Substrate Specificity Prediction of Enzymes and its Applications to Nonribosomal Peptide Synthetases

Dissertation

der Fakultät für Informations- und Kognitionswissenschaften
der Eberhard-Karls-Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Dipl.-Biotech. Christian Rausch

aus Neuendettelsau

Tübingen
2007

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sowie die Online-Version dieses Buchs sind im Internet über `http://dnb.d-nb.de` abrufbar.

Bibliographic Information of the German National Library (Deutsche Nationalbibliothek)
The German National Library (Deutsche Nationalbibliothek) lists this publication in the German National Bibliography; detailed bibliographic data as well as the online version of this book are available via `http://dnb.d-nb.de`.

Tag der mündlichen Qualifikation:	19.07.2007
Dekan:	Prof. Dr. Michael Diehl
1. Berichterstatter:	Prof. Dr. Daniel H. Huson
2. Berichterstatter:	Prof. Dr. Wolfgang Wohlleben

Erklärung

Hiermit erkläre ich, dass ich diese Schrift selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die im Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben der Quellen kenntlich gemacht sind.

Tübingen, Juni 2007 *Christian Rausch*

Zusammenfassung

Nichtribosomal synthetisierte Peptide (NRP) und Polyketide (PK) stellen eine vielfältige Gruppe von Naturstoffen dar, zu der Antibiotika, Arzneimittel gegen Krebs, Entzündungshemmer, Immunosuppressiva, Metallkomplexbildner und andere Moleküle mit interessanten Eigenschaften gehören.

Die ständige Nachfrage nach neuen Wirkstoffen und die wachsende Anzahl noch nicht erforschter Protein-Sequenzen aus Genom-Projekten verlangen nach besseren Methoden, um neuartige NRP-Synthetasen (NRPS) und PK-Synthasen (PKS) automatisiert in den Protein-Datenbanken aufzuspüren und die Zusammensetzung ihrer Produkte effizient vorherzusagen.

Neben der Suche nach neuartigen biologisch aktiven Molekülen ist man auch bestrebt, durch die gezielte Modifikation bekannter NRPS/PKS Biosynthese-Cluster maßgeschneiderte Produkte zu entwerfen. Diese Strategie ist umso effizienter, je besser Positionen bzw. Segmente in den Enzymen vorhergesagt werden können, die mutiert bzw. rekombiniert werden müssen, um neue Substanzen zu erhalten.

In dieser Arbeit wurden Methoden entwickelt und etabliert, die diese beiden Ansätze unterstützen: Eine effiziente Suchstrategie mit *Profile Hidden Markov Models (pHMMs)* wird genutzt, die das gleichzeitige Auftreten bestimmter enzymatischer Domänen fordert, und es so erlaubt, NRPS und PKS in Protein-Sequenzen sicher aufzufinden.

Eine neue, auf maschinellem Lernen (Stützvektormaschinen) beruhende Strategie wurde entwickelt, mit der vorhergesagt werden kann, welche Bausteine (in der Regel Aminosäuren) in NRPS von Adenylierungsdomänen ausgewählt werden, um im Folgenden in das Produkt eingebaut zu werden. Dadurch wird es möglich, auf die Zusammensetzung des synthetisierten Produkts zu schließen. Diese neue Methode wurde in dem Programm NRPSpredictor implementiert und steht kostenlos über `www-ab.informatik.uni-tuebingen.de/software/NRPSpredictor` zur Verfügung.

Die NRPS Kondensationsdomänen verbinden die von den Adenylierungsdomänen ausgewählten Aminosäuren durch Ausbildung einer Peptidbindung zu einem Peptidstrang und erzeugen je nach ihrer funktionellen Variante (Subtyp) unterschiedliche Produktgeometrien. In einer umfassenden Studie der evolutionären Beziehungen dieser Subtypen wurden charakteristische Sequenz-Motive und -Positionen aufgedeckt, in denen sich die verschiedenen Varianten unterscheiden. Eine automatisierte Vorhersage der funktionellen Subtypen der Kondensationsdomäne wird durch die erstellten *pHMMs* ermöglicht. Die ermittelten subtypspezifischen Positionen sind hilfreich für die gezielte Einführung von Mutationen, um einen Subtyp in einen anderen

zu überführen mit der Absicht, neuartige Produkte zu erhalten.

Desweiteren wurden die Möglichkeiten der Strukturbioinformatik untersucht und *Molecular Modeling* und *Docking* Simulationen durchgeführt, um die Spezifität von Adenylierungsdomänen sowie die Auswirkungen gezielter Punkt-Mutationen auf die Bindungspräferenzen der Adenylierungsdomänen vorherzusagen.

Die in dieser Arbeit eingeführten Methoden sind nutzbar für die Vorhersage der Spezifitäten bzw. der funktionellen Subtypen anderer Enzyme unter bestimmten Voraussetzungen, insbesondere genügend hoher Sequenzähnlichkeit zwischen den verschiedenen Gruppen, so dass über multiple Sequenz-Alignments homologe Positionen ermittelt werden können.

Abstract

Nonribosomal peptides (NRPs) and polyketides (PKs) are a diverse group of natural products comprising molecules with antibiotic, antitumoral, anti-inflammatory, immunosuppressing, metal chelating and other interesting properties. The steady demand for novel drugs and the increasing number of uncharacterized protein sequences issued from genome projects call for better methods to automatically detect novel NRP synthetases (NRPSs) and PK synthases (PKSs) in the protein databases, and to predict the composition of their products efficiently.

Besides the search for novel biologically active molecules, research also tries to obtain tailored products by the rational manipulation of known NRPS/PKS biosynthesis clusters. This strategy will become more efficient, as we are better able to predict positions to be mutated or segments to be recombined in these enzymes.

In this thesis, we develop and establish methods that are helpful for both strategies: predicting new and manipulating known products.

To detect NRPSs and PKSs efficiently in protein sequences, we use a search strategy with profile Hidden Markov Models (pHHMs) that requires the simultaneous occurrence of certain enzymatic domains specific for these enzymes.

We present a new machine learning (Support Vector Machine)-based strategy to predict which building blocks (mainly amino acids) are selected for incorporation by so-called Adenylation (A) domains in NRPSs. Thus, it becomes possible to infer the composition of the synthesized product. This new method is implemented in the program NRPSpredictor and is freely accessible via `www-ab.informatik.uni-tuebingen.de/software` → `NRPSpredictor`.

The NRPS Condensation (C) domains catalyze the bond formation between the amino acids (that were previously selected by the A domains) and may produce different product geometries according to their functional variant (subtype). In a comprehensive evolutionary study of these subtypes, we reveal characteristic sequence motifs and positions in which the unequal variants differ. We make available some pHHMs, which facilitate the automated prediction of the functional C domain subtypes. The determined subtype-specific positions will be helpful for the directed mutagenesis to turn one subtype into another with the goal of obtaining novel products.

Moreover, we explore possibilities of structural bioinformatics using molecular modeling and docking simulations to predict the specificity of A domains. These simulations also allow for the study of directed point-mutations in

these domains.

The methods introduced in this work are applicable to predicting the specificities of functional subtypes of other enzymes under certain conditions; in particular, a sufficiently high sequence similarity between the different groups is required to be able to determine homologous positions via a multiple sequence alignment.

Acknowledgments

First of all, I want to thank Prof. Daniel Huson, my advisor. He gave me incredible freedom in my choice of research activities and provided excellent conditions, candid support and encouragement during this work. Thanks to him, I learned to work autonomously in scientific research. I am also indebted to Prof. Wolfgang Wohlleben, my co-advisor who sparked my interest in NRPS and PKS enzymes. His experience was a great help in deciding what would be interesting and also helpful for the scientific community. I am deeply grateful to Prof. Oliver Kohlbacher for many fruitful and inspiring discussions, helpful comments and ideas. Very special thanks go to Tilmann Weber and Evi Stegmann, who always had time for discussions in this exemplary collaboration.

My time at the Sand really was pleasant thanks to all workmates of the research group *Algorithms in Bioinformatics*, namely Marine Gaudefroy-Bergmann, Olaf Delgado Friedrichs, Stefan Henz, Tobias Dezulian, Tobias Klöpper, Daniel Richter, Alexander Auch and Regula Rupp (thanks for proofreading parts of this thesis) as well as my other bioinformatics colleagues: Kay Nieselt, Muriel Quenzer, Stephan Steigele, Janko Dietzsch, Annette Höglund, Pierre Dönnes, and Marc Sturm (in order of appearance) among many others. In particular, I am very glad that Tobias Dezulian shared my office. He is a stupendously bright guy in every respect, and both humorous and serious conversations with him were a great joy and help. Very special thanks go to Ilka Hoof for a very good time, and a very fruitful and inspiring collaboration during her Master's thesis and as a student assistant.

Furthermore, my warm thanks go to all the students I had the chance to work with during their Bachelor's or Master's thesis projects, or as student assistants: Kristina Hug, Esther Rheinbay, Alex Thielen, Markus Zimmermann, Christian Rödelsperger, Andreas Biegert, Beatrix Weber, Marion Renner, Marc Röttig and Vassilena Gaykova.

Tausend Dank also to my family, especially to my parents Rudolf and Hedwig and to my friends who helpfully accompany me on this long and winding road of life. Last, but definitely not least, I would like to thank my wife Eva Merel-Rausch for encouraging, supporting and powering me during the whole time I was working at this thesis (and else). She and our little daughter Juliette, they show me every day how beautiful life is:

You are the sunrays that color my life.

Contents

Chapter 1

Introduction and Motivation

1.1 History of Antibiotics

The first antimicrobial drug, Salvarsan, was invented by the German-Jewish physician Paul Ehrlich in 1910. Although it is active only against a narrow range of germs, it could be used to combat a number of diseases like the then wide-spread syphilis, and represented a large step forwards for medicine [Parascandola, 2002].

In 1935, the German Gerhard Domagk invented Prontosil a member of the sulfonamide drug family, widely known as sulfa drugs, which are competitive inhibitors of the folate synthesis in bacteria, making them the first broad spectrum antibiotics [Parascandola, 2002]. In combination with Trimethoprim, a dihydrofolate reductase inhibitor, sulfonamides are still used today, e.g. under the name Cotrimoxazole.

In 1929, the Scot Alexander Fleming discovered the bactericidal action of a filamentous mold, *Penicillium chrysogenum*, contaminating his bacteria cultures. Fleming named the active functional substance penicillin. After impressive results on infected animals and later human subjects, and with the beginn of World War II, penicillin was soon produced on a large scale, with the US leading the way [Parascandola, 2002].

The discovery of the β-lactam antibiotic penicillin was the herald for the "golden age" of natural product discovery from the 1940s to the 1950s [Clardy et al., 2006] when many bio-active substances were discovered, including the following antibiotics and their respective classes: tetracycline (a polyketide), chloramphenicol (a phenylpropanoid), streptomycin (an aminoglycoside), erythromycin (a macrolid), vancomycin (a glycopeptide), ciprofloxacin (a synthetic quinolone, thus not a natural product) and pristinamycin (a streptogramin) [von Nussbaum et al., 2006].

In the second half of the 20^{th} century, the need for new antibiotics has been mainly met by semi-synthetic or totally synthetic improvement of these natural product *lead structures*, and the research activity of pharmaceutical companies in the discovery of new natural products declined [von Nussbaum et al., 2006].

1.2 Decreasing Effectiveness of Antibiotics

Because each application of antibiotics inevitably selects for resistant bacteria, insensitive strains become more and more of a problem, especially if pathogenic germs acquire multiple resistances against several antibiotics that can normally be used for their prevention.

The much celebrated benefits of antibiotics in the last century often led to their too overhasty prescription – it is not rare that antibiotics are administered "prophylactically" [many sources, e.g. Malhotra-Kumar et al., 2007].

Many studies show that about three out of four patients get a prescription for antibiotics when they consult a physician with a common cold (cough, coryza, hoarseness), although it is known that these symptoms are mostly due to a viral infection. Thereby, many physicians want to protect the virally weakened body against an additional bacterial infection. However, for a long time, such prophylactical therapy has been proven to be inadequate [Malhotra-Kumar et al., 2007]. In their recent study, Malhotra-Kumar et al. [2007] have shown a direct correlation of administration of antibiotics and the emergence of resistant bacteria in human.

Fortunately, it is coming into public awareness that antibiotics should be taken with caution. In France, for example, the state health insurance [Assurance Maladie, 2007] started a campaign in the media to sensitize people to the prudent use of antibiotics.

Another important reason why more and more resistant bacteria emerge is that certain antibiotics have been used as growth promoters in animal feeds in subtherapeutic levels since the 1950s [Zimmerman, 1986]. It is estimated that the amount of antibiotics used for animals is at least as large or even ten times larger than the amount used for human therapy [Wegener et al., 1999; Roberts, 2002], with the largest amount being used as growth promoters.

In the struggle of emerging resistance, some antibiotics like vancomycin [McCormick et al., discovered 1956] were regarded as last-resort antibiotics as they were efficient against already resistant pathogens like methicillin resistant *Staphylococcus aureus* strains (MRSA) that make up already approximately 40% or more of clinical *S. aureus* isolates in industrial countries [Appelbaum, 2006].

However, the first pathogen, *Enterococcus faecium*, resistant to vancomycin appeared in 1986 [Johnson et al., 1990], and an MRSA strain with reduced sensitivity to vancomycin appeared in 1996 [Hiramatsu et al., 1997]; six years later, in 2002, the first vancomycin resistant MRSA was discovered [Centers for Disease Control and Prevention (CDC), 2002; Chang et al., 2003], carrying the *vanHAX* resistance gene naturally present in vancomycin producers.

The appearance of vancomycin resistant MRSA has been anticipated, as the conjugation of the resistance genes from enterococci to *S. aureus* had been demonstrated in the laboratory [Noble et al., 1992]. A possible and likely route of these resistance genes to human pathogens is the transmission of resistant prokaryotes from production animals via the food chain to humans.

Indeed, avoparcin (Fig 1.1), a structural analog of vancomycin, has been

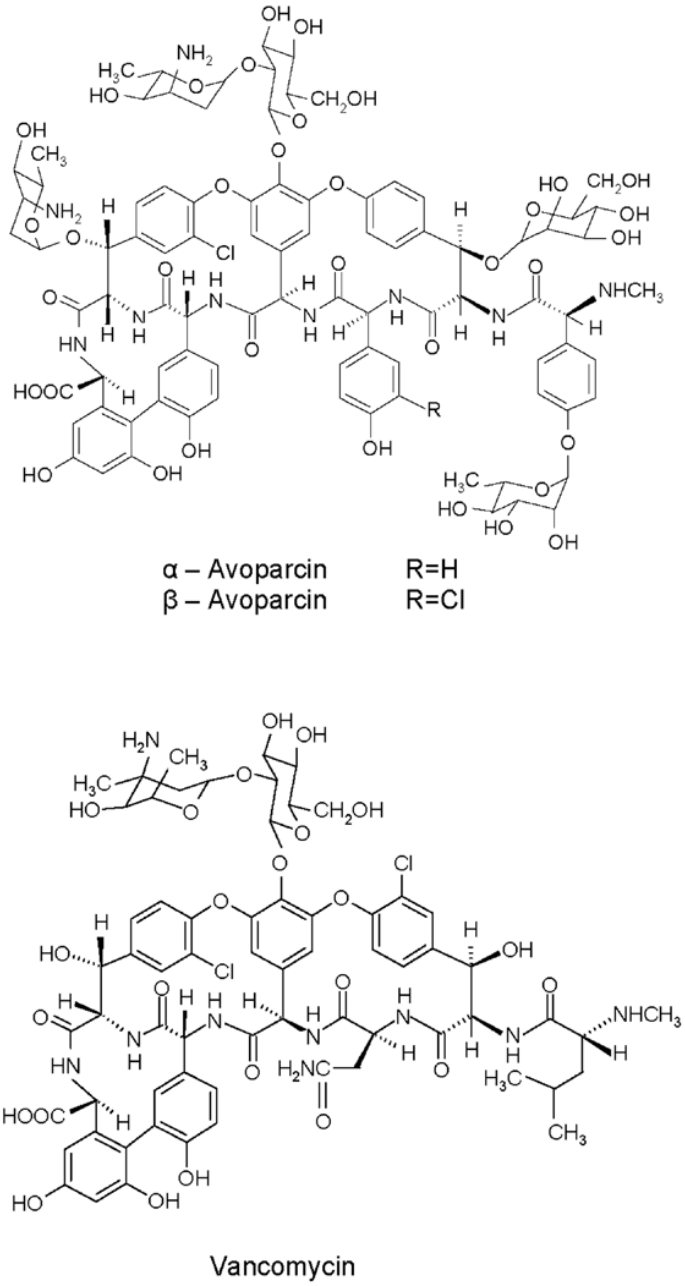

Figure 1.1: Chemical structures of the glycopeptide antibiotics avoparcin and vancomycin. Avoparcin had been used for more than two decades in many industrial countries as a growth promoter in animal husbandry, ignoring that it is structurally so similar to vancomycin that bacterial resistance to one of the two confers resistance to the other [Witte, 1998]. As animal commensal bacteria were steadily exposed to low concentrations of avoparcin and this drug had been found to be contaminated with genomic DNA of the avoparcin producer (*Amycolatopsis coloradensis* NRRL 3218) that was highly similar to the *vanHAX* gene cluster which confers vancomycin resistance, it is likely that resistance genes were spread through the food chain to human pathogens [Lu et al., 2004, image source: *idem*].

used in many countries (except in the USA and Canada because of possible carcinogenic effects [McDonald et al., 1997]) as a growth promoter in animal husbandry (it was licensed in the European Community 1975 and banned in all EU countries 1997, Donnelly et al. [1996]). Resistance to one of the two glycopeptides confers resistance to the other [Witte, 1998]. The consumption of avoparcin per year for use in animals was in the order of 100 to 1000 times higher than that of vancomycin for human use (numbers refer to Denmark and Australia, respectively [Witte, 1998]). Lu et al. [2004] report that they found substantial amounts of DNA of the avoparcin producer *Amycolatopsis coloradensis* NRPL 3218 in the drug carrying a homolog of the *vanHAX* cluster, which confers resistance to vancomycin.

Given the competence for DNA uptake of many gastro-intestinal bacteria [Lorenz and Wackernagel, 1994; Bertolla and Simonet, 1999] and that continuous sublethal concentrations of antibiotics favor the development of resistance [Grassi et al., 1980], it seems likely that intestinal bacteria have 'just' acquired resistance from the drug thought to inhibit their growth.

Although the USA, the EU and most industrial countries have now prohibited the use of antibiotics as growth promoters since 2005/2006 [Paul Ehrlich Society, 2006], emerging countries like China and Chile have not yet done so and are facing problems due to rising resistance in human pathogens, notably due to the use of quinolones amongst other antibiotics in animal husbandry [Cabello, 2006].

Motivated by these rising problems and recent studies which demonstrate that the usage of antibacterial growth promoters virtually brings no net economic gain [Collignon, 2004; Graham et al., 2007] (not counting increased costs in the health sector), the remaining countries will hopefully ban antibacterial growth promoters in the near future.

1.3 The Need for New Antibiotics

Because antibiotics have been used extensively for more than half a century in human and animal therapy, and as animal growth promoters, a drastic decline in the efficiency of many established antibiotics due to development of resistance in bacteria is the consequence. Additionally, the number of new antibiotics introduced on the market has become dramatically low. Both trends necessarily require reinforced efforts to discover and/or develop new antibiotics.

1.4 Possible Strategies for Discovering New Antibiotical Drugs

Since the genome sequence of *Haemophilus influenzae* was published in 1995 [Fleischmann et al., 1995], the number of sequencing projects has grown exponentially. The Genomes OnLine Database (GOLD) [Liolios et al., 2006] currently lists 438 finished and 1089 ongoing bacterial sequencing projects, and 37 finished and 59 ongoing archaeal sequencing projects (as of March

13, 2007). The number of metagenome sequencing projects is 2 finished and 73 ongoing. The genomes of virtually all severe human pathogens have been sequenced, and the sequence information of multiple strains or related subspecies of already sequenced pathogens (often with varying virulence) already does or will help to understand pathogenicity better with the help of comparative genomics.

The classic empirical approach that was applied in the "golden age" of natural product discovery started with the empirical observation of bacterial growth inhibition by colonies of microbes from environmental samples, followed by the exclusion of cytotoxic compounds in eukaryotic counterscreens, then the determination of the antibacterial spectrum, and determination of the mode of action in the pathogen and the eukaryotic cell [Freiberg and Brötz-Oesterhelt, 2005].

With the sequence information and annotation of bacterial model organisms and pathogens, complemented by the growing genome annotation of mammalian genomes including the human genome, this classical empirical approach could be complemented by a rational, target-directed antibacterial drug discovery strategy in the 1990s [Freiberg and Brötz-Oesterhelt, 2005]. Central to this approach is the belief that bacterial genomes harbor a variety of hitherto unexploited targets with the potential for being potent and selective antibiotics against a broad spectrum of bacterial pathogens [Payne et al., 2004; Allsop and Illingworth, 2002; McDevitt et al., 2002; Schmid, 2001]. Potential targets for new antibiotics are gene products that are conserved and essential in a broad number of pathogens but lack a close homolog in humans.

Unfortunately, despite significant efforts, only few genomics-derived compounds are currently in clinical development or in later preclinical stages. The reasons why many target-based screening approaches have failed to produce good lead structures are diverse (see Freiberg and Brötz-Oesterhelt [2005] and references therein). Important problems are, according to Freiberg and Brötz-Oesterhelt: *1.* high-throughput screening (HTS) of chemical libraries on purified enzymes that represent potential targets yielded hits that often lacked cell penetration; *2.* most compounds in large synthetic libraries were often too hydrophobic or too simple in structure to provide a good starting point for antibiotics; and *3.* target "screenability", according to HTS criteria, was sometimes more important than target quality.

At the same time, recent technology advances have led to a renaissance in the discovery of naturally produced antibiotics from microbial sources [Clardy et al., 2006].

In reinforced efforts, new antibiotics are being looked for in new sources (for example, actinomycetales, from which more than two-thirds of known secondary metabolites with antibacterial activity are derived [Challis and Hopwood, 2003], cyanobacteria and uncultured bacteria) and old sources (for example, streptomycetes, accounting for 70-80 % of the secondary metabolites produced by actinomycetes [Challis and Hopwood, 2003]).

The renewed discovery of natural antibiotics has two main motivations (quoting Clardy et al. [2006]):

first, "as antibiotics often reach their targets by transport rather than

diffusion, antibiotic candidates benefit from having structural features rarely found in the synthetic libraries of 'drug-like' molecules used in most high-throughput screening facilities"; and

second, "the well-established ability to discover useful antibiotics from natural sources suggests that continued efforts are likely to be fruitful".

Clardy et al. [2006] review several authors that have made estimations on the frequency of antibiotics in actinomycetes. Although 25% of the strains in a random soil sample are antibiotic producers, most of them produce known antibiotics classes (with the streptotricin, streptomycin and tetracycline class being the most abundant). Vancomycin and erythromycin would be rediscovered once in 70 000 and 200 000 strains, respectively. According to Baltz [2005], daptomycin (on the market since 2003) was found in one of 10 million actinomycete cultures screened.

In addition, Baltz estimates that less than one part in 10^{12} of the earth's soil surface has been screened for actinomycetes.

To further exploit this dormant reservoir of antibiotics, it will be necessary to screen 10^8-10^9 strains per year. Baltz points out that this will require a combination of high-throughput screening by modern technologies, selection against the most common antibiotics, methods to enrich rare and slow-growing actinomycetes, a prodigious microbial collecting and culturing effort, and combinatorial biosynthesis in streptomycetes (that is, manipulating the synthesis of antibiotics in their producer by genetic engineering).

Baltz [2006] reports that they have engineered an *E. coli* K12 strain – carrying 15 antibiotic-resistance genes – that will facilitate the sensitivity screening for novel antibiotics classes as none of the common broad spectrum antibiotics produced by actinomycetes will affect it.

1.5 Motivation and Scope of This Thesis

Bioinformatics already plays an important role in the discovery of new antibiotics and new drugs in general. Bioinformatics helps evaluating the huge amount of data generated by high-throughput experiments, helps comparing different entities of data (e.g. DNA or protein sequences) to determine their degree of similarity, and allows for predictions and simulations based on the principles and models that could be derived from the evaluation of the data itself or by theoretical approaches.

The classical empirical approach to discover new antibiotics described above is now complemented by bioinformatics in several ways.

As has already been mentioned before, the target-directed drug discovery strategy where computational comparative genomics is used to find possible drug targets and bioinformatics is employed to evaluate the data from high-throughput experiments.

But the new ultrafast DNA sequencing techniques (for an overview, see Kling [2003]) like pyrosequencing (reviewed by Ahmadian et al. [2006]) allow for the sequencing of a bacterial genome in about four days instead of one month using the Sanger method [Bonetta, 2006; Sanger et al., 1977], which is about to revolutionize modern biology, including drug discovery. When

a bacterial species, e.g. from an environmental sample, has been found to produce an interesting drug, its genome sequence can be determined quickly, opening the possibility for a bioinformatical analysis of the drug's anabolic pathway by the assignment (annotation) of putative functions to predicted gene products, which will facilitate further molecular biotechnological manipulations.

Computational comparative genomics takes a huge advantage of the accelerated speed in DNA sequencing as well.

By sequencing several strains of the same bacterial species that vary in drug resistance and/or virulence, it will be possible to understand the underlying molecular principles better. Sequencing strains that produce similar secondary metabolites – but with slight structural differences – will accelerate the elucidation of their anabolism.

The new DNA sequencing techniques also stimulate a new field, metagenomics, the study and exploitation of several genomes in parallel. All the DNA extracted from environmental samples (e.g. soil, sea water, deep sea vents) can be sequenced, analyzed for the presence of novel species [Venter et al., 2004; Poinar et al., 2006] and, more importantly, in drug discovery, for the presence of new enzymes with potentially new functions. For example, Venter et al. [2004] identified 1.2 million previously unknown genes in 1 billion base pairs sequenced from samples collected from the Sargasso Sea.

The technologic progress in DNA sequencing techniques has not been accompanied by comparable advances in the techniques available to experimentally determine the function of a gene. This is why the gap between the number of experimentally annotated and non-annotated sequences in the databases is growing. With this growing gap, the need for accurate function predictions of uncharacterized genes is growing steadily in order to reduce the number of "wet" lab experiments needed.

Even if computational prediction can often not completely replace the biological experiment, the number of hypotheses to be tested can be reduced in many cases.

To predict the function or at least the functional class of an unknown protein, the most successful strategy is to look for homologous annotated sequences in the databases, and assume that the unknown protein putatively carries out the same function. Two sequences are "homologous" if they share a common evolutionary origin [Berg et al., 2002], which is usually implied by a "sufficiently high" sequence similarity or sequence identity. As a rule-of-thumb, 30% sequence identity within a protein sequence alignment of length $\sim$150 amino acids [Rost, 1999] is just still "sufficiently high" but below 30% sequence identity, the so-called "twilight zone", where two sequences might share such similarity by chance without being related, starts.

Once the functional class of a protein has been determined (computationally or biochemically), it is often important to know exactly which substrate is processed and exactly what the product looks like.

In natural product research, compounds called polyketides (PKs) and non-ribosomal peptides (NRPs) play an important role, as many of the drugs applied today belong to this class of molecules. According to Borchardt [1999], 1 compound of 5000 molecules of typical synthetic chemical libraries

will become a drug, compared to 1 in 100 polyketides. The more than 40 polyketide drugs - including antibiotics, immunosuppressants, cholesterol-lowering agents, antifungals, and cancer chemotherapeutics - have a sales volume of more than US$15 billion per year.

The enzymes that synthesize PKs and NRPs (which have a similar pharmaceutical importance) are called polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs).

How NRPSs and PKSs can be identified automatically, given unannotated protein sequences, and how to predict their putative products, especially those synthesized by NRPSs, are the central questions for which existing approaches are discussed and new approaches are introduced within this doctoral thesis.

We provide new ideas, approaches, solutions and tools to predict the composition and order of NRPs, and highlight possible methods of further improvement. We discuss how these strategies could be applied to PKSs, and generalized for the prediction of functional subtypes and substrate specificities of other enzymes.

We want to contribute to the efforts of finding and understanding the synthesis of novel natural products which are highly valuable because of their high potential to be active as drugs against various diseases.

The burning need for new antibiotics discussed above, together with the rapid increase of non-annotated gene sequences in the databases stemming from genome and metagenome projects that hold hidden "treasures" to be discovered, is a great motivation for this work.

1.6 Overview on the Chapters of This Thesis

In Chapter 2, entitled *Biological Background*, we will discuss the architecture and functioning of NRPSs, give more examples of their products and compare them to the related PKSs. In Chapter 3, the *Technical Background* of this thesis, the reader will learn more about the different approaches and algorithms that are of fundamental importance for the whole dissertation. Materials and methods relevant to individual chapters are presented directly in their chapters in dedicated sections.

The subsequent chapters report on the most important results of this thesis project. Chapter 4 reports on a new strategy to predict the kind and the order of the building blocks (amino acids and the like) assembled by NRPSs, and Chapter 5 is about how the building blocks are connected with each other, including possible modifications they are subjected to at this step.

Chapter 6 reports the prediction of the building block selection in NRPS using structural bioinformatics approaches, and gives an outlook how this strategy could be continued and improved.

Chapter 7 concludes with the new predictive methods presented and their impact on other problems, followed by an outlook what bioinformatical challenges need to be solved in the NRPS and PKS field.

Chapter 2

Biological Background

2.1 Non-ribosomal Peptide Synthetases (NRPSs)

Non-ribosomal peptide synthetases (NRPSs) are large multimodular enzymes that synthesize a wide range of biologically active natural peptide compounds, of which many are pharmacologically important. A rich collection of them are used as drugs like antibiotics (e.g. penicillin and vancomycin), anti-tumorals and cytostatics (e.g. bleomycin), anti-inflamatorials and immunosuppressants (e.g. cyclosporin A), toxins (for example α-amanitin, the toxin of the mushroom *Amanita phalloides* (death cap)), or siderophores (iron chelators, e.g. yersiniabactin from *Yersinia pestis*). Scientifically, it is a challenge to discover how these structurally complex macromolecules are synthesized by the concerted interworking of the multi-domain proteins NRPS and polyketide synthases (PKS) that synthesize a peptide or ketide backbone with several other modifying and "decorating" enzymes (halogenases, glycosyl transferases etc.).

NRPSs belong to the family of megasynthetases, which are among the largest known enzymes with molecular weights of up to $\sim$ 2.3 MDa ($\sim$ 21,000 residues) [Wiest et al., 2002]. They possess several modules, each of which contains a set of enzymatic domains that, in their specificity, number, and organization, determine the primary structure of the corresponding peptide products. For recent reviews on NRPS, see, for example, Fischbach and Walsh [2006], Sieber and Marahiel [2005], or Lautru and Challis [2004].

A complete module contains at least three enzymatic domains, one Condensation (C) domain, one Adenylation (A) domain, and one Thiolation (T) domain (illustrated in Fig 2.1). By each module, one amino acid (or hydroxy acid) is appended to the peptide which is being synthesized. The relatively small T domain (8-10 kDa compared to 50 kDa of C and A domains) is post-translationally modified at a highly conserved serine, to which a phosphopantetheinyl (4'-PPant) group is attached by a phosphopantetheinyltransferase (PPTase, see Fig. 2.2).

The A domains are specific for a certain amino acid which they activate by adenylation (see Fig. 2.3 A). This unstable amino acid-adenylate is subsequently transferred to the downstream T domain onto its 4'-PPant cofactor, why the T domain may also be called peptidyl carrier protein (PCP).

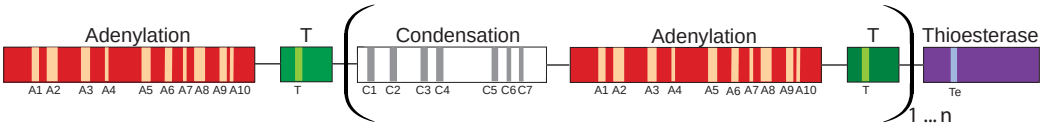

Figure 2.1: A minimal NRPS. The three enzymatic domains for Condensation (C), Adenylation (A), and Thiolation (T) form one complete minimal module. An NRPS assembly line which synthesizes one non-ribosomal peptide usually consists of several NRPS proteins. The first NRPS (usually) starts with an *initiation* module (A–T). Adenylation domains specifically bind a certain amino acid, activate it with an ATP forming an amino acyl adenylate and transfer it to a thiol group of the downstream T domain (see Fig. 2.3 A). The C domain catalyzes the peptide bond formation between the amino acid (or already synthesized peptide) which is tethered like on a pivot arm to 4'PPant of the upstream T domain (donor) and the amino acid at the downstream T domain (acceptor) (see Fig. 2.3 B; owing to its function, the T domain may also be called peptidyl carrier protein, PCP). This process is repeated until the peptide is passed onto the last T domain of an assembly line. Frequently, there are several NRPS enzymes that act sequentially in concert to synthesize one NRP. The last domain is usually a Thioesterase domain which cleaves the completed peptide off the last T domain. The number of repeats (n) in one NRPS is frequently below or around 10. The longest known NRPS synthesizes the 18 amino acid peptaibol peptide antibiotic in the fungus *Trichoderma virens* [Wiest et al., 2002]. The design of this representation is derived from Schracke [2005] with kind permission.

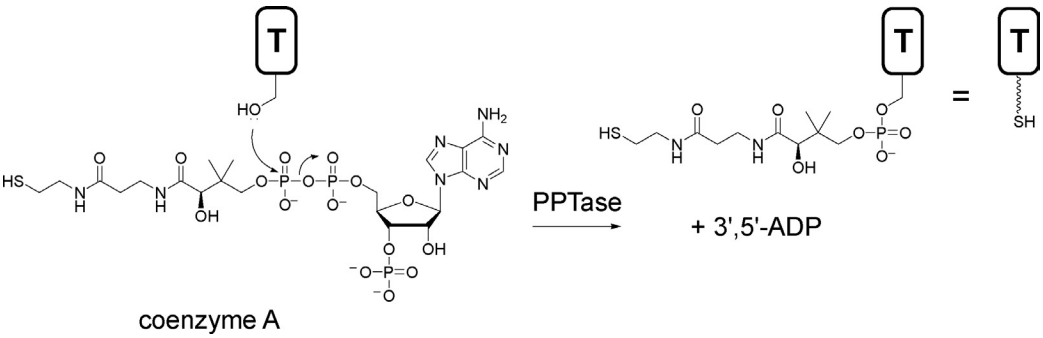

Figure 2.2: Posttranslational modification of T domains by phosphopantetheinyltransferases (PPTases) [Lambalot et al., 1996]. Enzymes of this class catalyze the transfer of phosphopantetheine from coenzyme A to a conserved serine in the T domain. Reprinted with kind permission from Fischbach and Walsh, ©2006 American Chemical Society.

The T domain changes conformations [Koglin et al., 2006] to "shuttle" the amino acid tethered to the 4'-PPant arm from the A domain to the next upstream (and then the next downstream) C domain, where the condensation reaction is catalyzed between one amino acid (or peptide) bound to the (upstream) *donor* T domain (T_1 in Fig. 2.3 B) and one amino acid bound to the (downstream) *acceptor* T domain (T_2 in Fig. 2.3 B), onto which the elongated product is transfered. The growing peptide is thus handed over by the T domains from one module to the next.

When the synthesized peptide arrives at the last T domain, it is cleaved off and released by a Thioesterase (TE) domain at the C-terminus of the NRPS (see Fig. 2.4). As in PKSs, TE domains in NRPS can be hydrolytic or cyclizing [Keating et al., 2001; Kohli and Walsh, 2003], liberating a linear or cyclic product [Samel et al., 2006].

In a few known NRPSs, the C-terminal TE domain is substituted by a Reductase (RE) domain, as occurs in safracin biosynthesis [Velasco et al., 2005], which catalyzes concomitant aldehyde formation and chain release.

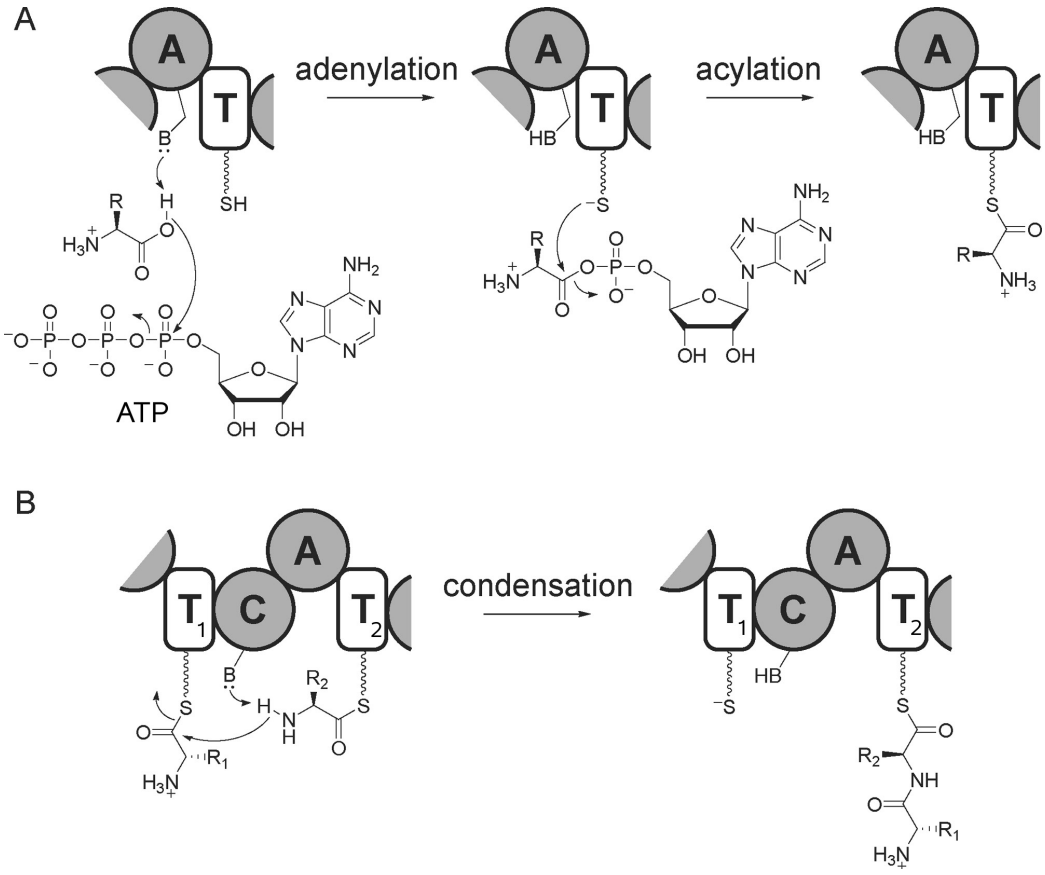

Figure 2.3: Chemical reactions catalyzed by the A and C domains. (A) The A domain catalyzes the adenylation of the amino acid to be incorporated and its subsequent acylation to the downstream T domain. (B) The C domain catalyzes C-N bond formation between the electrophilic upstream peptidyl-*S*-T_1 and the nucleophilic downstream aminoacyl-*S*-T_2. Reprinted with kind permission from Fischbach and Walsh, ©2006 American Chemical Society.

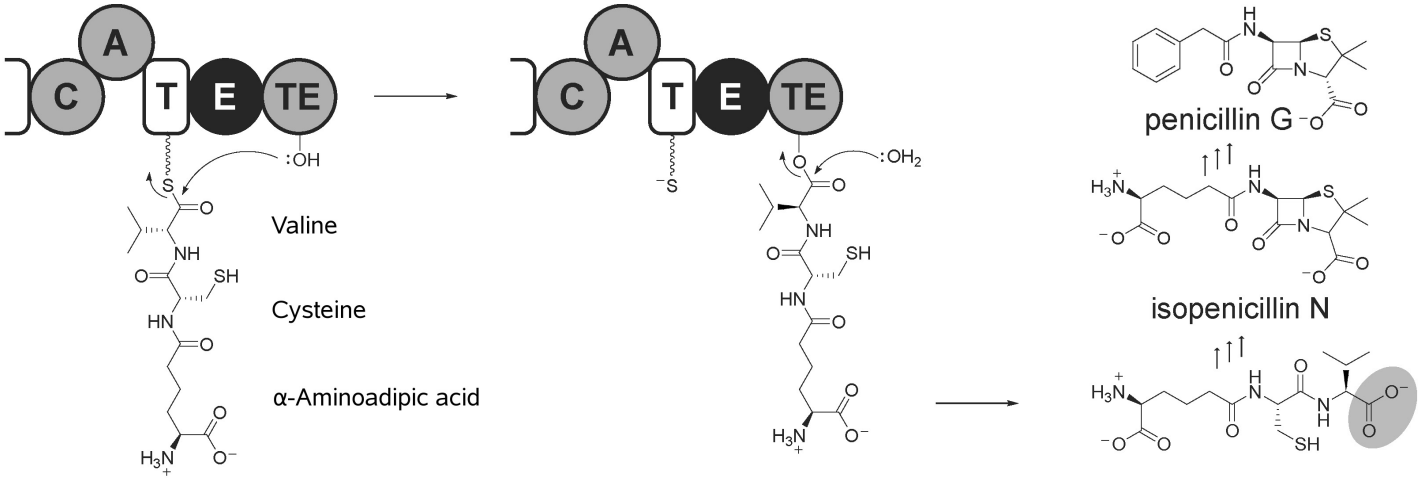

Figure 2.4: Chain termination by the Thioesterase (TE) domain in the example of penicillin synthesis. The TE domain of the ACV synthetase first acylates itself with the tripeptidyl group via a conserved serine and then catalyzes hydrolytic release. Prior to release, the valine is epimerized to D-valine by an Epimerization domain. The tripeptide is subsequently oxidatively cyclized to isopenicillin N by IPN synthase. In following steps, α-aminoadipic acid is replaced in a transacylation reaction by phenyl-acetic acid. Reprinted with kind permission from Fischbach and Walsh, ©2006 American Chemical Society.

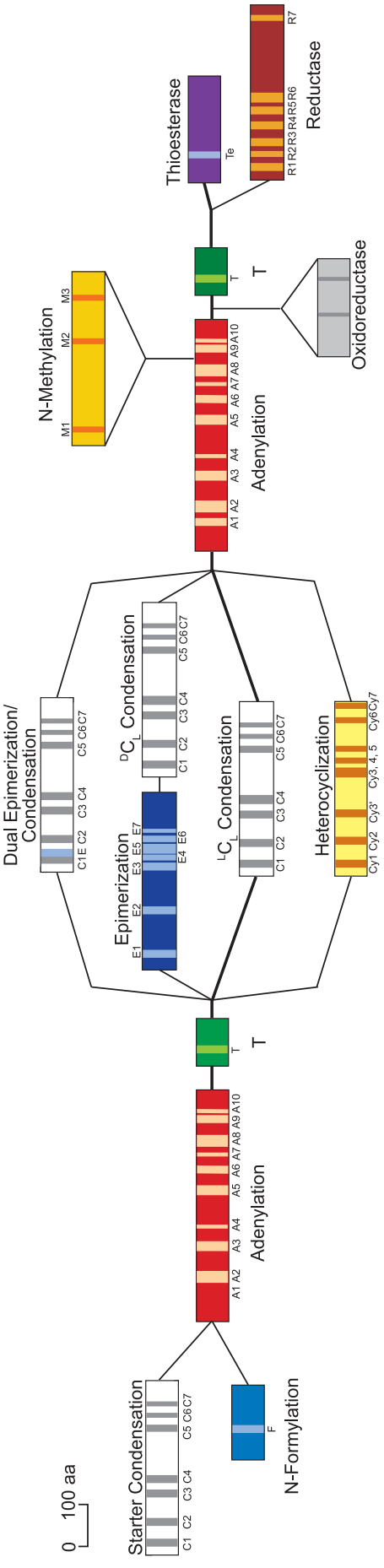

Figure 2.5: Optional and alternative domains, that may be found in NRPSs besides C, A, T and Thioesterase (TE) domains: Starter Condensation domains may be found at the N-terminus of the first NRPS in an assembly line; they acylate the first amino acid with a β-hydroxy-carboxylic acid (see Chap. 5). Similarly, the rare Formylation domains formylate the first amino acid [Schönafinger et al., 2006]. Different C domain functional variants and homologs exist (for details, see this chapter and Chap. 5). The Optional N-Methylation and Oxidoreductase domains will modify amino acids while they are appended to the growing peptide (more details in the text). The design of this representation is derived from Schracke [2005] with kind permission.

Typically, several NRPSs are involved in the synthesis of one NRP. Normally, the first module (called the *initiation* module) in the NRPS assembly line is a two-domain A–T module. In some cases, the first module has a three-domain C–A–T organization where the so-called Starter C domain acylates the first amino acid with a β-hydroxy-carboxylic acid (see Chap. 5 for details). More optional domains are shown in Fig. 2.5: The C domain can be preceeded by an Epimerization (E) domain, which changes the stereo-configuration of the amino acid bound to the upstream T domain. The C domain right after the E domain must thus be able to catalyze the peptide bond formation of a D-amino acid with an L-amino acid. Such C domains are called $^{\mathrm{D}}\mathrm{C}_{\mathrm{L}}$ domains as opposed to the more frequent $^{\mathrm{L}}\mathrm{C}_{\mathrm{L}}$ domains. Alternatively, a so-called Dual E/C domain can perform the same reactions as an E domain with a subsequent $^{\mathrm{D}}\mathrm{C}_{\mathrm{L}}$ domain. Heterocyclization domains can replace ordinary C domains, and catalyze the condensation reaction and a subsequent cyclization of amino acids (cysteine, serine or threonine) with an amide-nitrogen of the peptide "backbone", resulting in oxazolines (e.g. in vibriobactin) and thiazolines (e.g. in bacitracin); these can be further oxidized or reduced by Oxidoreductase domains [Du et al., 2000; Sieber and Marahiel, 2005]. N-Methylation domains, which are typically found to be integrated in A domains between their A8 and A9 motif, transfer a methyl group on the amino group of the amino acid that is about to be integrated [Patel and Walsh, 2001]. Furthermore, halogenation or hydroxylation may be mediated by specialized free-standing enzymes [Vaillancourt et al., 2005].

Occasionally, dehydration is performed on serines, resulting in dehydroalanine [Tillett et al., 2000]. Further modifications – glycosylation or phosphorylation – are usually performed by so-called "decorating" enzymes, usually clustered in proximity to the NRPS genes on the chromosome [Sieber and Marahiel, 2005].

Although the multi-domain proteins NRPS and PKS are also found in fungal and plant genomes, most of the known sequences stem from bacteria. The bacterial order *Actinomycetales* is known for the wealth of secondary metabolites produced by its members and comprises, among others, the *Streptomyces* species and mycobacteria. The majority of all currently known antibiotics and other therapeutic compounds are derived from streptomycetes [Brédy, 2005]. Many members of corynebacteria and mycobacteria are human pathogens which produce toxins as secondary metabolites. The structural and functional diversity of non-ribosomal peptides and their increased insensitivity to peptidases, incontrast to ribosomally synthesized peptides, arises from the incorporation of unusual amino acids (both proteinogenic and non-proteinogenic amino acids (e.g. ornithine), including D-amino acids) and the diverse modifications of the building blocks either directly by the NRPS assembly line or in the postprocessing by specialized enzymes (as detailed above).

2.2 Classification of NRPSs

NRPS can be classified according to their domain and module composition with respect to their product composition. Three types (A, B and C) can be differentiated: Linear NRPSs (type A), iterative NRPSs (type B) and non-linear NRPSs (type C) [Mootz et al., 2002b].

Most of the NRPSs that have been characterized biochemically to date belong to type A, e.g. surfactin, bacitracin, vancomycin and daptomycin. Here, the number of modules equals the number of amino acids in the products. Such NRPSs are typically composed according to the scheme A–T–(C–A–T)$_{n-1}$–TE as portrayed in Fig. 2.1 and resemble a linear assembly line that produces a product NRP of length n. Modifying domains may appear in any of the modules.

Prominent examples for type B NRPS are gramicidin S and enniatin, or the siderophores bacillibactin and enterobactin. In these iterative NRPSs, different domains or whole modules are used repeatedly. For example, at the gramicidin S synthesis, five modules are passed twice during the product synthesis, whereas the first pentapeptide stays fixed to the TE domain until the second round of synthesis is completed. Then the TE domain connects the first pentapeptide to the second one in a head-to-tail manner to form a cyclic homodimer [Schracke, 2005].

To date, type A and B are not distinguishable from each other based on their primary structure (peptide sequence and domain order) alone [Schracke, 2005].

The third type (C) includes all those NRPSs that cannot be classified into type A or B because of an unusual domain and/or module structure. Syringomycin, bleomycin or the siderophores yersiniabactin and vibriobactin are examples of type C.

2.3 Comparison of NRPSs to Polyketide Synthases (PKSs)

PKSs and NRPSs share the same general logic. In PKS assembly lines, the monomers are acyl-CoA thioesters (e.g., acetyl-CoA, malonyl-CoA, methylmalonyl-CoA), which are primary metabolites in the microbial producer cells. In analogy to the three core domains of a minimal NRPS module (C–A–T), there are three core domains in a minimal PKS module that carry out specific monomer recognition and binding (by the Acyltransferase (AT) domain) and elongation (by the Ketosynthase (KS) domain) of the growing polyketide chain which is tethered to a T domain: AT–T–(KS–AT–T)$_{n-1}$–TE is hence the prototype of the simplest PKS.

An important difference is that the monomers (acyl-CoA thioesters) are already available in an activated form thus the task of the AT domain is only to bind the acyl-CoA thioester specifically and mediate its transthiolation to the 4'PPant arm of the T domain. Note that ACP (acyl carrier protein) is synonymous with PKS T domains, as PCP is for NRPS T domains.

Whereas NRPS elongations involve C–N bond formation as an amide (peptide) link is forged in each condensation step, PKS elongations form C–C bonds via Claisen condensations [Fischbach and Walsh, 2006].

There are three types (I, II and III) of PKSs, but only the first two types correspond to each other directly; the third types are particular. The type III PKS subgroup is distinguished from types I and II by the direct use of malonyl-CoA rather than via malonyl-S-pantetheinyl-T. We refer to Austin and Noel's review [2003] on type III PKSs for details.

Like NRPS systems, several optional modifying domains are frequently found in PKS assembly lines. Three additional catalytic domains – invariably present in type I and II Fatty Acid Synthases (FASs) but optional in type I and II PKSs – are Ketoreductase (KR), Dehydratase (DH) and Enoylreductase (ER) domains. Exactly like in FASs, they operate sequentially: KR $\rightarrow$ DH $\rightarrow$ ER. The KR domain first reduces the β-ketoacyl-S-T which results from the KS-mediated condensation, then the DH domain dehydrates to α,β-enoyl-S-T, and in the final step, the ER domain reduces the conjugated olefin to the saturated acyl-S-T. In this manner, these three enzymatic groups catalyze the complete reduction of the keto group (>=O) via an alcohol group (>–OH) and a conjugated double bound to the alkyl (>). In fatty acid synthesis, these three steps result in an elongation by a CH_2–CH_2 unit. But in PKS assembly lines, the last reduction step (ER), the second and the last (DH $\rightarrow$ ER), or all three may be missing, therefore the combinatorial number of possible products increases dramatically (four possible oxidation states for each integrated monomer). Additionally, a broad selection of PK starter units (incorporated by the first AT domain), modifying PKS domains, the possibility of PKS-NRPS hybrid assembly lines and several post-assembly-line tailoring enzymes increase the "space" of chemical molecules that can/could be produced by these megasynthases. For a recent review on the logic, machinery, and mechanisms of PKS and NRPS, see Fischbach and Walsh [2006].

Chapter 3

Technical Background

This chapter describes the general concepts, methods and materials of great importance for this whole thesis. The projects presented in Chapters 4, 5 and 6 have own *Materials and Methods* Sections that report on the material and methods relevant only to the specific topic of each chapter.

The following section (3.1) on optimization theory is necessary for the complete understanding of Support Vector Machines (SVMs) described in Section 3.2. Readers who only want to get the basic notions of SVMs might directly skip to Section 3.2, have a look at Figures 3.3 and 3.4, and read the summary of SVMs in Section 3.2.8 on page 36.

3.1 Optimization Theory

In this section, we will introduce Lagrange multipliers and Karush-Kuhn-Tucker (KKT) conditions (Section 3.1.2) which will be used to construct the optimal separating hyperplane of Support Vector Machines presented in Section 3.2. KKT conditions and Lagrange multipliers are important concepts in constrained optimization problems of differentiable functions, which will be discussed in Section 3.1.2.

3.1.1 Unconstrained Optimization

First, we want to outline a simple unconstrained case: Suppose we are given a (differentiable) function $f(x_1, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R}$ where we want to find its extremum (its maximum or minimum). Given a definition domain $D(f)$, a point $\bar{x} \in D(f) \subseteq \mathbb{R}^n$ is called a *(global) minimum* of f on $D(f)$ if $f(\bar{x}) \leq f(x) \quad \forall x \in D(f)$ and is called a *(global) maximum* of f on $D(f)$ if $f(\bar{x}) \geq f(x) \quad \forall x \in D(f)$. If this is true only for a neighborhood $U \subseteq D(f)$ around $\bar{x}$, the point $\bar{x}$ is called a *local minimum* or a *local maximum*, respectively. A necessary condition for an extremum point $\bar{x}$ is that the gradient $\nabla f(\bar{x})$ is zero, or, equivalently, each of the partial derivatives is zero. The problem $\min_{x \in D(f)} f(x)$ is called an *unconstrained optimization* problem. The function f is called the *objective function*. A point $\bar{x} \in D(f)$ with $\nabla f(\bar{x}) = 0$ is also called a *stationary point* of f. A property of a stationary point $\bar{x}$ is that

the tangent plane of f is horizontal at $\bar{x}$ (in $\mathbb{R}^2$ and analogous in higher dimensions). This implies that besides the possibility of being a local or a global minimum or maximum, such a point may also be a saddle point (that is, an inflection point with a horizontal tangent). An important property of extrema (as opposed to saddle points) is that the curvature is not zero at such points. If f is globally *convex*, which means that the curvature does not change its sign for all $x \in \mathbb{R}^n$, then the Hessian matrix (square matrix of second partial derivatives of f) is positive semi-definite for all x (and vice versa), which implies that there is a global minimum at $\bar{x}$. Analogously, the same is true for a global maximum of globally concave functions for which the Hessian matrix $H(x)$ is always negative semi-definite. If the Hessian matrix $H(\bar{x})$ is positive definite (or negative definite) only in a neighborhood around $\bar{x}$, then we cannot decide whether $\bar{x}$ is a global or local minimum (or maximum if $H(\bar{x})$ is negative definite). If $H(\bar{x})$ is indefinite then there is a saddle point at $\bar{x}$. In practice, however, it is often easier to calculate the function values around $\bar{x}$ to decide whether there is a minimum, maximum or saddle point than to compute the definiteness of the Hessian matrix.

3.1.2 Constrained Optimization

As in the unconstrained case, there is the function $f(x)$ to minimize (by multiplying the function by -1, we can turn all minima into maxima and vice versa). We are only interested in a restricted set of points in $\mathbb{R}^n$ that satisfy a given set of side conditions called constraints.
The problem

$$\min f(x)$$

under the constraints:

$$\begin{array}{rcl} g_1(x) & \leq & 0 \\ & \vdots & \\ g_m(x) & \leq & 0 \\ h_1(x) & = & 0 \\ & \vdots & \\ h_p(x) & = & 0 \\ x & \in & \mathbb{R}^n \end{array}$$

is called a *constrained optimization problem*. The set of all vectors x that satisfy all constraints is called the *feasible region*.

Theorem 3.1 Necessary optimality conditions for constrained problems (Karush-Kuhn-Tucker conditions)
Suppose the following five conditions are fulfilled:

1. The functions
   $f(x), g_1(x) \leq 0, \ldots, g_m(x) \leq 0, h_1(x) = 0, \ldots, h_p(x) = 0 : \mathbb{R}^n \to \mathbb{R}$
   are given.

2. The point $\bar{x} \in \mathbb{R}^n$ is a local minimum of f on the *feasible region* $M := \{x \in \mathbb{R}^n \mid g_1(x) \leq 0, \ \ldots, \ g_m(x) \leq 0, \ h_1(x) = 0, \ \ldots, \ h_p(x) = 0\}$.

3. The functions $f, g_1, \ldots, g_m$ are differentiable at $\bar{x}$ and the functions $h_1, \ldots, h_p$ are continuously differentiable at $\bar{x}$.

4. The system of vectors $\nabla h_1(\bar{x}), \ldots, \nabla h_p(\bar{x})$ is linearly independent.

5. There exists a vector $y \in \mathbb{R}^n$ with $\nabla g_i(\bar{x})^T \cdot y < 0 \quad \forall i \in I(\bar{x})$ and $\nabla h_i(\bar{x})^T \cdot y = 0 \quad \forall i \in \{1, \ldots, p\}$, with $I(\bar{x}) := \{i \in 1, \ldots, m | g_i(\bar{x}) = 0\}$.

Then there exist multipliers $\alpha_i \geq 0$ $(i \in I(\bar{x}))$ and $\beta_1, \ldots, \beta_p \in \mathbb{R}$ with

$$\nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{i=1}^{p} \beta_i \nabla h_i(\bar{x}) = \vec{0}.$$

The necessary optimality conditions 1-5 in Theorem 3.1 are called the *Karush-Kuhn-Tucker conditions* (*KKT condition*, a result obtained independently by W. Karush [1939], F. John [1948], and by H.W. Kuhn and J.W. Tucker [1951]; see Fiacco and McCormick [1987]). If there are only equality restrictions, then the optimality condition is also called the *Lagrange multipliers rule*. The coefficients α_i $(i \in I(\bar{x}))$, $\beta_i, \ldots, \beta_p$ are called *Lagrange multipliers*. The function

$$\mathscr{L}(x, \alpha, \beta) = f(x) + \sum_{i \in I(\bar{x})} \alpha_i g_i(x) + \sum_{i=1}^{p} \beta_i h_i(x)$$

is called the *(generalized) Lagrangian Function*.

Theorem 3.2 Sufficient optimality condition for constrained problems (Karush-Kuhn-Tucker theory):
Suppose we are given the objective function f with the constraining functions specified in Theorem 3.1 with the additional requirement that f is convex and the constraints $g_1, \ldots, g_m, h_1, \ldots, h_p$ are affine-linear. Necessary and sufficient conditions for $\bar{x}$ to be an optimum are the existence of α, β such that

$$\begin{aligned}
\frac{\partial}{\partial x}\mathscr{L}(\bar{x}, \alpha, \beta) &= 0; \\
\frac{\partial}{\partial \beta}\mathscr{L}(\bar{x}, \alpha, \beta) &= 0; \\
\alpha_i g_i(\bar{x}) &= 0 \quad \forall i \in I(\bar{x}); \\
g_i(\bar{x}) &\leq 0 \quad \forall i \in I(\bar{x}); \\
\alpha_i &\geq 0 \quad \forall i \in I(\bar{x}); \\
h_i(\bar{x}) &= 0 \quad \forall i \in 1, \ldots, p; \\
\beta_i &\in \mathbb{R} \quad \forall i \in 1, \ldots, p;
\end{aligned}$$

with $I(\bar{x}) := \{i \in 1, \ldots, m | g_i(\bar{x}) = 0\}$.

Proof: See [Fletcher, 1987, p. 218]. □

Active Constraints

The concept of an *active constraint* is important in the context of Lagrangian & KKT theory and also relevant for the Support Vector Machine theory. A constraint g_i is called *active* at x if $g_i(x) = 0$. Consequently, any constraint is active at x if x is at the boundary of its feasible region. This implies that equality constraints are always active. Active constraints at the optimal solution $\bar{x}$ are of particular interest. If the set of active constraints at $\bar{x}$ $\mathscr{A}(\bar{x}) = \{g_i \mid g_i(\bar{x}) = 0\}$ is known, then the remaining (inactive) constraints can be ignored locally.

Example of an Optimization Problem with One Equality Constraint

In the two-dimensional case, it is often possible to solve constrained problems graphically. The following example will help us to better understand the idea of the Lagrangian theory. Let us take a function $f(x, y)$ that we want to maximize, as well as the constraint $g(x, y) - c = 0$ with a given constant c. We can draw a contour plot of f as depicted in Fig 3.1. A contour line or level curve of a function indicates where the function has a certain value, like the contour lines on a topographical map indicating the altitude.

The point set defined by $g(x, y) - c = 0$ is depicted as a green curve in Fig 3.1 and visualizes the feasible region of the given constraint optimization problem, which actually means that our constraint optimum must lie on g's contour line $g(x, y) = c$.

Let us assume that Fig 3.1 depicts a topographical map that shows the contour lines of a hill $f(x, y)$ and the course of a tunnel which runs at constant height $0 = g(x, y) - c$. The constrained problem is thus to find where the hill has its highest elevation over the tunnel, which is not necessarily its summit (the global maximum). Assuming we follow the tunnel (green curve, $g(x, y) - c = 0$) on the (2-dimensional) map, we will cross many contour lines of f. Recall that the gradient (∇) of a function is a vector that points towards the steepest ascent of the function and lies in the plane of the input values of the function (in this case, the x-y plane in which our map lies). We will continue walking along $g = c$ as long as we still advance in the direction of the gradient (in other words, the direction vector of our walk can be decomposed into a positive non-zero component collinear to the gradient at the point in which we cross f's contour line and one component vertical to it). The hill has its highest elevation at the point where $g = c$ touches a contour line of f tangentially. This point is the maximum of the constrained optimization problem. Geometrically, we can interpret this tangent condition by saying that the gradients of f and $g - c$ are parallel vectors at the constrained maximum. We introduce a scalar β and obtain $\nabla f(\bar{x}) + \beta \nabla g = \vec{0}$, which corresponds to the multipliers rule.

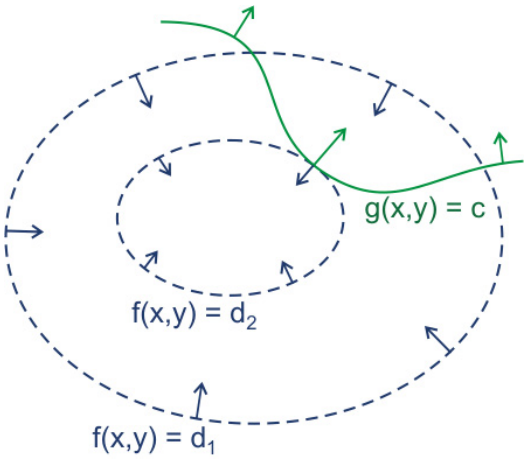

Figure 3.1: The concentric dashed ellipses (in blue) are contour lines (level curves) of the function $f(x, y)$ marking function values ("altitudes") of d_1 and d_2, respectively. The green curve corresponds to the feasible region of the constrained optimization problem given by the contour line of function g ($g(x, y) = c$). The solution must be a point for which $g(x, y) = c$ is satisfied and $f(x, y)$ is maximal, which is the case at the point where the gradients (depicted as arrows) on the contour lines of both functions are collinear. At this point, the level curves $g(x, y) = c$ and $f(x, y) = d_2$ touch tangentially. Image source: Wikipedia [2007].

3.2 Support Vector Machines

Support Vector Machines (SVMs) find their application in Chapter 4 on the specificity prediction of Adenylation domains in NRPS. Here, we introduce their principle theory.

Basically, SVMs are binary classifiers, which means that they can be used as a decision function that will return "yes" or "no" for a given input data point.
Vapnik and colleagues [Vapnik and Lerner, 1963; Vapnik and Chervonenkis, 1964] laid the foundations of SVMs in the 1960s by defining the *optimal hyperplane algorithm* (discussed in Section 3.2.4) for optimally separable data. In the 1990s, SVMs received growing scientific interest after a publication by Boser, Guyon and Vapnik [1992] on SVMs with kernels (see Section 3.2.5) and by Cortes and Vapnik [1995] on SVMs that can handle errors in the data sets (presented in Section 3.2.4) which turned SVMs into a very powerful and flexible tool for the classification of real-world data.

In the subsequent years, SVMs have been applied to a growing number of problems, including (but not limited to) particle identification, object recognition, text categorization (e.g. spam filters), hand-written character recognition, image classification, voice recognition and face detection. Soon SVMs were also applied very broadly in the field of bioinformatics to pattern recognition problems, including protein remote homology detection, microarray gene expression analysis, recognition of translation start sites, functional classification of promoter regions, prediction of protein-protein interactions, and peptide identification from mass spectrometry data [Noble, 2004]. This success is due to the excellent performance of SVMs compared to

other machine-learning algorithms. Their importance is still growing as they are being improved and adapted to new fields; to give one cutting-edge example, gene prediction with all predictive steps based on SVMs (see publications by Gunnar Rätsch [2007, `www.fml.mpg.de/raetsch`]).

The goal of this section is to introduce the theory of SVMs to a readership not familiar with machine-learning. Because it is still a relatively young field, the amount of introductory and didactically well prepared literature is still very limited. Research papers developing or applying SVMs usually refer to previous literature which is often mathematically abstract. The two books by Vladimir Vapnik (*The Nature of Statistical Learning Theory* [1995] and *Statistical Learning Theory* [1998]), which are often given as references for SVMs, present a general high-level introduction to statistical inference including SVMs. Chris Burges [1998] published a tutorial on SVMs, and Cristianini and Shawe-Taylor [2000] authored a book, *An Introduction to Support Vector Machines*; these are recommended reading. Furthermore, Schölkopf et al. [2004] edited a book on *Kernel Methods in Computational Biology* which is especially interesting in the context of this thesis in bioinformatics. All these books and tutorials convey a very good overview over the SVM theory to the reader but either they lack mathematical details in optimization theory and kernel techniques (like Burges [1998]), or the basic ideas of SVMs are explained only after one hundred pages or more [Vapnik, 1998; Cristianini and Shawe-Taylor, 2000]. For me, the lecture by Markowetz [2003] was a great help in my initial understanding of the idea of SVMs. Markowetz also introduced SVMs in his impressive Master's thesis [Markowetz, 2001] which is very well written and recommended reading.

In the following section, SVMs are introduced in a similar way as done by Markowetz [2001], Cristianini and Shawe-Taylor [2000] and Burges [1998] but with an effort to be as comprehensible as possible. For example, intermediate steps are explicitely written if they facilitate understanding.

3.2.1 Learning from Examples

A Support Vector Machine (SVM) is a function that assigns each input value to a *positive* or *negative class*; we also say it assigns a *positive* or *negative label*. A typical example would be the classification of text, e.g. the decision whether an email is spam or not. Here each email is encoded in one (high-dimensional binary) vector where each component represents the presence (e.g. $+1$) or absence (e.g. -1) of a certain word or word root in the email. To obtain a well adapted SVM, it has to be trained on data points whose labels are known, the so-called training data; in our example, this would be emails that have been sorted into spam and non-spam by a knowledgeable person, also called the *expert* or the *supervisor* in the context of machine learning. We can represent the training data as a set

$$\mathcal{X} = \{(x_1, y_1), \ldots, (x_n, y_n) : x_i \in \mathbb{R}^d, y_i \in \{-1, +1\}\},$$

where x_i are the data points and y_i their label, which can be either -1 or $+1$. The decision function $f_{\mathcal{X}} : \mathbb{R}^d \to \{-1, +1\}$ maps the input vectors x_i to the negative or positive class.

3.2.2 Generalization Ability: Performance on the Test Data

We want the SVM to generalize from the training examples to the whole range of observations (as well as possible). The quality of an SVM will be measured on how well it can classify new data that did not belong to the training set. Ideally, these test data should represent the complete diversity. This ability to achieve a small error rate (also called a small loss) on test data is termed *generalization ability*. The goal in learning theory is thus to maximize the generalization ability of the classifier, or, equivalently, to minimize the so-called risk functional (for more details on risk minimization, see Vapnik [1995, p. 72]).

3.2.3 Capacity: Performance on the Training Data

The *capacity* describes the ability of the machine to learn a given training set without error and measures the *richness* or *flexibility* of the function class. Chris Burges gives a nice example to illustrate *capacity* and *generalization ability*: "A machine with too much capacity is like a botanist with a photographic memory [i.e. very high capacity] who, when presented with a new tree, concludes that it is not a tree because it has a different number of leaves from anything she has seen before; a machine with too little capacity is like the botanist's lazy brother, who declares that if it's green, it's a tree. Neither can generalize well." [Burges, 1998, p. 2]. The problem arising from too much capacity is called *overfitting*, the other extreme is called *underfitting* and is illustrated in Fig 3.2.

3.2.4 Linear SVMs

To separate data points in the $\mathbb{R}^2$ into two classes, a simple and intuitive way is to construct a separating straight line, and a separating plane in $\mathbb{R}^3$. In higher-dimensional space we talk about hyperplanes. A separating hyperplane is defined by its *normal vector* w and its *offset* b (the distance by which the plane is displaced from the origin of the co-ordinate system):

$$\text{Hyperplane } \mathcal{H} = \{x | \langle w, x \rangle + b = 0\}$$

with $w \in \mathbb{R}^d, b \in \mathbb{R}$ and $\langle \cdot, \cdot \rangle$ denoting the dot product or scalar product (exchangeable expressions). If we construct the separating hyperplane so that w points towards the positive points, the decision function

$$f(x) = \text{sign}(\langle w, x \rangle + b)$$

will return $+1$ for points lying on the positive side of the hyperplane and -1 for points on the negative side. Obviously, there are many possible ways to place a hyperplane that will separate the two classes. Hence, we look for the optimal separating hyperplane (OSH). A basic assumption of learning from examples is that new data points are believed to lie close to or in-between the known training data. Therefore, the OSH should allow small deviations in

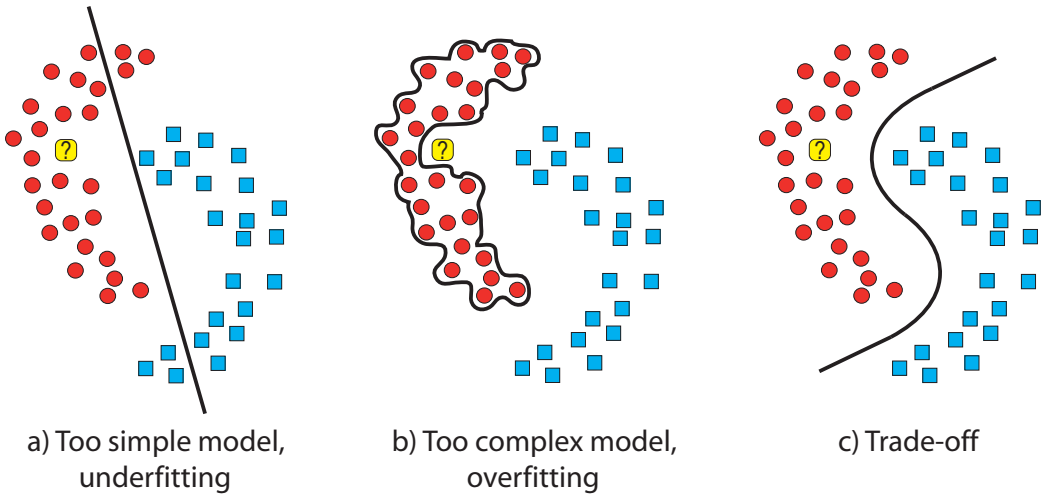

Figure 3.2: Illustration of overfitting and underfitting. Training data are shown in the shape of two intertwining crescents (positive data points are red dots; negative data points are blue squares). The yellow box with the question mark stands for a data point with an unknown label. Intuitively, we would classify the new data point within the red crescent. A very simple linear model (a) might correctly classify the new data point but might have errors in the training data. The danger of a very complex model (b) that quasi learns the training data by rote is that it will be too intolerant of small acceptable deviations in the data; of the illustrated case (b), the model would misclassify the new data point. Model (c) represents the desirable trade-off with a good generalization that classifies the new data point correctly.

the data and be in the middle of the structures of the positive and negative data clouds. Fig. 3.3 shows the optimal separating hyperplane, and a bad separating hyperplane that would misclassify test examples even very close to the training data.

When we look at Fig.3.3, the optimal separating hyperplane in (a) has a large margin. The concept of maximal margin hyperplanes was introduced by Vapnik and Lerner [1963], and Vapnik and Chervonenkis [1964] based on the intuition that the larger the margin, the better the generalization ability. In the following section, we will describe how optimal separating hyperplanes (OSHs) can be constructed efficiently. First, we will consider linear separable datasets and then show how these results can be generalized to allow for errors on the training set.

Optimal Separating Hyperplanes for Linearly Separable Data

Definition 3.1 We call a training set $\mathcal{X} = \{(x_1, y_1), \ldots, (x_n, y_n) : x_i \in \mathbb{R}^d, y_i \in \{-1, +1\}\}$ ***separable*** by a hyperplane $\langle w, x\rangle + b = 0$ if both a unit vector w ($\|w\| = 1$) and a constant b exists so that the following inequalities are fulfilled (in which case we talk about a separating hyperplane):

$$\langle w, x_i\rangle + b > 0 \quad \text{if} \quad y_i = +1 \tag{3.1}$$

$$\langle w, x_i\rangle + b < 0 \quad \text{if} \quad y_i = -1 \tag{3.2}$$

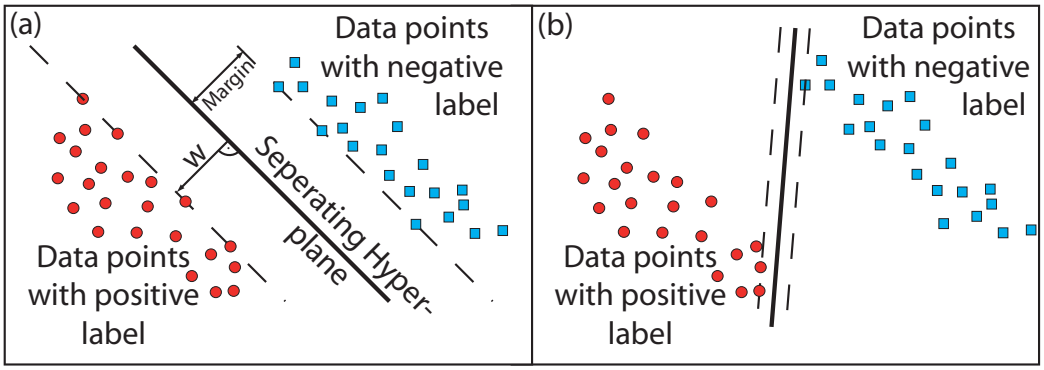

Figure 3.3: Optimal (a) and bad (b) separating hyperplane. The optimal separating hyperplane in (a) clearly separates the data much better than hyperplane (b). The margin in (b) is almost zero compared to the large margin in (a).

Definition 3.2 Recall that the distance of a point x_i to the hyperplane $H = \{x | \langle w, x \rangle + b = 0\}$ is:

$$d_{x_i}(w, b) = y_i(\langle w, x_i \rangle + b)$$

The ***margin*** $\gamma_{\mathcal{S}}(w, b)$ of a set of vectors $\mathcal{S}$ is defined as the minimum distance from H to the vectors in $\mathcal{S}$:

$$\gamma_{\mathcal{S}}(w, b) = \min_{x_i \in \mathcal{S}} d_{x_i}(w, b)$$

Formulating the Optimization Problem to Find the OSH
Given the training set $\mathcal{X}$, we thus want to find the separating hyperplane that maximizes the margin, the so-called *optimal separating hyperplane* (OSH) or *maximal margin hyperplane*. According to this, we need to find the unit vector w and the constant b that maximize the margin of the training set $\gamma_{\mathcal{X}}(w, b)$:

$$\begin{array}{lll} \text{maximize} & \gamma_{\mathcal{X}}(w, b) & \\ \text{subject to} & \gamma_{\mathcal{X}}(w, b) & > 0 \\ & \|w\|^2 = 1 & \end{array} \tag{3.3}$$

It is algorithmically difficult to solve the optimization problem in 3.3 because the constraints are non-linear and the objective function is non-linear and non-quadratic. Therefore, an equivalent formulation of the problem is needed; to construct it, we will take advantage of the scalability of the parameters of the hyperplane equation: The equations $\langle w, x \rangle + b = 0$ and $\langle cw, x \rangle + cb = 0$ (with $c \neq 0$ and $c \in \mathbb{R}$) describe the same hyperplane. It can be shown that the OSH itself is unique [Vapnik, 1998, p. 402] but using a scaling factor c, we obtain an infinite number of equivalent representations of the same OSH.

With γ being the size of the margin (positive values for the margin towards the positive data points, negative values towards the negative data points), consider the following:

$$\begin{aligned}(\langle w, x_i\rangle + b) &\geq \quad \gamma \qquad && \forall i \in I_+ = \{i \mid y_i = +1\} \\ (\langle w, x_j\rangle + b) &\leq \quad -\gamma \qquad && \forall j \in I_- = \{j \mid y_j = -1\}.\end{aligned} \tag{3.4}$$

With a scaling factor c, this can be transformed to:

$$\begin{aligned}c(\langle w, x_i\rangle + b) &\geq \quad c\gamma \qquad && \forall i \in I_+ \\ c(\langle w, x_j\rangle + b) &\leq \quad -c\gamma \qquad && \forall j \in I_-\end{aligned} \tag{3.5}$$

We substitute with $w^* := cw$, $b^* := cb$, scale aptly to obtain $c\gamma = 1 \iff c = \frac{1}{\gamma} = \|w^*\|$ and also use the y-values of the data points (that are $+1$ or -1 according to the class). Thus we obtain:

$$y_i(\langle w^*, x_i\rangle + b^*) \geq \quad 1 \qquad \forall i = 1, \ldots, n. \tag{3.6}$$

If we scale back to the plane description using the unit vector $w = \frac{w^*}{\|w^*\|}$

$$y_i(\langle \frac{w^*}{\|w^*\|}, x_i\rangle + b) \geq \quad \tfrac{1}{\|w^*\|} \qquad \forall i = 1, \ldots, n,$$

we can see that data points that have exactly the distance of the margin to the hyperplane have the distance $\frac{1}{\|w^*\|}$. Consequently, our optimization problem is to maximize $\frac{1}{\|w^*\|}$, or equivalently to minimize $\|w^*\|$, which is in turn equivalent to minimizing $\|w^*\|^2$. For cosmetic reasons, we will minimize $\frac{1}{2}\|w^*\|^2$, which will not change the solution:

$$\begin{aligned}&\text{minimize} && \tfrac{1}{2}\|w^*\|^2 \\ &\text{subject to} && 1 - y_i(\langle w^*, x_i\rangle + b^*) \leq 0 \qquad i = 1, \ldots, n.\end{aligned} \tag{3.7}$$

This is a quadratic optimization problem with linear constraints. As described in Section 3.1.2 on constrained optimization, the Lagrange method can be used to solve such problems. The problem is convex as the objective function is convex and the constraints describe a convex feasible region. Therefore, we may introduce Lagrange multipliers $\alpha_i \geq 0$ and combine the optimization problem in a Lagrangian function $\mathscr{L}(w^*, b^*, \alpha)$:

$$\mathscr{L}(w^*, b^*, \alpha) = \frac{1}{2}\|w^*\|^2 + \sum_{i=1}^{n} \alpha_i[1 - y_i(\langle w^*, x_i\rangle + b^*)].$$

We need to calculate the derivatives of $\mathscr{L}(w^*, b^*, \alpha)$ with respect to w^*, b^* and α:

$$\text{From } \frac{\partial}{\partial b^*}\mathscr{L}(w^*, b^*, \alpha) = 0 \qquad \text{we obtain:} \qquad \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{3.8}$$

$$\text{From } \frac{\partial}{\partial w^*}\mathscr{L}(w^*, b^*, \alpha) = 0 \qquad \text{we obtain:} \qquad w^* = \sum_{i=1}^{n} \alpha_i y_i x_i. \tag{3.9}$$

If we substitute 3.8 and 3.9 into the Lagrangian $\mathscr{L}$, it can be transformed as follows and we obtain the so-called *dual problem*:

$$\begin{aligned}
\mathscr{L} &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i [1 - y_i(\langle w^*, x_i\rangle + b^*)] && (3.10)\\
&= \frac{1}{2}\Big(\sum_{i=1}^{n} \alpha_i y_i x_i\Big)\Big(\sum_{j=1}^{n} \alpha_j y_j x_j\Big) + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i b^* y_i \\
&\quad + \sum_{i=1}^{n} \Big[\Big(\alpha_i y_i x_i\Big)\Big(\sum_{j=1}^{n} \alpha_j y_j x_j\Big)\Big] \\
&= \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle + \sum_{i=1}^{n} \alpha_i - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle. && (3.11)
\end{aligned}$$

The optimization problem can thus be formulated:
Find multipliers which

$$\begin{aligned}
&\text{maximize} && \mathscr{L}(\alpha) = \textstyle\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle && (3.12)\\
&\text{subject to} && \alpha_i \geq 0 \qquad i = 1, \ldots, n, && (3.13)\\
& && \textstyle\sum_{i=1}^{n} \alpha_i y_i = 0. && (3.14)
\end{aligned}$$

To solve the dual problem, let us first look at the addends that the double sum will add up:

$$\begin{bmatrix}
\alpha_1\alpha_1 y_1 y_1 \langle x_1, x_1\rangle & \cdots & \alpha_1\alpha_n y_1 y_n \langle x_1, x_n\rangle \\
\vdots & \ddots & \vdots \\
\alpha_n\alpha_1 y_n y_1 \langle x_n, x_1\rangle & \cdots & \alpha_n\alpha_n y_n y_n \langle x_n, x_n\rangle
\end{bmatrix}$$

To get the extremum of the dual problem (3.12), we have to substitute one α_i in 3.12, e.g. α_1, using the equality condition 3.14. In the next step, the partial derivatives $\frac{\partial}{\partial \alpha_i}$ for each α_i have to be determined which will lead to a system of n equations with quadratic terms for each α_i. This quadratic program (QP) can be solved efficiently with algorithms called "QP solvers".

Once we have found the coefficients α_i^* that solve the dual problem, we obtain the w^* of the OSH with the maximal margin:

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i. \tag{3.15}$$

The value of b^* did not appear in the dual problem but can be derived from the constraints 3.6 of the initial optimization problem:

$$b^* = -\frac{\max_{y_i=-1}\big(\langle w^*, x_i\rangle\big) + \min_{y_i=1}\big(\langle w^*, x_i\rangle\big)}{2} \tag{3.16}$$

Support Vectors

At this point, we have all necessary parameters to write down the decision function needed to predict the classification of a new data point x_{new}:

$$f(x_{new}) = \text{sign}(\langle w^*, x_{new}\rangle + b^*) \tag{3.17}$$

$$= \text{sign}\big(\sum_{i=1}^{n} \alpha_i y_i \langle x_i, x_{new}\rangle + b^*\big). \tag{3.18}$$

The optimal solution satisfies the KKT conditions of Theorem 3.2 on page 19:

$$\alpha_i^*[y_i(\langle w^*, x_i\rangle + b^*) - 1] = 0 \qquad \forall i.$$

This equation implies that for a given data point x_i, *either* the corresponding α_i must be zero *or* the term in squared brackets, which is exactly zero if the point x_i lies "on the margin", on the so-called ***margin hyperplane*** (i.e. has distance $\frac{1}{\|w^*\|}$ to the OSH). Those data points are called Support Vectors. They alone determine the position and orientation of the hyperplane; the influence of the other points is zero. Only the support vectors have $\alpha_i \neq 0$. One could even move the other points around (without crossing the margin hyperplanes) and recalculate the hyperplane and would obtain the identical one with the same points as support vectors. Note that some points might lie on the margin hyperplane but these will not be support vectors because both α_i and $y_i(\langle w^*, x_i\rangle + b^*) - 1$ equal zero.

Optimal Separating Hyperplanes for Linearly Non-separable Data

The strategy of finding the optimal separating hyperplane will fail on most real world data. This can have two possible reasons:

1. In principle, the data would be linearly separable but some noise in the data makes it impossible to find one OSH without errors.

2. The data can not be classified by a hyperplane, as a (more complex) curved hypersurface is necessary (non-linear separable data).

Let us look at the first case first; the second will be discussed in Section 3.2.5 (Non-linear SVMs). The major shortcoming of the OSH is that it does not allow for classification errors. To overcome this problem, the constraints 3.6 must be relaxed. We will introduce the positive slack variables ξ_i $(i = 1, \ldots, n)$ in the constraints to penalize suboptimal and misclassifications:

$$y_i(\langle w, x_i\rangle + b) - 1 + \xi_i \geq 0 \qquad \forall i = 1, \ldots, n. \tag{3.19}$$

The slack variables ξ_i measures the distance of a point that lies on the wrong side of its margin hyperplane. We can differentiate three different cases:

$$\begin{aligned}
\xi_i \geq 1 \quad &\Longleftrightarrow \quad \text{the point lies beyond the OSH and is thus misclassified } (y_i(\langle w, x_i\rangle + b) < 0);\\
0 < \xi_i < 1 \quad &\Longleftrightarrow \quad x_i \text{ is classified correctly, } \textit{but} \text{ lies inside the margin;}\\
\xi_i = 0 \quad &\Longleftrightarrow \quad x_i \text{ is classified correctly, } \textit{and} \text{ lies outside on or outside the margin boundary.}
\end{aligned}$$

Because the value for ξ_i of a single data point expresses its classification error, $\sum_{i=1}^{n} \xi_i$ is an upper bound of the total training error. The optimization problem is now different because we want to maximize the margin and minimize the total training error at the same time. The optimization problem of the linear separable case (3.7) is reformulated accordingly:

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}\|w^*\|^2 + C\sum_{i=1}^{n} \xi_i^k \qquad k \in \mathbb{N} \\ \text{subject to} & 1 - y_i(\langle w^*, x_i\rangle + b^*) - \xi_i \leq 0 \qquad i = 1, \ldots, l. \\ & \xi_i \geq 0 \qquad i = 1, \ldots, n. \end{array} \tag{3.20}$$

The parameter C is the *error weight* which penalizes suboptimal and misclassifications. To find the optimal C, one needs to vary its value across a wide range and determine the classification quality by cross-validation. Again we choose to introduce Lagrange multipliers to reformulate this optimization problem. This is possible as the problem is convex for any positive integer k, and for $k = 1$ and $k = 2$, it is also a quadratic programming problem. Because this approach accepts errors, it is often called the *Soft Margin Generalization* of the OSH as opposed to the *Hard Margin* OSH where no errors were tolerated.

1-Norm Soft Margin

For $k = 1$, we obtain the following Lagrangian for this optimization problem:

$$\mathscr{L}(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i[1 - y_i(\langle w, x_i\rangle + b) - \xi_i] - \sum_{i=1}^{n} \beta_i \xi_i$$

with $\alpha_i \geq 0$ and $\beta_i \geq 0$. The Lagrange multipliers β_i ensure $\xi_i \geq 0$. If we differentiate with respect to w, ξ and β, we obtain:

$$\begin{aligned} \frac{\partial \mathscr{L}(w, b, \xi, \alpha, \beta)}{\partial w} &= w - \textstyle\sum_{i=1}^{n} \alpha_i y_i x_i &= 0 \\ \frac{\partial \mathscr{L}(w, b, \xi, \alpha, \beta)}{\partial b} &= \textstyle\sum_{i=1}^{n} \alpha_i y_i &= 0 \\ \frac{\partial \mathscr{L}(w, b, \xi, \alpha, \beta)}{\partial \xi} &= C - \alpha_i - \beta_i &= 0 \end{aligned}$$

If we substitute these relations back into the objective function, we get:

$$\begin{array}{rl} \text{maximize} & \mathscr{L}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle, \\ \text{subject to} & 0 \leq \alpha_i \leq C, \quad 0 \leq \beta_i \leq C, \quad C - \alpha_i - \beta_i = 0, \\ & \sum_{i=1}^{n} \alpha_i y_i = 0, \quad i = 1, \ldots, n. \end{array} \tag{3.21}$$

Note that the only difference from the corresponding optimization problem of the linear separable case (3.12) at page 27 is that α_i and β_i are

upper-bounded by C. Taking $\beta_i = \alpha_i - C$ into account, we can write the KKT conditions as follows:

$$\begin{aligned} \alpha_i[y_i(\langle w, x_i\rangle + b) - 1 + \xi_i] = 0 \qquad & \forall i. \\ \xi_i(\alpha_i - C) = 0 \qquad & \forall i. \end{aligned}$$

From these two conditions, we can conclude that non-zero (=active) slack variables can only be obtained for $\alpha_i = C$. Points x_i with $\alpha_i = C$ have a distance less than $\frac{1}{\|w\|}$, $\frac{1-\xi_i}{\|w\|}$ to be precise, from the hyperplane. Points with $\xi_i \neq 0$, thus $0 < \alpha_i < C$ lie on one of the two margin hyperplanes. Note that by setting C to infinity we can describe the hard margin with the formulae used for the soft margin.

Now we have described the *1-Norm Soft Margin*, the *2-Norm Soft Margin* for the case $k = 2$ in the optimization problem 3.20 can be solved accordingly and results in an optimization problem which is very similar to 3.21 of the 1-Norm Soft Margin. For details, please refer to Markowetz [2001, p. 43] or Cristianini and Shawe-Taylor [2000, p. 105].

3.2.5 Non-linear SVMs

Many real world data cannot be separated linearly in a reasonable way, not even by using soft margins. In most cases, the process by which the data were generated simply cannot be approximated by a linear function. A loophole is to employ a function Φ, the *feature map*, which maps the data points x_i of the *data space* $\mathcal{L}$ to the *feature space* $\mathcal{H}$ where a linear separation is possible.

$$\begin{aligned} \Phi : \mathbb{R}^d &\rightarrow \mathcal{H} \\ x_i \in \mathcal{L} &\rightarrow \Phi(x_i) \in \mathcal{H} \end{aligned}$$

The "ideal" feature space $\mathcal{H}$, the one that allows for a linear separation of the data, may often have a much higher dimension than the data space $\mathcal{L}$. (We employ the mnemonic $\mathcal{L}$ and $\mathcal{H}$ to remember the *l*ow and *h*igh dimensionalities.) The feature space $\mathcal{H}$ must be a Hilbert space, which is a vector space in which a dot product (scalar product) is defined and has notions of distance and of angle (for a definition, see Levitan [2002]). Chris Burges wrote the following on Hilbert space in his tutorial on SVMs:

> "The literature on SVMs usually refers to the space $\mathcal{H}$ as a Hilbert space .... You can think of a Hilbert space as a generalization of Euclidean space that behaves in a gentlemanly fashion. Specifically, it is any linear space, with an inner product defined, which is also complete with respect to the corresponding norm (that is, any Cauchy sequence of points converges to a point in the space). Some authors (e.g. Kolmogorov and Fomin [1970]) also require that it be separable (that is, it must have a countable subset whose closure is the space itself), and some (e.g. Halmos [1967]) don't. It is a generalization mainly because its inner product can be any inner product, not just the scalar ("dot")

> product used here (and in Euclidean spaces in general). It is interesting that the older mathematical literature (e.g. Kolmogorov and Fomin [1970]) also required that Hilbert spaces be infinite-dimensional, and that mathematicians are quite happy defining infinite-dimensional Euclidean spaces. Research on Hilbert spaces centers on operators in those spaces, since the basic properties have long since been worked out. Since some people understandably blanch at the mention of Hilbert spaces, I decided to use the term Euclidean throughout this tutorial" [Burges, 1998].

We used the optimization problem 3.12 to obtain the coefficients α_i^*, which we will use to obtain the w^* and b^* of the OSH with the maximal margin (3.15 and 3.16), and the decision function 3.18. These contain the training points x_i only in the form of dot products. This means that we will need to compute dot products like $\langle \Phi(x_p), \Phi(x_q) \rangle$. This can be computationally difficult or impossible if the dimension of $\mathcal{H}$ becomes too large (or infinite) as the quadratic programs (to determine the αs) become complex.

We now consider a simple example.

Example: Given a training set $\mathcal{X} = (x_i, y_i)$ of points in $\mathbb{R}^2$ with labels $+1$ of -1, set $\mathcal{X} = \left\{ \binom{-1}{1}, +1 \right), \left(\binom{0}{1}, -1 \right), \left(\binom{1}{1}, +1 \right) \}$. As can be seen from Fig. 3.4, the three points cannot be separated by a hyperplane (i.e. a straight line) in $\mathbb{R}^2$. We apply now the mapping

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3, \quad x_i = (x_{i\mathbf{1}}, x_{i\mathbf{2}})^t \mapsto (x_{i\mathbf{1}}^2, \ \sqrt{2} x_{i\mathbf{1}} x_{i\mathbf{2}}, \ x_{i\mathbf{2}}^2)^t.$$

(Note that the bold subscripts refer to vector components). Fig. 3.4 shows the entire mapping of data in $\mathcal{L}$ defined on the square $[-1, 1] \times [-1, 1] \in \mathbb{R}^2$. This figure also helps us to better imagine what this mapping Φ actually does: The image of Φ may live in a space of very high dimension, but it is just a (possibly very distorted) surface whose *intrinsic dimension* is just that of $\mathcal{L}$ ("intrinsic dimension" means the number of parameters required to specify a point on the manifold [Burges, 1998]).

The mapping of the three training points will yield $((1, -\sqrt{2}, 1)^t, 1)$, $((0, 0, 1)^t, -1)$ and $((1, \sqrt{2}, 1)^t, 1)$, which are marked in Figure 3.4 as red dots for the positive points and blue squares for the negative point.

The Lagrange multipliers can be determined as $\alpha = (1, 2, 1)$, from which we can derive $w^* = \sum_{i=1}^{3} [\alpha_i y_i \Phi(x_i)] = (2, 0, 0)^t$ and the unit vector $w = (1, 0, 0)^t$. With Equation 3.16, we obtain $b = -0.5$. The maximal margin hyperplane in the feature space is thus a plane parallel to the $x_{\mathbf{2}} \times x_{\mathbf{3}}$ plane at distance 0.5 in positive $x_{\mathbf{1}}$ direction (plane is not shown in Fig. 3.4 but its position is indicated by the dashed intersection lines). We realize that the learning task can be solved very easily in $\mathbb{R}^3$.

But how does the corresponding decision surface look like in $\mathbb{R}^2$? There is no direct way to obtain the decision function in $\mathcal{L}$ from the one in $\mathcal{H}$. Even if we succeed in determining an inverse mapping function Φ', the hyperplane in $\mathcal{H}$, like in our example, might contain points with no correspondence in $\mathcal{L}$. Instead of that, we presume that a given arbitrary point x from $\mathcal{L}$ that lies on the decision boundary in $\mathcal{L}$ also has distance 0 from the decision boundary

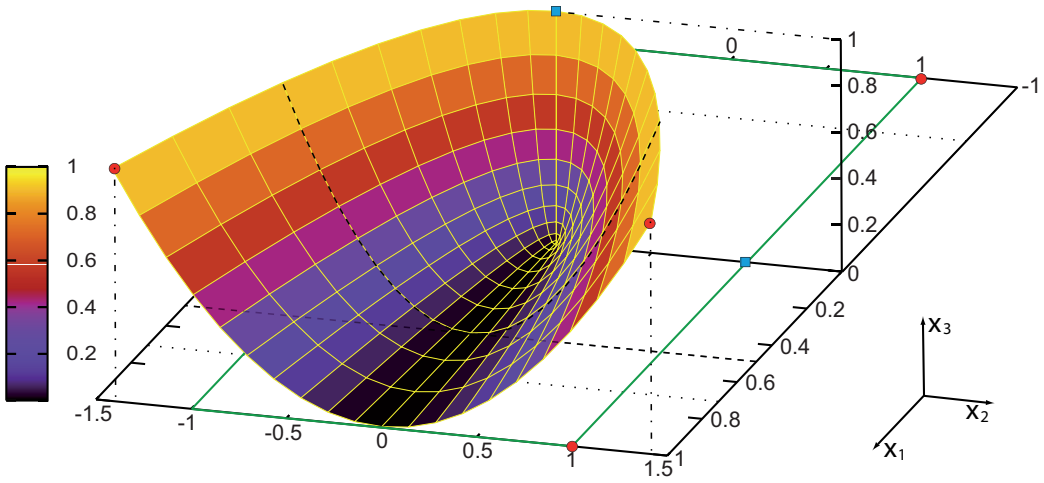

Figure 3.4: Graphical representation of the points enclosed in the green bordered square $[-1,1] \times [-1,1]$ in $\mathbb{R}^2$ and their mapping $\Phi : (x_1, x_2)^t \mapsto (x_1^2, \sqrt{2}x_1x_2, x_2^2)^t$ as a colored surface in $\mathbb{R}^3$. An example of three points is illustrated: Two red dots $\binom{1}{1}$ and $\binom{-1}{1}$ and one blue square $\binom{0}{1}$ which are inseparable in $\mathbb{R}^2$. When mapped to the feature space $\mathcal{H}$ in $\mathbb{R}^3$, the three points become separable by a hyperplane whose position is indicated by the dashed intersection lines. The dotted lines indicate the decision boundaries in $\mathbb{R}^2$ that can be derived from the optimal separating hyperplane in $\mathbb{R}^3$ (for more details refer to the text).

in $\mathcal{H}$ when mapped to the feature space $\mathcal{H}$.

$$\begin{aligned} \langle w, \Phi(x) \rangle + b &= 0 \\ \sum_{i=1}^{n} (\alpha_i y_i \langle x_i, \Phi(x) \rangle) + b &= 0 \end{aligned}$$

For the example we obtain:

$$\begin{aligned} & \sum_{i=1}^{3} (\alpha_i y_i \langle x_i, \Phi(x) \rangle) - 0.5 = 0 \\ \Leftrightarrow \quad & \frac{1}{2}(x_{\mathbf{1}}^2 - 2x_{\mathbf{1}}x_{\mathbf{2}} + x_{\mathbf{2}}^2) - x_{\mathbf{2}}^2 + \frac{1}{2}(x_{\mathbf{1}}^2 + 2x_{\mathbf{1}}x_{\mathbf{2}} + x_{\mathbf{2}}^2) = 0 \\ \Leftrightarrow \quad & x_{\mathbf{1}}^2 = \frac{1}{2} \\ \Leftrightarrow \quad & x_{\mathbf{1}} = \pm\frac{1}{2}\sqrt{2} \end{aligned}$$

As we can see, the hyperplane in $\mathbb{R}^3$ corresponds to two separating straight lines in $\mathbb{R}^2$, indicated as dotted lines in Fig. 3.4.

The Kernel Trick

As we can see in the last example and as we stated right before it, we always encounter the mapping function in dot products like $\langle \Phi(x_p), \Phi(x_q) \rangle$. Depending on the chosen Φ, $\mathcal{H}$ might possibly be high- or even infinite-dimensional,

so working with Φ directly and calculating dot products of mapped points x_i will be difficult. An important observation to overcome this problem will be demonstrated in the following:
Consider the two points $p = (p_1, p_2)$ and $q = (q_1, q_2)$ and apply the mapping Φ on p and q: $(p_1, p_2) \mapsto (p_1^2, \sqrt{2}p_1p_2, p_2^2)$ and calculate the dot product thereon:

$$\begin{aligned}\langle \Phi(p), \Phi(q) \rangle &= (p_1^2, \sqrt{2}p_1p_2, p_2^2)(q_1^2, \sqrt{2}q_1q_2, q_2^2)^t \\ &= p_1^2q_1^2 \ + \ 2 \ p_1q_1 \ p_2q_2 \ + \ p_2^2q_2^2 \\ &= (p_1q_1 \ + p_2q_2)^2 \\ &= \langle p, q \rangle^2 \ =: \ k(p, q)\end{aligned}$$

We see in the end, that instead of calculating this particular mapping followed by the dot product, we can equivalently calculate the dot product of the original points and square it. Thus we can calculate the dot product $\langle \Phi(p), \Phi(q) \rangle$ without applying the function Φ. Such functions that are equivalent to mapping followed by a dot product in $\mathcal{H}$ are called *kernel functions*, typically denoted by k.

So why is the usage of a kernel function a trick?
Vert et al. [2004] proposes: "Any algorithm for vectorial data that can be expressed only in terms of dot products between vectors can be performed implicitly in the feature space associated with any kernel by replacing each dot product by a kernel evaluation." This means that we do not need to know what the feature space $\mathcal{H}$ actually looks like; we only need the kernel function, which returns us a measure of similarity. As we still could do all computations directly in $\mathcal{H}$, we always keep the possibility of a geometric interpretation of SVMs – in our case – by the optimal separating hyperplane. Thus SVMs are more transparent than e.g. artificial neural networks.

Data Representation – Relevance of Kernel Functions
Kernel functions help us to simplify the representation of data to be analyzed with data analysis methods. Most data analysis methods that do not use kernels, require that the data are somehow preprocessed and presented to the algorithm in a processable format. This preprocessing step can be understood as a mapping of the data; consequently, SVMs in a feature space share this general characteristic.

For example, a data set $\mathcal{S}$ of, say, three particular oligonucleotides could be represented in various ways, e.g. by the biomolecules' molecular mass, pK_i, melting temperature, GC-content, predicted secondary structure, frequency in genomes etc. But, let us assume that we choose a representation by a set $\Phi(\mathcal{S})$ containing the three oligonucleotides represented by sequences of letters that stand for the succession of their bases. This representation can then be processed by an algorithm that, for example, compares the sequences' pairwise similarity. (This example is discussed by Vert et al. [2004]).

Kernel methods solve the problem of data representation differently. Instead of representing each data point individually, the data are compared pairwise and their set of pairwise similarities is represented. This means that a real-valued "comparison function" $k : \mathcal{L} \times \mathcal{L} \to \mathbb{R}$ is employed and

the resulting $n \times n$ matrix K of pairwise comparisons $K_{i,j} = K(x_i, x_j)$ will represent the data set $\mathcal{S}$.

Significance The development of kernel functions k that represent data (by comparing them pairwise and returning a square matrix) and the development of algorithms (= kernel methods) that process such data representations can be pursued completely independently. This is a great advantage because the same algorithms can be applied on image data, molecules or sequences, once they have been processed with adequate kernel functions that all yield real-valued square matices. Additionally, the complexity of an algorithm that processes kernel matrices of $n \times n$ objects will always have to process only n^2 values independent of the number of measured values associated with each object. Moreover, the comparison of two objects and calculation of the corresponding square matrix often is an easier task than finding an explicit representation for each object x as a vector $\Phi(x) \in \mathbb{R}^p$. For example, there is no obvious way to represent nucleic acid sequences as vectors in a biologically relevant way; however, meaningful and well-established pairwise sequence comparison methods exist.

General Definition of Kernel Functions
What are the criteria that a function must fulfill to be a kernel function k?

Definition 3.3 A function $k : \mathcal{L} \times \mathcal{L} \to \mathbb{R}$ is called a positive (semi-)definite kernel if and only if it is

1. symmetric, that is, $k(x, x') = k(x', x)$ for any two objects $x, x' \in \mathcal{L}$, and

2. positive (semi-)definite, that is,

$$c^T K c = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$$

with matrix K of all elements $k(x_i, x_j)$, for any $n > 0$, any choice of n objects $x_1, \ldots, x_n \in \mathcal{L}$, any choice of vectors $c \in \mathbb{R}^d$, and any choice of numbers $c_1, \ldots, c_n \in \mathbb{R}$ respectively [Vert et al., 2004].

An "ideal" kernel function assigns a higher similarity score to any pair of objects that belong to the same class than it does to any pair of objects from different classes. This is the case if the implicit mapping by the kernel function brings similar objects close together and takes dissimilar objects apart from each other in the induced feature space.

Examples of frequently used kernels
Frequently used kernel functions are [Vapnik, 1995; Müller et al., 2001;

Schölkopf and Smola, 2002]:

$$\begin{aligned}
\text{Linear kernel} \quad & k(x_i, x_j) = \langle x_i, x_j \rangle \\
\text{Radial basis function (RBF) kernel} \quad & k(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma_0^2}\right) \\
\text{Polynomial kernel} \quad & k(x_i, x_j) = (s\langle x_i, x_j \rangle + c)^d \\
\text{Sigmoid kernel} \quad & k(x_i, x_j) = \tanh(s\langle x_i, x_j \rangle + c) \\
\text{Convex combinations of kernels} \quad & k(x_i, x_j) = \lambda_1 k_1(x_i, x_j) + \lambda_2 k_2(x_i, x_j) \\
\text{Normalization kernel} \quad & k(x_i, x_j) = \frac{k'(x_i, x_j)}{\sqrt{k'(x_i, x_i) k'(x_j, x_j)}}
\end{aligned}$$

where s, c, d and λ_i are kernel-specific parameters, $\sigma_0^2 = \text{mean}\|x_i - x_j\|^2$.

3.2.6 Transductive SVMs and their Relevance to Biological Datasets

SVMs are playing an increasingly important role in the field of computational biology. For an in-depth overview of the current research and applications to computational biology, see Schölkopf et al. [2003, 2004].

The classical SVMs presented in the preceding sections are "inductive" SVMs. There, the training data that are used to build the model should ideally cover the whole problem space; the model is then used to predict the labeling of new data points. In most biological datasets, the number of labeled data points is rather small, but a large number of unlabeled data points (e.g. unannotated proteins) is available. To take advantage of these additional unlabeled data, the so-called "transductive" SVMs (TSVMs) have been developed [Vapnik, 1998; Joachims, 1999a]. To address the problem of learning with unlabeled data (often called "semi-supervised" or "transductive learning problem"), TSVMs assume that the missing labels of the unlabeled data points are consistent with their positions in the hyperspace in two aspects: (i) nearby points and (ii) points on the same structure (typically referred to as a cluster or a manifold) are likely to share the same label [Zhou et al., 2004]; see Figure 3.5.

3.2.7 Performance Estimates of Learning Algorithms

To assess the accuracy of any classifying algorithm there are several statistics on the number of true positive, false positive, true negative and false negative predictions (TP, FP, TN, FN); see Baldi et al. [2000] for a review:

$$\text{Error rate} = err = (FP + FN)/(FP + FN + TP + TN)$$

$$\text{Recall} = \text{sensitivity } S_n = TP/(TP + FN)$$

$$\text{Precision} = \text{specificity } S_p = TP/(TP + FP)$$

$$\text{Matthews correlation coefficient } MCC = \sqrt{\frac{TP \cdot TN - FN \cdot FP}{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}.$$

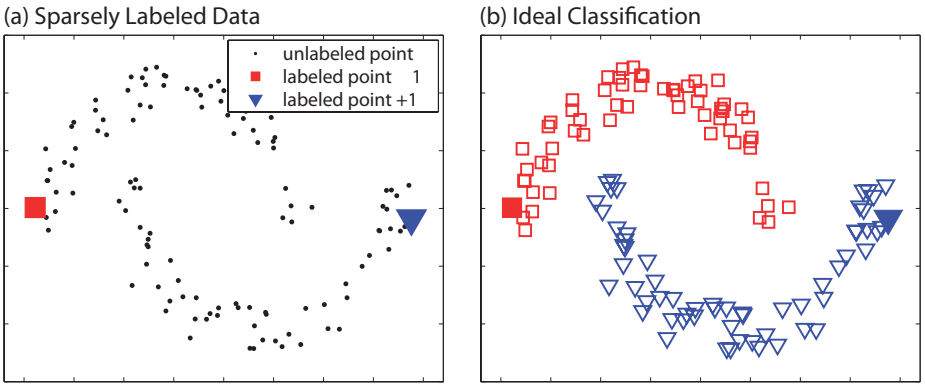

Figure 3.5: Illustration of the usefulness of unlabeled data points (e.g. representing unannotated protein sequences): If only the labeled data were used for model training, the separating hyperplane would just be a straight vertical line, separating the two labeled crescents poorly. But if one also takes the consistency of labeled and unlabeled data into account – as do transductive SVMs – i.e. that nearby points are likely to have the same label as points on the same structure (here two intertwining crescents), then the classification can be greatly enhanced. (Example taken from Zhou et al. [2004], a redrawing from Szummer and Jaakkola [2002]).

More precisely, the error rate gives the proportion of data points that are classified incorrectly, the recall gives the proportion of truly positive data points that are contained in the predicted positives, the precision specifies the proportion of TP in all data points predicted as positive. Matthews' correlation coefficient (MCC) uses all four numbers (*TP, TN, FP and FN*), is symmetric with respect to FP and FN, and may often provide a much more balanced evaluation of the prediction than the statistics given above [Baldi et al., 2000].
Two similar tests are widely used for determinating the above parameters: leave-one-out *(LOO)* tests and x-fold cross-validations (with x typically 3, 5, or 10). In a LOO test, the predictive model is trained on a dataset that has been reduced by one data point. The generated model is then used to give a prediction for the removed data point. The whole procedure is repeated for each single data point of the set. In a, say, 5-fold cross-validation, the dataset is divided randomly into five parts; one fifth of the dataset is removed; the model is trained on the rest, the so-called *training data*; and the prediction is made for the fifth, the so-called *test data*; and the procedure is repeated for every remaining fifth of the dataset.

3.2.8 Summary of Support Vector Machines

Let us briefly summarize the essential ideas of SVMs:
Assume that we are given a series of examples (e.g. measurement readings, protein sequences etc.), each associated with a number d of features (either numerical or binary values), then we can treat each example simply as a d-dimensional vector in a d-dimensional space $\mathcal{L}$. If we want to construct a binary classification of the examples, i.e. label each d-dimensional data point as "positive" or "negative", a simple and intuitive way would be to construct

a separating (hyper)plane which separates the positive and negative data points. If the data are linearly separable, a separating hyperplane can be found which leaves a maximal margin (a "clearance") between the two classes of data. The data points that are closest to the hyperplane are called support vectors. Once we have determined these points that "support the plane", we can write down a decision function that will assign a label to any new data point (+ or −).

If the data are not linearly separable then the kernel "trick" is used: Let us assume we first mapped the data to some other (possibly higher dimensional) space $\mathcal{H}$, using a mapping Φ. We can then determine the separating hyperplane in the hyperspace $\mathcal{H}$. We should always try to use a mapping function Φ that puts similar data points close to each other and allows for a linear separation in the hyperspace. Solving the equations for the optimal separating hyperplane in the hyperspace, we observe that all formulae only depend on the data through dot products in $\mathcal{H}$, i.e. on functions of the form $\Phi(x_i) \cdot \Phi(x_j)$. This encourages us to introduce a *kernel function* k, such that $k(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$, and use it as a similarity measure for x_i and x_j without explicitely knowing Φ nor the dimension of $\mathcal{H}$. As an overview, Fig. 3.6 depicts in a nutshell how SVMs with kernel functions work.

A Support Vector Machine is:

1. a hyperplane with a maximal separating margin in the feature space,

2. constructed in the input space via a kernel function,

3. used as a decision function to classify data into two classes.

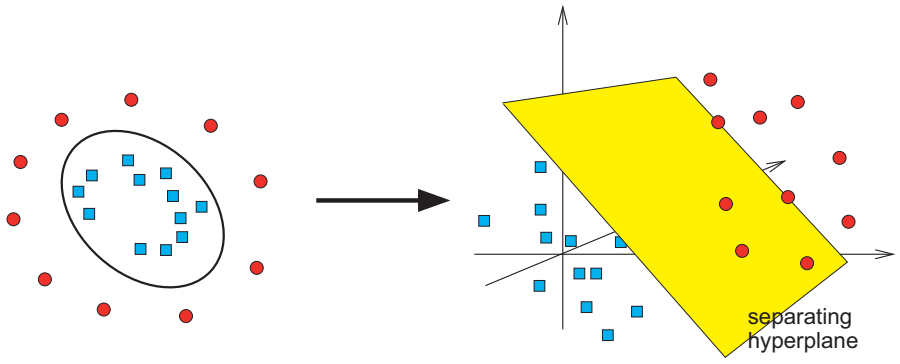

Figure 3.6: A non-linear separation of vectors (data points) in the *input space* (original space) is mediated by a kernel function. This is *equivalent* on a one-to-one basis to a linear separation of the vectors in the *feature space* using the dot product. Image source: Markowetz [2003].

3.3 Sequence Analysis and Comparison

3.3.1 BLAST

The Basic Local Alignment Search Tool (BLAST, Altschul et al. [1997]) is an algorithm for searching protein or nucleic acid sequences in a database that are similar to a given *query* sequence. BLAST is probably the most widely used program in bioinformatics. The most typical scenario where BLAST can be used is in the study of an unknown protein or gene sequence, where it is often very helpful to be able to find the most similar sequences. If the biological function of the best *hit* (the most similar sequence) is known, then it can often be informative for the protein/gene of unknown function. However it is necessary that the region of high similarity spans almost the entire length of the *target* and query sequences. For example, proteins that share only one or a few homologous domains might be involved in different processes and perform different functions. Nevertheless, if the user is aware of that, the BLAST results can also be helpful in this case. The major BLAST implementations are available from the NCBI [National Center for Biotechnology Information, McGinnis and Madden, 2004] and from Washington University in St. Louis [Gish, 2006] and include the following dedicated programs:

- `blastn` accepts a DNA query and searches it against a DNA database (specified by the user).
- `blastp` accepts a protein query to be compared against a selectable protein database.
- Position-specific iterated BLAST (PSI-BLAST) is a more recent BLAST version; it is used for detecting distant relatives of a protein. In the first phase, a *profile* (a PSSM, see next section) is derived from the alignment of the best hits in a normal `blastp` search. The protein database is then queried using this profile. Normally, the profile will be more general than the initial single query sequence and thus will return more matching proteins. A new profile is then created from the hits found and the process is repeated until convergence, when no more proteins are included at each iteration. More information about profiles can be found in the next section.
- `blastx` accepts a DNA query, translates it into all six reading frames and searches the corresponding six amino acid sequences against the protein database.
- `tblastn` searches a protein query against the DNA database which has been previously translated into all six reading frames.
- `tblastx` is suited for a DNA query and a DNA database but translates the query and the database into all six reading frames.

The translated BLAST versions are especially useful if one suspects a protein coding sequence in the query and/or target database sequences, knowing that

not all protein coding sequences have been correctly predicted and may be absent in the protein databases.

Like many search algorithms, the BLAST algorithm also employs the *seed-and-extend paradigm* to speed up searches. Given a query sequence of length L, BLAST extracts all contained words of length w (there are $L-w+1$ such words in the query). For amino-acid sequences, the default is $w = 3$ (shown in Fig. 3.7) and 11 for nucleotide sequences. The list of words is then expanded by all variations of each word (so-called neighborhood words) that have a score greater than a threshold T using a scoring matrix such as PAM250 [Dayhoff et al., 1978] or BLOSUM62 [Henikoff and Henikoff, 1992]. For typical parameter values, this results in about 50 words per residue of the query sequence [Pertsemlidis and Fondon, 2001] (part (a) in Fig. 3.7). Then, the high-scoring word list is compared to the sequence database and exact matches are identified (part (b) in Fig. 3.7). Finally, for each word match, the alignment is extended in both directions to generate alignments that score higher than the score threshold S (part (c) in Fig. 3.7). Besides the original publications on BLAST by Altschul et al. (1990; 1997), the tutorial by Pertsemlidis and Fondon [2001] on the principles, workings, applications and potential pitfalls of BLAST, and the book by Korf et al. [2003] are very recommended reading.

3.3.2 Detecting and Searching for Motifs in Protein Sequences

Position Specific Scoring Matrices (PSSMs)

A position specific scoring matrice (PSSM), also called a position weight matrix (PWM) or a profile, is a commonly used representation of motifs (patterns) in biological sequences. The concept of a sequence motif was first introduced by Doolittle [1981], and Gribskov et al. [1987] introduced the "profile", the first description of a PSSM. Here we present PSSMs as they are revised by Durbin et al. [1998]:

For protein sequences, a PSSM W is usually a 20 x l matrix, where l is the length of the motif and the rows correspond to the twenty proteinogenic amino acids. W can be generated from a set of aligned sequences without gaps. A matrix entry w_{kj} of W is calculated as the log-odds ratio of the observed and expected frequencies of residue a_k at position j in the motif $w_{kj} = \log \frac{\text{obs}_j(a_k)}{\exp(a_k)}$ where $\text{obs}_j(a_k)$ is the frequency of residue k at position j in the alignment and $\exp(a_k)$ is set to $\frac{1}{20}$ for all residues. This approach causes difficulties as soon as some observed frequency is set to zero, because this results in a matrix entry of minus infinity. Because of this, the PSSM is over-fitted to the training data that was used to build the PSSM, since whenever a certain residue is observed at a position where it did not occur in the training dataset, the over-fitted PSSM returns minus infinity and therefore makes it impossible to consider the subsequence an instance of the motif. This situation can be avoided by introducing pseudo-counts which are added to all observed frequencies. The simplest approach to this is the Laplace rule, according to which the value 1 is added to all counts to avoid zero entries

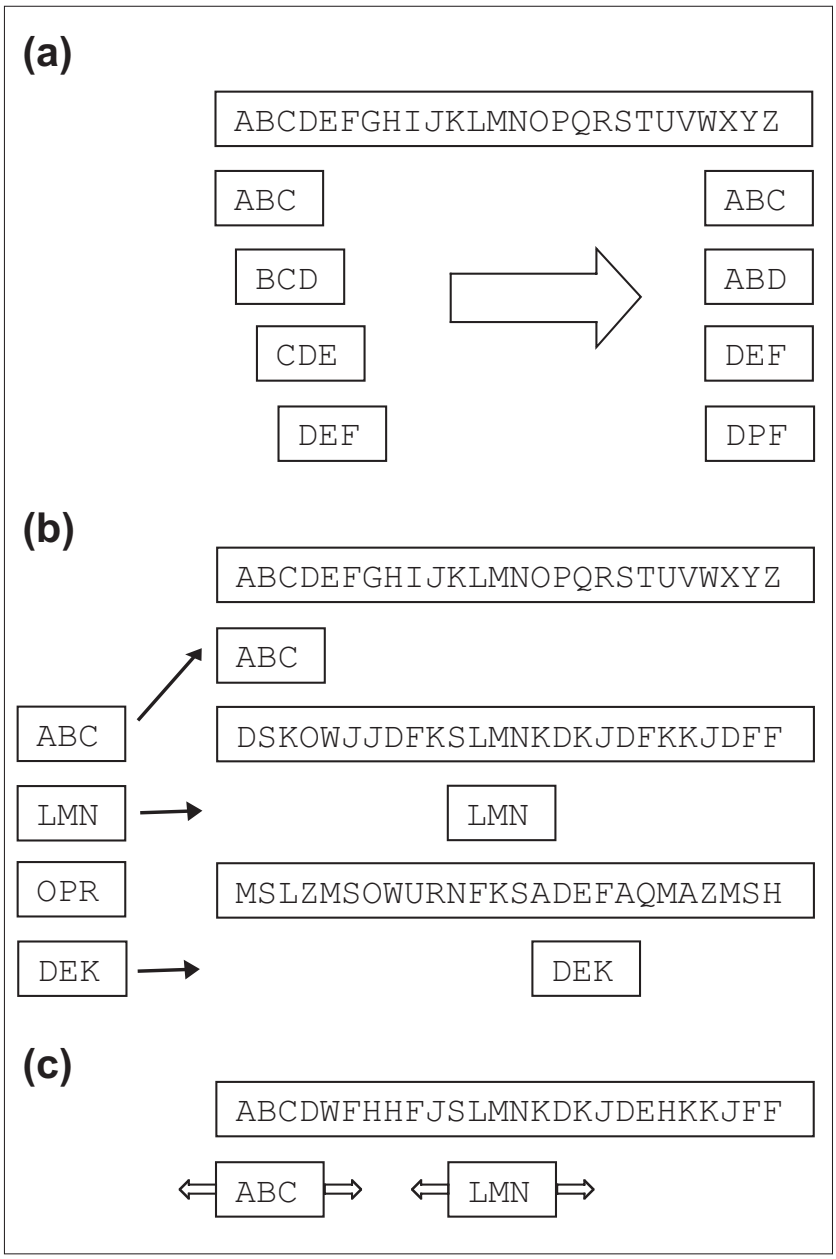

Figure 3.7: Illustration of the BLAST algorithm. (a) Given a query sequence of length L, BLAST derives a list of contained words of length w, where $w = 3$ for amino acid sequences (shown) and 11 for nucleotide sequences. This word list is then expanded to include all high-scoring matching words, keeping only those that score more than the neighborhood word score threshold T when scored using a scoring matrix such as PAM250 or BLOSUM62. For typical parameter values, this results in about 50 words per residue of the query sequence. (b) The high-scoring word list is compared to the sequence database and exact matches are identified. (c) For each word match, the alignment is extended in both directions to generate alignments that score higher than the score threshold S. Image source: Pertsemlidis and Fondon [2001].

[Durbin et al., 1998]. The motif can then be searched in a given sequence by evaluating the score $S = \sum_{j=0}^{l-1} w_{kj}$ starting from each position in the sequence between 0 and $N - l$, where N is the length of the sequence in which the motif is looked for.

Profile Hidden Markov Models for Sequence Families

A common strategy for the identification of a specific type of domain is to use profile Hidden Markov Models (pHMMs), which are statistical models extracted from multiple sequence alignments. In contrast to simple sequence motifs of fixed length, i.e. PSSMs (described in the previous paragraph), pHMMs are suited for identifying motifs that are interrupted by segments of variable length, and are used to characterize position-specific sequence similarities within a family of proteins. A collection of pHMMs for a wide array of domains and domain families is availabe from the database Pfam [Bateman et al., 2004] and TIGRFAMs [Haft et al., 2003]. The pHMM implementation HMMER [Durbin et al., 1998, `hmmer.janelia.org`] and self-written Perl and BioPerl scripts [`www.perl.org`; Stajich et al., 2002] were used to search for NRPS in the genome sequences and biosynthesis clusters, and to extract single domains from a given protein sequence.

Besides HMMER, SAM [Karplus et al., 1998; Karchin and Hughey, 1998] and Meta-MEME [Grundy et al., 1997] are also popular implementations of profile HMMs.

A concise introduction and tutorials to pHMMs can be found in the book by Durbin et al. [1998], on the SAM website [`www.soe.ucsc.edu/research/compbio/sam.html`] or on the author's professional web site linked at `en.wikipedia.org/wiki/Christian_Rausch`.

3.4 Phylogenetic Reconstruction

Phylogeny (or phylogenesis) is the origin and evolution of a set of taxa, usually a set of species. Phylogenetics is the science that has the goal of reconstructing the phylogeny of the species etc. under study. Phylogenetic approaches are applied in Chapter 5 where the evolutionary relationship of the Condensation domains of NRPS is reconstructed.

There are two principally different approaches for phylogenetic reconstruction, distance and character based reconstruction:

- *Distance based methods* first compute all pairwise distances for a given set of biological data (e.g. molecular sequences) and then compute a tree that represents these distances as closely as possible.

- Unlike distance based methods, which "condense" all sequence information into a single number (the pairwise distance), the *character based methods* attempt to infer the phylogeny based on all the individual characters (nucleotides, amino acids, or 1/0 for the presence or absence of phenotypic characteristics).

Both types of methods require an alignment of the sequences for which the phylogeny is to be reconstructed so that the corresponding positions are standing in the same column. The two main approaches that belong to the character based methods are *maximum parsimony* and *maximum likelihood*. In *maximum parsimony* based methods, we search for the best tree topology that minimizes the number of substitutions needed to explain the sites (alignment positions) of considered sequences. *Maximum likelihood* methods are probabilistic methods of inference. They use explicit models of molecular evolution and allow for rigorous statistical inference. However, like maximum parsimony methods, they are very computer intensive.

3.4.1 Character Based Methods

Maximum Parsimony

The maximum parsimony method in phylogenetic reconstruction follows a principle that is often applied in scientific model building: The simplest model that can explain all observations is mostly better than more complex models, as long as no further findings require the formulation of a new model. The most parsimonious tree is hence the one that explains the evolution of a set of aligned sequences by a minimal number of character changes (mutation events) in the sequences, originating from a common ancestor with mutations at each branching point of the tree. To determine the parsimony of a given tree, the different aligned sequences are assigned to the leaves of the tree. Then the so-called *Small Parsimony Problem* is to find how to label the interior nodes of a given tree and to calculate a scoring of the whole tree which will allow to compare the parsimony of different trees. Whereas the Small Parsimony Problem can be solved efficiently using the *Fitch algorithm* [Fitch, 1971], the problem of actually finding the tree with the lowest parsimony score for a given alignment of sequences is NP-hard. To solve this problem, the so-called *Large Parsimony Problem*, one would need to consider all $(2n-5)!! = 3 \cdot 5 \cdot 7 \cdot \ldots \cdot (2n-5)$ possible (unrooted) trees that all represent different relationships between n species (taxa). (Note that the double factorial grows very rapidly, e.g. for $n = 10$ taxa, $\approx 2 \cdot 10^6$ exist and for $n = 15$ taxa, $\approx 7.9 \cdot 10^{12}$ unrooted trees exist). The techniques applied to solve the Large Parsimony Problem are thus mostly heuristic algorithms which are not guaranteed to find the optimal solution. However, the *branch and bound* strategy can be applied to reduce computational costs for the calculation of the optimal solution. The idea behind branch and bound is to avoid calculating all possible trees. The tree is built up stepwise, successively adding leaf edges to the tree while each leaf represents a sequence. Keeping in mind that a new tree with one additional leaf edge can never have a lower parsimony score, and that one can agree on a maximal *global* parsimony score and ignore trees that one would obtain by adding more leaf edges. The maximal global parsimony score can be obtained from a tree calculated using a heuristic algorithm. For details on the application of branch and bound to parsimony, refer to Hendy and Penny [1982]. For an introduction to phylogenetic reconstruction using parsimony, see the lecture notes by Huson

[2007] and for more details see the book by Felsenstein [2004].

Maximum Likelihood Estimation (MLE)

Like maximum parsimony, MLE requires that the sequences $a_1, \ldots, a_n$ for which the phylogeny has to be reconstructed are given as a multiple alignment A. Moreover, a model of evolution M that assigns a certain probability to each possible nucleotide (or amino acid) substitution, has to be chosen. The goal of MLE methods in phylogenetic reconstruction is then to find a phylogenetic tree T with edge lengths ω that maximize the likelihood $\mathbb{P}(A \mid T, M)$ of generating the sequences $a_1, \ldots, a_n$ at the leaves of T. The chosen model of evolution is essential to obtain the correct tree. A simplistic model for protein sequences would be the Poisson model, which considers all changes between amino acids to occur at the same rate. However, programs implementing phylogenetic reconstruction using MLE use more advanced models, like an instantaneous rate matrix that is derived from an updated Dayhoff empirical substitution matrix [Dayhoff et al., 1978], called the JTT model of evolution [Jones et al., 1992]. As is the case for the maximum parsimony score of a single given tree (the Small Parsimony Problem), the likelihood of being the optimum tree can also be computed efficiently using Felsenstein's recursive algorithm (1973) for MLE. As with maximum parsimony, to find the optimal MLE tree, the entire tree space would have to be searched, which is NP-hard. Again, branch and bound techniques can be used to find an exact solution with a lowered computational burden. However, for larger taxa sets ($n > 20$) heuristic search techniques have to be used. For a digest on models of evolution and MLE, see Huson [2007] or for more details, consult Durbin et al. [1998, Chap. 8] or Felsenstein [2004].

3.4.2 Distance Based Methods

As stated above, the first step of distance based methods is to calculate all pairwise distances which are stored in the distance matrix. From the multiple sequence alignment of all sequences, the *normalized Hamming distance* may be calculated for each sequence pair. This simply counts all mismatching positions over all aligned positions (ignoring gap-only positions of pairs of sequences from the multiple alignment). To account for conservative substitutions in protein sequences, it is possible to count substitutions by amino acids with similar physico-chemical properties or amino acids with a positive score according to a substitution matrix (for example BLOSUM62 constructed by Henikoff and Henikoff [1992], the *de facto* standard for many protein alignment programs [Eddy, 2004]). Unless one is studying very closely related sequences, the Hamming distances need to be corrected using, for example, the Jukes-Cantor distance transformation [Jukes and Cantor, 1969] to account for several mutations that may have occurred at one site but would be underestimated by the Hamming distance alone. More sophisticated models of evolution than the Jukes-Cantor model exist. For example, more advanced models of DNA evolution differentiate between transitions (purines can be substituted by purines, and pyrimidines by pyrimidines) and transversions

(a purine is substituted by a pyrimidine or *vice versa*). Popular algorithms that reconstruct a tree from a given distance matrix are UPGMA [Sokal and Michener, 1958] and Neighbor-Joining [Saitou and Nei, 1987]. For details on distance based tree reconstruction methods, refer to Durbin et al. [1998, Chap. 7] and/or Felsenstein [2004, Chap. 11]. As well, Huson [2007] provides a concise presentation of the topic and its most important algorithms.

3.4.3 Modeling the Rate of Evolution at Different Sites

In some cases, the rate of amino acid substitution may be assumed to be the same for all positions in the alignment. In general, however, this does not reflect reality, since the substitution rate is usually higher at positions of lower functional importance. A more realistic model is achieved if the substitution rate is taken to vary among sites according to the gamma distribution [Gu and Zhang, 1997].

The variation of substitution rate r among sites can be described as $f(r) = \frac{\beta^\alpha}{\Gamma(\alpha)} \mathrm{e}^{-\beta} r^{\alpha-1}$ where the gamma function $\Gamma(\alpha)$ is defined by $\Gamma(\alpha) = \int_0^\infty \mathrm{e}^{-t} t^{\alpha-1} \mathrm{d}t$. In the gamma distribution, β is a scaling factor. The shape of the distribution is determined by the gamma parameter α. The larger α is, the weaker the rate variation: for $\alpha = \infty$, the rate is constant for all sites; for $\alpha > 20$, the distribution is bell-shaped, with most sites having intermediate rates and few showing very high or low rates; for $\alpha = 1$, the rate follows the exponential distribution, indicating that it varies extensively from site to site. If $\alpha < 1$, the rate distribution is skewed to the right, which implies that most variation comes from a few positions, while the other sites are practically invariant because they show a substitution rate close to zero [Nai and Kumar, 2000]. An appropriate α has to be estimated for each dataset.

Many methods of phylogenetic reconstruction offer an estimation of parameter α which determines the shape of the Γ distribution as an option. Typically, four gamma-rate categories can be chosen to approximate the distribution.

3.4.4 Assessing Tree Topologies With Bootstrapping

Bootstrapping is a resampling technique where data points, columns of the alignment in our case, are drawn from the dataset with replacement to form a new dataset of the same size. As a prerequisite, the columns of the alignment are assumed to evolve independently [Felsenstein, 2004]. This resampling is done a pre-defined number of times and a phylogenetic reconstruction method is applied to these multiple datasets.

In general, a topology is taken as reliable if tree reconstruction results in the same topology for at least 95% of the datasets generated by bootstrapping. This is a quite strict approach and it has been shown that subclades of a tree may be accepted as being significant if they are supported by only 70% of the trees [Hillis and Bull, 1993].

Chapter 4

Specificity Prediction of Adenylation Domains Using Transductive SVMs

4.1 Overview

In this chapter, we present a new support vector machine (SVM)-based approach to predict the substrate specificity of subtypes of a given protein sequence family (published 2005 [Rausch et al.]). We demonstrate the usefulness of this method on the example of aryl acid-activating and amino acid-activating Adenylation domains (A domains) of nonribosomal peptide synthetases (NRPS). The residues of gramicidin synthetase A that are 8 Å around the substrate amino acid and the corresponding positions of other Adenylation domain sequences with 397 known and unknown specificities were extracted and used to encode this *physico-chemical fingerprint* into normalized real-valued feature vectors based on the physico-chemical properties of the amino acids. The SVM software package SVM^{light} was used for training and classification, with transductive SVMs to take advantage of the information inherent in unlabeled data. Specificities for very similar substrates that frequently show cross-specificities were pooled to the so-called *composite specificities*, and predictive models were built for them. The reliability of the models was confirmed in cross-validations and in comparison with a currently used sequence-comparison-based method. When comparing the predictions for 1230 NRPS A domains that are currently detectable in UniProt, the new method was able to give a specificity prediction in an additional 18% of the cases compared with the old method. For 70% of the sequences, both methods agreed; for $< 6\%$, they did not, mainly on low-confidence predictions by the existing method. None of the predictive methods could infer any specificity for 2.4% of the sequences, suggesting completely new types of specificity. The new prediction method is implemented on a freely-usable webserver reachable at `www-ab.informatik.uni-tuebingen.de/software/NRPSpredictor`.

4.2 Motivation

The Adenylation domain (A domain) of NRPS specifically recognizes and activates one amino acid (or hydroxy acid) that will subsequently be appended to the nascent peptide chain by other NRPS domains (see Chapter 2 for details). Based on the crystal structure of the phenylalanine activating A domain of the NRPS gramicidin synthetase A (GrsA), Conti et al. [1997] determined ten residue positions that are crucial for substrate binding and catalysis. These residues are within a radius of $\sim$ 5.5 Å around the phenylalanine bound in the active site. The predictive method described by Stachelhaus et al. [1999] and Challis et al. [2000] is based on the high structural conservation of the binding pocket with a root mean square deviation (RMSD) of the C_α atoms of < 1 Å [di Vincenzo et al., 2005], reflected by a relatively high mutual sequence similarity of 26% to 56% [Marahiel et al., 1997] of NRPS A domains. Therefore, Stachelhaus et al. [1999] and Challis et al. [2000] concluded that the ten decisive residues of GrsA will line up with the corresponding positions of other A domains in a multiple sequence alignment, and can be extracted to form a 'specificity-confering code'. The specificity of uncharacterized A domains can then be inferred based on the 'code' of domains with known specificity [Challis et al., 2000] or based on consensus sequences for each specificity [Stachelhaus et al., 1999]. In this chapter, we present a new method for predicting the specificity of A domains by machine learning, using the *physico-chemical fingerprint* of the residues lining the active site of the enzymatic domain (8 Å around the bound substrate). The generality of the approach makes it applicable to the prediction of functional subspecificities of other classes of enzymes which share a conserved structure but catalyze different substrates (see the *Conclusion* at the end of this chapter at page 58). We use a state-of-the-art technique for encoding of residues into feature vectors for machine learning based on the physico-chemical properties of the amino acids, and use an up-to-date training dataset of A domains with known specificity that we have compiled from the literature.

4.3 Materials and Methods

4.3.1 Acquisition of a Collection of A Domains with Known Specificity

The HMMER package [Durbin et al., 1998, `hmmer.janelia.org`, see also 3.3.2 for an introduction] and self-written Perl scripts were used to search for NRPSs in the protein databases UniProt/TrEMBL/Swiss-Prot [Apweiler et al., 2004; Boeckmann et al., 2003], requiring the occurrence of a complete NRPS module with at least one Condensation domain, one A domain (AMP-binding) and one peptidyl carrier domain (Pfam [Bateman et al., 2004] accession numbers PF00668, PF00501, PF00550). The same software was also used to extract the AMP domains from NRPS sequences to generate Profile Hidden Markov Models (HMMs) of parts of domains and to extract certain

positions of subdomains that were aligned against HMMER profiles. The programs `ClustalW` [Thompson et al., 1994], `T-Coffee` [Notredame et al., 2000], and `MUSCLE` [Edgar, 2004a,b] were used for generating multiple sequence alignments that were then manually checked for good alignment of core sequences, structural 'anchors' and putative constituents of binding pockets. Specificity annotations of extracted A domains were either obtained directly from the literature or by following references (PubMed links, gene name, organism, authors etc.) given in database entries of proteins.

4.3.2 Extraction of Homologous Positions of A Domains

The "Biochemical Algorithms Library" (BALL) [Kohlbacher and Lenhof, 2000] and a simple `Python` script were used to extract residues that have at least one atom at a given distance from the bound phenylalanine in the GrsA-Phe crystal structure (PDB ID `1AMU`, [Conti et al., 1997]). In a multiple alignment of different A domains with the protein sequence of GrsA-Phe, the residue positions that lined up with certain residues in GrsA-Phe were extracted; we ensured that all extracted residues lie in conserved, gap-free segments to allow for a reliable inference of their structural and functional relevance (see Figure 4.1 for illustration).

4.3.3 Processing the Collection of A domains for Machine Learning

Starting with the current set of A domain sequences with known specificity (Section 4.3.1), the 34 residues at a distance of up to 8 Å from the bound phenylalanine in GrsA were extracted and duplicate sequences were removed; sequences with similar specificities (see *Results and Discussion* in this chapter) were clustered to *composite specificities*. Clusters comprising fewer than five sequences were discarded.

4.3.4 SVMs

For the classification of data, Transductive SVMs (TSVMs) were used (see Section 3.2.6 for details). The performance of the generated models was evaluated applying the statistics described in Section 3.2.7.

Initially, we evaluated our models for the different *composite specificities* using both x-fold cross-validation (3-, 5-, and 10-fold, each repeated three times with randomized splits) and LOO. Since both tests yielded extremely similar results, here we report only on the results of the LOO, the most fine grained form of cross-validation. A more thorough evaluation of the accuracy would require two levels of cross-validation (i.e. nested cross-validation) [Markowetz and Spang, 2005]. However, as the models considered here are relatively simple and do not allow for strong fitting of the data, using a straight LOO test is sufficient for our purposes.

Feature Representation Based on Physico-Chemical Properties of Amino Acids

From each A domain, we extracted a signature of 34 amino acids. This consisted of all residues with at least one atom ≤ 8 Å from the bound substrate. Residues of the A10 core motif (`NGK, K=Lys517`) [Marahiel et al., 1997] were not included because they are extremely highly conserved and do not vary between different specificities. We encoded each amino acid by normalized real values representing their physico-chemical properties. We used amino acid indices from AAindex [Kawashima and Kanehisa, 2000; Tomii and Kanehisa, 1996; Nakai et al., 1988] and Neumaier et al. [1999] to describe:

- the number of hydrogen bond donors [Fauchere et al., 1988],
- polarity (three different indices) [Zimmerman et al., 1968; Radzicka and Wolfenden, 1988; Grantham, 1974],
- volume [Tsai et al., 1999],
- secondary structure preferences for beta-turns, beta-sheets and alpha-helices [Chou and Fasman, 1978],
- hydrophobicity with a three-dimensional vector [Neumaier et al., 1999], and
- the isoelectric point [Zimmerman et al., 1968].

We standardized the values in such a way that the interval of ± 1 standard deviation (calculated from the value distribution of each AAindex file) was projected onto the interval of ± 1.
$StandardizedValue = \frac{IndexValue - MeanIndexValue}{StandardDeviation}$, and thus obtained a vector of 408 features for each A domain. The choice of these properties is discussed in the Section *Results and Discussion* of this Chapter (4.4.4).

SVM Implementation

In this study, we used the program package SVM^{light} [Joachims, 1999b, `svmlight.joachims.org`] for training SVM models on data and classification of data. This program also implements algorithms for training large transductive SVMs (TSVMs). The algorithm proceeds by solving a sequence of optimization problems, lower-bounding the solution using a form of local search. For details, see Joachims [1999a]. SVM^{light} can efficiently compute LOO testing; LOO provides "almost unbiased" estimates for error rate, recall (= sensitivity S_n), and precision (= specificity S_p) [`svmlight.joachims.org`].

Choice of the Optimal Kernel Function and Parameters

SVM^{light} provides linear, polynomial, radial basis (RBF) and sigmoid kernel functions. Two parameters, C and j, need to be chosen independently from the choice of the kernel function. The parameter C is the penalty

that is assigned to erroneous training points that cannot be classified correctly. If the features are normalized as described above, one can put $C = 1$ as a starting point for a grid search around this value (in this study $C \in \{\frac{1}{32}, \frac{1}{16}, \ldots, 1, 2, \ldots, 32\}$). The cost-factor j determines how training errors on positives examples outweigh errors on negative examples (see Section 3.2 for details). The usual initial estimation j_0 (see Morik et al. [1999]) of the cost-factors by the proportion of negative to positive training examples was also used in this study, with values of j_0 in the order of 10, depending on the ratio of the dataset. To determine the optimal value for j, a grid search was applied as well with $j \in \{\frac{1}{32}j_0, \frac{1}{16}j_0, ..., 1j_0, 2j_0, ..., 32j_0\}$. The non-linear kernel functions have additional parameters. The RBF kernel function has an additional parameter σ, with $\sigma^2 \approx$mean$(||x_i - x_j||^2)$, that is approximately the mean of the squared Euclidian distances of all pairs of data points. To be precise, SVM^{light} uses a parameter γ for the RBF kernel, with $\gamma = \frac{1}{2\sigma^2}$. The approximation given above can then be used as starting point for a grid search to find the best value for σ^2. In this study, the same factors as for the optimization of C and j were used, multiplied by the initial approximation of γ.

Multiclass Problem

After having trained the SVM models for each *composite specificity*, it is necessary to combine the predictions of all models to one single prediction for the "large" and "small" clusters. The most widely used method (according to Vert et al. [2004]) is to combine the scores (= distance from the classified point to the hyperplane) by a *max rule*: the SVM that outputs the largest score is used to assign the specificity to the unknown sequence. If all single SVMs return "negative", then no final prediction will be possible. This does not necessarily mean that the unknown sequence has a very "exotic" specificity, but possibly that the single model of the actual specificity might give a false negative answer. Because the quality of the single models differs, we decided to multiply the scores by the squared MCC value of the model. As the MCC is a quality measure close to 1 for very good models, and decreases with the reliability of the models, this allows for a reasonable scaling of the scores. In the relatively rare case of several "positive" answers, the one with the highest scaled score will be used in the evaluation of the overall predictive error of the combined model. However, the predictive program (NRPSpredictor) will list all models that return a positive value.

4.4 Results and Discussion

4.4.1 A Current Set of Annotated Specificities

Because the large majority of NRPS sequences deposited in public sequence databases are poorly annotated, and the annotation quality and syntax differs from author to author, keyword-based search strategies in an automated manner are infeasible. Therefore, we first manually collected all 160 A domain sequences used by Stachelhaus et al. [1999]. We then scanned the

Specificity	Occurrence	Specif.	Occur.	Specif.	Occur.
3-me-Glu	1	Dhb	15	Phe	11
4pPro	1	Dhpg	8	Phg	1
Aad	10	Dht	4	Pip	5
Abu	2	D-lyserg	1	Pro	16
Aeo	1	Gln	8	Sal	2
Ala	34	Glu	12	Ser	22
Ala-b	3	Gly	12	Ser-Thr	2
Ala-d	1	His	1	Tcl	1
Alaninol	1	Hpg	19	Thr	24
Arg	5	Hyv-d	1	Trp	3
Asn	14	Ile	11	Tyr	14
Asp	12	Iva	7	Val	27
Bht	7	Leu	31	Valhyphaa	1
Bmt	1	Lys	5	Vol	1
Cys	23	Lys-b	2		
Dab	4	Orn	10		

Table 4.1: Distribution of the 397 Adenylation domains with known specificity on their substrates: Besides the proteinogenic amino acids in three letter code there are the following known rare specificities: *3-me-Glu* 3-methyl-glutamate, *4pPro* 4-propyl-proline, *Aad* 2-amino-adipic acid, *Abu* 2-amino-butyric acid, *Aeo* 2-amino-9,10-epoxy-8-oxodecanoic acid, *Ala-b* β-alanine, *Ala-d* D-alanine, *Alaninol*, *Bht* beta-hydroxy-tyrosine, *Bmt* (4R)-4[(E)-2-butenyl]-4-methyl-L-threonine, *Dab* 2,4-diamino-butyric acid, *Dhb* 2,3-dihydroxy-benzoic acid, *Dhpg* = *Dpg* 3,5-dihydroxy-phenyl-glycine, *Dht* dehydro-threonine = Dhbu = 2,3-dehydroaminobutyric acid, *D-lyserg* D-lysergic acid, *Hpg* 4-hydoxy-phenyl-glycine, *Hyv-d* 2-hydroxy-valeric acid, *Iva* iso-valine, *Lys-b* β-lysine, *Orn* ornithine, *Phg* phenyl-glycine, *Pip* pipecolic acid, *Sal* salicylic acid, *Tcl* (4S)-5,5,5-trichloro-leucine, *Valhyphaa* valine or hydrophobic aa, *Vol* valinol.

UniProt/TrEMBL/Swiss-Prot protein database [Apweiler et al., 2004; Boeckmann et al., 2003] with profile HMMs for complete NRPS modules. We required modules to be complete (one Condensation, one A and one T (pp-binding) domain), so we could avoid extracting very similar enzymes, such as acyl-CoA ligases. For the 245 detected sequences, we followed the PubMed [`www.pubmed.gov`] literature references in the UniProt entry or tried to find the associated articles via PubMed (searching for gene name, organism, authors etc.). Thus we were able to find 227 additional A domain sequences. We joined this dataset with the sequences of J. Ravel's NRPS BLAST server [Challis et al., 2000] and finally obtained a set of 397 A domains with known specificity (fully listed in the Supplementary Data). We required the specificity annotation to be based on experimental evidence, either by an ATP-PP_i-exchange reaction [Miller and Lipman, 1973] or, when the specificity was inferred by the co-linearity rule based on the ordered composition of the peptide product, that the inference was confirmed by an unambiguous match with a known "specificity-conferring code" [Stachelhaus et al., 1999; Challis et al., 2000] of another A domain. The number of occurrences of the different specificities in these 397 A domain sequences are depicted in Table 4.1.

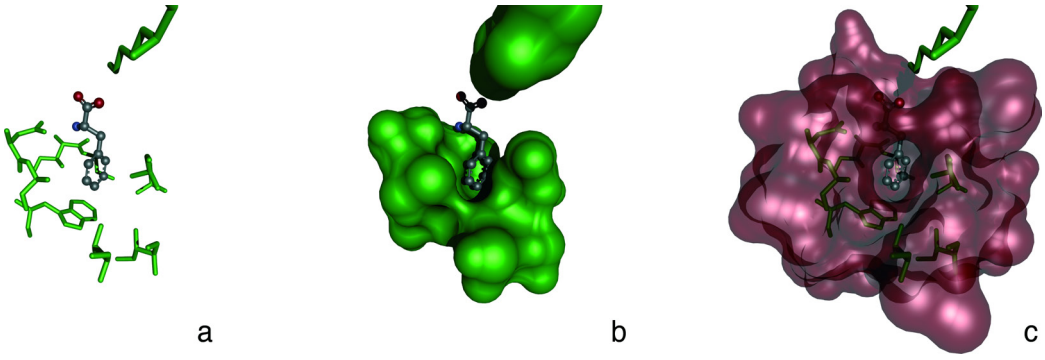

Figure 4.1: Illustration of the residues that have been taken into account for the predictive methods by Stachelhaus et al. [1999] and the method presented here. A phenylalanine is depicted with the residues of a gramicidin synthetase A activation domain which constitute the active site and are at a certain distance to the bound substrate (PHE). (**a**) The ten residues (green) that are in direct contact with the substrate phenylalanine (ball and stick representation) are shown. These 10 residues are the basis for the specificity prediction method by Stachelhaus et al. [1999]. (**b**) The same ten residues as in (a) are depicted but in the space filling representation. (**c**) The residues in green (at a distance of up to 5.5 Å from phenylalanine) are surrounded by all 34 residues (purple) at a distance of up to 8 Å from phenylalanine. The predictive method described here is based on these 34 amino acids and encodes them by their physico-chemical properties. Representations were created using BALLView [Moll et al., 2005, 2006].

4.4.2 Inferring Functional and Structural Relevance of Residues in a Structurally Conserved Context

When comparing firefly luciferase (another AMP-binding enzyme that activates luciferin) and GrsA, a structure-based alignment reveals that 67% of the alpha-carbon positions are conserved to within 3 Å. The RMSD is 2.6 Å, although both enzymes share only 16% sequence similarity [Stachelhaus et al., 1999; Conti et al., 1997; di Vincenzo et al., 2005]. However, the RMSD, calculated over the C_α atoms enclosed in a sphere of radius 9 Å centered at the GrsA residue Asp235 in the active site, is 0.95 Å [di Vincenzo et al., 2005]. Owing to the much higher similarity between GrsA and other NRPS A domains (between 30 and 80% [Turgay et al., 1992]), the conformation of their mainchains is likely to be even more similar, particularly around the substrate binding pocket. Therefore, in a multiple sequence alignment of other NRPS A domains with GrsA, those residues that align with the residues that line the active site can be expected to be involved in the specific substrate recognition and binding of the homologous A domain. To make sure that we included all residue positions that might have an interaction with the substrate, or might be influenced by or adapted to the residues that interact directly with the substrate, we decided to extract all residues up to a distance of 8 Å from the substrate in GrsA. A *steric cell* of 8 Å was likewise used by Lilien et al. [2004] for an energy simulation of the GrsA active site. In Fig. 4.1, we illustrate the residues at a distance up to 5.5 Å and 8 Å in direct and indirect contact with the substrate phenylalanine, respectively.

4.4.3 Clustering of Sequences with Similar Specificities

For a reliable prediction of specificities, the ideal is to have a training set of sequences for each distinct specificity. In reality we often find A domains with considerably high side specificities that either lead to alternate peptide products that differ at the corresponding position, such as in the case of tyrocidine: in the tyrocidine biosynthesis operon (*Bacillus brevis*, TYCB_BREPA, Mootz and Marahiel, 1997), the A domain TycB_m3 activates L-tryptophan with 100% relative activity (in an ATP-PP_i-exchange reaction), and L-phenylalanine with 48%, but is annotated as L-phenylalanine activating because (D-)phenylalanine is found in the product. It is also possible that in biochemical specificity tests (ATP-PP_i-exchange reaction with recombinant A domains), a considerable side specificity might be detected but the alternative substrate is not incorporated *in vivo* because of, for example, sterical reasons in the further processing of the nascent peptide: the A domain of BarD of the barbamide (*bar*) biosynthetic gene cluster has 100% specificity for leucine and valine, and 80% for trichloroleucine, but the *in vivo* incorporation of valine was experimentally excluded [Chang et al., 2002]. Because considerably high side specificity might exist, we addressed this problem by clustering specificities for amino acids with very similar physico-chemical properties. For this clustering we also took observations of Challis et al. [2000] into account. They analyzed the predicted binding pockets of most A domains known to date. Based on the "code" of eight amino acids closest to the substrate, they pointed out that specificities for physico-chemically similar substrates often only differ in single residues [Challis et al., 2000]. An experimentally verified example is the directed mutagenesis of Ala322Gly in GrsA increasing its specificity to Trp [Stachelhaus et al., 1999]. We decided to consider two different kinds of clusterings: grouping specificities into a few large clusters and into more small clusters. Forming larger clusters, i. e. putting together more closely related specificities into one *composite specificity* has the advantages of (i) obtaining larger positive datasets for SVM training (yielding models that are more reliable), (ii) covering a larger spectrum of sequence variations, (iii) covering a larger subspace in the hyperspace, (iv) lowering the risk of overfitting, and finally, (v) allowing for recognition of new substrates that are very similar to the substrate specificities in the cluster. However, forming smaller clusters by clustering similar specificities only where necessary (e. g. Phe/Trp, see above) has the advantage of allowing for more concrete/precise predictions, but at a higher risk of overfitting owing to a reduced number of positive training data. Table 4.2 and Figure 4.2 illustrate which specificities have been clustered.

Large Clusters		Small Clusters	
Gly (12), Ala (20), Val (22), Leu (22), Ile (7), Abu (2), Iva (7)	apolar, aliphatic side chains	Gly (12), Ala (20)	tiny size, hydrophilic, transition to aliphatic
		Val (22), Leu (22), Ile (7), Abu (2), Iva (7)	aliphatic, branched hydrophobic side chain
Ser (13), Thr (16), Ser/Thr (1), Dhpg (7), Hpg (13)	aliphatic chain or phenyl group with -OH	Ser (13)	Serine specific
		Thr (16)	Threonine specific
		Dhpg (7), Hpg (13)	polar, uncharged (hydroxy-phenyl)
Phe (11), Trp (3), Phg (1), Tyr (12), Bht (6)	aromatic side chain	Phe (11), Trp (3)	unpolar aromatic ring
		Tyr (12), Bht (6)	polar aromatic ring
Asp (8), Asn (13), Glu (9), Gln (6), Aad (7)	aliphatic chain ending with H-bond donor	Asp (8), Asn (13)	Asp-Asn-hydrogen bond acceptor
		Glu (9), Gln (6)	Glu-Gln-hydrogen bond acceptor
		Aad (7)	2-amino-adipic acid
Cys (17)	polar, uncharged (aliphatic chain with -SH group at the end)	-	-
Orn (8), Lys (3), Arg (5)	long positively charged side chain (aliphatic chain with -NH_2 group at the end)	Orn (8)	Orn and hydroxy-Orn specific
		Arg (5)	Arg-specific
Pro (16), Pip (4)	cyclic aliphatic chain with polar -NH_2^+ group	Pro (16)	Pro-specific
Dhb (9), Sal (2)	hydroxy-benzoic acid derivates (no amino group)	No small cluster, no separation possible	-

Table 4.2: Clustering amino acids with similar physico-chemical properties and/or similar substrate binding pockets [Challis et al., 2000] into *composite specificities*. The numbers in parenthesis denote the counts of domains with unique 8 Å sequences. Please note that the division of large into small clusters was not always possible owing to the small amount of available training data. Please also see Figure 4.2.

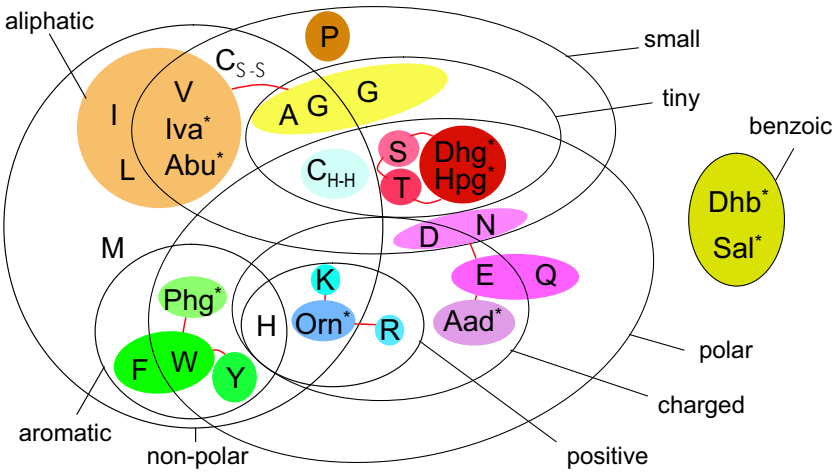

Figure 4.2: Venn diagram grouping amino acids by common physico-chemical properties according to Taylor [1986]. The colored sets show how similar amino acids have been clustered to *composite specificities* of A domains. To get larger clusters, several smaller clusters were joined as indicated by red lines connecting colored sets. This clustering is based on conclusions by Challis et al. [2000] on cross-specificities of A domains and our own groupings according to physical-chemical properties. An asterisk indicates rare non-proteinogenic amino acids. For abbreviations, see Table 4.1.

4.4.4 SVMs: Particularities

Feature Representation Based on Physico-Chemical Properties of Amino Acids

From each A domain, we extracted the signature of 34 amino acids at ≤ 8 Å from the bound substrate (see the *Materials and Methods* Section in 4.3). Each amino acid was encoded by 12 different values representing its physico-chemical properties, obtaining a vector of 408 features for each sequence. Chemical properties chosen were the number of hydrogen bond donors, the polarity and the hydrophobicity of the residues, and the isoelectric point; physical properties were volume and preferences to appear in different secondary structures. We chose these properties because they are the key factors in deciding how well a given substrate might bind to the defined set of residues and thus make sense, biologically, chemically and physically. If the positions of the active site residues are given and fixed in space (as we conclude they are here) then these properties describe the inside of the active site keyhole well. If for any reason (e.g. a very large / different substrate) the binding pocket structure is altered, then we expect to see residues that have a different secondary structure preference at the positions that we extract from the profile alignment. Therefore, it also makes sense to encode the secondary structure preferences.

SVM Implementation

In this study, we based our predictions on SVMs that implement the intuitive idea of separating two data "clouds" by a geometric plane (see *Materials and Methods* Section at 4.3 for details), as implemented in SVM^{light}. We used an innovative variant of SVMs, so-called transductive SVMs, that not only take the labeled training data into account but also integrate unlabeled data, in our case, sequences with unknown specificity. We tried different kernel functions in our experiments, including linear, polynomial, radial basis and sigmoid functions. In a grid search, we determined the optimal kernel parameters using SVM^{light}'s built-in leave-one-out test functionality. For linear and radial kernel functions (RBF) we got the best results (for error rate, specificity, sensitivity and MCC), varying from case to case. When the linear kernel was equally good or better, then we preferred it over the RBF kernel for simplicity of the models, otherwise we chose the RBF kernel. After the determination of the optimal kernel function and parameters, we gathered 646 uncharacterized A domain sequences from UniProt [Apweiler et al., 2004], as described in *Materials and Methods* at 4.3.
For each cluster of composite specificity, we prepared a feature file with the sequences belonging to this specificity labeled +, all other sequences with different but known specificity labeled −, and the uncharacterized sequences labeled 0 (i.e. unlabeled). We used SVM^{light} in transductive mode to build models. With a self-written Perl script, we ran LOO cross-validation to check error rate, recall (sensitivity), precision (specificity) and MCC. For each cluster, we trained a TSVM, as described above, to obtain a model for each composite specificity.

SVM Quality Assessment

The number of available positive training data points varied between the different clusters and was sometimes quite small. Although the quality of models in machine learning depends on the amount of training data available, previous findings show that, besides the highly conserved overall structure of the binding pocket common to all A domains, the composition of residues lining the active site of sequences with the same specificity are even more conserved [Stachelhaus et al., 1999], which should allow one to obtain relatively good models. In a rigorous quality assessment of the generated models (Table 4.3), we could show that most SVM models yield good to very good results (Matthews correlation coefficients 0.85-1). Some yield very poor results, such as the SVM with the *composite specificity* for Phe=Trp=Phg=Tyr=Bht ($MCC = 0.85$) or for Gly=Ala ($MCC = 0.84$). An explanation for the low performance of the model for very large aromatic amino acids could be that there exist a few, but spatially very different configurations of the binding pocket, for which it is impossible to generate one discriminative model. The problems with the glycine/alanine model could lie in the small size of the substrates; as Challis et al. [2000] already suggested, there might be many degenerate solutions to activate these substrates. Similarly, the quality of the model for proline specificity is poor. As Lautru and Challis [2004] pointed out, only 4-5 residues at the top of the selectivity pocket are likely to be in direct contact with proline's relatively compact side chain, based on homology modelings of the binding pockets. We obtained bad performance for models that aimed at distinguishing between phenylalanine/tyrosine and all other amino acids, because there are sequences known, like tyrocidine synthetase TycC_M3, with specificities for both Tyr and Trp, and others that have a specificity for Phe and Trp. A tabular overview of all results of the quality assessment of the models is shown in Table 4.3. As the predictive quality of the models was estimated by LOO tests on the set of sequences with known specificity (training data), one needs to check if the test data (sequences with unknown specificities) are drawn from the same distribution. To check this, we compared the mean pairwise distance of the training data with the mean pairwise distance between training and test data. The mean of the Euclidian distance within the training data was 18.9 (SD: 2.9) and the mean distance between training and test data was 18.8 (SD: 2.5). Because both distributions are very similar, it is safe to assume that the performance of our models on the test data will be similarly good. To finally obtain one model for all "large" and all "small" clusters, we score the results of different models using the returned distance of the data point to the hyperplane multiplied by the square of the MCC. This scaling makes sense because the MCC reflects the reliability of each model (see 4.3.4). Our "large" clusters cover 282 of the 300 specificities; the "small" clusters cover 273 sequences. We ran a LOO test on both multi-class models. The "large" cluster model gave 260 correct predictions and 30 incorrect predictions, and ten times, it gave no prediction, corresponding to a total error rate of 13%, or 7.8% on the sequences that the models were trained for. The "small" cluster model gave 231 correct predictions and 44 incorrect predictions, and did not decide for 25

sequences, corresponding to an error rate of 23%, or 15%, respectively. Given the set of 300 unique 8 Å signature sequences, we also evaluated the performance of a sequence-based model that used the 34 amino acid signatures. To get a first overview of the clustering of the 34 amino acid sequences, we built a phylogenetic tree (using a Maximum Likelihood method (the principle is described in Section 3.4.1; implementation used: IQPNNI by Vinh and von Haeseler [2004])), visualized with SplitsTree [Huson and Bryant, 2006, `www.splitstree.org`]. For the tree, see the *Supplementary Data* Section in 4.7. When we analyze the tree, we see – overall – a clustering of similar specificities. Looking at details, we detect some "incompatible" specificities in some subtrees, e.g. in one where most A domains of the fungus *Trichoderma virens TEX1* gene (Uniprot accession no. Q8NJX1) cluster despite their different specificities. The reason might be that by increasing the number of amino acid positions from 8 or 10 to 34, we also capture more of the species' phylogenetic signal. We also tested the performance of a BLAST search [Altschul et al., 1997; for introductory information see Section 3.3.1] using the 300 sequences with known specificity as our database. Using the closest BLAST match to infer the specificity, 233 sequences would have been annotated correctly, corresponding to an error rate of 22.3%. This indicates that BLAST could be helpful especially for rare specificities and, therefore, we plan to integrate it in a future version of the NRPSpredictor. However, the BLAST strategy is inferior to the SVM strategy because it cannot build a generalizing SVM model for a specificity, but only finds the closest sequence(s). To assess the accuracy of the predictions on "new" sequences, which are not very similar to the others with known specificity, we re-trained models only with sequences with a certain minimum distance and still got acceptable results (see the *Supplementary Data* Section in 4.7). To further examine the reliability and usefulness of our new method, we applied our prediction program to all 1230 Adenylation domains in the June 2005 version of UniProt [Apweiler et al., 2004] (the proteins were extracted from the database as described in Section *Materials and Methods* of this chapter at 4.3). We compared the consistency of our predictions with the predictions based on the "specificity-conferring code". (To automate this method by Stachelhaus et al., we automatically extracted the code of the 10 amino acids and scored it against the collection of 10 amino acid codes of known specificities, requiring the identity of at least seven of the ten positions for a "match"). For 70% of the sequences, both predictors gave consistent predictions, which underlines the usefulness of our approach. The new SVM-based method could predict the specificities for 18% of the sequences, where the sequence-based method by Stachelhaus et al. [1999] cannot. However, there are 2.4% for which neither method gives a prediction. For 1.5%, only the traditional method could give a prediction. About 8.8% of the sequences are inconsistently classified by the old and the new method; of them, 3% are rare specificities that the SVMs were not trained for.

Specificity of SVM	**positive training points**	**kernel type**	**leave-one-out cross-validation** *Error*	S_n	S_p	MCC	**quality of SVM**
large clusters	282 labeled and 664 unlabeled data points (18+646)						
Dhb=Sal	11	l	0.4	100	92	96	++
Asp=Asn=Glu=Gln=Aad	43	r	1.4	100	91	95	++
Pro=Pip	20	r	0.7	90	100	95	++
Cys	17	r	0.7	100	89	94	++
Ser=Thr=Dhpg=Dpg=Hpg	50	r	2.5	96	91	92	++
Gly=Ala=Val=Leu=Ile=Abu=Iva	92	r	4.3	95	93	90	+
Orn=Lys=Arg	16	l	0.7	88	88	87	+
Phe=Trp=Phg=Tyr=Bht	33	r	3.2	88	85	85	0
small clusters	273 labeled and 673 unlabeled data points (27 + 646)						
Dhb=Sal	11	l	0	100	100	100	++
Aad	7	l	0	100	100	100	++
Glu=Gln	15	l	0	100	100	100	++
Dhpg=Dpg=Hpg	20	l	0.4	100	95	97	++
Ser	13	l	0.4	92	100	96	++
Cys	17	l	0.7	100	89	94	++
Thr	16	l	0.7	94	94	93	++
Pro	16	r	0.7	94	94	93	++
Asp=Asn	21	l	1.1	90	95	92	++
Val=Leu=Ile=Abu=Iva	60	l	2.9	92	95	91	+
Orn	8	l	0.7	88	88	87	+
Gly=Ala	32	l	3.3	81	90	84	0
Tyr	18	r	2.2	94	77	84	0
Arg	5	l	0.7	80	80	80	0
Phe=Trp	14	l	3.7	57	67	60	-

Table 4.3: Results of cross-validating the different SVMs by leave-one-out. The more training data that are available, the more reliable the trained predictive models are. The "quality of SVM" in the last column therefore is a qualitative measure for the MCC. Kernel type l stands for linear kernel; r stands for radial basis function kernel. Error rate, sensitivity (S_n), specificity (S_p) and Mathews correlation coefficient (MCC) are given in percent.

An illustration of these comparisons are shown in Figure 4.3. If we accept only ≥80% matches for a positive "Stachelhaus" prediction, we observe that the number of sequences for which no predictor can say anything increases by 2.5% and the specificities that can only be predicted by the TSVMs increase by 8%. We also observe that the number of inconsistent predictions drops by 6.5%, the number of sequences only predicted by the Stachelhaus method drops by 1% and the number of consistent predictions decreases by 4%. We interpret this observation by saying that the Stachelhaus predictions at 70% are less reliable and give rise to more inconsistent predictions.

4.5 Conclusion

During the past decade, SVM-based machine learning has been extensively applied within the field of bioinformatics, e.g. to the classification of genes and proteins, predictions along the DNA or protein strand and microarray gene expression, and to other problems (for a recent review, see, for example, Noble [2004]). Here we describe a new application of SVMs to functional subtyping of the substrate specificities of a class of enzymes based on the physico-chemical fingerprint of the residues that form the substrate binding pocket. To take advantage of the abundant amount of unannotated data, we use an implementation of TSVMs [Joachims, 1999a] first introduced by Vapnik in 1998. TSVMs have been shown to be superior to inductive SVMs in a similar application, the prediction of receptor binding compounds based on three-dimensional properties of the molecule [Schölkopf et al., 2003; Weston et al., 2003], where also a large number of unlabeled data were avaliable. Our results prove a high reliability of the predictions, even though the currently available amount of training data is relatively low, leaving room for further improvement with a growing number of annotated A domains. When applying our method and the sequence-based method [Stachelhaus et al., 1999; Challis et al., 2000] to a set of over one thousand Adenylation domains currently detectable in UniProt [Apweiler et al., 2004], in summary, the new method can predict the specificities for 18% more sequences than the old one, while being consistent within the 70% that both methods predict. For 2.4% of the sequences, none of the methods can make any prediction. Moreover, the inconsistent predictions, where both methods disagree, have a large amount of "Stachelhaus" predictions at 70% identity. This illustrated that there is still a large amount of sequences for which a prediction is very uncertain or impossible. Interestingly, we can observe that in those "difficult" sequences, the ratio of eukaryotic sequences is more than two times higher than it is on average, indicating that the eukaryotic A domains might have developed alternative substrate binding patterns. In cases where both methods give consistent predictions, the method by Stachelhaus et al. gives a more concrete prediction, since it decides for one specificity, whereas our method decides for one *composite specificity* that usually stands for more than one substrate. Nevertheless, we would like to emphasize that the combination of the "old" and our method gives a new powerful prediction tool that can be directly used by the scientists working in the field. Our results confirm

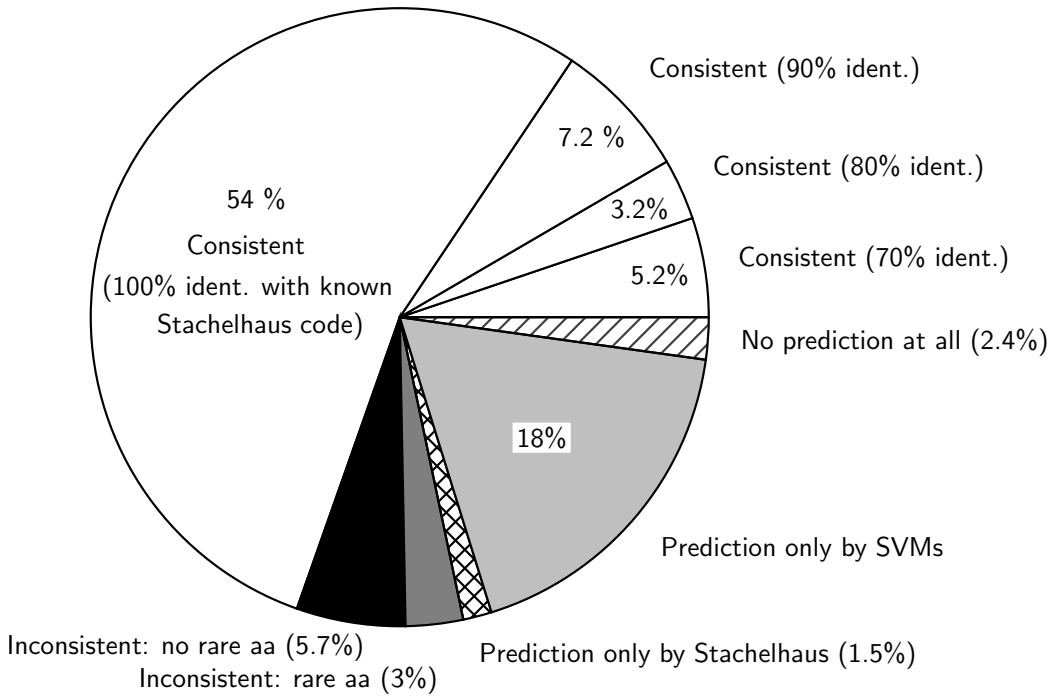

Figure 4.3: Results of a comparison of the new SVM-based method with the sequence-based prediction method based on the "specificity-confering code" by Stachelhaus et al. [1999] and Challis et al. [2000]: (For simplicity, we refer to the latter as the "Stachelhaus method"). Of 1230 Adenylation domains (with HMMER automatically extracted from the June 2005 version of UniProt), 70% or 858 obtained consistent predictions by both predictors (white sectors). For most of these consistent predictions – 54% of the total (or 666) – the Stachelhaus method was based on an exact match with a known "specificity-conferring code"; the others had at least a 70% match. To 2.4% (or 29 sequences), none of the predictors could assign any specificity (no match $\geq$ 70%, diagonal hatches). Eighteen percent or 217 sequences could be classified only by the SVMs and not by the Stachelhaus method (light gray sector), and 18 A domains (1.5%) could not be classified by the SVMs but by the Stachelhaus method (cross-hatched); two of them are rare specificities. The Stachelhaus predictions for the rest are mainly based on 70% matches to known specificity "codes". For 108 sequences (8.8%), the predictions were inconsistent but 38 of them (3% of the total, gray sector) had matches to rare amino acids that were not used for training the SVMs. The remaining 70 incompatible predictions were mainly based on $\leq 80\%$ identity matches with known "specificity-conferring codes" (black sector).

the applicability of the SVM-based strategy for substrate specificity prediction, and that it should also be considered for the prediction of the subtypes of other enzymes e.g. nucleotidyl cyclases, protein kinases, lactate/malate dehydrogenases and trypsin-like serine proteases, a selection used by Hannenhalli and Russell [2000].

4.6 Availability of the Program

An implementation of the described method, called NRPSpredictor, is freely available to the NRPS community as an online resource via our server reachable at:

`www-ab.informatik.uni-tuebingen.de/software/NRPSpredictor`.

The web-interface (see Fig. 4.4) allows one to upload or paste in the (multi-)fasta file(s) of the protein sequence(s) to be analyzed. The Adenylation domains are automatically extracted, as well as the residues of the "specificity-confering code" and the residues 8 Å around the substrate. The predictions of each model for each cluster of composite specificity are given, as well as the best matches of the "specificity-confering code" to known specificities. The results are presented as an HTML output as well as a short text-based report and a tabular output that can be viewed with a spreadsheet program (Fig. 4.4). The HTML interface is dynamically generated by a web framework [Biegert et al., 2006]. A sophisticated, job tracking facility based on a database allows monitoring the progress of several jobs running simultaneously. Also, by providing a fixed URL for each submitted job, the user still can access the results several days (currently nine) later. A detailed discussion of the web interface's architecture is described by Biegert et al. [2006]. The command-line version of the NRPSpredictor, which is internally called via the web interface, consits of several Perl scripts and uses the program packages HMMER and SVM^{light}. A free copy can be obtained from the author on request [Rausch et al., 2005].

4.7 Supplementary Data

A table and text files of all 397 annotated A domain sequences gathered and used in this study as training data, a phylogenetic tree of their 34 amino acid signature sequences, all automatic predictions for 1230 A domains of the June 2005 version of UniProt, the kernel parameters used for model training and the results of a test of the effect of sequence redundancy reduction are freely available at NAR Online on the website of the article [Rausch et al., 2005].

4.8 Acknowledgments

For the research project presented in this chapter, we would like to thank Andreas Wietzorrek (Department for Microbiology/Biotechnology, University of

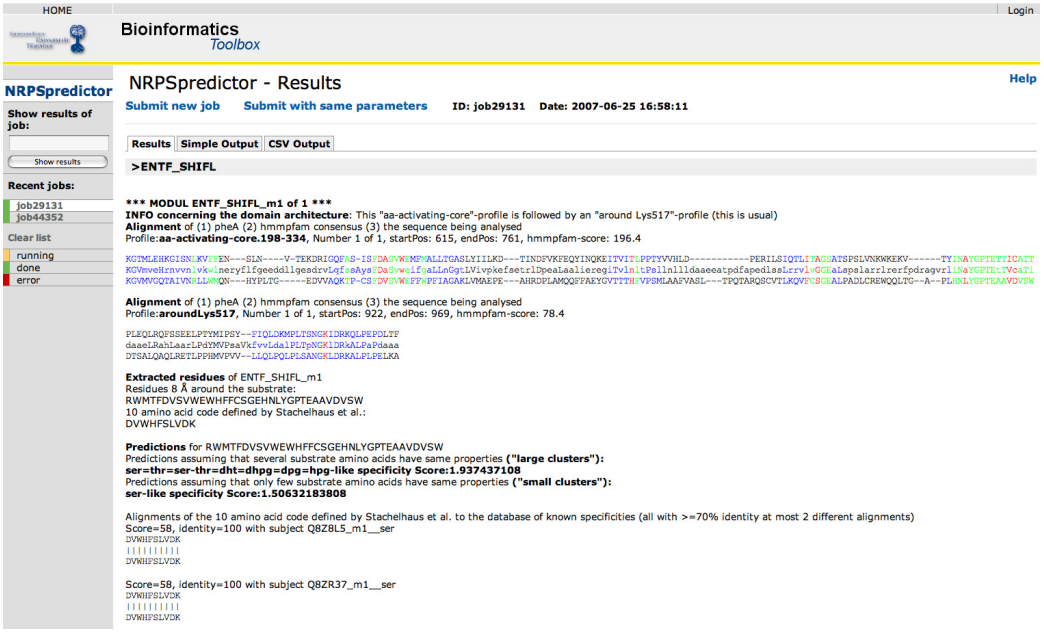

Figure 4.4: Screenshot of NRPSpredictor's web frontend showing the **Results page** of the analysis of enterobactin synthetase component F (UniProt [Wu et al., 2006] entry name ENTF_SHIFL). The report contains one section for each detected A domain in the sequence, beginning with an alignment of the core region (pos. 198-334) and the region around the conserved Lys517 of the GrsA A domain, the consensus sequence of the profile HMM used for the sequence extraction and the corresponding region of the query sequence. Conserved "anchor" regions are colored in the alignment, the 10 positions defined by Stachelhaus et al. [1999] are highlighted in red and positions additionally contained in an 8 Å sphere around the active site center of `1AMU` (GrsA A domain) are highlighted in green. In the following subsection, the extracted residues 8 Å around the substrate and the 10 amino acid code are listed. The *Predictions* subsection lists the classifiers that returned a positive score for a "large cluster" (assuming that several amino acids have the same properties) and those that returned a positive score for a "small cluster" (assuming that only few amino acids have the same properties). Note that a positive score means that the corresponding SVM has classified the data point as positive. The given score is the product of the point's distance to the hyperplane times the squared MCC confidence value of the SVM. The last part of the results page gives the alignments of the 10 residue code with the 10 residue codes of all sequences with known specificity in the database, requiring at least 7/10 identical residues. The pairwise alignments are scored with a BLOSUM62 substitution matrix [Henikoff and Henikoff, 1992; Eddy, 2004]. The **Simple Output** and **CSV Output** pages give the results in a condensed text form without the alignments. The latter output is in tabular form and notes possible inconsistencies between the SVM based prediction and the prediction by the "Stachelhaus method". The **Input page** (not shown) allows pasting in or uploading of one or more NRPS sequences in plain or FASTA format. An automatically generated Job-ID is proposed which can be changed and which allows the user to identify the job in the job tracking menu on the left-hand side. Once the job is started, an individual URL is generated for each Job-ID, which can be bookmarked to pick up the results later. The NRPSpredictor is reachable via `www-ab.informatik.uni-tuebingen.de/software/NRPSpredictor`.

Tübingen, Germany) for help with the collection of annotated A domains, Jacques Ravel (TIGR, Rockville, MD, USA) for kindly providing us with the sequences of the NRPS BLAST server, Timothy Davison (Max Planck Institute for Biological Cybernetics, Tübingen, Germany) for valuable discussion, and Andreas Biegert (University of Tübingen, Germany), who, as part of his Bachelor's thesis project, implemented the toolbox webfrontend, where the NRPSpredictor is integrated.

Chapter 5

Phylogeny, Evolution and Functional Subtypes of Condensation Domains

5.1 Overview

The peptide bond formation in NRPS is catalyzed by the Condensation (C) domain. Various functional subtypes of the C domain exist: An $^{L}C_{L}$ domain catalyzes a peptide bond between two L-amino acids, a $^{D}C_{L}$ domain links an L-amino acid to a growing peptide ending with a D-amino acid, a Starter C domain (first denominated and classified as a separate subtype here) acylates the first amino acid with a β-hydroxy-carboxylic acid (typically a β-hydroxyl fatty acid), and Heterocyclization (Cyc) domains catalyze both peptide bond formation and subsequent cyclization of cysteine, serine or threonine residues. The homologous Epimerization (E) domain flips the chirality of the last amino acid in the growing peptide; Dual E/C domains catalyze both epimerization and condensation.

In this chapter, we report on the reconstruction of the phylogenetic relationship of NRPS C domain subtypes and analyze in detail the sequence motifs of recently discovered subtypes (Dual E/C, $^{D}C_{L}$ and Starter domains) and their characteristic sequence differences, mutually and in comparison with $^{L}C_{L}$ domains. Based on their phylogeny and the comparison of their sequence motifs, $^{L}C_{L}$ and Starter domains appear to be more closely related to each other than to other subtypes, though pronounced differences in some segments of the protein account for the unequal donor substrates (amino vs. β-hydroxy-carboxylic acid). Furthermore, on the basis of phylogeny and the comparison of sequence motifs, we conclude that Dual E/C and $^{D}C_{L}$ domains share a common ancestor. In the same way, the evolutionary origin of a C domain of unknown function in glycopeptide (GP) NRPSs can be determined to be an $^{L}C_{L}$ domain. In the case of two GP C domains which are most similar to $^{D}C_{L}$ but which have $^{L}C_{L}$ activity, we postulate convergent evolution.

We systematize all C domain subtypes including the novel Starter C domain. With our results, it will be easier to decide the subtype of unknown

C domains as we provide profile Hidden Markov Models (pHMMs) for the sequence motifs as well as for the entire sequences. The determined specificity conferring positions will be helpful for the mutation of one subtype into another, e.g. turning $^{D}C_{L}$ to $^{L}C_{L}$, which can be a useful step for obtaining novel products.

5.2 Background and Motivation

As depicted in Fig. 2.1, besides the Adenylation (A) domain and the Thiolation (T) domain, the third of the three compulsory domains in NRPS is the Condensation (C) domain, which catalyzes the elongation reaction of the peptidyl chain tethered to the phosphopantetheinyl arm of the upstream T domain to the amino acid bound to the downstream T domain [reviewed by Lautru and Challis, 2004]. This is why the first module of an NRPS usually does not contain a C domain, but only the second module has the domains C–A–T. The exceptions are C domains, which we name *Starter C* domains; these acylate the first amino acid with a fatty acid (with a β-hydroxy-carboxylic acid actually, as we will discuss below). Chain elongation is terminated by the action of a thioesterase (TE) domain, which is usually the final domain of the last module in the assembly line. (For more details, refer to Chapter 2, *Biological Background*).

In this chapter, we report on the functional variants (subtypes) and homologs of the Condensation (C) domain of NRPS. All C domain sequences of this study were extracted from NRPS that were detected in all available completely sequenced bacterial genomes and a comprehensive collection of annotated biosynthesis clusters. Besides A domains (and thioesterase II domains; see Sieber and Marahiel [2005]) C domains also show specificity for their substrates (see below). An in-depth deep understanding of their function is thus crucial for re-engineering NRPS to produce novel bioactive compounds. In practice, it has been shown that it is possible to engineer synthetic systems for the production of novel products: Stachelhaus et al. [1995] demonstrated that domain swapping, which is the recombination of domain-coding regions of desired specificity to a synthetic fusion protein, worked to create new variants of surfactin and is thus one possibility, although only one amino acid position in the product was varied, which did not alter its activity, and the total yield was very low (0.5 % of the wild-type yield).

Because C domains have been shown to have non-negligible specificity for the amino acid that is activated by the downstream A domain, swapping whole modules or insertion/deletion seems to be more promising, provided that the integrity of the functional domains is carefully maintained and the modules are dissected in their linker regions [Mootz et al., 2000, 2002a]. Nevertheless, reduced catalytic efficiency and product yield is a serious problem. A less invasive strategy involves the manipulation of the domains' specificity by point mutations as demonstrated by Eppelmann et al. [2002] for the A domain. Therefore, an in-depth knowledge of all functional subtypes and homologs of the C domains is indispensable. In this chapter, we reconstruct their phylogeny and reveal the sequence motifs of all subtypes and homologs,

and their mutual differences. The insights gained will be helpful in future attempts to turn one sub-specificity into another, e.g. changing the stereoselectivity of the C domain.

Furthermore, we have analyzed the C domains and Epimerization (E) domains of glycopeptide NRPS. In these proteins, two Condensation domains preceded by former (now inactive) Epimerization domains have gained opposite stereoselectivity, probably due to convergent evolution, for which we accumulate evidence. Additionally, we discuss the origin of a C domain (often referred to as an X* domain) at the C-terminus of glycopeptide NRPS, which is thought to be inactive.

Current Knowledge of Subtypes $^{L}C_{L}$, $^{D}C_{L}$, Cyc, and Dual E/C

The C domain has two binding sites: one for the electrophilic donor substrate (the acyl group of the growing chain) and one for the nucleophilic acceptor substrate (the activated amino acid). The condensation reaction involves catalysis of a nucleophilic attack by the amino group of the aminoacyl adenylate bound to the downstream T domain on the acyl group of the growing peptide chain which is bound to the upstream T domain [Finking and Marahiel, 2004; Sieber and Marahiel, 2005; see Fig. 2.3]. The acceptor site of the C domain was shown to exhibit a strong stereoselectivity and significant side chain selectivity. The selectivity towards a specific side chain seems to be less pronounced at the donor site which, however, exhibits strong stereoselectivity [Lautru and Challis, 2004].

In particular, C domains succeeding an E domain are expected to show specificity towards the configuration (L or D) of the C-terminal residue that is bound at the donor site because the preceding E domain does not specifically catalyze the epimerization from L to D but provides a mixture of configurations. It is the role of the C domain to select the correct enantiomer [Finking and Marahiel, 2004]. Moreover, the C domain represents some kind of selectivity filter in that it supports the selection of the correct downstream nucleophile and prevents product mixtures [Sieber and Marahiel, 2005].

C domains immediately downstream of E domains were shown to be D-specific for the upstream donor and L-specific for the downstream acceptor, thus catalyzing the condensation reaction between a D- and an L-residue. These C domains were termed $^{D}C_{L}$-catalysts because of this behavior [Clugston et al., 2003].

Accordingly, $^{L}C_{L}$-catalysts promote the condensation of two L-amino acids. Both $^{L}C_{L}$- and $^{D}C_{L}$-catalysts possess a conserved His-motif in their active site. The consensus sequence of this motif is HHxxxDG where x denotes any residue (see Fig. 5.1, motif 3, or the magnification in Fig. 5.3). The second His-residue seems to be essential for the catalytic function of the domain [Sieber and Marahiel, 2005].

As a third type of C domain, so-called Dual Epimerization/Condensation (E/C) domains have recently been identified. This finding was based on the observation of NRPS which had products that contained D-residues although

the NRPS itself did not show an E domain in the corresponding module. Biochemical experiments supported the hypothesis that Dual E/C domains exist which are $^{D}C_{L}$-catalysts with epimerase activity [Balibar et al., 2005]. In the assembly line, a Dual E/C domain follows directly after a C-A-T module which activates and incorporates an L-amino acid. The module which contains the Dual domain also activates an L-amino acid. Then the Dual domain catalyzes the epimerization of the L-residue into D configuration and subsequently promotes the condensation of those two residues. In addition to the active site His-motif which is found in all C domains, Dual E/C domains exhibit a second His-motif, HHxxxxxGD, which is located close to the N-terminus of the domain [Balibar et al., 2005] (It is partly located on motifs C1 & C2; see Fig. 5.1 or for a magnification Fig. 5.4).

C domains may be replaced by Heterocyclization (Cyc) domains which catalyze both peptide bond formation, and subsequent cyclization of cysteine (Cys), serine (Ser) and threonine (Thr) residues. The five-membered heterocyclic rings which result from this reaction are important for chelating metals or interaction with proteins, DNA or RNA. Cyc domains are structurally related to C domains and are supposed to be evolutionary specialized C domains [Sieber and Marahiel, 2005]. In Cyc domains, however, the active site His-motif is replaced by another conserved motif, DxxxxD (see Fig. 5.3). Keating et al. [2002] found that the aspartate (Asp, D) residues are critical for both condensation and heterocyclization.

5.3 Results and Discussion

5.3.1 Collected C Domain Sequence Data and Their Phylogenetic Tree

A total of 481 Condensation domains (including their homologs, Epimerazation and Heterocyclization domains) were extracted from 182 (non-identical) NRPS and 31 NRPS/PKS hybrid sequences found in 62 bacterial genomes out of the 256 bacterial genomes screened, employing pHMMs as described in Section 5.5, *Materials and Methods* (Note that only one genome was considered for our analysis if sequences of several strains of the same species were available, which reduced the number of genomes containing NRPS or 'hybrid NRPS/PKS' from 62 to 43). Altogether, 108 C domains were obtained from 42 NRPS sequences from gene clusters downloaded from the UniProt database. After removing doublets, all 525 non-identical C domains and homologs obtained were multiply aligned and phylogenetic trees were built. The resulting tree topology was clearly dominated by the functional categories that are known for C domains (as described in the previous section), rather than species phylogeny or substrate specificity alone. The four main functions are: *1.* condensation performed by ordinary C domains; *2.* condensation and subsequent heterocyclization catalyzed by Heterocyclization (Cyc) domains; *3.* epimerization followed by condensation, both of which are catalyzed by a Dual E/C domain; *4.* Starter domains (see below) which are found on initiation (= first) modules and acylate the subsequent amino acid.

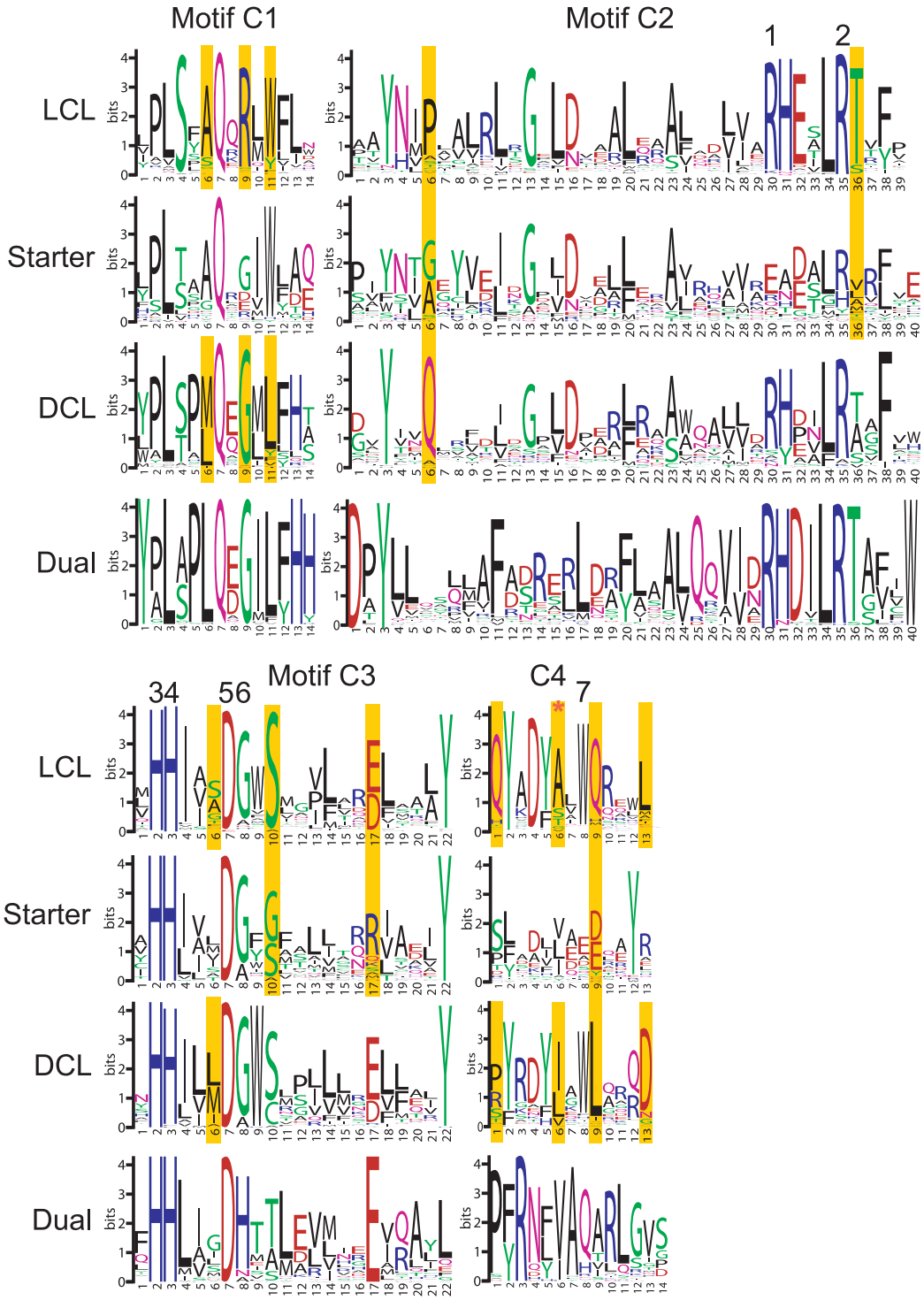

Figure 5.1: Core motifs C1 through C3 of C domain subtypes $^{L}C_{L}$, Starter, $^{D}C_{L}$ and Dual E/C domains. Compared to Marahiel et al. [1997], motifs are extended in both directions to include more significantly conserved positions. Yellow bars indicate significant specificity determining positions between $^{L}C_{L}$, Starter and $^{D}C_{L}$ domains; those with red stars on top are the most significant positions. Numbers above the letter stacks indicate residues of functional and structural importance referred to in Subsection 5.3.4 "Key Residues in Condensation Domains" and Table 5.1. Motifs C4 through C7 are shown in Fig. 5.2.

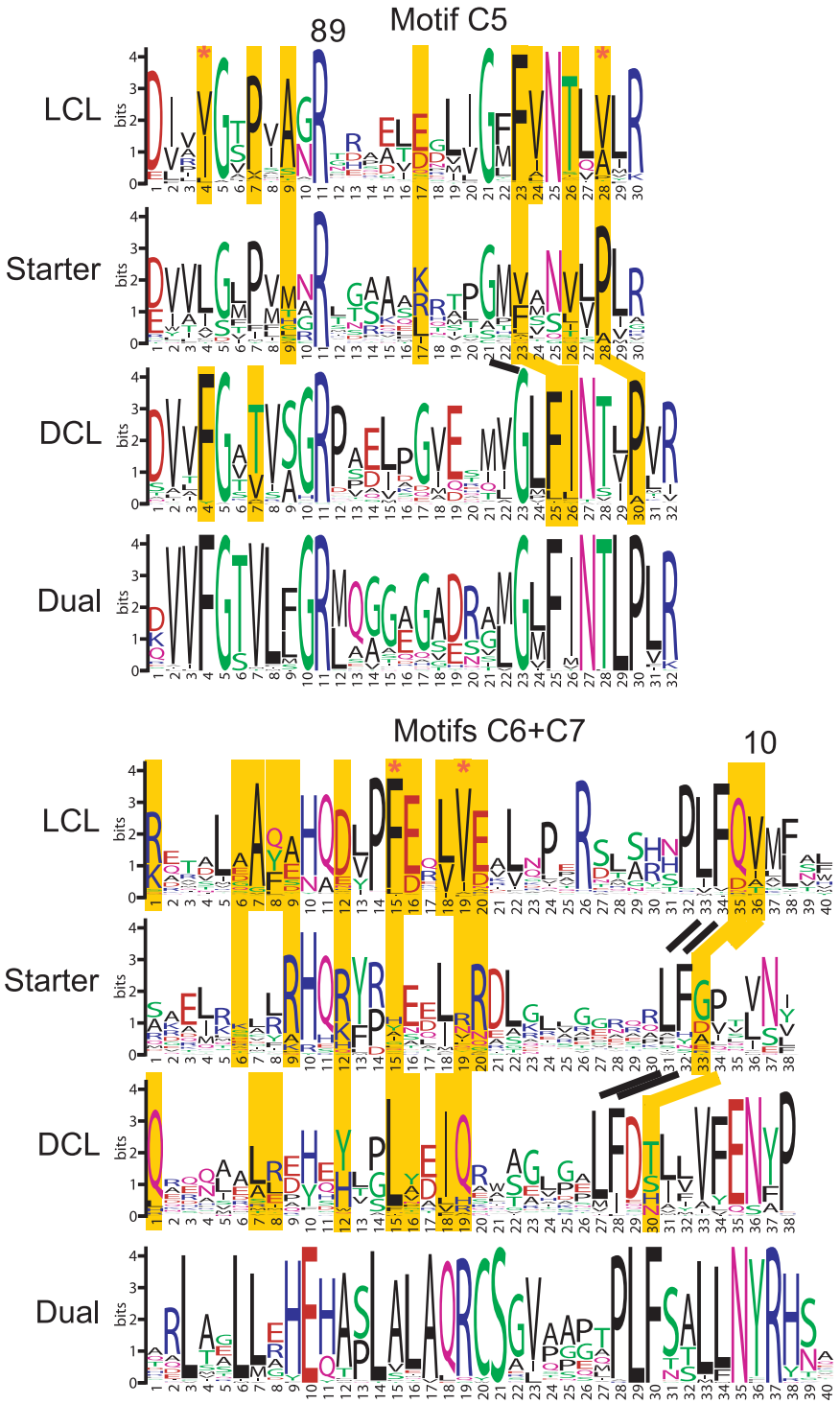

Figure 5.2: Core motifs C4 through C7 of C domain subtypes $^{L}C_{L}$, Starter, $^{D}C_{L}$ and Dual E/C domains. Also see Fig. 5.1.

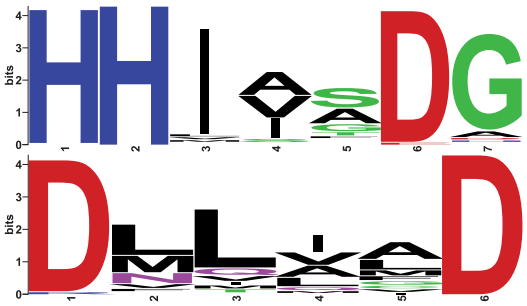

Figure 5.3: Sequence logos representing the PSSMs which were constructed for the active site motif C3 (starting ∼ at pos. 133) of the $^{L}C_{L}$ Condensation domain (top, based on 238 sequences) and the Heterocyclization domain (bottom, based on 45 sequences).

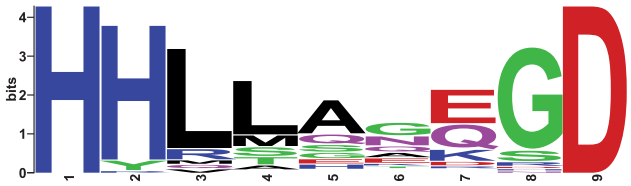

Figure 5.4: Sequence logo representation of the PSSM which was generated for the N-terminal His-motif found in Dual C/E domains at pos. ∼ 16 (based on 56 sequences).

Ordinary C domains may further be classified into $^{L}C_{L}$-catalysts and $^{D}C_{L}$-catalysts according to the stereochemistry of their substrates. The existence of all these functional subtypes is reflected by the phylogeny. Fig. 5.5 shows a phylogenetic tree for subsets of each C domain subtype, as the whole tree of 525 taxa is far too large to be displayed here (see Supplementary files 5.1 and 5.2). The tree of all taxa showed a similar topology, perfectly reflecting the functional categories.

5.3.2 Description of a New C Domain Subtype: The Starter C Domain

When analyzing the Condensation (C) domain phylogeny, it became apparent that some domains did not cluster with the known C domain subtypes. A closer look at the location of these deviating C domains revealed that all of them were the very first C domain of the corresponding NRPS assembly line. The remaining C domains of these assembly lines appeared in other subtrees in the phylogeny.

Included in this set of starter C domains are those stemming from the biosynthesis clusters for the lipopeptides surfactin [Arima et al., 1968], lichenysin [Horowitz and Griffin, 1991], fengycin [Tosato et al., 1997] and arthrofactin [Morikawa et al., 1993]. These lipopeptides are characterized by a β-hydroxyl fatty acid which is connected to the first amino acid of the peptide chain [Konz et al., 1999].

The peptide synthetases involved in the production of these lipopeptides all have a C domain as their very first domain. This C domain is supposed

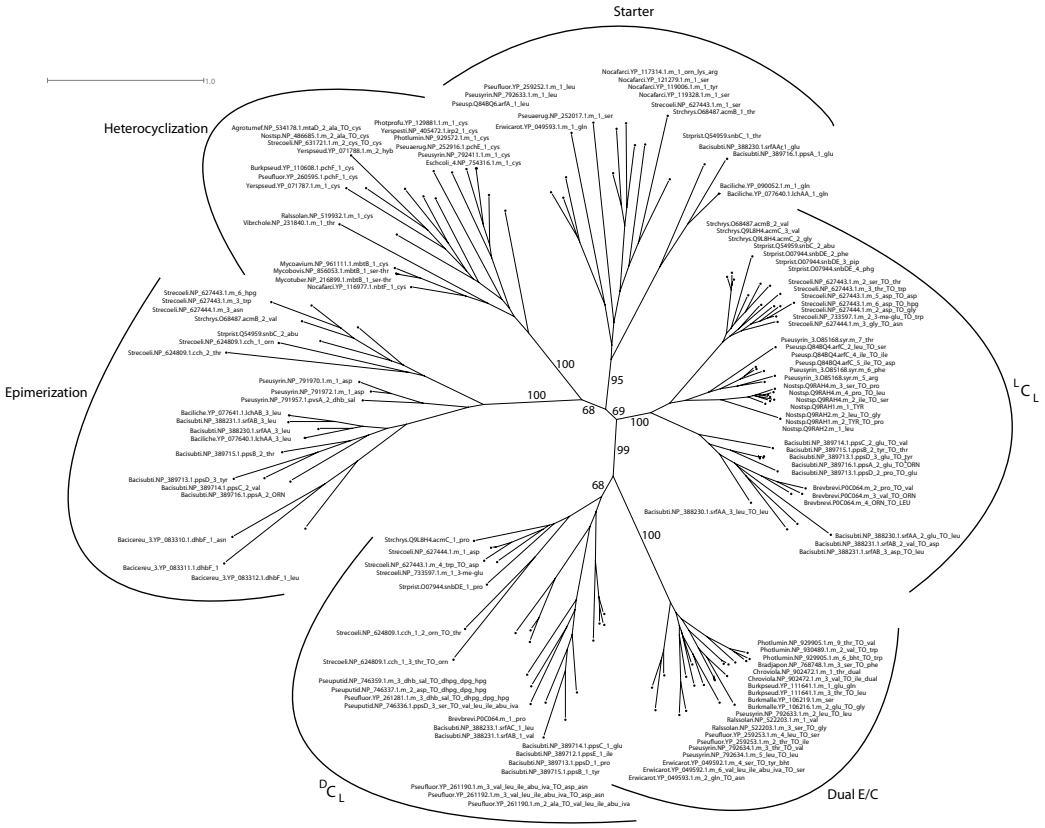

Figure 5.5: Phylogenetic tree of all C subtypes ($^{L}C_{L}$, $^{D}C_{L}$, Starter, Dual E/C, Epimerization and Heterocyclization domains). The phylogeny was reconstructed using `phyml`, employing the JTT model of amino acid substitution and a gamma-distributed rate variation with four categories. The support values are based on 100-fold bootstrapping. For further analysis, the different subtypes were examined separately. While Cyc and Dual E/C domains could be identified by means of their characteristic sequence motifs (see Section Materials and Methods/Predicting Functional Subtypes), $^{L}C_{L}$- and $^{D}C_{L}$-catalysts were either distinguished according to their domain structure or by their position in the phylogenetic tree. By this, 275 domains of all 525 C domains were classified as being $^{L}C_{L}$-catalysts, 69 were $^{D}C_{L}$-catalysts and 42 were Starter C domains (see next section).

to serve as an acceptor for a fatty acid which is transferred from an acyltransferase [Konz et al., 1999]. This acylation process has also been observed for surfactin [Cosmina et al., 1993] and fengycin biosynthesis [Tognoni et al., 1995]. Moreover, common to the Starter C domains of these biosynthesis clusters is their low sequence similarity to the remaining C domains of the same biosynthesis cluster [Konz et al., 1999].

The same has been observed for the synthesis of the acidic lipopeptide CDA in *Streptomyces coelicolor* [Hojati et al., 2002] and the recently identified lipopeptide produced by protein NP_960354.1 of *Mycobacterium avium* [Eckstein et al., 2006].

The Starter C domain of the pristinamycin cluster appears to diverge from this pattern at the first view. The C domain is the first domain of the polypeptide SnbC but the biosynthesis of pristinamycin is initiated by SnbA, which contains an A domain that activates 3-hydroxypicolinic acid (3-hydroxypyridine-2-carboxylic acid, "2-hydroxy-6-azabenzoate") but lacks a T domain [de Crécy-Lagard et al., 1997]. SnbA is homologous to EntE, which contains an A domain specific for 2,3-dihydroxybenzoate (DHB) and which is involved in the biosynthesis of enterobactin [Rusnak et al., 1989]. A similar organization can be found in actinomycin biosynthesis. The process is initiated by AcmA, which activates 4-methyl-3-hydroxyanthranilic acid (MHA, 4-methyl-3-hydroxy-2-aminobenzoate) [Schauwecker et al., 1998]. In conclusion, what the C domains of SnbC, AcmB and EntF have in common is that they catalyze bond formation between a derivative of salicylic acid (2-hydroxy-benzoate) and an α-amino acid. Assured by the fact that these Starter C domains match significantly well with the pHMM built from the Starter C domain sequences that process β-hydroxy fatty acids, we compared salicylic acid with β-hydroxy fatty acids. Because both are β-hydroxycarboxylic acids with no amino-substituent at the α position, as α-amino acids would have, we assume that this is the structural characteristic recognized by the prototype of Starter C domains. The pHMM built from all Starter C domains in our dataset (together with the pHMMs of the other domains) presents a powerful instrument for exploring and understanding tricky NRPS domain-product relations (for references to the files, refer to the Supplementary Data Section (5.7).

Note that Formylation domains as found, for example, at the N-terminus of linear gramicidin synthetase subunit A [Schönafinger et al., 2006] are not C domains but belong to the Pfam "formyl transferase" domain family.

5.3.3 Characteristic Sequence Motifs of $^{L}C_{L}$, $^{D}C_{L}$, Starter C Domains and Dual E/C Domains

The different core motifs in Condensation domains have first been described by de Crécy-Lagard et al. [1995] and recompiled by Marahiel et al. [1997] but have never been updated since then. The core motifs of the C domain homologs, the Epimerization and Heterocyclization domains are listed in the publication by Marahiel et al. [1997] but the sequence motifs of the recently discovered $^{D}C_{L}$ domains [Clugston et al., 2003; Luo et al., 2002] as well as the

Dual E/C domains [Balibar et al., 2005] have never been comprehensively analyzed. Moreover the Starter C domain has not yet been recognized in the literature as a proper separate subtype.

The sequence motifs represented in Figures 5.1 and 5.2 are an improvement on the C domain core motif consensus sequences published by Marahiel et al. [1997] which, at that time, were based on much fewer sequences and did not differentiate between the C domain subtypes. The motifs are represented as sequence logos [Crooks et al., 2004] which make it easier to identify variably conserved positions compared to simple consensus sequences. We adhere to the core motifs identified by Marahiel et al. [1997], and also show the surrounding "landscape" if there are highly conserved positions nearby, especially if they are important for distinguishing between the C domain subtypes. The motifs were built on the basis of 40 verified and 198 predicted $^{L}C_{L}$ sequences, in which "predicted" means that they were classified based purely on their position in the phylogenetic tree while "verified" sequences were checked individually, taking into account their position in the succession of neighboring NRPS domains, the presence of discriminative unique motifs (see the *Materials and Methods* Section in 5.5) and/or literature information. For the $^{D}C_{L}$ motifs, 23 verified and 46 predicted sequences were used; 7 verified and 35 predicted sequences were used for the Starter domains; 9 verified and 47 predicted sequences were used for the Dual E/C domains.

5.3.4 Key Residues in Condensation Domains Derived from the Literature

Based on three publications, four residues are likely to be essential for the catalytic activity of the C domain. The most important residue is the 2nd His of the active site His-motif [Stachelhaus et al., 1998]. Furthermore, six residues have been identified as being structurally important or as playing a role in correct folding of the domain. In the following, these residues are presented, grouped by their role; *the numbering is according to their linear occurrence on the peptide; see Figures 5.1 and 5.2.* This information is also presented in Table 5.1 where the sites are sorted by their relative position in the domain.

Residues Important for *Correct Folding*

#2 Arg67 (R) in TycB1 [Bergendahl et al., 2002]

#3 His146 in TycB1 (1st His of active site His-motif) [Bergendahl et al., 2002]

#7 Trp202 (W) in TycB1 [Bergendahl et al., 2002]

Residues of Importance for *Catalytic* Activity of the Domain

#4 His 126 (2nd His of the active site His-motif) in VibH [Keating et al., 2002; Roche and Walsh, 2003; Bergendahl et al., 2002]

#9 Trp264 (W) is catalytically important in VibH according to Keating et al. [2002], but the corresponding position is not conserved in any of the C domain subtypes $^{L}C_{L}$, $^{D}C_{L}$ or Starter.

#10 Asn335 (N) in VibH [Roche and Walsh, 2003]

#6 Gly131 (G of the active site His-motif) in VibH [Roche and Walsh, 2003]

Residues of *Structural* Importance

#1 Arg62 (R) in TycB1 [Bergendahl et al., 2002]

#5 Asp130 (D) in VibH [Keating et al., 2002; Roche and Walsh, 2003; Bergendahl et al., 2002]

#8 Arg263 (R) in VibH [Keating et al., 2002] = Arg278 (R) in EntF [Roche and Walsh, 2003]

No. in Fig. 5.1 & 5.2	Importance:	Position is homologous to:
1	structure	Arg62 (R) in TycB1 [Bergendahl et al., 2002]
2	folding	Arg67 (R) in TycB1 [Bergendahl et al., 2002]
3	folding	His146 in TycB1 (1st His of active site His-motif) [Bergendahl et al., 2002]
4	catalytic activity	His126 (2nd His of the active site His-motif) in VibH [Keating et al., 2002; Roche and Walsh, 2003; Bergendahl et al., 2002]
5	structure	Asp130 (D) in VibH [Keating et al., 2002; Roche and Walsh, 2003; Bergendahl et al., 2002]
6	catalytic activity	Gly131 (G of the active site His-motif) in VibH [Roche and Walsh, 2003]
7	folding	Trp202 (W) in TycB1 [Bergendahl et al., 2002]
8	structure	Arg263 (R) in VibH = Arg278 (R) in EntF [Keating et al., 2002; Roche and Walsh, 2003]
9	catalytic activity	Trp264 (W) in VibH according to Keating et al. [2002], but absent in $^{L}C_{L}$, $^{D}C_{L}$ and Starter C domains
10	catalytic activity	Asn335 (N) in VibH [Roche and Walsh, 2003]

Table 5.1: Residues of importance for catalytic activity, structure or correct folding. Residues for which the importance has been previously determined are listed, giving their numbers, their role and the bibliographic reference of the appropriate mutation study. The numbering is according to the numbering in Figures 5.1 and 5.2.

5.3.5 $^{L}C_{L}$ vs. $^{D}C_{L}$

$^{L}C_{L}$ and $^{D}C_{L}$ domains do not differ significantly in any of the residues identified as being of catalytic or structural importance (except residues No. 9 and No. 10). However, using methods described in Section 5.5, *Materials and Methods*, on page 80, 20 positions in which $^{L}C_{L}$ and $^{D}C_{L}$ have significant differences according to SDPpred [Kalinina et al., 2004] could be detected, plus 5 additional high scoring positions within the extended motifs according to FRpred [Fischer et al., 2006]. When comparing the different motifs, motif C4 differs noticeably between $^{L}C_{L}$ and $^{D}C_{L}$ subtypes. The same is true for the region downstream of C4 (after the mutually very conserved TRP at pos. 184 in VibH coordinates) where a moderately conserved motif LPxDxxRP is seen in $^{L}C_{L}$ which is completely absent in $^{D}C_{L}$ (see Supplementary file 5.3).

5.3.6 $^{L}C_{L}$ vs. Starter domain

While not being conserved at residues No. 5, No. 7, No. 9, and No. 10, all remaining 6 functionally important residues are highly conserved throughout the putative Starter domains. When comparing $^{L}C_{L}$ and Starter domains, 18 discriminative positions were found by SDPpred and 5 more were found in the motifs by FRpred. Those positions are highlighted in Figures 5.1 and 5.2. Common to these residues is the fact that they seem to be highly conserved among extender ($=^{L}C_{L}$) domains but show no conservation among Starter C domains. When we compare C domain sequence motifs, it is apparent that motifs C2 and C4, despite being well conserved in $^{L}C_{L}$, are unconserved in Starter domains, which presumably can be explained by the much broader structural range of substrates processed by Starter domains.

5.3.7 What the Phylogeny Tells Us about the Relationship of $^{D}C_{L}$ vs. Dual E/C and $^{L}C_{L}$ vs. Starter Domains

The phylogenetic trees in Figures 5.5 and 5.6 and the tree on 525 taxa in the Supplementary file 5.2 show that Dual E/C and $^{D}C_{L}$ domains share a common ancestor. From the phylogenetic trees it also appears that the closest common ancestor of $^{L}C_{L}$ domains and the closest common ancestor of Starter C domains are more closely related to each other than to other subtypes, based on the distances in the trees. However, these observations are not supported by an edge with high bootstrap value separating the $^{L}C_{L}$/Starter C domain subtrees from the other subtrees. On the other hand, comparing sequence motifs confirms this assumption, though pronounced differences in some segments of the protein (especially in motifs C2 and C3, as can be seen in Fig. 5.1) account for the unequal donor substrates (amino vs. β-hydroxycarboxylic acid).

To get more certainty on the assumed relation of $^{D}C_{L}$ versus Dual E/C domains and $^{L}C_{L}$ versus Starter domains, we tested the reliability of the phylogenies depicted in Fig. 5.5 and Fig. 5.6 by repeating the reconstruction on biased profile alignments. These biased alignments were generated by

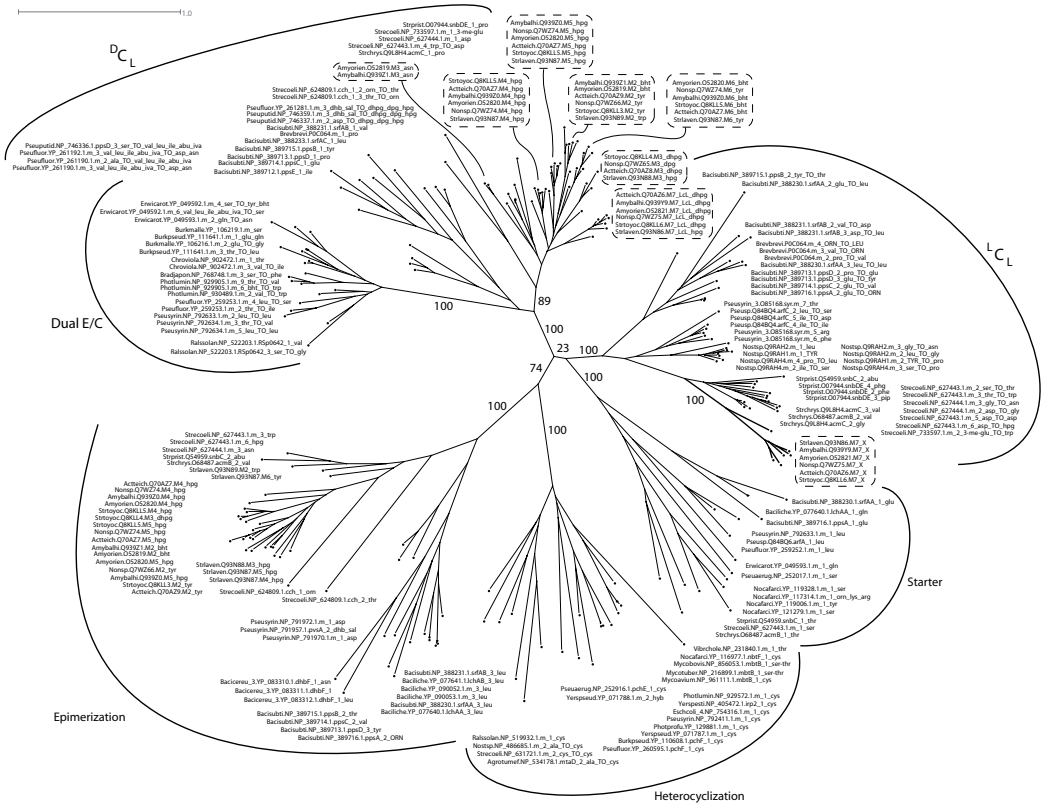

Figure 5.6: Phylogenetic trees of all C subtypes ($^{L}C_{L}$, $^{D}C_{L}$, Starter, Dual E/C, Epimerization and Heterocyclization domains). Compared to Fig. 5.5, this tree additionally includes all C domains of glycopeptide antibiotic biosynthesis clusters (in dashed boxes in the upper part of the tree).

producing MUSCLE profile-profile alignments in a step-wise manner, assuming evolutionary relationships of the different domain subtypes that are contradictory to what the original trees suggest. The topology of the resulting trees still supports the shared ancestry of $^{L}C_{L}$ and Starter C domains as well as of Dual E/C and $^{D}C_{L}$ domains. In addition, we generated an alignment using DIALIGN [Morgenstern, 1999], which is a non-progressive alignment method, and subsequently reconstructed a PHYML-tree based on this alignment. Here also, the Dual E/C and $^{D}C_{L}$ domains are grouped together, as are $^{L}C_{L}$ and Starter C domains.

Especially in motif C5, Dual E/C and $^{D}C_{L}$ domains are very similar to each other and dissimilar to $^{L}C_{L}$ and Starter domains. This observation of the relationship between the four subtypes is consistent with the stereochemistry of the substrates, bearing in mind that Dual E/C domains function as $^{D}C_{L}$ because the substrate L-amino acid is first epimerized by the intrinsic epimerization activity of the domain [Balibar et al., 2005].

Within the subtrees of $^{D}C_{L}$ and $^{L}C_{L}$ domains, the tree topology reflects the species phylogeny of the bacteria rather than substrate specificity of any kind. We analyzed this by reconstructing phylogenies for $^{D}C_{L}$ domains and $^{L}C_{L}$ domains separately to be able to see the topology within these subtypes in more detail (data not shown). The reconstructed phylogenies did not give any evidence that would support the hypothesis that C domains

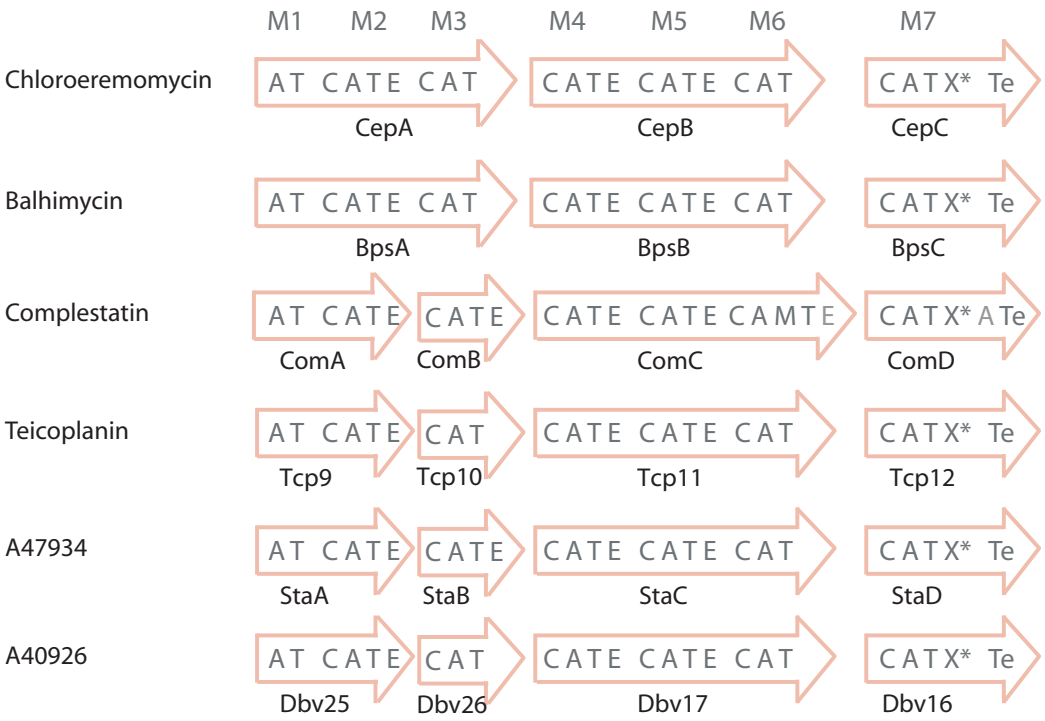

Figure 5.7: Modular organization of NRPSs involved in glycopeptide synthesis. Domains marked in light gray (Completstatin) are inactive and corrupt. Moreover, E domains in ComB and StaB are also thought to be inactive.

cluster according to their specificity towards the condensated amino acids. This analysis, however, is based on the complete C domain sequence. A strategy to investigate whether C domains exhibit substrate specificity would involve predicting putative specificity determining positions using entropy and/or conservation based approaches (e.g. SDPpred, FRpred), or inferring of putative active site residues by homology with the VibH structure (as done by Rausch et al. [2005] for the Adenylation domain).

5.3.8 Enigmatic NRPSs of Glycopeptide Antibiotics

Glycopeptide antibiotics are a subgroup of nonribosomal peptide antibiotics of which the best known representatives are probably vancomycin and teicoplanin. To date, all identified glycopeptide antibiotics are produced by actinomycetes. They interrupt cell wall formation of gram-positive bacteria by binding to the D-Ala-D-Ala termini of the growing peptidoglycan, thereby inhibiting the transpeptidation reaction. All glycopeptide antibiotics consist of a heptapeptide backbone which is synthesized by NRPS. Modification reactions involve extensive cross-linking of the aromatic side chains to rigidify the molecule [Bischoff et al., 2001a,b]. The modular organization of some NRPSs which were identified in glycopeptide-producing actinomycetes are depicted in Fig. 5.7.

All these NRPSs comprise seven modules. They show an identical domain composition, with the exceptions of module M3 in the A47934 (*sta*), and M3 and M6 in complestatin (*com*) clusters which contain an E domain

not present in the other clusters. The M3-E domain, however, is assumed to be inactive [Sosio et al., 2003], while the presence of an E domain in *com* M6 has not been reported elsewhere so far. We were able to detect it with an `hmmpfam` scan using our specific E domain pHMM (available as Supplementary file 5.5). All six NRPSs contain a domain X* of unknown function. Until now, it has been characterized as an atypical C or E domain but its role in glycopeptide synthesis remains to be clarified. In general, it is assumed that the stereochemistry of an NRPS product can be predicted from its domain structure. In the case of the known glycopeptides, the domain organization implies the stereochemistry NH_2-L-D-L-D-D-L-L-COOH, provided that the E in module M3 is inactive and that the X* domain does not function as an E domain. This stereochemistry is inconsistent with the chemically determined structure of the products: NH_2-D-D-L-D-D-L-L-COOH [Sosio et al., 2003]. The assumption is that the A domain of the first module activates a D-amino acid. For the *cep* cluster, however, Trauger and Walsh [2000] showed that the A domain of M1 prefers L-Leu over D-Leu in a 6:1 ratio; but on the other hand, they could not show which stereoisomer is processed further. This suggests the existence of an unknown E domain that acts on the L-Leu activated by M1. With the discovery of Dual E/C domains, a new possible strategy arises for the incorporation of a D-residue by the first module. However, no Dual E/C domain could be detected in any of the available glyco-NRPSs. Alternatively, one could imagine an external racemase as is found in the cyclosporin cluster [Hoffmann et al., 1994], which provides a D-Leu that can be incorporated directly.

Having gained knowledge about the differences between $^{L}C_{L}$, Starter and $^{D}C_{L}$ domains as described above, we examined all glyco-NRPSs. When we reconstructed the phylogeny of C domains including all homologous domains from glyco-NRPSs, it was staggering to find that all C domains were clustered in the $^{D}C_{L}$ subtree and the X* domain clustered in the $^{L}C_{L}$ subtree (see Fig. 5.6). This finding could be confirmed by analyzing all instances of the C domain motifs found in these domains. How could this be interpreted, given the fact that M4 and M7 C domains clearly act as $^{L}C_{L}$ domains, as we can tell by the stereochemistry of the products? Our hypothesis is that those C domains are former $^{D}C_{L}$ domains that have developed $^{L}C_{L}$ activity by convergent evolution. Accumulating supportive evidence is possible: When we look at the phylogeny of the C domains, the sequences of the *com* cluster from *Streptomyces lavendulae* are always most distant from the others and more closely related to the hypothetical common ancestor, implying that they can serve as a model for the archetype of glyco-C domains. It is likely that in the archetype, all C domains were true $^{D}C_{L}$ catalysts, supposing that the E domains which are still present in *com* modules M4 and M7 were still active.

In a similar way, we can trace back the origin of the X* domain: in the *com* cluster (and only there) it is followed by remnants of an Adenylation domain (which has several larger insertions and deletions; see Supplementary file 5.4). This tells us that the X* domain used to be the first domain of a new module followed by an Adenylation domain.

The assumption that the diverged C domains of modules M4 and M7

would have adopted mutations at positions that we have previously determined as "specificity determining positions" was disproved. Probably, a few spontaneous mutations in the $^{D}C_{L}$ domains relaxed the stereo-selectivity; supposing that this altered stereochemistry of the product resulted in a highly selective advantage (arising from a vancomycin-like product), the loss of the functional E domains in M3 and M6 would have been a selective gain. Comparing all M4 and/or all M7 C domains with all $^{D}C_{L}$ domains using SDPpred did not reveal any significant positions; comparing them against the other glyco-C domains gave thirty positions. As all glyco-C domains are very closely related and differences between them might also reflect substrate selectivity (not only stereo-selectivity) or different inter-domain interacting residues, we cannot decide which of them confers the altered stereo-selectivity. One point to notice, however, is a (positively charged) His in all M4 glyco-C domains at position 6 in the extended motif C2 where an (uncharged polar) Gln is highly conserved in other $^{D}C_{L}$ domains. This position has also been selected by `FRpred` as a significant (=subtyping) position. The other positions do not represent mutations in highly conserved residues (data not shown). It would be necessary to check their significance experimentally with mutation studies. It would also be helpful to compare the aberrant sequences with more glyco-C domains, but others are – unfortunately – not publicly available.

However, although we could not discover which altered positions are responsible for the functional shift from $^{D}C_{L}$ to $^{L}C_{L}$ in glyco-C domains, interesting experimental questions can be formulated based on our findings. For example, one could think of mutational studies with the goal of altering the stereo-selectivity of a $^{D}C_{L}$ domain and to determine the relevant residues experimentally. A starting point could be, for example, the M6 C domain of any glyco-NRPS. This C domain is the last to incorporate a D-amino acid and is preceded by an active E domain. Its position towards the end of the assembly line might be an advantage in the attempt to turn its function into a $^{L}C_{L}$ domain (together with the inactivation of the preceding E domain), because the altered product would only need to be processed by one further module. Thus, the risk of a lowered efficiency (lowered yield) of the assembly line would be reduced, because less domains would be in contact with the structurally modified product.

5.3.9 Glycopeptide-AB Module M7 vs. $^{L}C_{L}$

The second His of the His-motif in motif C3 which is important for catalysis is replaced by Arg (R). Also, the Gly of the His-motif is not present but is replaced by Arg in all but one X* domain. Note, however, that while the second active site His is invariant in C domains, Gly138 is not.

SDPpred predicted 13 specificity determining residues when comparing M7-X* to $^{L}C_{L}$-domains of *Streptomyces* species. Only three of these coincide with residues of functional importance: His126, Arg278 and Asn335. Furthermore, a C terminal region could be detected in which M7-X* and $^{L}C_{L}$ differ strikingly.

The concordance of M7-X* with the most highly conserved residues of Streptomycete $^{L}C_{L}$ domains supports the phylogenetically based suggestion that M7-X* is an inactive $^{L}C_{L}$ domain.

5.4 Conclusion

In this chapter, we present the evolutionary relationship of homologs of the NRPS Condensation domain, which include enzymatic domains catalyzing Epimerization, Heterocyclization, Condensation and Epimerization with subsequent Condensation in one domain (called the Dual E/C domain). The Condensation domain itself appears in three subtypes according to the stereochemistry of the substrates catalyzed: $^{L}C_{L}$ domains, which condense two L-aminoacids, $^{D}C_{L}$ domains, which condense a D-amino acid (N-terminal part of the growing peptide) with an L-amino acid, and Starter C domains (an expression that we coin here) which connect a β-hydroxy-carboxylic acid (e.g. β-hydroxyl fatty acid) with an L-amino acid. The phylogeny of C domain homologs is reconstructed using NRPS sequences (including hybrid NRPS) from completely sequenced genomes (43 genomes contained NRPSs) and selected biosynthesis clusters, involving 525 non-identical C domain sequences. The sequence motifs of $^{L}C_{L}$, $^{D}C_{L}$ and Starter domains have been extracted and are presented as sequence logos: for $^{L}C_{L}$ domains, this represents an update of consensus sequences published by Marahiel et al. [1997]; $^{D}C_{L}$ and Starter domain motifs are analyzed and mutually compared for the first time. For comparison, we also present the homologous motifs for Dual E/C domains, which were first described by Balibar et al. [2005].

We have investigated the "mysterious" evolutionary origin of C domains in glycopeptide antibiotic synthesis clusters and have discovered that two of the six C domains present in these glyco-NRPSs appear in the $^{D}C_{L}$ subtree of the phylogenetic tree and show all $^{D}C_{L}$ sequence motifs, although they clearly have $^{L}C_{L}$ activity. This suggests that they might be an example of convergent evolution. Even though this is probably a rare event, its possibility has to be kept in mind when uncharacterized C domains are to be classified, e.g. using pHMMs provided as Supplementary files 5.5-5.7 (see Section 5.7). Furthermore, we found that a C domain-like segment of glyco-NRPS, called X*, is related to the $^{L}C_{L}$ domains and is followed by remnants of an A domain, implying an additional complete module in the ancestor of glyco-NRPS.

Roongsawang et al. [2005] have already performed a study of the phylogeny of C domains which compares the three C domain subtypes. However, this study shows no awareness of the Dual E/C domain, which has since been discovered. Moreover, we used a much more comprehensive dataset of C domain subsequences (525, as opposed to Roongasawang et al.'s 162) compiled from all complete bacterial genomes and biosynthesis clusters. Because Roongasawang et al. omitted Dual E/C domains, their conclusions need to be revised, as we have shown.

5.5 Materials and Methods

5.5.1 Genomes and Sequences

The protein sequences and GenBank entries for all completely sequenced bacterial genomes available to date were obtained from the NCBI FTP site [`ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria`]. In total, the genomes of 256 bacterial species were downloaded and screened for NRPS protein sequences (including NRPS/PKS hybrids). Additional protein sequences of PKS and NRPS which are part of known secondary metabolite biosynthesis clusters were obtained from the UniProt database [Wu et al., 2006]. NRPSs were retrieved from 14 known biosynthesis clusters, of which 13 came from *Actinomycetes* and one from *Pseudomonas* (see Supplementary file 5.8).

5.5.2 Identification NRPSs in Protein Databases and Extraction of Their Enzymatic Domains

We employed the strategy of searching for the concurrent occurrence of several profile Hidden Markov Models to gather the multidomain protein NRPSs and NRPS/PKS hybrids from the protein databases and to extract their enzymatic domains (see *Technical Background*, Section 3.3.2 on page 41 for details).

To identify a protein sequence as an NRPS, the occurrence of at least one complete NRPS module with one C domain, one A domain and one T domain was required (Pfam accession numbers PF00668, PF00501 and PF00550), with an E-value threshold of 0.1 (thus we accepted missing free-standing starter modules containing only A and T domains, or had to add free-standing starter modules manually, as in the case of the biosynthesis clusters).

The Pfam pHMM Condensation (PF00668) recognizes both the Condensation (C) and Epimerization (E) domains of NRPS. The intention, however, is to be able to distinguish between these two domain types. Therefore C domain and E domain specific pHMMs were generated from a multiple sequence alignment (MSA) of Epimerization domains and non-Epimerization domains, both of which were recognized by the Pfam C pHMM. To obtain a set of Epimerization domains, all NRPS sequences with complete modules were extracted from all bacterial protein sequences in the Uniprot database [Wu et al., 2006] as described above. Whenever two consecutive C domains followed by an A domain were detected with Pfam pHMMs, the "first C" domain was extracted. That way, we obtained a set consisting mainly of E domains (151 of 237 sequences). By phylogenetic subtyping (as described below), we determined the E domain sequences from the phylogenetic tree of the "first C" domains, which were forming a distinct subtree. The E and non-E sequences were aligned with `MUSCLE` [Edgar, 2004a,b], and specific pHMMs were build for them with `hmmbuild` and `hmmcalibrate` from the HMMER package (As a control, it was not possible to detect E domains in the 771 "second C" domains). The domain sequence covered by our own pHMMs for C and E domains is identical with that of the Pfam Con-

densation pHMM; in other words, it extends from four positions before our extended C1 motif to the fourth position after the extended C5 motif (these motifs were first revealed by de Crécy-Lagard et al. [1995] and reviewed by Marahiel et al. [1997]). Phylogenetic reconstruction is always based on this part of the C domain (see Figures 5.1 and 5.2). To extract the complete N-terminal part of the C domains, we followed the dissections applied by Roche and Walsh [2003] and checked the secondary structure with Quick2D [`toolkit.tuebingen.mpg.de/quick2_d`] of the MPI Bioinformatics Toolkit [Biegert et al., 2006].

5.5.3 Generation of Multiple Sequence Alignments

The quality of a reconstructed phylogenetic tree crucially depends on the underlying multiple sequence alignment. All sequence alignments in our study were generated using `MUSCLE` [Edgar, 2004a,b]. The alignment algorithm can be divided into three stages. First, a progressive alignment is built based on a UPGMA guide-tree. In the second stage, the underlying guide-tree is iteratively improved, yielding a new progressive alignment. The third stage involves refinement of the tree: Based on the tree, bipartitions of the dataset are produced and their profiles are extracted and realigned to each other. Thus, the finally generated alignment is not solely based on a single guide-tree, which is why we can rule out that the phylogenies reconstructed on the basis of these alignments merely reflect the guide-tree used in the first step of the algorithm.

5.5.4 Predicting Substrate Specificity

C domains catalyze the condensation of two amino acids, thus, they have two binding sites: the acceptor and the donor site. To be able to investigate whether the substrate specificity of one of these sites influences the phylogeny of the domain, the specificity of the upstream and downstream A domains in the assembly line was predicted with the NRPSpredictor (described in Chapter 4 and in [Rausch et al., 2005]) and stored for each C domain.

5.5.5 Predicting Functional Subtypes

Functional subtypes may be distinguished on the basis of sequence features, domain architecture or clustering behavior during tree reconstruction. Condensation and Heterocyclization domains may be recognized by the sequence motif they exhibit at their active site. The occurrence of a sequence motif within a longer sequence can be detected with the help of a position specific scoring matrix (PSSM) (see Section 3.3.2).

PSSMs were generated and applied for detecting the active site His-motif of the C domain and the DxxxxD-motif of the Heterocyclization domain. These were used to discriminate between the two subtypes. The His-motif was built from 86 sequences and the Cyc motif from 15 sequences. Sequence logos representing the information content of these two datasets are shown

in Fig. 5.3. All sequence logos were created with the application WebLogo [Crooks et al., 2004].

The PSSMs were only applied to a region of 100 residues, which was expected to contain the active site. In addition, a PSSM was generated for the N-terminal His-motif found in Dual E/C domains. It was constructed from 55 sequences which had been identified as Dual E/C domains by their clustering behavior in the phylogeny and by additional visual inspection of the alignment. The PSSM was applied for validation purposes to make sure that this N-terminal His-motif was unique to Dual E/C domains and could not be found in any other C domain subtype. A sequence logo representation of it is depicted in Fig. 5.4.

Predicting whether a C domain is a $^{L}C_{L}$- or a $^{D}C_{L}$-catalyst was established according to the observed domain organization of the modules in an NRPS sequence ($^{D}C_{L}$-catalysts were first described by Luo et al. [2002]). It is assumed that the role of a module with the domain structure C-A-T-E is the activation and epimerization of a residue that is in the L-stereo-configuration with the intention of incorporating a D residue into the final product. Alongside this, a C domain directly following an E domain is expected to be selective for residues in D-configuration, which is why it was assigned to the $^{D}C_{L}$-type. All other C domains were assumed to be $^{L}C_{L}$-catalysts. Classification as a $^{D}C_{L}$-catalyst is supposed to be fairly reliable. A false positive should only occur if the preceding epimerase turns out to be nonfunctional. The $^{L}C_{L}$ classification, however, is prone to errors when the respective C domain is the very first (N-terminal) domain in the protein. In this case, the type of condensation reaction can only be determined if the order in which the proteins act in the assembly line is known. To overcome this problem, we checked all assignments with the classification suggested by the phylogeny.

If the order of the subunits is unknown, temporarily incorrect assignments can only be revised later in the analysis.

5.5.6 Analysis of Multiple Sequence Alignments for Specificity Determining Positions

In a set of homologous enzymes, we may find subsets that all contain sequences with one distinct substrate specificity. These subsets of common function are called subtypes and often vary at certain positions, whereas the same positions may be conserved within a given subtype. Li et al. [2003] call these specificity-determining residues (SDR); Kalinina et al. [2004] refer to them as specificity determining positions (SDP). To determine SDPs from an alignment, calculating each column's mutual information is a possible way, as described by Li et al. [2003] and Kalinina et al. [2004]. For the reseach project presented in this chapter, SDPs were determined using the freely accessible SDPpred server [Kalinina et al., 2004]. Here, the mutual information is based on so-called smoothed frequencies, which allow residues with similar physico-chemical properties to be substituted. In addition to that, the significance of the mutual information of each position is estimated by calculating Z-scores and evaluating their significance. Predictions by SDP-

pred were compared with the highest scoring positions predicted by FRpred [Fischer et al., 2006, `toolkit.tuebingen.mpg.de/frpred`], which combines a mutual information term with a conservation score.

5.5.7 Reconstruction of Phylogenetic Trees

Several methods were applied for reconstructing phylogenetic trees from the multiple sequence alignments that were generated for each domain type. Trees presented in this chapter were reconstructed using protein sequences, as amino acid sequences are preferred to nucleotide sequences because they are more conserved and are not influenced by compositional bias like G+C content and codon usage. In addition, the mathematical model for the evolutionary change of amino acid sequences is much simpler than that of nucleotide sequences, which reduces the risk that the phylogeny is based on wrong evolutionary assumptions, since just a suitable substitution matrix has to be selected [Nai and Kumar, 2000]. The amino acid substitution matrix employed in this study was the Jones-Taylor-Thornton (JTT) matrix [Jones et al., 1992]. (A short overview of methods for phylogenetic reconstruction can be found in the *Biological Background* in Section 3.4.) To model the substitution rate, which is usually higher at positions of lower functional importance, the gamma distribution has been used [Gu and Zhang, 1997, for technical details, see Section 3.4.3].

Apart from `PHYLIP` [Felsenstein, 2006], all methods used in this study offer an estimation of the parameter α which determines the shape of the Γ distribution as an option. Whenever a gamma distributed rate variation was assumed, four gamma-rate categories were used to approximate the distribution. Several tree reconstruction methods were applied to each dataset to determine whether different methods yield different topologies, which in turn would indicate that the proposed topologies are unreliable. As a distance-based method, the Neighbor-Joining (NJ) method [Saitou and Nei, 1987] was applied. The distances were calculated with the program `protdist` and NJ was performed with `neighbor`, both available from the `PHYLIP` package. For NJ, only uniform substitution rates were used. As a maximum likelihood method, the programs `IQPNNI` [Vinh and von Haeseler, 2004] and `PHYML` [Guindon and Gascuel, 2003] were applied.

Bootstrapping [Felsenstein, 1985] was performed to test the reliability of the topologies (see *Technical Background* in Section 3.4.4 for the principles of bootstrapping).

Using the `PHYLIP` package, bootstrap datasets were generated with `seqboot` and used as input data for `neighbor`. `PHYML` also offers an option that allows a bootstrap analysis of the original data. This results in a set of trees which can be visualized as a *consensus network* using `SplitsTree4` [Huson and Bryant, 2006]. Specifying a cutoff value allows a clearer view of the bootstrap tree/network where only those edges which are supported by boostrap values higher than the cutoff are included.

5.5.8 Detection of Sequence Motifs and Their Representation

The program `meme` [Bailey and Elkan, 1994, `meme.sdsc.edu`] was used to detect the sequence motifs in C domains. `Meme` discovers one or more motifs in a collection of unaligned DNA or protein sequences. The C domain subtypes were aligned using `MUSCLE` [Edgar, 2004a,b], the multiple alignments were visualized using `JalView` [Clamp et al., 2004] and the motifs found by `meme` were extracted (cut out). It was ascertained that the C domain motifs described by Sieber and Marahiel [2005] were included as well as remarkable sequence positions in proximity to the motifs, such as single conserved residues or positions which were important for discerning the subtypes. The dissected motif sequences were used to create pHHMs with `HMMER` and also to create sequence logos using `seqlogo` by Crooks et al. [2004]. Sequence logos were preferred over consensus sequences, as they provide a more precise description of sequence similarity and reveal significant features of the alignment which are otherwise difficult to perceive. For sequence logos, positions with $> 10\%$ gaps were removed. The sequence logos of all C domain motifs created with `seqlogo` are available online as Supplementary file 5.9.

5.6 Contribution

The first ideas for the research project described in this chapter arose during the Master's thesis project of Ilka Hoof [Hoof, 2006] that I supervised. After Ilka had finished her Master's thesis, she continued collaborating with me on the project as a student assistant. Ilka gathered the sequences and constructed and analyzed the phylogenetic trees. I analyzed the subtype determining residues, constructed and interpreted the sequence logos, and continued the investigations on the glyco-NRPS. I wrote our findings in a manuscript for publication with the participation of Ilka in several sections. The paper has been published [Rausch et al., 2007]. I have reworked and adapted the publication text for this chapter.

5.7 Supplementary Data

Supplementary data for this chapter are freely available online at the website of the publication by Rausch et al. [2007, `www.biomedcentral.com/1471-2148/7/78`].

Supplementary file 5.1 — Phylogenetic tree of all 525 C domain sequences in this study, reconstructed using `phyml`

File name: all525C_E_tree.nex.zip; zipped Nexus file (file name extension .nex.zip, to be unpacked and opened with SplitsTree [Huson and Bryant [2006], `www.splitstree.org`]).

Supplementary file 5.2 — Phylogenetic tree of all 525 C domain sequences in this study, reconstructed using `phyml`

File name: all525C_E_tree.pdf; PDF file.

Supplementary file 5.3 — Comparison of the logos generated from the pHMMs for the thee subtypes ($^{L}C_{L}$, Starter and $^{D}C_{L}$ domains) using `LogoMat-P` [Schuster-Böckler et al., 2004]

File name: HMMLogos_LCL_Starter_DCL.pdf; PDF file.

Supplementary file 5.4 — HMMER outputs of glyco-NRPS: fossils in ComC and ComD

File name: GP-fossils.zip; ZIP file containing two text files.

Supplementary file 5.5 — Profile HMMs of the four complete C domain subtypes ($^{L}C_{L}$, Starter, $^{D}C_{L}$, Dual) which can be used to detect and distinguish between the subtypes.

File name: Condensation-hmms.hmm.zip; zipped text file (file name extension .hmm to be used with the program package HMMER [`hmmer.janelia.org`].

Supplementary file 5.6 — Aligned full length Condensation domains in this study

File name: complete_aligned_Cdoms.zip; zipped sequence file (aligned protein sequences in FASTA format).

Supplementary file 5.7 — Profile HMMs of all seven motifs of all subtypes ($^{L}C_{L}$, Starter, $^{D}C_{L}$, Dual)

File name: Motifs_LCLstarterDCLdual.hmm.zip; zipped text file (file name extension .hmm to be used with the program package HMMER `hmmer.janelia.org`.

Supplementary file 5.8 — List of NRPSs from known biosynthesis clusters used in this study

File name: known_NRPS_used.pdf; PDF file.

Supplementary file 5.9 — Sequence logos of all C domain motifs created with `weblogo` (Crooks et al. (2004)

File name: allLogos.zip; ZIP file containing image files in the PNG file format.

Chapter 6

Structural Bioinformatics of the NRPS Adenylation Domain: An Outlook

6.1 Overview and Motivation

In this chapter, we explore how structural bioinformatics can help us to understand and predict the substrate preference of uncharacterized Adenylation domains better. Here, our ultimate goal is to predict the specificity of an A domain by building a structural model of it, and to use *virtual screening* to find out which proteinogenic and non-proteinogenic amino acids and other aryl acids could be its potential substrate(s). This approach is especially attractive as it does not depend on A domain sequences with annotated specificity, in contrast to the machine learning approach presented in Chapter 4 or the phylogenetic subtyping of Chapter 5.

6.2 Results and Discussion

6.2.1 Homology Modeling of NRPS Adenylation Domains

The first idea was to build a homology model for the uncharacterized A domain (*target*) using the structure of the gramicidin synthetase A phenylalanine adenylating domain (GrsA-PheA, PDB code `1AMU`) as template, which could then be used in a molecular docking simulation as a *macromolecule*, using a series of NRP building blocks as *ligands*.

We chose NosD1 (involved in nostopeptolide synthesis in the cyanobacterium *Nostoc* sp. GSV224) as a template that is known to be TYR-specific, which had been determined biochemically by an ATP-PP_i-exchange reaction using purified recombinant protein [Hoffmann et al., 2003]. (For simplicity, we will talk about GrsA and NosD1 in the following but we always refer to their first Adenylation domains). NosD1 appeared to be a good test candidate to start with because it has a high, though typical, sequence similarity to GrsA (from motif A3 to A7 [Marahiel et al., 1997] it has 46% sequence iden-

tity and 61% similarity). NosD1 also has no gaps in the alignment with the "core" region from pos. 198-334 (with respect to `1AMU`) and only one difference in the ten amino acid "Stachelhaus code" [Stachelhaus et al., 1999; Challis et al., 2000] compared to GrsA. (SER instead of TRP at pos. 239). In the NosD1 structure model obtained with MODELLER 8 [Eswar et al., 2006] as described in the MODELLER Tutorial [`salilab.org/modeller/tutorial`] (side chain prediction with SCWRL3 [Cantutescu et al., 2003]; see *Materials and Methods* in 6.3.2), all active site residues were found to be displaced in one direction by a few Ångströms. However, in a realignment of the two PDB structures (res. 235-517 and 219-521 in GrsA and NosD1 respectively) with FATCAT (a rigid pairwise structure alignment) [Ye and Godzik, 2003, 2004], the two structures were found to be *significantly* similar, with an RMSD of 0.42 Å. Fig. 6.1 depicts the superimposed ten active site positions of GrsA and the NosD1 model, and Fig. 6.2 also shows the superimposed backbones of the two structures (res. 235-517).

Homology models of Adenylation domains using the GrsA A domain (`1AMU`) as their template have been reported previously [Ackerley et al., 2003; Lautru and Challis, 2004; di Vincenzo et al., 2005; Schwecke et al., 2006]. The goal of those publications was manual or semi-automatic docking of the putative substrates of the modeled structures. The purpose of the model constructed here however, was to find methods of fully automated docking and scoring of the best fitting substrates.

The similarity between the modeled structure of NosD1 and the structure of GrsA is obvious from the very low RMSD and visual inspection (see Figures 6.1 and 6.2) and the substrate binding could be readily studied with such a model. However, the objective of MODELLER is to build a structural model satisfying the spatial restraints of the whole protein. If one tries to model a (target) structure which is more distantly related to the template structure one has to expect that, in order to model exterior (and often functionally unimportant) loops, thus obtaining a higher overall score, the active site pocket might be slightly deformed, although this might not correspond to its real conformation. Consequently, one has to be particularly careful when docking to homology-modeled structures.

The three related structures, GrsA, firefly luciferase and DhbE (the latter two catalyze very different substrates), all share 16% sequence identity on average and have an average RMSD over the C_α atoms of the entire superimposed structures of 2.6 Å. However, the average RMSD, calculated over the C_α atoms enclosed in a sphere of radius 9 Å centered at the GrsA residue Asp235 in the active site, is 0.95 Å [di Vincenzo et al., 2005]. In fact, it is a known phenomenon that the active sites of enzymes tend to be structurally more conserved during evolution (reported, for example, by Irving et al. [2001]). This is the reason, why di Vincenzo et al. [2005] were able to perform docking simulations on homology models built at a sequence identity to the template of only 25-30% (eukaryotic freestanding Adenylation domains).

Based on the almost exact conservation of the active site topology in our modeling experiment and the findings by Irving et al. and di Vincenzo et al., and the information from Kohlbacher, we decided to perform docking

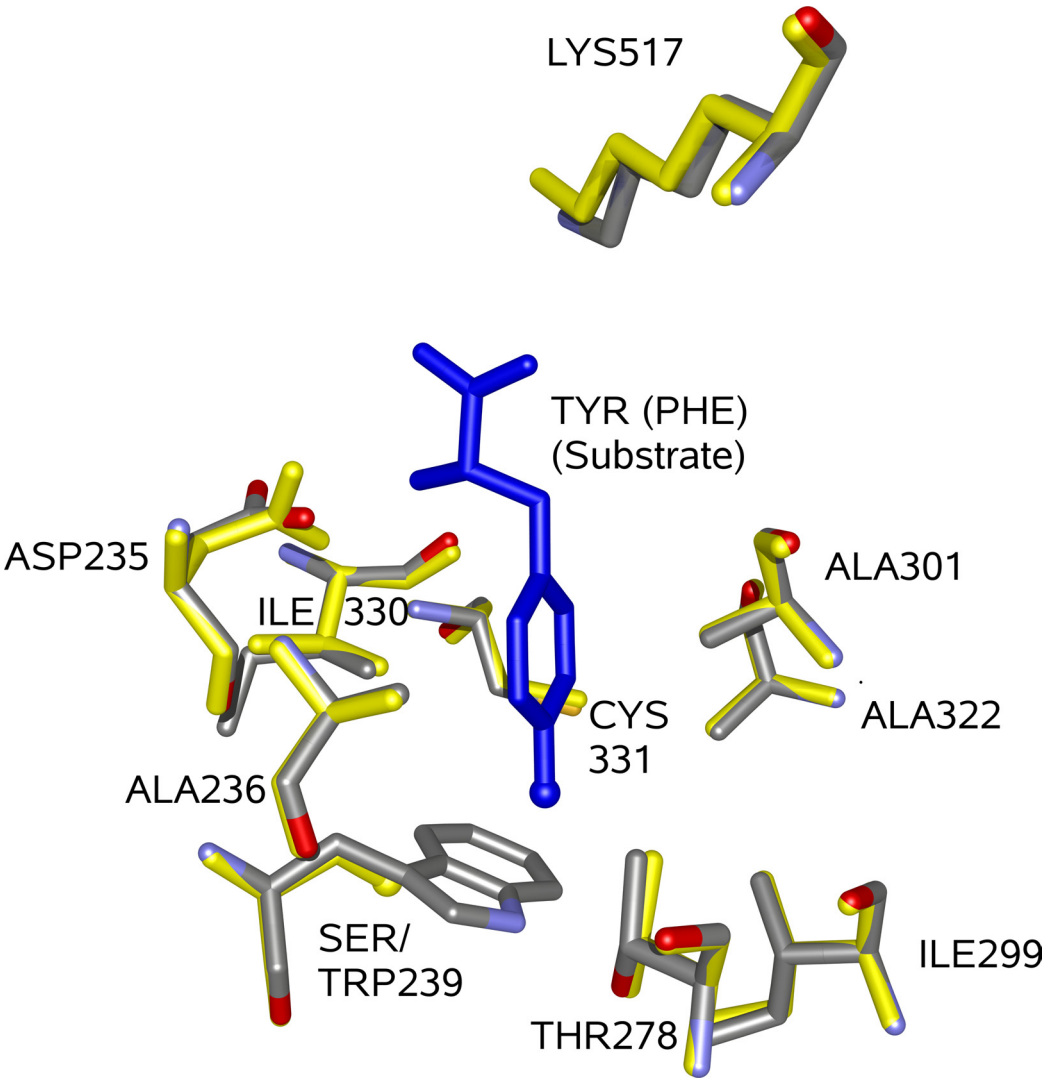

Figure 6.1: Superposition of the ten active site residues of GrsA (colored by atom type) and the NosD1 structure model (yellow). The depicted substrate (blue) is a tyrosine activated by NosD1, while GrsA activates phenylalanine. Where GrsA has a tryptophan (pos. 239), NosD1 has a serine that probably forms a hydrogen bridge with the hydroxyl group of the bound tyrosine. (Both hydroxyl groups are shown in ball representation for illustration).

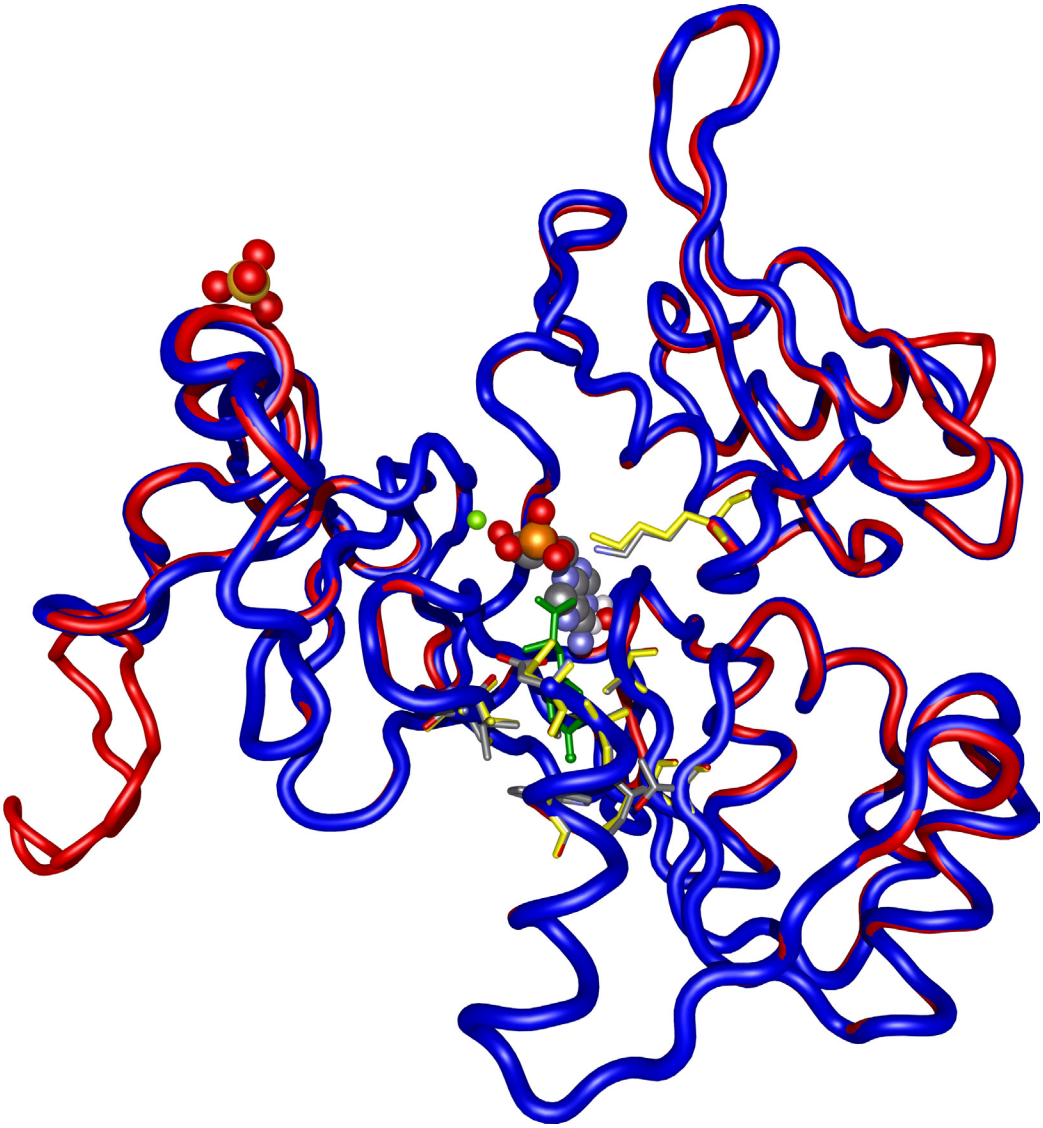

Figure 6.2: Superposition of positions 235 to 517 of GrsA (blue) and the modeled NosD1 A domain (red), which include the ten active site residues (which are represented as in Fig. 6.1 with the bound substrate in green). Both structures, depicted as tubes, deviate only by 0.42 Å RMSD. The hetero-atoms forming ATP, Mg^{2+}, SO_4^{2-} and a buried water molecule are shown as space-filling van der Waals models.

	wild type	T278M/A301G mutant
PHE	-31.9 kJ/mol	-31.1 kJ/mol
LEU	-24.6 kJ/mol	-23.9 kJ/mol

Table 6.1: Estimated free energies of binding between phenylalanine and leucine, and the GrsA wild type structure and its T278M/A301G mutant. According to the biochemical experiment [Stachelhaus et al., 1999], PHE has a higher affinity for the wild type structure than for the mutant structure and vice versa for LEU. The free binding energy estimation did thus not return the expected relations for the docking experiment using the mutant structure – the obtained energy values are the same within the error margins for the two macromolecules only depending on the ligand. Different possible explanations for these results are discussed in the text. However, the binding topology obtained from the docking experiments is expected to be very close to the actual one, based Fig. 6.3 and Fig. 6.4.

experiments directly on the GrsA structure, into which we introduced *in silico* mutations to "simulate" other A domains (see next section).

6.2.2 Molecular Docking Simulations on *in silico* Mutated GrsA A Domains Using AutoDock

We used the atom-based docking simulation program AutoDock 3 [Morris et al., 1998] for our molecular docking studies about the structure of GrsA (for more information about this program and on the parameters used, see the *Materials and Methods* section of this chapter). We decided to construct a double mutant GrsA as done by Stachelhaus et al. [1999] in wet lab experiments (Thr278→Met/Ala301→Gly). Stachelhaus et al. [1999] have predicted and proven biochemically that the mutant specifically activates Leu at 100% relative activity and Phe only at 40%, compared to 10% vs. 100% relative activity of the wild type (at comparable absolute activities for the preferred substrate). The single mutations were introduced with BALLView [Moll et al., 2005, 2006] (see the *Materials and Methods* Section of this chapter for details).

Four docking experiments have been undertaken: Phenylalanine (from the wild type structure) and leucine have been docked onto the wild type structure and the mutant structure (Thr278→Met/Ala301→Gly), respectively. The estimations of the free energies of binding – presented in Table 6.1 – are plausible for the docking experiment on the wild type structure even though one would expect a greater difference in the values. The energies obtained for the docking simulations on the mutant structure, however, are identical to the values obtained for the simulations on the wild type structure (within the error margins of docking experiments). These results do not correspond to the biochemical experiments by Stachelhaus et al. [1999].

There are several possible explanations for these findings, including propositions for future investigations: As we can see from Figures 6.3 and 6.4 we can be confident that our docking experiments yield the right binding topolo-

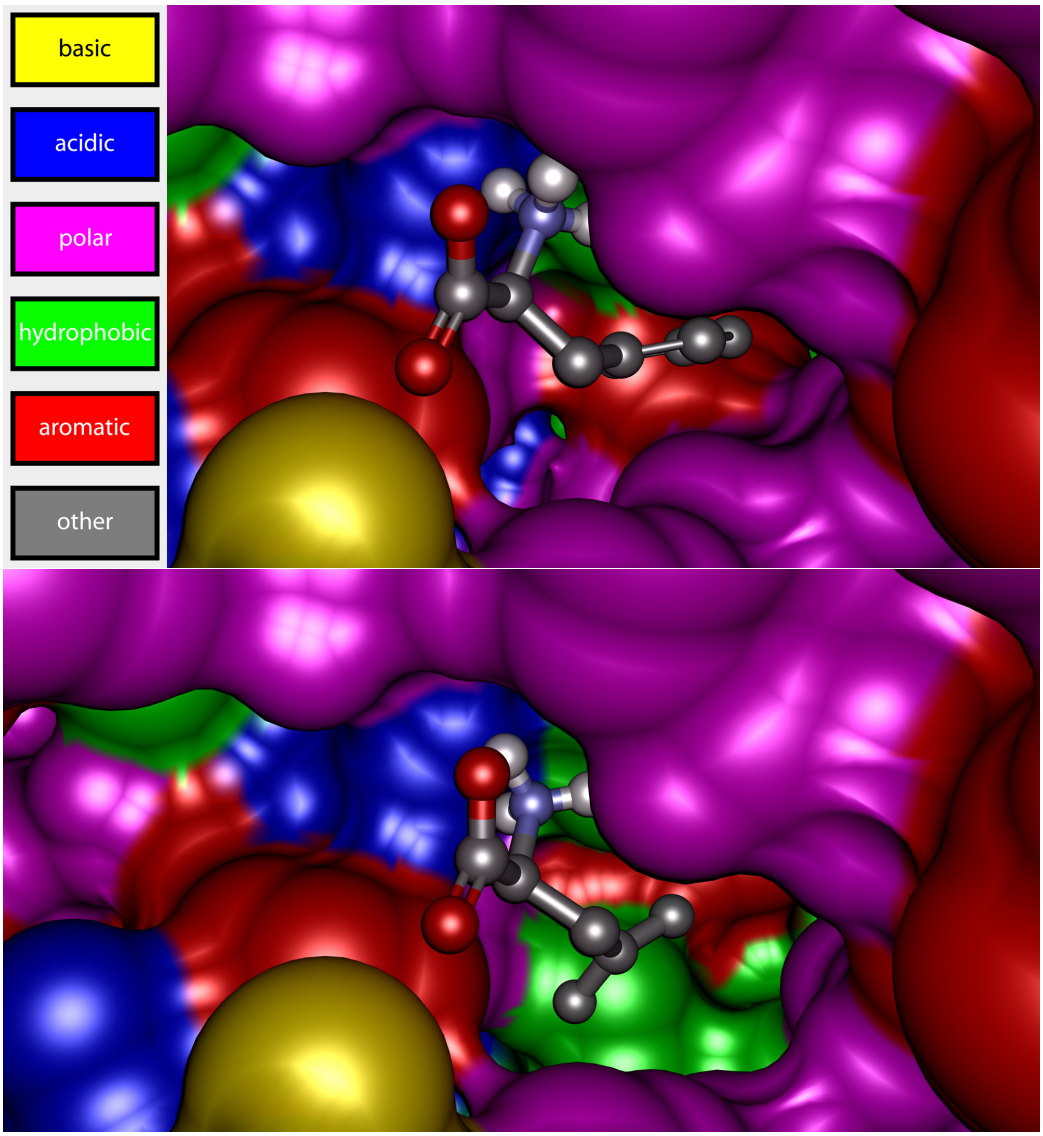

Figure 6.3: Visualization of the results of the docking experiments. The upper image shows the docked phenylalanine in the wild type GrsA Adenylation domain. Below, the image shows the result of the docking experiment with leucine and the T278M/A301G mutant structure. The carboxyl group of the bound amino acid always forms a salt bridge to the lysine at position 517 in GrsA (in the foreground in yellow) and the amino group forms a salt bridge to the aspartate 235 (in the background in blue). Behind the phenyl ring of the bound PHE in the upper picture, one can see the aromatic (red) tryptophan 239 which – in the lower picture – is partly covered by the hydrophobic methionine 278 in the mutant structure. The alanine 301 in the wild type structure is located directly below the phyenyl ring of the ligand (colored in magenta as a polar side chain); its replacement, glycine 301 is visibly smaller in the mutant structure.

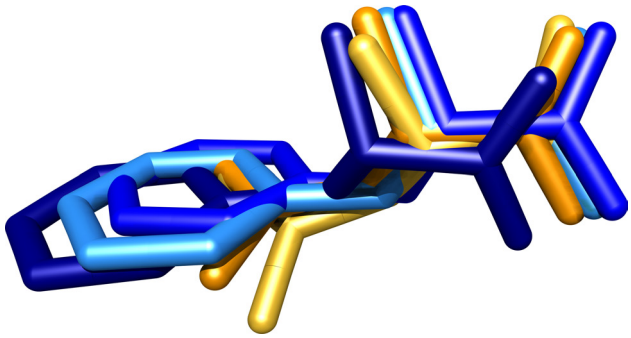

Figure 6.4: Stick models of the docked ligands phenylalanine (blue tones) and leucine (yellow/orange). The dark-blue PHE is shown how it is actually found in the wild type GrsA crystal. The positions of all backbone atoms overlap very well, which reflects the postulated conserved binding mode of all amino acids in Adenylation domain homologs by salt bridges formed to the conserved Asp235 and Lys517. Note that the docking program (AutoDock) useses random start positions for the ligands at the beginning of the procedure.

gies of the substrates, at least in a good approximation. The carboxyl group of the docked amino acid is always positioned in a way that it can form a salt bridge to the Lys517 (in Fig. 6.3 in the foreground in yellow), and the amino group can form a salt bridge to the Asp235 (blue in the background). The peptide backbones of the docked substrate amino acids are in the same orientation as the backbone of the PHE bound in the wild type crystal structure (see Fig. 6.4). All these observations suggest that the docking simulation is successful. However, our results would suggest that the T278M/A301G mutant has still a higher affinity for PHE, which is wrong according to Stachelhaus et al.'s results. We want to recall that during the docking simulation the side chains of the macromolecule are kept fixed and only the ligand is flexible. Of course, this is not optimal because the side chains in the active site will adapt to the ligand during the binding process. Moreover, the function for the estimated free energy of binding calculated by AutoDock is the result of a trade-off between computational efficiency and precision [Morris et al., 1998]. Thus it might return imprecise results. Consequently, we need an approach which allows us to model conformational flexibility in the macromolecule, at least in its active site, and we need a reliable scoring function for the binding affinities and catalytic efficiencies of the wild type and mutant enzymes that we will construct. Ideas for achieving this will be discussed in the next section.

6.2.3 Using SCWRL3 to Model the Side Chain Conformations in the Active Site of Wild Type and Mutated GrsA A Domains with Docked Substrates

AutoDock 3, which was used for the docking studies presented in the previous section, did not allow for the modeling of conformational flexibility in selected sidechains in the target macromolecule. Autodock 4, which now of-

fers this possibility, was not available during the course of this study, so we decided to use SCWRL3 [Cantutescu et al., 2003] for our simulations instead. SCWRL3 is a program which predicts the conformations of the side-chains of the amino acids of a protein structure, given the coordinates of the protein's peptide backbone (the C_α atoms). For a brief introduction to the strategy of SCWRL3, see the *Materials and Methods* section of this chapter; for full details, see Cantutescu et al. [2003]. The reason why we tried SCWRL3 for a docking simulation is the advantage that both the ligand and the macromolecule consist of amino acids (we ignored the fact that the ligands can be non-proteinogenic or carboxylic acids). Moreover, the "peptide backbone" of the ligand can be regarded as it would be in a normal SCWRL3 run. Therefore, we simply added the ligand (PHE, LEU or another ligand) after the C-terminal amino acid in the original pdb file to make it part of the main chain of the protein. The docking process with SCWRL3, with the backbone atoms of the amino acid being kept still, took only a few seconds on a normal PC compared to ten minutes up to one hour for a complete AutoDock run.

We have inspected the orientations of the amino acid side chains of the ligand and the macromolecule in detail. When we prepared the macromolecule for the docking with AutoDock, we had removed the phenyl ring of the bound PHE, leaving only the coordinates of an alanine as a "frame" during the SCWRL3 run to avoid a bias towards the conformation of the active site of the wild type GrsA domain with the bound PHE. We have now superimposed the coordinates of the wild type PDB coordinates and the coordinates obtained from the docking with SCWRL3 and AutoDock. Fig. 6.5 illustrates exemplarily the bound ligands (PHE) and the active site TRP239 of the wild type macromolecule. It becomes obvious that the simplification to keep the active site side chains fixed during the docking procedure is largely suboptimal: The prediction of the active site and PHE ligand side chains with SCWRL3 returned a conformation (yellow in Fig. 6.5) which is very similar to the one found in the actual crystal structure (`1AMU`, blue) and different from the conformation used in the AutoDock docking experiment. It actually appears that this conformation is an example of T-shaped π-π stacking; the phenyl ring stands perpendicular, slightly tilted over the indole ring of the tryptophan residue of the macromolecule at a distance of 3.5 Å as expected for this kind of aromatic stacking [Thomas et al., 2002]. One more indication that the active site conformation used for docking with AutoDock is sub-optimal.

The strategy to employ a side chain prediction (SCWRL3) is promising because it is much faster than a docking with AutoDock and it allows for side chain flexibility of the ligand and the macromolecule residues. Future work will be to develop a more complex energy function than the one implemented in SCWRL3, which is currently been done in Oliver Kohlbacher's working group and to look for a suitable energy function to evaluate the catalytic activity of the ligand-macromolecule complexes obtained by SCWRL3.

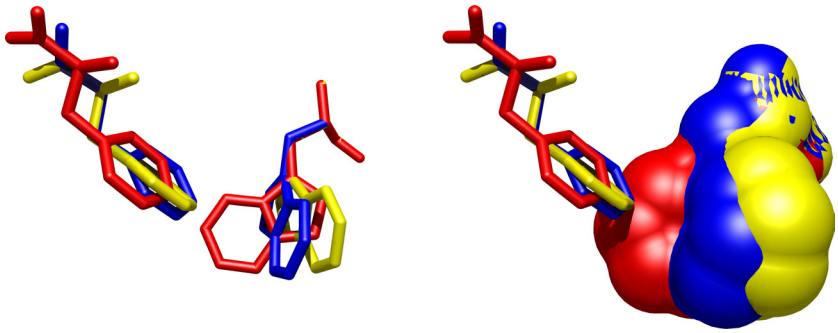

Figure 6.5: Three superimposed visualizations of the phenylalanine ligand (as a stick model) and tryptophan 239 of the macromolecule (as stick model on the left, as space-filling model on the right). Blue: coordinates of `1AMU`. Yellow: After running SCWRL3 on the coordinates of `1AMU` with the phenylalanine ligand appended to the peptide chain of the macromolecule so that SCWRL3 can also optimize its side chain orientation. Red: The ligand as positioned after docking with AutoDock, and the macromolecule TRP side chain as obtained by an SCWRL3 side chain prediction while leaving the PHE of `1AMU` without the phenyl ring in the active site. The phenyl ring of the phenylalanine and the indole ring of the tryptophan apparently form a T-shaped π-π stacking according to their distance (3.5 Å) and orientation.

6.3 Materials and Methods

6.3.1 Introducing Point Mutations *in silico*

Using BALLView

The side chains of the wild type that differ from the mutant structure were removed, the name of the residue was changed in the properties menu and hydrogens were added to all side chains, a process in which BALLView reconstructs incomplete side chains using the most frequent rotamer which does not interfere with other side chains.

Using SCWRL3

SCWRL3 [Cantutescu et al., 2003] offers a different way of replacing certain amino acid side chains but preserving the coordinates of the backbone. The program accepts a template structure and the sequence of a target structure as input. Side chains at positions where the template and target differ are replaced according to the given sequence information. The side chain orientations of these mutant positions, as well as of preserved positions written in upper case in the sequence file, are modeled by the program (see next section).

6.3.2 Side-Chain Conformation Prediction Using SCWRL3

Methods for homology modeling of protein structures, *ab initio* protein structure prediction and protein design applications typically predict the coordinates of the peptide backbone (the C_α atoms). In the next step, a fast and

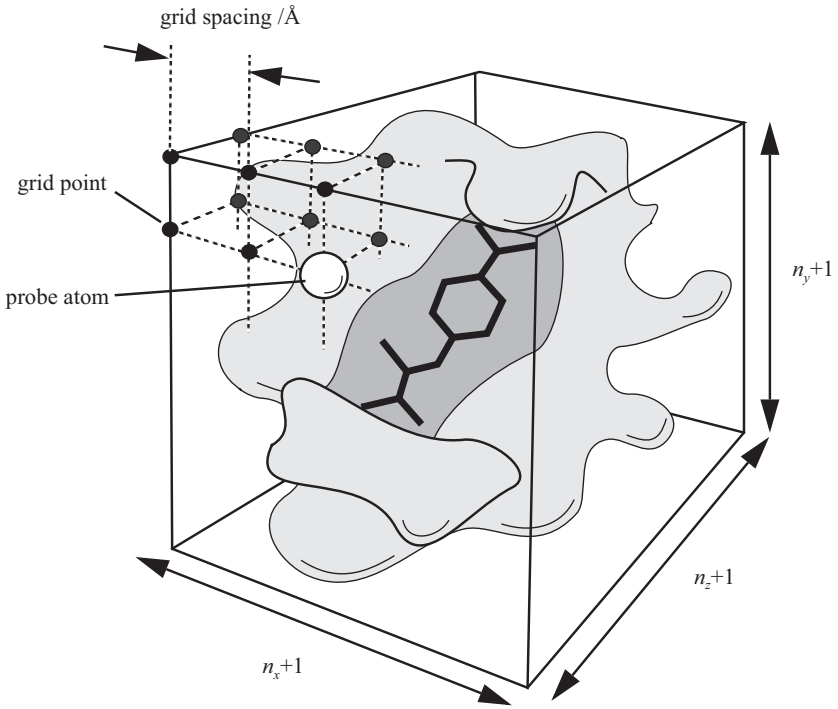

Figure 6.6: The main features of a grid map: The ligand can be seen in the center of the grid map, buried inside the active site of the protein. In the case shown, the grid map encompasses the whole protein. The grid spacing is the same in all three dimensions. Reprinted with kind permission from Morris et al. [2001].

accurate side-chain conformation prediction is needed. SCWRL3 is an algorithm that solves the combinatorial problem encountered in the side-chain prediction problem by using results from graph theory. In this method, side chains are represented as vertices in an undirected graph. Any two residues that have rotamers with nonzero interaction energies are considered to have an edge in the graph. The resulting graph can be partitioned into connected subgraphs with no edges between them. These subgraphs can in turn be broken into biconnected components, which are graphs that cannot be disconnected by the removal of a single vertex. The combinatorial problem is reduced to finding the minimum energy of these small biconnected components and combining the results to identify the global minimum energy conformation [Cantutescu et al., 2003].

6.3.3 Molecular Docking Simulations with AutoDock

Principles of AutoDock

AutoDock3 [Morris et al., 1998] is an atom-based molecular docking simulation program. AutoDock achieves a rapid energy evaluation by precalculating grid-based molecular affinity potentials. For each atom type in the substrate molecule, a separate three-dimensional grid map is generated by assigning the energy of the interaction of a single probe atom to the grid point. Fig. 6.6 illustrates the main features of a grid map.

In a similar way, a grid for the electrostatic potential is calculated. For

the ligand, a rigid root is defined from which rotatable bonds extend. The program implements different search methods (simulated annealing, genetic search algorithms and local search), from which the Lamarckian genetic algorithm was used in this study. This algorithm starts the docking process with a random population of a limited number of individuals. These individuals represent molecules with uniformly distributed random values for torsion angles, quaternions and translation vectors. The values for torsion angles, quaternions and translation vectors represent the genes of an individual and can be inherited by the upcoming population. Some of the ligands undergo a local search before the energy of each individual is calculated to determine how many offspring it will produce in the following generation. Finally, a two-point crossover and mutations are performed on random members of the population, resulting in new ligand positions and conformations. Additionally, elitism is normally used to let some individuals with the best energies survive unchanged into the next generation.

Parameters Used with AutoDock in the Presented Study

The macromolecule was prepared as follows: First, the side chain conformation was predicted using SCWRL3, setting the coordinates of the bound PHE (having the phenyl ring removed) in the structure as frame file. Then we followed the recommendations in AutoDock's User Guide: Polar hydrogens were added using `protonate` and the energetically optimal orientation of the hydrogen bonds was predicted using `pol_h` and the configuration file `PROTON_INFO.kollua_polH` from the Amber 7 package of molecular simulation programs [Case et al., 2005]. Finally, partial charges (Kollman charges) and solvation parameters were added using `q.kollua` and `addsol` from the AutoDock package [Morris et al., 1998]. The ligand coordinates were taken from the GrsA structure (`1AMU`); the PHE structure (ligand) was either taken as is, or changed into another amino acid with BALLView as described in the previous section.

The grid spacing was 0.2 Å with 40 intervals in the x- and y- dimension and 84 in the z-dimension, providing a grid box which included the entire binding site of the enzyme, and enough space for the ligand translational and rotational walk. The grid was centered at (32.20 Å, 98.48 Å, 33.23 Å), the active site pocket center (determined with AutoDockTools [`autodock.scripps.edu` → AutoDockTools]).

For each docking simulation, 200 runs were performed with a maximum number of 27 000 genetic algorithm operations, generated on a single population of 100 individuals. The maximum number of energy evaluations was set to 250 000. Other parameters for the docking were: a random starting position and conformation, a maximal mutation of 2 Å in translation and 50° in rotations, an elitism of 1, a mutation rate of 0.02, a crossover rate of 0.8 and a local search rate of 0.06. Simulations were ranked according to the estimated free energy of binding between the protein and the ligand, a summation of intermolecular energy terms, and the torsional free energy.

6.4 Conclusion and Outlook

We have demonstrated that techniques available from structural bioinformatics can provide useful insights into the molecular functioning of Adenylation domains of which the structure is unknown. Necessary steps in the future will be to look for better energy functions to evaluate complexes of Adenylation domains with bound substrates obtained by docking or by prediction of the side chains with programs like SCWRL3, with the goal of improving the prediction of the principally activated substrate. Therefore it will be necessary to apply different available program packages to predict the binding affinities, for example, to try the Amber 9 package [Case et al., 2005] or Liaison [Zhou et al., 2001; Schrödinger, Inc., 2003]. It would also be worthwhile to compare SCWRL3, which uses a simple repulsive steric energy term, with a program that uses a more sophisticated energy function such as a piece of software which is currently being developed by Nora Toussaint in Oliver Kohlbacher's group. Moreover, one should try the recently released AutoDock 4 docking program [`autodock.scripps.edu`] which now allows side chains in the macromolecule to be flexible and uses a new free-energy scoring function. The docking program Glide [Friesner et al., 2004; Halgren et al., 2004] from Schrödinger, Inc. (Portland, OG, USA) and other docking programs could be used for comparison.

Additionally it would be worthwhile to do more homology modeling of A domains with MODELLER 9 and other programs, for example Prime [Schrödinger, Inc.], putting a high emphasis on the preservation of the geometry of the binding pocket.

Once we will have established a working model for predicting the binding affinity, we will need to simulate the kinetics of the whole adenylation process with the goal of getting realistic estimations of the substrate turnover rates. Because the part of the active site that catalyzes the adenylation is highly conserved in all subtypes and is (supposably) independent from the part that coordinates the ligand amino acid side chain, there is hope that the turnover rate can be estimated from the binding affinity.

Chapter 7

General Conclusions

Chapter 7 concludes on the new predictive methods presented and their impact on and applicability to other problems, followed by an outlook what bioinformatical challenges need to be solved in the NRPS and PKS field.

7.1 Concluding Remarks on the Results

In this thesis, we have presented two important steps towards predicting the ordered composition of novel non-ribosomal peptides (NRPs) based on the sequence of their synthetases (NRPSs): First, we use machine learning (SVMs) to predict which amino acid is selected by a given Adenylation (A) domain for incorporation into the NRP (Chapter 4). We implement this approach in our free program NRPSpredictor. Then, by means of phylogenetic functional subtyping and profile Hidden Markov Models which we make available, we are able to predict the subtype of the following Condensation (C) domain which allows us to determine the stereo-configuration of the incorporated amino acids (Chapter 5). The knowledge of the exact subtype of a C domain may also be informative for determining the order in which several NRPSs in one biosynthesis cluster act in concert, if it is found at an N-terminus of an NRPS. If the C domain is a Starter C domain, its NRPS will be the first in the assembly line. A $^{\mathrm{D}}\mathrm{C}_{\mathrm{L}}$ domain is expected to succeed an E domain, an $^{\mathrm{L}}\mathrm{C}_{\mathrm{L}}$ domain is not. If the NRP product is known, then the building block-to-domain assignment is further facilitated.

In Chapter 6, we have highlighted that predicting the A domain specificity with the aid of structural bioinformatics techniques (molecular modeling and docking) can be very helpful, especially if the NRPSpredictor gives no prediction or only at low confidence (which can be expected for rare specificities, or domains which exhibit an alternative binding mode which is often observed for eukaryotic sequences). However, more work is necessary to obtain more meaningful binding energy estimations.

7.2 Impact and Applicability of the Developed Methods on the Substrate Specificity Prediction of Enzymes

The general strategies that we have pinpointed in this thesis are applicable to the prediction of functional subtypes of other enzyme families under certain requirements: The concept of machine learning of the physico-chemical fingerprint as implemented by the NRPSpredictor (Chapter 4) can be applied provided that the protein sequences of the different subtypes share a sufficiently high sequence identity to justify the assumption that the active site topology is conserved to make sure that homologous positions contribute to the specificity/functional subtype in an analogous manner. The most important factor is a high sequence conservation within the parts that constitute the active site. For an estimation of the relationship between sequential and structural similarity, refer to Rost [1999]. Ideally, one would need at least one resolved structure of the proteins of the homologous family, which will be used to determine the residues within a certain radius of the active site. If no structure is available, the functionally important (subtyping) positions may be inferred using entropy and/or conservation based approaches like those described by Fischer et al. [2006] and Kalinina et al. [2004].

Several possible applications have been listed in Section 4.5. In his Master's thesis project, Marc Röttig [2006] applied the (generalized) NRPSpredictor strategy successfully to different subtypes of glycosyltransferases and is currently further developing this "Active Site Classification" (ASC) method [Röttig et al., 2007].

Functional classifications may also be facilitated with phylogenetic methods and the detection of sequence motifs as shown in Chapter 5 for the C domains. However, the trees obtained by phylogenetic reconstruction may simply reflect the species phylogeny if the functional subtyping signal is only carried by a few positions in the peptide as is the case with the A domain. In a comprehensive analysis of the phylogeny of PKS and NRPS domains [Hoof, 2006], we could observe that both signals are frequently superimposed.

7.3 Future Challenges in NRPS/PKS Research

The biosynthetic factories of NRPS, PKS and post-assembly-line tailoring enzymes are still far from being fully understood. Molecular structures are now available for most NRPS/PKS domains, which greatly helps the understanding of their molecular functioning. Even so, the individual domains still bring surprises like the recently discovered Dual E/C domains [Balibar et al., 2005]. But understanding the substrate selection of and the communication between the domains is the great challenge ahead. The publication by Koglin et al. [2006] that elucidated the "shuttle" function of the T domain, which interacts with the A, C, E and TE domains, and the publications by Minowa et al. [2007] and Thattai et al. [2007] that explored the co-evolution of interacting C- and N-terminal domains in NRPS/PKS and PKS (respectively)

are important for answering this question. But, more bioinformatics and wet-lab studies are still necessary to find out which positions/regions in the protein are relevant for the efficient recognition and transfer of the correct substrate to the next domains/modules. In preparation for these statistical and biochemical studies, it will be necessary to establish a comprehensive analysis tool and a database of PKSs, NRPSs and their domains with annotated functions and specificities. Such a tool, combined with an annotated database, could integrate all currently available predictive methods of the NRPS/PKS field and would be very helpful for researchers in the field.

The common way to predict the substrate of the AT domain in PKS is currently to look for distinctive sequence motifs (see Haydock et al. [2005] and references therein). Using the recently published structure of the KS–AT didomain [Tang et al., 2006] with the NRPSpredictor strategy, one could implement an automated prediction for the AT domain substrate specificity. In an analogous manner, one should try to further elucidate the substrate specificity of KS, C and TE domains using the available structures and statistic evaluations of their sequence alignments for identifying specificity determining positions. Of course, additional crystal structures of NRPS and PKS domains, especially co-crystallizations of interacting domains, would boost our understanding of the molecular mechanisms. This would also allow us to further improve the predictive methods further.

The more we increase our understanding of those machineries that are "both elegant and admirably efficient" (citing Fischbach and Walsh [2006]), the more successful will our attempts be to engineer NRPS/PKS systems that produce novel compounds with interesting properties.

Appendix A

Publications

1. Olaf Delgado Friedrichs, Aaron L. Halpern, Ross Lippert, *Christian Rausch*, Stephan C. Schuster, Daniel H. Huson. **Syntenic Layout of Two Assemblies of Related Genomes.** In *Proceedings of the German Conference on Bioinformatics 2004 in Bielefeld*, vol. 53 of Lecture Notes in Informatics, Gesellschaft für Informatik, Germany, pages 3–12.

   To facilitate research in comparative genomics, sequencing projects are increasingly aimed at assembling the genomes of closely related organisms. Given two incomplete assemblies of two related genomes, the question arises how to use the similarity of the two sequences to obtain a better ordering and orientation of both assemblies. In this paper, we formalize this question as the Optimal Syntenic Layout problem, show that it is in general NP-hard, but that it can be solved well in practice using an algorithm based on maximal graph matching. We illustrate the problem using different assemblies of two strains of *Bdellovibrio bacteriovorus*.

2. *Christian Rausch*, Tilmann Weber, Oliver Kohlbacher, Wolfgang Wohlleben and Daniel H. Huson. **Specificity Prediction of Adenylation Domains in Nonribosomal Peptide Synthetases (NRPS) Using Transductive Support Vector Machines (TSVMs).** *Nucleic Acids Research* (2005), volume 33, pages 5799-5808.

   We present a new support vector machine (SVM)- based approach to predict the substrate specificity of subtypes of a given protein sequence family. We demonstrate the usefulness of this method on the example of aryl acid-activating and amino acid-activating Adenylation domains (A domains) of nonribosomal peptide synthetases (NRPS). The residues of gramicidin synthetase A that are 8 Å around the substrate amino acid and corresponding positions of other Adenylation domain sequences with 397 known and unknown specificities were extracted and used to encode this physico-chemical fingerprint into normalized real-valued feature vectors based on

the physico-chemical properties of the amino acids. The SVM software package SVM^{light} was used for training and classification, with transductive SVMs to take advantage of the information inherent in unlabeled data. Specificities for very similar substrates that frequently show cross-specificities were pooled to the so-called *composite specificities* and predictive models were built for them. The reliability of the models was confirmed in cross-validations and in comparison with a currently used sequence-comparison- based method. When comparing the predictions for 1230 NRPS A domains that are currently detectable in UniProt, the new method was able to give a specificity prediction in an additional 18% of the cases compared with the old method. For 70% of the sequences both methods agreed, for $< 6\%$ they did not, mainly on low-confidence predictions by the existing method. None of the predictive methods could infer any specificity for 2.4% of the sequences, suggesting completely new types of specificity.

3. Efthimia Stegmann, *Christian Rausch*, Sigrid Stockert, Daniel Burkert and Wolfgang Wohlleben. **The Small MbtH-like Protein Encoded by an Internal Gene of the Balhimycin Biosynthetic Gene Cluster is not Required for Glycopeptide Production.** *FEMS Microbiology Letters* (2006), volume 262, pages 85–92.

   The balhimycin biosynthetic gene cluster of the glycopeptide producer *Amycolatopsis balhimycina* includes a gene (*orf1*) with unknown function. *orf1* shows high similarity to the *mbtH* gene from *Mycobacterium tuberculosis*. In almost all nonribosomal peptide synthetase (NRPS) biosynthetic gene clusters, we could identify a small mbtH-like gene whose function in peptide biosynthesis is not known. The mbtH-like gene is always colocalized with the NRPS genes; however, it does not have a specific position in the gene cluster. In all glycopeptide biosynthetic gene clusters the *orf1*-like gene is always located downstream of the gene encoding the last module of the NRPS. We inactivated the orf1 gene in *A. balhimycina* by generating a deletion mutant. The balhimycin production is not affected in the *orf1*-deletion mutant and is indistinguishable from that of the wild type. For the first time, we show that the inactivation of an *mbtH*-like gene does not impair the biosynthesis of a nonribosomal peptide.

4. Lalitha Voggu, Steffen Schlag, Raja Biswas, Ralf Rosenstein, *Christian Rausch*, and Friedrich Götz. **Microevolution of Cytochrome *bd* Oxidase in Staphylococci and Its Implication in Resistance to Respiratory Toxins Released by *Pseudomonas*.** *Journal of Bacteriology* (2006), volume 188, pages 8079–8086.

   Pseudomonas aeruginosa and *Staphylococcus aureus* are op-

portunistic pathogens and frequently coinfect the lungs of cystic fibrosis patients. *P. aeruginosa* secretes an arsenal of small respiratory inhibitors, like pyocyanin, hydrogen cyanide, or quinoline N-oxides, that may act against the commensal flora as well as host cells. Here, we show that with respect to their susceptibility to these respiratory inhibitors, staphylococcal species can be divided into two groups: the sensitive group, comprised of pathogenic species such as S. aureus and S. epidermidis, and the resistant group, represented by nonpathogenic species such as *S. carnosus*, *S. piscifermentans*, and *S. gallinarum*. The resistance in the latter group of species was due to *cydAB* genes that encode a pyocyanin- and cyanide-insensitive cytochrome *bd* quinol oxidase. By exchanging *cydB* in *S. aureus* with the *S. carnosus*-specific *cydB*, we could demonstrate that CydB determines resistance. The resistant or sensitive phenotype was based on structural alterations in CydB, which is part of CydAB, the cytochrome *bd* quinol oxidase. CydB represents a prime example of both microevolution and the asymmetric pattern of evolutionary change.

5. *Christian Rausch*, Ilka Hoof, Tilmann Weber, Wolfgang Wohlleben and Daniel H. Huson. **Phylogenetic Analysis of Condensation Domains in NRPS Sheds Light on Their Functional Evolution.** *BMC Evolutionary Biology* (2007), volume 7, page 78.

Background: Non-ribosomal peptide synthetases (NRPSs) are large multimodular enzymes that synthesize a wide range of biologically active natural peptide compounds, of which many are pharmacologically important. Peptide bond formation is catalyzed by the Condensation (C) domain. Various functional subtypes of the C domain exist: An ${}^{L}C_{L}$ domain catalyzes a peptide bond between two L-amino acids, a ${}^{D}C_{L}$ domain links an L-amino acid to a growing peptide ending with a D-amino acid, a Starter C domain (first denominated and classified as a separate subtype here) acylates the first amino acid with a β-hydroxy-carboxylic acid (typically a β-hydroxyl fatty acid), and Heterocyclization (Cyc) domains catalyze both peptide bond formation and subsequent cyclization of cysteine, serine or threonine residues. The homologous Epimerization (E) domain flips the chirality of the last amino acid in the growing peptide; Dual E/C domains catalyze both epimerization and condensation. **Results:** In this paper, we report on the reconstruction of the phylogenetic relationship of NRPS C domain subtypes and analyze in detail the sequence motifs of recently discovered subtypes (Dual E/C, ${}^{D}C_{L}$ and Starter domains) and their characteristic sequence differences, mutually and in comparison with

$^{L}C_{L}$ domains. Based on their phylogeny and the comparison of their sequence motifs, $^{L}C_{L}$ and Starter domains appear to be more closely related to each other than to other subtypes, though pronounced differences in some segments of the protein account for the unequal donor substrates (amino vs. β-hydroxy-carboxylic acid). Furthermore, on the basis of phylogeny and the comparison of sequence motifs, we conclude that Dual E/C and $^{D}C_{L}$ domains share a common ancestor. In the same way, the evolutionary origin of a C domain of unknown function in glycopeptide (GP) NRPSs can be determined to be an $^{L}C_{L}$ domain. In the case of two GP C domains which are most similar to $^{D}C_{L}$ but which have $^{L}C_{L}$ activity, we postulate convergent evolution. **Conclusions:** We systematize all C domain subtypes including the novel Starter C domain. With our results, it will be easier to decide the subtype of unknown C domains as we provide profile Hidden Markov Models (pHMMs) for the sequence motifs as well as for the entire sequences. The determined specificity conferring positions will be helpful for the mutation of one subtype into another, e.g. turning $^{D}C_{L}$ to $^{L}C_{L}$, which can be a useful step for obtaining novel products.

6. Daniel H. Huson, Tobias Dezulian, Markus Franz, *Christian Rausch*, Daniel C. Richter and Regula Rupp (all authors have contributed equally). **Dendroscope: An Interactive Viewer for Large Phylogenetic Trees.** *BMC Bioinformatics* (2007), accepted.

Background: Research in evolution requires software for visualizing and editing phylogenetic trees, for increasingly very large datasets, such as arise in expression analysis or metagenomics, for example. It would be desirable to have a program that provides these services in an efficient and user-friendly way, and that can be easily installed and run on all major operating systems. Although a large number of tree visualization tools are freely available, some as a part of more comprehensive analysis packages, all have drawbacks in one or more domains. They either lack some of the standard tree visualization techniques or basic graphics and editing features, or they are restricted to small trees containing only tens of thousands of taxa. Moreover, many programs are difficult to install or are not available for all common operating systems. **Results:** We have developed a new program, Dendroscope, for the interactive visualization and navigation of phylogenetic trees. The program provides all standard tree visualizations and is optimized to run interactively on trees containing hundreds of thousands of taxa. The program provides tree editing and graphics export capabilities. To support the inspection of large trees, Dendroscope offers

a magnification tool. The software is written in Java 1.4 and and installers are provided for Linux/Unix, MacOS X and Windows XP. **Conclusions:** Dendroscope is a user-friendly program for visualizing and navigating phylogenetic trees, for both small and large datasets.

Bibliography

D. F. Ackerley, T. T. Caradoc-Davies, and I. L. Lamont. Substrate specificity of the nonribosomal peptide synthetase pvdd from pseudomonas aeruginosa. *J Bacteriol*, 185(9):2848–2855, 2003.

A. Ahmadian, M. Ehn, and S. Hober. Pyrosequencing: history, biochemistry and future. *Clin Chim Acta*, 363(1-2):83–94, 2006. doi: 10.1016/j.cccn.2005.04.038.

A. Allsop and R. Illingworth. The impact of genomics and related technologies on the search for new antibiotics. *J Appl Microbiol*, 92(1):7–12, 2002.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, 1990. doi: 10.1006/jmbi.1990.9999.

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. *Nucleic Acids Res*, 25(17):3389–3402, 1997.

P. C. Appelbaum. MRSA – the tip of the iceberg. *Clin Microbiol Infect*, 12 suppl 2:3–10, 2006. doi: 10.1111/j.1469-0691.2006.01402.x.

R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh. Uniprot: the universal protein knowledgebase. *Nucleic Acids Res*, 32 Database issue:D115–D119, 2004.

K. Arima, A. Kakinuma, and G. Tamura. Surfactin, a crystalline peptidelipid surfactant produced by *Bacillus subtilis*: isolation, characterization and its inhibition of fibrin clot formation. *Biochem Biophys Res Commun*, 31(3):488–494, 1968.

Assurance Maladie. Les antibiotiques c'est pas automatique, 2007. URL `www.antibiotiquespasautomatiques.com`.

M. B. Austin and J. P. Noel. The chalcone synthase superfamily of type III polyketide synthases. *Nat Prod Rep*, 20(1):79–110, 2003.

T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 2:28–36, 1994.

P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

C. J. Balibar, F. H. Vaillancourt, and C. T. Walsh. Generation of D amino acid residues in assembly of arthrofactin by dual condensation/epimerization domains. *Chem Biol*, 12(11):1189–1200, 2005. doi: 10.1016/j.chembiol.2005.08.010.

R. Baltz. Antibiotic discovery from actinomycetes: will a renaissance follow the decline and fall? *SIM News*, 55:186196, 2005.

R. H. Baltz. Marcel Faber roundtable: is our antibiotic pipeline unproductive because of starvation, constipation or lack of inspiration? *J Ind Microbiol Biotechnol*, 33(7):507–513, 2006. doi: 10.1007/s10295-005-0077-9.

A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. The Pfam protein families database. *Nucleic Acids Res*, 32 Database issue:D138–D141, 2004.

J. M. Berg, J. L. Tymoczko, and L. Stryer. *Biochemistry*. W.H. Freeman, New York, NY, USA, 5th edition, 2002.

V. Bergendahl, U. Linne, and M. A. Marahiel. Mutational analysis of the C-domain in nonribosomal peptide synthesis. *Eur J Biochem*, 269(2):620–629, 2002.

F. Bertolla and P. Simonet. Horizontal gene transfers in the environment: natural transformation as a putative process for gene transfers between transgenic plants and microorganisms. *Res Microbiol*, 150(6):375–384, 1999.

A. Biegert, C. Mayer, M. Remmert, J. Söding, and A. N. Lupas. The MPI bioinformatics toolkit for protein sequence analysis. *Nucleic Acids Res*, 34 (Web Server issue):W335–W339, 2006. doi: 10.1093/nar/gkl217.

D. Bischoff, S. Pelzer, B. Bister, G. J. Nicholson, S. Stockert, M. Schirle, W. Wohlleben, G. Jung, and R. D. Süssmuth. The biosynthesis of vancomycin-type glycopeptide antibiotics – the order of the cyclization steps. *Angew Chem Int Ed Engl*, 40(24):4688–4691, 2001a.

D. Bischoff, S. Pelzer, A. Höltzel, G. J. Nicholson, S. Stockert, W. Wohlleben, G. Jung, and R. D. Süssmuth. The biosynthesis of vancomycin-type glycopeptide antibiotics – new insights into the cyclization steps. *Angew Chem Int Ed Engl*, 40(9):1693–1696, 2001b.

B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, 31(1):365–370, 2003.

L. Bonetta. Genome sequencing in the fast lane. *Nat Methods*, 3(2):141–147, 2006. doi: 10.1038/nmeth0206-141.

J. K. Borchardt. Combinatorial biosynthesis: Panning for pharmaceutical gold. *Modern Drug Discovery*, 2(4):22–29, 1999. URL `pubs.acs.org/hotartcl/mdd/99/aug/panning.html`.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on computational learning theory*, pages 144–152, New York, NY, USA, 1992. ACM Press. doi: 10.1145/130385.130401.

J. Brédy. Bioactive microbial metabolites. *J Antibiot (Tokyo)*, 58(1):1–26, 2005.

C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998. URL `research.microsoft.com/~cburges/papers/SVMTutorial.pdf`.

F. C. Cabello. Heavy use of prophylactic antibiotics in aquaculture: a growing problem for human and animal health and for the environment. *Environ Microbiol*, 8(7):1137–1144, 2006. doi: 10.1111/j.1462-2920.2006.01054.x.

A. A. Cantutescu, A. A. Shelenkov, and R. L. D. Jr. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, 12(9):20012014, 2003.

D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The amber biomolecular simulation programs. *J Comput Chem*, 26(16):1668–1688, 2005. doi: 10.1002/jcc.20290. URL `amber.scripps.edu`.

Centers for Disease Control and Prevention (CDC). Vancomycin-resistant *Staphylococcus aureus* – Pennsylvania, 2002. *MMWR Morb Mortal Wkly Rep*, 51(40):902, 2002.

G. L. Challis and D. A. Hopwood. Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by streptomyces species. *Proc Natl Acad Sci USA*, 100 suppl 2:14555–14561, 2003. doi: 10.1073/pnas.1934677100.

G. L. Challis, J. Ravel, and C. A. Townsend. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem Biol*, 7(3):211–224, 2000.

S. Chang, D. M. Sievert, J. C. Hageman, M. L. Boulton, F. C. Tenover, F. P. Downes, S. Shah, J. T. Rudrik, G. R. Pupp, W. J. Brown, D. Cardo, S. K. Fridkin, and Vancomycin-Resistant *Staphylococcus aureus* Investigative Team. Infection with vancomycin-resistant *Staphylococcus aureus* containing the *vanA* resistance gene. *N Engl J Med*, 348(14):1342–1347, 2003.

Z. Chang, P. Flatt, W. H. Gerwick, V. A. Nguyen, C. L. Willis, and D. H. Sherman. The barbamide biosynthetic gene cluster: a novel marine cyanobacterial system of mixed polyketide synthase (PKS)-non-ribosomal peptide synthetase (NRPS) origin involving an unusual trichloroleucyl starter unit. *Gene*, 296(1-2):235–247, 2002.

P. Y. Chou and G. D. Fasman. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol*, 47:45–148, 1978.

M. Clamp, J. Cuff, S. M. Searle, and G. J. Barton. The Jalview Java alignment editor. *Bioinformatics*, 20(3):426–427, 2004. doi: 10.1093/bioinformatics/btg430.

J. Clardy, M. A. Fischbach, and C. T. Walsh. New antibiotics from bacterial natural products. *Nat Biotechnol*, 24(12):1541–1550, 2006. doi: 10.1038/nbt1266.

S. L. Clugston, S. A. Sieber, M. A. Marahiel, and C. T. Walsh. Chirality of peptide bond-forming condensation domains in nonribosomal peptide synthetases: the C5 domain of tyrocidine synthetase is a $^{D}C_{L}$ catalyst. *Biochemistry*, 42(41):12095–12104, 2003. doi: 10.1021/bi035090+.

P. Collignon. Antibiotic growth promoters. *J Antimicrob Chemother*, 54 (1):272; author reply 276–272; author reply 278, 2004. doi: 10.1093/jac/dkh266.

E. Conti, T. Stachelhaus, M. A. Marahiel, and P. Brick. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J*, 16(14):4174–4183, 1997.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995. URL `citeseer.ist.psu.edu/cortes95supportvector.html`.

P. Cosmina, F. Rodriguez, F. de Ferra, G. Grandi, M. Perego, G. Venema, and D. van Sinderen. Sequence and analysis of the genetic locus responsible for surfactin synthesis in *Bacillus subtilis*. *Mol Microbiol*, 8(5):821–831, 1993.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, UK, 2000.

G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. Weblogo: a sequence logo generator. *Genome Res*, 14(6):1188–1190, 2004. doi: 10.1101/gr.849004.

M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, suppl. 3, pages 345–352. Natl Biomed Res Found, Washington, DC, USA, 1978.

V. de Crécy-Lagard, P. Marlière, and W. Saurin. Multienzymatic non ribosomal peptide biosynthesis: identification of the functional domains catalysing peptide elongation and epimerisation. *C R Acad Sci III*, 318 (9):927–936, 1995.

V. de Crécy-Lagard, V. Blanc, P. Gil, L. Naudin, S. Lorenzon, A. Famechon, N. Bamas-Jacques, J. Crouzet, and D. Thibaut. Pristinamycin I biosynthesis in *Streptomyces pristinaespiralis*: molecular characterization of the first two structural peptide synthetase genes. *J Bacteriol*, 179(3):705–713, 1997.

L. di Vincenzo, I. Grgurina, and S. Pascarella. *In silico* analysis of the adenylation domains of the freestanding enzymes belonging to the eucaryotic nonribosomal peptide synthetase-like family. *FEBS J*, 272:929–941, 2005.

J. P. Donnelly, A. Voss, W. Witte, and B. E. Murray. Does the use in animals of antimicrobial agents, including glycopeptide antibiotics, influence the efficacy of antimicrobial therapy in humans? *J Antimicrob Chemother*, 37 (2):389–392, 1996.

R. F. Doolittle. Similar amino acid sequences: chance or common ancestry? *Science*, 214(4517):149–159, 1981.

L. Du, M. Chen, C. Sánchez, and B. Shen. An oxidation domain in the BlmIII non-ribosomal peptide synthetase probably catalyzing thiazole formation in the biosynthesis of the anti-tumor drug bleomycin in *Streptomyces verticillus* ATCC15003. *FEMS Microbiol Lett*, 189(2):171–175, 2000.

R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK, 1998.

T. Eckstein, S. Chandrasekaran, S. Mahapatra, M. McNeil, D. Chatterjee, C. Rithner, P. Ryan, B. J.T., and J. Inamine. A major cell wall lipopeptide of *Mycobacterium avium* subspecies *paratuberculosis*. *J Biol Chem*, 281(8): 5209–5215, 2006.

S. R. Eddy. Where did the BLOSUM62 alignment score matrix come from? *Nat Biotechnol*, 22(8):1035–1036, 2004. doi: 10.1038/nbt0804-1035.

R. C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1):113, 2004a. doi: 10.1186/1471-2105-5-113.

R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–1797, 2004b. doi: 10.1093/nar/gkh340.

K. Eppelmann, T. Stachelhaus, and M. A. Marahiel. Exploitation of the selectivity-conferring code of nonribosomal peptide synthetases for the rational design of novel peptide antibiotics. *Biochemistry*, 41(30):9718–9726, 2002.

N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, and A. Sali. Comparative protein structure modeling using Modeller. *Curr Prot Bioinformatics*, unit 5.6, 2006. doi: 10.1002/0471250953.bi0506s15. URL `bioinfo.cipf.es/marcius/BBibTeX/pdfs/20061201_Narayanan_etal_CPBI-III.pdf`.

J. L. Fauchere, M. Charton, L. B. Kier, A. Verloop, and V. Pliska. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int J Pept Protein Res*, 32(4):269–278, 1988.

J. Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst Zool*, 22: 240–249, 1973.

J. Felsenstein. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39:783–791, 1985.

J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, MA, USA, 2004.

J. Felsenstein. *PHYLIP (PHYLogeny Inference Package) version 3.66 (distributed by the author)*. Department of Genome Sciences, University of Washington, Seattle, WA, USA, 2006. URL `evolution.genetics.washington.edu/phylip.html`.

A. V. Fiacco and G. P. McCormick. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. Society for Industrial & Applied Mathematics, 1987.

R. Finking and M. A. Marahiel. Biosynthesis of nonribosomal peptides. *Annu Rev Microbiol*, 58:453–488, 2004. doi: 10.1146/annurev.micro.58.030603.123615.

M. A. Fischbach and C. T. Walsh. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chem Rev*, 106(8):3468–3496, 2006. doi: 10.1021/cr0503097.

J. D. Fischer, J. Ponjavic, O. Kohlbacher, A. N. Lupas, and J. Söding. FRpred – a package for prediction of functional residues in protein multiple sequence alignments. In *Proceedings of the German Conference in Bioinformatics 2006 – Poster Abstracts*, 2006. URL `toolkit.tuebingen.mpg.de/frpred`.

W. Fitch. Toward defining the course of evolution: minimum change of specified tree topology. *Syst Zoology*, 20:406–416, 1971.

R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223):496–512, 1995.

R. Fletcher. *Practical methods of optimization*. Wiley Interscience, Chichester, NY, USA, 2nd edition, 1987.

C. Freiberg and H. Brötz-Oesterhelt. Functional genomics in antibacterial drug discovery. *Drug Discov Today*, 10(13):927–935, 2005. doi: 10.1016/S1359-6446(05)03474-4.

R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*, 47(7):1739–1749, 2004. doi: 10.1021/jm0306430.

W. Gish. Washington University BLAST (WU-BLAST) 2.0, 2006. URL `blast.wustl.edu`.

J. P. Graham, J. J. Boland, and E. Silbergeld. Growth promoting antibiotics in food animal production: an economic analysis. *Public Health Rep*, 122 (1):79–87, 2007. URL `www.publichealthreports.org/userfiles/122_1/13_PHR122-1_79-87.pdf`.

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

G. G. Grassi, A. Ferrara, A. Navone, and P. Sala. Effect of subinhibitory concentrations of antibiotics on the emergence of drug resistant bacteria *in vitro*. *J Antimicrob Chemother*, 6(2):217–223, 1980.

M. Gribskov, A. D. McLachlan, and D. Eisenberg. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci USA*, 84(13):4355–4358, 1987.

W. N. Grundy, T. L. Bailey, C. P. Elkan, and M. E. Baker. Meta-MEME: motif-based hidden Markov models of protein families. *Comput Appl Biosci*, 13(4):397–406, 1997. URL `metameme.sdsc.edu`.

X. Gu and J. Zhang. A simple method for estimating the parameter of substitution rate variation among sites. *Mol Biol Evol*, 14(11):1106–1113, 1997.

S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5):696–704, 2003.

D. H. Haft, J. D. Selengut, and O. White. The TIGRFAMs database of protein families. *Nucleic Acids Res*, 31(1):371–373, 2003.

T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, and J. L. Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem*, 47(7):1750–1759, 2004. doi: 10.1021/jm030644s.

P. R. Halmos. *A Hilbert Space Problem Book*. D. van Nostrand Company, Princeton, NJ, USA, 1967.

S. Hannenhalli and R. Russell. Analysis and prediction of functional subtypes from protein sequence alignments. *J Mol Biol*, 303(1):61–76, 2000. doi: 10.1006/jmbi.2000.4036.

S. F. Haydock, A. N. Appleyard, T. Mironenko, J. Lester, N. Scott, and P. F. Leadlay. Organization of the biosynthetic gene cluster for the macrolide concanamycin a in *Streptomyces neyagawaensis* ATCC 27449. *Microbiology*, 151(Pt 10):3161–3169, 2005. doi: 10.1099/mic.0.28194-0.

M. D. Hendy and D. Penny. Branch and bound algorithms to determine minimal evolutionary trees. *Math Biosci*, 59:277–290, 1982.

S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA*, 89(22):10915–10919, 1992.

D. M. Hillis and J. J. Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst Biol*, 42(2):182–192, 1993. doi: 10.2307/2992540.

K. Hiramatsu, H. Hanaki, T. Ino, K. Yabuta, T. Oguri, and F. C. Tenover. Methicillin-resistant *Staphylococcus aureus* clinical strain with reduced vancomycin susceptibility. *J Antimicrob Chemother*, 40(1):135–136, 1997.

D. Hoffmann, J. M. Hevel, R. E. Moore, and B. S. Moore. Sequence analysis and biochemical characterization of the nostopeptolide a biosynthetic gene cluster from *Nostoc* sp. GSV224. *Gene*, 311:171–180, 2003.

K. Hoffmann, E. Schneider-Scherzer, H. Kleinkauf, and R. Zocher. Purification and characterization of eucaryotic alanine racemase acting as key enzyme in cyclosporin biosynthesis. *J Biol Chem*, 269(17):12710–12714, 1994.

Z. Hojati, C. Milne, B. Harvey, L. Gordon, M. Borg, F. Flett, B. Wilkinson, P. Sidebottom, B. Rudd, M. Hayes, C. Smith, and J. Micklefield. Structure, biosynthetic origin, and engineered biosynthesis of calcium-dependent antibiotics from *Streptomyces coelicolor*. *Chem Biol*, 9(11):1175–1187, 2002.

I. Hoof. Phylogenetic analysis of multi-domain proteins in the bacterial order actinomycetales. Master's thesis, Wilhelm-Schickard-Institute for Informatics, University of Tübingen, 2006.

S. Horowitz and W. Griffin. Structural analysis of *Bacillus licheniformis* 86 surfactant. *J Ind Microbiol*, 7(1):45–52, 1991.

D. H. Huson. Phylogeny. In *Lecture Notes on Algorithms in Bioinformatics*. Daniel H. Huson, 2007. URL `www-ab.informatik.uni-tuebingen.de/teaching`.

D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*, 23(2):254–267, 2006. doi: 10.1093/molbev/msj030.

J. Irving, J. Whisstock, and A. Lesk. Protein structural alignments and functional genomics. *Proteins Struct Func Genet*, 42:378382, 2001.

T. Joachims. Transductive inference for text classification using Support Vector Machines. In I. Bratko and S. Dzeroski, editors, *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pages 200–209, Bled, Slovenia, 1999a. Morgan Kaufmann Publishers, San Francisco, CA. URL `www-ai.cs.uni-dortmund.de/DOKUMENTE/joachims_99c.ps.gz`.

T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods*, pages 169–184. MIT Press, Cambridge, MA, USA, 1999b.

A. P. Johnson, A. H. Uttley, N. Woodford, and R. C. George. Resistance to vancomycin and teicoplanin: an emerging clinical problem. *Clin Microbiol Rev*, 3(3):280–291, 1990.

D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, 8 (3):275–282, 1992.

T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, 1969.

O. V. Kalinina, A. A. Mironov, M. S. Gelfand, and A. B. Rakhmaninova. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci*, 13(2):443–456, 2004. doi: 10.1110/ps.03191704.

R. Karchin and R. Hughey. Weighting hidden Markov models for maximum discrimination. *Bioinformatics*, 14(9):772–782, 1998. URL `www.cse.ucsc.edu/research/compbio/sam.html`.

K. Karplus, C. Barrett, and R. Hughey. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856, 1998.

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

T. A. Keating, D. E. Ehmann, R. M. Kohli, C. G. Marshall, J. W. Trauger, and C. T. Walsh. Chain termination steps in nonribosomal peptide synthetase assembly lines: directed acyl-S-enzyme breakdown in antibiotic and siderophore biosynthesis. *Chembiochem*, 2(2):99–107, 2001.

T. A. Keating, C. G. Marshall, C. T. Walsh, and A. E. Keating. The structure of VibH represents nonribosomal peptide synthetase condensation, cyclization and epimerization domains. *Nat Struct Biol*, 9(7):522–526, 2002. doi: 10.1038/nsb810.

J. Kling. Ultrafast DNA sequencing. *Nat Biotechnol*, 21(12):1425–1427, 2003. doi: 10.1038/nbt1203-1425.

A. Koglin, M. R. Mofid, F. Löhr, B. Schäfer, V. V. Rogov, M.-M. Blum, T. Mittag, M. A. Marahiel, F. Bernhard, and V. Dötsch. Conformational switches modulate protein interactions in peptide antibiotic synthetases. *Science*, 312(5771):273–276, 2006. doi: 10.1126/science.1122928.

O. Kohlbacher and H. Lenhof. BALL–rapid software prototyping in computational molecular biology. Biochemicals Algorithms Library. *Bioinformatics*, 16(9):815–824, 2000.

R. M. Kohli and C. T. Walsh. Enzymology of acyl chain macrocyclization in natural product biosynthesis. *Chem Commun (Camb)*, (3):297–307, 2003.

A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis*. Prentice-Hall, Englewood Cliffs, NJ, USA, 1970.

D. Konz, S. Doekel, and M. Marahiel. Molecular and biochemical characterization of the protein template controlling biosynthesis of the lipopeptide lichenysin. *J Bacteriol*, 181(1):133–140, 1999.

I. Korf, M. Yandell, and J. Bedell. *BLAST. Basic Local Alignment Search Tool*. O'Reilly Media, Sebastopol, CA, USA, 2003.

R. H. Lambalot, A. M. Gehring, R. S. Flugel, P. Zuber, M. LaCelle, M. A. Marahiel, R. Reid, C. Khosla, and C. T. Walsh. A new enzyme superfamily – the phosphopantetheinyl transferases. *Chem Biol*, 3(11):923–936, 1996.

S. Lautru and G. L. Challis. Substrate recognition by nonribosomal peptide synthetase multi-enzymes. *Microbiology*, 150(Pt 6):1629–1636, 2004. doi: 10.1099/mic.0.26837-0.

B. M. Levitan. Hilbert space. In M. Hazewinkel, editor, *Encyclopaedia of Mathematics (online edition)*, page `eom.springer.de/H/h047380.htm`. Springer, 2002.

L. Li, E. I. Shakhnovich, and L. A. Mirny. Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. *Proc Natl Acad Sci USA*, 100(8):4463–4468, 2003. doi: 10.1073/pnas.0737647100.

R. H. Lilien, B. W. Stevens, A. C. Anderson, and B. R. Donald. A novel ensemble-based scoring and search algorithm for protein redesign, and its application to modify the substrate specificity of the gramicidin synthetase A phenylalanine adenylation enzyme. In D. Gusfield, P. Bourne, S. Istrail, P. Pevzner, and M. Waterman, editors, *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 46–57, San Diego, CA, USA, 2004. ACM Press, New York, NY, USA. doi: 10.1145/974614.974622.

K. Liolios, N. Tavernarakis, P. Hugenholtz, and N. C. Kyrpides. The genomes on line database (gold) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res*, 34(Database issue):D332–D334, 2006. doi: 10.1093/nar/gkj145.

M. G. Lorenz and W. Wackernagel. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol Rev*, 58(3):563–602, 1994.

K. Lu, R. Asano, and J. Davies. Antimicrobial resistance gene delivery in animal feeds. *Emerg Infect Dis*, 10(4):679–683, 2004.

L. Luo, R. M. Kohli, M. Onishi, U. Linne, M. A. Marahiel, and C. T. Walsh. Timing of epimerization and condensation reactions in nonribosomal peptide assembly lines: kinetic analysis of phenylalanine activating elongation modules of tyrocidine synthetase B. *Biochemistry*, 41(29):9184–9196, 2002.

S. Malhotra-Kumar, C. Lammens, S. Coenen, K. van Herck, and H. Goossens. Effect of azithromycin and clarithromycin therapy on pharyngeal carriage of macrolide-resistant streptococci in healthy volunteers: a randomised, double-blind, placebo-controlled study. *Lancet*, 369(9560):482–490, 2007. doi: 10.1016/S0140-6736(07)60235-9.

M. Marahiel, T. Stachelhaus, and H. Mootz. Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chem Rev*, 97(7):2651–2674, 1997.

F. Markowetz. *Support Vector Machines in Bioinformatics*. Master's thesis, Mathematics Department, University of Heidelberg, 2001. URL `citeseer.ist.psu.edu/markowetz01support.html`.

F. Markowetz. Klassifikation mit Support Vector Machines. Chapter 16 of the lecture "Genomische Datenanalyse". *Max-Planck-Institute for Molecular Genetics*, 2003. URL `lectures.molgen.mpg.de/statistik03/docs/Kapitel_16.pdf`.

F. Markowetz and R. Spang. Molecular diagnosis. Classification, model selection and performance evaluation. *Methods Inf Med*, 44(3):438–443, 2005. doi: 10.1267/METH05030438.

M. H. McCormick, W. M. Stark, G. E. Pittenger, R. C. Pittenger, and J. M. McGuire. Vancomycin, a new antibiotic. I. Chemical and biologic properties. In H. Welch and F. Marti-Ibanez, editors, *Antibiotics annual 1955-1956*, pages 606–611. Medical Encyclopedia Inc, New York, 1956.

D. McDevitt, D. J. Payne, D. J. Holmes, and M. Rosenberg. Novel targets for the future development of antibacterial agents. *J Appl Microbiol*, 92 Suppl:28S–34S, 2002.

L. C. McDonald, M. J. Kuehnert, F. C. Tenover, and W. R. Jarvis. Vancomycin-resistant enterococci outside the health-care setting: prevalence, sources, and public health implications. *Emerg Infect Dis*, 3(3): 311–317, 1997.

S. McGinnis and T. L. Madden. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, 32(Web Server issue): W20–W25, 2004. doi: 10.1093/nar/gkh435.

G. Miller and M. Lipman. Release of infectious Epstein-Barr virus by transformed marmoset leukocytes. *Proc Natl Acad Sci USA*, 70(1):190–194, 1973.

Y. Minowa, M. Araki, and M. Kanehisa. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J Mol Biol*, 2007. doi: 10.1016/j.jmb.2007.02.099.

A. Moll, A. Hildebrandt, H.-P. Lenhof, and O. Kohlbacher. BALLView: an object-oriented molecular visualization and modeling framework. *J Comput Aided Mol Des*, 19(11):791–800, 2005. doi: 10.1007/s10822-005-9027-x.

A. Moll, A. Hildebrandt, H.-P. Lenhof, and O. Kohlbacher. BALLView: a tool for research and education in molecular modeling. *Bioinformatics*, 22 (3):365–366, 2006. doi: 10.1093/bioinformatics/bti818.

H. D. Mootz and M. A. Marahiel. The tyrocidine biosynthesis operon of *Bacillus brevis*: complete nucleotide sequence and biochemical characterization of functional internal adenylation domains. *J Bacteriol*, 179(21): 6843–6850, 1997.

H. D. Mootz, D. Schwarzer, and M. A. Marahiel. Construction of hybrid peptide synthetases by module and domain fusions. *Proc Natl Acad Sci USA*, 97(11):5848–5853, 2000. doi: 10.1073/pnas.100075897.

H. D. Mootz, N. Kessler, U. Linne, K. Eppelmann, D. Schwarzer, and M. A. Marahiel. Decreasing the ring size of a cyclic nonribosomal peptide antibiotic by in-frame module deletion in the biosynthetic genes. *J Am Chem Soc*, 124(37):10980–10981, 2002a.

H. D. Mootz, D. Schwarzer, and M. A. Marahiel. Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *Chembiochem*, 3(6):490–504, 2002b. doi: 3.0.CO;2-N.

B. Morgenstern. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15(3):211–218, 1999. doi: 10.1093/bioinformatics/15.3.211.

K. Morik, P. Brockhausen, and T. Joachims. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML '99)*, pages 268–277, Bled, Slovenia, 1999. Morgan Kaufmann Publishers, San Francisco, CA. ISBN 1-55860-612-2.

M. Morikawa, H. Daido, T. Takao, S. Murata, Y. Shimonishi, and T. Imanaka. A new lipopeptide biosurfactant produced by *Arthrobacter* sp. strain MIS38. *J Bacteriol*, 175(20):6459–6466, 1993.

G. Morris, D. Goodsell, and R. Halliday. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem*, 19:1639–1662, 1998.

G. M. Morris, D. S. Goodsell, R. Huey, W. E. Hart, S. Halliday, R. Belew, and A. J. Olson. Autodock 3 user's guide. *The Scripps Research Institute*, 2001. URL `autodock.scripps.edu/faqs-help/manual`.

K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12:181–201, 2001.

M. Nai and S. Kumar. *Biological Sequence Analysis*. Oxford University Press Inc, USA, 2000.

K. Nakai, A. Kidera, and M. Kanehisa. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng , Des Sel*, 2 (2):93–100, 1988.

A. Neumaier, W. Huyer, and E. Bornberg-Bauer. Hydrophobicity analysis of amino acids. *Faculty of Mathematics, University of Vienna*, 1999. URL `www.mat.univie.ac.at/~neum/software/protein/aminoacids.html`.

W. C. Noble, Z. Virani, and R. G. Cree. Co-transfer of vancomycin and other resistance genes from *Enterococcus faecalis* NCTC 12201 to *Staphylococcus aureus*. *FEMS Microbiol Lett*, 72(2):195–198, 1992.

W. S. Noble. Support Vector Machine applications in computational biology. In B. Schölkopf, K. Tsuda, and J. Vert, editors, *Kernel Methods in Computational Biology*, chapter 3, pages 71–92. MIT Press, Cambridge, MA, USA, 2004. URL `noble.gs.washington.edu/papers/noble_support.pdf`.

C. Notredame, D. G. Higgins, and J. Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1): 205–217, 2000.

J. Parascandola. From germs to genes: trends in drug therapy, 1852-2002. *Pharm Hist*, 44(1):3–11, 2002.

H. M. Patel and C. T. Walsh. *In vitro* reconstitution of the *Pseudomonas aeruginosa* nonribosomal peptide synthesis of pyochelin: characterization of backbone tailoring thiazoline reductase and N-methyltransferase activities. *Biochemistry*, 40(30):9023–9031, 2001.

Paul Ehrlich Society. Antibiotikaverbot in der Tiermast verhindert Resistenzen bei Durchfallerregern. *Aktuelles der Paul-Ehrlich-Gesellschaft für Chemotherapie e.V.*, 184: `www.p--e--g.org/aktuelles/184`, 2006.

D. J. Payne, M. N. Gwynn, D. J. Holmes, and M. Rosenberg. Genomic approaches to antibacterial discovery. *Methods Mol Biol*, 266:231–259, 2004. doi: 10.1385/1-59259-763-7:231.

A. Pertsemlidis and J. W. Fondon. Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome Biol*, 2(10):REVIEWS2002, 2001.

H. N. Poinar, C. Schwarz, J. Qi, B. Shapiro, R. D. E. Macphee, B. Buigues, A. Tikhonov, D. H. Huson, L. P. Tomsho, A. Auch, M. Rampp, W. Miller, and S. C. Schuster. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*, 311(5759):392–394, 2006. doi: 10.1126/science.1123360.

A. Radzicka and R. Wolfenden. Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *J Biochem*, 27:1664–1670, 1988.

G. Rätsch. Gunnar Rätsch's group at the Friedrich Miescher Laboratory of the Max Planck Society, 2007. URL `www.fml.mpg.de/raetsch`.

C. Rausch, T. Weber, O. Kohlbacher, W. Wohlleben, and D. H. Huson. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res*, 33(18):5799–5808, 2005. doi: 10.1093/nar/gki885.

C. Rausch, I. Hoof, T. Weber, W. Wohlleben, and D. H. Huson. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol Biol*, 7:78, 2007. doi: 10.1186/1471-2148-7-78.

W. C. Roberts. Facts and ideas from anywhere – antibiotic-resistant organisms grown from retail meat samples. *Proc (Bayl Univ Med Cent)*, 15(1): 107–118, 2002.

E. D. Roche and C. T. Walsh. Dissection of the EntF condensation domain boundary and active site residues in nonribosomal peptide synthesis. *Biochemistry*, 42(5):1334–1344, 2003. doi: 10.1021/bi026867m.

N. Roongsawang, S. P. Lim, K. Washio, K. Takano, S. Kanaya, and M. Morikawa. Phylogenetic analysis of condensation domains in the nonribosomal peptide synthetases. *FEMS Microbiol Lett*, 252(1):143–151, 2005. doi: 10.1016/j.femsle.2005.08.041.

B. Rost. Twilight zone of protein sequence alignments. *Protein Eng*, 12(2): 85–94, 1999.

M. Röttig. *Klassifikation von Enzymen mittels Support Vector Machines und Partial Least Squares*. Master's thesis, Wilhelm-Schickard-Institute for Informatics, University of Tübingen, 2006.

M. Röttig, C. Rausch, D. H. Huson, and O. Kohlbacher. Active Site Classification (ASC) – a webserver for classifcation of enzyme substrate specificity. In *Proceedings of the 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) – Poster Abstracts*, 2007.

F. W. Rusnak, S. Faraci, and C. T. Walsh. Subcloning, expression, and purification of the enterobactin biosynthetic enzyme 2,3-dihydroxybenzoate-AMP ligase: demonstration of enzyme-bound (2,3-dihydroxybenzoyl)adenylate product. *Biochemistry*, 28(17):6827–6835, 1989.

N. Saitou and M. Nei. The Neighbor-Joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, 1987.

S. A. Samel, B. Wagner, M. A. Marahiel, and L.-O. Essen. The thioesterase domain of the fengycin biosynthesis cluster: a structural base for the macrocyclization of a non-ribosomal lipopeptide. *J Mol Biol*, 359(4):876–889, 2006. doi: 10.1016/j.jmb.2006.03.062.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, 74(12):5463–5467, 1977.

F. Schauwecker, F. Pfennig, W. Schröder, and U. Keller. Molecular cloning of the actinomycin synthetase gene cluster from *Streptomyces chrysomallus* and functional heterologous expression of the gene encoding actinomycin synthetase II. *J Bacteriol*, 180(9):2468–2474, 1998.

M. Schmid. Microbial genomics – new targets, new drugs. *Expert Opin Ther Targets*, 5(4):465–475, 2001. doi: 10.1517/14728222.5.4.465.

B. Schölkopf and A. J. Smola, editors. *Learning with Kernels*. MIT Press, 2002.

B. Schölkopf, I. Guyon, and J. Weston. Statistical learning and kernel methods in bioinformatics. In P. Frasconi and R. Shamir, editors, *Artificial Intelligence and Heuristic Methods in Bioinformatics*, volume 183 of *NATO Science Series: Computer & Systems Sciences*, pages 1–21. IOS Press, Amsterdam, the Netherlands, 2003. URL `www.clopinet.com/isabelle/Papers/kerbioinfo.pdf`.

B. Schölkopf, K. Tsuda, and J. Vert, editors. *Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA, USA, 2004.

G. Schönafinger, N. Schracke, U. Linne, and M. A. Marahiel. Formylation domain: an essential modifying enzyme for the nonribosomal biosynthesis of linear gramicidin. *J Am Chem Soc*, 128(23):7406–7407, 2006. doi: 10.1021/ja0611240.

N. Schracke. *Die molekulare Logik der nichtribosomalen Peptidsynthetasen: Identifizierung und biochemische Charakterisierung der Biosynthesegene für Gramicidin A*. PhD thesis, Fachbereich Biologie, Philipps-Universität Marburg, Germany, 2005. URL `archiv.ub.uni-marburg.de/diss/z2005/0092`.

Schrödinger, Inc., editor. *FirstDiscovery Technical Notes*. Schrödinger, Inc., Portland, OG, USA, 2003. URL `www.schrodinger.com->Products->Liaison`.

B. Schuster-Böckler, J. Schultz, and S. Rahmann. HMM logos for visualization of protein families. *BMC Bioinformatics*, 5:7, 2004. doi: 10.1186/1471-2105-5-7.

T. Schwecke, K. Göttling, P. Durek, I. D. nas, N. F. Käufer, S. Zock-Emmenthal, E. Staub, T. Neuhof, R. Dieckmann, and H. von Döhren. Nonribosomal peptide synthesis in *Schizosaccharomyces pombe* and the architectures of ferrichrome-type siderophore synthetases in fungi. *Chem Bio Chem*, 7(4):612622, 2006. doi: 10.1002/cbic.200500301.

S. A. Sieber and M. A. Marahiel. Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics. *Chem Rev*, 105 (2):715–738, 2005. doi: 10.1021/cr0301191.

R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438, 1958.

M. Sosio, S. Stinchi, F. Beltrametti, A. Lazzarini, and S. Donadio. The gene cluster for the biosynthesis of the glycopeptide antibiotic A40926 by *Nonomuraea* species. *Chem Biol*, 10(6):541–549, 2003.

T. Stachelhaus, A. Schneider, and M. A. Marahiel. Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. *Science*, 269(5220):69–72, 1995.

T. Stachelhaus, H. Mootz, V. Bergendahl, and M. Marahiel. Peptide bond formation in nonribosomal peptide biosynthesis. Catalytic role of the condensation domain. *J Biol Chem*, 273(35):22773–22781, 1998.

T. Stachelhaus, H. D. Mootz, and M. A. Marahiel. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol*, 6(8):493–505, 1999.

J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. R. Gilbert, I. Korf, H. Lapp, H. Lehväslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–1618, 2002. doi: 10.1101/gr.361602.

M. Szummer and T. Jaakkola. Partially labeled classification with Markov random walks. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, USA, 2002. MIT Press.

Y. Tang, C.-Y. Kim, I. I. Mathews, D. E. Cane, and C. Khosla. The 2.7-Ångström crystal structure of a 194-kDa homodimeric fragment of the 6-deoxyerythronolide B synthase. *Proc Natl Acad Sci USA*, 103(30):11124–11129, 2006. doi: 10.1073/pnas.0601924103.

W. R. Taylor. The classification of amino acid conservation. *J Theor Biol*, 119(2):205–218, 1986.

M. Thattai, Y. Burak, and B. I. Shraiman. The origins of specificity in polyketide synthase protein interactions. *PLoS Comp Biol*, submitted, 2007.

A. Thomas, R. Meurisse, and R. Brasseur. Aromatic side-chain interactions in proteins. II. Near- and far-sequence Phe-X pairs. *Proteins*, 48(4):635–644, 2002. doi: 10.1002/prot.10191.

J. Thompson, D. Higgins, and T. Gibson. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680, 1994.

D. Tillett, E. Dittmann, M. Erhard, H. von Döhren, T. Börner, and B. A. Neilan. Structural organization of microcystin biosynthesis in *Microcystis aeruginosa* PCC7806: an integrated peptide-polyketide synthetase system. *Chem Biol*, 7(10):753–764, 2000.

A. Tognoni, E. Franchi, C. Magistrelli, E. Colombo, P. Cosmina, and G. Grandi. A putative new peptide synthase operon in *Bacillus subtilis*: partial characterization. *Microbiology*, 141(Pt 3):645–648, 1995.

K. Tomii and M. Kanehisa. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng , Des Sel*, 9(1):27–36, 1996.

V. Tosato, A. Albertini, M. Zotti, S. Sonda, and C. Bruschi. Sequence completion, identification and definition of the fengycin operon in *Bacillus subtilis* 168. *Microbiology*, 143(Pt 11):3443–3450, 1997.

J. W. Trauger and C. T. Walsh. Heterologous expression in *Escherichia coli* of the first module of the nonribosomal peptide synthetase for chloroeremomycin, a vancomycin-type glycopeptide antibiotic. *Proc Natl Acad Sci USA*, 97(7):3112–3117, 2000.

J. Tsai, R. Taylor, C. Chothia, and M. Gerstein. The packing density in proteins: standard radii and volumes. *J Mol Biol*, 290(1):253–266, 1999.

K. Turgay, M. Krause, and M. A. Marahiel. Four homologous domains in the primary structure of GrsB are related to domains in a superfamily of adenylate-forming enzymes. *Mol Microbiol*, 6(4):529–546, 2743–2744, 1992.

F. H. Vaillancourt, E. Yeh, D. A. Vosburg, S. E. O'Connor, and C. T. Walsh. Cryptic chlorination by a non-haem iron enzyme during cyclopropyl amino acid biosynthesis. *Nature*, 436(7054):1191–1194, 2005. doi: 10.1038/nature03797.

V. Vapnik and A. Chervonenkis. A note on one class of perceptrons. *Automation and Remote Control*, 25, 1964.

V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24, 1963.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, NY, USA, 1995.

V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, NY, USA, 1998.

A. Velasco, P. Acebo, A. Gomez, C. Schleissner, P. Rodríguez, T. Aparicio, S. Conde, R. Muñoz, F. de la Calle, J. L. Garcia, and J. M. Sánchez-Puelles. Molecular characterization of the safracin biosynthetic pathway from *Pseudomonas fluorescens* A2-2: designing new cytotoxic compounds. *Mol Microbiol*, 56(1):144–154, 2005. doi: 10.1111/j.1365-2958.2004.04433.x.

J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, 2004. doi: 10.1126/science.1093857.

J.-P. Vert, K. Tsuda, and B. Schölkopf. A primer on kernel methods. In B. Schölkopf, K. Tsuda, and J. Vert, editors, *Kernel Methods in Computational Biology*, chapter 2, pages 35–70. MIT Press, Cambridge, MA, USA, 2004. URL `www.kyb.mpg.de/publications/pdfs/pdf2549.pdf`.

L. S. Vinh and A. von Haeseler. IQPNNI: moving fast through tree space and stopping in time. *Mol Biol Evol*, 21(8):1565–71, 2004. doi: 10.1093/molbev/msh176.

F. von Nussbaum, M. Brands, B. Hinzen, S. Weigand, and D. Häbich. Antibacterial natural products in medicinal chemistry – exodus or revival? *Angew Chem Int Ed Engl*, 45(31):5072–5129, 2006. doi: 10.1002/anie.200600350.

H. C. Wegener, F. M. Aarestrup, L. B. Jensen, A. M. Hammerum, and F. Bager. Use of antimicrobial growth promoters in food animals and *Enterococcus faecium* resistance to therapeutic antimicrobial drugs in europe. *Emerg Infect Dis*, 5(3):329–335, 1999.

J. Weston, F. Pérez-Cruz, O. Bousquet, O. Chapelle, A. Elisseeff, and B. Schölkopf. Feature selection and transduction for prediction of molecular bioactivity for drug design. *Bioinformatics*, 19(6):764–771, 2003.

A. Wiest, D. Grzegorski, B.-W. Xu, C. Goulard, S. Rebuffat, D. J. Ebbole, B. Bodo, and C. Kenerley. Identification of peptaibols from *Trichoderma virens* and cloning of a peptaibol synthetase. *J Biol Chem*, 277(23):20862–20868, 2002. doi: 10.1074/jbc.M201654200.

Wikipedia. Image:Lagrange_multiplier.png – Wikipedia, the free encyclopedia, 2007. URL `en.wikipedia.org/wiki/Image:Lagrange_multiplier.png`. [Online; accessed 23 February 2007].

W. Witte. Medical consequences of antibiotic use in agriculture. *Science*, 279(5353):996–997, 1998.

C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*, 34(Database issue):D187–D191, 2006. doi: 10.1093/nar/gkj161.

Y. Ye and A. Godzik. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, 19 Suppl 2:II246–II255, 2003.

Y. Ye and A. Godzik. Fatcat: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res*, 32(Web Server issue):W582–W585, 2004. doi: 10.1093/nar/gkh430. URL `fatcat.burnham.org`.

D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 16:321–328, 2004.

R. Zhou, R. A. Friesner, A. Ghosh, R. andRizzo, W. L. Jorgensen, and R. M. Levy. New linear interaction method for binding affinity calculations using a continuum solvent model. *J PhysChem B*, 105:10388–10397, 2001.

D. Zimmerman. Role of subtherapeutic levels of antimicrobials in pig production. *J Anim Sci*, 62(suppl. 3):6–17, 1986.

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

Lebens- und Bildungsweg

Christian Rausch, geboren am 10. April 1976 in Neuendettelsau

1982 - 1986	Besuch der Grundschule in Windsbach
1986 - 1995	Besuch des Johann-Sebastian-Bach-Gymnasiums in Windsbach (Humanistisches Gymnasium)
06/1995	Abitur, Note: sehr gut (1,5)
07/1995 - 06/1996	Grundwehrdienst in Budel/NL (Luftwaffenausbildungsregiment 2) und Feuchtwangen (Fernmelderegiment 72)
11/1996 - 10/1998	Studium des Chemieingenieurwesens an der Universität Erlangen-Nürnberg in Erlangen
10/1998	Vordiplom, Note: gut (2,1)
09/1998 - 09/2001	Studium der Biotechnologie an der École Supérieure de Biotechnologie de Strasbourg (Universität Louis Pasteur) in Strasbourg, Frankreich
01/2001 - 09/2001	Diplomarbeit bei New England Biolabs, Inc. in Beverly, Massachusetts, USA mit dem Titel: *Cloning and Characterization of the Restriction-Modification Systems AspCNI and BspCNI* (Betreuung durch Rick Morgan und Prof. Jean-Marc Jeltsch)
09/2001	Diplom in Biotechnologie / Diplôme d'Ingénieur en Biotechnologie, Note: sehr gut (1,3)
10/2001 - 12/2001	Reise nach Neukaledonien
seit 03/2002	Wissenschaftlicher Mitarbeiter bei Prof. Dr. Daniel H. Huson, Universität Tübingen, Institut für Informatik, Arbeitsbereich Algorithmen der Bioinformatik. Anfertigung einer Dissertation mit dem Titel: *Substrate Specificity Prediction of Enzymes and its Applications to Nonribosomal Peptide Synthetases* Ko-Betreuung durch Prof. Dr. Wolfgang Wohlleben, Mikrobiologisches Institut, Lehrstuhl Mikrobiologie/Biotechnologie