second enzyme. Figure 9.20 sketches this possibility.

2. There are two different complexes. One of those complexes activates the transcription of an enzyme and the other complex activates the transcription of a receptor protein. Then, the enzyme is responsible to build a certain metabolite. Afterwards, this metabolite forms a third complex with the built receptor protein. Finally, the third complex activates the formation of pyocyanin. Figure 9.21 sketches this possibility.

3. There is a complex that activates the transcription of an enzyme. Then, the enzyme is responsible to build a certain metabolite. Afterwards, this metabolite forms a second complex. This complex activates the formation of pyocyanin. Figure 9.22 sketches this possibility. Obviously, this way is a special case of the second case.

Hereby, PqsR forms one of the two first complexes together with some ligand and PqsE the other one. The autoinducer represents one of the two ligands. The missing components may be some functionally uncharacterized proteins. As the two homologous *phzA–G* operons are required for the production of the precursor phenazine–1–carboxylic acid and the enzymes PhzM and PhzS are responsible for the conversion of phenazine–1–carboxylic acid into pyocyanin [112], a direct or indirect linkage of the unknown proteins to this biosynthesis pathway is probable. Further, the proteins MvaT and MvaU are also necessary to build pyocyanin [96] such that a connection of the unknown proteins to those is also possible.

Hence, it seems that the formation of pyocyanin is regulated by the proteins PqsR and PqsE. For this, PqsE acts as a regulatory protein and an autoinducer is used as ligand.

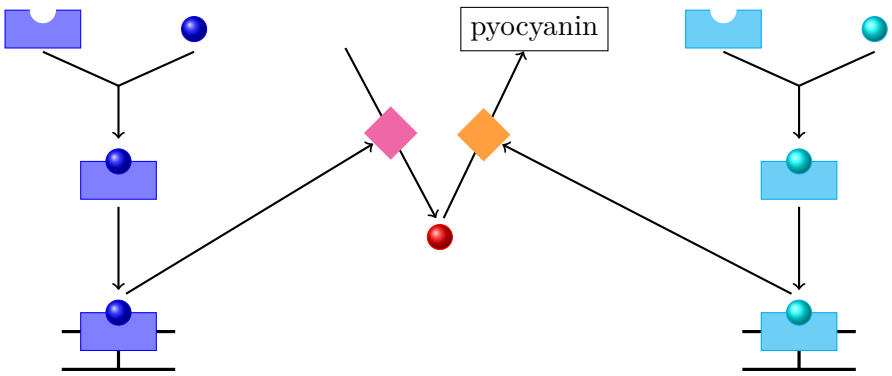


Figure 9.20.: **Interpretation possibility 1 of reactions Γ_{21} and Γ_{22}:** A first complex (shown in blue) activates the synthase (shown as magenta colored square) of a precursor (shown as red ball) which is converted by a second enzyme (orange square) into pyocyanin. The transcription of the second enzyme is up–regulated by a second complex (colored in cyan).

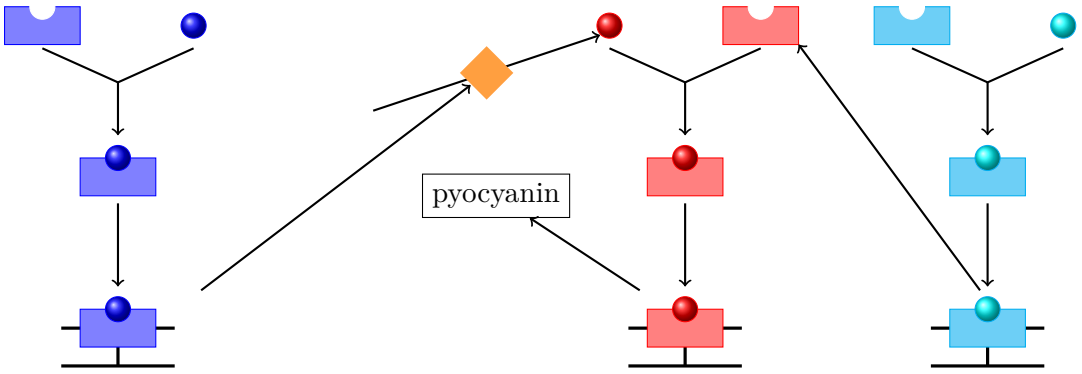


Figure 9.21.: **Interpretation possibility 2 of reactions Γ_{21} and Γ_{22}:** A first complex (shown in blue) activates the synthase (shown as orange colored square) of a ligand (shown as red ball) which binds to a receptor (also colored in red). The transcription of the receptor is up–regulated by the second complex (colored in cyan). The third complex activates the pyocyanin formation.

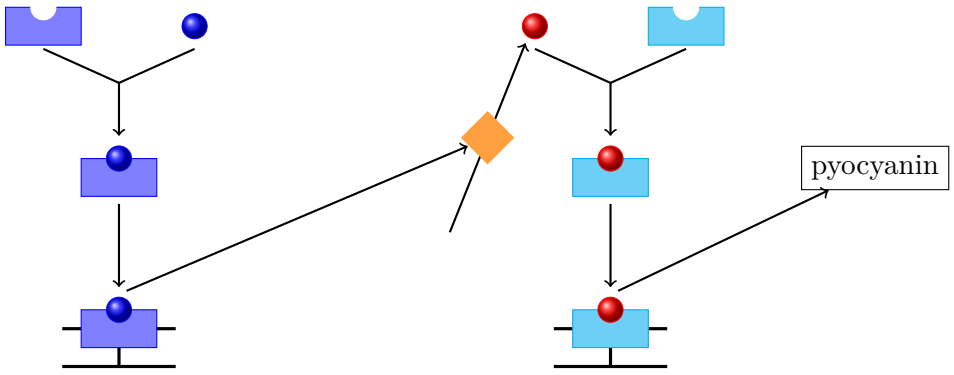


Figure 9.22.: **Interpretation possibility 3 of reactions Γ_{21} and Γ_{22}:** A first complex (shown in blue) activates the synthase (shown as orange colored square) of a ligand (shown as red ball) which binds to a second receptor (colored in cyan). The pyocyanin formation is up–regulated by the second complex.

9.2.8. Mixed Cell Cultures

Here, we studied the influence of single mutants that are resistant against drugs, such as Quorum sensing inhibitors, on the entire bacterial culture. For example, we tested how large must be the fraction of mutants such that an infection is dangerous for the host. For this, we analyzed the transport effect of PqsBCD$^-$ knock–out mutants on the pyocyanin formation. Figure 9.23 shows that the more PqsBCD$^-$ mutants are in the culture the lower is the average pyocyanin level. As PqsBCD$^-$ mutants are not able to produce HHQ and PQS, no pyocyanin is

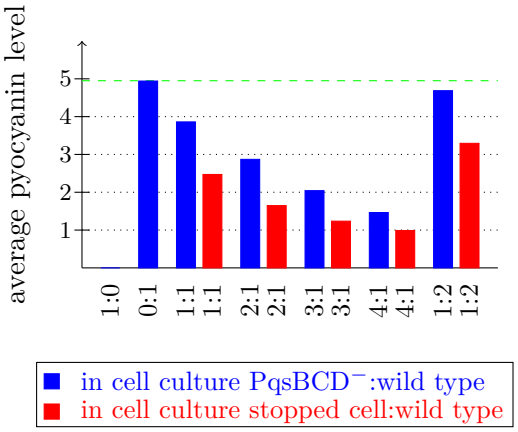


Figure 9.23.: **Influence of autoinducer exchange in mixed cell cultures** on the average level of pyocyanin. The levels were averaged over all starting cells. A stopped cell forms nothing. The dashed green colored line denotes the average theoretical maximum level Υ.

built in these mutants. However, the wild type forms enough autoinducers that are transported to the environment. Then, the mutants can use these autoinducers to activate the pyocyanin formation. Therefore, the average pyocyanin levels of mixed cell cultures (shown in blue) are higher than a culture in which the remaining cells produce nothing (shown in red).

9.2.9. Random Mutations

A simulation in which the growth parameters and the conversion frequencies of the network may be changed by mutations with a mutation rate ν of 25 showed that the growth rate κ was modified about 23 times, the maximum number of cell divisions δ_{max} was changed about twelve times, and the network was updated about eight times. Here, the total number of cells were clearly changed. Further, the standard deviation of species levels achieved by different random numbers was detectably higher than the corresponding standard deviation in a simulation without any mutations. For example, the autoinducer levels were stronger affected by mutations than the levels of the virulence factors.

Moreover, the start point of individual cells was completely flexible in the case of randomly occurring mutations, whereas it followed fixed exponential rules otherwise.

10. Summary

An extended multi–level logical formalism with a minimum number of required parameters was used to simulate the Quorum sensing in the multi–resistant pathogen *Pseudomonas aeruginosa*. For this, we included transport, growth, and randomly occurring mutation processes. The average level of autoinducers and virulence factors formed by the wild type was compared to knock–out and gain–of–function mutants as well as to Quorum sensing inhibitors.

After a transient effect of about 100 time steps, the system behaves independent of the initial system state. For statistical reasons, a total simulation time of 600 was chosen. A parameter scan over all required parameters showed that the autoinducer formation is independent of the growth processes and degradation influences the behavior only when it happens very frequently. In contrast, the frequency to convert HHQ to PQS is very important for the internal and external autoinducer levels. By comparing the simulation results experimental data, the frequency was set to 55%.

The wild type formed external autoinducers AI–1 and AI–2 as well as virulence factors in huge amounts. Further, the external levels of the autoinducers HHQ and PQS were also relatively large. In agreement with literature, a knock–out of the proteins $LasI^-$, $LasR^-$, $PqsA^-$, $PqsBCD^-$, or $PqsR^-$, which are all involved in Quorum sensing, stopped the production of PQS and the virulence factor pyocyanin. However, the $PqsE^-$ mutant produced pyocyanin at a too high level in our model.

The considered PqsBCD inhibitors that directly block the autoinducer biosynthesis clearly decreased the external HHQ and PQS levels as well as the pyocyanin level with a dependence on the inhibition level. Antagonists of the activating receptor PqsR were not able to reduce the level of pyocyanin except for a very high inhibition level, which is again a contradiction to literature.

Since the pyocyanin formation of our model does not fulfill all requirements, other network topologies were discussed. The present study suggests that PqsR and PqsE activate a part of the pyocyanin biosynthesis acting as a receptor and together with an autoinducer representing a ligand.

In a mixed cell culture that contains wild type cells and mutants that are alone not able to form pyocyanin, the average level of pyocyanin was higher than the wild type cells can produce, i.e., the autoinducers were built in the wild type, then transported into the mutants, and there used for pyocyanin formation.

Part IV.

Conclusions

11. Functional Prediction and Classification of Membrane Proteins

Since transport processes are important for organisms and the effort of experiments to annotate membrane proteins is very high, computational tools are required to functionally classify them.

11.1. Summary

In this study, we considered membrane proteins from the Transporter Classification Database with a focus on the subfamilies 1.A.1, 2.A.1, and 3.A.1 as well as *Arabidopsis thaliana* transporters and carriers with a focus on the substrate classes amino acids, oligopeptides, phosphates, and hexoses. Several features related to the amino acid composition were used to statistically analyze and functionally classify the data sets. The features may include neighborhood correlations, physicochemical properties, and conservation profiles. Alternatively, the amino acid composition may be split into different sequence regions depending on the transmembrane localization or on conservation levels. For this, statistical tests, such as analysis of variance and Wilcoxon–Mann–Whitney, and classical machine learning techniques, such as support vector machines, principal component analysis, and hierarchical clustering, were applied. Additionally, a ranking method was developed to handle small data sets in the context of functional classification and prediction purely based on amino acid sequences. Here, a combined ranking is generated using the Euclidean distance as similarity measure between a considered protein and members of a certain functional class. Hereby, the functional class is either represented as an average search profile or as individual compositions depending on the underlying data set. Commonly used quality measurements were estimated to measure the prediction quality and further compared to random classifications by t–test and a simple integer value ξ. Moreover, the connection of loop occurrence and length with the function of a membrane protein was investigated using so–called multiple binary string alignments. In these alignments, a non–transmembrane region is mapped to the corresponding non–transmembrane region of another protein and a transmembrane region to another transmembrane region, respectively.

Based on results for some proteins with known three–dimensional structures and homology searches by BLAST, we showed that the amino acid composition is a suitable feature to functionally annotate proteins of the considered data sets. Further, clearly more and stronger

significant differences between proteins with different function were identified in transmembrane segments than in the full sequence. The ranking method resulted in quite high quality measures. In the case of the *Arabidopsis thaliana* subsets, it achieved a sensitivity and an accuracy of more than 75% for the original amino acid composition and more than 80% for extended amino acid compositions. In contrast, randomly generated subsets gave clearly lower accuracies and sensitivities of less than 60% and around 30–35%, respectively. Overall, the profile–based amino acid composition (MSA–AAC) yielded the best results. Filtering between transmembrane and non–transmembrane regions could also improve the classification quality, whereby this is strongly pronounced for sets that are homogeneous in their transmembrane segment distribution. Here, the amino acid composition that only takes transmembrane segments into account behaved best. Nevertheless, in the absence of three–dimensional structures, the performance is strongly based on the quality of topology prediction methods due to missing three–dimensional information.

11.2. Outlook

Our tool makes suggestions for possible substrates and is able to narrow down the large search space of experiments. Obviously, it is possible to further refine the features. However, we suggest that sequence–based methods should ideally be combined with structure–based methods to exploit the increasing number of resolved three–dimensional structures. Moreover, proteins with the same cellular function are often co–expressed. Therefore, co–expression analysis can improve the annotation of unknown proteins besides sequence– or structure–based features.

As a superfamily primarily describes the corresponding transport mechanism instead of the substrate itself, the question whether substrate predictions are also possible between different organisms remains unsolved. For this, a positive set of a well–studied organism can be used to fish in a full data set of an unannotated organism. Preliminary results suggested that this works even between the not directly related organisms *Arabidopsis thaliana* and *Escherichia coli*. In comparison, we assume that a cross–species annotation is feasible between *Arabidopsis thaliana* and crop plants, such as *Oryza sativa* or *Zea mays*, due to the genomic similarity. In future, this could be improved by including statistical knowledge about special characteristics in the amino acid composition of the considered organisms. Based on homology and without comprising any organism specificities, Barghash and Helms functionally assigned membrane transporters between the different model organisms *Escherichia coli*, *Saccaromyces cereviseae*, and *Arabidopsis thaliana* [11].

Instead of multiple binary string alignments, it is possible to study loop length patterns in a principal component analysis. Here, each loop represents a feature that contains the corresponding loop length. To avoid problems caused by different numbers of transmembrane segments, at the end of each protein, loops with size zero can be added. Nevertheless, one can expect that proteins containing the same number of transmembrane segments tend to be clustered together.

12. Quorum Sensing in *Pseudomonas aeruginosa*

The pathogenicity and the multi–resistance against most current antibiotics of *Pseudomonas aeruginosa* provide a strong motivation for research on Quorum sensing that regulates the formation of virulence factors.

12.1. Summary

Due to a lack of experimental data, we decided to use a mostly parameter–free discrete rule–based approach to model the Quorum sensing in *Pseudomonas aeruginosa*. At first, a simple Boolean model was applied for an attractor analysis of the independent positive feedback loops. Since the simplifications of a Boolean network appeared too drastic to simulate the hierarchical connection of the three Quorum sensing systems in *Pseudomonas aeruginosa*, a multi–level logical formalism was extended and adopted to the considered pathways. In this connection, thresholds allow a concentration dependence in the signaling processes. Those enzymatic and regulatory propagations in a single cell were embedded into a system that represents a complete bacterial culture. Here, transport processes between individual cells and the common environment enable the bacterial communication, which is an important task in Quorum sensing. Further, growth processes and randomly occurring mutations were modeled. Besides the wild type network, knock–out and gain–of–function mutants, which are described in literature, as well as Quorum sensing inhibitors were considered.

After a detailed scan of parameters and initialization levels, the switching–on behavior was analyzed. The levels of the two autoinducers HHQ and PQS as well as the level of the virulence factor pyocyanin, which were formed in huge amounts by the wild type, could be reduced by inhibitors of the autoinducer biosynthesis enzymes PqsB, PqsC, or PqsD. In the case of antagonists of the receptor PqsR, the pyocyanin level is only reduced at high inhibition levels. As this is a contradiction to literature and as $PqsE^{-}$ knock–out mutants produce too much pyocyanin, it seems that the pyocyanin biosynthesis is incomplete. Therefore, several other network topologies were studied in detail. They suggest that PqsR and PqsE may activate as receptors the pyocyanin biosynthesis and an autoinducer behaves as a ligand. Further, we showed that PqsBCD inhibitors were more reliable to block the formation of HHQ and PQS, whereas PqsR antagonists seem to be more suitable in reducing the pyocyanin level.

Additionally, a mixed cell culture with wild type cells and mutants was considered. Those mutants are alone not able to form HHQ, PQS, and pyocyanin. This culture formed more pyocyanin than it is possible for the contained wild type cells. Thus, the wild type built the autoinducers, which were transported into the mutants. Then, the autoinducers activate the pyocyanin biosynthesis in the mutant.

12.2. Outlook

Our current understanding of virulence and resistance appears annoyingly incomplete. Therefore, a computational study is useful to show how resistance may develop in *Pseudomonas aeruginosa* and how resistance may be avoided against drugs. For this, mutations have to be considered, antibiotics must be added to the network, and the selection pressure has to be simulated. However, experimental data are required about how frequently mutations occur depending on certain conditions, such as nutrients and medication, and sequencing data to detect what kind of mutations happen. Then, it appears possible to determine the advantage of Quorum sensing inhibitors in comparison to growth inhibiting antibiotics that cause resistances due to an increased selection pressure. Further, it appears possible to analyze the impact of resistance against a Quorum sensing inhibitor. This could be followed by a prediction of a certain medication that avoids resistance development.

Part V.

Bibliography

[1] J. Abramson and E.M. Wright. Structure and function of Na+-symporters with inverted repeats. Curr. Opin. Struct. Biol., 19:425–432, 2009.

[2] A.M. Albus, E.C. Pesci, L.J. Runyen-Janecky, SE West, and B.H. Iglewski. Vfr controls quorum sensing in Pseudomonas aeruginosa. J. Bacteriol., 179:3928–3935, 1997.

[3] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. J. Mol. Biol., 215:403–410, 1990.

[4] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Ac. Res., 25:3389–3402, 1997.

[5] K. Anguige, J.R. King, and J.P. Ward. A multi-phase mathematical model of quorum sensing in a maturing Pseudomonas aeruginosa biofilm. Math. Biosciences, 203:240–276, 2006.

[6] K. Anguige, J.R. King, J.P. Ward, and P. Williams. Mathematical modelling of therapies targeted at bacterial quorum sensing. Math. Biosciences, 192:39–83, 2004.

[7] L.C.M. Antunes, R.B.R. Ferreira, M.M.C. Buckner, and B.B. Finlay. Quorum sensing in bacterial virulence. Microbiol., 156:2271–2282, 2010.

[8] P. Argos, J.K. Rao, and P.A. Hargrave. Structural Prediction of Membrane-Bound Proteins. Eur. J. Biochem., 128:565–575, 1982.

[9] H. Atlan, F. Fogelman-Soulie, J. Salomon, and G. Weisbuch. Random boolean networks. Cybernetics and System, 12:103–121, 1981.

[10] I. Bahar, A.R. Atilgan, R.L. Jernigan, and B. Erman. Understanding the recognition of protein structural classes by amino acid composition. Proteins, 29:172–185, 1997.

[11] A. Barghash and V. Helms. Transferring functional annotations of membrane transporters on the basis of sequence similarity and sequence motifs. in progress, 2013.

[12] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. Nucl. Ac. Res., 28:235–242, 2000.

[13] F. Bernardini, M. Gheorghe, and N. Krasnogor. Quorum sensing P systems. Theo. Comp. Sci., 371:20–33, 2007.

[14] A. Bernsel, H. Viklund, J. Falk, E. Lindahl, G. Von Heijne, and A. Elofsson. Prediction of membrane-protein topology from first principles. PNAS, 105:7177–7181, 2008.

[15] A. Bernsel, H. Viklund, A. Hennerdal, and A. Elofsson. TOPCONS: consensus prediction of membrane protein topology. Nucl. Ac. Res., pages W465–W468, 2009.

[16] C.M. Bishop. Pattern Recognition and Machine Learning, pages 325–358. Springer, 2006.

[17] D.R. Bush. Proton-coupled sugar and amino acid transporters in plants. Ann. Rev. Plant Physiol. Plant Mol. Biol., 44:513–542, 1993.

[18] W.A. Catterall. Structure and function of voltage-gated ion channels. Trends Neurosc., 16:500–506, 1993.

[19] A. Chakrabartty, T. Kortemme, and R.L. Baldwin. Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. Prot. Sci., 3:843–852, 1994.

[20] A. Chakrabartty, J.A. Schellman, and R.L. Baldwin. Large differences in the helix propensities of alanine and glycine. Nature, 351:586–588, 1991.

[21] A.K. Chamberlain and J.U. Bowie. Asymmetric amino acid compositions of transmembrane β-strands. Prot. Sci., 13:2270–2274, 2004.

[22] D.L. Chopp, M.J. Kirisits, B. Moran, and M.R. Parsek. A mathematical model of quorum sensing in a growing bacterial biofilm. J. Ind. Microbiol. Biotech., 29:339–346, 2002.

[23] K.C. Chou. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem. Biophys. Res. Comm., 278:477–483, 2000.

[24] K.C. Chou. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins, 43:246–255, 2001.

[25] K.C. Chou. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics, 21:10–19, 2005.

[26] K.C. Chou and H.B. Shen. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem. Biophys. Res. Com., 360:339–345, 2007.

[27] P.Y. Chou and G.D. Fasman. Empirical predictions of protein conformation. Ann. Rev. Biochem., 47:251–276, 1978.

[28] J.W. Costerton, P.S. Stewart, and E.P. Greenberg. Bacterial biofilms: a common cause of persistent infections. Science, 284:1318–1322, 1999.

[29] J.O. Dada and P. Mendes. Multi-scale modelling and simulation in systems biology. Integr. Biol., 3:86–96, 2011.

[30] G.L. Daikos, V.T. Lolans, and G.G. Jackson. Alterations in outer membrane proteins of Pseudomonas aeruginosa associated with selective resistance to quinolones. Antimicrob. Agents Chemother., 32:785–787, 1988.

[31] A.L. Davidson, E. Dassa, C. Orelle, and J. Chen. Structure, function, and evolution of bacterial ATP-binding cassette systems. Microbiol. Mol. Biol. Rev., 72:317–364, 2008.

[32] M.N. Davies, A. Secker, A.A. Freitas, M. Mendao, J. Timmis, and D.R. Flower. On the hierarchical classification of G protein-coupled receptors. Bioinformatics, 23:3113–3118, 2007.

[33] M.N. Davies, A. Secker, M. Halling-Brown, D.S. Moss, A.A. Freitas, J. Timmis, E. Clark, and D.R. Flower. GPCRTree: online hierarchical classification of GPCR function. BMC Res. Notes, 1:67, 2008.

[34] R.J.P. Dawson and K.P. Locher. Structure of a bacterial multidrug ABC transporter. Nature, 443:180–185, 2006.

[35] S. de Bentzmann and P. Plésiat. The Pseudomonas aeruginosa opportunistic pathogen and human infections. Environ. Microbiol., 13:1655–1665, 2011.

[36] B. De Hertogh, E. Carvajal, E. Talla, B. Dujon, P. Baret, and A. Goffeau. Phylogenetic classification of transporters and other membrane proteins from Saccharomyces cerevisiae. Funct. Integr. Genom., 2:154–170, 2002.

[37] T. De Kievit, P.C. Seed, J. Nezezon, L. Passador, and B.H. Iglewski. RsaL, a novel repressor of virulence gene expression in Pseudomonas aeruginosa. J. Bacteriol., 181:2175–2184, 1999.

[38] T.R. De Kievit and B.H. Iglewski. Bacterial quorum sensing in pathogenic relationships. Infect. Immun., 68:4839–4849, 2000.

[39] E. Déziel, F. Lépine, S. Milot, J. He, M.N. Mindrinos, R.G. Tompkins, and L.G. Rahme. Analysis of Pseudomonas aeruginosa 4-hydroxy-2-alkylquinolines (HAQs) reveals a role for 4-hydroxy-2-heptylquinoline in cell-to-cell communication. PNAS, 101:1339–1344, 2004.

[40] S.P. Diggle, K. Winzer, S.R. Chhabra, K.E. Worrall, M. Cámara, and P. Williams. The Pseudomonas aeruginosa quinolone signal molecule overcomes the cell density-dependency of the quorum sensing hierarchy, regulates rhl-dependent genes at the onset of stationary phase and can be produced in the absence of LasR. Mol. Microbiol., 50:29–43, 2003.

[41] J.D. Dockery and J.P. Keener. A mathematical model for quorum sensing in Pseudomonas aeruginosa. Bull. Math. Biol., 63:95–116, 2001.

[42] Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J. Mol. Biol., 347:827–839, 2005.

[43] S.R. Eddy. A new generation of homology search tools based on probabilistic inference. In Genome Inform, volume 23, pages 205–211, 2009.

[44] M. Eilers, S.C. Shekar, T. Shieh, S.O. Smith, and P.J. Fleming. Internal packing of helical membrane proteins. PNAS, 97:5796–5801, 2000.

[45] R.P. Elrod and A.C. Braun. Pseudomonas aeruginosa: its role as a plant pathogen. J. Bacteriol., 44:633–645, 1942.

[46] D.M. Engelman. Membranes are more mosaic than fluid. Nature, 438:578–580, 2005.

[47] G.L. Fan and Q.Z. Li. Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. J. Theo. Biol., 304:88–95, 2012.

[48] J.M. Farrow III, Z.M. Sund, M.L. Ellison, D.S. Wade, J.P. Coleman, and E.C. Pesci. PqsE functions independently of PqsR-Pseudomonas quinolone signal and enhances the rhl quorum-sensing system. J. Bacteriol., 190:7043–7051, 2008.

[49] A. Fauré, A. Naldi, F. Lopez, C. Chaouiya, A. Ciliberto, and D. Thieffry. Modular logical modelling of the budding yeast cell cycle. Mol. BioSyst., 5:1787–1796, 2009.

[50] J.A. Fozard, M. Lees, J.R. King, and B.S. Logan. Inhibition of quorum sensing in a computational biofilm simulation. Biosystems, 109:105–114, 2012.

[51] M.R. Frederick, C. Kuttler, B.A. Hense, and H.J. Eberl. A mathematical model of quorum sensing regulated EPS production in biofilm communities. Theo. Biol. Med. Model., 8, 2011.

[52] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. J. Comp. Biol., 7:601–620, 2000.

[53] W.C. Fuqua, Stephen C. Winans, and E.P. Greenberg. Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators. J. Bacteriol., 176:269–275, 1994.

[54] L.A. Gallagher, S.L. McKnight, M.S. Kuznetsova, E.C. Pesci, and C. Manoil. Functions required for extracellular quinolone signaling by Pseudomonas aeruginosa. J. Bacteriol., 184:6472–6480, 2002.

[55] M.J. Gambello and B.H. Iglewski. Cloning and characterization of the Pseudomonas aeruginosa lasR gene, a transcriptional activator of elastase expression. J. Bacteriol., 173:3000–3009, 1991.

[56] Z. Genfa, X. Xinhua, and Z. Chun-Ting. A weighting method for predicting protein structural class from amino acid composition. Eur. J. Biochem., 210:747–749, 1992.

[57] H. Genrich, R. Küffner, and K. Voss. Executable Petri net models for the analysis of metabolic pathways. Int. J. STTT, 3:394–404, 2001.

[58] G.D. Geske, R.J. Wezeman, A.P. Siegel, and E. Helen. Small molecule inhibitors of bacterial quorum sensing and biofilm formation. J. Am. Chem. Soc., 127:12762–12763, 2005.

[59] A.G. Gonzalez, A. Naldi, L. Sanchez, D. Thieffry, and C. Chaouiya. GINsim: A software suite for the qualitative modelling, simulation and analysis of regulatory networks. Biosystems, 84:91–100, 2006.

[60] G.W. Gould and G.D. Holman. The glucose transporter family: structure, function and tissue-specific expression. J. Biochem., 295:329–341, 1993.

[61] J.K. Griffith, M.E. Baker, D.A. Rouch, M.G.P. Page, R.A. Skurray, I.T. Paulsen, K.F. Chater, S.A. Baldwin, and P.J.F. Henderson. Membrane transport proteins: implications of sequence comparisons. Curr. Opin. Cell Biol., 4:684–695, 1992.

[62] M.M. Gromiha and Y. Yabuki. Functional discrimination of membrane proteins using machine learning techniques. BMC Bioinformatics, 9:135, 2008.

[63] T. Handorf and E. Klipp. Modeling mechanistic biological networks: An advanced Boolean approach. Bioinformatics, 28:557–563, 2012.

[64] M. Hayat, A. Khan, and M. Yeasin. Prediction of membrane proteins using split amino acid and ensemble classification. Amino Acids, 42:2447–2460, 2012.

[65] S. Hayat, Y. Park, and V. Helms. Statistical analysis and exposure status classification of transmembrane beta barrel residues. Comp. Biol. Chem., 35:96–107, 2011.

[66] R. Hedrich and J.I. Schroeder. The physiology of ion channels and electrogenic pumps in higher plants. Ann. Rev. Plant Biol., 40:539–569, 1989.

[67] S. Heeb, M.P. Fletcher, S.R. Chhabra, S.P. Diggle, P. Williams, and M. Cámara. Quinolones: from antibiotics to autoinducers. FEMS Microbiol. Rev., 35:247–274, 2011.

[68] S. Henkel, T. Nägele, I. Hörmiller, T. Sauter, O. Sawodny, M. Ederer, and A.G. Heyer. A systems biology approach to analyse leaf carbohydrate metabolism in Arabidopsis thaliana. J. Bioinformatics Systems Biol., 2011:1–10, 2011.

[69] M. Hentzer, H. Wu, J.B. Andersen, K. Riedel, T.B. Rasmussen, N. Bagge, N. Kumar, M.A. Schembri, Z. Song, P. Kristoffersen, et al. Attenuation of Pseudomonas aeruginosa virulence by quorum sensing inhibitors. EMBO, 22:3803–3815, 2003.

[70] G.Z. Hertz and G.D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics, 15:563–577, 1999.

[71] H. Hotelling. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol., 24:498–520, 1933.

[72] Y. Huang, J. Cai, L. Ji, and Y. Li. Classifying G-protein coupled receptors with bagging classification tree. Comp. Biol. Chem., 28:275–280, 2004.

[73] G.R. Iversen and H. Norpoth. Analysis of variance, volume 1. Sage Publications, Incorporated, 1987.

[74] M.M. Javadpour, M. Eilers, M. Groesbeek, and S.O. Smith. Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association. Biophys. J., 77:1609–1618, 1999.

[75] M.L. Jennings. Topography of membrane proteins. Ann. Rev. Biochem., 58:999–1025, 1989.

[76] Y. Jiang, A. Lee, J. Chen, V. Ruta, M. Cadene, B.T. Chait, and R. MacKinnon. X-ray structure of a voltage-dependent K+ channel. Nature, 423:33–41, 2003.

[77] Y. Jiang, J. Pjesivac-Grbovic, C. Cantrell, and J.P. Freyer. A multiscale model for avascular tumor growth. Biophysical J., 89:3884–3894, 2005.

[78] D.T. Jones. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics, 23:538, 2007.

[79] D.T. Jones, W.R. Taylor, and J.M. Thornton. A mutation data matrix for transmembrane proteins. FEBS Letters, 339:269–275, 1994.

[80] D. Juretic, L. Zoranic, and D. Zucic. Basic charge clusters and predictions of membrane protein topology. J. Chem. Inf. Comp. Sci, 42:620–632, 2002.

[81] K. Katoh, K. Misawa, K. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucl. Ac. Res., 30:3059–3066, 2002.

[82] S.A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. J. Theo. Biol., 22:437–467, 1969.

[83] M. Kesarwani, R. Hazan, J. He, Y.A. Que, Y. Apidianakis, B. Lesic, G. Xiao, V. Dekimpe, S. Milot, E. Deziel, et al. A Quorum Sensing Regulated Small Volatile Molecule Reduces Acute Virulence and Promotes Chronic Infection Phenotypes. PLoS Pathogens, 7:e1002192, 2011.

[84] J.A. Killian and G. von Heijne. How proteins adapt to a membrane-water interface. Trends Biochem. Sci., 25:429–434, 2000.

[85] T. Klein, C. Henn, J.C. de Jong, C. Zimmer, B. Kirsch, C.K. Maurer, D. Pistorius, R. Müller, A. Steinbach, and R.W. Hartmann. Identification of Small-Molecule Antagonists of the Pseudomonas aeruginosa Transcriptional Regulator PqsR: Biophysically Guided Hit Discovery and Optimization. ACS Chem. Biol., 7:1496–1501, 2012.

[86] A.J. Koerber, J.R. King, J.P. Ward, P. Williams, J.M. Croft, and R.E. Sockett. A mathematical model of partial-thickness burn-wound infection by Pseudomonas aeruginosa: quorum sensing and the build-up to invasion. Bull. Math. Biol., 64:239–259, 2002.

[87] K. Krickeberg and H. Ziezold. Stochastische Methoden, page 170. Springer, 1994.

[88] A. Krogh, B.È. Larsson, G. Von Heijne, E.L.L. Sonnhammer, et al. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol., 305:567–580, 2001.

[89] P.W. Kuchel. Schaum's outline of theory and problems of biochemistry. Schaum's Outline Series, 1997.

[90] C. Kuttler and B.A. Hense. Interplay of two quorum sensing regulation systems of Vibrio fischeri. J. Theo. Biol., 251:167–180, 2008.

[91] H. Lähdesmäki, S. Hautaniemi, I. Shmulevich, and O. Yli-Harja. Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks. Signal Processing, 86:814–834, 2006.

[92] M.G. Lamarche and E. Déziel. MexEF-OprN Efflux Pump Exports the Pseudomonas Quinolone Signal (PQS) Precursor HHQ (4-hydroxy-2-heptylquinoline). PLoS one, 6:e24310, 2011.

[93] G.W. Lau, D.J. Hassett, H. Ran, and F. Kong. The role of pyocyanin in Pseudomonas aeruginosa infection. Trends Mol. Med., 10:599–606, 2004.

[94] T.J. Lee, I. Paulsen, and P. Karp. Annotation-based inference of transporter function. Bioinformatics, 24:i259–i267, 2008.

[95] G.D. Leonard, T. Fojo, and S.E. Bates. The role of ABC transporters in clinical practice. Oncologist, 8:411–424, 2003.

[96] C. Li, H. Wally, S.J. Miller, and C.D. Lu. The multifaceted proteins MvaT and MvaU, members of the H-NS family, control arginine metabolism, pyocyanin synthesis, and prophage activation in Pseudomonas aeruginosa PAO1. J. Bacteriol., 191:6211–6218, 2009.

[97] F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang. The yeast cell-cycle network is robustly designed. PNAS, 101:4781–4786, 2004.

[98] H. Li, V.A. Benedito, M.K. Udvardi, and P.X. Zhao. TransportTP: A two-phase classification approach for membrane transporter prediction and characterization. BMC Bioinformatics, 10:418, 2009.

[99] H. Li, X. Dai, and X. Zhao. A nearest neighbor approach for automated transporter prediction and categorization from protein sequences. Bioinformatics, 24:1129–1136, 2008.

[100] H. Lin and Q.Z. Li. Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J. Comp. Chem., 28:1463–1466, 2007.

[101] J. Lin, S. Huang, and Q. Zhang. Outer membrane proteins: key players for bacterial adaptation in host niches. Microbes Infect., 4:325–331, 2002.

[102] S. Lin and J. Ding. Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA studies. Biometrics, 65:9–18, 2009.

[103] Q. Liu, J. Cui, Q. Yang, and Y. Xu. In-silico prediction of blood-secretory human proteins using a ranking algorithm. BMC Bioinformatics, 11:250, 2010.

[104] M.A. Lomize, A.L. Lomize, I.D. Pogozheva, and H.I. Mosberg. OPM: orientations of proteins in membranes database. Bioinformatics, 22:623–625, 2006.

[105] C. Lu, B. Kirsch, C. Zimmer, J.C. de Jong, C. Henn, C.K. Maurer, M. Müsken, S. Häussler, A. Steinbach, and R.W. Hartmann. Discovery of Antagonists of PqsR, a Key Player in 2-Alkyl-4-quinolone-Dependent Quorum Sensing in Pseudomonas aeruginosa. Chem. Biol., 19:381–390, 2012.

[106] R.M. Maier and G. Soberon-Chavez. Pseudomonas aeruginosa rhamnolipids: biosynthesis and potential applications. Appl. Microbiol. Biotechnol., 54:625–633, 2000.

[107] H.B. Mann and D.R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat., 18:50–60, 1947.

[108] A.R. Manolescu, K. Witkowska, A. Kinnaird, T. Cessford, and C. Cheeseman. Facilitated hexose transporters: new perspectives on form and function. Physiology, 22:234–240, 2007.

[109] M.D. Marger and M.H. Saier. A major superfamily of transmembrane facilitators that catalyse uniport, symport and antiport. Trends Biochem. Sci., 18:13–20, 1993.

[110] A. Marsico, A. Henschel, C. Winter, A. Tuukkanen, B. Vassilev, K. Scheubert, and M. Schroeder. Structural fragment clustering reveals novel structural and functional motifs in alpha-helical transmembrane proteins. BMC Bioinformatics, 11:204, 2010.

[111] P.L. Martelli, P. Fariselli, and R. Casadio. An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. Bioinformatics, 19:i205–i211, 2003.

[112] D.V. Mavrodi, R.F. Bonsall, S.M. Delaney, M.J. Soule, G. Phillips, and L.S. Thomashow. Functional analysis of genes for biosynthesis of pyocyanin and phenazine-1-carboxamide from Pseudomonas aeruginosa PAO1. J. Bacteriol., 183:6454–6465, 2001.

[113] S.L. McKnight, B.H. Iglewski, and E.C. Pesci. The Pseudomonas quinolone signal regulates rhl quorum sensing in Pseudomonas aeruginosa. J. Bacteriol., 182:2702–2708, 2000.

[114] G. Medina, K. Juárez, and G. Soberón-Chávez. The Pseudomonas aeruginosa rhlAB operon is not expressed during the logarithmic phase of growth even in the presence of its activator RhlR and the autoinducer N-butyryl-homoserine lactone. J. Bacteriol., 185:377–380, 2003.

[115] P. Melke, P. Sahlin, A. Levchenko, and H. Jönsson. A cell-based model for quorum sensing in heterogeneous bacterial colonies. PLoS Comp. Biol., 6:e1000819, 2010.

[116] L. Mendoza and E.R. Alvarez-Buylla. Dynamics of the Genetic Regulatory Network for Arabidopsis thaliana Flower Morphogenesis. J. Theo. Biol., 193:307–319, 1998.

[117] J. Metzger, N.S. Schaadt, S. Hayat, and V. Helms. Predicting Structural and Functional Properties of Membrane Proteins from Protein Sequence. Ann. Reports Comp. Chem., page 39, 2011.

[118] M.B. Miller and B.L. Bassler. Quorum sensing in bacteria. Ann. Rev. Microbiol., 55:165–199, 2001.

[119] K.H. Nealson, T. Platt, and J. Hastings. Cellular control of the synthesis and activity of the bacterial luminescent system. J. Bacteriol., 104:313–322, 1970.

[120] H. Nikaido. Outer membrane barrier as a mechanism of antimicrobial resistance. Antimicrob. Agents Chemother., 33:1831, 1989.

[121] K. Nishikawa, Y. Kubota, and T. Ooi. Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. J. Biochem., 94:981–995, 1983.

[122] T. Nugent and D.T. Jones. Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics, 10:159, 2009.

[123] U.A. Ochsner, A. Fiechter, and J. Reiser. Isolation, characterization, and expression in Escherichia coli of the Pseudomonas aeruginosa rhlAB genes encoding a rhamnosyltransferase involved in rhamnolipid biosurfactant synthesis. J. Biol. Chem., 269:19787–19795, 1994.

[124] U.A. Ochsner and J. Reiser. Autoinducer-mediated regulation of rhamnolipid biosurfactant synthesis in Pseudomonas aeruginosa. PNAS, 92:6424–6428, 1995.

[125] Joji M Otaki and Stuart Firestein. Length analyses of mammalian G-protein-coupled receptors. J. Theo. Biol., 211:77–100, 2001.

[126] Y.Y. Ou and S.A. Chen. Using Efficient RBF Networks to Classify Transport Proteins Based on PSSM Profiles and Biochemical Properties. Bio-Inspired Systems: Comp. Ambient Intelligence, pages 869–876, 2009.

[127] Y.Y. Ou, S.A. Chen, and M.M. Gromiha. Classification of transporters using efficient radial basis function networks with position-specific scoring matrices and biochemical properties. Proteins, 78:1789–1797, 2010.

[128] Y.Y. Ou, S.A. Chen, and M.M. Gromiha. Prediction of membrane spanning segments and topology in β-barrel membrane proteins at better accuracy. J. Comp. Chem., 31:217–223, 2010.

[129] J.P. Overington, B. Al-Lazikani, and A.L. Hopkins. How many drug targets are there? Nat. Rev. Drug Disc., 5:993–996, 2006.

[130] S.S. Pao, I.T. Paulsen, and M.H. Saier. Major facilitator superfamily. Microbiol. Mol. Biol. Rev., 62:1–34, 1998.

[131] K.J. Park and M. Kanehisa. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics, 19:1656–1663, 2003.

[132] Y. Park, S. Hayat, and V. Helms. Prediction of the burial status of transmembrane residues of helical membrane proteins. BMC Bioinformatics, 8:302, 2007.

[133] Y. Park and V. Helms. On the derivation of propensity scales for predicting exposed transmembrane residues of helical membrane proteins. Bioinformatics, 23:701–708, 2007.

[134] Y. Park and V. Helms. Prediction of the translocon-mediated membrane insertion free energies of protein sequences. Bioinformatics, 24:1271–1277, 2008.

[135] M.R. Parsek, D.L. Val, B.L. Hanzelka, J.E. Cronan, and E.P. Greenberg. Acyl homoserine-lactone quorum-sensing signal generation. PNAS, 96:4360–4365, 1999.

[136] L. Passador, J.M. Cook, M.J. Gambello, L. Rust, and B.H. Iglewski. Expression of Pseudomonas aeruginosa virulence genes requires cell-to-cell communication. Science, 260:1127–1130, 1993.

[137] J.P. Pearson, L. Passador, B.H. Iglewski, and E.P. Greenberg. A second N-acylhomoserine lactone signal produced by Pseudomonas aeruginosa. PNAS, 92:1490, 1995.

[138] J.P. Pearson, C. Van Delden, and B.H. Iglewski. Active efflux and diffusion are involved in transport of Pseudomonas aeruginosa cell-to-cell signals. J. Bacteriol., 181:1203–1210, 1999.

[139] K. Pearson. On lines and planes of closest fit to systems of points in space. Phil. Mag., 2:559–572, 1901.

[140] E.C. Pesci, J.B.J. Milbank, J.P. Pearson, S. McKnight, A.S. Kende, E.P. Greenberg, and B.H. Iglewski. Quinolone signaling in the cell-to-cell communication system of Pseudomonas aeruginosa. PNAS, 96:11229–11234, 1999.

[141] E.C. Pesci, J.P. Pearson, P.C. Seed, and B.H. Iglewski. Regulation of las and rhl quorum sensing in Pseudomonas aeruginosa. J. Bacteriol., 179:3127–3132, 1997.

[142] D. Pistorius, A. Ullrich, S. Lucas, R.W. Hartmann, U. Kazmaier, and R. Müller. Biosynthesis of 2-Alkyl-4 (1H)-Quinolones in Pseudomonas aeruginosa: Potential for Therapeutic Interference with Pathogenicity. ChemBioChem, 12:850–853, 2011.

[143] M. Punta and Y. Ofran. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. PLoS Comp. Biol, 4:e1000160, 2008.

[144] P. Raman, V. Cherezov, and M. Caffrey. The membrane protein data bank. Cell. Mol. Life Sci., 63:36–51, 2006.

[145] G. Rampioni, C. Pustelny, M.P. Fletcher, V.J. Wright, M. Bruce, K.P. Rumbaugh, S. Heeb, M. Cámara, and P. Williams. Transcriptomic analysis reveals a global alkyl-quinolone-independent regulatory role for PqsE in facilitating the environmental adaptation of Pseudomonas aeruginosa to plant and animal hosts. Environ. Microbiol., 12:1659–1673, 2010.

[146] T.B. Rasmussen, M. Givskov, et al. Quorum-sensing inhibitors as anti-pathogenic drugs. IJMM, 296:149–161, 2006.

[147] C. Reimmann, M. Beyeler, A. Latifi, H. Winteler, M. Foglino, A. Lazdunski, and D. Haas. The global activator GacA of Pseudomonas aeruginosa PAO positively controls the production of the autoinducer N-butyryl-homoserine lactone and the formation of the virulence factors pyocyanin, cyanide, and lipase. Mol. Microbiol., 24:309–319, 1997.

[148] L. Reinhold and A. Kaplan. Membrane transport of sugars and amino acids. Ann. Rev. Plant Physiol., 35:45–83, 1984.

[149] Q. Ren, K.H. Kang, and I.T. Paulsen. TransportDB: a relational database of cellular membrane transport systems. Nucl. Ac. Res., 32:D284–D288, 2004.

[150] A. Rodríguez, D. Sosa, L. Torres, B. Molina, S. Frías, and L. Mendoza. A Boolean network model of the FA/BRCA pathway. Bioinformatics, 28:858–866, 2012.

[151] D.M. Rosenbaum, S.G.F. Rasmussen, and B.K. Kobilka. The structure and function of G-protein-coupled receptors. Nature, 459:356–363, 2009.

[152] B. Rost and C. Sander. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins, 19:55–72, 1994.

[153] M.H. Saier. Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. Microbiol. Rev., 58:71–93, 1994.

[154] M.H. Saier. Genome archeology leading to the characterization and classification of transport proteins. Curr. Opin. Microbiol., 2:555–561, 1999.

[155] M.H. Saier. A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol. Rev., 64:354–411, 2000.

[156] M.H. Saier, C.V. Tran, and R.D. Barabote. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. Nucl. Ac. Res., 34:D181–D186, 2006.

[157] L. Sanchez and D. Thieffry. A Logical Analysis of the Drosophila Gap-gene System. J. Theo. Biol., 211:115–141, 2001.

[158] N.S. Schaadt, J. Christoph, and V. Helms. Classifying Substrate Specificities of Membrane Transporters from *Arabidopsis thaliana*. J. Chem. Inf. Model, 50:1899–1905, 2010.

[159] N.S. Schaadt and V. Helms. Functional classification of membrane transporters and channels based on filtered TM/non-TM amino acid composition. Biopolymers, 97:558–567, 2012.

[160] N.S. Schaadt, A. Steinbach, R.W. Hartmann, and V. Helms. Rule-based regulatory and metabolic model for Quorum sensing in P. aeruginosa. submitted to BMC Systems Biol., 2013.

[161] A.A. Schäffer, L. Aravind, T.L. Madden, S. Shavirin, J.L. Spouge, Y.I. Wolf, E.V. Koonin, and S.F. Altschul. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucl. Ac. Res., 29:2994–3005, 2001.

[162] H. Scheffe. The analysis of variance., 1959.

[163] M. Schiffer, C.H. Chang, and FJ Stevens. The function of tryptophan residues in membrane proteins. Prot. Eng., 5:213–214, 1992.

[164] G.E. Schulz. β-barrel membrane proteins. Curr. Opin. Struct. Biol., 10:443–447, 2000.

[165] G.E. Schulz. The structure of bacterial outer membrane proteins. BBA-Biomembranes, 1565:308–317, 2002.

[166] S. Schulze, S. Köster, U. Geldmacher, A.C.T. van Scheltinga, and W. Kühlbrandt. Structural basis of Na^+-independent and cooperative substrate/product antiport in CaiT. Nature, 467:233–236, 2010.

[167] R. Schwacke, A. Schneider, E. van der Graaff, K. Fischer, E. Catoni, M. Desimone, W.B. Frommer, U.I. Flugge, and R. Kunze. ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. Plant Physiol., 131:16–26, 2003.

[168] A. Secker, M.N. Davies, A.A. Freitas, E.B. Clark, J. Timmis, and D.R. Flower. Hierarchical classification of G-protein-coupled receptors with data-driven selection of attributes and classifiers. Internat. J. Data Mining Bioinformatics, 4:191–210, 2010.

[169] P.C. Seed, L. Passador, and B.H. Iglewski. Activation of the Pseudomonas aeruginosa lasI gene by LasR and the Pseudomonas autoinducer PAI: an autoinduction regulatory hierarchy. J. Bacteriol., 177:654–659, 1995.

[170] A. Senes, M. Gerstein, and D.M. Engelman. Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with [beta]-branched residues at neighboring positions. J. Mol. Biol., 296:921–936, 2000.

[171] S.S. Shapiro and M.B. Wilk. An analysis of variance test for normality (complete samples). Biometrika, 52:591–611, 1965.

[172] H.B. Shen and K.C. Chou. PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. Anal. Biochem., 373:386–388, 2007.

[173] I. Shmulevich, E.R. Dougherty, S. Kim, and W. Zhang. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics, 18:261–274, 2002.

[174] A. Silvescu and V. Honavar. Temporal boolean network models of genetic networks and their inference from gene expression time series. Complex Systems, 13:61–78, 2001.

[175] SJ Singer and G.L. Nicolson. The fluid mosaic model of the structure of cell membranes. Science, 175:720–731, 1972.

[176] M.E. Skindersoe, M. Alhede, R. Phipps, L. Yang, P.O. Jensen, T.B. Rasmussen, T. Bjarnsholt, T. Tolker-Nielsen, N. Høiby, and M. Givskov. Effects of antibiotics on quorum sensing in Pseudomonas aeruginosa. Antimicrob. Agents Chemother., 52:3648–3663, 2008.

[177] K.M. Smith, Y. Bu, H. Suga, et al. Induction and inhibition of Pseudomonas aeruginosa quorum sensing by synthetic autoinducer analogs. Chem. Biol., 10:81–89, 2003.

[178] L. Steggles, R. Banks, and A. Wipat. Modelling and analysing genetic networks: From Boolean networks to Petri nets. In Comp. Meth. Systems Biol., pages 127–141. Springer, 2006.

[179] J.B. Stock, A.M. Stock, and J.M. Mottonen. Signal transduction in bacteria. Nature, 344:395–400, 1990.

[180] M.P. Storz, C.K. Maurer, C. Zimmer, N. Wagner, C. Brengel, J.C. de Jong, S. Lucas, M. Müsken, S. Häussler, A. Steinbach, and R.W. Hartmann. Validation of PqsD as anti-biofilm target in Pseudomonas aeruginosa by development of small molecule inhibitors. J. Am. Chem. Soc., 134:16143–16146, 2012.

[181] H. Suga and K.M. Smith. Molecular mechanisms of bacterial quorum sensing as a new drug target. Curr. Opin. Chem. Biol., 7:586–591, 2003.

[182] Yoshiaki Sugiyama, Natalia Polulyakh, and Toshio Shimizu. Identification of transmembrane protein functions by binary topology patterns. Prot. Eng., 16:479–488, 2003.

[183] L.K. Tamm, A. Arora, and J.H. Kleinschmidt. Structure and assembly of β-barrel membrane proteins. J. Biol. Chem., 276:32399–32402, 2001.

[184] S. Tan, H.T. Tan, and M. Chung. Membrane proteins and membrane proteomics. Proteomics, 8:3924–3932, 2008.

[185] R. Thomas. Regulatory networks seen as asynchronous automata: a logical description. J. Theo. Biol., 153:1–23, 1991.

[186] J.D. Thompson, D.G. Higgins, and T.J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl. Ac. Res., 22:4673–4680, 1994.

[187] K.D. Tsirigos, A. Hennerdal, L. Käll, and A. Elofsson. A guideline to proteome-wide α-helical membrane protein topology predictions. Proteomics, 12:2282–2294, 2012.

[188] M.B. Ulmschneider and M.S.P. Sansom. Amino acid distributions in integral membrane protein structures. Biochim. Biophys. Acta, 1512:1–14, 2001.

[189] C. Van Delden and B.H. Iglewski. Cell-to-cell signaling and Pseudomonas aeruginosa infections. Emerg. Infect. Dis., 4:551–560, 1998.

[190] N. Verstraeten, K. Braeken, B. Debkumari, M. Fauvart, J. Fransaer, J. Vermant, and J. Michiels. Living on a surface: swarming and biofilm formation. Trends Microbiol., 16:496–506, 2008.

[191] H. Viklund and A. Elofsson. Best α-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. Prot. Sci., 13:1908–1917, 2004.

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

[193] H. Viklund, E. Granseth, and A. Elofsson. Structural classification and prediction of reentrant Regions in α-helical transmembrane proteins: application to complete genomes. J. Mol. Biol., 361:591–603, 2006.

[194] J. Vogt and G.E. Schulz. The structure of the outer membrane protein OmpX from Escherichia coli reveals possible mechanisms of virulence. Structure, 7:1301–1309, 1999.

[195] G. von Heijne. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO, 5:3021–3027, 1986.

[196] G. Von Heijne. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J. Mol. Biol., 225:487–494, 1992.

[197] G. Von Heijne and C. Manoil. Membrane proteins: from sequence to structure. Prot. Eng., 4:109–112, 1990.

[198] E. Wallin and G.V. Heijne. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Prot. Sci., 7:1029–1038, 1998.

[199] C. Wang, S. Li, L. Xi, H. Liu, and X. Yao. Accurate prediction of the burial status of transmembrane residues of α-helix membrane protein by incorporating the structural and physicochemical features. Amino Acids, 40:991–1002, 2011.

[200] J. Wang, Y. Li, Q. Wang, X. You, J. Man, C. Wang, and X. Gao. ProClusEnsem: Predicting membrane protein types by fusing different modes of pseudo amino acid composition. Computers Biol. Med., 42:564–574, 2012.

[201] J.P. Ward, J.R. King, A.J. Koerber, P. Williams, J.M. Croft, and R.E. Sockett. Mathematical modelling of quorum sensing in bacteria. Math. Med. Biol., 18:263–292, 2001.

[202] C.M. Waters and B.L. Bassler. Quorum sensing: cell-to-cell communication in bacteria. Ann. Rev. Cell Dev. Biol., 21:319–346, 2005.

[203] Frank Wilcoxon. Individual comparisons by ranking methods. Biometrics Bull., 1:80–83, 1945.

[204] C.N. Wilder, S.P. Diggle, and M. Schuster. Cooperation and cheating in Pseudomonas aeruginosa: the roles of the las, rhl and pqs quorum-sensing systems. ISME J., 5:1332–1343, 2011.

[205] G. Xiao, J. He, and L.G. Rahme. Mutation analysis of the Pseudomonas aeruginosa mvfR and pqsABCDE gene promoters demonstrates complex quorum-sensing circuitry. Microbiol., 152:1679–1686, 2006.

[206] X. Xiao, P. Wang, and K.C. Chou. GPCR-CA: A cellular automaton image approach for predicting G-protein–coupled receptor functional classes. J. Comp. Chem., 30:1414–1423, 2009.

[207] L. Yang, M.T. Rybtke, T.H. Jakobsen, M. Hentzer, T. Bjarnsholt, M. Givskov, and T. Tolker-Nielsen. Computer-aided identification of recognized drugs as Pseudomonas aeruginosa quorum-sensing inhibitors. Antimicrob. Agents Chemother., 53:2432–2443, 2009.

[208] W.M. Yau, W.C. Wimley, K. Gawrisch, and S.H. White. The preference of tryptophan for membrane interfaces. Biochemistry, 37:14713–14718, 1998.

[209] D. Yernool, O. Boudker, Y. Jin, and E. Gouaux. Structure of a glutamate transporter homologue from Pyrococcus horikoshii. Nature, 431:811–818, 2004.

[210] S. Yu, V. Jensen, J. Seeliger, I. Feldmann, S. Weber, E. Schleicher, S. Häussler, and W. Blankenfeldt. Structure elucidation and preliminary assessment of hydrolase activity of PqsE, the Pseudomonas quinolone signal (PQS) response protein. Biochem., 48:10298–10307, 2009.

[211] X. Yu, X. Zheng, T. Liu, Y. Dou, and J. Wang. Predicting subcellular location of apoptosis proteins with pseudo amino acid composition: approach from amino acid substitution matrix and auto covariance transformation. Amino Acids, 42:1619–1625, 2012.

[212] R. Zimmermann, S. Eyrisch, M. Ahmad, and V. Helms. Protein translocation across the ER membrane. BBA-Biomembranes, 1808:912–924, 2011.

Part VI.

Appendix

A. Functional Classification of Membrane Proteins

A.1. Characteristics of Amino Acids

amino acid	hydrophobicity	hydrophilicity	mass	pK_{COOH}	pK_{NH_2}	pI
A	0.62	-0.5	15	2.35	9.87	6.11
C	0.29	-1.0	47	1.71	10.78	5.02
D	-0.90	3.0	59	1.88	9.60	2.98
E	-0.74	3.0	73	2.19	9.67	3.08
F	1.19	-2.5	91	2.58	9.24	5.91
G	0.48	0.0	1	2.34	9.60	6.06
H	-0.40	-0.5	82	1.78	8.97	7.64
I	1.38	-1.8	57	2.32	9.76	6.04
K	-1.50	3.0	73	2.20	8.90	9.47
L	1.06	-1.8	57	2.36	9.60	6.04
M	0.64	-1.3	75	2.28	9.21	5.74
N	-0.78	0.2	58	2.18	9.09	10.76
P	0.12	0.0	42	1.99	10.60	6.30
Q	-0.85	0.2	72	2.17	9.13	5.65
R	-2.53	3.0	101	2.18	9.09	10.76
S	-0.18	0.3	31	2.21	9.15	5.68
T	-0.05	-0.4	45	2.15	9.12	5.60
V	1.08	-1.5	43	2.29	9.74	6.02
W	0.81	-3.4	130	2.38	9.39	5.88
Y	0.26	-2.3	107	2.20	9.11	5.63

Table A.1.: **Values of the used physicochemical characteristics** for all amino acids given in one–letter code as described by Shen *et al.* [172].

A.2. Statistics of Membrane Proteins

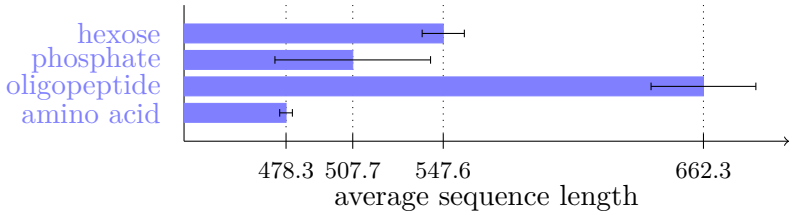


Figure A.1.: **Sequence length:** Shown are the averages with their standard deviation in *Arabidopsis thaliana* positive sets.

Pair of positive sets	amino acids (p–values)
amino acid – oligopeptide	aliphatic (0.1235), aromatic (0.4651), charged (0.0520), positive (0.2890), large (0.0203)
amino acid – phosphate	aromatic (0.0335), large (0.4607), small (0.7023)
amino acid – hexose	aliphatic (0.1368), polar (0.0211), charged (0.2537), positive (0.1196), large (0.0762), tiny (0.8164)
oligopeptide – phosphate	aliphatic (0.0225), aromatic (0.0288)
oligopeptide – hexose	charged (0.4866), positive (0.0206), negative (0.3862), small (0.5839)
phosphate – hexose	aromatic (0.0388), large (0.0478), small (0.0477)

Table A.2.: **Not significant ANOVA p–values** corresponding to A.3 considering the full sequence for Gaussian distributed sets.

Property	amino acid	oligopeptide	phosphate	hexose	*A. thaliana*
full sequence					
hydrophobic	0.637	0.628	0.606	0.607	0.592
aliphatic	0.265	0.256	0.239	0.276	0.264
aromatic	0.147	0.151	0.137	0.123	0.120
polar	0.426	0.446	0.435	0.406	0.448
charged	0.150	0.161	0.162	0.156	0.181
positive	0.090	0.094	0.098	0.085	0.103
negative	0.059	0.067	0.064	0.071	0.079
large	0.205	0.216	0.201	0.184	0.191
small	0.519	0.501	0.522	0.505	0.505
tiny	0.264	0.241	0.285	0.266	0.261
TMH areas					
hydrophobic	0.761	0.759	0.738	0.741	0.737
aliphatic	0.342	0.332	0.303	0.372	0.361
aromatic	0.163	0.179	0.174	0.142	0.141
polar	0.316	0.324	0.304	0.280	0.311
positive	0.034	0.029	0.039	0.020	0.042
negative	0.016	0.018	0.019	0.011	0.028
large	0.177	0.195	0.196	0.157	0.167
tiny	0.295	0.271	0.315	0.282	0.278
non–TMH areas					
hydrophobic	0.483	0.506	0.490	0.476	0.484
aliphatic	0.169	0.185	0.179	0.181	0.196
aromatic	0.127	0.126	0.110	0.104	0.104
polar	0.562	0.560	0.557	0.532	0.552
positive	0.160	0.154	0.154	0.152	0.150
negative	0.016	0.018	0.019	0.011	0.028
large	0.239	0.236	0.212	0.214	0.210
tiny	0.226	0.213	0.254	0.247	0.241

Table A.3.: **Average characteristic frequencies** based on physicochemical properties for the *Arabidopsis thaliana* data sets. Red colored entries denote a difference from the value of the full *Arabidopsis thaliana* data set by more than 25%, blue by more than 15%, and green by more than 10%. The considered groups of physicochemical properties are defined in Table 6.1 in Section 6.2.2.

Property	H.1.A.1	H.2.A.1	H.3.A.1	H.TCDB
full sequence				
hydrophobic	0.597	0.631	0.612	0.586
aliphatic	0.276	0.273	0.130	0.258
aromatic	0.125	0.128	0.115	0.114
polar	0.460	0.401	0.428	0.454
positive	0.115	0.086	0.099	0.109
negative	0.094	0.065	0.080	0.089
large	0.204	0.187	0.188	0.190
tiny	0.218	0.271	0.240	0.244
TMH areas				
hydrophobic	0.786	0.745	0.749	0.725
aliphatic	0.429	0.352	0.381	0.361
aromatic	0.150	0.147	0.129	0.132
polar	0.267	0.276	0.301	0.313
positive	0.039	0.030	0.041	0.052
negative	0.029	0.018	0.039	0.040
large	0.179	0.163	0.155	0.162
tiny	0.213	0.306	0.255	0.267
non–TMH areas				
hydrophobic	0.517	0.518	0.542	0.528
aliphatic	0.214	0.195	0.227	0.213
aromatic	0.115	0.113	0.108	0.107
polar	0.539	0.526	0.496	0.514
positive	0.148	0.144	0.130	0.135
negative	0.122	0.109	0.099	0.108
large	0.215	0.213	0.206	0.202
tiny	0.219	0.240	0.233	0.236

Table A.4.: **Average characteristic frequencies** based on physicochemical properties for the H.TCDB data sets. Red colored entries denote a difference from the value of the full H.TCDB data set by more than 25%, blue by more than 15%, and green by more than 10%. The considered groups of physicochemical properties are defined in Table 6.1 in Section 6.2.2.

Property	B.1.A.1	B.2.A.1	B.3.A.1	B.TCDB
full sequence				
hydrophobic	0.545	0.603	0.538	0.535
aliphatic	0.233	0.249	0.228	0.219
aromatic	0.128	0.134	0.095	0.110
polar	0.506	0.441	0.489	0.508
positive	0.128	0.088	0.121	0.121
negative	0.103	0.075	0.113	0.109
large	0.207	0.193	0.187	0.193
tiny	0.233	0.277	0.244	0.244
TMB areas				
hydrophobic	0.746	0.798	0.671	0.674
aliphatic	0.374	0.371	0.323	0.315
aromatic	0.139	0.192	0.130	0.157
polar	0.365	0.286	0.388	0.403
positive	0.080	0.033	0.098	0.095
negative	0.035	0.010	0.058	0.053
large	0.234	0.213	0.199	0.218
tiny	0.197	0.258	0.234	0.233
non–TMB areas				
hydrophobic	0.461	0.541	0.485	0.474
aliphatic	0.176	0.209	0.190	0.178
aromatic	0.103	0.115	0.081	0.088
polar	0.564	0.493	0.529	0.554
positive	0.149	0.106	0.130	0.132
negative	0.132	0.097	0.135	0.134
large	0.195	0.187	0.183	0.182
tiny	0.248	0.282	0.247	0.249

Table A.5.: **Average characteristic frequencies** based on physicochemical properties for the B.TCDB data sets. Red colored entries denote a difference from the value of the full B.TCDB data set by more than 25%, blue by more than 15%, and green by more than 10%. The considered groups of physicochemical properties are defined in Table 6.1 in Section 6.2.2.

Pair of positive sets	amino acids (p–values)
full sequence	
amino acid – oligopeptide	hydrophobic (0.5644), aliphatic (0.1303), aromatic (0.4713), polar (0.0105), positive (0.2962), negative (0.0662), large (0.0307), tiny (0.0025)
amino acid – phosphate	aliphatic (0.0044), aromatic (0.0528), polar (0.4526), positive (0.1138), negative (0.4292), large (0.4768), tiny (0.0057)
amino acid – hexose	aliphatic (0.2542), aromatic (0.0050), polar (0.0317), positive (0.0722), negative (0.0050), large (0.0317), tiny (0.8608)
oligopeptide – phosphate	hydrophobic (0.0361), aliphatic (0.0097), aromatic (0.0565), polar (0.0396), positive (0.7198), negative (0.3749), large (0.0069)
oligopeptide – hexose	hydrophobic (0.0383), aliphatic (0.0028), aromatic (0.0041), positive (0.0226), negative (0.3052), tiny (0.0089)
phosphate – hexose	hydrophobic (0.7297), aromatic (0.0761), polar (0.0156), positive (0.0022), negative (0.2401), large (0.0761), tiny (0.0226)
TMH areas	
amino acid – oligopeptide	hydrophobic (0.8289), aliphatic (0.1130), aromatic (0.0476), polar (0.4074), positive (0.0307), negative (0.8289), large (0.0069)
amino acid – phosphate	aromatic (0.3184), polar (0.2444), positive (0.1240), negative (0.4543), large (0.0340), tiny (0.0047)
amino acid – hexose	hydrophobic (0.0057), aromatic (0.0066), negative (0.0159), large (0.0016), tiny (0.0485)
oligopeptide – phosphate	hydrophobic (0.0015), aliphatic (0.0017), aromatic (0.2497), polar (0.0741), positive (0.0020), negative (0.7811), large (0.7209)
oligopeptide – hexose	hydrophobic (0.0069), positive (0.0089), negative (0.0113), tiny (0.3464)
phosphate – hexose	hydrophobic (0.0466), aromatic (0.0042), polar (0.0198), negative (0.0153)

Table A.6.: **Not significant Wilcoxon–Mann–Whitney p–values** corresponding to A.3 considering the full sequence and TMHs in the *Arabidopsis thaliana* data sets.

B. Quorum Sensing in *Pseudomonas aeruginosa*

B.1. Network

Reaction	Type	Reference
GacA → LasR	activation, transcription + translation	[147]
GacA → RhlR	activation, transcription + translation	[147]
Vfr → LasR	activation, transcription + translation	[2]
AI–1 + LasR → C1	assoziation	[169]
C1 → C1:G1	activation	[169]
C1:G1 → LasI	transcription + translation	[169]
C1:G1 → RsaL	transcription + translation	[37]
C1:G1 → AI–1	degradation	- - - - -
RsaL → RsaL:G	activation	[37]
RsaL →	degradation	- - - - -
RsaL:G −| LasI	inhibition	[37]
LasI → AI–1	enzymatic reaction (formation)	[136]
AI–1 ↔ AI–1_out	active efflux (MexAB–OprM)	[138]
AI–1 →	degradation	- - - - -
AI–1 + RhlR → C4	assoziation	[141]
C1 → C1:G2	activation	[141]
C1:G2 → RhlR	transcription + translation	[141]
AI–2 −| C4	blocking	[141]
C1:G2 → RhlI	transcription + translation	- - - - -
AI–2 + RhlR → C2	assoziation	[124]
C2 → C2:G2	activation	[124]
C2:G2 → RhlI	transcription + translation	[124]
C2:G2 → AI–2	degradation	- - - - -
RhlI → AI–2	enzymatic reaction (formation)	[124]
AI–2 ↔ AI–2_out	passive diffusion	[138]
AI–2 →	degradation	- - - - -

C1 → C1:G3	activation	[145]
C1:G3 → PqsR	transcription + translation	[145]
C1:G3 → PqsH	transcription + translation	[145]
HHQ ↔ HHQ_out	active efflux (MexEF–OprN)	[92]
HHQ + PqsH → PQS	enzymatic reaction (formation)	[39]
PQS + PqsR → C3	assoziation	[39]
C3 → C3:G3	activation	[39]
C3:G3 → PQS	degradation	- - - - -
C3:G3 → PqsABCDE	transcription + translation	[39]
PqsA + PqsBCD → HHQ	enzymatic reaction (formation)	[39]
PQS ↔ PQS_out	transport	
PQS →	degradation	- - - - -
HHQ + PqsR → C5	assoziation	[205]
C5 → C5:G3	activation	[205]
C5:G3 → PqsR + HHQ	degradation	- - - - -
C5:G3 → PqsABCDE	transcription + translation	[205]
PqsA → DHQ	enzymatic reaction (formation)	
DHQ ↔ DHQ_out	transport	
DHQ ↔	degradation	- - - - -
C1:G1 → LasB	transcription + translation	[55]
C2:G2 → LasB	transcription + translation	[204]
C3:G3 → LasB	transcription + translation	[204]
LasB →	degradation	- - - - -
C2:G2 → RhlAB	transcription + translation	[114, 123]
C2:G2 → RhlC	transcription + translation	
RhlAB → Rhm1	enzymatic reaction (formation)	[106, 123]
Rhm1 + RhlC → Rhm2	enzymatic reaction (formation)	[106, 123]
Rhm2 →	degradation	- - - - -
C3:G3 → pyocyanin	whole biosynthesis	
PqsE → pyocyanin	whole biosynthesis	
pyocyanin →	degradation	- - - - -

Table B.1.: **References for individual reactions** in the three Quorum sensing systems and for virulence factor formation of *Pseudomonas aeruginosa*.

B.2. Logical Trajectories

time step	HHQ	PQS	C3	C5	C3:G3	C5:G3	PqsA	PqsBCD	PqsE
10	0	0	0	0	1	1	0	0	0
11	1	1	0	0	0	0	1	1	1
12	2	1	0	0	0	0	0	0	0
13	2	1	0	0	0	0	0	0	0
14	2	1	0	0	0	0	0	0	0
15	2	1	0	0	0	0	0	0	0
16	0	2	0	1	0	0	0	0	0
17	0	2	0	0	0	1	0	0	0
18	1	2	0	0	0	0	1	1	1
19	2	1	1	0	0	0	0	0	0
20	0	1	0	1	1	0	0	0	0
21	0	2	0	0	0	1	1	1	1
22	2	1	1	0	0	0	1	1	1
23	2	1	0	1	0	0	0	0	0
24	0	2	0	1	0	1	0	0	0
25	1	1	1	0	0	1	1	1	1
26	2	2	0	0	0	0	1	1	1
27	1	2	1	0	0	0	0	0	0
28	0	2	1	0	1	0	0	0	0
29	0	2	1	0	1	0	1	1	1
30	1	2	1	0	0	0	1	1	1
31	1	2	1	0	0	0	0	0	0
32	0	2	1	0	1	0	0	0	0
33	0	2	1	0	1	0	1	1	1
34	1	2	1	0	0	0	1	1	1
35	2	1	1	0	0	0	0	0	0
36	1	1	0	1	1	0	0	0	0
37	1	2	0	0	0	1	1	1	1
38	2	2	1	0	0	0	1	1	1
39	1	2	1	0	0	0	0	0	0
40	0	1	1	0	1	0	0	0	0
41	0	2	0	0	1	0	1	1	1
42	1	2	1	0	0	0	1	1	1
43	2	1	1	0	0	0	0	0	0
44	1	1	0	1	1	0	0	0	0
45	0	2	0	0	0	1	1	1	1
46	2	2	1	0	0	0	1	1	1
47	1	2	1	0	0	0	0	0	0

48	0	2	1	0	1	0	0	0	0
49	0	2	1	0	1	0	1	1	1
50	1	2	1	0	0	0	1	1	1
51	2	1	1	0	0	0	0	0	0
52	0	2	0	1	1	0	0	0	0
53	0	2	1	0	0	1	1	1	1
54	2	1	1	0	0	0	1	1	1
55	2	1	0	1	0	0	0	0	0
56	0	2	0	1	0	1	0	0	0
57	1	1	1	0	0	1	1	1	1
58	2	1	0	0	0	0	1	1	1
59	3	1	0	1	0	0	0	0	0
60	1	1	0	1	0	1	0	0	0
61	1	2	0	0	0	1	1	1	1
62	2	1	1	0	0	0	1	1	1
63	3	1	0	1	0	0	0	0	0
64	1	2	0	1	0	1	0	0	0
65	1	2	1	0	0	1	1	1	1
66	2	2	1	0	0	0	1	1	1
67	1	2	1	0	0	0	0	0	0
68	0	2	1	0	1	0	0	0	0
69	0	2	1	0	1	0	1	1	1
70	1	2	1	0	0	0	1	1	1
71	2	1	1	0	0	0	0	0	0
72	0	2	0	1	1	0	0	0	0
73	0	2	1	0	0	1	1	1	1
74	2	1	1	0	0	0	1	1	1
75	2	1	0	1	0	0	0	0	0
76	1	1	0	1	0	1	0	0	0
77	1	2	0	0	0	1	1	1	1
78	2	1	1	0	0	0	1	1	1
79	2	1	0	1	0	0	0	0	0
80	0	1	0	1	0	1	0	0	0
81	1	1	0	0	0	1	1	1	1
82	2	2	0	0	0	0	1	1	1
83	1	2	1	0	0	0	0	0	0
84	1	1	1	0	1	0	0	0	0
85	0	2	0	0	1	0	1	1	1
86	1	3	1	0	0	0	1	1	1
87	1	2	1	0	0	0	0	0	0
88	0	3	1	0	1	0	0	0	0

89	0	2	1	0	1	0	1	1	1
90	1	3	1	0	0	0	1	1	1
91	2	2	1	0	0	0	0	0	0
92	1	1	1	0	1	0	0	0	0
93	0	2	0	0	1	0	1	1	1
94	1	2	1	0	0	0	1	1	1
95	1	2	1	0	0	0	0	0	0
96	1	1	1	0	1	0	0	0	0
97	1	2	0	0	1	0	1	1	1
98	1	2	1	0	0	0	1	1	1
99	1	3	1	0	0	0	0	0	0
100	0	2	1	0	1	0	0	0	0
101	0	2	1	0	1	0	1	1	1
102	1	2	1	0	0	0	1	1	1
103	1	2	1	0	0	0	0	0	0
104	0	2	1	0	1	0	0	0	0
105	0	2	1	0	1	0	1	1	1
106	1	2	1	0	0	0	1	1	1
107	1	2	1	0	0	0	0	0	0
108	0	2	1	0	1	0	0	0	0
109	0	2	1	0	1	0	1	1	1
110	1	2	1	0	0	0	1	1	1
111	1	2	1	0	0	0	0	0	0
112	0	2	1	0	1	0	0	0	0
113	0	3	1	0	1	0	1	1	1
114	1	2	1	0	0	0	1	1	1
115	2	2	1	0	0	0	0	0	0
116	1	1	1	0	1	0	0	0	0
117	0	2	0	0	1	0	1	1	1
118	1	3	1	0	0	0	1	1	1
119	1	2	1	0	0	0	0	0	0
120	0	1	1	0	1	0	0	0	0
121	0	2	0	0	1	0	1	1	1
122	1	3	1	0	0	0	1	1	1
123	1	2	1	0	0	0	0	0	0
124	0	2	1	0	1	0	0	0	0
125	0	2	1	0	1	0	1	1	1
126	1	2	1	0	0	0	1	1	1
127	1	2	1	0	0	0	0	0	0
128	1	1	1	0	1	0	0	0	0
129	0	2	0	0	1	0	1	1	1

130	1	2	1	0	0	0	1	1	1
131	1	2	1	0	0	0	0	0	0
132	1	1	1	0	1	0	0	0	0
133	0	2	0	0	1	0	1	1	1
134	1	2	1	0	0	0	1	1	1
135	2	1	1	0	0	0	0	0	0
136	1	1	0	1	1	0	0	0	0
137	0	2	0	0	0	1	1	1	1
138	2	1	1	0	0	0	1	1	1
139	1	2	0	1	0	0	0	0	0
140	1	0	1	0	0	1	0	0	0
141	2	0	0	0	1	0	1	1	1
142	1	2	0	1	0	0	1	1	1
143	3	1	1	0	0	0	0	0	0
144	1	2	0	1	1	0	0	0	0
145	1	2	1	0	0	1	1	1	1
146	2	2	1	0	0	0	1	1	1
147	1	2	1	0	0	0	0	0	0
148	1	1	1	0	1	0	0	0	0
149	0	2	0	0	1	0	1	1	1
150	1	2	1	0	0	0	1	1	1

Table B.2.: **Example trajectory** for species in the *pqs* system of a wild type cell. Listed are the levels from time interval 10 to time 150. The used parameters are given in Table 9.1 in Section 9.2.2, the considered network is illustrated in Figure 8.8 in Section 8.4.1, and the system is initialized as explained in Section 9.2.1. When a degradation takes place the time is colored in orange. Multilevel nodes with six possible states are shown in blue and Boolean nodes in red.

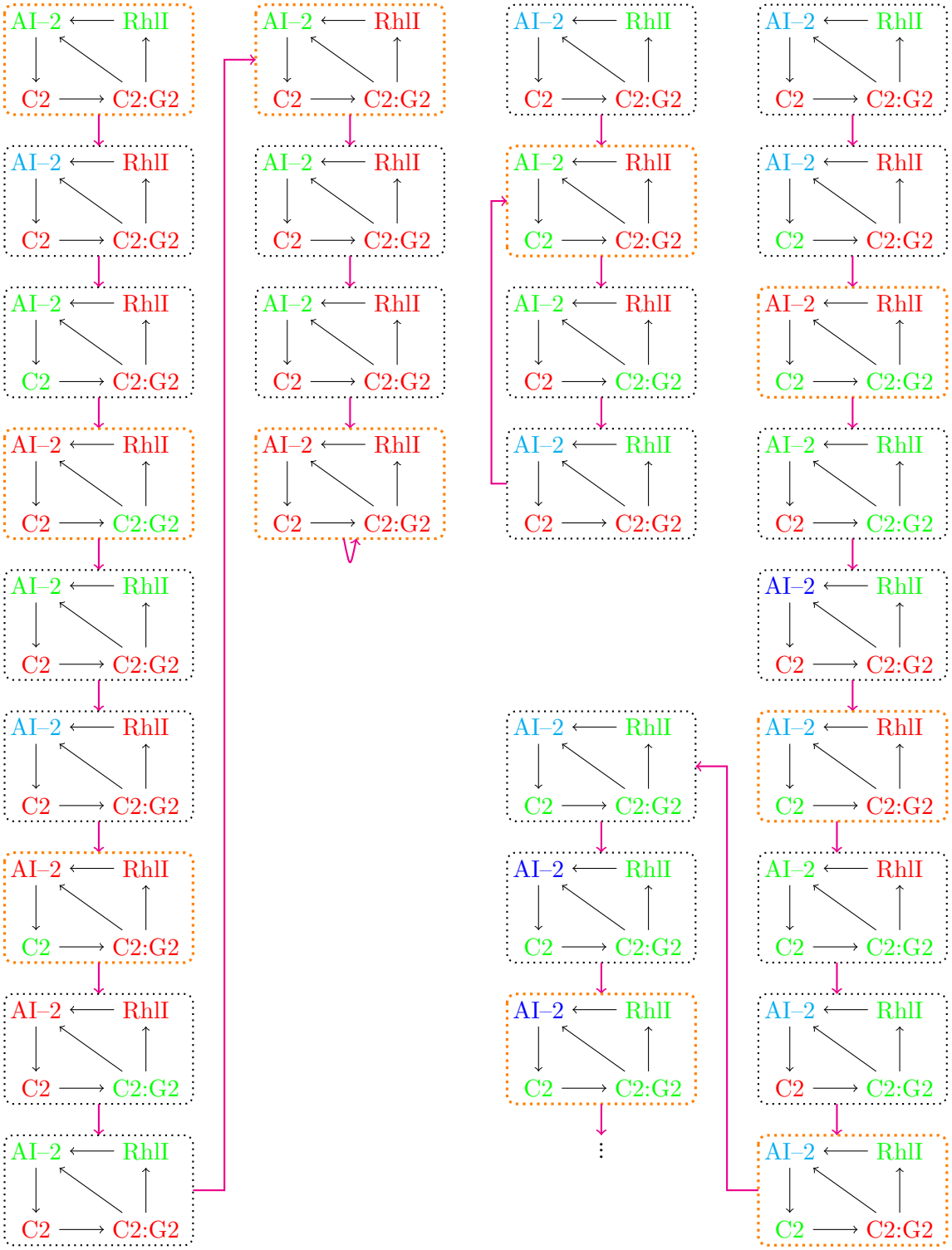


Figure B.1.: **Influence of degradation frequency on *rhl* system:** Shown is a simplified, positive feedback loop of the *rhl* system. Species colored in red represent a level of zero, in green a level of one, in cyan a level of two, and in blue a level of three. Arrows colored in magenta denote the update of the system state to the next time step and cells labeled in orange mark the time point at which a degradation happens. Here, $\varpi = \frac{1}{3}$.

B.3. User Manual for Extended Logical Formalism

Input Files

The considered networks are required as file with filename "conditions_<string>.txt", whereby <string> represents the character that has to be listed in the command line. The file contains for each species a line with its label and its edges:

$$>\text{Species: a[activated by], d[inhibited by],}$$
$$\text{r[degradated], t[transported to]; } initialState/\text{N}=\mathcal{M}$$

Here, each reaction should be given separately in the form (A land B) or (A land C{ϵ}). For stochastically modeled reactions, "and "random(ζ)"" must be added to the corresponding reaction description. An example input file for the wild type illustrated in Figure 8.8 in Section 8.4.1 is shown in Listing B.1.

```
1  >PqsH: a[K1G3], d[], r[], t[]; 0/N=2
2  >K1G1: a[K1], d[], r[], t[]; 1/N=2
3  >K1G2: a[K1], d[], r[], t[]; 0/N=2
4  >K1G3: a[K1], d[], r[], t[]; 0/N=2
5  >K2G2: a[K2], d[], r[], t[]; 0/N=2
6  >K3G3: a[K3], d[PqsE], r[], t[]; 1/N=2
7  >K5G3: a[K5], d[PqsE], r[], t[]; 1/N=2
8  >K1: a[AI1_in{2} and R1], d[], r[], t[]; 0/N=2
9  >K2: a[AI2_in{2} and R2], d[], r[], t[]; 0/N=2
10 >K3: a[PQS_in{2} and R3], d[], r[], t[]; 0/N=2
11 >K4: a[AI1_in{4} and R2], d[AI2_in{4}], r[], t[]; 0/N=2
12 >K5: a[HHQ_in{2} and R3], d[PQS_in{2}], r[], t[]; 0/N=2
13 >R1: a[GacA or Vfr or K1G1], d[], r[], t[]; 0/N=2
14 >R2: a[GacA or K1G2 or K2G2], d[K4], r[], t[]; 0/N=2
15 >R3: a[K1G3 or K3G3 or K5G3], d[], r[], t[]; 0/N=2
16 >GacA: a[GacA], d[], r[], t[]; 0/N=2
17 >Vfr: a[Vfr], d[], r[], t[]; 1/N=2
18 >RsaL: a[K1G1], d[], r[], t[]; 0/N=3
19 >RsaLG: a[RsaL{2}], d[], r[], t[]; 0/N=2
20 >LasB: a[K1G1 or K2G2 or K3G3], d[], r[], t[]; 0/N=6
21 >pyocyanin: a[K3G3 or PqsE], d[], r[], t[]; 0/N=6
22 >Rhm2: a[Rhm1 and RhlC], d[], r[], t[]; 0/N=6
23 >Rhm1: a[RhlAB], d[], r[], t[]; 0/N=2
24 >RhlAB: a[K2G2], d[], r[], t[]; 0/N=2
25 >RhlC: a[K2G2], d[], r[], t[]; 0/N=2
26 >RhlI: a[K2G2 or K1G2], d[], r[], t[]; 0/N=2
27 >LasI: a[K1G1], d[RsaLG], r[], t[]; 0/N=2
28 >PqsA: a[K3G3 or K5G3], d[], r[], t[]; 0/N=2
29 >PqsBCD: a[K3G3 or K5G3], d[], r[], t[]; 0/N=2
30 >PqsE: a[K3G3 or K5G3], d[], r[], t[]; 0/N=2
31 >HHQ_in: a[(PqsA and PqsBCD) or K5G3], d[], r[], t[HHQ_out{3}]; 0/N
       =6
```

```
>PQS_in: a[(HHQ_in and PqsH and "random(55)") or K3G3], d[], r[], t[
    PQS_out{3}]; 0/N=6
>DHQ_in: a[PqsA or K5G3], d[], r[], t[DHQ_out{3}]; 0/N=6
>AI2_in: a[RhlI or K2G2], d[], r[], t[AI2_out{3}]; 0/N=6
>AI1_in: a[LasI or K1G1], d[], r[], t[AI1_out{3}]; 0/N=6
>AI1_out: a[], d[], r[], t[AI1_in{1}]; 0/N=n
>AI2_out: a[], d[], r[], t[AI2_in{1}]; 0/N=n
>HHQ_out: a[], d[], r[], t[HHQ_in{1}]; 0/N=n
>PQS_out: a[], d[], r[], t[PQS_in{1}]; 0/N=n
>DHQ_out: a[], d[], r[], t[DHQ_in{1}]; 0/N=n
>trash: a[], d[], r[AI1_in or AI2_in or PQS_in or DHQ_in or RsaL or
    LasB or pyocyanin or Rhm2], t[]; 0/N=2
```

Listing B.1: **Input file for wild type**

Besides the used network, it is possible to use files with a list of random numbers.

Command Line Parameters

Table B.3 lists all required parameters with their default values.

Command line parameter	parameter / description	default value
-characters <string>	to list the input networks	wild type
-randomType <string>	**real** from random number generator, **pseudo** from file	pseudo
-randomNumbers <filename>	for pseudo	randomNumbers1.txt
-randomNumbersGaussian <filename>	for pseudo	randomGaussian1.txt
-printAllFiles <boolean>	for intermediate steps in output	false
-simulationTime <integer>	t_{max}	600
-sleepingTime <integer>	delay time of new cell	10
-maxCellDivisions <integer>	δ_{max}	5
-growthRate <integer>	$\frac{1}{\kappa}$	60
-growthRates <string> <integer>	to list a different κ for each network	wild type 60
-mortalityTime <integer>	ψ	t_{max}
-mutationRate <integer>	ν	17
-degradationTime <integer>	$\frac{1}{\varpi}$	20

Table B.3.: **Command line parameters**

Output Files

In general, the average values in the time interval from 100 to 600 of each species for every cell are listed in a file called averageValues.txt. Afterwards, the average values of a certain species can be averaged over several runs with different random numbers (statistics.txt). Additionally (-printAllFiles true), it is possible to write the levels at each time step in a file called species.txt in the folder of the corresponding cell.

C. Glossaries

List of Figures

List of Tables

List of Listings

List of Abbreviations

A. thaliana . . . . . *Arabidopsis thaliana*
P. aeruginosa . . . . *Pseudomonas aeruginosa*
3–oxo–C_{12}–HSL . . . *N*–3–oxododecanoyl–homoserine lactone
3D . . . . . three–dimensional (in space)
5–methyl–PCA . . . . 5–methylphenazine–1–carboxylic acid betaine

AAC . . . . . amino acid composition
ABC . . . . . ATP binding cassette
ACoA . . . . . anthraniloyl–coenzyme A
ANOVA . . . . . analysis of variance

BLAST . . . . . basic local alignment search tool

C_4–HSL . . . . . *N*–butyryl–homoserine lactone

DHQ 2,4–dihydroxy–quinoline

GPCR . . . . . G–protein coupled receptor

HHQ 4–hydroxy–2–heptyl–quinoline

LOOCV . . . . . leave one out cross validation

MAFFT . . . . . multiple alignment using fast Fourier transform
MBA . . . . . multiple binary string alignment
MFS . . . . . major facilitator superfamily
MSA . . . . . multiple sequence alignment

OPM . . . . . orientations of proteins in membranes

PAAC . . . . . pair amino acid composition
PCA . . . . . principal component analysis
PQS 2–heptyl–3–hydroxy–4–quinolone
PseAAC . . . . . Pseudo amino acid composition
PsePSSM . . . . . Pseudo position specific scoring matrix
PSI–BLAST . . . . . position–specific iterative basic local alignment search tool
PSSM . . . . . position specific scoring matrix

SAAC . . . . . split amino acid composition
SVM . . . . . support vector machine

TMB . . . . . transmembrane sheet
TMH . . . . . transmembrane helix
TMS . . . . . transmembrane segment

List of Notations

FN . . . . . false negative
FP . . . . . false positive
TN . . . . . true negative
TP . . . . . true positive

AI–1 . . . . . N–3–oxododecanoyl–homoserine lactone
AI–2 . . . . . N–butyryl–homoserine lactone

C1 . . . . . complex between N–3–oxododecanoyl–homoserine lactone and LasR
C2 . . . . . complex between N–butyryl–homoserine lactone and RhlR
C3 . . . . . complex between PQS and PqsR
C4 . . . . . complex between N–3–oxododecanoyl–homoserine lactone and RhlR
C5 . . . . . complex between HHQ and PqsR
Ci:Gj . . . . . complex i bind to operon j

negative set . . . . . subset complementary to a positive set

oAAC . . . . . original AAC

positive set . . . . . subset of the full data set in which all proteins have a similar function

Rhm1 . . . . . monorhamnolipid
Rhm2 . . . . . dirhamnolipid

List of Symbols

B . . . . . step function given in equation (8.6)
G . . . . . step function given in equation (8.7)
H_0 . . . . . null hypothesis
K . . . . . step function given in equation (8.5)
L . . . . . step function given in equation (8.3)
M . . . . . step function given in equation (8.8)
P . . . . . probability
T . . . . . correlation function
Φ . . . . . probability density function of standard normal distribution
χ . . . . . transport threshold
δ . . . . . cell divisions
ϵ . . . . . threshold
η . . . . . life time
γ . . . . . threshold
κ . . . . . growth rate
$\mathbf{S}$ . . . . . system states
$\mathbf{W}$ . . . . . weighting matrix
$\mathcal{M}$ . . . . . maximal possible level
$\mathcal{N}$ . . . . . network
$\mathfrak{d}$ . . . . . number of degradations
$\mathfrak{n}$ . . . . . normalization factor
$\mathfrak{p}$ . . . . . positive set
$\mathfrak{r}$ . . . . . random set
$\mathfrak{s}_{i,j}$ . . . . . PSSM entry
Γ . . . . . pyocyanin formation reactions
Υ . . . . . average theoretical maximum level
ν . . . . . mutation rate
ω . . . . . weighting factor
ρ . . . . . rank
τ . . . . . correlation factor
ϖ . . . . . degradation frequency
ξ . . . . . integer value given in equation (5.12)
ζ . . . . . conversion/occurring frequency
b . . . . . Gaussian distributed random variable
d . . . . . Euclidean distance

f . . . . . amino acid frequencies
n . . . . . number of nodes in a network
r . . . . . ranking
s . . . . . score
t . . . . . t–distribution, time
v . . . . . AAC
x . . . . . state
y . . . . . AAC