Aus dem Institut für Biochemie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Prediction of cleavage fragments generated by the proteasome

zur Erlangung des akademischen Grades
Doctor medicinae (Dr. med.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Justus Richard Pett

aus Berlin

Datum der Promotion: 09.12.2016

Table of Contents

Abstract

Introduction

The proteasome is a vital cell organelle, which generates the majority of antigenic peptides within the MHC I (major histocompatibility complex) pathway. Accordingly, a deeper understanding of its properties and behavior may lead to new developments in cancer therapy, vaccine design or the treatment of viral infections. The proteasome inhibitor bortezomib for example was one of the first FDA-approved drugs directly targeting the proteasome and is successfully used in the treatment of relapsed myeloma. Even though the proteasome's structure has been examined in detail, the factors and conditions relevant for its cleavage behavior still remain unclear for the most part.

Methods

This work aims to deepen the understanding of the proteasome's cleavage behavior using a machine learning approach: data of in vitro experiments gathered at the institute of biochemistry of the Charité Berlin was used as training data in order to learn a model classifying proteasomal cleavage products using a decision tree algorithm. The main advantage of the decision tree algorithm compared to other approaches like neural networks or support vector machines is the comprehensibility of its model: The decisions that make up the learned classification can be displayed in form of a tree or simple if-then-rules with good human readability. This way a model was created, which not only allows the prediction of fragments created by the proteasome but also makes it possible to understand, which properties of the substrate are important for the model's classification.

Results

28 different decision trees were created using various sets of training data as well as different sets of substrate attributes. Cross validation showed that the trees classified the training data correctly. The possibilities for validation with in vivo data are limited, since only data of CTL epitopes, which are no direct products of a proteasome's digestion

process, is available. Still validation of the decision trees with CTL epitope data gave plausible results.

No property or class of properties showed to be distinctly relevant for the proteasome's cleavage behavior. The different decision trees classified the data using a variety of different properties.

Abstrakt

Einleitung

Das Proteasom ist ein lebenswichtiges Zell-Organell, das die Mehrheit anitgener Peptide im MHC I (major histocompatibility complex) Pathway produziert. Dementsprechend bietet ein genaueres Verständnis seiner Eigenschaften und seines Verhaltens das Potenzial für neue Entwicklungen im Bereich der Therapie maligner und viraler Erkrankungen, sowie beim Design neuer Vakzine. Der Proteasom-Inhibitor Bortezomib war beispielsweise das erste zugelassene Medikament mit dem Proteasom als direkter Zielstruktur und wird erfolgreich in der Therapie des multiplen Myeloms angewandt. Auch wenn die Struktur des Proteasoms bereits ausführlich untersucht wurde, bleiben die Faktoren und Bedingungen, die das Schnittverhalten des Proteasoms beeinflussen, nach wie vor weithin unbekannt.

Methodik

Das Ziel dieser Arbeit besteht in der Untersuchung des Schnittverhaltens des Proteasoms mit Hilfe von Methoden des Machine Learnings: Daten von in vitro Experimenten, die am Institut für Biochemie der Charité durchgeführt wurden, dienten als Trainingsdaten, um ein Modell zur Klassifikation von Schnittprodukten des Proteasoms zu generieren. Hierfür kam ein Decision Tree (Entscheidungsbaum) Algorithmus zum Einsatz. Im Gegensatz zu anderen Verfahren wie neuronalen Netzen oder Support Vector Machines bieten Decision Trees den Vorteil, dass die Entscheidungen, die zur Klassifikation im Modell führen, in Form von Entscheidungsbäumen oder einfachen Wenn-Dann-Regeln dargestellt werden können. So wurde ein Modell erstellt, das nicht nur die Vorhersage von Schnittprodukten des Proteasoms erlaubt, sondern es auch ermöglicht, die für die Klassifikation relevanten Eigenschaften des Substrats zu identifizieren.

Ergebnisse

28 verschiedene Decision Trees wurden mit unterschiedlichen Trainings-Datensätzen und verschiedenen Sätzen von möglichen Attributen erzeugt. Mittels Cross Validation

wurde überprüft, dass die Trainingsdaten durch die generierten Bäume korrekt klassifiziert wurden. Eine Validierung mit in vitro Daten ist hingegen nur eingeschränkt möglich, da lediglich Daten zu T-Zell-Epitopen verfügbar sind. Dabei handelt es sich jedoch nicht um direkte Verdauprodukte des Proteasoms. Dennoch zeigte die Validierung der Decision Trees mit T-Zell-Epitopdaten plausible Ergebnisse.

Keine Eigenschaft oder Klasse von Eigenschaften des Substrats zeigte eine hervorstechende Bedeutung bei der Klassifikation von Schnittfragmenten. Die verschiedenen Decision Trees verwendeten eine Vielzahl unterschiedlicher Substrateigenschaften.

Introduction

The proteasome is an important cell organelle, which plays a vital role in a variety of cell functions, including the generation of the majority of antigenic peptides within the MHC I (major histocompatibility complex) pathway. Various types of new medical therapies like epitope-based peptide vaccines and antiviral- or oncological drugs target the MHC I pathway or the proteasome itself. Even though these treatments show promising results and may improve the treatment of a wide spectrum of diseases significantly in the future, the majority of these new drugs have not been officially approved yet and are still being evaluated in clinical trials. A lot of questions regarding the processing of peptides within the MHC I pathway still remain open. Gaining a better understanding of the proteasome as an important part of the MHC I pathway can therefore proof valuable for the design of these new treatments mentioned before.

In the following, the MHC I pathway will be introduced in more detail before the most important therapies targeting the proteasome and the MHC I pathway are discussed. Afterwards, the properties of the proteasome are explained in more detail and an overview over approaches to proteasomal cleavage site and fragment prediction is given.

MHC I pathway

The MHC I pathway enables cells to present fragments of intracellular proteins to cytotoxic T cells and includes all steps from generation of protein fragments to their presentation on the cell surface. The process is shown in Figure 1. Cytosolic proteins are marked for degradation by the ubiquitin-system and then processed by the proteasome generating short oligopeptides. Defective ribosomal products (DRiPs) are an important source during this process: DRiPs are newly synthesized polypeptides which are degraded again within minutes due to errors in translation or defects in post-translational folding (Schubert, Antón, et al. 2000).

While most of the fragments are cleaved even further into single amino acids and then reused for the assembly of new proteins, a part is transferred to the endoplasmatic reticulum by TAP (transporter associated with antigen presentation) and binds to MHC molecules. In every human being, MHC I and MHC II molecules are encoded by three gene locations each, which due to the diploid chromosome set results in 6 different MHC molecules per class. Each MHC molecule binds a unique set of peptides with an average length of 8-10 amino acids (H.-G. Rammensee, Friede, and Stevanović 1995). Multiple findings suggest that the proteasome is responsible for generating precursor peptides of 3–22 residues which contain the C-terminus of the final MHC I ligand while their N-terminus is trimmed by cellular aminopeptidases (Peter M Kloetzel 2004; Craiu et al. 1997; Mo et al. 1999; Cascio et al. 2001; Serwold and Shastri 1999). The resulting MHC complex is transported to the cell surface by the Golgi apparatus and presented to cytotoxic T1-Lymphocytes.

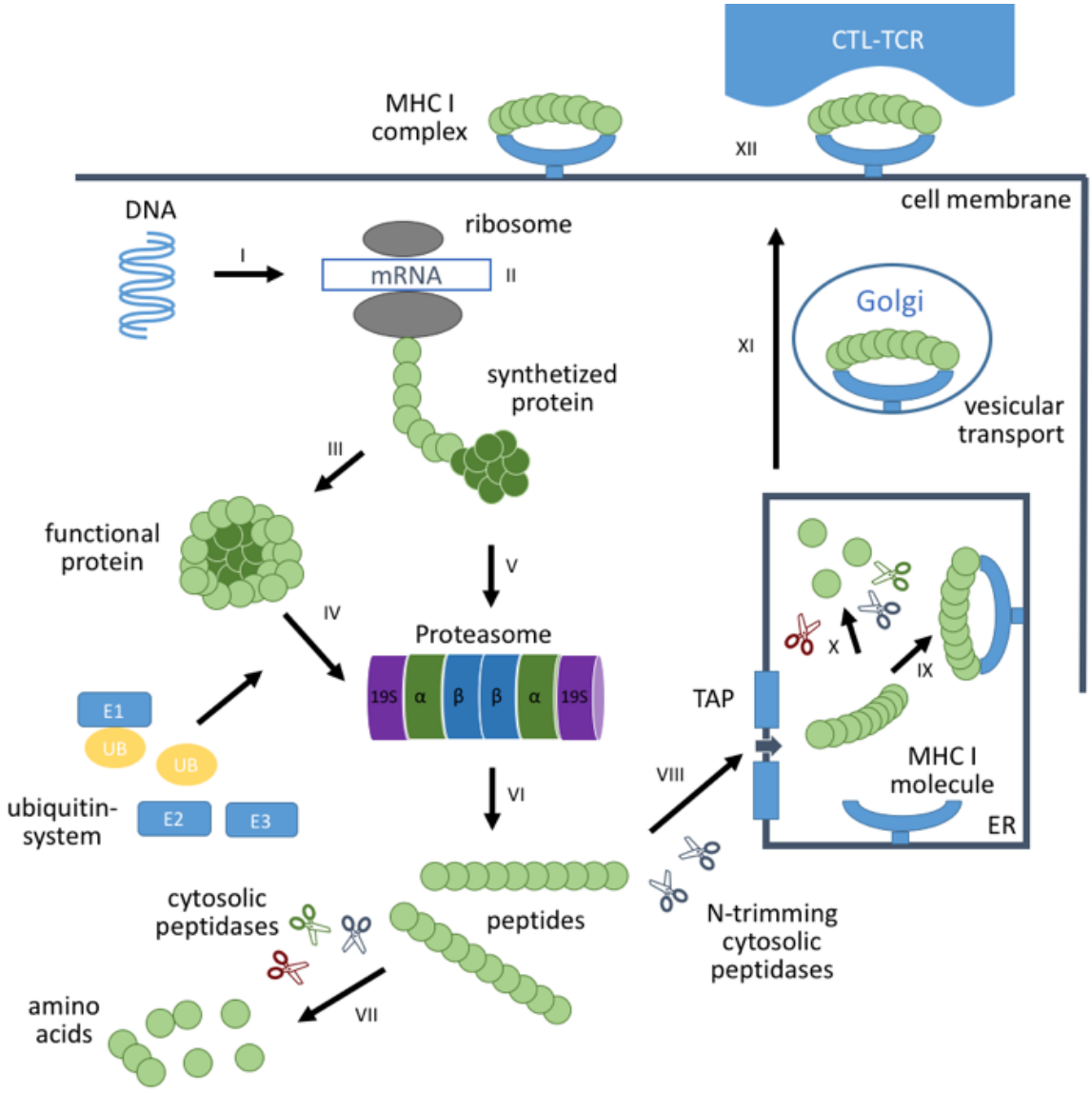


Figure 1: Overview of the MHC I pathway. I: Transcription of DNA to mRNA, splicing of mRNA and additional interactions on mRNA level. II: Translation of mRNA by the ribosome. III: Folding and post-translational modifications of the newly synthesized protein. IV: Degradation of functional proteins, in part assisted by the ubiquitin-system. V: Degradation of defective proteins (DRiPs, also assisted by the ubiquitin-system) or proteins in creation by the proteasome. VI: Creation of peptide fragments by the proteasome. VII: Hydrolytic cleavage of peptides into amino acids by cytosolic peptidases. VIII: Binding of cytosolic peptides to TAP and transport into the ER. IX: Binding of endoplasmatic peptides to MHC I molecules and building of MHC I complexes. X: Hydrolysis of peptides by endoplasmatic amino-peptidases, export of fragments into the cytosol. XI: Vesicular transport of MHC I complexes to the cell surface by the Golgi-apparatus. XII: Presentation of MHC I complexes on the cell surface and binding of cytotoxic T-lymphocytes. Adapted from (Bulik 2011)

Therapies targeting the proteasome and the MHC I pathway

The proteasome and the MHC-I pathway play an important role in various new treatment strategies, which are introduced in the following.

Proteasome inhibitors in cancer therapy

The proteasome itself is used as a target in cancer therapy. Proteasome inhibitors like bortezomib were first identified as drug candidates after studies showed that they induce apoptosis in leukemic cell lines (Shinohara et al. 1996; Imajohohmi et al. 1995). This effect was even observed in chemotherapy-resistant and radiation-resistant chronic lymphocytic leukemia cells. In addition, proteasome inhibitors have been shown to induce apoptosis preferentially in transformed cells (Delic et al. 1998).

Multiple mechanisms are responsible for the effect of proteasome inhibitors: they repress nuclear factor-κB (NF-κB), which plays an important role in angiogenesis, cell invasion and oncogenesis (R. Z. Orlowski and Baldwin 2002). Furthermore, proteasome inhibitors induce cell cycle arrest by interfering with timely degradation of cyclins and other cell cycle regulatory proteins. They are also able to stabilize proapoptotic proteins like p53 and Bax, while reducing levels of other antiapoptotic proteins like Bcl-2 (Rajkumar et al. 2005).

The proteasome inhibitor bortezomib was approved for treatment of relapsed/refractory myeloma in patients who have progressed past at least one prior regimen after a phase III study showed a better response rate in comparison with dexamethasone (Richardson et al. 2005; Richardson et al. 2007). It is also used in combination with various chemotherapeutics like carboplatin, docetaxel or melphalan in order to induce chemo sensitivity or overcome chemo resistance (Aghajanian et al. 2005; Messersmith et al. 2006; Berenson et al. 2006). Carfilzomib, a next generation proteasome inhibitor that unlike bortezomib binds to the proteasome irreversibly, was approved by the FDA in 2012 for patients with multiple myeloma who have received two prior therapies, including treatment with bortezomib, after a multicenter, open-label trial had shown an improved progression-free survival (Siegel et al. 2012).

Despite the proteasome's vital role in cellular homeostasis the toxicity of proteasome inhibitors proved to be manageable. Adverse events documented in the clinical trials include anemia, anorexia, constipation, dehydration, diarrhea, neutropenia, thrombocytopenia and neuropathy and have been shown to be transient and reversible. A better understanding of the underlying mechanisms might help to handle these effects, predict the efficacy or toxicity of the treatment and overcome resistance against proteasome inhibitors, which has especially been observed in solid tumors.

Viral infections

Many viruses have been reported to use different strategies in order to use the MHC I pathway for their own benefits.

One example is viral immune evasion: Viruses have developed different strategies for down-regulation of MHC I molecules in order to reduce antigen presentation and therefore survive inside cells causing latent or chronic infections (Furman, Ploegh, and others 2002). The human cytomegalovirus for example produces the protein US2, which induces dislocation of MHC I molecules from the endoplasmic reticulum to the cytoplasm, where they are polyubiquitinated and rapidly degraded by the proteasome (Shamu et al. 2001; Kikkert et al. 2001). The Epstein-Barr virus nuclear antigen 1 (EBNA1) contains Gly-Ala repeats that prevent viral protein degradation by the proteasome (Levitskaya et al. 1997).

Another viral abuse mechanism, used by some enveloped RNA-viruses, is related to viral progeny release and viral membrane envelopment (budding). Multiple studies were able to show that proteasomal inhibition reduces viral progeny release and viral infectivity (Patnaik, Chau, and Wills 2000; Strack et al. 2000; Schubert, Ott, et al. 2000).

Apoptosis is another process with involvement of the MHC I pathway that is abused by viruses in order to delay cell death during early viral infection to provide time for the production of high yields of progeny viruses. The tumor suppressor protein p53 plays an important role in this process and is therefore targeted by multiple viruses. The human papillomavirus for example produces protein E6, which builds a complex that targets p53 for polyubiquitination and degradation of the proteasome (M. Barry and McFadden 1998).

Vaccine design

While traditional vaccines consist of live attenuated or inactivated microorganisms, recent scientific and methodological developments now allow the creation of specific epitope-based vaccines, which open up new possibilities for the treatment of chronic viral diseases and cancer. Accordingly, a variety of vaccines for different indications is currently under development, including gastric cancer, HIV, Asthma, HCV, type 1 diabetes and many more (Purcell, McCluskey, and Rossjohn 2007). In order to identify new potential antigens, there is a great interest in the development of tools to predict proteasome cleavage products.

Epitope-based vaccines offer several advantages over other forms of vaccines: They do not contain infectious material, they can be produced relatively easily on a large scale and they can be stored freeze-dried without the need of a 'cold-chain' for distribution. Drawbacks on the other hand include the need to potently stimulate T cells in order to elicit an immunological response. Epitope-based vaccines also need to be tailored for a given human leukocyte antigen (HLA) haplotype, which is viable however thanks to newer technological advances (Singh-Jasuja, Emmerich, and Rammensee 2004). Furthermore, in many cases the problem can be reduced to nine HLA super types.

An alternative to creating HLA-specific vaccines is the creation of longer peptides with relevance for a broader range of different HLA allotypes. However, this approach relies on the processing of these longer peptides into shorter allele-specific peptides, which requires a detailed understanding of the MHC I pathway.

A better understanding of the MHC I pathway and the proteasomal cleavage behavior may also prove useful for treating immunoevasive pathogens, which often evolve mechanisms to avoid proteolysis by the proteasome and MHC I presentation. A possible solution for this problem might be the fusion of the corresponding antigens to ubiquitin (M. A. Barry, Lai, and Johnston 1995; Levitskaya et al. 1997).

Properties of the Proteasome

The proteasome is an intracellular multi-subunit protease, which is vital for cellular homeostasis. It is not only responsible for the removal of misfolded or malfunctioning proteins within the cell but also supplies the majority of antigenic peptides within the MHC-I pathway. Furthermore, the proteasome is involved in the cell cycle, the cell's stress response, cell-differentiation and metabolic adaptation. (Schwartz and Ciechanover 1999) (Coux, Tanaka, and Goldberg 1996).

Structure

The 26S proteasome consists of the proteolytically active 20S proteasome and two additional 19S regulator units that are ATP-dependently attached to its sides (J. M. Peters et al. 1993).

Four heptameric rings form the cylindrical structure of the 20S proteasome. While the outer rings, through which the substrate enters, consist of 7 α-subunits, the inner rings are formed by 7 β-subunits. The active sites of the proteolytically active subunits β1, β2 and β5 are single threonines located at their amino termini (Groll et al. 1997) (Löwe et al. 1995). The three subunits have different preferences: β1 exhibits a caspase-like, β2 a trypsin-like and β5 a chemotrypsin-like activity (M. Orlowski and Wilk 2000). In the presence of interferon-γ (IFN-γ) the three subunits are replaced by the homologous subunits β1i, β2i and β5i which form an 'immunoproteasome' upon de novo assembly that features a different cleavage specificity (Nandi et al. 1997).

The 19S unit is responsible for recognizing (Deveraux et al. 1994) (Young et al. 1998), deubiquitylating and unfolding the proteasome's ubiquitylated substrate before it is translocated to the 20S proteasome (Michael H Glickman and Ciechanover 2002). It features a 'base' and a 'lid' multisubunit component.

The base consists of six ATPase- and two non-ATPase subunits and binds to the 20S catalytic core (Michael H. Glickman et al. 1998). The ATPases have chaperone-like activity and help to unfold and channel the substrate into the 20S core (Braun et al. 1999; Strickland et al. 2000; M H Glickman et al. 1999).

The lid component binds to the side of the base particle and consists of nine non-ATPase subunits. Its major activity is proposed to be deubiquitylation (Verma et al. 2002; Yao and Cohen 2002; Guterman and Glickman 2004) and its subunits exhibit high homology to the COP9 signalosome complex, which is an essential regulator in various cellular processes (Michael H. Glickman et al. 1998). Additional regulators are discussed in the following.

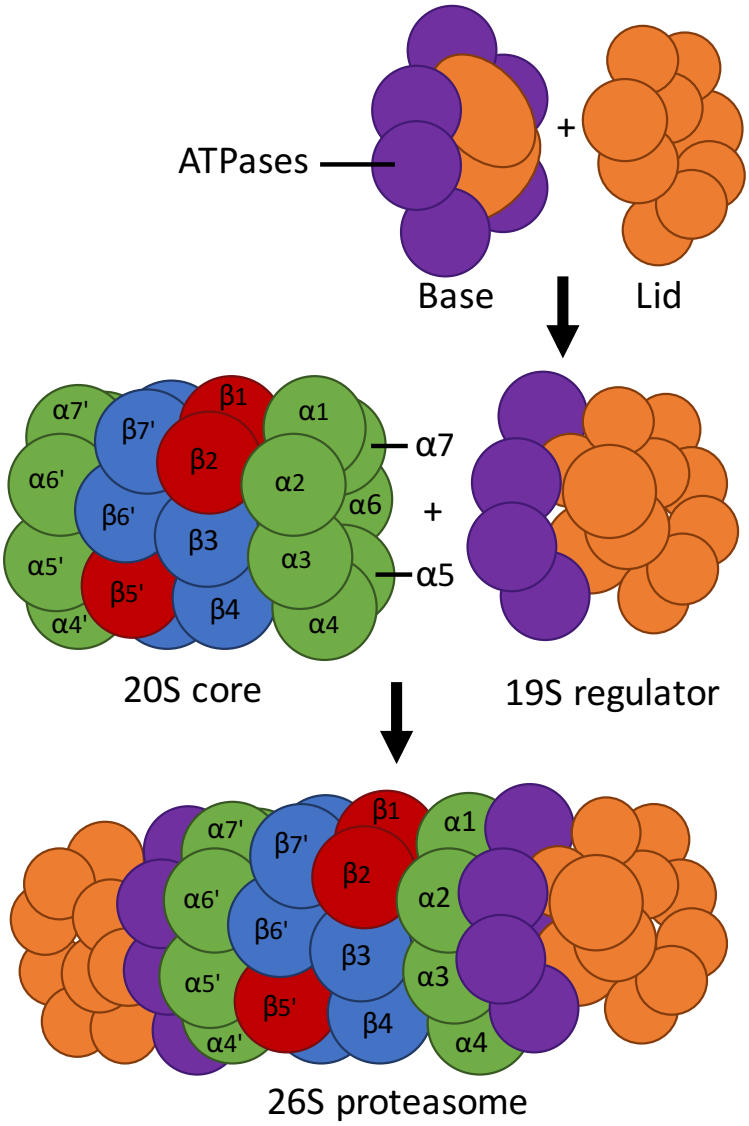


Figure 2 Structure of the 26S proteasome, adapted from (P M Kloetzel 2001)

The immunoproteasome

As mentioned before, the immunoproteasome is formed upon de novo assembly in presence of IFN-γ. Compared to the assembly of the constitutive proteasome, its assembly is accelerated by a factor of three to four but its half-life of 21h is also considerably shorter than the 120h observed for the constitutive proteasome. This high turnover is independent of the presence of cytokines and seems to serve as a transient early response during the early phase of an infection (Heink et al. 2005).

The question of how the immunoproteasome's cleavage behavior differs from that of the constitutive proteasome is not easily answered: One experiment showed, for example, that when HeLa cells were infected with vaccinia virus expressing the hepatitis B virus (HBV) core antigen, the epitope $HBVcAg_{141-151}$ was only presented after stimulation with IFN-γ (A. J. Sijts et al. 2000). This finding seems to support the assumption that the immunoproteasome generates a qualitatively different set of peptides. Highly sensitive analysis by mass spectrometry however revealed that the epitope was in fact also produced by the constitutive proteasome even though with greatly reduced efficiency. In combination with other similar observations, the immunoproteasome therefore seems to have a great quantitative effect on a given epitope (A. J. A. M. Sijts et al. 2000; Strehl and Heink 2005). Thus, effects on an immunological level become detectable after reaching a certain quantitative threshold.

In the majority of other experiments, the immunoproteasome had a positive effect on MHC class I antigen presentation (P M Kloetzel 2001; van Hall T et al. 2000; Schwarz et al. 2000; Van Kaer et al. 1994). At the same time, no findings for a negative effect of the immunoproteasome on epitope generation exist so far.

A large number of in vitro experiments combining mass spectrometry and high-performance liquid-chromatography showed that the immunoproteasome changes the cleavage site preference and therefore the relative amount of peptides being generated. It has to be taken into account however that the substrate turnover of the immunoproteasome is accelerated and that under in vitro conditions a peptide fragment might be more abundant either due to increased turnover or to altered cleavage site preferences. Still, it could be shown that the immunoproteasome has a high cleavage

preference for residues that represent the correct C-terminus of an MHC class I epitope (Strehl and Heink 2005). In addition, it preferably generates epitope precursor peptides with a more extended N-terminal sequence that will facilitate TAP transport (Cascio et al. 2001). Furthermore, the relative usage frequency of certain cleavage sites can greatly differ between the immunoproteasome and the constitutive proteasome depending on the surrounding amino acids (Strehl and Heink 2005).

The regulator PA28

Another component induced by IFN-γ is the 11S regulator PA28 (Chu-Ping, Slaughter, and DeMartino 1992), which attaches ATP-independently to the proteasome's outer α-rings. Expression of PA28 is not completely IFN-γ dependent however, since most tissues exhibit a constitutive, IFN-γ-independent expression as well.

The PA28 component consists of two subunits PA28α and PA28β, which form a ring-like structure (Soza et al. 1997). Binding of PA28 to the 20S core induces subtle conformational changes within the 20S complex that might alter the accessibility of the active site pockets or their binding affinity (Sun et al. 2002).

PA28 enhances the presentation of multiple viral antigens even in the absence of immunosubunits by increasing substrate affinity or the release of peptide product without changing the maximal activity of the enzyme complex (Stohwasser et al. 2000). In contrast to the immunoproteasome, PA28 seems to affect the generation of only a minor fraction of epitopes, considerably fewer studies for PA28 exist however.

While PA28 does not seem to induce new cleavage specificities, it enhances the usage frequency of certain preferred or minor cleavage sites (Sun et al. 2002). Similar to the immunoproteasome, it also greatly accelerates substrate turnover (Strehl and Heink 2005).

There is no experimental evidence that would suggest an additive or cooperative effect of PA28 and the immunoproteasome.

Gating

The N-terminal tails of the 20S α-subunits project into the proteasome's gate blocking access to the catalytic cavity in the absence of regulatory particles (Groll et al. 1997). When binding to the proteasome, PA28 causes the tails of the α-subunits to flip into the hollow core of the PA28 body.

While the opening width of the gate does not affect the proteasome's processing rate, which is determined by substrate binding to the 19S regulator (Thrower et al. 2000), it facilitates substrate entry and product exit through the otherwise closed gate, therefore decreasing the retention time of the substrate intermediates within the catalytic chamber (Stohwasser et al. 2000).

Initially it was suggested that an open conformation could result in the release of longer N-terminally extended peptides which were assumed to be more suitable for antigen presentation. However binding of the 19S regulator opens the gate completely as well and proteasomes formed by the 19S regulator and PA28 (so called hybrid proteasomes) show the same cleavage activity as the 26S proteasome (Kopp, Dahlmann, and Kuehn 2001; Hendil, Khan, and Tanaka 1998). The effect of PA28 on cleavage behavior seems therefore not to be the result of the open gate conformation.

Additional in vitro methods like the addition of low levels of sodium dodecyl sulfate (SDS) are also effective in opening the gate (Coux, Tanaka, and Goldberg 1996).

Peptide processing

The substrate's protein chains are unfolded and transported into the proteasome's core by the 19S regulator. Some findings also indicate that a partial re-folding of the substrate takes place within the core (Sharon et al. 2006). Detailed information about the spatial processes taking place within the proteasome is still lacking.

Multiple findings indicate that the sequence environment of the P1 residue affects the efficiency of epitope generation: Small amino acids like glycine or alanine at the P1' position increase the cleavage probability while other amino acids decrease it (Ossendorp et al. 1996; Beekman et al. 2000; Del Val et al. 1991). The positions P4-P7 affect proteasomal cleavage as well (A K Nussbaum et al. 1998).

Furthermore it was shown that proline residues within the substrate improve the cleavage efficiency (Shimbara et al. 1998).

Even though various cleavage site preferences exist, the proteasome exhibits a high degree of flexibility. Within a protein, almost every amino acid residue can serve as a cleavage site although cleavage efficiency varies due to the flanking residues (Beekman et al. 2000).

While the proteasome generates the C-terminus anchor residues of MHC class I epitopes as mentioned before, correct C-terminal cleavage site usage proves to be less robust than one might expect: Mutations resulting in substitution of only one amino acid flanking the correct C-terminal cleavage site can reduce epitope-generation significantly as was shown for epitopes of Moloney murine leukemia virus (MuLV), p53 and the immunodominant hepatitis C virus (HCV) (Beekman et al. 2000; Theobald et al. 1998; Seifert et al. 2004).

Approaches to cleavage site and fragment prediction

First attempts to model the proteasome's cleavage behavior were solely structure-based. Early findings had suggested that the distance between the active sites acted as a molecular ruler that determined the product length (Wenzel et al. 1994). In fact, the distance between neighboring active sites corresponds to the length of an octa- or nonapeptide in extended conformation (Löwe et al. 1995). Additional experiments however showed size variations that are difficult to explain by an exclusively geometry-based ruler (Kisselev, Akopian, and Goldberg 1998; a K. Nussbaum et al. 1998; Dolenc, Seemüller, and Baumeister 1998). Furthermore, it was observed that proteasomes with different numbers of active sites generated peptides with a very similar length distribution (a K. Nussbaum et al. 1998).

Subsequent models for fragment prediction therefore favor a sequence-based approach. In general, these approaches can be used to predict either cleavage sites or fragments. It is important to note that predicting cleavage sites only does not allow to infer the actually occurring peptide fragments as illustrated in Figure 3.

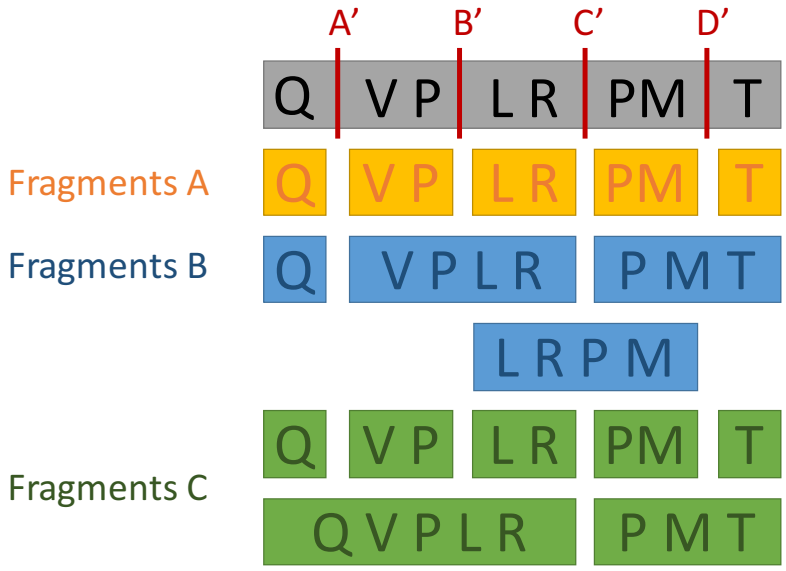


Figure 3: Predicting cleavage sites (A,' B', C', D') does not allow making definite predictions about the resulting fragments. The figure shows three possible sets of fragments that can be inferred from the four cleavage sites given.

In many experiments the number of fragments actually found differs significantly from the number of fragments theoretically possible by the cleavage sites detected. For example, in a digestion experiment with the yeast 20S proteasome and enolase 1 conducted by (A K Nussbaum et al. 1998), the cleavage sites detected would allow for a total of 81 fragments with a length between 9 and 11 amino acids. However only 18 fragments featuring this length were actually detected.

The majority of approaches, which are described in the following, predict cleavage sites only.

FragPredict: Statistical analysis and kinetic model for fragment prediction

MAPPP (MHC I antigenic peptide processing prediction) combines proteasome cleavage with MHC binding prediction. The part responsible for cleavage prediction is called FragPredict and consists of two algorithms: The first one identifies potential cleavage sites based on a statistical analysis of cleavage-determining amino acid motifs present around the scissile bond (Holzhütter, Frömmel, and Kloetzel 1999). The results serve as input for the second algorithm which provides predictions of major proteolytic fragments based on a kinetic model describing the time-dependent digestion of smaller peptide substrates (Holzhütter and Kloetzel 2000).

PAProC: Stochastic algorithm for cleavage site prediction

PAProC uses a stochastic hill climbing algorithm which inspects ten critical amino acid positions in order to predict cleavage sites based on cleavage data obtained in vitro (Kuttler et al. 2000; A. K. Nussbaum et al. 2001). The model assumes that the amino acids at the P1 and P1’ positions have the highest impact on the cleavage probability and learns affinity parameters for the amino acids at each position, which are independent from the state of the other positions.

NetChop: Cleavage site prediction using a neural network

NetChop uses a neural network for fragment prediction (Keşmir et al. 2002) (Nielsen et al. 2005). The network is trained using MHC I class ligands generated by the human proteasome as opposed to the in vitro datasets used in the previous approaches. As not

all fragments generated by the proteasome bind to MHC molecules however, MHC I class ligands represent only a subset of all cleavage products.

Comparison of FragPredict, PAProC and NetChop

Saxová et al. evaluated the three approaches mentioned before by measuring their ability to predict the C-terminal of a set of MHC class I ligands obtained from the SYFPEITHI database (Saxová et al. 2003; H. Rammensee et al. 1999). In their comparison, NetChop performed best even when applied to in vitro data, mainly because non-cleavage sites were predicted better than by the other two algorithms. In addition, as mentioned before, NetChop is the only approach trained with MHC class I ligand data. Table 1 and Table 2 show the performance of all three algorithms as measured by Saxová et al.

However, the fact that Saxová et al penalized the prediction of cleavage sites within a MHC class I ligand is arguable, because a cleavage site might not be used for every fragment being generated. Cleavage sites could be used in different combinations resulting in various fragments (also see Figure 3 for a more detailed explanation).

Method	N	Sensitivity (%)	Specifity (%)	CC
PAProC	217	45.6	30.0	-0.25
FragPredict	231	83.5	16.5	0.00
NetChop 1.0	231	39.8	46.3	-0.14
NetChop 2.0	231	73.6	42.4	0.16

Table 1: Performance of PAProc, FragPredict and NetChop on MHC class I ligands. Saxová et al. found NetChop to predict the C-terminal best of the algorithms examined. N: number of natural MHC ligands tested (less for PAProC because it requires a flanking region). CC: Correlation score that measures the algorithm's positive and negative performance as described in the paper. From (Saxová et al. 2003)

Method	Sensitivity (%)	Specifity (%)	CC
PAProC	46.4	64.7	0.10
FragPredict	72.1	41.4	0.12
NetChop 1.0	34.4	91.4	0.31
NetChop 2.0	57.4	76.4	0.32

Table 2: Performance of PAProC, FragPredict and NetChop on in vitro data. Saxová et al found NetChop to predict the C-terminal best of the algorithms examined. CC: Correlation score that measures the algorithm's positive and negative performance as described in the paper. From (Saxová et al. 2003)

Kinetic analysis of time-dependent product formation

Another approach quantifies cleavage rates using a kinetic proteasome model that incorporates the time-dependent changes of the amount of the peptides generated (B. Peters et al. 2002). The model incorporates a procession rate, which depends on the peptide length and a cleavage probability for each potential cleavage site. Model parameters are estimated for in vitro experiments of two different peptides by quantifying the intensity of the MS signals measured using experimental calibration curves and theoretically determined linear scaling functions. However, the model is mainly intended to examine differences between the cleavage behavior of the constitutive and immunoproteasome and provides evidence for an increased procession rate and some alterations of cleavage probabilities for a couple of restricted cleavage sites.

ProteaSMM: A scoring matrix for cleavage site prediction

Another approach models the whole MHC class I pathway including MHC binding and TAP transport based on in vitro digests of whole proteins (Tenzer et al. 2005). The method responsible for proteasomal cleavage prediction is named ProteaSMM and works with scoring matrices that assign scores to each amino acid located in a 10-residue window around the scissile bond. The cleavage probability of a specific site is determined by adding the score values of the surrounding amino acids. Different scoring matrices for the constitutive and immunoproteasome based on different training data are provided. The authors compare the prediction quality of their method with FragPredict, PAProC and NetChop using a custom set of in vitro data. For this dataset, ProteaSMM clearly outperforms the other methods. Interestingly, the immunoproteasome-specific scoring matrix outperforms the scoring matrix of the constitutive proteasome even on test data derived from constitutive proteasomes. In another comparison of the complete MHC class I pathway model with NetChop 2.0 using MHC I ligand data, the authors show that both methods reach the same level of prediction quality.

Pcleavage: Support vector machine for cleavage site prediction

Pcleavage uses a support vector machine trained with in vitro and MHC I ligand data for cleavage site prediction (Bhasin and Raghava 2005). The authors evaluated the prediction quality and found it to be comparable with that of NetChop. Like the neural network used in NetChop however, the support vector machine allows for no insight on which properties of the test data have an impact on the classification.

ProteaMAlg: Proteasome modeling algorithm

ProteaMAlg describes the proteasome's degradation dynamics using a system of ordinary differential equations (Mishto et al. 2008). The model considers processes like uptake and release of fragments into/from the proteasome as well as proteolytic cleavage of peptides inside the proteasome. In addition, the amino acid at each position of the substrate is incorporated in the model using substrate-specific cleavage strengths which can be determined either experimentally or using PAProc, NetChop or a similar prediction algorithm. The authors find that prediction of peptides is not possible with their or other existing statistical models. They can only describe the production of observed fragments from a specific substrate by fitting the model parameters to the observed data. New substrates provide entirely new parameter values. It could be shown however that both the substrate length and the amino acid composition affect the substrate cleavage strength and the overall substrate degradation rate. It was also shown that the generation of double cleavage products is favored in presence of PA28.

Scoring function for fragments

Another approach by Ginodi et al. assigns a score for the probability of a fragment to be generated by the proteasome instead of predicting cleavage sites (Ginodi et al. 2008). The scoring functions, which are distinct for the constitutive and immunoproteasome, assign a position-specific score to each amino acid within a given peptide as well as the flanking amino acids at its C- and N-terminus. The score values are learned from in vitro data using a simulated annealing process. Thus the probability that a given peptide is produced during cleavage is described as a linear combination of each amino acid's effect within the peptide and its flanking region.

Validation with multiple datasets including naturally processed epitopes taken from the SYFPEITHI database showed a specificity and sensitivity of over 70%. Depending on the training data, results were even better. Therefore, the authors find their algorithm to perform significantly better than all the approaches evaluated by Saxová et al., even though a direct comparison is admitted to be difficult, since the other methods predict cleavage sites instead of fragments.

Mass spectrometry

The foundation of any prediction algorithm is the experimental data available for training and/or evaluation purposes. The data used in this work was obtained in in-vitro digestion experiments. The digestion products were identified using mass spectrometry. The general process is depicted in Figure 4.

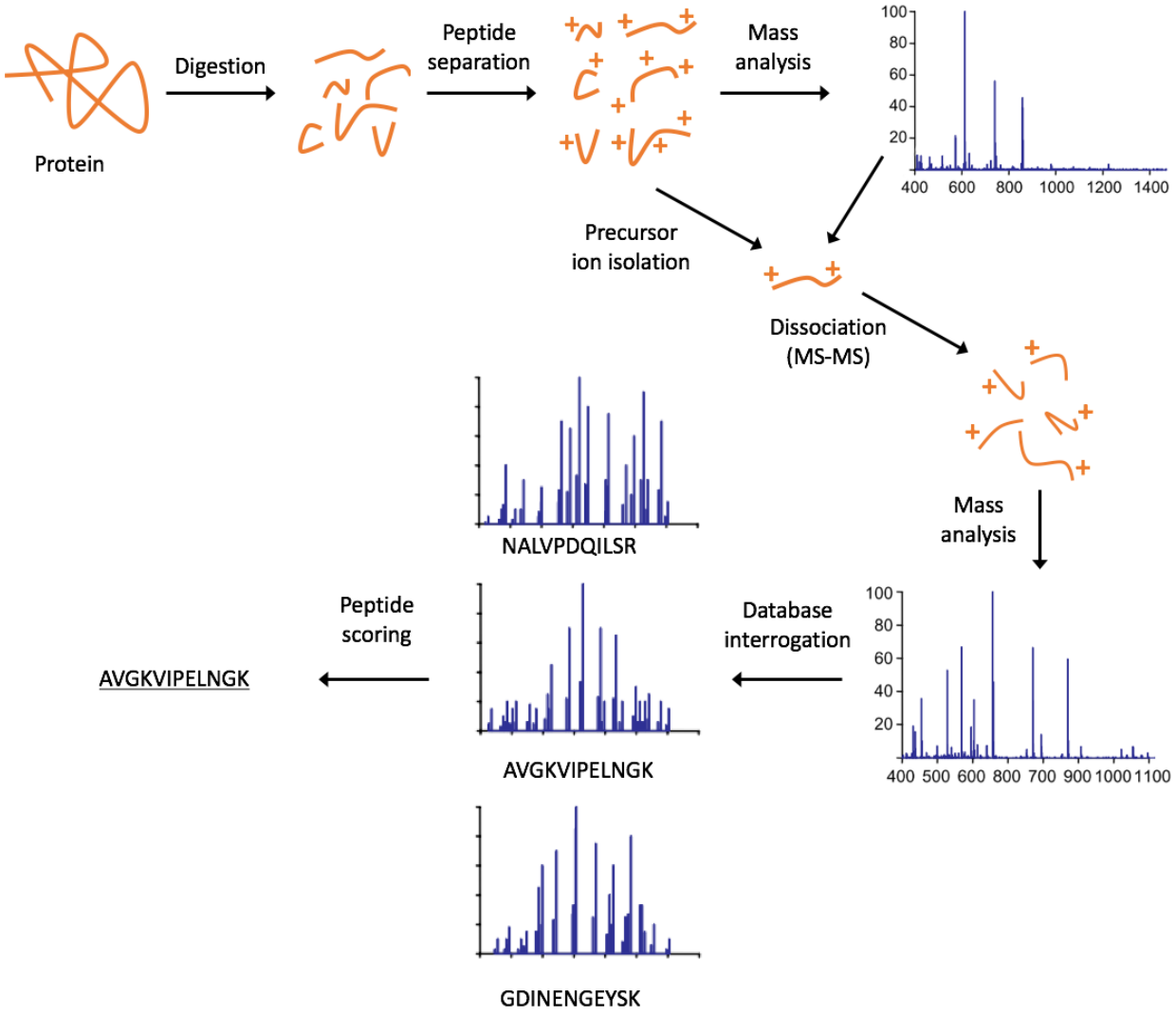


Figure 4: General process of in-vitro digestion experiments, which supplied the training data for this work. Proteins are digested by the proteasome and separated by high performance liquid chromatography. Peptide masses are measured and individual isolated peptides are subjected to MS-MS. The measured fragment ions combined with the peptide mass are used for peptide identification through the database of possible fragments. Picture adapted from (Kolker, Higdon, and Hogan 2006).

Mass spectrometry has become the standard method for analysing peptides over the past 20 years. Mass spectrometers measure ions and make measurements of mass-to-charge. There are two most commonly used ionization methods: electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI). Structural information about peptides can be obtained by fragmentation of peptides in two consecutive MS-measurements (MS-MS, or tandem MS).

The measured MS-MS spectra are usually analyzed using database search programs, which compare the observed MS-MS spectra with all candidate peptide MS-MS spectra that can occur according to the initial substrate. A key challenge for data analysis is to distinguish correct peptide identifications from incorrect ones. Accepting each database search result as correct would lead to a an abundance of false positives (Keller et al. 2002). Therefore, minimum score thresholds are usually used to reduce the number of positive identifications. Various confounding factors like noise, instrument under-sampling or low abundance signal suppression also need to be taken into consideration when trying to identify actually occurring peptides.

A database search-program used in this work, called Mass Spectrometry FileAnalyzer (further explained in the following chapter), also takes the substrate's time-dependent degradation into account in order to improve the reliability of its results.

Goals of this work

Considering the introductory remarks, this work tries to contribute to the greater goal of gaining a deeper understanding of the proteasome's cleavage behavior by:

1. Developing an improved approach for prediction of the proteasome's cleavage products using statistical methods, which requires
2. Establishing a suitable database created from digestion experiments conducted at the institute of biochemistry of the Charité Berlin. This in turn requires
3. Validating the software "FileAnalyzer" created by the Holzhütter working group, which was used to analyze the experiments' MS-data.

Methods

Dataset

Working groups under supervision of Prof. Kloetzel and Prof. Holzhütter of the institute of biochemistry of the Charité Berlin conducted a variety of experiments concerned with proteasome digestion between 2000 and 2011. The results of these experiments served as training data for the work of this thesis. Overall, there was data of experiments with 63 different substrates available. A complete list of all experiments included and their most important properties can be found in Appendix 1.

During the experiments, peptides were incubated with proteasomes of varying cell lines. The digestion products were separated using HPLC and then analyzed by mass spectrometry (MS). The MS raw data obtained was analyzed using two different methods: software-assisted manual evaluation and a fully automated approach using a software called “Mass Spectrometry File Analyzer” created by Dr. Andrean Goede of the institute of biochemistry of the Charité Berlin.

Internal instead of publicly available data was used because of its integrity and consistency: All experiments had been conducted in a homogeneous setting, with the same methods (MS, HPLC) and were well documented.

Software assisted manual evaluation

During software-assisted manual evaluation members of the Kloetzel working group analyzed and validated the raw mass spectrometry data using the MS Bioworks software suite, creating cleavage maps for each experiment containing all fragments detected with a high level of certainty (usually between 20 and 30), which was ensured by crosschecking the MS-data at multiple time points and iterations.

One cleavage map usually incorporated data of multiple mass spectrometry measurements, sometimes even of multiple experiment iterations (e.g. with and without an activator like PA28) and listed the fragments detected without any ranking. See Table 3 for a sample cleavage map.

Ion signal				Idx	RT	MW monoiso.		MS/MS		
(monoiso., m/z)					min	Da				
det	calc	D	z			det	calc	D		RT
										min
473,3	473,3	0	1	26-29	21,8	472,2	472,3	-0,1	2	22.1
472,1	471,3	0.8	2	4-12	23,6	942,2	940,6	1,8	2	23.8
966,1	966,6	- 0.5	1	11-19	23,8	965,1	965,6	-0,5	3	
572,1	571,8	+0.3	2	20-29	24,4	1142,2	1141,6	0,6	2	24.7
429,8	429,7	+0.1	2	20-27	26	857,5	857,5	0	2	25.9
679,1	678,9	+0.2	2	18-29	26,2	1356,1	1355,8	0,3	2	26.5
537,1	536,8	+0.3	2	18-27	27,4	1072,2	1071,6	0,6	2	29.0
563,6	563,4	+0.2	2	1-10	27,5	1125,3	1124,7	0,6	2	27.9
300,4	300,2	+0.2	2	1-5	27,9	598,7	598,4	0,3	2	27.9
575	574,8	+0.2	2	16-25	28,3	1148	1147,6	0,4	2	28.4
802,1	801,9	+0.2	2	16-29	28,6	1602,3	1601,9	0,3	2	28.9
479,1	479,3	-0.2	1	16-19	29,1	478,1	478,3	-0,2	2	29.3
648,5	648,4	+0.1	2	1-12	29,5	1295	1294,8	0,2	2	29.6
660	659,9	+0.1	2	16-27	29,6	1317,9	1317,7	0,2	2	29.7
506,2	506,3	+0.1	1	11-15	30,2	505,2	505,3	-0,1	2	30.4
517,1	516,8	+0.3	2	6-15	30,7	1032,1	1031,6	0,5	2	30.9
807,2	807	+0.2	2	1-15	31,5	1612,3	1612	0,3	2	32.0
640,1	639,9	+0.2	2	6-17	33	1278,2	1277,7	0,5	2	33.1
930,3	930,1	+0.2	2	1-17	33,9	1858,5	1858,1	0,4	2	34.2

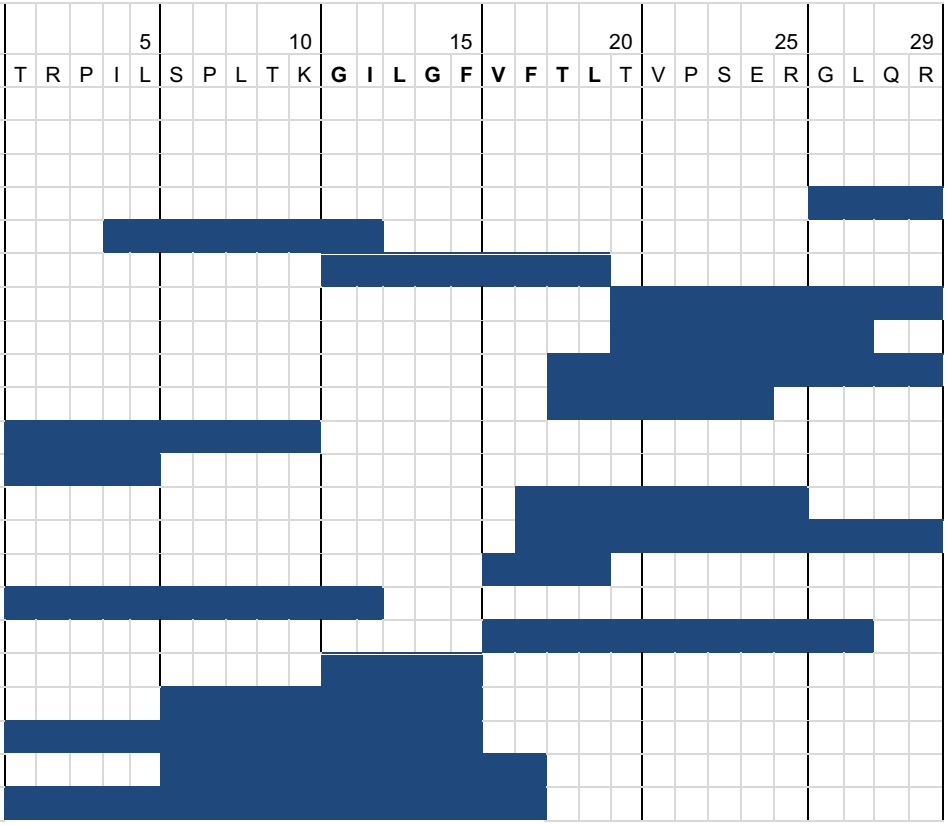


Table 3: Sample cleavage map created via software assisted manual evaluation (some additional information was left out for better readability). On the right top the substrate is shown (TRPILSPLTKGILGFVFTLTVPSERGLQR), below the fragments detected are displayed as blue blocks.

Fully automated approach with “Mass Spectrometry File Analyzer”

The mass spectrometry data was also analyzed using a custom software solution created by Dr. Andrean Goede named “Mass Spectrometry File Analyzer” (see Figure 5 and Figure 6 for sample screenshots). All mass spectrometry raw-data-files available for each individual experiment-iteration were analyzed separately and a list of fragments detected ranked by a probability score was obtained for each file.

While the manually created cleavage maps only listed fragments that were found with a high level of certainty, which was ensured by validating their occurrence in multiple mass spectrometry files, the Mass Spectrometry File Analyzer detected fragments with a higher level of sensitivity at the expense of specificity.

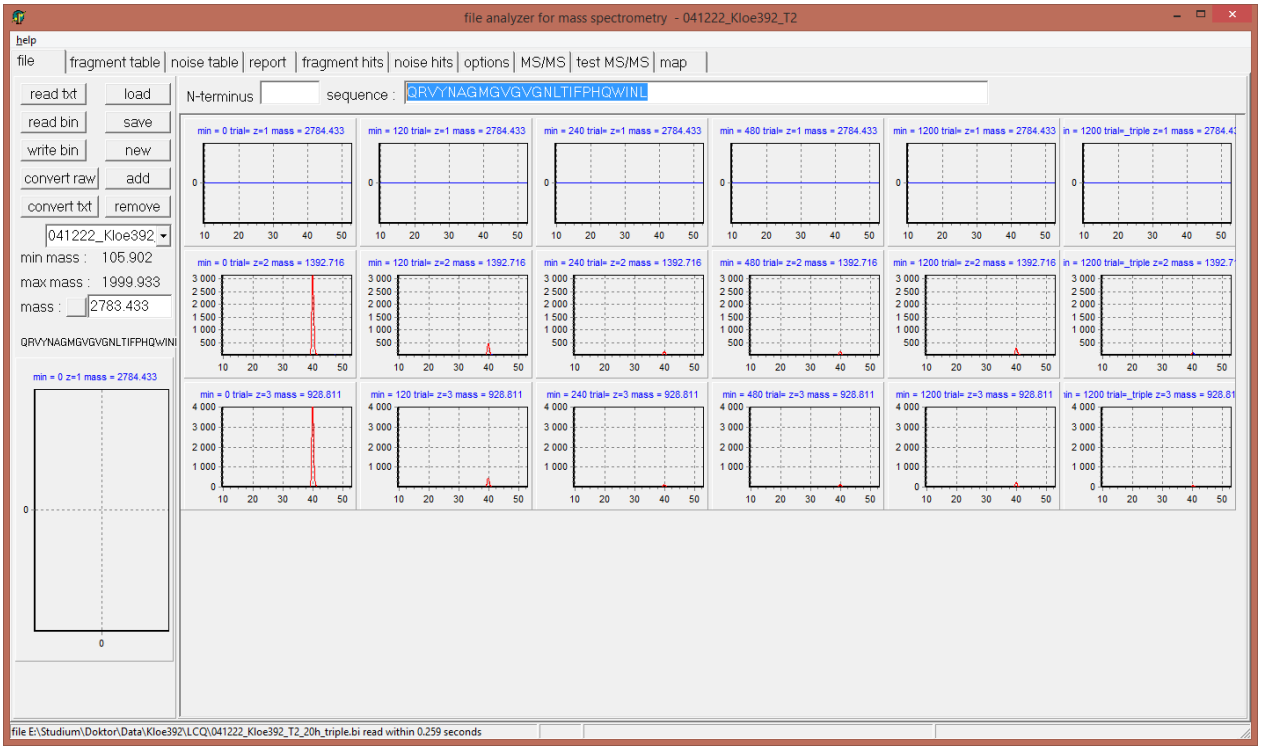


Figure 5: Screenshot from the "Mass Spectrometry File Analyzer" used for analysing MS raw data. In this screen the substrate sequence was entered in the upper textbox. Multiple mass spectrometry raw files of an experiment were selected and are shown at the bottom.

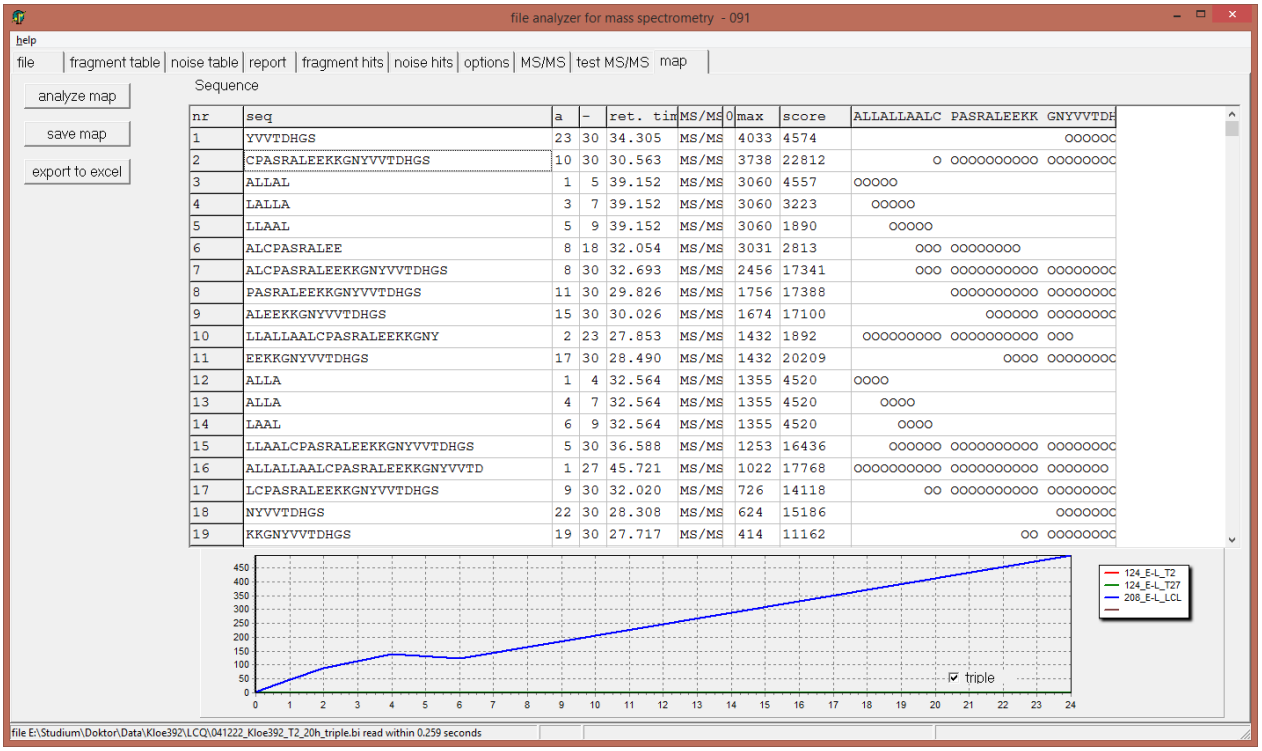


Figure 6: Fragment list retrieved from Mass Spectrometry File Analyzer. The list is sorted by a probability score, which is displayed next to the fragment's sequence.

Ensuring a distinct dataset

In order to prevent biased results caused by a culmination of similar substrates, a distinct dataset was built by filtering out similar peptides using the basic local alignment search tool (BLAST) provided by the National Library of Medicine (Johnson et al. 2008). Table 4 shows the settings used for the DELTA-BLAST algorithm. Each experiment was aligned with all other experiments of the dataset. All alignments with an Expect value below 1e-04 and a query cover above 80% were considered. The Expect value reflects the probability of a detected similarity to be random, while the query cover accounts for the portion of matching amino acids relative to the whole peptide.

In order to obtain the distinct dataset, the alignment with the largest query cover was selected and its experiment was removed from the dataset. If there were multiple alignments with the same query cover, the experiment with the most alignments to all other experiments was removed. This process was repeated until no more alignments above the threshold remained. The resulting distinct dataset contains 48 experiments (see Appendix 1 for a detailed list of all experiments).

Setting	Value
Max target sequences	100
Short queries	Enabled
Expect threshold	10
Word size	3
Max matches in a query range	0
Matrix	BLOSUM62
Gap Costs	Existence: 11 Extension: 1
Compositional adjustments	Conditional compositional score matrix adjustment

Table 4: BLAST settings used to identify similar sequences

Dataset subsets

During decision tree generation, various subsets of the training data were used. Table 5 shows an overview of these subsets.

Name	Description	# Fragments	Fragments detected
CMap	Data of all manually created cleavage maps	26131	1826 (7%)
CMap*	Data of all manually created cleavage maps for the set of distinct experiments (see above)	20838	1349 (6%)
FileAn	Data of all analyses performed with the Mass Spectrometry File Analyzer	124906	52049 (42%)
FileAn*	Data of all analyses performed with the Mass Spectrometry File Analyzer for the set of distinct experiments (see above)	103858	43522 (42%)

Table 5: Training data subsets used for decision tree generation. # Fragments: Overall number of fragments that could theoretically be derived from the substrates is included in the subset. Fragments detected: Actual number of fragments that were detected either by manual evaluation (in case of CMap) or by the File Analyzer (in case of FIleAn)

Decision tree

Pattern recognition and empirical learning from examples is a common task in today's biomedical sciences. While the majority of the algorithms used in the field provide very good and reliable results in most scenarios, many of them lack the possibility to easily read and understand the decisions that are relevant in order to obtain the resulting classification.

Decision tree learning as introduced by Quinlan (J. R. Quinlan 1986) in the form of the ID3 algorithm is a method of machine learning which allows to visualize the learned discrete-valued function as a tree or a set of if-then-rules, both with good human readability. It is one of the most widely used methods for inductive inference and is robust to noise while searching a completely expressive hypothesis space.

The main motivation for using decision trees in this work was the possibility to create a model, which does not work like a black box but whose rules and decisions are comprehensible. The goal was not to just model the training data as well as possible but also to identify relevant substrate properties which determine the cleavage process, thus gaining more insight into the inner workings of the proteasome.

Algorithm

The algorithm performs a top-down greedy search through the space of possible decision trees, evaluating at each step which attribute separates the training data best using a criterion that usually measures the expected reduction in entropy but may vary depending on the actual implementation.

Multiple variations and refinements of the decision tree learning algorithm exist, e.g. C4.5 (J R Quinlan 1993), GID3 (Cheng et al. 1988) or ASSISTANT (Cestnik, Kononenko, and Bratko 1987). The general approach for tree induction is mostly the same in all variants and is shown in Table 6. In this work an implementation of C4.5 with the gain ratio criterion and pruning as described in (J R Quinlan 1993) was used.

C4.5 (examples, targetAttribute, attributes)
examples: training data used to induce the tree
targetAttribute: attribute whose value is to be predicted by the tree
attributes: Set of attributes to be examined by the algorithm for classification
Returns a decision tree that classifies the given examples into the values of targetAttribute using the attributes supplied.

- Create a *Root* node for the tree
- If all *examples* have the same value v of *targetAttribute*, return the single-node tree *Root* with label = v
- If *attributes* is empty, return the single-node tree *Root* with label = most common value of *targetAttribute* in *examples*
- Otherwise begin
  - *A* = the attribute with the highest gain ratio
  - Set decision attribute of *Root* = *A*
  - For each possible value a_i of *A*
    - Add a new tree branch below *Root* corresponding to the test $A = a_i$
    - $examples_{ai}$ = subset of *examples* that have value a_i for *A*
    - If $examples_{ai}$ is empty
      - Then below this new branch add a new leaf node with label = most common value of *targetAttribute* in *examples*
      - Else below this new branch add subtree C4.5($examples_{ai}$, *targetAttribute*, *attributes* – {A})
- Return *Root*

Table 6: Summary of the decision tree algorithm C4.5. After tree induction, the tree is pruned in an additional step in order to avoid overfitting

Gain ratio criterion

In order to select the attribute that best classifies the training data in each step the original ID3 algorithm makes use of the information entropy:

$$Entropy(S) = \sum_{c \in C} -p_c \, log_2 p_c$$

where S is a set of samples, C the target classification and p_c the proportion of S belonging to the target class c.

Using the entropy measure, the gain criterion can be defined as follows:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

where Values(A) is the set of all possible values of attribute A and S_v the subset of samples in S, which have the attribute value v (S_v = { s ∈ S | A(s) = v }). ID3 selects the attribute with the highest information gain in each recursive step.

However, the gain criterion has a strong bias in favor of attributes with many attribute values. An extreme example would be a patient identification number in a medical diagnosis task. Since each subset would only contain a single case, Entropy(S_v) would become 0 for all subsets and Gain(S, A) would reach its maximum. While perfectly classifying the example data, this division would be rather useless regarding its predicting value. In order to rectify this bias, Quinlan introduced the gain ratio criterion in C4.5:

$$Gain\ ratio(S, A) = \frac{Gain(S, A)}{Split\ info(S, A)}$$

with

$$Split\ info(S, A) = -\sum_{v \in Values(A)} \frac{|S_v|}{|S|} log_2 \frac{|S_v|}{|S|}$$

Handling continuous attributes

An approach for handling continuous attribute values within decision trees was first introduced by Paterson and Niblett (Paterson and Niblett 1982): The samples in the examples set S are first sorted by their values of attribute A to be considered. These values {v_1, v_2 … v_n} can then be split into two subsets {v_1 …. v_i} and {v_{i+1} … v_n} by a single threshold value lying between v_i and v_{i+1}. There are thus only n – 1 possible splits on A, which can be examined with linear costs because the list of values is sorted. In C4.5, the threshold value is usually set to v_i.

Pruning

Since the decision tree is grown until it fits the training data as well as possible, the danger of overfitting the data is relatively high. Therefore, the decision tree is pruned after its creation in an additional step. A node in the tree is pruned by removing the node's subtree and making it a leaf node with the most common classification of all samples associated with the node. The approach used in C4.5 is called reduced error pruning: Starting from the bottom of the tree, each non-leaf node is examined. If replacing the node's subtree results in a lower predicted error rate, the node is pruned accordingly.

A node's error rate is estimated using the upper limit of the binomial proportion confidence interval U(E, N):

$$U(E,N) = \hat{p} \pm z\sqrt{\frac{1}{N}\hat{p}\,(1-\hat{p})}$$

where N is the number of training samples covered by a leaf, E the number of wrongly assigned samples within the leaf, $\hat{p}$ the proportion of successes in a Bernoulli trial process and z the $1 - \frac{1}{2}\,\alpha$ percentile of a standard normal distribution (C4.5 uses a 25% confidence level).

The error rate of a non-leaf node is given by the sum of predicted error of its child nodes. A major advantage of C4.5's reduced error pruning approach is that no part of the training data needs to be reserved for error estimation because the error is estimated heuristically. Figure 7 illustrates the process with a simplified example.

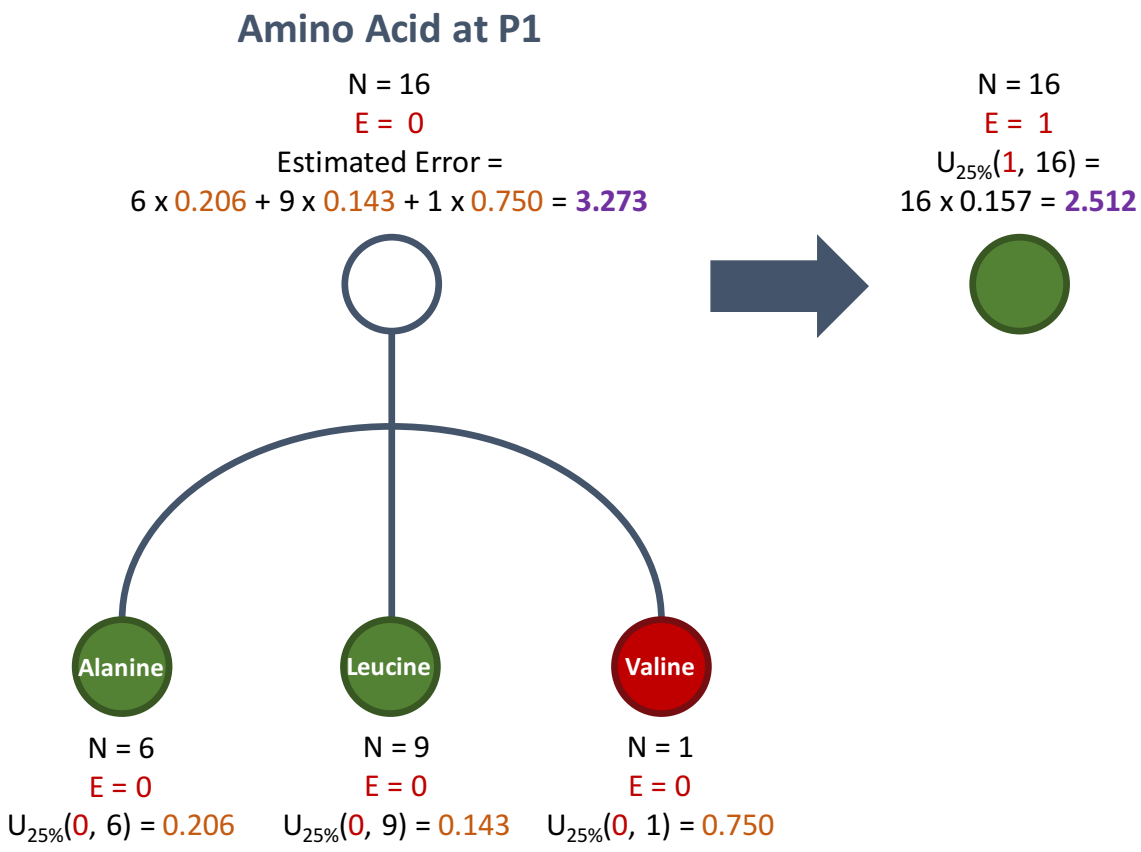


Figure 7: Simplified example for error pruning. The tree on the left has three leaf nodes, the first two of them (green) classifying a fragment as being products of the proteasome. The estimated error in the parent node is 3.273. If the three nodes are replaced by a single leaf node, the estimated error is only 2.512 (node on the right), thus pruning is performed on this node. Adapted from (J R Quinlan 1993).

Classification

A decision tree maps its input data to a discrete classification, which is associated with a probability score. Predicting multiple cleavage sites within a peptide is a rather unsuitable task for a decision tree, since the learned function is injective and could therefore only classify a single cleavage site at a time. Multiple cleavage sites would have to be modeled using multiple trees or using other workarounds. A binary classification of whole fragments into the classes “generated by the proteasome” and “not generated by the proteasome” however, is very well suited for a decision tree and results in a score for each fragment reflecting the probability of the fragment being created by the proteasome.

Attributes selected for decision tree creation

The hypothesis space searched by the decision tree algorithm is defined by the attributes that describe the training data. Selecting promising attributes is therefore critical and a variety of attribute sets, which is described in the following, was used in this study. Both positional constraints and physicochemical properties of the individual amino acids play an important role during substrate binding to the proteasome's active sites as illustrated in Figure 8.

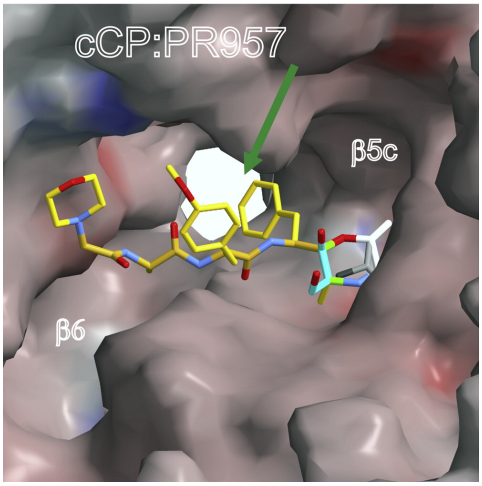


Figure 8: Conolly Surface Representation of the proteasome's β5c and β5i active sites in the presence of a substrate (in this case the epoxyketone inhibitor PR-957). Surface colors indicate positive and negative electrostatic potentials contoured from 50 kT/e (intense blue) to 50 kT/e (intense red). Thr1 is colored in white, and the substrate is highlighted in yellow. Reprinted from Cell, volume 148, issue 4, (Huber et al. 2012), with permission from Elsevier

Amino acid index database

The amino acid index database (AAIndex) contains a wide collection of published physicochemical and biological properties of amino acids (Nakai, Kidera, and Kanehisa 1988; Tomii and Kanehisa 1996; Shuichi Kawashima, Ogata, and Kanehisa 1999; S Kawashima and Kanehisa 2000). Currently it includes 544 different attributes. All of these attributes were used for decision tree creation.

In order to reduce the calculation duration for some decision trees and to avoid overfitting due to an abundance of properties, the attributes of the amino acid index database were also clustered using a maximum linkage cluster algorithm. Each attribute within the amino acid index database is defined by a vector of 20 values. The Pearson product-moment correlation coefficient was used as a distance measure between two vectors:

$$r(\vec{x},\vec{y}) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

with $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and x and y being distinct attribute vectors of the amino acid index database.

A cluster c is defined as a set of attribute vectors: $c = \{\vec{x},\vec{y},\vec{z}\}$ and the maximum linkage between two clusters is given by

$$maximumLinkage(c_1, c_2) = \max(r(x_{c1}, x_{c2})) \mid x_{c1} \in c_1 \wedge x_{c2} \in c_2$$

The algorithm was started with n = 544 clusters, each containing a single attribute vector. The two clusters with minimum maximumLinkage were determined and merged until the target count of ten clusters was reached. For each cluster, the attribute vector with minimum distance to all other vectors of the cluster was selected as representative:

$$representative(c) = x \mid x \in c \wedge \sum_{i=1}^{|c|} r(x, c_i) = min$$

Tomii et al. describe the same approach for clustering, however they define six logical clusters: alpha and turn propensities, beta propensity, composition, hydrophobicity, physicochemical properties and other properties (Tomii and Kanehisa 1996). Because six clusters did not seem to provide a sufficient selection of attributes to choose from for the decision tree algorithm, ten clusters were created for this work instead.

Table 7 shows an overview of the 10 clusters created for tree generation. A complete list of the clusters including all entries from the amino acid index database can be found in Appendix 2.

Cluster Name	Cluster Representative	Number of Attributes included
Representative Description		
Cluster 1	WERD780103	79
Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule		
Cluster 2	KHAG800101	67
The Kerr effect of amino acids in water		
Cluster 3	AURR980118	58
Helix capping		
Cluster 4	RACS820102	31
Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids		
Cluster 5	TANS770108	66
Statistical mechanical treatment of protein conformation. 5. A multiphasic model for specific-sequence copolymers of amino acids		
Cluster 6	YUTK870104	80
Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit		
Cluster 7	RICJ880102	29
Amino acid preferences for specific locations at the ends of alpha helices		
Cluster 8	QIAN880117	50
Predicting the secondary structure of globular proteins using neural network models		
Cluster 9	QIAN880138	44
Predicting the secondary structure of globular proteins using neural network models		
Cluster 10	KLEP840101	40
Prediction of protein function from sequence properties: Discriminant analysis of a data base		
		544

Table 7: AAIndex database clusters created for tree generation

Aggregated fragment attributes versus specific position attributes

Many of the eligible attributes (like hydrophobicity, mass or polarity) are properties of the single amino acids comprising the fragments. There are different possibilities for evaluating these attributes, which were both examined in this work.

Considering only specific positions

Only specific positions within the fragment can be considered, which seems promising, since various findings indicate that certain positions (like P1 or P4-P7) within the proteasome's substrate are especially relevant during the cleavage process (Ossendorp et al. 1996; Beekman et al. 2000; Del Val et al. 1991; A K Nussbaum et al. 1998). Since a fragment can result from a single or two consecutive cuts, the cleavage sites are distinguished by naming the site closer to the N-terminus the "tail" site as illustrated in Figure 9. An attribute's median value was used for positions that were not available in a fragment (e.g. P1 of a head-fragment).

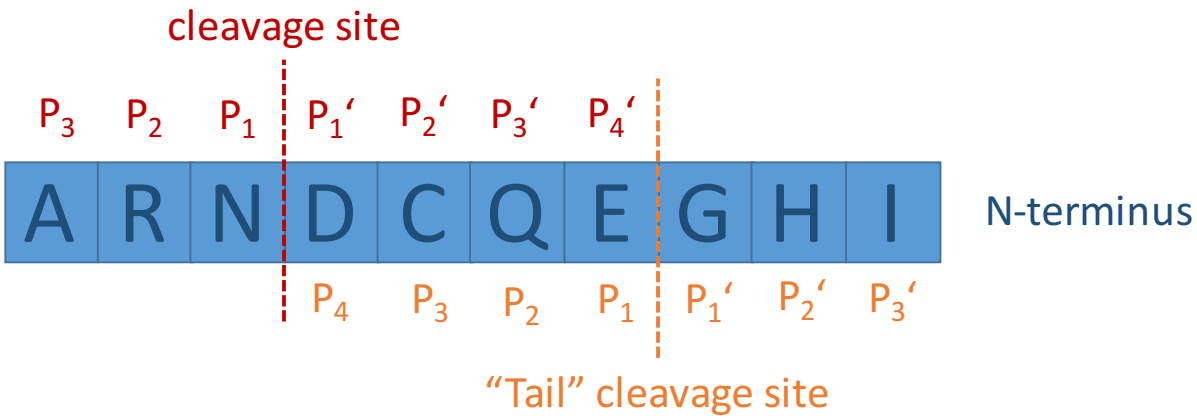


Figure 9: Naming conventions used for cleavage sites and amino acid positions

Calculating a summed value for the whole fragment

Alternatively, it is also possible to calculate a summed value by adding the attribute values of all amino acids comprising the fragment. Figure 10 shows an example evaluating the amino acids' hydrophobicity index (as described in (Argos, Rao, and Hargrave 1982)) of the fragment "ARN": When only considering the P1 position of the

tail cut, we obtain the hydrophobicity index of asparagine (N), which is 0.06. Calculating the summed value for the fragment, we obtain a value of 1.27 (hydrophobicity index of alanine, arginine and asparagine combined). It is important to note that by adding up the values of all amino acids within the fragment, the fragment's length is inherently contained in all attribute values and a short fragment containing amino acids with a high hydrophobicity might produce similar values like longer fragments with amino acids featuring a lower hydrophobicity. Still, this approach seems more promising than calculating the mean value of all amino acids within a fragment or similar, since this would result in similar values for a very short and a very long fragment as long as both consisted of amino acids with a similar hydrophobicity.

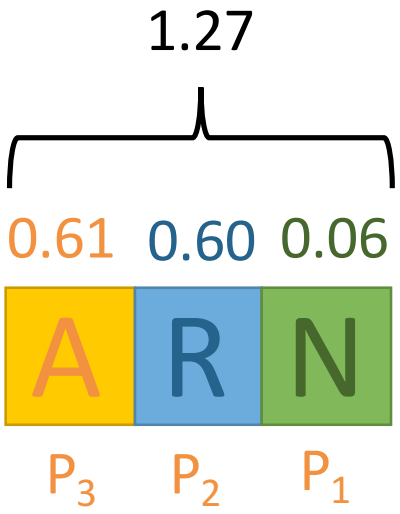


Figure 10: Evaluating the hydrophobicity index of the amino acids within a sample fragment: We can either consider amino acids at specific positions only or build a sum of all values together

Attribute sets used for decision tree generation

Table 8 shows the attribute sets used for decision tree generation.

Name	Description
AaCodesP1-P7	Amino acid one letter codes at positions P1/P1' to P7/P7' (head and tail cleavage site). In addition, the length of the fragment in amino acids was added as an attribute
AAIndexPm-Pn	Contains an attribute for each position Pm/Pm' to Pn/Pn' (head/tail) and each property in the AAIndex database
AAIndexFragment	Contains an attribute for each property in the AAIndex database returning the fragment's summed property value as described above
[set]#	The corresponding attribute set containing only the cluster representative properties of AAIndex as described above

Table 8: Attribute sets used for decision tree generation

Results

Software development

For decision tree creation and additional statistical computations an interactive web-based software-application was implemented, whose architecture is shortly presented in the following.

Database

The Microsoft Excel-based manual cleavage maps as well as the output of the Mass Spectrometry FileAnalyzer were imported into a SQL-based relational database. Its database diagram is shown in Figure 11. Each experiment is stored in the “Experiment” table, which stores the experiment’s date, the protein-sequence and some additional information. The experiments were usually conducted multiple times with multiple measurements taking place. While these measurings were evaluated together in order to obtain a single manual cleavage map, the FileAnalyzer analyzed each measuring separately. Both the manual cleavage maps and the FileAnalyzer results are stored in the “Analysis” table, where the “AnalysisTypeId” marks the origin of the corresponding list. The fragments of an analysis are stored in the “Fragment” table. The relational database allows for a flexible analysis of the dataset even for future questions.

In order to improve performance of the decision tree algorithm the data was also copied into a non-SQL database (using the MongoDB database runtime), which allowed quicker retrieval of certain data required during tree generation. The data of the amino acid index database (AAIndex), the corresponding clusters, data of the SYFPEITHI-database and the generated trees themselves were also stored in this database. Figure 12 shows all catalogs used in the non-SQL database.

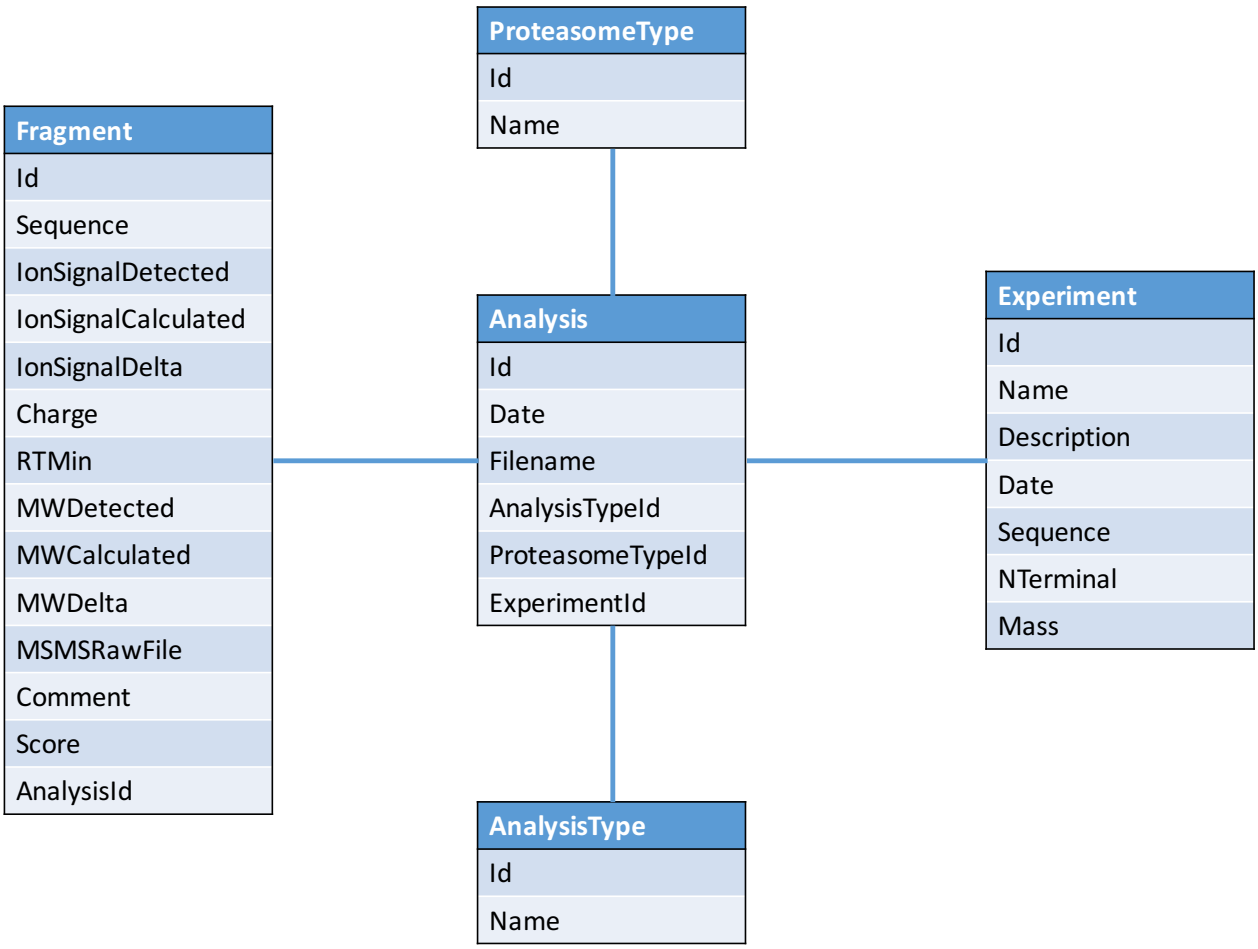


Figure 11: Entity-Relationship-Model of the database used in the implementation. An experiment was usually conducted involving multiple mass spectrometry measurements. These measurings were evaluated together in order to obtain a single manual cleavage map. Within the FileAnalyzer, each measuring was evaluated separately. Both the manual cleavage maps and the FileAnalyzer fragment lists are stored within the Analysis table (the field "AnalysisTypeId" marks their type).

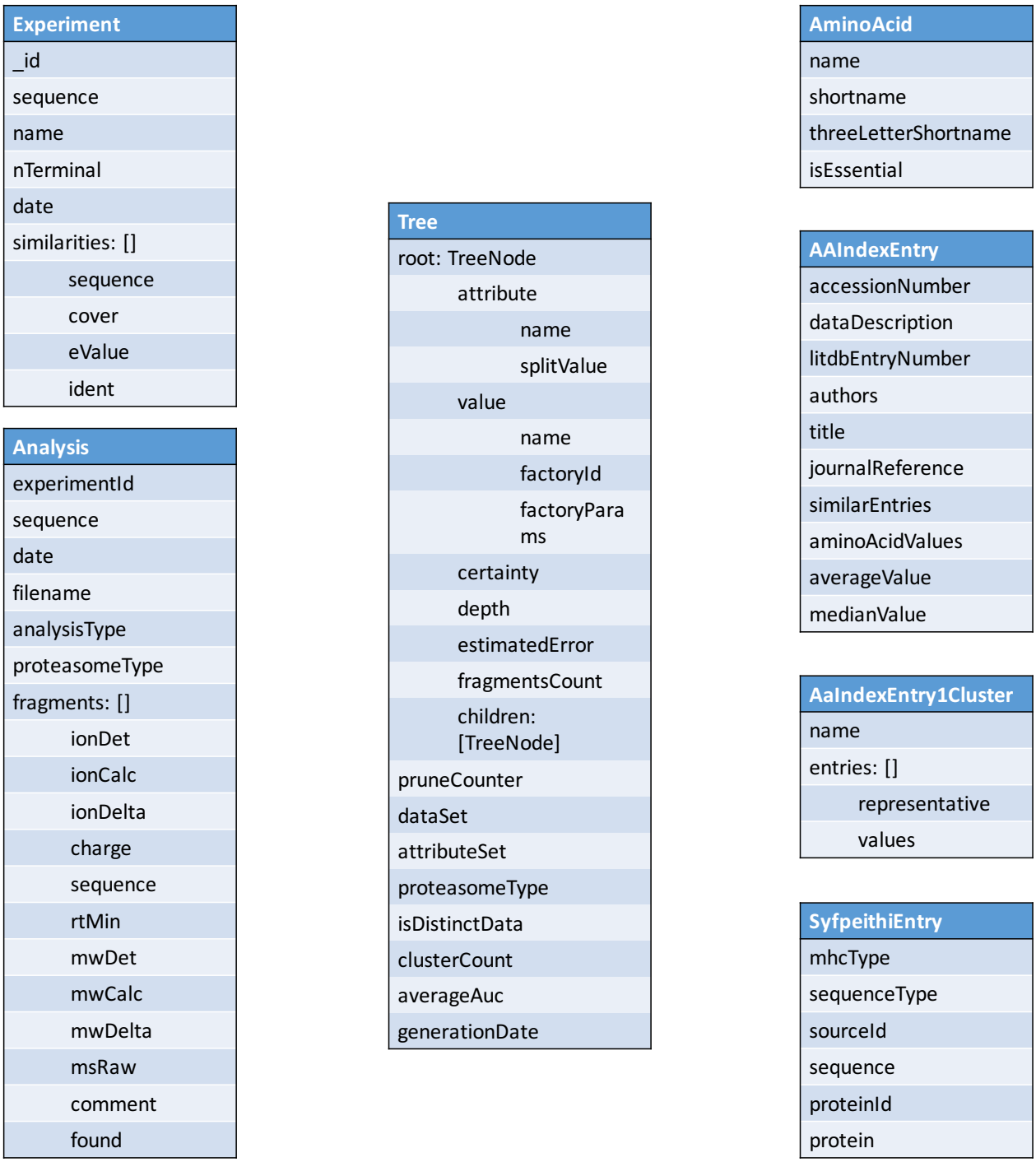


Figure 12: Schema of catalogs used in the non-SQL database. This database was used for the decision tree algorithm because it allowed for faster data retrieval and storage. The decision trees generated were saved in the tree catalog. Experiment data was stored in the Experiment and Analysis catalog. Data from the amino acid index database and the clusters created were stored in AAIndexEntry and AAIndexEntry1Cluster. General data about the amino acids was stored in AminoAcid, data from the SYFPEITHI database used for tree validation was stored in SyfpeithiEntry.

Interactive program

An interactive, web-based program was implemented, which allowed decision tree generation with varying parameters. The general architecture and frameworks used are shown in Figure 13. The main algorithm used for decision tree generation was implemented in JavaScript and is therefore exclusively running on the client computers. After retrieval of all necessary data from the server, which usually requires less than a minute, the actual tree generation is performed without any further server interaction on the corresponding client. Only after being finished, the resulting tree is send to the server and stored in the database again. This approach allowed for parallelized computation of multiple trees at once without major performance tradeoffs. Because the algorithm only requires one processor thread, even parallelized computation on a single client computer was possible with an average multicore client computer using a multithreaded web browser like Google Chrome.

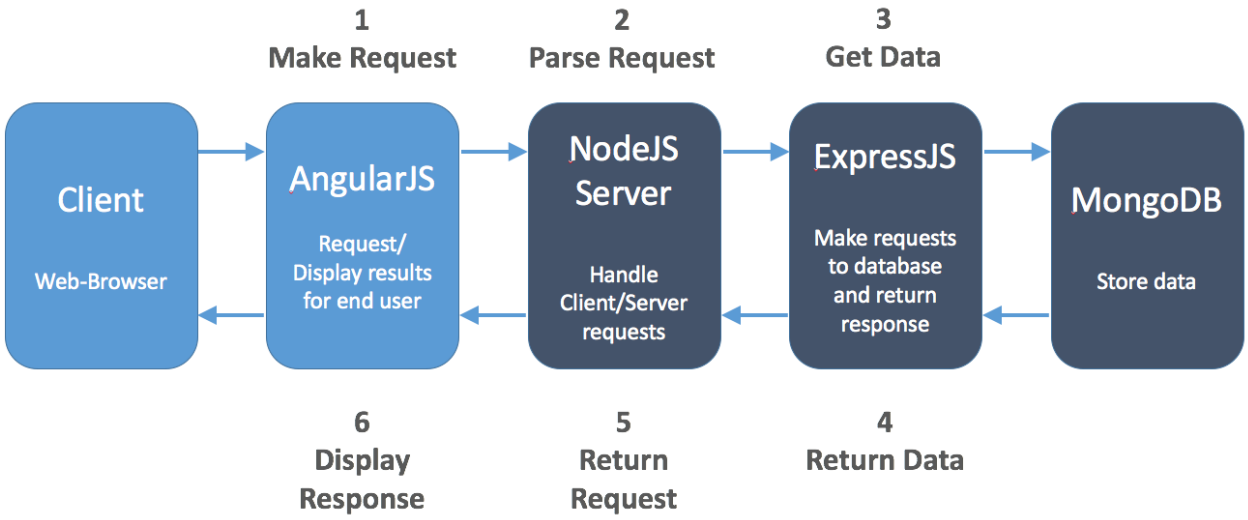


Figure 13: Architecture used for implementation of the interactive program. Components running on the server are shown in dark. The algorithm used for decision tree generation was completely implemented in JavaScript and was exclusively running on the client, which allowed for distributed computation on multiple client computers at once.

Figure 14 shows the dialog used for decision tree generation, which allows selecting different sets of training data as well as various settings for the decision tree algorithm.

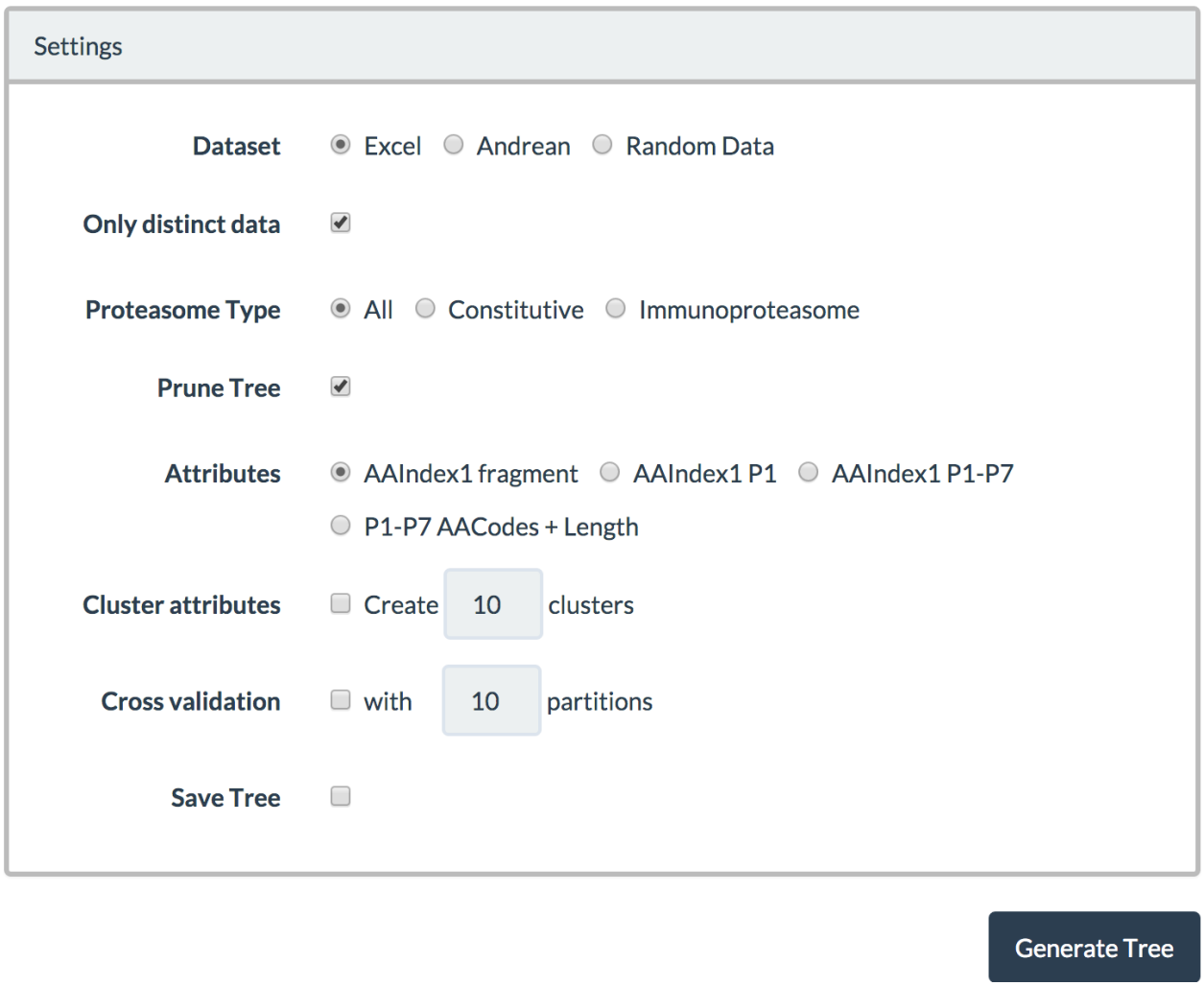


Figure 14: Dialog used for decision tree generation. The dialog allows selection of the dataset, proteasome type, attribute set as well as some other settings for the algorithm

Figure 15 shows another screenshot in which a decision tree is displayed. Once generated, a decision tree can be loaded within seconds and can be explored interactively by clicking its nodes. Each node displays information about the number of fragments as well as the current probability of a fragment being cut by the proteasome. In addition, the program also shows the results of cross validation either in table or in ROC curve form.

Additional parts of the program allow for generation of various statistics and analyses, which were also used in this work.

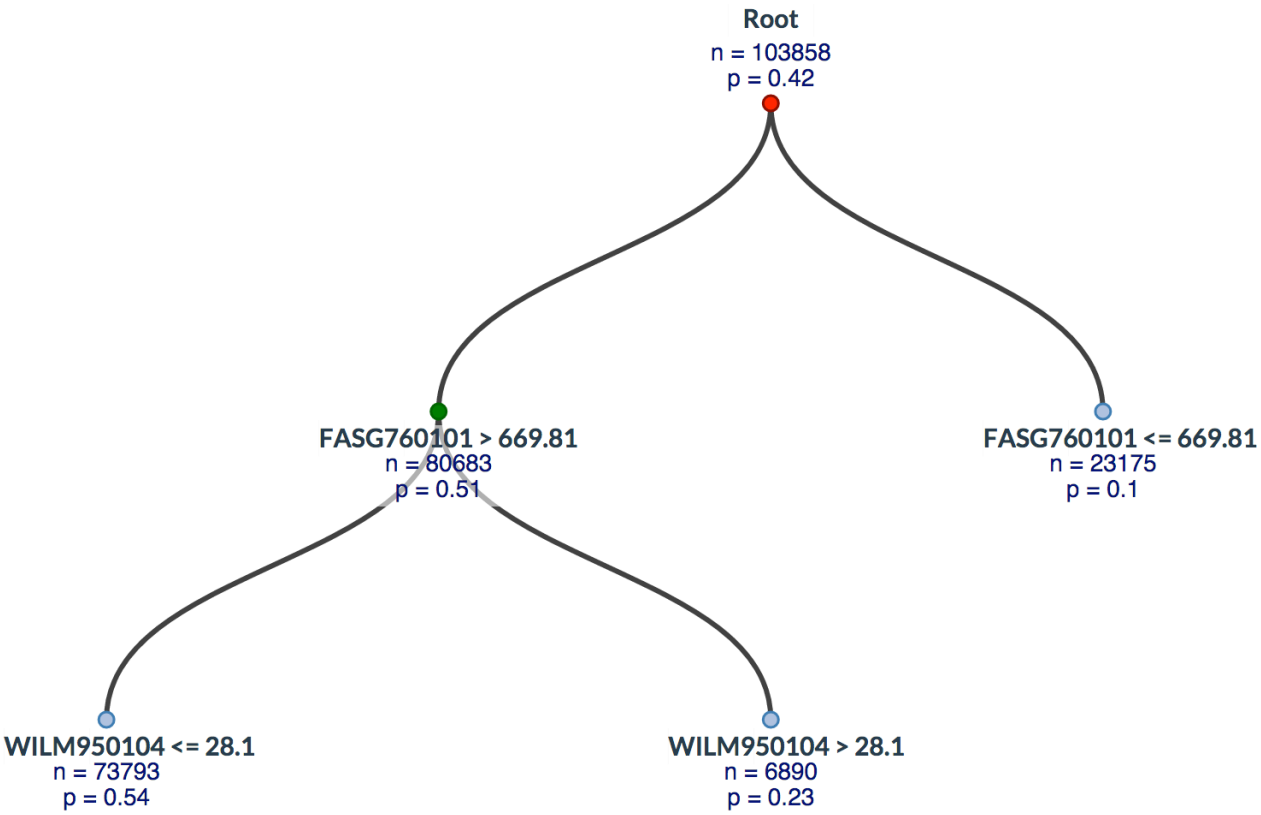


Figure 15: Decision tree within the program. Child nodes can be clicked and are expanded with an animation. Each node shows the number of fragments (n) and the probability of a fragment being created by the proteasome (p). Each level's nodes are sorted by probability p.

Validation of fragment lists created with the Mass Spectrometry File Analyzer

In order to validate the list of fragments retrieved using the Mass Spectrometry File Analyzer and to ensure it included the fragments with the highest level of certainty as listed in the corresponding manual cleavage map, receiver operating characteristics were computed using the manually created cleavage maps as reference: A fragment was counted as true positive (TP) if it occurred in the manual cleavage map, otherwise as false positive (FP). A receiver operating characteristic (ROC) was plotted for each fragment list obtained from the FileAnalyzer by sorting the list by its probability score and calculating the false positive and true positive rates. The area under the curve (AUC) was used as a quality measure for the File Analyzer's fragment list: the AUC value

becomes highest when all fragments of the manual cleavage map were ranked on top of the File Analyzer list. Figure 16 shows a ROC curve for a sample experiment.

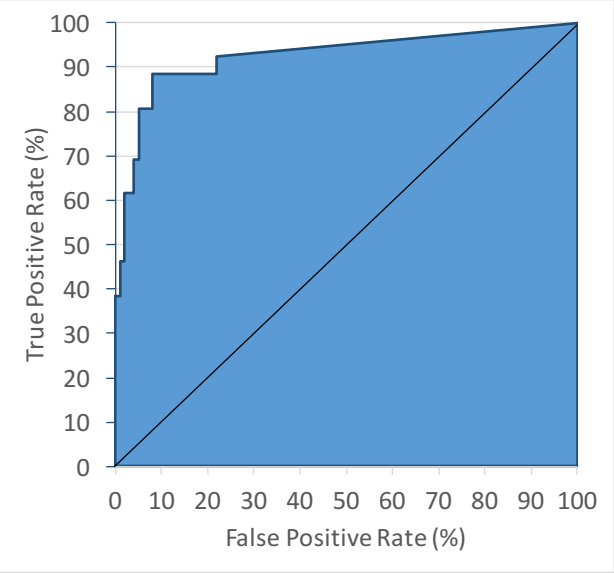


Figure 16: Receiver-operating-characteristic for the File Analyzer fragment list of experiment Kloe686 (AUC 0.92). The fragment list was sorted by the fragments' probability score. A fragment was counted as true positive if it was also listed in the corresponding manually created cleavage map, otherwise it was counted as false positive. The ROC curve was plotted with the false and true positive rates. The area under the curve (AUC) reaches its maximum if all fragments listed in the manually created cleavage map are listed on top of the fragment list.

A combined ROC-curve was created by merging all lists obtained from the Mass Spectrometry FileAnalyzer into an overall list and sorting the fragments again by their probability score. The AUC value for this combined ROC-curve was 0.74 (standard deviation from the AUC of all single ROC-curves: 0.12). The combined ROC-curve is shown in Figure 17.

As mentioned before, the Mass Spectrometry FileAnalyzer was less restrictive and usually detected more fragments than the ones listed in the manually created cleavage maps. However, from the steep rise of most of the ROC-curves we can deduce that the majority of fragments listed in the manually created cleavage maps was also ranked high in the lists obtained from the Mass Spectrometry FileAnalyzer.

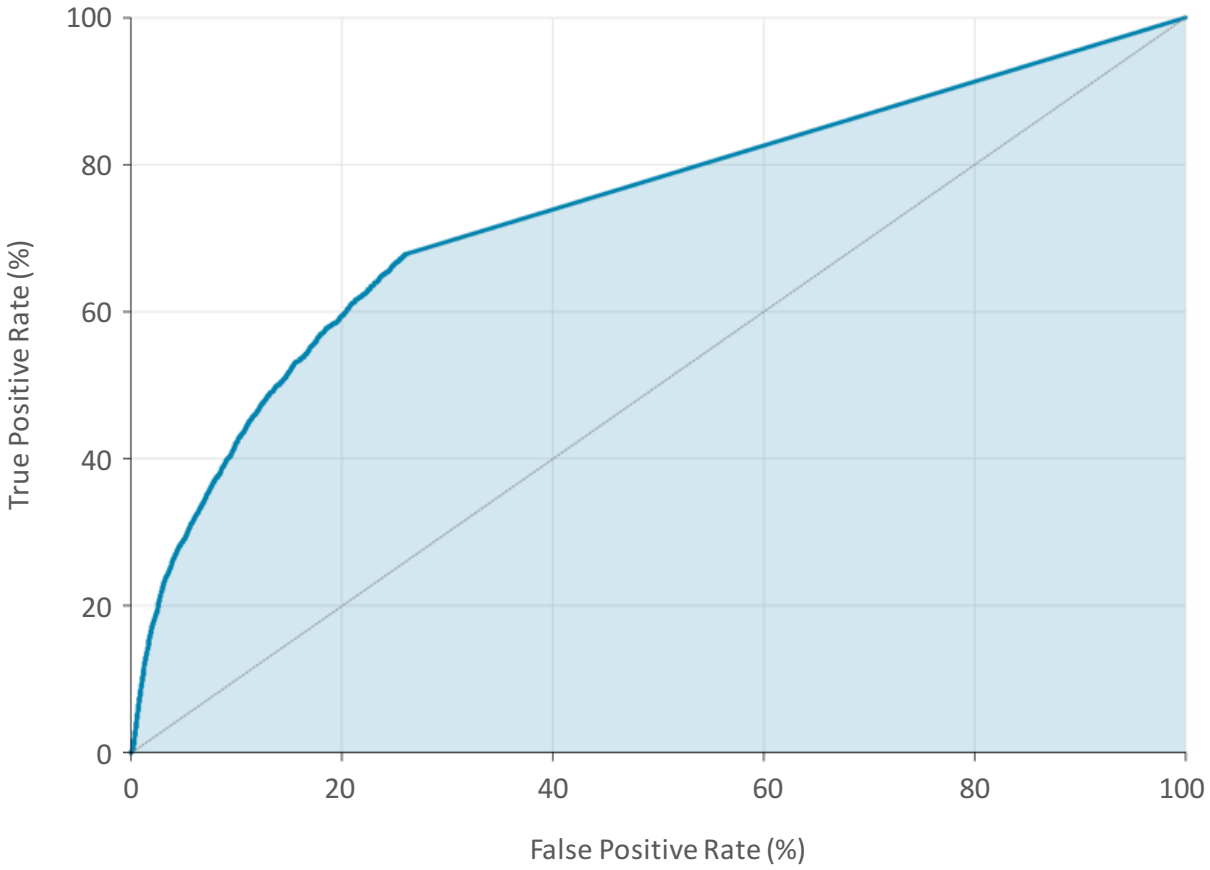


Figure 17: Combined receiver-operating-characteristic of all fragment lists obtained from the Mass Spectrometry File Analyzer (AUC 0.74). All fragment lists were merged into a combined list, whose fragments was sorted by the probability score computed by FileAnalyzer.

Properties of the training dataset

In the beginning, some general properties of the training dataset were evaluated in order to identify potential candidates for attributes. The following statistics include all fragments listed by either the manually created cleavage maps or by the Mass Spectrometry FileAnalyzer.

Distribution of amino acids

Table 9 shows the absolute and relative occurrence of amino acids in the peptides used as substrate for the experiments of the training data set.

Code	Name	Occurrence in training data		Occurrence in vertebrates*	Difference relative occurrence
		absolute	relative	relative	
A	Alanine	121	6.9	7.4	-0.5
C	Cysteine	24	1.4	3.3	-1.9
D	Aspartic acid	54	3.1	5.9	-2.8
E	Glutamic acid	74	4.2	5.8	-1.6
F	Phenylalanine	66	3.8	4.0	-0.2
G	Glycine	149	8.5	7.4	1.1
H	Histidine	42	2.4	2.9	-0.5
I	Isoleucine	86	4.9	3.8	1.1
K	Lysine	81	4.6	7.2	-2.6
L	Leucine	187	10.7	7.6	3.1
M	Methionine	50	2.9	1.8	1.1
N	Asparagine	81	4.6	4.4	0.2
P	Proline	104	5.9	5.0	0.9
Q	Glutamine	63	3.6	3.7	-0.1
R	Arginine	109	6.2	4.2	2.0
S	Serine	127	7.3	8.1	-0.8
T	Threonine	107	6.1	6.2	-0.1
V	Valine	130	7.4	6.8	0.6
W	Tryptophan	28	1.6	1.3	0.3
Y	Tyrosine	67	3.8	3.3	0.5

Table 9: Absolute and relative amino acid occurrence in the training dataset compared to relative observed occurrence in vertebrates (*source: (Dyer 1971))

P1/P1’ pairs

Table 10 and Table 11 show a matrix of all P1/P1’ pairs (including P1/P1’ tail) listed in the manual and the FileAnalyzer cleavage maps. The values shown in the matrix reflect the relative cut frequency and were computed by dividing the number of P1/P1’ sites by their overall occurrence within the training data: The sequence “AA” in Table 10, for example, occurred 12 times in the training data. 326 fragments originated from a cut between two alanine amino acids, which results in a value of 326 / 12 = 27.2. Values above or below the standard deviation were marked. In both matrices, a glutamine at the P1’ position seems to yield a higher cut frequency in multiple cases.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	27.2	25.5	34.8	27.2	29.0	29.1	36.3	24.7	27.5	28.3	30.2	22.0	39.1	25.8	34.0	24.7	30.1	29.8	17.5	28.0
C		27.0	27.0		22.3	**44.0**		29.0	26.0		23.0	25.0	30.0	23.0	29.5	26.0	26.0			27.5
D	32.7		25.0		24.7	34.8	30.0	26.0	36.7	30.0	23.0	38.3	25.4	31.0	29.5	24.4	35.5	44.5		33.8
E	23.3	29.0	**41.7**	26.3	37.0	27.5	29.0	25.5	33.8	32.6	23.0	24.5	27.0	27.8	28.8	28.3	30.5	25.6	18.7	28.0
F	36.3	29.0	25.0	21.0	25.0	23.7	23.0	21.5	35.3	26.3	24.0		35.5	**62.0**	38.0	31.5	27.1	30.2		
G	33.8	28.0	25.3	27.3	31.8	32.4	24.6	25.3	25.3	25.0	33.6	34.3	25.0	29.0	25.1	27.3	28.0	32.0	33.0	37.7
H	30.0	22.0			23.3	32.7		23.5		30.0		33.7	23.5	32.0	24.0	25.5	29.0	27.0		29.0
I	27.5		**44.0**	28.0	30.7	24.8	22.0	22.3	32.0	25.8		29.8	29.7	**61.0**	29.0	23.5	23.3	30.8		23.3
K	28.8	19.0	26.5	29.0	21.0	25.5	25.8	22.0	21.2	29.3	40.5	**48.0**	26.0		25.8	25.0	28.3	23.0		28.5
L	31.5	26.5	32.3	29.8	32.0	24.9	24.0	25.8	28.0	25.9	35.8	26.3	31.4	29.9	38.3	32.5	28.4	29.6		22.0
M			34.8	35.2	37.0	39.0		31.0	30.0	26.2	25.0	**44.5**	34.2	26.0	25.0	25.5	22.0	30.5		24.3
N	34.0	27.0	29.5		**44.0**	30.0	27.0	33.7	23.8	34.5	34.5	34.3	25.1	**60.0**	21.3	35.4	24.3	33.7		35.3
P	23.2			26.3	39.0	24.9	30.5	27.2	23.0	37.1	35.2	31.0	29.2	25.5	26.5	27.9	23.5	27.3	24.5	35.0
Q	26.0		32.8	27.0		33.8	24.0	22.0	29.0	26.8	**42.5**	34.5	27.0	**60.0**	29.3	29.0	22.5	31.1	38.3	28.7
R	28.4		29.0	29.5	24.2	32.4	32.7	28.0	23.3	28.8	26.0	36.8	28.6	23.6	35.6	31.8	29.0	29.8	24.7	
S	32.0		30.0	31.6	26.7	27.0	26.0	22.0	28.0	30.0		37.3	27.7	26.8	26.4	**43.7**	28.6	29.2	32.0	21.3
T	24.3	**43.0**	29.5	25.2	25.0	29.7	27.5	36.3	22.7	31.7	28.3	33.8	26.0	21.5	25.0	23.7	26.7	25.7	32.0	27.0
V	25.0	24.7	26.5		28.8	29.3	27.0	32.6	27.8	31.4	34.8	30.3	32.7	**45.3**	30.2	25.4	28.5	27.0	28.3	27.7
W	27.0			36.0		28.5		31.3	29.0	22.5	24.0	17.5	33.0	31.0	**13.0**	31.0	20.3			35.0
Y	30.0	28.0	25.0	28.5	29.0	25.9		35.0	23.8	25.2		29.0	19.7		29.0	24.3	34.0	25.4	30.0	26.0

Table 10: Cut frequency for P1/P1' pairs in the manual cleavage map dataset. Each row stands for a P1 position, while the columns show the corresponding P1' position. The values shown reflect the relative cut frequency and were computed by dividing the number of P1/P1' sites by their overall occurrence within the training data: The sequence "AA", for example, occurred 12 times in the training data. 326 fragments originated from a cut between two alanine amino acids, which results in a value of 326 / 12 = 27.2. Mean value 29.18, σ = 6.16. Values outside the standard deviation are marked green/red, values outside two times the standard deviation are marked even darker. If a pair did not occur in the test data, the field was left blank. Especially the glutamine (Q) column (P1') shows multiple clearly elevated cut frequencies.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	30.5	30.3	40.5	32.8	32.2	34.0	38.2	27.8	27.3	31.3	32.1	32.0	41.0	35.3	40.7	29.2	32.2	32.4	36.5	40.5
C		29.0	37.2		23.1	**48.0**		29.0	33.0		22.5	37.5	30.8	23.2	31.7	29.0	29.6			31.2
D	37.8		33.0		29.1	37.4	32.6	27.0	40.3	33.5	27.8	41.6	30.5	35.1	34.0	33.2	39.2	44.2		40.9
E	30.0	29.0	**47.9**	28.9	36.9	33.6	29.0	30.0	31.7	36.7	23.0	28.3	32.5	28.3	32.8	29.8	38.2	27.5	26.5	32.3
F	36.7	29.5	36.5	21.0	31.3	29.9	25.0	22.7	43.3	30.5	27.7		32.0	**67.5**	39.2	35.1	34.6	34.7		
G	37.7	29.5	30.1	32.7	35.7	33.1	28.2	30.4	31.0	27.8	35.4	32.4	29.0	33.5	29.3	32.9	37.1	34.7	37.8	39.4
H	30.0	23.0			28.1	38.4		26.6		40.6		40.2	26.8	29.9	29.1	35.2	40.0	34.4		30.8
I	32.6		**48.7**	28.5	29.4	28.0	23.3	22.4	38.5	30.9		29.1	32.0	**70.5**	33.0	27.3	27.0	33.7		26.0
K	39.6	19.0	31.9	32.5	35.5	30.0	30.9	23.0	22.9	31.0	31.9	**49.0**	30.3		28.8	25.3	34.3	27.9		34.0
L	36.0	33.2	36.3	32.5	34.1	29.4	36.5	27.1	31.3	28.9	39.7	28.9	31.6	33.4	42.2	36.9	31.2	37.9		32.9
M			33.3	41.9	37.8	40.2		**47.5**	34.8	29.2	30.0	**47.4**	34.9	26.5	26.0	34.9	22.5	33.5		28.3
N	35.7	31.3	32.3		38.3	37.8	39.0	29.8	29.1	35.8	36.2	37.3	27.6	**64.5**	32.1	39.1	24.5	35.4		37.9
P	24.3			33.3	38.2	29.5	30.5	31.6	23.8	38.0	36.7	33.7	34.2	25.9	30.5	31.8	28.0	32.6	27.8	36.6
Q	29.9		35.9	28.0		35.5	23.2	22.0	33.1	30.6	39.8	38.8	31.5	**68.0**	32.1	30.3	22.4	35.6	37.8	31.5
R	32.8		29.7	32.7	28.5	35.0	36.2	32.2	25.4	30.0	30.3	40.0	32.5	32.1	**47.4**	37.9	36.1	33.6	27.4	
S	34.7		33.1	32.5	32.8	30.1	40.0	23.7	33.0	33.7		38.9	31.7	31.5	31.8	**49.0**	34.5	30.6	37.3	23.7
T	29.1	**47.4**	31.1	32.6	26.6	36.3	35.8	35.5	29.7	34.8	30.9	35.5	27.8	27.7	29.6	26.9	31.9	31.4	38.4	36.4
V	32.3	34.3	32.9		33.0	29.9	36.8	36.6	34.3	35.9	39.9	32.5	35.7	44.8	36.5	29.0	33.8	31.5	32.2	28.2
W	32.8			39.3		36.8		32.2	35.1	22.7	28.8	34.7	36.6	36.6	**14.0**	34.2	25.9			41.3
Y	31.1	31.3	28.1	35.8	30.8	32.8		43.7	28.5	33.1		32.6	32.3		37.0	27.0	36.3	28.4	35.0	36.3

Table 11: Cut frequency for P1/P1' pairs in the FileAnalyzer dataset. Each row stands for a P1 position, while the columns show the corresponding P1' position. The values shown reflect the relative cut frequency and were computed by dividing the number of P1/P1' sites by their overall occurrence within the training data: The sequence "AA", for example, occurred 48 times in the training data. 1436 fragments originated from a cut between two alanine amino acids, which results in a value of 1436 / 48 = 30.5. Mean value 33.07, σ = 6.7. Values outside the standard deviation are marked green/red, values outside two times the standard deviation are marked even darker. If a pair did not occur in the test data the field was left blank. Especially the glutamine (Q) column (P1') shows multiple clearly elevated cut frequencies.

Decision trees

Decision trees were created using different subsets of the training data and different attributes. Figure 18 shows a sample output of the algorithm using the FileAn dataset and the AAIndexFragment attribute set.

Cross-validation

In order to measure the quality of a decision tree's classification, a score was determined using cross validation and a receiver-operating-characteristic. The algorithm, which is explained in more detail in Table 12, uses the following approach: The tree's training data was randomly split into ten groups of the same size. Nine of these groups were used as training data in order to build a decision tree. Each fragment of the remaining group was applied to the tree created, resulting in a probability value for the fragment being created by the proteasome. All fragments of the validation group were ordered by the probability value and added to a receiver-operating-characteristic. The area under the curve (AUC) was calculated to evaluate the tree's performance. The final tree was generated using all data, its quality described by the average AUC of all ten receiver-operating-characteristics computed during cross-validation.

CalculateTreeAUC(groupCount, examples, targetAttribute, attributes)
groupCount: number of groups that should be created for cross validation (a value of 10 was used in this work)
examples: training data used to induce the tree
targetAttribute: attribute whose value is to be predicted by the tree
attributes: Set of attributes to be examined by the algorithm for classification

Returns the average area under the curve (AUC) of all trees created during cross validation. Uses C4.5 for tree creation (see Table 6)

- *summedAUC* = *0*
- *groups* = Randomly split *examples* into *groupCount* groups of the same size
- For each *group* in *groups*
  - *trainingData* = *examples – group*
  - *tree* = C4.5(*trainingData*, *targetAttribute*, *attributes*)
  - For each *fragment* in group
    - *fragment.score* = tree.classifyFragment(fragment)
  - *sortedFragments* = *o*rder fragments by score (descending, highest score first)
  - *roc* = createROC(sortedFragments)
  - *summedAUC* = *summedAUC* + calculateAUC(*roc*)
- return summedAUC / groupCount

Table 12: Algorithm used to calculate a decision tree's average AUC value. The average AUC value was used as a score in order to measure the tree's classification quality. The final decision tree was created after cross validation using all data available.

Figure 19 shows a sample ROC curve of a single cross validation step performed when creating the tree shown in Figure 18.

A comprehensive list of all trees generated and their average area under the curve (AUC) values determined by cross validation is displayed in Table 13. In all cases, trees based on the FileAn dataset showed a higher average AUC value, which can be explained by the very restrictive selection of fragments in the CMap dataset (usually containing only 20-40 fragments): The very small number of positive samples is obviously harder to separate from the negative samples using the attributes provided.

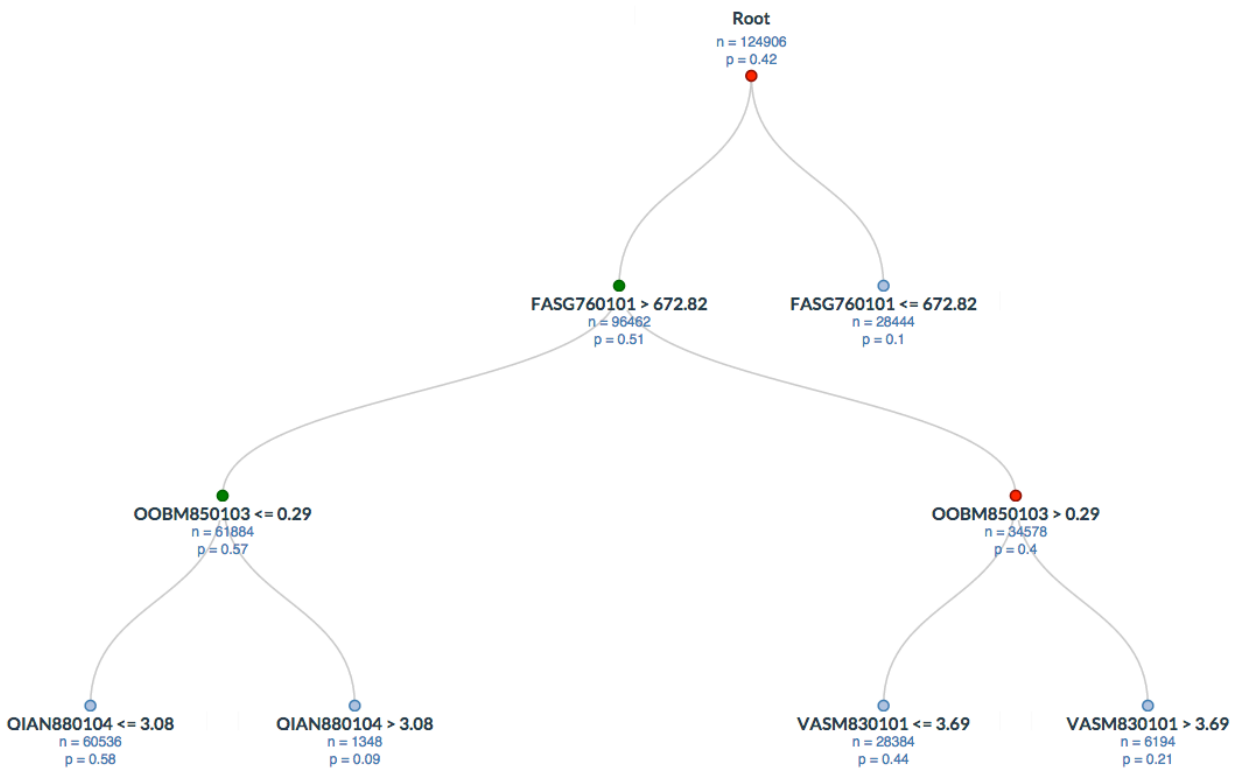


Figure 18: Sample output of the algorithm (dataset: FileAn, attribute set: AAIndexFragment)

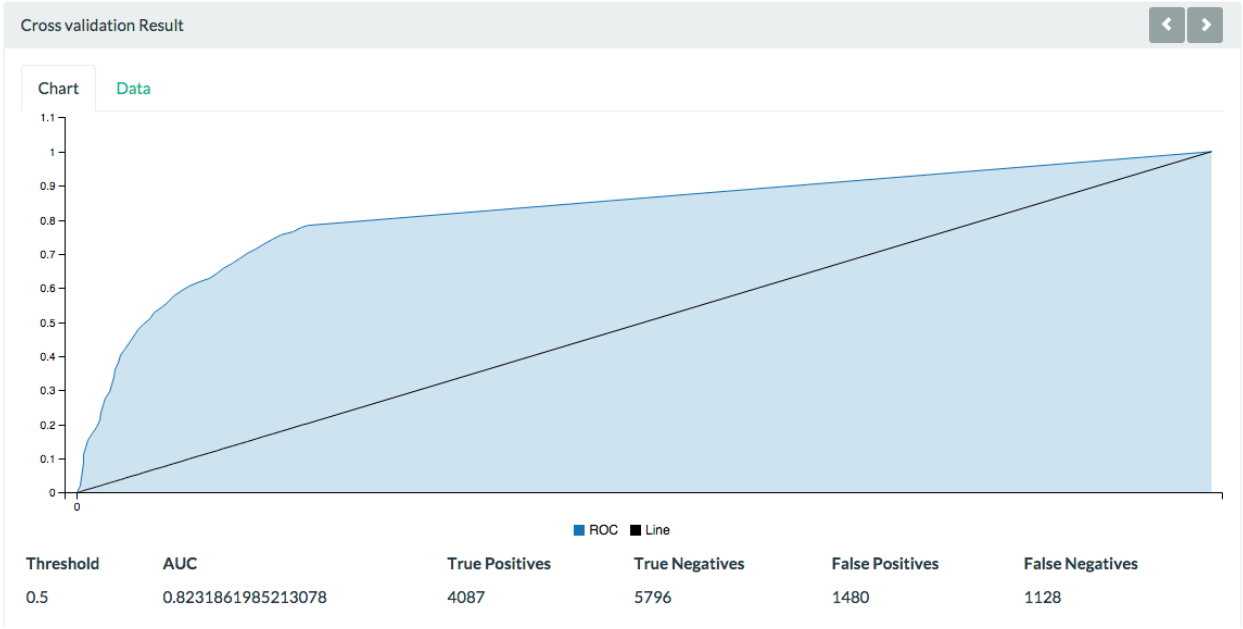


Figure 19: Sample ROC-curve created during cross validation of the tree shown in Figure 18

Attribute Set	Dataset	Average AUC
AACodesP1-P7	CMap	0.58
AACodesP1-P7	CMap*	0.57
AACodesP1-P7	FileAn	0.75
AACodesP1-P7	FileAn*	0.76
AAIndexP1	CMap	0.64
AAIndexP1	CMap*	0.62
AAIndexP1	FileAn	0.78
AAIndexP1	FileAn*	0.78
AAIndexP1#	CMap	0.64
AAIndexP1#	CMap*	0.62
AAIndexP1#	FileAn	0.78
AAIndexP1#	FileAn*	0.79
AAIndexP1-P7	CMap	0.65
AAIndexP1-P7	CMap*	0.65
AAIndexP1-P7	FileAn	0.82
AAIndexP1-P7	FileAn*	0.82
AAIndexP1-P7#	CMap	0.63
AAIndexP1-P7#	CMap*	0.62
AAIndexP1-P7#	FileAn	0.82
AAIndexP1-P7#	FileAn*	0.82
AAIndexFragment	CMap	0.60
AAIndexFragment	CMap*	0.57
AAIndexFragment	FileAn	0.82
AAIndexFragment	FileAn*	0.83
AAIndexFragment#	CMap	0.50
AAIndexFragment#	CMap*	0.50
AAIndexFragment#	FileAn	0.69
AAIndexFragment#	FileAn*	0.70

Table 13: List of all 28 decision trees created

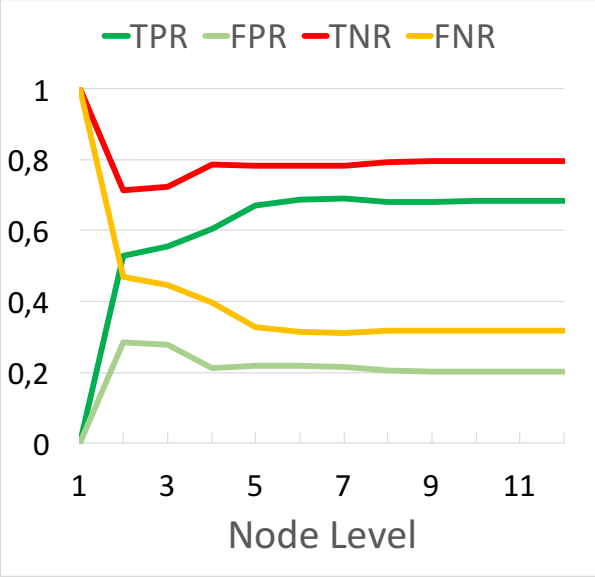
Fitting of training data

In addition to the average AUC value, we can also examine how well a tree fits the training data by looking at the true positive rate (TPR), true negative rate (TN), false positive rate (FPR) and false negative rate (FNR). It is especially interesting to see down to which node level we have to traverse the tree until we reach good rate values: If there are only a few important attributes, which decide if a fragment is created by the proteasome or not, these attributes should have been found by the tree and be associated with a node on top of the tree. We should therefore be able to observe a fast improvement of the TPR and TNR with each node level. However, if it takes many node levels until we reach a good TPR and TNR, this might indicate that there are no specific attributes relevant for the classification.

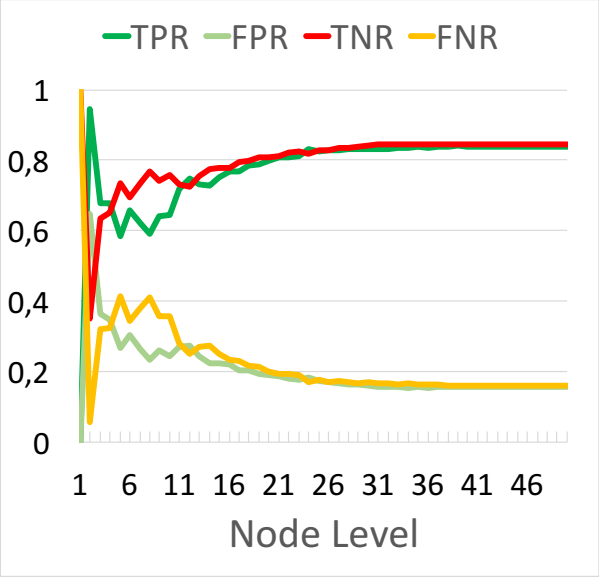
Figure 20 shows the TPR, TNR, FPR and FNR by node level for a selection of trees. Every tree has a TPR of 0 and a TNR of 1 at node level 1. This is due to the fact that all the datasets include more negative than positive samples. A tree with only one node therefore selects the classification of the majority of samples. The deeper we move down the tree, the more attributes were used in order to separate the positive from the negative samples, which usually leads to an increasing TPR.

While the trees created with the AACodesP1-P7 and AAIndexFragment attribute sets and the FileAn/FileAn* datasets showed a rather quick increase of the TPR, the remaining trees featured a rather slow increase, often requiring 10 node levels until reaching a TPR above 0.5. The trees for the attribute set AAIndexFragment# and the CMap dataset were not able to classify the training data correctly. Apparently, ten different attributes do not suffice in order to separate the small set of fragments detected in CMap from the remaining fragments.

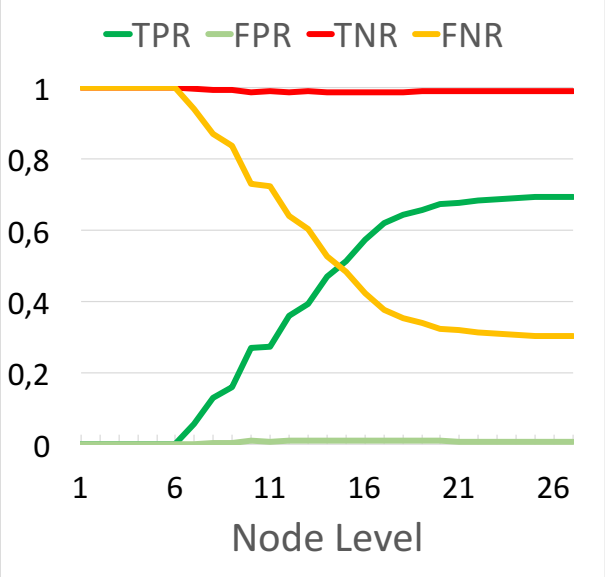
Tree 1
AACodesP1-P7, FileAn



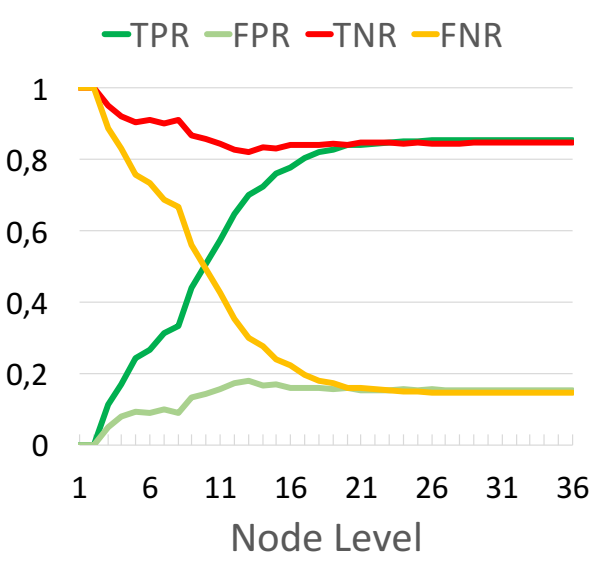
Tree 2
AAIndexFragment, FileAn



Tree 3
AAIndexP1, CMap



Tree 4
AAIndexP1-P7, FileAn



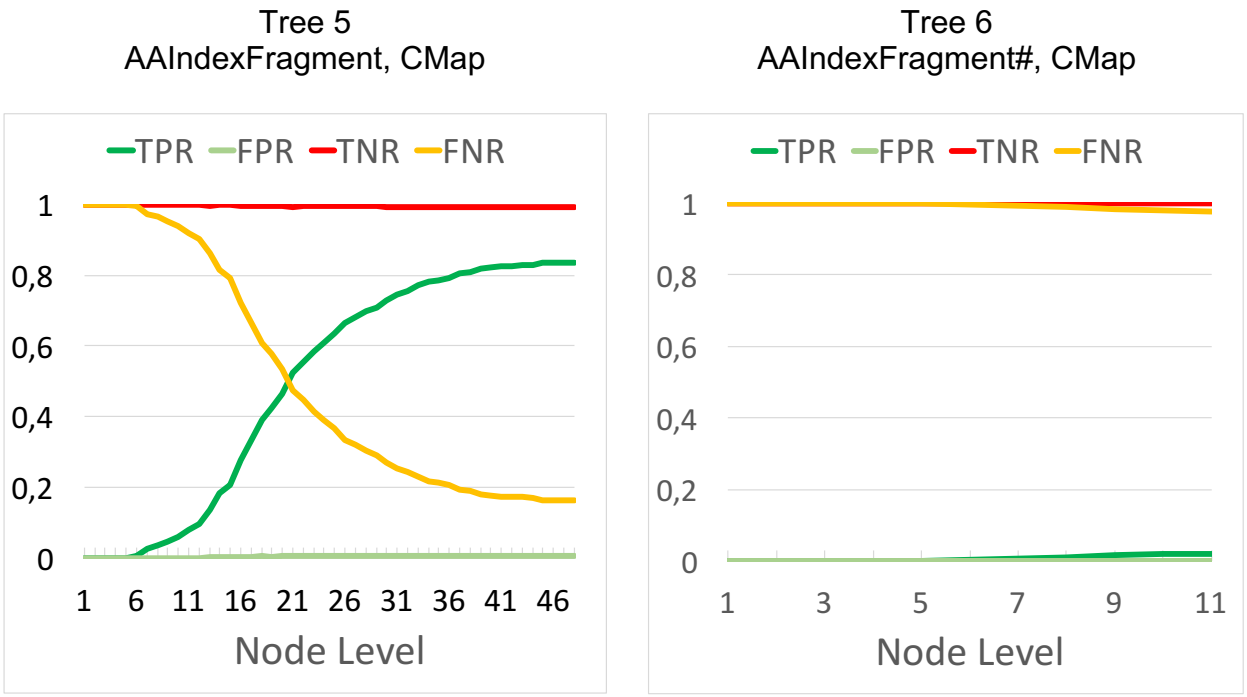


Figure 20: True positive rate (TPR), false positive rate (FPR), true negative rate (TNR) and false negative rate (FNR) (y-axis) by node level (x-axis) for selected trees. On node level 1, each tree features a TPR of 0 because the majority of samples in all datasets was negative. Trees 1 and 2: These trees show a rather fast increase of the TPR indicating the attributes associated with the first node levels are especially important. Trees 3-5: The trees shown here exemplify the majority of trees; in most cases the TPR is increasing steadily but slowly, which points to no attribute being of special importance. Bottom row: this tree is not able to classify the training data most likely due to the small share of positive samples in the CMap dataset and only ten attributes being supplied by the AAIndexFragment# attribute set

Most relevant attributes

In each recursive step the decision tree algorithm selects the attribute, which minimizes the information entropy, using the gain ratio criterion. Put differently, the algorithm selects the attribute that separates the current's node's data best into the different classifications. Thus the higher an attribute's position in a tree, the more relevant it is for a general decision about the question if a fragment is created by the proteasome. Admittedly, this thinking is a little simplified since a greedy algorithm was used, which is prone to running into local minima. Still, looking at the attributes of the first tree levels seems interesting. In the following, an overview is given of which attributes have been selected first by the decision tree algorithm.

Amino Acid Letter Codes

As discussed in the section before, the trees trained with the CMap datasets only achieved a true positive rate of 0.3. It can therefore be assumed that the fragments listed in the manually created cleavage maps cannot be separated from the remaining fragments by just looking at the amino acid letter codes at positions P1-P7. However, both trees (CMap and CMap*) selected the attribute "P1" as first attribute, which correlates with the importance of the P1 cleavage site as described in various sources (Ossendorp et al. 1996; Beekman et al. 2000; Del Val et al. 1991).

The trees trained with the FileAn and FileAn* datasets reach a better true positive rate of almost 0.7. Both select the attribute "Fragment Length" as first attribute. In multiple branches of both trees, a certainty of over 90% for a fragment being created by the proteasome is reached within 3 steps (see Figure 21 for an example).

Table 14 shows the first two levels of the AACodesP1-P7 tree trained with the FileAn dataset. For fragment lengths between 5 and 25 amino acids, P1 was selected as second attribute.

The results of the AACodesP1-P7 (FileAn*) tree were very similar and are therefore not described to the same extent as those of AACodesP1-P7 (FileAn).

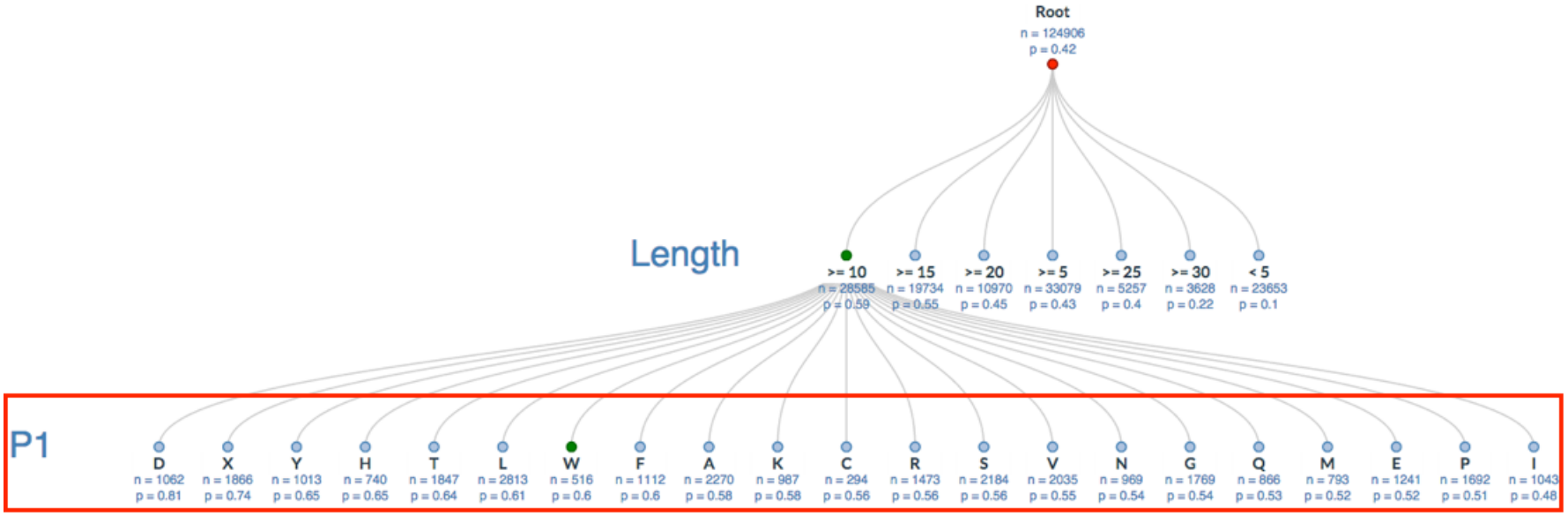


Figure 21: Decision tree with attribute set AACodesP1-P7 and FileAn dataset. Each level's nodes are sorted by their probability. In the first branch, a certainty of up to 0.81 (81%) is reached after three steps only. The red rectangle is referenced in Table 14.

Fragment length < 5 (n = 23653, p = 0.10)																					
P3 tail	**A**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**	**X**
n	1566	300	787	1082	915	1870	538	1003	901	2515	616	969	1442	848	1232	1662	1570	1774	440	848	775
p	0.08	0.08	0.10	0.11	0.11	0.07	0.12	0.08	0.09	0.14	0.09	0.08	0.08	0.10	0.13	0.09	0.12	0.12	0.17	0.09	0
Fragment length >= 5 and < 10 amino acids (n = 33079, p = 0.43)																					
P1	**A**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**	**X**
n	2377	392	1046	1337	1291	2326	768	1266	1076	3511	901	1384	1889	1237	1741	2424	2180	2619	524	1119	1670
p	0.43	0.46	**0.58**	0.34	0.47	0.40	0.44	0.33	0.36	0.46	0.40	0.40	0.33	0.39	0.46	0.42	0.42	0.45	0.40	0.48	**0.63**
Fragment length >= 10 and < 15 amino acids (n = 28585, p = 0.59)																					
P1	**A**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**	**X**
n	2270	294	1062	1241	1112	1769	740	1043	987	2813	793	969	1692	866	1473	2184	1847	2035	516	1013	1866
p	**0.58**	**0.56**	**0.81**	**0.52**	**0.58**	**0.54**	**0.64**	0.48	**0.58**	**0.60**	**0.52**	**0.54**	**0.51**	**0.53**	**0.56**	**0.56**	**0.61**	**0.55**	**0.60**	**0.65**	**0.74**
Fragment length >= 15 and < 20 amino acids (n = 19734, p = 0.55)																					
P1	**A**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**	**X**
n	1571	249	809	820	612	1252	451	800	758	1886	496	570	1237	511	992	1577	1200	1094	350	726	1773
p	**0.52**	0.49	**0.73**	**0.56**	**0.62**	**0.52**	**0.56**	0.47	**0.55**	**0.56**	0.39	0.43	0.47	0.47	**0.51**	**0.54**	**0.57**	**0.54**	0.42	**0.55**	**0.69**
Fragment length >= 20 and < 25 amino acids (n = 10970, p = 0.45)																					
P1	**A**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**	**X**
n	947	167	283	460	376	606	223	437	419	841	303	321	722	238	438	887	600	640	229	317	1516
p	0.42	**0.63**	**0.57**	0.48	**0.50**	0.39	**0.51**	0.32	0.41	0.40	0.24	0.34	0.43	0.30	0.37	0.47	0.39	0.39	0.28	0.48	**0.69**
Fragment length >= 25 and < 30 amino acids (n = 5257, p = 0.40)																					
P7	**A**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**	**X**
n	73	35	72	64	73	108	21	71	32	154	27	128	14	53	36	60	128	81	32	10	3985
p	0.16	0	0.13	0.36	0.18	0.14	0.14	0.06	0.28	0.30	0.04	0.13	0.14	0.21	0.08	0.27	0.12	0.11	0.16	0	0.48
Fragment length >= 30 (n = 3628, p = 0.22)																					
P1' tail	**A**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**	**X**
n	166	0	206	108	156	182	14	49	50	386	113	257	164	104	205	217	30	364	0	169	688
p	0.30	0	0.13	0.26	0.06	0.12	0.29	0.31	0.32	0.05	0.12	0.09	0.04	0.30	0.26	0.09	0.20	0.10	0	0.11	**0.57**

Table 14: Level 1 and 2 of tree AACodesP1-P7, FileAn. n: number of fragments included in node, p: probability of a fragment being created by the proteasome (values > 0.5 highlighted). X: No amino acid at corresponding position. The red rectangle is referenced in Figure 21.

AAIndex Attributes

The trees created with the AAIndex attribute sets are rather complex, since they include a multitude of different attributes and each attribute can be used in multiple sub trees. Furthermore, there are 24 trees created with AAIndex attributes. This multitude of results needs to be condensed into summaries, which can then be searched for patterns.

In the following tables, the AAIndex attribute clusters are color-coded. An attribute belonging to a certain cluster is shown in its corresponding cluster's color. A legend of the color codes is given in Table 15. The colors have been chosen arbitrarily, similar colors do not describe a similarity between individual clusters.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10

Table 15: Color coding used for the 10 clusters of AAIndex attributes. The colors have been chosen arbitrarily, similar colors do not suggest a similarity between individual clusters.

First Levels of decision trees

Table 16 shows a matrix of all trees and the attributes used within their first three node levels. The occurrence of each cluster in the matrix is summarized Figure 22. While the matrix does not reveal a very noticeable pattern, it can still be seen that cluster 6 (representative attribute: 'Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit') occurs three to four-times as often as other clusters.

Attribute Set	Dataset	Level 1	Level 2.1	Level 2.2	Level 3 (2.1)	Level 3 (2.1)	Level 3 (2.2)	Level 3 (2.2)
AAIndexP1	CMap	CORJ870108 P1	QIAN880133 P1	QIAN880116 P1 tail	FASG890101 P1	ZIMJ680102 P1' tail	ROSM880103 P1' tail	RICJ880113 P1' tail
AAIndexP1	CMap*	CORJ870108 P1	CORJ870107 P1	QIAN880116 P1 tail	KARP850101 P1 tail	RICJ880114 P1' tail	TANS770107 P1' tail	RICJ880113 P1' tail
AAIndexP1#	CMap	Cluster 1 P1	Cluster 8 P1	Cluster 8 P1 tail	Cluster 2 P1	Cluster 6 P1' tail	Cluster 9 P1	Cluster 8 P1' tail
AAIndexP1#	CMap*	Cluster 1 P1	Cluster 6 P1	Cluster 9 P1 tail	Cluster 5 P1	Cluster 9 P1	Cluster 9 P1	Cluster 6 P1' tail
AAIndexP1	FileAn	RADA880105 P1' tail	RADA880104 P1' tail	TANS770106 P1'	BASU050103 P1 tail	KIMC930101 P1 tail	RACS820107 P1	KUHL950101 P1
AAIndexP1	FileAn*	KRIW790101 P1	VHEG790101 P1	KHAG800101 P1'	FASG890101 P1	PONP800104 P1'	GEOR030107 P1' tail	ISOY800104 P1' tail
AAIndexP1#	FileAn	Cluster 6 P1	Cluster 1 P1	Cluster 6 P1'	Cluster 2 P1	Cluster 8 P1	Cluster 9 P1	Cluster 8 P1
AAIndexP1#	FileAn*	Cluster 6 P1	Cluster 1 P1	Cluster 6 P1'	Custer 2 P1	Cluster 8 P1	Cluster 9 P1	Cluster 7 P1'
AAIndexP1-P7	CMap	CORJ870108 P1	QIAN880133 P1	QIAN880116 P1 tail	FASG890101 P1	ZIMJ680102 P1' tail	BULH740102 P3	RACS820113 P2' tail
AAIndexP1-P7	CMap*	CORJ870108 P1	CORJ870107 P1	QIAN880116 P1 tail	KARP850101 P1 tail	RICJ880114 P1' tail	FINA910101 P3	RICJ880113 P1' tail
AAIndexP1-P7#	CMap	Cluster 1 P1	Cluster 8 P1	Cluster 1 P3	Cluster 2 P1	Cluster 7 P6	Cluster 6 P1	Cluster 6 P4
AAIndexP1-P7#	CMap*	Cluster 1 P1	Cluster 6 P1	Cluster 9 P1 tail	Cluster 5 P1	Cluster 9 P1	Cluster 9 P1	Cluster 6 P1' tail
AAIndexP1-P7	FileAn	RADA880105 P1' tail	RADA880104 P1' tail	TANS770105 P6 tail	RICJ880112 P4 tail	TANS770106 P7 tail	NAKH920106 P1 tail	GEOR030102 P3 tail
AAIndexP1-P7	FileAn*	RACS820107 P5 tail	ISOY800107 P4' tail	WILM950104 P6 tail	CHOP780206 P4 tail	RACS820106 P3 tail	WOLS870103 P5 tail	PARS000102 P4 tail
AAIndexP1-P7#	FileAn	Cluster 4 P2	Cluster 6 P6	Cluster 4 P7 tail	Cluster 6 P1	Cluster 3 P1	Cluster 3 P5'	Cluster 4 P6' tail
AAIndexP1-P7#	FileAn*	Cluster 4 P2	Cluster 6 P6	Cluster 4 P7 tail	Cluster 6 P1	Cluster 3 P1	Cluster 6 P6 tail	Cluster 4 P6' tail
AAIndexFragment	CMap	ZHOH040101	WERD780103	GRAR740103	TAKK010101	MEEJ800101	NAKH900105	WILM950101
AAIndexFragment	CMap*	ZHOH040102	WERD780103	WOLS870103	TAKK010101	MEEJ800101	GUYH850104	
AAIndexFragment#	CMap	Cluster 3	Cluster 1	Cluster 4	Cluster 6	Cluster 5	Cluster 9	Cluster 8
AAIndexFragment#	CMap*	Cluster 3	Cluster 1	Cluster 4	Cluster 7	Cluster 9	Cluster 5	
AAIndexFragment	FileAn	FASG760101	OOBM850103	OOBM770105	QIAN880104	VASM830101	OOBM850102	KIMC930101
AAIndexFragment	FileAn*	FASG760101	WILM950104	OOBM770105	OOBM850103	NAKH900106	PALJ810113	KIMC930101
AAIndexFragment#	FileAn	Cluster 3	Cluster 6	Cluster 4	Cluster 10	Cluster 1	Cluster 5	Cluster 6
AAIndexFragment#	FileAn*	Cluster 3	Cluster 1	Cluster 4	Cluster 10	Cluster 10	Cluster 5	Cluster 6

Table 16: AAIndex attributes used in the first three levels of all trees generated. Attributes are color-coded according to the clusters they belong to.

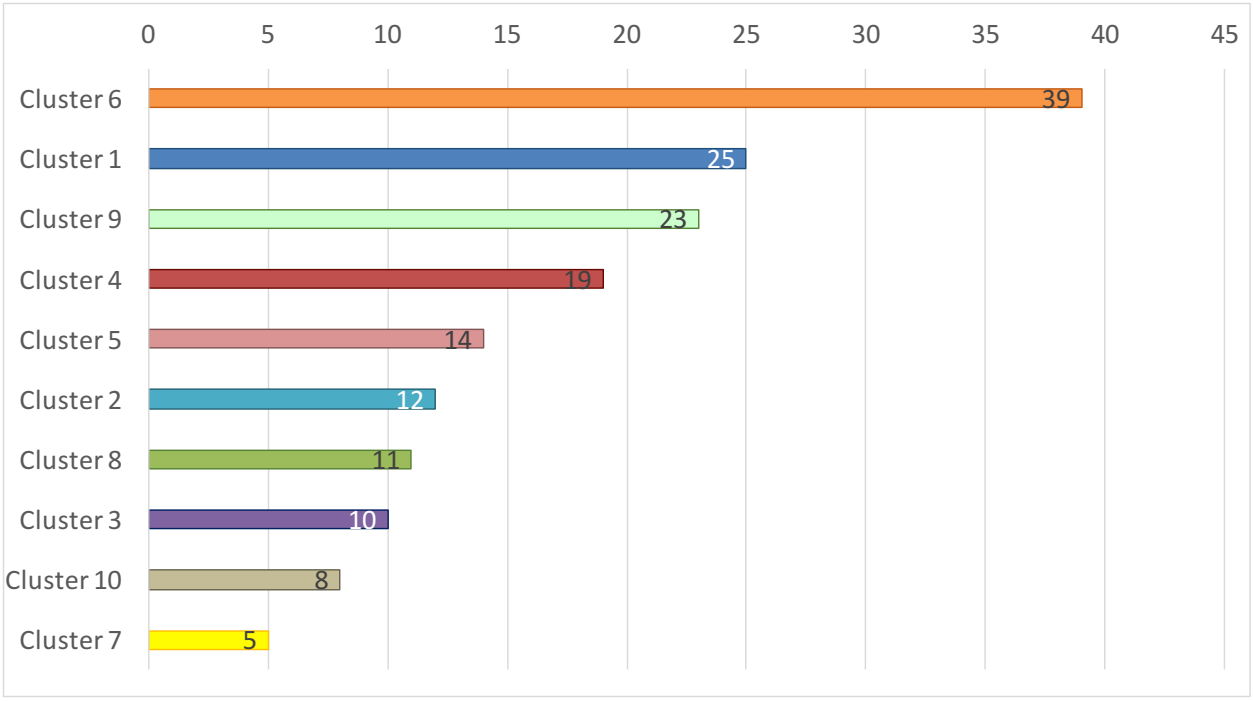


Figure 22: Cluster occurrence in the first three levels of all trees generated with the AAIndex attribute set (also see Table 16). Most attributes occurring within the first three levels belong to cluster 6, while only five attributes from cluster 7 are used. Even though the frequency of the clusters in between varies, no other cluster sticks out in particular.

Attribute overall information gain

The matrix shown in Table 17 shows each tree's attributes with the highest overall information gain (*oGain*), which was calculated by adding up the information gain in every node the attribute occurred in, while accounting for the number of fragments in the training data affected by the node:

$$oGain(attribute, tree) = \sum_{node \in Nodes} \frac{|node.fragments|}{|trainingData|} Gain(node)$$

where *Nodes* = { tree.nodes | node.attribute = attribute}.

Attribute Set	Dataset	1.	2.	3.	4.	5.	6.	7.	8.
AAIndexP1	CMap	FASG890101 P1' tail	FASG890101 P1'	FASG890101 P1 tail	FASG890101 P1	CORJ870108 P1	WOEC730101 P1' tail	QIAN880133 P1	ENGD860101 P1
AAIndexP1	CMap*	FASG890101 P1'	CORJ870107 P1	FASG890101 P1	CORJ870108 P1	FASG890101 P1 tail	FASG890101 P1' tail	ENGD860101 P1'	ENGD860101 P1 tail
AAIndexP1#	CMap	Cluster 2 P1' tail	Cluster 1 P1'	Cluster 1 P1' tail	Cluster 2 P1	Cluster 5 P1	Cluster 1 P1 tail	Cluster 2 P1'	Cluster 2 P1 tail
AAIndexP1#	CMap*	Cluster 1 P1'	Cluster 1 P1 tail	Cluster 1 P1' tail	Cluster 2 P1' tail	Cluster 5 P1	Cluster 3 P1'	Cluster 2 P1	Cluster 2 P1'
AAIndexP1	FileAn	FASG890101 P1 tail	FASG890101 P1' tail	FASG890101 P1'	FASG890101 P1	ENGD860101 P1'	ENGD860101 P1' tail	ENGD860101 P1 tail	ENGD860101 P1
AAIndexP1	FileAn*	FASG890101 P1 tail	FASG890101 P1' tail	FASG890101 P1'	FASG890101 P1	ENGD860101 P1 tail	ENGD860101 P1' tail	ENGD860101 P1'	ENGD860101 P1
AAIndexP1#	FileAn	Cluster 1 P1 tail	Cluster 1 P1' tail	Cluster 2 P1' tail	Cluster 1 P1'	Cluster 2 P1 tail	Cluster 5 P1 tail	Cluster 1 P1	Cluster 5 P1' tail
AAIndexP1#	FileAn*	Cluster 1 P1' tail	Cluster 1 P1'	Cluster 1 P1 tail	Cluster 2 P1 tail	Cluster 2 P1' tail	Cluster 2 P1	Cluster 1 P1	Cluster 2 P1'
AAIndexP1-P7	CMap	CORJ870108 P1	WOEC730101 P1' tail	QIAN880133 P1	ZIMJ680102 P1' tail	FASG890101 P7' tail	FASG890101 P1	FASG890101 P7 tail	QIAN880133 P1 tail
AAIndexP1-P7	CMap*	CORJ870107 P1	CORJ870108 P1	RICJ880115 P1' tail	RICJ880114 P1' tail	CORJ870102 P1' tail	PTIO830101 P1 tail	QIAN880116 P1 tail	ZHOH040103 P1 tail
AAIndexP1-P7#	CMap	Cluster 1 P1	Cluster 1 P7 tail	Cluster 5 P1	Cluster 1 P7' tail	Cluster 8 P1	Cluster 2 P1	Cluster 1 P7'	Cluster 6 P1 tail
AAIndexP1-P7#	CMap*	Cluster 5 P1	Cluster 1 P1	Cluster 6 P1	Cluster 5 P1' tail	Cluster 1 P7' tail	Cluster 1 P7 tail	Cluster 1 P7'	Cluster 6 P1' tail
AAIndexP1-P7	FileAn	FASG890101 P7' tail	FASG890101 P7'	FASG890101 P7 tail	FASG890101 P7	FASG890101 P6'	FASG890101 P6' tail	FASG890101 P6 tail	FASG890101 P6
AAIndexP1-P7	FileAn*	FASG890101 P7' tail	FASG890101 P7 tail	FASG890101 P7'	FASG890101 P7	FASG890101 P6'	FASG890101 P6' tail	FASG890101 P6 tail	FASG890101 P5'
AAIndexP1-P7#	FileAn	Cluster 1 P7' tail	Cluster 1 P7 tail	Cluster 1 P7'	Cluster 1 P7	Cluster 1 P6 tail	Cluster 1 P6' tail	Cluster 1 P6'	Cluster 1 P5'
AAIndexP1-P7#	FileAn*	Cluster 1 P7' tail	Cluster 1 P7 tail	Cluster 1 P7'	Cluster 1 P7	Cluster 1 P6'	Cluster 1 P6 tail	Cluster 1 P6' tail	Cluster 1 P5 tail
AAIndexFragment	CMap	ZHOH040101	FASG890101	ENGD860101	WERD780103	QIAN880125	SNEP660102	MIYS990105	SNEP660103
AAIndexFragment	CMap*	ZHOH040102	FASG890101	ENGD860101	WERD780103	ROBB760109	SNEP660104	MIYS990105	SUYM030101
AAIndexFragment#	CMap	Cluster 7	Cluster 8	Cluster 4	Cluster 10	Cluster 9	Cluster 3	Cluster 6	Cluster 2
AAIndexFragment#	CMap*	Cluster 5	Cluster 2	Cluster 4	Cluster 6	Cluster 8	Cluster 10	Cluster 3	Cluster 1
AAIndexFragment	FileAn	FASG760101	FASG890101	ENGD860101	CORJ870107	OOBM770105	MIYS990105	OOBM850103	MIYS990104
AAIndexFragment	FileAn*	FASG760101	FASG890101	ENGD860101	MIYS990105	CORJ870107	WILM950104	OOBM770105	OOBM850103
AAIndexFragment#	FileAn	Cluster 5	Cluster 2	Cluster 9	Cluster 3	Cluster 1	Cluster 7	Cluster 8	Cluster 10
AAIndexFragment#	FileAn*	Cluster 7	Cluster 3	Cluster 8	Cluster 6	Cluster 10	Cluster 4	Cluster 9	Cluster 2

Table 17: Attributes with highest overall information gain (oGain) in each tree (top eight attributes). Attributes are color-coded according to the clusters they belong to.

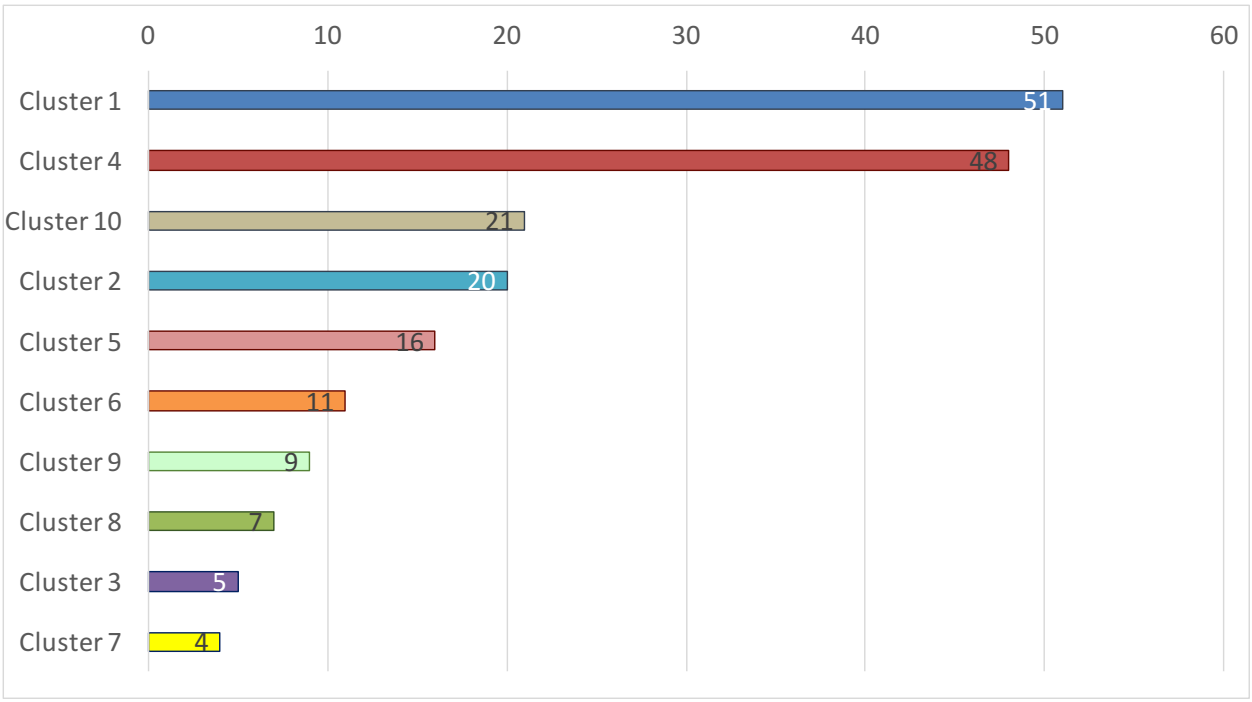


Figure 23: Cluster occurrence of the top eight attributes with the highest information gain in all decision trees generated with an AAIndex based attribute set (also see).

As can be seen, Cluster 6 loses its distinctive role when sorting the attributes by their information gain. While the trees created with the AAIndexP1-P7 attribute set stand out because of their consistent occurrence of the same attribute/cluster within all of the top eight attributes, the results are not conclusive: While in the unclustered version (AAIndexP1-P7), cluster 4 plays the most important role, cluster 1 is chosen in the clustered version (AAIndexP1-P7#).

In summary, the trees created with AAIndex-based attributes show rather inconclusive results: Even though some of the clusters stand out, they do not do so consistently in all trees and/or both types of analyses (attributes sorted by nodes vs. attributes sorted by oGain).

Tree validation with data of an enolase digestion experiment

Data of an enolase digestion experiment (A K Nussbaum et al. 1998) was applied to the decision trees and receiving operator characteristics were calculated to measure the trees' prediction quality.

Table 18 shows how the trees performed predicting the fragments of the experiment. In summary, the results are rather unsatisfactory. The majority of trees achieve an area under the curve near to 0.5. The best performing tree (AACodesP1-P7, FileAn*) reaches an AUC of 0.64. There are multiple explanations for these poor results: First, all substrates in the training data were significantly shorter (average: 55 amino acids) compared to the enolase substrate (437 amino acids). Furthermore, in proportion to the long substrate, only a very small number of cleavage products (81) was detected in the enolase experiment. The training data used in this work featured a far bigger ratio of true positives, especially within the FileAn datasets. This might lead to the decision trees exhibiting a high rate of false positives when applied to the enolase experiment data.

Attribute-Set	Dataset	AUC	TP	TN	FP	FN
AACodesP1-P7	CMap	0.50	2	93229	1958	79
AACodesP1-P7	Cmap*	0.49	1	91860	3327	80
AACodesP1-P7	FileAn	0.64	29	87855	7332	52
AACodesP1-P7	FileAn*	0.64	29	88534	6653	52
AAIndexP1	CMap	0.48	5	84417	10770	76
AAIndexP1	CMap*	0.47	5	83688	11499	76
AAIndexP1	FileAn	0.52	32	61639	33548	49
AAIndexP1	FileAn*	0.55	32	64583	30604	49
AAIndexP1#	CMap	0.48	5	85775	9412	76
AAIndexP1#	CMap*	0.47	5	83536	11651	76
AAIndexP1#	FileAn	0.55	34	61047	34140	47
AAIndexP1#	FileAn*	0.53	32	62560	32627	49
AAIndexP1-P7	CMap	0.45	1	84844	10343	80
AAIndexP1-P7	CMap*	0.47	5	82741	12446	76
AAIndexP1-P7	FileAn	0.47	32	51572	43615	49
AAIndexP1-P7	FileAn*	0.39	22	47772	47415	59
AAIndexP1-P7#	CMap	0.42	4	74914	20273	77
AAIndexP1-P7#	CMap*	0.49	5	85979	9208	76
AAIndexP1-P7#	FileAn	0.46	27	54841	40346	54
AAIndexP1-P7#	FileAn*	0.52	27	60116	35071	54
AAIndexFragment	CMap*	0.29	9	40676	54511	72
AAIndexFragment	CMap	0.36	10	54302	40885	71
AAIndexFragment	FileAn*	0.36	38	28863	66324	43
AAIndexFragment	FileAn	0.37	29	38608	56579	52
AAIndexFragment#	CMap*	0.21	0	40684	54503	81
AAIndexFragment#	CMap	0.34	0	64155	31032	81
AAIndexFragment#	FileAn*	0.22	43	22892	72295	38
AAIndexFragment#	FileAn	0.19	29	24639	70548	52

Table 18: Results of validation with data of an enolase digestion experiment. The area under the curve (AUC) is near to 0.5 for most of the trees, which reflects a poor prediction performance. The best performing trees (marked green) reach an AUC of 0.64. The trees without any true positive hit (AAIndexFragment#, CMap) had already shown poor results during cross validation as discussed before.

Tree validation with MHC I ligand data

In order to validate the quality of prediction achieved by each individual decision tree, all 4026 MHC-I ligands (HLA-A, HLA-B and HLA-C) contained in the SYFPEITHI database (at date 05/11/2015) were applied to the trees (H. Rammensee et al. 1999).

Using MHC-I ligand data for evaluating proteasome prediction algorithms is common but nevertheless not an optimal solution, because MHC ligands are the result of a process

that involves more steps than only proteasomal digestion, like TAP-transport or N-terminal nibbling. Only testing the MHC ligands themselves is therefore not sufficient: We also need to test fragments with the same C-terminus that are longer than the ligand. Furthermore, since the trees were trained with in vitro data, we cannot rule out that the trees are more likely to predict smaller fragments, since fragments that might have continued their way in the MHC I pathway under in vivo conditions might have been cleaved further under in vitro conditions.

When evaluating the MHC ligand data, all smaller and larger fragments with the same C-terminus were therefore also tested.

Table 19 shows the results of the tests with MHC I ligand data obtained from the SYFPEITHI database. “Any fragment found” shows, if for a given MHC ligand a decision tree found at least one smaller/larger fragment or the ligand itself, while “exact or larger found” shows for each MHC ligand if the ligand itself or at least one larger fragment was found. The following columns show how many of the MHC I ligands, smaller and larger fragments were found in total.

Even though the “exact or larger found” rate is not very specific, since in most cases it includes a multitude of fragments, it reflects the only definite conclusion that can be drawn from MHC ligand data: if a decision tree is not able to detect an MHC I ligand or an N-terminally extended version of the ligand, this is a strong indication for a poor prediction performance of the tree. With values of 21.8% and 16.1% the first two trees (AACodesP1-P7, trained with the CMap dataset) fall into this category. The remaining trees show a good detection rate of 80% and above. The trees trained with the FileAn dataset even show a rate above 90%, which can be explained by the larger share of positive samples in this dataset.

Attribute Set	Dataset	Any fragment found			Exact or larger found			Smaller found			Exact found			Larger found		
		found	missed	%	found	missed	%	found	missed	%	found	missed	%	found	missed	%
AACodesP1-P7	CMap	941	3085	23.4	876	3150	**21.8**	308	33158	0.9	81	3945	2.0	10785	1280115	0.8
AACodesP1-P7	CMap*	700	3326	17.4	650	3376	**16.1**	310	33156	0.9	74	3952	1.8	9971	1280929	0.8
AACodesP1-P7	FileAn	3903	123	96.9	3875	151	96.2	4716	28750	14.1	1328	2698	33.0	74714	1216186	5.8
AACodesP1-P7	FileAn*	3880	146	96.4	3840	186	95.4	5465	28001	16.3	1406	2620	34.9	61899	1229001	4.8
AAIndex1 P1	CMap	3454	572	85.8	3368	658	83.7	1318	32148	3.9	346	3680	8.6	54074	1236826	4.2
AAIndex1 P1	CMap*	3511	515	87.2	3438	588	85.4	1494	31972	4.5	336	3690	8.3	56711	1234189	4.4
AAIndex1 P1#	CMap	3379	647	83.9	3297	729	81.9	1631	31835	4.9	363	3663	9.0	61632	1229268	4.8
AAIndex1 P1#	CMap*	3477	549	86.4	3393	633	84.3	1461	32005	4.4	314	3712	7.8	59293	1231607	4.6
AAIndex1 P1	FileAn	4020	6	99.9	3974	52	98.7	10612	22854	31.7	1432	2594	35.6	404048	886852	31.3
AAIndex1 P1	FileAn*	4016	10	99.8	3977	49	98.8	10437	23029	31.2	1392	2634	34.6	411485	879415	31.9
AAIndex1 P1#	FileAn	4006	20	99.5	3957	69	98.3	10461	23005	31.3	1414	2612	35.1	405274	885626	31.4
AAIndex1 P1#	FileAn*	4015	11	99.7	3973	53	98.7	10218	23248	30.5	1342	2684	33.3	395812	895088	30.7
AAIndex1 P1-P7	CMap	3328	698	82.7	3232	794	80.3	1515	31951	4.5	355	3671	8.8	61608	1229292	4.8
AAIndex1 P1-P7	CMap*	3324	702	82.6	3233	793	80.3	1486	31980	4.4	345	3681	8.6	57917	1232983	4.5
AAIndex1 P1-P7#	CMap	3267	759	81.1	3172	854	78.8	1814	31652	5.4	325	3701	8.1	77420	1213480	6.0
AAIndex1 P1-P7#	CMap*	3353	673	83.3	3265	761	81.1	1661	31805	5.0	326	3700	8.1	71874	1219026	5.6
AAIndex1 P1-P7	FileAn	4025	1	99.9	3997	29	99.3	9247	24219	27.6	1294	2732	32.1	387025	903875	30.0
AAIndex1 P1-P7	FileAn*	4022	4	99.9	3978	48	98.8	9268	24198	27.7	1191	2835	29.6	387642	903258	30.0
AAIndex1 P1-P7#	FileAn	3990	36	99.1	3942	84	97.9	9376	24090	28.0	1354	2672	33.6	412366	878534	31.9
AAIndex1 P1-P7#	FileAn*	3978	48	98.8	3927	99	97.5	9005	24461	26.9	1219	2807	30.3	380388	910512	29.5
AAIndex1 Fragment	CMap	3907	119	97.0	3847	179	95.6	1782	31684	5.3	594	3432	14.8	994847	296053	77.1
AAIndex1 Fragment	CMap*	3893	133	96.7	3828	198	95.1	1873	31593	5.6	579	3447	14.4	938969	351931	72.7
AAIndex1 Fragment#	CMap	3609	417	89.6	3596	430	89.3	70	33396	0.2	0	4026	0.0	1036448	254452	80.3
AAIndex1 Fragment#	CMap*	3496	530	86.8	3495	531	86.8	29	33437	0.1	4	4022	0.1	1043148	247752	80.8
AAIndex1 Fragment	FileAn	4018	8	99.8	3978	48	98.8	7451	26015	22.3	2021	2005	50.2	807361	483539	62.5
AAIndex1 Fragment	FileAn*	4017	9	99.8	3985	41	99.0	7237	26229	21.6	1988	2038	49.4	1190483	100417	92.2
AAIndex1 Fragment#	FileAn	4006	20	99.5	3998	28	99.3	4660	28806	13.9	2752	1274	68.4	1204390	86510	93.3
AAIndex1 Fragment#	FileAn*	4017	9	99.8	4012	14	99.7	7289	26177	21.8	3377	649	83.9	795462	495438	61.6

Table 19: Result of decision tree validation using SYFPEITHI MHC I ligand data. "Any fragment found": The MHC ligand itself, a cleave product of the ligand (smaller), or a larger fragment containing the ligand was detected as a proteasome product by the decision tree. "Exact or larger found", "Smaller found", "Exact found" and "Larger found" accordingly. Especially the "Exact or larger found" category is of interest, since the proteasome produces cleavage products that are usually longer than the final MHC I ligand due to N-terminal nibbling. The first two trees show a poor prediction performance for the MHC I ligand data: Only 21.8%/16.1% of the ligands or a longer precursor fragment are predicted (marked red). The remaining trees show a good prediction performance however.

Comparing the decision trees with other prediction methods

Comparing the decision trees' performance with the results of other prediction methods, especially the ones presented in the introduction chapter, is difficult because most other methods predict cleavage sites instead of fragments. Therefore, it would have been especially interesting to compare the decision trees' performance with the algorithm developed by Ginodi et al., which predicts fragments as well (Ginodi et al. 2008). However, neither the test data set used in the publication, nor the algorithm's implementation was available at the time of writing.

Therefore, the decision trees were evaluated against the dataset of CTL epitopes used by Saxova et al. for their comparison of NetChop, FragPredict and PAProC (Saxová et al. 2003). Since NetChop and PAProC predict cleavage sites, the authors define the following classifications in order to compare the algorithms:

- True positive (TP): if the prediction at the C-terminal, Pc, is above the algorithm's threshold.
- False negative (FN): if Pc is below the threshold.
- True negative (TN): if no cleavages are predicted within the epitope (excluding the C-terminal residue) or if the predicted cleavage sites within the epitope are less likely than at the C-terminal (i.e. less than Pc and the threshold).
- False positive (FP): if there is at least one predicted cleavage site within the epitope which is more likely than at the C-terminal (i.e. higher than Pc)

These classifications do not really fit for a fragment prediction method like the decision trees, however: even if a decision tree would predict a fragment that would imply cleavage within a CTL-epitope, this is not relevant as long as the tree also predicts the epitope, too, or an N-terminally extended fragment. The definitions for true negatives and false positives therefore do not apply.

Thus it was only examined how many epitopes in the validation dataset were detected by each decision tree. The results are shown in Table 20. It is also shown how the prediction rate improves when fragments extended at the N-terminus are included. The best performing tree predicts 84% of the CTL-epitopes exactly. It is not possible

however to put this rate into perspective by measuring the tree's specificity, given CTL epitope data, because the trees do not predict CTL epitopes but cleavage products of the proteasome. Since most cleavage products are not forwarded to the MHC I pathway, predicting a fragment that is not compatible with any CTL epitope does not necessarily imply a mistake by the algorithm.

This might also pose a problem in Saxova et al.'s original comparison, who found NetChop to be the best performing algorithm of the three examined. This algorithm, however, is also the only one, which was trained with CTL epitope data.

Tree		N + 0	N + 1	N + 2	N + 3	N + 4	N + 5	N + 6	N + 7	N + 8	N + 9	N + 10
AACodesP1-P7	CMap	5 (0,02)	8 (0,03)	10 (0,04)	13 (0,06)	13 (0,06)	16 (0,07)	18 (0,08)	18 (0,08)	20 (0,09)	20 (0,09)	23 (0,10)
AACodesP1-P7	CMap*	3 (0,01)	6 (0,03)	8 (0,03)	13 (0,06)	13 (0,06)	13 (0,06)	15 (0,06)	15 (0,06)	17 (0,07)	17 (0,07)	26 (0,11)
AACodesP1-P7	FileAn	77 (0,33)	120 (0,52)	128 (0,55)	129 (0,56)	132 (0,57)	145 (0,63)	156 (0,68)	158 (0,68)	158 (0,68)	158 (0,68)	160 (0,69)
AACodesP1-P7	FileAn*	79 (0,34)	126 (0,55)	132 (0,57)	135 (0,58)	137 (0,59)	150 (0,65)	169 (0,73)	171 (0,74)	171 (0,74)	172 (0,74)	174 (0,75)
AAIndex1 P1	CMap	7 (0,03)	14 (0,06)	19 (0,08)	23 (0,10)	24 (0,10)	26 (0,11)	26 (0,11)	32 (0,14)	38 (0,16)	44 (0,19)	89 (0,39)
AAIndex1 P1	CMap*	7 (0,03)	20 (0,09)	25 (0,11)	34 (0,15)	41 (0,18)	44 (0,19)	48 (0,21)	53 (0,23)	56 (0,24)	60 (0,26)	101 (0,44)
AAIndex1 P1#	CMap	9 (0,04)	15 (0,06)	19 (0,08)	26 (0,11)	30 (0,13)	37 (0,16)	41 (0,18)	43 (0,19)	45 (0,19)	48 (0,21)	86 (0,37)
AAIndex1 P1#	CMap*	15 (0,06)	25 (0,11)	32 (0,14)	35 (0,15)	45 (0,19)	52 (0,23)	54 (0,23)	55 (0,24)	59 (0,26)	64 (0,28)	105 (0,45)
AAIndex1 P1	FileAn	68 (0,29)	108 (0,47)	134 (0,58)	152 (0,66)	181 (0,78)	190 (0,82)	192 (0,83)	200 (0,87)	204 (0,88)	210 (0,91)	216 (0,94)
AAIndex1 P1	FileAn*	77 (0,33)	112 (0,48)	144 (0,62)	159 (0,69)	171 (0,74)	184 (0,80)	190 (0,82)	196 (0,85)	199 (0,86)	204 (0,88)	213 (0,92)
AAIndex1 P1#	FileAn	61 (0,26)	105 (0,45)	137 (0,59)	149 (0,65)	171 (0,74)	182 (0,79)	185 (0,80)	194 (0,84)	199 (0,86)	201 (0,87)	209 (0,90)
AAIndex1 P1#	FileAn*	77 (0,33)	108 (0,47)	143 (0,62)	156 (0,68)	171 (0,74)	189 (0,82)	195 (0,84)	200 (0,87)	206 (0,89)	209 (0,90)	212 (0,92)
AAIndex1 P1-P7	CMap	23 (0,10)	34 (0,15)	36 (0,16)	41 (0,18)	46 (0,20)	50 (0,22)	54 (0,23)	62 (0,27)	65 (0,28)	71 (0,31)	98 (0,42)
AAIndex1 P1-P7	CMap*	16 (0,07)	26 (0,11)	29 (0,13)	30 (0,13)	39 (0,17)	40 (0,17)	47 (0,20)	52 (0,23)	53 (0,23)	56 (0,24)	91 (0,39)
AAIndex1 P1-P7#	CMap	14 (0,06)	19 (0,08)	26 (0,11)	31 (0,13)	49 (0,21)	54 (0,23)	59 (0,26)	69 (0,30)	73 (0,32)	86 (0,37)	106 (0,46)
AAIndex1 P1-P7#	CMap*	22 (0,10)	38 (0,16)	43 (0,19)	52 (0,23)	59 (0,26)	66 (0,29)	73 (0,32)	77 (0,33)	83 (0,36)	86 (0,37)	100 (0,43)
AAIndex1 P1-P7	FileAn	76 (0,33)	112 (0,48)	136 (0,59)	161 (0,70)	175 (0,76)	192 (0,83)	205 (0,89)	209 (0,90)	213 (0,92)	219 (0,95)	227 (0,98)
AAIndex1 P1-P7	FileAn*	57 (0,25)	105 (0,45)	138 (0,60)	152 (0,66)	173 (0,75)	188 (0,81)	193 (0,84)	202 (0,87)	207 (0,90)	212 (0,92)	217 (0,94)
AAIndex1 P1-P7#	FileAn	62 (0,27)	97 (0,42)	119 (0,52)	140 (0,61)	155 (0,67)	174 (0,75)	185 (0,80)	191 (0,83)	195 (0,84)	203 (0,88)	207 (0,90)
AAIndex1 P1-P7#	FileAn*	57 (0,25)	96 (0,42)	121 (0,52)	140 (0,61)	151 (0,65)	163 (0,71)	173 (0,75)	181 (0,78)	189 (0,82)	191 (0,83)	196 (0,85)
AAIndex1 Fragment	CMap	27 (0,12)	44 (0,19)	59 (0,26)	76 (0,33)	89 (0,39)	93 (0,40)	103 (0,45)	111 (0,48)	117 (0,51)	119 (0,52)	123 (0,53)
AAIndex1 Fragment	CMap*	29 (0,13)	59 (0,26)	74 (0,32)	85 (0,37)	95 (0,41)	104 (0,45)	112 (0,48)	123 (0,53)	133 (0,58)	143 (0,62)	148 (0,64)
AAIndex1 Fragment#	CMap	0 (0,00)	0 (0,00)	1 (0,00)	1 (0,00)	1 (0,00)	1 (0,00)	1 (0,00)	2 (0,01)	2 (0,01)	4 (0,02)	4 (0,02)
AAIndex1 Fragment#	CMap*	0 (0,00)	0 (0,00)	0 (0,00)	1 (0,00)	2 (0,01)	4 (0,02)	5 (0,02)	5 (0,02)	5 (0,02)	5 (0,02)	6 (0,03)
AAIndex1 Fragment	FileAn	102 (0,44)	149 (0,65)	171 (0,74)	197 (0,85)	202 (0,87)	207 (0,90)	211 (0,91)	213 (0,92)	217 (0,94)	218 (0,94)	219 (0,95)
AAIndex1 Fragment	FileAn*	106 (0,46)	146 (0,63)	168 (0,73)	184 (0,80)	198 (0,86)	206 (0,89)	208 (0,90)	212 (0,92)	213 (0,92)	213 (0,92)	218 (0,94)
AAIndex1 Fragment#	FileAn	156 (0,68)	185 (0,80)	202 (0,87)	211 (0,91)	214 (0,93)	218 (0,94)	219 (0,95)	219 (0,95)	219 (0,95)	221 (0,96)	221 (0,96)
AAIndex1 Fragment#	FileAn*	193 (0,84)	202 (0,87)	211 (0,91)	218 (0,94)	221 (0,96)	222 (0,96)	223 (0,97)	224 (0,97)	224 (0,97)	225 (0,97)	225 (0,97)

Table 20: Results for tree validation with the data set of Saxová et al. (Saxová et al. 2003). The matrix shows how many epitopes with length N were detected by each tree (percentage of all epitopes given in brackets). Since the proteasome's cleavage products are trimmed at the N-terminus it was also examined how the prediction improves, if extended fragments (N + x) are included. The detection rates of succeeding columns are added up, i. e. the column N + 1 counts all detections of the epitope itself or the epitope extended by one amino acid.

Discussion

Summary

In order to gain a deeper understanding of the proteasome's cleavage behavior, a model for prediction of its cleavage products was developed in this work. The decision tree algorithm was selected as approach for fragment prediction because it allows insight into the decision process that leads to its final classification.

A new database of in vitro training data for the model was compiled from existing experimental data from the institute of biochemistry at the Charité Berlin. Two datasets were created for training: a more restrictive, manually validated one (CMap) and one containing a larger number of proteasomal cleavage products, detected by the software "FileAnalyzer" (FileAn). The validity of the FileAnalyzer data was verified using the manually validated data as reference.

Different sets of attributes were used for decision tree induction. The amino acid index database served as a source for the majority of these attributes.

Cross-validation of the decision trees

Cross validation was used in order to evaluate the trees' ability to classify the training data. With an average AUC between 0.57 and 0.83, true positive rates mostly above 0.7 and true negative rates above 0.8, most of the trees performed well in this task. In general, trees trained with the FileAn dataset showed better results. This might be explained by the CMap dataset being too restrictive and therefore providing a very small number of positive samples, which could not be separated as good from the negative ones using the attributes provided. The problem gets especially noticeable at the two trees trained with CMap using the Amino Acid Letter Codes attribute set, which only reach a true positive rate of 0.3.

Relevance of amino acid attributes and positions

Amino acid letter codes

Decision trees generated with the amino acid letter code attribute set performed well only when trained with the FileAn dataset. In these trees, the attributes “fragment length” and “P1” were selected first. Attributes selected at the third level, however, varied and seemed primarily to be selected because they fit the training data best.

At the position P1, aspartic acid (D) especially stands out as an indicator for a fragment being cleaved by the proteasome with a probability up to 0.8. This finding fits the results of Tenzer et al. (Tenzer et al. 2005), who reported a high score for aspartic acid in their scoring matrix for the constitutive proteasome, a similar importance of lysine (L) as described by Tenzer et al. as well could, however, not be confirmed by the tree.

AAIndex attributes

All decision trees generated with a set of attributes derived from the amino acid index database performed rather well fitting the training data. Trees trained with the FileAn dataset consistently showed higher average AUC values and true positive/negative rates. A big hope of this study did not come true, however: No attribute or cluster of attributes with unambiguously high relevance for the proteasomal cleavage process could be identified. There is no clear dominance of one or a few cluster colors in Table 16 or Table 17. There are multiple possible explanations for these unclear results: 1. The wrong set of attributes was used (this matter is discussed in more detail in the following section) 2. The factors relevant for the cleavage process are too complex to be revealed as a sequence of a few decisions 3. In the unclustered attribute sets, too many attributes were provided and so an overfitting of the training data occurred, hiding the actually relevant types. 4. The clustered attribute sets provided too few attributes or hid the relevant attributes within their corresponding clusters.

Validation with MHC I ligand data

As described before in the “Methods”-section, validation of the decision trees using MHC I ligand data is more a compromise for lack of any better suited data for validation.

All decision trees except for AACodesP1-P7 (CMap/CMap*) were able to identify more than 80% of MHC I ligands or at least one N-terminally extended fragment. This finding serves as a sanity check for the implementation and confirms that the in vitro training data is to a certain extend also applicable to in vivo data.

The MHC I pathway still remains complex and its substrate is modified in multiple steps and ways, including N-terminal nibbling (Kisselev et al. 1999) or, for example, CTL epitopes, which originated from the fusion of two segments located at either end of the antigen, resulting in an epitope that cannot be identified within the original substrate (Hanada, Yewdell, and Yang 2004).

Potential sources of error

In the following, problems and restrictions of this work are identified and discussed.

Mass spectrometry and data set

The training data used in this study was obtained from in vitro experiments that were analyzed with mass spectrometry. A peptide's properties affect its detection by mass spectrometry, which itself may lead to biased results. Especially small fragments may not be detected reliably by mass spectrometry. Within the last ten years, the quality and resolution of mass spectrometry instruments has improved significantly, making scientists realize that a considerable amount of peptides has been missed in past experiments.

In vitro data

The training data originates from in vitro experiments, which must be taken into consideration when trying to draw conclusions regarding the in vivo process. Under in vitro conditions there is an abundance of substrate, which is processed by the proteasome. Its cleavage products remain available as substrate for further digestions and are neither removed by cytosolic peptidases nor transported away by TAP.

In vivo data, on the other hand, does not allow an isolated examination of the proteasome. The CTL epitope data available is the result of the MHC I pathway, a far more complex process including multiple steps and systems. This makes validation of a

model learned from in vitro data virtually impossible, because subsequent modifications to the in vivo cleavage products after they have left the proteasome cannot be identified.

Validation with other in vitro data available, as performed with the enolase digestion data, showed another problem: Even though the training dataset used was adequately large, especially compared to the data of similar publications, the training data itself may still bias the model. The experiments, which provided the data, mainly focused on oligopeptides with a length of about 50 amino acids. For really universal deductions a far greater dataset would be required.

Attribute sets

While the evaluation of all amino acid properties described in the amino acid index database seemed promising, the lack of unambiguous results indicates that the attribute sets used were not able to reveal a certain logic behind the proteasomal cleavage process. There are two possible explanations:

1. The correct logic was identified by one of the trees while all other trees do not describe the correct logic. The correct solution is therefore hidden amidst the wrong ones and cannot be identified.
2. The attribute sets used did not contain the information relevant to the decision process. In this case, additional experiments might provide clearer results. For example, further candidates of attribute sets could include spatial information about the substrate and its position relative to the catalytic sites during the cleavage process.

Limitations of sequence-based methods

All models created in this work are sequence-based only, meaning that they only rely on the substrate's sequence and the properties of its amino acid sequence irrespective of the peptides' steric conformations. While various methods for in silico docking experiments exist, their benefit for this work would at least be questionable: the shortness of the oligopeptides examined allows these molecules to change their conformation rather freely.

Blending different proteasome types

The training data contained experiments with proteasomes of different cell lines (T2, T27 etc.) and different states (constitutive proteasome, immunoproteasome, PA28), which are not differentiated in the model in order to keep the data base large enough. However, a cell may use different proteasomes at the same time, which means the model can still represent a biological reality (Brooks et al. 2000).

Bibliography

Aghajanian, Carol, D. S. Dizon, P. Sabbatini, J. J. Raizer, J. Dupont, and D. R. Spriggs. 2005. “Phase I Trial of Bortezomib and Carboplatin in Recurrent Ovarian or Primary Peritoneal Cancer.” *Journal of Clinical Oncology* 23 (25): 5943–49. doi:10.1200/JCO.2005.16.006.

Argos, P, J K Rao, and P A Hargrave. 1982. “Structural Prediction of Membrane-Bound Proteins.” *European Journal of Biochemistry / FEBS* 128 (2-3): 565–75.

Barry, M A, W C Lai, and S A Johnston. 1995. “Protection against Mycoplasma Infection Using Expression-Library Immunization.” *Nature.* 377 (6550): 632–35. doi:10.1038/377632a0.

Barry, M, and G McFadden. 1998. “Apoptosis Regulators from DNA Viruses.” *Current Opinion in Immunology* 10 (4): 422–30. doi:10.1016/S0952-7915(98)80116-7.

Beekman, N J, P A van Veelen, T van Hall, A Neisig, A Sijts, M Camps, P M Kloetzel, J J Neefjes, C J Melief, and F Ossendorp. 2000. “Abrogation of CTL Epitope Processing by Single Amino Acid Substitution Flanking the C-Terminal Proteasome Cleavage Site.” *Journal of Immunology (Baltimore, Md. : 1950)* 164 (4): 1898–1905.

Berenson, James R, Hank H Yang, Karen Sadler, Supol G Jarutirasarn, Robert a Vescio, Russell Mapes, Matthew Purner, et al. 2006. “Phase I/II Trial Assessing Bortezomib and Melphalan Combination Therapy for the Treatment of Patients with Relapsed or Refractory Multiple Myeloma.” *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology* 24 (6): 937–44. doi:10.1200/JCO.2005.03.2383.

Bhasin, Manoj, and G. P S Raghava. 2005. “Pcleavage: An SVM Based Method for Prediction of Constitutive Proteasome and Immunoproteasome Cleavage Sites in Antigenic Sequences.” *Nucleic Acids Research* 33 (SUPPL. 2).

Braun, B C, M Glickman, R Kraft, B Dahlmann, P M Kloetzel, D Finley, and M Schmidt. 1999. “The Base of the Proteasome Regulatory Particle Exhibits Chaperone-like Activity.” *Nature Cell Biology* 1 (4): 221–26.

Brooks, P, G Fuertes, R Z Murray, S Bose, E Knecht, M C Rechsteiner, K B Hendil, K Tanaka, J Dyson, and J Rivett. 2000. “Subcellular Localization of Proteasomes and Their Regulatory Complexes in Mammalian Cells.” *The Biochemical Journal* 346 Pt 1: 155–61. doi:10.1042/0264-6021:3460155.

Bulik, S. 2011. “Theoretische Untersuchungen Zur MHC I Antigenpräsentation.”

Cascio, Paolo, Craig Hilton, Alexei F. Kisselev, Kenneth L. Rock, and Alfred L. Goldberg. 2001. “26S Proteasomes and Immunoproteasomes Produce Mainly N-Extended Versions of an Antigenic Peptide.” *EMBO Journal* 20 (10): 2357–66.

Cestnik, Bojan, Igor Kononenko, and Ivan Bratko. 1987. “ASSISTANT 86: A Knowledge-

Elicitation Tool for Sophisticated Users." In *Proc.\ of the 2nd European Working Session on Learning*, 31–45. Sigma Press.

Cheng, Jie, Usama M Fayyad, Keki B Irani, and Zhaogang Qian. 1988. "Improved Decision Trees: A Generalized Version of id3." In *Proc. Fifth Int. Conf. Machine Learning*, 100–107.

Chu-Ping, Ma, Clive A. Slaughter, and George N. DeMartino. 1992. "Identification, Purification, and Characterization of a Protein Activator (PA28) of the 20 S Proteasome (macropain)." *Journal of Biological Chemistry* 267 (15): 10515–23.

Coux, O, K Tanaka, and A L Goldberg. 1996. "Structure and Functions of the 20S and 26S Proteasomes." *Annual Review of Biochemistry* 65: 801–47.

Craiu, A, T Akopian, A Goldberg, and K L Rock. 1997. "Two Distinct Proteolytic Processes in the Generation of a Major Histocompatibility Complex Class I-Presented Peptide." *Proceedings of the National Academy of Sciences of the United States of America* 94 (20): 10850–55.

Del Val, M, H J Schlicht, T Ruppert, M J Reddehase, and U H Koszinowski. 1991. "Efficient Processing of an Antigenic Sequence for Presentation by MHC Class I Molecules Depends on Its Neighboring Residues in the Protein." *Cell* 66 (6): 1145–53.

Delic, J, P Masdehors, S Omura, J M Cosset, J Dumont, J L Binet, and H Magdelénat. 1998. "The Proteasome Inhibitor Lactacystin Induces Apoptosis and Sensitizes Chemo- and Radioresistant Human Chronic Lymphocytic Leukaemia Lymphocytes to TNF-Alpha-Initiated Apoptosis." *British Journal of Cancer* 77 (7): 1103–7. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2150120&tool=pmcentrez&rendertype=abstract.

Deveraux, Q., V. Ustrell, C. Pickart, and M. Rechsteiner. 1994. "A 26 S Protease Subunit That Binds Ubiquitin Conjugates." *Journal of Biological Chemistry* 269 (10): 7059–61.

Dolenc, I, E Seemüller, and W Baumeister. 1998. "Decelerated Degradation of Short Peptides by the 20S Proteasome." *FEBS Letters* 434 (3): 357–61. doi:S0014-5793(98)01010-2 [pii].

Dyer, K F. 1971. "The Quiet Revolution: A New Synthesis of Biological Knowledge." *Journal of Biological Education* 5 (1). Taylor & Francis: 15–24.

Furman, Margo H, Hidde L Ploegh, and others. 2002. "Lessons from Viral Manipulation of Protein Disposal Pathways." *The Journal of Clinical Investigation* 110 (110 (7)). Am Soc Clin Investig: 875–79.

Ginodi, Ido, Tal Vider-Shalit, Lea Tsaban, and Yoram Louzoun. 2008. "Precise Score for the Prediction of Peptides Cleaved by the Proteasome." *Bioinformatics* 24 (4): 477–83.

Glickman, M H, D M Rubin, H Fu, C N Larsen, O Coux, I Wefes, G Pfeifer, et al. 1999. "Functional Analysis of the Proteasome Regulatory Particle." *Molecular Biology*

Reports 26 (1-2): 21–28.

Glickman, Michael H, and Aaron Ciechanover. 2002. “The Ubiquitin-Proteasome Proteolytic Pathway: Destruction for the Sake of Construction.” *Physiological Reviews* 82 (2): 373–428.

Glickman, Michael H., David M. Rubin, Olivier Coux, Inge Wefes, Günter Pfeifer, Zdenka Cjeka, Wolfgang Baumeister, Victor A. Fried, and Daniel Finley. 1998. “A Subcomplex of the Proteasome Regulatory Particle Required for Ubiquitin-Conjugate Degradation and Related to the COP9-Signalosome and eIF3.” *Cell* 94 (5): 615–23.

Groll, M, L Ditzel, J Löwe, D Stock, M Bochtler, H D Bartunik, and R Huber. 1997. “Structure of 20S Proteasome from Yeast at 2.4 A Resolution.” *Nature* 386 (6624): 463–71.

Guterman, Adi, and Michael H. Glickman. 2004. “Complementary Roles for Rpn11 and Ubp6 in Deubiquitination and Proteolysis by the Proteasome.” *Journal of Biological Chemistry* 279 (3): 1729–38.

Hanada, Ken-Ichi, Jonathan W Yewdell, and James C Yang. 2004. “Immune Recognition of a Human Renal Cancer Antigen through Post-Translational Protein Splicing.” *Nature* 427 (6971): 252–56. doi:10.1038/nature02240.

Heink, Sylvia, Daniela Ludwig, Peter-M Kloetzel, and Elke Krüger. 2005. “IFN-Gamma-Induced Immune Adaptation of the Proteasome System Is an Accelerated and Transient Response.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (26): 9241–46.

Hendil, K B, S Khan, and K Tanaka. 1998. “Simultaneous Binding of PA28 and PA700 Activators to 20 S Proteasomes.” *The Biochemical Journal* 332 (Pt 3: 749–54.

Holzhütter, H G, C Frömmel, and P M Kloetzel. 1999. “A Theoretical Approach towards the Identification of Cleavage-Determining Amino Acid Motifs of the 20 S Proteasome.” *Journal of Molecular Biology* 286 (4): 1251–65.

Holzhütter, H G, and P M Kloetzel. 2000. “A Kinetic Model of Vertebrate 20S Proteasome Accounting for the Generation of Major Proteolytic Fragments from Oligomeric Peptide Substrates.” *Biophysical Journal* 79 (3): 1196–1205.

Huber, Eva M., Michael Basler, Ricarda Schwab, Wolfgang Heinemeyer, Christopher J. Kirk, Marcus Groettrup, and Michael Groll. 2012. “Immuno- and Constitutive Proteasome Crystal Structures Reveal Differences in Substrate and Inhibitor Specificity.” *Cell* 148 (4): 727–38. doi:10.1016/j.cell.2011.12.030.

Imajohohmi, S., T. Kawaguchi, S. Sugiyama, K. Tanaka, S. Omura, and H. Kikuchi. 1995. “Lactacystin, a Specific Inhibitor of the Proteasome, Induces Apoptosis in Human Monoblast U937 Cells.” *Biochemical and Biophysical Research Communications* 217 (3): 1070–77. doi:10.1006/bbrc.1995.2878.

Johnson, Mark, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L. Madden. 2008. “NCBI BLAST: A Better Web Interface.” *Nucleic Acids*

Research 36 (Web Server issue).

Kawashima, S, and M Kanehisa. 2000. “AAindex: Amino Acid Index Database.” *Nucleic Acids Research* 28 (1): 374. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102411&tool=pmcentrez&rendertype=abstract.

Kawashima, Shuichi, Hiroyuki Ogata, and Minoru Kanehisa. 1999. “AAindex: Amino Acid Index Database.” *Nucleic Acids Research*.

Keller, A., S. Purvine, A.I. Nesvizhskii, S. Stolyar, D.R. Goodlett, and E. Kolker. 2002. “Experimental Protein Mixture for Validating Tandem Mass Spectral Analysis.” *Journal of Integrative Biology* 6 (2): 207–12. doi:10.3168/jds.2011-4192.

Keşmir, Can, Alexander K Nussbaum, Hansjörg Schild, Vincent Detours, and Søren Brunak. 2002. “Prediction of Proteasome Cleavage Motifs by Neural Networks.” *Protein Engineering* 15 (4): 287–96.

Kikkert, Marjolein, Gerco Hassink, Martine Barel, Christian Hirsch, F J van der Wal, and E Wiertz. 2001. “Ubiquitination Is Essential for Human Cytomegalovirus US11-Mediated Dislocation of MHC Class I Molecules from the Endoplasmic Reticulum to the Cytosol.” *The Biochemical Journal* 358 (Pt 2): 369–77. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1222069&tool=pmcentrez&rendertype=abstract\nhttp://www.biochemj.org/bj/358/bj3580369.htm.

Kisselev, Alexei F., Tatos N. Akopian, and Alfred L. Goldberg. 1998. “Range of Sizes of Peptide Products Generated during Degradation of Different Proteins by Archaeal Proteasomes.” *Journal of Biological Chemistry* 273 (4): 1982–89. doi:10.1074/jbc.273.4.1982.

Kisselev, Alexei F., Tatos N. Akopian, Kee Min Woo, and Alfred L. Goldberg. 1999. “The Sizes of Peptides Generated from Protein by Mammalian 26 and 20 S Proteasomes. Implications for Understanding the Degradative Mechanism and Antigen Presentation.” *Journal of Biological Chemistry* 274 (6): 3363–71. doi:10.1074/jbc.274.6.3363.

Kloetzel, P M. 2001. “Antigen Processing by the Proteasome.” *Nature Reviews. Molecular Cell Biology* 2 (3): 179–87. doi:10.1038/35056572.

Kloetzel, Peter M. 2004. “Generation of Major Histocompatibility Complex Class I Antigens: Functional Interplay between Proteasomes and TPPII.” *Nature Immunology* 5 (7): 661–69. doi:10.1038/ni1090.

Kolker, Eugene, Roger Higdon, and Jason M. Hogan. 2006. “Protein Identification and Expression Analysis Using Mass Spectrometry.” *Trends in Microbiology*. doi:10.1016/j.tim.2006.03.005.

Kopp, F, B Dahlmann, and L Kuehn. 2001. “Reconstitution of Hybrid Proteasomes from Purified PA700-20 S Complexes and PA28alphabeta Activator: Ultrastructure and Peptidase Activities.” *Journal of Molecular Biology* 313 (3): 465–71.

Kuttler, C, A K Nussbaum, T P Dick, H G Rammensee, H Schild, and K P Hadeler.

2000. “An Algorithm for the Prediction of Proteasomal Cleavages.” *Journal of Molecular Biology* 298 (3): 417–29.

Levitskaya, Jelena, Anatoly Sharipo, Ainars Leonchiks, Aaron Ciechanover, and Maria G. Masucci. 1997. “Inhibition of Ubiquitin/proteasome-Dependent Protein Degradation by the Gly-Ala Repeat Domain of the Epstein–Barr Virus Nuclear Antigen 1.” *Proceedings of the National Academy of Sciences* 94 (23): 12616–21. http://www.pnas.org/content/94/23/12616\nhttp://www.ncbi.nlm.nih.gov/pubmed/9356498\nhttp://www.pnas.org/content/94/23/12616.full\nhttp://www.pnas.org/content/94/23/12616.full.pdf.

Löwe, J, D Stock, B Jap, P Zwickl, W Baumeister, and R Huber. 1995. “Crystal Structure of the 20S Proteasome from the Archaeon T. Acidophilum at 3.4 A Resolution.” *Science (New York, N.Y.)* 268 (5210): 533–39.

Messersmith, Wells a, Sharyn D Baker, Lance Lassiter, Rana a Sullivan, Kimberly Dinh, Virna I Almuete, John J Wright, Ross C Donehower, Michael a Carducci, and Deborah K Armstrong. 2006. “Phase I Trial of Bortezomib in Combination with Docetaxel in Patients with Advanced Solid Tumors.” *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research* 12 (4): 1270–75. doi:10.1158/1078-0432.CCR-05-1942.

Mishto, Michele, Fabio Luciani, Hermann-georg Holzhütter, Elena Bellavista, Aurelia Santoro, Kathrin Textoris-taube, Claudio Franceschi, Peter M Kloetzel, and Alexey Zaikin. 2008. “Modeling the in Vitro 20S Proteasome Activity : The Effect of PA28 – αβ and of the Sequence and Length of Polypeptides on the Degradation Kinetics,” 1607–17. doi:10.1016/j.jmb.2008.01.086.

Mo, X Y, P Cascio, K Lemerise, A L Goldberg, and K Rock. 1999. “Distinct Proteolytic Processes Generate the C and N Termini of MHC Class I-Binding Peptides.” *Journal of Immunology (Baltimore, Md. : 1950)* 163 (11): 5851–59.

Nakai, K, A Kidera, and M Kanehisa. 1988. “Cluster Analysis of Amino Acid Indices for Prediction of Protein Structure and Function.” *Protein Engineering* 2 (2): 93–100.

Nandi, Dipankar, Elaine Woodward, David B. Ginsburg, and John J. Monaco. 1997. “Intermediates in the Formation of Mouse 20S Proteasomes: Implications for the Assembly of Precursor β Subunits.” *EMBO Journal* 16 (17): 5363–75.

Nielsen, Morten, Claus Lundegaard, Ole Lund, and Can Keşmir. 2005. “The Role of the Proteasome in Generating Cytotoxic T-Cell Epitopes: Insights Obtained from Improved Predictions of Proteasomal Cleavage.” *Immunogenetics* 57 (1-2): 33–41.

Nussbaum, A K, T P Dick, W Keilholz, M Schirle, S Stevanović, K Dietz, W Heinemeyer, et al. 1998. “Cleavage Motifs of the Yeast 20S Proteasome Beta Subunits Deduced from Digests of Enolase 1.” *Proceedings of the National Academy of Sciences of the United States of America* 95 (21): 12504–9.

Nussbaum, a K, T P Dick, W Keilholz, M Schirle, S Stevanović, K Dietz, W Heinemeyer, et al. 1998. “Cleavage Motifs of the Yeast 20S Proteasome Beta Subunits Deduced from Digests of Enolase 1.” *Proceedings of the National Academy of Sciences of*

the United States of America 95 (21): 12504–9. doi:10.1073/pnas.95.21.12504.

Nussbaum, A. K., C. Kuttler, K. P. Hadeler, H. G. Rammensee, and H. Schild. 2001. “PAProC: A Prediction Algorithm for Proteasomal Cleavages Available on the WWW.” *Immunogenetics* 53 (2): 87–94.

Orlowski, M, and S Wilk. 2000. “Catalytic Activities of the 20 S Proteasome, a Multicatalytic Proteinase Complex.” *Archives of Biochemistry and Biophysics* 383 (1): 1–16.

Orlowski, Robert Z, and Albert S Baldwin. 2002. “NF-kappaB as a Therapeutic Target in Cancer.” *Trends in Molecular Medicine* 8 (8): 385–89. doi:10.1016/S1471-4914(02)02375-4.

Ossendorp, Ferry, Maren Eggers, Anne Neisig, Thomas Ruppert, Marcus Groettrup, Alice Sijts, Erica Mengedé, et al. 1996. “A Single Residue Exchange within a Viral CTL Epitope Alters Proteasome-Mediated Degradation Resulting in Lack of Antigen Presentation.” *Immunity* 5 (2): 115–24.

Paterson, A, and T B Niblett. 1982. “ACLS Manual.” *Edinburgh: Intelligent Terminals Ltd*.

Patnaik, A., V. Chau, and J. W. Wills. 2000. “Ubiquitin Is Part of the Retrovirus Budding Machinery.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (24): 13069–74. doi:10.1073/pnas.97.24.13069.

Peters, Björn, Katharina Janek, Ulrike Kuckelkorn, and Hermann Georg Holzhütter. 2002. “Assessment of Proteasomal Cleavage Probabilities from Kinetic Analysis of Time-Dependent Product Formation.” *Journal of Molecular Biology* 318 (3): 847–62.

Peters, J M, Z Cejka, J R Harris, J A Kleinschmidt, and W Baumeister. 1993. “Structural Features of the 26 S Proteasome Complex.” *Journal of Molecular Biology* 234 (4): 932–37.

Purcell, Anthony W, James McCluskey, and Jamie Rossjohn. 2007. “More than One Reason to Rethink the Use of Peptides in Vaccine Design.” *Nature Reviews. Drug Discovery* 6 (5): 404–14. doi:10.1038/nrd2224.

Quinlan, J R. 1993. *C4.5: Programs for Machine Learning*. Edited by Morgan Kaufmann. *Morgan Kaufmann San Mateo California*. Vol. 1. Morgan Kaufmann Series in {M}achine {L}earning. Morgan Kaufmann. http://portal.acm.org/citation.cfm?id=152181.

Quinlan, J. R. 1986. “Induction of Decision Trees.” *Machine Learning*. doi:10.1007/BF00116251.

Rajkumar, S Vincent, Paul G Richardson, Teru Hideshima, and Kenneth C Anderson. 2005. “Proteasome Inhibition as a Novel Therapeutic Target in Human Cancer.” *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology* 23 (3): 630–39. doi:10.1200/JCO.2005.11.030.

Rammensee, H, J Bachmann, N P Emmerich, O A Bachor, and S Stevanović. 1999. “SYFPEITHI: Database for MHC Ligands and Peptide Motifs.” *Immunogenetics* 50

(3-4): 213–19.

Rammensee, Hans-Georg, Thomas Friede, and Stefan Stevanović. 1995. “MHC Ligands and Peptide Motifs: First Listing.” *Immunogenetics* 41 (4). Springer: 178–228.

Richardson, Paul G, Pieter Sonneveld, Michael W Schuster, David Irwin, Edward A Stadtmauer, Thierry Facon, Jean-Luc Harousseau, et al. 2005. “Bortezomib or High-Dose Dexamethasone for Relapsed Multiple Myeloma.” *The New England Journal of Medicine* 352 (24): 2487–98. doi:10.1056/NEJMoa043445.

Richardson, Paul G., Pieter Sonneveld, Michael Schuster, David Irwin, Edward Stadtmauer, Thierry Facon, Jean Luc Harousseau, et al. 2007. “Extended Follow-up of a Phase 3 Trial in Relapsed Multiple Myeloma: Final Time-to-Event Results of the APEX Trial.” *Blood* 110 (10): 3557–60. doi:10.1182/blood-2006-08-036947.

Saxová, Patricia, Søren Buus, Søren Brunak, and Can Keşmir. 2003. “Predicting Proteasomal Cleavage Sites: A Comparison of Available Methods.” *International Immunology* 15 (7): 781–87.

Schubert, U, L C Antón, J Gibbs, C C Norbury, J W Yewdell, and J R Bennink. 2000. “Rapid Degradation of a Large Fraction of Newly Synthesized Proteins by Proteasomes.” *Nature* 404 (6779): 770–74.

Schubert, U, D E Ott, E N Chertova, R Welker, U Tessmer, M F Princiotta, J R Bennink, H G Krausslich, and J W Yewdell. 2000. “Proteasome Inhibition Interferes with Gag Polyprotein Processing, Release, and Maturation of HIV-1 and HIV-2.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (24): 13057–62. doi:10.1073/pnas.97.24.13057.

Schwartz, A L, and A Ciechanover. 1999. “The Ubiquitin-Proteasome Pathway and Pathogenesis of Human Diseases.” *Annual Review of Medicine* 50: 57–74. doi:10.1146/annurev.med.50.1.57.

Schwarz, K, M van Den Broek, S Kostka, R Kraft, A Soza, G Schmidtke, P M Kloetzel, and M Groettrup. 2000. “Overexpression of the Proteasome Subunits LMP2, LMP7, and MECL-1, but Not PA28 Alpha/beta, Enhances the Presentation of an Immunodominant Lymphocytic Choriomeningitis Virus T Cell Epitope.” *Journal of Immunology (Baltimore, Md. : 1950)* 165 (2): 768–78.

Seifert, Ulrike, Heike Liermann, Vito Racanelli, Anne Halenius, Manfred Wiese, Heiner Wedemeyer, Thomas Ruppert, et al. 2004. “Hepatitis C Virus Mutation Affects Proteasomal Epitope Processing.” *Journal of Clinical Investigation* 114 (2): 250–59.

Serwold, T, and N Shastri. 1999. “Specific Proteolytic Cleavages Limit the Diversity of the Pool of Peptides Available to MHC Class I Molecules in Living Cells.” *Journal of Immunology (Baltimore, Md. : 1950)* 162 (8): 4712–19.

Shamu, C E, D Flierman, H L Ploegh, T a Rapoport, and V Chau. 2001. “Polyubiquitination Is Required for US11-Dependent Movement of MHC Class I Heavy Chain from Endoplasmic Reticulum into Cytosol.” *Molecular Biology of the*

Cell 12 (8): 2546–55.

Sharon, Michal, Susanne Witt, Karin Felderer, Beate Rockel, Wolfgang Baumeister, and Carol V. Robinson. 2006. “20S Proteasomes Have the Potential to Keep Substrates in Store for Continual Degradation.” *Journal of Biological Chemistry* 281 (14): 9569–75.

Shimbara, Naoki, Kiyoko Ogawa, Yuko Hidaka, Hiroto Nakajima, Naoko Yamasaki, Shin Ichiro Niwa, Nobuyuki Tanahashi, and Keiji Tanaka. 1998. “Contribution of Proline Residue for Efficient Production of MHC Class I Ligands by Proteasomes.” *Journal of Biological Chemistry* 273 (36): 23062–71.

Shinohara, K, M Tomioka, H Nakano, S Toné, H Ito, and S Kawashima. 1996. “Apoptosis Induction Resulting from Proteasome Inhibition.” *The Biochemical Journal* 317 (Pt 2: 385–88. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1217499&tool=pmcentrez&rendertype=abstract.

Siegel, David S, Thomas Martin, Michael Wang, Ravi Vij, Andrzej J Jakubowiak, Sagar Lonial, Vishal Kukreti, et al. 2012. “A Phase 2 Study of Single-Agent Carfilzomib (PX-171-003-A1) in Patients with Relapsed and Refractory Multiple Myeloma.” *Blood* 120 (14): 2817–25. doi:10.1182/blood-2012-05-425934.

Sijts, A J, T Ruppert, B Rehermann, M Schmidt, U Koszinowski, and P M Kloetzel. 2000. “Efficient Generation of a Hepatitis B Virus Cytotoxic T Lymphocyte Epitope Requires the Structural Features of Immunoproteasomes.” *The Journal of Experimental Medicine* 191 (3): 503–14.

Sijts, Alice J A M, Sybille Standera, René E M Toes, Thomas Ruppert, Nico J C M Beekman, Peter A van Veelen, Ferry A Ossendorp, Cornelis J M Melief, and Peter M Kloetzel. 2000. “MHC Class I Antigen Processing of an Adenovirus CTL Epitope Is Linked to the Levels of Immunoproteasomes in Infected Cells.” *The Journal of Immunology* 164 (9). Am Assoc Immnol: 4500–4506.

Singh-Jasuja, Harpreet, N. P N Emmerich, and Hans Georg Rammensee. 2004. “The T??bingen Approach: Identification, Selection, and Validation of Tumor-Associated HLA Peptides for Cancer Therapy.” *Cancer Immunology, Immunotherapy* 53 (3): 187–95. doi:10.1007/s00262-003-0480-x.

Soza, Andrea, Christine Knuehl, Marcus Groettrup, Peter Henklein, Keiji Tanaka, and Peter M. Kloetzel. 1997. “Expression and Subcellular Localization of Mouse 20S Proteasome Activator Complex PA28.” *FEBS Letters* 413 (1): 27–34.

Stohwasser, Ralf, Ulrike Salzmann, Jan Giesebrecht, Peter Michael Kloetzel, and Hermann Georg Holzhütter. 2000. “Kinetic Evidences for Facilitation of Peptide Channelling by the Proteasome Activator PA28.” *European Journal of Biochemistry* 267 (20): 6221–30.

Strack, B., A. Calistri, M. A. Accola, G. Palu, and H. G. Gottlinger. 2000. “A Role for Ubiquitin Ligase Recruitment in Retrovirus Release.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (24): 13063–68.

doi:10.1073/pnas.97.24.13063.

Strehl, Britta, and Sylvia Heink. 2005. “Interferon- G , the Functional Plasticity of the Ubiquitin – Proteasome System , and MHC Class I Antigen Processing” 207: 19–30.

Strickland, Elizabeth, Kevin Hakala, Philip J. Thomas, and George N. DeMartino. 2000. “Recognition of Misfolding Proteins by PA700, the Regulatory Subcomplex of the 26 S Proteasome.” *Journal of Biological Chemistry* 275 (8): 5565–72.

Sun, Yuansheng, Alice J A M Sijts, Mingxia Song, Katharina Janek, Alexander K. Nussbaum, Sylvie Kral, Markus Schirle, et al. 2002. “Expression of the Proteasome Activator PA28 Rescues the Presentation of a Cytotoxic T Lymphocyte Epitope on Melanoma Cells.” *Cancer Research* 62 (10): 2875–82.

Tenzer, S., B. Peters, S. Bulik, O. Schoor, C. Lemmel, M. M. Schatz, P. M. Kloetzel, H. G. Rammensee, H. Schild, and H. G. Holzhütter. 2005. “Modeling the MHC Class I Pathway by Combining Predictions of Proteasomal Cleavage, TAP Transport and MHC Class I Binding.” *Cellular and Molecular Life Sciences* 62 (9): 1025–37.

Theobald, M, T Ruppert, U Kuckelkorn, J Hernandez, A Häussler, E A Ferreira, U Liewer, et al. 1998. “The Sequence Alteration Associated with a Mutational Hotspot in p53 Protects Cells from Lysis by Cytotoxic T Lymphocytes Specific for a Flanking Peptide Epitope.” *The Journal of Experimental Medicine* 188 (6): 1017–28.

Thrower, J S, L Hoffman, M Rechsteiner, and C M Pickart. 2000. “Recognition of the Polyubiquitin Proteolytic Signal.” *The EMBO Journal* 19 (1): 94–102.

Tomii, K, and M Kanehisa. 1996. “Analysis of Amino Acid Indices and Mutation Matrices for Sequence Comparison and Structure Prediction of Proteins.” *Protein Engineering* 9 (1): 27–36.

van Hall T, A Sijts, M Camps, R Offringa, C Melief, P M Kloetzel, and F Ossendorp. 2000. “Differential Influence on Cytotoxic T Lymphocyte Epitope Presentation by Controlled Expression of Either Proteasome Immunosubunits or PA28.” *The Journal of Experimental Medicine* 192 (4): 483–94.

Van Kaer, L, P G Ashton-Rickardt, M Eichelberger, M Gaczynska, K Nagashima, K L Rock, A L Goldberg, P C Doherty, and S Tonegawa. 1994. “Altered Peptidase and Viral-Specific T Cell Response in LMP2 Mutant Mice.” *Immunity* 1 (7): 533–41.

Verma, Rati, L Aravind, Robert Oania, W Hayes McDonald, John R Yates, Eugene V Koonin, and Raymond J Deshaies. 2002. “Role of Rpn11 Metalloprotease in Deubiquitination and Degradation by the 26S Proteasome.” *Science (New York, N.Y.)* 298 (5593): 611–15.

Wenzel, Thorsten, Christoph Eckerskorn, Friedrich Lottspeich, and Wolfgang Baumeister. 1994. “Existence of a Molecular Ruler in Proteasomes Suggested by Analysis of Degradation Products.” *FEBS Letters* 349 (2): 205–9. doi:10.1016/0014-5793(94)00665-2.

Yao, Tingting, and Robert E Cohen. 2002. “A Cryptic Protease Couples Deubiquitination and Degradation by the Proteasome.” *Nature* 419 (6905): 403–7.

Young, Patrick, Quinn Deveraux, Richard E. Beal, Cecile M. Pickart, and Martin Rechsteiner. 1998. “Characterization of Two Polyubiquitin Binding Sites in the 26 S Protease Subunit 5a.” *Journal of Biological Chemistry* 273 (10): 5461–67.

Appendix

Dataset used for decision tree learning

The training data was collected between 2000 and 2011 by working groups under supervision of Prof. Kloetzel and Prof. Holzhütter of the institute of biochemistry of the Charité Berlin. The following table shows the peptide sequence of each experiment (with the C-terminus on the left) and the number of fragments found by manual evaluation (as described in "Methods") at the first line. The following lines show the number of fragments found by automated evaluation using the FileAnalyzer tool and the area under the curve (AUC) of the receiver-operating-characteristic referring to the manually generated cleavage map (as described in "Methods"). Experiments belonging to the distinct dataset are displayed in black, all other experiments in gray.

Identifier	**Peptide sequence** Mass Spectrometry Raw file	**Cell Line**	**Fragments found**	**AUC**
E-L	**ALLALLAALCPASRALEEKKGNYVVTDHGS**		**40**	
	091124_E-L_T2.txt	T2	53	0.59
	091124_E-L_T27.txt	T27	46	0.63
	091208_E-L_LCL.txt	Unspecified	137	0.83
	091208_E-L_LCL_40ul.txt	Unspecified	64	0.77
	E_L.txt	Unspecified	151	0.83
ETV6-AML1-L	**MVSVSPPEEHAMPIGRIAECILGMNPSRDV**		**29**	
	091.txt	Unspecified	86	0.88
	091112_ETV6_T27.txt	T27	3	0.51
	091208_ETV-6_LCL.txt	Unspecified	88	0.82
HepB	**AYRPPNAPILSTLPETTVVRRRGRSPRRRTPS**		**2**	
	011112_H.txt	Unspecified	460	0.59
	011112_HA.txt	Unspecified	314	0.61
	011112_HB.txt	Unspecified	212	0.59
	011112_HC.txt	Unspecified	230	0.32
	011112_HD.txt	Unspecified	383	0.34
	011112_HE.txt	Unspecified	238	0.48
Kloe110	**TGSTAVPYGSFKHVDTRLQ**		**20**	
	030326_Kloe110.txt	Unspecified	181	0.88
	030326_Kloe110_Hela+Inf.txt	Unspecified	137	0.86
	030326_Kloe110_Hela.txt	Unspecified	134	0.81
	030326_Kloe110_TriMel.txt	Unspecified	148	0.84
Kloe111	**ELSWEDYLETGSTAVPYGSFKHVDTRLQNGFAPGMKL**		**44**	
	030326_Kloe111.txt	Unspecified	502	0.89
	030326_Kloe111_Hela+Inf.txt	Unspecified	360	0.78
	030326_Kloe111_Hela.txt	Unspecified	319	0.71
	030326_Kloe111_Trimel.txt	Unspecified	478	0.83
Kloe184	**ALEGFDKADGTLDSQVMSLHNLVHSFLNG**		**52**	
	0110xx_Kloe184_10.txt	Unspecified	534	0.66
	0110xx_Kloe184_11.txt	Unspecified	358	0.60
	0110xx_Kloe184_12.txt	Unspecified	544	0.59
	0110xx_Kloe184_7.txt	Unspecified	394	0.63
	0110xx_Kloe184_8.txt	Unspecified	589	0.65

	0110xx_Kloe184_9.txt	Unspecified	419	0.72
	0110xx_Kloe184.txt	Unspecified	472	0.74
	100216_Kloe184_LcL.txt	Unspecified	168	0.87
Kloe208	**ETVCDSLDDYNHLVTLCNGTYEGLLRR**		**27**	
	020516_Kloe208_10.txt	Unspecified	314	0.60
	020516_Kloe208_11.txt	Unspecified	387	0.60
	020516_Kloe208_12.txt	Unspecified	336	0.57
	020516_Kloe208_1.txt	Unspecified	477	0.60
	020516_Kloe208_7.txt	Unspecified	454	0.72
	020516_Kloe208_8.txt	Unspecified	396	0.60
	020516_Kloe208_9.txt	Unspecified	399	0.71
	020516_Kloe208.txt	Unspecified	543	0.69
	Kloe208_Deca.txt	Unspecified	472	0.73
	Kloe208_LCQ.txt	Unspecified	411	0.72
Kloe256	**TRPILSPLTKGILGFVFTLTVPSERGLQR**		**19**	
	020702_KLoe256_C_single.txt	Unspecified	149	0.72
	040923_Kloe256_NA.txt	Unspecified	47	0.55
	040923_Kloe256_NB.txt	Unspecified	60	0.62
	040923_Kloe256_NC.txt	Unspecified	262	0.62
	040923_Kloe256_ND.txt	Unspecified	310	0.57
	040923_Kloe256_NE.txt	Unspecified	6	0.52
	040923_Kloe256_NF.txt	Unspecified	6	0.49
	040923_Kloe256_N_1.txt	Unspecified	297	0.62
	040923_Kloe256_N_2.txt	Unspecified	33	0.57
Kloe258	**PSQGKGRGLSLSRFSWGAEGQRPGFGYG**		**35**	
	030616_Kloe258_GB.txt	Unspecified	195	0.70
	030820_Kloe258_GA.txt	Unspecified	203	0.66
	030820_Kloe258_GB.txt	Unspecified	273	0.70
	030820_Kloe258_GC.txt	Unspecified	291	0.67
	100316_Kloe258_LcL.txt	Unspecified	251	0.78
	1003xx_Kloe258_LCL_24h_60ul.txt	Unspecified	153	0.70
	100527_Kloe258_HBX.txt	Unspecified	221	0.75
Kloe260	**TESPSFSAGDNPPVLFSSDFRISGAPEKYESERR**		**23**	
	030820_Kloe258_BA.txt	Unspecified	461	0.51
	030820_Kloe258_BB.txt	Unspecified	442	0.49
	040831_Kloe260_BC.txt	Unspecified	513	0.51
	040831_Kloe260_BD.txt	Unspecified	444	0.49
Kloe272	**SRALVVTHTYLEPGPVTAQVVLQAAIPLTS**		**41**	
	030326_Kloe272_Hela+Inf.txt	Unspecified	301	0.75
	030326_Kloe272_Hela.txt	Unspecified	258	0.60
	030326_Kloe272_TriMel.txt	Unspecified	278	0.64
	030727_Kloe272_T2.txt	T2	484	0.70
	030727_Kloe272_T2_27.txt	T2	446	0.84
	100603_Kloe272_IE.txt	Unspecified	260	0.71
	100603_Kloe272_IF.txt	Unspecified	210	0.70
Kloe308	**GSWSQKRSFVYVWKTWGQYWQVLGGPVSGLSI**		**21**	
	030727_Kloe308_T2.txt	T2	188	0.79
	030727_Kloe308_T2_27.txt	T2	203	0.83
	1003xx_Kloe308_LCL_24h_60ul.txt	Unspecified	107	0.66
	100603_Kloe308_IC.txt	Unspecified	386	0.67
	100603_Kloe308_ID.txt	Unspecified	369	0.68
Kloe310	**AHSSSAFTITDQVPFSVSVSQLRALDGGNK**		**29**	
	030727_Kloe310_T2.txt	T2	192	0.54
	030727_Kloe310_T2_27.txt	T2	233	0.73
	041118_Kloe310_T2+PA28.txt	T2	211	0.53
	041118_Kloe310_T2_27+PA28.txt	T2	218	0.74
	041118_Kloe310_T2_27.txt	T2	225	0.70
	041118_Kloe310_T2.txt	T2	196	0.59
	090914_Kloe310_LCL.txt	Unspecified	375	0.66
	090914_Kloe310_T27.txt	T27	81	0.79
	090914_Kloe310_T2.txt	T2	73	0.67
	100527_Kloe310_HF.txt	Unspecified	413	0.62
Kloe334	**YYKDAASNSANRQDFTQDPGKFTE**		**20**	
	040728_Kloe334_20S+PA28.txt	20S	305	0.69

	040728_Kloe334_20S.txt	20S	302	0.65
	040728_Kloe334_i20S+PA28.txt	i20S	341	0.66
	040728_Kloe334_i20S.txt	i20S	268	0.58
Kloe336	**VYNAGMGVGVGNLTIFPHQWINLRTNNSATIVMPYTNSVPMDNMFR**		**21**	
	040505_Kloe336_20S.txt	20S	100	0.52
	040505_Kloe336_i20S.txt	i20S	118	0.51
Kloe337	**HQWINLRTNNSATIVMPYTNSVPMDNMFRHNNVTLMVIPFVPLDY**		**18**	
	031203_Kloe337_i20S.txt	i20S	116	0.72
	031218_Kloe337.txt	Unspecified	375	0.75
	040505_Kloe337_20S.txt	20S	278	0.84
	040505_Kloe337_i20S.txt	i20S	287	0.76
Kloe371	**PLELSEKNFQLNQDKMNFSTLRNIQGLFAPLKLQMEFKAVQQVQRLPFLSSSNLSLDVLRGN**		**50**	
	040221_Kloe371_20S.txt	20S	847	0.70
	040221_Kloe371_i20S.txt	i20S	609	0.80
Kloe392	**QRVYNAGMGVGVGNLTIFPHQWINL**		**20**	
	041222_Kloe392_T2+PA28.txt	T2	47	0.76
	041222_Kloe392_T2.txt	T2	45	0.53
	041222_Kloe392_T2_27+PA28.txt	T2	36	0.66
	041222_Kloe392_T2_27.txt	T2	60	0.63
Kloe393	**VMPYTNSVPMDNMFRHNNVTLMVIPFVPLDY**		**52**	
	080410_Kloe393_T2.txt	T2	148	0.68
	080410_Kloe393_T27.txt	T27	127	0.67
	080410_Kloe393_T27_PA28.txt	T27	220	0.58
	080410_Kloe393_T2_PA28.txt	T2	228	0.57
Kloe394	**ATIEQSAPSQSDQEQLFSNVQYFAHYCRKY**		**31**	
	080410_Kloe394_T2.txt	T2	92	0.71
	080410_Kloe394_T27.txt	T27	88	0.71
	080410_Kloe394_T27_PA28.txt	T27	198	0.61
	080410_Kloe394_T2_PA28.txt	T2	209	0.59
Kloe396	**AKGKSRLIEASSLNDSVAMRQTFGNL**		**22**	
	041222_Kloe396_T2+PA28.txt	T2	36	0.70
	041222_Kloe396_T2.txt	T2	45	0.87
	041222_Kloe396_T2_27+PA28.txt	T2	39	0.77
	041222_Kloe396_T2_27.txt	T2	39	0.70
Kloe400	**AMMKRNSSRVKTEYGEFTMLGIYDRWAV**		**22**	
	051130_Kloe400_T2.txt	T2	69	0.86
	051130_Kloe400_T2_27.txt	T2	87	0.91
	051130_Kloe400_T2_27_single.txt	T2	351	0.77
	051130_Kloe400_T2_single.txt	T2	279	0.87
	070618_Kloe400_T2.txt	T2	72	0.71
	070618_Kloe400_T27.txt	T27	60	0.74
Kloe409	**RPILSPLTKGILGFVFTLTVPSERGLQR**		**31**	
	050523_Kloe409_T27.txt	T27	287	0.69
	050523_Kloe409_T2.txt	T2	248	0.78
	051115_Kloe409_T2.txt	T2	286	0.78
	051115_Kloe409_T27.txt	T27	338	0.85
	1003xx_Kloe409.txt	Unspecified	311	0.80
Kloe426	**EVSGLEQLESIINFEKLTEWTS**		**14**	
	070510_Kloe426_LCQ_MEC_TET.txt	Unspecified	54	0.75
	070510_Kloe426_LCQ_T27.txt	T27	29	0.66
	070510_Kloe426_LCQ_T2.txt	T2	35	0.73
	070510_Kloe426_MEC.txt	Unspecified	4	0.52
	070510_Kloe426_MEC_TET.txt	Unspecified	5	0.52
	070510_Kloe426_T27.txt	T27	2	0.53
	070510_Kloe426_T2.txt	T2	2	0.53
Kloe427	**KDSTRTQINKVVRFDKLPGFGD**		**16**	
	070302_Kloe427_T2.txt	T2	108	0.79
	070302_Kloe427_T27.txt	T27	109	0.83
Kloe451	**CMKVFAQYILGADPLRVCSPSVDDLRA**		**26**	
	051220_Kloe451_T2.txt	T2	165	0.76
	051220_Kloe451_T2_27.txt	T2	135	0.82

Kloe453	**SEFCRVLCCYVLEETSVMLAKRPLITKPE**		**31**	
	051220_Kloe453_T2.txt	T2	65	0.76
	051220_Kloe453_T2_27.txt	T2	96	0.75
Kloe478	**KRWIILGLNKIVRMYSPVSILD**		**36**	
	060707_Kloe478_T2.txt	T2	111	0.84
	060707_Kloe478_T27.txt	T27	104	0.86
Kloe480	**KRWIILGLNKIVRMYSPVSILDIRQGP**		**32**	
	060619_Kloe480_T27.txt	T27	5	0.49
	060627_Kloe480_T2.txt	T2	91	0.79
	060627_Kloe480_T27.txt	T27	97	0.81
Kloe497	**GKNATGMEVGWYRSPFSRVVHLYRNGKDQDAEQA**		**44**	
	070126_Kloe497_MH.txt	Unspecified	141	0.55
	070126_Kloe497_MM.txt	Unspecified	114	0.58
	070126_MH_MOG_1.txt	Unspecified	430	0.77
	070126_MM_MOG_2.txt	Unspecified	448	0.73
Kloe582	**LSRKVAELVHFLLLKYRAR**		**27**	
	090914_Kloe582_T27.txt	T27	79	0.79
	090914_Kloe582_T2.txt	T2	65	0.76
Kloe585	**VSRQLRTKAWNRQLYPEWTEAQR**		**35**	
	0805xx_Kloe585_T27_A.txt	T27	357	0.82
	0805xx_Kloe585_T27_A_single.txt	T27	533	0.78
	0805xx_Kloe585_T2_A.txt	T2	314	0.77
	0806xx_Kloe585_EB.txt	Unspecified	342	0.72
	0806xx_Kloe585_HA.txt	Unspecified	340	0.68
Kloe614	**RRSGAAGAAVKGVGTMVMELIRMIKRGVNDRNF**		**40**	
	0811xx_Kloe614_Milz.txt	Unspecified	359	0.52
	0811xx_Kloe614_Muskel.txt	Unspecified	372	0.57
Kloe649	**AVNPGLLEVTSYKLKHI**		**9**	
	090728_Kloe649_LCQ_T2.txt	T2	167	0.92
	090728_Kloe649_LCQ_T27.txt	T27	167	0.96
	090728_Kloe649_T2.txt	T2	58	0.99
	090728_Kloe649_T27.txt	T27	65	0.99
Kloe652	**QLYPEWRTKAWNR**		**8**	
	090903_Kloe652_AB.txt	Unspecified	13	0.85
Kloe686	**NTYASKRGCSPRVKPQHISTHFLPRFK**		**26**	
	100216_Kloe686_LcL.txt	Unspecified	167	0.92
MDC-20	**FFLTPHRHRVSAINNYAQKLCTFSFL**		**47**	
	110207_MDC20_LA.txt	Unspecified	188	0.77
	110207_MDC20_LC.txt	Unspecified	226	0.78
	110210_MDC20_MA.txt	Unspecified	216	0.80
MDC-22	**LPKMDSVVYDFLKCMVYNIP**		**23**	
	110411_MDC22_TB.txt	Unspecified	60	0.78
	110411_MDC22_TD.txt	Unspecified	62	0.80
	110411_MDC22_TF.txt	Unspecified	47	0.71
	110530_MDC22_Lcl.txt	Unspecified	56	0.84
	110530_MDC22_T2.txt	T2	63	0.84
	110530_MDC22_T27.txt	T27	38	0.77
NOR1	**RWLLLGLNPLVGGGRLYSPTSILG**		**41**	
	060808_NOR1_T2.txt	T2	77	0.80
	060808_NOR1_T27.txt	T27	87	0.82
NOR10	**RLIYATRQLQRFAVNPGLLIT**		**27**	
	060808_NOR10_T27.txt	T27	101	0.78
	060808_NOR10_T2.txt	T2	112	0.82
	070424_T27_NOR10.txt	T27	88	0.89
	070424_T2_NOR10.txt	T2	90	0.84
	070905_T27_NOR10.txt	T27	110	0.77
	070905_T2_NOR10.txt	T2	107	0.81
NOR11	**MEPVDPRLEPWKHPGSQPKTACTNCYCK**		**13**	
	070424_T27_NOR11.txt	T27	115	0.85
	070424_T2_NOR11.txt	T2	104	0.87
	070911_T27_NOR11.txt	T27	456	0.84
	070911_T2_NOR11.txt	T2	191	0.84
NOR12	**FVIHRLEPWLHPGSQHITASTN**		**35**	
	070621_NOR12_T2.txt	T2	101	0.87

	070621_NOR12_T27.txt	T27	105	0.83
	0706xx_Nor12_T27.txt	T27	38	0.74
	0706xx_Nor12_T2.txt	T2	56	0.82
NOR13	**ALSEGATPQDLNTMLNTVGGHQAAMQML**		**12**	
	070621_NOR13_T2.txt	T2	32	0.64
	070621_NOR13_T27.txt	T27	36	0.68
	0706xx_Nor13_T27.txt	T27	21	0.78
	0706xx_Nor13_T2.txt	T2	22	0.78
NOR14	**YKLKHIVWASRELERFAVNPGLLEVTSEGC**		**55**	
	070703_NOR14_T2.txt	T2	159	0.76
	070703_NOR14_T27.txt	T27	115	0.67
	070911_T2_NOR14.txt	T2	178	0.75
	090701_NOR14_T27.txt	T27	226	0.69
	090914_NOR14_T27.txt	T27	104	0.69
	090914_NOR14_T2.txt	T2	117	0.71
	091119_NOR14_Lcl.txt	Unspecified	67	0.62
NOR15	**CFHCQVCFITKGLGISYGRKKRR**		**8**	
	070703_T27_NOR15.txt	T27	77	0.91
	070703_T2_NOR15.txt	T2	81	0.80
	070904_T27_NOR15.txt	T27	110	0.68
	070904_T2_NOR15.txt	T2	86	0.68
NOR2	**PEVIPMFSALSEGATPQDLNTMLNTVGGH**		**32**	
	060808_NOR2_T2.txt	T2	65	0.73
	060808_NOR2_T27.txt	T27	45	0.69
	090728_NOR2_T27.txt	T27	332	0.73
	090728_NOR2_T2.txt	T2	329	0.70
NOR3	**NNPPIPVGEIYKRWIILGLNKIV**		**30**	
	060808_NOR3_T2.txt	T2	79	0.85
	060808_NOR3_T27.txt	T27	81	0.86
NOR4	**RALGPAATLQTPWTASLGVG**		**33**	
	061128_HeLa_NOR4.txt	Unspecified	60	0.82
	061128_T27_NOR4.txt	T27	49	0.79
	061128_T2_NOR4.txt	T2	108	0.83
	070308_T27_NOR4.txt	T27	72	0.87
	070308_T2_NOR4.txt	T2	63	0.84
NOR41	**STAGLYVLFLTKGLSISYLGKK**		**27**	
	071016_T27_NOR41.txt	T27	90	0.78
	071016_T2_NOR41.txt	T2	77	0.81
NOR42	**CFHSQVSFITKGLGISYGRKKRR**		**23**	
	071016_T27_NOR42.txt	T27	101	0.72
	071016_T2_NOR42.txt	T2	68	0.68
NOR5	**RAIPIPAGTLLSGGGRAIYKRWAILG**		**33**	
	070308_NOR5_T2.txt	T2	151	0.76
	070308_T27_NOR5.txt	T27	126	0.78
NOR6	**KALGPAATLEEMMTACQGVGGPGH**		**18**	
	070308_T27_NOR6.txt	T27	40	0.68
	070308_T2_NOR6.txt	T2	22	0.70
NOR8	**YAIPQALNTLLNTVGGHQAA**		**19**	
	070424_T27_NOR8.txt	T27	65	0.85
	070424_T2_NOR8.txt	T2	53	0.79
	070703_T27_NOR8.txt	T27	77	0.82
	070703_T2_NOR8.txt	T2	53	0.87
	070904_T27_NOR8.txt	T27	71	0.79
	070904_T2_NOR8.txt	T2	71	0.71
NOR9	**YVLFLTKGLSISYLGKK**		**30**	
	070424_T27_NOR9.txt	T27	52	0.77
	070424_T2_NOR9.txt	T2	74	0.87
	070911_T27_NOR9.txt	T27	99	0.89
	070911_T2_NOR9.txt	T2	101	0.90
pp89	**RLMYDMYPHFMPTNLGPSEKRVWMS**		**35**	
	061019_Kloe1_PC.txt	Unspecified	48	0.60
	061019_Kloe1_RC.txt	Unspecified	33	0.59
	080108_Kloe1_T2.txt	T2	211	0.82
	080108_Kloe1_T27.txt	T27	188	0.73

	100527_Kloe1_HC.txt	Unspecified	275	0.72
	100527_Kloe1_HD.txt	Unspecified	310	0.74
Sei102	**VFTWPPWQAGILARNLVPMVATVQGQNLKYGEF**		**18**	
	050523_Sei102_T27.txt	T27	435	0.75
	050523_Sei102_T2.txt	T2	403	0.71
	051031_Sei102_T2.txt	T2	386	0.55
	051031_Sei102_T27.txt	T27	306	0.61
	060126_Sei102_T2_27.txt	T2	63	0.90
	060126_Sei102_T2.txt	T2	38	0.70
	060712_Sei102_T2_27.txt	T2	121	0.78
	060712_Sei102_T2.txt	T2	87	0.54
Sei104	**AELELAENREILKEPVHGVYYDPSKDLIAE**		**10**	
	050523_Sei104_T2.txt	T2	364	0.86
	050523_Sei104_T27.txt	T27	354	0.85
	0602xx_Sei104_T27.txt	T27	188	0.88
	0602xx_Sei104_T2.txt	T2	132	0.73
Sei164	**ISSIFSRIGDPALNMENITSGL**		**14**	
	091001_Sei164_T2.txt	T2	28	0.78
	091001_Sei164_T27.txt	T27	33	0.89
Sei5	**VIDTLTCGFADLMGYIPLVGAPLGGAARALAHGVRVLEDGVNYA**		**93**	
	080915_Sei5_20S_human.txt	20S	486	0.60
	080915_Sei5_20S_mouse.txt	20S	465	0.61
	080915_Sei5_20S_rat.txt	20S	541	0.68
Sei52	**VIDTLTCGFADAMGYIPLVGAPLGGAARALAHGVRVLEDGVNYA**		**57**	
	030206_Sei52_T2.txt	T2	235	0.75
	030206_Sei52_T27.txt	T27	178	0.55
	030206_Sei5_T27.txt	T27	208	0.53
	030206_Sei5_T2.txt	T2	254	0.69
Sei96	**KGHGHSYTTAEELAGIGILTVILGVL**		**11**	
	041208_Sei-96_gamma.txt	Unspecified	236	0.66
	041208_Sei-96_K0.txt	Unspecified	277	0.86
Sei97	**KGHGHSYTTAEEAAGIGILTVILGVL**		**25**	
	041208_Sei-97_gamma.txt	Unspecified	299	0.79
	041208_Sei-97_K0.txt	Unspecified	280	0.76
	0501xx_Sei-97_gamma.txt	Unspecified	106	0.70
	0501xx_Sei-97_K0.txt	Unspecified	119	0.66
Ste5	**AYISSVAYGRQVYLKLSTNSHSTKVKA**		**44**	
	100603_Ste5_IA.txt	Unspecified	452	0.65
	100603_Ste5_IB.txt	Unspecified	457	0.67

Appendix 1: Dataset used for decision tree learning

Amino acid index database clusters

Cluster representative shown first in bold text.

Cluster 1

WERD780103	**Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule**
MIYS990103	Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues
WOLS870101	Principal property values for six non-natural amino acids and their application to a structure-activity relationship for oxytocin peptide analogues
WOLS870102	Principal property values for six non-natural amino acids and their application to a structure-activity relationship for oxytocin peptide analogues
AURR980104	Helix capping
AVBF000103	Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins
JANJ780103	Conformation of amino acid side-chains in proteins

ROBB760101	Conformational properties of amino acid residues in globular proteins
RICJ880107	Amino acid preferences for specific locations at the ends of alpha helices
CHOP780209	Prediction of the secondary structure of proteins from their amino acid sequence
VASM830101	Computed conformational states of the 20 naturally occurring amino acid residues and of the prototype residue alpha-aminobutyric acid
NAKH920107	The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins
TANS770106	Statistical mechanical treatment of protein conformation. 5. A multiphasic model for specific-sequence copolymers of amino acids
PONP800108	Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins
MITS020101	Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces
CHOP780214	Prediction of the secondary structure of proteins from their amino acid sequence
RICJ880116	Amino acid preferences for specific locations at the ends of alpha helices
YUTK870102	Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit
MUNV940102	Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales
KARP850103	Prediction of chain flexibility in proteins
WILM950104	Physicochemical basis of amino acid hydrophobicity scales: evaluation of four new scales of amino acid hydrophobicity coefficients derived from RP-HPLC of peptides
ANDN920101	Peptide/protein structure analysis using the chemical shift index method: upfield alpha-CH values reveal dynamic helices and aL sites
PARS000101	Protein thermal stability: insights from atomic displacement parameters (B values)
RACS820110	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
ROSM880102	Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds
HUTJ700103	Heat capacities, absolute entropies, and entropies of formation of amino acids and related compounds
MEIH800101	Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids
DESM900102	A critical evaluation of the hydropathy profile of membrane proteins
TSAJ990101	The packing density in proteins: standard radii and volumes
QIAN880133	Predicting the secondary structure of globular proteins using neural network models
RACS820103	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
AURR980107	Helix capping
PALJ810102	Protein secondary structure
ROBB790101	Refined models for computer simulation of protein folding: Applications to the study of conserved secondary structure and flexible hinge points during the folding of pancreatic trypsin inhibitor
MONM990201	Turns in transmembrane helices: determination of the minimal length of a "helical hairpin" and derivation of a fine-grained turn propensity scale
WILM950103	Physicochemical basis of amino acid hydrophobicity scales: evaluation of four new scales of amino acid hydrophobicity coefficients derived from RP-HPLC of peptides
PONJ960101	Deviations from standard atomic volumes as a quality measure for protein crystal structures
GOLD730101	Contribution of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure
ROBB760109	Conformational properties of amino acid residues in globular proteins
DAYM780101	Composition of proteins
QIAN880122	Predicting the secondary structure of globular proteins using neural network models
ZIMJ680103	The characterization of amino acid sequences in proteins by statistical methods
AVBF000102	Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins
GUYH850101	Amino acid side-chain partition energies and distribution of residues in soluble proteins
RACS820104	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
GEIM800110	Amino acid preferences for secondary structure vary with protein class
OOBM850105	Optimization of amino acid parameters for correspondence of sequence to tertiary structures of proteuins
CHOP780205	Prediction of the secondary structure of proteins from their amino acid sequence
NAKH900111	Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
NAGK730101	Local analysis of the mechanism of protein folding. I. Prediction of helices, loops, and beta-structures from primary structure
BASU050103	Principal eigenvector of contact matrices and hydrophobicity profiles in prote
QIAN880106	Predicting the secondary structure of globular proteins using neural network models
RADA880101	Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution
MUNV940103	Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales
KUMS000103	Factors enhancing protein thermostability
PONP800107	Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins
QIAN880109	Predicting the secondary structure of globular proteins using neural network models
ZIMJ680101	The characterization of amino acid sequences in proteins by statistical methods
DAWD720101	

KRIW710101	Local interactions as structure determinant for globular proteins
ROSG850102	Hydrophobicity of amino acid residues in globular proteins
LIFS790103	Antiparallel and parallel beta-strands differ in amino acid residue preference
ARGP820103	Structural prediction of membrane-bound proteins
HUTJ700102	Heat capacities, absolute entropies, and entropies of formation of amino acids and related compounds
LIFS790102	Antiparallel and parallel beta-strands differ in amino acid residue preference
NAKH920103	The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins
MEEJ800102	Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition
QIAN880107	Predicting the secondary structure of globular proteins using neural network models
QIAN880132	Predicting the secondary structure of globular proteins using neural network models
FAUJ830101	Hydrophobic parameters pi of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides
LEVM760102	A simplified representation of protein conformations for rapid simulation of protein folfing
QIAN880103	Predicting the secondary structure of globular proteins using neural network models
CORJ870108	Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins
PALJ810112	Protein secondary structure
QIAN880126	Predicting the secondary structure of globular proteins using neural network models
WERD780103	Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule
BUNA790101	1H-nmr parameters of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH
WEBA780101	Genetic code correlations: Amino acids and their anticodon nucleotides
ISOY800108	Characterization of multiple bends in proteins
GUYH850105	Amino acid side-chain partition energies and distribution of residues in soluble proteins

Cluster 2

KHAG800101	**The Kerr effect of amino acids in water**
MEIH800102	Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids
NAKH920102	The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins
AURR980113	Helix capping
LIFS790101	Antiparallel and parallel beta-strands differ in amino acid residue preference
GEIM800109	Amino acid preferences for secondary structure vary with protein class
NAKH900108	Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
BEGF750103	Une methode statistique simple de prediction des conformations proteiques
CHAM820102	The structural dependence of amino acid hydrophobicity parameters
CHOC750101	Structural invariants in protein folding
FAUJ880110	Amino acid side chain parameters for correlation studies in biology and pharmacology
GRAR740102	Amino acid difference formula to help explain protein evolution
MUNV940101	Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales
ROSM880104	Hydrophilicity of Polar Amino Acid Side-chains is Markedly Reduced by Flanking Peptide Bonds
FINA910102	Physical reasons for secondary structure stability: alpha-helices in short peptides
FAUJ880112	Amino acid side chain parameters for correlation studies in biology and pharmacology
CEDJ970101	Relation between amino acid composition and cellular location of proteins
AVBF000101	Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins
QIAN880113	Predicting the secondary structure of globular proteins using neural network models
WARP780101	A survey of amino acid side-chain interactions in 21 proteins
CORJ870105	Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins
NADH010105	Prediction of protein surface accessibility with information theory
PALJ810111	Protein secondary structure
TANS770101	Statistical mechanical treatment of protein conformation. 5. A multiphasic model for specific-sequence copolymers of amino acids
CHAM830107	The dependence of the Chou-Fasman parameters on amino acid side chain structure
RICJ880117	Amino acid preferences for specific locations at the ends of alpha helices
CHOP780201	Prediction of the secondary structure of proteins from their amino acid sequence
CORJ870101	Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins
MEEJ810102	Factors affecting retention and resolution of peptides in high-performance liquid chromatography
FAUJ880105	Amino acid side chain parameters for correlation studies in biology and pharmacology
RICJ880110	Amino acid preferences for specific locations at the ends of alpha helices
QIAN880134	Predicting the secondary structure of globular proteins using neural network models
QIAN880124	Predicting the secondary structure of globular proteins using neural network models
WOLR790101	Water, protein folding, and the genetic code
FAUJ880104	Amino acid side chain parameters for correlation studies in biology and pharmacology
SWER830101	Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure
NADH010101	Prediction of protein surface accessibility with information theory

FAUJ880107	Amino acid side chain parameters for correlation studies in biology and pharmacology
AURR980114	Helix capping
PALJ810103	Protein secondary structure
QIAN880116	Predicting the secondary structure of globular proteins using neural network models
NAKH900112	Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
HARY940101	Volume changes on protein folding
MAXF760106	Status of empirical methods for the prediction of protein backbone topography
PALJ810113	Protein secondary structure
CIDH920105	Hydrophobicity and structural classes in proteins
COHE430101	
KHAG800101	The Kerr effect of amino acids in water
KRIW790102	Local interactions as a structure determinant for protein molecules: II
BUNA790103	1H-nmr parameters of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH
CHAM830106	The dependence of the Chou-Fasman parameters on amino acid side chain structure
CHOP780215	Prediction of the secondary structure of proteins from their amino acid sequence
RACS820101	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
PALJ810110	Protein secondary structure
AURR980116	Helix capping
NAGK730103	Local analysis of the mechanism of protein folding. I. Prediction of helices, loops, and beta-structures from primary structure
NAKH900101	Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
QIAN880118	Predicting the secondary structure of globular proteins using neural network models
ZHOH040102	Quantifying the effect of burial of amino acid residues on protein stability
JANJ790102	Surface and inside volumes in globular proteins
ONEK900102	A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids
CORJ870104	Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins
QIAN880104	Predicting the secondary structure of globular proteins using neural network models
ZASB820101	Measurement of relative hydrophobicity of amino acid side-chains by partition in an aqueous two-phase polymeric system: Hydrophobicity scale for non-polar and ionogenic side-chains
FASG760102	
YANJ020101	GEM: a Gaussian Evolutionary Method for predicting protein side-chain conformations
RICJ880103	Amino acid preferences for specific locations at the ends of alpha helices
AURR980119	Helix capping

Cluster 3

AURR980118	**Helix capping**
PONP930101	Hydrophobic characteristics of folded proteins
EISD860103	Solvation energy in protein folding and binding
FINA770101	Theory of protein molecule self-organization. II. A comparison of calculated thermodynamic parameters of local secondary structures with experiments
GEOR030102	An analysis of protein domain linkers: their classification and role in protein folding
NOZY710101	The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions
ROBB760108	Conformational properties of amino acid residues in globular proteins
CHOP780204	Prediction of the secondary structure of proteins from their amino acid sequence
FUKS010111	Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria
CHAM830101	The dependence of the Chou-Fasman parameters on amino acid side chain structure
GEOR030101	An analysis of protein domain linkers: their classification and role in protein folding
MEIH800103	Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids
RADA880103	Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution
VINM940102	Accuracy of protein flexibility predictions
NAKH900107	Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
PALJ810108	Protein secondary structure
OLSK800101	Internal residue criteria for predicting three-dimensional protein structures
JOND920102	The rapid generation of mutation data matrices from protein sequences
QIAN880112	Predicting the secondary structure of globular proteins using neural network models
MUNV940105	Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales
KIDA850101	Statistical Analysis of the Physical Properties of the 20 Naturally Occuring Amino Acids
SUEM840102	Helix-coil stability constants for the naturally occurring amino acids in water. 22. Histidine parameters from random poly{(hydroxybutyl)glutamine-co-L-histidine}
KANM800104	Local hydrophobicity stabilizes secondary structures in proteins
AURR980111	Helix capping
OOBM770103	An analysis of non-bonded energy of proteins
ISOY800103	Characterization of multiple bends in proteins

BROC820101	The isolation of peptides by high-performance liquid chromatography using predicted elution positions
JUKT750101	Amino acid composition of proteins: Selection against the genetic code
ROBB760102	Conformational properties of amino acid residues in globular proteins
AURR980118	Helix capping
ROSM880105	Hydrophilicity of Polar Amino Acid Side-chains is Markedly Reduced by Flanking Peptide Bonds
MCMT640101	
SNEP660101	Relations between chemical structure and biological activity in peptides
GEIM800106	Amino acid preferences for secondary structure vary with protein class
RACS820107	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
RADA880107	Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution
KARP850102	Prediction of chain flexibility in proteins
AURR980115	Helix capping
OOBM850104	Optimization of amino acid parameters for correspondence of sequence to tertiary structures of proteuins
VASM830103	Computed conformational states of the 20 naturally occurring amino acid residues and of the prototype residue alpha-aminobutyric acid
BURA740102	Analysis of conformations of amino acid residues and prediction of backbone topography in proteins
PRAM820101	Shape and surface features of globular proteins
RADA880108	Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution
QIAN880119	Predicting the secondary structure of globular proteins using neural network models
GEOR030106	An analysis of protein domain linkers: their classification and role in protein folding
WOLR810101	Affinties of amino acid side chains for solvent water
FODM020101	Occurrence, conformational features and amino acid propensities for the pi-helix
PALJ810116	Protein secondary structure
QIAN880127	Predicting the secondary structure of globular proteins using neural network models
PARJ860101	New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: Correlation of predicted surface residues with antigencity and x-ray-derived accessible sites
AURR980120	Helix capping
KOEP990102	Structure-based conformational preferences of amino acids
QIAN880108	Predicting the secondary structure of globular proteins using neural network models
QIAN880123	Predicting the secondary structure of globular proteins using neural network models
BIGC670101	On the average hydrophobicity of proteins and the relation between it and protein structure
NAKH920108	The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins
BAEK050101	Prediction of protein inter-domain linker regions by a hidden Markov model
RICJ880105	Amino acid preferences for specific locations at the ends of alpha helices
WERD780102	Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule

Cluster 4

RACS820102	**Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids**
MIYS850101	Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation
RACS820102	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
CASG920101	Structure-derived Hydrophobic Potential. Hydrophobic Potential Derived from X-ray Structures of Globular Proteins is able to Identify Native Folds
FASG760101	
ROBB760110	Conformational properties of amino acid residues in globular proteins
CORJ870102	Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins
CHOP780203	Prediction of the secondary structure of proteins from their amino acid sequence
EISD840101	Three-dimensional structure of membrane and surface proteins
BLAM930101	Structural basis of amino acid alpha helix propensity
ZHOH040103	Quantifying the effect of burial of amino acid residues on protein stability
KANM800101	Local hydrophobicity stabilizes secondary structures in proteins
ZIMJ680105	The characterization of amino acid sequences in proteins by statistical methods
DESM900101	A critical evaluation of the hydropathy profile of membrane proteins
GOLD730102	Contribution of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure
RICJ880114	Amino acid preferences for specific locations at the ends of alpha helices
FASG890101	Prediction of Protein Structure and the Principles of Protein Conformation
PALJ810104	Protein secondary structure
NAKH900106	Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
LEWP710101	Folding of polypeptide chains in proteins: A proposed mechanism for folding
FUKS010107	Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria

CIDH920101	Hydrophobicity and structural classes in proteins
CHOP780206	Prediction of the secondary structure of proteins from their amino acid sequence
YUTK870101	Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit
CEDJ970104	Relation between amino acid composition and cellular location of proteins
WILM950102	Physicochemical basis of amino acid hydrophobicity scales: evaluation of four new scales of amino acid hydrophobicity coefficients derived from RP-HPLC of peptides
QIAN880131	Predicting the secondary structure of globular proteins using neural network models
CHAM810101	Protein folding and the genetic code: An alternative quantitative model
MAXF760102	Status of empirical methods for the prediction of protein backbone topography
SUYM030101	DomCut: Prediction of inter-domain linker regions in amino acid sequences
JOND750102	Amino acid properties and side-chain orientation in proteins: A cross correlation approach
QIAN880139	Predicting the secondary structure of globular proteins using neural network models

Cluster 5

TANS770108	**Statistical mechanical treatment of protein conformation. 5. A multiphasic model for specific-sequence copolymers of amino acids**
WOEC730101	Evolution of genetic code
RADA880106	Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution
GEOR030104	An analysis of protein domain linkers: their classification and role in protein folding
VINM940103	Accuracy of protein flexibility predictions
NAKH900103	Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
PALJ810101	Protein secondary structure
BULH740102	Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues
FAUJ880109	Amino acid side chain parameters for correlation studies in biology and pharmacology
VASM830102	Computed conformational states of the 20 naturally occurring amino acid residues and of the prototype residue alpha-aminobutyric acid
PALJ810109	Protein secondary structure
PONP800101	Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins
TANS770108	Statistical mechanical treatment of protein conformation. 5. A multiphasic model for specific-sequence copolymers of amino acids
GUOD860101	Prediction of peptide retention times in reversed-phase high-performance liquid chromatography
FINA910103	Physical reasons for secondary structure stability: alpha-helices in short peptides
LEVM780103	Conformational preferences of amino acids in globular proteins
CHOC760104	The nature of the accessible and buried surfaces in proteins
NAKH920101	The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins
GRAR740101	Amino acid difference formula to help explain protein evolution
AURR980117	Helix capping
RACS770102	Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins
GEIM800105	Amino acid preferences for secondary structure vary with protein class
KANM800103	Local hydrophobicity stabilizes secondary structures in proteins
BASU050102	Principal eigenvector of contact matrices and hydrophobicity profiles in prote
CHOP780101	Empirical predictions of protein conformation
AVBF000107	Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins
TSAJ990102	The packing density in proteins: standard radii and volumes
KUHL950101	Atomic and residue hydrophilicity in the context of folded protein structures
MIYS990104	Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues
CHAM830108	The dependence of the Chou-Fasman parameters on amino acid side chain structure
VENT840101	Hydrophobicity parameters and the bitter taste of L-amino acids
AURR980109	Helix capping
PARS000102	Protein thermal stability: insights from atomic displacement parameters (B values)
RACS820114	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
JURD980101	Protein transmembrane structure: recognition and prediction by using hydrophobicity scales through preference functions
GEIM800102	Amino acid preferences for secondary structure vary with protein class
RACS820111	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
CHOP780211	Prediction of the secondary structure of proteins from their amino acid sequence
NAKH920106	The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins
LEVM760105	A simplified representation of protein conformations for rapid simulation of protein folfing
GEOR030107	An analysis of protein domain linkers: their classification and role in protein folding
NADH010104	Prediction of protein surface accessibility with information theory

PRAM820102	Shape and surface features of globular proteins
AVBF000108	Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins
FUKS010103	Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria
GEIM800101	Amino acid preferences for secondary structure vary with protein class
VHEG790101	Trans-membrane translocation of proteins: The direct transfer model
QIAN880120	Predicting the secondary structure of globular proteins using neural network models
JOND750101	Amino acid properties and side-chain orientation in proteins: A cross correlation approach
WERD780104	Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule
RICJ880109	Amino acid preferences for specific locations at the ends of alpha helices
FUKS010109	Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria
CORJ870107	Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins
ROBB760113	Conformational properties of amino acid residues in globular proteins
FUKS010102	Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria
CRAJ730102	The reverse turn as a polypeptide conformation in globular proteins
CHOC760102	The nature of the accessible and buried surfaces in proteins
LEVM760107	A simplified representation of protein conformations for rapid simulation of protein folfing
PUNT030102	A knowledge-based scale for amino acid membrane propensity
CHOP780216	Prediction of the secondary structure of proteins from their amino acid sequence
EISD860102	Solvation energy in protein folding and binding
PALJ810107	Protein secondary structure
BULH740101	Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues
TANS770102	Statistical mechanical treatment of protein conformation. 5. A multiphasic model for specific-sequence copolymers of amino acids
JANJ790101	Surface and inside volumes in globular proteins
TANS770109	Statistical mechanical treatment of protein conformation. 5. A multiphasic model for specific-sequence copolymers of amino acids
ONEK900101	A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids

Cluster 6

YUTK870104	**Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit**
NISK800101	Prediction of the surface-interior diagram of globular proteins by an empirical method
RICJ880112	Amino acid preferences for specific locations at the ends of alpha helices
AVBF000106	Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins
FUKS010110	Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria
PUNT030101	A knowledge-based scale for amino acid membrane propensity
QIAN880128	Predicting the secondary structure of globular proteins using neural network models
KIMC930101	Thermodynamic beta-sheet propensities measured using a zinc-finger host peptide
PTIO830102	Theory of protein secondary structure and algorithm of its prediction
CHAM820101	The structural dependence of amino acid hydrophobicity parameters
ISOY800106	Characterization of multiple bends in proteins
BEGF750101	Une methode statistique simple de prediction des conformations proteiques
HOPA770101	
NAKH900102	Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
OOBM850101	Optimization of amino acid parameters for correspondence of sequence to tertiary structures of proteuins
BIOV880102	Secondary structure prediction: combination of three different methods
CHAM830103	The dependence of the Chou-Fasman parameters on amino acid side chain structure
CHAM830105	The dependence of the Chou-Fasman parameters on amino acid side chain structure
LEVM780104	Conformational preferences of amino acids in globular proteins
QIAN880121	Predicting the secondary structure of globular proteins using neural network models
CHOP780210	Prediction of the secondary structure of proteins from their amino acid sequence
NAKH900104	Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
GEOR030109	An analysis of protein domain linkers: their classification and role in protein folding
KRIW790101	Local interactions as a structure determinant for protein molecules: II
CHOP780208	Prediction of the secondary structure of proteins from their amino acid sequence
PONP800105	Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins
LEVM760103	A simplified representation of protein conformations for rapid simulation of protein folfing
FUKS010112	Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria
ISOY800102	Characterization of multiple bends in proteins
ROBB760103	Conformational properties of amino acid residues in globular proteins
NAKH900110	Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
TANS770105	Statistical mechanical treatment of protein conformation. 5. A multiphasic model for specific-sequence copolymers of amino acids
FAUJ880106	Amino acid side chain parameters for correlation studies in biology and pharmacology
NADH010107	Prediction of protein surface accessibility with information theory
GEIM800103	Amino acid preferences for secondary structure vary with protein class

BASU050101	Principal eigenvector of contact matrices and hydrophobicity profiles in prote
LEVM760101	A simplified representation of protein conformations for rapid simulation of protein folfing
RACS820109	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
NAKH900109	Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
TANS770104	Statistical mechanical treatment of protein conformation. 5. A multiphasic model for specific-sequence copolymers of amino acids
ROSG850101	Hydrophobicity of amino acid residues in globular proteins
CHOC760103	The nature of the accessible and buried surfaces in proteins
AURR980110	Helix capping
MIYS990105	Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues
LEVM780106	Conformational preferences of amino acids in globular proteins
OOBM770104	An analysis of non-bonded energy of proteins
RADA880104	Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution
QIAN880130	Predicting the secondary structure of globular proteins using neural network models
TANS770103	Statistical mechanical treatment of protein conformation. 5. A multiphasic model for specific-sequence copolymers of amino acids
ROBB760106	Conformational properties of amino acid residues in globular proteins
ROBB760107	Conformational properties of amino acid residues in globular proteins
BROC820102	The isolation of peptides by high-performance liquid chromatography using predicted elution positions
ROBB760111	Conformational properties of amino acid residues in globular proteins
AURR980105	Helix capping
HUTJ700101	Heat capacities, absolute entropies, and entropies of formation of amino acids and related compounds
FAUJ880113	Amino acid side chain parameters for correlation studies in biology and pharmacology
JANJ780101	Conformation of amino acid side-chains in proteins
GUYH850102	Amino acid side-chain partition energies and distribution of residues in soluble proteins
PRAM820103	Shape and surface features of globular proteins
WILM950101	Physicochemical basis of amino acid hydrophobicity scales: evaluation of four new scales of amino acid hydrophobicity coefficients derived from RP-HPLC of peptides
FAUJ880102	Amino acid side chain parameters for correlation studies in biology and pharmacology
DAYM780201	A model of evolutionary change in proteins
ISOY800107	Characterization of multiple bends in proteins
BURA740101	Analysis of conformations of amino acid residues and prediction of backbone topography in proteins
BLAS910101	Development of Hydrophobicity Parameters to Analyze Proteins Which Bear Post- or Cotranslational Modifications
RADA880105	Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution
ZIMJ680102	The characterization of amino acid sequences in proteins by statistical methods
RACS770103	Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins
AVBF000104	Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins
CHOP780207	Prediction of the secondary structure of proteins from their amino acid sequence
CEDJ970102	Relation between amino acid composition and cellular location of proteins
MANP780101	Hydrophobic character of amino acid residues in globular proteins
WOLS870103	Principal property values for six non-natural amino acids and their application to a structure-activity relationship for oxytocin peptide analogues
KARP850101	Prediction of chain flexibility in proteins
YUTK870104	Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit
ROBB760104	Conformational properties of amino acid residues in globular proteins
FUKS010108	Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria
MAXF760105	Status of empirical methods for the prediction of protein backbone topography
VINM940104	Accuracy of protein flexibility predictions
CRAJ730103	The reverse turn as a polypeptide conformation in globular proteins
OOBM850102	Optimization of amino acid parameters for correspondence of sequence to tertiary structures of proteuins

Cluster 7

RICJ880102	**Amino acid preferences for specific locations at the ends of alpha helices**
BIOV880101	Secondary structure prediction: combination of three different methods
KYTJ820101	A simple method for displaying the hydropathic character of a protein
RACS820108	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
RICJ880102	Amino acid preferences for specific locations at the ends of alpha helices
ZHOH040101	Quantifying the effect of burial of amino acid residues on protein stability
CIDH920103	Hydrophobicity and structural classes in proteins
PRAM900104	The distribution of physical, chemical and conformational properties in signal and nascent peptides

CHAM830102	The dependence of the Chou-Fasman parameters on amino acid side chain structure
LAWE840101	A simple experimental model for hydrophobic interactions in proteins
LEVM780105	Conformational preferences of amino acids in globular proteins
ARGP820102	Structural prediction of membrane-bound proteins
FAUJ880103	Amino acid side chain parameters for correlation studies in biology and pharmacology
CORJ870106	Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins
LEVM780101	Conformational preferences of amino acids in globular proteins
NAKH900105	Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
GARJ730101	Coefficients de partage d'aminoacides, nucleobases, nucleosides et nucleotides dans un systeme solvant salin
AURR980103	Helix capping
ARGP820101	Structural prediction of membrane-bound proteins
QIAN880136	Predicting the secondary structure of globular proteins using neural network models
OOBM770101	An analysis of non-bonded energy of proteins
AURR980108	Helix capping
KUMS000102	Factors enhancing protein thermostability
NAGK730102	Local analysis of the mechanism of protein folding. I. Prediction of helices, loops, and beta-structures from primary structure
QIAN880101	Predicting the secondary structure of globular proteins using neural network models
RACS820105	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
RICJ880106	Amino acid preferences for specific locations at the ends of alpha helices
ROBB760112	Conformational properties of amino acid residues in globular proteins
FAUJ880108	Amino acid side chain parameters for correlation studies in biology and pharmacology
FAUJ880111	Amino acid side chain parameters for correlation studies in biology and pharmacology

Cluster 8

QIAN880117	**Predicting the secondary structure of globular proteins using neural network models**
CIDH920104	Hydrophobicity and structural classes in proteins
FASG760104	
QIAN880125	Predicting the secondary structure of globular proteins using neural network models
QIAN880117	Predicting the secondary structure of globular proteins using neural network models
MONM990101	A turn propensity scale for transmembrane helices
TAKK010101	A new scale for side-chain contribution to protein stability based on the empirical stability analysis of mutant proteins
PALJ810115	Protein secondary structure
NAKH900113	Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
AURR980106	Helix capping
OOBM770102	An analysis of non-bonded energy of proteins
QIAN880114	Predicting the secondary structure of globular proteins using neural network models
CRAJ730101	The reverse turn as a polypeptide conformation in globular proteins
LEVM780102	Conformational preferences of amino acids in globular proteins
PTIO830101	Theory of protein secondary structure and algorithm of its prediction
COSI940101	Macromolecular bioactivity: is it resonant interaction between macromolecules?--Theory and applications
RACS820106	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
FUKS010104	Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria
PONP800103	Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins
KOEP990101	Structure-based conformational preferences of amino acids
TANS770110	Statistical mechanical treatment of protein conformation. 5. A multiphasic model for specific-sequence copolymers of amino acids
YUTK870103	Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit
WIMW960101	Experimentally determined hydrophobicity scale for proteins at membrane interfaces
GEOR030103	An analysis of protein domain linkers: their classification and role in protein folding
KANM800102	Local hydrophobicity stabilizes secondary structures in proteins
MAXF760101	Status of empirical methods for the prediction of protein backbone topography
CHOP780213	Prediction of the secondary structure of proteins from their amino acid sequence
GEOR030105	An analysis of protein domain linkers: their classification and role in protein folding
LEVM760106	A simplified representation of protein conformations for rapid simulation of protein folfing
AURR980101	Helix capping
FUKS010106	Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria
FINA910104	Physical reasons for secondary structure stability: alpha-helices in short peptides
RACS770101	Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins
ISOY800105	Characterization of multiple bends in proteins
CEDJ970105	Relation between amino acid composition and cellular location of proteins
PRAM900101	The distribution of physical, chemical and conformational properties in signal and nascent peptides

QIAN880110	Predicting the secondary structure of globular proteins using neural network models
MIYS990102	Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues
RADA880102	Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution
GEIM800104	Amino acid preferences for secondary structure vary with protein class
MAXF760103	Status of empirical methods for the prediction of protein backbone topography
SUEM840101	Helix-coil stability constants for the naturally occurring amino acids in water. 22. Histidine parameters from random poly{(hydroxybutyl)glutamine-co-L-histidine}
JANJ780102	Conformation of amino acid side-chains in proteins
JUNJ780101	The genetic code as a periodic table
QIAN880102	Predicting the secondary structure of globular proteins using neural network models
ZIMJ680104	The characterization of amino acid sequences in proteins by statistical methods
VINM940101	Accuracy of protein flexibility predictions
CEDJ970103	Relation between amino acid composition and cellular location of proteins
QIAN880135	Predicting the secondary structure of globular proteins using neural network models
CHOC760101	The nature of the accessible and buried surfaces in proteins
COWR900101	Hydrophobicity indices for amino acid residues as determined by high-performance liquid chromatography

Cluster 9

QIAN880138	**Predicting the secondary structure of globular proteins using neural network models**
NISK860101	Radial locations of amino acid residues in a globular protein: Correlation with the sequence
EISD860101	Solvation energy in protein folding and binding
ISOY800101	Characterization of multiple bends in proteins
ROSM880103	Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds
NADH010106	Prediction of protein surface accessibility with information theory
AVBF000109	Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins
GRAR740103	Amino acid difference formula to help explain protein evolution
NAKH920105	The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins
ROBB760105	Conformational properties of amino acid residues in globular proteins
BHAR880101	Positional flexibilities of amino acid residues in globular proteins
KUMS000104	Factors enhancing protein thermostability
ISOY800104	Characterization of multiple bends in proteins
SIMZ760101	
FINA910101	Physical reasons for secondary structure stability: alpha-helices in short peptides
LEVM760104	A simplified representation of protein conformations for rapid simulation of protein folfing
QIAN880138	Predicting the secondary structure of globular proteins using neural network models
GUYH850104	Amino acid side-chain partition energies and distribution of residues in soluble proteins
VELV850101	Is it possible to analyze DNA and protein sequences by the method of digital signal processing?
NADH010103	Prediction of protein surface accessibility with information theory
QIAN880111	Predicting the secondary structure of globular proteins using neural network models
RICJ880115	Amino acid preferences for specific locations at the ends of alpha helices
GUYH850103	Amino acid side-chain partition energies and distribution of residues in soluble proteins
PALJ810105	Protein secondary structure
JOND920101	The rapid generation of mutation data matrices from protein sequences
JACR890101	The nature of the hydrophobic bonding of small peptides at the bilayer interface: implications for the insertion of transbilayer helices
PLIV810101	Partition coefficients of amino acids and hydrophobic parameters pi of their side-chains as measured by thin-layer chromatography
GEIM800107	Amino acid preferences for secondary structure vary with protein class
NAKH920104	The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins
GEOR030108	An analysis of protein domain linkers: their classification and role in protein folding
PALJ810114	Protein secondary structure
SNEP660103	Relations between chemical structure and biological activity in peptides
SNEP660102	Relations between chemical structure and biological activity in peptides
SNEP660104	Relations between chemical structure and biological activity in peptides
CORJ870103	Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins
CHOP780212	Prediction of the secondary structure of proteins from their amino acid sequence
RICJ880113	Amino acid preferences for specific locations at the ends of alpha helices
OOBM770105	An analysis of non-bonded energy of proteins
AURR980102	Helix capping
GEIM800111	Amino acid preferences for secondary structure vary with protein class
HOPT810101	Prediction of protein antigenic determinants from amino acid sequecces
RACS820113	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids

FAUJ880101	Amino acid side chain parameters for correlation studies in biology and pharmacology
PONP800104	Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins
PRAM900102	The distribution of physical, chemical and conformational properties in signal and nascent peptides

Cluster 10

KLEP840101	**Prediction of protein function from sequence properties: Discriminant analysis of a data base**
WERD780101	Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule
BEGF750102	Une methode statistique simple de prediction des conformations proteiques
KUMS000101	Factors enhancing protein thermostability
ROSM880101	Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds
MUNV940104	Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales
GEIM800108	Amino acid preferences for secondary structure vary with protein class
NADH010102	Prediction of protein surface accessibility with information theory
QIAN880129	Predicting the secondary structure of globular proteins using neural network models
RICJ880104	Amino acid preferences for specific locations at the ends of alpha helices
CIDH920102	Hydrophobicity and structural classes in proteins
PRAM900103	The distribution of physical, chemical and conformational properties in signal and nascent peptides
RICJ880101	Amino acid preferences for specific locations at the ends of alpha helices
OOBM850103	Optimization of amino acid parameters for correspondence of sequence to tertiary structures of proteuins
AURR980112	Helix capping
RACS820112	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
DIGM050101	A comparison of proteins from Pyrococcus furiosus and Pyrococcus abyssi: barophily in the physicochemical properties of amino acids and in the genetic code
FASG760105	
MAXF760104	Status of empirical methods for the prediction of protein backbone topography
FASG760103	
PONP800106	Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins
MIYS990101	Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues
MEEJ800101	Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition
QIAN880105	Predicting the secondary structure of globular proteins using neural network models
PONP800102	Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins
CHAM830104	The dependence of the Chou-Fasman parameters on amino acid side chain structure
KLEP840101	Prediction of protein function from sequence properties: Discriminant analysis of a data base
MEEJ810101	Factors affecting retention and resolution of peptides in high-performance liquid chromatography
PALJ810106	Protein secondary structure
FUKS010101	Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria
KRIW790103	Local interactions as a structure determinant for protein molecules: II
ENGD860101	Identifying Nonpolar Transbilayer Helices in Amino Acid Sequences of Membrane Proteins
BUNA790102	1H-nmr parameters of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH
QIAN880115	Predicting the secondary structure of globular proteins using neural network models
QIAN880137	Predicting the secondary structure of globular proteins using neural network models
TANS770107	Statistical mechanical treatment of protein conformation. 5. A multiphasic model for specific-sequence copolymers of amino acids
CHOP780202	Prediction of the secondary structure of proteins from their amino acid sequence
RICJ880108	Amino acid preferences for specific locations at the ends of alpha helices
RICJ880111	Amino acid preferences for specific locations at the ends of alpha helices
FUKS010105	Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria
AVBF000105	Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins

Appendix 2: Clustered entries of the amino acid index database. The representatives of each cluster are displayed in bold lettering

Eidesstattliche Versicherung

„Ich, Justus Richard Pett, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: “Prediction of cleavage fragments generated by the proteasome” selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.

Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche in korrekter Zitierung (siehe „Uniform Requirements for Manuscripts (URM)“ des ICMJE -*www.**icmje**.org*) kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) entsprechen den URM (s.o) und werden von mir verantwortet.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Betreuer/in, angegeben sind. Sämtliche Publikationen, die aus dieser Dissertation hervorgegangen sind und bei denen ich Autor bin, entsprechen den URM (s.o) und werden von mir verantwortet.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§156,161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

Datum Unterschrift

Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Danksagung

Mein besonderer Dank gilt meinem Doktorvater Prof. Holzhütter für seine stets freundliche und konstruktive Unterstützung. Ebenso danke ich Dr. Sascha Bulik, der mir als Betreuer jederzeit mit Rat und Tat zur Seite stand und mich mit seinen Ideen und Hinweisen stets in die richtige Richtung geleitet hat.

Weiterhin gilt mein Dank meinen Freunden und Kollegen bei Ascaion, die es mir ermöglicht haben, meinen wissenschaftlichen Interessen nachzugehen.

Ich danke meinen Eltern, dass sie mich stets liebevoll zu einem interessierten Menschen großgezogen haben. Meiner Mutter gilt insbesondere mein Dank für Ihre stetige Unterstützung in den letzten Jahren.

Nicht zu vergessen seien meine Schwiegereltern in Korea, die mir aus der Ferne mentale Untersützung geschickt haben.

Zuletzt und vor allem danke ich meiner Frau Sangah, die mich mit ihrer unendlichen, liebevollen und loyalen Unterstützung Tag für Tag glücklich macht.