Computational Immunology: From MHC-peptide Binding to Immunotherapy

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.-Ing. Pierre Dönnes
aus Skövde

Tübingen
2006

Tag der mündlichen Qualifikation:	15.09.2006
Dekan:	Prof. Dr. Michael Diehl
1. Berichtstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichtstatter:	Prof. Dr. Hans-Peter Lenhof (Universität des Saarlandes)
3. Berichtstatter:	Prof. Dr. Hans-Georg Rammensee

Abstract

The human immune system provides effective protection against invading pathogens and cancer. Soluble antibodies can directly bind to extracellular antigens, whereas other mechanisms are needed for the recognition of virally infected or cancerous cells. Intracellular proteins are digested into smaller peptides, which are then displayed on the cell surface bound to major histocompatibility (MHC) class I molecules. Cytotoxic T (Tc) cells play an important role in the immune system since they can recognize MHC-peptide complexes and eliminate infected or abnormal cells.

The intracellular events leading to MHC-peptide presentation are collectively known as antigen processing. There are three main steps in the antigen processing pathway; digestion of proteins into peptides by proteasomes in the cytosol, transport of peptides into the endoplasmic reticulum (ER) by the transporters associated with antigen processing (TAP), and MHC-peptide complex formation. A detailed understanding of these processes is a prerequisite for rational peptide vaccine design aiming to efficiently activate Tc cells. This has motivated the development of computational methods dealing with the different steps of the antigen processing pathway.

Methods predicting MHC-peptide binding with relatively good accuracy exists, however, there is room for improvements. Less is known about the two preceding steps, protein cleavage and peptide transport. There is a need of methods addressing these steps. Furthermore, there is a lack of methods that consider the whole antigen processing pathway in an integrative manner.

The first part of this thesis describes different methods for predicting MHC-peptide binding. Support vector machines and decision trees are used to study a wide range of different MHC alleles. In a comparative study the SVM-based method SVMHC shows better prediction accuracy compared to the well-known SYFPEITHI and BIMAS methods. Additionally, a consensus method for predicting peptide binding to HLA-A*0201 is presented. Novel methods for prediction of proteasomal cleavage and TAP transport are presented. These show improved prediction accuracy in comparison to existing methods. The prediction methods addressing the individual steps of the processing pathway are integrated in the WAPP (whole antigen processing pathway) method. WAPP shows increased accuracy of MHC-peptide binding prediction by filtering out peptides not likely to be generated by the proteasome or transported by TAP.

Immunotherapy has proven useful in cancer therapy during recent years. The promising results include several successful reports using MHC-binding peptides in order to activate the immune system. In cancer immunotherapy these peptides typically originate from tumor-specific antigens (TSAs) or tumor-associated antigens (TAAs). The second part of this thesis describes an integrative analysis system for cancer-related data. CAP is used to analyze the effects of genetic variation and gene expression levels for raising autoimmune responses in cancer. This provides insights into the characteristics of TSAs and TAAs. Furthermore, TSAs are analyzed for potential MHC-binding peptides.

In conclusion, the individual methods presented here show improvement when compared to other similar methods. The integrated method WAPP modeling the whole antigen processing pathway is the first of its kind and shows promising results. Finally the combination of CAP and SVMHC prove the usefulness of integrative analysis coupled to prediction tools for finding peptide immunotherapy candidates.

Kurzzusammenfassung

Das menschliche Immunsystem stellt eine effektive Barriere zum Schutz vor Pathogenen und Krebserkrankungen dar. Extrazellulärer Antigene werden von löslichen Antikörper erkannt, die die Basis der humoralen Immunantwort bilden. Die Erkennung intrazelluläre Antigene, die für eine Immunantwort auf Virusinfektionen oder Krebs notwendig ist, erfolgt mit Hilfe andere Mechanismen: intrazelluläre Proteine werden zu kleinen Peptiden verdaut, die, an die Moleküle des Histokompatibilitätskomplexes (MHC Klasse I) binden und auf der Zelloberfläche präsentiert werden. Zytotoxische T-Zellen (Tc) erkennen MHC-Peptid-Komplexe die von körperfremden oder veränderten Antigenen stammen und eliminieren diese infizierten oder abnormalen Zellen.

Das intrazelluläre Vorgang, der zur MHC-Peptid-Präsentation führt, wird als Antigenprozessierung bezeichnet. Drei Schritte im Rahmen der Antigenprozessierung sind von besonderer Bedeutung: Verdau, Transport und MHC-Bindung. Im ersten Schritt verdauen Proteasen des Cytosols Proteine zu kurzen Peptiden. Diese werden im zweiten Schritt aktiv ins endoplasmatische Retikulum (ER) transportiert. Dort binden sie schließlich spezifisch an MHC-Moleküle. Ein detailliertes Verständnis und eine theoretische Modellierung dieser Schritte ist Voraussetzung für den computergestützten Entwurf von peptidbasierten Impfstoffen.

Für die Vorhersage der MHC-Peptidbindung existiert eine Reihe von Verfahren mit guter Vorhersagegenauigkeit, die aber immer noch Raum für Verbesserungen bieten. Wesentlich weniger gut sind die beiden anderen Schritte (Verdau zu Peptiden und Transport ins ER) vorhersagbar. Darüber hinaus fehlen Methoden, diese drei Schritte zu einer integrierten Vorhersage der Antigenprozessierung zusammenzuführen.

Der erste Teil dieser Arbeit beschreibt die unterschiedlichen Methoden zur Vorhersage der MHC-Peptidbindung. Zur Vorhersage der Bindung an eine Reihe unterschiedlicher Allele kommen Supportvektormaschinen (SVMs) und Entscheidungsbäume zum Einsatz. Die SVM-basierte Methode SVMHC bietet eine bessere Vorhersagegenauigkeit als die bekannten Methoden SYFPEITHI und BIMAS. Diese lässt sich durch Konsensusmethoden noch weiter steigern, wie am Beispiel für das Allels HLA-A*0201 gezeigt wird. Auch für die Vorhersage des Verdaus in Peptide und den Transport ins ER werden Vorhersagemodelle vorgestellt. Diese zeigen ebenfalls deutlich verbesserte Vorhersagequalität als vergleichbare Methoden. Die drei Einzelvorhersagen (Verdau, Transport, Bindung) werden schließlich in einer inte-

grierten Vorhersage der gesamten Prozessierung zusammengeführt: WAPP (Whole Antigen Processing Pathway). WAPP zeichnet sich ebenfalls durch eine verbesserte Vorhersagegenauigkeit aus, insbesondere aufgrund seiner geringeren Rate an falsch positiven Vorhersagen. Im Gegensatz zu reinen MHC-basierten Methoden kann die Peptide, die nicht verdaut oder transportiert werden, erkannt und ausfiltriert werden.

Immuntherapie hat sich in den letzten Jahren als ein vielversprechender Weg in der Krebsbekämpfung herausgestellt. Dabei wurden zum Beispiel MHC-Bindende Peptide eingesetzt, um das Immunsystem gegen Krebszellen zu aktivieren. In der Krebsimmuntherapie stammen diese Peptide üblicherweise von tumorspezifischen Antigenen (TSAs) und tumorassoziierten Antigenen (TAAs). Der zweite Teil der Arbeit beschreibt ein integriertes System zur Analyse krebsrelevanter Datensätze zur Unterstützung der Immuntherapie. Dazu kommt die in dieser Arbeit entwickelte Methode zur Vorhersage der Antigenprozessierung wieder zum Einsatz. Integriert wird diese Vorhersage in das Analysewerkzeug CAP, das die Integration und Analyse heterogener krebsrelevanter Datensätze ermöglicht. CAP wird verwendet, um den Einfluss von genetischer Variabilität und Genexpression auf die Entstehung einer Immunantwort gegen Krebs zu untersuchen. Diese Daten erlauben die Identifizierung von TSAs und TAAs, die dann wieder mit Hilfe von SVMHC auf ihre Immunrelevanz untersucht werden können.

Zusammenfassend zeigen die hier entwickelten Methoden gegenüber vorher bekannten Methoden deutlich verbesserte Vorhersagegenauigkeit. Die integrierte Vorhersagemethode WAPP ist die erste ihrer Art und liefert vielversprechende Ergebnisse. Die Kombination von SVMHC und CAP zeigt den Nutzen der beiden Methoden für die Identifizierung von Peptiden für die Immuntherapie.

Acknowledgements

First of all I want to thank Professors Oliver Kohlbacher and Hans-Peter Lenhof for their support during this work. Hans-Peter was the one that convinced us that Saarbrücken would be a good place to do a PhD. Together with the rest of the people at ZBI he gave us a good start in Germany. Working with Oliver has been a privilege in many ways and it has been a couple of interesting years being involved in the development of the Division for Simulation of Biological Systems. I would also like to thank Prof. Hans-Georg Rammensee for reviewing the work presented in this thesis.

Many thanks to the whole SBS group (Andreas, Marc, Annette, Torsten, Jana, Muriel, Nico, and Nora), it is always good to have nice colleagues. A special thanks to Marc and Andreas, I am very glad that you also left Saarland to live abroad.

I would also like to acknowledge some students that I had the privilege to work with. My warm thanks to Lena, Jochen, Claas, Demet, Angela, Jan, Chris, Alex, Mathias, and Manfred.

Many thanks also to my family and friends who makes life so much more interesting.

Finally, I would like to thank Annette and Emmy, my two companions on this long and winding road. Without you I would be nothing, with you I am everything.

Contents

1. Introduction

In daily life we are constantly attacked by invading pathogens such as virus and bacteria. Furthermore, normal cells might undergo transformation into tumor cells. As in most other higher vertebrates, the immune system provides efficient protection (immunity) against most of these events. The human immune system is highly complex and involves many different organs and cell types, which can be split into a less specific (innate) and more specific (adaptive) parts. The innate part of the immune system includes anatomic barriers (skin and mucous membranes) and phagocytic barriers (cells of the immune system that "engulf" invading pathogens in a rather unspecific manner). The adaptive part of the immune system is responsible for highly controlled recognition of antigens and for initiating the appropriate response mechanisms. Adaptive responses have the characteristics of antigen specificity, diversity, memory, and discrimination between self and non-self. There are some differences in terms of adaptive responses depending on the origin of the antigen. Extracellular pathogens are mostly identified and destroyed by antibody-mediated mechanisms (humoral immunity), whereas virus-infected or malignant cells are mainly eliminated by cytotoxic T (Tc) cells (cellular immunity).

In both cancer and viral infection, foreign proteins are present within the cell. These proteins, as well as normal cellular proteins, are digested into smaller peptides by proteases. The peptides then bind to major histocompatibility (MHC) molecules and the MHC-peptide complex can then displayed on the cell surface. This process results in a fingerprint of the current cellular proteome. Under normal conditions only self-peptides are presented on the cell surface. On the other hand, if the cell is infected, virus-specific peptides will be displayed and may act as an activation signal for nearby Tc cells.

MHC molecules have been known to play an important role in graft rejection and T-cell activation for a long time. However, not until the beginning of the 1990's it became clear that this process is mediated by MHC-bound peptides. Since then a large number of MHC-binding peptides have been identified and several X-ray structures of MHC-peptide complexes have

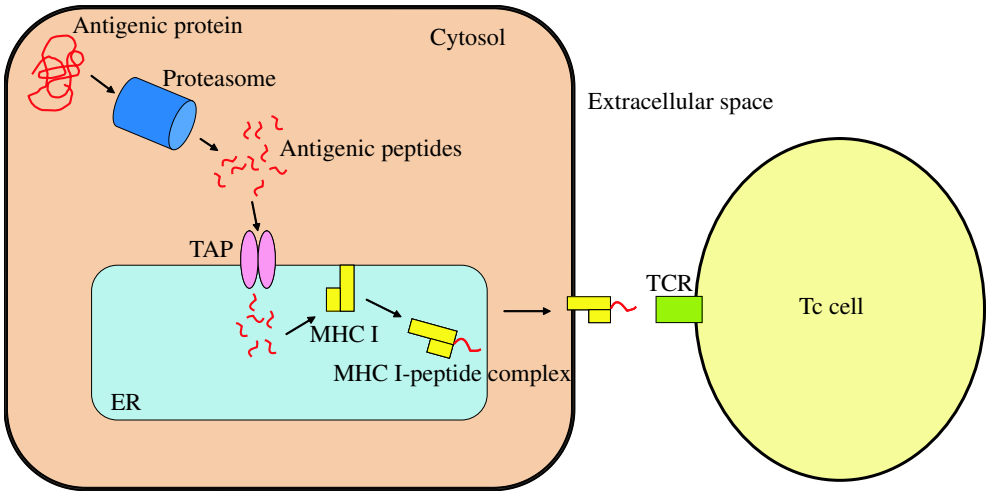

Figure 1.1.: An overview of the most important steps in the MHC class I antigen processing pathway. Antigenic proteins are cleaved into smaller peptides by proteasomes in the cytosol. The peptides can then be transported into the ER by TAP proteins. After MHC-peptide binding in the ER, the MHC-peptide complex is translocated to the cell surface where it can be recognized by a Tc cell, which binds to the MHC-peptide complex via a T-cell receptor (TCR).

been solved. These experimental data provide the basis for studying antigen processing and presentation.

Peptides originating from intracellular antigens bind to MHC class I molecules in the endoplasmic reticulum (ER). However, there are other intracellular events involved in the processing of these peptides before they actually bind to MHC molecules. Proteins in the cytosol are cleaved into smaller peptides by the proteasome, which is a huge protease complex mostly found in the cytosol. The peptides generated by proteasomal cleavage can then be transported into the ER, where they can bind to MHC molecules. The transporter associated with antigen processing (TAP), found in the ER membrane, can actively transport peptides from the cytosol into the ER. An overview of these most important events in the MHC class I antigen processing pathway is given in Fig. 1.1. A number of alternative processing mechanisms both for cleavage and transport have been presented. These include other cytosolic proteases and alternative transport routes into the ER, however, currently available experimental data shows that the proteasome and TAP play a key role in the antigen processing pathway.

In recent years, the potential of activating the immune system to fight both infectious disease and cancer have been explored. It has been known for more than 100 years that immune system activation by bacterial infection can lead to regression of solid human tumors,

but only recently these findings have resulted in clinical applications. Tc cells are especially important for immune responses in cancer. A reduced number of Tc cells typically imply a reduced protection against cancer [35]. When it became clear that MHC-binding peptides are a pre-requisite for Tc cell responses, several attempts have been made to use MHC-binding peptides as vaccines. Even though there are several problems that still need to be addressed, studies have shown an increased level of antigen-specific Tc cells from 0.1% to 2% of the total T cell population after treatment [35, 50, 208, 264]. These findings have motivated further analysis of MHC-binding peptides and have also led to the development of computational prediction methods. These aim to predict the MHC-binding peptides in a protein for a given MHC allele.

Proteasomal cleavage and TAP transport are also important for the processing and presentation of the MHC-binding peptides. Several studies have shown the importance of the proteasome to generate the correct peptides for MHC presentation [142, 228, 246]. The proteasome is mainly responsible for generating the correct C-terminal of these peptides, which has been considered by several approaches to predict proteasomal cleavage computationally.

A major problem faced by these approaches is that very little data is available for proteasome cleavage on the protein level. A wide range of studies have highlighted the importance of TAP in antigen processing [44, 120, 210, 265, 266]. Attempts have been made to predict the affinity between peptides and TAP, which have succeeded with reasonable accuracy. The results also show that the three N-terminal and C-terminal positions of the peptides are the most important for TAP affinity. The existing methods leave much room for improvements both regarding the underlying models and computational approaches.

It is desirable to find candidate proteins, from which peptides can be identified. Potential immunotherapy (vaccination) candidates can be derived from tumor-associated (TAA) or tumor-specific (TSA) antigens. Such antigens are of special interest, since they are mainly found within the tumor, reducing the risk of side effects. Unfortunately, only a limited number of TAAs and TSAs are currently characterized.

This thesis addresses three major questions regarding antigen processing and identification of immunotherapy candidates: (i) improved prediction of MHC-peptide binding, (ii) an integrated model of the major events in the MHC class I antigen processing pathway, and (iii) an analysis tool for cancer-related data that can be used to identify TAAs and TSAs.

Chapter 2 gives some theoretical background of both the underlying biology and computational methods used in this thesis. Some concepts of immunology are outlined and a special

focus is given to MHC-peptide interaction and the role of the immune system in cancer, motivating the work presented in this thesis from an immunological point of view. Furthermore, this chapter describes existing computational methods and available experimental data, and gives an overview of the major computational approaches. The machine learning methods support vector machines (SVMs) and decision trees (DTs) are described in detail, as well as the statistical measure employed to evaluate and compare the quality of prediction results.

MHC-peptide binding prediction, the final and most important step of the antigen processing pathway is the focus of Chapter 3. This is an area where many different computational methods have been proposed for predicting MHC-peptide affinity. Most existing methods use position-specific scoring matrices (PSSMs) for prediction, which assume an independent contribution to the overall binding energy from each amino acid of the peptide. Here, an SVM-based approach, SVMHC, is described that outperforms the frequently used BIMAS [201] and SYFPEITHI methods [217]. Several computational and statistical aspects for applying SVMs for this prediction task are discussed. This chapter also introduces a consensus prediction method for HLA-A*0201 binding peptides. Here the prediction results of three different methods are combined into a consensus score, further improving the prediction accuracy. Finally DTs and biophysical properties of amino acids can be used to construct a prediction method. The advantage with this approach is the biological interpretability of the decision rules generated from the DTs.

An integrated method for the whole antigen processing pathway, WAPP, is presented in Chapter 4. The major steps considered here are proteasomal cleavage, TAP transport, and MHC-peptide binding. A PSSM-based method, which is more stable than other artificial neural network previously presented, are used for proteasomal cleavage prediction. For prediction of TAP affinity, support vector regression (SVR) is used, resulting in the SVMTAP method which shows improved performance compared to the stabilized matrix method (SMM) presented by Peters *et al.* [204]. Finally, the methods for proteasomal cleavage and TAP transport are combined with SVMHC to form WAPP. The integration of the different prediction methods shows an improved performance for several MHC alleles.

Chapter 5 describes how heterogeneous cancer-related data can be integrated and analyzed in the CAP database. Analyzing data from different aspects of cancer, such as immunology and genetics, can facilitate in the identification of immunotherapy candidates. The technical aspects of integrating heterogeneous data, but also the data integrated into the database and prediction methods used to functionality annotate the data, are described in detail. A

large-scale analysis investigating the correlation of autoimmune responses in cancer to gene expression levels and genetic modifications is also carried out. The conclusion from this study is that gene expression levels seem to influence the immune response to certain cancer-associated genes, whereas no evidence of the influence of genetic variation could be found. Finally, the SVMHC method is used to analyze several TSAs, which serves as a qualitative validation of vaccine candidate identification process.

There are still many open questions regarding antigen processing and we are still far away from completely understanding cancer. The title of this thesis "From MHC-peptide binding to Immunotherapy" also reflects this transition from the reasonably well-defined problem of predicting MHC-binding peptide, to an area where many challenges still have to overcome. However, the work presented here clearly shows the usefulness of *in silico* methods in immunology going from MHC-peptide binding, over an integrated model of antigen processing, to identification of immunotherapy candidates.

2. Background and theory

This chapter introduces the most important biological and computational theory on which this thesis is based. The most fundamental aspects of immunology are described, but focus is of course kept on the parts and concepts important for this work. The interested reader can find more detailed introductions to immunology in several well-written textbooks [1, 98, 132, 231]. An overview of existing computational methods and available data is also given. This chapter also introduces the main computational and statistical methods used in this thesis. Some focus is given to the machine learning methods support vector machines (SVMs) and decision trees (DTs). Furthermore, statistical methods for evaluating the accuracy of prediction models are outlined.

2.1. Immunology

The field of immunology studies our protection (immunity) against foreign molecules, invading micro-organisms, and aberrant cells. These can all be recognized and eliminated by immune responses elicited by the various cells and molecules of the immune system. Early immunological studies were based on simple observations, whereas immunological process today can be described at the cellular or even molecular level.

The first documented observation of immunity goes all the way back to the Peloponnesian war, almost 450 years *BC*. Around this time, plague was tormenting Athens and Thucydides described how people who had recovered from the disease could nurse the sick without being affected again. The word **immunity** itself comes from the Latin word *immunis* meaning "exempt". In the 15th century, dried crust of smallpox was used to induce immunity by the Chinese and the Turks, something refined by Edward Jenner in the end of the 18th century. Jenner observed that the fluid from a cowpox pustule could induce immunity against smallpox and the first real **vaccine** was invented. Vaccination was further refined by Louis Pasteur who managed to use weakened forms of different pathogens, such as cholera and anthrax,

as vaccines. The experiments of Pasteur are often considered as the start of the field of immunology.

During the last 100 years, technological advances have made it possible to understand immunology in a different way. The importance of this knowledge is in some way reflected by the many Noble Prizes awarded to immunologists. Examples of important findings are the role of blood groups and the genes responsible for graft rejection.

The next sections give an overview of the immune system and describe adaptive immunity in detail. Some important aspects of the immune system in cancer and immunotherapy are also pointed out.

2.1.1. The immune system

The immune system is an effective defense system that has evolved in vertebrates as a defense against pathogens and cancer. The function of the immune system can be split into recognition and response. In recognition, the immune system should be able to recognize a wide variety of foreign cells and molecules, and at the same time be able to differentiate between them and the host's own cells and proteins [126]. Cells and molecules foreign to the host are referred to as antigens (Ags). Once an antigen has been recognized, the wide variety of organs, cells, and molecules have to induce the correct immune response to provide protection, i.e. immunity. The organs of the immune system can be classified into primary and secondary lymphoid tissues.

The primary lymphoid tissue is where the cells of the immune system develop and mature, whereas the secondary tissues function to trap antigens and provide an environment for efficient antigen recognition. White blood cells (leukocytes) constantly circulate the blood and the lymphoid system, where they play a central role in the specific and selective responses to antigens. The variety of the immune system enables an almost unlimited ability to recognize foreign invaders. All cells and molecules work together in an extremely dynamic network.

Immunity can be divided into **innate immunity** (natural immunity) and **adaptive immunity** (specific immunity), see Fig. 2.1. Innate immunity provides a first defense against infection and is not specific to certain pathogens. Components of innate immunity are anatomic barriers, physiological barriers, phagocytic/endocytic barriers, and inflammatory barriers [98]. Skin and mucous membranes are examples of anatomic barriers, whereas temperature and pH are physiological barriers. The phagocytic/endocytic barriers consist of cells that can internalize and break down foreign material. Cells involved in these processes are

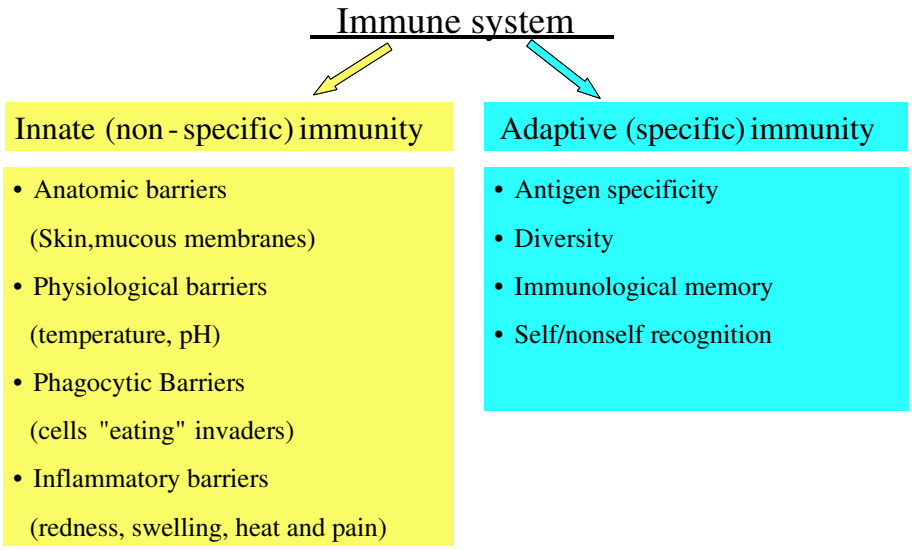

Figure 2.1.: A schematic overview of how the immune system can be split into innate and adaptive immunity. Innate immunity involves unspecific protection like anatomic and phagocytic barriers. Adaptive immunity is responsible for specific recognition and memory.

blood monocytes, neutrophils, and macrophages. Inflammation refers to the events occurring after tissue damage or infection leading to the release of various proteins attracting other immune cells, having antibacterial effects etc. Adaptive immunity on the other hand has the main features of antigen specificity, diversity, immunological memory, and self/non-self recognition. No further details are given on innate immunity, since the processes important for this thesis are part of the adaptive immune system.

2.1.2. Adaptive immunity

The four main characteristics of adaptive immunity were given in Fig. 2.1: The antigenic **specificity** allows for recognition of very small differences between antigens, where one amino acid difference between two proteins is enough. **Diversity** refers to the fact that the immune system can recognize more or less all foreign molecules. Both specificity and diversity are the effect of genetic recombination events generating an immense number of slightly different surface receptors on cells of the adaptive immune system. The ability of the immune system for a faster and heightened response when it is encountered with an antigen for a second time is called **immunologic memory**. Specialized memory cells carry out his function and triggering of these is the aim of vaccination. The last, but not the least important, feature of the adaptive immune system is its ability to **discriminate between self and**

non-self. Failure of this can lead to a variety of diseases such as Multiple Sclerosis (MS) and Rheumatoid Arthritis.

Adaptive immunity can be split into two branches, antibody-mediated and cell-mediated, that differ in both recognition and response to antigens. There are two types of lymphocytes, B lymphocytes (B cells) and T lymphocytes (T cells), responsible for these processes. B cells can directly recognize antigen by membrane-bound antibodies, whereas T cells only recognize antigenic peptides complexed with major histocompatibility (MHC) molecules on the surface of other cells. More details about MHC molecules and their characteristics in terms of peptide binding are given in Sect2.2. Furthermore a third category of cells of the adaptive immune system are specialized in "eating" antigen and presenting them to T cells. These are the antigen presenting cells (APCs). The most important APCs are macrophages and dendritic cells (DCs). Furthermore, B cells can also act as APCs. The major cell types are described in more detail below.

T lymphocytes

As mentioned before, T cells only recognize antigenic peptides bound to MHC molecules. There are also two major classes of T cells: cytotoxic T cells (Tc) and helper T cells (Th). Both of these cell types can recognize MHC-peptide complexes by means of T-cell receptors (TCRs). The two T cell types can be distinguished by two "cluster of differentiation" (CD) glycoproteins. Th has CD4 molecules that assist in the MHC-peptide-TCR binding and Tc has CD8 molecules. There are two classes of MHC molecules (class I and class II) involved in the activation of T cells. Tc cells recognize peptides from intracellular antigens bound to MHC class I molecules. MHC class I molecules are present on almost all nucleated cells and the peptides presented give a "fingerprint" of the cellular proteome. This enables detection of virus-infected and cancerous cells to which the Tc cells can induce apoptosis. Th cells recognize peptides originating from extracellular antigens that are presented by MHC class II molecules. The function of Th cells is to regulate the activity of other cells of the adaptive immunity and MHC class II molecules are almost exclusively found on APCs. The processes by which an antigenic peptide is presented by MHC molecules is a major focus of this thesis and these processes are described in detail in Sect. 2.3.

B lymphocytes

B lymphocytes express unique antigen binding receptors, antibodies, on their cell surface. Antibodies are glycoproteins consisting of two heavy and two light chains, held together by disulfide bonds. The N-terminal parts of the chains contain hypervariable regions make up the antigen binding sites. B cells that recognize an antigen can be activated, leading to rapid differentiation into memory B cells and plasma cells. Memory B cells are long-lived and continue to express the same membrane-bound antibody as its parent B cell. Plasma cells on the other hand can secrete soluble antibodies that help in the elimination of antigens. This type of antibody-mediated immune response is called *humoral* immunity. The plasma cells are rather short-lived, but it has been estimated that they might secrete up to 2000 antibody molecules per second. It has been estimated that there are between 10^7 to 10^9 different B cell clones possible due to genetic rearrangement in the antibody gene loci. This enables the antigenic diversity described earlier.

B cells can not only recognize and secrete antibodies towards specific antigens, they can also present antigens by MHC class II molecules in order to activate Th cells. Antigen uptake by means of membrane-bound antibodies is highly specific and enables effective delivery of antigen to the compartments used for antigen degradation [155, 299]. Furthermore, B cells are able to secrete a variety of co-stimulatory molecules needed for Th cell activation.

Macrophages

Macrophages can very effectively internalize antigens, up to 50 % of their surface area can be internalized in a single antigen uptake [267]. They express both MHC class I and MHC class II molecules, and can also produce different co-stimulatory molecules. However, the level of MHC class II molecules expressed are much lower than those of B cells and DCs [15, 175]. Further evidence suggests that macrophages are rather specialized in the clearance of antigens instead of presentation, which is their major function in innate immunity [2].

Dendritic cells

DCs are probably the most specialized APCs. They are enriched in regions of the immune system with a high concentration of T cells "waiting" to be activated [128, 176, 218]. They can also carry antigens to the secondary lymphoid organs [129]. Experiments in mice have show that DCs are needed for an adaptive immune response, something not true for

macrophages [137]. DCs can endocytose and present almost all forms of protein antigens on both MHC class I and MHC class II molecules. This type of effective "cross-presentation" is much more effective in DCs compared to other APCs [5, 67]. This is an important mechanism by which extracellular antigen are presented on MHC class I molecules or intracellular antigens presented on MHC class II molecules.

Main receptor types in adaptive immunity

There are four major molecules involved in the adaptive immune response described above:

- Membrane-bound antibodies on B-cells.
- T-cell receptors on T-cells
- MHC class I molecules present on all nucleated cells.
- MHC class II molecules on APCs.

A summary view of the most important receptor types in adaptive immunity can be seen in Fig. 2.2. Furthermore, this figure illustrates activation of both Tc and Th cells.

2.2. The major histocompatibility complex

As mentioned above, the T cells recognize antigenic peptides complexed with MHC molecules. This section focus on MHC molecules and gives an overview of their general role in the immune system. Furthermore, detailed descriptions of the structure and peptide-binding properties of MHC are given.

2.2.1. Overview and general aspects

The importance of MHC molecules was first discovered in the context of graft rejection. In the 1940s, Gorer and Snell discovered a set of genes involved in graft rejection. These genes were found to determine if a grafted tissue was accepted or not by the host and the term **histocompatibility genes** was coined. Snell was awarded the Noble Prize for these finding 1980. However, almost 20 years after the discovery of the MHC molecules, their only known function was in terms of graft rejection. In the 1960s and 1970s experiments showed that MHC molecules are important for an immune response against protein antigens, but the definitive proof of MHC restriction was given by Zinkernagel and Doherty in the 1970s [307].

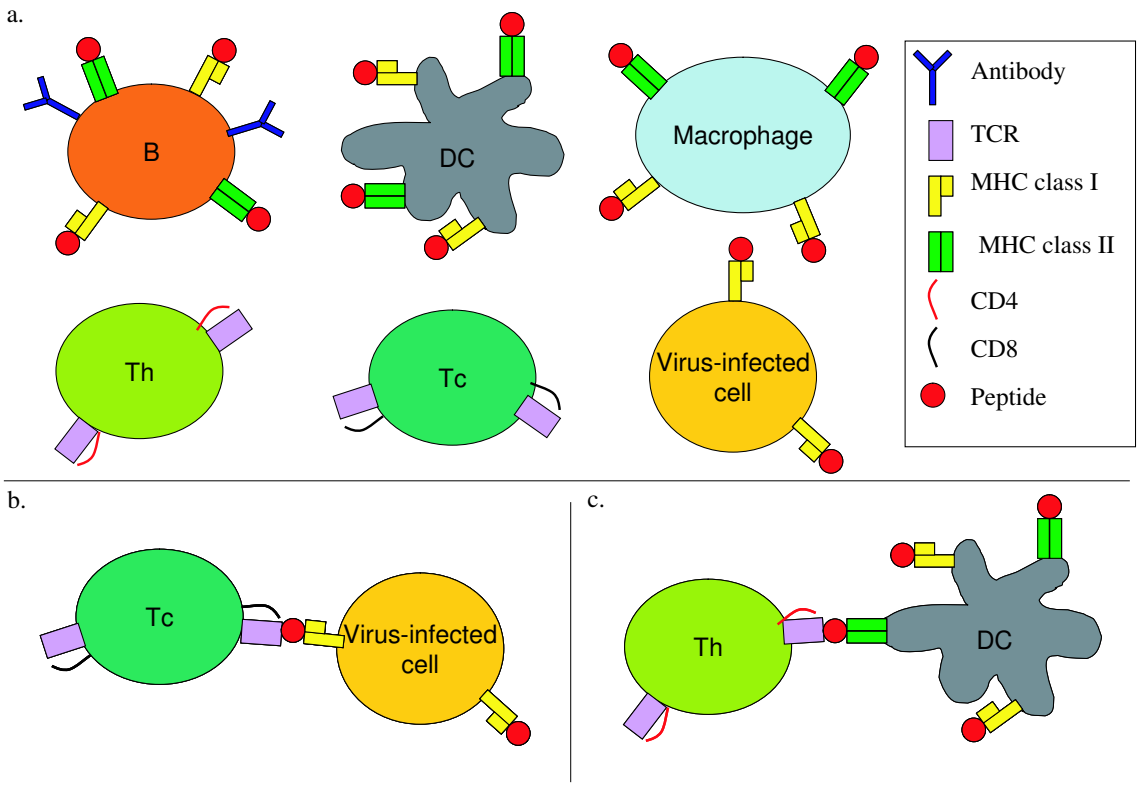

Figure 2.2.: (a.) The most important cell types and receptors involved in the adaptive immune system. Notice that the difference between Tc and Th cells are in their CD molecules. (b) Activation of Tc cells needs presentation of MHC class I molecules and involves an interaction of the MHC-peptide complex and TCR-CD8. In this example the antigenic peptide is presented by a virally infected cell. (c.) Specialized APCs, here a DC, can activate Th cells by presenting antigenic peptides bound to MHC class II molecules. Activation of Tc cells needs presentation of MHC class I molecules and involves an interaction of the MHC-peptide complex and TCR-CD8. For Th cells CD4 molecules are needed instead of CD8.

These experiments showed that Tc cells generated in a virus-infected mouse only elicit an immune response in hosts expressing the same MHC molecules as the animal from which they were generated.

The genes encoding MHC molecules are located on chromosome 6 in humans [294] (also referred to as human leukocyte antigen (HLAs)). In humans, three gene loci called A, B and C encode MHC class I molecules. Three other regions called DP, DQ and DR encode MHC II genes. There is also a region coding MHC class III genes which are proteins associated with the immune response, but not in the specific presentation to T cells. The MHC genes are highly polymorphic and each variant of a gene is called an allele. There are over 1200 HLA I alleles and over 700 HLA II alleles currently available in the IMGT/HLA sequence database [225]. Individuals expressing different MHC alleles are called **allogenic**.

The first crystal structure of an MHC molecule was solved by Wiley and colleagues [25] in 1987. However, at this point it was merely observed that the binding groove contained a collection of atoms and it was not clear that peptides can bind to MHC. In 1991, the group of H.-G. Rammensee presented evidence that MHC molecules bind smaller peptides and that these even have specific binding motifs [84, 235]. After this a wide variety of MHC-binding peptides have been identified and many structures of MHC-peptide complexes have been solved. Both classes of MHC molecules have an extracellular peptide-binding domain that is anchored in the cell membrane. Estimates suggest that several different MHC class I and class II molecules present over 10,000 different peptides at a level higher than 1 fmol per 10^8 cells [81, 125, 160]. The total number of MHC molecules per cell is thought to be in the range of 50,000 to 100,000 [93] and about $2{\cdot}10^6$ peptides are generated per second [211]. Attempts have been made to find out the number of MHC-peptide complexes needed for Tc-cell activation. However, this is a hard task since many factors like TCR-MHC affinity, MHC-peptide affinity, and availability of the peptide play a role. While some studies suggested that about 100-400 complexes are needed [51, 68, 105], more recent studies suggest that only three to five complexes are needed per cell [36, 272] and in the most extreme case a single complex is suggested to be enough [273].

The specific structure and peptide binding function of both MHC class I and class II molecules are now described.

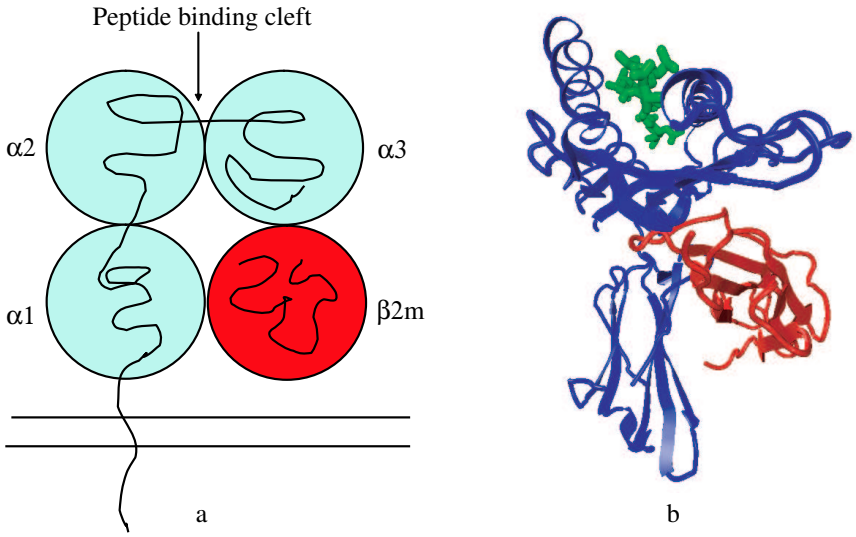

Figure 2.3.: (a.) A schematic overview of an MHC-peptide complex. The α-chain consist of three domains non-covalently associated with the β_2-microglobulin chain. The peptide biding cleft is formed between the $\alpha 1$ and $\alpha 2$ domains. (b.) The crystal structure of an MHC class I molecule (HLA-A*0201) with a bound peptide (PDB code: 1HHJ [165]).

2.2.2. MHC class I structure

MHC class I molecules consist of two non-covalently bound chains: the α chain (42-47 kDa) and the smaller β_2-microglobulin chain (β_2m) (12 kD). A schematic view and a crystal structure of an MHC molecule can be seen in Fig. 2.3. The α chain has about 75% of its total length in the extracellular space and only a small part is anchored in the cell membrane or cytosol. The α chain can be divided into three domains where the $\alpha 1$ and $\alpha 2$ domains, each about 90 amino acids long, form the peptide binding groove. The peptide binding groove is formed by eight anti-parallel β-sheets surrounded by two α-helices. The size of the binding groove (25 x 10 x 11 Å) can accommodate peptides with a length ranging from 8 to 11 amino acids. The ends of the binding groove are closed which makes it impossible for peptides to extend on either end. The polymorphic residues are mainly contained in the $\alpha 1$ and $\alpha 2$ domains where they contribute to a certain peptide-binding preference, meaning that different alleles typically bind different sets of peptides. Each individual expresses six different class I alleles in total (two allelic variants of the HLA-A, HLA-B, and HLA-C genes respectively). The $\alpha 3$ domain has a binding site for CD8, a surface molecule of Tc cells, but also contain about 25 hydrophobic amino acids that extend through the plasma membrane and a 30 residue long cytosolic part. The β_2-m domain is not encoded in the MHC locus and is invariant between all class I molecules.

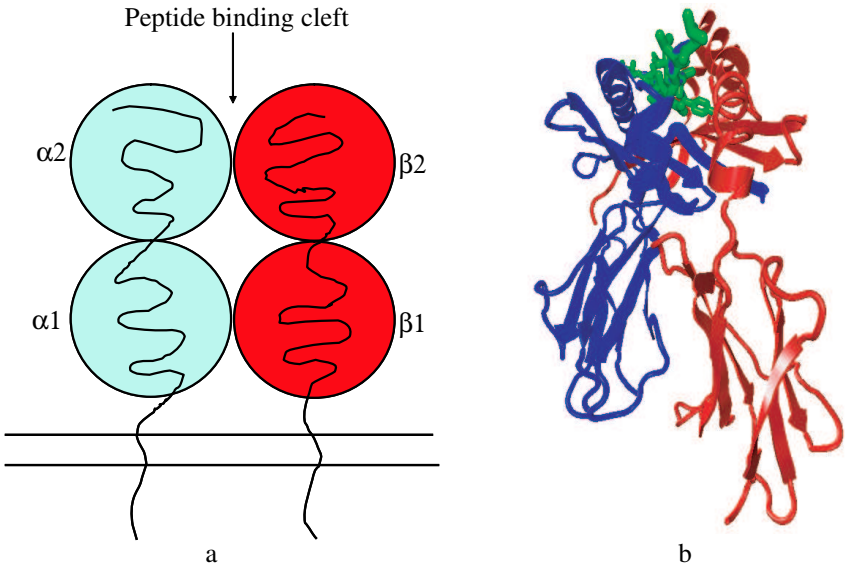

Figure 2.4.: (a.) An overview of the structure of an MHC class II molecule. It consists of an α and a β chain (both anchored in the membrane), which together form the peptide binding groove. b. The crystal structure of an MHC class II molecule (MHC Ia-G7) with a bound peptide (PDB code: 1JK8 [159])

2.2.3. MHC class II structure

MHC class II molecules consist of two non-covalently bound chains, the α chain (32-34 kD) and the β (29-32kD) chain, both encoded in the polymorphic MHC locus. An overview of the MHC class II structure can be seen in Fig. 2.4. The peptide binding groove is formed by the N-terminal ends of the α1 and β1 chains and the floor of the binding groove is formed by β-sheets from both chains. In contrast to MHC class I molecules, MHC class II have open ends of the binding groove, meaning that peptides can be longer and form "floppy" ends. This enables MHC class II molecules to bind peptides up to a length of about 30 amino acids, but the actual MHC-peptide interaction takes place in the binding groove and involves about nine amino acids of the peptide. The β2 domain contains a binding site for the CD4 surface molecules found on Th cells. The chains of MHC class II molecules mostly pair with chains from the same loci (e.g. DQα with DQβ), but they can also pair with chains from the other alleles (e.g. DQα can pair with DRβ). This pairing mechanism produces many variants of MHC class II molecules and the total number typically ranges between 10 and 20 in each individual. Both chains contain a membrane-spanning region of about 25 amino acids and a small cytoplasmic region.

MHC-peptide binding

A lot of effort has been put into trying to understand the mechanisms of MHC-peptide interaction and its role in T-cell activation. In principle this knowledge could be used to design optimal MHC-binding peptides that can elicit T-cell responses. The characteristics of MHC-peptide interaction has been analyzed by several different experimental approaches:

- Naturally presented peptides can be stripped of cells and analyzed by Edman degradation or mass spectrometry. This gives a qualitative representation of the peptides binding to a certain MHC allele (i.e. providing information about the amino acid sequence of the peptides).

- The affinity of MHC-peptide interactions can be estimated with competitive binding assays. This gives a quantitative view of the interaction.

- X-ray crystallography experiments gives insights into the interactions on an atomic level. At this level structural insights to the MHC-peptide interaction can be obtained.

- T-cell activation studies give immunological important insight into MHC-peptide binding. If T-cell activation occurs, this means that the peptide binds to MHC and can interact with the TCR.

The data from these different experiments provide the basis for understanding MHC-peptide interaction and also highlight the differences between class I and class II MHC molecules. The peptide binding domains of a MHC class I and a MHC class II molecule can be seen in Fig. 2.5. In the case of class I, the peptide is bound into a closed binding groove ("bathtub"), whereas in the class II case the ends of the binding groove are open and the peptide be extended at both ends ("hotdog"). In both cases the interactions in the complex are non-covalent.

MHC class I molecules typically bind peptides with a length of eight to ten amino acids. Peptides binding to a certain MHC class I molecule usually have conserved residues in some positions, referred to as anchor residues. These are deeply buried in well-defined binding pockets of the MHC molecule. Figure 2.6 shows a set of superimposed peptides extracted from crystal structures of HLA-A*0201-peptide complexes. It can be clearly seen that the conserved anchor residues are the ones in closest contact with the MHC molecule, whereas the mid-section residues bulges out from the binding groove.

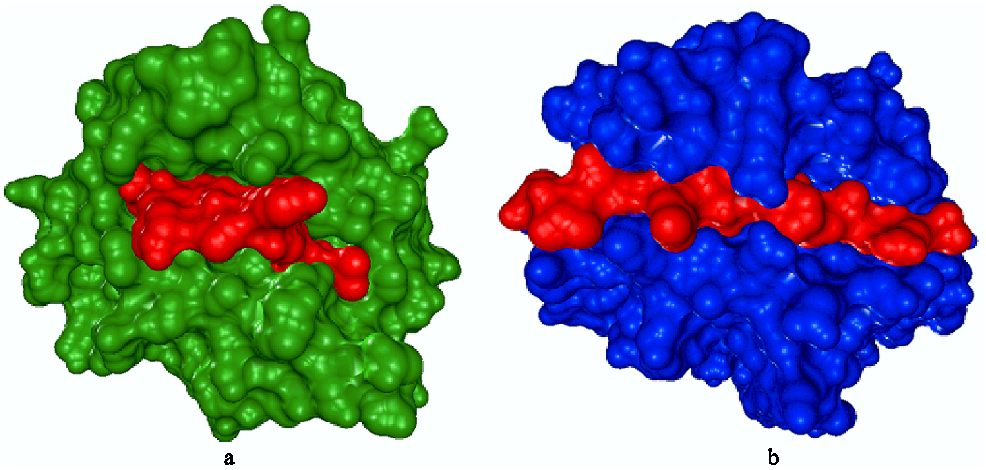

Figure 2.5.: (a.) MHC class I molecules have a binding cleft that is closed at both ends, meaning that only peptides with a limited length can bind. (b.) MHC class II molecules are open at both ends which means that the peptides can have "floppy ends" that are not in direct contact with the MHC molecules itself.

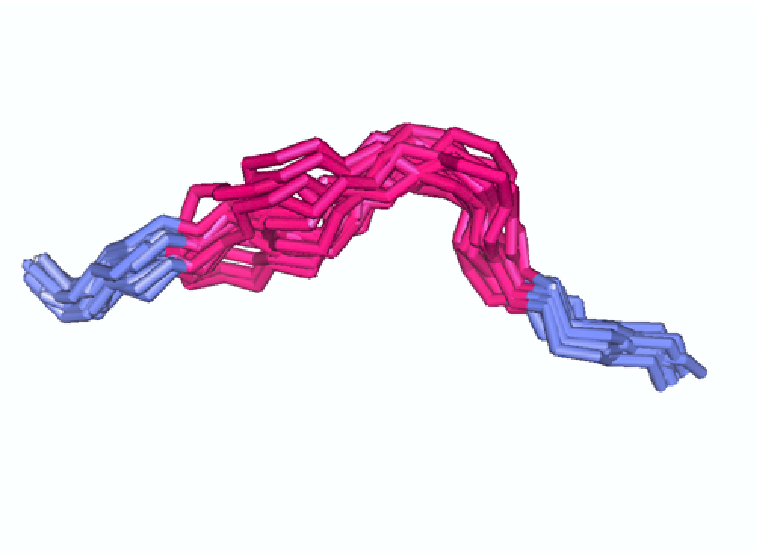

Figure 2.6.: Superpositioning of a set of HLA-A*0201 complexes binding nine amino acid long peptides [52]. It can be clearly seen that the backbone of these peptides is highly conserved in both terminal ends (blue) and a bit more variable in the middle part (pink). The superposition was done using the $\alpha 1$ and $\alpha 2$ domains of the structures.

MHC class II molecules bind peptides with lengths ranging from 10 to 30 amino acids. Most class II alleles do not have deep binding pockets, instead the MHC-peptide interaction is distributed along the whole peptide. These interactions are mainly hydrogen bonds between the peptide backbone and the α-helices surrounding the binding groove. Even though MHC class II binding peptides are rather long, the main interactions occur within a core of a similar length to that of MHC class I peptides.

Studies have also been made to elucidate the role of water in MHC-peptide interactions, assumed to be most prominent in the MHC class I case. Smith *et al.* classified water in the MHC-peptide interface into three categories: fixed, semi-fixed or variable [261]. Fixed water molecules are typically invariant between MHC molecules e.g. waters involved in the hydrogen bond network at the peptide N-terminus. Semi-fixed waters can be replaced by atoms from the MHC molecule or peptide and are considered to add variability in terms of preferred peptide types of a certain MHC molecule. Variable waters are found in the space between the peptide and the floor of the binding groove, indirectly anchoring the peptide main chain to the MHC molecule.

2.3. Antigen processing

Antigen processing refers to the mechanism by which peptides originating from antigenic proteins are finally displayed on the cell surface by MHC molecules. As described previously, T-cells can not recognize whole antigenic proteins and there are several processing events involved in generating the peptides presented by MHC molecules needed for T cell activation. The lengths of the peptides presented vary between MHC class I (8-10 amino acids) and MHC class II molecules (15-25). Studies have shown that even short peptides, such as in the class I case, in most cases represent a unique signature of a protein [42]. This uniqueness of the peptides enables the immune system to discriminate between self and foreign.

Two different processing pathways can be differentiated dependent on the origin of the antigen. Intracellular antigens are processed and presented by MHC class I molecules, whereas extracellular antigens are presented by MHC class II molecules. The following sections describe the processing of both intracellular and extracellular antigens. The focus will however be on the processing of intracellular antigens since this is the focus of this thesis.

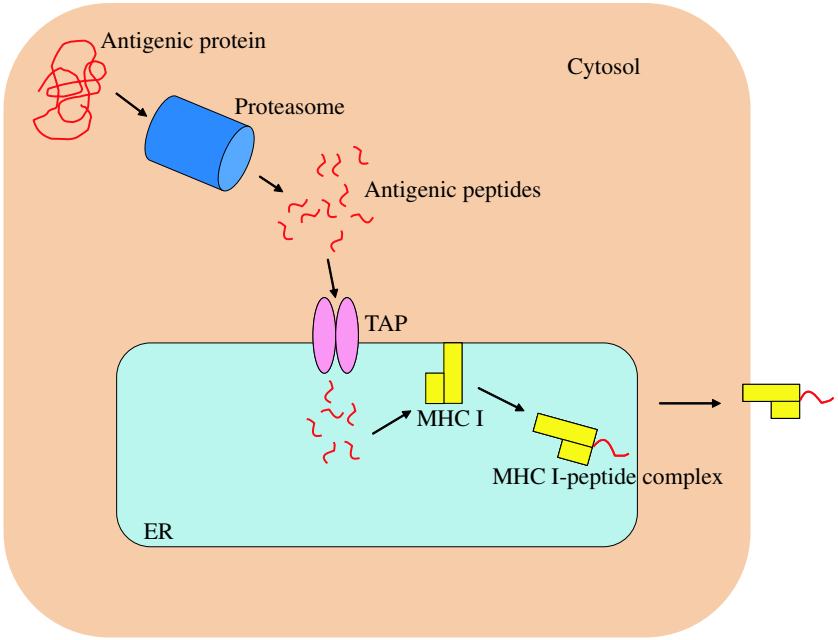

Figure 2.7.: An overview of the processing of intracellular antigens. Cytosolic proteins are cleaved into smaller peptides by the proteasome and can be transported into the ER by TAP. In the ER the peptides can associate with MHC molecules and subsequently be translocated to the cell surface for recognition by Tc cells.

2.3.1. Processing of intracellular antigens

Intracellular antigen processing involves three major steps: (i) protein cleavage into smaller peptides in the cytosol, (ii) peptide transport into the ER, and (iii) MHC peptide association. It is thought that the majority of the cytosolic antigens are cleaved by the proteasome and the smaller peptides, that then are transported into ER by TAP where they can to MHC molecules, see Fig. 2.7 for an overview. However, several alternative events have been suggested for both cleavage and transport. The following sections will describe the main steps of proteasomal cleavage and TAP transport in detail, followed by a section about alternative events.

The proteasome

Proteolytic events, cleavage of longer proteins into peptides and amino acids, occur in all cellular compartments and play a major role in cellular homeostasis. The variety of biological processes regulated by proteolytic enzymes includes degradation of misfolded proteins [83] and transcription factor activation [199]. Eukaryotes have several proteolytic systems e.g. lysosomal proteases, calpains, and proteasomes. The proteasome also generates peptides that can be presented by MHC molecules [228]. The importance of the proteasome in terms of MHC

presentation of antigenic peptides has been shown by using proteasome inhibitors, leading to a reduction of the amount of MHC-binding peptides presented on the cell surface [228]. Furthermore, *in vitro* studies of proteasomal degradation have shown its capability of generating known MHC-binding peptides from longer peptides [142, 246].

The process of protein degradation is ATP-dependent and is highly conserved from yeast to mammals [97]. There are two major forms of the proteasome, the 20S and 26S proteasomes. The 20S proteasome contains the proteolytic sites and is a barrel shaped molecule consisting of 28 subunits stacked into four rings [100], see Fig. 2.8. The subunits are evolutionary related and fall into two categories, α and β, dependent on sequence similarity. The outer rings consist of the subunits α1-α7, which form the substrate entry "gate". The two inner rings, β1-β7, each containing three catalytic sites. These catalytic sites have been described to have trypsin-like (cleavage after positively charged amino acids), chymotrypsin-like (cleavage after large hydrophobic residues), and peptidylglutamyl-peptide hydrolytic-like activity (cleavage after acidic amino acids) [285]. The activity of all three catalytic subunits rely on an N-terminal threonine, which has been proven by single-point mutation studies [30, 249].

The 26S proteasome consists of the 20S proteasome associated with two regulatory 19S particles. The 19S particle (also known as PA700) has two multi-subdomains consisting of six ATPase and one binding domain interacting with the 20S proteasome. The function of the ATPase domains is thought to be unfolding of the proteins in a chaperone-like way [33, 95].

Recent studies provide evidence that the 20S proteasome can be found in two different forms, the constitutive- (c20S) and the immuno- (i20S) proteasomes. During IFN-γ stimulation, three of the β-subunits can be replaced by three β_i subunits [227]. Two of these subunits, LMP-2 ($\beta 1_i$) and LMP-7 ($\beta 5_i$), are encoded next to the TAP1 and TAP2 loci [96, 140, 170]. The third subunit MECL-1 ($\beta 2_i$) is not encoded within the MHC locus [114]. Some efforts have been made to elucidate the specific role of these subunits and mice missing the LMP-7 subunit were found to have defect presentation of some antigens and reduced MHC class I expression [184]. Furthermore, presentation of some antigens was also reduced in mice missing the LMP-2 subunit [292]. In general the immuno-proteasome has been found to enhance cleavage after basic and hydrophobic amino acids and to inhibit cleavage after acidic residues [4, 76, 92].

Most proteasomal substrates are marked for degradation by ubiquitination and a major source of proteins are defective ribosomal products (DRiPs), shown to be rapidly ubiquitinated and degraded by the proteasome [243]. The proteasome been shown responsible for

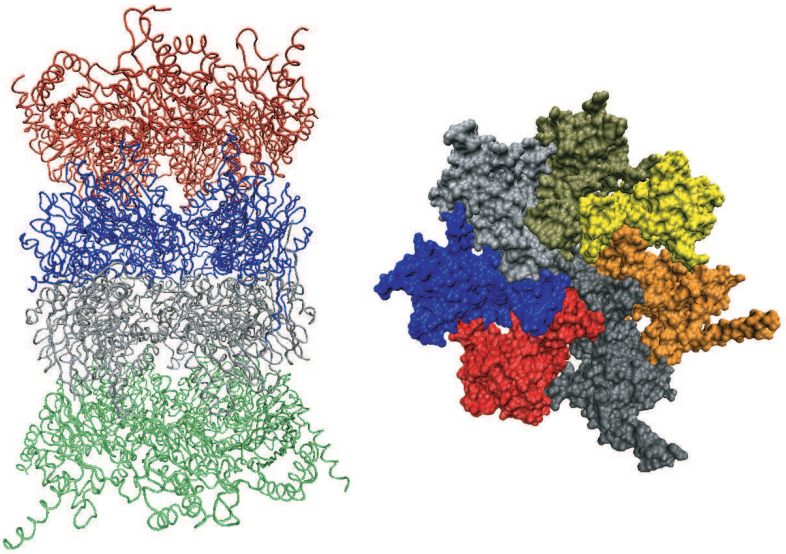

Figure 2.8.: (a.) The proteasome with the four subunits forming the barrel-shaped protease complex in different colors. The catalytic sites are within the two inner β subunits. (b.) One of the α rings of the proteasome highlighting the seven different subunits. The entry into the proteasome is in the center, but in the conformation shown here the opening is closed.(PDB code: 1IRU)

generating the correct C-termini of MHC class I binding peptides in several studies [60, 190], but there are also several examples where the proteasome generates the correct N-termini as well [162]. The proteasome generates peptides with lengths ranging from 3-30 amino acids, but the majority of peptides are between 6-11 amino acids in length [79, 144].

A number of experimental studies have been conducted to elucidate the cleavage specificity of the proteasome. Several experiments have shown that hydrophobic and aromatic amino acids are preferred around position P1 at the cleavage sites [189, 190, 196], whereas residues such as P, Q, and K are disfavored.

Proteasomal splicing

Most identified MHC-binding peptides can be mapped back to a contiguous sequence in their respective source protein. However, a recently published report by Hanada *et al.* showed an example of an HLA-A*03 epitope generated from two non-contiguous parts of its source protein [104]. Several experimental tests were done and it became clear that the actual splicing occurs on the peptide level. Post-translational splicing has previously only been identified in lower organisms and plants. It had previously been identified that a renal cell carcinoma (RCC)-specific Tc clone (C2) was correlated with (in terms of activation) the expression of

the FGF-5 protein [103]. At first they managed to specify the immunodominant peptide to a shorter region of the proteins (amino acids 161-220), but synthetic 9-11mers of this longer peptide failed in T-cell activation. Further studies proved that the peptide (NTYASPRFK) was generated from the fragments NTYAS (amino acids 172-176) and PRFK (amino acids 217-220). Another report of a post-translational modification of a T-cell epitope was presented by Vigneron *et al.* [296]. An isolated T-cell clone was activated by the melanoma-specific Pmel17 protein, but identification of the responsible peptide failed. Experiments identified the peptide RTKQLYPEW to be responsible for T-cell activation. This peptide is also post-translationally generated and *in vitro* experiments have shown that it can be generated from the 13mer RTKAWNRQLYPEW in a mixture with proteasomes. This leads to the conclusion that the proteasome can generated MHC-binding peptides by means of protein splicing. In comparison to the FGF-5 case, the excised fragment is much smaller in this case (only four amino acids). Some theories for the mechanism of proteasomal splicing were also presented, but the details are still unknown.

The presence of MHC-peptides generated by proteasomal splicing increase the complexity of epitope identification. It has to be pointed out that the level on which this occurs *in vivo* is not known. There are MHC-binding peptides reported in literature and in databases without a known source protein, these can serve as a starting point for the identification of spliced peptide candidates. Furthermore, there is a need for new mass spectrometry approaches to identify these peptides, since all algorithms based on theoretical comparison of spectra and do not take splicing into account.

The transporters associated with antigen processing (TAP)

Peptides generated in the cytosol have to be transported into the ER in order to bind to MHC class I molecules. A major proportion of the peptides finally binding to MHC molecules are thought to cross the ER membrane by means of TAP. Several studies have shown that loss of TAP function leads to a loss of cell surface expression of MHC class I molecules [44, 120]. Further evidence of the importance of TAP for MHC-peptide presentation was shown by transfecting TAP-negative cell lines with TAP1 and TAP2, restoring the antigen presenting function [210, 265, 266]

TAP is a heterodimeric transmembrane protein, consisting of the subunits TAP1 and TAP2, and belongs to the family of ATP-binding cassette (ABC) proteins. Both TAP1 and TAP2 have a transmembrane domain and a nucleotide-binding domain between which a transloca-

tion pore is formed. Sequence alignment of the TAP1 and TAP2 proteins shows a stretch of about 200 amino acids with high homology. In this region the so called Walker A (P loop) and Walker B motifs that form the ATP-binding cassette where Mg^{2+}-dependant ATP hydrolysis occurs. It is also thought that the "EAA" sequence of the nucleotide-binding domain interacts with the transmembrane domain [58]. Experiments have been made to map the peptide-binding site of the TAP complex. Peptide photo-crosslinking followed by trypsin/bromocyan digestion and immunoprecipitation showed that both TAP1 and TAP2 have similar peptide binding sites [193]. The cytosolic loops between TM5 and TM6 together with the C-terminal stretch after TM6 were found to form the binding site, a theory further supported by deletion experiments resulting in a loss of peptide transport [224]. Both TAP1 and TAP2 are polymorph in all species examined so far, something influencing the peptide specificity [109, 209]. However, it is thought that polymorphisms within the human population have little effect on the peptide specificity [87, 181].

Several experimental attempts to study the substrate specificity of TAP have been made. The first attempt used a trapping of peptide within the ER by glycosylation [187]), revealing information about the amino acid preference and length of the transported peptides. Peptides with lengths ranging from 8-16 amino acids bind TAP with equal affinity [291], but actual translocation is most efficient for peptides with lengths ranging from 8-12 amino acids [150]. Transport was also proven for peptides of lengths up to 40 amino acids, although with less efficiency. Furthermore a correlation between TAP affinity and transport rates of peptides has been observed [101]. The contribution of certain positions of a peptide in the interaction with TAP has been systematically determined using combinatorial peptide libraries [284]. This study shows that the three N-terminal and the C-terminal residues are critical for binding, whereas the amino acids in the other positions are less important.

Details about the transport mechanism of TAP still need to be resolved, but the knowledge gained so far allows for a simple model of peptide binding and translocation. Peptides in the cytosol bind to TAP and induce a structural re-organization leading to the hydrolysis of ATP. The peptide is then translocated through the membrane. It has also been proven that the ATPase activity of TAP is tightly coupled to peptide binding, possibly preventing the waste of ATP without peptide transport [99]. A problem with studying the effects of ATP is the existence of ER export systems that in an ATP-dependent manner export peptides from the ER into the cytosol [149, 229, 247].

MHC-peptide binding

MHC-peptide binding occurs in the ER with the assistance of a number of associated molecules, chaperones. Calnexin is a chaperone facilitating the folding of the MHC α-chain. This molecule is released upon binding of the β_2m subunit and the whole MHC molecules associated with the calreticulin molecule instead. Tapasin is also a very important molecule brining TAP and MHC together to enable peptide binding [182]. The molecule ERp57 is also involved in this loading process [158]. The MHC-peptide complex is fairly stable and dissociate from calreticulin and tapasin. More detail about the actual MHC-peptide interaction were given in Sect. 2.2.

Alternative processing events

Luckey *et al.* showed that for some MHC alleles, a significant amount of peptides were generated even in the presence of proteasome inhibitors [163]. These results clearly indicate an important effect of other cytosolic proteases [21, 78]. TPPII is one such protease that has important effects in the trimming of proteasomal degradation products [222]. A further example points out the importance of TPPII in the generation of a known HIV-Nef(73-82) epitope [250]. TPPII prefers peptide substrates longer that 15 amino acids, something matching most *in vivo* proteasomal substrates [222]. It also seems like TPPII can have both N-terminal activity cleaving of two to three amino acids [281], but also cleave substrates longer than nine amino acids [112, 222]. Generation of such long fragments could also generate new C-termini of MHC-binding peptides. RNAi knockout of TPPII has shown a downregulation of MHC class I expression, indicating its important role in antigen processing [222]. Some cytosolic peptidases are also likely to destroy potential MHC-binding peptides by cleavage. TOP and LAP are two such examples where increased concentrations reduce MHC class I expression [221, 304].

Some alternative ways of peptide-transport into the ER have also been suggested. Lautscham *et al.* described TAP-independent transport of hydrophobic peptides [157] and suggested that these might enter the ER by passive diffusion or by an unknown transport protein within the ER membrane. Furthermore, they pointed out that many known MHC-binding peptides are derived from protein signal sequences and suggested Sec61 as a potential transporter. MHC-binding peptides can also enter the ER as longer pre-cursors. One indication of this are the MHC alleles preferring Pro residues in position two of the peptide. Such peptides

are not effectively transported by TAP, but two ER-aminopeptidases, ERAP1 and ERAAP, can trim longer peptides in order to generate such peptides [251, 252]. Peptides not binding to MHC molecules in the ER are actively transported out to the cytosol again [255]. These peptides can be further trimmed in the cytosol and subsequently enter the ER again [229].

2.3.2. Processing of extracellular antigens

Processing of extracellular antigens is mainly performed by specialized APCs in the secondary lymphoid tissues. These express MHC class II molecules and are able to secrete co-stimulatory signals required for T-cell activation. APCs are mainly dendritic cells (DCs), macrophages, and B cells [94]. DCs typically produce a high concentration of co-stimulatory molecules and express a high level of MHC molecules when activated. Macrophages can phagocytose large particles from e.g. bacteria or parasites. B cells use their membrane-bound antibodies to endocytose antigens that can be processed and presented by MHC class II molecules. DCs and macrophages can recognize general structures such as mannose residues on bacterial cell walls. B cells on the other hand are able to capture low concentration antigen by means of high affinity antibody-antigen interaction. This means that a wide variety of different antigens can be processed and presented.

An overview of the events involved in the processing of extracellular antigens can illustrated in Fig. 2.9. First, the antigens are endocytosed or phagocytosed and form the early endosome. The early endosomes have a slightly acidic pH and contain proteolytic enzymes. The later endosomes can fuse with the lysosomes, both having an increased acidity compared to the early endosomes. The lysosomes contain a mixture of different proteases, of which a few have been fully characterized. For example a variety of thiol and aspartyl proteases (Cathepsins) have been identified in the lysosomes. MHC-peptide binding occurs in the MIIC (MHC class II compartment) that does not contain any of the markers of the late endosomes or lysosomes [146, 206]. The MHC class II molecules arrive to these compartments associated with the **Invariant chain** (Ii, CD74). Ii binds to MHC class II molecules in the ER and prevents peptide binding at this stage [61, 300]. The region of Ii (residues 81-104) preventing binding is called the CLIP region [62]. Furthermore, Ii has chaperone effects and also help in the routing of MHC class II to the MIIC compartments. MHC peptide binding can occur after the release of CLIP by a combined effect of proteases and the acidic environment [151], or after the removal of CLIP by HLA-DM molecules [226]. HLA-DM is a protein specialized in removing the medium affinity CLIP peptides from MHC class II molecules, allowing binding

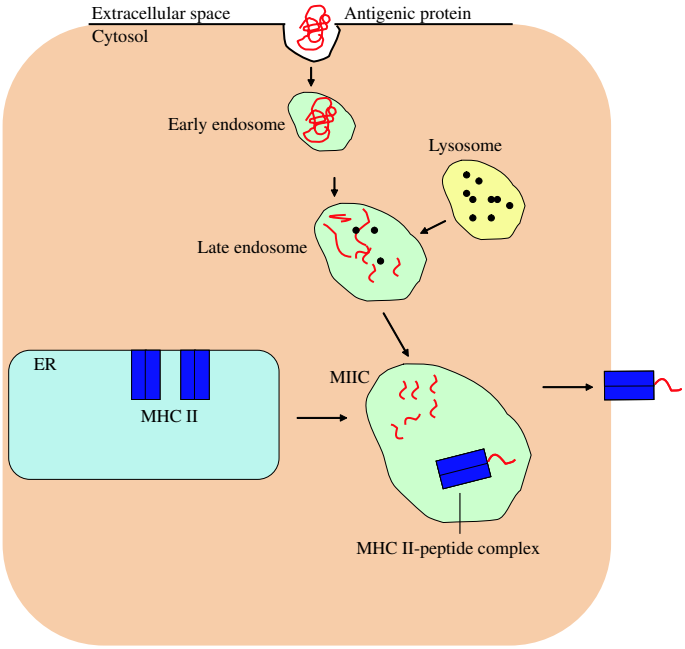

Figure 2.9.: An overview of the processing of extracellular antigens. Antigenic proteins that are taken up pass several acid compartments, where they are degraded into smaller peptides. In these association with MHC class II molecules also occur and the MHC-peptide complex is subsequently transported to the cell surface.

of high affinity antigenic peptides.

2.4. The immune system in cancer

Reports from the beginning of 2005 showed that cancer has past heart disease as the major cause of death in the USA. Cancer is initiated by normal cells escaping cell growth control mechanisms, leading to a clone of modified cells, a tumor. Benign tumors do not invade the surrounding tissue and these can in many cases be treated successfully. Malignant tumors on the other hand usually continue to grow and to invade other tissue. They also exhibit the ability of metastasis, where a small part of the tumor is carried (e.g. by blood) to another tissue. This gives rise to a secondary tumor that can be far away from its origin. The immune system functions to monitor the body and destroy modified self-cells. However, in many cases this monitoring process fails. Tumor immunology tries to identify tumor markers for diagnosis and prognostics [3], and to find ways to activate the immune system to recognize and kill cancer cells (immunotherapy). Early studies in melanoma investigated the T-cell response and specificity [233, 293]. It soon became clear that most tumors, if not all, express T-cell antigens that can be specifically targeted [233].

The following sections describe the most important groups of genes involved in cancer and how targets for immunotherapy can be identified. This is followed by an overview of databases containing cancer-related data and a motivation why integrative data-analysis is needed to understand the underlying mechanisms of cancer and for the development of effective cancer vaccines.

2.4.1. Types of cancer-related genes

Genes that are able to induce cancer are called **oncogenes**. Most of the known oncogenes have a function in the regulation of cellular growth and can be divided into three main categories. The first category contains genes that induce cellular proliferation, such as growth factors or transcription factors. Some examples are the well-known signal transducers src and *abl* or the transcription factors jun and myc. The second category of genes is those inhibiting cellular proliferation. The best example here is p53, which plays a central role in the development of many different cancer types. The third category of genes is those involved in apoptosis (programmed cell death), where bcl-2 is a good example.

We now have an overview of the types of **oncogenes** that might induce cancer and the question now is how the immune system can protect against tumors. The immune system recognizes antigens from tumors, which can be classified into **tumor-specific antigens (TSA)** and **tumor-associated antigens (TAA)**. TSAs can be found exclusively on tumor cells and can be the results of certain mutations or the results of modified processing of MHC class I restricted antigens, resulting in a unique set of T-cell epitopes presented. TAAs can be proteins that are only expressed during fetal development and not in adults. There are many examples of the expression of such genes in various cancer-types. TAAs can also be proteins having low expression levels under normal conditions that are being overexpressed in the tumor.

There are many different approaches to identify new tumor-related antigens. Large-scale DNA microarrys have revolutionized this by scanning the expression levels of many thousands of genes in experiment. A large number of differentially expressed TAA have been identified in such experiments, sometimes improving the clinical classification of a certain cancer type [77, 90, 262, 287]. Another method for detecting antigens causing a humoral immune response in cancer is the serological analysis of recombinant cDNA expression libraries (SEREX) method [283]. This method uses patient serum together with a cDNA expression library in order to isolate proteins to which auto-antibodies can be detected and has been

used to study many different cancer types.

A wide number of cancer-specific MHC-binding peptides have also been identified. The experimental procedure typically consist of a stripping of MHC-binding peptides from a cell culture, followed by HPLC-MS for sequencing. Alternatively, Edman degradation is used for the sequencing. In many cases an identified MHC-binding peptide is tested for its ability to elicit T-cell activation. In some cases the peptides arise from overexpressed proteins, but also from single point mutations [19, 49, 59] or from frame shifts [123, 223, 248].

2.4.2. Cancer immunotherapy

Cancer immunotherapy in general means that parts of the immune system is used to fight tumors. There is a difference between active and passive therapies, where active ones stimulate the body's own immune system, whereas passive rely on immune system components (e.g. monoclonal antibodies). Studies in melanoma-bearing mice showed that injection of whole melanoma cells could make the tumor disappear in more than 40% of the cases. DCs cultured with tumor fragments can also be used to activate the immune system. Studies in mouse have shown that these are able to activate both Th and Tc cells to recognize TSAs.

The key issue in the approaches described above is the presentation of peptides by MHC molecules. Many attempts have also been made using shorter peptides to activate the immune system. Early studies of both hematological malignancies and melanoma showed that introducing tumor-specific peptides as a vaccine is both safe and feasible [121, 188]. However, one general problem of the peptide-vaccines is their inability to raise clinical responses *in vivo*.

Several attempts have also been made to use monoclonal antibodies for cancer immunotherapy. Levy *et al.* successfully treated a patient with long progressed B-cell lymphoma. Since this cancer type is B-cell specific, all tumor cells express the same surface-bound antibody. Monoclonal antibodies were raised against this "tumor-specific" antibody and successfully used for therapy. The problem here is that one would need to raise specific antibodies for every new patient. More promising is the use of growth-factor receptors as target. Many tumors show an overexpression of HER2, an epidermal-growth-factor, and an antibody (Herceptin) against this protein is used in the treatment of breast cancer.

A major challenge in terms of immunotherapy is the identification of candidate proteins or peptides that can be used to trigger the immune system. Although the Herceptin antibody can be used for breast cancer treatment, the use of monoclonal antibodies is not that widely

applied. T-cell activation seems more promising in many aspects, but the identification of vaccine candidates is still a problem. In order to find vaccine candidates it is useful to consider all available information regarding a certain cancer type. A good candidate might for example be exclusively expressed in the tumor types (a TSA), have a high expression level, and several presented T-cell epitopes. The next section describes some cancer-related databases where this type of information can be found. Chapter 5 of this thesis also deals with the integration of cancer-related data and the identification of cancer vaccine candidates.

2.4.3. Cancer-related databases

There are numerous databases containing cancer-related information. Typically these focus on one type of data and one aspect such as cancer genetics or cancer immunology. Examples of genetic aspects of cancer are the Mitelman database for chromosomal aberrations in cancer [179], the SNP500Cancer databases for single nucleotide polymorphisms in cancer [198], and Cancer GeneticsWeb giving a broader view of altered or mutated genes.

Several databases also deal with immunology-related cancer information. An example here are the SEREX databases and its successor the cancer immunome database (CIDB) that focus on auto-antigens detected by the SEREX method [283].

Some databases have a focus on T-cell epitopes related to cancer. The SYFPEITHI [217] database described earlier lists MHC-binding peptides that are naturally processed and presented by T-cells. Some of these are also labeled as cancer related, although no further information is available. The cancer immunity database on the other hand gives a more comprehensive overview of MHC-binding peptides and their relation to different cancer types [277].

In addition to the databases introduced above, comprehensive genomic and proteomic data is available from databases such as Swiss-Prot [27], NCBI, and Locus Link [212].

Although all this data is freely available online, the integrative analysis and identification of vaccine candidates is not trivial. The data usually have differences in both data format (syntactic differences) and in the meaning of the data items (semantic differences) [48] The CAP database presented in this thesis was developed to overcome these difficulties [74] and to create a unified view of the data. CAP integrates different cancer-related databases and enables complex queries on the data. It can be used to get an understanding of some general principles underlying cancer development, but also to identify vaccine candidates. The development of CAP is discussed in Chapter 5.

2.5. Prediction methods for antigen processing and presentation

There have been many attempts to model and predict the processes involved in antigen processing. Most focus has been on the prediction of MHC binding, since this can be considered the most specific step. In terms of processing of extracellular antigens this is currently the only step that can be modeled, since very little is known about e.g. the proteases of the lysosome. For intracellular processing, proteasomal cleavage and TAP transport can also be predicted. This section will describe prediction methods for MHC binding, proteasomal cleavage, and TAP transport. Many computational approaches for MHC binding prediction are similar for both class I and class II, hence these are described first. Proteasomal cleavage and TAP transport involved in MHC class I antigen processing are described subsequently, followed by an overview of attempts made so far to combine methods.

2.6. Prediction of MHC-peptide binding

Prediction of MHC-binding peptides is an area where many "standard" bioinformatics methods have been applied. Most attempts have been on the prediction of MHC class I binding peptides, since these have a defined length. MHC class II-peptide prediction typically involves some alignment method in order to extract the binding cores of the peptide, which are then used for prediction. The prediction methods can be classified into sequence-based and structure-based. The following sections will describe a number of approaches for prediction of MHC class I binding peptides.

2.6.1. Simple motifs and PSSMs

Even before a large number of MHC-binding peptides had been characterized, people suggested that there were common sequence pattern in T cell epitopes. Rothbard suggested that the common pattern consisted of a charged residue or glycine followed by two hydrophobic residues. This general pattern was later extended and used for prediction [234]. Another method suggested by Margalit *et al.* searches for amphiphatic helices and defined those as probable T helper cell antigenic sites [169]. Sette *et al.* further investigated sequence motifs by means of binding assays [253]. The motifs suggested here were also simple e.g. the IE^d motif was defined as 'basic-basic-noncharged-basic'. In 1991, Falk *et al.* a number of self-peptides eluted from MHC molecules that underlined the importance of anchor residues for

MHC binding [85]. These findings were then used to define allele-specific motifs for prediction [238, 200]. An example is the HLA-A*0201 allele were peptides of a length of nine amino acids often have a Leu in position two and a Val in position nine. The search motif used for scanning proteins for potential binding peptides would be $XLXXXXXXV$, were X matches any amino acid.

The simple sequence motifs have been extended to position-specific scoring matrix methods (PSSMs). Here amino acid-specific scores are given for each position of the peptide in a matrix. A PSSM will thus have i columns corresponding to number of amino acids in the peptide sequence and j rows corresponding to the number of amino acids considered (20 if all naturally occurring amino acids are considered). In the case of a 9mer peptide the size of the matrix will be 9x20=180 elements. The score of a peptide is calculated as the sum of position-specific scores:

$$S = \sum_i s_{i,j} \tag{2.1}$$

where $s_{i,j}$ is amino acid-specific score for amino acid i of the peptide. Two of the best known prediction methods, BIMAS [201] and SYFPEITHI [217] fall into the category of PSSMs. The matrices from these methods are derived from experimental measurements in the case of BIMAS and expert knowledge in the SYFPEITHI case. Other PSSM-based methods derive matrices from sequence alignments, putting the whole problem into a statistically more well-defined framework. More details about the different methods are given below.

BIMAS

The BIMAS prediction method was presented by Parker *et al.* and the position-specific scores are derived from stability measurements of $\beta 2m$ dissociation rates. The results of this study show that the contribution of a certain amino acid side-chain is in many cases independent of the overall peptide sequence, underlining the use PSSMs to obtain overall affinity measures. The initial approach presented considered 9mer peptides binding to HLA-A2 and 154 peptides were used to determine the 180 coefficients of the PSSM. The half-life ($t_{1/2}$) of $\beta 2m$ was used in to define an error function:

$$err = ln(t_{1/2}) - ln(P1 \cdot P2 \cdot P3 \cdot P4 \cdot P5 \cdot P6 \cdot P7 \cdot P8 \cdot P9 \cdot Constant) \tag{2.2}$$

where P1, P2 etc corresponds to the position-specific amino acid scores of the peptide (each

position has 20 scores for the 20 amino acids). The scores in all positions were normalized by dividing by the score of Ala in that position, meaning that there are 172 independent terms of the matrix plus the *Constant*.

In order to reduce the dimensionality of the problem only position considered relevant for binding were considered in the optimization procedure. Here 82 variables were selected of which 40 came from position 2 and 9 of the peptides (the anchor positions). The other position were selected by picking cases where a single amino acid change between two peptides showed a large difference in dissociation and from some other specific criteria outlined in the original publication [201]. The coefficients can the be found by solving the system of equations.

Coefficients have also been constructed for a wide range of different HLA molecules and a prediction server is publicly available.

SYFPEITHI

The SYFPEITHI prediction method is based on expert knowledge of MHC-peptide binding motifs and the matrices used are derived by hand. Only naturally processed and presented peptides are considered for matrix construction and the focus is put on anchor and auxiliary anchor positions. Ideal anchors are given a score of 10, unusual anchors a score between 6 and 8, and auxiliary anchor scores between 4 and 6. Furthermore, frequently occurring amino acids are given scores between 1 and 4, whereas amino acids having negative effect on binding are given scores from -3 to -1. An example of the HLA-A*0201 matrix from SYFPEITHI can be seen in Table 2.1.

PSSMs from sequence alignment

A number of prediction methods have also been presented where a PSSM is constructed from aligned MHC-binding peptides [219, 305, 220]. This is a standard procedure in bioinformatics and it has been used for e.g. the identification of transcription factor binding sites [268]. The advantage of such methods is that they are mathematically well defined and the actual matrices used for prediction are not derived by hand. In a set of aligned sequences, the frequency of each amino acid in each position of the alignment can easily be calculated. A PSSM is then often a logarithmic transformation of the frequency matrix. Usually the background (or prior) probability of the amino acids are also considered. The score of amino

Table 2.1.: The SYFPEITHI HLA-A*0201 matrix.

	1	2	3	4	5	6	7	8	9
A	**2**	**4**	**2**	0	0	0	**2**	**1**	**4**
C	0	0	0	0	0	0	0	0	0
D	**-1**	0	0	**1**	0	0	0	0	0
E	**-3**	0	**-1**	**2**	0	0	0	**2**	0
F	**1**	0	**1**	**-1**	**1**	0	0	0	0
G	**1**	0	0	**2**	**2**	0	0	**1**	0
H	0	0	0	0	0	0	**1**	0	0
I	**2**	**8**	**2**	0	0	**4**	0	0	**8**
K	**1**	0	**-1**	0	**1**	0	**-1**	**2**	0
L	**2**	**10**	**2**	0	**1**	**4**	**1**	0	**10**
M	0	**8**	**1**	0	0	0	0	0	**4**
N	0	0	**1**	0	0	0	**1**	0	0
P	**-3**	0	0	**2**	**1**	0	**1**	0	0
Q	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0
S	**2**	0	0	0	0	0	0	**2**	0
T	0	**4**	**-1**	0	0	**2**	0	**2**	**4**
V	**1**	**4**	0	0	0	**4**	**2**	0	**10**
W	0	0	**1**	0	0	0	0	0	0
X	0	0	0	0	0	0	0	0	0
Y	**2**	0	**1**	**-1**	**1**	0	**1**	0	0

acid j in positions i of the alignment can be described as:

$$s_{i,j} = \log \frac{f_{i,j}}{b_j} \tag{2.3}$$

where $f_{i,j}$ is the frequency of amino acid j in position i and $b_{i,j}$ is the background frequency of amino acid j. There are different ways to obtain the background frequencies. In some cases equal background probability is considered for all amino acids, whereas in other cases a protein database (like SwissProt) is used to calculate amino acid-specific backgrounds. Most approaches also use a form of pseudocount correction as presented by Henikoff and Henikoff [110].

2.6.2. Machine-learning methods

One major drawback with the prediction methods presented so far is that they assume an independent contribution to the binding affinity of each amino acid, not considering the neighboring amino acids. Parker *et al.* concluded that this holds in many cases, but they also found sequence where this assumption could not explain binding [201]. Machine learning

methods such as artificial neural networks (ANNs) or support vector machines (SVMs) allows for the generation of a model taking amino acid correlations of the peptide into account. I wide range of different ANN methods have been presented [43, 102, 119, 192]. In most cases a standard back-propagation network is used and different numbers of hidden layers and training cycles investigated. Hidden Markov Models (HMMs) have also been used for prediction. In a study presented by Mamitsuka, HMMs performed approximately 2-15 % better than backpropagation ANN for prediction [167]. In this thesis support vector machines (SVMs) and decision trees are introduced for prediction of MHC binding peptides.

All methods described so far concern prediction of MHC class I binding of peptides of a certain length. MHC class II molecules bind longer peptides and the identified MHC binding peptides usually differ in length. However, the actual binding core of these peptide corresponds to approximately nine amino acids. Several methods have been presented for MHC class II binding using these binding cores [38, 166, 191, 270]. These methods are not described in more detail since the focus here is MHC class I binding.

2.6.3. Structure-based methods

Structure-based methods for prediction of MHC class I binding peptides includes docking [161, 282], molecular dynamics studies [230], and threading [7, 8, 244]. The advantage of these methods is that they can be applied given a single starting structure of a MHC-peptide complex. The drawback is that there are not many different MHC alleles for which a structure is known. Furthermore, even if a structure is known the accuracies of the structure-based methods are not convincing.

Rognan *et al.* have presented a method using molecular dynamics simulations and a new force field (FRESNO) for prediction [230]. An initial structure given a new peptide is built using information from a database of MHC-peptide complexes with known structure. First the anchors are placed into the MHC molecule, followed by a loop search procedure for the middle part of the peptide. The structure is then energy minimized. The binding energy of the peptide is then estimated using the FRESNO force field. A fair estimate of the predictive power of this procedure is hard give, since it was tested on very few sequences.

Altuvia *et al.* have investigated threading of MHC peptides using statistical pairwise potentials [7, 8, 244]. In comparison to docking and molecular dynamics methods, threading is much less CPU intensive and can easily be applied on a genome scale. The threading procedure depends on two determinants: (1) the definition of contact residues between the

MHC molecule and the peptide and (2) the choice of pairwise interaction potential table. A contact table for the MHC-peptide complex can be obtained by defining two residues to be in contact according to some distance criteria, e.g. two residues are in contact if their $C\alpha$ atoms are closer than 7 $\AA$ from each other. For a given MHC-peptide complex, such a contact list can be constructed for each amino acid of the peptide. In order to "thread" a new peptide, it is assumed that the new peptide make the same contacts with the MHC molecules as the original one. The binding energy of the new peptide is then calculated by summing up individual pairwise contact potentials. In comparison to docking and molecular dynamics methods, threading is much less CPU intensive and can easily be applied on a genome scale.

2.6.4. MHC-peptide databases

There are several databases containing information about MHC molecules and their binding peptides. The SYFPEITHI database [217] contains over 4,500 peptide sequences known to bind MHC class I and MHC class II molecules. This database is highly curated and only contains naturally processed MHC-binding peptides. Another database containing a both naturally processed peptides and data from binding experiments is the MHCPEP database [38], which is now a part of the FIMM database [241, 242]. MHCPEP contains about 13,000 peptides. Another database focusing on quantitative binding data is the AntiJen database, which is an improved version of the Jenpep database [26, 173] (no information about the total number of MHC-binding peptides in the database is given). Some specialized MHC peptide databases also exist, such as the HIV database [279] and the HCV database [306]. Furthermore, the Protein Data Bank (PDB) contains data of MHC-peptide and MHC-peptide-TCR complexes [23], which are the starting point of structure-based prediction methods.

2.7. Prediction of proteasomal cleavage and TAP transport

MHC-peptide binding is considered the most specific step in the MHC class I restricted antigen processing pathway. Two other important processes are proteasomal cleavage and TAP transport. The following sections outlines the most important prediction methods presented for these processes.

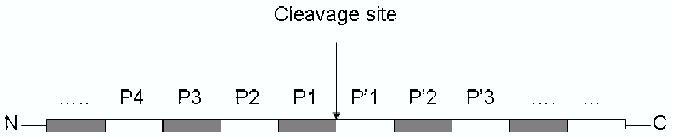

Figure 2.10.: This figure illustrated the nomenclature of a proteasome cleavage site and the flanking amino acids of such a site.

2.7.1. Proteasomal cleavage prediction

The proteasome, introduced in Sect. 2.3.1, mainly cleaves cytosolic proteins into smaller peptides. Data of proteasomal cleavage is available in the form of longer peptide or whole proteins with verified proteasomal cleavage sites. Several computational approaches to elucidate the cleavage specificity of the proteasome have been presented. They are all based on these experimentally verified cleavage sites within peptide/protein substrates and analyses the flanking regions. The nomenclature of a proteasomal cleavage site and the surrounding amino acids can be seen in Fig. 2.10.

Holzhütter *et al.* [118] used a statistical method to analyze the cleavage sites found in a set of peptide substrates with lengths ranging from 22 to 30 amino acids. From this analysis cleavage-determining amino acid motifs (CDAAMs) could be identified and incorporated into the FRAGPREDICT prediction method. A final set of ten different CDAAMs were said to represent the cleavage specificity of the proteasome. The ten CDAAMs accounts for cleavage of one to five different groups of peptide bonds and the accuracy of prediction reached 93 %. In each motif, the important residues for cleavage were found within five residues from the cleavage site. A total of seven peptide with length ranging from 22 to 30 amino acids were used in this study. The total length of all substrates is 181 residues with a total of 118 cleavage sites.

PAProC is a method utilizing a stochastic hill-climbing algorithm (evolutionary algorithm) for prediction of proteasomal cleavage sites [154, 197]. The data used for algorithm training comes from digestion experiments with the enolase protein. This protein is 436 amino acids long and there are 117 identified cleavages generated by the human proteasome. In this study it is also shown that amino acids more than 6 amino acids away from a cleavage site have no effect, whereas a prominent role is played by the two closest amino acids for a given cleavage site. The effect of the two closest amino acids is modeled with an affinity parameter $\alpha_1(X_1, X_{1'})$. Each of the other positions $P_i, i \ldots, k$ (or $P_{i'}, i' = 2, \ldots, m$) have an affinity

$\alpha_i(X_i)$ ($\alpha_{i'}(X_{i'})$). The model is additive and the final cleavage affinity for a cleavage site is:

$$\delta = \alpha_1(X_1, X_{1'}) + \sum_{i=2}^{k} \alpha_i(X_i) + \sum_{i=2}^{m} \alpha_{i'}(X_{i'}) \tag{2.4}$$

Given a set of affinities, cleavage sites within the enolase protein are predicted and compared to the experimental results. An objective function used for optimization was defined as:

$$F = K \cdot \text{the number of missing cuts} + \text{the number of superfluous cuts} \tag{2.5}$$

where $K = 2$ was used in most cases. Give a starting set of affinities, the performance F_0 of the algorithm is calculated. The affinities are then subject to a random perturbation, and the overall performance F_1 is calculated. If $F_1 \leq F_0$, the new affinity parameters are stored and a new perturbation is carried out, otherwise the old affinity parameters are kept. In the case of human proteasome, positions $P6 - P1', P4'$ were considered giving a total of 480 parameters (20 x 20 for the $P1$ and $P1'$ positions and 4 x 20 for the $P5 - P2$ positions. The number of iteration steps used for the human proteasome was 275,000. The algorithm almost perfectly reproduce the training data, but the results in this thesis will points out the bad performance of PAProC on data not used for training.

An ANN method, NetChop, has also been presented for proteasomal cleavage [141]. There are two different sets of training data used by NetChop, verified cleavage sites within proteins and naturally processed MHC ligands. MHC class I ligands for studying proteasomal cleavage was previously introduced by Altuvia and Margalit [6]. The network used is a standard feed-forward network and the final network 29 hidden neurons. The data used comes from the same enolase cleavages used by PAProC and digestion experiments of the β-casein protein. For algorithm training 19 flanking residues (9 on each side) if a cleavage site were used.

2.7.2. TAP transport prediction

Experiments estimating the peptide binding affinities to TAP have been used to elucidate the binding specificity of TAP. A very general peptide binding motif of TAP was presented by van Endert *et al.* [290]. Their general conclusion was that TAP binding is in general stronger for MHC binding peptides than for non MHC-binding peptides. Uebel *et al.* used combinatorial peptide libraries to identify the recognition principle of TAP [284]. Their findings showed that TAP is highly specific with peptide affinities spanning three orders of magnitude. These

early studies revealed some general properties of TAP, but several prediction methods based on machine learning and statistical analysis have also been presented.

The first machine learning method proposed used ANNs for prediction [39, 65] and focused on peptides with a length of nine amino acids. A dataset of peptides with experimentally verified TAP affinities was used for training. A standard backpropagation network with two hidden layers was used and a correlation of 0.73 was reached. From this study it was also pointed out that the three N-terminal and the C-terminal residues of the peptides are most important for binding.

Peters *et al.* used the stabilized matrix method (SMM) for prediction of TAP affinity [204]. The dataset used was more or less the same as used in the ANN approach described above and a correlation of 0.78 between predicted and experimental values was obtained. Furthermore, prediction of peptides longer that nine amino acids was also tested, by studying only the part of the 9x20 scoring matrix that corresponds to the three N-terminal and the C-terminal amino acids. By this approach and a weighting of the N-terminal scores, they predicted the binding affinity of peptides longer than 9 amino acids. In an attempt to combine TAP prediction and MHC binding predictions for HLA-A*0201, they found only a marginal increase in prediction accuracy.

An approach using a network of SVMs has also been presented [24]. In this study amino acid-specific properties are used together with normal sequence encoding is used to boost the prediction performance. Using a network of SVM trained on the same data might lead to overestimation of prediction performance. The use of SVMs for prediction of TAP affinity is discussed in more detail in Chapter 4.

2.8. Combined prediction of the whole antigen processing pathway

One major part of this thesis deals with modeling of the whole antigen processing pathway of MHC class I binding peptides. The WAPP method presented here is the first method enabling an integrative modeling of the whole antigen processing pathway [75]. In the meantime two other methods have been presented for modeling the whole antigen processing pathway.

Larsen *et al.* have introduced the NetCTL method [156]. NetCTL is based on the NetChop method for proteasomal cleavage [141], the SMM method presented by Peters *et al.* for TAP transport [204], and a neural network for prediction of MHC-peptide binding. The scores each method are then scaled into the same range and combined into an overall score. Here the

proteasomal score is weighted by 0.05 and the TAP score with 0.1. Even though this can not be directly translated into the contribution of each method to the overall score (stated by the authors), the improvements of the combined approach is only marginal compared to MHC-peptide binding prediction alone. A further attempt for modeling the overall processing pathway has been described by Tenzer *et al.* [275]. This method is based on a modified version of the SMM method for proteasomal cleavage [204, 205], the same matrix method as NetCTL for TAP transport [204], and a number of ARB matrices for MHC-peptide binding prediction [147, 148, 257, 258, 259]. As for NetCTL, the separate scores are combined into an overall score without additional weighting. Both the NetCTL and the method by Tenzer *et al.* are compared to the WAPP method in Chapter 4.

2.9. Machine learning

Learning from observations and experiments is important in biology. In most cases a set of observation or measurements (*features* or *attributes*) can be associated with a certain outcome, e.g. gene expression levels can be associated to a clinical outcome. In many cases no explicit model can be constructed that explains the input/output mapping. Here machine learning can be applied with the aim to learn an approximate model from the data. Machine learning methods try to learn the input/output mapping using theories from statistics, optimization, signal processing etc. The field of molecular biology has been described as tailor-made for machine learning approaches [254], where a vast amount of data is available but no (or little) theory.

Many different types of machine learning methods have been described in literature and these can broadly be split into *supervised* and *unsupervised*. Unsupervised algorithms include e.g. hierarchical clustering and self-organizing maps, where the aim is to find clusters of similar data points. The focus here is on supervised methods, where a set of labeled training examples is available. This means that the starting point of the learning procedure is a dataset where a set of features are associated with a certain outcome. Typically the function mapping a set of input features to a certain output is hard to derive. This might be due to a number of reasons such as no good underlying theory or noisy measurements. The aim of the machine learning algorithm is to learn a function that approximates the input/output functionality with high reliability.

The following sections describe the basics of the machine learning methods support vector

machines (SVMs) and decision trees (DTs). Furthermore, the measures used to evaluate the accuracy of the methods presented in this thesis are outlined.

2.9.1. Support Vector Machines (SVMs)

SVM-based methods have gained interest over recent years and have been applied to many tasks in bioinformatics such as protein subcellular localization prediction [116, 117, 122], gene expression analysis [37], and classification of cancer types [91]. One advantage with SVMs compared to neural networks is the relatively small number of tunable parameters and the optimization problem solved is a convex quadratic function giving a global, usually unique, solution [41]. The following sections give a brief overview to SVM classification and SVM regression. A more complete coverage of SVMs has been given in several books that can be recommended to the interested reader [63, 240, 295].

SVM classification

The typical case for SVM classification is a set of input attribute vectors belonging to one of two classes (represented as +1 or -1). SVMs use linear functions to separate data from different classes and the concept of linearly separable data will first be introduced, followed by an overview of how kernel functions can be used to map non-linearly separable data into a new feature space where linear decision functions can be used to separate the data.

Linear SVMs Given a set of N pairs $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n))$,$x_i \in \mathbf{R}^p$ and $y_i \in \{-1, +1\}$, the aim of the SVM is to find a function:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}_i + b = 0 \tag{2.6}$$

where w is a unit vector ($\|w\| = 1$). If $f(\mathbf{x}) \geq 0$ for all examples of the positive class and $f(x) \leq 0$ for all negative examples, the function separates the two classes. A classification rule, $G(x)$ can then be defined as:

$$G(\mathbf{x}) = sign[\mathbf{w} \cdot \mathbf{x}_i + b] \tag{2.7}$$

where $f(\mathbf{x})$ gives the signed distance to the separating hyperplane. The geometric interpretation of this in 2D is a hyperplane (a line in 2D) that separates the datapoints of the two classes, see Fig. 2.11.

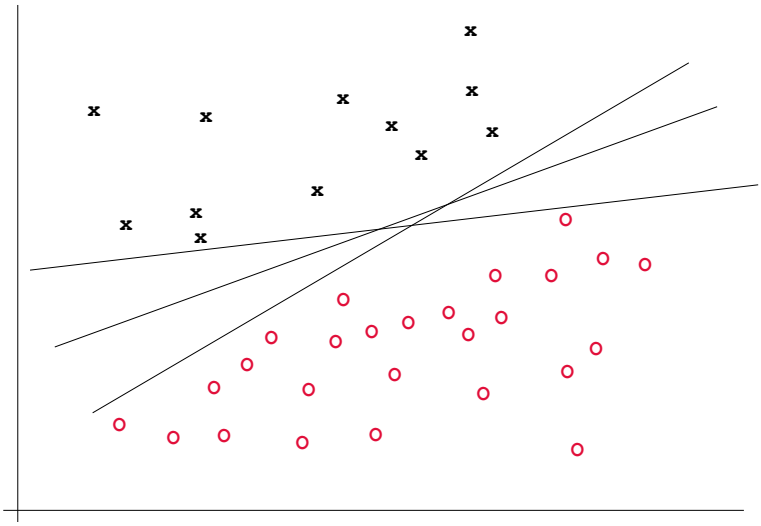

Figure 2.11.: This example shows how data in two dimensions from two classes can be separated by a hyperplane (a line in this case). However, the question remains how the optimal separating plane can be found.

However, there are several lines that fulfill the task of separating the data and the question remains how the optimal one is chosen. The best choice will be the hyperplane (a line in our example) that maximizes the distance to the closest point of each class, this plane is called the optimal separating hyperplane (OSH). Points on such a hyperplane fulfill the criteria $\mathbf{w} \cdot \mathbf{x}_i + b = 0$, where $\mathbf{w}$ is the normal of the hyperplane and $|b|/||w||$ is the perpendicular distance to the origin. The distance from the hyperplane to the closest positive example is defined as d^+ and the corresponding distance for the closest negative example is d^-. The sum of the shortest distances to each class is called the margin ($d^+ + d^-$), see Fig 2.12. The aim of the SVM is to find the separating hyperplane with the largest margin, which then by definition is the OSH. If the data is separable, it is possible to chose a scaling of $\mathbf{w}$ and b such that:

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1 \qquad \text{for } i \in \{i | y_i = +1\} \tag{2.8}$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1 \qquad \text{for } i \in \{i | y_i = -1\} \tag{2.9}$$

which can be combined into

$$y_i(\mathbf{w} \cdot \mathbf{x}_i) - 1 \geq 0 \qquad \forall i, 1 \leq i \leq N \tag{2.10}$$

This also means that two perpendicular hyperplanes to the OSH H1 and H2 can be defined as: H1 : $\mathbf{w} \cdot \mathbf{x}_i + b = 1$ with the normal $\mathbf{w}$ and the distance: $|1 - b|/||\mathbf{w}||$ from the origin. The

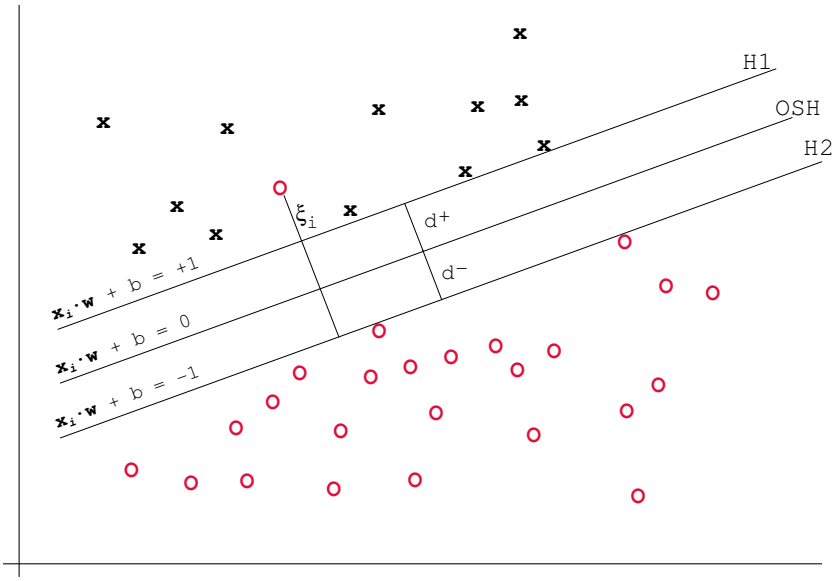

Figure 2.12.: Introducing slack variable ξ_i in the optimization problems allows for classification of data that is not perfectly separable.

corresponding is true for the points on H2 : $\mathbf{w} \cdot \mathbf{x}_i + b = -1$ with the distance $|-1-b|/||\mathbf{w}||$ to the origin. The points lying on the hyperplanes H1 and H2 are called support vectors. The margin in this case is $2/||\mathbf{w}||$, hence the optimal OSH can be found by minimizing $||\mathbf{w}||^2$ under the constraints given above.

In most real-life cases the data is not perfectly separable and we would like to relax the constraints we have on our optimization problem. This can be done by introducing slack variables (ξ_i) [57], see Fig 2.12. This gives a new formulation of the optimization problem:

$$\text{Min } \frac{1}{2}||\mathbf{w}||^2 + C\sum \xi_i \tag{2.11}$$

having the following constrains:

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq 1 - \xi_i \qquad \text{for } i \in \{i|y_i = +1\} \tag{2.12}$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1 + \xi_i \qquad \text{for } i \in \{i|y_i = -1\} \tag{2.13}$$

$$\xi_i \geq 0, \quad \forall i, 1 \leq i \leq N \tag{2.14}$$

which can be combined into:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{2.15}$$

This means that we have a convex optimization problem (quadratic criterion with linear inequality constraints). Notable is also that the parameter C in Eq. 2.11 can be used to weigh the penalty for misclassified examples. An effective way to solve such problems is to introduce Lagrange multipliers α_i, $i = 1, ..., N$. For constraints of the form $c_i \geq 0$, the constraint equations are multiplied by positive Lagrange multipliers and subtracted from the objective function. The Lagrange multipliers for equality constraints are unconstrained [40]. This gives the following primal function to minimize:

$$L_P = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\{y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i\} - \sum_{i=1}^{N}\gamma_i\xi_i \tag{2.16}$$

The optimization is carried out with respect to $\mathbf{w}$, b and ξ_i. Setting the respective derivatives to zero give the following:

$$0 = \sum_{i=1}^{N}\alpha_i y_i \tag{2.17}$$

$$\mathbf{w} = \sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i \tag{2.18}$$

$$\alpha_i = \gamma - \mu_i \tag{2.19}$$

By substituting the equations above into the primal, the Wolfe dual objective function can be obtained [89]:

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i}^{N}\sum_{j}^{N}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i \cdot \mathbf{x}_j) \tag{2.20}$$

The maximization of the dual is a simpler programming task than the primal and can be solved by standard optimization techniques [106]. In addition to the constraints above, the Karush-Kuhn-Tucker condition to the optimization problem includes the following constraints:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i) - 1 + \xi_i \geq 0 \tag{2.21}$$

$$\alpha_i\{y_i(\mathbf{x}_i + b) - 1 + \xi_i\} = 0 \tag{2.22}$$

$$\gamma_i\xi_i = 0 \tag{2.23}$$

Taken together all these constraints uniquely gives the solution to both the primal and dual problem, which is:

$$\hat{\mathbf{w}} = \sum_{i}^{N}\hat{\alpha}_i y_i \mathbf{x}_i \tag{2.24}$$

Here the coefficient α is only zero for observations where the constraints in Eg. X are met. These observation are in fact the support vectors and $\hat{\mathbf{w}}$ is represented by these alone. Any of the support vectors can be used to determine $\hat{b}$. This gives the following decision function that can be used for classification:

$$\hat{G}(\mathbf{x}) = sign[\hat{\mathbf{w}} \cdot \mathbf{x}_i + \hat{b}] \tag{2.25}$$

Nonlinear SVMs In many real-world examples, linear functions are not enough to separate the data. However, Boser *et al.* found a way to solve this by the use of a kernel function [29]. A kernel functions maps the data from the input space into a new feature space where linear functions can be used to separate the data. A simple geometric interpretation of how a kernel function can be used to map data into a space where linear functions can be used for classification is given in Fig. 2.13. In the SVM optimization function, the input data only occurs in the form of dot products. If we can define a kernel mapping such that $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$ the optimization problem would be:

$$L_D = \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \tag{2.26}$$

The crucial feature of the kernel K is that it is used only in the training step and $\phi(x)$ does not even have to be known explicitly. Examples of two frequently used kernels are the radial-basis function and the polynomial, Eq. 2.27 and Eq. 2.28 respectively:

$$K(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2) \tag{2.27}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d \tag{2.28}$$

where γ and d are kernel-specific parameters.

SVM regression

The theory of classification SVMs can be used for regression tasks as well. Here the aim is to find a function that map the data from the input domain to a real-valued output. Given such a function, the error for a given datapoint is referred to as the residual of the output. This gives an estimate of the accuracy of the function and the aim is of course to have small

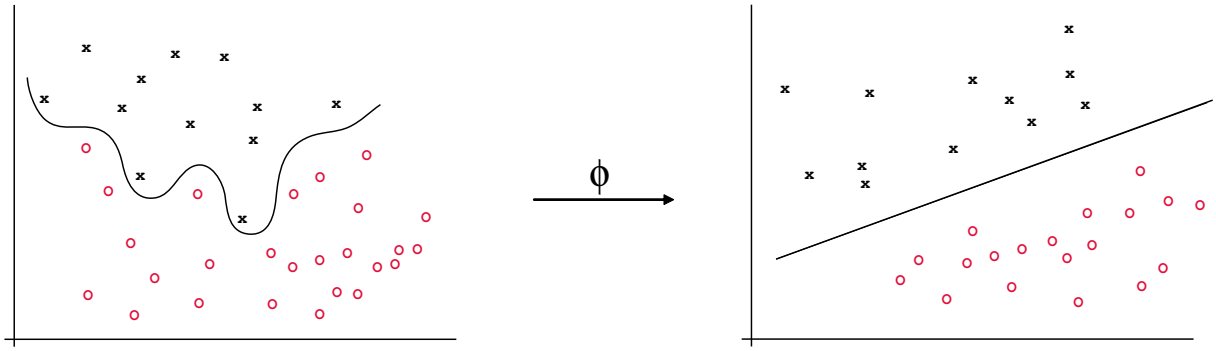

Figure 2.13.: An example of how a kernel function can be used to map data into a space where linear functions can be used to discriminate the two classes.

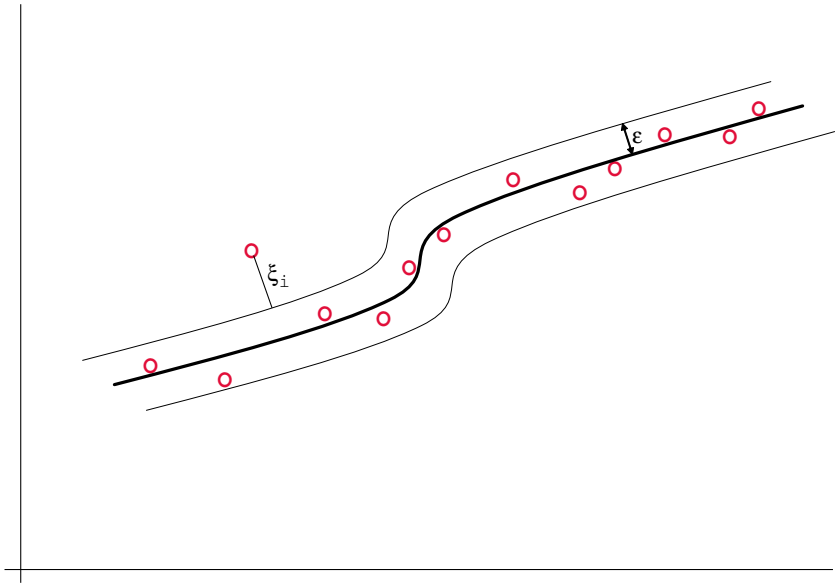

Figure 2.14.: An example of the insensitive band *varepsilon* for an SVM regression problem.

residuals. In least-squares regression for example the aim is to minimize the sum of squares of the residuals. In order to optimize the generalization bounds for an SVM-regression problem, a loss function that ignores errors within a certain distance from the true value is needed. This is done by introducing an insensitive band, referred to as ε-insensitive band. For points within this band the error ξ is zero, see Fig 2.14.

The optimization problem that has to be solved can then be formulated as:

$$\min \|\mathbf{w}\|^2 + C\sum_{i=1}^{N}(\xi_i^2 + \hat{\xi}_i^2) \tag{2.29}$$

subject to

$$(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle) + b) - y_i \leq \varepsilon + \xi_i, i = 1, \ldots, N,$$

$$y_i - (\langle \mathbf{w} \cdot \mathbf{x}_i \rangle) + b) \leq \varepsilon + \hat{\xi}_i, i = 1, \ldots, N,$$

$$\xi_i, \hat{\xi}_i \geq 0, \quad \forall i, 1 \leq i \leq N \tag{2.30}$$

where two new slack variables $(\xi, \hat{\xi})$ have been introduced. One of the slack variables is for values larger than ε and the other is for values lower than ε. The Lagrange formulation of the problem is then [63]:

$$\max \sum_{i=1}^{N} y_i(\hat{\alpha}_i - \alpha_i) - \varepsilon \sum_{i=1}^{N}(\hat{\alpha}_i + \alpha_i) - \frac{1}{2}\sum_{i=1}^{N} y_i(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)(\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle + \frac{1}{C}\delta_{ij})$$

subject to

$$\sum_{i=1}^{N}(\hat{\alpha}_i - \alpha_i) = 0,$$

$$\hat{\alpha} \geq 0, \alpha_i \geq 0, 1 \leq i \leq N \tag{2.31}$$

By some simple substitutions and introducing a kernel function, the optimization problem that needs to be solved looks pretty similar to the classification case:

$$\max W(\alpha) = \sum_{i=1}^{N} y_i\alpha_i - \varepsilon \sum_{i=1}^{N} |\alpha_i| - \frac{1}{2}\sum_{i=1}^{N} \alpha_i\alpha_j(K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{C}\delta_{ij})$$

subject to

$$\sum_{i=1}^{N} \alpha_i = 0 \tag{2.32}$$

As similar procedure as for classification can be used to find the maximum of this function.

SVM implementation

There are several different SVM implementations available, but here the SVM^{light} package was used. SVM^{light} is an implementation of Vapnik's Support Vector Machine for the problem of pattern recognition [295]. The optimization algorithm used in SVM^{light} has adjustable memory requirements and can handle problems with many thousands of support vectors efficiently [136]. The problem with general off-the-shelf optimization techniques is that they become infeasible in their time and memory requirements if the learning task is hard. The implementation of SVM^{light} is an SVM learner which addresses the problem of large tasks, which makes large scale training more practical. The memory requirements are linear with the number of training examples and with the number of support vectors. Nevertheless, the

algorithm gains from additional storage space, since a caching strategy allows an elegant trade-off between training time and memory consumption [136]. More information and other SVM implementations can be found at: http://www.kernel-machines.org/.

2.9.2. Decision trees

A decision tree is a directed acyclic graph where the leaves represent classifications and the branches features that leads to a certain classification [177]. There are two major algorithms existing for generating decision trees, CART [34] and C4.5/C5.0 [215]. Both methods recursively split the input data according to an attribute value test, where the aim is to find the attribute that give the maximum separation of the data. One major difference between the methods lies in the measure of performance gain given a certain split. The CART method uses the Gini index for this purpose, whereas C4.5 uses the entropy gain. In this thesis C4.5 and C5.0 were used, which both build on the ID3 algorithm developed by Quinlan. The principles of ID3 are briefly described in order to clarify the construction of a decision tree. Given a dataset of n examples and m attributes $(x_{ij} \dots x_{nm}, y_i)$, the aim is to create a decision tree based on the attributes that correctly classify the data. The root node of the tree is selected by searching for the attribute that best separate the training examples into the prediction classes y_i. In ID3, C4.5 and C5.0 this attribute is found by calculating the entropy gain given a split according to a certain feature. Given a collection (S) of n classes (c) the entropy (E) is defined as:

$$E(S) = \sum_{i=1}^{n} p(c_i) \log_2 p(c_i) \tag{2.33}$$

where $p(c)$ is the proportion of S belonging to class c_i. The entropy gain (E_G) given a split using attribute A is defined as:

$$E_G(S, A) = E(S) - \sum_{i=1}^{n} \frac{|S_j|}{S} log_2 \frac{|S_j|}{S} \tag{2.34}$$

where S_j is the subset of S where attribute A has the value j. Once the best attribute has been found, the training examples are portioned according their values of that attribute and the algorithms then try to split these subsets according to the maximal information gains. This procedure is repeated recursively with some stop criteria (e.g. when all examples after a split belong to one class). In this work decision rules were generated from the decision trees. These rules then be easily applied for prediction purposes and are easy to interpret. More

detail and examples of decision rules generated are given in Sect. 3.3.

2.10. Performance evaluation

The aim of the machine learning methods introduced above is to extract general patterns and trends from the data which can be formulated into a model. The model can then subsequently be used for prediction. When machine learning approaches are applied, it is important to remember some features of model complexity in respect of the training and test errors. A highly complex model can usually reproduce the training data perfectly, whereas the performance on data not included in the training procedure is poor. This problem is usually called overfitting and in general the training error is not a good measure of the test error [106]. The following sections describe the performance statistics measures used in this thesis and how cross-validation can be used to get an estimate of the predictive power of models.

2.10.1. Performance measures

Many different measures can be used for prediction accuracy, e.g. raw percentages, quadratic error measures, and correlation coefficients. The most frequently used measures throughout this thesis is the Matthews correlation coefficient (MCC) [172] and the standard Pearson correlation coefficient. In some cases the ranks of predicted values are also interesting to compare to the rank of experimentally measured values and here the Spearman rank correlation is used. The raw percentage correct prediction is not used widely in this thesis, since this might be misleading in many cases. Assume that the ratio between positive:negative examples in the test set is 1:9. If the classification always classifies all test data as negative, the percentage correct predictions would be 90%. The prediction method is of course not good even if the percentage correct predictions in this case is high (no positives are found at all).

Matthews correlation coefficient and related measures: The typical classification task is to assign an example into one of two classes. Here we will refer to these as the 'positive' and 'negative' classes. Four variables are defined and used for this purpose: true positives (TP) - the number of binders predicted as such, true negatives (TN) - the number of non-binders predicted as such, false positives (FP) - the number of predicted binders that actually are non-binders, and false negative (FN) - the number of predicted non-binders actually that

actually are binders. From these values the Matthews correlation coefficient (MCC) can be defined as:

$$\text{MCC}=\frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TN+FN)(TN+FP)(TP+FN)(TP+FP)}} \tag{2.35}$$

A perfect correlation between predicted and real values would give an MCC of 1, random predictions an MCC of 0, and anti-correlated predictions a value of -1. Furthermore the specificity (SP) and sensitivity (SE) of the prediction can be defined as:

$$\text{SP}=\frac{TP}{TP+FP} \tag{2.36}$$

$$\text{SE}=\frac{TP}{TP+FN} \tag{2.37}$$

In medical statistics specificity is often use to describe the prediction of negative examples (TN/(TN+FP)), which is better described as the sensitivity of the negative category [14].

Spearmans rank correlation: The Spearman rank correlation is a modified version of the standard correlation and compares the ranks between two data sets. There might be "gaps" in the prediction scores and this type of statistics will give a measure on how good the methods are in producing a correct ranking of the data. Spearman's rank correlation (ρ) is defined as [56]:

$$\rho = \frac{\sum_{i=1}^{n} R(X_i)R(Y_i) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\sum R(X_i)^2 - n\left(\frac{n+1}{2}\right)}} \tag{2.38}$$

which is equivalent to the Pearson correlation calculated on the ranks and average ranks.

2.10.2. Cross-validation

Cross-validation is a procedure where parts of the training data taken out from the training procedure and used for testing the performance. By doing this a pretty fair estimate of the methods ability to generalize can be obtained. The idea of cross-validation is to split the training set randomly into N subsets. N-1 sets are then used for training and the remaining set is used for testing the performance [106]. The procedure is repeated for all the N possible subsets omitted from the training procedure. One drawback of this approach is that the training procedure must be repeated N times. The choice of N often depends on data set size, but in many cases the same N as in similar methods is used for comparison reasons.

3. Prediction of MHC class I binding peptides

The most extensively studied step of the antigen processing pathway is MHC-peptide binding. In comparison to proteasomal cleavage and TAP transport, a lot of experimental data is available for different aspects of MHC-peptide binding and T-cell activation. Qualitative data includes sequences of naturally processed peptides, whereas quantitative data includes peptides with experimentally measured binding affinity to a certain MHC allele. Furthermore, crystal structures of many MHC-peptide complexes are known, which serve as a starting point for structure-based prediction.

Only a very limited number of peptides from a protein usually bind a certain MHC allele [303]. The aim of MHC-peptide prediction methods is to identify these peptides, which will reduce the number of peptides that have to be investigated experimentally. The typical output of the prediction methods is a ranked list of peptides, where the top one is the most likely to bind the MHC allele of interest.

Here different sequence-based methods for prediction of MHC-binding peptides are investigated. The first focus of this chapter is how SVMs can be applied for prediction of MHC-binding peptides. SVMs make it possible to circumvent the assumption used by PSSM methods that each amino acid of the peptide contributes to the overall binding energy independent of its surrounding amino acids. In a comparison to the SYFPEITHI and BIMAS methods, the SVM method SVMHC shows better prediction accuracy for most alleles. Some computational aspects regarding dataset homology and number of sequences needed for SVM training is also investigated.

The next approach presented is a consensus method for HLA-A*0201 using all the prediction methods mentioned above. The consensus score is generated by considering the score distributions of sets of semi-quantitative data (high, medium, and low affinity binders) from the MHCPEP database. A score is assigned to a new peptide according to its probability to belong to one of the binding strength classes and the scores from all methods are then summed up into a consensus score.

The third main part of this chapter discusses the use of alternative sequence representation and decision trees for prediction. The aim here is to move away from the "black box" SVMs and to investigate if simple biochemical properties of the amino acids can be used to qualitatively describe MHC-peptide binding.

3.1. SVMHC

SVMHC is the first method described for prediction of MHC class I binding peptides using support vector machines (SVMs). The idea of using SVMs and sparse binary representation of peptides for prediction was introduced in a previous work [71]. This early work only considered peptides from the MHCPEP database and compared the use of SVMs and ANNs. The work presented here extends this by considering peptide homology, number of peptides needed for prediction, and naturally processed peptides from the SYFPEITHI database. Furthermore, comparative studies to other prediction methods have been done. For a complete overview of SVMHC, this chapter describes all steps of the development process, including data extraction and representation.

3.1.1. Data and data representation

The data used to develop SVMHC was extracted from the MHCPEP [38] and SYFPEITHI databases [217] described earlier. The peptides were extracted from the databases and grouped into allele-specific data sets. Some peptides contain undetermined amino acids (labeled 'X') and such peptides were removed from the data sets. Unfortunately, there are very few experimentally verified examples of peptides that do not bind to a particular MHC (or at least very few are published). Therefore, the non-binding training examples were extracted randomly from the ENSEMBL database of human proteins [124]. Protein sequences from the ENSEMBL database were chopped up into the length of interest and known MHC-peptides were removed. Obviously, there is a risk that some of the non-binders actually do bind, but since less than 1% of the peptides are expected to bind any given MHC molecule [167], this is a valid approach. Details about data set sizes are given in the following sections.

The peptides were represented using binary "sparse" representation. This means that a binary string is assigned to each amino acid in the peptide. Each binary string consists of 20 positions with zeros in all places except for one. An example illustrating how a peptide is represented using binary sparse encoding is given in Fig. 3.1.

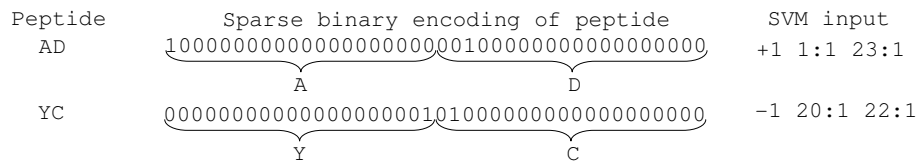

Figure 3.1.: This figure illustrates how the two dipeptides "AD" and "YC" are represented using binary sparse encoding. Furthermore, it is shown how the peptides are encoded for the SVM software, where "+1" represent a binder and "-1" a non-binder. The SVM software reads data in the form: *class feature1:value feature2:values*, but for those *feature:value* pairs now given in the input file, the software will consider the value of these features to be zero.

3.1.2. SVM training and evaluation

The linear, polynomial, and radial-basis function (rbf) kernels were all optimized during training. The parameter space was explored using a grid search, where all parameter combinations were tested given the start, stop, and step size of each parameter.

The prediction performance of SVMHC was compared to that of the SYFPEITHI [217] and BIMAS [201] prediction methods. A thorough comparison was made for those alleles that could be predicted by all three methods.

3.1.3. Amount of data needed for SVM training

The numbers of known binders vary significantly between different MHC alleles. One problem faced with a learning algorithm trying to deduce a model from a given data set is to estimate the number of training examples needed for creating a reliable model. To investigate the effects of training data set size, three MHC alleles for which a reasonable number of known MHC-binding peptides are known were used. For each of these alleles a test data set consisting of 20 binders and 40 non-binders were extracted and used throughout the study. The sizes of the training data sets were then varied including different numbers of known binders. The ratio between binders and non-binders in the training data sets were kept at a 1:2 ratio in order to keep this factor as constant as possible. For each dataset a grid search for the optimal SVM parameters was done. Figure 3.2 shows the MCC for varying numbers of binders included in the training data set for the three alleles HLA-A*0201, HLA-A*03, and HLA-B*2705. A significant improvement of the MCC can be observed as the number of binders was increased up to about 20 for each of the alleles. Furthermore, a small increase can be observed as the number of binders increase up to 50. The conclusion that can be drawn from this study is that about 20 peptides is a reasonable minimum number of peptides

needed for training in order to obtain a reasonable prediction model.

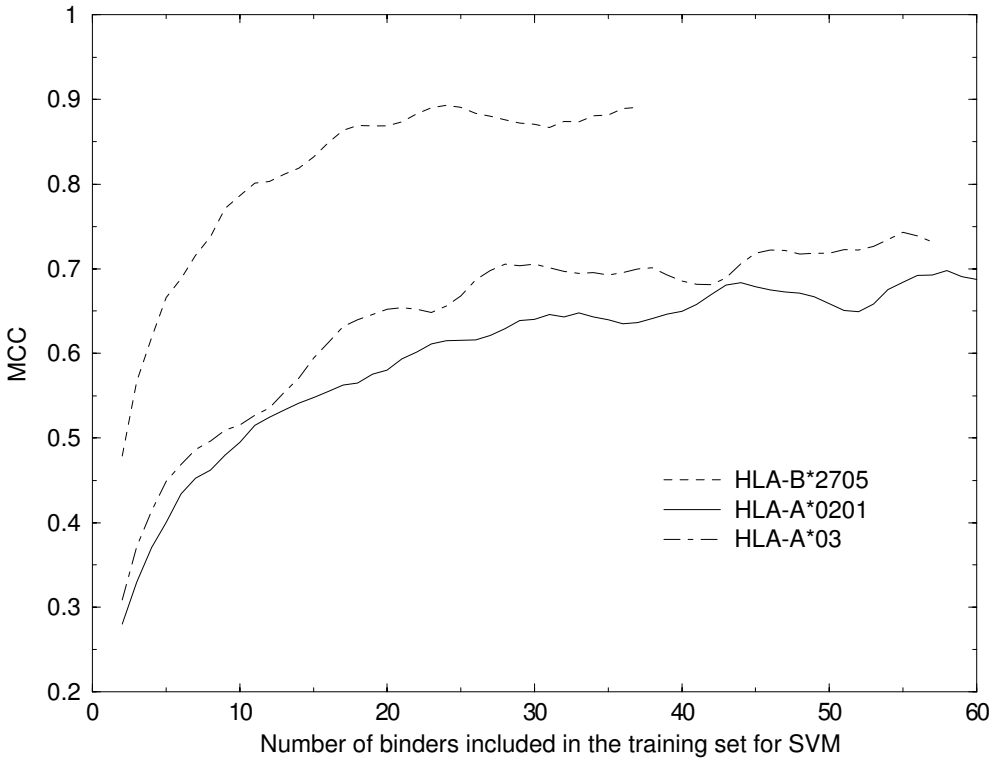

Figure 3.2.: Performance of SVMHC for the three MHC alleles HLA-A*0201, HLA-A*03, and HLA-B*2705, measured by the Matthews correlation coefficient (MCC) as a function of the amount of training data used. For all sizes of training data sets, the test set was kept constant and no part of the training data was a part of the test data.

3.1.4. Redundancy/homology reduction

Algorithms like SVMs and ANNs typically have a large number of free parameters and the use of too homologous datapoints might give misleading results. This is a frequently occurring problem in bioinformatics that is often being neglected by researchers and peer-reviewers. It is a trivial task for e.g. an SVM or a network-based model to almost perfectly reproduce the training data. Different alleles were studied in order to investigate the effects of homology reduction. Peptides in the training data was only allowed a maximum pairwise sequence identity to all other peptides. The results from this study can be seen in Fig. 3.3. As expected there is a small increase in performance as the number of maximum identity is increased. However, the effects of this reduction is pretty small and by removing identical sequences from the data an over-estimation of the prediction performance should be prevented.

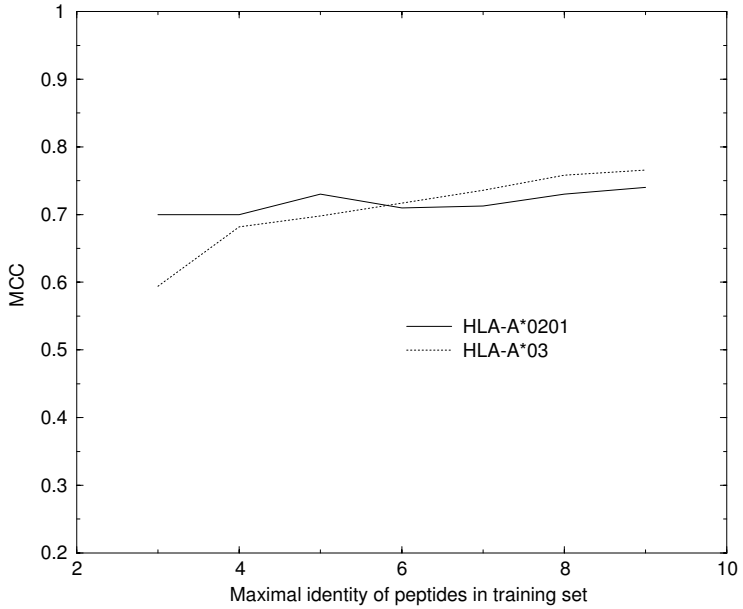

Figure 3.3.: The effect of sequence identity reduction on prediction accuracy. The accuracy increase is only small considering the identity of sequences within the data, indicating that the SVM models have learned general patterns of the data.

3.1.5. SVMHC training results

The first version of SVMHC was published in 2002 [72] and offered prediction for 26 HLA alleles from the MHCPEP database and six HLA alleles based on SYFPEITHI data. The performance values obtained for data from the MHCPEP database can be seen in Table 3.1. For some alleles the prediction accuracies are very high with MCCs over 0.90 and most MCCs are above 0.70. The MHCPEP database is no longer updated, whereas the SYFPEITHI database has undergone several updates. Hence, SVMHC has also been trained on later releases of SYFPEITHI and now also include some of the most common mouse alleles. Table 3.2 gives and overview of the training results from the two SVMHC versions based on SYFPEITHI data. Version 1 corresponds to the results published in 2002 [72]. The first version of SVMHC only allowed a maximal identity of six positions between peptides in the training data, which is not the case for the second version. One main reason for removing similar peptides is that the MHCPEP database contains peptides from alanine scans, meaning they have a large number of similar sequences only changed in one position, which is not the case for SYFPEITHI data. For some alleles the performance gets better when more data is available, but this is not true for all cases. The reason for this is most probably that the new data gives a better description of the 'peptide range' a certain MHC allele will bind. The data for some alleles is highly biased towards certain motif patterns that have been used to identify candidates for

Table 3.1.: Prediction performance for SVMHC based on data from the MHCPEP database. Results are available for 26 different HLA molecules with a total of 31 HLA/peptide length (Mer) combinations. As can be seen the majority of the data from the MHCPEP database consists of 9mer peptides.

MHC	**Mer**	**# binders**	**MCC**	**Kernel**
HLA-A*01	9	28	0.95	lin
HLA-A*1101	9	40	0.74	poly
HLA-A*11	9	46	0.75	rbf
HLA-A*11	10	21	0.59	poly
HLA-A*02	9	118	0.76	poly
HLA-A*02	10	35	0.65	poly
HLA-A*2402	9	73	0.90	poly
HLA-A*03	9	73	0.76	rbf
HLA-A*0201	9	184	0.73	rbf
HLA-A*0201	10	96	0.78	poly
HLA-A*3301	9	32	0.72	lin
HLA-A*0301	9	38	0.72	rbf
HLA-A*0301	10	32	0.77	lin
HLA-A*31	9	39	0.79	poly
HLA-A*6801	9	42	0.84	poly
HLA-B*07	9	32	0.95	lin
HLA-B*08	9	26	0.77	poly
HLA-B*2705	9	41	0.93	lin
HLA-B*3501	9	67	0.93	lin
HLA-B*3501	10	34	0.96	poly
HLA-B*35	9	23	0.71	lin
HLA-B*2703	9	22	0.90	lin
HLA-B*5301	9	41	0.95	lin
HLA-B*27	9	34	0.91	rbf
HLA-B*2706	9	20	0.93	lin
HLA-B*51	9	67	0.82	poly
HLA-B*5102	9	29	0.79	poly
HLA-B*0702	9	52	0.96	poly
HLA-B*5103	9	29	0.84	rbf
HLA-B*5401	9	42	0.98	lin
HLA-B*5101	9	35	0.89	lin

Table 3.2.: Prediction performance for SVMHC based on data from the SYFPEITHI database. Results are presented for two SVMHC version and Ver. 2.0 of SVMHC covers a much wider range of peptides.

		Version 1			Version 2		
MHC	**Mer**	**# binders**	**MCC**	**Kernel**	**# binders**	**MCC**	**Kernel**
H2-Db	9	-	-	-	32	0.88	lin
H2-Kb	8	-	-	-	46	0.92	rbf
H2-Kd	9	-	-	-	38	0.94	lin
H2-Kk	8	-	-	-	23	0.91	lin
HLA-A*0201	9	113	0.78	rbf	241	0.87	rbf
HLA-A*0201	10	40	0.70	poly	59	0.69	lin
HLA-A*01	9	28	0.96	lin	50	0.96	rbf
HLA-A*03	9	73	0.80	lin	78	0.89	rbf
HLA-A*03	10	-	-	-	25	0.79	lin
HLA-A*1101	9	-	-	-	30	0.93	lin
HLA-A*2402	9	-	-	-	28	0.84	rbf
HLA-A*24	9	-	-	-	36	0.92	rbf
HLA-A*25	10	-	-	-	20	0.82	rbf
HLA-B*07	9	23	0.93	lin	44	0.81	lin
HLA-B*08	9	25	0.79	lin	32	0.81	lin
HLA-B*1501	9	-	-	-	62	0.87	rbf
HLA-B*1501	10	-	-	-	22	0.86	lin
HLA-B*1801	8	-	-	-	49	0.83	lin
HLA-B*1801	9	-	-	-	70	0.93	lin
HLA-B*2705	9	29	1.00	lin	45	0.98	lin
HLA-B*27	9	-	-	-	47	0.97	lin
HLA-B*3701	9	-	-	-	23	0.84	lin
HLA-B*44	9	-	-	-	26	0.94	rbf
HLA-B*5101	9	-	-	-	33	0.91	rbf
HLA-Cw*0401	9	-	-	-	54	0.93	rbf

experimental verification. New techniques enable more peptides to be extracted from a cell culture and this seems to reveal new motif patterns.

3.1.6. SVMHC benchmarking results

Comparison of new prediction methods to the current state-of-the-art methods is important. Hence, SVMHC was compared to the BIMAS and SYFPEITHI prediction methods, see Table 3.3 (results using SVMHC Version 1 [72]). SVMHC performs better than the other two methods for all alleles except for HLA-B*08, where BIMAS shows the best performance. For one prediction method, the difference in performance is rather large in some cases. This can be explained by the difference in the peptide sequences found to by a certain MHC allele.

For example all HLA-B*2705 peptides have and Arg in position two of the peptide. This is a rather dominant motif compared to other alleles which allow a broader range of amino acids as anchors. It is not possible from this study to determine which of the BIMAS and SYFPEITHI methods that is the best one, since this is rather allele dependent. Some more aspects of these results are given at the end of this chapter.

Table 3.3.: Comparison results of SVMHC, SYFPEITHI (SYF) and BIMAS. The tables shows the allele-specific Matthews correlation coefficients (MCC) of each method.

MHC	Mer	# binders	SVMHC	SYFPEITHI	BIMAS
HLA-A*0201	9	113	0.78	0.77	0.77
HLA-A*0201	10	40	0.70	0.61	0.61
HLA-A*01	9	28	0.96	0.93	0.96
HLA-A*03	9	73	0.80	0.73	0.71
HLA-B*08	9	25	0.79	0.79	0.82
HLA-B*2705	9	29	1.00	0.92	0.93
Avg	-	-	0.84	0.79	0.80

Both SVMHC and SYFPEITHI have been updated since this initial comparison was conducted. Hence an additional comparison of the SVMHC version 2 and SYFPEITHI (matrices from March 2006) was done, see Table 3.4. This comparison is done exclusively for peptides with a length of nine amino acids and the amount of data used is much larger than in the previous comparison. The cutoff used for the SYFPEITHI method is half the maximum score for the relevant matrix as described by Schuler *et al.* [245].

Table 3.4.: Comparison results of SVMHC version 2 and SYFPEITHI (matrices from March 2006). The tables shows the allele-specific Matthews correlation coefficients (MCC) of each method.

MHC	# binders	SVMHC	SYFPEITHI
HLA-A*0201	241	0.87	0.83
HLA-A*01	50	0.96	0.94
HLA-A*03	78	0.89	0.75
HLA-B*08	32	0.81	0.81
HLA-B*2705	45	0.98	0.92
Avg	-	0.90	0.85

3.1.7. Quantitative prediction

The presented accuracies for SVMHC concern the classification task of binders versus non-binders, i.e. a qualitative prediction. However, for some alleles enough quantitative binding

data is available for a statistical evaluation. Most quantitative data come from assays measuring the binding affinity of peptides to an MHC molecules using radiolabeled standard probes, as described by Sette *et al.* [138, 153, 237]. A large set of HLA-A*0201 peptides with experimentally verified binding energies were used to evaluate the correlation between experiment and prediction for the SVMHC, SYFPEITHI, and BIMAS methods. The scores from the BIMAS methods are estimated $t_{1/2}$ dissociation rates of the β_2-microglobulin subunit which is in a logarithmic relationship to the binding energy. Hence the logarithm of the BIMAS scores were used. Two measures of performance were used in this study, the standard Pearson correlation and the Spearman rank correlation.

The results from this study can be seen in Table 3.5 and a plot of experimentally versus predicted values can be seen in Fig. 3.4. The BIMAS log values show slightly better results compared to SVMHC. There is not much difference between the Spearman and Pearson correlation for each method, except for the BIMAS comparison using the raw output scores.

Table 3.5.: Correlation coefficients for the quantitative data set of HLA-A*0201 binders.

Method	**Spearman**	**Pearson**
BIMAS	-0.52	-0.51
SYFPEITHI	-0.33	-0.35
SVMHC	-0.45	-0.47

3.1.8. The SVMHC prediction server

SVMHC has been implemented as a prediction server and can be accessed at http://www-bs.informatik.uni-tuebingen.de/SVMHC/. The current version allows prediction of 26 alleles based on MHCPEP data and 21 alleles based on SYFPEITHI data. Furthermore, MHC class II prediction is available using the matrices presented by Hammer et al. [270]. These matrices are used by the TEPITOPE software and have been proven effective in identifying T-cell epitopes in the melanoma related protein gp100 [53] and MAGE-3 [168]. SVMHC currently contains 50 HLA-DR alleles.

There are several different output formats available from SVMHC, both graphical and tabular. Figure 3.5 shows a sequence-summary plot for eight different alleles of the a lentiviral protein (Swiss-Prot id: P05917). Amino acids that are part of a predicted MHC-binding peptide are colored red and the starting position of each peptide is colored blue. This enables a fast identification of promiscuous epitopes, i.e. epitopes shared over several different MHC alleles. Peptides that can bind a wide range of different MHC alleles are good candidates for

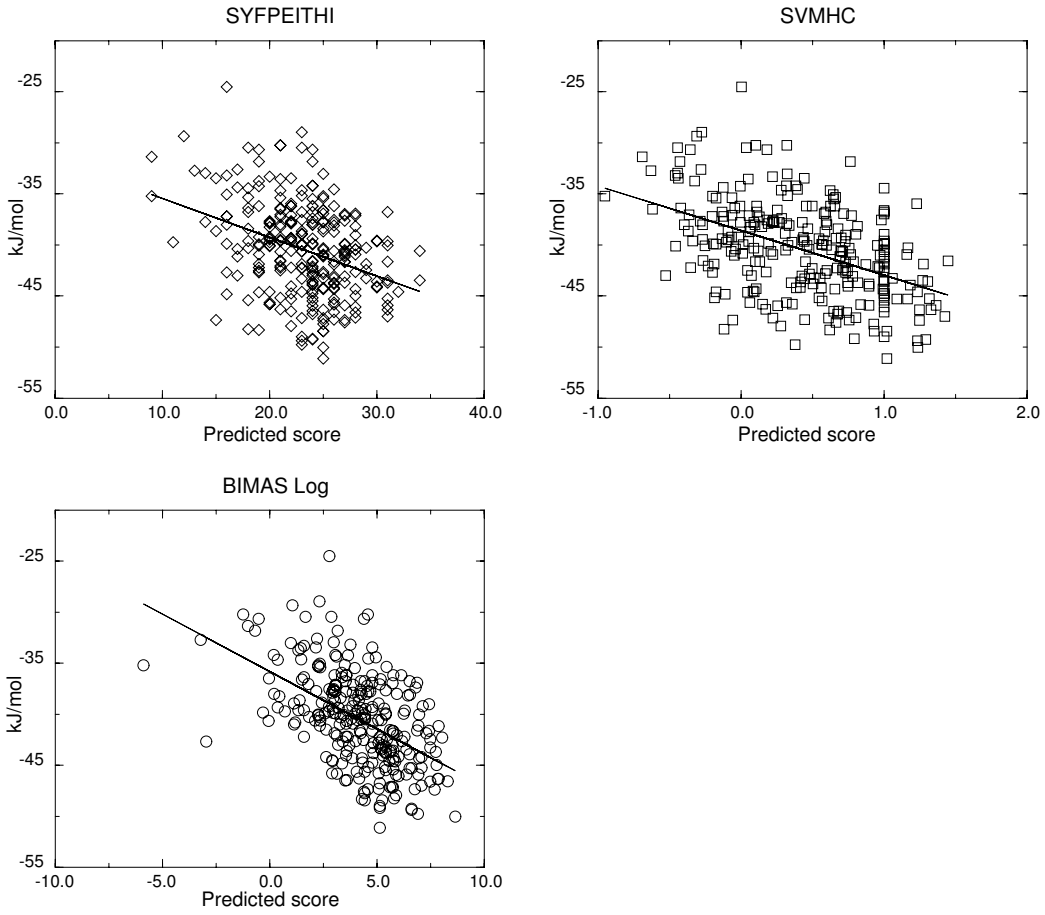

Figure 3.4.: Comparison of prediction scores (x-axes) and experimentally verified binding energies (y-axes).

vaccines, covering a wide range of the population.

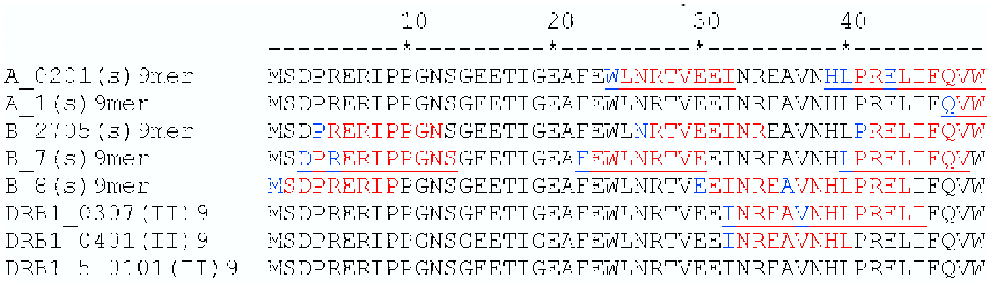

Figure 3.5.: Prediction results from the SVMHC prediction server. This type of plot enables simple identification of promiscuous epitopes.

3.2. Consensus prediction of HLA-A*0201 binding peptides

Methods for MHC-binding prediction usually differ in both data used and computational approach. One strategy often used in bioinformatics is to combine different prediction servers into a consensus score. This has successfully been done for e.g. protein secondary structure prediction [64], transmembrane helix prediction [301] and protein structure prediction [164]. Most approaches use something like a majority-wins schema or use a more sophisticated way to combine the scores. This section presents results for a consensus prediction methods for HLA-A*0201 binding peptides using the SVMHC, SYFPEITHI, and BIMAS methods. There are many different ways in which a consensus method can be constructed, e.g. one could sum the ranks from all three methods for each peptide from a query protein and use this as a consensus score. However, here a more sophisticated approach is presented, based on score distributions from the different prediction methods. The MHCPEP database contains semi-quantitative binding information and groups the peptides into three classes based on binding strength (high, medium, low binders). By studying how the scores of each of these classes given a prediction method is distributed, a peptide can be assigned to a class given a certain score, i.e. given a certain score and the score distribution of the binding-strength classes, to which class does the peptide (score) belong.

3.2.1. Materials and methods

The MHCPEP database contains semi-quantitative binding data for pretty many peptides. The binding strengths of these peptides are classified into low, medium or high. All HLA-A*0201 peptide where this information was available were extracted from the database and non-binders were extracted as described for the SVMHC method, giving four data sets with peptides grouped according to binding strength. The aim now is to assign a new peptide to one of the classes based on its prediction score alone. This means that we have four models $M_m, m = 1, ..., 4$ corresponding to the four binding-strength classes. The question is now: given the score of a new peptide is by which probability it belongs to one of the classes. A simple way to do this is to group the scores into intervals and calculate the fraction of peptides covered. An example of this for the SYFPEITHI prediction method on the data from MHCPEP can be seen in Fig 3.6. Peptides belonging to the Low-binder class are more frequent in lower score ranges, although some High-binders also have low scores. The same type of distribution plots can be done for non-binders and the class of a new peptide (score) is

simply assigned to the class which is most prominent in the relevant interval. This is done for all prediction methods and the peptide is assigned a score from each method corresponding to its class (0 to 3 for non-binder to high binder). The consensus score of a peptide is then the sum of scores from all prediction methods (scores in the range 0 to 9).

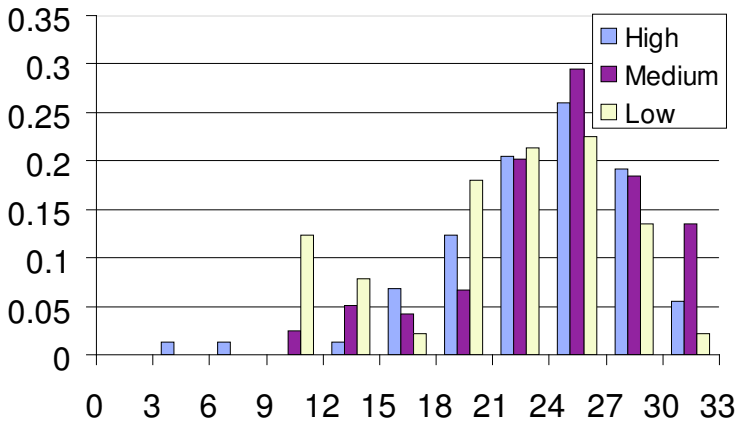

Figure 3.6.: The fraction of peptides within certain score intervals as predicted by the SYFPEITHI method for the semi-quantitative data from the MHCPEP database.

3.2.2. Results and interpretation

The consensus method was compared to the SVMHC, the BIMAS, and the SYFPEIHTI methods. In this comparison all classes of binders (low, medium, and high) were considered just binders and the classification between binders and non-binders was investigated. The results from this comparison can be seen in Table 3.6.

Table 3.6.: Prediction results for the BIMAS, SYFPEITHI, SVMHC, and Consensus methods. The Consensus method shows better prediction performance compared to the other methods.

Method	MCC	SP	SE
BIMAS	0.66	0.75	0.75
SYFEITHI	0.70	0.81	0.73
SVMHC	0.71	0.89	0.67
Consensus	**0.73**	**0.87**	**0.87**

These results clearly show that the best performance is achieved with the Consensus method. Furthermore, a specificity/sensitivity plot for the prediction can be seen in Fig. 3.7. From this plot it can also be seen that the Consensus method performs better than the other

methods and especially has a range of improved specificity.

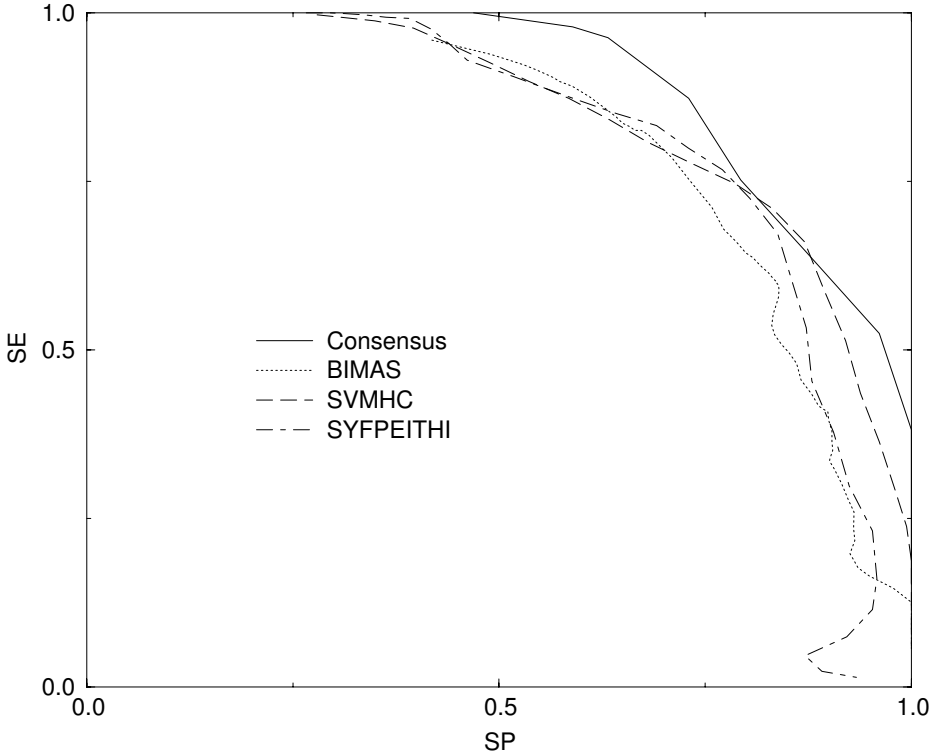

Figure 3.7.: Sensitivity/specificity plot for the SVMHC, BIMAS, and SYFPEITHI methods compared to that of the consensus method.

3.3. Decision trees and amino acid-specific properties for prediction of MHC class I binding peptides

MHC-peptide binding is highly dependent on the amino acids in the anchor positions. Earlier in this chapter, the successful use of SVMs for the prediction of MHC-peptide binding was described in detail and the SVMHC method was presented. The amino acids in all methods described for MHC-binding prediction are described as an alphabet consisting of 20 letters. However, the MHC-peptide interaction is of course dependent on the biochemical properties of these amino acids and there are no explicit similarity described by the 20 letter encoding. One way to describe the biochemical properties of the amino acids is to use the AAindex database [139] . AAindex contains different types of amino acid properties, e.g. hydrophobicity and accessible surface area, that can be used to represent the peptides. Here DTs are used together with amino acid properties to investigate the underlying properties of MHC-peptide interaction. DTs are used since the decision rules generated can be easily interpreted

and give some basic ideas about the MHC-peptide interaction.

3.3.1. Materials and methods

Experimentally verified MHC binding peptides with a length of nine amino acids for the HLA-A*0201, HLA-B*08, and HLA-B*2705 alleles were obtained from the SYFPEITHI database [217]. This gave a total of 241 HLA-A*0201, 45 HLA-B*2705, and 32 HLA-B*08 peptides. A dataset of non-binders was created by extracting peptides from existing proteins in the Ensembl database [124] as described for SVMHC. The number of non-binders used in the training data was twice the size of known binders for each allele. Duplicate entries were removed from the data sets and peptides were encoded using amino acid-specific properties (e.g. hydrophobicity, side chain volume, and absolute entropy) from the AAindex database [139]. Many of the different features in the AAindex database and hence a set of 24 representative features were used throughout this study, see Table 3.7.

Table 3.7.: The features from the AAindex database used in the DT study.

Name	Definition	AAindex accession code
AbsEntropy	Absolute entropy	HUTJ700102
Accessibility	Information value for accessibility	BIOV880101
AccSurfaceArea	Average accessible surface area	JANJ780101
AlphaHelix	Alpha-helix indices	GEIM800101
BetaTurn	Conformational parameter of beta-turn	BEGF750103
BetaStruct	Conformational parameter of beta-structure	BEGF750102
BurResidue	Percentage of buried residues	JANJ780102
ExpResidue	Percentage of exposed residues	JANJ780103
HydrophobDir	Direction of hydrophobic moment	EISD860103
Interaction	Side chain interaction parameter	KRIW710101
LengthSideChain	Length of the side chain	FAUJ880104
MaxWidthSideChain	Maximum width of the side chain	FAUJ880106
MinWidthSideChain	Minimum width of the side chain	FAUJ880105
MolWeight	Molecular weight	FASG760101
NetCharge	Net charge	KLEP840101
NonBindEnergy	Average non-bonded energy per atom	OOBM770101
PosCharge	Positive charge	FAUJ880111
SideChainGyr	Radius of gyration of side chain	LEVM760105
SideChainVol	Side chain volume	KRIW790103
SideOrient	Average side chain orientation angle	MEIH800103
Stability	Side-chain contribution to protein stability	TAKK010101
SurrHydro	Average gain in surrounding hydrophobicity	PONP800102
SurrResidue	Average number of surrounding residues	PONP800108
WaalsVolume	Normalized van der Waals volume	FAUJ880103

The well known C4.5 and C5.0 software packages [216] were used to create rulesets that consist of unordered collections of if-then rules. Rulesets were preferred instead of decision trees since they give clear descriptions of the rules associated with a certain class. In cases where more than one rule applies, the C5.0 program takes the confidence value of each rule into account to calculate a total vote for each class. Furthermore, there is a default class that is used when none of the rules in the ruleset is applicable. The aim with the sampling procedure is to find amino acid patterns, in terms of biochemical properties describing MHC-peptide binding.

In a preliminary analysis every amino acid property was evaluated separately considering all peptide positions. The rulesets generated were then searched for key positions which were used for the subsequent analysis (in order to reduce the search space). In the next step an extensive search for feature/positions combinations was conducted. In most cases a limited number of peptide positions and amino acid features gave the best prediction results. The prediction performance was evaluated using Matthews correlation coefficient [172] and fivefold cross-validation. The ruleset method was also compared against the SYFPEITHI and BIMAS methods by applying the same data and statistics to the online versions of these methods.

3.3.2. Results and interpretation

The performance for the best combination of peptide positions and amino acid properties found for each allele are presented in Table 3.8. The general conclusion from these results is that only a limited number of peptide positions and amino acid properties are needed to describe MHC-peptide binding. This also means that the dimensionality of the classification problem can be reduced, since there is no need to take all sequence positions into account. We also investigated the effect of data splits into training and test sets for the best results, giving only small differences to the results presented here.

Figure 3.8 shows the rulesets generated for HLA-B*2705 (a) and HLA-B*08 (b) using the whole datasets. The rules for HLA-B*2705 are based on two peptide positions and two amino acid properties. The two features found to be the most important for this allele are *Stability* and *LengthSideChain*. *Stability* describes the contribution to protein stability from a certain side-chain [274], whereas the *LengthSideChain* property is a size descriptor for the amino acids [86]. *Stability* is a feature closely correlated to the hydrophobicity of an amino acid, which can be seen in position nine of the peptide where hydrophobic amino acids are

Table 3.8.: Results for the best prediction accuracy achieved for each of the studied MHC alleles. The table shows the statistics obtained from fivefold cross-validation, the positions of the peptide considered, and the amino acid properties used.

MHC	MCC	SP	SE	ACC	POS	Properties
HLA-A*0201	0.85	0.95	0.89	0.93	2,4,6,9	BetaStruct,SideChainGyr,BurResidue
HLA-B*2705	0.95	1.0	0.94	0.98	2,9	Stability,LengthSideChain
HLA-B*08	0.92	0.97	0.97	0.97	3,5,9	PosCharge,Stability,AccSurfaceArea

a.

```
Rule 1:
 Stability_9 > 0.2024793
 LengthSideChain_2 > 0.9756944
 ->  class epitope

Rule 2:
 LengthSideChain_2 <= 0.9756944
 ->  class non-epitope

Rule 3:
 Stability_9 <= 0.2024793
 ->  class non-epitope

Default class: non-epitope
```

b.

```
Rule 1:
 PosCharge_3 > 0
 AccSurfaceArea_9 <= 0.2057143
 ->  class epitope

Rule 2:
 PosCharge_5 > 0
 Stability_3 > 0.4917355
 ->  class epitope

Rule 3:
 AccSurfaceArea_9 > 0.2057143
 ->  class non-epitope

Rule 4:
 PosCharge_3 <= 0
 PosCharge_5 <= 0
 ->  class non-epitope

Rule 5:
 PosCharge_3 <= 0
 Stability_3 <= 0.4917355
 ->  class non-epitope

Default class: non-epitope
```

Figure 3.8.: The rulesets created for the HLA-B*2705 (a) and HLA-B*08 (b) alleles. The rules presented here were generated using the whole dataset for each allele.

preferred. It can be seen from the ruleset that HLA-B*2705 prefers amino acids with long side chains in position two of the peptide. This is important since the "binding pocket" of the MHC molecule is rather spacious. A small amino acid would not be able to fill the pocket. Space-filling effects like this are known to be important for protein structure stability and protein-ligand interaction [82, 127]. The small number of features needed to describe the relevant properties for binding gives a very compact model.

The ruleset for HLA-B*08 is based in three amino acid properties. *PosCharge* describes the charge of the amino acids [86] and *AccSurfaceArea* is a measure of how exposed a certain amino acid is to the solvent in known protein structures [133]. The importance of *PosCharge*

in position five of the peptide can clearly be seen (Rule 2 and Rule 4). This has been previously described in literature, where the positively charged amino acid lysine has been found in position five of the peptide [69]. *AccSurfaceArea* in position nine is also important for HLA-B*08 (Rule 1 and Rule 3) where small residues are preferred. Fig. 3.9 shows a cross-section of a HLA-B*08 molecule with a bound peptide. This figure clearly shows how the amino acids in positions 3, 5, and 9 of the peptide are deeply buried in the MHC molecule. These positions are in close contact with the MHC molecule and are crucial for binding, something also captured in the rulesets generated.

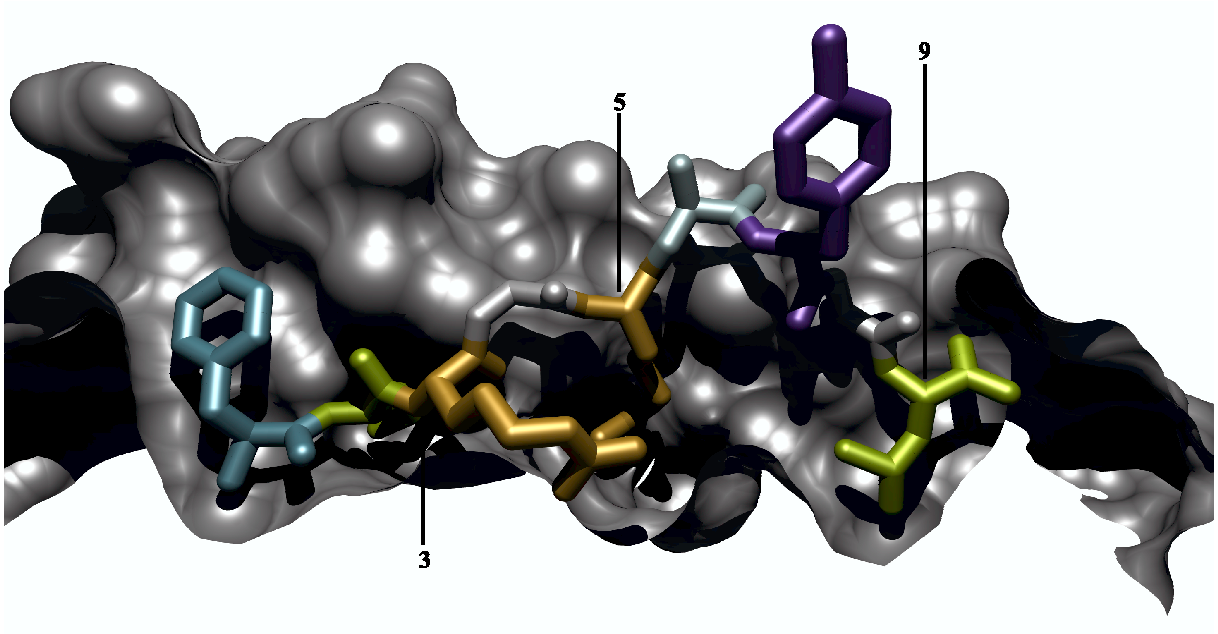

Figure 3.9.: This figure shows the cross-section an HLA-B*08 MHC molecule with a bound peptide(PDB code:1M05 [145]). Here the positions 3, 5, and 9 of the peptide can be seen to go deeply into the MHC molecule, something also reflected in the rulesets generated for HLA-B*08.

The results of the external methods SYFPEITHI and BIMAS can be seen in Table 3.9. SYFPEITHI performs better than BIMAS for the HLA-A*0201 and HLA-B*08 alleles, but is worse for HLA-B*2705. The ruleset method is better than both SYFPEITHI and BIMAS considering all alleles. The advantage of the ruleset method is that it finds the peptide positions and amino acid properties that best describe MHC-peptide interaction. In comparison to SVMHC, the DT method also performs well. Initial studies with SVMs based on only certain positions and amino acid features showed that this can improve the performance accuracy [271]. However, one drawback of DTs is that different optimal peptide positions and amino acid features are found for each MHC allele. Using all position and a limited number

Table 3.9.: Prediction performance of the SYFPEITHI and BIMAS methods for the three MHC alleles HLA-A*0201, HLA-B*08, and HLA-B*2705. The measures used for the performance evaluation are Matthew's correlation coefficient (MCC), specificity (SP), and sensitivity, (SE).

	A*0201			**B*2705**			**B*08**		
Method	MCC	SP	SE	MCC	SP	SE	MCC	SP	SE
SYFPEITHI	0.84	0.95	0.88	0.92	0.96	0.98	0.83	0.98	0.81
BIMAS	0.79	0.93	0.86	0.95	0.97	1.0	0.79	0.98	0.75

of features (e.g. the 24 used in this study) does not give good results.

3.4. General discussion

The results presented in this chapter show several different aspects of MHC-peptide binding. Using SVMs for prediction improves the overall accuracy and indicates that there are at least some peptides for which the independent contribution of each amino acid to the overall binding energy hypothesis does not hold. This was actually noted by Parker *et al.* in the paper on which the BIMAS method is based upon [201]. Many studies using PSSM turned the findings from Parker into their advantage, stating that most of the MHC-binding peptides can be described by a PSSM and did not investigate the fraction falsely predicted. The use of machine learning methods like SVMs or ANNs make is possible to model non-linear contributions of the amino acids, which is one reason for the improved performance. In most cases the improved performance of the SVM method is due to increased specificity of the prediction compared to the other methods, something that also have been pointed out previously [102]. The study of how the prediction methods perform on quantitative data is interesting. SVMHC and BIMAS perform significantly better in representing binding energies compared to SYFPEITHI. One reason for this is that the SYFPEITHI matrices have been tuned to predict naturally processed peptides. The data set used in the benchmark contains a mixture of synthetic and naturally processed peptides (although the binding energies of course are measured on synthetic peptides).

The consensus prediction method presented here shows promising results. Currently HLA-A*0201 is the only allele where enough experimental data is available for the approach presented here, but as more data become available, it might be interesting to try the same approach for other alleles as well. The overall increase in sensitivity means the consensus method successfully identifies binding peptides that are missed if the prediction methods are

used separately.

The DT approach gives some useful insights into MHC-peptide interaction. These can be used to give some qualitative description of MHC-peptide binding, e.g. peptides binding to a certain allele should have hydrophobic amino acids in position two and charged in position nine of the peptide. This can also be verified by manual inspection of MHC-peptide structures. The results also show that DTs are, from an algorithmic point of view, suitable for MHC-peptide binding prediction.

It is hard to "predict" the future of MHC-peptide prediction. One trend will definitely be to focus more on quantitative prediction of MHC-peptide affinity. This thesis presents some results for quantitative prediction and other both sequence based and structure-based attempts have been made. For most alleles the data needed is still lacking, but it is very likely that accurate prediction quantitative method for many different alleles will be presented in the future.

4. Modeling the whole MHC class I antigen processing pathway

MHC-peptide presentation on the surface of cells is crucial for the activation of Tc cells. The previous chapter described different methods for predicting MHC-peptide binding, but there are a number of other events involved in the processing and presentation of antigens not yet considered. The processing of proteins into smaller peptides is carried out by different proteases, where the proteasome is the most important for intracellular antigens. These are cleaved into smaller peptides in the cytosol and transported into the ER where they can bind MHC class I molecules (see Sect. 2.3.1 for an introduction to MHC class I antigen processing).

This chapter describes an integrated model of the whole processing pathway of MHC class I restricted antigens. The three steps modeled are proteasomal cleavage in the cytosol, TAP transport of peptides into the ER, and MHC-peptide binding. New methods for predicting these events are presented in detail and compared to existing methods. The new methods are then combined with SVMHC into a model of the whole antigen processing pathway (WAPP), see Fig 4.1 for a schematic overview. Furthermore, the prediction performance of WAPP is compared to NetCTL and the method presented by Tenzer *et al.* Finally, alternative events of the processing pathway are discussed with a focus on proteasomal splicing.

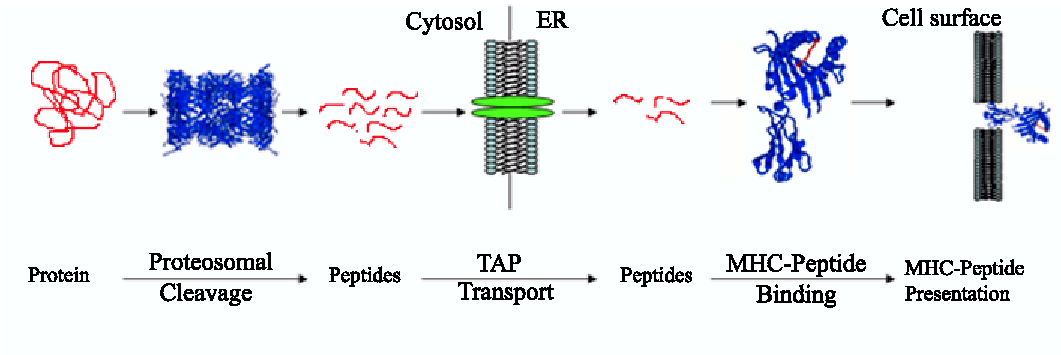

Figure 4.1.: A schematic overview of different processing steps modeled in WAPP.

4.1. Proteasomal cleavage prediction - the PCM method

The general structure and function of the proteasome was described in detail in Sect. 2.3.1 and prediction methods for proteasomal cleavage were discussed in Section 2.7.1. This section describes the development of the PCM (proteasomal cleavage matrix) method and compares the method to other methods for proteasomal cleavage prediction.

4.1.1. Materials and methods

Data for proteasomal cleavage is pretty sparse and whole protein digestion experiments have currently been done on only three proteins; β-casein [80], enolase [195], and the prion protein [276]. The experimental protocol can in a simplified way be described by three steps: (i) incubate the protein with purified proteasomes, (ii) extract the peptide fragments generated by proteasomal cleavage and characterize these with mass spectrometry, and (iii) map the fragments back to the protein sequence and deduce the cleavage sites. In principle it is also possible to deduce quantitative information from the peptide concentrations observed. However, this information is not available for all data and hence only the qualitative measure cleavage/non-cleavage was used.

Cleavage site information from the three proteins was used to create PSSMs (referred to as PCMs). In principle the flanking sequences of cleavage sites are extracted and aligned in order to create a PCM. The general way to construct a PSSM was described in Sect 2.6.1 and here the score $s_{i,j}$ of amino acid (i) in position j of the alignment is defined as:

$$s_{i,j} = ln\frac{(n_{i,j} + p_i)/(N+1)}{p_i} \approx ln(f_{i,j}/p_{i,j}) \tag{4.1}$$

where $f_{i,j}$ is the frequency of amino acids i at position j and $p_{i,j}$ is the prior probability of amino acid i in position j [113].

Different window sizes were tested for the extraction of peptides surrounding the cleavage sites. It is expected that the closest amino acids, especially P1 and P1', will influence the cleavage the most. The final version of the PCM method uses four N-terminal and two C-terminal amino acids around each cleavage site. The priors used are based on the amino acid composition of the source proteins. The score for a new sequence is calculated as the sum of individual position-specific scores for the amino acid in the sequence.

Since the data concerning proteasomal cleavage is very sparse, several different PCMs were generated for comparison reasons. Some of the previously presented methods used parts of

their training data to estimate prediction accuracy, making it hard to really estimate the true prediction performance. Here data from each protein was left out from the training procedure and then used to estimate the prediction accuracy. Matrices based on all three proteins as well as combination of two sets were used for performance evaluation. A comparison of the PCM based on the enolase and casein proteins, PCM(E+C), can be used to compare the performance of all methods for the prion protein. The cutoffs for distinguishing between cleavage and non-cleavage sites were chosen at maximum MCC. All three proteins were submitted to the prediction servers MAPPP, PAProC, and NetChop for benchmarking. The experimentally verified binding sites were compared to the prediction scores and performance measures of the predictions were calculated. The MAPPP prediction server offers only one type of proteasomal cleavage prediction, but both PAProC and NetChop provide several options for prediction. Three different models from the PAProC server were used for prediction: N1-N3. The N1 model is based on cleavages in enolase, the N2 model is based on cleavages in enolase and ovalbumin, and the N3 model is based on cleavages of enolase and a different set of ovalbumin cleavages. Two different types of networks were used from the NetChop server, 20S and C-term 2.0. The 20S network was trained on *in vitro* degradation of the enolase and casein proteins, whereas the C-term 2.0 network was trained on MHC ligands. For MAPPP and NetChop a cutoff of 0.5 was used. This is the default cut-off used by the MAPPP server and a recent study by the developers of NetChop used 0.5 to discriminate between cleavage and non-cleavage sites [239].

4.1.2. Results and discussion

The results of the PCM method can be seen in Table 4.1, showing results for PCMs based on different types of training data. As a measure of prediction performance, MCC, SP, and SE were used. The average total accuracy for the three proteins of the PCM method, when no training data was used for evaluation, reaches 65%. While the overall MCCs are not all that impressive (ranging from 0.18 to 0.32), our method is fairly robust.

The robustness of the PCM method was also compared to that of other methods for proteasomal cleavage site prediction. The methods in the comparative study were PAProC N1, PAProC N2, PAProC N3 [154, 197], Netchop 20S, Netchop C-term 2.0 [141], and MAPPP [118]. In each case, all proteins that were not included in the respective method's training set were used for assessing their prediction performance. It turns out, that all methods are considerably less robust than the PCM method. While PAProC and Netchop 20S achieve very high

Table 4.1.: Prediction performance results of proteasomal cleavage site prediction the PCM method. Several different PCMs were constructed in order to compare methods.

	Enolase (E)			**Casein (C)**			**Prion (P)**		
Method	MCC	SP	SE	MCC	SP	SE	MCC	SP	SE
PCM(ALL)	0.54	0.74	0.57	0.51	0.58	0.67	0.40	0.5 8	0.77
PCM(E+C)	0.59	0.64	0.80	0.50	0.69	0.51	0.18	0.51	0.44
PCM(E+P)	0.26	0.36	0.84	0.32	0.46	0.49	0.26	0.50	0.75
PCM(C+P)	0.19	0.35	0.69	0.51	0.67	0.53	0.46	0.62	0.78

Table 4.2.: Results from the proteasomal cleavage site prediction using already existing methods (values in brackets) are prediction performances for data used in method development and should not be used to compare methods. A large difference in performance can be seen for data contained in the training set versus data not in the training set for the PAProC and NetChop methods.

	Enolase (E)			**Casein (C)**			**Prion (P)**		
Method	MCC	SP	SE	MCC	SP	SE	MCC	SP	SE
MAPPP	0.09	0.30	0.75	0.12	0.24	0.77	0.03	0.40	0.56
PAProC N1	(0.95)	(0.98)	(0.95)	0.17	0.29	0.53	-0.03	0.36	0.33
PAProC N2	(0.95)	(0.97)	(0.97)	0.27	0.35	0.63	0.07	0.44	0.34
PAProC N3	(0.94)	(0.96)	(0.96)	0.15	0.27	0.56	0.08	0.44	0.39
Netchop 20S	(0.88)	(0.85)	(0.99)	(0.76)	(0.71)	(0.93)	0.12	0.47	0.41
Netchop C	0.18	0.37	0.49	0.18	0.29	0.58	0.07	0.44	0.33

MCC values (0.88-0.95) on the proteins they were trained on (enolase for PAProC, enolase and casein for Netchop 20S), their performance drops to values between -0.03 and 0.07 when validated with other proteins. This is a clear indication of overfitting on the training data. MAPPP and Netchop C-term 2.0 are clearly more robust, but their prediction performance is well below the performance of our method (see Table 4.2).

The average total accuracies of all external methods, when no training data was used for evaluation, was also calculated. The MAPPP method has an average accuracy of 47%, PAProC 60%, and Netchop 61% (PCM 65%). We therefore conclude that PCM combines comparable or slightly better prediction accuracy with improved robustness. Our PCM method also allows the easy extraction of proteasomal cleavage motifs based on amino acid preferences in a specific position. The three proteolytic sites of the proteasome have been described as having trypsin-like, chymotrypsin-like, and peptidylglutamylpeptide hydrolytic (PGPH) activity [285]. Figure 4.2 shows the preferences for specific amino acids at positions surrounding the cleavage site. This figure has been prepared from the values of the PCM derived from all three proteins and thus reflects the current knowledge on the cleavage preference of the

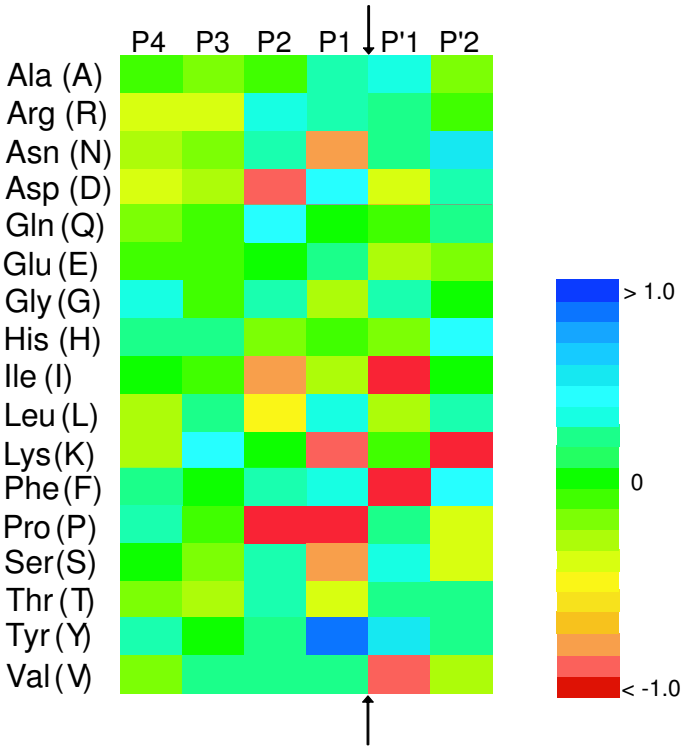

Figure 4.2.: The effect of specific amino acids on proteasomal cleavage. High values (blue) contribute to proteasomal cleavage, whereas red inhibit cleavage. The cleavage occurs between the P1 and P'1 positions (Met, Cys, and Trp have been omitted due to insufficient statistical basis).

proteasome. The real values for each amino acid can be seen in Table 4.3. Tryptic activity would imply cleavage immediately after Lys and Arg, however, we observe only Arg to be favorable, whereas Lys seems to have a negative effect on cleavage probability. Chymotryptic activity (cleavage after Phe, Tyr, and Trp) and PGPH activity (cleavage after Asp and Glu) is quite obvious from the very favorable values for these amino acids. In addition, we observe very unfavorable effects of Pro on the two positions immediately preceding the cleavage site and Val, Ile, and Phe immediately following the cleavage site. Due to the low abundance of Met, Cys, and Trp in the source proteins, the effects seen for these three amino acids is not included since they might be artifacts of the analysis.

4.2. Prediction of TAP affinity, SVMTAP

Peptides in the cytosol have to be transported into the ER for binding to MHC class I molecules. A key-player in this process are the TAP proteins in the ER membrane and understanding the mechanism of TAP transport can be used in modeling the whole antigen

Table 4.3.: The PCM matrix generated using data from the casein, enolase, and prion proteins. Cleavage occurs between the P1 and P'1 amino acids.

	Sequence position					
Amino acids	**P4**	**P3**	**P2**	**P1**	**P'1**	**P'2**
A	-0.07	-0.18	-0.01	0.22	0.30	-0.18
R	-0.31	-0.31	0.38	0.29	0.20	-0.02
N	-0.24	-0.12	0.24	-0.71	0.16	0.57
D	-0.34	-0.23	-0.81	0.41	-0.34	0.28
C	0.16	-5.49	0.16	-5.49	-5.49	-5.49
Q	-0.17	-0.08	0.43	0.07	-0.08	0.14
E	-0.03	-0.03	0.04	0.16	-0.27	-0.18
G	0.32	-0.09	0.22	-0.29	0.22	0.08
H	0.12	0.12	-0.16	-0.01	-0.16	0.44
I	0.05	-0.04	-0.73	-0.27	-1.64	0.05
L	-0.20	0.13	-0.42	0.38	-0.27	0.22
K	-0.23	0.42	0.04	-0.85	-0.02	-1.40
M	0.22	0.22	0.22	0.37	0.63	-0.47
F	0.13	0.01	0.23	0.33	-0.96	0.41
P	0.22	-0.02	-1.29	-1.29	0.18	-0.34
S	0.04	-0.16	0.20	-0.71	0.34	-0.40
T	-0.12	-0.23	0.28	-0.48	0.14	0.14
W	0.31	0.31	-1.29	1.00	0.31	0.64
Y	0.27	0.05	0.16	0.80	0.53	0.16
V	-0.14	0.18	0.13	0.18	-0.96	-0.20

processing pathway. The following sections will describe the data and methods used to develop the SVMTAP prediction method for TAP affinity. A comparative study against other methods will also be carried out. The prediction accuracy when only parts of the peptides are used is also investigated. Furthermore an analysis of TAP affinity of naturally processed MHC binding peptides is carried out.

4.2.1. Materials and Methods

Quantitative binding data

Very little data is available for actual TAP transport and available methods focus on prediction of TAP affinity, where quantitative data is available. The data used to develop the SVMTAP method consists of 446 9mer peptides with experimentally verified TAP affinity[1]. The binding affinities have been obtained from Sf9 insect cells overexpressing human TAP proteins in a competitive binding experiment using radiolabeled peptides, for more details see [291]. Sparse binary representation of the data was used and the binding affinity was represented as $lnIC_{50}$. Previous studies have pointed out that the three N-terminal and the C-terminal residues are the most important for peptide binding to TAP [284]. To investigate this we used only these four sequence positions in addition to the whole sequence.

Naturally processed peptides

A set of naturally processed MHC-binding peptides were extracted from the SYFPEITHI database in order to study the differences in TAP affinity between alleles. Several reports have described TAP-dependent and TAP-independent alleles, both experimentally [66, 157, 290] and computationally [39]. Using this type of data for analysis of TAP affinity should reveal differences between TAP-sufficient and TAP-insufficient alleles.

SVR training and evaluation

SVR training and evaluation was done using the SVM^{light} software. Several kernels were tested, but the linear kernel gave the best results. Leave-one-out cross-validation was used in order to evaluate the prediction performance of the method. The performance measure used was the Pearson correlation between predicted and experimentally verified binding affinities. The SVMTAP method was also compared to that of the SMM method from Peters *et al.* [204].

[1]The data was supplied by Peter van Endert, INSERM 580, Institut Necker, Paris, France

4.2.2. SVMTAP results and interpretation

Quantitative data and benchmarking

A plot of predicted binding affinities versus the experimentally verified binding affinities can be seen in Fig. 4.3. Here all positions of the peptide sequences have been used for training and the correlation coefficient between experimental and predicted values are 0.82. There are some outliers and the detection limit for the experimental procedure can also be seen.

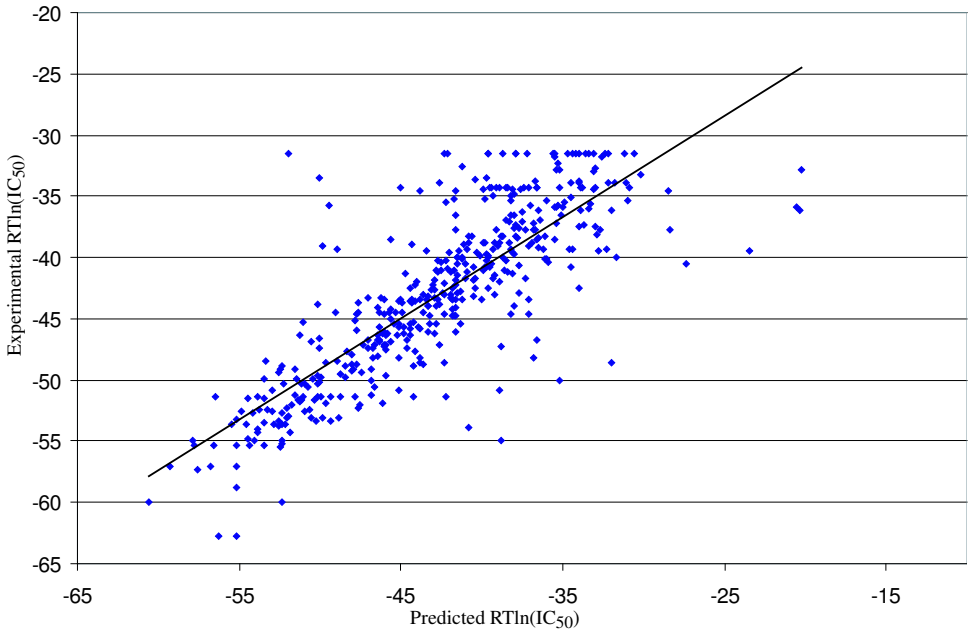

Figure 4.3.: Predicted binding affinities plotted against experimentally verified affinities for the SVMTAP method. The correlation of predicted and experimental values is 0.82.

The corresponding plot using only the three N-terminal and the C-terminal residues can be seen in Fig. 4.4. From these results one can conclude that much information regarding the actual affinity is found in the terminal residues, giving a correlation of 0.75 between predicted and experimental values.

Figure 4.5 shows the two corresponding plots when the Peters_matrix is used for prediction.

The SVMTAP method outperforms the Peters_matrix method is both cases. When only the terminals of the peptides are used for prediction, the difference is pretty huge, and SVMTAP_1239 performs almost as good as the Peters_matrix using the whole sequence.

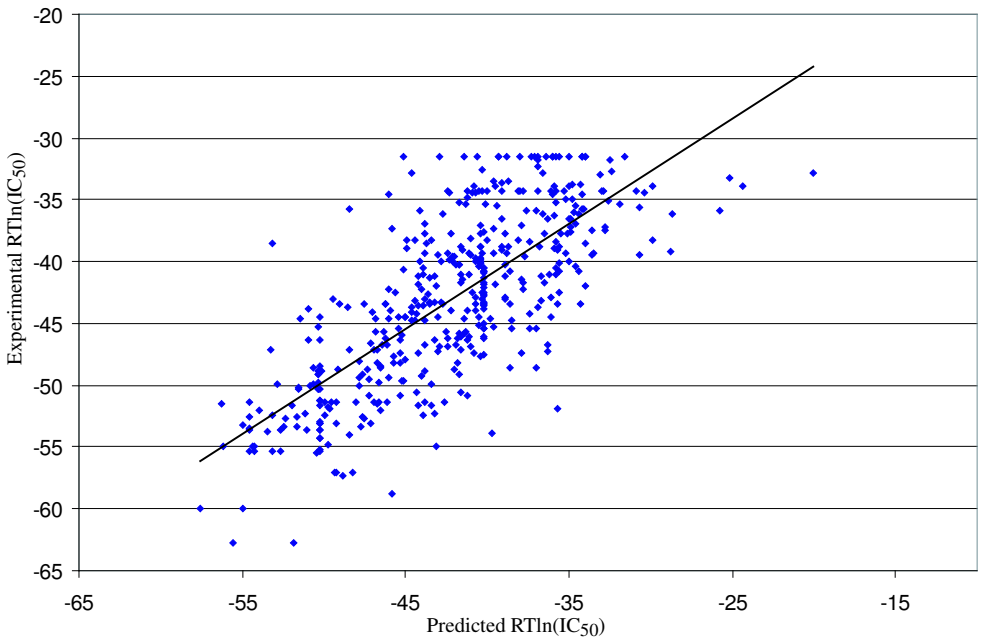

Figure 4.4.: Predicted binding affinities plotted against experimentally verified affinities for the SVMTAP method using the three N-terminal and the C-terminal positions. The correlation of predicted and experimental values is 0.75.

Analysis of SYFPEITHI peptides

Naturally processed and presented MHC-peptides were extracted from the SYFPEITHI database and their TAP affinities were predicted with SVMTAP. The results from this study can be seen in Fig. 4.6. A clear distinction can be observed in this plot where one group of alleles shows a trend of higher TAP affinity than the others. This once again verifies the theory that the alleles can be split into TAP-efficient and TAP-inefficient in terms of TAP transport.

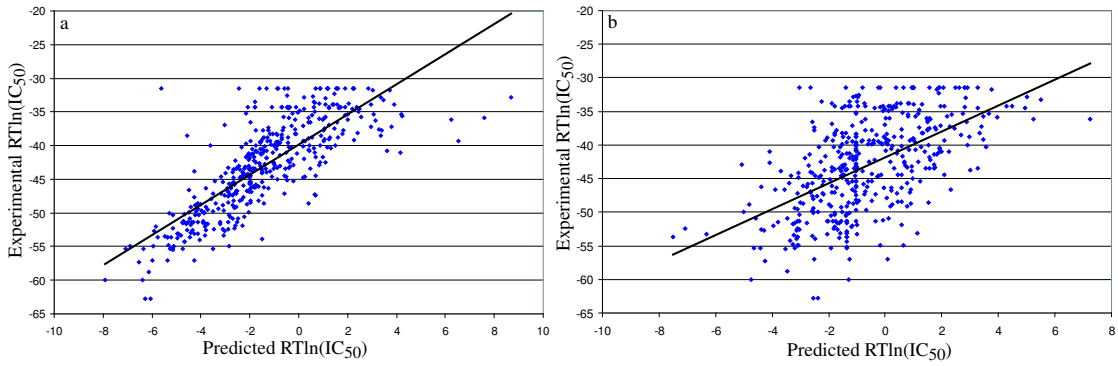

Figure 4.5.: Predicted affinities plotted against experimentally verified affinities for the Peters_matrix method. a. The correlation for scores based on the whole peptides sequence has a correlation of 0.79. b. Using only the three N-terminal and C-terminal residues gives a correlation of 0.56.

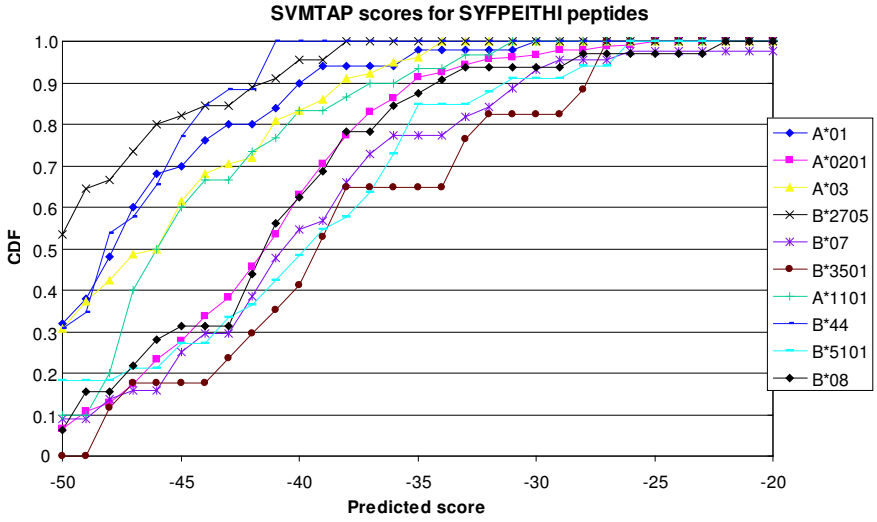

Figure 4.6.: The predicted SVMTAP scores represented as allele-specific cumulative distribution (CDF) curves. Going from low to higher scores, the CDF represents the fraction of all peptides from a given allele that have score equal or lower to a certain value (good binder have large negative values).

4.3. An integrated model of the processing events (WAPP)

The PCM and SVMTAP methods were combined with SVMHC in order model the whole MHC class I antigen processing pathway (WAPP) [75]. The hypothesis motivating this model is that a large fraction of all MHC-binding peptides originates from proteins that are cleaved by the proteasome into smaller peptides, which are then transported into the ER by the TAP proteins. With this in consideration, candidate MHC binding peptides can be filtered for peptides not generated by the proteasome or not transported by TAP. This approach was evaluate using peptides from the HLA-A*0201, HLA-A*01, HLA-A*03, and HLA-B*2705 alleles. The following section describes how the separate prediction methods were integrated and evaluated for prediction performance.

4.3.1. Materials and methods

Data sets

MHC binding peptides from the HLA-A*0201, HLA-A*01, HLA-A*03, and HLA-B*2705 alleles were extracted from the SYFPEITHI database. Peptides with a length of nine amino acids were used for further analysis. The SYFPEITHI database also contain information about the source protein if the peptides and these were also extracted. Peptides were no source protein could be found were discarded from the analysis. Both TAP and MHC prediction considers peptides with a length of nine amino acids, but C-terminal extended peptides were used for proteasomal cleavage prediction. This means that datasets of C-terminal extended peptides were extracted for each allele. The numbers of sequences used were HLA-A*0201 (96), HLA-A*01 (36), HLA-A*03 (47), and HLA-B*2705 (71). A set of extended non-binders was generated as described in Sect 3.1.1.

Combination of prediction methods

All peptides were predicted with the PCM, SVMTAP, and SVMHC methods. In the cases of SVMTAP and SVMHC the peptides with a length of nine amino acids were used, whereas the C-terminal cleavage score for each peptide was predicted using the C-terminal extended peptides. This gives a prediction score for proteasomal cleavage of the C-terminal of a MHC-binding peptide considering the flanking region from its source proteins. The scores obtained from the SVMTAP and PCM methods were then used to filter out peptides that are not correctly cleaved by the proteasome or transported by TAP. Peptides with scores lower than

Table 4.4.: Prediction accuracies for the four alleles using different approaches. The performance using WAPP is better compared to all other approaches except for the HLA-A*01 allele. The increase in specificity of WAPP compared to SVMHC is significant in most cases.

	SVMHC			**WAPP**			**PC+MHC**	**TAP+MHC**
Allele	MCC	SP	SE	MCC	SP	SE	MCC	MCC
HLA-A*0201	0.68	0.78	0.78	0.74	0.86	0.78	0.71	0.71
HLA-B*2705	0.85	0.76	1.00	0.88	0.82	1.00	0.86	0.86
HLA-A*01	0.92	0.94	0.96	0.93	0.95	0.98	0.93	0.93
HLA-A*03	0.80	0.84	0.90	0.82	0.92	0.89	0.81	0.81

a certain threshold were removed from the candidate list (as describe for SVMTAP affinity), but these were chosen conservatively in order to not remove known binders. The final cutoffs chosen for HLA-A*0201 were -4.8 for proteasomal cleavage and -27 for TAP affinity. The corresponding values for HLA-B*2705, HLA-A*01, and HLA-A*03 were -2.0 and -35.

4.3.2. Results and interpretation

We found a significantly improved performance of WAPP over SVMHC alone (MCC increases from 0.68 to 0.74 for HLA-A*0201, from 0.85 to 0.88 for HLA-B*2705, and from 0.80 to 0.82 for HLA-A*03), see Table 4.4. This improvement is mostly due to a smaller number of false positives, i.e. peptides that could bind to MHC, but are either not cleaved by the proteasome or not transported by TAP. The improvement for HLA-A*01 is somewhat less, however the overall prediction accuracy for this allele is very high.

The best performance is achieved when both proteasomal cleavage and TAP filtering is used. The largest increase in prediction performance is achieved for the HLA-A*0201 allele where the MCC increase from 0.68 to 0.74. Using either proteasomal cleavage or TAP as filter shows worse results for the HLA-A*0201, HLA-B*2705, and HLA-A*03 alleles. The main feature of the combined approach is a reduction in false positives in the prediction, i.e. removal of peptides that actually could bind to MHC, but are unlikely to be generated by the proteasome or transported by TAP. The specificity increases from 0.78 to 0.86 for the HLA-A*0201 allele and from 0.76 to 0.82 for HLA-B*2705. The only allele that shows slightly different results is HLA-A*01. The peptides binding to this alleles almost exclusively have a Tyr in position nine. This means that a high specificity can be obtained by MHC prediction alone, however it should be pointed out that the prediction accuracy is not negatively influenced by taking proteasomal cleavage and TAP transport into account.

A plot underlining the use of SVMTAP as a filter can be seen in Fig. 4.7, showing the CDF for HLA-A*0201, HLA-B*2705 and non-binding peptides. A difference can clearly be seen between the three classes, where the known HLA binders show a higher affinity for TAP than the non-binders. Both experimental and computational studies have previously shown that HLA-B*2705 peptides have a high TAP affinity, whereas HLA-A*0201 have relatively low TAP affinity [39, 290]. This difference is to be expected, as HLA-A*0201 is a TAP-inefficient allele, whereas HLA-B*2705 is TAP-efficient.

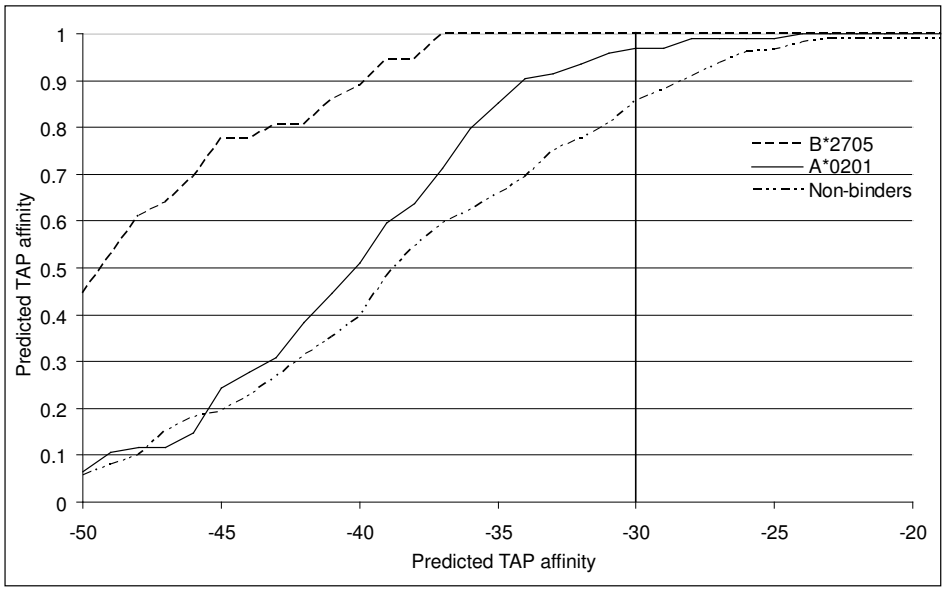

Figure 4.7.: Predicted TAP affinity for the HLA-A*0201 and HLA-B*2705 data sets, represented as a cumulative distribution functions (CDFs) going from high to low affinity binders. The value of the CDF corresponds to the fraction of data with values below a given TAP affinity. A clear difference in the distribution of TAP affinity can be seen between known epitopes and non-epitopes. Only a small fraction of the known epitopes has a TAP affinity higher than -30 (corresponding to an IC_{50} of 5,000 nM).

4.4. Comparison of WAPP and competing methods

The prediction performance of WAPP was also compared to the NetCTL [156] and the method presented by Tenzer *et al.* [275]. A comparison of this kind is not trivial, since the data is very sparse and has often been used to develop the individual methods. However, a set of peptides binding to three different alleles recently published in the SYFPEITHI database was used for this purpose. It is possible that the two external methods have used some of the peptides for method development, but no part of WAPP is based on any of this data. Ten

peptides from each of the three MHC alleles HLA-A*0201, HLA-A*01, and HLA-A*03 was used for comparison. The source proteins of all peptides respectively were submitted to the three prediction methods and the rank of the known MHC-binding peptide was calculated. The results are presented in Table 4.5.

From this comparative study a few things can be seen. In general the methods are all performing pretty well in finding the known epitope sequences within the source proteins. The WAPP method gives the best rank for most of the peptides, the NetCTL method is the second best, and the Tenzer method performs worse. It should however once again be pointed out that the benchmark dataset is small and a dataset of several hundreds of peptides from different alleles would be preferred for a better statistical evaluation.

4.5. Proteasomal splicing - SpliPep

Proteasomal splicing was described in Sect. 2.3.1. A good starting point for finding possibly spliced candidates are in databases of identified MHC-binding peptides and literature. Some peptides reported have not been mapped to any protein sequence and one reason for this might be proteasomal splicing. The tool SpliPep was developed for identifying peptides generated from two non-contiguous parts of a source protein. Given the sequence of a peptide, SpliPep can search databases for potential source proteins.

4.5.1. Implementation

The problem of finding a spliced peptide within a protein sequence can is in principle a easy string matching problem. The peptide sequence S of length l can be split into S_1 and S_2. If both S_1 and S_2 can be found as non-overlapping parts of the whole protein sequence, generation of the peptide by proteasomal splicing is theoretically possible. Simple regular expressions were used for the matching of sequences. In principle a search for one of the subsequence is done first and only if a hit is found, a search for the second sequence is also done. If the length of a peptide S is L, $S_1 = S(1, l)$ and $S_2 = S(l, L - l) \forall l = 1, \ldots, L$. The most effective way to conduct the search is to start with the longest sequence. Single amino acids and dipeptides frequently occur within a protein sequence, whereas longer peptides are rather unique. Hence, the search always start with the longest sub-sequence and only if a hit is found a search using the second sub-sequence is done.

Several databases can be used for searching and specifically searching for only the viral

Table 4.5.: Comparison of WAPP, NetCTL and the method by Tenzer *et al.* For each peptide the best rank is shown in bold allowing a qualitative comparison of the different methods.

Peptide	Protein SwissProt ID	Allele	WAPP	NetCTL	Tenzer
VALEFALHL	Q8TDN4	A*0201	31	39	**8**
ALLDKLYAL	Q9NV31	A*0201	3	**1**	**1**
TLSDLRVYL	Q9BYN0	A*0201	**2**	3	6
FVHDLVLYL	P53675	A*0201	27	**11**	13
RLASYLDRV	P05783	A*0201	**1**	2	2
ALATLIHQV	Q9UBW8	A*0201	**1**	**1**	**1**
VLAEVPTQL	Q99829	A*0201	**2**	5	4
LLDRFLATV	Q140943	A*0201	**1**	3	3
VLFGLLREV	Q92620	A*0201	**1**	**1**	19
RLASYLDKV	P35527	A*0201	**1**	4	3
YTSDYFISY	P14921	A*01	**1**	2	14
AIDQLHLEY	O43707	A*01	**1**	**1**	**5**
HLDLGILYY	Q9H3H5	A*01	**1**	**1**	**1**
TSPSQSLFY	Q8IV72	A*01	**1**	**1**	3
GTDELRLLY	Q9Y4W2	A*01	**1**	**1**	2
ELEDSTLRY	Q6S383	A*01	**1**	6	5
DTDHYFLRY	Q969N2	A*01	**1**	**1**	3
VTEIDQDKY	P21333	A*01	**1**	**1**	**8**
FIDASRLVY	P35221	A*01	**1**	**1**	**1**
YTAVVPLVY	P01591	A*01	**1**	**1**	2
KLFDKLLEY	Q9BZZ5	A*03	**1**	4	2
TSALPIIQK	Q99541	A*03	**1**	3	5
KLYEMILKR	P56559	A*03	**1**	**1**	2
SLFSRLFGK	P18085	A*03	**1**	3	4
RLLEMILNK	P52895	A*03	**1**	3	8
KLADFGLAR	P11802	A*03	**1**	**1**	4
KVYENYPTY	P35659	A*03	**1**	6	**1**
TLADLLALR	Q96DT5	A*03	**26**	130	270
RVHAYIISY	Q9NZN4	A*03	**1**	2	4
KLFIGGLSF	P09651	A*03	**3**	10	9

subset of protein from Swiss-Prot is of course more effective than a whole database search.

4.5.2. Results and interpretation

To prove the usefulness of SpliPep, the peptides found by Vigneron *et. al.* (RTKQLYPEW) and Hanada *et. al.* (NTYASPRFK) were used. The results from this search can be seen in Fig. 4.8.

Query peptide: **RTKQLYPEW**

Protein:	PM17_HUMAN
Description:	Melanocyte protein Pmel 17 precursor (Melanocyte lineage-specificantigen GP100) (Melanoma-associated ME20 antigen) (ME20M/ME20S)(ME20-M/ME20-S) (95 kDa melanocyte-specific secreted glycoprotein).
Fragment 1:	RTK(40-42)
Fragment 2:	QLYPEW(47-52)
Distance fragments:	4 amino acids
Protein:	NP_008859.1
Description:	silver homolog; Melanocyte protein mel 17; Pmel 17; Pmel17; Silver,mouse, homolog of; silver (mouse homolog)-like; silver (mousehomolog) like [Homo sapiens].
Fragment 1:	RTK(40-42)
Fragment 2:	QLYPEW(47-52)
Distance fragments:	4 amino acids

Query peptide: **NTYASPRFK**

Protein:	FGF5_HUMAN
Description:	Fibroblast growth factor-5 precursor (FGF-5) (HBGF-5) (Smag-82).
Fragment 1:	NTYAS(172-176)
Fragment 2:	PRFK(217-220)
Distance fragments:	40 amino acids
Protein:	NP_004455.1
Description:	fibroblast growth factor 5 isoform 1 precursor; fibroblast growthfactor 5S [Homo sapiens].
Fragment 1:	NTYAS(172-176)
Fragment 2:	PRFK(217-220)
Distance fragments:	40 amino acids

Figure 4.8.: Search results of SpliPep for the peptides RTKQLYPEW and NTYASPRF. The source proteins of the respective peptides are found in both NCBI RefSeq and Swiss-Prot databases.

All human databases were used for the search and there is one hit from Swiss-Prot and one hit from NCBI RefSeq [213, 214] for each peptide. Both hits correspond to the proteins reported in the original publications. RTKQLYPEW has an interspersed fragment of four amino acids, whereas NTYASPRFK has a longer interspersed fragment of 40 amino acids.

A SpliPep search was also done for a HLA-A*03 epitope, SQNFPGSQK, identified in melanoma patients [115]. No source protein has previously been identified for this peptide. SpliPep finds three potential source proteins of the SQN-peptide. Two of these are PR-domain zinc finger proteins, PRD7 and PRD9, involved in transcription. These proteins are very similar and the length of the interspersed fragment is 11 amino acids in both cases, see Fig. 4.9.

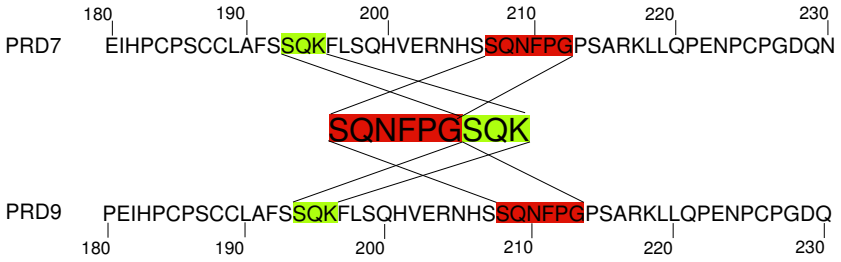

Figure 4.9.: Search results of the previously unmapped HLA-A3 peptide SQNFPGSQK identified from a melanoma patient. The peptide is successfully mapped to two distinct proteins PR-domain containing proteins. In both cases the length of the interspersed fragment is 11 amino acids.

A third potential source protein of the SQN-peptide was also found. This is the Pmel17 protein which also is the source of the Vigneron *et al.* peptide [296]. Here the SQN-peptide can be generated from position 89-95 (NFPGSQK) and positions 113-114 (SQ), see Fig. 4.10.

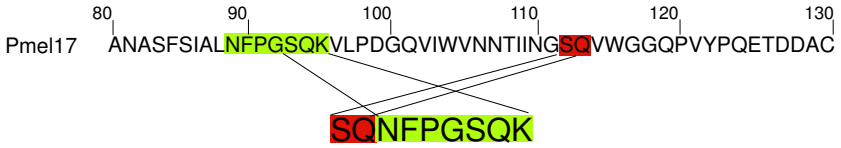

Figure 4.10.: A more detailed view of the Pmel17 protein. An interspersed fragment of 17 amino acids has to be spliced away in order to generate the SQN-peptide.

Since proteasomal splicing has been verified for other peptides from the Pmel17 protein, one has to consider this a more likely source than the PR-domain proteins. In both cases the length of the interspersed fragment is interesting. Protein splicing is known to occur in lower organisms, where inteins are autocatalyticly removed from the sequence and the flanking regions are ligated [202]. The inteins have a rather specific domain structure and the shortest one found so far has a length of 134 amino acids [203]. One can only speculate about the influence of the process on antigen processing as a whole. Currently there are only two cases known. Our collaborators are currently investigating the SQN-peptide experimentally.

4.6. Discussion

This chapter presented an integrated prediction method, WAPP, for the major events in the processing pathway of MHC class I antigens. WAPP mimics the series of biological events by predicting peptides with a proteasomal cleavage site at the C-terminus, moderate to high affinity to TAP, and an affinity to MHC. The three steps modeled are generally thought to be the major determinants in class I antigen processing, although several alternative processing events have been described in literature.

Luckey *et al.* showed that for some MHC alleles, a significant amount of peptides were generated even in the presences of proteasome inhibitors [163]. These results clearly indicate an important effect of other cytosolic proteases [21, 78]. TPPII is one such protease that has important effects in the trimming of proteasomal degradation products [222]. A further example points out the importance of TPPII in the generation of a known HIV-Nef(73-82) epitope [250]. Some alternative ways of peptide-transport into the ER have also been suggested. Lautscham *et al.* described TAP-independent transport of hydrophobic peptides [157] and suggested that these might enter the ER by passive diffusion or by an unknown transport protein within the ER membrane. Furthermore, they pointed out that many known MHC-binding peptides are derived from protein signal-sequences and suggested Sec61 as a potential transporter.

A recent study showed that peptides for some MHC alleles have a low TAP affinity [207]. We also observe this for the HLA-A*0201 and HLA-B*2705 alleles, described as TAP-inefficient and TAP-efficient, respectively. It is likely that some of the TAP-inefficient alleles utilize the routes described by Lautcham *et al.*, but it is still possible to combine TAP and MHC prediction to reduce the number of false positives. The overall increase in performance obtained by adding TAP affinity prediction and proteasomal cleavage site prediction to MHC binding prediction is significant, although these steps are clearly less specific than MHC binding itself. Thus, improved overall performance for a combined model can only be achieved through high-quality models for proteasomal cleavage and TAP affinity. Previous attempts to combine the different steps yielded only a small increase in performance combining TAP prediction with MHC binding predictions and even a decrease in performance if proteasomal cleavage was predicted together with MHC binding [204]. At least for the case of proteasomal cleavage, we argue that this might be largely due to an overfitting of the cleavage models, as insufficient data was available. The existing methods for proteasomal

cleavage prediction, NetChop and PAProC, can reproduce their training data with high accuracy, while their performance on external validation data is much lower. This implies an overfitting of the model, which typically results in lower generality of the models. Our PCM method presented has thus been carefully designed to be more robust at the cost of slightly reduced accuracy on the training set. The robustness, however, turns out to be key to a successful combination with the other prediction steps. Future challenges in the prediction of proteasomal cleavage are likely to include splicing events [104, 296]. Splicing within the proteasome can generate a peptide from two non-contiguous part of its source protein. The mechanisms underlying proteasomal splicing is not fully understood and currently there is not enough data available to model this in the predictions. Prediction of TAP transport by SVMTAP shows an increase in performance compared to the Peters_matrix method. It is also likely that some of the peptides transported into the ER have extended N-terminals that can be trimmed by ER aminopeptidases [251]. Peters *et al.* used parts of the matrix for predicting peptides longer than nine amino acids. They explored a weighting of the N-terminal scores in order to improve prediction. For some alleles the weighting improved accuracy, whereas the effect was negative in other cases. It should also be pointed out that the study of the relationship of TAP affinity and TAP transport was done using a library of nine amino acid long peptides [101]. The problem of predicting TAP affinity for peptides longer than nine amino acids is still unsolved and more data is needed for a thorough investigation. In summary, WAPP shows improved prediction performance for the four MHC alleles using an integrated approach including the three major processing steps. Furthermore, WAPP shows better prediction accuracy compared to competing methods. Whole-pathway predictions will hopefully improve the rational design of epitope-driven vaccines in the future. WAPP increases the prediction specificity and hence reduces the number of peptides that has to be tested experimentally. Future improvements on the prediction will largely be data-driven, as the lack of data for TAP transport and proteasomal cleavage are currently the issues limiting the predictive power.

5. Integrative analysis of cancer-related data

The prediction methods presented in the two previous chapters can be used to identify candidates for peptide-based vaccines. However, the starting point of the methods is a target protein sequence. For development of a vaccine against viruses or bacteria, one would focus on proteins originating from the pathogen. For development of tumor vaccines on the other hand, the identification of candidate proteins is a challenging task. Here one would like to find TAAs or TSAs for a certain cancer type that subsequently can be used for vaccine development.

There are much cancer-related data available in different databases. Here "databases" refer to everything from spreadsheets to object-oriented database implementations. Considering different aspects of cancer, e.g. immunology and genetics, could lead to identification of proteins well-suited for vaccine development. However, joining data from different types of experiments and fields of research is an often underestimated problem.

This chapter presents an integrated approach, CAP, for the analysis of cancer-related data. CAP joins data from heterogeneous data sources and enables statistical evaluation of the data. In a sense this goes in the direction of "cancer systems biology", since CAP enables integrated analysis of different aspects of cancer research. One focus of CAP is the analysis of cancer related protein, identified with the SEREX method [283], causing an auto-immune response in cancer patients. A large-scale study was done to investigate if autoimmune responses in cancer are related to gene expression levels or genetic alterations.

The first sections of this chapter describe the data sources and data model used in CAP. This is followed by a description of the prediction methods and statistical analysis tools integrated in CAP. The last part of the chapter shows the results from the integrative analysis of the autoimmune-related data of CAP. Furthermore, some examples are given of how the prediction methods presented earlier in this thesis can be used to identify peptide vaccine candidates.

5.1. CAP content

The aim of CAP was to create a database that could be used to store and analyze data from several different disciplines of cancer research. Most of the data from our collaborators focused on tumor antigens causing an autoimmune response in cancer patients, detected with the serological analysis of recombinant cDNA expression libraries (SEREX) method [283]. Furthermore, data concerning mutations and mRNA expression levels were put into the database. In addition to the rather cancer-specific data, resources like Swiss-Prot [27], RefSeq [213, 214], and Locus Link [213] were used to obtain general information about the genes and proteins. The data collected from own experiments and external databases were functionally annotated using e.g. prediction of protein subcellular localization and protein function. The underlying data model of CAP is designed to be easily extendible for new data types. The data sources and prediction methods used in CAP are described in detail below. An overview of how the different data sources are combined is given in Fig. 5.1.

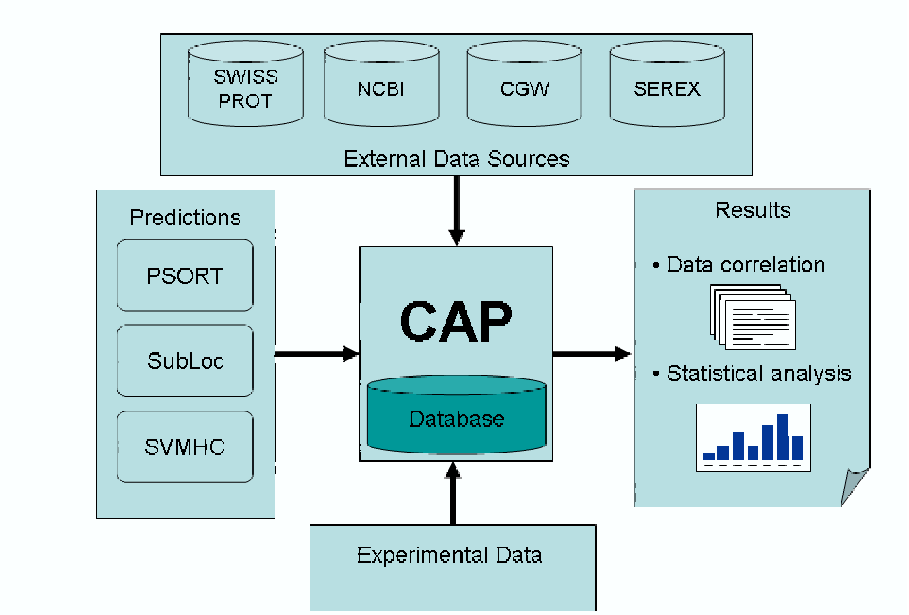

Figure 5.1.: An overview of the different data sources integrated in CAP. The data comes from external databases, own experimental data, and prediction methods. CAP enables analysis of the data in an integrative manner and offer on-the-fly statistics of large data sets.

5.1.1. Data sources

All data from publicly available data sources was updated in December 2003.

RefSeq

Two major resources of well-annotated genes and proteins are the Swiss-Prot and RefSeq databases. RefSeq aims to provide a comprehensive and non-redundant set of both nucleotide and protein sequences. It is a stable reference source for gene identification and characterization, polymorphisms, expression, and comparative analysis. Protein and nucleotide sequences are explicitly linked and there is an ongoing curation of the databases and reviewed entries are marked as such. The sequence data of RefSeq is well-validated, e.g. it is checked that the genomic sequence from a region really matches a given mRNA sequence. The database is also kept consistent with the LocusLink database records.

Swiss-Prot

Swiss-Prot is a highly curated database providing well-annotated protein sequences. Examples of Swiss-Prot information is function, subcellular localization, post-translational modifications, and domain structure. The sequences are annotated according to information obtained from original publications describing a sequence, review articles describing e.g. a whole family of proteins, and expert knowledge from external reviewers. Swiss-Prot also provides cross-references to about 60 other databases. Swiss-Prot is now part of the UniProt database [10].

SEREX data

Much of the cancer-related information in CAP come from SEREX experiments. This can be used to identify tumor antigens that elicit an autoimmune response in cancer patients. Typically, cDNA expression libraries are constructed from fresh tumor samples and cloned into *lambda* phage expression vectors. The recombinant proteins expressed during the lytic infection can be transferred to nitrocellulose membranes, which are subsequently incubated with patient serum. The clones reactive with the patient sera are then cloned into monoclonality allowing for characterization by DNA sequencing. Four main types of antigens have been identified. The first group contains known tumor antigen, e.g. MAGE-1 and tyrosinase. The second group contains autoantigens associated with well-studied autoimmune diseases. A third group of antigens are homologous to well-characterized genes, but have not been previously known to elicit an immune response and the fourth group contains unknown antigens with no homologous genes. SEREX-related information in CAP originates

from experiments by our collaborators [18, 31, 32, 54, 55, 70, 88, 108, 183, 269], as well as the SEREX database [280]. The SEREX database itself is currently in a 'read-only' mode, however, it is now a part of the Cancer Immunome Database (CIDB) [278].

Additional cancer-related data was obtained from the Cancer GenticsWeb (CGW) database containing information about abnormal and mutated genes. CGW provides a comprehensive list of genes associated with a specific cancer types. Information was also obtained from the Mitelman databases of chromosomal aberrations and from the NCI60 database. An overview of the most important data sources used by CAP is given in Table 5.1.

Table 5.1.: The major data sources used in CAP and their URLs.

Data source	**URL**
Cancer GeneticsWeb (CGW)	http://www.cancerindex.org/geneweb/
Cancer Immunome DB (CID)	http://www2.licr.org/CancerImmunomeDB/
LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/
Mitelman	http://cgap.nci.nih.gov/Chromosomes/Mitelman/
NCBI	http://www.ncbi.nlm.nih.gov/
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq/
SEREX	http://www.licr.org/SEREX.html
Swiss-Prot	http://www.ebi.ac.uk/swissprot/
NCI60	http://genome-www.stanford.edu/nci60/

5.1.2. Prediction methods

Several different prediction methods have been integrated into CAP in order to functionally annotate sequences. These involve protein subcellular localization, protein function, and T-cell epitopes.

Protein subcellular location

Knowing the subcellular localization of a protein is a good starting point for finding out its function. However, only a subset of all proteins has an annotated localization, hence prediction methods are needed for studies on a proteome scale. There are many different approaches for protein subcellular localization prediction, reviewed in [73]. The different methods can be split into a few different groups based on the underlying biological assumption on which they are based. The methods used in CAP are PSORT [186] and an SVM approach utilizing the overall amino acid composition of the proteins [116, 122]. PSORT was the first method allowing for prediction of more than ten different subcellular localizations. A new

prediction system, MultiLoc [117], has recently been developed in our group and is likely to soon be incorporated into CAP.

Protein function

Prediction of protein function is harder than predicting subcellular localization. For protein function prediction the ProtFun method was used [134, 135]. The annotation taken from these predictions involves functional category, enzyme class, and gene ontology category.

T-cell epitopes

Putative T-cell epitopes are predicted with the SVMHC method, described in Sect. 3.1. This enables fast identification of putative T-cell epitopes from protein sequences in CAP.

5.2. Data modeling

Modeling of biological data is not a trivial task, since it can be extremely heterogeneous. An important requirement of the data model of CAP is that it should be easy to extend and to integrate new types of data in the future. In order to obtain a unified view of the data, both differences in data format (syntactic differences) and differences in the meaning of data (semantic differences) have to be considered [48]. The core of our data model consists of sequences and annotations to these sequences. A generalization of the model can be seen in Fig. 5.2. The Unified Modeling Language (UML) [236] was used for data modeling, giving a well-defined data model that can be easily updated.

The data model is centered around protein and nucleic acid sequences, that are of either experimental or reference types. A reference sequence is usually obtained from the RefSeq or Swiss-Prot databases, whereas an experimental sequence might come from a SEREX experiment. An important aspect of annotations is the source from which they were obtained. A protein might for example have an annotated function and the source of that annotation might be Swiss-Prot or a prediction method. Knowing the source of an annotation is important since this gives a certain meaning in terms of reliability, need for updates etc. In general experimentally verified annotations are considered more reliable than predicted ones.

The different sequence types are connected in a way that also provides extra meaning. An experimental nucleic acid sequence can for example be mapped to a reference sequence by a BLAST search, which in turn is connected to a protein sequence. These connections make

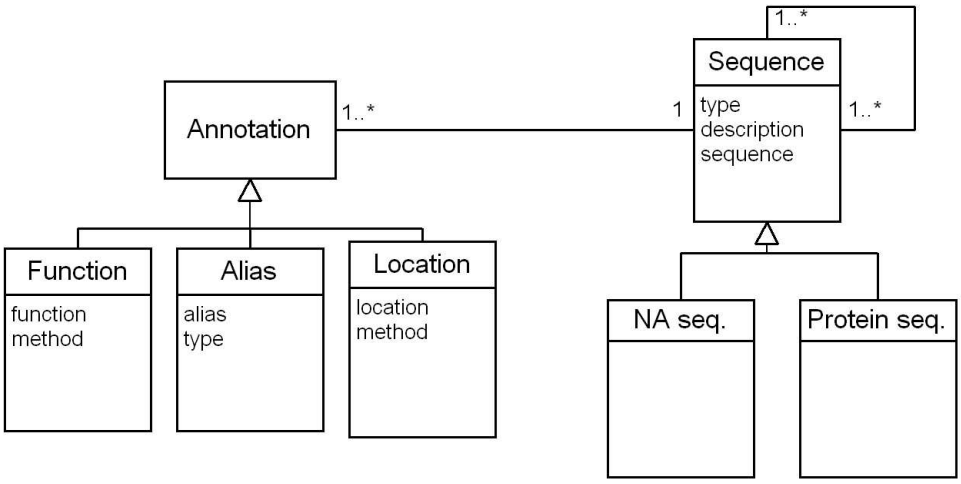

Figure 5.2.: A part of the CAP data model. The sequence class is a generalization of the NA and Protein classes. Sequences are associated with Annotations. The annotation examples here are Function, Alias, and Location. The model itself is very flexible, since new information can be incorporated by simply adding a new annotation subclass.

complex queries of the database possible, since several experimental sequences from different types of experiments might be connected to the same reference sequence. An example of how subcellular location of proteins related to a set of experimentally found NA sequences can be extracted can be seen in Fig. 5.3. It can also be seen how data in CAP is linked to external resources.

For privacy reasons the data model also enables the user to mark sequence and sequence-related information as private. This information is then only visible for the user, but he can still profit from all the analysis tools of CAP. The concept of limited data access is important if CAP will be extended to contain clinical (patient) data. However, this feature is also useful for researcher analyzing unpublished data.

5.3. Data analysis tools

One aim of CAP was to provide tools for data analysis and statistical evaluation. One important aspect is the many different search possibilities a user have when trying to access the stored data. There are numerous ways from free text search to search for specific cancer-related information implemented in CAP. An example of the search form for SEREX sequences can be seen in Fig. 5.4. The options for searching include sequence description, expression level, or cancer type. Several different search-fields can also be combined into

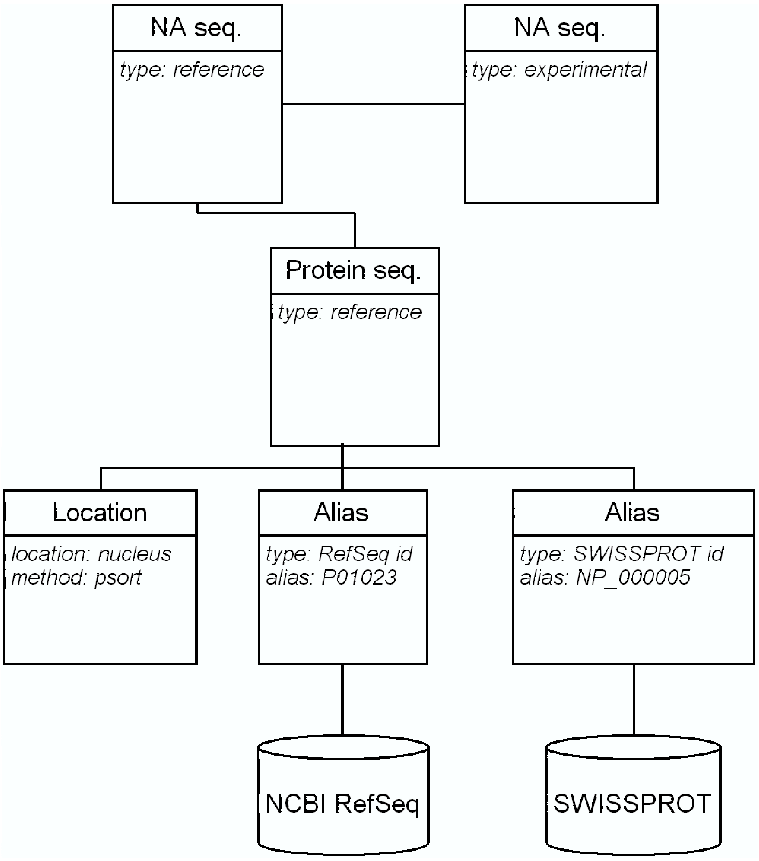

Figure 5.3.: Extracting information from related sequences and linking to external databases is a key feature of CAP. An experimental NA sequence is typically related to a reference NA sequence. The NA sequence itself is connected to a PROT reference sequence. In this example, the Protein sequence has three annotations, one Location and two Aliases. The protein location for a sequence related to a NA sequence can easily be extracted, since all that information is connected. By using databases identifiers as aliases, it is possible to create hyperlinks to external databases. Linking to external data sources enables for fast extended information retrieval.

more refined search.

Single sequences in CAP can be grouped into user-defined data sets. These data sets can then be used for statistical evaluation and analysis, such as chromosomal distribution of genes, protein function, and subcellular localization. To illustrate the use of CAP and data set of nucleic acid sequences obtained from SEREX experiments for renal cell carcinoma (RCC) was created. Using the tools for editing data sets, a data set of reference nucleic acid sequences and one of protein sequences were created. A summary view of a subset of the protein sequences can be seen in Fig. 5.5.

This view shows information about the data set in table format and the user can specify exactly what to display. Statistical analysis of the subcellular localization of the RCC proteins

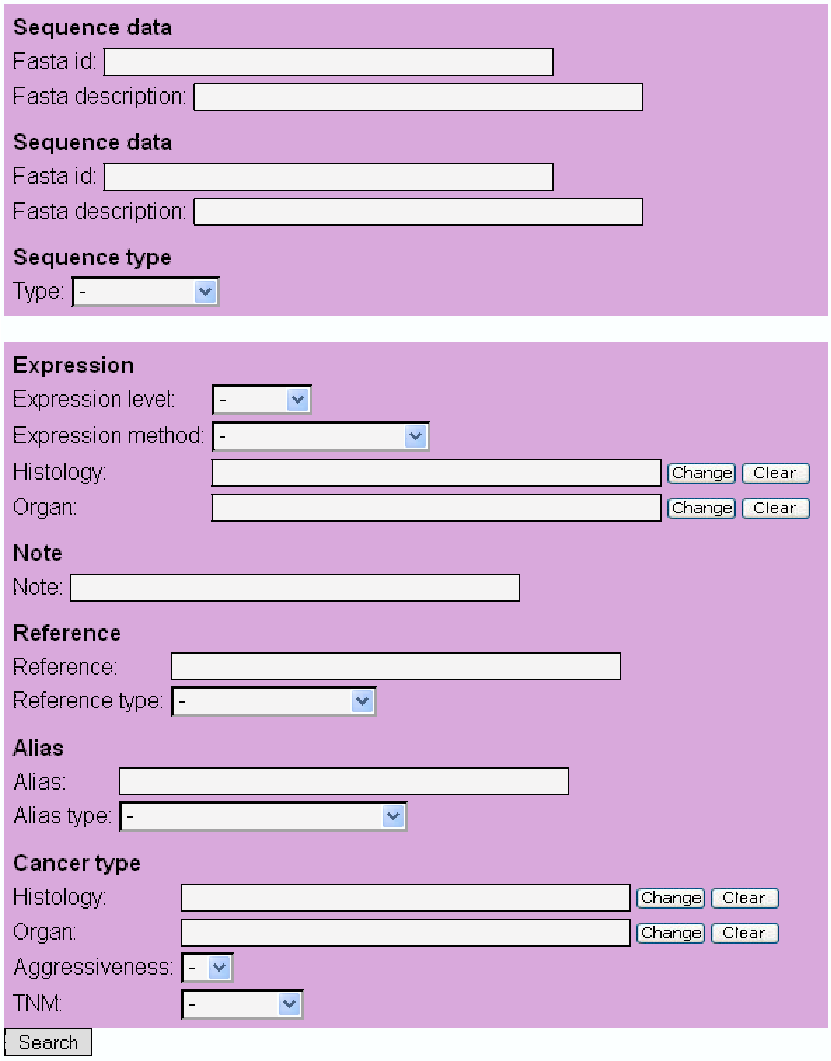

Figure 5.4.: The SEREX sequence search form of CAP.

sequence info			function		location	
fasta id	description	type	function	method	location	method
gi|8923204|ref|NP_060185.1|	zinc finger protein 3 [Homo sapiens]	refseq	Cell_envelope	ProtFun_Category	Cytoplasm	PSORT
gi|17465036|ref|XP_037759.2|	similar to KIAA0376 [Homo sapiens]	refseq	Cell_envelope	ProtFun_Category	Mitochondrial matrix space	PSORT
gi|27500858|ref|XP_114002.3|	similar to CG1676-PA [Drosophila melanogaster] [Homo sapiens]	refseq	Cell_envelope	ProtFun_Category	Peroxisome (microbody)	PSORT
gi|27465783|ref|XP_027116.6|	similar to hypothetical protein FLJ25555 [Homo sapiens]	refseq			Cytoplasm	PSORT
gi|4759006|ref|NP_004694.1|	rabaptin-5 [Homo sapiens]	refseq			Nucleus	PSORT
gi|7706216|ref|NP_056958.1|	H-2K binding factor-2 [Homo sapiens]	refseq	Cell_envelope	ProtFun_Category	Cytoplasm	PSORT
gi|7705837|ref|NP_057205.1|	NY-REN-45 antigen [Homo sapiens]	refseq			Nucleus	PSORT
gi|22550104|ref|NP_115971.2|	ubiquitin specific protease [Homo sapiens]	refseq	Cell_envelope	ProtFun_Category	Plasma membrane (cytoplasmic)	PSORT

Figure 5.5.: An overview of the RCC data in table format as generated by CAP. Here general sequence information in addition to function and location is shown.

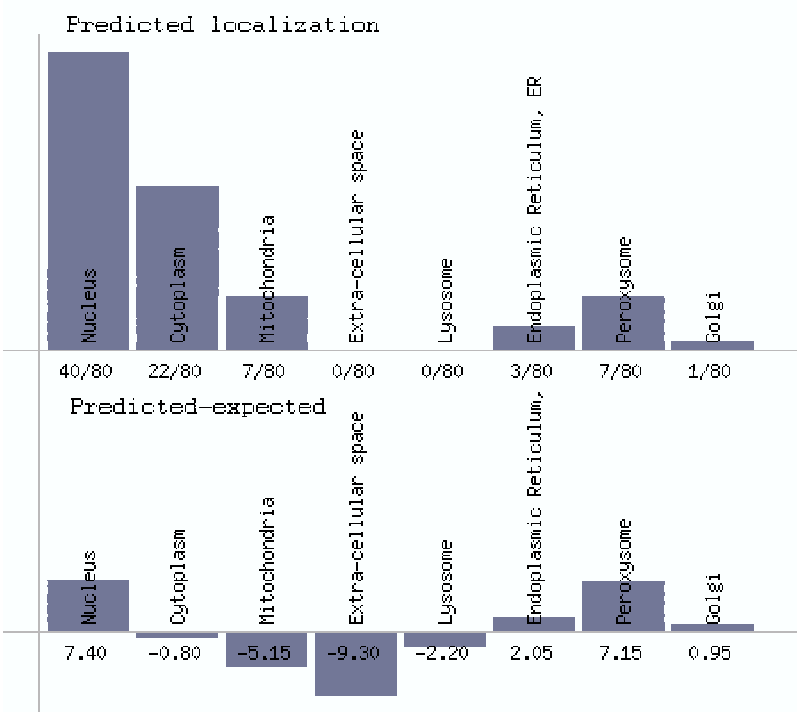

Figure 5.6.: Predicted subcellular localizations of the RCC proteins. The upper graph shows the number of sequences from the whole data set that was predicted to be located in a certain localization. The lower graph shows a similar plot, but here the expected values, given the localization of all proteins in the proteome, are subtracted from the predicted values.

can be seen in Fig. 5.6. The graph labeled *Distribution* is the plain count of sequences from the data set belonging to a certain subcellular location. The *Expected distribution* is an estimation of the whole-proteome distribution of locations as described in [185]. The third graph, *Distribution - Expected distribution*, shows the difference between the two previous graphs. As can be seen in the plain distribution a lot of the sequence are predicted to be located in the nucleus. The normalized graph, on the other hand, shows that this is not too surprising, since a rather large portion of all proteins are expected to be found in the nucleus.

A graph displaying the protein function of the proteins is shown in Fig. 5.7. According to the Prot_Fun category prediction, many of the sequence have a function in the cell envelope. The GeneOntology analysis, however, assigns many of the proteins into the classes growth factors, transcription, and transcription_regulation. Statistics can also be done for chromosomal distribution of the genes. In some cases it might be interesting to take a closer look at genes on a certain chromosome, for example if chromosomal breakpoints are known to occur in the cancer type of interest.

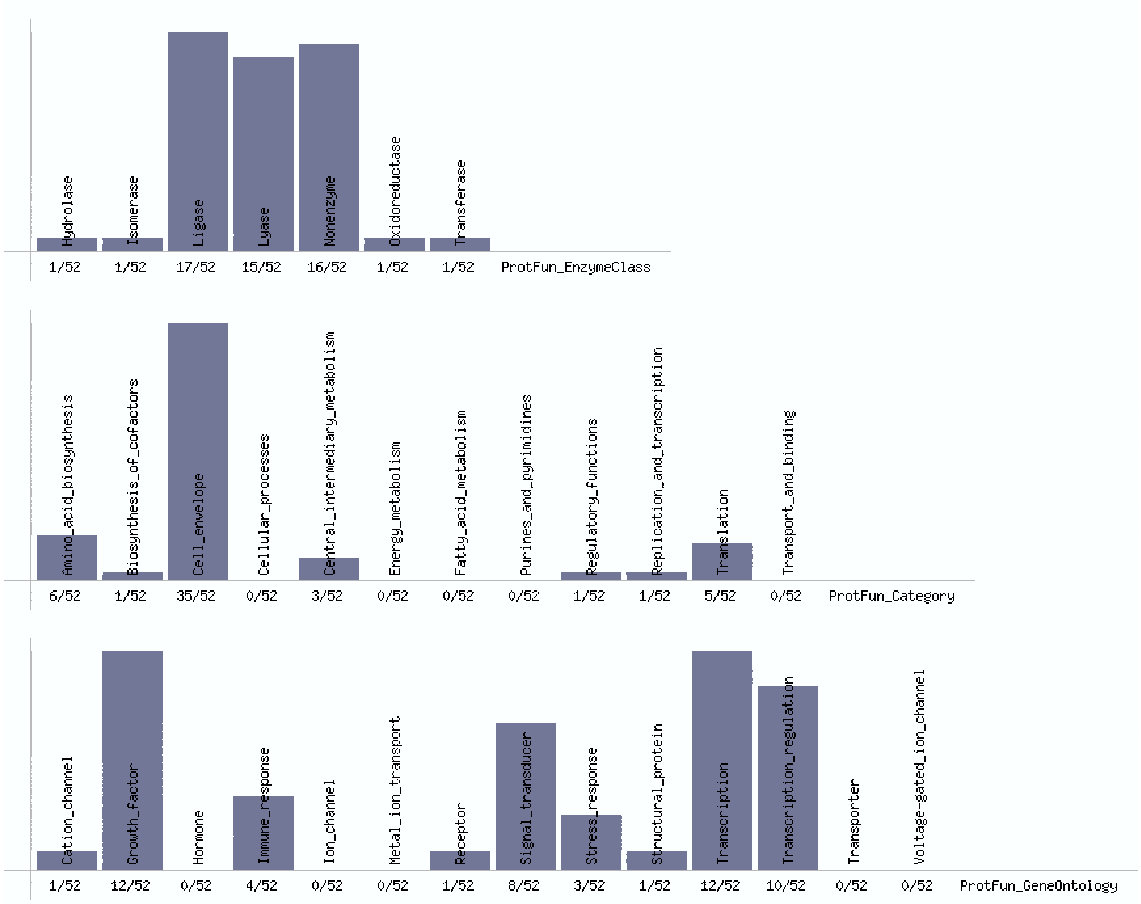

Figure 5.7.: Predicted protein function of the RCC proteins from the three different ProtFun categories.

5.4. Integrative analysis results

As an example of the usefulness of CAP, a large-scale analysis of data regarding genetic changes, gene expression levels, and autoimmune responses was done. Genes in CAP found by SEREX experiments were analyzed for genetic alterations and expression. This might give insights about the correlation of autoimmune responses and cancer genetics. A total of 1500 genes were extracted from CAP of which 19 variants could be found in CGW. In this step no consideration of cancer types was taken. In the next step genes that were found in the same cancer type according to CGW and SEREX were analyzed. Seven genes were found in this analysis, including two genes carrying specific mutations or polymorphisms, TP53 and GSTT1 (Glutathione S-transferase Theta 1). TP53 has previously been found to cause immune responses in primary colon carcinoma and in breast carcinoma, both known to carry TP53 mutations [9, 263]. Mutations in TP53 have also been found in a large number of other tumor types in patients that do not have antibodies against TP53. Furthermore, there

are several known MHC class I binding peptide derived from TP53 and has been suggested as a possible target for immunotherapy [13, 16, 232]. As for GSTT1, antibody responses occur in patients with breast cancer. This tumor is associated with specific GSTT1 polymorphisms [180]. However, this type of polymorphisms also occurs in other tumors including head and neck cancer without an antibody response [47]. Other examples of genes include NME2/NME1 (protein NM23B/A expressed in non-metastatic cells 2/1), HSPCA (heat shock 90kD protein 1), Ki-67 (MKI67), and MIF (macrophage migration inhibitory factor). NME1 and NME2 have been reported as immunogenic and overexpressed in malignant colon carcinoma [171], HSPCA in renal cell carcinoma [194], Ki-67 in melanoma [107, 111] and MIF in melanoma [256]. These genes might be interesting for further analysis in order to investigate if the mutations occur in epitope regions.

For analyzing expression levels, gene expression profile data from the NCI60 microarray project (http://genome-www.stanford.edu/nci60/) was used. In this project, cDNA microarrays were used to explore the variation of gene expression in 8,000 genes from 60 cancer cell lines. These 60 cell lines are also used by the National Cancer Institute for screening potential cancer drugs. The expression data provided by NCI includes fluorescence ratios, normalized against a pool of 12 cancer cell lines. Genes that show at least a two-fold increase in expression levels are considered to be overexpressed. The criteria that a certain gene must have measured expression levels in at least four of the sixty cell lines was also added. This gives expression levels for 319 CAP genes.

Independent of cancer type, 277 (87%) of the genes were found to be overexpressed in at least one cell line. Out of these 277 genes, 69 were found to have an overexpression in at least 10% of all evaluated cell lines. A more cancer-specific analysis was also done. The 60 cancer clones in the NCI60 data can be grouped into a number of cancer types, e.g. melanoma, breast cancer or colon cancer. Expression levels for genes found in the same cancer type in both SEREX experiments and the NCI60 data were extracted. The criteria for selection was that at least three tumor specific cell lines show overexpression, giving a total of 13 genes. The genes and SEREX-related information is presented in Table 5.2. These genes are once again interesting for further analysis, since they might give insight into the correlation of expression levels and autoimmune responses.

Table 5.2.: SEREX-related information for the 13 genes found in the same cancer type in both SEREX experiments and the NCI60 data. For a number of genes several related SEREX clones were found.

Abbreviation	Gene name	RefSeq id	SEREX clone	SEREX tumor
Melanoma				
COL9A3	collagen, type IX, alpha 3	NM_001853	Hom.TsMe3-89	melanoma
HEXB	hexosaminidase B (beta polypeptide)	NM_000521	Hom.TsMe2-12	melanoma
RRBP1	ribosome-binding protein 1 homologe 180kDa	NM_004587	TE53	unclassifiable
			TM-76	melanoma
SLC2A11	solute carrier family 2 (facilitated glucose	NM_030807	Mz19-64	melanoma
	transporter), member 11		NY-SAR-47	fibrosarcoma
TIMP	tissue inhibitor of metalloproteinase 3	NM_000362	Mz19-3	melanoma
Breast cancer				
P8	p8 protein (candidate of metastasis 1)	NM_012385	NY-BR-89	malignant breast
TP53	tumor protein p53 (Li-Fraumeni syndrome)	NM_000546	NY-Co-13	colorectal adenocarcinoma
			NY-BR-94	malignant breast
			NW-F14	malignant colon
			NW-F93	malignant colon
CENPF	centromer protein F, 350/400ka (mitosin)	NM_016343	MOC-SW-139	malignant colon
			MOC-SW-18	malignant colon
			MOC-SW-151	malignant colon
			NGO-Br-7	malignant breast
			MO-TES-148	malignant colon
			NY-ESO-11	esophageal cancer
			NGO-Pr-24	malignant prostate
			NY-BR-69	malignant breast
GBP1	guanylate binding protein 1, interferon-inducible, 67 kDa	NM_002053	NGO-Br-40	malignant breast
Colon cancer				
SCNN1A	sodium channel, nonvoltage-gated 1 alpha	NM_001038	NW-CD35b	adenocarcinoma colon
AP1G2	adaptor-related protein, complex 1, gamma 2 subunit	NM_003917	NW-SW15	adenocarcinoma colon
Lung cancer				
TRAP1	heat shock protein 75	NM_016292	LC19	malignant lung
Renal cancer				
PFKFB3	6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase 3	NM_004566	NY-REN-56	malignant kidney

Table 5.3.: Prediction of known TSA-peptides using SVMHC.

Protein	Peptide	Allele	SVMHC score	Reference
NY-ESO-1	SLLMWITQC	A*0201	0.54	[45, 130, 286]
NY-ESO-1	APRGVRMAV	B*07	0.57	[260]
NY-ESO-1	MPFATPMEA	B*35	0.03	[22]
NY-ESO-1	MPFATPMEA	B*51	-0.03	[131]
CTNNA1	FIDASRLVY	A*01	0.53	[152]
CTNNA1	LQHPDVAAY	B*1501	1.19	[152]
CTNNA1	NEQDLGIQY	B*44/B*18	0.16/0.74	[152]
hTERT	ILAKFLHWL	A*0201	0.94	[298]
hTERT	RLVDDFLLV	A*0201	1.17	[178]
hTERT	KLFGVLRLK	A*03	1.36	[297]
hTERT	VYAETKHFL	A*24	0.24	[11]
hTERT	VYGFVRACL	A*24	0.48	[11]

5.5. Finding candidates for T-cell based immunotherapy

Cancer immunotherapy candidates should fulfill a number of criteria. The immune response raised should be specific against the tumor, not induce autoimmunity, have a long-term effect etc. Both TSAs and TAA can be used for cancer immunotherapy and there is a variety of methods by which these can be detected. The SEREX method described above is one such method, but expression analysis or MHC-peptide sequencing of cancer-tissue samples can also be used. This further motivates the use integrative data analysis such as offered by CAP. Three well-characterized TSAs are used to exemplify the use of CAP and the methods developed in this thesis for the identification of peptide candidates for immunotherapy.

The first example is the well-characterized NY-ESO-1 protein [46] identified in esophageal squamous cell carcinoma (CAP id: 6007). However, reverse transcription-PCR analysis showed NY-ESO-1 mRNA expression many different tumor types including melanoma, breast cancer, bladder cancer, prostate cancer, and hepatocellular carcinoma. NY-ESO-1 belongs to the cancer testis (CT) antigens (proteins normally not expressed in tissues other than testis) and other well-known CTs are MAGE [289], BAGE [28], and GAGE [288]. NY-ESO-1 also contains several identified T-cell epitopes, see Table 5.3. Here it can be seen that three of the four peptides are successfully identified by SVMHC and the fourth peptide has a score very close to '0' (which is the cutoff for binding/non-binding),

The second example comes from a renal cell carcinoma (RCC). A whole range of interesting immunotherapy candidates were recently presented by Krüger *et al.* [152]. Here 13 specimen of RCC were used to identify a large number of MHC binding peptides. Most of these originate

from normal self-proteins, but put together with gene expression analysis some interesting candidates can be found. One example is the α-catenin 1 protein (CTNNA1) which was found to be overexpressed in most patients. This gene can also be found in CAP (CAP id: 6865) and it has been identified with the SEREX method. Hence, this is another good example of a gene found to be overexpressed, causing an autoimmune response, and having several identified Tc epitopes. Three MHC-binding peptides have been identified from CTNNA1 and prediction results using SVMHC are presented in Table 5.3. In this case SVMHC correctly identifies all three peptides.

The third TSA studied here is the Telomerase Reverse Transcriptase (hTERT) (CAP id: 26818). This protein is expressed in more than 85% of all tumors, but only rarely in normal cells [298, 143]. Several MHC-binding peptides have been identified from hTERT and investigated experimentally. Experimental studies using the HLA-A*0201 restricted epitope ILAKFLHWL showed that Tc cells specific to this epitope could kill a wide range of different hTERT$^+$ cell lines [298]. This peptide and several others are correctly identified by SVMHC, see Table 5.3.

5.6. CAP discussion

The CAP database is a good example of how heterogeneous data from different types of experimental procedure can be integrated and analyzed. New efforts in molecular biology have been made to describe biological entities in a structured manner. A promising example here is the gene ontology aiming to describe gene and gene product in a structured way [12]. Furthermore, there is a never ending flood of prediction methods that can be used to annotate protein sequences. The underlying data model of CAP is designed in a way that these types of data can easily be incorporated.

The use of CAP was illustrated by the analysis of genes causing autoimmune responses in different cancer types. This analysis gives an overall view of how immune responses in cancer relate to gene expression and genetic modifications. CAP also enables statistical evaluation of user-defined data sets, something that can facilitate hypothesis regarding different aspects of cancer research.

One hypothesis is that immunogenic antigens might stem from genes that are altered by tumor-specific mutations or have a changed expression profile in a certain tumor, reviewed by Meese and Comtesse [174]. Here CAP was to merge information from the fields of genetics

and immunology. From available data, we see no evidence that genetic alterations, such as mutations or polymorphisms, cause immune responses in cancer. In terms of variations in expression, indications can be seen that overexpression contributes to the antibody responses against tumor antigens. The majority of the 319 genes are actually found to be overexpressed in the NCI60 data set. This result might be somewhat biased from the selection of genes tested for expression levels. The NCI60 data set was designed to explore the variation in gene expression among different cancer types.

It might be misleading to turn these findings into general rules concerning immune responses in cancer. This study was rather conducted to show CAP makes this kind of analysis possible by integrating different sources of data.

Statistics on specific data sets might help to understand the mechanisms behind certain cancer types. There are many reports on the correlations between cancers and chromosomal aberrations. One example is the changed expression patterns of genes in ovarian carcinomas [20]. These genes show reduced expression in the 3p25.5-3p21.31 region and increased expression of genes from 3q13.33-3q28. CAP can be a useful tool in the identification of such chromosomal regions. Many cancer types have disrupted protein and signaling pathways. One example is the retinoblastoma protein pathway, reviewed in [17]. Analysis of protein function and subcellular location are important steps in the identification of such pathways. CAP provides tools for both finding protein functional families associated to certain cancers as well as analysis of protein subcellular location for sets of sequences.

The final part of this chapter showed how SVMHC can be used to identify cancer vaccine candidates. This was done by analyzing some TAAs and TSAs identified by CAP that also contain experimentally verified MHC-binding peptides. Peptides originating from proteins specifically found in certain tumors are interesting for immunotherapeutic purposes.

6. Discussion and concluding remarks

This work describes computational methods for modeling antigen processing and for finding immunotherapy candidates. The first part focused on MHC-peptide binding alone, where several approaches for MHC-peptide binding prediction were presented. In the following chapter antigen processing as an integrated pathway was analyzed. The novel WAPP method was presented and new methods for proteasomal cleavage and TAP transport were described in detail. Finally the integration of cancer-related data and identification of immunotherapy candidates was described. Here the concept of integrating heterogeneous cancer-related data for was introduced. The CAP analysis tool developed was used for identifying TSAs and TAAs, which were further analyzed of cancer peptide vaccine candidates.

Several different methods for prediction of MHC class I binding peptides have been presented. The first methods were based on expert knowledge regarding certain anchor positions of the peptides. These were followed by PSSM-based methods like BIMAS and SYFPEITHI. The main drawback of PSSM-based methods is that they assume an independent contribution to the overall binding energy from each amino acid of the peptide. The SVM-based method presented here can circumvent this and allow for a non-linear model. One might argue that the "black box" SVMs do not give an easily interpretable explanation for binding, but in many cases this is not the focus. The most important aspect is that MHC-binding peptides are predicted with high accuracy, and in this sense the SVM-based method, SVMHC, outperforms other PSSM-based methods. In a comparison of between the SVMHC, SYFPEITHI, and BIMAS method for six different MHC alleles, SVMHC had an average MCC of 0.84 compared to 0.79 and 0.80 for SYFPEITHI and BIMAS respectively. Consensus prediction of MHC-binding peptides is also interesting and here such a method was presented for HLA-A*0201. When more data becomes available this method is easily extendible to further alleles. Using DTs and biochemical properties of the amino acids also give interesting results. Here, simple interpretable rules can be generated explaining different aspects of MHC-peptide interaction.

Antigen processing is a highly complex machinery and hypotheses regarding different parts of the pathway are presented on a weekly basis. The WAPP method presented in this thesis clearly shows that proteasomal cleavage and TAP transport can be successfully taken into consideration, when modeling the overall antigen processing pathway of MHC class I-restricted antigens. This have implication both in basic immunology research and in rational vaccine design. Depending on the way a certain vaccine is administrated, processing events can also be important in order to get sufficient MHC-peptide presentation.

The PCM method for proteasomal cleavage prediction presented in this thesis shows improved performance and stability compared to previously presented methods. The PCM method avoids overfitting to the training data, which is the major problem of related PAProC and NetChop methods. The average accuracy of the PCM method is 65% which can be compared to 47% for the MAPPP method, 47% for the PAProC method, and 61% for the NetChop method. Incorporating a higher complexity model for proteasomal cleavage into WAPP might be useful when more experimental data become available.

There are alternative events possible for both protein cleavage and peptide transport. The proteasome is not the only protease that cleaves proteins in the cytosol. Another major player here seems to be the protease TPPII. It seems as if a number of alleles bind peptides that are not effectively generated and some evidence exist that TPPII is involved in the generation of these [222]. These findings would explain why several alleles can bind peptides with a Lys residue in the C-terminus, which are typically not generated by the proteasome. There is not enough data available for TPPII in order to create a prediction model as for the proteasome.

The SVMTAP method shows a good prediction performance, with a correlation of 0.82 between the experimentally measured and predicted values. The presented results highlight the importance of the three N-terminal and the C-terminal amino acids to the binding affinity. This is further verified, since using only these peptide positions gives a correlation of 0.79 between experimentally measured and predicted values. A high quality data set of peptides with a length of nine amino acids was used in this study. Since most of the peptides presented by MHC molecules have exactly this length, WAPP focuses on nine amino acid long peptides. However, some findings suggest that peptides might enter the ER as longer precursors and further trimmed in the ER by ER-peptidases [251, 252]. For extended peptides there is once again a lack of data and no convincing model of the transport and subsequent processing of peptide in the ER have been presented so far. SVMTAP was also used to analyze the differences in TAP affinity of peptides from different MHC alleles. These results suggest two

sets of alleles, one TAP-efficient and one TAP-inefficient allele in terms of TAP transport.

The methods for proteasomal cleavage (PCM), TAP transport (SVMTAP), and MHC binding (SVMHC) were integrated in the WAPP method. For the four alleles analyzed with WAPP, improved performance in comparison to MHC binding prediction alone was observed. Furthermore, WAPP outperforms the other currently existing methods for whole pathway prediction. The accumulation of new data will definitely improve the modeling and accuracy assessment of integrative models of the whole processing pathway.

The focus of the final results chapter of this thesis describes analysis of cancer-related data. The MHC-peptide prediction and whole pathway modeling chapters includes a lot of comparative studies on large data sets, where it is fairly straight forward to determine the best method. The focus of the chapter dealing with integrative analysis of cancer-related data was different. Here the main question is how one can generated a data model that accommodates and enables analysis of heterogeneous data from different sources. The verification that this can be done successfully was the subsequent large-scale analysis conducted. Furthermore, it was shown that several TSAs identified by CAP could be used to identify peptide vaccine candidates. The need for integrated databases and analysis tools like CAP will continue to increase. New technical advances constantly produce data that makes it possible to understand new dimensions of cancer research. Without systems that enable storage of these data in a very structured way, much useful information will be lost.

Databases like CAP can supply us with target proteins for immunotherapy. Coupled to accurate prediction tools, fast identification of peptide-vaccine candidates is possible. There are of many hurdles to overcome regarding T-cell based immunotherapy. In principle it would probably be optimal to activate both B-cells and Th cells in addition to Tc cells. However, prediction of B-cell epitopes is a much trickier task that prediction of T-cell epitopes. In selecting vaccine candidates it might also be good to consider similarity to normal proteins in order to avoid allergic reactions. There are also several open questions regarding how to optimally select a set of candidate peptides in order to cover a wide range of alleles. Assuming that immunogenicity can be related to MHC-peptide binding, it should be possible formulate an optimization function using different constraints, describing the optimal set of candidate peptides. In the ideal case, it might even be possible to construct personalized peptide-vaccine, depending on the HLA type of a patient. Furthermore, other types of data, such as gene expression, might be used to reduce the risk of allergy or to optimally select target proteins.

Immunology is very complex field and new technologies constantly provide useful insights into the function of the immune system. In a recent review article, Jonathan Yewdell describes seven dirty little secrets (DLSs) of antigen processing [302]. He concludes that in order to understand antigen processing and immunology in general it is necessary to understand the interactions between trillions of cells that make up the organism. By describing different DLSs, Yewdell underlines that there are still many "if's" and loose assumptions regarding antigen processing. In some respects the human brain is definitely limited: "Why should this organ, selected to facilitate our survival in the macro-world of lions-and-tigers-and-bears, be equipped to fully understand the micro- and nano-world of molecules-and-atoms-and-waves?". However, it is easy to agree with Yewdell that this is not discouraging in any way, since it implies that there will always be an unlimited source of intriguing questions and hypotheses to investigate.

Bibliography

[1] Abbas, A. K., Lichtman, A. H., and Pober, J. *Cellular and molecular immunology.* W.B Saunders Company, Philadelphia, USA, 2000.

[2] Aderem, A., and Underhill, D. M. Mechanisms of phagocytosis in macrophages. *Annu Rev Immunol 17* (1999), 593–623.

[3] Admon, A., Barnea, E., and Ziv, T. Tumor antigens and proteomics from the point of view of the major histocompatibility complex peptides. *Mol Cell Proteomics 2*, 6 (Jun 2003), 388–398.

[4] Aki, M., Shimbara, N., Takashina, M., Akiyama, K., Kagawa, S., Tamura, T., Tanahashi, N., Yoshimura, T., Tanaka, K., and Ichihara, A. Interferon-gamma induces different subunit organizations and functional diversity of proteasomes. *J Biochem (Tokyo) 115*, 2 (Feb 1994), 257–269.

[5] Albert, M. L., Pearce, S. F., Francisco, L. M., Sauter, B., Roy, P., Silverstein, R. L., and Bhardwaj, N. Immature dendritic cells phagocytose apoptotic cells via $\alpha_v\beta_5$ and CD36, and cross-present antigens to cytotoxic T lymphocytes. *J Exp Med 188*, 7 (Oct 1998), 1359–1368.

[6] Altuvia, Y., and Margalit, H. Sequence signals for generation of antigenic peptides by the proteasome: implications for proteasomal cleavage mechanism. *J. Mol. Biol. 295* (2000), 879–890.

[7] Altuvia, Y., Schueler, O., and Margalit, H. Ranking potential binding peptides to MHC molecules by a computational threading approach. *J Mol Biol 249*, 2 (Jun 1995), 244–250.

[8] Altuvia, Y., Sette, A., Sidney, J., Southwood, S., and Margalit, H. A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. *Hum Immunol 58*, 1 (Nov 1997), 1–11.

[9] Angelopoulou, K., Yu, H., Bharaj, B., Giai, M., and Diamandis, E. p53 gene mutation, tumor p53 protein overexpression, and serum p53 autoantibody generation in patients with breast cancer. *Clin. Biochem. 33*, 1 (2000), 53–62.

[10] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L.-S. L. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res 32*, Database issue (Jan 2004), 115–119.

[11] Arai, J., Yasukawa, M., Ohminami, H., Kakimoto, M., Hasegawa, A., and Fujita, S. Identification of human telomerase reverse transcriptase-derived peptides that induce HLA-A24-restricted antileukemia cytotoxic T lymphocytes. *Blood 97*, 9 (May 2001), 2903–2907.

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

[13] Azuma, K., Shichijo, S., Maeda, Y., Nakatsura, T., Nonaka, Y., Fujii, T., Koike, K., and Itoh, K. Mutated p53 gene encodes a nonmutated epitope recognized by HLA-B*4601-restricted and tumor cell-reactive CTLs at tumor site. *Cancer Res 63*, 4 (Feb 2003), 854–858.

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

[15] Banchereau, J., and Steinman, R. M. Dendritic cells and the control of immunity. *Nature 392*, 6673 (Mar 1998), 245–252.

[16] Barfoed, A. M., Petersen, T. R., Kirkin, A. F., Thor Straten, P., Claesson, M. H., and Zeuthen, J. Cytotoxic T-lymphocyte clones, established by stimulation with the HLA-A2 binding p5365-73 wild type peptide loaded on dendritic cells In vitro, specifically recognize and lyse HLA-A2 tumour cells overexpressing the p53 protein. *Scand J Immunol 51*, 2 (Feb 2000), 128–133.

[17] Bartek, J., Bartkova, J., and Lukas, J. The retinoblastoma protein pathway in cell cycle control and cancer. *Exp. Cell. Res. 237*, 1 (1997), 1–6.

[18] Bauer, C., Diesinger, I., Brass, N., Steinhart, H., Iro, H., and Meese, E. Translation initiation factor eIF-4G is immunogenic, overexpressed, and amplified in patients with squamous cell lung carcinoma. *Cancer 92*, 4 (2001), 822–829.

[19] Baurain, J. F., Colau, D., van Baren, N., Landry, C., Martelange, V., Vikkula, M., Boon, T., and Coulie, P. G. High frequency of autologous anti-melanoma CTL directed against an antigen generated by a point mutation in a new helicase gene. *J Immunol 164*, 11 (Jun 2000), 6057–6066.

[20] Bayani, J., Brenton, J., Macgregor, P., Beheshti, B., Albert, M., Nallainathan, D., Karaskova, J., Rosen, B., Murphy, J., Laframboise, S., Zanke, B., and Squire, J. Parallel analysis of sporadic primary ovarian carcinomas by spectral karyotyping, comparative genomic hybridization, and expression microarrays. *Cancer Res. 62*, 12 (2002), 3466–3476.

[21] Beninga, J., Rock, K., and Goldberg, A. Interferon-gamma can stimulate post-proteasomal trimming of the N terminus of an antigenic peptide by inducing leucine aminopeptidase. *J. Biol. Chem. 273*, 30 (1998), 18734–18742.

[22] Benlalam, H., Linard, B., Guilloux, Y., Moreau-Aubry, A., Derre, L., Diez, E., Dreno, B., Jotereau, F., and Labarriere, N. Identification of five new HLA-B*3501-restricted epitopes derived from common melanoma-associated antigens, spontaneously recognized by tumor-infiltrating lymphocytes. *J Immunol 171*, 11 (Dec 2003), 6283–6289.

[23] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic Acids Res. 28*, 1 (2000), 235–242.

[24] Bhasin, M., and Raghava, G. P. S. Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Sci. 13*, 3 (2004), 596–607.

[25] Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L., and Wiley, D. C. The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature 329*, 6139 (Oct 1987), 512–518.

[26] Blythe, M. J., Doytchinova, I. A., and Flower, D. R. JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics 18*, 3 (Mar 2002), 434–439.

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

[28] Boel, P., Wildmann, C., Sensi, M. L., Brasseur, R., Renauld, J. C., Coulie, P., Boon, T., and van der Bruggen, P. BAGE: a new gene encoding an antigen recognized on human melanomas by cytolytic T lymphocytes. *Immunity 2*, 2 (Feb 1995), 167–175.

[29] Boser, B. E., Guyon, I., and Vapnik, V. A training algorithm for optimal margin classifiers. In *Computational Learning Theory* (1992), pp. 144–152.

[30] Brannigan, J. A., Dodson, G., Duggleby, H. J., Moody, P. C., Smith, J. L., Tomchick, D. R., and Murzin, A. G. A protein catalytic framework with an N-terminal nucleophile is capable of self-activation. *Nature 378*, 6555 (Nov 1995), 416–419.

[31] Brass, N., Heckel, D., Sahin, U., Pfreundschuh, M., Sybrecht, G., and Meese, E. Translation initiation factor eif-4γ is encoded by an amplified gene and induces an immune response in squamous cell lung carcinoma. *Hum. Mol. Genet. 6*, 1 (1997), 33–39.

[32] Brass, N., Racz, A., Bauer, C., Heckel, D., Sybrecht, G., and Meese, E. Role of amplified genes in the production of autoantibodies. *Blood 93*, 7 (1999), 2158–2166.

[33] Braun, B. C., Glickman, M., Kraft, R., Dahlmann, B., Kloetzel, P. M., Finley, D., and Schmidt, M. The base of the proteasome regulatory particle exhibits chaperone-like activity. *Nat Cell Biol 1*, 4 (Aug 1999), 221–226.

[34] Breiman, L., Friedman, J., Olshen, R., and Stone, C. *Classification and regression trees.* Wadsworth and Brooks, 1984.

[35] Bronte, V., Apolloni, E., Ronca, R., Zamboni, P., Overwijk, W. W., Surman, D. R., Restifo, N. P., and Zanovello, P. Genetic vaccination with "self" tyrosinase-related protein 2 causes melanoma eradication but not vitiligo. *Cancer Res 60*, 2 (Jan 2000), 253–258.

[36] Brower, R. C., England, R., Takeshita, T., Kozlowski, S., Margulies, D. H., Berzofsky, J. A., and Delisi, C. Minimal requirements for peptide mediated activation of CD8+ CTL. *Mol Immunol 31*, 16 (Nov 1994), 1285–1293.

[37] Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. J., and Haussler, D. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A. 97* (2000), 262–267.

[38] Brusic, V., Rudy, G., Honeyman, G., Hammer, J., and Harrison, L. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics 14*, 2 (1998), 121–130.

[39] Brusic, V., van Endert, P., Zeleznikow, J., Daniel, S., Hammer, J., and Petrovsky, N. A neural network model approach to the study of human TAP transporter. *In Silico Biol. 1*, 2 (1999), 109–121.

[40] Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery 2*, 2 (1998), 121–167.

[41] Burges, C. J. C., and Crisp, D. J. Uniqueness of the SVM solution. In *NIPS99* (1999).

[42] Burroughs, N., de Boer, R., and Kesmir, C. Discriminating self from nonself with short peptides from large proteomes. *Immunogenetics 56*, 5 (2004), 311–320.

[43] Buus, S., Lauemoller, S. L., Worning, P., Kesmir, C., Frimurer, T., Corbet, S., Fomsgaard, A., Hilden, J., Holm, A., and Brunak, S. Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens 62*, 5 (Nov 2003), 378–384.

[44] Cerundolo, V., Alexander, J., Anderson, K., Lamb, C., Cresswell, P., McMichael, A., Gotch, F., and Townsend, A. Presentation of viral antigen controlled by a gene in the major histocompatibility complex. *Nature 345*, 6274 (May 1990), 449–452.

[45] Chen, J. L., Dunbar, P. R., Gileadi, U., Jager, E., Gnjatic, S., Nagata, Y., Stockert, E., Panicali, D. L., Chen, Y. T., Knuth, A., Old, L. J., and Cerundolo, V. Identification of NY-ESO-1 peptide analogues capable of improved stimulation of tumor-reactive CTL. *J Immunol 165*, 2 (Jul 2000), 948–955.

[46] Chen, Y. T., Scanlan, M. J., Sahin, U., Tureci, O., Gure, A. O., Tsang, S., Williamson, B., Stockert, E., Pfreundschuh, M., and Old, L. J. A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening. *Proc Natl Acad Sci U S A 94*, 5 (Mar 1997), 1914–1918.

[47] Cheng, L., Sturgis, E., Eicher, S., Char, D., Spitz, M., and Wei, Q. Glutathione-S-transferase polymorphisms and risk of squamous-cell carcinoma of the head and neck. *Int. J. Cancer. 84*, 3 (1999), 220–224.

[48] Cheung, K., Nadkarni, P., and Shin, D. A metadata approach to query interoperation between molecular biology databases. *Bioinformatics 14*, 6 (1998), 486–497.

[49] Chiari, R., Foury, F., De Plaen, E., Baurain, J. F., Thonnard, J., and Coulie, P. G. Two antigens recognized by autologous cytolytic T lymphocytes on a melanoma result from a single point mutation in an essential housekeeping gene. *Cancer Res 59*, 22 (Nov 1999), 5785–5792.

[50] Chiong, B., Wong, R., Lee, P., Delto, J., Scotland, R., Lau, R., and Weber, J. Characterization of long-term effector-memory T-cell responses in patients with resected high-risk melanoma receiving a melanoma Peptide vaccine. *J Immunother 27*, 5 (Sep 2004), 368–379.

[51] Christinck, E. R., Luscher, M. A., Barber, B. H., and Williams, D. B. Peptide binding to class I MHC on living cells and quantitation of complexes required for CTL lysis. *Nature 352*, 6330 (Jul 1991), 67–70.

[52] Claassen, M. Molecular modeling of peptide ligands in HLA-A*0201 complexes. Master's thesis, Eberhard-Karls-Universität Tübingen, Germany, 2004.

[53] Cochlovius, B., Stassar, M., Christ, O., Raddrizzani, L., Hammer, J., Mytilineos, I., and Zoller, M. In vitro and in vivo induction of a Th cell response toward peptides of the melanoma-associated glycoprotein 100 protein selected by the TEPITOPE program. *J Immunol 165*, 8 (Oct 2000), 4731–4741.

[54] Comtesse, N., Heckel, D., Racz, A., Brass, N., Glass, B., and Meese, E. Five novel immunogenic antigens in meningioma: cloning, expression analysis, and chromosomal mapping. *Clin. Cancer Res. 5*, 11 (1999), 3560–3568.

[55] Comtesse, N., Niedermayer, I., Glass, B., Heckel, D., Maldener, E., Nastainczyk, W., Feiden, W., and Meese, E. MGEA6 is tumor-specific overexpressed and frequently recognized by patient-serum antibodies. *Oncogene 21*, 2 (2002), 239–247.

[56] Conover, W. J. *Practical nonparametric statistics, 3rd ed.* John Wiley & Sons, New York, USA, 1999.

[57] Cortes, C., and Vapnik, V. Support-vector networks. *Machine Learning 20*, 3 (1995), 273–297.

[58] Cotten, J. F., Ostedgaard, L. S., Carson, M. R., and Welsh, M. J. Effect of cystic fibrosis-associated mutations in the fourth intracellular loop of cystic fibrosis transmembrane conductance regulator. *J Biol Chem 271*, 35 (Aug 1996), 21279–21284.

[59] Coulie, P. G., Lehmann, F., Lethe, B., Herman, J., Lurquin, C., Andrawiss, M., and Boon, T. A mutated intron sequence codes for an antigenic peptide recognized by cytolytic T lymphocytes on a human melanoma. *Proc Natl Acad Sci U S A 92*, 17 (Aug 1995), 7976–7980.

[60] Craiu, A., Akopian, T., Goldberg, A., and Rock, K. Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide. *Proc. Natl. Acad. Sci. USA 94*, 20 (1997), 10850–10855.

[61] Cresswell, P. Assembly, transport, and function of MHC class II molecules. *Annu Rev Immunol 12* (1994), 259–293.

[62] Cresswell, P. Invariant chain structure and MHC class II function. *Cell 84*, 4 (Feb 1996), 505–507.

[63] Cristianini, N., and Shawe-Taylor, J. *An Introduction to Support vector machines and other kernel-based learning methods.* Cambridge University Press, The Edinburgh Building, Cambridge, UK, 2000.

[64] Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M., and Barton, G. J. JPred: a consensus secondary structure prediction server. *Bioinformatics 14*, 10 (1998), 892–893.

[65] Daniel, S., Brusic, V., Caillat-Zucman, S., Petrovsky, N., Harrison, L., Riganelli, D., Sinigaglia, F., Gallazzi, F., Hammer, J., and van Endert, P. M. Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules. *J Immunol 161*, 2 (Jul 1998), 617–624.

[66] de la Salle, H., Houssaint, E., Peyrat, M. A., Arnold, D., Salamero, J., Pinczon, D., Stevanovic, S., Bausinger, H., Fricker, D., Gomard, E., Biddison, W., Lehner, P., UytdeHaag, F., Sasportes, M., Donato, L., Rammensee, H. G., Cazenave, J. P., Hanau, D., Tongio, M. M.,

and Bonneville, M. Human peptide transporter deficiency: importance of HLA-B in the presentation of TAP-independent EBV antigens. *J Immunol 158*, 10 (May 1997), 4555–4563.

[67] Delamarre, L., Holcombe, H., and Mellman, I. Presentation of exogenous antigens on major histocompatibility complex (MHC) class I and MHC class II molecules is differentially regulated during dendritic cell maturation. *J Exp Med 198*, 1 (Jul 2003), 111–122.

[68] Demotz, S., Grey, H. M., and Sette, A. The minimal number of class II MHC-antigen complexes needed for T cell activation. *Science 249*, 4972 (Aug 1990), 1028–1030.

[69] DiBrino, M., Parker, K. C., Shiloach, J., Turner, R. V., Tsuchida, T., Garfield, M., Biddison, W. E., and Coligan, J. E. Endogenous peptides with distinct amino acid anchor residue motifs bind to HLA-A1 and HLA-B8. *J Immunol 152*, 2 (1994), 620–631.

[70] Diesinger, I., Bauer, C., Brass, N., Schaefers, H., Comtesse, N., Sybrecht, G., and Meese, E. Toward a more complete recognition of immunoreactive antigens in squamous cell lung carcinoma. *Int. J. Cancer 102*, 4 (2002), 372–378.

[71] Dönnes, P. Prediction of MHC class I binding peptides, using a machine learning approach. Master's thesis, University of Linköping, Sweden, 2001.

[72] Dönnes, P., and Elofsson, A. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics. 3*, 1 (2002), 25.

[73] Dönnes, P., and Höglund, A. Predicting Protein Subcellular Localization: Past, Present, and Future. *Genomics, Proteomics, and Bioinformatics 2*, 4 (2004).

[74] Dönnes, P., Höglund, A., Sturm, M., Comtesse, N., Backes, C., Meese, E., Kohlbacher, O., and Lenhof, H.-P. Integrative analysis of cancer-related data using CAP. *FASEB J 18*, 12 (Sep 2004), 1465–1467.

[75] Dönnes, P., and Kohlbacher, O. Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Sci 14*, 8 (Aug 2005), 2132–2140.

[76] Driscoll, J., Brown, M. G., Finley, D., and Monaco, J. J. MHC-linked LMP gene products specifically alter peptidase activities of the proteasome. *Nature 365*, 6443 (Sep 1993), 262–264.

[77] Dyrskjot, L., Thykjaer, T., Kruhoffer, M., Jensen, J. L., Marcussen, N., Hamilton-Dutoit, S., Wolf, H., and Orntoft, T. F. Identifying distinct classes of bladder carcinoma using microarrays. *Nat Genet 33*, 1 (Jan 2003), 90–96.

[78] E, G., Pfeifer, G., Wilm, M., Lucchiari-Hartz, M., Baumeister, W., Eichmann, K., and Niedermann, G. A giant protease with potential to substitute for some functions of the proteasome. *Science 283*, 5404 (1999), 978–981.

[79] Ehring, B., Meyer, T. H., Eckerskorn, C., Lottspeich, F., and Tampe, R. Effects of major-histocompatibility-complex-encoded subunits on the peptidase and proteolytic activities of human 20S proteasomes. Cleavage of proteins and antigenic peptides. *Eur J Biochem 235*, 1-2 (Jan 1996), 404–415.

[80] Emmerich, N. P., Nussbaum, A. K., Stevanovic, S., Priemer, M., Toes, R. E., Rammensee, H.-G., and Schild, H. The human 26 S and 20 S proteasomes generate overlapping but different sets of peptide fragments from a model protein substrate. *J. Biol. Chem. 275* (2000), 21140–21148.

[81] Engelhard, V. H., Appella, E., Benjamin, D. C., Bodnar, W. M., Cox, A. L., Chen, Y., Henderson, R. A., Huczko, E. L., Michel, H., and Sakaguichi, K. Mass spectrometric analysis of peptides associated with the human class I MHC molecules HLA-A2.1 and HLA-B7 and identification of structural features that determine binding. *Chem Immunol 57* (1993), 39–62.

[82] Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P., and Matthews, B. W. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science 255*, 5041 (Jan 1992), 178–183.

[83] Etlinger, J. D., and Goldberg, A. L. A soluble ATP-dependent proteolytic system responsible for the degradation of abnormal proteins in reticulocytes. *Proc Natl Acad Sci U S A 74*, 1 (Jan 1977), 54–58.

[84] Falk, K., Rotzschke, O., Stevanovic, S., Jung, G., and HG, R. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Science 351*, 6234 (1991), 290–296.

[85] Falk, K., Rötzschke, O., Stevanovic, S., Jung, G., and Rammensee, H. G. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature 351*, 6324 (1991), 290–296.

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

[87] Faucz, F. R., Probst, C. M., and Petzl-Erler, M. L. Polymorphism of LMP2, TAP1, LMP7 and TAP2 in Brazilian Amerindians and Caucasoids: implications for the evolution of allelic and haplotypic diversity. *Eur J Immunogenet 27*, 1 (Feb 2000), 5–16.

[88] Fischer, U., Struss, A., Hemmer, D., Pallasch, C., Steudel, W., and Meese, E. Glioma-expressed antigen 2 (GLEA2): a novel protein that can elicit immune responses in glioblastoma patients and some controls. *Clin. Exp. Immunol. 126*, 2 (2001), 206–213.

[89] Fletcher, R. *Practical Methods of Optimization*. John Wiley & Sons, New York, USA, 1987.

[90] Frederiksen, C. M., Knudsen, S., Laurberg, S., and Orntoft, T. F. Classification of Dukes' B and C colorectal cancers using expression arrays. *J Cancer Res Clin Oncol 129*, 5 (May 2003), 263–271.

[91] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics 16*, 10 (Oct 2000), 906–914. Evaluation Studies.

[92] Gaczynska, M., Rock, K. L., and Goldberg, A. L. Gamma-interferon and expression of MHC genes regulate peptide hydrolysis by proteasomes. *Nature 365*, 6443 (1993), 264–267.

[93] Gakamsky, D. M., Davis, D. M., Strominger, J. L., and Pecht, I. Assembly and dissociation of human leukocyte antigen (HLA)-A2 studied by real-time fluorescence resonance energy transfer. *Biochemistry 39*, 36 (Sep 2000), 11163–11169.

[94] Geuze, H. J. The role of endosomes and lysosomes in MHC class II functioning. *Immunol Today 19*, 6 (Jun 1998), 282–287.

[95] Glickman, M. H., Rubin, D. M., Fu, H., Larsen, C. N., Coux, O., Wefes, I., Pfeifer, G., Cjeka, Z., Vierstra, R., Baumeister, W., Fried, V., and Finley, D. Functional analysis of the proteasome regulatory particle. *Mol Biol Rep 26*, 1-2 (Apr 1999), 21–28.

[96] Glynne, R., Powis, S. H., Beck, S., Kelly, A., Kerr, L. A., and Trowsdale, J. A proteasome-related gene between the two ABC transporter loci in the class II region of the human MHC. *Nature 353*, 6342 (Sep 1991), 357–360.

[97] Goldberg, A. L., and St John, A. C. Intracellular protein degradation in mammalian and bacterial cells: Part 2. *Annu Rev Biochem. 45* (1976), 747–803.

[98] Goldsby, R. A., Kindt, T. J., Osborne, B. A., and Kuby, J. *Immunology 5th ed.* W.H. Freeman and company, Basingstoke, USA, 2003.

[99] Gorbulev, S., Abele, R., and Tampe, R. Allosteric crosstalk between peptide-binding, transport, and ATP hydrolysis of the ABC transporter TAP. *Proc Natl Acad Sci U S A 98*, 7 (Mar 2001), 3732–3737.

[100] Groll, M., Ditzel, L., Lowe, J., Stock, D., Bochtker, M., Bartunik, H. D., and Huber, R. Structure of the 20 S proteasome from yeast at 2.4 *Å* resolution. *Nature 386*, 6624 (1997), 463–471.

[101] Gubler, B., Daniel, S., Armandola, E. A., Hammer, J., Caillat-Zucman, S., and van Endert, P. M. Substrate selection by transporters associated with antigen processing occurs during peptide binding to TAP. *Mol. Immunol. 35*, 8 (1998), 427–433.

[102] Gulukota, K., Sidney, J., Sette, A., and DeLisi, C. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol. 267* (1997), 1258–1267.

[103] Hanada, K., Perry-Lalley, D. M., Ohnmacht, G. A., Bettinotti, M. P., and Yang, J. C. Identification of fibroblast growth factor-5 as an overexpressed antigen in multiple human adenocarcinomas. *Cancer Res 61*, 14 (Jul 2001), 5511–5516.

[104] Hanada, K., Yewdell, J., and Yang, J. Immune recognition of a human renal cancer antigen through post-translational protein splicing. *Nature 427*, 6971 (2004), 252–256.

[105] Harding, C. V., and Unanue, E. R. Quantitation of antigen-presenting cell MHC class II/peptide complexes necessary for T-cell stimulation. *Nature 346*, 6284 (Aug 1990), 574–576.

[106] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning.* Springer-Verlag, New York, USA, 2001.

[107] Hazan, C., Melzer, K., Panageas, K., Li, E., Kamino, H., Kopf, A., Cordon-Cardo, C., Osman, I., and Polsky, D. Evaluation of the proliferation marker MIB-1 in the prognosis of cutaneous malignant melanoma. *Cancer 95*, 3 (2002), 634–640.

[108] Heckel, D., Comtesse, N., Brass, N., Blin, N., Zang, K., and Meese, E. Novel immunogenic antigen homologous to hyaluronidase in meningioma. *Hum. Mol. Genet. 7*, 12 (1998), 1859–1872.

[109] Heemels, M. T., Schumacher, T. N., Wonigeit, K., and Ploegh, H. L. Peptide translocation by variants of the transporter associated with antigen processing. *Science 262*, 5142 (Dec 1993), 2059–2063.

[110] Henikoff, J. G., and Henikoff, S. Using substitution probabilities to improve position-specific scoring matrices. *Comput Appl Biosci 12*, 2 (Apr 1996), 135–143.

[111] Henrique, R., Azevedo, R., Bento, M., Domingues, J., Silva, C., and Jeronimo, C. Prognostic value of Ki-67 expression in localized cutaneous malignant melanoma. *J. Am. Acad. Dermatol. 43*, 6 (2000), 991–1000.

[112] Herberts, C., Reits, E., and Neefjes, J. Proteases, proteases and proteases for presentation. *Nat Immunol 4*, 4 (Apr 2003), 306–308. Comment.

[113] Hertz, G. Z., and Stormo, G. D. Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics 15*, 7-8 (1999), 563–577.

[114] Hisamatsu, H., Shimbara, N., Saito, Y., Kristensen, P., Hendil, K. B., Fujiwara, T., Takahashi, E., Tanahashi, N., Tamura, T., Ichihara, A., and Tanaka, K. Newly identified pair of proteasomal subunits regulated reciprocally by interferon γ. *J Exp Med 183*, 4 (Apr 1996), 1807–1816.

[115] Hogan, K. T., Coppola, M. A., Gatlin, C. L., Thompson, L. W., Shabanowitz, J., Hunt, D. F., Engelhard, V. H., Slingluff, C. L. J., and Ross, M. M. Identification of a shared epitope recognized by melanoma-specific, HLA-A3-restricted cytotoxic T lymphocytes. *Immunol. Lett. 90*, 2 (2003), 131–135.

[116] Höglund, A., Dönnes, P., Adolph, H., and Kohlbacher, O. From prediction of subcellular localization to functional classification: Discrimination of DNA-packing and other nuclear proteins. *Online Journal of Bioinformatics 6*, 1 (2005), 51–64.

[117] Höglund, A., Dönnes, P., Blum, T., Adolph, H., and Kohlbacher, O. Using N-terminal targeting sequences, amino acid composition, and sequence motifs for predicting protein subcellular localization. In *Lecture Notes in Informatics, German conference on Bioinformatics 2005* (2005), A. Torda, S. Kurtz, and M. Rarey, Eds., Gesellchaft für Informatik (GI), pp. 45–59.

[118] Holzhutter, H. G., Frommel, C., and Kloetzel, P. M. A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20 S proteasome. *J. Mol. Biol. 286*, 4 (1999), 1251–1265.

[119] Honeyman, M. C., Brusic, V., L, S. N., and Harrison, L. C. Neural network-based prediction of candidate T-cell epitopes. *Nature Biotechnology 16* (1998), 966–969.

[120] Hosken, N. A., and Bevan, M. J. Defective presentation of endogenous antigen by a cell line expressing class I molecules. *Science 248*, 4953 (Apr 1990), 367–370.

[121] Hsu, F. J., Benike, C., Fagnoni, F., Liles, T. M., Czerwinski, D., Taidi, B., Engleman, E. G., and Levy, R. Vaccination of patients with B-cell lymphoma using autologous antigen-pulsed dendritic cells. *Nat Med 2*, 1 (Jan 1996), 52–58. Case Reports.

[122] Hua, S., and Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics. 17*, 8 (2001), 721–728.

[123] Huang, J., El-Gamil, M., Dudley, M. E., Li, Y. F., Rosenberg, S. A., and Robbins, P. F. T cells associated with tumor regression recognize frameshifted products of the CDKN2A tumor suppressor gene locus and a mutated HLA class I gene product. *J Immunol 172*, 10 (May 2004), 6057–6064.

[124] Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., and Clamp, M. The ensembl genome database project. *Nucleic Acids Research 30*, 1 (2002), 38–41.

[125] Huczko, E. L., Bodnar, W. M., Benjamin, D., Sakaguchi, K., Zhu, N. Z., Shabanowitz, J., Henderson, R. A., Appella, E., Hunt, D. F., and Engelhard, V. H. Characteristics of endogenous peptides eluted from the class I MHC molecule HLA-B7 determined by mass spectrometry and computer modeling. *J Immunol 151*, 5 (Sep 1993), 2572–2587.

[126] Huston, D. P. The biology of the immune system. *JAMA 278*, 22 (Dec 1997), 1804–1814.

[127] Ishikawa, K., Nakamura, H., Morikawa, K., and Kanaya, S. Stabilization of Escherichia coli ribonuclease HI by cavity-filling mutations within a hydrophobic core. *Biochemistry 32*, 24 (Jun 1993), 6171–6178.

[128] Itano, A. A., and Jenkins, M. K. Antigen presentation to naive CD4 T cells in the lymph node. *Nat Immunol 4*, 8 (Aug 2003), 733–739.

[129] Itano, A. A., McSorley, S. J., Reinhardt, R. L., Ehst, B. D., Ingulli, E., Rudensky, A. Y., and Jenkins, M. K. Distinct dendritic cell populations sequentially present antigen to CD4 T cells and stimulate different aspects of cell-mediated immunity. *Immunity 19*, 1 (Jul 2003), 47–57.

[130] Jager, E., Chen, Y. T., Drijfhout, J. W., Karbach, J., Ringhoffer, M., Jager, D., Arand, M., Wada, H., Noguchi, Y., Stockert, E., Old, L. J., and Knuth, A. Simultaneous humoral and cellular immune response against cancer-testis antigen NY-ESO-1: definition of human histocompatibility leukocyte antigen (HLA)-A2-binding peptide epitopes. *J Exp Med 187*, 2 (Jan 1998), 265–270.

[131] Jager, E., Karbach, J., Gnjatic, S., Jager, D., Maeurer, M., Atmaca, A., Arand, M., Skipper, J., Stockert, E., Chen, Y.-T., Old, L. J., and Knuth, A. Identification of a naturally processed NY-ESO-1 peptide recognized by CD8+ T cells in the context of HLA-B51. *Cancer Immun 2* (Sep 2002), 12.

[132] Janeway, C. A., Travers, P., and Walport. *Immunobiology: the immune system in health and disease 4th ed.* Elsevier Science, London, UK, 1999.

[133] Janin, J., and Wodak, S. Conformation of amino acid side-chains in proteins. *J Mol Biol 125*, 3 (Nov 1978), 357–386.

[134] Jensen, L., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H., Rapacki, K., Workman, C., Andersen, C., Knudsen, S., Krogh, A., Valencia, A., and Brunak, S. Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol. 319*, 5 (2002), 1257–1265.

[135] Jensen, L., Gupta, R., Staerfeldt, H., and Brunak, S. Prediction of human protein function according to Gene Ontology categories. *Bioinformatics 19*, 5 (2003), 635–642.

[136] Joachims, T. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning.*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT Press, Cambridge Massachusetts, London, England, 1998.

[137] Jung, S., Unutmaz, D., Wong, P., Sano, G.-I., De los Santos, K., Sparwasser, T., Wu, S., Vuthoori, S., Ko, K., Zavala, F., Pamer, E. G., Littman, D. R., and Lang, R. A. In vivo depletion of CD11c(+) dendritic cells abrogates priming of CD8(+) T cells by exogenous cell-associated antigens. *Immunity 17*, 2 (Aug 2002), 211–220.

[138] Kast, W. M., Brandt, R. M., Sidney, J., Drijfhout, J. W., Kubo, R. T., Grey, H. M., Melief, C. J., and Sette, A. Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins. *J Immunol 152*, 8 (Apr 1994), 3904–3912.

[139] Kawashima, S., Ogata, H., and Kanehisa, M. AAindex: Amino Acid Index Database. *Nucleic Acids Res 27*, 1 (1999), 368–369.

[140] Kelly, A., Powis, S. H., Glynne, R., Radley, E., Beck, S., and Trowsdale, J. Second proteasome-related gene in the human MHC class II region. *Nature 353*, 6345 (Oct 1991), 667–668.

[141] Kesmir, C., Nussbaum, A. K., Schild, H., Detours, V., and Brunak, S. Prediction of proteasome cleavage motifs by neural networks. *Prot. Eng. 15*, 4 (2002), 287–296.

[142] Kessler, J. H., Beekman, N. J., Bres-Vloemans, S. A., Verdijk, P., van Veelen, P. A., Kloosterman-Joosten, A. M., Vissers, D. C., ten Bosch, G. J., Kester, M. G., Sijts, A., Wouter Drijfhout, J., Ossendorp, F., Offringa, R., and Melief, C. J. Efficient identification of novel HLA-A(*)0201-presented cytotoxic T lymphocyte epitopes in the widely expressed tumor antigen PRAME by proteasome-mediated digestion analysis. *J Exp Med 193*, 1 (Jan 2001), 73–88.

[143] Kim, N. W., Piatyszek, M. A., Prowse, K. R., Harley, C. B., West, M. D., Ho, P. L., Coviello, G. M., Wright, W. E., Weinrich, S. L., and Shay, J. W. Specific association of human telomerase activity with immortal cells and cancer. *Science 266*, 5193 (Dec 1994), 2011–2015.

[144] Kisselev, A. F., Akopian, T. N., Woo, K. M., and Goldberg, A. L. The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes. Implications for understanding the degradative mechanism and antigen presentation. *J Biol Chem 274*, 6 (Feb 1999), 3363–3371.

[145] Kjer-Nielsen, L., Clements, C. S., Brooks, A. G., Purcell, A. W., Fontes, M. R., McCluskey, J., and Rossjohn, J. The structure of HLA-B8 complexed to an immunodominant viral determinant: peptide-induced conformational changes and a mode of MHC class I dimerization. *J Immunol 169*, 9 (Nov 2002), 5153–5160.

[146] Kleijmeer, M. J., Morkowski, S., Griffith, J. M., Rudensky, A. Y., and Geuze, H. J. Major histocompatibility complex class II compartments in human and mouse B lymphoblasts represent conventional endocytic compartments. *J Cell Biol 139*, 3 (Nov 1997), 639–649.

[147] Kondo, A., Sidney, J., Southwood, S., del Guercio, M. F., Appella, E., Sakamoto, H., Celis, E., Grey, H. M., Chesnut, R. W., Kubo, R. T., and Sette, A. Prominent roles of secondary anchor residues in peptide binding to HLA-A24 human class I molecules. *J Immunol 155*, 9 (Nov 1995), 4307–4312.

[148] Kondo, A., Sidney, J., Southwood, S., del Guercio, M. F., Appella, E., Sakamoto, H., Grey, H. M., Celis, E., Chesnut, R. W., Kubo, R. T., and Sette, A. Two distinct HLA-A*0101-specific submotifs illustrate alternative peptide binding modes. *Immunogenetics 45*, 4 (1997), 249–258.

[149] Koopmann, J. O., Albring, J., Huter, E., Bulbuc, N., Spee, P., Neefjes, J., Hammerling, G. J., and Momburg, F. Export of antigenic peptides from the endoplasmic reticulum intersects with retrograde protein translocation through the Sec61p channel. *Immunity 13*, 1 (Jul 2000), 117–127.

[150] Koopmann, J. O., Post, M., Neefjes, J. J., Hämmerling, G. J., and Momburg, F. Translocation of long peptides by transporter associated with antigen processing (TAP). *Eur. J. Immunol. 26*, 8 (1996), 1720–1728.

[151] Kropshofer, H., Vogt, A. B., Stern, L. J., and Hammerling, G. J. Self-release of CLIP in peptide loading of HLA-DR molecules. *Science 270*, 5240 (Nov 1995), 1357–1359.

[152] Kruger, T., Schoor, O., Lemmel, C., Kraemer, B., Reichle, C., Dengjel, J., Weinschenk, T., Muller, M., Hennenlotter, J., Stenzl, A., Rammensee, H.-G., and Stevanovic, S. Lessons to be learned from primary renal cell carcinomas: novel tumor antigens and HLA ligands for immunotherapy. *Cancer Immunol Immunother. 54* (Sep 2005), 826–836.

[153] Kubo, R. T., Sette, A., Grey, H. M., Appella, E., Sakaguchi, K., Zhu, N. Z., Arnott, D., Sherman, N., Shabanowitz, J., and Michel, H. Definition of specific peptide motifs for four major HLA-A alleles. *J Immunol 152*, 8 (Apr 1994), 3913–3924.

[154] Kuttler, C., Nussbaum, A. K., Dick, T. P., Rammensee, H.-G., Schild, H., and Hadeler, K.-P. An algorithm for the prediction of proteasomal cleavages. *J. Mol. Biol. 298* (2000), 417–429.

[155] Lanzavecchia, A. Receptor-mediated antigen uptake and its effect on antigen presentation to class II-restricted T lymphocytes. *Annu Rev Immunol 8* (1990), 773–793.

[156] Larsen, M. V., Lundegaard, C., Lamberth, K., Buus, S., Brunak, S., Lund, O., and Nielsen, M. An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur J Immunol 35*, 8 (2005), 2295–2303.

[157] Lautscham, G., Rickinson, A., and Blake, N. TAP-independent antigen presentation on MHC class I molecules: lessons from Epstein-Barr virus. *Microbes Infect 5*, 4 (Apr 2003), 291–299.

[158] Leach, M. R., Cohen-Doyle, M. F., Thomas, D. Y., and Williams, D. B. Localization of the lectin, ERp57 binding, and polypeptide binding sites of calnexin and calreticulin. *J Biol Chem 277*, 33 (Aug 2002), 29686–29697.

[159] Lee, K. H., Wucherpfennig, K. W., and Wiley, D. C. Structure of a human insulin peptide-HLA-DQ8 complex and susceptibility to type 1 diabetes. *Nat Immunol 2*, 6 (Jun 2001), 501–507.

[160] Lippolis, J. D., White, F. M., Marto, J. A., Luckey, C. J., Bullock, T. N. J., Shabanowitz, J., Hunt, D. F., and Engelhard, V. H. Analysis of MHC class II antigen processing by quantitation of peptides that constitute nested sets. *J Immunol 169*, 9 (Nov 2002), 5089–5097.

[161] Logean, A., and Rognan, D. Recovery of known T-cell epitopes by computational scanning of a viral genome. *J Comput Aided Mol Des 16*, 4 (Apr 2002), 229–243.

[162] Lucchiari-Hartz, M., van Endert, P. M., Lauvau, G., Maier, R., Meyerhans, A., Mann, D., Eichmann, K., and Niedermann, G. Cytotoxic T lymphocyte epitopes of HIV-1 Nef: Generation of multiple definitive major histocompatibility complex class I ligands by proteasomes. *J Exp Med 191*, 2 (Jan 2000), 239–252.

[163] Luckey, C. J., Marto, J. A., Partridge, M., Hall, E., White, F. M., Lippolis, J. D., Shabanowitz, J., Hunt, D. F., and Engelhard, V. H. Differences in the expression of human class I

MHC alleles and their associated peptides in the presence of proteasome inhibitors. *J Immunol 167*, 3 (Aug 2001), 1212–1221.

[164] Lundstrom, J., Rychlewski, L., Bujnicki, J., and Elofsson, A. Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci 10*, 11 (Nov 2001), 2354–2362.

[165] Madden, D. R., Garboczi, D. N., and Wiley, D. C. The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2. *Cell 75*, 4 (Nov 1993), 693–708.

[166] Mallios, R. R. Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm. *Bioinformatics 17*, 10 (Oct 2001), 942–948.

[167] Mamitsuka, H. Predicting peptides that bind to MHC molecules using supervised learning of hidden markov models. *Proteins: Structure, Function and Genetics 33* (1998), 460–474.

[168] Manici, S., Sturniolo, T., Imro, M. A., Hammer, J., Sinigaglia, F., Noppen, C., Spagnoli, G., Mazzi, B., Bellone, M., Dellabona, P., and Protti, M. P. Melanoma cells present a MAGE-3 epitope to CD4(+) cytotoxic T cells in association with histocompatibility leukocyte antigen DR11. *J Exp Med 189*, 5 (Mar 1999), 871–876.

[169] Margalit, H., Spouge, J. L., Cornette, J. L., Cease, K. B., Delisi, C., and Berzofsky, J. A. Prediction of immunodominant helper T cell antigenic sites from the primary sequence. *J Immunol 138*, 7 (1987), 2213–2229.

[170] Martinez, C. K., and Monaco, J. J. Homology of proteasome subunits to a major histocompatibility complex-linked LMP gene. *Nature 353*, 6345 (Oct 1991), 664–667.

[171] Martinez, J., Prevot, S., Nordlinger, B., Nguyen, T., Lacarriere, Y., Munier, A., Lascu, I., Vaillant, J., Capeau, J., and Lacombe, M. Overexpression of nm23-H1 and nm23-H2 genes in colorectal carcinomas and loss of nm23-H1 expression in advanced tumour stages. *Gut 35*, 5 (1995), 712–720.

[172] Matthews, B. W. Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta. 405* (1975), 442–451.

[173] McSparron, H., Blythe, M. J., Zygouri, C., Doytchinova, I. A., and Flower, D. R. JenPep: a novel computational information resource for immunobiology and vaccinology. *J Chem Inf Comput Sci 43*, 4 (Jul 2003), 1276–1287.

[174] Meese, E., and Comtesse, N. Cancer genetics and tumor antigens: time for a combined view? *Genes Chromosomes Cancer 33*, 2 (2002), 107–113.

[175] Mellman, I., Turley, S. J., and Steinman, R. M. Antigen processing for amateurs and professionals. *Trends Cell Biol 8*, 6 (Jun 1998), 231–237.

[176] Mempel, T. R., Henrickson, S. E., and Von Andrian, U. H. T-cell priming by dendritic cells in lymph nodes occurs in three distinct phases. *Nature 427*, 6970 (Jan 2004), 154–159.

[177] Menzies, T., and Hu, Y. Data mining for very busy people. *IEEE Computer 36*, 11 (2003), 22–29.

[178] Minev, B., Hipp, J., Firat, H., Schmidt, J. D., Langlade-Demoyen, P., and Zanetti, M. Cytotoxic T cell immunity against telomerase reverse transcriptase in humans. *Proc Natl Acad Sci U S A 97*, 9 (Apr 2000), 4796–4801.

[179] Mitelman F, J. B., and F, M. Mitelman Database of Chromosome Aberrations in Cancer. http://cgap.nci.nih.gov/Chromosomes/Mitleman, 2005.

[180] Mitrunen, K., and Hirvonen, A. Molecular epidemiology of sporadic breast cancer. The role of polymorphic genes involved in oestrogen biosynthesis and metabolism. *Mutat. Res. 544*, 1 (2003), 9–41.

[181] Moins-Teisserenc, H., Semana, G., Alizadeh, M., Loiseau, P., Bobrynina, V., Deschamps, I., Edan, G., Birebent, B., Genetet, B., and Sabouraud, O. TAP2 gene polymorphism contributes to genetic susceptibility to multiple sclerosis. *Hum Immunol 42*, 3 (Mar 1995), 195–202.

[182] Momburg, F., and Tan, P. Tapasin-the keystone of the loading complex optimizing peptide binding by MHC class I molecules in the endoplasmic reticulum. *Mol Immunol 39*, 3-4 (Oct 2002), 217–233.

[183] Monz, D., Munnia, A., Comtesse, N., Fischer, U., Steudel, W., Feiden, W., Glass, B., and EU, M. Novel tankyrase-related gene detected with meningioma-specific sera. *Clin. Cancer Res. 7*, 1 (2001), 113–119.

[184] Morel, S., Levy, F., Burlet-Schiltz, O., Brasseur, F., Probst-Kepper, M., Peitrequin, A. L., Monsarrat, B., Van Velthoven, R., Cerottini, J. C., Boon, T., Gairin, J. E., and Van den Eynde, B. J. Processing of some antigens by the standard proteasome but not by the immunoproteasome results in poor presentation by dendritic cells. *Immunity 12*, 1 (Jan 2000), 107–117.

[185] Nair, R., and Rost, B. Sequence conserved for subcellular localization. *Protein Science 11*, 12 (2002), 2836–2847.

[186] Nakai, K., and Kanehisa, M. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics. 14*, 4 (1992), 897–911.

[187] Neefjes, J. J., Momburg, F., and Hammerling, G. J. Selective and ATP-dependent translocation of peptides by the MHC-encoded transporter. *Science 261*, 5122 (Aug 1993), 769–771.

[188] Nestle, F. O., Alijagic, S., Gilliet, M., Sun, Y., Grabbe, S., Dummer, R., Burg, G., and Schadendorf, D. Vaccination of melanoma patients with peptide- or tumor lysate-pulsed dendritic cells. *Nat Med 4*, 3 (Mar 1998), 328–332. Clinical Trial.

[189] Niedermann, G., Grimm, R., Geier, E., Maurer, M., Realini, C., Gartmann, C., Soll, J., Omura, S., Rechsteiner, M. C., Baumeister, W., and Eichmann, K. Potential immunocompetence of proteolytic fragments produced by proteasomes before evolution of the vertebrate immune system. *J Exp Med 186*, 2 (Jul 1997), 209–220.

[190] Niedermann, G., King, G., Butz, S., Birsner, U., Grimm, R., Shabanowitz, J., Hunt, D. F., and Eichmann, K. The proteolytic fragments generated by vertebrate proteasomes: structural relationships to major histocompatibility complex class I binding peptides. *Proc Natl Acad Sci U S A 93*, 16 (Aug 1996), 8572–8577.

[191] Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S., Lamberth, K., Buus, S., Brunak, S., and Lund, O. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics 20*, 9 (Jun 2004), 1388–1397. Evaluation Studies.

[192] Nielsen, M., Lundegaard, C., Worning, P., Lauemoller, S. L., Lamberth, K., Buus, S., Brunak, S., and Lund, O. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci 12*, 5 (May 2003), 1007–1017.

[193] Nijenhuis, M., and Hammerling, G. J. Multiple regions of the transporter associated with antigen processing (TAP) contribute to its peptide binding site. *J Immunol 157*, 12 (Dec 1996), 5467–5477.

[194] Nishie, A., Masuda, K., Otsubo, M., Migita, T., Tsuneyoshi, M., Kohno, K., Shuin, T., Naito, S., Ono, M., and Kuwano, M. High expression of the Cap43 gene in infiltrating macrophages of human renal cell carcinomas. *Clin. Cancer Res. 7*, 7 (2001), 2145–2151.

[195] Nussbaum, A. K. *From the test tube to the World Wide Web*. PhD thesis, Eberhard-Karls-Universität Tübingen, 2001.

[196] Nussbaum, A. K., Dick, T. P., Keilholz, W., Schirle, M., Stevanovic, S., Dietz, K., Heinemeyer, W., Groll, M., Wolf, D. H., Huber, R., Rammensee, H. G., and Schild, H. Cleavage motifs of the yeast 20S proteasome beta subunits deduced from digests of enolase 1. *Proc Natl Acad Sci U S A 95*, 21 (Oct 1998), 12504–12509.

[197] Nussbaum, A. K., Kuttler, C., Hadeler, K.-P., Rammensee, H.-G., and Schild, H. PAProC: A prediction algorithm for proteasomal cleavages available on the WWW. *Immunogenetics 53* (2001), 87–94.

[198] Packer, B. R., Yeager, M., Staats, B., Welch, R., Crenshaw, A., Kiley, M., Eckert, A., Beerman, M., Miller, E., Bergen, A., Rothman, N., Strausberg, R., and Chanock, S. J. SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes. *Nucleic Acids Res 32*, Database issue (Jan 2004), 528–532.

[199] Palombella, V. J., Rando, O. J., Goldberg, A. L., and Maniatis, T. The ubiquitin-proteasome pathway is required for processing the NF-kappa B1 precursor protein and the activation of NF-kappa B. *Cell 78*, 5 (Sep 1994), 773–785.

[200] Pamer, E. G., Harty, J. T., and Bevan, M. J. Precise prediction of a dominant class I MHC-restricted epitope of Listeria monocytogenes. *Nature 353*, 6347 (1991), 852–855.

[201] Parker, K. C., Bednarek, M. A., and Coligan, J. E. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol. 152* (1994), 163–175.

[202] Paulus, H. Protein splicing and related forms of protein autoprocessing. *Annu. Rev. Biochem. 69* (2000), 447–496.

[203] Perler, F. B. InBase: the intein database. *Nucleic Acids Res. 30*, 1 (2002), 383–384.

[204] Peters, B., Bulik, S., Tampe, R., Van Endert, P., and Holzhutter, H. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J. Immunol. 171*, 4 (2003), 1741–1749.

[205] Peters, B., Tong, W., Sidney, J., Sette, A., and Weng, Z. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics 19*, 14 (Sep 2003), 1765–1772. Evaluation Studies.

[206] Peters, P. J., Raposo, G., Neefjes, J. J., Oorschot, V., Leijendekker, R. L., Geuze, H. J., and Ploegh, H. L. Major histocompatibility complex class II compartments in human B lymphoblastoid cells are distinct from early endosomes. *J Exp Med 182*, 2 (Aug 1995), 325–334.

[207] Petrovsky, N., and Brusic, V. Virtual models of the HLA class I antigen processing pathway. *Methods 34*, 4 (Dec 2004), 429–435.

[208] Powell, D. J. J., and Rosenberg, S. A. Phenotypic and functional maturation of tumor antigen-reactive CD8+ T lymphocytes in patients undergoing multiple course peptide vaccination. *J Immunother 27*, 1 (Jan 2004), 36–47. Evaluation Studies.

[209] Powis, S. H., Mockridge, I., Kelly, A., Kerr, L. A., Glynne, R., Gileadi, U., Beck, S., and Trowsdale, J. Polymorphism in a second ABC transporter gene located within the class II region of the human major histocompatibility complex. *Proc Natl Acad Sci U S A 89*, 4 (Feb 1992), 1463–1467.

[210] Powis, S. J., Deverson, E. V., Coadwell, W. J., Ciruela, A., Huskisson, N. S., Smith, H., Butcher, G. W., and Howard, J. C. Effect of polymorphism of an MHC-linked transporter on the peptides assembled in a class I molecule. *Nature 357*, 6375 (May 1992), 211–215.

[211] Princiotta, M. F., Finzi, D., Qian, S.-B., Gibbs, J., Schuchmann, S., Buttgereit, F., Bennink, J. R., and Yewdell, J. W. Quantitating protein synthesis, degradation, and endogenous antigen processing. *Immunity 18*, 3 (Mar 2003), 343–354.

[212] Pruitt, K., and Maglott, D. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res. 29*, 1 (2001), 137–140.

[213] Pruitt, K. D., Katz, K. S., Sicotte, H., and Maglott, D. R. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet 16*, 1 (Jan 2000), 44–47.

[214] Pruitt, K. D., Tatusova, T., and Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res 33*, Database issue (Jan 2005), 501–504.

[215] Quinlan, J. R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco, 1993.

[216] Quinlan, R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco, USA, 1993.

[217] Rammensee, H.-G., Bachman, J., Philipp, N., Emmerich, N., Bachor, O. A., and Stevanovic, S. SYFPEITHI: a database for MHC ligands and peptide motifs. *Immunogenetics 50* (1997), 213–219.

[218] Randolph, G. J. Dendritic cell migration to lymph nodes: cytokines, chemokines, and lipid mediators. *Semin Immunol 13*, 5 (Oct 2001), 267–274.

[219] Reche, P. A., Glutting, J.-P., and Reinherz, E. L. Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol 63*, 9 (Sep 2002), 701–709.

[220] Reche, P. A., Glutting, J.-P., Zhang, H., and Reinherz, E. L. Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics 56*, 6 (Sep 2004), 405–419.

[221] Reits, E., Griekspoor, A., Neijssen, J., Groothuis, T., Jalink, K., van Veelen, P., Janssen, H., Calafat, J., Drijfhout, J. W., and Neefjes, J. Peptide diffusion, protection, and degradation

in nuclear and cytoplasmic compartments before antigen presentation by MHC class I. *Immunity 18*, 1 (Jan 2003), 97–108.

[222] Reits, E., Neijssen, J., Herberts, C., Benckhuijsen, W., Janssen, L., Drijfhout, J. W., and Neefjes, J. A major role for TPPII in trimming proteasomal degradation products for MHC class I antigen presentation. *Immunity 20*, 4 (Apr 2004), 495–506.

[223] Ripberger, E., Linnebacher, M., Schwitalle, Y., Gebert, J., and von Knebel Doeberitz, M. Identification of an HLA-A0201-restricted CTL epitope generated by a tumor-specific frameshift mutation in a coding microsatellite of the OGT gene. *J Clin Immunol 23*, 5 (Sep 2003), 415–423.

[224] Ritz, U., Momburg, F., Pircher, H. P., Strand, D., Huber, C., and Seliger, B. Identification of sequences in the human peptide transporter subunit TAP1 required for transporter associated with antigen processing (TAP) function. *Int Immunol 13*, 1 (Jan 2001), 31–41.

[225] Robinson, J., Waller, M. J., Parham, P., de Groot, N., Bontrop, R., Kennedy, L. J., Stoehr, P., and Marsh, S. G. E. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res 31*, 1 (Jan 2003), 311–314.

[226] Roche, P. A. HLA-DM: an in vivo facilitator of MHC class II peptide loading. *Immunity 3*, 3 (Sep 1995), 259–262.

[227] Rock, K. L., and Goldberg, A. L. Degradation of cell proteins and the generation of MHC class I-presented peptides. *Annu. Rev. Immunol. 17* (1999), 739–779.

[228] Rock, K. L., Gramm, C., Rothstein, L., Clark, K., Stein, R., Dick, L., Hwang, D., and Goldberg, A. L. Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on mhc class i molecules. *Cell 78*, 5 (1994), 761–771.

[229] Roelse, J., Gromme, M., Momburg, F., Hammerling, G., and Neefjes, J. Trimming of TAP-translocated peptides in the endoplasmic reticulum and in the cytosol during recycling. *J Exp Med 180*, 5 (Nov 1994), 1591–1597.

[230] Rognan, D., Scapozza, L., Folkers, G., and Daser, A. Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes. *Biochemistry 33*, 38 (1994), 11476–11485.

[231] Roitt, I. M., Brostoff, J., and Male, D. K. *Immunology*. C.V. Mosby Company, St. Louis, USA, 1999.

[232] Ropke, M., Hald, J., Guldberg, P., Zeuthen, J., Norgaard, L., Fugger, L., Svejgaard, A., Van der Burg, S., Nijman, H. W., Melief, C. J., and Claesson, M. H. Spontaneous human squamous cell carcinomas are killed by a human cytotoxic T lymphocyte clone recognizing a wild-type p53-derived peptide. *Proc Natl Acad Sci U S A 93*, 25 (Dec 1996), 14704–14707.

[233] Rosenberg, S. A. Cancer vaccines based on the identification of genes encoding cancer regression antigens. *Immunol Today 18*, 4 (Apr 1997), 175–182.

[234] Rothbard, J. B., and Taylor, W. R. A sequence pattern common to T cell epitopes. *EMBO J 7*, 1 (1988), 93–100.

[235] Rotzschke, O., Falk, K., Deres, K., Schild, H., Norda, M., Metzger, J., Jung, G., and Rammensee, H. G. Isolation and analysis of naturally processed viral peptides as recognized by cytotoxic T cells. *Nature 348*, 6298 (Nov 1990), 252–254.

[236] Rumbaugh, J., Jacobson, I., and Booch, G. *The Unified Modelling Language Reference Manual.* Addison-Wesley, New York, USA, 1999.

[237] Ruppert, J., Sidney, J., Celis, E., Kubo, R. T., Grey, H. M., and Sette, A. Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell 74*, 5 (Sep 1993), 929–937.

[238] Rötzschke, O., Falk, K., Stevanovic, S., Jung, G., Walden, P., and Rammensee, H. G. Exact prediction of a natural T cell epitope. *Eur J Immunol 21*, 11 (1991), 2891–2894.

[239] Saxova, P., Buus, S., Brunak, S., and Kesmir, C. Predicting proteasomal cleavage sites: a comparison of available methods. *Int. Immunol. 15*, 7 (2003), 781–787.

[240] Schölkopf, B., and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[241] Schonbach, C., Koh, J. L., Sheng, X., Wong, L., and Brusic, V. FIMM, a database of functional molecular immunology. *Nucleic Acids Res 28*, 1 (Jan 2000), 222–224.

[242] Schonbach, C., Koh, J. L. Y., Flower, D. R., Wong, L., and Brusic, V. FIMM, a database of functional molecular immunology: update 2002. *Nucleic Acids Res 30*, 1 (Jan 2002), 226–229.

[243] Schubert, U., Anton, L. C., Gibbs, J., Norbury, C. C., Yewdell, J. W., and Bennink, J. R. Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature 404*, 6779 (Apr 2000), 770–774.

[244] Schueler-Furman, O., Altuvia, Y., and Sette, A. Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Science 9* (2000), 1838–1846.

[245] Schuler, M., Dönnes, P., Nastke, M.-D., Kohlbacher, O., Rammensee, H.-G., and Stevanovic, S. SNEP: SNP-derived Epitope Prediction program for minor H antigens. *Immunogenetics 57*, 11 (2005), 816–820.

[246] Schultz, E. S., Chapiro, J., Lurquin, C., Claverol, S., Burlet-Schiltz, O., Warnier, G., Russo, V., Morel, S., Levy, F., Boon, T., Van den Eynde, B. J., and van der Bruggen, P. The production of a new MAGE-3 peptide presented to cytolytic T lymphocytes by HLA-B40 requires the immunoproteasome. *J Exp Med 195*, 4 (Feb 2002), 391–399.

[247] Schumacher, T. N., Kantesaria, D. V., Heemels, M. T., Ashton-Rickardt, P. G., Shepherd, J. C., Fruh, K., Yang, Y., Peterson, P. A., Tonegawa, S., and Ploegh, H. L. Peptide length and sequence specificity of the mouse TAP1/TAP2 translocator. *J Exp Med 179*, 2 (Feb 1994), 533–540.

[248] Schwitalle, Y., Linnebacher, M., Ripberger, E., Gebert, J., and von Knebel Doeberitz, M. Immunogenic peptides generated by frameshift mutations in DNA mismatch repair-deficient cancer cells. *Cancer Immun 4* (Nov 2004), 14.

[249] Seemuller, E., Lupas, A., Stock, D., Lowe, J., Huber, R., and Baumeister, W. Proteasome from Thermoplasma acidophilum: a threonine protease. *Science 268*, 5210 (Apr 1995), 579–582. Comment.

[250] Seifert, U., Maranon, C., Shmueli, A., Desoutter, J.-F., Wesoloski, L., Janek, K., Henklein, P., Diescher, S., Andrieu, M., de la Salle, H., Weinschenk, T., Schild, H., Laderach, D., Galy, A., Haas, G., Kloetzel, P.-M., Reiss, Y., and Hosmalin, A. An essential role for tripeptidyl peptidase in the generation of an MHC class I epitope. *Nat Immunol 4*, 4 (Apr 2003), 375–379.

[251] Serwold, T., Gaw, S., and Shastri, N. ER aminopeptidases generate a unique pool of peptides for MHC class I molecules. *Nat. Immunol. 2*, 7 (2001), 644–651.

[252] Serwold, T., Gonzalez, F., Kim, J., Jacob, R., and Shastri, N. ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. *Nature 419*, 6906 (Oct 2002), 480–483.

[253] Sette, A., Buus, S., Appella, E., Smith, J. A., Chesnut, R., Miles, C., Colon, S. M., and Grey, H. M. Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc Natl Acad Sci U S A 86*, 9 (1989), 3296–3300.

[254] Shavlik, J., Hunter, L., and Searls, D. Introduction. *Mach. Learn. 21*, 1-2 (1995), 5–9.

[255] Shepherd, J. C., Schumacher, T. N., Ashton-Rickardt, P. G., Imaeda, S., Ploegh, H. L., Janeway, C. A. J., and Tonegawa, S. TAP1-dependent peptide translocation in vitro is ATP dependent and peptide selective. *Cell 74*, 3 (Aug 1993), 577–584.

[256] Shimizu, T., Abe, R., Nakamura, H., Ohkawara, A., Suzuki, M., and Nishihira, J. High expression of macrophage migration inhibitory factor in human melanoma cells and its role in tumor cell growth and angiogenesis. *Biochem. Biophys. Res. Commun. 264*, 3 (1999), 751–758.

[257] Sidney, J., Grey, H. M., Southwood, S., Celis, E., Wentworth, P. A., del Guercio, M. F., Kubo, R. T., Chesnut, R. W., and Sette, A. Definition of an HLA-A3-like supermotif demonstrates the overlapping peptide-binding repertoires of common HLA molecules. *Hum Immunol 45*, 2 (Feb 1996), 79–93.

[258] Sidney, J., Southwood, S., del Guercio, M. F., Grey, H. M., Chesnut, R. W., Kubo, R. T., and Sette, A. Specificity and degeneracy in peptide binding to HLA-B7-like class I molecules. *J Immunol 157*, 8 (Oct 1996), 3480–3490.

[259] Sidney, J., Southwood, S., Pasquetto, V., and Sette, A. Simultaneous prediction of binding capacity for multiple molecules of the HLA B44 supertype. *J Immunol 171*, 11 (Dec 2003), 5964–5974.

[260] Slager, E. H., van der Minne, C. E., Goudsmit, J., van Oers, J. M. M., Kostense, S., Havenga, M. J. E., Osanto, S., and Griffioen, M. Induction of CAMEL/NY-ESO-ORF2-specific CD8+ T cells upon stimulation with dendritic cells infected with a modified Ad5 vector expressing a chimeric Ad5/35 fiber. *Cancer Gene Ther 11*, 3 (Mar 2004), 227–236.

[261] Smith, K. J., Reid, S. W., Harlos, K., McMichael, A. J., Stuart, D. I., Bell, J. I., and Jones, E. Y. Bound water structure and polymorphic amino acids act together to allow the binding of different peptides to MHC class I HLA-B53. *Immunity 4*, 3 (Mar 1996), 215–228.

[262] Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L., and Liu, E. T. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A 100*, 18 (Sep 2003), 10393–10398.

[263] Soussi, T. p53 Antibodies in the sera of patients with various types of cancer: a review. *Cancer Res. 60*, 7 (2000), 1777–1788.

[264] Speiser, D. E., Pittet, M. J., Rimoldi, D., Guillaume, P., Luescher, I. F., Lienard, D., Lejeune, F., Cerottini, J.-C., and Romero, P. Evaluation of melanoma vaccines with molecularly defined antigens by ex vivo monitoring of tumor-specific T cells. *Semin Cancer Biol 13*, 6 (Dec 2003), 461–472.

[265] Spies, T., Cerundolo, V., Colonna, M., Cresswell, P., Townsend, A., and DeMars, R. Presentation of viral antigen by MHC class I molecules is dependent on a putative peptide transporter heterodimer. *Nature 355*, 6361 (Feb 1992), 644–646.

[266] Spies, T., and DeMars, R. Restored expression of major histocompatibility class I molecules by gene transfer of a putative peptide transporter. *Nature 351*, 6324 (May 1991), 323–324.

[267] Steinman, R. M., Mellman, I. S., Muller, W. A., and Cohn, Z. A. Endocytosis and the recycling of plasma membrane. *J Cell Biol 96*, 1 (Jan 1983), 1–27.

[268] Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics 16*, 1 (Jan 2000), 16–23. Historical Article.

[269] Struss, A., Romeike, B., Munnia, A., Nastainczyk, W., Steudel, W., Konig, J., Ohgaki, H., Feiden, W., Fischer, U., and Meese, E. PHF3-specific antibody responses in over 60% of patients with glioblastoma multiforme. *Oncogene 20*, 31 (2001), 4107–4114.

[270] Sturniolo, T., Bono, E., Ding, J., Raddrizzani, L., Tuereci, O., Sahin, U., Braxenthaler, M., Gallazzi, F., Protti, M. P., Sinigaglia, F., and Hammer, J. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol 17*, 6 (Jun 1999), 555–561.

[271] Supper, J. Predicting MHC class I binding peptides based on amino acid properties using decision trees and support vector machines. Tech. rep., 2005.

[272] Sykulev, Y., Cohen, R. J., and Eisen, H. N. The law of mass action governs antigen-stimulated cytolytic activity of CD8+ cytotoxic T lymphocytes. *Proc Natl Acad Sci U S A 92*, 26 (Dec 1995), 11990–11992.

[273] Sykulev, Y., Joo, M., Vturina, I., Tsomides, T. J., and Eisen, H. N. Evidence that a single peptide-MHC complex on a target cell can elicit a cytolytic T cell response. *Immunity 4*, 6 (Jun 1996), 565–571.

[274] Takano, K., and Yutani, K. A new scale for side-chain contribution to protein stability based on the empirical stability analysis of mutant proteins. *Protein Eng 14*, 8 (Aug 2001), 525–528.

[275] Tenzer, S., Peters, B., Bulik, S., Schoor, O., Lemmel, C., Schatz, M. M., Kloetzel, P.-M., Rammensee, H.-G., Schild, H., and Holzhutter, H.-G. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell Mol Life Sci 62*, 9 (2005), 1025–1037.

[276] Tenzer, S., Stoltze, L., Schönfisch, B., Dengjel, J., Müller, M., Stevanovic, S., Rammensee, H.-G., and Schild, H. Quantitative analysis of prion-protein degradation by constitutive and immuno-

20S proteasomes indicates differences correlated with disesase susceptibility. *J. Immunol. 172*, 2 (2004), 1083–1091.

[277] The cancer immunity peptide database. http://www.cancerimmunity.org/peptidedatabase/Tcellepitopes.htm.

[278] The cancer immunome database. http://www2.licr.org/CancerImmunomeDB/.

[279] The hiv database. http://hiv-web.lanl.gov/content/immunology/index.html.

[280] The serex database. http://www.licr.org/SEREX.html.

[281] Tomkinson, B. Tripeptidyl peptidases: enzymes that count. *Trends Biochem Sci 24*, 9 (Sep 1999), 355–359.

[282] Tong, J. C., Tan, T. W., and Ranganathan, S. Modeling the structure of bound peptide ligands to major histocompatibility complex. *Protein Sci 13*, 9 (Sep 2004), 2523–2532.

[283] Türeci, O., Sahin, U., and Pfreundschuh, M. Serological analysis of human tumor antigens: molecular definition and implications. *Molecular Medicine Today 3*, 8 (1997), 342–349.

[284] Uebel, S., Kraas, W., Kienle, K.-H., Wiesmüller, Jung, G., and Tampe, R. Recognition principle of the TAP transporter disclosed by combinatorial peptide libraries. *Proc. Natl. Acad. Sci. USA 94* (1997), 8976–8981.

[285] Uebel, S., and Tampe, R. Specificity of the proteasome and TAP transporter. *Curr. Opin. Immunol. 11* (1999).

[286] Valmori, D., Dutoit, V., Lienard, D., Rimoldi, D., Pittet, M. J., Champagne, P., Ellefsen, K., Sahin, U., Speiser, D., Lejeune, F., Cerottini, J. C., and Romero, P. Naturally occurring human lymphocyte antigen-A2 restricted CD8+ T-cell response to the cancer testis antigen NY-ESO-1 in melanoma patients. *Cancer Res 60*, 16 (Aug 2000), 4499–4506.

[287] van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med 347*, 25 (Dec 2002), 1999–2009. Evaluation Studies.

[288] Van den Eynde, B., Peeters, O., De Backer, O., Gaugler, B., Lucas, S., and Boon, T. A new family of genes coding for an antigen recognized by autologous cytolytic T lymphocytes on a human melanoma. *J Exp Med 182*, 3 (Sep 1995), 689–698.

[289] van der Bruggen, P., Traversari, C., Chomez, P., Lurquin, C., De Plaen, E., Van den Eynde, B., Knuth, A., and Boon, T. A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science 254*, 5038 (Dec 1991), 1643–1647.

[290] van Endert, P., Riganelli, D., Greco, G., Fleischhauer, K., Sidney, J., Sette, A., and JF, B. The peptide-binding motif for the human transporter associated with antigen processing. *J. Exp. Med. 182*, 6 (1995), 1883–1895.

[291] van Endert, P. M., Tampe, R., Meyer, T. H., Tisch, R., Bach, J. F., and McDevitt, H. A sequential model for peptide binding and transport by the transporters associated with antigen processing. *Immunity 1* (1994), 491–500.

[292] Van Kaer, L., Ashton-Rickardt, P. G., Eichelberger, M., Gaczynska, M., Nagashima, K., Rock, K. L., Goldberg, A. L., Doherty, P. C., and Tonegawa, S. Altered peptidase and viral-specific T cell response in LMP2 mutant mice. *Immunity 1*, 7 (Oct 1994), 533–541.

[293] Van Pel, A., van der Bruggen, P., Coulie, P. G., Brichard, V. G., Lethe, B., van den Eynde, B., Uyttenhove, C., Renauld, J. C., and Boon, T. Genes coding for tumor antigens recognized by cytolytic T lymphocytes. *Immunol Rev 145* (Jun 1995), 229–250.

[294] van Someren, H., Westerveld, A., Hagemeijer, A., Mees, J. R., Meera Khan, P., and Zaalberg, O. B. Human antigen and enzyme markers in man-Chinese hamster somatic cell hybrids: evidence for synteny between the HL-A, PGM3, ME1, and IPO-B loci. *Proc Natl Acad Sci U S A 71*, 3 (Mar 1974), 962–965.

[295] Vapnik, V. N. *The Nature of Statistical Learning Theory*. Wiley, New York, USA, 1999.

[296] Vigneron, N., Stroobant, V., Chapiro, J., Ooms, A., Degiovanni, G., Morel, S., van der Bruggen, P., Boon, T., and Van den Eynde, B. An antigenic peptide produced by peptide splicing in the proteasome. *Science 304*, 5670 (2004), 587–590.

[297] Vonderheide, R. H., Anderson, K. S., Hahn, W. C., Butler, M. O., Schultze, J. L., and Nadler, L. M. Characterization of HLA-A3-restricted cytotoxic T lymphocytes reactive against the widely expressed tumor antigen telomerase. *Clin Cancer Res 7*, 11 (Nov 2001), 3343–3348.

[298] Vonderheide, R. H., Hahn, W. C., Schultze, J. L., and Nadler, L. M. The telomerase catalytic subunit is a widely expressed tumor-associated antigen recognized by cytotoxic T lymphocytes. *Immunity 10*, 6 (Jun 1999), 673–679.

[299] Watts, C. Capture and processing of exogenous antigens for presentation on MHC molecules. *Annu Rev Immunol 15* (1997), 821–850.

[300] Wolf, P. R., and Ploegh, H. L. How MHC class II molecules acquire peptide cargo: biosynthesis and trafficking through the endocytic pathway. *Annu Rev Cell Dev Biol 11* (1995), 267–306.

[301] Xia, J.-X., Ikeda, M., and Shimizu, T. Conpred_elite: a highly reliable approach to transmembrane topology prediction. *Comput. Biol. Chem. 28* (2003), 51–60.

[302] Yewdell, J. W. The seven dirty little secrets of major histocompatibility complex class I antigen processing. *Immunol Rev 207* (Oct 2005), 8–18.

[303] Yewdell, J. W., and Bennink, J. R. Mechanisms of viral interference with MHC class I antigen processing and presentation. *Annu Rev Cell Dev Biol 15* (1999), 579–606.

[304] York, I. A., Mo, A. X. Y., Lemerise, K., Zeng, W., Shen, Y., Abraham, C. R., Saric, T., Goldberg, A. L., and Rock, K. L. The cytosolic endopeptidase, thimet oligopeptidase, destroys antigenic peptides and limits the extent of MHC class I antigen presentation. *Immunity 18*, 3 (Mar 2003), 429–440.

[305] Yu, K., Petrovsky, N., Schonbach, C., Koh, J. Y. L., and Brusic, V. Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol Med 8*, 3 (Mar 2002), 137–148.

[306] Yusim, K., Richardson, R., Tao, N., Szinger, J., Funkhouser, R., Korber, B., and Kuiken, C. The Los Alamos Hepatitis C Immunology Database. *Applied Bioinformatics* (2005 (in press)).

[307] Zinkernagel, R. M., and Doherty, P. C. H-2 compatability requirement for T-cell-mediated lysis of target cells infected with lymphocytic choriomeningitis virus. Different cytotoxic T-cell specificities are associated with structures coded for in H-2K or H-2D;. *J Exp Med 141*, 6 (Jun 1975), 1427–1436.

A. Abbreviations

APC	antigen presenting cell
ANN	artificial neural network
β_2m	β_2 microglobulin
DC	dendritic cell
DriP	defective ribosomal produce
DT	decision tree
ER	endoplasmic reticulum
HLA	human leukocyte antigen
MCC	Matthews correlation coefficient
MHC	major histocompatibility complex
PDB	Protein Data Bank
RCC	renal cell carcinoma
SE	sensitivity
SP	specificity
SVM	support vector machine
TAA	tumor-associated antigen
Tc	cytotoxic T cell
TCR	T-cell receptor
Th	helper T cell
TSA	tumor-specific antigen

B. Curriculum Vitae

Pierre Robert Dönnes

Heinrichsweg 7

D-720 74 Tübingen, Germany

doennes@informatik.uni-tuebingen.de

Education

2002-2006	PhD studies in Computer Science / Bioinformatics, University of Tübingen and Saarland University, Germany
1997-2001	Master of Science Program, Biotechnology, University of Linköping, Sweden.
1999-2000	Studies at the University of Salford, Manchester, UK. Main subjects: genetics, physiology, and immunology.
1993-1996	High School, Natural Science Program, Ållebergsgymnasiet Falköping, Sweden

Work experience

2003-present	Research assistant at the Div. for Simulation of Biological System, University of Tübingen, Tübingen, Germany
2002-2003	Research assistant at the Center for Bioinformatics, Saarbrücken, Germany
2001	Project employee at the Stockholm Bioinformatics Center, Stockholm, Sweden

Publications

1. Dönnes, P, and Elofsson, A (2002). Prediction of MHC class I binding peptides, using SVMHC, *BMC Bioinformatics*, 3(25).

2. Dönnes, P, and Höglund, A (2004). Predicting Protein Subcellular Localization: Past, Present, and Future, *Genomics Proteomics Bioinformatics* 2(4):209–215.
3. **Dönnes, P, Höglund, A**, Sturm, M, Comtesse, N, Backes, C, Meese, E, Kohlbacher, O, and Lenhof, H (2004). Integrative analysis of cancer-related data using CAP, *FASEB Journal* 18(12):1465-1467.
4. Höglund, A, Dönnes, P, Blum, T, Adolph, H, and Kohlbacher, O (2005). Using N-terminal targeting sequences, amino acid composition, and sequence motifs for predicting protein subcellular localization, *In: Proceedings of the German Conference on Bioinformatics (GCB 2005), GI,*: 45–59.
5. **Schuler, MM, Dönnes, P**, Nastke, M, Kohlbacher, O, Rammensee, H, and Stevanovic, S (2005). SNEP: SNP-derived Epitope Prediction program for minor H antigens, *Immunogenetics* 57(11):816-820
6. Dönnes, P, and Kohlbacher, O (2005). Integrated modelling of the major events in the MHC class I antigen processing pathway, *Protein Sci.* 14(8):2132–2140.
7. **Höglund, A, Dönnes**, P, Adolph, H, and Kohlbacher, O (2005). From prediction of subcellular localization to functional classification: Discrimination of DNA-packing and other nuclear proteins, *Online Journal of Bioinformatics* 6(1):51-64.
8. **Supper, J, Dönnes, P**, and Kohlbacher, O (2005). Analysis of MHC-Peptide Binding Using Amino Acid Property-Based Decision Rules, *Springer Lecture Notes in Computer Science (LNCS)* 3686:446-453
9. Höglund, A, Blum, T, Brady, S, Dönnes, P, Miguel, JS, Rocheford, M, Kohlbacher, O, and Shatkay, H (2006). Significantly improved prediction of subcellular localization by integrating text and protein sequence data, *In: Proceedings of the Pacific Symposium on Biocomputing (PSB 2006)*
10. Höglund, A, Dönnes, P, Blum, T, Adolph, H, and Kohlbacher, O (2006). MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs, and amino acid composition, *Bioinformatics, in press.*
11. Dönnes, P and Kohlbacher, O (2006). SVMHC: a server for prediction of MHC-binding peptides *Nucleic Acids Research, in press.*

Index