Universität Greifswald

Mathematisch-Naturwissenschaftliche Fakultät

Bachelorarbeit

Vorhersage von MHC-I-restringierten T-Zell-Epitopen auf *Staphylococcus aureus* mittels Automatisierung bekannter Prädiktionsalgorithmen

Zur Erlangung des akademischen Grades

Bachelor of Science

Angefertigt am Institut für Immunologie und Transfusionsmedizin der Universitätsmedizin Greifswald

Unter Betreuung von Prof. Dr. med. Barbara M. Bröker

Eingereicht von

Cedric Mahncke

Greifswald, März 2021

Inhaltsverzeichnis

Inhaltsverzeichnis ..... **ii**

Abkürzungsverzeichnis ..... **iv**

Überblick über die Arbeit ..... **vi**

1 Einleitung **1**

1.1 Immunologie 1
1.1.1 Das Immunsystem 1
1.1.2 T-Zellen 1
1.1.3 Antigenpräsentation 2
1.1.4 Aktivierung von $CD8^+$-T-Zellen 5

1.2 *Staphylococcus aureus* 7

1.3 Bioinformatik 11
1.3.1 *UniProt Knowledgebase* 11
1.3.2 Epitopvorhersage 12
1.3.3 SYFPEITHI 12

1.4 Künstliche Neuronale Netzwerke 14
1.4.1 NetMHCpan 15
1.4.2 Vergleich der Algorithmen 18

2 Verwandte Arbeiten **19**

3 Zielsetzung **20**

4 Methoden **21**

4.1 Abrufen der Proteindaten 21
4.1.1 Theorie 21
4.1.2 Implementierung 22

4.2 Vorhersagen von Epitopen 23
4.2.1 Theorie 23
4.2.2 Implementierung 24

4.3 Auswertung 25

5 Ergebnisse **27**

5.1 *Staphylococcus aureus* 27
5.1.1 Quantitativ 27
5.1.2 Qualitativ 29

5.2 Vergleichsdaten 31
5.2.1 Quantitativ 31
5.2.2 Qualitativ 33

5.3 Algorithmus 34

6 Diskussion **37**

6.1 Methoden 37

6.2 Bioinformatik 37
6.3 Vergleich der Epitopdaten innerhalb der Spezies *Staphylococcus aureus* 38
6.4 Vergleich der Epitopdaten ausgewählter Stämme von *S. aureus* und Viren 39

7 Ausblick 41
7.1 Bioinformatik 41
7.2 Biologie 41

8 Fazit 43

9 Literaturverzeichnis 44

10 Anhang 49
10.1 Abbildungen 49
10.2 Datenverfügbarkeit 56

Danksagung 58

Abkürzungsverzeichnis

Abkürzung	Bedeutung
ANN	*Artificial neural network*, künstliches neuronales Netzwerk
ANOVA	Einfaktorielle Varianzanalyse
APC	Antigenpräsentierende Zelle
API	*Application programming interface*, Programmierschnittstelle
ARS	Antibiotika-Resistenz-Surveillance
AS	Aminosäure
AUC	*Area under the curve*, Fläche unter der Kurve
BA	Bindungsaffinität
CD	*Cluster of differentiation*
$CD3^+$	CD3-positiv
CTL	*Cytotoxic t lymphocyte*, zytotoxische T-Zelle
DRiP	Defekte ribosomale Produkte
Dict	*Dictionary*
EBI	*European Bioinformatics Institute*
EL	*Eluted ligands*, gelöste Liganden
FPR	*False positive rate*, Falsch-Positiv-Rate
Glp	*Glycerophosphodiester phosphodiesterase*
GUI	*Graphics user interface*, grafische Benutzeroberfläche
GUMC	*Georgetown University Medical Center*
HLA	*Human leukocyte antigen*, humanes Leukozytenantigen
ID	Identifikationsnummer
IEDB	*Immune Epitope Database*
IFN	Interferon
Ig	Immunglobulin
IL	Interleukin
IQR	*Interquartile range*, Interquartilsabstand
MA	Multiallelisch
MHC	*Major histocompatibility complex*, Haupthistokompatibilitätskomplex
MRSA	Multi-Resistenter *Staphylococcus aureus*
MS	Massenspektrometrie

Mu50	*Mu50 / ATCC 700699*
N/A	*Not available*, nicht verfügbar
NET	*Neutrophile extracellular traps*, neutrophile extrazelluläre Fallen
PBMC	*Peripheral blood mononuclear cells*, mononukleäre Zellen des periphären Blutes
PBP	Penicillin bindendes Protein
PLC	*Peptide-loading complex*, Peptid-Beladungskomplex
PSGL	P-Selektin Glykoprotein Ligand
PSSM	*Position specific scoring matrix*, Positionsspezifische Bewertungsmatrix
Pwv.	Paarweise verschieden
RKI	Robert Koch-Institut
ROC	*Receiver operating characteristics*, Isosensitivitätskurve
S. aureus	*Staphylococcus aureus*
SARI	Surveillance der Antibiotika-Anwendung und bakteriellen Resistenzen auf Intensivstationen
Sbi	*Staphylococcus aureus binder of IgG*
SE	*Standard error*, Standardfehler
SIB	*Swiss Institute of Bioinformatics*
SIRS	Systemisches Inflammatorisches Response Syndrom
SM	*Score matrix*, Bewertungsmatrix
sod	Superoxiddismutasen
SpA	*Staphylococcal protein A*
Spl	*Serin protease*
SSL	*Staphylococcal superantigen-like protein*
TAP	*Transporter associated with antigen processing*
TC	*T-cell*, T-Zelle
TCR	*T-cell receptor*, T-Zell-Rezeptor
Th-Zelle	T-Helferzelle
TPR	*True positive rate*, Richtig-Positiv-Rate
Treg	Regulatorische T-Zelle
UniProt	*Universal Protein Resource*
URL	*Uniform resource locator*, einheitlicher Ressourcenzeiger

Überblick über die Arbeit

In dieser Arbeit werden T-Zell-Epitopdaten mehrerer Subproteomen von *Staphylococcus aureus* (*S. aureus*) automatisch zusammengetragen. Die untersuchten Proteome beziehen sich einerseits auf die funktionellen Lokalisationen von *S. aureus*-Proteinen, andererseits auf zwei *S. aureus*-Stämme, die dann mit Mikroben-stämmen von *Influenza A* und *SARS-CoV-2* verglichen werden. Die Epitopdaten basieren auf mathematisch vorhergesagten Neunersequenzen zu dem MHC-Allel HLA-A*02:01. Anhand der Epitopdichte werden die Datensätze vergleichend analysiert. Dadurch sollen mögliche Ansätze zur weiteren Untersuchung der humanen Immunantwort auf *S. aureus* geschaffen werden. Die Vergleiche zu Epitopdaten der Viren *Influenza A* und *SARS-CoV-2* setzen die Erkenntnisse in einen bekannten Kontext.

Dazu wird ein Algorithmus zur Automatisierung der Arbeitsschritte entwickelt. In Anbetracht der großen Datensätze spart die Methode viel Arbeit ein. Die zugrunde liegende Proteindatenbank wird die *Universal Protein Resource* sein. Die Proteinsequenzen werden automatisch an die Webserver der Prädiktionsalgorithmen SYFPEITHI und NetMHCpan weitergeleitet. Schließlich werden zusammenhängende Datensätze über Proteome und zugehörige Epitopvorhersagen produziert.

1 Einleitung

1.1 Immunologie

1.1.1 Das Immunsystem

Das Immunsystem schützt den Körper vor fehlerhaften körpereigenen Zellen (z.B. Tumorzellen), fremden Substanzen (z.B. Toxine) und invasiven Mikroorganismen (z.B. *S. aureus)*. Die beteiligten Organe, Zelltypen und Effektormoleküle sind in zwei immunologische Kompartimente unterteilt.

Der innate oder unspezifische Teil erkennt konservierte Strukturen und reagiert unmittelbar mit einer immer steten Antwort. Nicht unmittelbar, dafür hochspezifisch ist die Antwort des adaptiven Immunsystems. Die hier beteiligten Zellen erkennen auch bis dato unbekannte Strukturen und bilden daraufhin ein expositionsbedingtes Immungedächtnis aus. Beide sind wiederum in humorale und zelluläre Bestandteile unterteilt (Bröker et al. 2019).

1.1.2 T-Zellen

T-Zellen (*t-cell*, TC) bilden einen zentralen Teil der zellulären adaptiven Immunantwort. Im Knochenmark bildet sich aus der pluripotenten hämatopoetischen Stammzelle die lymphatische Vorläuferzelle. Die Vorläuferzelle der TC migriert früh in den namensgebenden Thymus und reift dort weiter. Da die Zelle noch keine Effektorfunktion hat, heißt sie naiv.

Die naive T-Zelle ist mit Rezeptoren einer einzigartigen Spezifität ausgestattet. Die Verteilung der Rezeptoren unterliegt der Klon-Selektionstheorie (Burnet 1959). Somit werden bei Kontakt einer T-Zelle mit einem Antigen nicht alle, sondern nur die für dieses Antigen spezifischen T-Zellen aktiviert. Der Kontakt wird dabei zwischen Rezeptormolekülen auf der T-Zelle (*t-cell receptor*, TCR) und einer antigenpräsentierenden Zelle (APC) hergestellt, die das Antigen auf ihrer Oberfläche präsentiert. Im Durchschnitt ist von einer Aktivierung nur eine aus 10^4 – 10^5 Zellen betroffen (Abdurrahman et al. 2020), welche dann durch Proliferation und Differenzierung zu einer von zwei Subklassen reift.

Jede Subklasse von T-Zellen hat verschiedene Erkennungsmerkmale und Funktionen. Das Unterscheidungsmerkmal ist das Oberflächenprofil aus Molekülen, *cluster of differentiation* (CD) genannt, das je nach Aktivierungs- und Differenzierungszustand charakteristisch ausgebildet wird. Die spezifischen CDs wurden laufend nummeriert (Bröker et al. 2019). T-Zellen im Allgemeinen unterscheiden sich von anderen Lymphozyten durch den Oberflächenproteinkomplex CD3, sie sind CD3-positiv ($CD3^+$). CD3 ist ein im Komplex mit dem TCR für die Aktivierung der TC wichtiger Corezeptor (Murphy und Weaver 2018). Funktionell gliedern sich die T-Zellen weiter in T-Helferzellen (Th-Zellen) und Zytotoxische T-Zellen (*cytotoxic t lymphocyte*, CTL), die $CD4^+$ respektive $CD8^+$ sind, und regulatorische T-Zellen (Treg).

$CD4^+$ TCs aktivieren unter anderem B-Zellen und modulieren die Immunantwort. Mittels des Corezeptors CD4 erkennen die Th-Zellen intrazellulär prozessierte Antigenfragmente auf professionellen APCs wie Makrophagen, B-Zellen und dendritischen Zellen (Andersen et al. 2006). Die Fragmente liegen in einem Komplex mit dem Membranglykoprotein MHC-II (*major histocompatibility complex*, Haupthistokompatibilitätskomplex) vor.

$CD8^+$ TCs hingegen zerstören jegliche für den Körper pathogen erscheinenden körpereigenen Zellen. Dafür werden die Erkennungsmerkmale von fast allen Zellen exprimiert, sodass bei Kontakt und signifikanten Gefahrensignalen verschiedene Effektormoleküle rekrutiert werden können. In diesem Fall bildet MHC-I den Kontaktpartner auf den Zielzellen. CTLs sekretieren nach Aktivierung unter anderem folgende immunogene Proteine: Granzyme, die die Freisetzung von Apoptose-einleitenden Zytokinen erhöhen, Perforin, das den Transmembrantransport von Granzymen unterstützt und das Zytokin Interferon-gamma (IFN-γ), welches die Erkennung der Zielzelle vom Organismus verstärkt und Makrophagen aktiviert.

1.1.3 Antigenpräsentation

T-Zellen erkennen Antigene aus allen Zellkompartimenten auf der Oberfläche der Zelle. Genauer gesagt bindet der TCR an einen Komplex aus MHC und einem Peptid geringer Länge, das aus einem oder mehreren Teilen eines Proteins intrazellulär prozessiert wurde, einem Epitop. Dadurch können auch Bestandteile aus dem Zytoplasma, zytoplasmatischen Kompartimenten und dem Extrazellularraum erkannt werden. Solche zytoplasmatische, auch endogene Antigene sind funktionell ausgediente Proteine und defekte ribosomale Produkte (DRiPs), welche etwa 30 % aller synthetisierten Proteine ausmachen (Murphy und Weaver 2018). Die Peptide werden nach der Prozessierung mit MHC-I /-II als Komplex an die Zelloberfläche transportiert und den antigenerkennenden Zellen präsentiert (Neefjes et al. 2011). MHC-Moleküle werden in einem Bereich des Genoms kodiert, der hochvariabel und histologisch ausschlaggebend für die Kompatibilität eines Transplantats ist. Die Menge aller von einer Zelle präsentierten Peptide heißt Immunopeptidom.

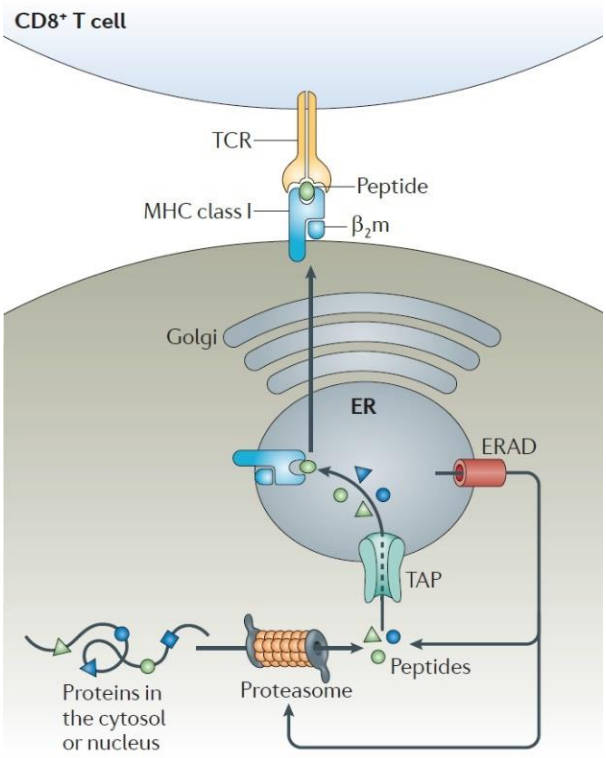

Abbildung 1.1: **Der Weg der Antigenpräsentation durch MHC-I**: Proteine werden im Proteasom in Peptidfragmente gespalten. Die Fragmente werden in das endoplasmatische Retikulum (ER) transportiert und auf MHC-I geladen. Dieser Komplex wird durch den Golgi-Apparat an die Oberfläche transportiert, wo er das Peptid präsentiert (Neefjes et al. 2011).

MHC-Moleküle sind die Träger des Epitops und präsentieren dieses als ihren Liganden. Die MHC-Moleküle werden funktionell in zwei Typen unterschieden: Typ 1, der Peptide aus dem Intrazellularraum (endogene) präsentiert und Typ 2 für Peptide aus Proteinen, die über Vesikel aus dem Extrazellularraum migrieren (exogene) und dann prozessiert werden. MHC-I wird von allen nukleären Zellen exprimiert, ist also sehr häufig und umso diverser. Es präsentiert Peptide für die CTL, wohingegen MHC-II nur auf professionellen APCs vorkommt und Peptide für Th-Zellen präsentiert.

Die Prozessierung der Proteine läuft in mehreren zytosolischen Kompartimenten ab (Abbildung 1.1). Im zytoplasmatischen Proteasom werden sie zu Peptidfragmenten abgebaut. Dabei können die Epitope kontinuierlich sein, also eine Folge direkt aufeinanderfolgende Aminosäuren (AS), oder diskontinuierlich aus mehreren Teilen zusammengesetzt sein, die innerhalb der Proteinsequenz auch weit voneinander entfernt liegen können. Im ER bindet MHC-I mehrere Proteine, die die Beladung mit dem Epitop unterstützen. Die Chaperone Calreticulin und Calnexin und der Cofaktor ERp57 stabilisieren und schützen das MHC-Molekül bis zur Bindung des Peptids. Dieser Peptid-Beladungs-Komplex (*peptide-loading complex*, PLC) ist an Tapasin gebunden, das wiederrum den Peptidtransporter TAP (*transporter associated with antigen processing*) bindet. TAP schleust das Peptid in das ER. Sobald ein Peptidfragment durch TAP in das ER transloziert und an MHC-I gebunden

wurde, separiert sich der MHC-I-Peptid-Komplex von TAP, ERp57 und den Chaperonen und wird durch den Golgi-Apparat an die Zelloberfläche transportiert. Dort präsentiert MHC-I das Peptid als Epitop in den Extrazellularraum für CTLs. Exogene Antigene werden mittels Endosomen und Lysosomen, die auch endosomale Proteasen enthalten, in die Zelle aufgenommen. Die Degradation und Beladung von MHC-II findet weiterhin in den Vesikeln statt. Das fertige Vesikel gelangt dann an die Zelloberfläche, wo der MHC-II-Peptid-Komplex das Epitop membranassoziiert für Th-Zellen präsentiert (Neefjes et al. 2011).

Die MHC-Moleküle präsentieren Peptide mit unterschiedlichen Eigenschaften. Die Epitope auf MHC-I sind 8 – 9 Aminosäuren lang (Okta-, Nonamere). Einerseits ist die Länge beschränkt durch den TAP-Komplex, der Peptide der Länge 8 – 16 bevorzugt transportiert. Diese müssen zusätzlich basische beziehungsweise hydrophobe Restgruppen an den Aminosäuren am Carboxylende des Peptids haben. Darüber hinaus stabilisiert das MHC-I-Molekül das Peptid an seinen Enden mit Wasserstoffbrücken und ionischen Bindungen an den unveränderlichen Bereichen der Bindungsstelle. Peptide, die länger sind, ragen über die Bindungsfurche von MHC-I hinaus – der Überhang wird von Exopeptidasen des ER abgespalten. MHC vom Typ II hingegen stabilisiert das Peptid über seine gesamte Kontaktfläche, statt nur an den Enden. Da die Bindung zudem in Vesikeln stattfindet und somit TAP-unabhängig ist, ist die Bindung von deutlich längeren Peptiden von mindestens 13 Aminosäuren möglich (Murphy und Weaver 2018; Neefjes et al. 2011).

Die genetische Region der MHCs ist von größter Variabilität. Im Humangenom HLA (*human leukocyte antigen*) genannt, besteht die Region aus jeweils drei Genloci, HLA-A, -B und -C für MHC-I und HLA-P, -Q und -R für MHC-II, sie ist polygen. Innerhalb dieser Loci befinden sich mehrere Allele, die mit Zahlen bezeichnet werden (beispielsweise HLA-A*02:01). Sie ist also polymorph, und zwar von höchstem bisher von humanen Genen bekannten Grad. Der Genpool einiger MHC-I- und -II-Gene umfasst über 1000 Allele (Murphy und Weaver 2018), wobei alle relativ häufig vorkommen und somit die Wahrscheinlichkeit für einen heterozygoten MHC-Locus sehr groß ist. Der große Polymorphismus und die codominante Expression beider Allele können bei Vererbung die Vielfalt der MHC-Moleküle noch weiter erhöhen (Murphy und Weaver 2018).

Die Peptidbindungsfurche ist dabei an den Stellen der Verankerungen des Peptids der variabelste Ort. So unterscheiden sich selbst Produkte eines einzelnen Allels um bis zu 20 Aminosäuren. Eine Gruppe solcher Ankerreste in einem MHC-Molekül wird Sequenzmotiv genannt. Je nach Ausstattung des Motivs mit AS werden verschiedene Epitope gebunden, was auch die Epitope höchst heterogen macht. Dieses Motiv lässt sich identifizieren und darauf aufbauend die Peptide innerhalb eines Proteins, die dafür spezifisch binden können. Folglich kann man dann abschätzen, welcher Teil eines Proteins von MHC präsentiert wird und eine T-Zell-Antwort auslöst (Janeway, JR. et al. 2001).

1.1.4 Aktivierung von $CD8^+$-T-Zellen

Naive, also inaktive $CD8^+$ TCs werden über verschiedene Costimulationswege aktiviert. Während sie durch die sekundären lymphatischen Organe wandern, interagieren CTLs ständig mit professionellen APCs. Sie prüfen, ob ihr Rezeptor für das präsentierte Peptid spezifisch ist. Falls die TC eine solche Zelle findet, bindet sie über den MHC-I-Peptid-Komplex an diese und die Aktivierung wird initiiert. Reife dendritische Zellen können die CTL ausreichend stark costimulieren, da CD80 oder -86 an CD28 auf der TC binden können. Diese Selbstcostimulation verstärkt die Signalkaskade. Ansonsten ist für die Costimulation eine Th-Zelle nötig. Die Bindung von $CD4^+$ TCs stimuliert die sonst zu geringe Freisetzung von CD40 und ermöglicht die Proliferation und Differentiation der CTL. Das geschieht in den sog. Proliferationszonen (Milz, Lymphknoten und -follikel) (Murphy und Weaver 2018).

Die Aktivierung von CTLs wird durch zwei Signale vermittelt. Das erste Signal wird freigesetzt, wenn der TCR an den MHC-Komplex bindet und die Bindung des CD8-Moleküls an MHC initiiert (Signal 1, Antigenerkennung). Daraufhin binden Rezeptoren der naiven CTL, wie CD28 und CD70, als Liganden an Costimulatoren, wie B7. Diese Bindungen induzieren die Proliferation der Zelle und bilden Signal 2 (Costimulatorisches Signal). Daraufhin wird in der CTL die Synthese des Zytokins Interleukin-2 (IL2) aktiviert, es wird freigesetzt und bindet wiederum an den IL2-Rezeptor auf der CTL, was die Proliferation unterstützt und die Differentiation einleitet. Ist der costimulatorische Effekt der Ligandenbindung nicht ausreichend für die Auslösung der Proliferation und von Signal 2, kann eine Th-Zelle über die Ligandenbindung CD40-CD40L die nötige Aktivität in der APC auslösen (Murphy und Weaver 2018).

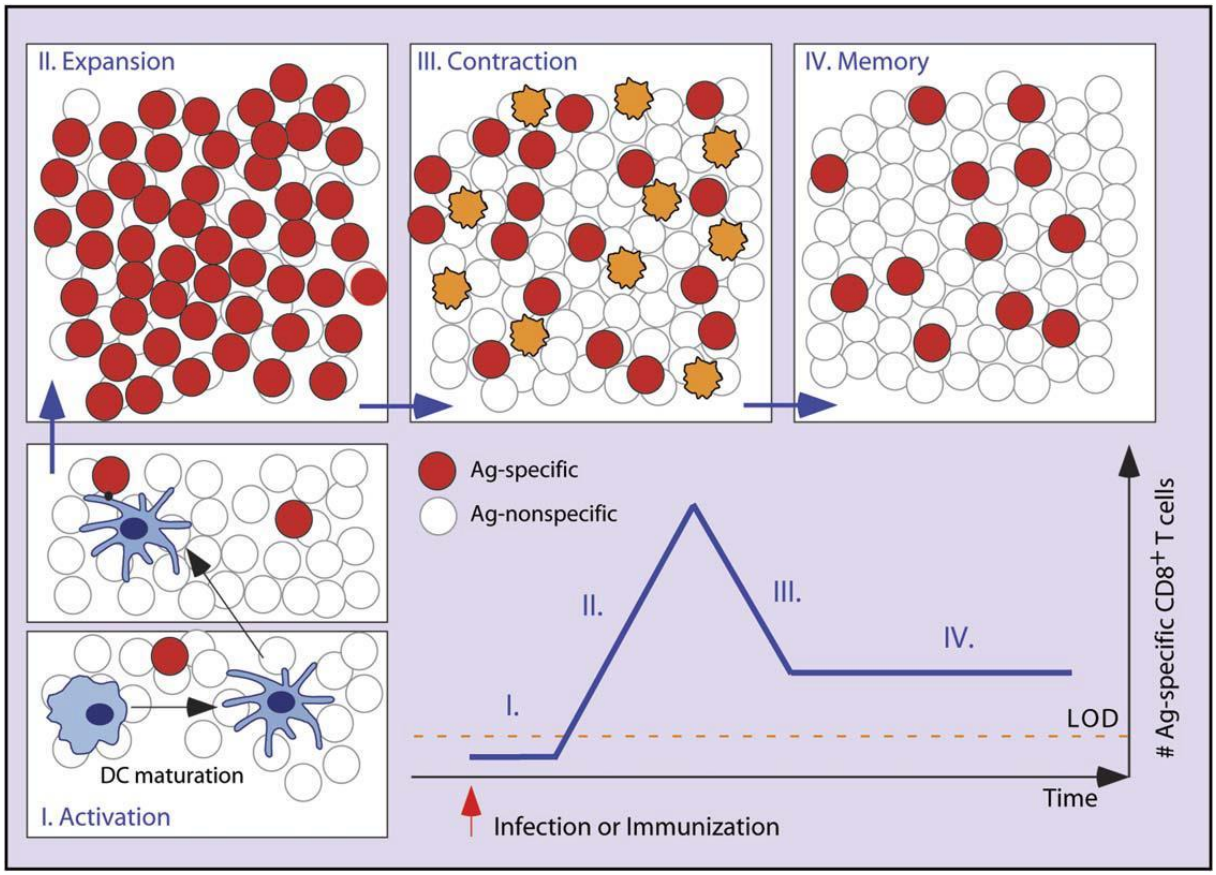

Abbildung 1.2: **Entwicklung von CD8⁺ T-Zellen nach Infektion oder Vakzinierung**: Von einer DC präsentierte antigene Peptide aktivieren (I.) CD8⁺ T-Zellen und bringen sie zur Expansion und Differenzierung (II.). 5-10 % der am Höhepunkt der Expansion verfügbaren T-Zellen überleben die Kontraktion (III.) und bilden die spezifischen Gedächtnis-T-Zellen (Haring et al. 2006).

Die CTL ist nach ihrer Aktivierung (Phase I) effektiv zytotoxisch und durchläuft drei weitere Phasen (Abbildung 1.2). Die aktivierte CTL vermehrt sich nun über klonale Expansion, wobei die Anzahl spezifischer CTLs in der ersten Woche um mehr als das 10.000-fache ansteigt (Expansion, Phase II) (Haring et al. 2006; Kaech et al. 2002). Durch die klonale Expansion haben alle Nachkommen die gleiche Antigenspezifität. Die aktivierten CTLs binden jetzt an alle Zellen über unspezifische Adhäsionsmoleküle und prüfen sie dann auf die Antigenspezifität. Ist diese negativ, trennen sich die Zellen wieder, bei spezifischer Erkennung entsteht jedoch eine stabile Paarung mit der Zielzelle. Daraufhin setzt die nun aktivierte CTL ohne Costimulation die unter 1.1.2 beschriebenen Effektormoleküle Granzyme, Perforine und IFN-γ frei. Somit werden verstärkt Peptide präsentiert und die immunologische Aktivität im Bereich der aktiven CTL erhöht (Wagner und Dannecker 2007).

In der folgenden Woche werden 90 % – 95 % der Effektorzellen eliminiert. Das gibt Raum und Ressourcen für andere CD8⁺ TC-Antworten frei (Kontraktion, Phase III) (Haring et al. 2006; Sprent und Tough 2001). Die überlebenden Zellen bilden anschließend das CD8⁺-T-Zell-Gedächtnis (Erinnerung, Phase IV). Dabei wird angenommen, dass nicht etwa eine stete Antigenpräsentation durch professionelle APC nötig ist, sondern die initiale Aktivierung die drei Phasen in Gang setzt und diese im

Folgenden weitestgehend unabhängig weiterer Präsentation auf APCs ablaufen (Haring et al. 2006; Kaech und Ahmed 2001; Mercado et al. 2000; van Stipdonk et al. 2001).

Aufgrund des Zusammenhanges des Epitops mit dem TCR, MHC-Molekül und dem Protein sind Epitope in der spezifischen Immunabwehr von zentraler Bedeutung und potentiell großem Nutzen.

1.2 *Staphylococcus aureus*

S. aureus ist ein weit verbreitetes kommensales, dennoch potentiell pathogenes Bakterium. Es hat einen fakultativ anaeroben Stoffwechsel und bildet als Teil der Gattung Staphylokokken runde Körper, die zu Haufen organisiert sind, sich weder aktiv bewegen können, noch Sporen bilden und grampositiv sind. Von anderen Staphylokokken-Spezies unterscheidet sich *S. aureus* durch aktive Koagulase (Taylor und Unakal 2019). Neben Nahrungsmitteln und Gewässern besiedelt *S. aureus* warmblütige Tiere, in 20 % – 30 % aller Menschen besiedelt *S. aureus* üblicherweise die Haut, Schleimhäute und die oberen Atemwege (van Belkum et al. 2009). Immunsupprimierte und hospitalisierte Menschen, sowie deren Kontaktpersonen, wie Pflegepersonal sind sogar zu 80 % von *S. aureus* bevölkert. Dabei sind 15 % – 20 % permanenter und 50 % – 70 % passagerer Natur (Abeck 2018). Das Bakterium ist über direkten Kontakt und Infektionsträger wie Instrumente und Lebensmittel übertragbar (Taylor und Unakal 2019).

S. aureus ist als Erreger für einen Großteil invasiver Infektionen und Toxin-bedingter Krankheiten ursächlich (*Tabelle 1.1*). Dabei kann es jedes Organ befallen, wobei kein anderes Pathogen alleine für mehr Infektionen ursächlich ist (Archer 1998). In Europa sind allein die Antibiotika-resistenten Stämme jährlich im Schnitt für 170.000 Infektionen mit über 5.000 Toden und über 1.000.000 Hospitalisierungstagen verantwortlich. In Deutschland liegt die Zahl für MRSA mit 132.000 niedriger, allerdings macht MRSA nur 18 % – 20 % der pathogenen *S. aureus*-Stämme aus. Die Gesamtzahl an *S. aureus*-Infektionen in deutschen Krankenhäusern beträgt somit knapp 700.000 (Köck et al. 2011).

Die Pathogenese verläuft in fünf Phasen. Sie beginnt dabei mit der Kolonisierung des Bakteriums im oder am Wirt. Bei einer Verletzung des Gewebes und Eindringen von *S. aureus* kann es anschließend zu einer lokalen Infektion kommen. Ist die Immunantwort nur unzureichend, verbreitet sich der Erreger invasiv systemisch im Körper, teils begleitet von Sepsis. Das kann metastasierende Infektionen auslösen und am Ende in Toxikose des Körpers enden. Die Gefahr für den Wirt geht primär von toxischen Produkten des *S. aureus* aus, die verschiedene Syndrome, wie das Systemische Inflammatorische Response Syndrom (SIRS) und als Folge einen septischen Schock auslösen können. Nach der Metastasierung steigt die Mortalität umso stärker an (Archer 1998).

Aktuelle Statistik meldepflichtiger Infektionskrankheiten, Deutschland
2. Woche 2018 (Datenstand: 31. Januar 2018)

Krankheit	2018 2. Woche	2018 1.–2. Woche	2017 1.–2. Woche	2017 1.–52. Woche
Adenovirus-Konjunktivitis	23	43	29	703
Brucellose	1	1	1	40
Chikungunyavirus-Erkrankung	1	1	2	32
Clostridium-difficile-Erkrankung, schwere Verlaufsform	74	123	104	2.767
Creutzfeldt-Jakob-Krankheit *	0	0	2	68
Denguefieber	8	13	8	626
FSME	0	0	2	476
Hämolytisch-urämisches Syndrom (HUS)	0	0	5	95
Haemophilus influenzae, invasive Infektion	34	65	54	803
Hantavirus-Erkrankung	2	9	24	1.697
Hepatitis D	0	0	0	31
Hepatitis E	70	108	71	2.911
Influenza	2.346	3.600	5.515	93.517
Legionellose	22	45	31	1.270
Leptospirose	2	5	2	126
Listeriose	15	31	31	762
Methicillin-resistenter *Staphylococcus aureus* (MRSA), invasive Infektion	39	94	111	2.663
Ornithose	0	0	1	11
Paratyphus	0	1	0	42
Q-Fieber	1	1	0	107
Trichinellose	0	0	0	2
Tularämie	0	0	0	47
Typhus abdominalis	0	0	0	78

* Übermittelte Fälle insgesamt, bisher kein Fall einer vCJK

Tabelle 1.1 **Statistik meldepflichtiger Infektionskrankheiten in Deutschland**: Zum Stand 31.01.2018 nahmen MRSA-Infektionen tendenziell ab, machen aber dennoch den viertgrößten Anteil der meldepflichtigen Infektionskrankheiten aus (Robert Koch-Institut 2018).

Resistenzen machen *S. aureus* zum klinisch relevantesten Erreger. Resistente Stämme gegen die gängigsten Antibiotika (Multi-Resistenter *Staphylococcus aureus*, MRSA) machten unter allen klinischen *S. aureus*-Isolaten, die von der Antibiotika-Resistenz-Surveillance (ARS) am Robert-Koch-Institut (RKI) 2016 erfasst wurden 10,6 % aus. Die Surveillance der Antibiotika-Anwendung und bakteriellen Resistenzen auf Intensivstationen (SARI) erfasste in Deutschland sogar eine MRSA-Rate von 21,3 %. Derzeit ist eine Abnahme (2010: 16 % (ARS), 2011: 27,2 % (SARI)) der MRSA-Prävalenz aufgrund steigender Hygienemaßnahmen und konsequenter Meldepflicht zu beobachten. Dennoch ist die Inzidenz mit 3,8 in Deutschland hoch, wobei lokal hohe Unterschiede auftreten. So sank die Inzidenz in Baden-Württemberg 2016 auf 1,4, stieg aber in Sachsen-Anhalt auf 8,7 (Robert Koch-Institut 2018).

Ursächlich für die Resistenz gegenüber Penicillinen ist das *mecA* Gen in der *Staphylococcal chromosomal cassette* Region. Es codiert ein Penicillin bindendes Protein (PBP). Das PBP-2a bindet β-Lactam, also Penicillin und von Penicillin abgeleitete Antibiotika und so können MRSA-Stämme auch in Gegenwart von Antibiotika leben (Taylor und Unakal 2019).

S. aureus verfügt über vielfältige Mechanismen der Immunevasion. *S. aureus* schränkt die Mobilität von Neutrophilen durch Sekretion von *staphylococcal superantigen-like 5* (SSL5) ein. SSL5 bindet das für die Bewegung wichtige Protein PSGL-1 (P-Selektin Glykoprotein Ligand), das sonst an Zelladhäsionsmoleküle bindet (Abbildung 1.3). Über diverse weitere SSL-Proteine moduliert *S. aureus* die Aktivierung und Chemotaxis von Neutrophilen. Von Neutrophilen aufgespannte Netze aus Chromatin (*neutrophile extracellular traps*, NET) degradiert *S. aureus* mittels Nuc, einer sekretierten Nuclease. Das Pigmentprotein *Staphyloxanthin* und die Superoxiddismutasen sodA und sodM beeinträchtigen entscheidend die Sauerstoff-abhängigen Zelltodmechanismen. Auch gegen Sauerstoff-unabhängige Apoptose sekretiert *S. aureus* Moleküle (Jong et al. 2019).

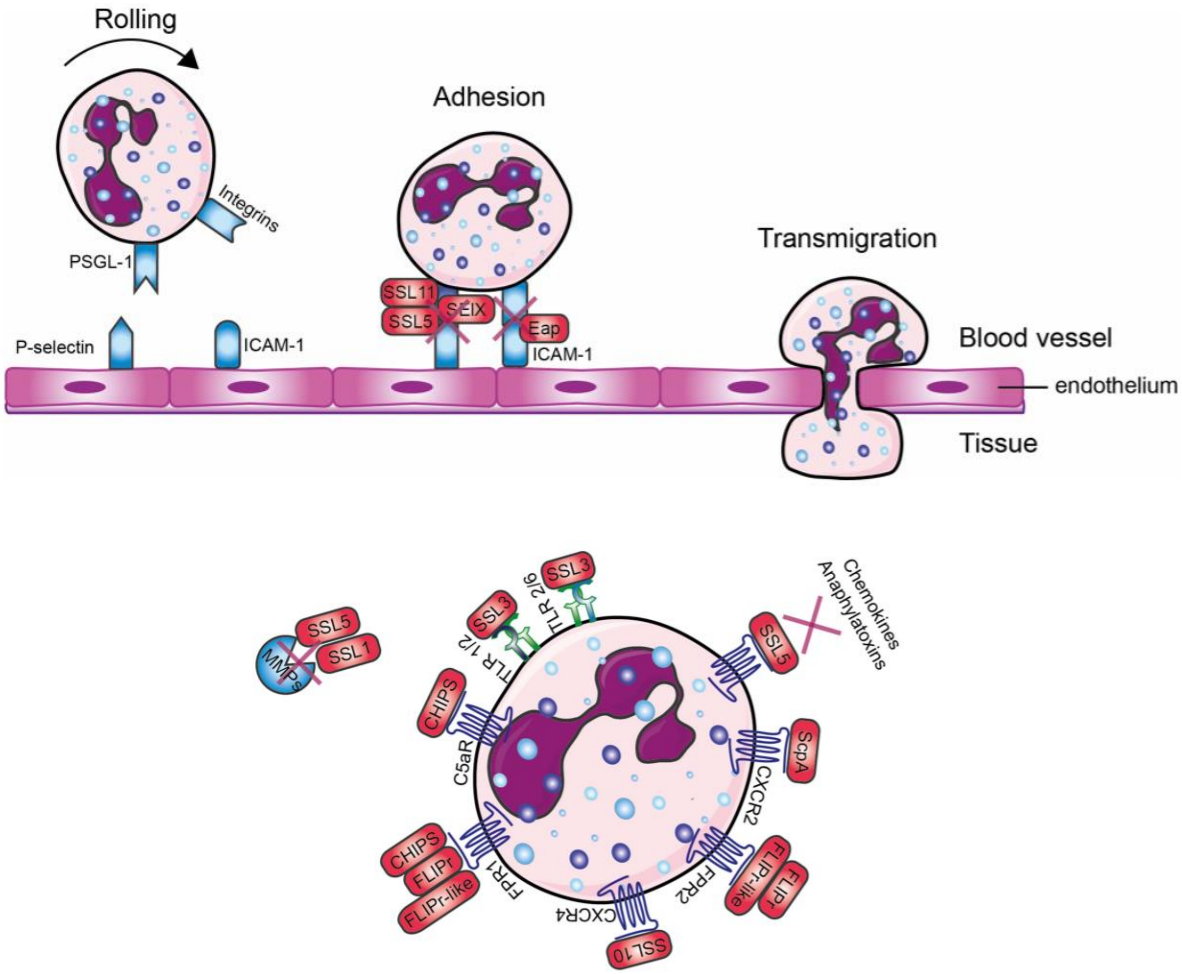

Abbildung 1.3: **Evasion der neutrophilen Antwort**: Oben: Proteine des *S. aureus* (rot) verhindern die Bindung von Zelladhäsionsmolekülen der Wirtszellen (blau) und somit die Migration an den Infektionsort. Unten: Proteine des *S. aureus* (rot) verhindern das Priming, die Chemotaxis und Aktivierung von Neutrophilen durch Besetzen der Bindungsstellen von Wirtsproteinen (blau) (Jong et al. 2019).

Während der Kolonisierung umgeht *S. aureus* die Immunantwort unter anderem durch *Staphylococcal protein A* (SpA) und *Staphylococcus aureus binder of IgG* (Sbi). Diese befinden sich auf der Oberfläche von *S. aureus* und binden an Antikörper, die für die Opsonierung für Makrophagen essenziell wären. Dann können die Komplexe durch Hydrolyse von der Membran gelöst und freigesetzt werden. SpA und Sbi binden die Fc Region des IgG, SpA zusätzlich die Fab-Region des IgM. Darüber hinaus kann *S. aureus* intrazellulär überleben. Dies wird ermöglicht durch Fibrinogen-bindende Proteine, Clumping-Faktoren und Teichonsäuren, die an Integrine binden und daraufhin in die Zelle eindringen (Taylor und Unakal 2019; Hauck et al. 2006). Das ermöglicht ein endogenes chronisches Reservoir, um so eine Immunantwort oder Behandlung zu überdauern und wieder zu kolonisieren.

Darüber hinaus produziert *S. aureus* Superantigene. Das heißt die Antigene aktivieren die T-Zell-Antwort antigenunspezifisch, indem sie eine Kreuzvernetzung der variablen β-Kette des TCR mit MHC-II-Molekülen herstellen. Die Folge ist eine polyklonale Vermehrung der T-Zellen und unkontrollierte

Zytokinsynthese, was zu einer überschwingenden und unspezifischen Immunreaktion führt; bis zu 20 % aller T-Zellen können so von einem Antigen aktiviert werden (Murphy und Weaver 2018). Davon sind Epithel- und Endothelzellen, Keratinozyten, Osteoklasten und sowohl professionell, als auch nicht-professionell phagozytierende Zellen betroffen (Bröker et al. 2016).

CD8+ TCs könnten von entscheidender Bedeutung gegen chronische Infektionen sein. Sie sind es, die intrazelluläre Antigene erkennen und den direkten Zelltod einleiten können. Dennoch sind die Auswirkungen von CTLs bei einer Infektion mit *S. aureus* noch relativ unbekannt. Grundlegend sind TCs wichtig, die einen Anstieg von IFN-γ bewirken. Dieses Zytokin bildet den Großteil der Immunantwort, wenn *S. aureus* intrazellulär persistiert (Bröker et al. 2016). Bisherige Versuche ein Vakzin zu designen zielten auf Oberflächen-Antigene von *S. aureus* ab und schlugen fehl. Resultierend aus den Studien wird der Ansatz diskutiert, nicht das Bakterium selber, sondern die Toxine als wichtige Virulenzfaktoren für die Kolonisierung von *S. aureus* als Ziele für Vakzine zu untersuchen (Miller et al. 2020).

Der *S. aureus*-Stamm *USA300* ist virulenter, resistenter gegen die humane und antibiotische Abwehr und bewirkt stärkere Zellschädigung am Wirt als andere Stämme (Strauß et al. 2017). Er ist weit verbreitet in Westeuropa und Nordamerika und dort von großer Bedeutung in Kohorten-assoziierten Ausbrüchen in Kliniken, Sportmannschaften und ähnlichen. Daher ist *USA300* bereits Gegenstand vieler wissenschaftlicher Untersuchungen und wird hier vergleichend betrachtet. MHC-I könnte von Nutzen sein, wenn *S. aureus* intrazellulär persistiert. Die Produkte von *S. aureus* würden bei intrazellulärem Überleben als Antigene prozessiert und die Peptidteile von MHC-I auf der Oberfläche der Wirtszelle für CTLs sichtbar gemacht werden.

1.3 Bioinformatik

1.3.1 *UniProt Knowledgebase*

Die *Universal Protein Knowledgebase* bietet eine große Quelle für Proteinsequenzen. Sie ist als Datenbank Teil der *Universal Protein Resource* (UniProt), einem Zusammenschluss des *European Bioinformatics Institute* (EBI), *Swiss Institute of Bioinformatics* (SIB) und des *Georgetown University Medical Center* (GUMC) mit deren Datenbanken *TrEMBL*, *Swiss-Prot* und *PIR.* Fast 210 Millionen Proteinsequenzen können zurzeit frei zugänglich abgerufen und genutzt werden, wobei teils grundlegende Informationen wie Länge, Gewicht oder Stamm, aber auch funktionelle Daten, wie der Bereich des Signalpeptids, verfügbar sind. Aus der ursprünglich manuellen Datenbank *Swiss-Prot* des SIB wurden über 560.000 Einträge übernommen. Darüber hinaus werden mit *TrEMBL* bis heute über 209 Millionen Proteinsequenzen als Translationsprodukte automatisch aus den Gen-Datenbanken *EMBL* (Europa), *GenBank* (USA), und *DDBJ* (Japan) hinzugefügt. Diese bilden die große Mehrheit,

müssen allerdings von Experten überprüft und dann als *reviewed* markiert in die *Swiss-Prot* übernommen werden. Zusammen bilden die Datenbanken die *Knowledgebase*, die die Basis der UniProt darstellt (The UniProt Consortium 2020a, 2020b).

Neben den Sequenzen enthalten die Einträge der UniProt weitere detaillierte Informationen. So lassen sich aus der Datenbank alle Proteine auf einmal abrufen, die mit einem Organismus, zum Beispiel *S. aureus* assoziiert sind, unter anderem SplB und GlpQ. Darüber hinaus lässt sich die Lokalisation insofern spezifizieren, sodass beispielsweise nur noch Proteine angezeigt werden, die sich in oder an der Zellwand von *S. aureus* befinden. Es können auch mehrere Filter kombiniert und je Filter mehrere Argumente eingetragen werden. Somit können auf einen Abruf alle in der UniProt eingetragenen Proteine anzeigt werden, die mit *S. aureus* assoziiert sind und in der Zellmembran, Zellwand, dem Zytoplasma oder sekretiert vorkommen.

Die UniProt bietet eine Programmierschnittstelle, um Daten mit einem Programmcode runterzuladen. Die Programmierschnittstelle (*Application Programming Interface*, API) stellt dabei den Code zur Verfügung, der die Kommunikation zwischen dem Anwender und den Anwendungen oder Datenbanken ermöglicht. Somit sind die Daten der UniProt nicht nur manuell über das Web-Interface verfügbar, sondern auch in eigens geschriebenen Programmen, die Anfragen mit den entsprechenden Parametern stellen.

1.3.2 Epitopvorhersage

Für eine Proteinsequenz lassen sich T-Zell-Epitope vorhersagen. Die Grundlage dessen bilden Datenbanken über Peptidsequenzen bekannter Epitope, welche an MHC binden und Strukturanalysedaten von MHC. Die einzelnen Epitopsequenzen wurden durch experimentelle Epitopkartierung entdeckt und manuell aus Publikationen in die Datenbanken eingetragen. Kontinuierliche Epitope können über Stimulation von mononukleären Zellen des peripheren Blutes (*peripheral blood mononuclear cells*, PBMC) mit synthetischen Antigen-Peptiden und anschließender Messung der IFN-γ-Sekretion mittels Durchflußzytometrie identifiziert werden. Diskontinuierliche wurden mittels Röntgen-Kristallstrukturanalyse oder Kernspinresonanzspektroskopie identifiziert (Schuler et al. 2007). Einerseits gibt es Algorithmen, welche die Vorhersage anhand von Bewertungen (*scores*) basierend auf einer Bewertungsmatrix (*score matrix*, SM) treffen. Andererseits werden Vorhersagen auch durch maschinelles Lernen, speziell künstliche neuronale Netzwerke (a*rtificial neural networks*, ANN) getroffen.

1.3.3 SYFPEITHI

SYFPEITHI ist eine Epitopdatenbank und einer der ältesten Epitop-Vorhersagealgorithmen. Der Algorithmus wurde von Prof. H.-G. Rammensee am Institut für Zellbiologie der Universität Tübingen

und O. A. Bachor von *EMS Medicinal and Scientific Dataprocessing* entwickelt und basiert auf einer Bewertungsmatrix. Er betrachtet jede einzelne AS und ihre Position in dem zu analysierenden Peptid und ordnet ihr einen Score zu, das heißt, die SM ist eine positionsspezifische SM (*position specific score matrix*, PSSM). Je höher der somit erreichte *Score* eines Peptids ist, desto wahrscheinlicher ist es, dass es das von dem gewählten MHC-Molekül präsentierte Peptid ist. Mit einer Zuverlässigkeit des Algorithmus von 80 % sollte das wahrscheinlichste Epitop in den obersten 2 % der vorhergesagten Epitope zu finden sein (Schuler et al. 2007; Rammensee et al. 1999).

Die PSSM, die SYFPEITHI zugrunde liegt, hat die 20 Aminosäuren als Zeilenindizes und jede Spalte stellt nach ihrem Index eine Position im Epitop dar (Abbildung 1.4). So hat jede AS einen positionsbezogenen Wert (Abbildung 1.5, Abbildung 1.6). Für häufig an Ankerpositionen vorkommende AS wird der Score 10 vergeben, 8 für AS, die an der betrachteten Position, außerhalb von Ankerstellen, häufig in den Epitopen der Datenbank vorkommen. Hochfrequente AS an Hilfsankerpositionen werden mit 6 und weniger häufige mit 4 bewertet. Positionsunabhängig werden die Werte 1-4 für die Häufigkeit in individuellen Sequenzen vergeben. AS, die an einer bestimmten Position bekanntermaßen nicht vorkommen, werden mit negativen Werten von -1 bis -3 bewertet. Der Score von SYFPETHI kann für das Allel HLA-A*02:01 maximal 36 erreichen.

	Position 1 **2** 3 4 5 6 7 8 **9**
Anchor residues	**H** **L**
Preferred residues	I E P A V V V F Y A D G I R R M T S E P M P E G G N K T E K R A Q V

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

Abbildung 1.4 **Bewertung von AS nach Positionen im Epitop**: Links: Einordnung der verschiedenen Aminosäuren an spezifischen Positionen im Epitop. Rechts: Bewertungsmatrix für Aminosäuren für HLA-B*15:10 in Nonameren. Von SYFPEITHI zum Stand 1999 (Rammensee et al. 1999).

Position 1 2 3 4 5 6 7 8 9	Source of peptides	Reference	Remark
anchors or auxiliary anchors			
H L		Seeger et al. 1999, Prilliman et al. 1999	
preferred residues			
I E P A V V V F			
Y A D G I R R			
T S E P M P E			
G Y G N K M T			
E I K R R L A			
K Q V			
F V			
L			
V			
T			
G			
D			
N			
M			

Abbildung 1.5: **Einordnung verschiedener Aminosäuren an den Positionen im Epitop von SYFPEITHI**: Zum Stand 04.03.2021 für HLA-B*15:10 und Nonamere. Anker (fett) werden mit 10 bewertet, Hilfspositionen mit 6 und die weiteren individuell nach Häufigkeit (syfpeithi.de).

Position 1 2 3 4 5 6 7 8 9	Source of peptides	Reference	Remark
anchors or auxiliary anchors			
L V V		Falk et al. 1991	
M L			
preferred residues			
E K			
K			
other residues			
I A G I I A E			
L Y P K L Y S			
F F D Y T H			
K P T N			
M M G			
Y S F			
V R V			
H			

Abbildung 1.6: **Einordnung verschiedener Aminosäuren an den Positionen im Epitop von SYFPEITHI**: Zum Stand 04.03.2021 für HLA-A*02:01 und Nonamere. Anker (fett) werden mit 10 bewertet, Hilfspositionen mit 6 und die weiteren individuell nach Häufigkeit (syfpeithi.de).

Der Algorithmus summiert die Werte jeder Aminosäure eines Epitops und berechnet so die Bindungswahrscheinlichkeit. Die übergebene Sequenz wird in alle möglichen zusammenhängenden Peptidfragmente festgelegter Größe geteilt (hier Nonamere). Jedem potenziellen Epitop wird dann seine Bindungswahrscheinlichkeit zugeordnet und zusammen mit der Position der ersten AS in der Gesamtsequenz ausgegeben.

1.4 Künstliche Neuronale Netzwerke

Ein künstliches neuronales Netzwerk simuliert das zentrale Nervensystem und kann lernen. Verschieden definierte Neurone sind in Schichten organisiert und untereinander zur Kommunikation weit verzweigt. Solch eine Verzweigung heißt Gewicht und beeinflusst die Daten, die über die

Verbindung weitergegeben werden. Ein Neuron besteht dabei aus einer Aktivierungsfunktion, häufig Sigmoid- oder Tangens-hyperbolicus-Funktion, die eingehende Werte verarbeitet und den zukünftigen Zustand des Neurons bestimmt. Und einer Ausgabefunktion, durch die das Neuron in einem neu bestimmten Zustand aktiviert wird, wenn das Ergebnis einen Schwellenwert überschreitet. Die erste Schicht, die Eingabeebene, empfängt externe Variablen und gibt sie an die nachfolgenden versteckten Neuronen weiter. Diese führen die oben genannten Berechnungen mit jeweils eigens definierten Funktionen durch. Am Ende werden die Daten der versteckten Schicht an die Ausgabeebene gegeben, die die Daten in der gewünschten Form ausgeben (Abbildung 1.7).

Das ANN muss für Informationsgehalt und präzise Aussagen trainiert werden. Zu Beginn sind den Gewichten zufällige Werte zugeordnet. Das Training in diesem Fall umfasst die Bearbeitung von Fällen mit bekanntem Ergebnis (*supervised learning*) und anschließende Korrektur des berechneten Ergebnisses. Das Korrektursignal wird dabei rückwärts durch das Netzwerk gegeben und die Gewichte auf dem Weg angepasst. Somit werden sukzessive die Gewichte von der Ausgabeschicht bis zu den Eingabeneuronen durch das sog. *Backpropagation*-Lernverfahren angepasst. Das Prinzip des rückgekoppelten Feedback-Netzwerks ist natürlichen, lernenden Systemen nachempfunden (Traeger et al. 2003). Weitere Möglichkeiten ein ANN zu trainieren sind *unsupervised learning*, wobei das Netz nur aufgrund der Eingabedaten trainiert wird, *reinforced learning*, bei welchem das ANN eigenständig Entscheidungen trifft und nach Bewertung derselben, positive oder negative Belohnungen erfährt und sich somit anpasst und stochastisches Lernen, das auf der zufälligen Suche von Parametern für eine vermutete Ausgabe basiert.

1.4.1 NetMHCpan

NetMHCpan ist ein Algorithmus, der ein ANN für die Vorhersage von Epitopen nutzt. Er wurde von Prof. Morton Nielsen am *Department of Bio and Health Informatics* der Technischen Universität Dänemark entwickelt. NetMHCpan lernt, wie SYFPEITHI, auf Grundlage von Daten über die Bindungsaffinität (BA). Darüber hinaus werden auch Cluster-Analysen von Massenspektrometrie-Daten (MS) gelöster Liganden (*eluted ligands*, EL) aus dem Immunopeptidom mit MHC-Molekülen genutzt. Dadurch können auch Informationen aus dem Präsentationsprozess, die über die bloße Bindung hinaus gehen, für die Vorhersage genutzt werden. Die Kombination wird durch das *Framework* (Gerüst) *NNAlign* ermöglicht. Die Bewertung eines Epitops von NetMHCpan entspricht dem prozentualen Rang des Scores verglichen mit einer Menge zufälliger natürlicher Peptide und reicht daher von nahezu 100 für sehr schlecht und nahezu 0 für sehr gut bewertete Epitope.

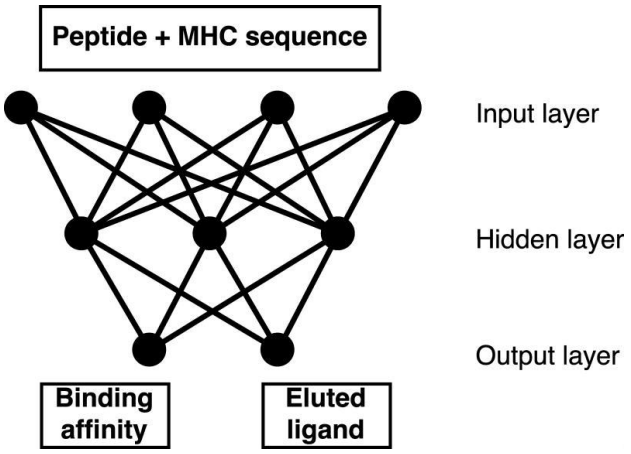

Abbildung 1.7 **Struktur des neuronalen Netzwerks hinter NetMHCpan**: Die Eingabe bilden die Sequenzen des Peptids und des MHC-Moleküls. Eine versteckte Schicht führt die Berechnungen und Bewertungen aus. Die Ausgabe umfasst Vorhersagedaten über BA- oder EL-Daten oder beide in Kombination (Jurtz et al. 2017).

NNAlign führt während des Trainings das Clustering aus. Dadurch können EL-Daten zu verwandten MHCs, die bis dahin nicht eindeutig zu unterscheiden waren, nun einzelnen MHC-Spezifitäten zugeordnet werden. Somit können die BA- und EL-Daten daraufhin kombiniert betrachtet zu einer besseren Vorhersage führen, da die Größe des Trainingsdatensatzes erhöht wird. Zudem wird die Möglichkeit geschaffen, bisher unbekannte und somit in den BA-Daten nicht-existente Bindungsmotive zu entdecken (Jurtz et al. 2017). Die Informationen zu den MHC-Allelen wurden als eine Pseudo-Sequenz der Länge 34 AS nach dem Muster in Abbildung 1.8 gespeichert und in Abbildung 1.9 visualisiert. Die Erweiterung von *NNAlign* um den Zusatz *_MA* von der Version NetMHCpan_4.0 an bedeutet dabei, dass das Framework nicht nur MS-Daten über Peptide, die zu jeweils einem MHC-Allel annotiert sind, verarbeiten kann, sondern auch über sogenanntes Pseudo-Labelling auch multiallelische (MA) Ligandendaten (Reynisson et al. 2020).

Das ANN hat 43 Eingabe-, und einen Ausgabeknoten, je nach Ausgabemethode BA oder EL (Abbildung 1.7). Die Eingabe bilden die neun Peptid-AS und 34 MHC-Reste, wobei die Sequenzen als Folge von 19 Nullen und einer eins, nach BLOSUM50 verschlüsselt oder gemischt, also das Peptid nach der ersten Variante und die MHC-Pseudo-Sequenz nach BLOSUM50 verschlüsselt ist. Das ANN wurde mittels fünffacher Kreuzvalidierung trainiert und die Gewichte wurden mittels *Backpropagation* angepasst. Der Datensatz wurde dazu in fünf Teilmengen geteilt und es wurde fünfmal mit jeweils einer anderen Teilmenge als Referenzdatensatz trainiert und validiert.

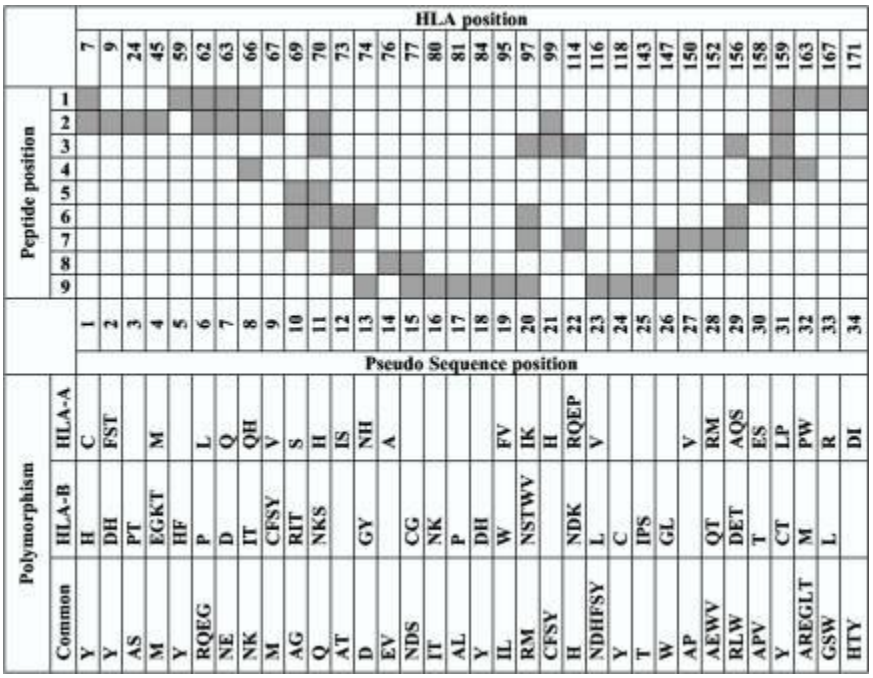

Abbildung 1.8: **Einordnung verschiedener Aminosäuren an den Positionen im Epitop von NetMHCpan:** Nach Stand 2007 wurden oben die Positionen der HLA-Reste (nach ImMunoGeneTics) gegenüber den Positionen im Peptid als grau markiert, wenn an diesen Positionen Peptid-HLA-Kontakt vermutet wird. Unten ist der Polymorphismus in der Pseudo-Sequenz nach Positionen anhand der jeweils verbreiteten AS dargestellt (Nielsen et al. 2007).

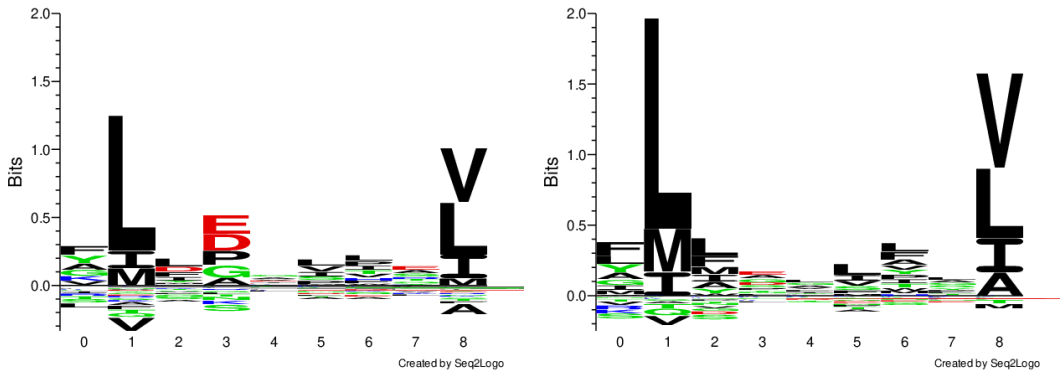

Abbildung 1.9: **AS-Bindungsmotive für HLA-A*02:01 von NetMHCpan**: Zum Stand 19.03.2021 sind links die natürlich präsentierten Liganden und rechts die MHC-Bindungsdaten zu jeder Position im Epitop bewertet (*Motif Viewer*, http://www.cbs.dtu.dk/services/NetMHCpan/logos_ps.php).

Die Webserver der Vorhersageprogramme lassen sich aus eigens geschriebenem Quelltext erreichen. Für NetMHCpan bietet die *Immune Epitope Database* (IEDB), auf deren Datensätzen das ANN trainiert wurde, den Quellcode-Zugang über ein API an. SYFPEITHI hingegen bietet keine solche Schnittstelle zu den Online-Services an. Hier kann die Eingabe mittels des URL gemacht und die Vorhersagedaten im Webformat HTML heruntergeladen werden.

1.4.2 Vergleich der Algorithmen

Die Prädiktionsalgorithmen lassen sich über die Falsch-Positiv- und Richtig-Positiv-Rate (*false positive rate*, FPR; *true positive rate*, TPR) miteinander vergleichen. Die Isosensitivitätskurve (*receiver operating characteristic*, ROC) trägt die falsch positiven (FPR) auf der x-Achse gegen die richtig positiven vorhergesagten Epitope (TPR) auf der y-Achse auf. Je schneller und steiler die Funktion ansteigt, desto größer ist die Fläche unter der Kurve (*area under the curve*, AUC) und umso besser sind die Vorhersagen des Algorithmus, wobei der Wert 1 für eine perfekte Vorhersage steht.

Anhand eines Maus-Modells mit dem Vaccinia-Virus wurden alle aktuell öffentlich verfügbaren Vorhersagetools von Paul et al. 2020 verglichen. NetMHCpan bietet in der Version 4.0 die beste Performance mit einer durchschnittlichen AUC über beide betrachteten Allele und alle Epitoplängen (7 – 13 AS) von 0.983. SYFPEITHI hat eine durchschnittliche AUC von 0.961, wobei nur Vorhersagen für Okta- und Nonamere gemacht wurden. Mit diesem Wert allein ist es die beste Performance der linearen Algorithmen, allerdings bieten andere mehr Epitoplängen für die Vorhersage. Es zeigt sich, dass Methoden basierend auf ANNs im Durchschnitt besser performen, aber aufgrund der komplexeren Struktur mehr Rechenzeit benötigen. Für eine Sequenz der Länge 1000 AS brauchten Matrix-basierte Algorithmen im Mittel 2.07 Sekunden (SYFPEITHI: 0.99 s) und ANN-basierte 6.06 Sekunden (NetMHCpan-*4.0*: 8.53 s) (Paul et al. 2020).

2 Verwandte Arbeiten

Vor dem Hintergrund der geringen Datenlage über *S. aureus*-spezifische CTLs wurden bereits zwei Ansätze für die Untersuchung dergleichen erarbeitet. Daniel Mrochen vom Institut für Immunologie der Universität Greifwald hat PBMCs *in-vitro* mit *S. aureus*-Antigenen stimuliert. Er nutzte die rekombinanten Proteine *serin protease SplB* und *glycerophosphodiester phosphodiesterase* (*GlpQ*). Daraufhin konnte er humane, murine und bovine CTLs nachweisen, die reaktiv für *S. aureus* waren.

Der zweite Ansatz bediente sich der mathematischen Prädiktionsalgorithmen. Dr. Clemens Cammann von Greifswalder Friedrich Loeffler-Institut für Medizinische Mikrobiologie nutzte SYFPEITHI und NetMHCpan, um nach T-Zell-Epitopen in *S. aureus*-Proteinen zu suchen. Er nutzte die *serin proteasen SplA – SplF* (also unter anderem *SplB*, wie Daniel Mrochen) und führte die Vorhersagen für HLA-A*02:01 und eine Epitoplänge von 9 AS aus. Nach Abgleichen der Vorhersagen wurde einerseits deutlich, dass T-Zell-Epitope vorhergesagt werden. Allerdings nur wenige, verglichen mit bisherigen Erfahrungswerten viraler Epitopvorhersagen. Mit mehreren überlappenden Epitopen jeweils hoher Bindungswahrscheinlichkeit in der Region von Position 5 bis 23 der AS-Sequenz, fiel andererseits das *Leader*- oder Signalpeptid besonders auf.

Die Stimulation von PBMCs mit jeweils einem Antigen, wie von Daniel Mrochen durchgeführt oder auch die manuelle Eingabe der fraglichen Peptidsequenzen in Vorhersageprogramme sind im Einzelfall nützliche Methoden, aber in jedem Fall zeit- und arbeitsintensiv und auf wenige Proteine pro Durchgang beschränkt.

3 Zielsetzung

In dieser Arbeit soll ein Algorithmus für die Zusammenstellung von Proteomdaten und deren vorhergesagter Epitope erstellt werden. Die den Proteomen zugrundeliegende Datenbank wird UniProt sein. Über deren Parameter werden die Proteome in diesem Algorithmus spezifiziert. Die Vorhersagen der Epitope werden anhand der AS-Sequenzen von den online frei zugänglichen Prädiktionsalgorithmen SYFPEITHI und NetMHCpan durchgeführt. Diese werden hier basierend auf dem MHC-Allel HLA-A*02:01 und einer Epitoplänge von 9 AS durchgeführt.

Der Algorithmus wird dann für verschiedene Subproteome, unter anderem von *S. aureus* angewendet. Im ersten Schritt trägt der Algorithmus Informationen zum gesamten notierten Proteom von *S. aureus* zusammen. Daraufhin werden auch Daten zu den *S. aureus*-Subproteomen der Lokalisationen Extrazellularraum, Zellmembran, Zellwand und Zytoplasma erstellt. Die Lokalisationen beziehen sich dabei auf den Funktionsort der Proteine. Im letzten Schritt werden die Datensätze für die *S. aureus*-Stämme *USA300* und *Mu50 / ATCC 700699* (*Mu50*), die *Influenza*-Stämme *Puerto Rico/8/1934, New Zealand:South Canterbury/35/2000*, *Russia:St.Petersburg/8/2006* und *USA:Texas/12/2007* und das Corona-Virus *SARS-CoV-2* erstellt.

Die gewonnen Daten werden auf Gemeinsamkeiten, Unterschiede und Korrelationen untersucht. Die Untersuchungen basieren zum Großteil auf der Epitopdichte, die Aufschluss über die Immunogenität geben soll. Die Epitopdichte wird einerseits bezüglich der Gesamtsequenz und andererseits, aufgrund der Aussagen Dr. Clemens Cammans, bezüglich des Signalpeptids der Proteine betrachtet. Darauf aufbauend sollen zwei Thesen untersucht werden: (1.) Die Signalpeptide stellen aufgrund des gehäuften Vorkommens von T-Zell-Epitopen Sequenzen konservierten Charakters dar und (2.) Die T-Zell-Epitopdichte von *S. aureus* ist vermutlich insgesamt gering und lässt sich als Immunevasionsmechanismus interpretieren.

Der Algorithmus soll über die Arbeit hinaus zu nutzen sein. Insofern sollen die Parameter und Spezifikationen der Datensätze und Vorhersagen nicht im Quellcode festgeschrieben sein. Es soll die individuelle Eingabe ermöglicht werden, sodass die gesamte Reichweite der Datenbank UniProt genutzt werden kann. Darüber hinaus sollen für die Vorhersagen jedes verfügbare MHC-Allel zu den verschiedenen Epitoplängen betrachtet werden können. Damit soll der Algorithmus universell je nach Fragestellung einsetzbar sein.

4 Methoden

Für die Implementierung wurde die Programmiersprache Python und die Entwicklungsumgebung PyCharm verwendet und für die Auswertung der Daten R mit der Umgebung RStudio. Als API wurde *requests* genutzt und für die Datenaufbereitung *numpy, collections*, *operator* und *re*. Die Ausgabe der Daten in ein Tabellendokument wurde mit *xlsxwriter* realisiert und für das Kommandozeileninterface *argparse* genutzt. Das Paket *time* ermöglichte die Zeitmessung der Iterierungen und damit die Ausgabe einer groben Abschätzung der voraussichtlichen Laufzeit. Mittels des Pakets *sys* konnte eine sich aktualisierende prozentuale Fortschrittsanzeige implementiert werden. In R wurden die Pakete *ggplot2*, *grid*, *gridExtra*, *gtable*, *plyr, purrr*, *readxl*, *rlist* und *stringr* genutzt.

4.1 Abrufen der Proteindaten

4.1.1 Theorie

Die Anfrage an UniProt für den Abruf der Proteindaten wurde durch mehrere Parameter spezifiziert. Dabei waren einzelne Sequenzen mehrfach unter verschiedenen Identifikationsnummern (IDs) eingetragen, wenn sie in mehreren Stämmen der Spezies vorkommen. Daher wurde die Möglichkeit geschaffen, je nach Zielsetzung die doppelten Sequenzen zu filtern, um keine Verfälschung der Ergebnisse zu bekommen.

Es wurden mehrere Datensätze zu *S. aureus* und vier Lokalisationen erstellt. Ein Datensatz enthielt alle in UniProt notierten und überprüften *S. aureus*-spezifischen Proteine, die paarweise verschieden (pwv.) waren, das Proteom. Anhand dessen ließen sich Merkmale, Unterschiede und Korrelationen innerhalb des gesamten Organismus *S. aureus* untersuchen. Ein weiterer enthielt alle *S. aureus*-spezifischen Proteinsequenzen, die assoziiert mit der Zellmembran oder Zellwand, zytoplasmatisch oder sezerniert, also extrazellulär lokalisiert waren und deren Einträge geprüft waren. In diesem sind die Daten jeweils innerhalb einer Lokalisation um redundante Sequenzen bereinigt, um aussagekräftige Untersuchungen zu den Sequenzen, Korrelationen und Varianzunterschieden zwischen den Lokalisationen durchführen zu können.

Weitere Datensätze wurden zu verschiedenen Spezies erstellt. Um die Daten zu *S. aureus* einschätzen zu können, wurden Epitopdaten einerseits zu den *Influenza A*-Stämmen des Subtyps *H1N1 Puerto Rico/8/1934, New Zealand:South Canterbury/35/2000, Russia:St.Petersburg/8/2006* und *USA:Texas/UR06-0195/2007* und andererseits zu *SARS-CoV-2* erstellt. Von *S. aureus* werden für den Vergleich die Stämme *USA300* und *Mu50* genutzt. Der Parameter *locations* wurde nicht gesetzt.

Die Ausgabe wurde auf für die Analysen wichtige Daten beschränkt. Die geforderten Datenfelder im Datensatz waren die ID, Protein- und Genname, Sequenzlänge und -masse, Abstammung (nur der letzte Eintrag des Taxonomiebaums), Länge des Signalpeptids und AS-Sequenz. Um die weitere

Bearbeitung der Daten zu ermöglichen war die Rückgabe Tabulator-separiert gegeben. Von Interesse war im Datensatz über die vier Kompartimente, ob Sequenzen mehrfach, also an mehreren Lokalisationen, vorkommen. Dazu wurden die Häufigkeiten der Sequenzen im gesamten Datensatz und die Parameter der Abfrage als Deckblatt zusammen mit dem vollständigen Datensatz in einem Tabellendokument ausgegeben.

4.1.2 Implementierung

Die Anfrage anhand eines Suchauftrags fand in der Funktion *exec_uniProt* statt. Der erste Suchauftrag (*query*) setzte sich aus den Parametern *organism* (= *staphylococcus aureus*), und *reviewed* (= *yes*) zusammen und die Frage nach Reduzierung um die redundanten Sequenzen auf der Kommandozeile wurde bestätigt (*Delete redundant sequences* = *yes*). Die zweite Anfrage wurde um *location*-Parameter (= *cell_membrane*, *cell_wall, cytoplasm, secreted*) erweitert, wobei je Parameter eine *query* erstellt und die Abfrage ausgeführt wurde.

Die Datenfelder wurden durch den Parameter *columns* spezifiziert. Dieser erhielt die Argumente *id*, *protein_name*, *gene_name*, *length*, *mass*, *lineage(ALL)*, *feature(SIGNAL)* und *sequence* für die oben in gleicher Reihenfolge genannten Kategorien. Über *format* (=*tab*) war das Ausgabeformat festgelegt. UniProt bietet eine genaue Anleitung, wie die *query* und *columns* zu übergeben sind (https://www.uniprot.org/help/api_queries, (The UniProt Consortium 2020c). Die Parameter wurden als *Dictionary* (Dict) unter den *Keys* (Schlüssel) *query*, *format* und *columns* über *GET* an UniProt gegeben. Pro *query* wurden die zurückgegebenen Daten als Liste unter einem Schlüssel, welcher der Lokalisation entsprach, im Datensatz gespeichert. Wurde eine schon gespeicherte Sequenz wieder registriert, wurden die IDs lokalisationsunabhängig zusammen gespeichert, damit Vorhersagen nicht doppelt ausgeführt werden, sondern auf die Daten der ersten Sequenz zurückgegriffen werden kann. In diesen Anfragen wurden die zur doppelten Sequenz gehörigen Daten gelöscht.

Die Ausgabe erfolgte als Tabellendokument im Format xlsx. Die Häufigkeiten der Sequenzen wurden anhand der Funktionen *Counter*, *operator* und *sort* berechnet und absteigend sortiert. Schließlich wurde mit der Funktion *print_sheet* eine Arbeitsmappe erstellt und die Häufigkeiten mit den *query*-Parametern und einem Vermerk auf die Methode UniProt auf der ersten Seite eingetragen. Auf den folgenden Seiten wurden je Lokalisation die Proteindaten spaltenweise nach den Datenfeldern in *columns* formatiert und der Link zum entsprechenden UniProt-Eintrag eingetragen. Am Ende wurden die Sequenzen und der Speicher mit den IDs äquivalenter Sequenzen an die Vorhersagefunktionen übergeben.

4.2 Vorhersagen von Epitopen

4.2.1 Theorie

Für jede Sequenz in dem Datensatz wurde eine Anfrage an NetMHCpan und SYFPEITHI gestellt. Die Vorhersagen wurden für das Allel HLA-A*02:01 und Nonamere gemacht. Die Rückgabe umfasste zu jedem Epitop Daten über die Position, die genaue Sequenz und den *Score*. Zudem wurde ein Grenzwert für den *Score* eingeführt, da die Prädiktionsalgorithmen alle zusammenhängenden Mere der gewählten Länge als Epitop betrachten und bewerten. Hier von Interesse waren allerdings nur solche dessen Bindung zu dem gewählten MHC-Allel wahrscheinlich ist.

Da NetMHCpan nach Paul et al. deutlich mehr Laufzeit benötigen würde, wurden die Vorhersagen von SYFPEITHI zuerst ausgeführt. Dies ermöglichte die Arbeit am SYFPEITHI-Datensatz, während die Vorhersagen durch NetMHCpan noch liefen und sparte somit Zeit ein.

Zur Bewertung der Vorhersagen wurde die Epitopdichte genutzt. Die Dichte der Epitope wurde über die gesamte Sequenz und innerhalb des Signalpeptids nach Gaseitsiwe et al. wie folgt berechnet:

$$D(s) = \frac{E_{bindend}(s)}{E_{gesamt}(s)}$$

Formel 1: **Epitopdichte**

Die Dichte D berechnete sich in Abhängigkeit der Sequenz s als Quotient der Anzahl der besser als dem Grenzwert bewerteten Epitope $E_{bindend}$ in s und der Anzahl aller möglichen Epitope E_{gesamt} innerhalb s (Gaseitsiwe et al. 2010).

$$E_{gesamt}(s) = L_{Sequenz}(s) - L_{Epitop} + 1$$

Formel 2: **Anzahl aller möglichen zusammenhängenden Epitope**

Die Gesamtzahl der Epitope E_{gesamt} in einer Sequenz s, also die Anzahl aller zusammenhängender Mere der Länge L_{Epitop} berechnete sich als Differenz aus der Sequenzlänge $L_{Sequenz}$ und der um eine Stelle reduzierten Epitoplänge L_{Epitop}. Im Falle der Signal-Epitopdichte entsprach $L_{Sequenz}$ der Länge des Signalpeptids und $E_{bindend}$ der Anzahl von Epitopen, deren Beginn innerhalb des Signalpeptids lag.

Die Ausgabe entsprach dem Format der Proteindaten. Es wurde eine Seite mit den Parametern der Vorhersage und Häufigkeiten der Epitope und darauffolgend Seiten über die Vorhersagedaten nach Lokalisationen auf den folgenden Seiten produziert. Die Datenfelder waren hier ID, Proteinname, Epitopdichte über die Gesamtsequenz und das Signalpeptid, der Link zum UniProt-Eintrag des Proteins der Vorhersage und schließlich die Epitope nach *Score* geordnet. Je ein Dokument wurde für einen der Prädiktionsalgorithmen erstellt.

4.2.2 Implementierung

Die Funktionen *exec_netmhcpan* und *exec_syfpeithi* bekamen mehrere Datensätze von UniProt für die Vorhersagen übergeben. Einerseits die Sequenzen für die Vorhersage und Längen der Signalpeptide für die Berechnung der Signal-Epitopdichte. Andererseits die IDs zu äquivalenten Sequenzen, sodass zu den redundanten keine neue Vorhersage getätigt werden musste sondern auf die bestehenden Daten zurückgegriffen und so Laufzeit eingespart werden konnte. Über die Sequenzen iterierend wurde jeweils über das RESTful-Interface der IEDB (NetMHCpan) beziehungsweise den URL (SYFPEITHI) eine POST-Anfrage mit den Parametern *allele* (=*HLA-A*02:01*), *length* (=*9*) und *sequence_text* (=SEQUENZ) an die Webserver gestellt. Die IEDB forderte zusätzlich den Parameter *method* (=NetMHCpan), da noch weitere Prädiktionsalgorithmen angeboten werden. Die Web-Version des SYFPEITHI-Algorithmus hatte ein Eingabelimit für die Sequenzlänge. Der Eingabemaske der SYFPEITHI-Website nach sollte dieses Limit bei 2048 AS liegen, allerdings konnten für diese Länge auch keine Vorhersagen ausgeführt werden (syfpeithi.de 2012). Daher wurde nach mehreren Versuchen die Länge 2000 AS als funktionell ausgemacht. Sequenzen, die länger als 2000 AS waren, wurden an dieser Stelle abgeschnitten und nur der vordere Teil, der auch das Signalpeptid enthält, betrachtet. Die Grenzwerte für die Übernahme betrugen 3.0 (NetMHCpan) und 20 (SYFPEITHI), Epitope mit einem schlechteren Score wurden nicht übernommen.

Die Rückgaben mussten zur Weiterverarbeitung aufbereitet werden. Die Rückgabe von SYFPEITHI wurde in der Funktion *request_handle* von HTML-Auszeichnungen bereinigt und die geforderten Daten über Position, Peptid und Score extrahiert. Die von NetMHCpan liegt im tab-separierten Format vor, sodass die Daten über Indexierung einfach zugänglich waren. Falls aufgrund eines Fehlers in der Datenübertragung die Rückgabe fehlerhaft war, wurde ein Vermerk mit der Lokalisation und ID auf der Konsole ausgegeben und der entsprechende Eintrag im Datensatz mit *Bad Answer* (Ungültige Antwort) markiert. Die aufbereiteten Daten wurden abermals unter einem für die Lokalisation spezifischen Schlüssel im Datensatz gespeichert und um die Proteinnamen und Epitopdichten erweitert. Wenn die entsprechenden Daten aufgrund von fehlerhafter Rückgabe oder nicht vorhandenen Signalpeptids nicht zur Verfügung standen, wurde entsprechend *N/A* (*not available*, nicht verfügbar) gespeichert.

In der Ausgabe der Vorhersagen wurden mehrere Markierungen vorgenommen. Die Häufigkeiten werden wie die der Proteinsequenzen berechnet. Das Format und die Reihenfolge der Daten entsprachen der Ausgabe der Proteindaten und die Häufigkeiten und Parameter bildeten wieder die erste Seite. Nachfolgend wurden die Epitopdaten zu einem Protein je Zeile mit den oben genannten Datenfeldern in eine Tabelle je Lokalisation geschrieben. Darüber hinaus wurde die ID eines Proteins grün unterlegt, wenn die Epitopdichte oberhalb des dritten Quantils, also unter den höchsten 25 % aller Dichten, in der jeweiligen Lokalisation lag. Eine rote Unterlegung der Namens- und Dichtefelder

und einen Vermerk über die Kürzung bekamen die an Stelle 2000 gekappten Sequenzen. In grüner Schrift wurden Epitope hervorgehoben, die innerhalb der Signalpeptid lagen.

4.3 Auswertung

Die verschiedenen Datensätze wurden auf Ausreißer, Korrelationen und Varianzunterschiede untersucht. Ausreißer waren standardgemäß definiert als Werte, die um das 1,5-fache des Interquartilsabstandes ($1{,}5 * IQR$) unterhalb des ersten beziehungsweise oberhalb des dritten Quartils lagen. Den Bereich zwischen dem ersten und dritten Quartil markierte in den Boxplots (Abbildung 4.1) das Rechteck, den Bereich $1{,}5 * IQR$ markieren die Antennen und Ausreißer sind als die Punkte außerhalb der Antennen gekennzeichnet. Der Boxplot zeigt darüber hinaus den Median als zu den Antennen senkrechte Linie innerhalb der Box an. Der Median ist gegenüber dem Mittelwert unempfindlich für Ausreißer in den Daten.

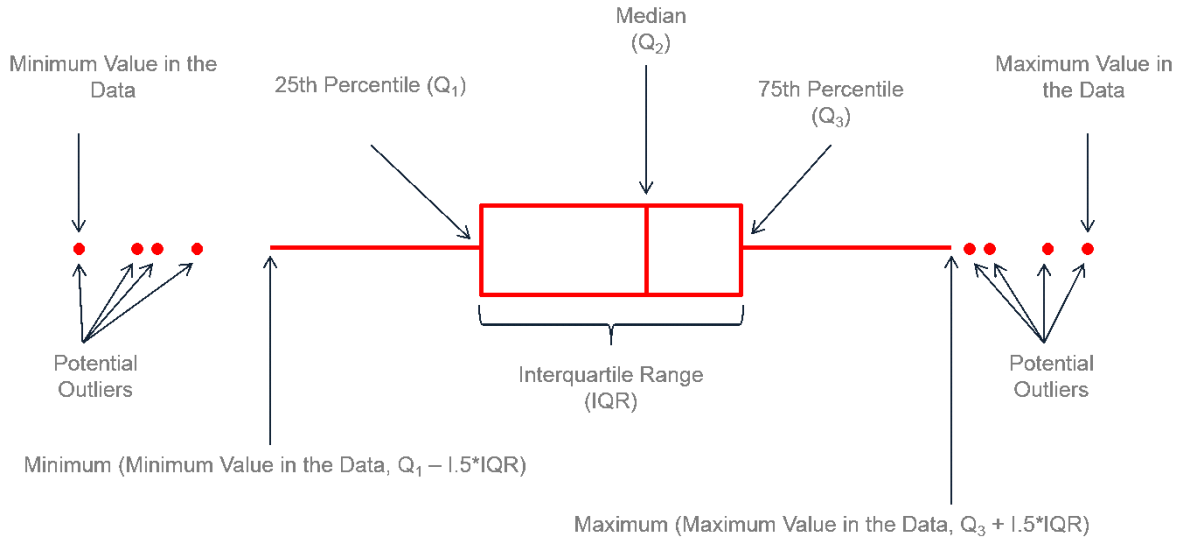

Abbildung 4.1 **Anatomie eines Boxplots**: Die Box beschreibt den Bereich zwischen dem ersten und dritten Quantil, also die mittleren 50 % aller Werte. Dieser Bereich heißt Interquartilsabstand (IQR). In der Box ist der Median, der Zentralwert eingezeichnet. Die Antennen markieren den Bereich des 1,5-fachen des IQR. Werte außerhalb dessen werden standardgemäß als Ausreißer bezeichnet (Coleman 2015).

Die Zusammenhänge in den numerischen Sequenzdaten wurden über die verteilungsunabhängige Rangkorrelationsanalyse nach Spearman (in R: `cor.test(…)` mit dem Parameter `method="spearman"`) wie nach Formel 3 berechnet.

$$r_s = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X}\, \sigma_{\mathrm{rg}_Y}}$$

Formel 3 **Rangkorrelationskoeffizient nach Spearman**: $cov(rg_X, rg_Y)$ ist die Kovarianz und σ_{rg_X} beziehungsweise σ_{rg_Y} sind die Standardabweichungen der gerankten Zufallsvariablen.

Unterschiede zwischen Gruppen, wie einerseits den Lokalisationen von *S. aureus* und andererseits den Stämmen in den Vergleichsdaten, wurden mittels einfaktorieller Varianzanalyse (ANOVA, in R: `aov()`) untersucht. Studien zeigten, dass die ANOVA nicht nur robust gegen Verletzungen der Normalitätsvoraussetzung in den Daten ist, sondern die Art der Verteilung keinen signifikanten Einfluss auf die Ergebnisse der ANOVA hat (Schmider et al. 2010; Blanca et al. 2017). Die paarweisen Unterschiede wurden durch den Post-hoc-Test nach Tukey berechnet (Formel 4.1, in R: `TukeyHSD(aov(…))`) und mittels der Funktion `tuk_plot(TukeyHSD(…), …)` visualisiert. Letztere ist Nathan Days Funktion im Forum *StackOverflow.com* (https://stackoverflow.com/questions/60794019/how-to-edit-a-tukey-test-plot-in-r) mit Erlaubnis nachempfunden und für die Markierung der Unterschiede mit signifikantem P-Wert angepasst. Die Funktion in R nutzt für ungleiche Gruppengrößen automatisch die Anpassung des Tests nach Tukey und Kramer (Formel 5). Auch diese ist robust gegenüber Verletzungen der Normalität in dem Datensatz (Driscoll 1996). Die Einschätzung der Korrelations- und Varianzanalysenwerte erfolgt nach (Bortz und Döring 2006, S. 606).

$$q = \frac{\mu_A - \mu_B}{SE}$$

$$SE = \frac{\sigma}{\sqrt{n}}$$

Formel 4 **Tukey's Test**: Oben: μ sind die Mittelwerte der zu vergleichenden Datensätze, wobei μ_A der größere von beiden ist und SE der Standardfehler der Summe der Mittelwerte (4.1). Unten: Standardfehler für Zufallsvariablen gleicher Größe n und Varianz σ (4.2).

$$SE = \sqrt{\frac{\sigma^2}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Formel 5 **Anpassung des Standardfehlers nach Kramer**: Der Standardfehler (*Standard error*, SE) Mit n_i und n_j als Größe der Zufallsvariablen und Varianz σ.

5 Ergebnisse

5.1 *Staphylococcus aureus*

5.1.1 Quantitativ

Insgesamt konnten 3222 pwv., im Mittel 409 AS lange Sequenzen aus dem Proteom *S. aureus* von UniProt runtergeladen werden. Unter den Subproteomen war das Zytoplasma das deutlich größte, gefolgt von der Zellmembran und dem Extrazellularraum. Die Zellwand bildete das Kompartiment der geringsten Anzahl an Sequenzen mit weniger als einem Zehntel des Zytoplasmas. Nur die zytoplasmatischen Proteine lagen mit der mittleren Länge unter der aller pwv. Sequenzen, beeinflussten den Mittelwert des Proteoms durch ihre große Anzahl aber stärker als die der anderen Subproteome. In Abbildung 10.1 (siehe Anhang) wird allerdings deutlich, dass die Mittelwerte der Sequenzlängen des Proteoms und der Zellmembran stark durch die teils gleichen Ausreißer beeinflusst wurden. Die für Ausreißer weniger empfindlichen Mediane lagen fast alle innerhalb eines Bereiches von 320 bis 341. Einzig die Zellwandigen wiesen trotz des hohen Mittelwerts keinen großen Einfluss durch Ausreißer auf, ihr Median lag mit fast dem Dreifachen des Proteoms deutlich erhöht.

Zu 384 Sequenzen war ein Signalpeptid angegeben, das im Mittel 32 AS lang waren. Das entsprach allen Zellwand-Proteinen, fast allen Extrazellulären und einem Teil der Zellmembran-Proteine. Die Zytoplasmatischen hatten aufgrund ihrer Funktion außerhalb von Zellorganellen kein Signalpeptid. Auffallend sind die deutlich längeren Proteine der Zellwand und die deutlich kürzeren der Zellmembran. Ein positiver Zusammenhang zwischen der Sequenz- und Signallänge scheint wahrscheinlich, einzig die Zellmembran-Proteine mit den deutlich kürzeren Signalpeptiden sprechen dagegen (Tabelle 5.1).

	Anz. Sequenzen	**Länge [AS]**	**Anz. Signalpeptide**	**Signallänge [AS]**
Proteom	3222	409	384	32
Zellwand	86	1008	86	46
Zytoplasma	903	378	0	N/A
Extrazellularraum	263	551	237	36
Zellmembran	672	500	113	23

Tabelle 5.1 **Zusammenfassung der Proteom-Daten zu *S. aureus***: Über den gesamten geprüften UniProt-Datensatz paarweise verschiedener *S. aureus*-Sequenzen (i), die zellwandigen (ii), zytoplasmatischen (iii), sezernierten (iv) und Zellmembran-Proteine (v). Die Anzahlen entsprechen den absoluten Beträgen im Datensatz notierter Sequenzen. Die Längen sind als Mittelwerte in der Einheit Aminosäuren (AS) berechnet. Für nicht vorhandene Signalpeptide ist *N/A* notiert.

	Anz. Epitope	Epitopdichte	Anz. Signalepitope	Signal-Epitopdichte
Proteom	16	0.045	4	0.17
Zellwand	15	0.018	2	0.061
Zytoplasma	13	0.037	0	N/A
Extrazellularraum	10	0.026	3	0.112
Zellmembran	31	0.083	4	0.287

A

	Anz. Epitope	Epitopdichte	Anz. Signalepitope	Signal-Epitopdichte	Gekappte Seq.
Proteom	16	0.046	4	0.172	18
Zellwand	16	0.02	3	0.094	7
Zytoplasma	14	0.041	0	N/A	0
Extrazellularraum	10	0.027	3	0.12	7
Zellmembran	27	0.075	4	0.286	9

B

Tabelle 5.2 **Zusammenfassung der Prädiktions-Daten zu *S. aureus***: Über die Datensätze aus Tabelle 5.1. A: NetMHCpan, B: SYFPEITHI. Die gekappten Sequenzen sind als absolute Werte angegeben, die restlichen als Mittelwerte. Das Limit für die Sequenzlänge in SYFPEITHI ist 2000, längere Sequenzen wurden an dieser Stelle gekappt. Die Daten solcher Lokalisationen über die gesamte Sequenz sind nur unter Vorbehalt des Einflusses gekappter Sequenzen zu betrachten. Die Signalpeptide sind davon nicht betroffen. Nicht zu berechnende Daten sind mit *N/A* notiert.

Zu den Sequenzen der Zellmembran wurden etwa bis zu zweimal mehr Epitope vorhergesagt. Auffallend war auch die besonders niedrige Anzahl der Epitope extrazellulärer Proteine. Beide Extremwerte äußerten sich jeweils auch in der höchsten beziehungsweise einer niedrigen Epitopdichte. Die Epitopanzahlen der Zellwand und des Zytoplasmas unterschieden sich nur wenig vom Proteom, wobei in diesem Vergleich die Epitopdichte der Zellwand als insgesamt niedrigste herausstach. Die Boxplots ergaben für NetMHCpan (A) ähnliche Verhältnisse für die Mediane und offenbarten, dass unter den hohen Dichten der Zellmembran-Proteine keine Ausreißer waren. Im Proteom hingegen waren entsprechende Sequenzen hoher Epitopdichte Ausreißer (siehe Anhang, Abbildung 10.2).

Ähnlich verhielt es sich mit den Epitopdaten innerhalb des Signalpeptids. Die Signal-Epitopdichte der Zellmembran ist sogar etwa bis zu viermal höher als die der Zellwand, die wieder die deutlich niedrigste aufweist. Auch die Mediane entsprechen den obigen Verhältnissen (siehe Anhang, Abbildung 10.3). Innerhalb der jeweiligen Proteome lagen die Epitopdichten des Signalpeptids ausnahmslos deutlich über denen der Gesamtsequenz.

Die Epitopdaten über die gesamte Sequenz in SYFPEITHI waren nur bedingt belastbar. Und zwar unter Berücksichtigung des Einflusses des Eingabelimits, da Sequenzen länger als 2000 AS auf diese Länge zugeschnitten worden sind und der hintere Teil in der Vorhersage unbeachtet geblieben ist. Die Unterschiede in Anzahl und Dichte der Epitope zwischen NetMHCpan und SYFPEITHI und die Verhältnisse von gekappten Sequenzen zu Sequenzen insgesamt waren niedrig. Einzig die Daten zu

den Zellmembran-Proteinen bedürfen näherer Untersuchung, da die Differenzen hier etwas höher sind. Im Zweifel sind hier nur die NetMHCpan-Daten vorerst zu verwerten (Tabelle 5.2)

5.1.2 Qualitativ

		Korrelation [-1,1]	P-Wert
Länge ~ Masse		0.9986	<0.05
Länge ~ Signallänge		0.4962	<0.05
Länge ~ Anz. Epitope	**N**	0.7346	<0.05
Länge ~ Anz. Epitope	**S**	0.7526	<0.05
Länge ~ Epitopdichte	**N**	-0.0039	0.83
Länge ~ Epitopdichte	**S**	-0.027	0.13
Länge ~ Signal-Epitopdichte	**N**	-0.4368	<0.05
Länge ~ Signal-Epitopdichte	**S**	-0.4989	<0.05
Signallänge ~ Signal-Epitopdichte	**N**	-0.7379	<0.05
Signallänge ~ Signal-Epitopdichte	**S**	-0.6883	<0.05
Anz. Epitope ~ Epitopdichte	**N**	0.6031	<0.05
Anz. Epitope ~ Epitopdichte	**S**	0.5786	<0.05
Anz. Epitope ~ Signal-Epitopdichte	**N**	-0.079	0.12
Anz. Epitope ~ Signal-Epitopdichte	**S**	-0.2104	<0.05
Epitopdichte ~ Signal-Epitopdichte	**N**	0.6301	<0.05
Epitopdichte ~ Signal-Epitopdichte	**S**	0.5104	<0.05

Tabelle 5.3 **Korrelationen zwischen Sequenz- und Vorhersagedaten aller pwv. *S. aureus*-Sequenzen**: Spearman-Korrelation als Rangkorrelationsanalyse für nicht normalverteilte Daten mit Korrelationswert (1. Spalte) und zugehörigem P-Wert (2. Spalte). Zeilen mit dem Vermerk N beziehen sich auf NetMHCpan-Daten, S auf SYFPEITHI-Daten. Korrelationswerte sind auf die vierte Nachkommastelle gerundet. Das Signifikanzniveau ist 5 %.

Die Spearman-Rangkorrelationstests ergaben erwartungsgemäße Werte. Die nur fast perfekte Korrelation der Sequenzlänge und -masse ist auf die unterschiedlichen Massen der Aminosäuren zurückzuführen, die allerdings alle mit Wert 1 in die Länge zählen. Mit der Sequenzlänge wuchsen auch die Zahl der Epitope mit starkem und, um die Vermutung zu Tabelle 5.1 zu bestätigen, das Signalpeptid mit hohem mittlerem Effekt. Im Gegensatz dazu korrelierte die Signalpeptidlänge mit der Signal-Epitopdichte stark negativ und somit auch die Sequenzlänge mit der Signal-Epitopdichte mit mittlerem Effekt, das heißt längere Sequenzen hatten geringere Signal-Epitopdichten (Tabelle 5.3).

Erwartungsgemäß wuchs auch die Epitopdichte mit der Anzahl der Epitope. Ebenso nicht überraschen korrelierte auch die Epitopdichte der Gesamtsequenz mit der des Signalpeptids stark, da die Gesamtsequenz hier das Signalpeptid enthält. Weitere Schlüsse über den Einfluss des Signalpeptids würden Betrachtungen der Restsequenz ohne Signal erlauben. Die leicht negative Korrelation der Anzahl der Epitope und Signal-Epitopdichte war nur für SYFPEITHI signifikant. Dieser Zusammenhang musste wieder unter dem Vorbehalt des Einflusses gekappter Sequenzen betrachtet werden, da die

Differenz zu dem Wert der NetMHCpan-Daten nicht gering war (Tabelle 5.3). Die Einschätzung der Korrelationswerte erfolgte nach (Bortz und Döring 2006).

		F-Wert	P-Wert
Länge		27	<0.05
Masse		25.2	<0.05
Signallänge		91.6	<0.05
Anz. Epitope	**N**	144.1	<0.05
Anz. Epitope	**S**	142.7	<0.05
Epitopdichte	**N**	344.2	<0.05
Epitopdichte	**S**	302.9	<0.05
Signal-Epitopdichte	**N**	160.5	<0.05
Signal-Epitopdichte	**S**	132.8	<0.05

Tabelle 5.4 **Varianzanalyse ANOVA der Sequenz- und Vorhersagedaten mit Bezug auf die Lokalisation**: Zeilen mit dem Vermerk N beziehen sich auf NetMHCpan-Daten, S auf SYFPEITHI-Daten. F-Wert als Indikator für die Größe der Unterschiede zwischen den Lokalisationen und P-Wert für die Signifikanz der ANOVA. Das Signifikanzniveau ist 5 %.

Nach Tabelle 5.4 gab es unter den Proteomdaten in allen betrachteten Werten signifikante Unterschiede zwischen Lokalisationen. Darunter waren weniger starke in der Länge und Masse der Gesamtsequenz, wobei sich paarweise nur die extrazellulären von den Proteinen der Zellmembran nicht signifikant unterscheiden ließen (siehe Anhang, Abbildung 10.4).

Sehr starke Unterschiede waren in der Epitopdichte, Signal-Epitopdichte und Anzahl der Epitope. Die Zellmembran-Daten unterscheiden sich paarweise am meisten und immer signifikant von den anderen. Im Hinblick auf die hohen Epitopanzahlen dieser Daten waren sie in NetMHCpan die einzig signifikant differenzierbaren (siehe Anhang, Abbildung 10.4, mittig-links (m.-l.)). Derselbe Vergleich für SYFPEITHI (siehe Anhang, Abbildung 10.4, u.-l.) stellte auch die Vergleiche mit extrazellulären Sequenzen signifikant dar, diese waren allerdings kleiner und die Daten aufgrund des Kappens potentiell weniger belastbar.

Bezüglich der Epitopdichte waren einzig die Extrazellulären und Zellwandigen nicht signifikant zu unterscheiden (Abbildung 10.4, m.-m. und m.-u.). Die Vergleiche zu den Zellmembran-Daten zeigten wieder große Differenzen. Ähnlich war es bei der Signal-Epitopdichte, wobei nur der Vergleich der extrazellulären Proteine zu denen der Zellwand in SYFPEITHI nicht signifikant war.

Insgesamt wurden den Zellwand-Proteinen deutlich mehr Epitope vorhergesagt. Das führte zu signifikant stark erhöhten Epitopdichten bezüglich der gesamten Sequenz und des Signalpeptids. Auch

die Zytoplasmatischen sind hier hervorzuheben als Proteine höherer Epitopdichte, allerdings fehlt hier die Signalsequenz, die potenziell von Bedeutung ist.

5.2 Vergleichsdaten

5.2.1 Quantitativ

	Anz. Sequenzen	Länge [AS]	Anz. Signalpeptide	Signallänge [AS]
S. aureus, USA300	677	334	53	30
S. aureus, Mu50	941	340	75	30
SARS-CoV-2	16	900	3	14
H1N1, Puerto Rico / 1934	13	408	1	17
H1N1, S. Canterbury / 2000	12	398	1	17
H1N1, St.Petersburg / 2006	12	400	1	17
H1N1, Texas / 2007	12	397	1	17

Tabelle 5.5 **Zusammenfassung der Proteom-Daten der Mikroben-Stämme**: Zu den *S. aureus*-Stämmen *USA300* (i) und *Mu50* (ii), *SARS-CoV-2* (iii) und den *Influenza A*-Stämmen *Puerto Rico/8/1934*, *South Canterbury/35/2000*, *St.Petersburg/8/2006*, *Texas/UR06-0195/2007* (iv-vii). Die Anzahlen entsprechen den absoluten Beträgen im Datensatz notierter Sequenzen. Die Längen sind als Mittelwerte in der Einheit Aminosäuren (AS) berechnet.

Die Datensätze zu den Proteomen der Virenstämmen waren deutlich kleiner als die der Proteome der Bakterienstämme (Tabelle 5.5). Die bakteriellen Stämme umfassten fast das bis zu 80-fache der Sequenzen der viralen Stämme. Allerdings waren die bakteriellen Sequenzen insgesamt im Mittel kürzer als die viralen. Die Sequenzlänge von *SARS-CoV-2* fiel als besonders lang auf. Sie ist auf die Polyproteine *Replikase-Polyprotein 1a* (4405 AS) und *Replikase-Polyprotein 1ab* (7096 AS) zurückzuführen, die, bevor sie funktionell sind, in mehrere Nichtstrukturproteinen gespalten werden. Hingegen war der Median der Sequenzlängen (siehe Anhang, Abbildung 10.5) für *SARS-CoV-2* um das Zwei- bis Vierfache niedriger als der Median der anderen Stämme.

Auch die Mittelwerte der *S. aureus*-Stämme waren durch Ausreißer nach oben beeinflusst. Insgesamt zeigte sich hier, dass fast alle Mediane (außer *Puerto Rico*) unter den Mittelwerten der Sequenzlängen liegen. Das heißt, die Sequenzen, die länger waren als der Mittelwert, waren geringer in der Anzahl als die kürzeren. Allerdings beeinflussten sie den Mittelwert stärker, da sie umso länger waren.

Im Gegensatz zur Gesamtsequenz waren die Signalpeptide von *SARS-CoV-2* im Mittel die kürzesten. Die der *Influenza A*-Stämme waren nur unwesentlich länger, aber die insgesamt kürzesten *S. aureus*-Sequenzen wiesen mit mehr als der doppelten Länge die größten Signalpeptide auf. Allerdings beziehen sich die Mittelwerte hier auf drei beziehungsweise ein einziges notiertes Signalpeptid. Das

Verhältnis von notierten Signalpeptiden zu Sequenzen insgesamt war beim neuen Coronavirus am größten, dann folgten die Bakterienstämme, die *Influenza*-Stämme hatten das niedrigste Verhältnis.

A

	Anz. Epitope	Epitopdichte	Anz. Signalepitope	Signal-Epitopdichte
S. aureus, USA300	13	0.046	4	0.207
S. aureus, Mu50	14	0.046	4	0.19
SARS-CoV-2	50	0.108	2	0.389
H1N1, Puerto Rico / 1934	14	0.038	3	0.375
H1N1, S. Canterbury / 2000	14	0.035	4	0.5
H1N1, St.Petersburg / 2006	14	0.04	4	0.5
H1N1, Texas / 2007	13	0.034	4	0.5

B

	Anz. Epitope	Epitopdichte	Anz. Signalepitope	Signal-Epitopdichte	Gekappte Seq.
S. aureus, USA300	13	0.047	4	0.217	2
S. aureus, Mu50	14	0.047	4	0.191	3
SARS-CoV-2	19	0.103	2	0.5	2
H1N1, Puerto Rico / 1934	13	0.045	2	0.25	0
H1N1, S. Canterbury / 2000	13	0.04	0	0	0
H1N1, St.Petersburg / 2006	13	0.046	0	0	0
H1N1, Texas / 2007	13	0.039	0	0	0

Tabelle 5.6 **Zusammenfassung der Prädiktions-Daten der Mikroben-Stämme**: Über die Datensätze aus Tabelle 5.5. A: NetMHCpan, B: SYFPEITHI. Die gekappten Sequenzen sind als absolute Werte angegeben, die restlichen als Mittelwerte. Das Limit für die Sequenzlänge in SYFPEITHI ist 2000, längere Sequenzen wurden an dieser Stelle gekappt. Die Daten solcher Stämme über die gesamte Sequenz sind nur unter Vorbehalt des Einflusses gekappter Sequenzen zu betrachten. Die Signalpeptide sind davon nicht betroffen.

Die Anzahlen vorhergesagter Epitope der *S. aureus*- und *Influenza*-Stämme unterschieden sich kaum. Mit den ähnlichen Sequenzlängen aus Tabelle 5.5 resultierten in Tabelle 5.6 auch ähnliche Epitopdichten für die Gesamtsequenzen. Zu *SARS-CoV-2* wurden von NetMHCpan fast viermal so viele Epitope vorhergesagt. Der große Einfluss der Polypeptide (Abbildung 10.8) wurde hier insofern deutlich, als dass SYFPEITHI, das nur die gekürzten Sequenzen behandelte, weniger als halb so viele Epitope vorhersagte. Das Verhältnis gekappter zu Sequenzen insgesamt war bei diesem Virenstamm mit 1:8 sehr groß, weshalb die Daten über die Gesamtsequenz von SYFPEITHI zu *SARS-CoV-2* weniger belastbar waren. Die Epitopdichten hingegen unterschieden sich kaum untereinander und waren mehr als doppelt so hoch wie die von *S. aureus* und *Influenza A*. Dabei wurden die Mittelwerte der Epitopdichten der *S. aureus*-Stämme noch durch viele Ausreißer nach oben hin beeinflusst (Abbildung 10.6).

Auch hier waren die Epitopdichten der Signalpeptide deutlich erhöht zu denen der Gesamtsequenzen (Tabelle 5.6). So betrug die Epitopdichte des Signalpeptids von *SARS-CoV-2* in den Daten von SYFPEITHI fast das Fünffache der der Gesamtsequenz. Die Signal-Epitopdichten der *Influenza A*-Stämme waren ebenfalls sehr hoch in NetMHCpan. Die Vermutung der besonders immunogenen Regionen wurde auch in den Viren-Stämmen bestärkt.

Bezüglich der *Influenza A*-Daten bestand eine Diskrepanz zwischen NetMHCpan und SYFPEITHI. Die Signalpeptid-Daten basierten auf nur einem notieren Peptid und zu diesem sagte NetMHCpan Epitope vorher und SYFPEITHI nicht. Aufgrund der Höhe der Signal-Epitopdichte wurde die Diskrepanz umso deutlicher. Die geringe Anzahl an Signalpeptiden in allen Viren bedingte den starken bis vollkommenen Einfluss einzelner Werte (siehe Anhang, Abbildung 10.7) und die große Varianz der Signal-Epitopdaten zwischen NetMHCpan und SYFPEITHI, entsprechend waren diese Daten kaum belastbar.

5.2.2 Qualitativ

		USA300	*Mu50*	*SARS-CoV-2*	*Puerto Rico*	*S. Canterbury*	*St.Petersburg*	*Texas*	
Länge ~ Masse		0.999	1	1	0.99	0.99	0.99	0.99	**Korrelation**
		<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	**P-Wert**
Länge ~ Signallänge		0.194	0.37	-1	N/A	N/A	N/A	N/A	
		0.165	<0.05	<0.05	N/A	N/A	N/A	N/A	
Länge ~ Anz. Epitope	**N**	0.764	0.76	0.65	0.93	0.94	0.91	0.93	
		<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	
Länge ~ Anz. Epitope	**S**	0.779	0.78	0.72	0.9	0.88	0.86	0.91	
		<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	
Länge ~ Epitopdichte	**N**	0.0842	0.057	-0.69	-0.044	0.039	-0.25	0.28	
		<0.05	0.078	<0.05	0.89	0.91	0.44	0.38	
Länge ~ Epitopdichte	**S**	0.0208	0.013	-0.75	-0.59	-0.24	-0.67	-0.22	
		0.59	0.68	<0.05	<0.05	0.45	<0.05	0.5	
Länge ~ Signal-Epitopdichte	**N**	-0.185	-0.32	-0.5	N/A	N/A	N/A	N/A	
		0.185	<0.05	0.67	N/A	N/A	N/A	N/A	
Länge ~ Signal-Epitopdichte	**S**	-0.278	-0.4	0.5	N/A	N/A	N/A	N/A	
		<0.05	<0.05	0.67	N/A	N/A	N/A	N/A	
Anz. Epitope ~ Epitopdichte	**N**	0.625	0.62	-0.018	0.18	0.24	0.046	0.48	
		<0.05	<0.05	0.95	0.56	0.46	0.89	0.12	
Anz. Epitope ~ Epitopdichte	**S**	0.575	0.57	-0.16	-0.27	0.063	-0.25	0.035	
		<0.05	<0.05	0.56	0.37	0.85	0.43	0.91	
Anz. Epitope ~ Signal-Epitopdichte	**N**	-0.0267	0.077	0	N/A	N/A	N/A	N/A	
		0.85	0.51	1	N/A	N/A	N/A	N/A	
Anz. Epitope ~ Signal-Epitopdichte	**S**	-0.148	-0.037	0.87	N/A	N/A	N/A	N/A	
		0.289	0.75	0.33	N/A	N/A	N/A	N/A	
Epitopdichte ~ Signal-Epitopdichte	**N**	0.46	0.61	0.87	N/A	N/A	N/A	N/A	
		<0.05	<0.05	0.33	N/A	N/A	N/A	N/A	
Epitopdichte ~ Signal-Epitopdichte	**S**	0.354	0.54	0.87	N/A	N/A	N/A	N/A	
		<0.05	<0.05	0.33	N/A	N/A	N/A	N/A	

Tabelle 5.7 **Korrelationen zwischen Sequenz- und Vorhersagedaten der Vergleichsstämme**: Spearman-Korrelation als Rangkorrelationsanalyse für nicht normalverteilte Daten mit Korrelationswert (1. Zeile) und zugehörigem P-Wert (2. Zeile). Zeilen mit dem Vermerk N beziehen sich auf NetMHCpan-Daten, S auf SYFPEITHI-Daten. Das Signifikanzniveau ist 5 %.

Viele der schon in Tabelle 5.3 beobachteten Korrelationen traten hier wieder auf. Einzig die stark negative Korrelation der Sequenz- und Signalpeptidlänge von *SARS-CoV-2* entsprach nicht der Erwartung, allerdings musste hier wieder beachtet werden, dass die Analyse nur auf drei Sequenzen beruhte. Aufgrund der teils geringen Anzahl notierter Sequenzen und Signalpeptide waren viele Korrelationsanalysen nicht möglich (*N/A*), vor allem für die *Influenza A*-Stämme, oder nicht signifikant.

Eindeutig stark und zu erwarten waren wieder die Werte der Korrelationen von Sequenzlänge mit Masse und mit der Anzahl der Epitope. Die Epitopdichte der *S. aureus*-Stämme korrelierte scheinbar positiv, entgegen der Erwartung nach Tabelle 5.3, allerdings war der Effekt nicht nennenswert. Die schon in Tabelle 5.3 aufgetretene negative Korrelation der Sequenzlänge mit Signal-Epitopdichte konnte hier für *Mu50* doppelt und für *USA300* in SYFPEITHI bestätigt werden. Ebenso zweimal doppelt die Epitopdichte jeweils mit der Anzahl der Epitope und Signal-Epitopdichte.

		F-Wert	P-Wert
Länge		4.85	<0.05
Masse		5.06	<0.05
Signallänge		1.75	0.114
Anz. Epitope	**N**	10.22	<0.05
Anz. Epitope	**S**	0.71	0.643
Epitopdichte	**N**	8.31	<0.05
Epitopdichte	**S**	8.08	<0.05
Signal-Epitopdichte	**N**	3.02	<0.05
Signal-Epitopdichte	**S**	3.06	<0.05

Tabelle 5.8 **Varianzanalyse ANOVA der Sequenz- und Vorhersagedaten mit Bezug auf die Vergleichsstämme**: Zeilen mit dem Vermerk N beziehen sich auf NetMHCpan-Daten, S auf SYFPEITHI-Daten. F-Wert als Indikator für die Größe der Unterschiede zwischen den Lokalisationen und P-Wert als Indikator für die Signifikanz der ANOVA. Das Signifikanzniveau ist 5 %.

Die Varianzanalyse und der Post-Hoc-Test offenbarten nicht so große Unterschiede unter den Mikroben-Stämmen (Tabelle 5.8). Die Analysen der Signallänge und der Anzahl von Epitopen in SYFPEITHI waren global nicht signifikant, womit die paarweisen Tests automatisch nicht mehr belastbar waren. Letztere ist dabei unter Umständen wieder durch das Kappen der Polyproteine beeinflusst. Global signifikant war der Vergleich der Signal-Epitopdichten in NetMHCpan zwar, paarweise waren nach Tukey aber keine Unterschiede wesentlich (siehe Anhang, Abbildung 10.9).

Wenn überhaupt, dann waren in den verbleibenden Analysen nur die Vergleiche mit *SARS-CoV-2* signifikant. Die Unterschiede in Sequenzlänge und -masse, Epitopanzahl in NetMHCpan und der Epitopdichten waren dabei groß. In der Signal-Epitopdichte unterschied sich *SARS-CoV-2* nur von den *S. aureus*-Stämmen in SYFPEITHI signifikant. Zwischen den *Influenza*- und *S. aureus*-Stämmen waren keine Varianzanalysen mit P-Wert geringer als 0,05 (Tabelle 5.8, siehe Anhang, Abbildung 10.9). Darauf aufbauend lassen sich also keine Unterschiede in Erkennung oder Bekämpfung der Mikroben durch das Immunsystem vermuten.

5.3 Algorithmus

Der finale Algorithmus war für jegliche Proteindaten aus UniProt anwendbar. Gestartet über die Eingabeaufforderung führte der Algorithmus bei Kommandozeilenparameter `-h` (`--help`) die

Hilfsfunktion aus. Bei `-e` (`--example`) den Abruf und die Vorhersage der Epitope der Proteindaten zu den vier Lokalisationen wie in 5.1 beschrieben und erstellte schließlich die Ausgabedateien *example_sa_uniprot.xlsx*, *example_sa_syfpeithi.xlsx* und *example_sa_netmhcpan.xlsx*, wobei *sa* hier für *S. aureus* stand.

Bei Kommandozeilenparameter `-f` (`--full`) wurden die Vorhersagen für beliebige Daten gemacht. Die Kategorien der Parameter der *query* waren festgelegt auf *organism* und *reviewed*. Dabei musste die Eingabe des Organismus auf der Konsole so gewählt werden, dass sie im letzten Eintrag des Taxonomiebaumes eines Proteins in UniProt exakt vorkommt. Stammte ein Protein etwa vom *S. aureus*-Stamm *USA300 / TCH1516* ab, dann war der letzte UniProt-Eintrag der Taxonomie *Staphylococcus aureus (strain USA300 / TCH1516)*. Dieses Protein wäre von einer Anfrage mit Organismus-Parameter *Staphylococcus aureus* oder *strain USA300* gespeichert worden, mit dem Organismus-Parameter *Staphylococcus aureus (strain USA300)* aufgrund der zweiten Klammer allerdings nicht.

Optional zu setzen waren Parameter für die Lokalisation (*locations*). Als Format war alternativ zu *tab* auch *fasta* zu wählen, wobei dann nur die Sequenzen in einer FastA-Datei ausgegeben und keine weiteren Daten oder Vorhersagen abgefragt wurden. Die Datenfelder im *tab*-Format waren weiterhin auf die oben genannten festgelegt. Zudem bestand für jeden der drei Datensätze die Möglichkeit der Eingabe eines individuellen Dateinamens. Ebenso wurden die Parameter der Vorhersagen, *allele*, *length* und der Score-Grenzwert für die Übernahme der Epitope über die Konsole eingegeben.

Es war somit möglich Epitopvorhersagen und -analysen zu jeglichem Organismus und jeder Lokalisation durchzuführen, die in UniProt notiert sind. Darüber hinaus waren die Vorhersagen über HLA-A*02:01 hinaus für jedes bei NetMHCpan und SYFPEITHI verfügbare MHC-Allel, auch von anderen Spezies als dem Menschen, möglich.

Die Vorhersage von NetMHCpan dauerte deutlich länger als die von SYFPEITHI. Das von Paul et al. 2020 beschriebene Verhältnis des Zeitaufwandes von 6:1 traf hier mindestens zu. Aufgrund dessen, dass die Ausführung von SYFPEITHI der von NetMHCpan vorgelagert war, konnte schon vor Beendigung der gesamten Ausführung mit dem Datensatz gearbeitet werden. Dies wurde vor allem während der Ausführung des Datensatzes über das *S. aureus*-Proteom nützlich, da die Vorhersagen mit NetMHCpan über drei Stunden dauerten.

Nicht alle Anfragen an die IEDB-Server waren erfolgreich. Das konnten einerseits unvollständige Antworten sein. Andererseits konnte es bei größeren Datensätzen vorkommen, dass der Zugang serverseitig temporär eingeschränkt wurde. Alle in diesem Zeitraum erfolgten Anfragen blieben erfolglos. In diesem Fall wurden auf der Kommandozeile und im Dokument ein Vermerk über eine

unvollständige Antwort der Server (*Bad Answer*) und die zugehörige ID des Proteins geschrieben. Am Ende sollte die Anzahl der fehlgeschlagenen Anfragen gegen die der erfolgreichen abgewogen und entschieden werden, ob das Ergebnis davon unbeeinflusst bleibt oder die Vorhersagen wiederholt werden müssen.

6 Diskussion

6.1 Methoden

Die gewählten Parameter erstellten ein möglichst großes und verlässliches Bild der Thematik. Unter allen in UniProt für *S. aureus* notierten Proteinen machen die überprüften, belastbaren Einträge 10.227 von insgesamt 137.957 aus. Davon sind 5.642 einer oder mehreren Lokalisationen zugeordnet. Rund 99,5 % (5.613 von 5.642) dieser Proteine sind aus den vier gewählten Lokalisationen (Zellmembran, Zellwand, Zytoplasma und sekretiert). Das gewählte MHC-Allel HLA-A*02:01 ist das in Europa häufigste Allel und für dieses haben Nonamere die höchste Bindungsaffinität (Gonzalez-Galarza et al. 2020; Eupedia.com 2015; Trolle et al. 2016). Die Grenzwerte wurden so gewählt, dass einerseits nur Epitope gespeichert werden, die eine realistische Bindungswahrscheinlichkeit haben, und andererseits die Anzahlen vorhergesagter Epitope zwischen den beiden Algorithmen in etwa übereinstimmen (siehe Anhang, Abbildung 10.10).

6.2 Bioinformatik

UniProt ist eine der größten Datenbanken für Proteine von Lebewesen und Viren. Insofern war sie die wahrscheinlich beste Wahl der Datenbank als Grundlage für den quantitativen Ansatz des Algorithmus, der auch in Zukunft für verschiedene Fragestellungen zu unterschiedlichen Organismen genutzt werden soll. Für *S. aureus* allerdings existieren andere, spezialisierte Datenbanken, wie *Aureo*Wiki (Fuchs et al. 2018). Diese könnten für den einzelnen Organismus detailliertere Informationen bieten und somit tiefgründigere Analysen erlauben. Für diesen quantitativen Ansatz ist die automatische Annotation von grundlegenden Merkmalen von UniProt allerdings eher von Nutzen, als die wahrscheinlich ausführlichere aber langsamer fortschreitende von *Aureo*Wiki.

Ein Problem, das keine der Datenbanken lösen kann, sind die in der Antigenpräsentation bedeutenden DRiPs. Mit 30 – 70 % aller synthetisierten Proteine machen diese einen großen Anteil derer aus, die zu Epitopen prozessiert werden. Die Datenbanken beinhalten allerdings nur funktionale Proteine, sodass aus den veränderten Stellen der DRiPs entstehende Epitope hier nicht betrachtet werden können.

Die Prädiktionsalgorithmen ergänzten sich in den Stärken und Schwächen. NetMHCpan bietet nach Paul et al. die beste Genauigkeit in den Vorhersagen, auch aufgrund des Einbeziehens von MHC-Liganddaten. Die Struktur macht den Algorithmus allerdings langsam, was bei einem quantitativen Ansatz wie diesem problematisch sein kann. Zudem kommt es während der Anfragen teils zu unvollständigen Rückgaben. SYFPEITHI, basierend auf einer Scoring-Matrix ist deutlich schneller (Paul et al. 2020). Im Falle der viralen Datensätze über nicht mehr als 20 Proteine ist die Laufzeit unerheblich. Für den Datensatz der 3222 *S. aureus*-Proteine allerdings wird die Vorhersage von SYFPEITHI in weniger als zehn Minuten und NetMHCpan in über einer Stunde ausgeführt. Gerade für potenziell instabile

Internetverbindungen, den schnellen Zugang zu Informationen oder die schließlich doppelte Datenlage kann die vorgelagerte Ausführung von SYFPEITHI nützlich sein.

Die Nutzung der Web-Server für die Vorhersagen hatte Nachteile. Eine angesprochene instabile Internetverbindung oder hohe Auslastung der Bandbreiten und Server konnten die Performance des Algorithmus direkt beeinflussen oder sogar komplett funktionslos machen. Die Integration des verfügbaren Quellcodes der beiden Algorithmen würde diese Probleme überwinden. Allerdings ist zumindest NetMHCpan nur unter Linux auszuführen und die Verteilung der Algorithmen auf Systemen mehrerer Nutzer könnte andere, ungünstigere Lizenzbestimmungen bedeuten. Ein weiterer Nachteil besteht im Eingabelimit der SYFPEITHI, wodurch die Daten nur bedingt belastbar sind. Die jetzige Methode, Sequenzen zu kappen ist nicht zielführend. Eine Lösungsmöglichkeit wäre die Teilung der Sequenzen mit einer Überlappung und Zusammenführen der Ergebnisse, ganz nach dem Teile-und-Herrsche-Verfahren. Die Teile müssten dabei mindestens um die Hälfte der gewählten Epitoplänge überlappen, um potenzielle Epitope auf der Schnittstelle nicht zu verlieren.

Die Grenzwerte der Scores für die Übernahme der Epitope sind auf die jeweilige Situation anzupassen. Hier wurden relativ niedrige Grenzwerte gewählt, da der Ansatz ein quantitativer war und eine große Datenlage als Ausgangspunkt weiterer Untersuchungen produziert werden sollte. Hier wurden also eine große Abdeckung der Epitope gefordert und dafür falsch-positive Epitope toleriert. Für andere Fragestellungen könnte aber eine hohe Spezifität mit möglichst wenigen falsch-positiven Ergebnissen besser geeignet sein (Paul et al. 2020).

6.3 Vergleich der Epitopdaten innerhalb der Spezies *Staphylococcus aureus*

Die Unterschiede in den Epitopdaten könnten unterschiedliche Immunogenitäten bedeuten. Die Proteine der Zellmembran zeigten teils mehr als zweifach erhöhte Epitopdichten in Abbildung 10.2 (siehe Anhang). Eine höhere Epitopdichte kann generell mehr aktive TCs bedeuten. Und so bedingt nach Cosma und Eisenlohr eine höhere Epitopdichte in der frühen Phase einer Infektion eine größere Anzahl an Gedächtnis-TCs. Somit sind hier die Ausreißer-Proteine mit auffallend hoher Epitopdichte von besonderem Interesse (siehe Anhang, Abbildung 10.2, Abbildung 10.3). Es ist also möglich, dass Proteine der Zellmembran eine deutlich stärkere Immunantwort in der Quantität auslösen und auch eher zu einem Immungedächtnis führen, was gerade in der Impfstoffforschung von Interesse wäre.

Zudem kann die Epitopdichte die Immunodominanz eines Epitops beeinflussen (Cosma und Eisenlohr 2019), wobei hier die Dichte eines einzelnen Epitops entscheidend ist, die sich anhand der Anzahl dieses Epitops auf der ersten Seite der Vorhersage-Datensätze berechnen ließe. Es sind also auch die anderen Kompartimente nicht zu vernachlässigen, da hier noch keine Analysen bezüglich der Anzahlen einzelner Epitope gemacht wurden. Hinzu kommt, dass eukaryotische Zellen keine Zellwand besitzen

und somit Epitope Zellwand-assoziierter Proteine einzigartig für eben diese Proteine sein können und nicht etwa auch von körpereigenen Proteinen prozessiert werden.

Eine hohe Epitopdichte innerhalb des Signalpeptids könnte einen größeren Einfluss haben als die der gesamten Sequenz. Signalpeptide werden während oder nach der Translokation zu ihrem Bestimmungsort abgespalten und durch Proteasen abgebaut (Heijne 1998). Somit werden potenziell von jedem synthetisierten Protein mit Signalpeptid Epitope produziert, anstatt nur von alternden und dysfunktionalen DRiPs. Dabei haben sich *in vitro* Peptide aus der Signalregion der Proteinsequenzen von *mycobacterium tuberculosis* als besonders immunogen für $CD8^+$ TCs herausgestellt (Kovjazin et al. 2011). Sequenzen hoher Signal-Epitopdichte, wie die der Extrazellulären oder der Zellmembran, könnten somit aufgrund der hohen Anzahl unmittelbar degradierter Peptidketten von höherer Immunogenität sein. Darüber hinaus konnte die Hypothese, Signal- oder Leadersequenzen hätten einen konservierten Character, verstärkt werden. Die Signalpeptide wiesen zumeist deutlich höhere Epitopdichten auf als die Gesamtsequenz (Tabelle 5.2,Tabelle *5.6*).

Die Zellmembran ist also insgesamt die potenziell immunogenste Region. Die hohe Anzahl vorhergesagter Epitope könnte an einer Verbindung zu den Anker-AS des MHC-Moleküls liegen. So wie die Zellmembran als Lipiddoppelschicht ein hydrophobes Kompartiment ist, sind auch die Anker-AS *Isoleucin*, *Leucin*, *Methionin* und *Valin* (Abbildung 1.9) hydrophobe Aminosäuren (Pacheco et al. 2017), deren Seitenketten eben solche Wechselwirkungen ermöglichen. Unter dieser Annahme sind kürzere Sequenzen mit Signalpeptid wahrscheinlich immunogener, da die Signal-Epitopdichte mit zunehmender Sequenzlänge stark abnimmt (Tabelle 5.3).

Der stammweise Vergleich von Sequenz- und Epitopdaten zu *S. aureus* wäre nicht belastbar gewesen. Die Datensätze der einzelnen Stämme reichten von weniger als zehn Proteinen bis zu mehr als 900. Diese Diskrepanz macht weitere Untersuchungen auf dem gesamten Datensatz diesbezüglich invalide. Stattdessen wäre eine Auswahl ähnlich gut annotierter Stämme möglich, wobei auch hier die Verteilung der Proteine vorher untersucht werden müsste.

6.4 Vergleich der Epitopdaten ausgewählter Stämme von *S. aureus* und Viren

Basierend auf den Epitopdichten war kein Unterschied zwischen *S. aureus* und *Influenza A* festzustellen. Bezüglich der von *S. aureus* bekannten Evasionsmechanismen wurden keine verringerten Epitopdichten verglichen mit denen der *Influenza A*-Stämme gefunden (Tabelle 5.6). Zwar waren die Signal-Epitopdichten letzterer in NetMHCpan stark erhöht, aber ebenso waren sie in SYFPEITHI stark erniedrigt, was insgesamt auf die dünne Datenlage zurückzuführen und nicht aussagekräftig ist. Ein möglicher Ansatz für weitere Untersuchungen der Epitope könnte sich hier bieten, da die Vorhersagen von NetMHCpan im Vergleich etwas genauer sind (Paul et al. 2020).

SARS-CoV-2 könnte vom Immunsystem leichter zu erkennen sein. Die Virusproteine haben mehrfach erhöhte Epitopdichten, also mehr Ziele für CTLs, zu den anderen betrachteten Organismen (Tabelle *5.6*, siehe Anhang, Abbildung 10.6, Abbildung 10.7). Der Einfluss der großen Polyproteine ist hierbei nicht eindeutig, da es weiter untersucht werden müsste, wann das Polyprotein in die einzelnen gespalten wird. Es könnte sein, dass Epitope auf den Schnittstellen oder in Regionen, die gar nicht in den MHC-Präsentationsweg gelangen liegen und somit das Ergebnis hier verfälscht wird. In den schon besser bekannten Datensätzen zu *S. aureus* und *Influenza A* sind keine Polyproteine notiert. Das Coronavirus bereitet trotz der hohen Epitopdichten diverse Komplikationen in der Immunabwehr und hat eine deutlich erhöhte Letalität (Piroth et al. 2021). Die Komplikationen können teils zu Autoimmunreaktionen führen, die dann nicht mehr von den T-Zell-Epitopen auf *SARS-CoV-2* abhängen. Ob ein Zusammenhang besteht, erfordert dahingehend weitere Untersuchungen.

Einen anderen Ansatz die vergleichsweise hohe Epitopdichte zu erklären, bietet die Coevolutionstheorie, nach der Organismen auch bezüglich ihrer Epitope der natürlichen Selektion unterliegen. Somit bestünde für *S. aureus* und *Influenza A* aufgrund der langen Zeit der Koexistenz und -evolution die Möglichkeit sich während dieser Zeit an den Wirt Mensch angepasst zu haben. *Influenza A* wurde symptomatisch schon etwa 460-370 vor Christus von Hippokrates beschrieben und brach seit 1900 dreimal pandemisch aus (Shahab und Glezen 1994). *S. aureus* wurde 1884 identifiziert, seit den 1940er Jahren mit Penicillin erfolgreich behandelt und entwickelte schon in den 1960er Jahren die erste Resistenz gegen Antibiotika (Feng et al. 2008). Allerdings besiedelt *S. aureus* als Kommensale des menschlichen Mikrobioms den Menschen wohl schon deutlich länger. Diese lange Coevolution besteht für *SARS-CoV-2* nicht, dass zu diesem Zeitpunkt erst seit knapp 1,5 Jahren im Menschen nachgewiesen wird. Trotz dessen verbreitet sich der Wildtyp des Virus seit Beginn an pandemisch und inzwischen auch Varianten, die virulenter sind oder reduzierter auf Antikörper reagieren (Robert Koch-Institut 2021). Ungünstig für den Menschen wäre es, würden sich die coevolutionären Prozesse so auf bestimmte Proteine und Epitope auswirken, dass aktuell entwickelte Therapien und Impfstoffe nicht mehr wirksam sind.

7 Ausblick

7.1 Bioinformatik

Die Datensätze bieten noch mehr Informationen. Die Validität der SYFPEITHI-Daten sollte zum Beispiel mittels der vorgeschlagenen Teile-und-Herrsche-Methode gewährleistet werden. Darüber hinaus lassen sich aus den einzelnen Einträgen zu jeden Epitop noch mehr Informationen über die Scores, Positionen und Sequenzen bereitstellen und analysieren.

Der Algorithmus kann als alleinstehendes ausführbares Programm von großem Nutzen sein. Da hiermit Vorhersagen zu Proteinen ganzer Datensätze automatisch und in geringer Zeit durchgeführt werden können, stellt der Algorithmus eine erhebliche Zeit- und Aufwandsersparnis dar. Für eine anwenderfreundliche Nutzung muss das Programm aber noch erweitert werden. Der Aufbau eines nachvollziehbaren Fehlermanagements und die Umgehung möglicher Fehler ohne den Absturz des Programms zu provozieren sind essenziell und die nächsten Schritte in der Entwicklung.

Darüber hinaus vereinfacht eine grafische Oberfläche (*graphics user interface*, GUI) die Bedienung. Gerade Anwendern, die nicht über die nötigen Kenntnisse zur Nutzung eines Programmcodes über die Konsole verfügen, würde das GUI helfen. Eben solche dürften in der Zielgruppe überwiegen. Die Entwicklungsumgebung *QT Creator* beispielsweise bietet mit dem Paket *PyQt* die Möglichkeit ein GUI für Algorithmen in der Sprache Python zu erstellen.

Die Nutzung des Algorithmus in der Zielgruppe setze allerdings auch voraus, dass Python und die nötigen Pakete installiert sind. Abhilfe können hier Programme wie *PyInstaller* bieten, die automatisch aus einer Python-Datei und den zugehörigen Paketen eine unter dem jeweiligen Betriebssystem direkt ausführbare Datei erstellen. *PyInstaller* unterstützt dabei die aktuellsten Versionen von Python und viele externe Pakete, wie *numpy*, *PyQt* und *requests*, die hier grundlegen sind. Unter bereits installiertem Python kann bereits jetzt eine Installationsdatei für die nötigen externen Pakete bereitgestellt werden.

Es besteht die Möglichkeit weiterführende Analysen zu den Epitopen automatisiert durchzuführen. Die IEDB bietet über das schon von NetMHCpan genutzte API Werkzeuge zur Analyse der Konservierung, des Clustering oder Ansprechen der Immunantwort in verschiedenen Populationen an. Die Integration solcher Analysen könnte die Auswahl besonders interessanter Proteine und Epitope erheblich vereinfachen. Dieser Schritt ist bis jetzt im Anschluss an den Algorithmus manuell auszuführen.

7.2 Biologie

Weitere Ausführungen des Algorithmus zu unterschiedlichen MHC-Allelen würden Erkenntnisse über günstige oder ungünstige MHC-Kombinationen bringen. Dadurch ließen sich Kombinationen

untersuchen, die potenziell die Abwehr von *S. aureus* begünstigen oder verschlechtern, was direkt die persistente Kolonisierung beeinflusst. So konnte für Mäuse bereits gezeigt werden, dass von Individuen mit dem Haplotyp H-2^d eine Immunantwort gegen *S. aureus* ausgelöst wird, aber nicht von solchen mit H-2^b. Allerdings wurde von Mäusen mit H-2^d eine Vakzin-induzierte T-Zell-Antwort aufgrund der kompetitiven Bindung der eigenen T-Zellen zum MHC verhindert (Si et al. 2020). Die Möglichkeit der Vorhersage für mehrere HLA-Allele innerhalb einer Ausführung würde ein noch größeres Bild liefern.

Ausgewählte Proteine und deren Epitope können labortechnisch weiter untersucht werden. Die Zellmembran als Kompartiment und Ausreißer jeder Lokalisation wurden bereits als aussichtsreiche Gegenstände zukünftiger Untersuchungen vorgeschlagen. Gerade die Ausreißer müssen allerdings überprüft werden, da die Prädiktionsalgorithmen keine Gegebenheiten in der Zelle oder dem Kompartiment miteinbeziehen können. Die Vorhersagen sind also nicht immer realistisch, sondern teils fehlerhaft. Es muss daher untersucht werden, ob die Epitope unter Realbedingungen exprimiert werden. Darüber hinaus können als solche bestätigte Epitope und deren MHC-Komplexe auf die Affinität und Avidität der Bindungen zu T-Zellen und die Immunodominanz hin labortechnisch untersucht werden.

Epitope die besonders häufig vorhergesagt werden sind ebenfalls von besonderem Interesse. Die Epitopdichten beziehen sich hier nur auf paarweise verschiedene Epitope. Allerdings könnte ein Epitop, das von vielen verschiedenen Proteinen, die selbst keine auffällige Epitopdichte haben, exprimiert wird deutlich häufiger präsentiert werden. Die Immunodominanz dieses Epitops würde also steigen und somit die Bedeutung in der Antigenpräsentation. Insofern bietet auch die Liste der Häufigkeiten jedes Epitops auf der ersten Seite jeder Ausgabe einen aussichtsreichen Ansatz für weitere Untersuchungen.

Die aktuelle Situation bedingt ein besonderes Interesse an der Antigenpräsentation von *SARS-CoV-2*. Die Kapazitäten für die Prüfung der Proteineinträge des Virus wurden von UniProt bereits erhöht (The UniProt Consortium 2020b). Dennoch sind zum 19.03.21 nur die Proteine des Wildtyps geprüft notiert. Aufgrund des Umfangs des Datensatzes von 16 Proteinen je Stamm könnten sich hier Analysen mit den ungeprüften Einträgen anbieten, die dann eigenhändig validiert werden müssten. Die aktuell starke Ausbreitung der virulenteren Mutation B.1.1.7 oder Reduktion der Antikörperwirksamkeit der Mutationen B.1.315 und P.1 bekräftigt das Interesse und die Notwendigkeit (Robert Koch-Institut 2021).

8 Fazit

Es wurde ein universell einsetzbarer Algorithmus zum Zusammentragen von Proteindaten mit zugehörigen Epitopdaten erstellt. Der Zugang zur grundlegenden Proteindatenbank UniProt ist uneingeschränkt und die Spezifizierung der Proteome an der Kommandozeile über den Organismus und die Lokalisation möglich. Die Nutzung von SYFPEITHI ist noch durch die Begrenzung der Sequenzlänge eingeschränkt und die von NetMHCpan durch die mangelnde Zuverlässigkeit der Antworten des Servers. Beide Probleme ließen sich durch die Implementierung des Quellcodes beheben, ebenso würde das die Laufzeit verbessern und Unabhängigkeit von der Internetverbindung schaffen. Die grafische Oberfläche und Übersetzung in ein alleinstehendes Programm würden den Algorithmus auch universell verfügbar machen. In vollem Umfang zu nutzen ist er schon unter installiertem Python und entsprechender Pakete.

Zu allen eingangs geforderten Subproteomen und Mikroben-Stämmen konnten die Protein- und Epitopdatensätze erstellt werden. In jeder Analyse konnten signifikante und aussagekräftige Ergebnisse produziert werden. Verglichen mit dem Gesamtproteom von *S. aureus* stellt die Zellmembran das einzige Kompartiment höherer Epitopdichten dar. Somit sind die assoziierten Proteine die aussichtsreichsten Kandidaten nachfolgender Untersuchungen möglicher Immunantworten. Die Vermutung, Signalpeptide seien konservierte Strukturen des adaptiven Immunsystems, konnte untermauert werden und liefert ebenso Ansätze weiterer Untersuchungen. *S. aureus* zeigt im Vergleich mit *Influenza A* keine Anzeichen von Immunevasionsmechanismen. Im Vergleich mit *SARS-CoV-2* offenbart sich allerdings der Mechanismus des Reduzierens der Epitopanzahl in *S. aureus* und *Influenza A* als wahrscheinliche Folge der langen Coevolution mit dem Menschen.

Insgesamt konnte ein Algorithmus entwickelt werden, der große Datensätze produziert, deren Erstellung bisher mit viel Arbeitsaufwand verbunden war. Gleichsam bieten die Datensätze viel Raum für Analysen und daraufhin nicht wenige Ansätze zu Untersuchung der Immunogenität von Organismen, Proteomen und einzelnen Peptidsequenzen.

9 Literaturverzeichnis

Abdurrahman, Goran; Schmiedeke, Frieder; Bachert, Claus; Bröker, Barbara M.; Holtfreter, Silva (2020): Allergy—A New Role for T Cell Superantigens of Staphylococcus aureus? In: *Toxins* 12 (3), S. 176. DOI: 10.3390/toxins12030176.

Abeck, Dietrich (2018): Staphylokokken- und Streptokokkeninfektionen der Haut. In: Gerd Plewig, Thomas Ruzicka, Roland Kaufmann und Michael Hertl (Hg.): Braun-Falco's Dermatologie, Venerologie und Allergologie. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 1–28.

Andersen, Mads Hald; Schrama, David; Thor Straten, Per; Becker, Jürgen C. (2006): Cytotoxic T cells. In: *The Journal of investigative dermatology* 126 (1), S. 32–41. DOI: 10.1038/sj.jid.5700001.

Archer, Gordon L. (1998): Staphylococcus aureus: A Well-Armed Pathogen. In: *Clinical Infectious Diseases* 26 (5), S. 1179–1181. Online verfügbar unter http://www.jstor.org/stable/4481569.

Blanca, María J.; Alarcón, Rafael; Arnau, Jaume; Bono, Roser; Bendayan, Rebecca (2017): Non-normal data: Is ANOVA still a valid option? In: *Psicothema* 29 (4), S. 552–557. DOI: 10.7334/psicothema2016.383.

Bortz, Jürgen; Döring, Nicola (2006): Forschungsmethoden und Evaluation. 4. Aufl. Berlin, Heidelberg: Springer Berlin Heidelberg.

Bröker, Barbara; Schütt, Christine; Fleischer, Bernhard; VISUV (2019): Grundwissen Immunologie. 4th ed. Berlin, Heidelberg: Springer Berlin / Heidelberg.

Bröker, Barbara M.; Mrochen, Daniel; Péton, Vincent (2016): The T Cell Response to Staphylococcus aureus. In: *Pathogens (Basel, Switzerland)* 5 (1). DOI: 10.3390/pathogens5010031.

Burnet, F. M. (1959): The clonal selection theory of acquired immunity ; the Abraham Flexner lectures of Vanderbilt University. 1. Aufl. Nashville, London: Vanderbilt University Press Tennessee; Cambridge University Press. Online verfügbar unter https://books.google.de/books?id=SZncuQEACAAJ.

Coleman, Denise (2015): Box Plot with Minitab. Hg. v. Lean Sigma Corporation, zuletzt aktualisiert am 22.12.2015, zuletzt geprüft am 28.03.2021.

Cosma, Gabriela L.; Eisenlohr, Laurence C. (2019): Impact of epitope density on CD8+ T cell development and function. In: *Molecular immunology* 113, S. 120–125. DOI: 10.1016/j.molimm.2019.03.010.

Driscoll, Wade C. (1996): Robustness of the ANOVA and Tukey-Kramer statistical tests. In: *Computers & Industrial Engineering* 31 (1-2), S. 265–268. DOI: 10.1016/0360-8352(96)00127-1.

Eupedia.com (2015): Distribution of HLA-A allels by country. Hg. v. Eupedia.com. Online verfügbar unter https://eupedia.com/genetics/HLA-A_allele_frequencies_by_country.shtml, zuletzt aktualisiert am 04.2015, zuletzt geprüft am 16.03.2021.

Feng, Ye; Chen, Chih-Jung; Su, Lin-Hui; Hu, Songnian; Yu, Jun; Chiu, Cheng-Hsun (2008): Evolution and pathogenesis of Staphylococcus aureus: lessons learned from genotyping and comparative genomics. In: *FEMS microbiology reviews* 32 (1), S. 23–37. DOI: 10.1111/j.1574-6976.2007.00086.x.

Fuchs, Stephan; Mehlan, Henry; Bernhardt, Jörg; Hennig, André; Michalik, Stephan; Surmann, Kristin et al. (2018): AureoWiki - The repository of the Staphylococcus aureus research and annotation community. In: *International journal of medical microbiology : IJMM* 308 (6), S. 558–568. DOI: 10.1016/j.ijmm.2017.11.011.

Gaseitsiwe, Simani; Valentini, Davide; Mahdavifar, Shahnaz; Reilly, Marie; Ehrnst, Anneka; Maeurer, Markus (2010): Peptide microarray-based identification of Mycobacterium tuberculosis epitope binding to HLA-DRB1*0101, DRB1*1501, and DRB1*0401. In: *Clinical and vaccine immunology : CVI* 17 (1), S. 168–175. DOI: 10.1128/CVI.00208-09.

Gonzalez-Galarza, Faviel F.; McCabe, Antony; Santos, Eduardo J. Melo Dos; Jones, James; Takeshita, Louise; Ortega-Rivera, Nestor D. et al. (2020): Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. In: *Nucleic acids research* 48 (D1), D783-D788. DOI: 10.1093/nar/gkz1029.

Haring, Jodie S.; Badovinac, Vladimir P.; Harty, John T. (2006): Inflaming the CD8+ T cell response. In: *Immunity* 25 (1), S. 19–29. DOI: 10.1016/j.immuni.2006.07.001.

Hauck, Christof R.; Agerer, Franziska; Muenzner, Petra; Schmitter, Tim (2006): Cellular adhesion molecules as targets for bacterial infection. In: *European journal of cell biology* 85 (3-4), S. 235–242. DOI: 10.1016/j.ejcb.2005.08.002.

Heijne, G. von (1998): Life and death of a signal peptide. In: *Nature* 396 (6707), 111, 113. DOI: 10.1038/24036.

Janeway, Charles A., JR.; Tavers, Paul.; Walport Mark; Shlomchik, Mark J.;. (2001): Immunebiology: The Immune System in Health and Disease. 5. Aufl. New York, NY: Garland Science.

Jong, Nienke W. M. de; van Kessel, Kok P. M.; van Strijp, Jos A. G. (2019): Immune Evasion by Staphylococcus aureus. In: *Microbiology spectrum* 7 (2). DOI: 10.1128/microbiolspec.GPP3-0061-2019.

Jurtz, Vanessa; Paul, Sinu; Andreatta, Massimo; Marcatili, Paolo; Peters, Bjoern; Nielsen, Morten (2017): NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. In: *Journal of immunology (Baltimore, Md. : 1950)* 199 (9), S. 3360–3368. DOI: 10.4049/jimmunol.1700893.

Kaech, S. M.; Ahmed, R. (2001): Memory CD8+ T cell differentiation: initial antigen encounter triggers a developmental program in naïve cells. In: *Nature immunology* 2 (5), S. 415–422. DOI: 10.1038/87720.

Kaech, Susan M.; Wherry, E. John; Ahmed, Raft (2002): Effector and memory T-cell differentiation: implications for vaccine development. In: *Nature reviews. Immunology* 2 (4), S. 251–262. DOI: 10.1038/nri778.

Köck, Robin; Mellmann, Alexander; Schaumburg, Frieder; Friedrich, Alexander W.; Kipp, Frank; Becker, Karsten (2011): The epidemiology of methicillin-resistant Staphylococcus aureus (MRSA) in Germany. In: *Deutsches Arzteblatt international* 108 (45), S. 761–767. DOI: 10.3238/arztebl.2011.0761.

Kovjazin, Riva; Volovitz, Ilan; Daon, Yair; Vider-Shalit, Tal; Azran, Roy; Tsaban, Lea et al. (2011): Signal peptides and trans-membrane regions are broadly immunogenic and have high CD8+ T cell epitope densities: Implications for vaccine development. In: *Molecular immunology* 48 (8), S. 1009–1018. DOI: 10.1016/j.molimm.2011.01.006.

Mercado, R.; Vijh, S.; Allen, S. E.; Kerksiek, K.; Pilip, I. M.; Pamer, E. G. (2000): Early programming of T cell populations responding to bacterial infection. In: *Journal of immunology (Baltimore, Md. : 1950)* 165 (12), S. 6833–6839. DOI: 10.4049/jimmunol.165.12.6833.

Miller, Lloyd S.; Fowler, Vance G.; Shukla, Sanjay K.; Rose, Warren E.; Proctor, Richard A. (2020): Development of a vaccine against Staphylococcus aureus invasive infections: Evidence based on

human immunity, genetics and bacterial evasion mechanisms. In: *FEMS microbiology reviews* 44 (1), S. 123–153. DOI: 10.1093/femsre/fuz030.

Murphy, Kenneth M.; Weaver, Casey (2018): Janeway Immunologie. 9. Auflage. Berlin, Germany: Springer Spektrum (Lehrbuch).

Neefjes, Jacques; Jongsma, Marlieke L. M.; Paul, Petra; Bakke, Oddmund (2011): Towards a systems understanding of MHC class I and MHC class II antigen presentation. In: *Nature reviews. Immunology* 11 (12), S. 823–836. DOI: 10.1038/nri3084.

Nielsen, Morten; Lundegaard, Claus; Blicher, Thomas; Lamberth, Kasper; Harndahl, Mikkel; Justesen, Sune et al. (2007): NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. In: *PloS one* 2 (8), e796. DOI: 10.1371/journal.pone.0000796.

Pacheco, Shaun; Fung, Shan-Yu; Liu, Mingyao (2017): Solubility of Hydrophobic Compounds in Aqueous Solution Using Combinations of Self-assembling Peptide and Amino Acid. In: *Journal of visualized experiments : JoVE* (127). DOI: 10.3791/56158.

Paul, Sinu; Croft, Nathan P.; Purcell, Anthony W.; Tscharke, David C.; Sette, Alessandro; Nielsen, Morten; Peters, Bjoern (2020): Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. In: *PLoS computational biology* 16 (5), e1007757. DOI: 10.1371/journal.pcbi.1007757.

Piroth, Lionel; Cottenet, Jonathan; Mariet, Anne-Sophie; Bonniaud, Philippe; Blot, Mathieu; Tubert-Bitter, Pascale; Quantin, Catherine (2021): Comparison of the characteristics, morbidity, and mortality of COVID-19 and seasonal influenza: a nationwide, population-based retrospective cohort study. In: *The Lancet Respiratory Medicine* 9 (3), S. 251–259. DOI: 10.1016/S2213-2600(20)30527-0.

Rammensee, H.; Bachmann, J.; Emmerich, N. P.; Bachor, O. A.; Stevanović, S. (1999): SYFPEITHI: database for MHC ligands and peptide motifs. In: *Immunogenetics* 50 (3-4), S. 213–219. DOI: 10.1007/s002510050595.

Reynisson, Birkir; Alvarez, Bruno; Paul, Sinu; Peters, Bjoern; Nielsen, Morten (2020): NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. In: *Nucleic acids research* 48 (W1), W449-W454. DOI: 10.1093/nar/gkaa379.

Robert Koch-Institut (2018): Eigenschaften, Häufigkeit und Verbreitung von MRSA in Deutschland – Update 2015/2016.

Robert Koch-Institut (2021): Übersicht und Empfehlungen zu besorgniserregenden SARS-CoV-2-Virusvarianten (VOC). Hg. v. Robert Koch-Institut. Online verfügbar unter https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Virusvariante.html;jsessionid=BCF1707DB2FA1B453DDCE44FD13BA86A.internet092?nn=2386228, zuletzt aktualisiert am 15.03.2021, zuletzt geprüft am 19.03.2021.

Schmider, Emanuel; Ziegler, Matthias; Danay, Erik; Beyer, Luzi; Bühner, Markus (2010): Is It Really Robust? In: *Methodology* 6 (4), S. 147–151. DOI: 10.1027/1614-2241/a000016.

Schuler, Mathias M.; Nastke, Maria-Dorothea; Stevanovikć, Stefan (2007): SYFPEITHI: database for searching and T-cell epitope prediction. In: *Methods in molecular biology (Clifton, N.J.)* 409, S. 75–93. DOI: 10.1007/978-1-60327-118-9_5.

Shahab, Shamsa Z.; Glezen, W. Paul (1994): Influenza Virus. In: Isaac Schiff und Bernard Gonik (Hg.): Viral Diseases in Pregnancy. New York, NY: Springer New York (Clinical Perspectives in Obstetrics and Gynecology), S. 215–223.

Si, Youhui; Zhao, Fan; Beesetty, Pavani; Weiskopf, Daniela; Li, Zhaotao; Tian, Qiaomu et al. (2020): Inhibition of protective immunity against Staphylococcus aureus infection by MHC-restricted immunodominance is overcome by vaccination. In: *Science advances* 6 (14), eaaw7713. DOI: 10.1126/sciadv.aaw7713.

Sprent, J.; Tough, D. F. (2001): T cell death and memory. In: *Science (New York, N.Y.)* 293 (5528), S. 245–248. DOI: 10.1126/science.1062416.

Strauß, Lena; Stegger, Marc; Akpaka, Patrick Eberechi; Alabi, Abraham; Breurec, Sebastien; Coombs, Geoffrey et al. (2017): Origin, evolution, and global transmission of community-acquired Staphylococcus aureus ST8. In: *Proceedings of the National Academy of Sciences of the United States of America* 114 (49), E10596-E10604. DOI: 10.1073/pnas.1702472114.

syfpeithi.de: FindYourMotif. Online verfügbar unter http://www.syfpeithi.de/bin/MHCServer.dll/FindYourMotif?HLA_TYPE=HLA-A*02%3A01&AASequence=&OP1=AND&select1=002&content1=&OP2=AND&select2=004&content2=&OP3=AND&select3=003&content3=&OP4=AND, zuletzt geprüft am 28.03.2021.

syfpeithi.de (2012): Epitope Prediction. Online verfügbar unter http://www.syfpeithi.de/bin/MHCServer.dll/EpitopePrediction.htm, zuletzt aktualisiert am 27.08.2012, zuletzt geprüft am 25.03.2021.

Taylor, Tracey A.; Unakal, Chandrashekhar G. (2019): Staphylococcus Aureus. Treasure Island (FL): StatPearls Publishing. Online verfügbar unter http://europepmc.org/books/NBK441868.

The UniProt Consortium (2020a): About UniProt. Online verfügbar unter https://www.uniprot.org/help/about, zuletzt geprüft am 17.12.2020.

The UniProt Consortium (2020b): UniProt: the universal protein knowledgebase in 2021. In: *Nucleic acids research*. DOI: 10.1093/nar/gkaa1100.

The UniProt Consortium (2020c): Programmatic access. Retrieving entries via queries. Online verfügbar unter https://www.uniprot.org/help/api_queries, zuletzt aktualisiert am 16.04.2020, zuletzt geprüft am 11.02.2021.

Traeger, M.; Eberhart, A.; Geldner, G.; Morin, A. M.; Putzke, C.; Wulf, H.; Eberhart, L. H. (2003): Künstliche neuronale Netze. Theorie und Anwendungen in der Anästhesie, Intensiv- und Notfallmedizin. In: *Der Anaesthesist* 52 (11), S. 1055–1061. DOI: 10.1007/s00101-003-0576-x.

Trolle, Thomas; McMurtrey, Curtis P.; Sidney, John; Bardet, Wilfried; Osborn, Sean C.; Kaever, Thomas et al. (2016): The Length Distribution of Class I-Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele-Specific Binding Preference. In: *Journal of immunology (Baltimore, Md. : 1950)* 196 (4), S. 1480–1487. DOI: 10.4049/jimmunol.1501721.

van Belkum, Alex; Melles, Damian C.; Nouwen, Jan; van Leeuwen, Willem B.; van Wamel, Willem; Vos, Margreet C. et al. (2009): Co-evolutionary aspects of human colonisation and infection by Staphylococcus aureus. In: *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* 9 (1), S. 32–47. DOI: 10.1016/j.meegid.2008.09.012.

van Stipdonk, M. J.; Lemmens, E. E.; Schoenberger, S. P. (2001): Naïve CTLs require a single brief period of antigenic stimulation for clonal expansion and differentiation. In: *Nature immunology* 2 (5), S. 423–429. DOI: 10.1038/87730.

Wagner, Norbert; Dannecker, Günther (2007): Pädiatrische Rheumatologie. Heidelberg: Springer Medizin.

10 Anhang

10.1 Abbildungen

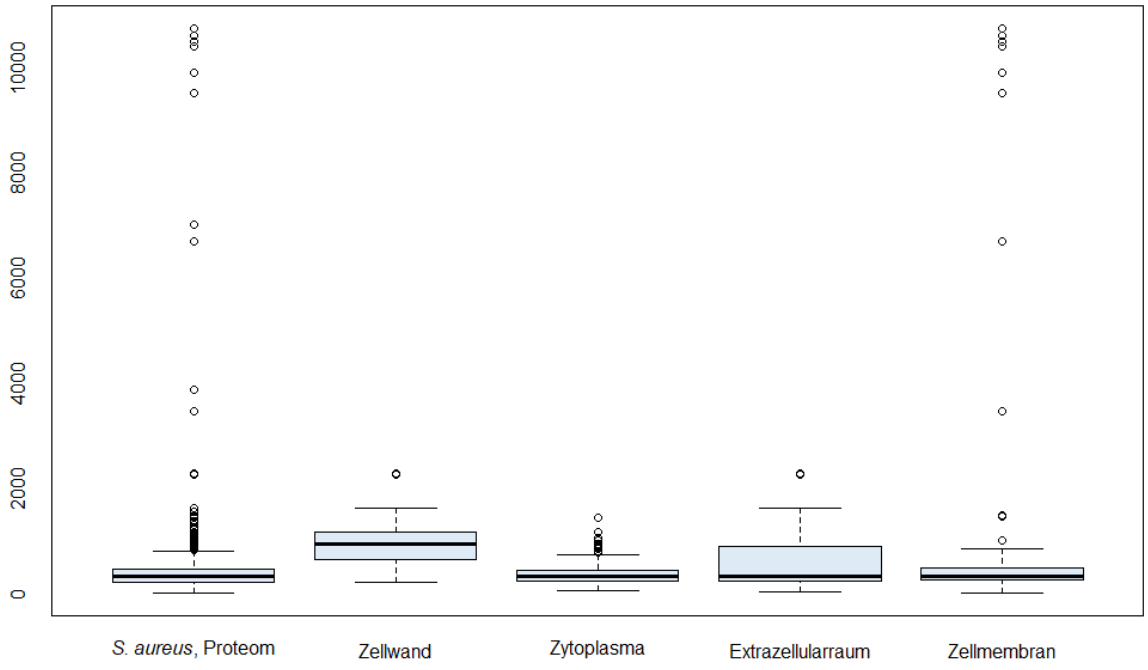

Abbildung 10.1 **Boxplots der Sequenzlängen von *S. aureus***: Anhand der Boxplots sind Ausreißer in den Daten von *S. aureus* im gesamten Proteom (i) und den betrachteten Kompartimenten (ii-v) gut zu erkennen.

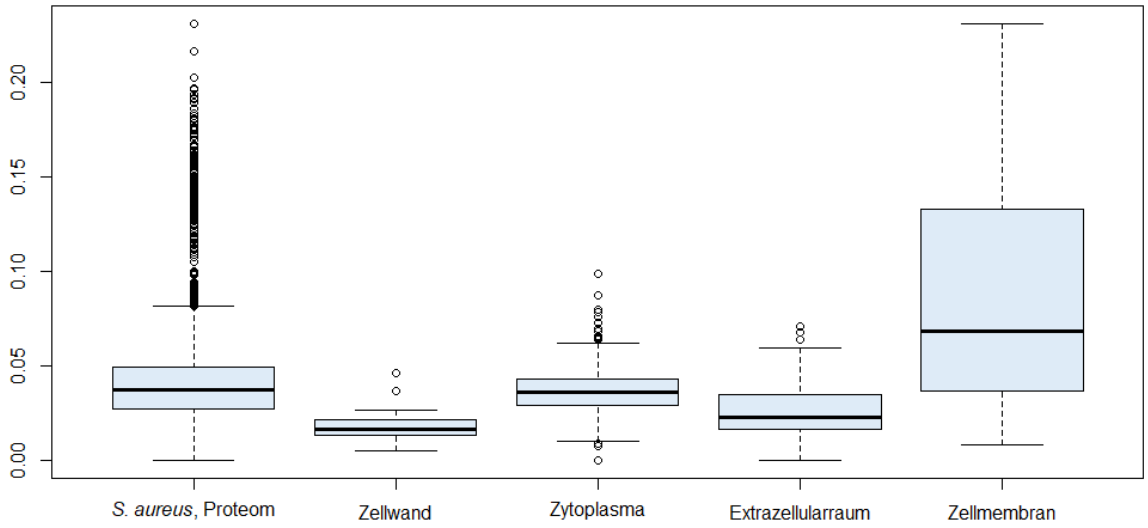

Abbildung 10.2 **Boxplots der Epitopdichten von *S. aureus* von NetMHCpan**: Anhand der Boxplots sind Ausreißer in den Daten von *S. aureus* im gesamten Proteom (i) und den betrachteten Kompartimenten (ii-v) gut zu erkennen.

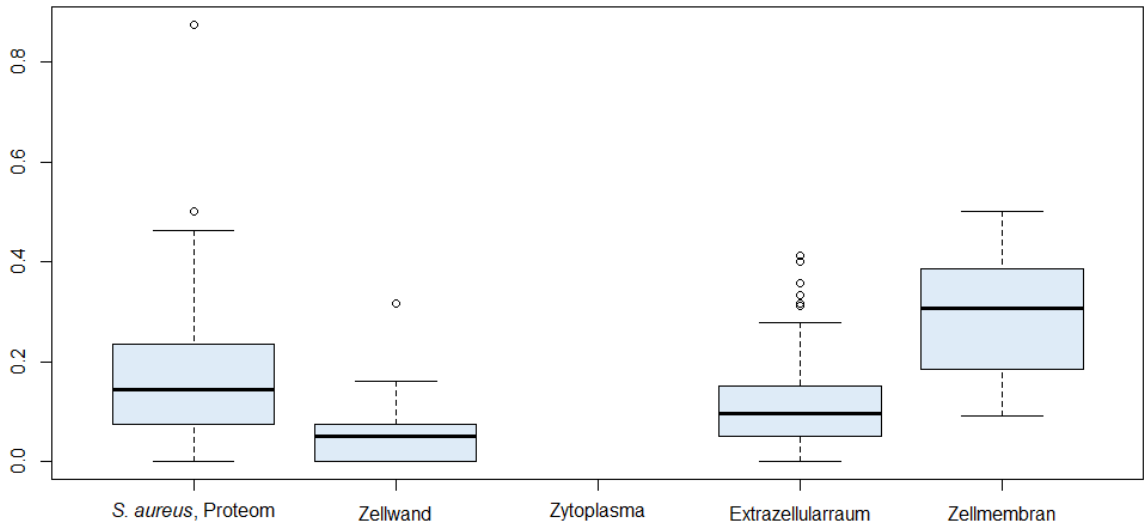

Abbildung 10.3 **Boxplots der Signal-Epitopdichten von *S. aureus* von NetMHCpan**: Anhand der Boxplots sind Ausreißer in den Daten von *S. aureus* im gesamten Proteom (i) und den betrachteten Kompartimenten (ii-v) gut zu erkennen.

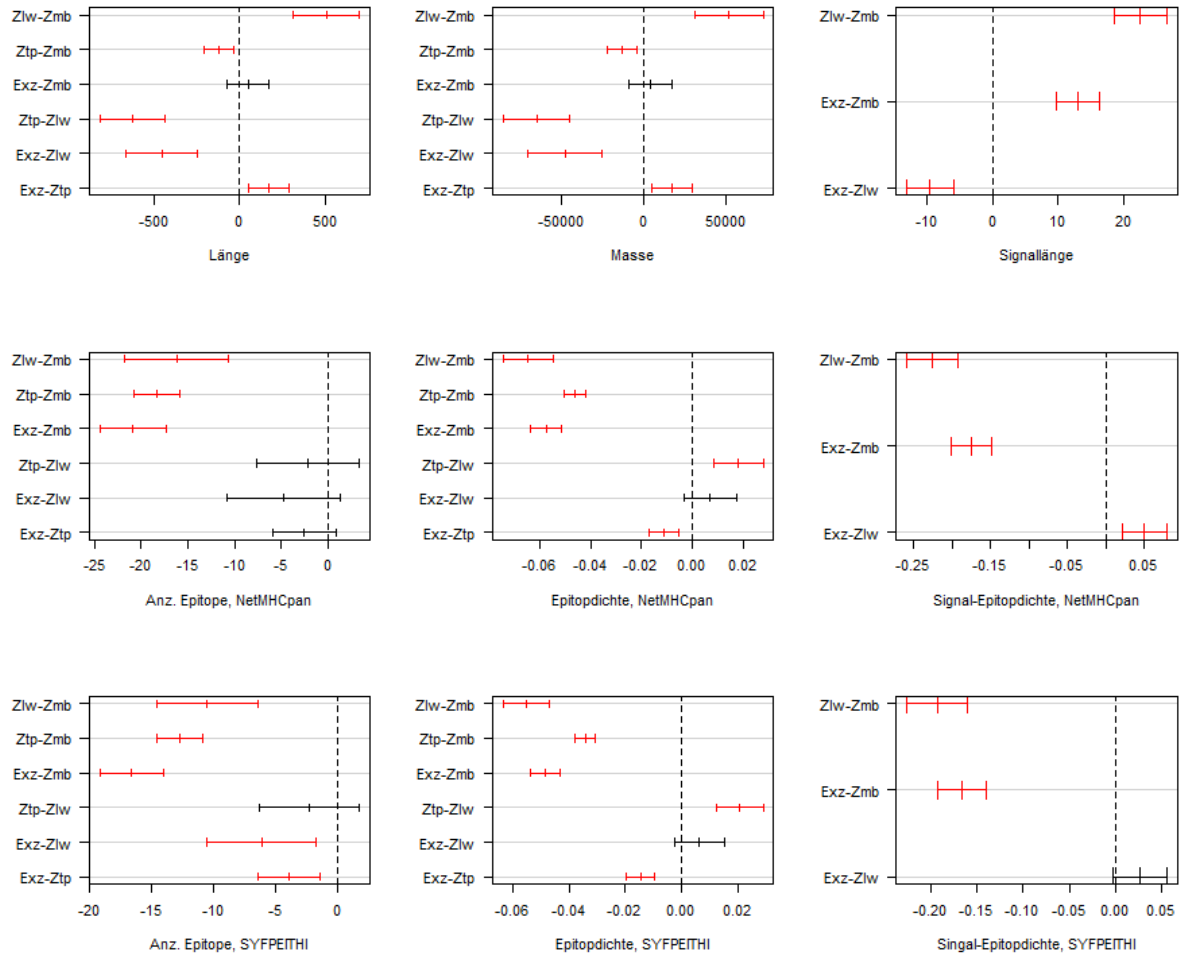

Abbildung 10.4 **Visualisierung des Post-Hoc-Test der ANOVA aus Tabelle 5.5 nach Tukey**: Paarweise Unterschiede der ANOVAs zwischen den Lokalisationen (Zellmembran = Zmb, Zellwand = Zlw, Zytoplasma = Ztp, Extrazellularraum = Exz). Das Konfidenzniveau ist 95 %, signifikante sind rot markiert.

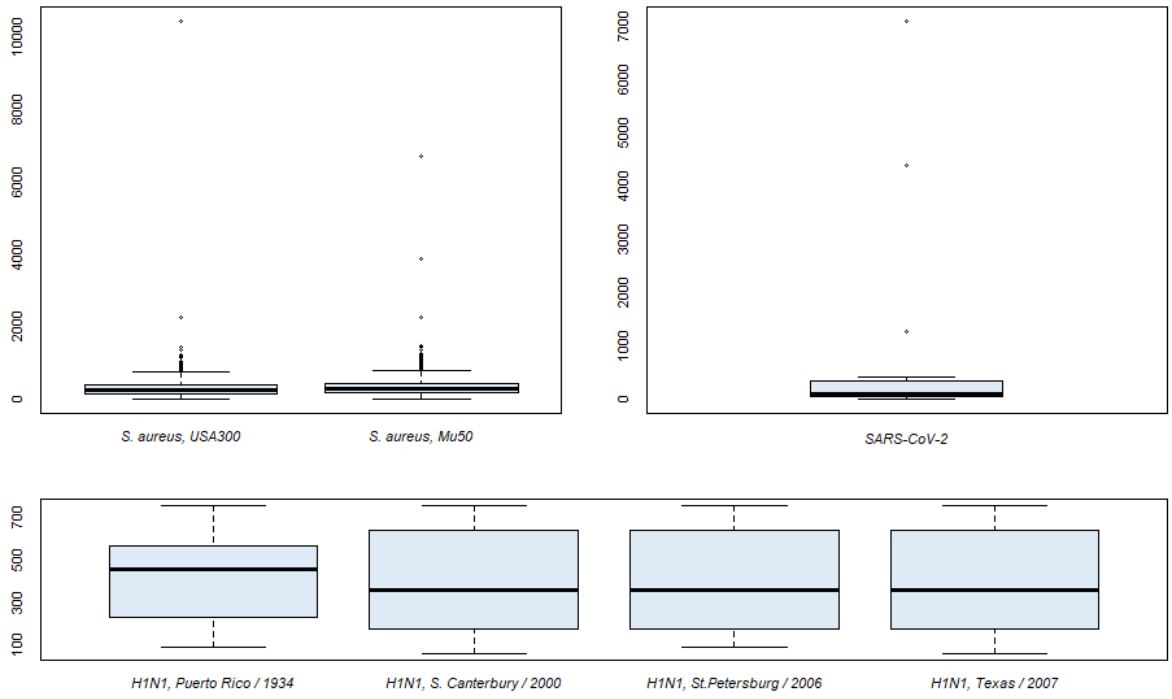

Abbildung 10.5 **Boxplots der Sequenzlängen der Vergleichsstämme**: Anhand der Boxplots sind Ausreißer in den Daten von *S. aureus*, *USA300* (i) und *Mu50* (ii) sowie *SARS-CoV-2* (iii) gut zu erkennen. Die betrachteten *Influenza A*-Stämme *Puerto Rico/8/1934*, *South Canterbury/35/2000*, *St.Petersburg/8/2006*, *Texas/UR06-0195/2007* (iv-vii) weisen keine Ausreißer auf.

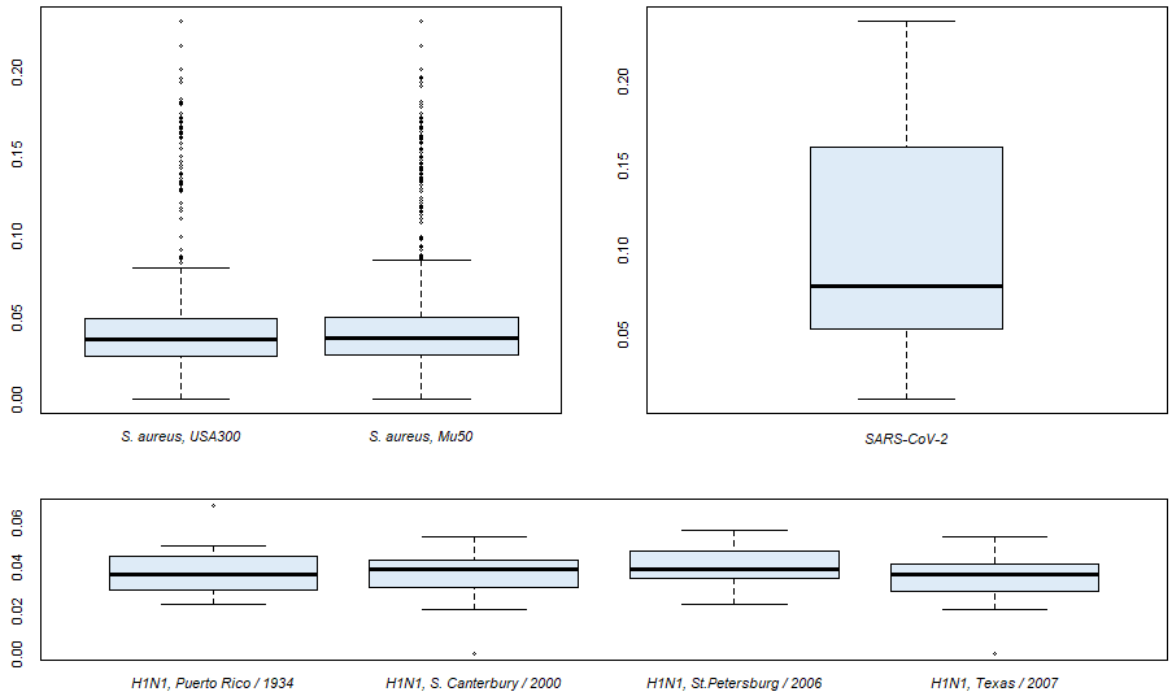

Abbildung 10.6 **Boxplots der Epitopdichten der Vergleichsstämme von NetMHCpan**: Anhand der Boxplots sind Ausreißer in den Daten von *S. aureus*, *USA300* (i) und *Mu50* (ii) sowie den Influenza A-Stämmen *Puerto Rico/8/1934*, *South*

Canterbury/35/2000 und *Texas/UR06-0195/2007* (iv, v, vii) gut zu erkennen. Der *Influenza A*-Stamm *St.Petersburg/8/2006* weist keine Ausreißer auf.

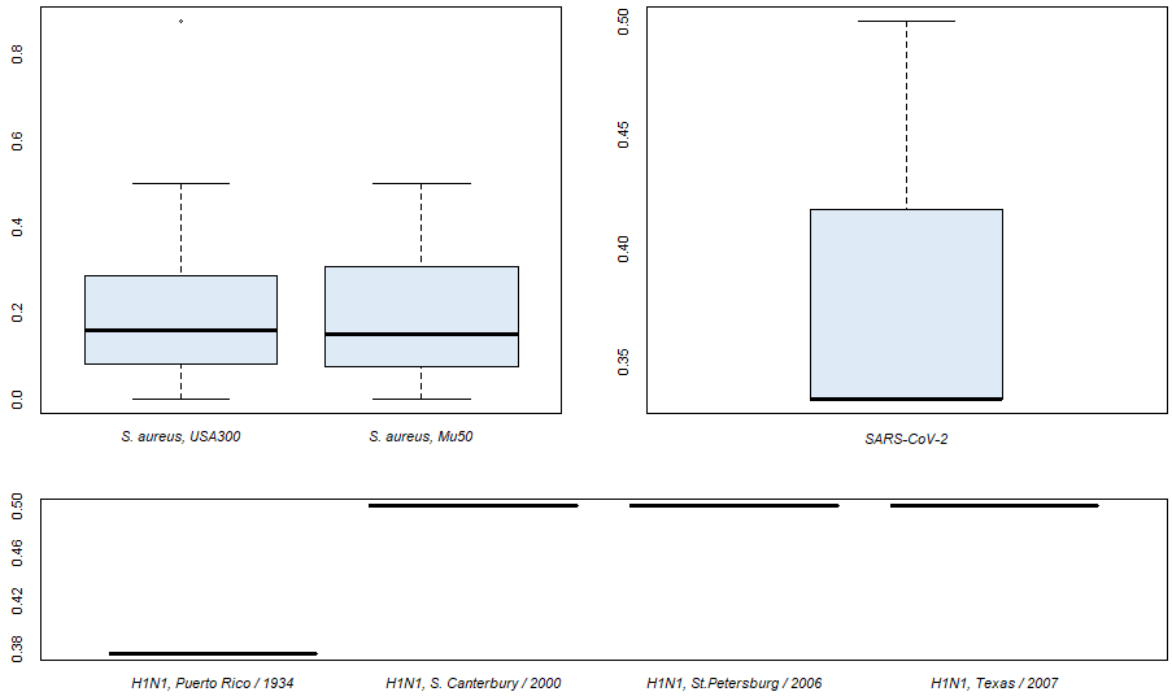

Abbildung 10.7 **Boxplots der Signal-Epitopdichten der Vergleichsstämme von NetMHCpan**: Anhand der Boxplots sind Ausreißer in den Daten von *S. aureus*, *USA300* (i) gut zu erkennen. Die restlichen betrachteten Stämme (ii-vii) weisen keine Ausreißer auf.

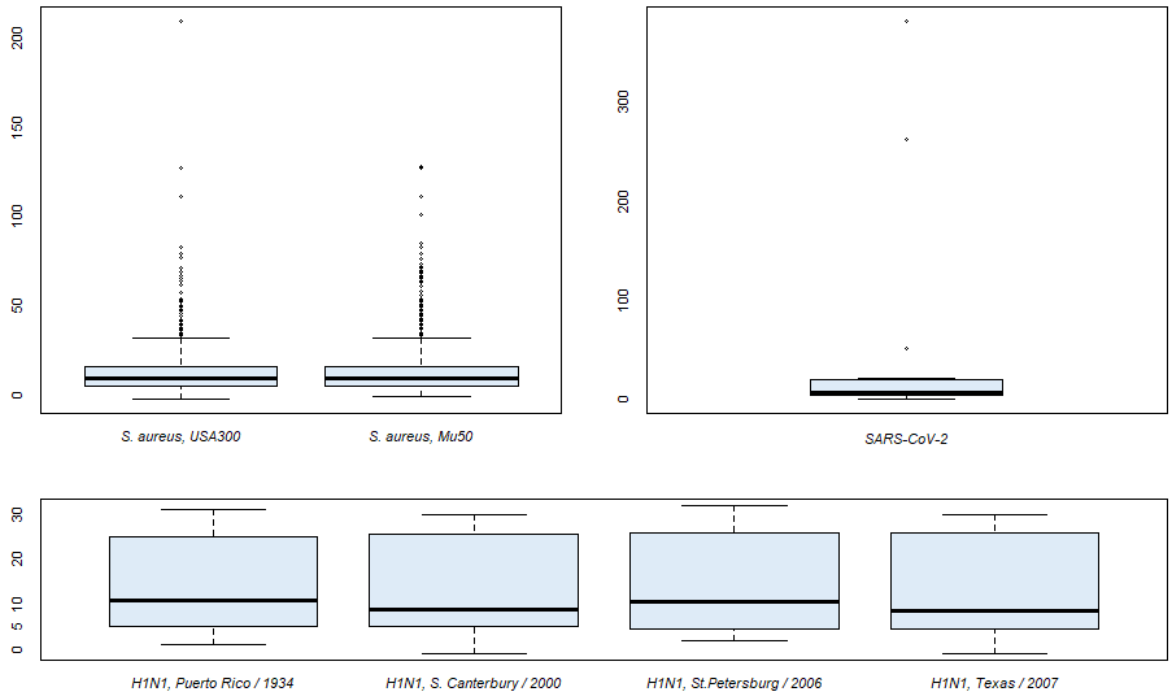

Abbildung 10.8 **Boxplots der Anzahlen vorhergesagter Epitope pro Sequenz der Vergleichsstämme von NetMHCpan**: Anhand der Boxplots sind Ausreißer in den Daten von *S. aureus*, *USA300* (i) und *Mu50* (ii) sowie *SARS-CoV-2* (iii) gut zu erkennen. Die restlichen betrachteten Stämme (iv-vii) weisen keine Ausreißer auf.

Abbildung 10.9 **Visualisierung des Post-Hoc-Test der ANOVA aus Tabelle 5.8 nach Tukey**: Paarweise Unterschiede der ANOVAs zwischen den Stämmen. Das Konfidenzniveau ist 95 %, signifikante sind rot markiert. Abkürzungen: *USA300* (US), *Mu50* (MU), *SARS-CoV-2* (CV), *Puerto Rico* (PR), *S. Canterbury* (SC), *St.Petersburg* (SP) und *Texas* (TX).

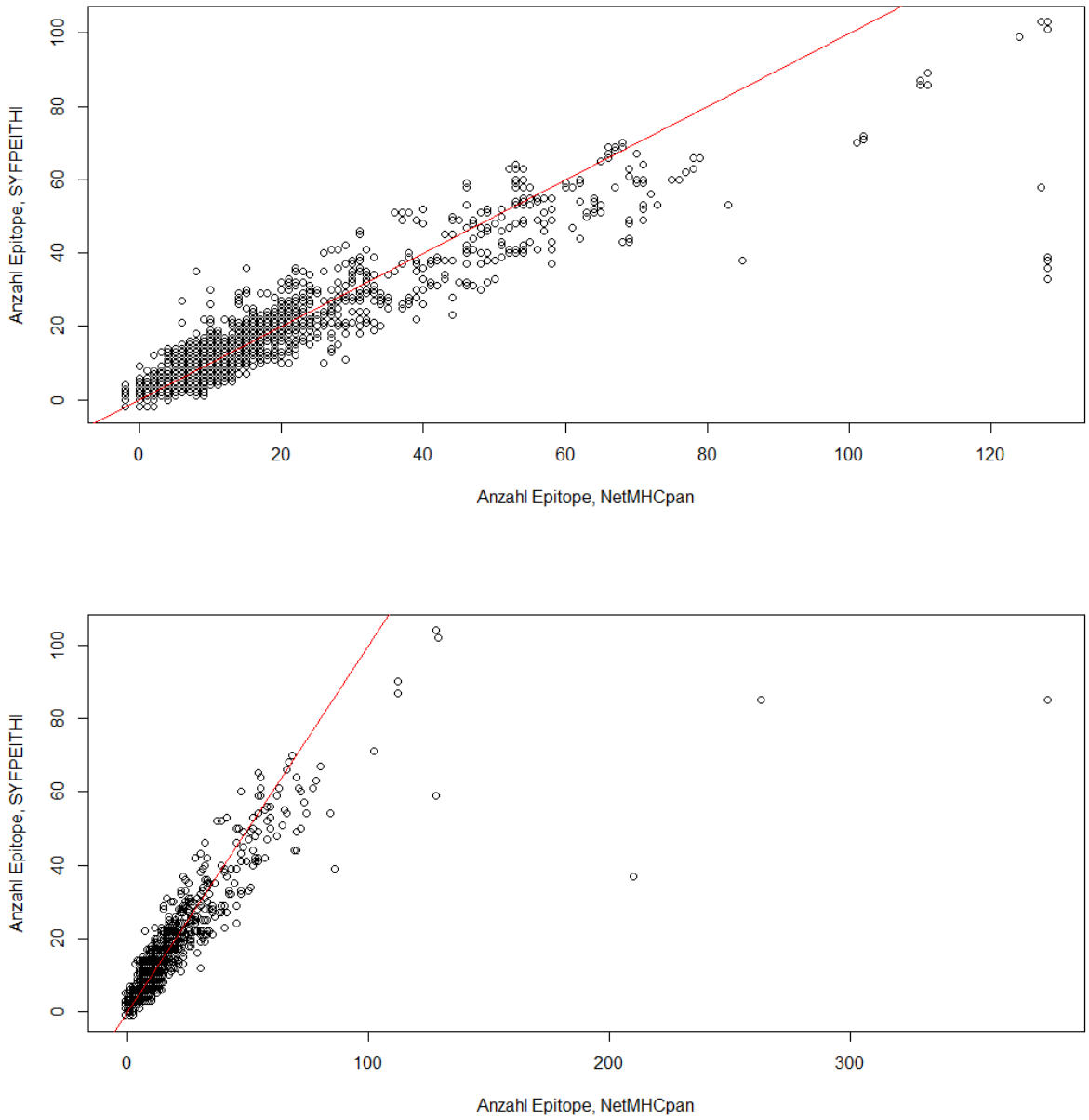

Abbildung 10.10 **Epitopanzahlen aus NetMHCpan und SYFPEITHI**: Oben: Die Sequenzen der vier Lokalisationen von *S. aureus*. Unten: Die Sequenzen der sieben Stämme, die verglichen wurden. Punkte auf der roten Linie, der 1. Winkelhalbierenden, haben die gleiche Anzahl an Epitopen in beiden Methoden.

10.2 Datenverfügbarkeit

Alle beschriebenen Daten und Quellcodes werden online zur Verfügung gestellt. Unter dem Profil *cmahncke* und gleichnamigen Speicherort sind die Ordner Code, R_Code und Results erstellt worden (zu erreichen unter https://www.github.com/cmahncke/cmahncke). Der erste Ordner enthält alle Python-Quellcode-Dateien, die für die Ausführung des hier beschriebenen Algorithmus benötigt werden inklusive der Installationsdatei für externe Module und Pakete. Der Ordner enthält auch eine *README*-Datei, die den Code, die Installation und die Ausführung erklärt. Der zweite Ordner enthält

alle R-Quellcode-Dateien, die für die statistische Auswertung und Erstellung der Grafiken und Tabellen genutzt wurden. Dabei enthält die Datei *staphAureus_data_analysis.R* alle Analysen, Grafiken und Tabellen zu den Subproteomen von *S. aureus*. *data_comparison.R* enthält die Grafiken und Tabellen der Vergleiche von *S. aureus* mit *Influenza A* und *SARS-CoV-2*, wobei dafür die jeweiligen R-Dateien die Analysen enthalten und vorher in der gleichen *Workspace* ausgeführt werden müssen. *data_analysis_functions.R* enthält ausgelagerte Funktionen, die für die Analysen notwendig sind, die Datei muss also als erstes in der *Workspace* ausgeführt werden. Der dritte Ordner enthält alle erstellten Tabellendokumente mit folgenden Vorsilben, wobei zu jeder eine UniProt-, SYFPEITHI- und NetMHCpan-Datei erstellt wurde:

Datei	Spezies	Beschreibung
cov2_	*SARS-CoV-2*	Proteom des Wildtyps
example_sa_	*S. aureus*	Subproteome der Spezies zu den Lokalisationen Extrazellularraum, Zellmembran, Zellwand, und Zytoplasma
mu50_	*S. aureus*	Proteom des Stammes *Mu50 / ATCC 700699*
pur_	*Influenza A*	Proteom des Stammes *A/Puerto Rico/8/1934*
sa_distinct_	*S. aureus*	Proteom der Spezies
scb_	*Influenza A*	Proteom des Stammes *A/New Zealand:South Canterbury/35/2000*,
stp_	*Influenza A*	Proteom des Stammes *A/Russia:St.Petersburg/8/2006*
tex_	*Influenza A*	Proteom des Stammes *A/USA:Texas/12/2007*
usa300_	*S. aureus*	Proteom des Stammes *USA300*

Danksagung

Der Dank gilt:

Prof. Dr. med. Barbara Bröker

Prof. Dr. Mario Stanke

Frieder Schmiedeke

Dr. Clemens Cammann

Prof. Dr. Lars Kaderali