Johann Wolfgang Goethe-Universität Frankfurt am Main
Fachbereich Chemische und Pharmazeutische Wissenschaften

Institut für Organische Chemie und Chemische Biologie

Diplomarbeit

Merkmalsextraktion mitochondrialer Targetingsequenzen in *Plasmodium falciparum*

Von	Cand. Chem. Andreas Bender Geboren am 20. August 1976 in Berlin
Betreut durch	Prof. Dr. Gisbert Schneider
Im Zeitraum	11. April 2002 bis 30. September 2002 an der Johann Wolfgang Goethe-Universität In Frankfurt am Main

Abstract

The malaria causing protozoan *Plasmodium falciparum (P. falciparum)* contains mitochondrial genes encoded in its nuclear genome. With the recent sequence completion of its genome, it is desirable to have software tools at hand for prediction of subcellular locations for all proteins. Established tools for the prediction of mitochondrial transit peptides like MitoProtII and TargetP were shown to perform poorly when applied to *P. falciparum* sequences. Therefore, methods specifically designed for this organism had to be developed. Nuclear-encoded mitochondrial protein precursors of *P. falciparum* were analyzed by statistical methods, principal component analysis, self-organizing maps and supervised neural networks and compared to those of other eukaryotes. Two types of descriptions were used, namely relative amino acid frequencies and 19 physicochemical properties. A general distinct amino acid usage pattern has been found in *P. falciparum,* compared to that of other organisms. Glycine, Alanine, Proline and Arginine are underrepresented, whereas Isoleucine, Tyrosine, Asparagine and Lysine are overrepresented, compared to the Swiss-Prot database, Version 36. These patterns were, with variations, also observed in all targeting sequences considered. Using Principal Component Analysis and Self-Organizing Maps, cytosolic N-terminal sequences showed considerable differences to mitochondrial, extracellular and apicoplastical targeting sequences, where the latter were difficult to distinguish from each other. A neural network system (PlasMit) for prediction of mitochondrial transit peptides in *P. falciparum* was developed based on the relative amino acid frequency in the first 24 N-terminal amino acids, yielding a Matthews correlation coefficient of 0.74 (86% correct prediction) in a 20-fold cross-validation study. This system predicted 2449 (24%) mitochondrial genes, based on 10276 predicted open reading frames in the *P. falciparum* genome. A network with the same topology has been trained to give a lower numer of false positive sequences in the training set. This second, more stringent network achieved a Mathews correlation coefficient of 0.51 (84% correct prediction) in a 10-fold cross-validation study. It predicted 903 (8.8%) mitochondrial genes, based on 10276 predicted open reading frames in the *P. falciparum* genome.

Danksagung

Mein Dank gilt Gisbert Schneider von der Universität Frankfurt für seine inhaltlich wertvolle und persönlich herzliche Betreuung. Ich danke Giel van Dooren von der Universität Melbourne, Australien, für die Zusammenstellung der Sequenzen aus *P. falciparum* und viele hilfreiche, biochemische Kommentare. Stuart Ralph, Universität Melbourne, danke ich für die Zusammenstellung der Open Reading Frames aus *P. falciparum*. Für hilfreiche Kommentare bin ich Geoffrey McFadden, Universität Melbourne, Australien dankbar. Auch der CallistoGen AG, insbesondere Dietmar Gundel und Paul Wrede, danke ich für ihre Unterstützung bei der vorliegenden Arbeit.

1. EINFÜHRUNG 8

1.1. Allgemeines, Zielsetzung 8

1.2. Die Biologie von *Plasmodium falciparum* 11

1.2.1. Taxonomie und Lebenszyklus 11

1.2.2. Zellaufbau 13

2. MATERIAL UND METHODEN 14

2.1. Zusammenstellung der Aminosäuresequenzen 14

2.1.1. Sequenzen aus Eukaryonten mit Ausnahme von *P. falciparum* 14

2.1.2. Sequenzen aus *P. falciparum* 16

2.1.3. Vorhergesagte Open Reading Frames aus *P. falciparum* 18

2.2. Datenaufbereitung 19

2.3. Angewandte Analysemethoden 26

2.3.1. Hauptkomponentenanalyse (Principal Component Analysis, PCA) 26

2.3.2. Selbstorganisierende Kohonen-Karte (Self-organizing map, SOM) 28

2.3.2.1. Allgemeines 28

2.3.2.2. Lernalgorithmus 29

2.3.3. Überwachtes Neuronales Netz (fully connected feed-forward ANN) 31

2.3.3.1. Allgemeines 31

2.3.3.2. Lernalgorithmus 32

2.4. Vorhersage von mitochondrialen Transitpeptiden mit etablierten Vorhersagemethoden 35

2.4.1. MitoProtII 35

2.4.2. TargetP 35

3. ERGEBNISSE 36

3.1. Vorhersage mitochondrialer Transitpeptide mit etablierten Methoden 36

3.1.1. MitoProtII 36

3.1.2. TargetP 36

3.2. Aminosäurehäufigkeiten 38

3.2.1. Vergleich *P. falciparum* und SwissProt, Version 36 38

3.2.2. Einzelne Aminosäuren in N-terminalen Abschnitten 39

3.2.3. Aminosäureklassen 47

3.2.4. Kullback-Leibler-Distanzen 51

3.2.5. Hauptkomponentenanalyse 54
3.2.6. Self-Organizing-Map 58

3.3. 19-dimensionaler Eigenschaftsraum 62
3.3.1. Allgemeines 62
3.3.2. Hauptkomponentenanalyse 62
3.3.3. Self-Organizing-Map 66

3.4. Ermittlung der optimalen Struktur des Neuronalen Netzes 70
3.4.1. Variation der Anzahl Hidden Neuronen 70
3.4.1.1. Variation der Anzahl Hidden Neuronen im Aminosäurehäufigkeitsraum 71
3.4.1.2.Variation der Anzahl Hidden Neuronen im 19-dimensionalen Eigenschaftsraum 76
3.4.2. Variation anderer Netzparameter 82
3.4.3. Genetische Variablenselektion 83
3.4.4. Training der optimalen Netze mit allen Daten 85
3.4.5. Optimierung auf möglichst wenig falsch-positive Vorhersagen 87

4. DISKUSSION 91

4.1. Vorhersage mit etablierten Methoden 91
4.1.1. MitoProtII 91
4.1.2. TargetP 91

4.2. Aminosäurehäufigkeiten / PCA / SOM 93

4.3. Ermittlung der optimalen Struktur des Neuronalen Netzes 96
4.3.1. Variation der Neuronenanzahl 96
4.3.2. Variation der Trainingsparameter 96
4.3.3. Fehlerinterpretation 98

4.4. Genome Scanning *P. falciparum* 100

5. AUSBLICK 102

6. ZUSAMMENFASSUNG 103

7. ANHANG 105

7.1. *P. falciparum* – positiver Datensatz 105

7.2. *P. falciparum* – negativer Datensatz 108

7.3. Quellcode des fully connected feed-forward ANN 117

8. LEBENSLAUF 124

9. EIDESSTATTLICHE ERKLÄRUNG 125

10. REFERENZEN 126

Abkürzungsverzeichnis

ANN – Artificial Neural Network, künstliches neuronales, mehrlagiges Perzeptron
cc – Mathews-Koeffizient
mTP – mitochondriales Transitpeptid
PCA – Principal Component Analysis, Hauptkomponentenanalyse
SOM – Self-Organizing-Map, Selbstorganisierende Karte

P: Anzahl korrekt positiv vorhergesagter Sequenzen
N: Anzahl korrekt negativ vorhergesagter Sequenzen
O: Anzahl negativer, falsch-positiv vorhergesagter Sequenzen
U: Anzahl positiver, falsch-negativer vorhergesagter Sequenzen

Sensitivität = P / (P + U)
Selektivität = P / (P + O)

Mathews-Koeffizient[1] $MCC = \frac{PN - OU}{\sqrt{(N+U)(N+O)(P+U)(P+O)}}$

Kullback-Leibler-Distanz $d = \sum_k p_k \log_2(\frac{p_k}{q_k})$, wobei q_k die unterliegende Häufigkeitsverteilung, die p_k die damit zu vergleichende Häufigkeitsverteilung bedeuten. Die Kullback-Leibler-Distanz wird auch als relative Entropie bezeichnet.

1. Einführung

1.1. Allgemeines, Zielsetzung

Mitochondrien eukaryontischer Zellen besitzen ihre eigene DNA, RNA sowie Ribosome zur Transkription und Translation[2]. So codiert beispielsweise die menschliche mitochondriale DNA 13 Proteine, 2 rRNAs sowie 22 tRNAs[3]. Der Ursprung der Mitochondrien liegt in einer Endosymbiose genannten Einverleibung von prokaryontischen Zellen in eine eukaryontische Wirtszelle[4]. Diese Annahme wird dadurch gestützt, dass der Translationsapparat der Mitochondrien prokaryontische Züge aufweist. Die Wirtszelle, die ansonsten nur in anaeroben Prozessen energiereiche Bausteine – Nucleotidtriphosphate - herstellen kann, wird durch die Mitochondrien in die Lage versetzt, diesen Prozess auch aerob auszuführen. Amöben, die als eukaryontischer Organismus keine Mitochondrien besitzen, benötigen daher einen aeroben Gast zu ihrer Existenz.

Während der Evolution fand (und findet weiterhin) ein Transfer von Genfunktionen von mitochondrialer DNA in die Kern-DNA der Wirtszelle statt. Ein Grund dafür könnte die geringere Mutationsrate der kerncodierten Gene gegenüber in den Mitochondrien lokalisierten Genen sein[5]. Kerncodierte Gene besitzen ebenso gegenüber dem mütterlicherseits asexuell vererbten mitochondrialen Genom durch die sexuelle Rekombination evolutionäre Vorteile. Ein analoger Prozess des Gentransfers in den Kern der Wirtszelle findet sich auch bei anderen Organellen[4], z.B. bei Apicoplasten („apicomplexan plastid“) in *Plasmodium falciparum (P. falciparum)* und *Toxoplasma gondii (T. gondii)*.

Nach der Translation der Proteine im Cytoplasma müssen diese in das entsprechende Zielkompartiment transportiert werden. Die Adressierung wird im Falle der für die Mitochondrien bestimmten Genprodukte über mitochondriale Transitsequenzen vorgenommen, die in den meisten Fällen durch Signalpeptidasen nach dem Import in die Mitochondrien abgespalten werden[6]. Der Großteil der Transitsequenzen ist N-terminal lokalisiert, es existieren jedoch auch C-terminale und interne Transitpeptide[7]. Im Falle mitochondrialer Vorläuferproteine werden diese in die mitochondriale Matrix transportiert, wo die Mitochondriale Processing Peptidase (MPP) das Transitpeptid entfernt (ebd.). An diesen Schritt kann sich die weitere Abspaltung eines Oktapeptids durch die Mitochondriale Intermediate Peptidase (MIP) anschließen (ebd.). Eine alternative Variante findet sich bei der Lokalisierung im Inner-Membrane-Space, in

diesem Fall existiert ein zweiter Teil des Transitpeptids, der die Lokalisierung im Subkompartiment bestimmt (ebd.). Falls vorhanden, wird dieser zweite Teil des Transitpeptids durch die Mitochondriale Inner-Membrane-Peptidase (IMP) abgespalten (ebd.). In Abbildung 1 sind Proteine unterschiedlicher Lokalisationen in ihrem Aufbau gegenübergestellt.

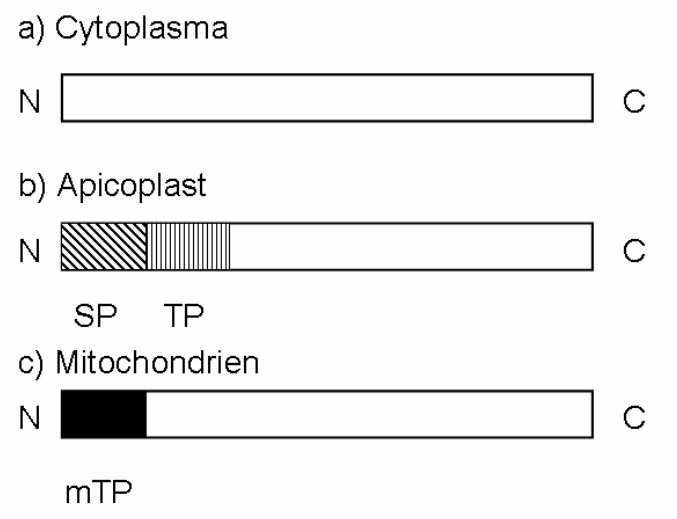

Abbildung 1 – Unterschiedlicher Aufbau von cytoplasmatischen Proteinen und apicoplastischen und mitochondrialen Proteinen, die jeweils unterschiedliche Targetingsignale zum Transport an ihren Zielort besitzen. Apicoplastische Targetingpeptide bestehen aus Signalpeptid (SP) und Transitpeptid (TP), mitochondriale Proteine besitzen ein mitochondriales Transitpeptid (mTP).

Typische mitochondriale Transitpeptide (mTPs) besitzen einen höheren Anteil an Arginin, Alanin und Serin und einen geringeren Anteil an den negativ geladenen Aminosäuren Asparaginsäure und Glutaminsäure[7]. Auf Basis der relativen Aminosäurehäufigkeiten am N-terminalen Ende der Peptide wurden bereits mehrere Algorithmen zur Vorhersage der Lokalisation des Proteins innerhalb der Zelle implementiert. So nimmt zum Beispiel MitoProtII[8] die Unterscheidung zwischen mitochondrial und nicht-mitochondrial lokalisierten Proteinen vor, PSORT[9] unterscheidet zwischen mehreren möglichen Zellkompartimenten. Das Programm TargetP[8] unterscheidet vier Zellkompartimente und verwendet ein neuronales Netz zur Klassifikation von Daten.

Im Falle von *P. falciparum* allerdings wurde deren Vorhersagesicherheit noch nicht bewertet und ist auch nicht durch einfachen Analogieschluss zu gewährleisten. Durch ein bei *P. falciparum* zusätzlich vorhandenes Plastid, den Apicoplasten („apicomplexan plastid"), wurde das Adressierungssystem des Organismus', die Targetingpeptide,

möglicherweise einem zusätzlichen Anpassungsdruck unterworfen, der zu einer Änderung der herkömmlichen Adressierungsschemata geführt haben könnte.

In dieser Arbeit

a) sollte die Anwendbarkeit bekannter Algorithmen, MitoProtII und TargetP, zur Vorhersage mitochondrialer Transitpeptide auf den Organismus *P. falciparum* überprüft werden,

b) sollten, im Falle einer Nichtanwendbarkeit der bisherigen Methoden, Unterschiede von mitochondrialen Transitpeptiden in *P. falciparum* zu solchen anderer Organismen beschrieben werden und eine neue Methode zur Vorhersage der Lokalisation von Peptiden anhand derer Aminosäuresequenzen entwickelt werden,

c) sollten durch eine Anwendung auf die vorhergesagten Open Reading Frames des Genoms von *P. falciparum* Anteil und Anzahl mitochondrialer Transitpeptide bestimmt werden.

1.2. Die Biologie von *Plasmodium falciparum*

1.2.1. Taxonomie und Lebenszyklus

P. falciparum ist eine Protozoe, ein parasitischer, eukaryontisch aufgebauter Einzeller, der zur Klasse der Sporozoen (Sporentierchen) gehört. Grund für die Klassifikation als Sporozoe sind die Abwesenheit von Bewegungsorganellen sowie eine in zwei Wirtsorganismen ablaufende Reproduktion, die in eine asexuelle Phase (Schizogenie) im Menschen und eine sexuelle Phase (Gamogenie und Sporogenie) in der Anopholesmücke unterteilt ist. Damit gehört *P. falciparum* zur Gruppe der Mehrfachwirt-Parasiten, dessen Endwirt, den Ort der sexuellen Vermehrung, die Anopholesmücke und dessen Zwischenwirt der Mensch darstellt.

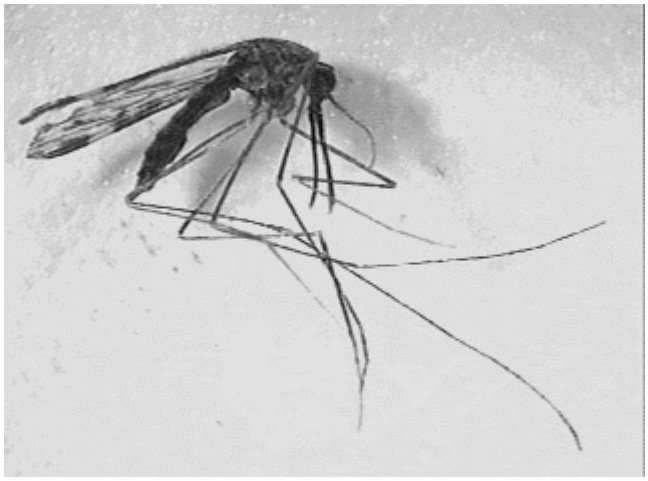

Abbildung 2 – Bild der weiblichen Anopholesmücke, die für die Übertragung von Malaria verantwortlich ist (Quelle: *anaesthetist.com*)

Der Zellzyklus dieses Organismus' beginnt mit dem Stich der weiblichen Anopholesmücke (Abbildung 2), wodurch Sporozoiten (Abbildung 3), Endprodukte der geschlechtlichen Vermehrung, in die Blutbahn des Menschen injiziert werden.

Abbildung 3 – Sporozoiten von *Plasmodium falciparum* (Quelle: *anaesthetist.com*)

Diese gelangen in die Leber und dringen in Hepatozyten, die eigentlichen Leberzellen ein. Durch asexuelle Vermehrung entstehen dort vielkernige Zellen, die sog. Schizonten (Abbildung 4).

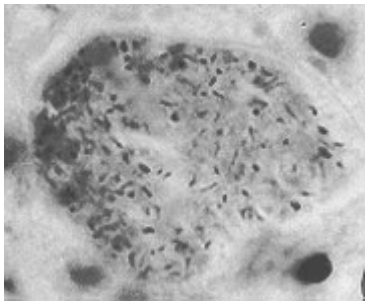

Abbildung 4 – Multinukleare Schizonten in den Hepatozyten (Quelle: *anaesthetist.com*)

Diese besitzen in einer Zelle mehrere, bis zu einigen 10^3 Zellkerne. Durch wiederholte cytoplasmatische Teilung entstehen nach Platzen der Schizonten einkernige Merozoiten, die in die Blutlaufbahn freigesetzt werden. Die freigesetzten Merozoiten (Abbildung 5)

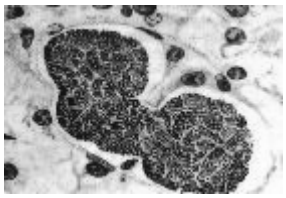

Abbildung 5 – Ein Schizont spaltet einen Merozoiten ab (Quelle: *anaesthetist.com*)

befallen nun die Erythrozyten, die roten Blutkörperchen (Abbildung 6). In diesen findet wiederum asexuelle Vermehrung, Schizogenie, statt.

Abbildung 6 – Erythrozyt, von Merozoiten (dunkler Kern) befallen (Quelle: *anaesthetist.com*)

Das auch hier erfolgende Platzen der Schizonten führt zu einem Massenzerfall von Erythrocyten und als körperliches Symptom zu einem Fieberschub. Nach mehreren Generationen asexueller Vermehrung entstehen aus einigen Plasmodien

Geschlechtsformen, männliche Mikrogameten und weibliche Makrogameten. Diese sterben nach einiger Zeit im Organismus ab, oder sie werden von einem neuen blutsaugenden Anopholes-Weibchen aufgenommen, in dem die sexuelle Phase (Gamogenie) mit anschließender Vermehrung (Sporogenie) abläuft. Damit beginnt der beschriebe Zyklus von neuem.
Details zum Lebenszyklus finden sich auf der Internetseite der Malaria Foundation International unter http://www.malaria.org/whatismalaria.html.

1.2.2. Zellaufbau

Als Eukaryont besitzt auch *P. falciparum* die die typischen Bestandteile Zellkern, Cytoplasma mit Organellen sowie eine Plasmamembran. Zusätzlich befindet sich im Cytoplasma ein Apicoplast genanntes Plastid, das, im Gegensatz zu den Plastiden in Pflanzenzellen, keine Funktion der Photosynthese ausführt. Die Funktionen dieser Organelle sind bis heute nicht geklärt, es ist jedoch bekannt dass sie eine große Anzahl kerncodierter Proteine importiert[10]. Die Nährstoffaufnahme erfolgt auf endozytische Weise, also durch Umhüllung des Nährstoffs. Dessen Abbau erfolgt in den Verdauungsvakuolen, Reserven werden in Speichervakuolen gelagert. Unverdaubares Material wird durch Exocytose, über exozytosische Vesikel, aus der Zelle ausgeschleust.

2. Material und Methoden

2.1. Zusammenstellung der Aminosäuresequenzen

2.1.1. Sequenzen aus Eukaryonten mit Ausnahme von *P. falciparum*

Zur Sequenzzusammenstellung von kerncodierten, in den Mitochondrien, im Cytoplasma und in der extrazellulären Matrix lokalisierten Proteinen wurde das „Sequence Retrieval System“ (SRS)[11] des European Bioinformatics Institute verwendet. Als Datenbank fand die SWISS-PROT[12] im Release 40.15 vom 16. April 2002 mit 107523 indexierten Einträgen Anwendung. Die hier wiedergegebenen Queries entsprechen der im SRS verwendeten Syntax.

Zur Zusammenstellung kerncodierter, mitochondrialer Proteine wurden folgende Parameter verwendet:
"((([swissprot-AllText:euka*] ! [swissprot-AllText:Plasmo*]) & ((([swissprot-FtKey:transit] & [swissprot-FtLength#1:1000000]) & ([swissprot-FtDescription:mito*] ! ((([swissprot-FtDescription:putative*] | [swissprot-FtDescription:similarity*]) | [swissprot-FtDescription:potential*]) | [swissprot-FtDescription:non_ter*])))) > parent)) ! (([swissprot-FtKey:mito*] | [swissprot-FtKey:non_ter*]) > parent)) ". Diese Anfrage lieferte 422 kerncodierte, in die Mitochondrien transportierte Proteine von Eukaryonten mit Ausnahme von *P. falciparum*.

Zur Zusammenstellung kerncodierter, im Cytosol der Zelle befindlicher Proteine wurde folgende Abfrage formuliert:
"((([swissprot-AllText:euka*] ! (((([swissprot-AllText:Plasmodium*] | [swissprot-AllText:non_ter*]) | [swissprot-AllText:init_met*]) | [swissprot-AllText:conflict*]) | [swissprot-AllText:potential*])) & [swissprot-EntryName:*e*]) & (([swissprot-CommentType:subcellular location] & ([swissprot-Comment:cytoplasm*] ! [swissprot-Comment:similarity*])) > parent)) ". Diese Anfrage lieferte 729 Einträge. Die Bedingung eines „e“ im Namen des Eintrags wurde zur Reduzierung der Trefferanzahl eingefügt, andernfalls wurden 2887 Treffer erzielt, die sich (siehe 2.2.) nicht in sinnvoller Zeit paarweise zur Redundanzreduktion ausrichten ließen.

Die dritte Gruppe bestand aus kerncodierten und aus der Zelle ausgeschleusten, also extrazellulären, Proteinen. Hier wurde folgende Abfrage benutzt: "(((([swissprot-AllText:euka*] ! (([swissprot-AllText:Plasmodium*] | [swissprot-AllText:non_ter*]) | [swissprot-AllText:init_met*])) & (([swissprot-CommentType:subcellular location] & [swissprot-Comment:extracellular*]) > parent)) & (([swissprot-FtKey:signal] & [swissprot-FtLength#1:1000000]) > parent)) ! ((([swissprot-AllText:euka*] ! (([swissprot-AllText:Plsmodium*] | [swissprot-AllText:non_ter*]) | [swissprot-AllText:init_met*])) & (([swissprot-CommentType:subcellular location] & ([swissprot-Comment:extracellular*] & [swissprot-Comment:similarity*])) > parent)) & ((([swissprot-FtKey:signal] & [swissprot-FtLength#1:1000000]) & ((([swissprot-FtDescription:potential*] | [swissprot-FtDescription:similarity*]) | [swissprot-FtDescription:alignment*]) | [swissprot-FtDescription:putative*])) > parent))) ". Diese Abfrage führte zu 656 die Bedingungen erfüllenden Sequenzen.

2.1.2. Sequenzen aus *P. falciparum*

Proteinsequenzen aus *P. falciparum* ließen sich nicht auf dem bei anderen Eukaryonten beschrittenen Wege zusammenstellen, da in der SWISS-PROT-Datenbank zu wenige Sequenzen dieses Organismus' vorhanden waren. Die Zusammenstellung sowohl der positiven als auch der negativen Datensätze von *P. falciparum* erfolgte von Giel van Dooren an der Universität Melbourne/Australien. Als positive Datensätze werden im Folgenden diejenigen Aminosäuresequenzen mit (vermutetem) mitochondrialen Transitpeptid bezeichnet, als negative Datensätze diejenigen ohne mitochondriales Transitpeptid.

Die Sequenzen wurden in diesem Fall nach folgenden Kriterien in die Liste kerncodierter, mitochondrial lokalisierter Proteine aufgenommen:

a) Sie sind homolog zu Proteinen ausschließlich oder gewöhnlich in den Mitochondrien anderer Organismen lokalisierter Proteine. Dazu gehören Proteine des Zitronensäurezyklus, der Elektronentransportkette, der Biosynthese von Ubichinon, einige am Abbau von Aminosäuren oder an der Biosynthese von Haem beteiligte Proteine, mitochondriale Transportproteine und mitochondriale RNA-prozessierende Enzyme
oder

b) Sie verzweigen bei der Konstruktion eines phylogenetischen Baumes von bekannten mitochondrialen Proteinen ab (was zum Beispiel für EF-Tu, cpn60 und hsp70 der Fall ist)
oder

c) Sie haben, wenn sie mit bakteriellen Proteinen ausgerichtet werden, N-terminale Erweiterungen und besitzen homologe Proteine, die in anderen Organismen in den Mitochondrien lokalisiert sind
oder

d) Sie wurden experimentell als mitochondriale Proteine bestimmt.

Mit diesen Kriterien konnte ein Datensatz von 40 kerncodierten, in den Mitochondrien lokalisierten Proteinen zusammengestellt werden.

Als negativer Datensatz wurden 135 Proteine zusammengestellt, die im Kern von *P. falciparum* codiert sind und im Cytoplasma oder im Apicoplasten lokalisiert sind oder aus der Zelle ausgeschleust werden. Hier wurden ebenfalls Analogieschlüsse zu anderen Organismen verwendet.

Eine Auflistung aller verwendeten Datensätze findet sich im Anhang sowie im Internet unter http://www.modlab.de.

2.1.3. Vorhergesagte Open Reading Frames aus *P. falciparum*

Dieser Datensatz wurde von Stuart Ralph von der University of Melbourne, Australien, zusammengestellt. Wie in der Version vom 24. Oktober 2001 der PlasmoDB[13] enthalten, wurden mit Hilfe der Programme glimmerM[14], fullphat und genefinder[15] insgesamt 20542, zum Teil identische Open Reading Frames vorhergesagt. Aus diesen 20542 potentiellen Proteinen wurden aus Plausibilitätsgründen diejenigen aussortiert, die nicht mit Methionin beginnen. Um identische vorhergesagte ORFs auszusortieren, wurden die jeweils ersten 20 Aminosäuren der Sequenzen verglichen und Duplikate aussortiert. Dadurch ergab sich eine Datenbasis von 10276 potentiellen Proteinen.

Die verwendeten vorhergesagten Open Reading Frames sowie die vorhergesagten mitochondrialen Transitpeptide finden sich im Internet unter http://www.modlab.de.

2.2. Datenaufbereitung

Um mögliche Unterschiede zwischen mitochondrialen Transitpeptiden in P. falciparum und anderen Organismen zu erhalten, wurden die Längen bekannter Transitpeptide bestimmt und die Häufigkeit in Abhängigkeit von der Länge ermittelt. Die Längenverteilung der Längen von Transitpeptiden mit bekanntem Schnittpunkt der Proteinase ist, getrennt nach 22 Proteinen aus *P. falciparum* und 422 aus anderen Eukaryonten, in Abbildung 7 dargestellt.

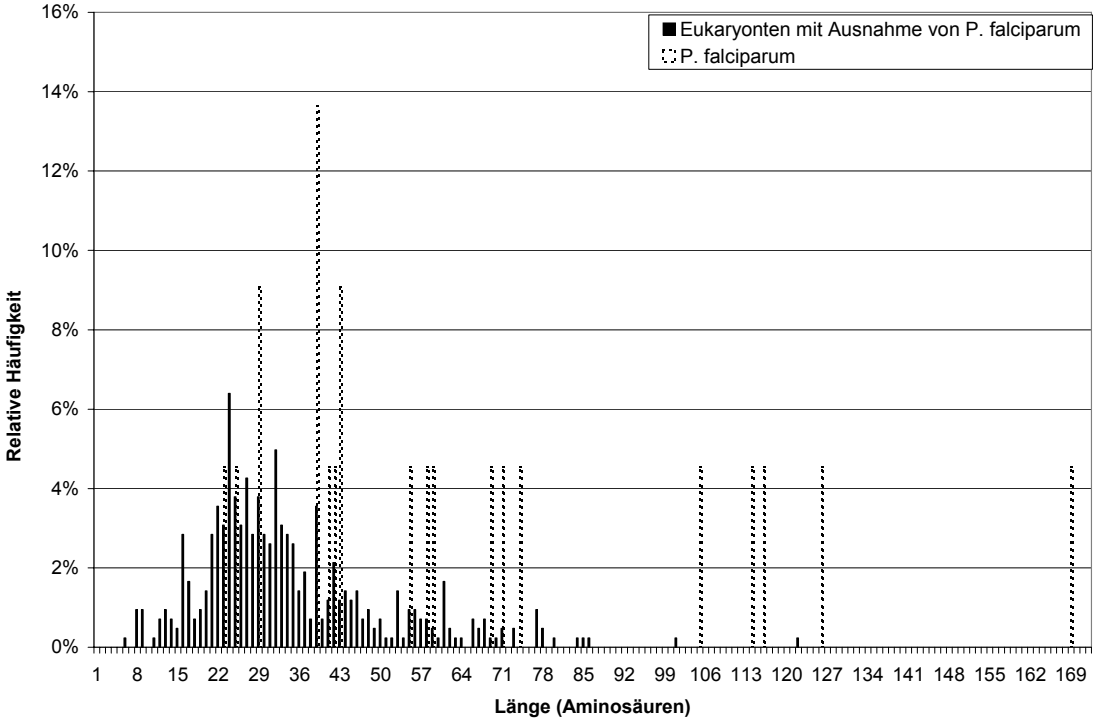

Abbildung 7 – Längenverteilung mitochondrialer Transitpeptide in *P. falciparum* und anderen Eukaryonten. Die Länge der 422 mitochondrialen Transitpeptide streut über einen weiten Längenbereich, was ein Grund dafür ist, dass sie sich nur schwer alignen lassen.

Die Längen der mitochondrialen Transitpeptide von Eukaryonten mit Ausnahme von *P. falciparum* streut weit um einen Mittelwert von 34,4 (σ = 5,2) Aminosäuren. Das 1. Quartil dieser Längenverteilung liegt bei 24, der Median bei 31 und das 3. Quartil bei 42 Aminosäuren.

Die Längenverteilung mitochondrialer Transitpeptide in *P. falciparum* wurde aufgrund der geringen Datenbasis nicht statistisch erfasst, sie ist jedoch, wie in Abbildung 7 sichtbar, offensichtlich unterschiedlich zu der in anderen Eukaryonten.

Analog wurde die Längenverteilung der Signalpeptide extrazellulärer Proteine in Eukaryonten mit Ausnahme von *P. falciparum* bestimmt (Abbildung 8). Diese

unterscheidet sich mit einer klar definierten unteren Längebegrenzung, einem Mittelwert von 21,2 Aminsäuren und einer Standardabweichung von 5,2 Aminosäuren von der Längenverteilung mitochondrialer Transitpeptide.

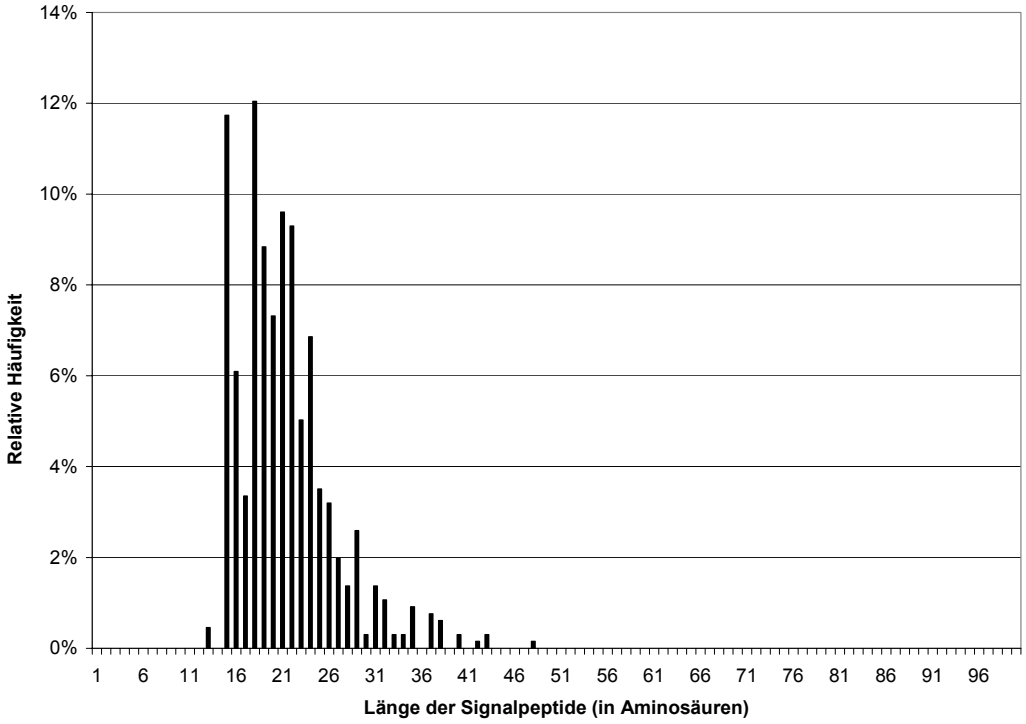

Abbildung 8 – Längenverteilung von Signalpeptiden in Eukaryonten mit Ausnahme von *P. falciparum.* Die Verteilung weicht deutlich von der in Abbildung 7 gezeigten Längenverteilung mitochondrialer Transitpeptide ab.

Die anschließend vollzogene Aufbereitung der Sequenzdaten erfolgte zur Redundanzreduktion. Da in unserer Datenzusammenstellung der nicht-Plasmodium-Sequenzen bisher keine homologen Proteine unterschiedlicher Eukaryonten explizit ausgeschlossen wurden, mussten wir diesen Schritt nun durchführen, um eine möglichst unbeeinflusste Datenzusammenstellung zu erhalten.

Im ersten Schritt musste die Frage entschieden werden, ob (wenn bekannt) die tatsächliche, experimentell bestimmte Länge der Transitpeptide oder eine konstante Länge verwendet werden sollte. Es war aus vorhergehenden Arbeiten bekannt dass sich komplette mitochondriale Transitsequenzen aufgrund ihrer differierenden Länge nur schwer alignen lassen. Es existieren prinzipiell die Möglichkeiten, nur Transitpeptide mit bekannter Länge zu verwenden oder alle Sequenzen auf eine (oder mehrere) einheitliche Längen zuzuschneiden. In einem ersten Schritt wurde versucht, die mitochondrialen Transitpeptide mit bekannter Länge aus Eukaryonten mit Ausnahme

von *Plasmodium* zu alignen. Die relative Häufigkeit paarweise ausgerichteter Aminosäuren gegen die Länge des Alignment ist in Abbildung 9 dargestellt.

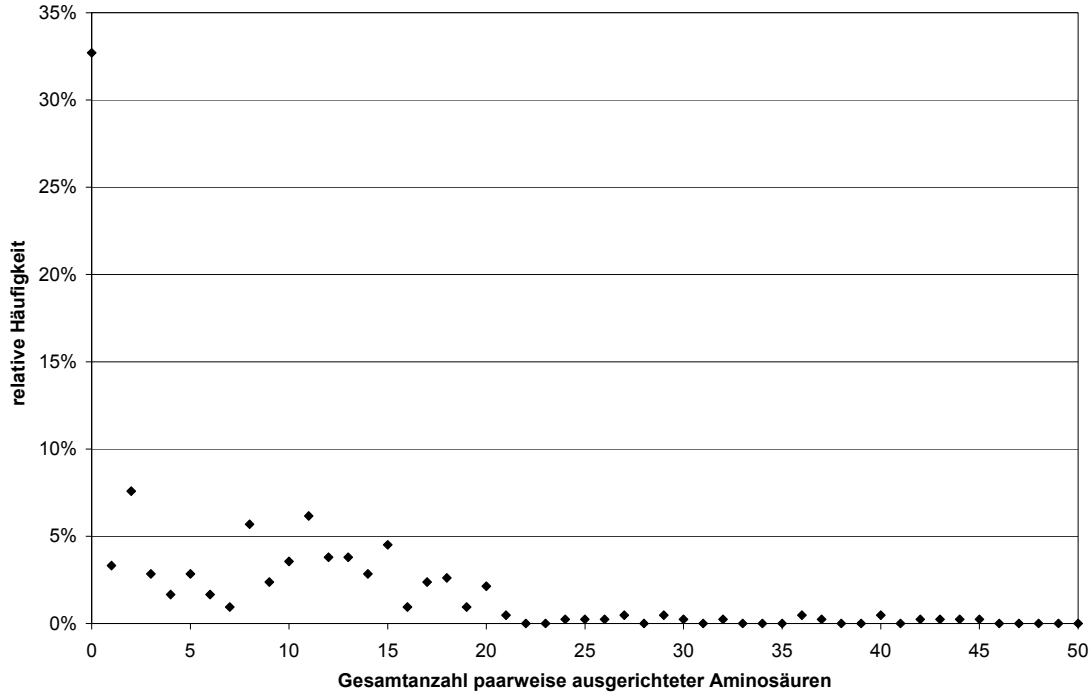

Abbildung 9 – Relative Häufigkeit der Gesamtanzahl paarweise ausgerichteter Aminosäuren unter 422 mitochondrialen Transitpeptiden. Nur wenige der 422 mitochondrialen Transitpeptide ließen sich über eine größere Zahl Aminosäuren ausrichten.

Es ist deutlich ersichtlich, dass in der weitaus größten Zahl der Alignments nur eine sehr kleine Gesamtzahl Aminosäuren paarweise ausgerichtet werden konnte (der Median lag bei 11 Aminosäuren, der Mittelwert bei 11,2 (Standardabweichung 8,3) Aminosäuren.

Da eine Redundanzreduktion jedoch zwingend war, konnte die Analyse kompletter Transitpeptide nicht durchgeführt werden. Daraus ergab sich, dass sämtliche Peptidsequenzen mit fixer Länge vom N-Terminus zurechtgeschnitten wurden. Die optimale Anzahl Aminosäuren wurde dabei ausgehend von der Längenverteilung mitochondrialer Transitsequenzen in Eukaryonten (mit Ausnahme von *Plasmodium*) bestimmt (Abbildung 7). Ein derartiges Vorgehen erschien angemessen, da das Ziel dieser Arbeit ja die Vorhersage der Anzahl *mitochondrialer* Transitsequenzen in *P. falciparum* war. Der Median 422 eukaryontischer mitochondrialer Transitsequenzen wurde (siehe oben) zu 31 Aminosäuren bestimmt, das 1. Quartil zu 24 und das 3.

Quartil zu 42 Aminosäuren (Mittelwert 34,4, Standardabweichung 5,2 Aminosäuren). Sämtliche Datensätze wurden diese Längen angepasst.
Diese Datensätze gleicher Sequenzlänge wurden zum Zweck der Datenreduktion mit ClustalW[16] ausgerichtet. Verwendet wurden dabei die Standardparameter. Nach dem Alignment wurde mit dem in ClustalW im European Bioinformatics Institute integrierten JalView[17] eine Redundanzreduktion durchgeführt. Dabei wurden alle Peptidsequenzen mit einer Aminosäureidentität über einem wählbaren Grenzwert aussortiert. Die Anzahl der verbleibenden Sequenzen wurde als Funktion der minimal nötigen Sequenzdifferenz aufgetragen (Abbildung 10).

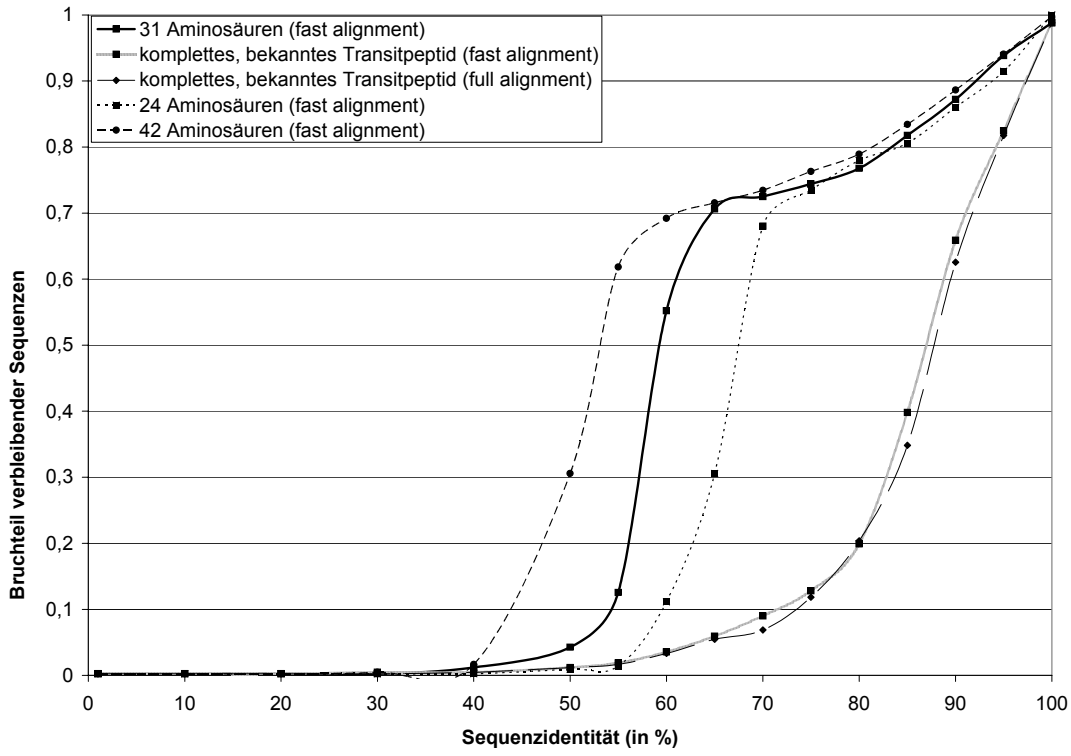

Abbildung 10 – Anzahl bei einer Redundanzreduktion verbleibender mitochondrialer N-terminaler Abschnitte eukaryontischer Proteinsequenzen. Dieses Diagramm liest sich von der rechten Seite beginnend – wenn die nötigen Sequenzunterschiede abnehmen, werden erst homologe Sequenzen, später dann nicht verwandte Peptidsequenzen aussortiert. Durch das vollständige Aussortieren homologer Sequenzen kommt die Plateauphase im rechten Drittel des Diagramms zustande.

Innerhalb der Sequenzen sämtlicher eukaryontischer Organismen mit Ausnahme von *Plasmodium*, siehe Abbildung 10, war eine deutliche Redundanz sichtbar – schon bei hohen erlaubten Identitätsanteilen nahm die Zahl verbleibender Sequenzen ab. Dies

liegt, wie oben erwähnt, an homologen Sequenzen, die zu diesem Zeitpunkt noch nicht ausgeschlossen waren. Die Sequenzen wurden jeweils auf eine Ähnlichkeit geringer als das Ende der Plateauphase reduziert. Damit ergaben sich die in Tabelle 1 genannten Datensatzgrößen. Ordinalzahlen geben jeweils eine Anzahl Sequenzen an, Prozentwerte die im Höchstfall erlaubte paarweise Sequenzidentität als Parameter für die Redundanzreduktion mit JalView.

Tabelle 1 – Anzahl erhaltener Sequenzen in den nicht-Plasmodium-Datensätzen nach Redundanzreduktion

	Mitochondriale Proteine (vorher: 422)		Cytoplasmatische Proteine (vorher: 729)		Extrazelluläre Proteine (vorher: 655)	
	Ähnlich-keit	Nachher	Ähnlich-keit	Nachher	Ähnlich-keit	Nachher
1. Quartil (24 AAs)	70%	282	55%	226	70%	215
2. Quartil (31 AAs)	65%	297	65%	369	65%	226
3. Quartil (42 AAs)	55%	258	55%	198	70%	261

Die in Abbildung 11 dargestellte Redundanzreduktion bei *Plasmodium* lieferte ein anderes Resultat. Hier wurde bei z.B. 61 cytoplasmischen Proteinen bis zu einer erlaubten Sequenzidentität von 50% nur eine einzige als redundant aus dem Datensatz entfernt.

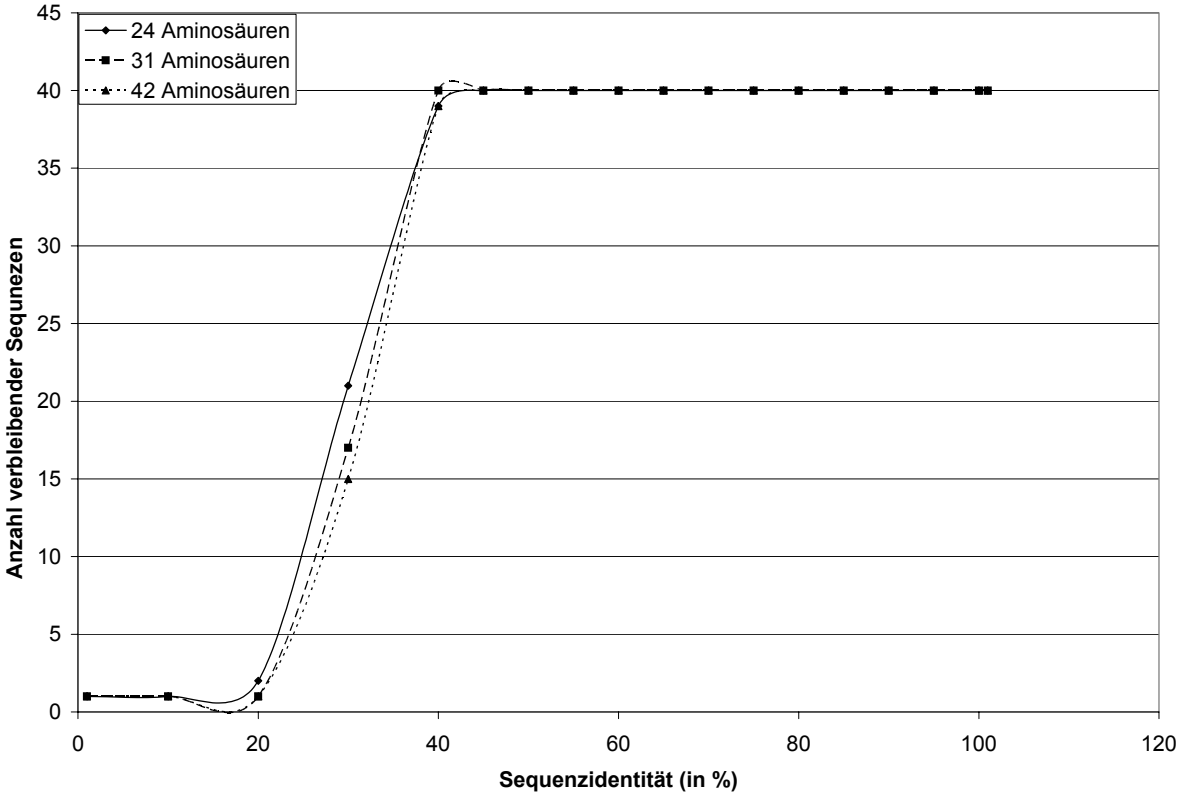

Abbildung 11 - Anzahl bei einer Redundanzreduktion verbleibender mitochondrialer N-terminaler Abschnitte eukaryontischer Proteinsequenzen. Die mitochondrialen Transitpeptide in Plasmodium sind sich untereinander eher unähnlich.

Der Grund folgt aus dem bei der Interpretation von Abbildung 10 genannten, da hier ja nur Sequenzen eines einzigen Eukaryonten vorlagen.

Bei den Proteindatensätzen aus *Plasmodium* wurden, da kaum paarweise Identitäten oberhalb eines Wertes von etwa 30% vorhanden waren, die kompletten Datensätze verwendet. Somit wurden die in Tabelle 2 angegebenen Datensatzgrößen für alle Sequenzlängen erhalten.

Tabelle 2 - Anzahl erhaltener Sequenzen an Plasmodium-Datensätzen ohne (nicht notwendige) Redundanzreduktion

Lokalisierung	Mitochondriale Proteine	Cytoplasmatische Proteine	Extrazelluläre Proteine	Apicoplast-Proteine
Anzahl	40	61	51	53

Bis zu diesem Zeitpunkt lagen die Aminosäuresequenzen in drei Datensätzen unterschiedlicher Länge vor. Als geeignete Eingangsdaten sowohl für die Hauptkomponentenanalyse und auch für die neuronalen Netze werden jedoch numerische Datenvektoren benötigt. Die Peptidsequenzen wurden in zweierlei Weise aufbereitet.

Einerseits wurde jedem 24, 31 und 42 Aminosäuren langem N-terminalen Abschnitt jeder Sequenz ein 20-dimensionaler Vektor seiner relativen Aminosäurehäufigkeiten zugeordnet. Andererseits wurde jeder Sequenz ein 19-dimensionaler Vektor voneinander unabhängiger, relativer Aminosäureeigenschaften zugeordnet. Die Darstellung im 19-dimensionale Raum physikochemischer Eigenschaften ist ihrerseits eine Hauptkomponentenanalyse[18] von 434 tabellierten Aminosäureeigenschaften[19]. Die erste Hauptkomponente dieses Raumes korelliert stark mit dem Hydrophobizitätsindex der Aminosäuren, Komponente zwei mit der Helixneigung, Komponente drei mit Volumen und Aminosäurehäufigkeit und Komponente vier mit der Polarität[20]. Durch die Orthogonalität der Achsen sollte diese Darstellung mit nur geringem Datenverlust eine Darstellung der Sequenzeigenschaften in einem möglichst niederdimensionalen Raum gestatten, was insbesondere für das Training neuronaler Netze von Vorteil sein sollte.

2.3. Angewandte Analysemethoden

2.3.1. Hauptkomponentenanalyse (Principal Component Analysis, PCA)

Die Hauptkomponentenanalyse ermöglicht die – im Regelfall fehlerbehaftete – Projektion hochdimensionaler Datenpunkte in einen Raum geringerer Dimension durch eine Neuwahl des Koordinatensystems[21]. Ausgenutzt wird dabei eine mögliche Redundanz - eine Korrelation - der in den Koordinaten enthaltenen Informationen.
In der Hauptkomponentenanalyse wird meistens sukzessive durch Rotation des Koordinatensystems die jeweils nächste Achse maximaler Varianz bestimmt (Varimax-Methode). Diese abgeleiteten Achsen maximaler Varianz, oder Hauptkomponenten, sind Linearkombinationen der ursprünglichen Variablen und stehen orthogonal zueinander. Die Repräsentation in diesem neuen Koordinatensystem erhält noch die ursprüngliche relative Lage der Datenpunkte zueinander, während die absoluten Koordinaten (jetzt: Faktorladungen) sich aufgrund des neuen Bezugssystems unterscheiden. Bei Repräsentation der ursprünglichen n-dimensionalen Datenpunkte im neuen, ebenfalls n-dimensionalen Raum bleiben alle ursprünglichen Informationen erhalten. Im neuen Koordinatensystem nimmt jedoch die in jeder neuen Hauptkomponente erklärte Varianz ab, so dass je nach Art Eingangsdaten die relative Lage der Datenpunkte zueinander schon in einem m-dimensionalen Raum (mit $m < n$) mit einem geringeren Fehler als in der ursprünglichen Darstellung abgebildet werden kann. In diesem Fall werden nur die ersten m Hauptkomponenten bei der Datenreproduktion berücksichtigt. Bei einer Darstellung der ersten zwei Hauptkomponenten in einem zweidimensionalen Koordinatensystem erreicht man so die mit linearen Abbildungen maximal mögliche Auftrennung der Datenpunkte.
An die Hauptkomponentenanalyse kann sich abermals eine Rotation der Koordinatenachsen anschließen, die diesmal nicht zur Reduktion der Dimensionalität der Daten, sondern einer erhöhten Aussagekraft der gewählten Achsen führen soll. Prinzipiell können diese Rotationsverfahren entweder ein orthogonales oder ein schiefwinkliges Koordinatensystem benutzen, wobei die Achsen im ersten Fall voneinander unabhängig bleiben, im zweiten Fall aber – prinzipiell – eine bessere Korrelation einer Varianz mit einer versteckten Variablen liefern könnten. In allen Fällen wird eine Funktion der Faktorladungen maximiert. Der bekannteste orthogonale Algorithmus ist Kaiser's Varimax-Methode[22], bei der die normalisierte Summe der Varianz der quadrierten Faktorladungen maximiert wird. Dies ist die oben bereits

angesprochene Methode, bei der die Achse maximaler Varianz ermittelt wird. Andere orthogonale Rotationen sind die Quartimax-, Equimax- und Parsimax- Rotationen, bei denen die Varianz der quadrierten Faktorladungen maximiert wird. Anders ausgedrückt, versucht der Varimax-Algorithmus die Faktorladungen möglichst nah an -1, 0 oder 1 zu bringen. Die Quartimax-Methode versucht eine hohe Ladung auf einem Faktor und möglicht niedrige auf allen anderen zu erreichen. Equimax versucht diese beiden Ziele gleichzeitig zu verwirklichen. Falls diese orthogonalen Rotationen keine befriedigende Interpretation der Achsen zulassen, kann noch eine schiefwinklige Transformation, z. B. die Promax- oder die Harris-Kaiser-Rotation durchgeführt werden. Darauf soll hier jedoch nicht weiter eingegangen werden.

2.3.2. Selbstorganisierende Kohonen-Karte (Self-organizing map, SOM)

2.3.2.1. Allgemeines

Die selbstorganisierende Kohonen-Karte[23,24] ist ein Spezialfall unüberwachter neuronaler Netze, deren Ausgabeneuronen in einer Matrix angeordnet sind (siehe Abbildung 12).

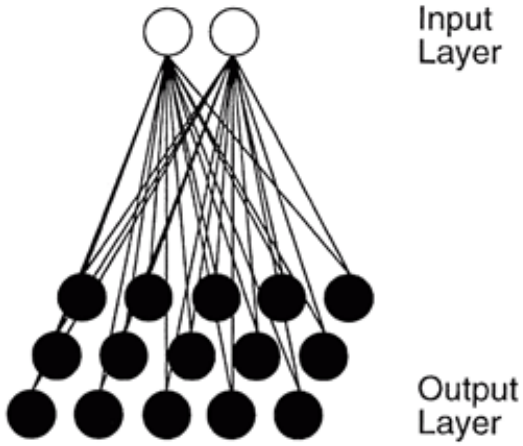

Abbildung 12 – Eine selbstorganisierende Kohonen-Karte mit in Matrixform angeordneten Ausgabeneuronen

Die Bezeichnung „unüberwachtes“ Lernen rührt daher, dass das Netz selber die Abbildung des Eingabevektors auf ein aktiviertes Outputneuron vornimmt. Im Gegensatz dazu wird im Falle des überwachten Lernens gleichzeitig ein Input- und Outputvektor vorgegeben und das Netz lernt durch eine Anpassung der Gewichte eine optimale Abbildungsfunktion. Daraus ergibt sich, dass im überwachten Lernen die Ausgabewerte in vorher zu definierende Kategorien einsortiert werden, während beim unüberwachten Lernen das Netz selbst eine Klassifikation „lernt“.

Die Anzahl der Inputneuronen und der zu jedem Outputneuron führenden Gewichtsfaktoren entspricht der Dimension der Inputdaten. Die Anzahl möglicher Datencluster entspricht der Anzahl Outputneuronen, wobei nicht in jedem Fall alle verfügbaren Outputneuronen genutzt werden und auch die Zuordnung intuitiv zu unterschiedlichen Klassen gehörender Daten auf ein und dasselbe Neuron vorkommen kann.

Kohonen-Netzwerke weichen von der „winner take all“ - Strategie anderer Netzwerke ab, bei der nur das jeweilige Gewinner-Neuron seine Gewichte adaptiert[25]. In Kohonen-

Netzwerken sind die Outputneuronen in einer niedrigdimensionalen Geometrie angeordnet (bei der hier verwendeten selbstorganisierenden Kohonen-Karte beispielsweise in zwei Dimensionen, in einer Fläche). Während des Trainings adaptieren neben dem Gewinnerneuron auch zu einem kleineren Anteil benachbarte Neuronen ihre Gewichtsvektoren, so dass die Aktivierung benachbarter Outputneuronen, zumindest im angestrebten Idealfall, durch einen ähnlichen Inputvektor hervorgerufen wird. Damit sind Kohonen-Karten ein sinnvolles Tool zur Datenklassifikation, „ähnliche“ bzw. in der jeweiligen Repräsentation für das Netzt ähnlich erscheinende Daten werden näher beieinander abgelegt als nur entfernter ähnliche Daten. Für jeden angelegten Inputvektor feuert jeweils nur ein Outputneuron, also wird mit jedem Inputvektor eine eindeutige Klassifikation vorgenommen. Um fehlerhafte Klassifikationen im Randbereich eines planaren Outputlayers zu minimieren, kann die Verwendung toroidaler Geometrie des Outputlayers sinnvoll sein[26]. Diese Option besaß das hier verwendete, vom Programm Statistica[27] generierte SOM jedoch nicht.
Details zu Kohonenkarten finden sich u.a. in Zupan/Gasteiger, „Neural Networks in Chemistry and Drug Discovery“, Wiley (1999).

2.3.2.2. Lernalgorithmus

Der Lernalgorithmus der Selbstorganisierenden Karte ist ein *kompetitives* Verfahren, es wird also zunächst nur ein sogenanntes „Gewinnerneuron“ selektiert, dessen Gewichtsvektor angepasst wird. (Erst in einem zweiten Schritt werden auch die Gewichte der Umgebungsneuronen korrigiert). Dazu wird entweder dasjenige Neuron bestimmt, dessen Gewichtsvektor dem des Inputvektors (gemäß Least-Square-Error) am ähnlichsten ist (Formel 1), oder dasjenige, das den maximalen Output produziert (gemäß Formel 2).

$$out_c \leftarrow \min\left(\sum_{i=1}^{m}(x_{si} - w_{ji})^2\right)$$

Formel 1 – Formel zur Ermittlung des Gewinnerneurons nach der Methode der Ähnlichkeit von Input- und Gewichtsvektor

$$out_c \leftarrow \max(out_j) = \max\left(\sum_{i=1}^{m} w_{ji} x_{si}\right)$$

Formel 2 – Formel zur Ermittlung des Gewinnerneurons nach der Methode des maximalen Outputs

Danach berechnet sich der neue Gewichtsvektor der Neuronen im ersten Fall nach

$$w_{ji}^{(new)} = w_{ji}^{(old)} + \eta(t)a(d_c - d_j)\left(x_i - w_{ji}^{(old)}\right)$$

Formel 3 – Berechnung der angepassten Gewichte im ersten Fall, bei der Ermittlung des Gewinnerneurons nach der Methode der Ähnlichkeit von Input- und Gewichtvektor

bzw., bei Auswahl des Gewinnerneurons nach der Methode des maximalen Outputs, nach

$$w_{ji}^{(new)} = w_{ji}^{(old)} + \eta(t)a(d_c - d_j)\left(1 - x_i w_{ji}^{(old)}\right)$$

Formel 4 – Berechnung der angepassten Gewichte im zweiten Fall, bei der Ermittlung des Gewinnerneurons nach der Methode des maximalen Outputs

Hier stellt der Parameter η die Lernrate dar, die eine schnelle oder weniger schnelle Änderung des jeweiligen Gewichtsvektors beeinflusst. Die Lernrate ist zumeist auch zeitabhängig, d.h. in den ersten Epochen findet eine schnellere Adaption als in späteren Epochen statt. Das Symbol $a(d_c\text{-}d_j)$ stellt eine Funktion dar, die von der Distanz zwischen dem Gewinnerneuron d_c und dem jeweiligen Neurons d_j abhängt, dessen Gewichte angepasst werden. Typische Funktionen sind eine Konstante, eine Dreiecksfunktion oder die Mexican-Hat-Funktion. In letzterem Fall werden nahe liegende Nachbarn stark in die gleiche Richtung wie das Gewinnerneuron adaptiert, etwas weiter entfernte Nachbarn in geringerem Ausmaß in die *Gegenrichtung*. Damit kann in manchen Fällen durch Kontrast-Verstärkung eine bessere Trennung verschiedener Datencluster erreicht werden. Hier wurde eine Stufenfunktion mit während des Trainings verkleinertem Nachbarschaftsgebiet verwendet.

2.3.3. Überwachtes Neuronales Netz (fully connected feed-forward ANN)

2.3.3.1. Allgemeines

Überwachte Neuronale Netze besitzen, im Gegensatz zu unüberwachten Netzen wie der oben erwähnten Kohonen-Karte, während des Trainings einen für jeden Trainingssatz festgelegten Inputvektor und einen zugehörigen Outputvektor. Sie approximieren beliebig komplizierte Input-Output-Funktionen mit Hilfe mehrer nichtlinearer Funktionen. Cybenko fand heraus, dass mit einem dreilagigen Feed-Forward-Netz wie in Abbildung 13, hyperbolischer Aktivierungsfunktion im Hidden Layer und linearer Aktivierungsfunktion im Output Layer jede beliebige kontinuierliche Funktion mit beliebiger Genauigkeit angenähert werden kann[28].
Während des Trainings werden die Gewichte der die Neuronenschichten verbindenden Gewichtsvektoren angepasst, um den zu einem Inputvektor gehörigen Outputvektor mit möglichst geringem Fehler wiederzugeben.

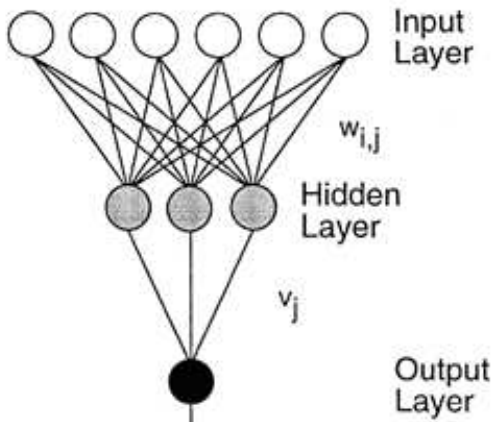

Abbildung 13 – Struktur eines überwachten neuronalen Netzes mit drei Neuronen im „Hidden Layer“

Das Inputlayer bei überwachten neuronalen Netzen ist aufgebaut wie bei unüberwachten Netzen. Zwischen Input- und Outputlayer findet sich hier jedoch eine versteckte („hidden layer“) Neuronenschicht mit (zumeist) sigmoidaler Aktivierungsfunktion. Der genaue Lernalgorithmus ist in Abschnitt 2.3.3.2. erläutert.

Bei überwachten neuronalen Netzen ist insbesondere die Anzahl Datenvektoren im Verhältnis zur Anzahl zu adjustierender Gewichte zu beachten, um kein unterbestimmtes System zu erhalten. Nach Andrea und Kalayeh[29] sollte dieses

Verhältnis mindestens zwei betragen, nach Manallack und Livingston[30] sollte es über drei liegen, um „Überlernen“ zu vermeiden. Als Überlernen bezeichnet man eine übermäßige Anpassung der Gewichte an die Trainingsdaten, die bei den Testdaten aus genau diesem Grund schlechtere Ergebnisse liefert.
Essentiell für eine erfolgreiche Anwendung überwachter Netze ist die Auswahl einer geeigneten Netztopologie, also einer geeigneten Anzahl Layer und Neuronen innerhalb der Layer.

In diesem Fall fand (Vorgriff auf Kapitel 3, Ergebnis) ein dreilagiges Perzeptron mit drei Neuronen im Hidden Layer Anwendung. Die Aktivierungsfunktionen waren von hyperbolischer Form

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

im Hidden Layer sowie von logistischer Form

$$f(x) = \frac{1}{1 + e^{x}}$$

im Output Layer. Als Lernalgorithmus für das optimierte Netz optimaler Topologie (siehe Ergebnisteil) wurde der Backpropagation-Algorithmus[31,32,33,34] verwendet, der im Folgenden beschrieben wird und zur Gruppe der überwachten Lernmethoden gehört.

2.3.3.2. Lernalgorithmus

Als Lernstrategie wurde der Backpropagation-Algorithmus[31,32,33,34] verwendet. Der Name rührt von einer Anwendung der Korrekturformeln vom letzten Layer, dem Output Layer, sequenziell rückwärts („back“) bis zum Input Layer des Netzes her. Dieses Vorgehen gehört zur Gruppe der Gradientensuchverfahren, bei denen die erste Ableitung der Fehlerfunktion für die optimale Anpassung der Gewichtsvektoren (in Richtung des maximalen Gefälles des Fehlers) verwendet wird. Ein Ablaufdiagramm zum Algorithmus ist in Abbildung 14 dargestellt.

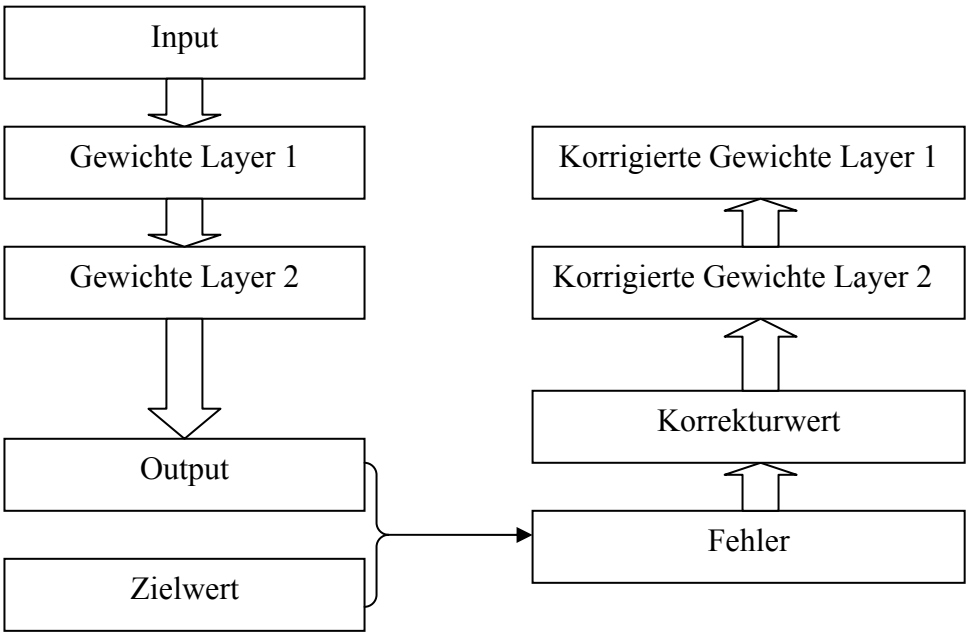

Abbildung 14 – Beim Backpropagation-Algorithmus werden die Gewichte der einzelnen Layer „von hinten nach vorne“ angepasst

Die Anpassung der Gewichte kann jeweils nach Präsentation jedes einzelnen Datensatzes (Batch-Verfahren) oder nach der Präsentation des gesamten Samples erfolgen, wobei in den meisten Fällen – so auch hier - die Gewichte nach jeder einzelnen Präsentation angepasst werden.
Der Backpropagation-Algorithmus benutzt die so genannte Delta-Regel der folgenden Form:

ΔParameter = Lernrate η * (Ist-Wert – Soll-Wert) * Inputwert.

Unter Benutzung der üblichen Symbolik stellt sie sich in folgender Form dar:

$$\Delta w_{ji}^{l} = \eta \delta_{j}^{l} out_{i}^{l-1} + \mu \Delta w_{ji}^{l(previous)}$$

Formel 5 - Die Delta-Regel unter Anwendung des beim Backpropagation-Algorithmus angewendeten Formalismus

Die Lernrate η ist ein Parameter, der die Geschwindigkeit der Adaption an die präsentierten Daten beeinflusst. Sein Gegenspieler ist das Momentum (oder der Impuls) μ, das angibt, wie stark der Wert der vorangegangenen Gewichtsadaption in die neue Berechnung einfließen soll. Die δ^{l}_{j} stellen die Abweichung des Ist-Ausgabewertes vom

Soll-Ausgabewert und damit den Fehler dar. Die mit out_i^{l-1} benannten Größen sind die Ausgaben des jeweils vorhergehenden Neuronenlayers, wobei ja der Output des Layers (l-1) den Input des Layers l darstellt. Δw_{ji}^{l} ist der anzupassende i-te Gewichtswert des Layers l von Neuron j.

Für den Fehler im Ausgabelayer ergibt sich folgende Formel

$$\delta_j^{last} = \left(y_j - out_j^{last}\right) out_j^{last} \left(1 - out_j^{last}\right)$$

Formel 6 – Formel zur Berechnung des Fehlers im Ausgabelayer eines Perzeptronen-Netzes

und für den Fehler in allen anderen Layern die Formel

$$\delta_j^l = \left(\sum_{k=1}^{r} \delta_k^{l+1} w_{kj}^{l+1}\right) out_j^l \left(1 - out_j^l\right)$$

Formel 7 – Formel zur Berechnung des Fehlers in anderen Layern eines Perzeptronen-Netzes

Substitution von Formel 7 in Formel 5 ergibt folgende allgemeine Formel zur direkten Berechnung der Gewichtskorrektur für jedes einzelne Neuron:

$$\Delta w_{ji}^l = \eta \left(\sum \delta_k^{l+1} w_{kj}^{l+1}\right) out_j^l \left(1 - out_j^l\right) out_i^{l-1} + \mu \Delta w_{ji}^{l(previous)}$$

Formel 8 - Allgemeine Formel zur Gewichtskorrektur beim Backpropagation-Algorithmus

2.4. Vorhersage von mitochondrialen Transitpeptiden mit etablierten Vorhersagemethoden

2.4.1. MitoProtII

Dieses Programm, MitoProtII[8], nimmt eine binäre Klassifikation in mitochondriale und andere Sequenzen vor, ist also mit dem hier entwickelten Ansatz vergleichbar. Allerdings kommt kein neuronales Netz zur Anwendung, sondern es werden 47 physikochemische Parameter, zumeist über die gesamte vorliegende Sequenz, berechnet und eine Trennungsgerade (bzw. hochdimensionale Fläche) zur Klassifikation errechnet. Zu den wichtigsten physikochemischen Trenngrößen zählen, in dieser Reihenfolge, die Aminosäurezusammensetzung im N-terminalen Abschnitt und die höchste Hydrophobizität – innerhalb eines 17-Aminosäure-Fensters -, die nach vier verschiedenen Skalen ermittelt wird.
Bei Verwendung menschlicher Aminosäuresequenzen wurde ein Mathews-Koeffizient von 0,46 bei einer Sensitivität von 0,92 und einer Spezifität von 0,27 erzielt[35].

2.4.2. TargetP

Das Tool TargetP[35] wendet, wie in dieser Arbeit, ebenfalls neuronale Netze auf die N-terminalen Abschnitte von Aminosäuresequenzen an. Es existieren in der Pflanzenversion separate Netze für die möglichen Lokalisierungen Cytoplasma, Mitochondrien, Chloroplast und Export aus der Zelle. In der für nicht-pflanzliche Lebewesen ausgelegten Version entfällt das Netz für den Chloroplasten. Die verschiedenen Netze berechnen für alle Aufenthaltsorte einen Score aus den ersten 100 N-terminalen Aminosäuren. Dabei wird jeder Aminosäure an einer bestimmten Position ein Punktwert zugewiesen. Das Netz, das im Ausgabeneuron die höchste Aktivierung besitzt, zeigt die wahrscheinlichste Lokalisierung des Peptids an.
Bei Verwendung menschlicher Sequenzen wurde ein Mathews-Koeffizient von 0,66 bei einer Sensitivität von 0,95 und einer Spezifität von 0,49 erzielt. Das auf Pflanzen spezialisierte Netz erzielte einen Mathews-Koeffizienten von 0,52 bei einer Sensitivität von 0,65 und einer Spezifität von 0,48[7].

3. Ergebnisse

Die Anwendbarkeit bekannter Algorithmen, MitoProtII und TargetP, zur Vorhersage mitochondrialer Transitpeptide in *P. falciparum* wurde mit Aminosäuresequenzen bekannter Lokalisation überprüft. Durch Vergleich der Aminosäurehäufigkeiten in mitochondrialen Transitpeptiden von *P. falciparum* und anderen Eukaryonten wurden Unterschiede beider Sequenzgruppen aufgezeigt. In zwei Repräsentationen, namentlich im 20-dimensionalen Aminosäurehäufigkeitsraum und im 19-dimensionalen physikochemischen Eigenschaftsraum, wurden die Hauptkomponentenanalyse und Selbstorganisierende Karten zur Datenvisualisierung angewandt. Mit Hilfe von Aminosäuresequenzen bekannter Lokalisation wurde die Netztopologie des innerhalb der Rahmenbedingungen optimalen dreilagigen Perzeptronennetzes zur Lokalisationsvorhersage ermittelt. Dieses neuronale Netz wurde auf die vorhergesagten Open Reading Frames des Genoms von *P. falciparum* angewandt.

3.1. Vorhersage mitochondrialer Transitpeptide mit etablierten Methoden

3.1.1. MitoProtII

175 Aminosäuresequenzen bekannter Lokalisation (40 mitochondriale, 135 nicht-mitochondriale Sequenzen) wurden mit MitoProtII, Version1.0a4[36] analysiert. Das Programm gab für jede Sequenz deren Wahrscheinlichkeit an, ein mitochondriales Transitpeptid zu beinhalten. Bei einer Wahrscheinlichkeit größer oder gleich 0,5 wurde die Vorhersage als positiv aufgefasst, bei einer Wahrscheinlichkeit kleiner 0,5 als negativ.

Von den 40 Proteinen mit mitochondrialem Transitpeptid wurden 32 als korrekt positiv (P) erkannt, acht der Sequenzen wurden als falsch-negativ (U) eingeteilt. Von den 135 negativen Sequenzen wurden 99 als korrekt negativ erkannt (N), 36 wurden als falsch positiv (O) klassifiziert. Damit ergibt sich ein Matthews Koeffizient[1] von 0,49.

3.1.2. TargetP

175 Aminosäuresequenzen bekannter Lokalisation (40 mitochondriale, 135 nicht-mitochondriale Sequenzen) wurden in TargetP[37] eingegeben und im ersten Durchgang die Optionen „Plant“ sowie „no cutoffs“ (um jeder Sequenz eine Lokalisation zuzuweisen) gewählt. Von den 40 Proteinen mit mitochondrialem Transitpeptid werden 22 als korrekt positiv klassifiziert (P), 18 werden als falsch-negativ (U) eingeteilt. Von

den 135 negativen Sequenzen werden 130 als korrekt negativ erkannt (N), 5 werden falsch-positiv klassifiziert (O). Damit ergibt sich ein Mathews-Koeffizient von 0,60. Anschließend wurden die Sequenzen mit der Option „non Plant“ und „no cutoffs“ analysiert. Hier wurden von den 40 Sequenzen mit vermutlich mitochondrialem Transitpeptid 14 korrekt als positiv erkannt (P) und von den 135 andernorts lokalisierten Proteinen 130 als korrekt negativ (N) eingestuft. 26 Sequenzen wurden falsch negativ (U) und 5 als falsch positiv (O) klassifiziert. Dies ergibt einen Mathews-Koeffizienten von 0,42. Zu berücksichtigen ist, dass dieses Vorhersageprogramm vier Lokalisierungen vorhersagen kann, hier jedoch nur die Binärunterscheidung in mitochondriale / andere Proteine verwendet wurde.

Es lässt sich feststellen, dass beide betrachtete Vorhersagesysteme deutliche Schwächen in der Klassifikation aufweisen. MitoProtII ist wenig selektiv, klassifiziert also viele negative Sequenzen als falsch-positiv. TargetP dagegen ist wenig sensitiv, erkennt also viele der positiven Sequenzen nicht und klassifiziert sie als falsch-negativ. Somit ist die Entwicklung eines auf *P. falciparum* trainierten Vorhersageprogramms nötig. Dies soll in der vorliegenden Arbeit geschehen.

3.2. Aminosäurehäufigkeiten

3.2.1. Vergleich *P. falciparum* und SwissProt, Version 36

Es wurden die Aminosäurehäufigkeiten in den codierenden Regionen des Genoms von *P. falciparum* mit den Aminosäurehäufigkeiten der SwissProt verglichen. Die Daten aus *P. falciparum* wurden aus dem Codon Usage Table der PlasmoDB[38]-Datenbank errechnet, der annotierte codierende Regionen von Chromosom 2 und 3 erhielt. Als Vergleichswert diente die Aminosäurehäufigkeit in der SwissProt-Datenbank in Version 36. Mit Hilfe der Abbildung 15 sollen also generelle Unterschiede in der Aminosäurebenutzung zwischen *P. falciparum* und anderen Organismen aufgezeigt werden. Die SwissProt-Datenbank wurde als ein über alle bekannte Genome gemittelter Referenzwert ausgewählt.

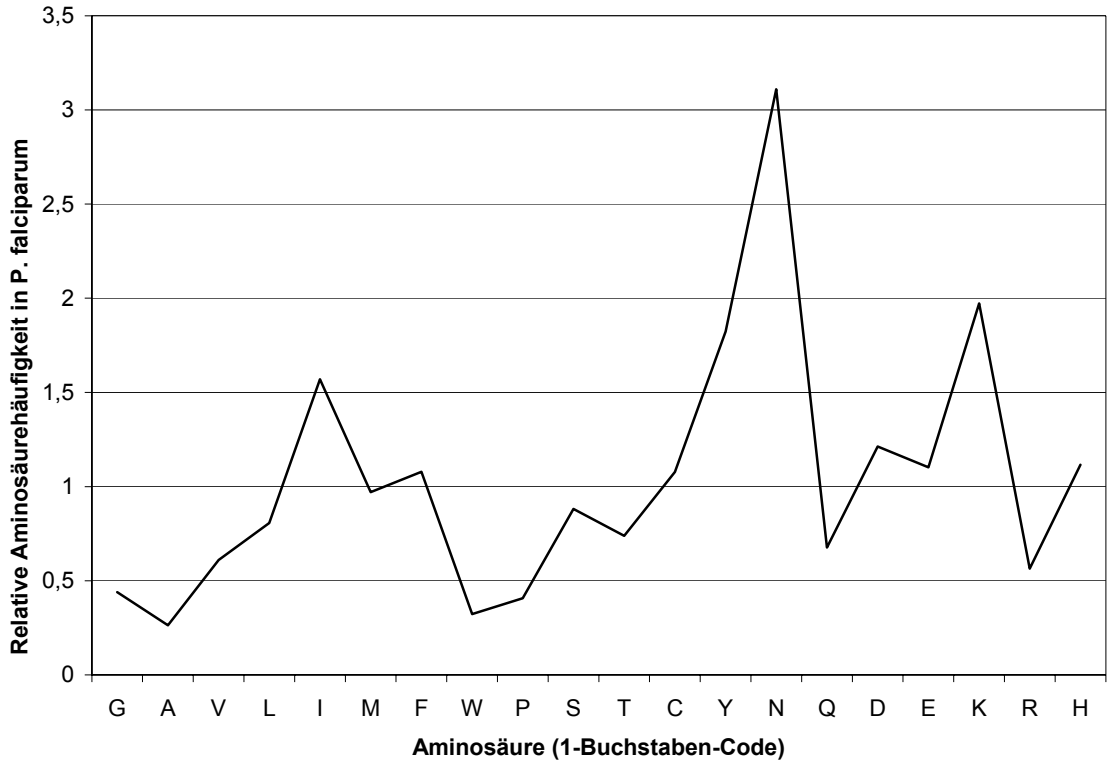

Abbildung 15 – Aminosäurehäufigkeitsverteilung in *P. falciparum*, relativ zu Aminosäurehäufigkeiten in der SwissProt, Version 36. Werte über 1 zeigen einen häufigeren Einbau der Aminosäure in *P. falciparum* als im durchschnittlichen Protein der SwissProt Version 36 an, Werte unter 1 eine weniger häufige Benutzung.

Glycin, Alanin, Prolin und Arginin werden wesentlich seltener (Quotient < 0,6) als in der SwissProt verwendet, Isoleucin, Tyrosin, Asparagin und Lysin wesentlich häufiger (Quotient > 1,5) als in der Referenzdatenbank.

3.2.2. Einzelne Aminosäuren in N-terminalen Abschnitten

In den Abbildungen 16-22 sind die relativen Aminosäurehäufigkeiten der 20 proteinogenen Aminosäuren dargestellt. Als Referenz wurde in jeder Grafik die relative Aminosäurehäufigkeit der Swiss-Prot-Datenbank in der Version 36 als durchgezogene Linie dargestellt. In Abbildung 16 ist die Aminosäurehäufigkeitsverteilung der ersten 24 N-terminalen Aminosäuren mitochondrialer Proteine aus Eukaryonten mit Ausnahme von *P. falciparum* dargestellt. Im Vergleich dazu ist die durchschnittliche Aminosäurehäufigkeit der SwissProt, Version 36, dargestellt. In den Diagrammen sind Median, 25% und 75% Quartile (Box), Non-Outlier-Range (innerhalb 75% Quartil +/- 1,5 · Box; Whiskers), Outliers (innerhalb 75% Quartil +/- 3 · Box; Kreise) und Extremwerte (darüber und darunter; Sternchen) aufgetragen.
Primäre physikochemische Eigenschaften wurden entsprechend der folgenden Tabelle zugeordnet.

Tabelle 3 - Den Aminosäuren zugeordnete primäre physikochemische Eigenschaften. Diese Einteilung dient lediglich einer groben Klassifizierung.

Eigenschaft	Klein	Hydrophob	Negative Ladung	Positive Ladung	Polar
Aminosäuren	A,G,P,S,T	C,L,I,V,M,W,F,Y	E,D	K,R,H	N,Q

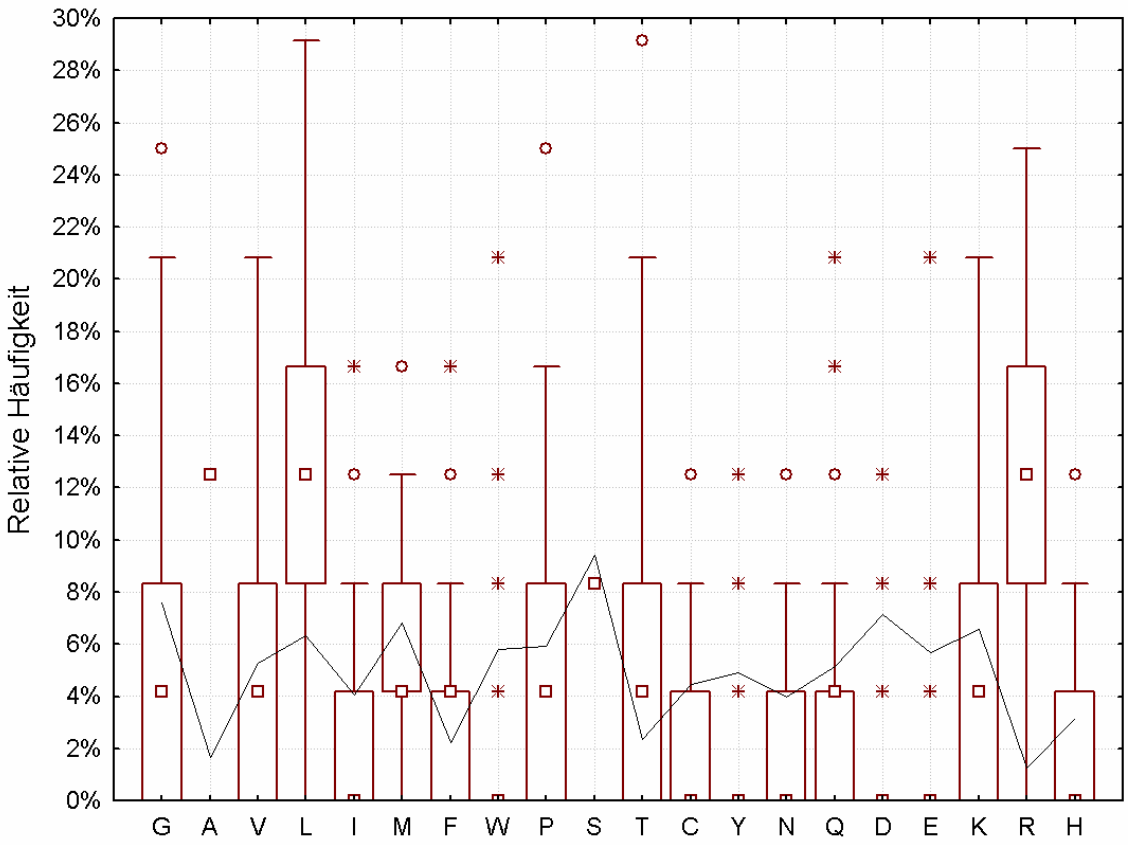

Abbildung 16 – Aminosäurehäufigkeitsverteilung mitochondrialer Transitpeptide in Eukaryonten außer *P. falciparum*. Zum Vergleich ist als durchgezogene Linie die Verteilung der SwissProt V36 angegeben.

Im Vergleich zur Swiss-Prot-Datenbank ist ein deutlich höherer Anteil der hydrophoben Aminosäuren Alanin (13% vs. 2%) und Leucin (13% vs. 6%) festzustellen. (Alle Zahlenangaben im Text sind Mittelwerte, da keine Median- und Quartilwerte für die SwissProt-Datenbank vorlagen. Der Referenzwert der SwissProt Version 36 steht in den folgenden Vergleichen stets an zweiter Stelle.). Ebenso liegt der Anteil an Arginin deutlich über dem Durchschnitt der SwissProt (12% vs. 1,2%). Isoleucin (3% vs. 4%), Tryptophan (1,1% vs. 5%) und negativ geladene Aminosäuren sind dagegen durchweg seltener als in der Referenzdatenbank vertreten. Dies ist in Übereinstimmung mit früheren Untersuchungen[7] . In Abbildung 17 ist die Aminosäurehäufigkeitsverteilung der ersten 24 N-terminalen Aminosäuren cytoplasmatischer Proteine aus Eukaryonten mit Ausnahme von *P. falciparum* dargestellt.

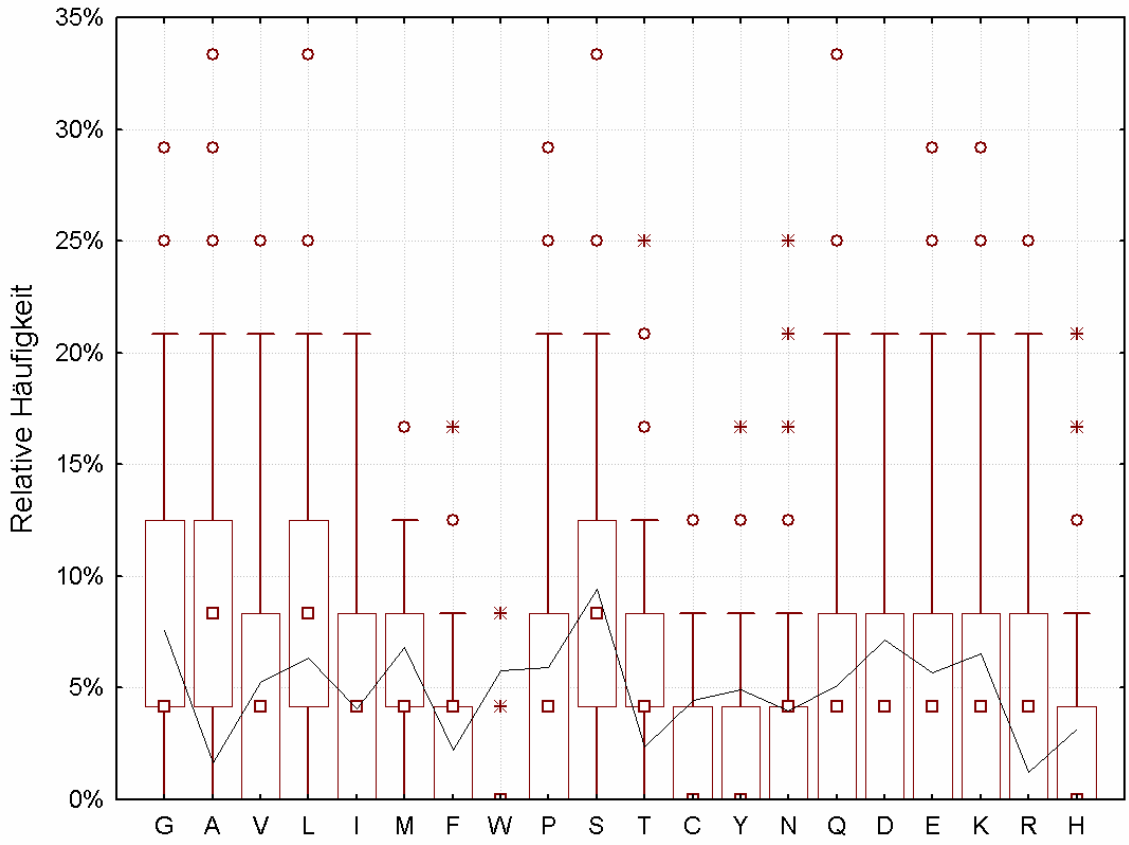

Abbildung 17 – Aminosäurehäufigkeitsverteilung cytoplasmatischer Proteine in Eukaryonten mit Ausnahme von *P. falciparum*. Diese Proteine weisen in etwa die Aminosäurehäufigkeitsverteilung der als durchgezogene Linie dargestellten SwissProt V36 auf.

Die Aminosäurehäufigkeitsverteilung cytoplasmatischer Proteine liegt recht nah am Durchschnittswert der SwissProt. Auffällige Abweichungen sind lediglich der höhere Gehalt an Alanin (8% vs. 2%) und der geringere Gehalt an Cystein (1,4% vs. 4%), Tryptophan (0,9% vs. 6%) und Tyrosin (2% vs. 5%).
In Abbildung 18 folgt die analoge Darstellung für extrazelluläre Proteine.

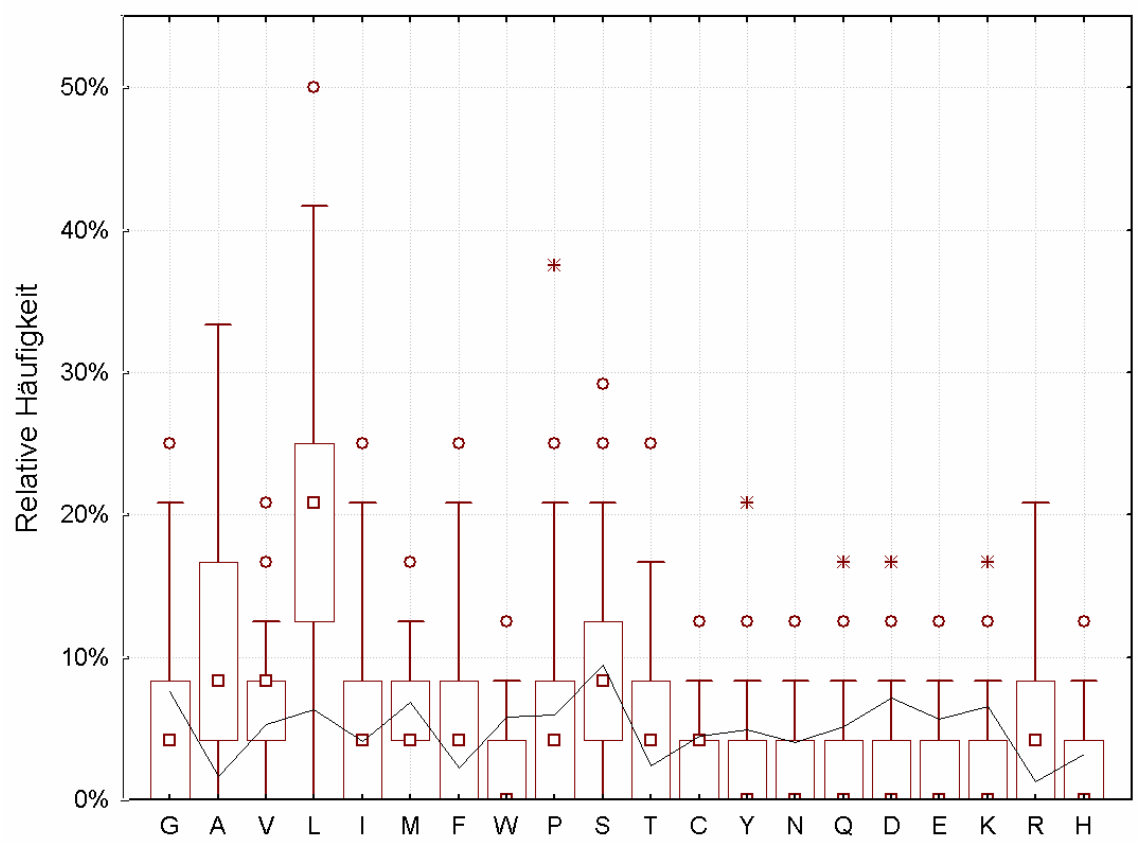

Abbildung 18 – Aminosäurehäufigkeitsverteilung extrazellulärer Proteine in Eukaryonten mit Ausnahme von *P. falciparum.* Extrazelluläre Proteine besitzen einen deutlich größeren Anteil hydrophober Aminosäuren als die Referenzdatenbank SwissProt V36 (durchgezogene Linie).

Extrazelluläre Proteine enthalten – ähnlich wie in mitochondrialen Proteinen – in den ersten 24 Aminosäuren ebenfalls deutlich mehr der hydrophoben Aminosäuren Alanin (11% vs. 2%) und Leucin (20% vs. 6%) als die SwissProt-Datenbank in Version 36. Gleichzeitig liegt der Anteil an den meisten (mit Ausnahme von Serin und Threonin) polaren und allen negativ geladenen Aminosäuren unter dem der SwissProt. Diese Merkmale sind charakteristisch für Signalpeptide[6], die einen gewünschten Export aus der Zelle signalisieren.

In den folgenden Diagrammen ist der analoge Sachverhalt in den Datensätzen aus *Plasmodium falciparum* dargestellt. Abbildung 19 zeigt die relative Aminosäurehäufigkeitsverteilung innerhalb der ersten 24 Aminosäuren in mitochondrialen Transitpeptiden von *Plasmodium falciparum*.

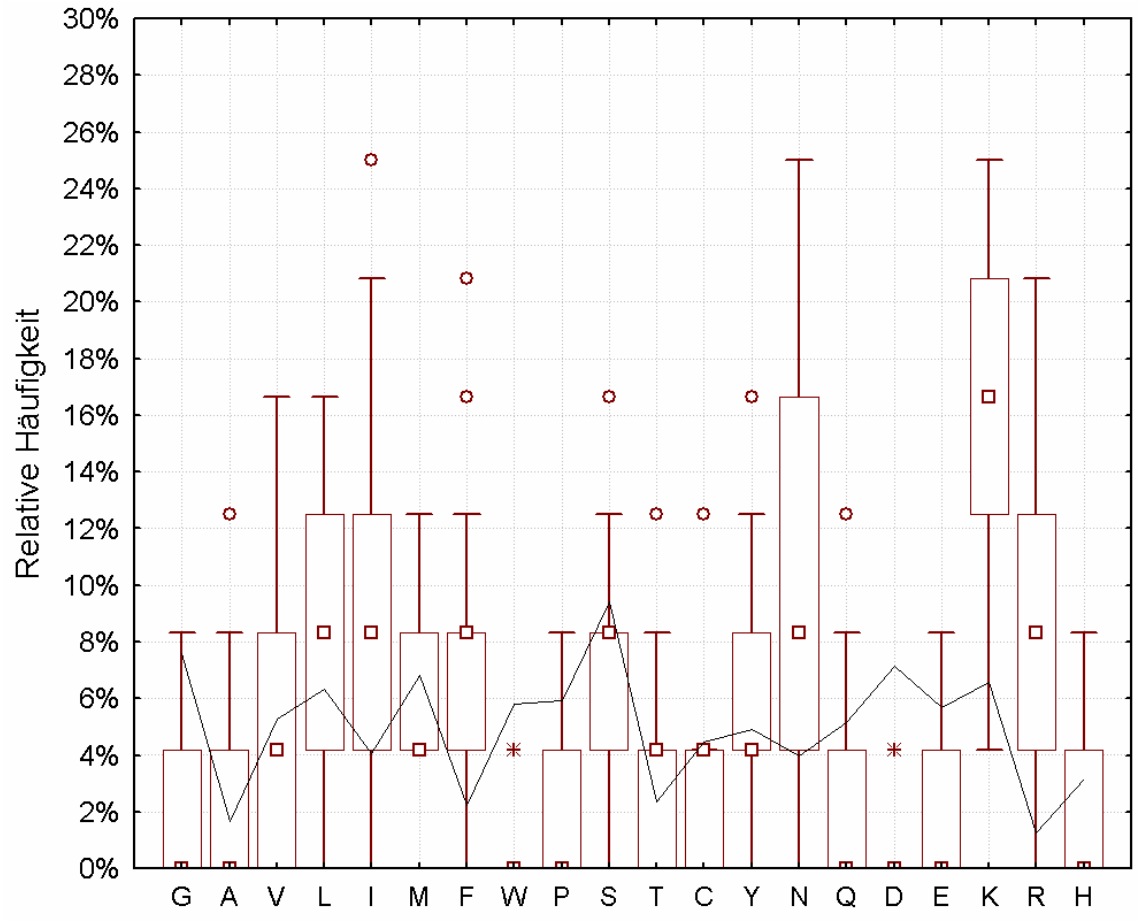

Abbildung 19 – Aminosäurehäufigkeitsverteilung mitochondrialer Proteine in *P. falciparum*. Mitochondriale N-terminale Abschnitte zeigen die für typische Häufung positiv geladener Aminosäuren. Als Vergleich ist die Verteilung des SwissProt V36 (durchgezogene Linie) dargestellt.

Die mitochondrialen Transitpeptide weisen zwar, wie bei anderen Eukaryonten, auch hier einen höheren Anteil der positiv geladenen Aminosäuren auf – allerdings bevorzugt *P. falciparum* Lysin gegenüber Arginin deutlich (16% gegenüber 8%, verglichen mit 4% zu 12% in anderen Eukaryonten). Dieses Ergebnis ist in Übereinstimmung mit der generellen Präferenz von *P. falciparum* für Lysin, siehe Abbildung 15. Hydrophobe Aminosäuren werden etwas weniger häufig, Asparagin als polare Aminosäure häufiger (10% vs. 2%) als in anderen Eukaryonten eingebaut. Auch die häufige Verwendung von Asparagin trifft für das Gesamtgenom von *P. falciparum* zu. Tyrosin wird mitochondrialen Transitpeptiden nur etwa so häufig wie in der SwissProt-Datenbank verwendet, obwohl es im Gesamtgenom von *P. falciparum* etwa doppelt so häufig vorhanden ist In Abbildung 20 ist die entsprechende Verteilung für cytoplasmatische Proteine wiedergegeben.

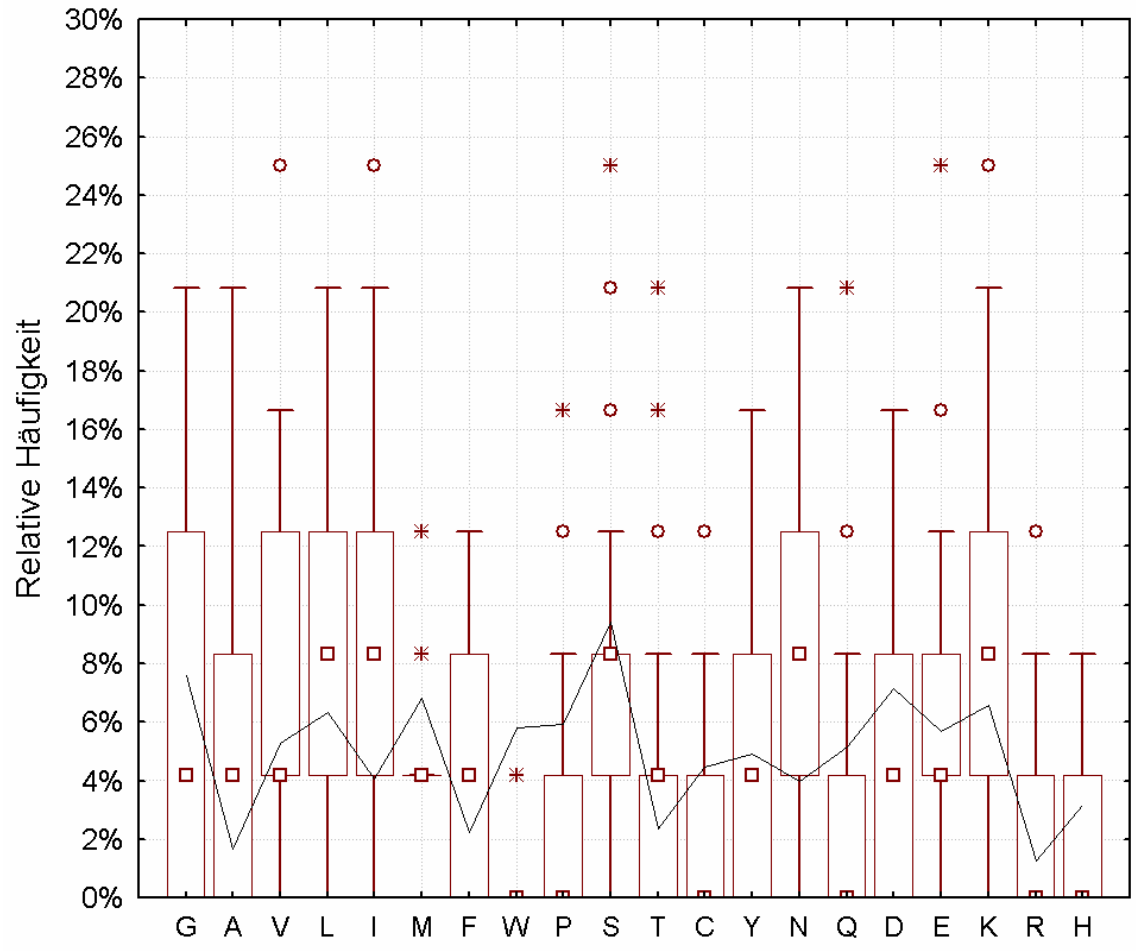

Abbildung 20 – Aminosäurehäufigkeitsverteilung cytoplasmatischer Proteine in *P. falciparum*. In *P. falciparum* orientiert sich, wie auch in anderen Eukaryonten, die Aminosäurehäufigkeitsverteilung cytoplasmatischer Proteine an derjenigen der SwissProt V36 (durchgezogene Linie).

Die Aminosäurenkomposition cytoplasmatischer Proteine in *P. falciparum* orientiert sich ungefähr an der Aminosäureverteilung der SwissProt. Dies entspricht tendenziell dem Sachverhalt in anderen Eukaryonten. Auch die bei anderen Eukaryonten sichtbare weniger häufige Verwendung von Tryptophan und Cystein, von Prolin und Glutamin findet sich hier wieder. Allerdings sind insgesamt stärkere Abweichungen von der Referenz als bei anderen Eukaryonten zu erkennen. Tyrosin wird auch hier nur mit der gleichen relativen Häufigkeit wie in der SwissProt angefunden, obschon es im Gesamtgenom etwa doppelt so häufig erwartet wird (Abbildung 15). In Abbildung 21 ist die Aminosäurehäufigkeitsverteilung für extrazelluläre Proteine von *P. falciparum* wiedergegeben.

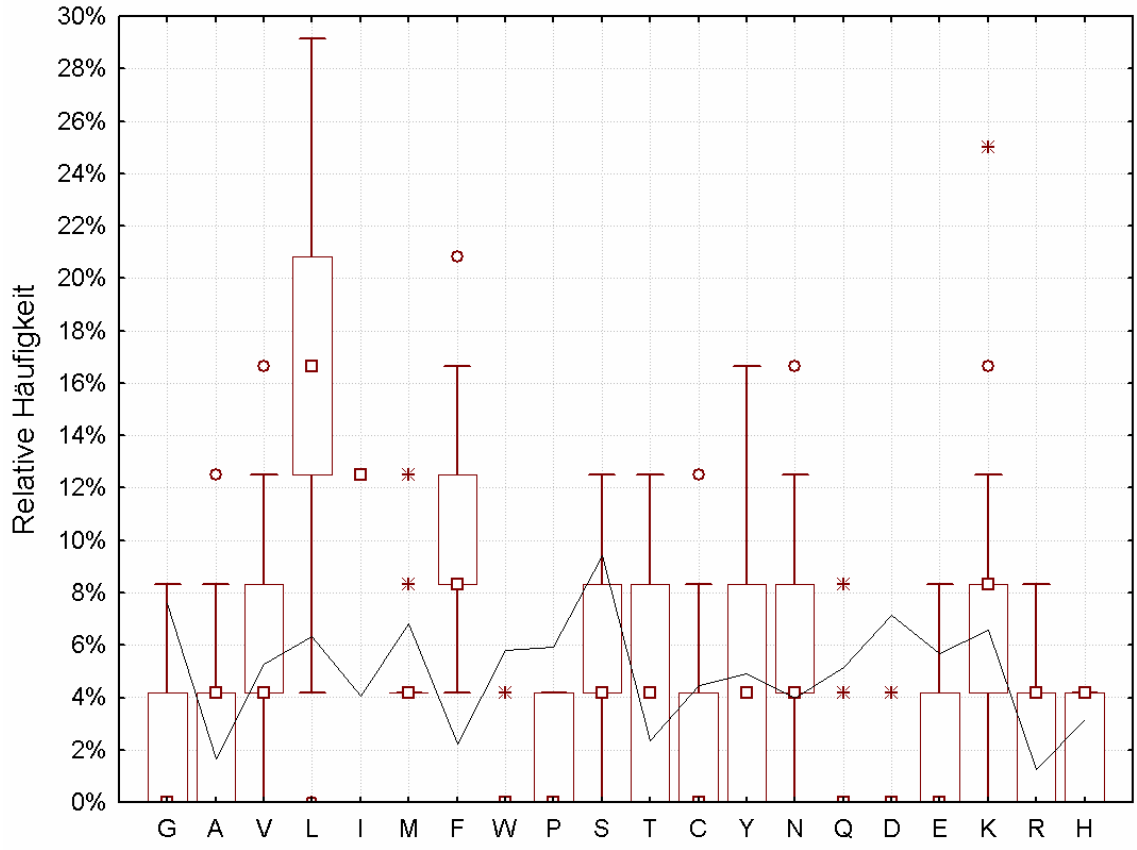

Abbildung 21 – Aminosäurehäufigkeitsverteilung extrazellulärer Proteine in *P. falciparum.* Ebenso wie andere Eukaryonten verwendet *P. falciparum* bei extrazellulären Signalpeptiden bevorzugt hydrophobe und wenig negativ geladene Aminosäuren.

Plasmodium lässt eine ähnliche Tendenz wie andere Eukaryonten gegenüber der SwissProt erkennen – der Gebrauch hydrophober Aminosäuren ist in beiden Fällen erhöht. Allerdings bevorzugen andere Eukaryonten Alanin (11% vs. 3% in *P. falciparum*) und Leucin (20% vs. 15%) als hydrophobe Residuen.
Im Falle von *P. falciparum* wird Isoleucin (13% vs. 5% in anderen Eukaryonten) und Phenylalanin (12% vs. 5% in anderen Eukaryonten) häufiger eingebaut. Negativ geladene Residuen finden sich ebenso selten wie bei anderen Eukaryonten, in beiden Datensätzen liegt die Häufigkeit unterhalb desjenigen in der Referenzdatenbank. Auffällig ist die seltene Verwendung der polaren Aminosäure Asparagin, die im Gesamtgenom etwa dreimal so häufig wie in der SwissProt erwartet wird, hier im Signalpeptid aber nicht die gewünschten hydrophoben Eigenschaften besitzt. In Abbildung 22 ist das analoge Diagramm für Proteine wiedergegeben, die für den Apicoplasten bestimmt sind und somit sowohl ein N-terminales Signalpeptid als auch ein darauf folgendes Transitpeptid besitzen[39].

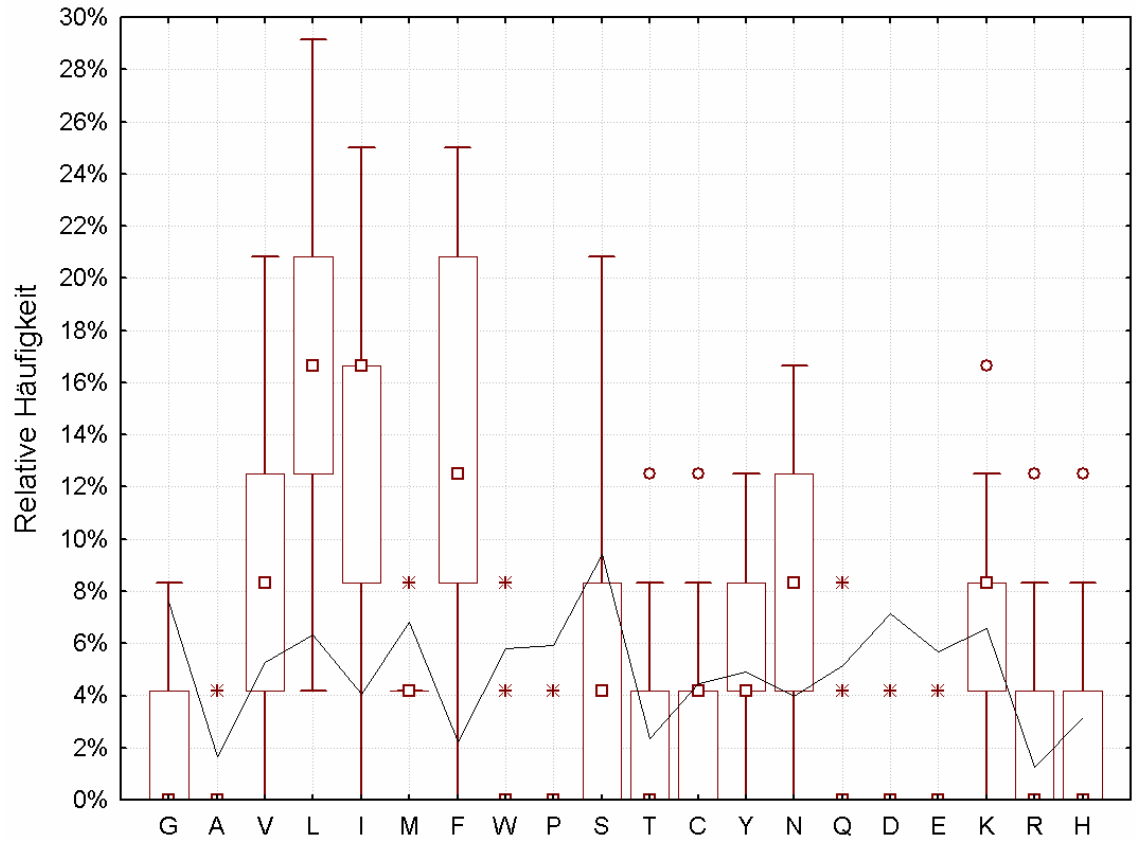

Abbildung 22 – Aminosäurehäufigkeitsverteilung apicoplastischer Proteine in *P. falciparum*. Apicoplasten - Targetingsequenzen verwenden ebenfalls häufig hydrophobe und selten negativ geladene Aminosäuren.

Apicoplasten - Targetingsequenzen weichen ebenfalls durch verstärkten Einbau hydrophober Aminosäuren von der SwissProt ab – dies ähnelt dem Fall in mitochondrialen Transit- und extrazellulären Signalpeptiden in beiden Datensätzen. Leucin und Isoleucin werden häufiger als in der SwissProt-Datenbank und auch häufiger als im Gesamtgenom von *P. falciparum* eingebaut. Dies ist verständlich, da das Targetingsignal für den Apicoplasten aus einem N-terminalen Signalpeptid und einem darauf folgenden Transitpeptid besteht. Ebenfalls werden weniger negativ geladene Bausteine, Glutamat und Aspartylsäure, eingebaut. Allerdings liegt die Verwendung positiv geladener Aminosäuren hier nur im Schnitt der SwissProt – dies ist ein Unterschied zu mitochondrialen und extrazellulären Sequenzen. Lysin wird gegenüber Arginin als positiv geladene Aminosäure bevorzugt, dies ist in Übereinstimmung mit der Bevorzugung dieser Aminosäure im Gesamtgenom von *P. falciparum*.

3.2.3. Aminosäureklassen

Um einen besseren Eindruck der zusammengestellten Sequenzdaten zu erhalten, wurden deren relative Aminosäurehäufigkeiten der folgenden Tabelle gemäß gruppiert und grafisch dargestellt.

Tabelle 4 - Gruppierung der Aminosäuren nach ähnlichen physikochemischen Eigenschaften

Eigenschaft	Klein	Hydrophob	Negative Ladung	Positive Ladung	Polar
Aminosäuren	A,G,P,S,T	C,L,I,V,M,W,F,Y	E,D	K,R,H	N,Q

In Abbildung 23 findet sich eine Übersicht über die Aminosäurehäufigkeiten von Proteinen in verschiedenen Kompartimenten, zusätzlich unterteilt nach solchen aus *P. falciparum* und solchen aus anderen Eukaryonten.

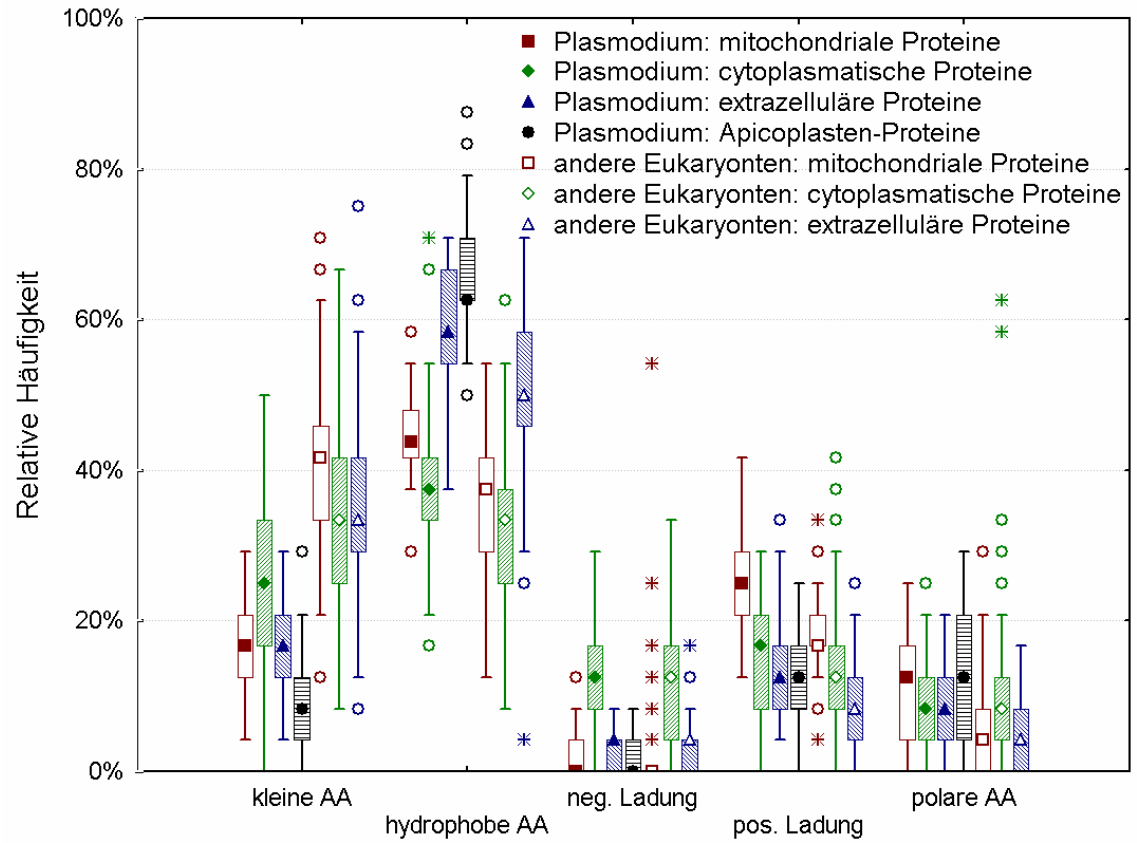

Abbildung 23 - Relative Aminosäurehäufigkeiten nach Aminosäuregruppen in *P. falciparum* und anderen Eukaryonten

Im Vergleich von Proteinen aus *P. falciparum* und anderen Eukaryonten fällt auf, dass *Plasmodium*

- allgemein einen geringeren Anteil „kleiner“ Aminosäuren benutzt
- in mitochondrialen Proteinen mehr „positive“ Aminosäuren einbaut
- Apicoplasten-Proteine mit Abstand den größten Anteil an hydrophoben Aminosäuren besitzen.

Zur besseren Veranschaulichung werden im Folgenden Untergruppen der Daten aus Abbildung 23 beschrieben. In Abbildung 24 sind die relativen Häufigkeiten der Aminosäuregruppen von Proteinen aus *P. falciparum* dargestellt. Sämtliche angegebenen Aminosäurehäufigkeiten beziehen sich auf die 24 N-terminalen Aminosäuren.

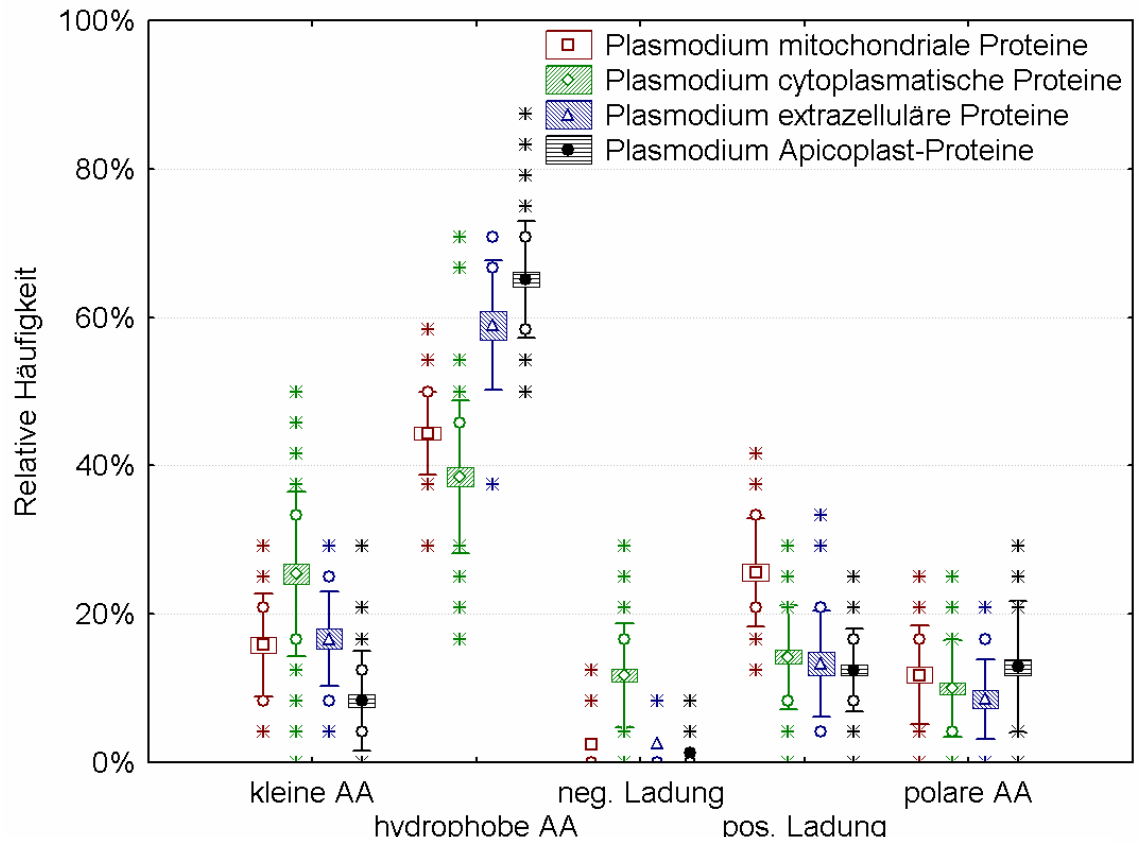

Abbildung 24 - Relative Aminosäurehäufigkeiten innerhalb der ersten 24 N-terminalen Aminosäuren nach Gruppen in P. falciparum

In *P. falciparum* fällt auf, dass

- mitochondriale Transitpeptide typischerweise wenig negativ und viele positiv geladen Aminosäuren benutzen (das entspricht dem aus anderen Eukaryonten bekannten Fall)
- Apicoplasten-Proteine als herausragendes Merkmal einen Anteil von im Schnitt 2/3 hydrophoben Aminosäuren besitzen, extrazelluläre Proteine besitzen einen nahezu ebensogroßen Anteil an hydrophoben Bausteinen
- cytoplasmatische Proteine als einzige einen nennenswerten Anteil negativer Aminosäuren besitzen.

Analog sind in Abbildung 25 die relativen Aminosäurehäufigkeiten aus anderen Eukaryonten dargestellt.

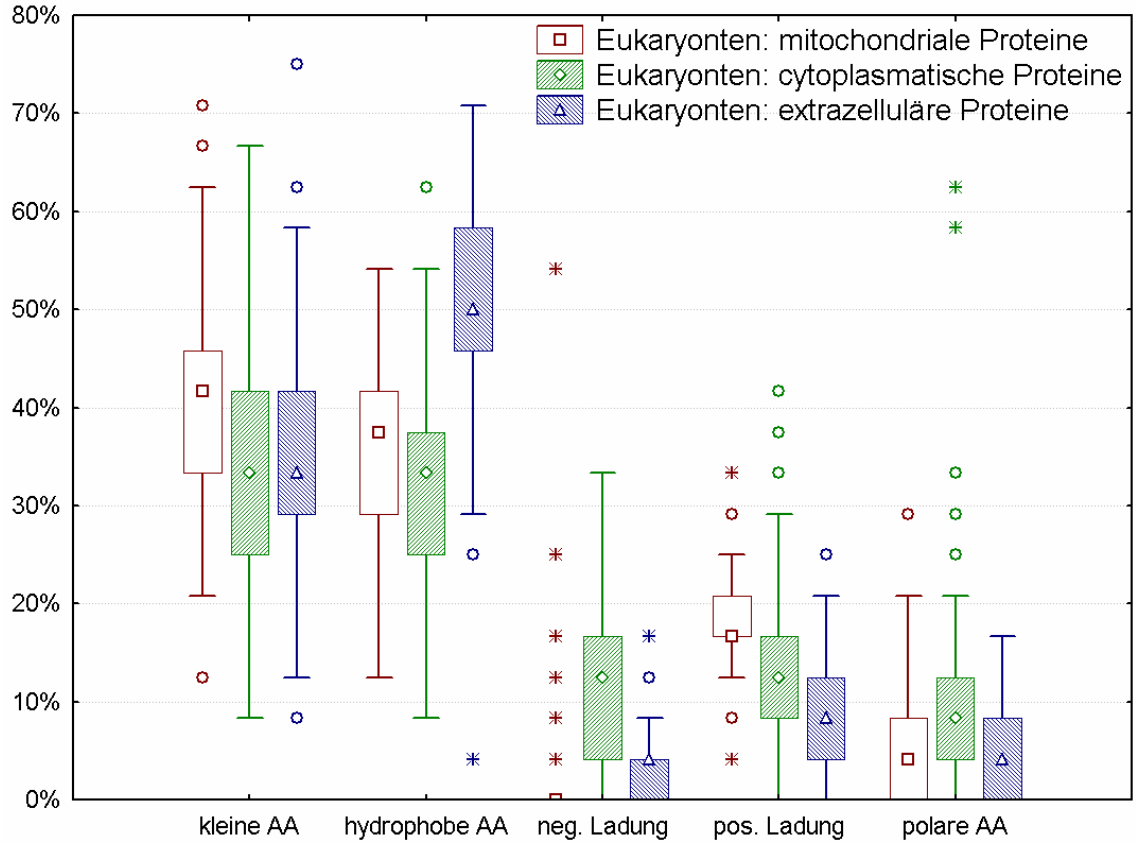

Abbildung 25 - Relative Aminosäurehäufigkeiten innerhalb der ersten 24 N-terminalen Aminosäuren nach Gruppen in anderen Eukaryonten

In Abbildung 25 fällt auf, dass

- mitochondriale und extrazelluläre Sequenzen weniger negativ geladene und polare Aminosäuren als cytoplasmatische Proteine besitzen,
- sie mehr hydrophobe und kleine Aminosäuren als cytoplasmatische Proteine besitzen
- Mitochondriale mehr und extrazelluläre Proteine weniger positiv geladene Aminosäuren als das cytoplasmatische Durchschnittsprotein besitzen.

Da hier ein Algorithmus zur Vorhersage von kerncodierten, in die Mitochondrien transportierten Proteinen entwickelt werden sollte, ist im folgenden noch einmal der Vergleich von solchen Proteinen in *P. falciparum* und anderen Eukaryonten aufgetragen. Durch die in Abbildung 26 wiedergegebenen Differenzen zwischen *Plasmodium*-Transitpeptiden und denen aus anderen Organismen erklärt sich auch, dass

für andere Eukaryonten entwickelte Programme zur Lokalisierungsvorhersage im Falle von *P. falciparum* schlechtere Ergebnisse erzielen.

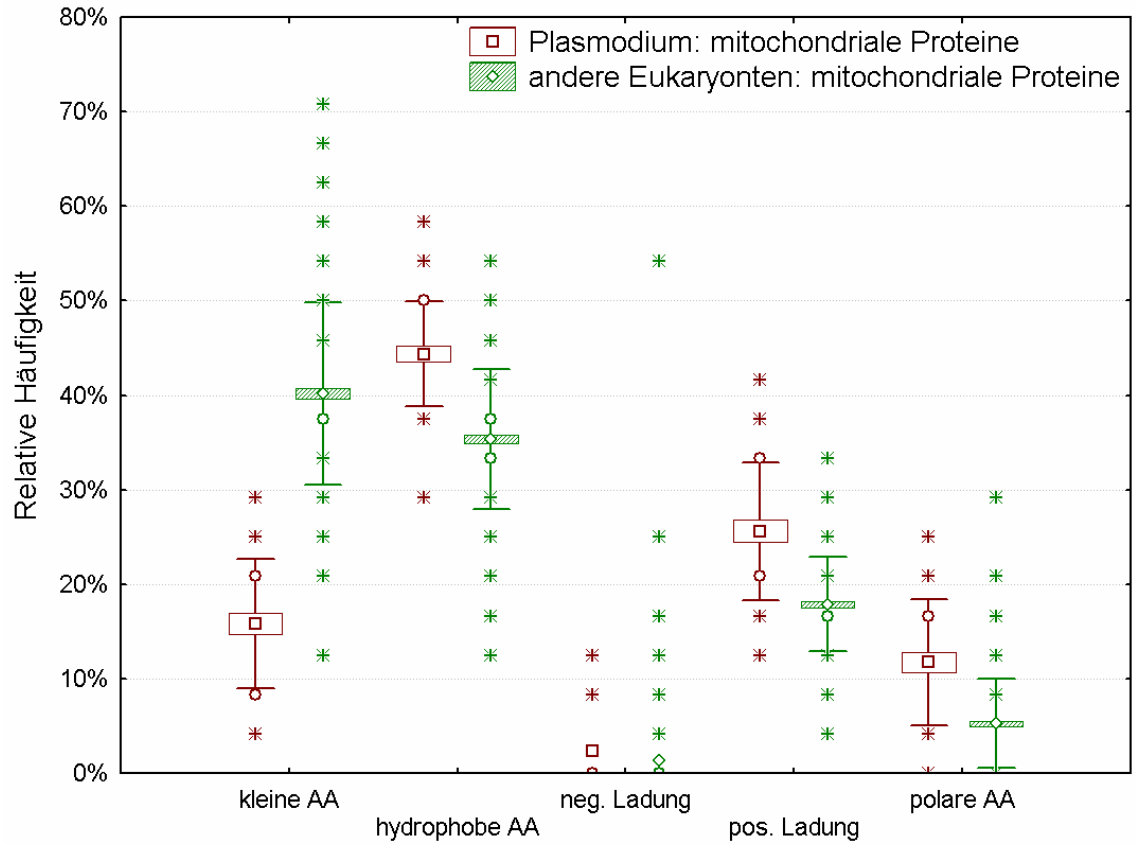

Abbildung 26- Gegenüberstellung der relativen Aminosäurehäufigkeiten von mitochondrialen Proteinen in P. falciparum und anderen Eukaryonten

Mitochondriale Transitpeptide in *P. falciparum,* wie in Abbildung 26 dargestellt, unterscheiden sich von solchen in anderen Eukaryonten durch

- weniger kleine Aminosäuren (mittlerer Anteil von 16% gegenüber 40%)
- mehr hydrophobe, positiv geladene und polare Aminosäuren

Negativ geladene Aminosäuren kommen in beiden Fällen nur selten (in weniger als einem Viertel der Sequenzen) vor.

3.2.4. Kullback-Leibler-Distanzen

Auf Basis der Kullback-Leibler-Distanzen wurde die Distanz der relativen Aminosäurehäufigkeitsverteilung von mitochondrialen, cytoplasmatischen, sekretorischen und – im Fall von *P. falciparum* – auch Apicoplast-Proteinen innerhalb der ersten 24, 31 und 42 N-terminalen Aminosäuren zur Verteilung in der SwissProt-Datenbank, Version 36 berechnet. In Abbildung 27 ist die Kullback-Leibler-Distanz

von Proteinen von Eukaryonten mit Ausnahme von *P. falciparum* als Funktion der Länge des N-terminalen Abschnitts aufgetragen.

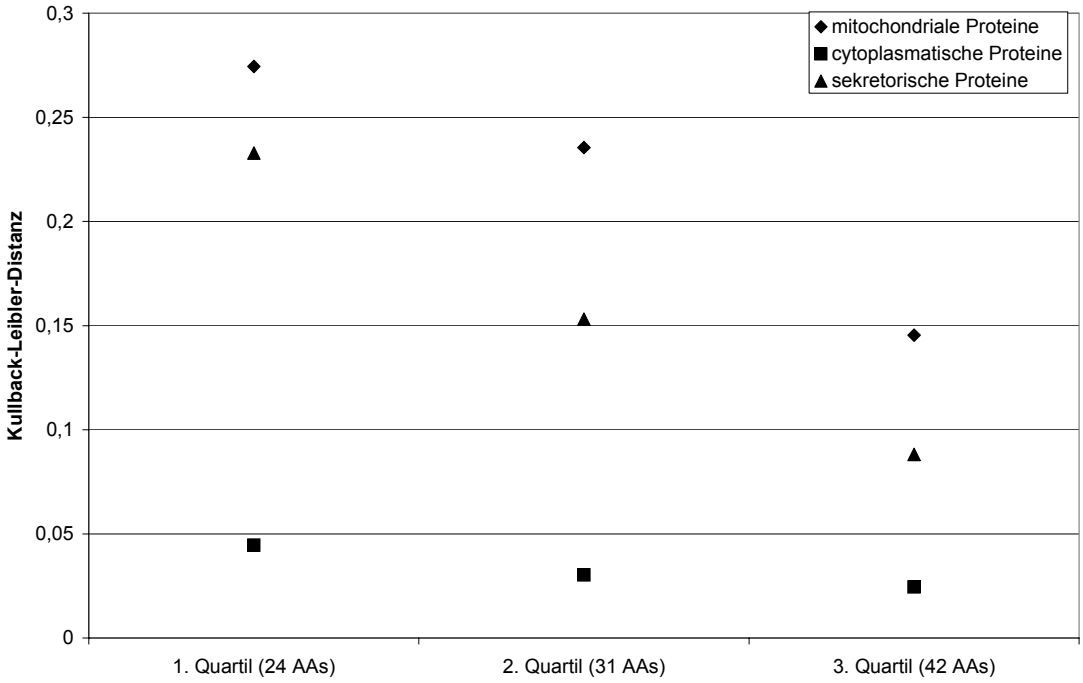

Abbildung 27– Kullback-Leibler-Distanz der Aminosäurehäufigkeitsverteilung von eukaryontischen Proteinen unterschiedlicher Kompartimente zur Referenz SwissProt V36. Die Distanz nimmt mit zunehmender Länge des N-terminales Abschnitts ab

Mitochondriale Transitpeptide zeigen die größte Distanz zur Referenz, der Swiss-Prot-Datenbank in Version 36, sekretorische Signalpeptide ähnlich große Distanz. In beiden Fällen nimmt sie mit N-terminaler Länge jedoch ab. Cytoplasmatische Proteine zeigen eine wesentlich geringere Distanz, die mit Variation des N-terminalen Abschnitts in etwa gleich bleibt.

Die in *P. falciparum* errechneten Distanzen sind in Abbildung 28 dargestellt.

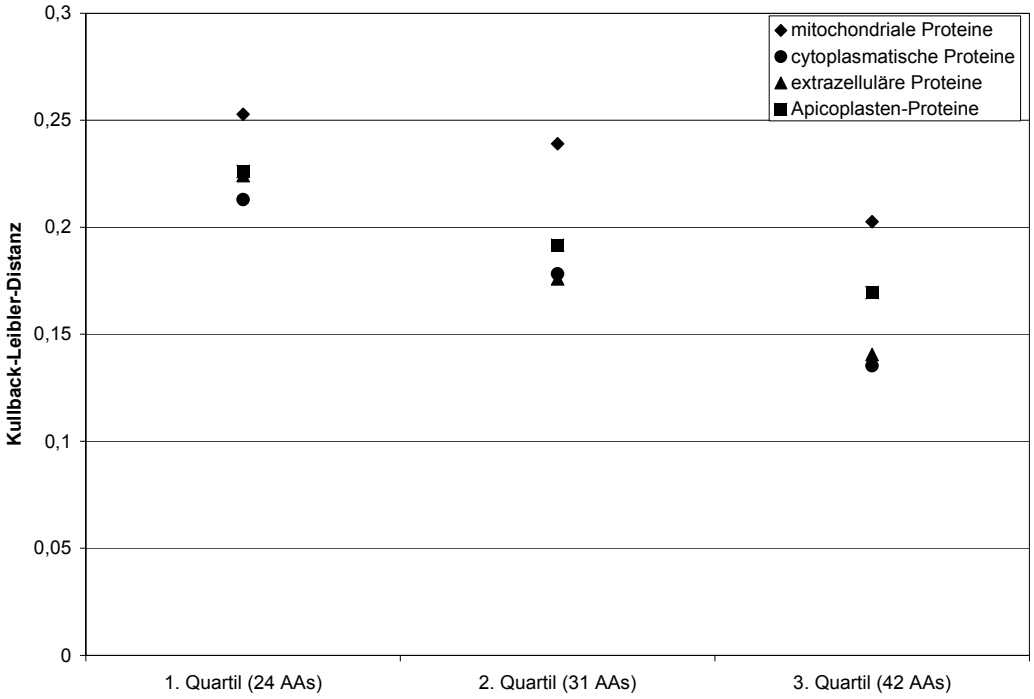

Abbildung 28 – Kullback-Leibler-Distanz der Aminosäurehäufigkeitsverteilung von Proteinen unterschiedlicher Kompartimente aus *P. falciparum* zur Referenz SwissProt V36. Proteine von Plasmodium weisen eine größere Distanz zur SwissProt auf als Proteine anderer Eukaryonten. Dies ist auch bei cytoplasmatischen Proteinen der Fall.

Hier ist in den Fällen aller betrachteter Zellkompartimente eine mit der Länge des N-terminalen Abschnitts abnehmende Distanz zur Referenz zu sehen. Auffällig ist hier, dass selbst cytoplasmatische Proteine zum einen eine fast ebenso große Distanz zur SwissProt aufweisen wie andere Proteine, zum anderen dass diese Distanz auch hier mit zunehmender Größe des N-terminalen Abschnitts abnimmt. Eine durchgängige Distanz zur Verteilung in der Swiss-Prot-Datenbank stimmt mit dem in Abbildung 15 festgestellten, durchgängig unterschiedlichen Aminosäuregebrauch in *P. falciparum* und anderen Eukaryonten überein. Die Abnahme der Kullback-Leibler-Distanz kann dadurch allerdings nicht erklärt werden. Als Begründung könnte die völlige Neuordnung der Targetingmechanismen in *P. falciparum* aufgrund seines Apicoplasten angeführt werden, so dass auch cytoplasmatische Proteine ihren N-terminalen Abschnitt anpassen mussten.

Sowohl in *P. falciparum* als auch in anderen Eukaryonten beobachten wir bei Aminosäuresequenzen aller Lokalisationen eine Abnahme der Distanz zur SwissProt-Datenbank mit zunehmender N-terminaler Abschnittslänge. Kurze N-terminale

Abschnitte weisen also größere Unterschiede zur Datenbank auf, was einen Hinweis darauf gibt, dass 24 Aminosäuren lange Sequenzabschnitte zur Vorhersage der Lokalisation am besten geeignet sein könnten.

3.2.5. Hauptkomponentenanalyse

Diese Analysemethode wurde zur Veranschaulichung der erhaltenen Datensätze gewählt.In Abbildung 29 sind die ersten beiden Hauptkomponenten einer PCA der Aminosäurehäufigkeiten der ersten 24 N-terminalen Aminosäuren aufgetragen, die die Lage der Sequenzen in allen Kompartimenten in *P. falciparum* und anderen Eukaryonten einander gegenüberstellt.

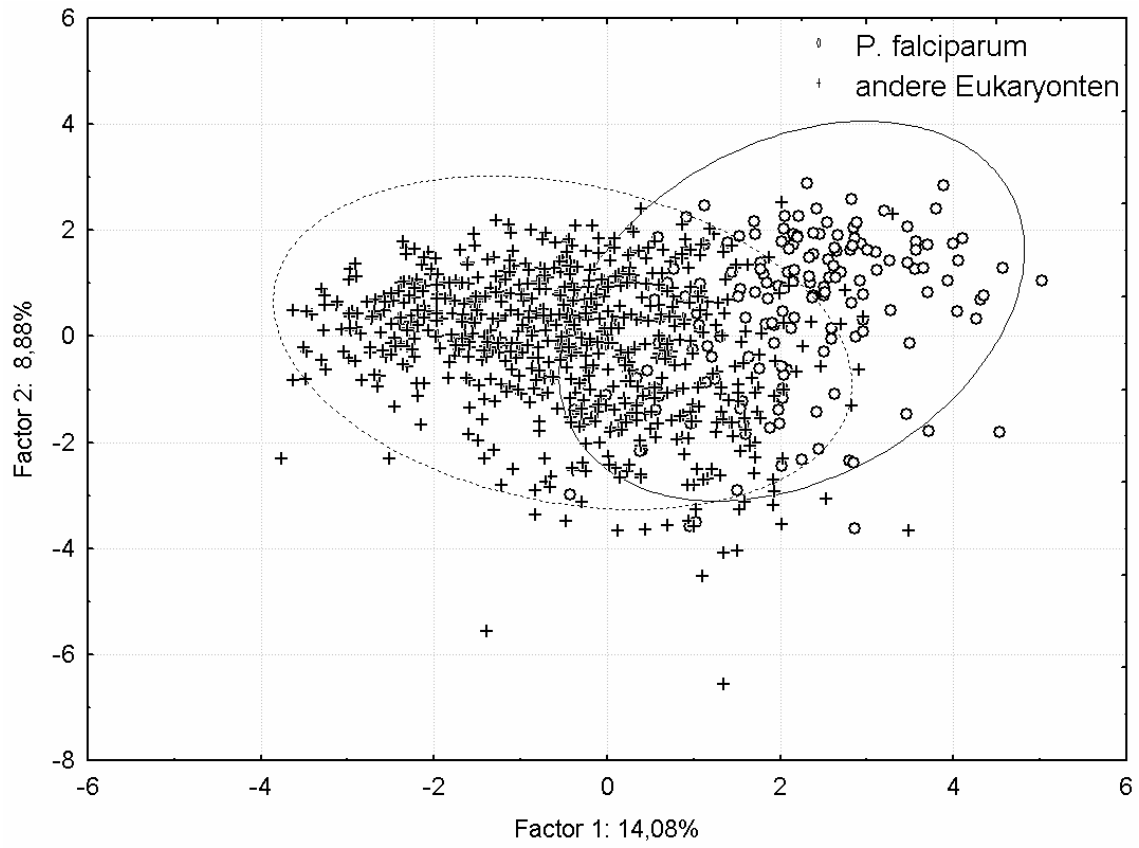

Abbildung 29 – Hauptkomponentenanalyse der Aminosäurehäufigkeiten in Proteinen aus *P. falciparum* (o) und solchen aus andere Eukaryonten (+). Eine tendenzielle Trennung zwischen Proteinen aus Plasmodium und solchen aus anderen Eukaryonten ist zu erkennen.

Schon bei 23,0% erklärter Varianz ist eine deutliche Trennungstendenz zwischen N-terminalen Abschnitten aus *P. falciparum* und anderen Eukaryonten zu sehen. Es existiert jedoch ebenso ein deutlicher Überlappungsbereich beider Gruppen, der die Hinzunahme weiterer Dimensionen zur besseren Klassifikation notwendig erscheinen lässt. Es existiert keine ausgeprägte 1. Hauptachse. Die etwa 14% erklärte Varianz

durch den ersten Faktor sind, im Vergleich zu unten beschriebenen anderen Datensätzen, ein mittlerer Wert.

Wie trennen sich nun die in den einzelnen Kompartimenten lokalisierten Sequenzen in anderen Eukaryonten? Diese Fragestellung wird in Abbildung 30 beantwortet.

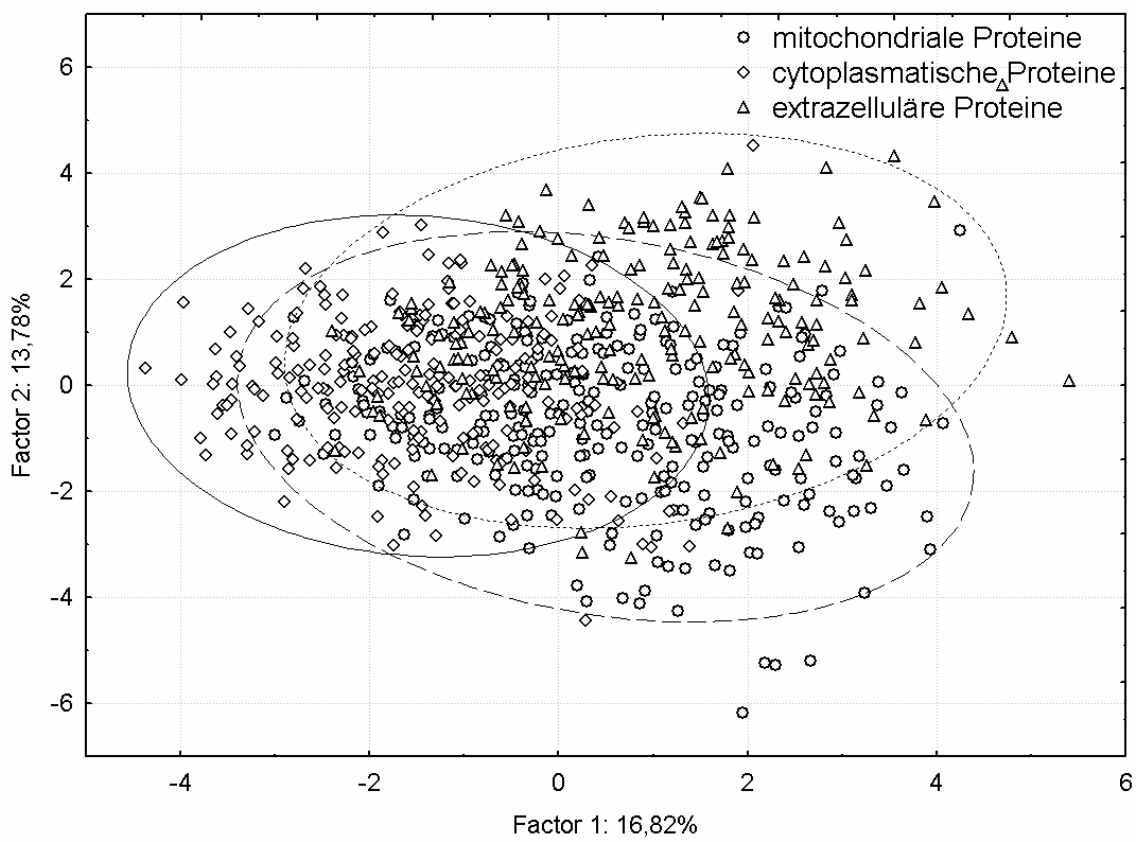

Abbildung 30 – Hauptkomponentenanalyse der Aminosäurehäufigkeiten von eukaryontischen Proteinen verschiedener Lokalisationen. Wir sehen eine deutliche Überlappung zwischen den Proteinen verschiedener Lokalisationen.

In diesem Fall erklären die ersten beiden Hauptkomponenten 30,6% der Varianz. Wie in Abbildung 30 ersichtlich ist, bestehen offenbar tendenzielle Unterschiede in der Aminosäurehäufigkeit unterschiedlich lokalisierter Proteine in eukaryontischen Organismen. Allerdings kann aufgrund dieser Projektion nicht eindeutig auf eine bestimmte Lokalisation einer N-terminalen Aminosäuresequenz geschlossen werden, da die Bereiche, in denen sich Sequenzen einer Lokalisation abgebildet werden, stark überlappen. Wie verhält es sich mit der in Abbildung 31 gezeigten Trennbarkeit in *P. falciparum*?

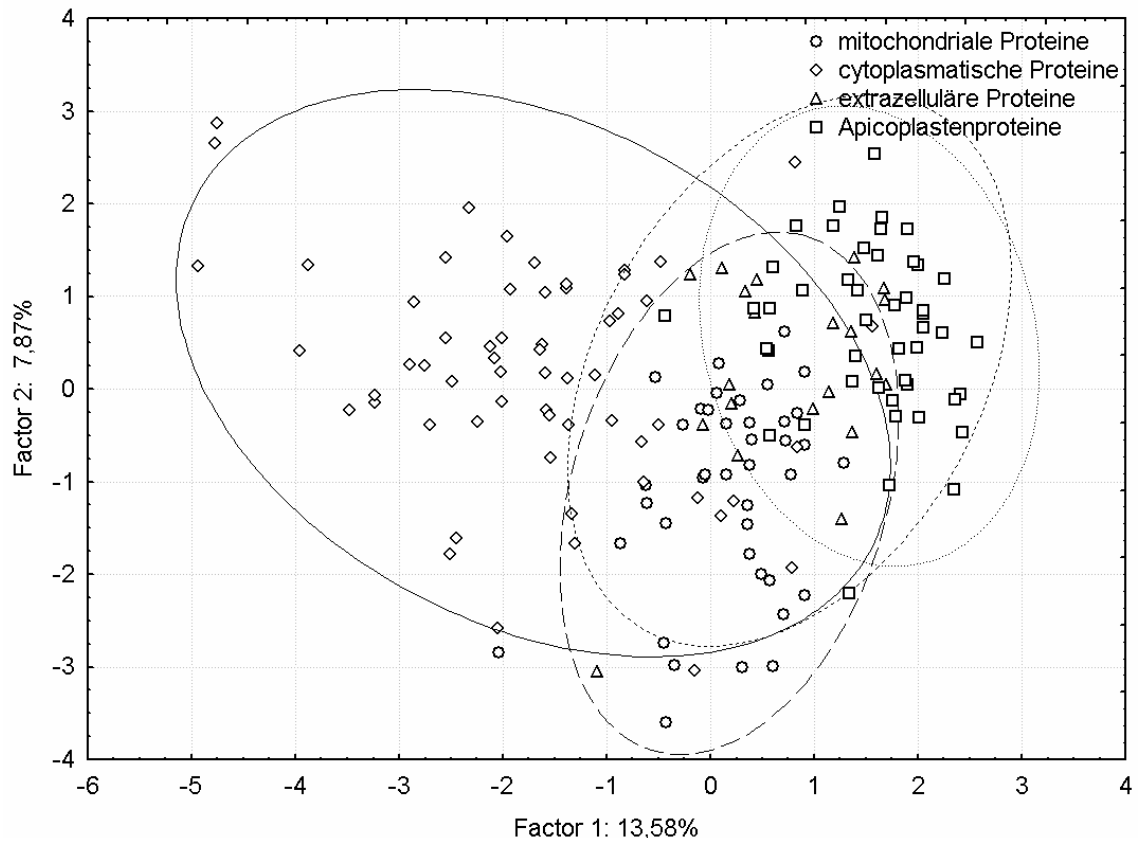

Abbildung 31 – Hauptkomponentenanalyse der Aminosäurehäufigkeiten für Proteine unterschiedlicher Lokalisationen aus *P. falciparum.* Cytoplasmatische Proteine in Plasmodium unterscheiden sich stark von Proteinen der anderen drei Lokalisationen.

Die ersten beiden Hauptachsen erklären in Abbildung 31 21,5% der Gesamtvarianz. In diesem Fall besteht ein großer Unterschied zwischen der Gruppe der cytoplasmatischen und anderen Proteinen. Eine Trennung innerhalb dieser zweiten Gruppe ist kaum möglich, da extrazelluläre, mitochondriale und apicoplastische N-terminale Abschnitte in dieser Projektion stark überlappen. In dieser Projektion scheinen deutliche Gemeinsamkeiten zwischen Apicoplasten-, mitochondrialen und extrazellulären Sequenzen zu bestehen. Dies ist in Übereinstimmung mit der Literatur, da die Targetingpeptide von Proteinen dieser Lokalisationen miteinander starke Ähnlichkeiten – u.a. positive Nettoladung, höherer Anteil hydrophober Aminosäuren – aufweisen[7].

Wie stellen sich nun mitochondriale Transitpeptide aus *P. falciparum* und anderen Eukaryonten gegenüber dar (Abbildung 32)?

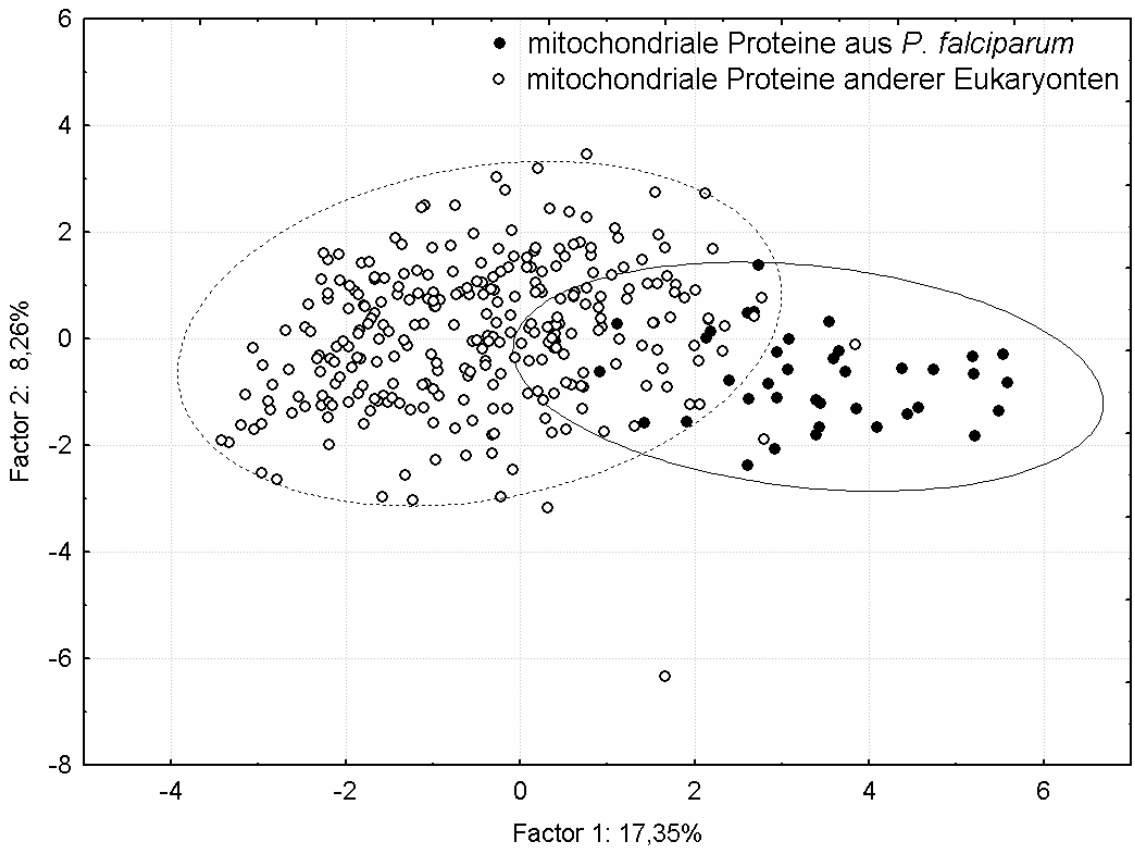

Abbildung 32 - Hauptkomponentenanalyse der Aminosäurehäufigkeiten mitochondrialer Transitpeptide aus *P. falciparum* und solchen aus anderen Eukaryonten. Mitochondriale Transitpeptide aus Plasmodium falciparum und anderen Eukaryonten zeigen deutliche Unterschiede.

Die ersten beiden Hauptkomponenten erklären eine Varianz von 25,6. In dieser Projektion ist ein deutlicher Unterschied zwischen mitochondrialen Proteinen aus *P. falciparum* und anderen Eukaryonten sichtbar. Dies bestätigt die im Abschnitt der Aminosäurehäufigkeiten (Abbildung 16, Abbildung 19) erkennbaren unterschiedlichen Gebrauch von einzelnen Aminosäuren. Ebenso weist diese Darstellung darauf hin, dass bereits etablierte Methoden zur Vorhersage mitochondrialer Transitpeptide in Eukaryonten in *P. falciparum* aufgrund des unterschiedlichen Aminosäuregebrauchs keine optimalen Ergebnisse erzielen könnten.

3.2.6. Self-Organizing-Map

Die Selbstorganisierende Karte (Self-Organizing-Map, SOM) hat die Fähigkeit, Daten zu gruppieren ohne vorgegebene Kategorien im Lernprozess zur Hand haben zu müssen. Jeder Dateneingabevektor führt zur Aktivierung genau eines Neurons im Ausgabelayer, und das Netz lernt (im Idealfall), sehr ähnliche Eingabevektoren auf dasselbe und ähnliche Eingabevektoren auf benachbarte Ausgabeneuronen abzubilden. Hier wurde die SOM mit der gleichen Intention wie die Hauptkomponentenanalyse verwendet, und zwar mit dem Ziel, Veranschaulichungen der Ähnlichkeit oder Verschiedenheit der Aminosäuresequenzen unterschiedlicher Lokalisationen darzustellen. Daraus lässt sich beispielsweise eine Vermutung über das Ausmaß der Trennbarkeit der Datensätze durch mehrlagige Perzeptronen oder andere Methoden in den jeweils verwendeten Datenrepräsentationen treffen. Hier wurde als Datenrepräsentation die relative Aminosäurehäufigkeit verwendet, in einem späteren Abschnitt (Kapitel 3.3.3) die physikochemischen Eigenschaften.
Im Falle der aus Eukaryonten mit Ausnahme von *P. falciparum* zusammengestellten Sequenzen ergaben sich die in Abbildung 33 gezeigten Aktivierungen der Kohonen-Karte.

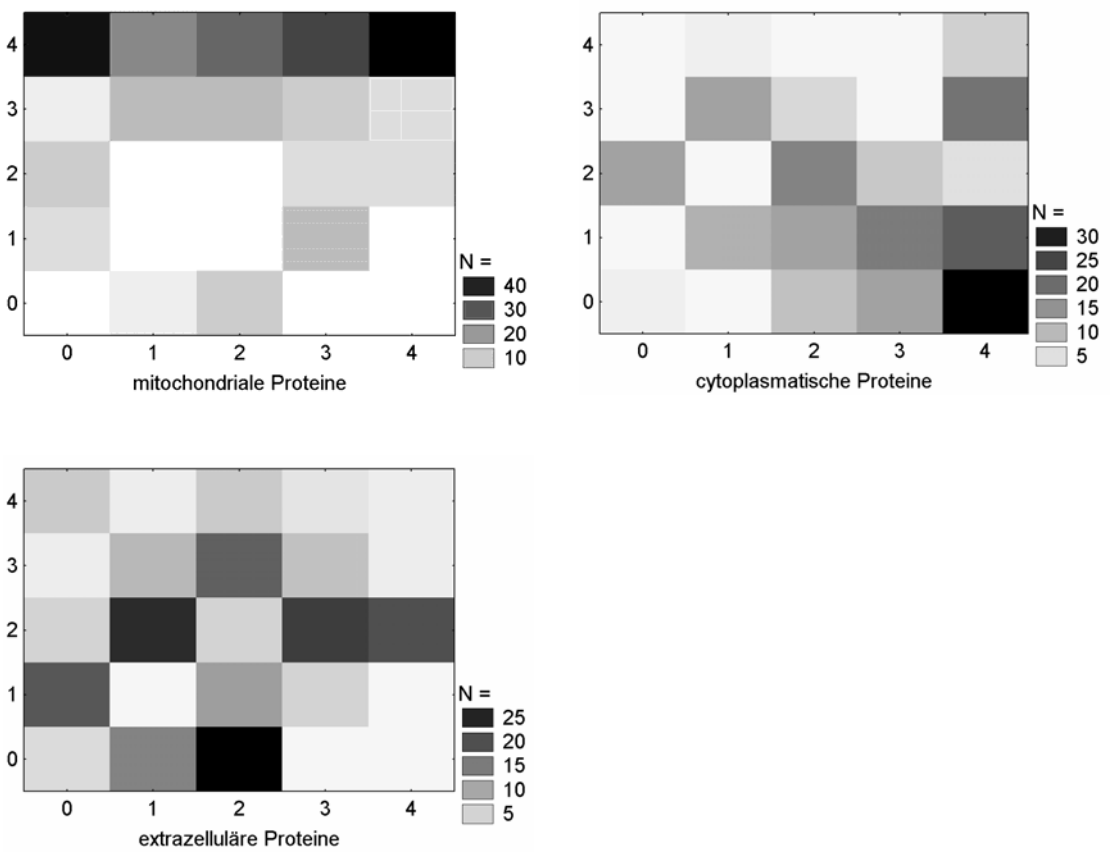

Abbildung 33 – Klassifizierung von 175 N-terminalen, eukaryontischen Sequenzabschnitten mit einer Länge von 24 Aminosäuren mit Hilfe einer selbstorganisierenden Karte (SOM). Mit dieser Methode werden Unterschiede in der verwendeten Repräsentation der Aminosäurehäufigkeiten gut erkennbar. N ist die Anzahl der das jeweilige Neuron aktivierenden N-terminalen Aminosäureabschnitte.

Im Fall mitochondrialer Transitpeptide erfolgt die Aktivierung im Ausgabelayer zum Großteil von Neuron (0;4) bis Neuron (4;4). (Das Tupel stellt in dieser Reihenfolge X- und Y-Koordinate des Neurons dar.) Auf diese Neuronen werden 172 der 282 mitochondrialen Transitpeptide abgebildet. Fälschlich werden diese Neuronen von 14 cytoplasmatischen und 19 extrazellulären Proteinen aktiviert. Cytoplasmatische Proteine aktivieren einen anderen Bereich des Ausgabelayers; die vier Neuronen mit den Koordinaten (3,4; 0,1) erfassen 87 der 226 cytoplasmatische Proteine und werden gleichzeitig von 14 mitochondrialen und 6 extrazellulären Proteinen aktiviert. Extrazelluläre Proteine werden über einen weiteren Bereich der Ausgabeneuronen gestreut; die Neuronen (2; 0-3), (1; 2) und (3; 2) erfassen 109 der 215 extrazellulären Datensätze und gleichzeitig 34 extrazelluläre und 49 cytoplasmatische Proteine.

Im Falle von *P. falciparum* ergaben sich die folgenden Aktivierungen.

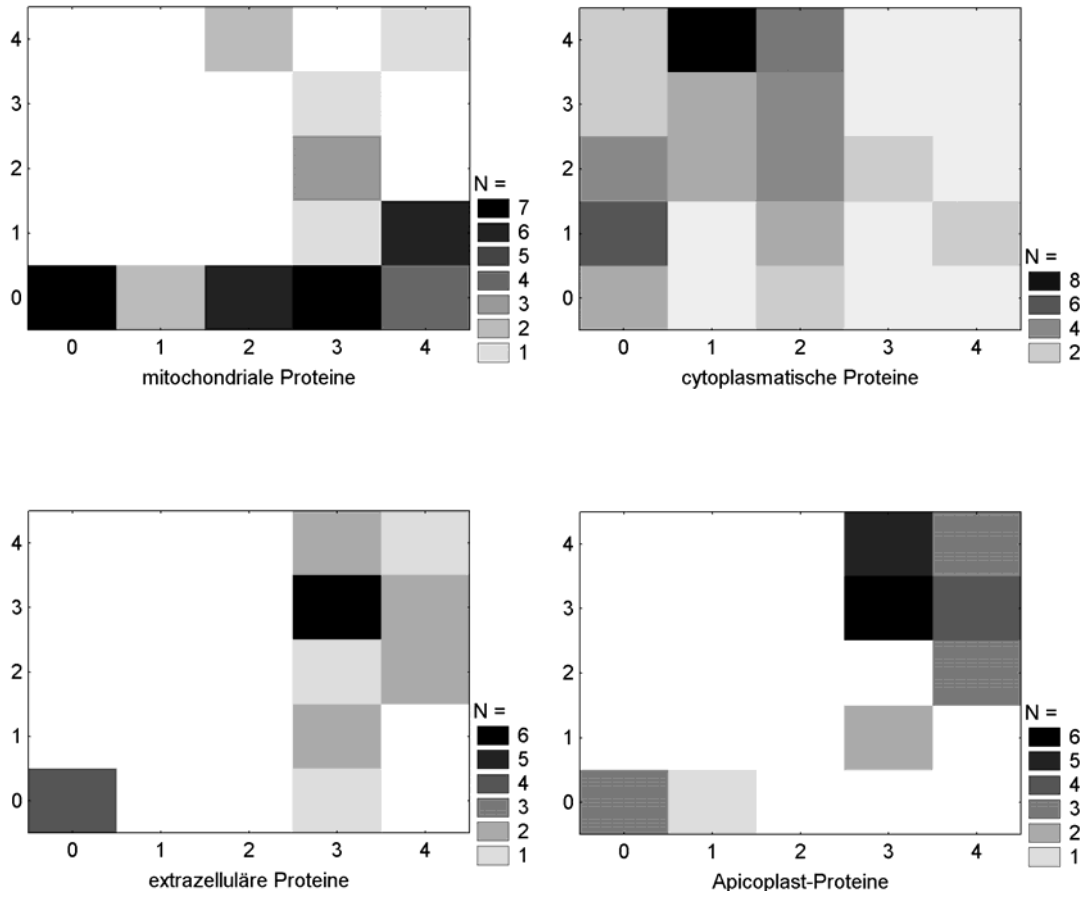

Abbildung 34 – Klassifizierung von 175 N-terminalen Sequenzabschnitten aus *P. falciparum* mit einer Länge von 24 Aminosäuren mit Hilfe einer selbstorganisierenden Karte (SOM). Die Sequenzen extrazellulärer Proteine und diejenigen apicoplastischer Proteinen können in dieser Darstellung nur schwer getrennt werden. N ist die Anzahl der das jeweilige Neuron aktivierenden N-terminalen Aminosäureabschnitte.

Hier werden 23 der 40 mitochondrialen Transitpeptide auf 5 der Ausgabeneuronen abgebildet – auf die Neuronen (1-4; 0) sowie das rechts gelegene Neuron der darüberliegenden Reihe, Neuron (4; 1). Auf diese Neuronen werden auch 6 cytoplasmatische Proteine, ein extrazelluläres Protein, jedoch kein Apicoplastenprotein abgelegt. Die cytoplasmatischen Proteine streuen weit über die Neuronen des Ausgabelayers; auf einem in der linken oberen Ecke befindlichen Rechteck der zwölf Neuronen (0-2; 1-4) werden 46 der 61 cytoplasmatischen, jedoch nur zwei mitochondriale und weder extrazelluläre noch Apicoplastensequenzen abgelegt. Extrazelluläre und Apicoplastenproteine werden praktisch den gleichen Ausgabeneuronen zugeordnet; hier ist also, mit dieser Darstellungsmethode, kaum eine Trennung möglich. Dieses Ergebnis deckt sich mit der schlechten Trennbarkeit in der

Hauptkomponentenanalyse. Grund ist eine Ähnlichkeit der Funktion und auch des Aufbaus der mitochondrialen Transit- und der apicoplastischen Targetingpeptide. Letztere bestehen ebenfalls aus einem Signal- und einem Transitpeptid und weisen ähnliche Aminosäurezusammensetzung, wie beispielsweise eine positive Nettoladung und einen höheren Anteil hydrophober Aminosäuren gegenüber cytoplasmatischen Proteinen auf.

Die in der SOM gezeigten Ergebnisse decken sich gut mit den aus der PCA gewonnenen Erkenntnissen. Cytoplasmatische N-terminale Abschnitte einerseits und Abschnitte extrazellulärer, mitochondrialer und im Fall von *P. falciparum* auch apicoplastischer Aminosäuresequenzen andererseits bilden zwei deutlich unterschiedliche, jedoch nicht vollständig voneinander trennbare Gruppen von Sequenzen. Die Trennbarkeit innerhalb der Gruppe extrazellulärer, mitochondrialer und im Fall von *P. falciparum* auch apicoplastischer Sequenzen ist weder durch Projektion der ersten beiden Hauptachsen bei der Hauptkomponentenanalyse noch durch Anwendung einer Selbstorganisierenden Karte möglich. Diese Ergebnisse beziehen sich auf die in diesem Abschnitt behandele Datenrepräsentation im Aminosäurehäufigkeitsraum.

3.3. 19-dimensionaler Eigenschaftsraum

3.3.1. Allgemeines

Hinter der Idee, die Information der N-terminalen Abschnitte in einem physikochemischen Eigenschaftsraum darzustellen, steckt die Annahme, dass ebendiese Eigenschaften zu einem Teil für die Erkennung der Targetingsignale in der Zelle verantwortlich sind oder zumindest mit der Erkennung innerhalb der Zelle korrelieren. Dies scheint plausibel, da beispielsweise mitochondriale Transitpeptide einen höheren Anteil positiv geladener und hydrophober Aminosäuren aufweisen als cytoplasmatische Proteine oder Referenzdatenbanken wie die SwissProt-Datenbank.
Es liegt eine sehr große Zahl von derzeit 434 tabellierten physikochemischen Eigenschaften vor[40]. Allerdings enthalten diese Eigenschaften teilweise miteinander korrelierte Informationen, was einerseits zu einer Informationsredundanz und andererseits aufgrund der hohen Dimensionalität der Daten zu Problemen beim Training neuronaler Netze führen würde. Daher wurde hier ein 19-dimensionaler Eigenschaftsraum gewählt, der durch Hauptkomponentenanalyse der 434 Eigenschaften gewonnen wurde[39].

3.3.2. Hauptkomponentenanalyse

Hier sollen die auch im Aminosäurehäufigkeitsraum durchgeführten Hauptkomponentenanalysen in einer neuen Kodierung – mit Hilfe physikochemischer Eigenschaften - wiederholt werden. Wie sieht es hier mit der Trennung von Proteinen aus *P. falciparum* und anderen Eukaryonten aus, die in Abbildung 35 gezeigt ist?

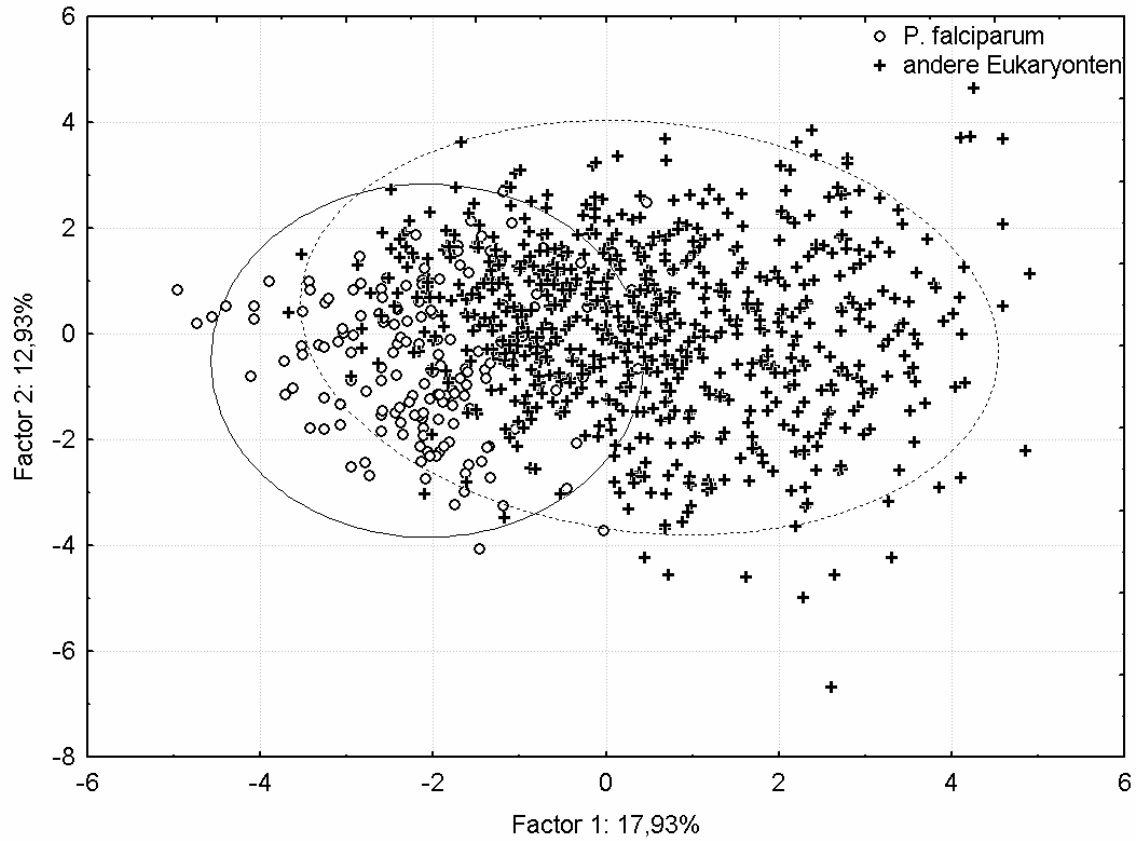

Abbildung 35 – Hauptkomponentenanalyse der 19-dimensionalen physikochemischen Eigenschaftsvektoren von Proteinen aus *P. falciparum* und solchen aus andere Eukaryonten. Es bestehen leichte Unterschiede zwischen Proteinen aus Plasmodium falciparum und denen aus anderen Eukaryonten, die Proteine anderer Eukaryonten streuen erwartungsgemäß über einen weiten Bereich.

Die ersten beiden Hauptachsen erklären 30,9% der Gesamtvarianz, was – gegenüber der entsprechenden Darstellung im Aminosäurehäufigkeitsraum (Abbildung 29) – einen recht großen Wert darstellt. Die Datensätze aus beiden Organismen überlappen sich in der Eigenschaftsprojektion jedoch stärker als in der Aminosäurehäufigkeitsprojektion, ein Grossteil der ersten Hauptachse trennt die Datensätze anderer Eukaryonten auf. Die PCA von eukaryontischen Sequenzen mit unterschiedlicher Lokalisierung ist in Abbildung 36 gegeben.

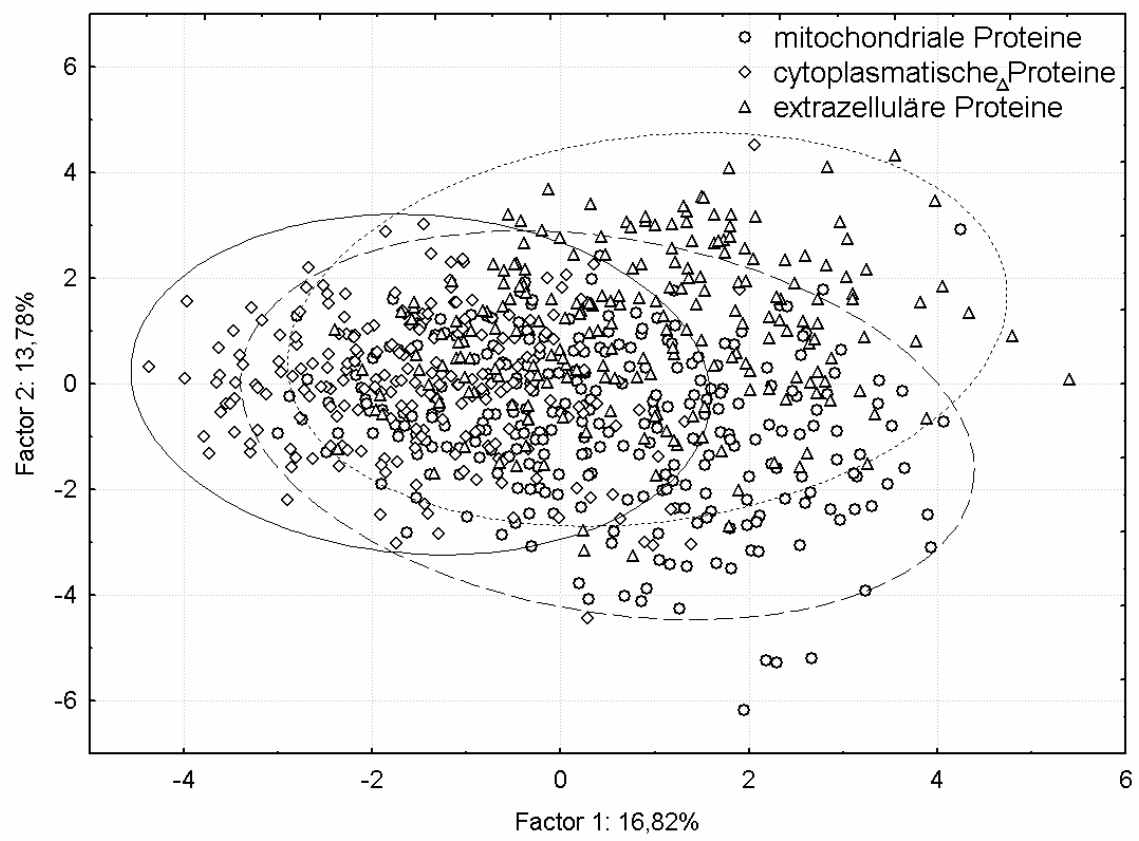

Abbildung 36 – Hauptkomponentenanalyse der 19-dimensionalen physikochemischen Eigenschaftsvektoren von Proteinen unterschiedlicher Lokalisationen verschiedener Eukaryonten. Proteine unterschiedlicher Lokalisierung besitzen eine in etwa ähnliche Aminosäurezusammensetzung und sind mit dieser Methode kaum zu unterscheiden.

In dieser Auftragung erklären die ersten beiden Hauptachsen 30,6% der totalen Varianz. Allerdings ist hier, im Fall der Aminosäurehäufigkeiten, kaum ein tendenzieller Unterschied zwischen den einzelnen Proteinlokalisierungen sondern große Unterschiede innerhalb der Gruppen festzustellen.

Wie gut lassen sich die Proteine unterschiedlicher Lokalisierung im Falle von *P. falciparum* auftrennen (Abbildung 37)?

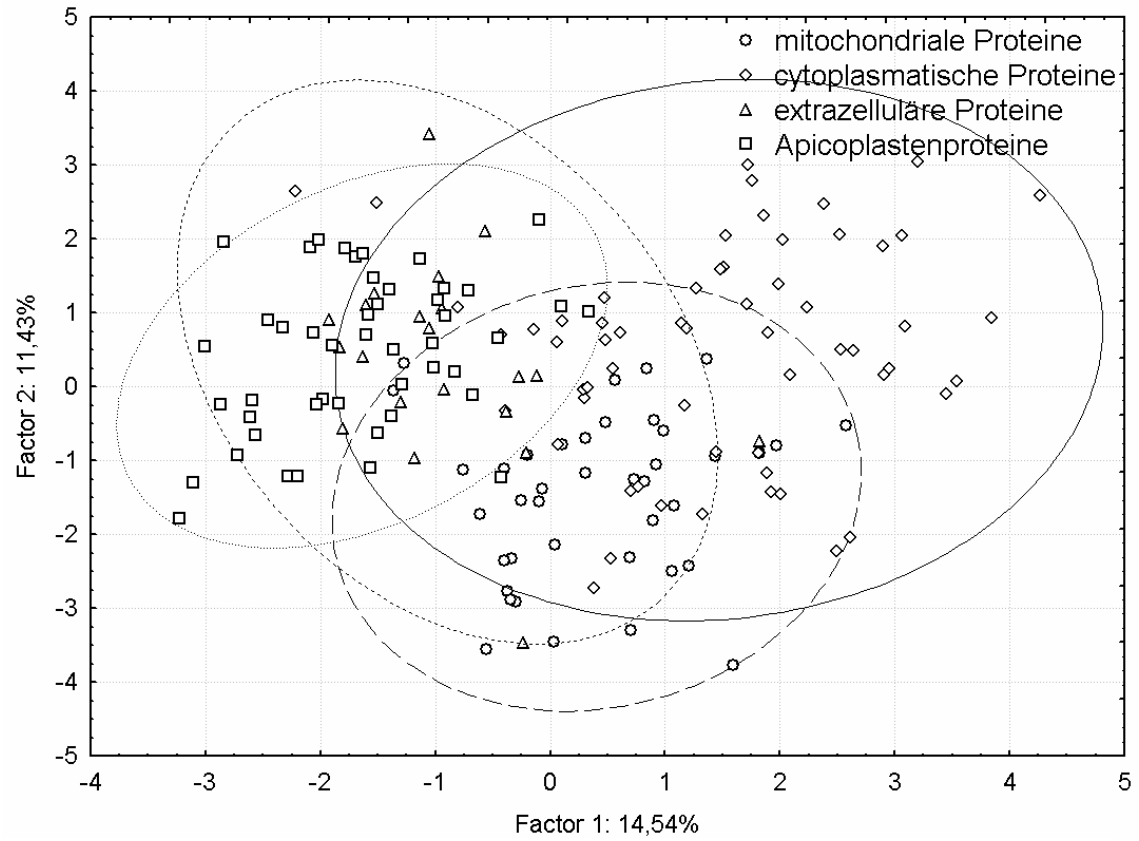

Abbildung 37 – Hauptkomponentenanalyse der 19-dimensionalen physikochemischen Eigenschaftsvektoren von Proteinen unterschiedlicher Lokalisationen aus *P. falciparum*. Cytoplasmatische Proteine unterscheiden sich stark von andernorts lokalisierten Proteinen.

Mit den ersten beiden Hauptachsen werden hier 26,0% der Gesamtvarianz erklärt. Die cytoplasmatischen Proteine unterscheiden sich stark von denjenigen anderer Kompartimente. Mitochondriale Transitpeptide und extrazelluläre Sequenzen überlappen auch hier deutlich.

Wie trennen sich die mitochondrialen Transitpeptide von *P. falciparum* und anderen Eukaryonten (Abbildung 38)?

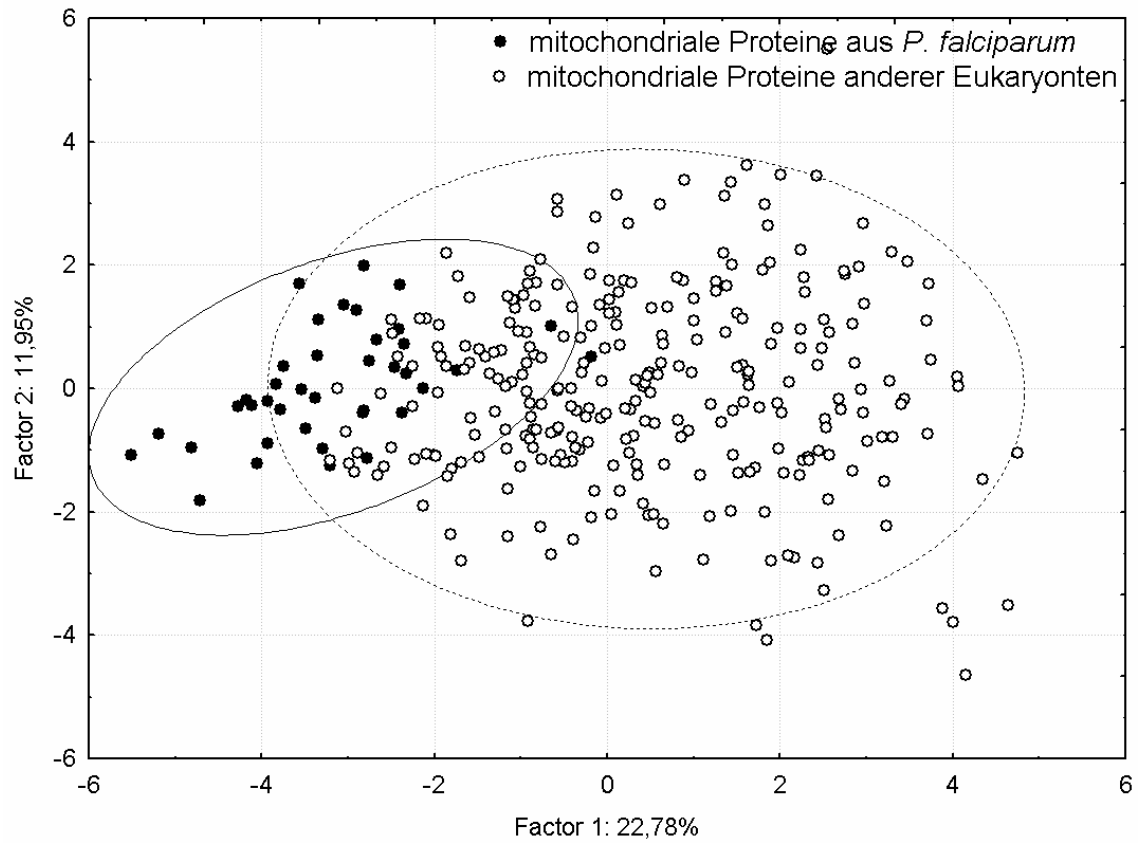

Abbildung 38 – Hauptkomponentenanalyse der 19-dimensionalen physikochemischen Eigenschaftsvektoren von mitochondrialen Transitpeptiden aus *P. falciparum* und solchen aus andere Eukaryonten. Es ist eine unvollständige Trennung der beiden Gruppen zu erkennen.

Der durch die ersten beiden Hauptachsen erklärte Varianzwert von 34,7% ist sehr hoch, wobei die erste Hauptachse recht genau eine Achse zwischen Sequenzen aus *P. falciparum* und Sequenzen anderer Eukaryonten darstellt. Wie in der Darstellung durch Aminosäurehäufigkeiten (Abbildung 32) ist eine tendenzielle, aber nicht vollständige Trennung zwischen Sequenzen aus *P. falciparum* und solchen aus anderen Eukaryonten zu erkennen.

3.3.3. Self-Organizing-Map

Im Falle der Aminosäuresequenzen aus Eukaryonten mit Ausnahme von *P. falciparum* ergaben sich auf den Selbstorganisierenden Karten folgende Aktivierungsmuster. Allgemein ist in dieser Datenrepräsentation eine gute Klassifizierung möglich.

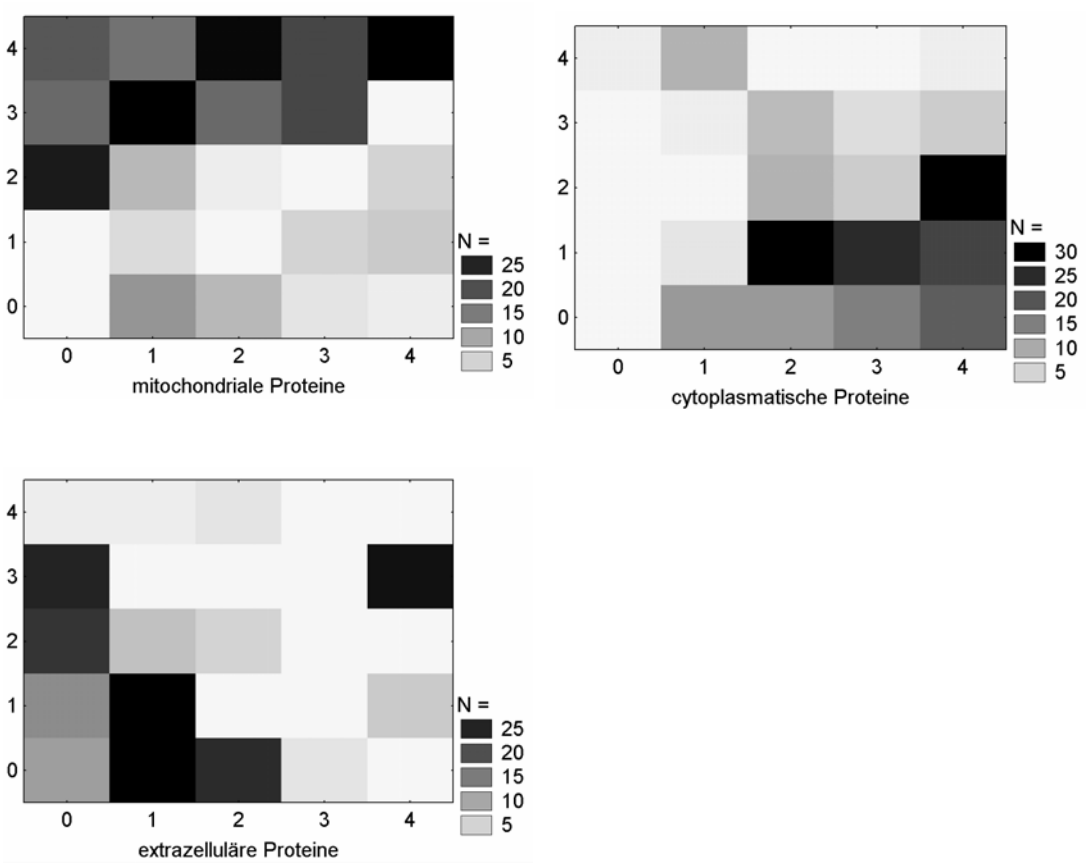

Abbildung 39 - Klassifizierung von 175 N-terminalen, eukaryontischen Sequenzabschnitten mit einer Länge von 24 Aminosäuren mit Hilfe einer selbstorganisierenden Karte (SOM). Mit dieser Methode werden Unterschiede in der verwendeten Repräsentation der 19-dimensionalen physikochemischen Eigenschaften ebenfalls gut erkennbar.

In den 10 Neuronen (0-4; 4), (0-3; 3) und (0; 2) werden 223 der 282 mitochondrialen Transitpeptide erkannt. 30 cytoplasmatische und 58 extrazelluläre Sequenzen aktivieren ebenfalls diese Neuronen im Ausgabelayer. Die hier recht unten angeordneten, dunklen Ausgabeneuronen (1-4; 0), (2-4; 1) und (4; 2) werden durch 126 der 226 cytoplasmatischen Proteine aktiviert. Ebenso werden dort 19 mitochondriale und 8 extrazelluläre Proteine fehlklassifiziert. Die fünf Neuronen (0-2; 0), (0-1; 1) sowie (4; 3) erkennen 133 der 215 extrazellulären Sequenzen. 25 mitochondriale und 34 cytoplasmatische Proteine aktivieren ebenfalls eins dieser Ausgabeneuronen.

Im Falle der *P. falciparum*-Sequenzen ergaben sich die folgenden Aktivierungen der Outputneuronen:

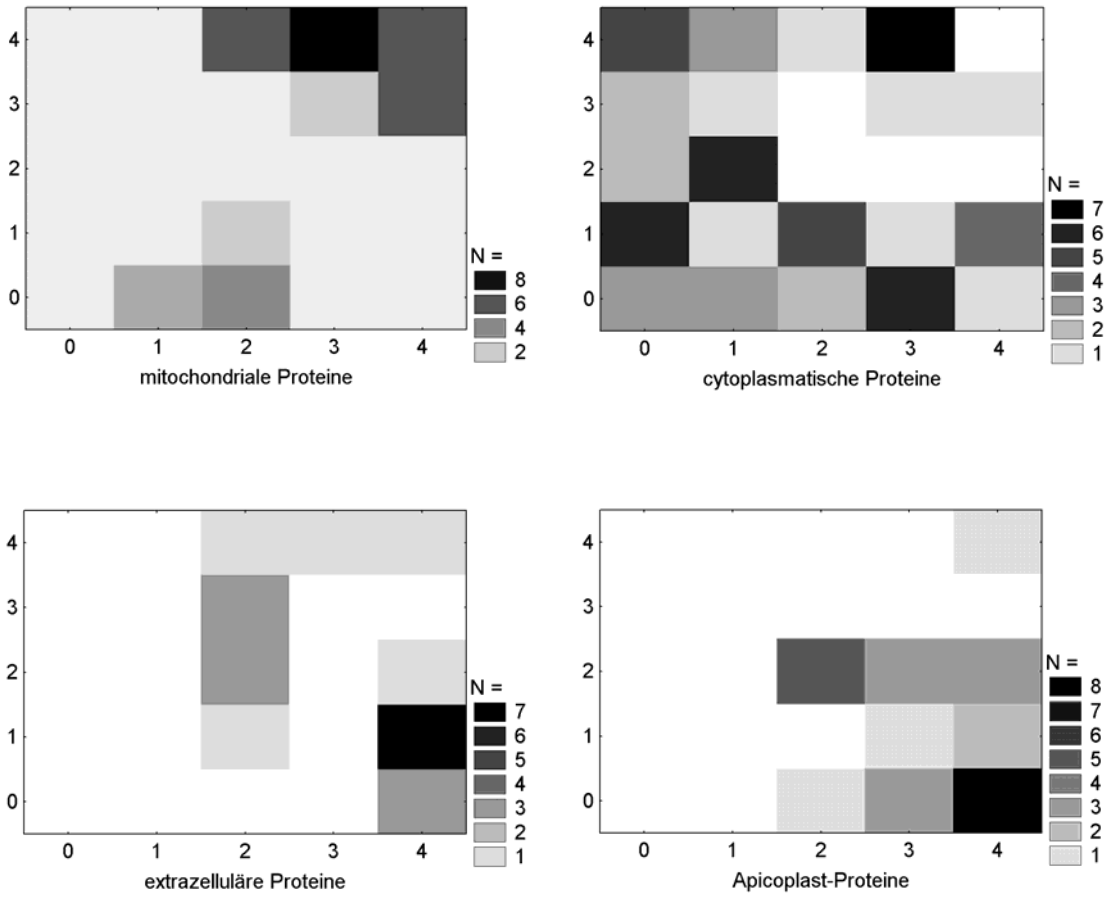

Abbildung 40– Klassifizierung von 175 N-terminalen Sequenzabschnitten aus *P. falciparum* mit einer Länge von 24 Aminosäuren mit Hilfe einer selbstorganisierenden Karte (SOM). Die Sequenzen extrazellulärer Proteine und diejenigen apicoplastischer Proteinen können in dieser Darstellung nur schwer getrennt werden.

Die vier dunkel markierten Ausgabeneuronen (4; 3) und (2-4; 4) erkennen 27 der 61 mitochondrialen Proteine. Fehlerhaft werden diesen Neuronen auch 9 cytoplasmatische, drei extrazelluläre und ein Apicoplast-Protein zugeordnet. Die 16 Ausgabeneuronen (0-3; 0-3) erkennen 38 der 61 cytoplasmatischen Sequenzen. Außerdem werden als falsch-positiv 12 mitochondriale, 7 extrazelluläre und 13 Apicoplast-Sequenzen erkannt. Extrazelluläre und Apicoplastenproteine werden zu einem großen Teil den gleichen Ausgabeneuronen zugeordnet; hier ist also, mit dieser Darstellungsmethode, kaum eine Trennung möglich. Dies entspricht dem Ergebnis bei Nutzung der Aminosäurehäufigkeiten als Inputvektoren (Abbildung 33 und Abbildung 34).

Die Ergebnisse der Veranschaulichung mittels Hauptkomponentenanalyse und Selbstorganisierender Karte, bei Repräsentation der Daten durch relative Aminosäurehäufigkeit und 19-dimensionalen physikochemischen Eigenschaftsraum, weisen allesamt in eine ähnliche Richtung. Es ist stets ein tendenzieller Unterschied zwischen cytoplasmatischen N-terminalen Abschnitten auf der einen Seite und extrazellulären, mitochondrialen und im Fall von *P. falciparum* auch apicoplastischen N-terminalen Abschnitten auf der anderen Seite zu erkennen. Allerdings ist die Trennung beider Gruppen nie vollständig, eine Zuordnung mittels dieser Methoden wäre nicht eindeutig möglich. Wir erwarten daher zwar eine gute Unterscheidung zwischen den Gruppen, eine Verwechslung zwischen mitochondrialen, extrazellulären und im Fall von *P. falciparum* auch apicoplastischen Sequenzen scheint eher möglich zu sein als die Missklassifizierung einer zu diesen Gruppen gehörenden Aminosäuresequenzen mit einer cytoplasmatischen, N-terminalen Sequenz.
Aus Abschnitt 3.2.4., der zeigte dass im Aminosäurehäufigkeitsraum die größte Kullback-Leibler-Distanz zwischen der SwissProt als Referenzdatenbank und 24 Aminosäuren kurzen Abschnitten besteht, folgern wir die Erwartung dass kurze N-terminale Abschnitte auch die besten Ergebnisse bei Anwendung neuronaler Netze zur Datenklassifikation liefern. Da die genannte Distanz jedoch auch mit 31 und 42 Aminosäuren langen Abschnitten zwar abnahm, jedoch prinzipiell vorhanden blieb, scheint auch mit längeren Abschnitten eine Klassifikation durchführbar zu sein.

3.4. Ermittlung der optimalen Struktur des Neuronalen Netzes

3.4.1. Variation der Anzahl Hidden Neuronen

Ein wesentlicher Parameter eines neuronalen Netzes ist dessen Netztopologie. Für das hier vorliegende Problem wurde ein dreilagiges Perzeptron verwendet, das ein Layer aus Hidden Neuronen besaß. Als entscheidender Parameter sollte nun in diesem Teil der Arbeit die optimale Anzahl Hidden Neuronen herausgefunden werden. Hierzu wurde diese bei ansonsten gleichen Parametern von 1-20, 30, 50, 100 bis auf 500 Neuronen variiert. Es wurden im Programm Statistica die folgenden Standardparameter gewählt (die englischen Begriffe wurden der Reproduzierbarkeit wegen übernommen):

- Multilayer-Perzeptron, 1 Hidden Layer
- Classification Error Function: Sum Squared
- Learning Algorithm: Back Propagation, max. 10000 Epochen, Lernrate 0,01
- Initialisierung Random/Uniform, Min.:0, Max.:1
- Stopping Condition: Minimum improvement in Error 0,001 / 1000 Epochen im Select-Datensatz
- Classification: Assign to Highest Threshold
- Decay Factors: none
- Adjust Learning Rate: off
- Shuffle Presentation order each Epoch: on
- Sampling: 10faches Random Resampling (“Kreuzvalidierung”) mit 89 Datensätzen im Trainingssatz und jeweils 43 im Select- und Testdatensatz

Als Übersicht sind in Abbildung 41 sämtliche Klassifikationsergebnisse als Übersicht in einem Diagramm dargestellt. Teile daraus werden später gesondert behandelt. Aus dieser Übersicht zogen wir die Information, dass wenig qualitative Unterschiede (ausgedrückt als Mathews-Koeffizient) zwischen

- der Anzahl Hidden Neuronen in gewissen Grenzen (2-30)
- der Repräsentation (Aminosäurehäufigkeiten/19-dimensionaler Eigenschaftsraum) und
- Länge des N-terminalen Abschnitts (24,31 oder 42 Aminosäuren)

bestehen.

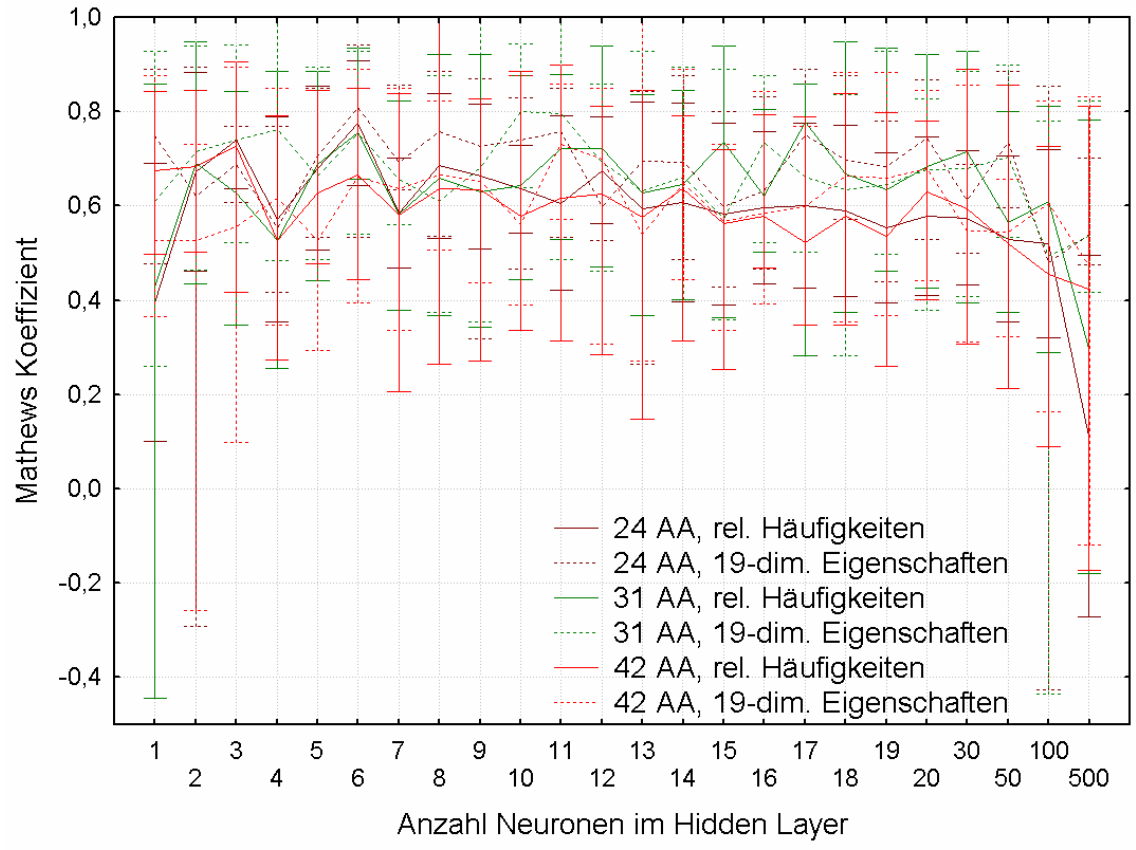

Abbildung 41 – Übersicht über die im Testsatz erzielten Mathews Koeffizienten bei Variation der Anzahl Neuronen von 1 bis 500, Variation des N-terminalen Abschnitts von 24 über 31 zu 42 und Repräsentation der Daten im 20-dimensionalen Aminosäurehäufigkeitsraum und im 19-dimensionalen physikochemischen Eigenschaftsraum. Die Klassifikationsgüte ist – abgesehen von extrem vielen und extrem wenigen Neuronen in Hidden Layer – weitgehend unabhängig von Art der Datenrepräsentation und Länge des verwendeten N-terminalen Abschnitts.

3.4.1.1. Variation der Anzahl Hidden Neuronen im Aminosäurehäufigkeitsraum

In Abbildung 42 sind die Ergebnisse einer 10fachen Kreuzvalidierung mit den relativen Aminosäurehäufigkeiten der ersten 24 Aminosäuren als Inputvektor gezeigt. Tendenziell liefern wenige (2-6) Hidden Neuronen bessere, jedoch von der genauen Anzahl Neuronen stark abhängige Ergebnisse als umfangreichere Hidden Layer. Im Falle eines Neurons sowie von 500 Neuronen im Hidden Layer werden deutlich schlechtere Ergebnisse als bei einer mittleren (2-100) Anzahl Neuronen erzielt. Die genauen Ergebnisse sind in der darauf folgenden Tabelle 5 wiedergegeben.

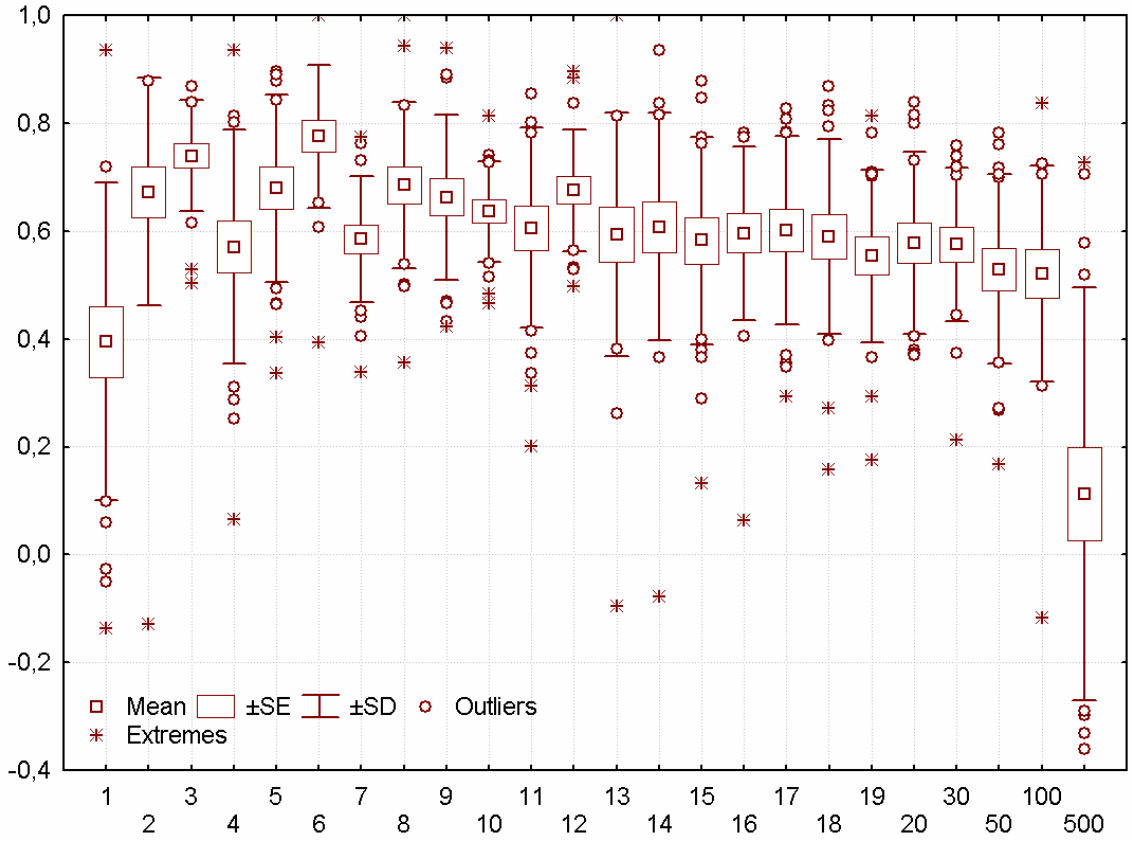

Abbildung 42 – Erzielte Klassifikationsgüte im Testsatz als Mathews-Koeffizient bei Verwendung von 24 Aminosäuren langen Abschnitten, Darstellung im Aminosäurehäufigkeitsraum und Variation der Anzahl Hidden Neuronen von 1 bis 500. Wenige Neuronen im Hidden Layer liefern tendenziell eine besser Klassifikation als umfangreichere versteckte Schichten.

Tabelle 5 - Im Testsatz erzielter Mathews-Koeffizient und Standardabweichung bei einer 10fachen Kreuzvalidierung unter Verwendung der ersten 24 N-terminalen Aminosäuren im Aminosäurehäufigkeitsraum

Neuronen	1	2	3	4	5	6	7	8	9	10	11	12
cc	0,40	0,67	0,74	0,57	0,68	0,78	0,59	0,69	0,66	0,64	0,61	0,68
σ^2	0,30	0,21	0,10	0,22	0,17	0,13	0,12	0,15	0,15	0,09	0,19	0,11

Neuronen	13	14	15	16	17	18	19	20	30	50	100	500
cc	0,59	0,61	0,58	0,60	0,60	0,59	0,55	0,58	0,58	0,53	0,52	0,11
σ^2	0,23	0,21	0,19	0,16	0,17	0,18	0,16	0,17	0,14	0,18	0,20	0,38

Unter Benutzung der ersten 31 Aminosäuren ergibt sich das in Abbildung 43 gezeigte Bild.

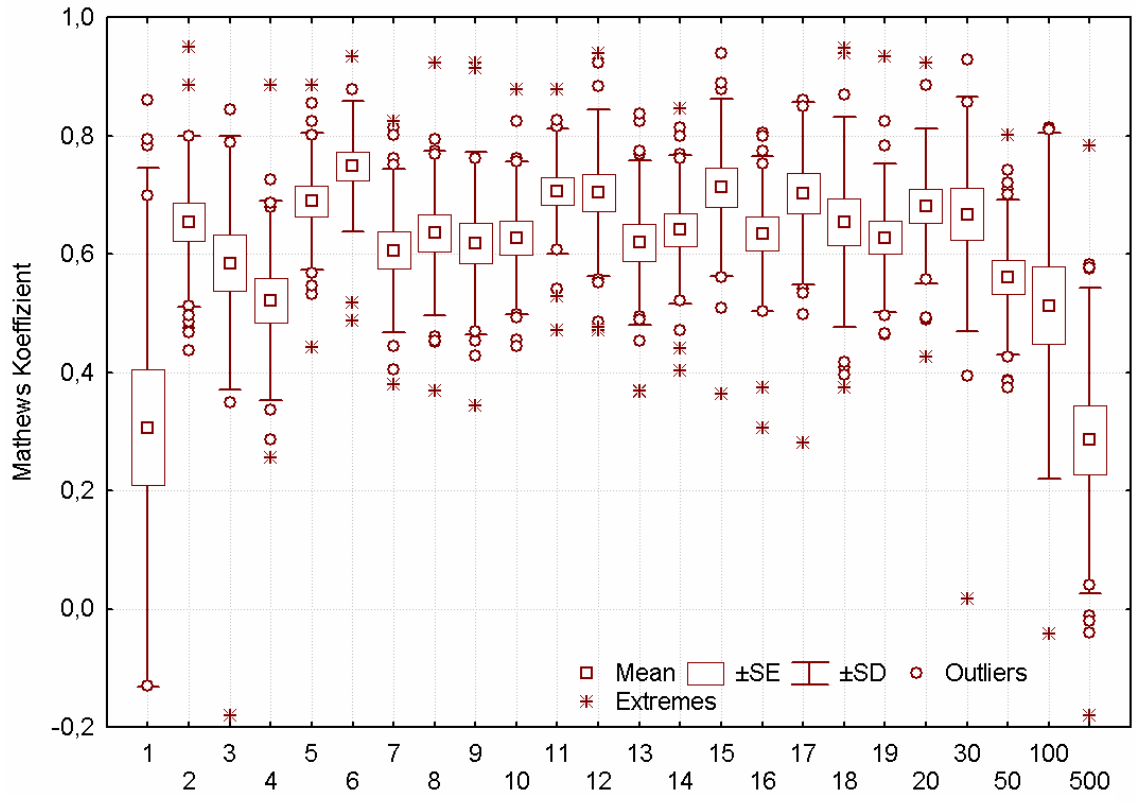

Abbildung 43 – Erzielte Klassifikationsgüte im Testsatz als Mathews-Koeffizient bei Verwendung von 31 Aminosäuren langen Abschnitten, Darstellung im Aminosäurehäufigkeitsraum und Variation der Anzahl Hidden Neuronen von 1 bis 500. Hier sind starke Schwankungen der Klassifikationsgüte bei Variation der Neuronenanzahl zu erkennen.

Tabelle 6- Im Testsatz erzielter Mathews-Koeffizient und Standardabweichung bei einer 10fachen Kreuzvalidierung unter Verwendung der ersten 31 N-terminalen Aminosäuren im Aminosäurehäufigkeitsraum

Neuronen	1	2	3	4	5	6	7	8	9	10	11	12
cc	0,31	0,65	0,58	0,52	0,69	0,75	0,61	0,64	0,62	0,63	0,71	0,70
σ^2	0,44	0,14	0,21	0,17	0,12	0,11	0,14	0,14	0,15	0,13	0,11	0,14

Neuronen	13	14	15	16	17	18	19	20	30	50	100	500
cc	0,62	0,64	0,71	0,63	0,70	0,65	0,63	0,68	0,67	0,56	0,51	0,29
σ^2	0,14	0,13	0,15	0,13	0,15	0,18	0,13	0,13	0,20	0,13	0,29	0,26

Hier zeigt sich keine einheitliche Tendenz, außer einem allgemeinen Qualitätsabfall bei sehr vielen (100, 500) und sehr wenigen (1) Neuronen. Die schlechte Klassifizierung

des Testdatensatzes in diesen Fällen extrem vieler und extrem weniger Neuronen im Hidden Layer spiegelt die ungenügende Parametrisierbarkeit der Trennfunktion wieder – bei zu wenigen Neuronen – bzw. das Überlernen auf die Trainingsdaten bei zu vielen Neuronen.

Das Ergebnis, wen man von den ersten 42 Aminosäuren ausgeht, ist in der folgenden Abbildung 44 gezeigt.
Abermals ist keine einheitliche Tendenz erkennbar. Die schlechte Klassifikationsgüte bei einer sehr großen (50, 100, 500) Anzahl Hidden Neuronen ist auch hier zu sehen. Bei Verwendung von nur einem Neuron im Hidden Layer ist eine erstaunlich gute Klassifikation möglich.

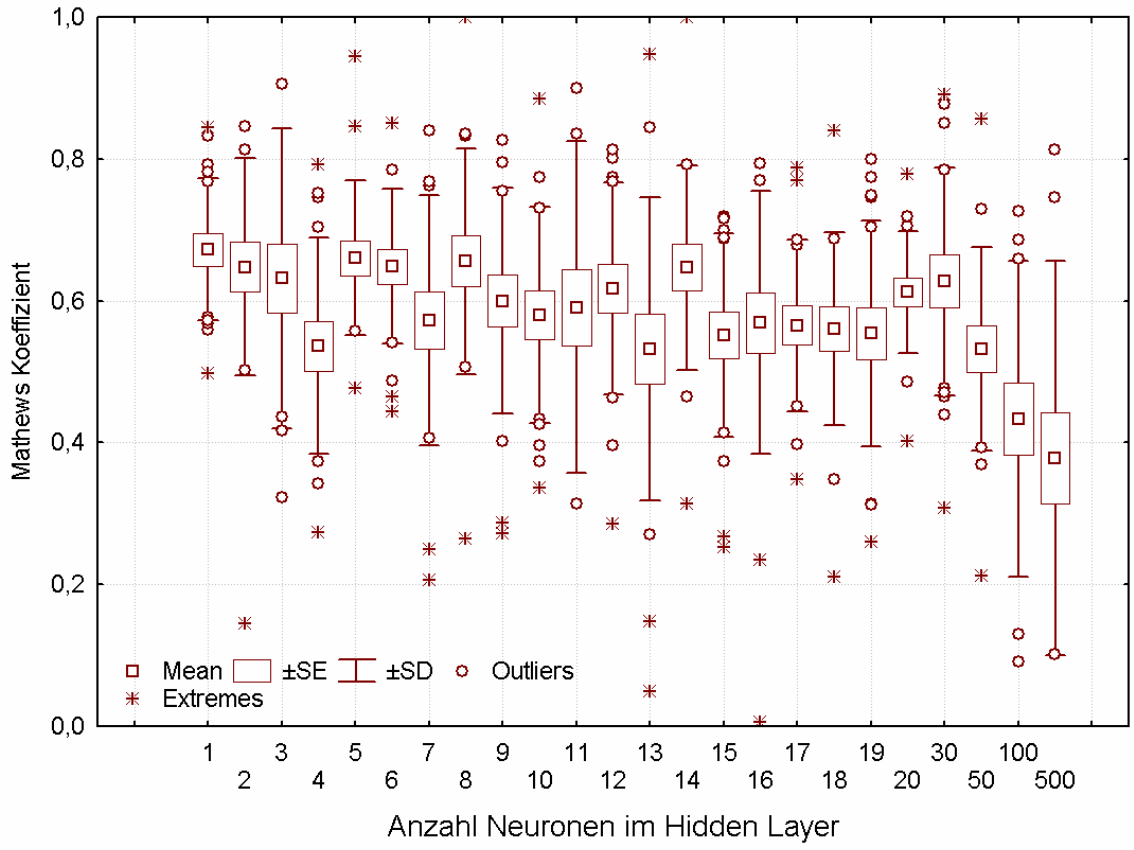

Abbildung 44 – Erzielte Klassifikationsgüte im Testsatz als Mathews-Koeffizient bei Verwendung von 42 Aminosäuren langen Abschnitten, Darstellung im Aminosäurehäufigkeitsraum und Variation der Anzahl Hidden Neuronen von 1 bis 500. Die Klassifikationsgüte ist schlechter als bei Verwendung kürzerer N-terminaler Abschnitte.

Tabelle 7- Im Testsatz erzielter Mathews-Koeffizient und Standardabweichung bei einer 10fachen Kreuzvalidierung unter Verwendung der ersten 42 N-terminalen Aminosäuren im Aminosäurehäufigkeitsraum

Neuronen	1	2	3	4	5	6	7	8	9	10	11	12
cc	0,67	0,65	0,63	0,54	0,66	0,65	0,57	0,66	0,60	0,58	0,59	0,62
σ^2	0,10	0,15	0,21	0,15	0,11	0,11	0,18	0,16	0,16	0,15	0,23	0,15

Neuronen	13	14	15	16	17	18	19	20	30	50	100	500
cc	0,53	0,65	0,55	0,57	0,57	0,56	0,55	0,61	0,63	0,53	0,43	0,38
σ^2	0,21	0,14	0,14	0,19	0,12	0,14	0,16	0,09	0,16	0,14	0,22	0,28

3.4.1.2.Variation der Anzahl Hidden Neuronen im 19-dimensionalen Eigenschaftsraum

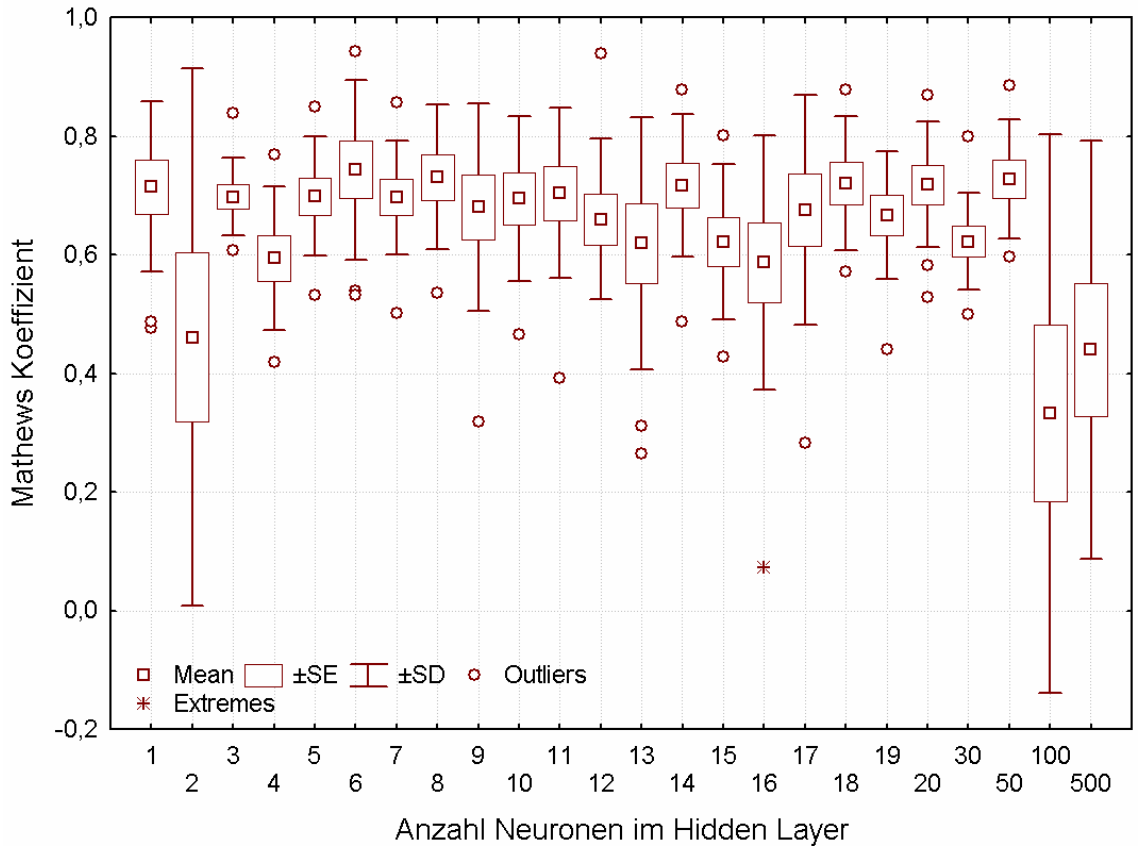

Abbildung 45 – Erzielte Klassifikationsgüte im Testsatz als Mathews-Koeffizient bei Verwendung von 24 Aminosäuren langen Abschnitten, Darstellung im physikochemischen Eigenschaftsraum und Variation der Anzahl Hidden Neuronen von 1 bis 500. Die Klassifikationsgüte ist ähnlich gut wie bei Darstellung im Aminosäurehäufigkeitsraum.

Unter Berücksichtigung der ersten 24 Aminosäuren ist die Klassifikationsgüte in Abbildung 45 gezeigt. Ähnlich wie bei Verwendung der ersten 24 Aminosäuren im Aminosäurehäufigkeitsraum sind starke Schwankungen bei einer kleinen Neuronenzahl zu beobachten. Sehr viele Neuronen im Hidden Layer führen auch hier zu einem Überlernen der Trainingsdaten, so dass die hier verwendeten Testdaten nur mit geringer Qualität klassifiziert werden können.

Tabelle 8 - Im Testsatz erzielter Mathews-Koeffizient und Standardabweichung bei einer 10fachen Kreuzvalidierung unter Verwendung der ersten 24 N-terminalen Aminosäuren im physikochemischen Eigenschaftsraum

Neuronen	1	2	3	4	5	6	7	8	9	10	11	12
cc	0,71	0,46	0,70	0,59	0,70	0,74	0,70	0,73	0,68	0,69	0,70	0,66
σ^2	0,14	0,45	0,06	0,12	0,10	0,15	0,10	0,12	0,17	0,14	0,14	0,14

Neuronen	13	14	15	16	17	18	19	20	30	50	100	500
cc	0,72	0,62	0,59	0,68	0,72	0,67	0,72	0,62	0,73	0,33	0,44	0,72
σ^2	0,12	0,13	0,21	0,19	0,11	0,11	0,11	0,08	0,10	0,47	0,35	0,12

Im Falle der Berücksichtigung der ersten 31 Aminosäuren stellt sich der Sachverhalt wie in Abbildung 46 dar. Hier ist zu beobachten, dass bis hin zu höheren Neuronenzahlen im Hidden Layer (1-17 Neuronen) starke Schwankungen mit der genauen Anzahl der Neuronen auftreten. Ebenfalls ist der Effekt des Überlernens bei hohen (100, 500) Neuronenzahlen zu beobachten.

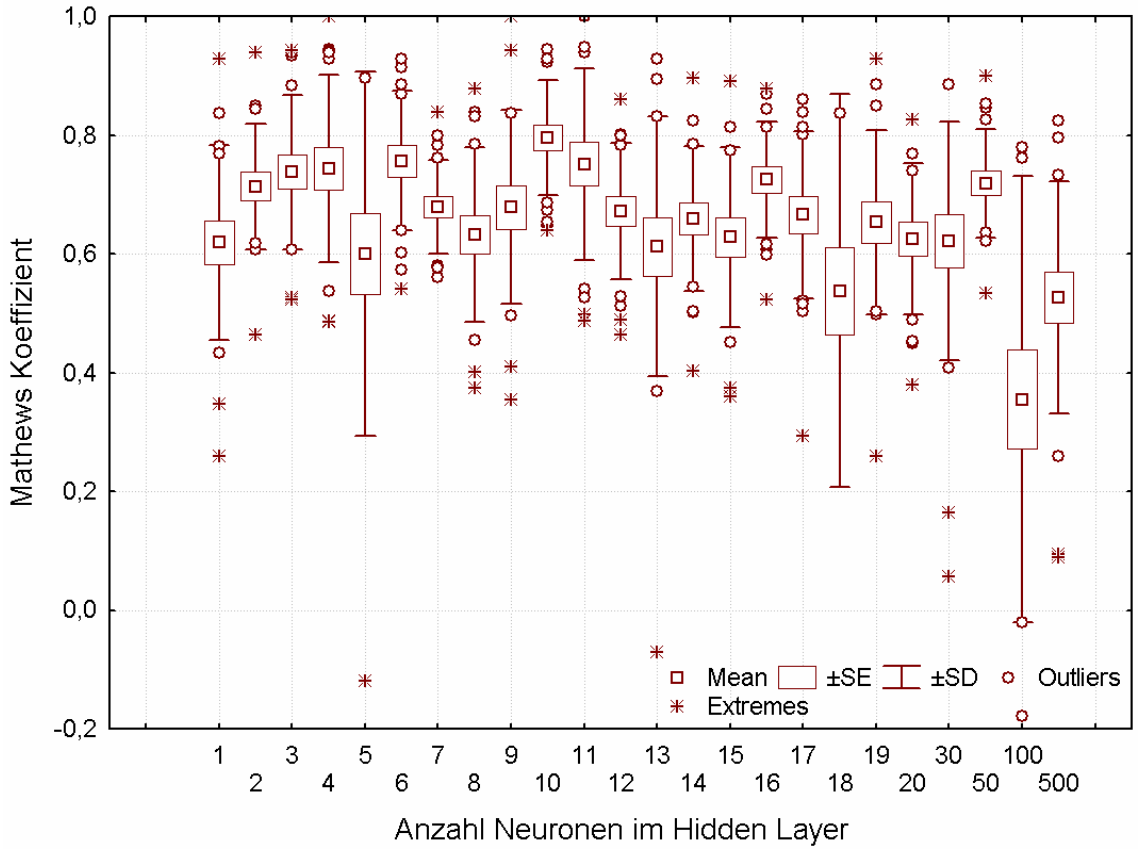

Abbildung 46– Erzielte Klassifikationsgüte im Testsatz als Mathews-Koeffizient bei Verwendung von 31 Aminosäuren langen Abschnitten, Darstellung im physikochemischen Eigenschaftsraum und Variation der Anzahl Hidden Neuronen von 1 bis 500. Die Klassifikationsgüte ist hier stark schwankend.

Tabelle 9 - Im Testsatz erzielter Mathews-Koeffizient und Standardabweichung bei einer 10fachen Kreuzvalidierung unter Verwendung der ersten 31 N-terminalen Aminosäuren im physikochemischen Eigenschaftsraum

Neuronen	1	2	3	4	5	6	7	8	9	10	11	12
cc	0,62	0,71	0,74	0,74	0,60	0,76	0,68	0,63	0,68	0,80	0,75	0,67
σ^2	0,14	0,45	0,06	0,12	0,10	0,15	0,10	0,12	0,17	0,14	0,14	0,14

Neuronen	13	14	15	16	17	18	19	20	30	50	100	500
cc	0,61	0,66	0,63	0,72	0,67	0,54	0,65	0,62	0,62	0,72	0,36	0,53
σ^2	0,12	0,13	0,21	0,19	0,11	0,11	0,11	0,08	0,10	0,47	0,35	0,12

Unter Einbeziehung der ersten 42 Aminosäuren erhalten wir die in Abbildung 47 gezeigten Ergebnisse. Bei geringen (1-2) Neuronenzahlen im Hidden Layer werden

schlechte Klassifikationsergebnisse erzielt. Die Klassifikationsgüte ist durchgängig schwankend und von der genauen Anzahl Neuronen abhängig. Den Effekt des Überlernens können wir auch hier beobachten.

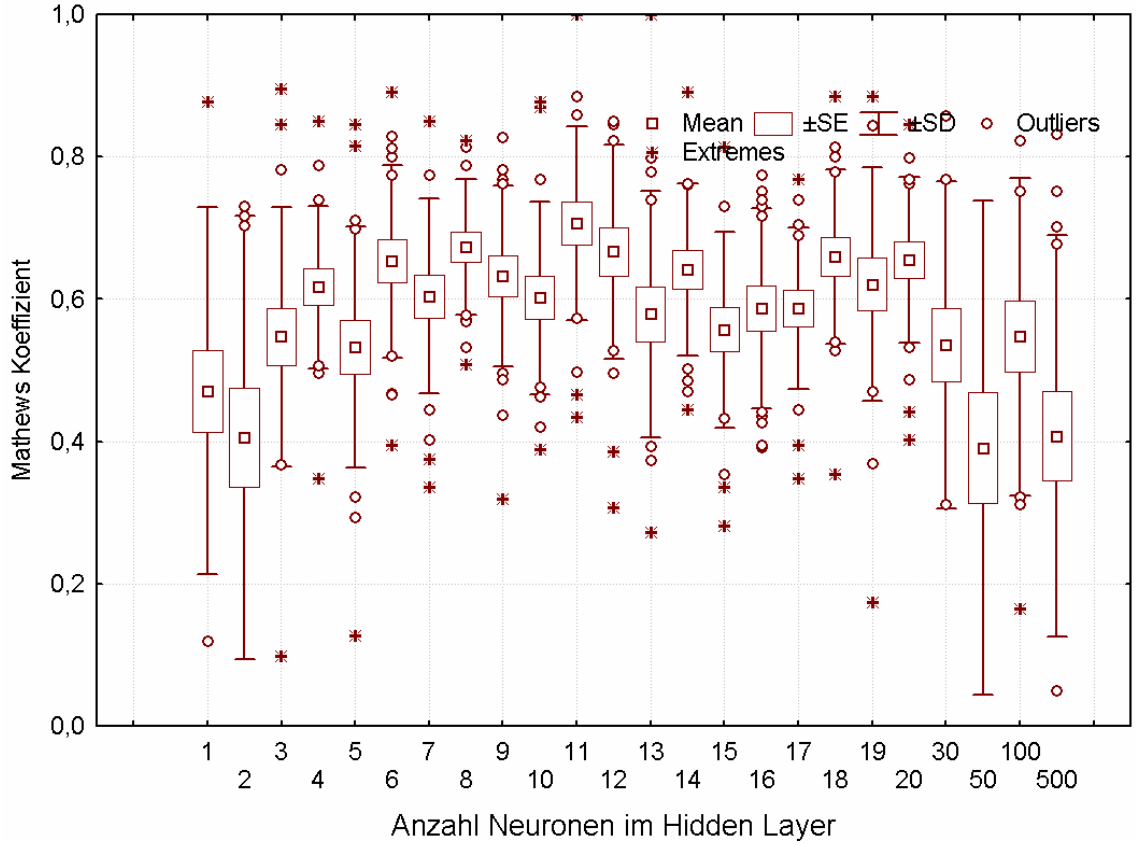

Abbildung 47 – Erzielte Klassifikationsgüte im Testsatz als Mathews-Koeffizient bei Verwendung von 42 Aminosäuren langen Abschnitten, Darstellung im physikochemischen Eigenschaftsraum und Variation der Anzahl Hidden Neuronen von 1 bis 500. Die Klassifikationsgüte ist hier stark schwankend und insgesamt schlechter als bei Verwendung kürzerer N-terminaler Abschnitte.

Tabelle 10 - Im Testsatz erzielter Mathews-Koeffizient und Standardabweichung bei einer 10fachen Kreuzvalidierung unter Verwendung der ersten 42 N-terminalen Aminosäuren im physikochemischen Eigenschaftsraum

Neuronen	1	2	3	4	5	6	7	8	9	10	11	12
cc	0,47	0,41	0,55	0,62	0,53	0,65	0,60	0,67	0,63	0,60	0,71	0,67
σ^2	0,26	0,31	0,18	0,11	0,17	0,13	0,14	0,10	0,13	0,14	0,14	0,15

Neuronen	13	14	15	16	17	18	19	20	30	50	100	500
cc	0,58	0,64	0,56	0,59	0,59	0,66	0,62	0,66	0,54	0,39	0,55	0,41
σ^2	0,17	0,12	0,14	0,14	0,11	0,12	0,16	0,12	0,23	0,35	0,22	0,28

Als Übersicht sind die erzielten Ergebnisse unter Verwendung 24, 31 und 42 Aminosäuren langer N-terminaler Abschnitte im Aminosäurehäufigkeitsraum in der folgenden Tabelle 11 dargestellt. Die bei jeder Abschnittslänge besten Ergebnisse in Bezug auf Mathews-Koeffizient und Standardabweichung sind unterstrichen, die als „bestes Netz“ gewählte Topologie fett markiert.

Tabelle 11 – Übersicht Trainings- und Testklassifikationsergebnisse im Aminosäurehäufigkeits-raum. Die jeweils besten Testergebnisse in jeder Spalte sind unterstrichen, die als Optimum gewählte Topologie fett markiert.

N-terminale Abschnitts-länge (AS)	24	24	24	24	31	31	31	31	42	42	42	42
Datensatz	train	train	test	test	train	train	test	test	train	train	test	test
Anzahl Neuronen	cc	σ^2	cc	σ^2	cc	σ^2	cc	σ^2	cc	σ^2	cc	σ^2
1	0,78	0,16	0,40	0,30	0,37	0,47	0,31	0,44	0,72	0,10	0,67	0,10
2	0,83	0,07	0,67	0,21	0,68	0,14	0,65	0,14	0,69	0,13	0,65	0,15
3	0,78	0,09	**0,74**	**0,10**	0,73	0,21	0,58	0,21	0,65	0,18	0,63	0,21
4	0,84	0,07	0,57	0,22	0,76	0,09	0,52	0,17	0,66	0,08	0,54	0,15
5	0,80	0,06	0,68	0,17	0,79	0,11	0,69	0,12	0,70	0,06	0,66	0,11
6	0,76	0,06	0,78	0,13	0,79	0,08	0,75	0,11	0,72	0,10	0,65	0,11
7	0,78	0,08	0,59	0,12	0,83	0,08	0,61	0,14	0,71	0,09	0,57	0,18
8	0,77	0,06	0,69	0,15	0,77	0,09	0,64	0,14	0,78	0,11	0,66	0,16
9	0,81	0,09	0,66	0,15	0,84	0,08	0,62	0,15	0,71	0,10	0,60	0,16
10	0,85	0,07	0,64	0,09	0,81	0,09	0,63	0,13	0,75	0,09	0,58	0,15
11	0,82	0,08	0,61	0,19	0,84	0,08	0,71	0,11	0,74	0,20	0,59	0,23
12	0,69	0,43	0,68	0,11	0,81	0,08	0,70	0,14	0,74	0,09	0,62	0,15
13	0,79	0,08	0,59	0,23	0,84	0,11	0,62	0,14	0,73	0,13	0,53	0,21
14	0,75	0,09	0,61	0,21	0,84	0,12	0,64	0,13	0,77	0,10	0,65	0,14
15	0,82	0,13	0,58	0,19	0,87	0,10	0,71	0,15	0,64	0,13	0,55	0,14
16	0,74	0,12	0,60	0,16	0,82	0,10	0,63	0,13	0,72	0,09	0,57	0,19
17	0,60	0,44	0,60	0,17	0,84	0,11	0,70	0,15	0,75	0,11	0,57	0,12
18	0,81	0,08	0,59	0,18	0,83	0,09	0,65	0,18	0,75	0,08	0,56	0,14
19	0,82	0,08	0,55	0,16	0,81	0,12	0,63	0,13	0,68	0,09	0,55	0,16
20	0,90	0,07	0,58	0,17	0,83	0,09	0,68	0,13	0,74	0,15	0,61	0,09
30	0,68	0,28	0,58	0,14	0,85	0,14	0,67	0,20	0,80	0,10	0,63	0,16

50	0,79	0,07	0,53	0,18	0,80	0,12	0,56	0,13	0,77	0,14	0,53	0,14
100	0,74	0,16	0,52	0,20	0,69	0,26	0,51	0,29	0,66	0,26	0,43	0,22
500	0,41	0,31	0,11	0,38	0,42	0,33	0,29	0,26	0,51	0,30	0,38	0,28

Als Referenztopologien wurden im Falle von 24 Aminosäuren sowohl im Aminosäurehäufigkeitsraum als auch im 19-dimensionalen Eigenschaftsraum die Netztopologie mit 3 Neuronen im Hidden Layer ausgewählt. Diese erfüllen die drei Kriterien

- wenig Neuronen für gute Verallgemeinerungsfähigkeit
- hohe Klassifierungsgüte und
- geringe Streuung

am ehesten. Die Klassifikationsergebnisse dieser Topologie im Aminosäurehäufigkeits-raum sind in Tabelle 11 fett markiert.

3.4.2. Variation anderer Netzparameter

Ausgehend vom dreilagigen Perzeptron mit 3 Neuronen im Hidden Layer wurde versucht, andere Parameter des Trainingsprozesses zu optimieren. Dazu wurde jeweils ein Parameter der Tabelle 12 variiert und alle anderen mit den oben angegebenen Bedingungen konstant gehalten. Es wurde eine 20fache Kreuzvalidierung mit ansonsten den oben angegebenen Parametern durchgeführt. Die Parameter der Lernalgorithmen waren die Standardparameter von Statistica. (Quick propagation: Learning rate 0,01, Acceleration 2; Delta bar delta: Learning rate 0,01, Increment 0,01, Decay 8, Smoothing 5). Sämtliche verwendeten Bezeichnungen entsprechen zugunsten der Reproduzierbarkeit den in Statistica verwendeten Namen.

Tabelle 12 – Einfluss der Variation von Lernparametern auf die Klassifikationsgüte des trainierten Netzes. Sämtliche Parameter und deren Werte tragen die im Programm Statistica verwendeten Bezeichnungen. Die angegebenen Mathews-Koeffizienten beziehen sich auf den Testsatz.

Variierter Parameter	Standard-parameter war	Parameter geändert zu	Mathews Koeffizient Testsatz (Standardabweichung)
Standardparameter			0,74 (0,10)
Classification Error Function	Sum squared	Entropy	0,59 (0,25)
Learning algorithm	Back propagation	Conjugate Gradient Descent	0,65 (0,19)
		Quasi-Newton	0,51 (0,30)
		Levenberg-Marquardt	0,21 (0,33)
		Quick propagation	0,71 (0,15)
		Delta bar Delta	0,66 (0,18)
Learning rate	0,01	0,001	0,58 (0,23)
Initialisation method	Uniform	Gaussian	0,71 (0,16)
End conditions – Minimal improvement	0,001 / 10000	0,0001/1000	0,72 (0,12)

Classification Threshold	Variabel	Fix 0,5	0,72 (0,12)
Weight Decay	Off	On: Factor 0,01 Scale factor 1	0,61 (0,35)
Learning rate	Fix	Adjust (Learning rate 0,01 -> 0,005, Momentum 0,3 -> 0,1)	0,67 (0,12)

Zusammenfassend lässt sich sagen, dass durch keine der Parameteränderungen eine Verbesserung der Klassifikation zu erreichen war. Daher wurden zum Training des optimalen Netzes die oben angegebenen Standardparameter genutzt. Mögliche Interpretationen der Resultate sind in der Diskussion zu finden.

3.4.3. Genetische Variablenselektion

In der Theorie sollten für die Klassifikation unwesentliche Variablen durch Anpassung der entsprechenden Gewichte (auf einen sehr kleinen Wert) „ausgeblendet“ werden. In der Praxis führt aber die völlige „Abschaltung“ unwesentlicher Variablen durch deren Nichtberücksichtigung of zu einer verbesserten Verallgemeinerungsfähigkeit des Netzes[41]. Hinzu kommt, dass durch einen Inputvektor geringerer Dimensionalität auch die Anzahl der anzupassenden Gewichte verringert wird, was für das Lernverhalten des Netzes von Vorteil ist.

Die hier gewählt genetische Variablenselektion[42] stellt einen sinnvollen Ansatz dar, um die Abhängigkeit des Klassifikationsergebnisses auch von einer Kombination von Eingangsvariablen zu berücksichtigen. Iterativ werden die in jeder Generation die die besten Ergebnisse liefernden Kombinationen an Inputvariablen miteinander kombiniert und einzelne Variablen mutiert (d.h. ein- oder ausgeschaltet). Nun werden abermals die besten Vertreter der Generation („Children“) ermittelt und als neue Elterngeneration („Parents“) verwendet. Durch einen der Evolution nachempfundenen Prozess nähert man sich einem lokalen, im optimalen Fall auch dem globalen Optimum an.

Die Variablenselektion wurde sowohl im Raum der Aminosäurehäufigkeiten als auch im 19-dimensionalen Eigenschaftsraum durchgeführt.

In beiden Datensätzen wurde mit dem Programm Statistica eine genetische Variablenselektion mit 100 Generationen und 100 Kindergenerationen („Children“) pro

Generation durchgeführt. Die Mutationsrate je Iteration wurde auf 1, die Rekombinationswahrscheinlichkeit auf 0,1 gesetzt.

Im Falle der Aminosäurehäufigkeiten wurden die Aminosäuren Cystein, Histidin, Glutamin, Serin, Threonin, Tryptophan und Tyrosin aussortiert. Es blieb damit ein 13-dimensionaler Aminosäurehäufigkeitsvektor als Input übrig. Eine Übersicht über durch die Variablenselektion aussortierte und erhalten bleibende Aminosäuren ist in Tabelle 13 gezeigt.

Tabelle 13 – Durch genetische Variablenselektion entfernte und erhalten bleibende Aminosäuren

Entfernte Aminosäuren	Erhalten bleibende Aminosäuren
Cystein	Alanin
Histidin	Arginin
Glutamin	Asparaginsäure
Serin	Asparagin
Threonin	Glutaminsäure
Tryptophan	Glutamin
Tyrosin	Glycin
	Isoleucin
	Leucin
	Methionin
	Phenylalanin
	Prolin
	Valin

Vergleich der Testperformance mit Standardparametern liefert zu dem vollständigen Vektor praktisch identische Klassifikationsgüte mit einem Mathews-Koeffizienten von cc=0,76 (Standardabweichung 0,06) in einer 10fachen Kreuzvalidierung. Mit dem vollständigen Inputvektor wurde ein Wert von cc=0,74 erreicht.
Im Falle des 19-dimensionalen Eigenschaftsraumes wurden, unter gleichen Parametern für die genetische Variablenselektion, die Variablen 9, 11, 16, 18 und 19 aussortiert. Es blieben 14 Dimensionen übrig, die bei einer Klassifizierung mit Standardparametern einen Mathews-Koeffizienten von cc=0,67 (Standardabweichung 0,11) erreichten.
Unter Benutzung des vollständigen Inputvektors wurde im Mittel ein Wert von cc=0,70 erreicht.

3.4.4. Training der optimalen Netze mit allen Daten

Als optimal wurde eine Netztopologie von drei Neuronen im Hidden Layer, kombiniert mit jeweils den relativen Aminosäurehäufigkeiten oder dem 19-dimensionalen Eigenschaftsvektor der ersten 24 Aminosäuren ermittelt. Die erzielten Klassifikationsgüten sind als Mathews-Koeffizienten in Tabelle 14 dargestellt.

Tabelle 14 – Erzielte Klassifikationsergebnisse der im Weiteren verwendeten neuronalen Netze im Trainings- und im Testsatz. Diese Netze wiesen nicht den besten Mathews-Koeffizienten aller Netze auf, besaßen jedoch zur gleichen Zeit eine gute Klassifikation, geringe Standardabweichung der Klassifikationsgüte und eine möglichst geringe Anzahl Neuronen im Hidden Layer auf.

	cc (Trainingssatz)	cc (Testsatz)
20-dimensionaler Aminosäurehäufigkeitsraum	0,78 (0,09)	0,74 (0,10)
19-dimensionaler Physikochemischer Eigenschaftsram	0,79 (0,07)	0,70 (0,06)

Diese beiden Netze wurden nun mit allen verfügbaren Daten im Trainingssatz trainiert. In einem ersten Schritt musste das Lernverhalten der Netze untersucht werden, um nach einer geeigneten Anzahl Epochen das Training abzubrechen. Zu diesem Zweck wurde der Datensatz von 175 Sequenzen stets in einen Trainingssatz von 89 zufällig ausgewählten Sequenzen, einen Select-Datensatz von 43 und einen Testdatensatz von ebenfalls 43 Sequenzen eingeteilt. Dem Netz wurden in jeder Iteration die Trainingsdaten präsentiert und der Lernfortschritt mit dem Select-Datensatz verfolgt. Um ein Überlernen auf die Trainingsdaten zu vermeiden, wurden für diese Zwecke getrennte Datensätze verwendet. Wenn keine (weniger als 0,1%) Verbesserung im Bruchteil korrekt klassifizierter Datensätze innerhalb 1000 Iteration eintrat, wurde das bis zu diesem Zeitpunkt optimale Netz und die zu seinem Training nötige Anzahl Lerniterationen ermittelt. Verwendet wurden die durch Variablenreduktion bereinigten Datensätze mit den Standard-Trainingsparametern.

Im Fall der Repräsentation der Daten im 13-dimensionalen Aminosäurehäufigkeitsraum (nach Variablenreduktion, siehe Tabelle 13) fand der Halt des Netztrainings mit dem genannten Abbruchkriterium (eine Verbesserung des Anteils korrekt klassifizierter Daten von 0,001 auf 1000 Epochen im Select-Datensatz) nach 281, 461, 413, 275, 358,

388, 456, 184, 290 und 1254 Epochen statt. Mit den im Durchschnitt 436 Epochen und 132 Datensätzen um Trainings- und 42 Datensätzen im Testdatensatz wurde eine 20fache Kreuzvalidierung durchgeführt. Diese liefert für das finale Netz einen Mathews Koeffizienten von cc=0,69 (Standardabweichung 0,13). Beim Training mit allen 175 Datensätzen erreichte das Netz einen Mathews Koeffizienten von 0,92 in der Reklassifikation. Die Sensitivität beträgt 98%, die Selektivität 91%. Es wurden die in Tabelle 15 dargestellten Ergebnisse erzielt.

Tabelle 15 – Klassifizierungsergebnisse des Netzes mit der bestmöglichen Klassifikationsgüte, basierend auf Aminosäurehäufigkeiten

Klassifikation als	Anzahl zugeordneter Aminosäuresequenzen	Name
Positiv	39	
Negativ	131	
Overpredicted	4	M1 Family Aminopeptidase Clathrin Coat Assembly Protein Vacuolar Proton Pumping Pyrophosphatase-2 Knob-associated histidine rich protein (KAHRP)
Underpredicted	1	Fumarase Class 1

Im Falle des 14-dimensionalen Eigenschaftsraumes fand der Halt des Netztrainings mit dem Standard-Abbruchkriterium (mindestens 0,001 Verbesserung auf 1000 Epochen im Select-Datensatz) nach 515, 993, 317, 843, 1002, 538, 396, 485, 659 und 247 Epochen statt. Mit den im Durchschnitt 600 Epochen und 132 Datensätzen um Trainings- und 42 Datensätzen im Testdatensatz wurde eine 20fache Kreuzvalidierung durchgeführt. Diese liefert für das finale Netz einen Mathews Koeffizienten von cc=0,76 (Standardabweichung 0,13). Beim Training mit allen 175 Aminosäuresequenzen erreichte das Netz einen Mathews Koeffizienten von 0,79, eine Sensitivität von 93% und eine Selektivität von 77%.

Tabelle 16– Klassifizierungsergebnisse des Netzes mit der bestmöglichen Klassifikationsgüte, basierend auf dem physikochemischen Eigenschaftsraum

Klassifikation als	Anzahl Datensätze	Namen
Positiv	37	
Negativ	124	
Overpredicted	11	M1 Family Aminopeptidase Thioredoxin Reductase Cytosolic Triosephosphate Isomerase Ubiquitin Conjugating Enzyme E2 Clathrin Coat Assembly Protein Vacuolar Proton Pumping Pyrophosphatase-2 Vacuolar ATPase Subunit B Knob-associated hisitidine rich protein (KAHRP) Epithelial membrane protein-3 (EMP3) Epidermical surface anitgen (ESA) Pyruvate Dehydrogenase 2
Underpredicted	3	Fumarase Class 1 50S Ribosomal Protein L2 Ubiquinone Biosynthesis Protein COQ4

Zusammenfassend lässt sich sagen, dass entgegen „chemischer Intuition“ unter Verwendung relativer Aminosäurehäufigkeiten der ersten 24 N-terminalen Aminosäuren bessere Ergebnisse als bei Repräsentation der Daten durch physikochemische Eigenschaften erzielt werden. Ein Vergleich mit etablierten Methoden ist in der Diskussion zu finden.

3.4.5. Optimierung auf möglichst wenig falsch-positive Vorhersagen

Im vorangegangen Abschnitt wurde der Klassifikationsschwellwert, der Aktivierung des Ausgabeneurons, die zwischen positiv und negativ klassifizierten Aminosäuresequenzen unterscheidet, unter gleichgewichteter Berücksichtigung von

falsch-positiven und falsch-negativen Datensätzen optimiert. Die relative Strafe von falsch-positiven zu falsch-negativen Datensätzen wurde also auf 1 gesetzt. Allerdings ist es für den Nutzer des Netzes vorteilhaft, möglichst wenig falsch-positive Ergebnisse zu erhalten. Daher wurde die genannte relative Strafe auf 3 gesetzt, womit nur noch eine (im Fall der Aminosäurehäufigkeiten) bzw. zwei (im Fall des Eigenschaftsraums) Sequenzen falsch-positiv vorhergesagt wurden. Die Trainingszeit entsprach mit 436 Präsentationen des Trainingssatzes derjenigen des vorangegangenen Abschnitts. Im Fall der Aminosäurehäufigkeiten wurden folgende Ergebnisse mit einem Mathews Koeffizienten von cc = 0,51 erzielt. Die Sensitivität wurde hier gegen erhöhte Selektivität eingetauscht – es konnte eine Sensitivität von 35% und eine Selektivität von 93% erreicht werden. Es werden zwar wesentlich weniger der mitochondrialen Transitpeptide vom neuronalen Netz erkannt, jedoch beschreibt eine vom Netz getroffene positive Aussage mit einer höheren Wahrscheinlichkeit tatsächlich eine mitochondriales Transitpeptid. Die Klassifikationsergebnisse sind in Tabelle 17 dargestellt.

Tabelle 17 – Klassifikationsergebnisse des auf eine geringe Anzahl falsch-positiver Datensätze trainierten Netzes bei Darstellung im Aminosäurehäufigkeitsraum

Klassifikation als	Anzahl Datensätze	Namen
P	14	
N	134	
O	1	Vacuolar Proton Pumping Pyrophosphatase-2
U	26	

Im Fall des 15-dimensionalen physikochemischen Eigenschaftsraumes nach Variablenselektion wurde ein Mathews Koeffizient von cc=0,49 erzielt. Die Sensitivität beträgt wie bei Benutzung der relativen Aminosäurehäufigkeiten ebenfalls 35%, die Selektivität 88%. Die Klassifikationsergebnisse sind in Tabelle 18 dargestellt.

Tabelle 18 – Klassifikationsergebnisse des auf eine geringe Anzahl falsch-positiver Datensätze trainierten Netzes bei Darstellung im physikochemischen Eigenschaftsraum

Klassifikation als	Anzahl Datensätze	Namen
P	14	
N	133	
O	2	Vacuolar Proton Pumping Pyrophosphatase-2 Knob-associated hisitidine rich protein (KAHRP)
U	26	

Abschließend lässt sich sagen, dass sich die Anzahl falsch-positiv klassifizierter Aminosäuresequenzen zwar senken lässt, dies aber deutlich auf Kosten der Sensitivität geht, also viele tatsächlich mitochondriale Transitpeptide vom Netz nicht erkannt werden.

3.5. Genome Scanning *P. falciparum*

Im vorangegangenen Abschnitt wurden vier neuronale Netze mit jeweils drei Neuronen im Hidden Layer trainiert. Als Eingabedaten wurden relative Aminosäurehäufigkeiten und physikochemische Eigenschaftsvektoren gewählt, und jedes dieser Netze wurde zum einen auf einen optimalen Mathews-Koeffizienten, zum anderen auf eine geringe Anzahl falsch-positiv klassifizierter Aminosäuresequenzen trainiert (größere relative Strafe für falsch-positive Vorhersagen, „stringentes Netz"). Damit erhielten wird vier Netze, die im Weiteren verwendet wurden.

Mit diesen Netzen wurden nun die vorhergesagten offenen Leserahmen (Open Reading Frames) aus *P. falciparum* gescannt. Es kamen folgende Resultate zustande.

Tabelle 19 – Anzahl der mit den verschiedenen Netzen vorhergesagten mitochondrialen Transitpeptide bei Verwendung der vorhergesagten Open Reading Frames des Genoms von *P. falciparum*

Verwendetes Netz	AA-Häufigkeiten, optimaler cc	Eigenschafts-raum, optimaler cc	AA-Häufigkeiten, stringentes Netz	Eigenschafts-raum, stringentes Netz
Mitochondriales Transitpeptid	**2449**	3564	903	1239
Andere Sequenz	7827	6712	9373	9037

706 der Sequenzen haben sowohl das „stringentere" Netz basierend auf Aminosäurehäufigkeiten als auch das „stringentere" Netz basierend auf den Eigenschaftsvektoren als mitochondriale Transitsequenz passiert. Diese Sequenzen können mit großer Wahrscheinlichkeit als mitochondriale Transitpeptide klassifiziert werden.

Als beste Schätzung – womit die Verwendung des Netzes, das in der Kreuzvalidierung den besten Mathews-Koeffizienten von im Mittel cc = 0,74 im Testsatz erzielt hat, gemeint ist - ergeben sich 2449 mitochondriale Transitpeptide (ausgehend von den 10276 Open Reading Frames ergibt dies 23,8% des Kerngenoms). Dieser Anteil liegt etwa doppelt so hoch wie bei anderen Eukaryonten (näheres siehe Diskussion).

4. Diskussion

4.1. Vorhersage mit etablierten Methoden

4.1.1. MitoProtII

Dieses Programm erzielte unter Verwendung der 175 Aminosäuresequenzen einen Mathews-Koeffizienten von cc = 0,49. Dieser Wert ähnelt dem von den Autoren dieser Methode erzielten Ergebnis von cc = 0,46 bei einer Kreuzvalidierung unter Verwendung menschlicher Aminosäuresequenzen. Diese Klassifikationsgüte ist, aufgrund der oben genannten Unterschiede in Bezug auf die Aminosäurezusammensetzung von N-terminalen Abschnitten von Proteinen aus *P. falciparum* und aus anderen Eukaryonten, bemerkenswert gut.
Auf der anderen Seite erkennt das Programm zwar 32 der 40 mitochondrialen Transitpeptide und besitzt so eine Sensitivität von 80%, kategorisiert aber auch 36 der nicht-mitochondrialen Transitpeptide als (falsch-)positiv. Somit ergibt sich zwar ein annehmbarer Mathews-Koeffizient, der aber durch die hohe Anzahl falsch-positiver gegenüber korrekt-positiver Datensätze (36 zu 32) nur eine Selektivität von weniger als 50% (47%) zur Folge hat. MitoProtII erkennt viele (80%) der mitochondrialen Transitpeptide, aber nur hinter weniger als der Hälfte (47%) der positiv klassifizierten Aminosäuresequenzen verbirgt sich tatsächlich ein mitochondriales Transitpeptid.

4.1.2. TargetP

Unter Anwendung von TargetP wurden deutlich bessere Klassifikationsergebnisse bei Anwendung der auf Pflanzenzellen („plant“ - Option aktiviert) trainierten Netze als bei Anwendung der „non-plant“ - Netze erzielt. In erstem Fall wurde ein Mathews-Koeffizient von 0,60 erhalten, gegenüber einem Wert von 0,42 bei Anwendung eines auf nicht-pflanzliche, eukaryontische Zellen trainierten neuronalen Netzes. Der Unterschied kommt durch eine bessere Erkennung der mitochondrialen Transitpeptide durch das mit Pflanzensequenzen trainierte neuronale Netze zustande – es wurden 22 der 40 Sequenzen erkannt, im non-plant-Fall nur 14 von 40 Sequenzen. Dies ergibt Sensitivitäten von 55% bzw. 35%. Bei in beiden Fällen fünf falsch-positiv klassifizierten Sequenzen ergibt sich eine Selektivität von 81% bzw. 14 74%.

Durch die besseren Klassifikationsergebnisse des auf pflanzliche Sequenzen trainierten Netzwerkes scheinen die hier vorliegenden Transitpeptide von *P. falciparum* eher

pflanzlichen als nicht-pflanzlichen Transitpeptiden zu ähneln. Dieses Ergebnis könnte damit zusammenhängen, dass *P. falciparum* einen Apicoplasten enthält, der zwar im Gegensatz zu pflanzlichen Plastiden wie z.B. den Chloroplasten keine Aufgaben der Photosynthese wahrnimmt. Allerdings könnte durch den Apicoplasten das interne Protein - Targetingsystem von *P. falciparum* einer Evolution unterworfen worden sein, die dessen Targetingsystem nun ähnlich dem von Pflanzenzellen werden lässt. Zusammenfassend lässt sich sagen, dass die Anwendung von Neuronalen Netzen im Fall des auf pflanzliche Sequenzen trainierten Netzwerkes von TargetP sowohl dem auf nicht-pflanzliche Sequenzen trainierten Netz als auch der Diskriminierungsfunktion von MitoProtII deutlich überlegen ist.

4.2. Aminosäurehäufigkeiten / PCA / SOM

Mitochondriale Transitpeptide aus *Plasmodium falciparum* weisen zwar, ebenso wie mitochondriale Transitpeptide anderer Eukaryonten, einen erhöhten Anteil positiv geladener Aminosäuren auf. Ein entscheidender Unterschied ist allerdings die Bevorzugung von Lysin durch *P. falciparum* (16% vs. 4% bei anderen Eukaryonten) im Vergleich zu Arginin, das bei anderen Eukaryonten bevorzugt wird (12% vs. 8% in *P. falciparum*).

Die Aminosäurehäufigkeitsverteilungen in *P. falciparum* folgen auch in den N-terminalen Abschnitten im Allgemeinen der in Abbildung 15 gezeigten Aminosäurehäufigkeitsverteilung im Gesamtgenom dieses Organismus'. Diese Abbildung ist in der folgenden Abbildung 48 wiederholt dargestellt.

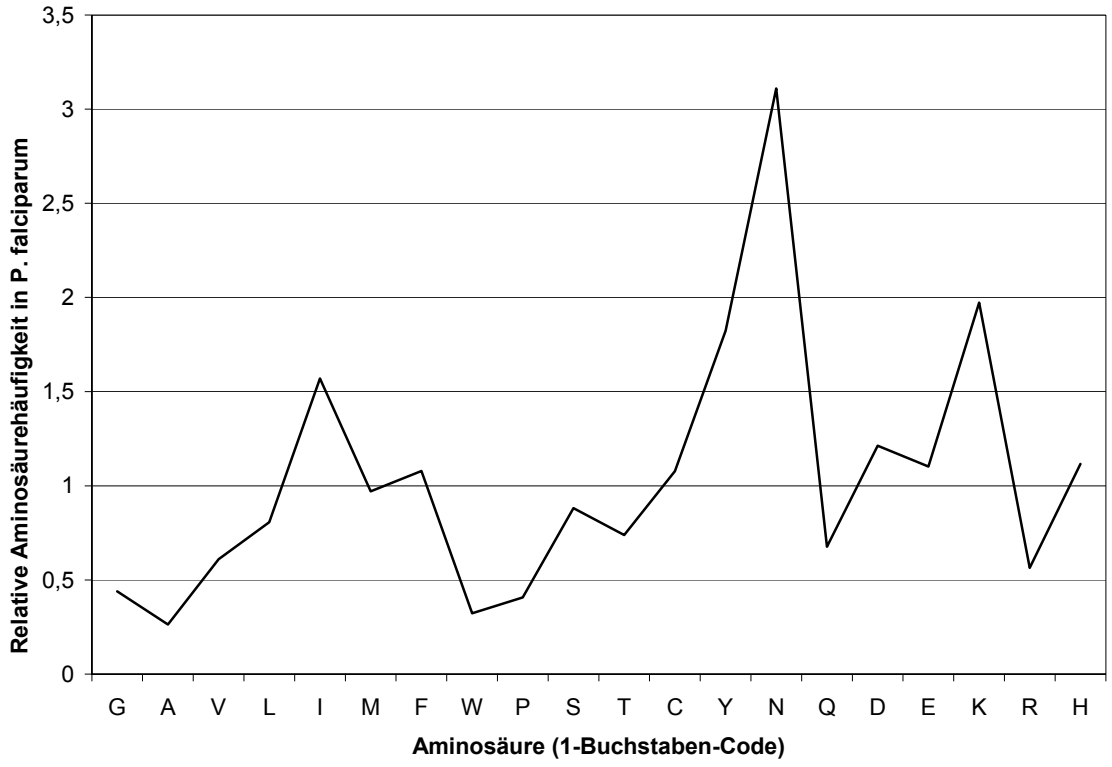

Abbildung 48 – Aminosäurehäufigkeitsverteilung in *P. falciparum*, relativ zu Aminosäurehäufigkeiten in der SwissProt, Version 36. Werte über 1 zeigen einen häufigeren Einbau der Aminosäure in *P. falciparum* als im durchschnittlichen Protein der SwissProt Version 36 an, Werte unter 1 eine weniger häufige Benutzung.

Einige Aminosäuren weichen in ihrer Häufigkeit im Genom von *P. falciparum* stark von der Häufigkeit in der SwissProt-Datenbank, Version 36, ab. Glycin, Alanin, Prolin und Arginin werden wesentlich seltener (Quotient < 0,6) als in der SwissProt

verwendet, Isoleucin, Tyrosin, Asparagin und Lysin wesentlich häufiger (Quotient > 1,5) als in der Referenzdatenbank.
Lobry hat für einen Satz von 59 Bakteriengenomen den Zusammenhang zwischen CG-Gehalt der DNA und relativer Aminosäurehäufigkeit der reifen Proteine untersucht und dabei in einigen Fällen Abhängigkeiten der beiden Größen voneinander festgestellt[43]. Ob diese Übereinstimmung hier wiederzufinden ist, sollte überprüft werden.
Der Gehalt an Guanin und Cytosin wurde aus der in der PlasmoDB-Datenbank angegebenen Codon Usage der codierenden Regionen der Chromosomen 2 und 3 bestimmt. Guanin und Cytosin (C+G) kommen in dieser Stichprobe auf 24,0%, Adenosin und Thymin (A+T) auf 76,0% der Nukleinsäuren. Dieser Wert von 24%, also nicht einmal einem Viertel der der Basen C+G, liegt am unteren Ende der von Lobry berichteten Anteile dieser Basen zwischen 25% und 75%, dort allerdings in bakteriellen Genomen.
Lobry stellte einen streng monotonen Verlauf der Interpolationsfunktion zwischen G+C-Gehalt und Aminosäurehäufigkeit in neun der 20 betrachteten proteinogenen Aminosäuren fest. Nur diese neun Fälle sollen im Folgenden betrachtet werden. Grund dafür ist, dass ein streng monotoner Funktionsverlauf eine durchgehende Tendenz in der Veränderung des Aminosäuregebrauchs widerspiegelt und somit verlässlichere Vorhersagen ermöglichen sollte.
Der G+C-Gehalt des Genoms von *P. falciparum* ist mit etwa 25% sehr niedrig. Da in den neun betrachteten Fällen eine starke, streng monotone Abhängigkeit der Aminosäurehäufigkeit vom G+C-Gehalt beobachtet werden konnte, erwarten wir

- eine hohe relative Häufigkeit im Falle streng monoton fallender Interpolationsfunktion sowie
- eine niedrige relative Häufigkeit im Falle streng monoton steigender Interpolationsfunktion.

In acht Fällen der betrachteten neun Aminosäuren, mit Ausnahme von Leucin, weicht der in *P. falciparum* beobachtete Gehalt bestimmter Residuen tatsächlich stark (Quotient der Häufigkeiten kleiner 0,6 oder größer 1,5) von der Häufigkeit in der SwissProt, Version 36 ab. Dieser Sachverhalt ist in der folgenden Tabelle 20 dargestellt. Damit kann die von Lobry bei bakteriellen Genomen beobachtete Abhängigkeit zwischen G+C-Gehalt und Aminosäurehäufigkeit im hier betrachteten Spezialfall bestätigt werden.

Tabelle 20 – Vergleich der von Lobry beobachteten Abhängigkeit der relativen Aminosäurehäufigkeit vom G+C-Gehalt des Genoms mit den im Gesamtgenom von *P. falciparum* beobachteten Aminosäurehäufigkeiten. Aufgrund des sehr geringen C+G-Gehalts des Genoms von *P. falciparum* wird bei einer steigenden Funktion eine seltene Verwendung der betreffenden Aminosäure erwartet, und vice versa. Einzig Leucin tritt trotz des hohen G+C-Gehalts in etwa so häufig wie in der SwissProt, Version 36, auf und damit seltener als erwartet.

Aminosäuren, deren Anteil am Proteom nach Lobry streng monoton vom C+G-Gehalt der DNA abhängt	Funktions-verlauf, streng monoton	Aminosäuren, die auffällig selten (Quotient < 0,6) oder auffällig häufig (Quotient >1,5) vertreten sind	Häufigkeit
Alanin	Steigend	Alanin	Selten
Arginin	Steigend	Arginin	Selten
Glycin	Steigend	Glycin	Selten
Prolin	Steigend	Prolin	Selten
Asparagin	Fallend	Asparagin	Häufig
Isoleucin	Fallend	Isoleucin	Häufig
Leucin	Fallend		
Lysin	Fallend	Lysin	Häufig
Tyrosin	Fallend	Tyrosin	Häufig

Den Hauptkomponentenanalysen sowie selbstorganisierenden Karten in Aminosäurehäufigkeits- ebenso wie im physikochemischen Eigenschaftsraum war eine tendenzielle, aber keinesfalls vollständige oder eindeutige Trennung zwischen den verschiedenen Datensatzgruppen unterschiedlicher Lokalisationen gemein. Die Unterschiede zwischen Proteinen aus *P. falciparum* und solchen aus anderen Organismen waren durchweg stärker ausgeprägt als diejenigen zwischen Proteinen unterschiedlicher Lokalisationen im selben Organismus. Dies ist auf die Berücksichtigung von nur zwei Komponenten im Falle der Hauptkomponentenanalysen sowie durch die große Varianz innerhalb der Datensätze im Falle der Selbstorganisierenden Karten zu erklären. Für eine Trennung im höherdimensionalen Raum – durch Berücksichtigung der jeweiligen vollständigen Inputvektoren – waren die Datensätze hinreichend unterschiedlich.

4.3. Ermittlung der optimalen Struktur des Neuronalen Netzes

4.3.1. Variation der Neuronenanzahl

Unter Verwendung von 24, 31 und 42 Aminosäuren langen N-terminalen Abschnitten und bei Repräsentation der Daten sowohl im Aminosäurehäufigkeitsraum als auch im physikochemischen Eigenschaftsraum werden bei sehr wenigen (1-2) und sehr vielen (100/500) verwendeten Neuronen im Hidden Layer schlechte Ergebnisse bei der Klassifizierung der Testdaten erzielt. Dies ist bei sehr wenigen Neuronen darauf zurückzuführen, dass die Funktionsapproximation nicht genügend variable Parameter zur Adaption an die Trainingsdaten besitzt. Die Testdaten werden durch die vorhergehende ungenügende Approximation ebenfalls nicht optimal klassifiziert. Bei sehr vielen Neuronen im Hidden Layer ist zwar eine sehr gute Anpassung an die Trainingsdaten möglich, allerdings verliert das Netz seine Verallgemeinerungsfähigkeit durch ein „Überlernen“ der Daten, d.h. durch eine exakte Anpassung der Klassifizierungsgrenze an die gelernten Daten. Dadurch ist bei sehr vielen Neuronen im Hidden Layer ebenfalls – wie in der Theorie erwartet – eine schlechtere Klassifikationsgüte, gemessen als Mathews-Koeffizient, festzustellen.

4.3.2. Variation der Trainingsparameter

Die Trainingsparameter wurden nicht systematisch, sondern nur stichprobenartig variiert. Ebenso wurde zu jedem Zeitpunkt nur ein Parameter geändert. Zur systematischen Optimierung existierten zum einen zu viele Parameter. Zum anderen erzielte auch keine einzelne Optimierung eine Verbesserung der Klassifizierung, so dass die Standardparameter offensichtlich recht sinnvoll gewählt wurden. Die Standardparameter ergaben einen Mathews-Koeffizienten von cc=0,74.
Eine Änderung der Fehlerklassifizierung von der Summe der Quadrate zur Optimierung nach der größten Wahrscheinlichkeit („Entropy“) liefert keinen Vorteil (cc=0,59). Grundlage für diesen Schritt wäre auch eine Verteilungsfunktion der Daten aus der Exponentialfunktionsfamilie (wie z.B. die Gauss’sche Normalverteilung). Diese Annahme ist theoretisch weder zu widerlegen noch zu begründen, im vorliegenden Fall liefert sie keinen Gewinn.
Die Verwendung des Conjugate Gradient Descent[44] als Trainingsmethode lieferte ebenfalls eine Verschlechterung des Mathews-Koeffizienten (cc=0,65). In der Theorie ist dieser Algorithmus zwar der Back-Propagation überlegen, wird aber gleichzeitig

auch erst bei einem größeren Anzahl (> mehrere 100) Gewichte und/oder multiplen Ausgabeneuronen empfohlen[45]. Beide Bedingungen waren hier nicht erfüllt.
Die Verwendung des Quasi-Newton-Algorithmus war ebenfalls nicht zu empfehlen (cc=0,51). Er ist zwar für weniger als mehrere hundert Gewichte empfohlen, allerdings auch für mehrere Ausgabeneuronen. Dieser Fall lag hier nicht vor.
Der Levenberg-Marquardt-Algorithmus lieferte einen sehr niedrigen Mathews-Koeffizienten (cc=0,21). Er ist empfohlen für Netze mit einem Ausgabeneuron, kleine Netze und die Verwendung der Fehlerfunktion nach der Summe der Quadarate. Diese Erfordernisse sind sämtlich erfüllt. Allerdings geht dieser Algorithmus vereinfachend von einer linearen Trennfunktion zur Ermittlung der Schrittrichtung aus – eine Bedingung, die hier offensichtlich nicht erfüllt war.
Quick-propagation führte zu praktisch der gleichen Qualität (cc=0,71) wie der standardmäßige Back-Propagation-Algorithmus. Der Unterschied zu Back-Propagation besteht darin, dass er in jeder Epoche den durchschnittlichen Gradienten der Fehlerfläche für alle Fälle berechnet und nur nach der gesamten Epoche die Gewichte adaptiert. Back-Propagation dagegen passt die Gewichte nach jedem präsentierten Datensatz an. Ausserdem geht Quick-Propagation von einer am Minimum lokal quadratischen Form der Fehleroberfläche aus – eine Bedingung, die i. allg. nicht erfüllt ist, aber auch bei leicht abweichenden Fällen gute Ergebnisse produziert[44].
Eine Änderung der Lernrate von 0,01 auf 0,001 führte zu einer Verschlechterung des Ergebnisses (cc=0,58). Vermutlich konnte der Algorithmus unter diesen Bedingungen nicht in sinnvoller Zeit an die Daten adaptieren.
Eine Änderung der Initialisierung der Gewichte von Gleichverteilung auf Gauß-Verteilung führte zu praktisch dem gleichen Ergebnis (cc=0,71) wie die Standardparameter und war daher ohne Einfluss auf das Ergebnis.
Eine Erhöhung des Abbruchkriteriums von einer minimalen Verbesserung des Prozentsatzes korrekt klassifizierter Datensätze von 0,001 in 1000 Epochen im zur Überprüfung des Lernfortschritts verwendeten Select-Datensatz auf 0,0001 in 1000 Epochen führte ebenfalls zu praktisch dem gleichen Ergebnis (cc=0,72) wie die Standardparameter. Die Fehleroberfläche war wohl nicht so flach, dass eine Verkleinerung der minimal nötigen Verbesserung von Vorteil ist.
Eine Änderung der Klassifikationsgrenze auf einen Wert von 0,5, anstelle der standardmäßigen Optimierung, führte ebenfalls zum gleichen Ergebnis (cc=0,72) wie

die Standardparameter. Die Optimierung der Grenzen schien demnach nicht zwingend notwendig zu sein.
Ein Aktivieren der Weight Decay-Funktion bestraft zu große Gewichte in den Neuronen. Allerdings schienen diese manchmal nötig zu sein – das Ergebnis verschlechterte sich beim Aktivieren der Funktion etwas (cc=0,61).
Ein epochenweises Anpassen (Verringern) der Lernrate verschlechterte ebenfalls den Mathews-Koeffizienten (auf cc=0,67). Dieses Ergebnis war konsistent mit der Verschlechterung bereits bei der anfänglichen Wahl einer kleinen Lernrate; die Gewichte passten sich zu langsam den zu lernenden Daten an.

4.3.3. Fehlerinterpretation

Das beste Netz, ein dreilagiges fully-connected feed-forward Neuronales Netz, sagte unter Benutzung der relativen Aminosäurehäufigkeiten als Inputvektoren vier Sequenzen falsch-positiv und eine Sequenz falsch negativ voraus. Hier sollen Gründe für die fehlerhafte Klassifizierung dieser Datensätze angesprochen werden. Die falsch klassifizierten Datensätze sind in Tabelle 21 wiedergegeben.

Tabelle 21 – Klassifizierungsergebnisse des Netzes mit der bestmöglichen Klassifikationsgüte, basierend auf Aminosäurehäufigkeiten

Klassifikation als	Anzahl Sequenzen	Namen
P	39	
N	131	
O	4	M1 Family Aminopeptidase Clathrin Coat Assembly Protein Vacuolar Proton Pumping Pyrophosphatase-2 Knob-associated hisitidine rich protein (KAHRP)
U	1	Fumarase Class 1

Die Lokalisierung der M1 Family Aminopeptidase ist, nach Diskussion mit Giel van Dooren, nicht eindeutig experimentell belegt[46] und kann daher nicht weiter diskutiert werden.

Das Clathrin Coat Assembly Protein ist zuständig für die Umhüllung der Vesikel mit dreilagigen Schichten, die hauptsächlich aus Clathrin bestehen und essentiell für die Zelle sind. Dieses Protein ist in anderen Organismen im Cytoplasma lokalisiert, was auch von der Funktion her Sinn macht. Daher ist seine Lokalisation aus bisher unbekannten Gründen falsch vorhergesagt.

Die Vacuolar Proton Pumping Pyrophosphatase-2 ist, laut Literaturstelle[47], nicht in den Mitochondrien lokalisiert. Allerdings besitzt sie eine große N-terminale Erweiterung im Vergleich zur Pyrophosphatase 1, diese Erweiterung scheint mitochondrialen Transitpeptiden zu ähneln.

Die letzte der vier falsch-positiven Sequenzen ist KAHRP, das zu einem Teil für die fatale Wirkung von *P. falciparum* verantwortlich ist. Dieses Protein sitzt auf der Oberfläche von infizierten Blutzellen und führt zu deren Verklumpung. Anders als andere sekretorische Proteine, die N-terminale Signalsequenzen besitzen, besitzt KAHRP ein internes Signalpeptid (persönliche Information von Giel van Dooren). Es befindet sich etwa 20 Aminosäuren vom N-Terminus entfernt. Dieses Signalpeptid ist für den Export aus der Zelle verantwortlich, direkt daran schließt sich eine Sequenz an, die für das Targeting auf die Oberfläche der roten Blutkörperchen verantwortlich ist. Die Region vor dem Signalpeptid ist sehr reich and basischen Aminosäuren und ähnelt nach Meinung des Neuronalen Netzes einem mitochondrialen Transitpeptid, was für die falsch-positive Vorhersage verantwortlich ist.

Die falsch-negativ vorhergesagte Fumarase der Klasse 1 tritt typischerweise in Bakterien auf. Die meisten bekannten Mitochondrien besitzen eine Fumarase der Klasse 2 – es scheint nicht ausgeschlossen, jedoch unwahrscheinlich, dass die hier positiv vorhergesagte Fumarase Klasse 1 tatsächlich mitochondrial lokalisiert ist (Gespräch mit Giel van Dooren).

4.4. Genome Scanning *P. falciparum*

Ausgehend von den 10276 Open Reading Frames ergaben sich als beste Schätzung – womit die Verwendung des Netzes mit dem besten Mathews-Koeffizienten von cc = 0,74 im Testsatz gemeint ist - 2449 mitochondriale Transitpeptide. Dies entspricht 23,8% des Kerngenoms. Als Mindestzahl, unter Benutzung restriktiver Netze, ergeben sich 706 derartige Peptide. Der erstgenannte Wert von fast 24% vermutlich mitochondrialer Transitpeptide erscheint im Vergleich zu anderen Organismen sehr hoch, der Unterschied beträgt etwa einen Faktor von zwei. Bäckerhefe (*Saccharomyces cerevisiae)* besitzt etwa 6000 Gene im Kerngenom, von denen etwa 500 vermutlich mitochondriale Transitpeptide darstellen[5] (etwas über 8%). *Arabidopsis thaliana*, eine Blütenpflanze aus der Familie der Senfpflanzen (wie Kohl und Rettich), besitzt etwa 11% mitochondrialer Transitpeptide im Kerngenom[48].

Grund für die hohe Zahl vorhergesagter mitochondrialer Sequenzen könnte eine für die Sequenzen des Gesamtgenoms nicht repräsentative Stichprobe an Trainingssequenzen sein. Die nicht-mitochondrialen („negativen“) Trainingssequenzen enthalten möglicherweise zum Großteil deutlich von mitochondrialen Transitpeptiden unterschiedliche Sequenzen, während die im Genom vorhandenen Sequenzen in einer schwerer zu unterscheidenden Grauzone liegen. Als Sequenz ergibt sich, dass eher den Vorhersagen des strikteren, auf relative Aminosäurehäufigkeiten trainierten Netzes Vertrauen geschenkt werden sollte.

Das verwendete neuronale Netz ist mit den verwendeten Open Reading Frames sowie vorhergesagten mitochondrialen und nicht-mitochondrialen Sequenzen im Internet unter http://www.modlab.de zu finden. Abbildung 49 zeigt die Bedienoberfläche des hier verwendeten neuronalen Netzes.

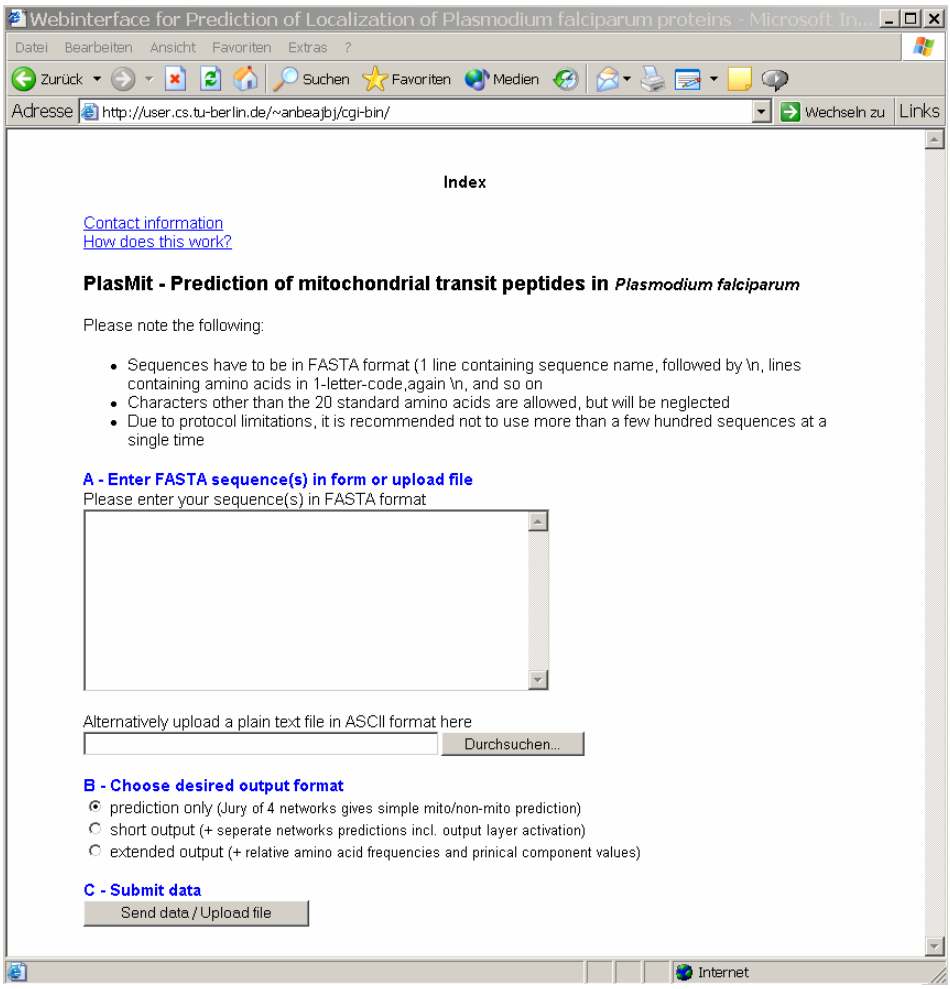

Abbildung 49 - Screenshot des Vorhersageprogramms PlasMit

5. Ausblick

Mit dieser Arbeit wurde ein funktionstüchtiges Vorhersageprogramm für mitochondriale Transitpeptide in *Plasmodium falciparum* entwickelt, das die Klassifikationsgüte bekannter Programme übertrifft. Die vorhergesagte Anzahl von positiven Datensätzen, angewendet auf das Kerngenom, erscheint sehr hoch, doppelt so hoch wie in anderen Fällen (Bäckerhefe und Arabidopsis). Daher ist den positiv vorhergesagten Datensätzen Aufmerksamkeit in Bezug auf eventuelle Gemeinsamkeiten zu widmen. Ebenso wäre festzustellen ob die positiv vorhergesagten Datensätze Gemeinsamkeiten mit den beim Training als falsch positiv in Erscheinung getretenen Datensätzen haben. Im Idealfall ist zu einem späteren Zeitpunkt die Übereinstimmung der vorhergesagten Lokalisierungen mit experimentellen Daten zu verifizieren. Möglich wäre die Erweiterung des Neuronalen Netzes auf die Vorhersage von vier Lokalisierungen – Mitochondrien, Cytoplasma, extrazellulär oder Apicoplast. Diese Versuche haben in vorläufigen Tests zufrieden stellende Ergebnisse gezeigt, häufig war aber eine Verwechslung mitochondrialer Transitpeptide mit solchen, die für den Apicoplasten bestimmt waren, zu beobachten.

6. Zusammenfassung

Der Malaria verursachende Organismus *Plasmodium falciparum (P. falciparum)* besitzt in seinem Kerngenom für die Mitochondrien bestimmte Proteine, die als Transportsignal ein mitochondriales Transitpeptid enthalten. Durch die kürzlich erfolgte Sequenzierung des Genoms von *P. falciparum* ist es wünschenswert, Vohersagealgorithmen für verschiedene Proteinlokalisationen zur Verfügung zu haben. Für andere Organismen etablierte Programme zur Vorhersage von mitochondrialen Transitpeptiden, MitoProtII und TargetP, lieferten bei Anwendung auf Sequenzen aus *P. falciparum* nur unbefriedigende Ergebnisse. MitoProtII erzielte in einer 20-fachen Kreuzvalidierung einen Mathews-Koeffizienten von cc = 0,49, TargetP erzielte in diesem Fall einen Mathews-Koeffizienten von cc = 0,60. TargetP erzielte für die Sequenzen aus *P. falciparum* nur eine Selektivität von 47%, MitoProtII nur eine Sensitivität von 35%. Dieser Ergebnisse haben die Entwicklung eines speziell auf *P. falciparum* trainierten Vorhersagemodells wünschenswert gemacht.

Kerncodierte mitochondriale Precursorproteine aus *P. falciparum* wurden mit statistischen Methoden, Hauptkomponentenanalyse, selbstorganisierenden Karten und überwachten neuronalen Netzen analysiert und mit solchen aus anderen Organismen verglichen. Zwei Repräsentationen der Datensätze wurden gewählt, Aminosäurehäufigkeiten und 19 physikochemische Eigenschaften. Ein grundsätzlich unterschiedlicher Aminosäuregebrauch konnte festgestellt werden. Glycin, Alanin, Prolin und Arginin werden in *P. falciparum* mit weniger als 60% der Häufigkeit in der Swiss-Prot-Datenbank, Version 36, verwendet. Isoleucin, Tyrosin, Asparagin und Lysin werden hingegen mit mehr als 150% der Häufigkeit in der Referenzdatenbank verwendet. Diese Häufigkeitsmuster wurden, mit Variationen, auch in allen Targetingsequenzen beobachtet.

In der Datenanalyse mittels Hauptkomponentenanalyse und selbstorganisierenden Karten ließen sich cytoplasmatische Proteine in beiden Repräsentationen klar von der Gruppe mitochondrialer, extrazellulärer und apicoplastischer Proteine trennen. Die Trennung innerhalb der zweiten Gruppe war weniger deutlich.

Ein neuronales Netz (PlasMit) zur Vorhersage mitochondrialer Transitpeptide in *P. falciparum* wurde entwickelt. Basierend auf der relativen Aminosäurehäufigkeitsverteilung innerhalb der ersten 24 N-terminalen Aminosäuren lieferte es einen Mathews- Korrelationskoeffizienten von 0,74 (86% korrekt vorhergesagte Sequenzen) in einer 20fachen Kreuzvalidierung. Dieses Netz sagte 2449

(24%) der 10276 vorhergesagten Open Reading Frames aus dem Genom von *P. falciparum* als mögliche mitochondrial lokalisierte Proteine voraus. Ein Netz mit identischer Topologie wurde auf eine geringere Anzahl falsch-positiver Vorhersagen trainiert und erzielte einen Mathews-Koeffizienten von 0,51 (84% korrekte Vorhersagen) in einer 10fachen Kreuzvalidierung. Dieses Netz sagte 903 (8,8%) potentielle mitochondriale Precursorproteine unter den 10276 vorhergesagten Open Reading Frames voraus.

Sämtliche Trainingsdatensätze, die Open Reading Frames des Genoms von *P. falciparum,* sowie das Netz, das den höchsten Mathews-Koeffizienten erzielt hat, sind per Web unter http://www.modlab.de, Menüpunkt PlasMit, erreichbar.

7. Anhang

7.1. *P. falciparum* – positiver Datensatz

Dieser Datensatz umfasst kerncodierte Proteine mit vermutlich mitochondrialem Transitpeptid. In Spalte 1 von Tabelle 22 ist der Name des Gens oder Proteins, in Spalte 2 die zugehörige Klassifikation und in Spalte 3 die Referenz oder eine zusätzliche Begründung für die Aufnahme in den Datensatz gegeben. Diese Liste wurde von Giel van Dooren, Plant Cell Biology Research Centre, School of Botany, University of Melbourne, Parkville, Victoria 3052, Australia erstellt und ohne Änderungen (mit Ausnahme der Spaltenüberschriften) übernommen. Zugriffsnummern konnten nur für bereits indexierte Einträge angegeben werden.

Tabelle 22 – Alphabetische sortierte Sequenzen des positiven Datensatzes

Name des Gens/Proteins	Klassifikation	Referenz / Begründung für Aufnahme in den Datensatz
50s rpl17	Translation	
50S RPL2	Translation	
50s rpl24	Translation	
ATP synthase alpha-SU	complex V	acc no: C71606
ATP synthase beta-SU	complex V	
ATP synthase delta-SU	complex V	
ATP synthase gamma-SU	complex V	
branched-chain alpha keto-acid DH E1 - alpha SU	amino acid degradation	
branched-chain alpha keto-acid DH E1 - beta SU	amino acid degradation	
citrate synthase	TCA cycle	
coQ synthesis methyltransferase (coq5)	CoQ biosynthesis	found in mdria of other orgs
coq2	CoQ biosynthesis	
coq4	CoQ biosynthesis	found in mdria of other orgs
coq8 = ubiquinol-cytochrome-c reductase assembly protein (ABC1)	CoQ biosynthesis	aka abc1; ref for yeast: Do etal JBC 276:18161(2001)
cytochrome c1	complex III	

delta aminolevulinic acid synthase	haem synthesis	AAC37294; MBP 75 (2), 271-276 (1996)
Dihydroorotate dehydrogenase	pyrimidine biosynthesis	localised to mdria - Biochim Biophys Acta 1995 Apr 13;1243(3):351-60
Dihydroxy hexaprenylbenzoate methyltransferase (DHHB-MTase)	CoQ biosynthesis	coq3 hom
DNA-directed RNA polymerase	Transcription	AAG00950
EF-Tu	Translation	mitochondrial homologue
elongation factor g (EF-G)	Translation	
fumerate hydratase class I	TCA cycle	class I found in Ecoli and Archaea, class II in yeast and Bact only
geranyl diphosphate synthase/prenyl transferase (coq1)	CoQ biosynthesis	GPP syn found in plastids, prenyl tase found in mdria
grpE	protein folding	binds to hsp70; matches at c-term; mdrial in other orgs (although a plastid form exists in Arabidopsis)
hsp60/cpn60	import	AAC47497; MBP 83 (2), 263-264 (1996)
HSP-70/DnaK	protein folding	branches with mdrial hsp70s on phylogenetic tree
isocitrate dehydrogenase (NADP-dependent)	TCA cycle???	best match to mitochondrial NADP-dep enzyme; no NAD+ hom in dbase
Lon protease homologue	misc: chaperone/protease	
mdrial serine hydroxymethyl transferase (SHMT)	AAdegrad/AA syn/1C metabolism	involved in Gly syn and breakdown and 1-carbon metabolism; and poss purine synth; note other SHMT homologues in dbase
mitochondrial intermediate processing peptidase	import	
mitochondrial phosphate carrier	transport	AAC47174.1; MBP 78:297 1996

MPP alpha SU	import	
MPP beta SU	import/complexIII	
prohibitin/BAP37 (PHB2 homologue)	respiratory chain assembly	localises to inner mdrial membrane in other orgs. could have "inner membrane" targeting sequence
rhodanese	AA degrad/CN detox	mdrial in other orgs, although in plants may be plastid; involved in cysteine degrad
rieske Fe-S protein 3	complex III	
rotenone-insensitive NADH DH	bind ADP/electron transport	possible inner membrane protein/ only found in plants/fungi
succinate dehydrogenase Fe-S subunit	TCA cycle/electron transport	
succinyl-CoA sythetase beta subunit	TCA cycle	strongest match to ATP-specific form, a comm feature of protozoan TCA cyc enzymes
valyl-trna synthetase	Translation	

7.2. *P. falciparum* – negativer Datensatz

Dieser Datensatz umfasst kerncodierte Proteine ohne mitochondriales Transitpeptid. In Spalte 1 von Tabelle 23 ist der Name des Gens oder Proteins, in Spalte 2 die Lokalisierung, in Spalte 3 die Zugriffsnummer und in Spalte 4 die Referenz gegeben. Diese Liste wurde von Giel van Dooren erstellt und ohne Änderungen (mit Ausnahme der Spaltenüberschriften) übernommen. Zugriffsnummern konnten nur für bereits indexierte Einträge angegeben werden.

Tabelle 23 – Alphabetisch sortierte Sequenzen des negativen Datensatzes

Name des Gens/Proteins	Lokalisie-rung	GeneBank Acc. No.	Referenz für Lokaliserung
ACIDIC RIBOSOMAL PHOSPHOPROTEIN PO	Cytoplasma	AAD10140	located in ribosomes of other organisms; Mol Biochem Parasitol. 1996 Nov 12;82(1):117-20
ACTIN1	Cytoplasma	A54496	Parasitology 1996 May;112 (Pt 5):451-7
ACYL-COA SYNTHETASE	Cytoplasma	AAD53966	J. Mol. Biol. 291 (1), 59-70 (1999)
ADENYLOSUCCINATE LYASE	Cytoplasma	AAF06822; note: accession no. is not complete at N-terminus - rest of sequence obtained from NCBI database	cytosolic in other organisms where characterised - eg in Dictyostelium: J Biol Chem 1991 Feb 5;266(4):2480-5
ARF	Cytoplasma	CAB02498	involved in ER-Golgi vesicle traffic as well as other membrane traffic
CELL DIVISION CONTROL PROTEIN 2	Cytoplasma	Q07785	Eur J Biochem 1994 Mar 15;220(3):693-701
CHORISMATE SYNTHASE	Cytoplasma	AAB63293	IFAs: Mol Microbiol 2001 Apr;40(1):65-75; and branches with cytosolic enzymes of other organisms, Nature 1999 Jan 21;397(6716):219-20

CLATHRIN COAT ASSEMBLY PROTEIN	Cytoplasma	AAC71950	in other organisms may associate with internal side of plasma membrane
CTP SYNTHASE	Cytoplasma	AAC36385	cytosolic in other organisms; J Biol Chem 1976 Jul 25;251(14):4372-8
CYTOSOLIC TRIOSEPHOSPHATE ISOMERASE	Cytoplasma	Q07412	most similar to cytosolic enzymes from other organisms
DYNAMIN1	Cytoplasma	AAK26820	cytosol in other organisms
DYNAMIN (DNM1)	Cytoplasma	N/A	cytosolic in other organisms
EF-1-ALPHA	Cytoplasma	S21909	branches with cytosolic protein on phylogenetic tree
ENOLASE	Cytoplasma	AAA18634	cytosolic in other organisms; branches with eukaryotic enolases on tree: Proc Natl Acad Sci U S A 2001 98(19):10745-50
FRUCTOSE BISPHOSPHATE ALDOLASE	Cytoplasma	A44942	Mol Biochem Parasitol 1990 Apr;40(1):1-12
GCN20-ABC TRANSPORTER-LIKE	Cytoplasma	AAB03649	Mol Biochem Parasitol 1998 Nov 30;97(1-2):81-95
GLUCOSE-6-PHOSPHATE DEHYDROGENASE	Cytoplasma	CAA52921	most similar to cytosolic enzymes from other organisms
GLUCOSE-6-PHOSPHATE ISOMERASE	Cytoplasma	P18240	most similar to cytosolic enzymes from other organisms
GLUTAMATE DEHYDROGENASE	Cytoplasma	CAA73390	indirect biochemical localization; Eur. J. Biochem. 258 (2), 813-819 (1998)
GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE	Cytoplasma	AAD10249	branches with cytoplasmic protein of chromalveolates
HEXOKINASE	Cytoplasma	A48457	Mol Biochem Parasitol. 1994 Jan;63(1):171-4
HISTONE DEACETYLASE	Cytoplasma	AAD22407	Mol. Biochem. Parasitol. 99 (1), 11-19 (1999)
HSP70	Cytoplasma	P11144	eg J Biol Chem 2002 Feb 8;277(6):3902-12
HYPOXANTHINE GUANINE XANTHINE PHOSPHORIBOSYLTRANSFERASE	Cytoplasma		

HYPOXANTHINE PHOSPHORIBOSYLTRANSFERASE	Cytoplasma	P07833	Exp Parasitol 1992 Feb;74(1):11-9 - but not via SP-targeted secretory pathway
IMP DEHYDROGENASE	Cytoplasma	AAD10256	cytosol in most other organisms where characterised - P.fal enzyme lacks typical plastid leader - therefore not plastid
INITIATION FACTOR EIF-4E	Cytoplasma		
LACTATE DEHYDROGENASE	Cytoplasma	Q27743	bacterial-like protein, but with no N-terminal extension
M1 FAMILY AMINOPEPTIDASE	Cytoplasma	O96935	Mol Biochem Parasitol 1998 Nov 30;97(1-2):149-60
MULTI-DRUG RESISTANCE PROTEIN (PFMDR)	Cytoplasma	P13568	Exp Parasitol 1995 Aug;81(1):1-8
MYOSIN A	Cytoplasma	AAD21242	di-basic motif in tail specifies PM targeting in toxo; Mol Biol Cell 2000 Apr;11(4):1385-400
MYOSIN D	Cytoplasma	AAK56302	cytosolic in other organisms
NSF	Cytoplasma	CAB10575	J Biol Chem 2001 May 4;276(18):15249-55
NT1 NUCLEOSIDE TRANSPORTER	Cytoplasma	AAF67611	J Biol Chem 2001 Nov 2;276(44):41095-9
PHOSPHOGLYCERATE KINASE	Cytoplasma	JU0475	cytosolic in other organisms
POLYUBIQUITIN	Cytoplasma	CAB59728	cytosolic in other organisms
PROTEIN KINASE	Cytoplasma	CAA58680	biochemical localisation; Mol Biochem Parasitol 1995 Jun;72(1-2):163-78
RAB1	Cytoplasma	AAB16753	for review of localisation in other orgs see eg Curr Opin Cell Biol 2001 Aug;13(4):500-11
RAB18	Cytoplasma	CAD27350	for review of localisation in other orgs see eg Curr Opin Cell Biol 2001 Aug;13(4):500-11
RAB2	Cytoplasma	CAC34627	for review of localisation in other orgs see eg Curr Opin Cell Biol 2001 Aug;13(4):500-11
RAB5B	Cytoplasma	CAD19466	for review of localisation in other orgs see eg Curr Opin Cell Biol 2001 Aug;13(4):500-11

RAB5C	Cytoplasma	CAD12439	for review of localisation in other orgs see eg Curr Opin Cell Biol 2001 Aug;13(4):500-11
RAB6	Cytoplasma	CAA63555	Mol Biochem Parasitol 1996 Sep;80(1):77-88; Mol. Biochem. Parasitol. 83 (1), 107-120 (1996)
RIBONUCLEOSIDE DIPHOSPHATE REDUCTASE LARGESU	Cytoplasma	P50647	cytosolic in other organisms eg EMBO J. 1984;3:863-867
RIBONUCLEOSIDE DIPHOSPHATE REDUCTASE SMALLSU	Cytoplasma	B49412	cytosolic in other organisms eg EMBO J. 1984;3:863-867
SAR1	Cytoplasma	AAF06723	Eur J Cell Biol. 1999 Jul;78(7):453-62
SEC31	Cytoplasma	NP_473056	MMDLIKIYKRNEIIKDYVNIYFDDEKSENSNKNLSYKILHIIGNGVYGVVYKADCLNNCSIVALKQTYQKSTRIFKEIEIMKKLKHPNIVKLKHAFYTSTNDGGVYVHMVMEYGNTDLASSLYYITLKDSKEFLKDSFDLDNYNYEDYDDNKSEKDMLPNDQQKSDFYYYRSLPNQNEEMTADEKSGIHPGNNNMIVQGMINSNNNNNRDNNGDNNSSNNNSSNNNSSNNNSSNNNSSNNNSSSSSSSSVLNDVH
SEC61 ALPHA	Cytoplasma	AAC38988	Mol. Biochem. Parasitol. 92 (1), 89-98 (1998)
SEC61 GAMMA	Cytoplasma	AAC71879	associated with sec61 alpha
THIOREDOXIN REDUCTASE	Cytoplasma	Q25861	most similar to cytosolic enzymes from other organisms
TRANLATION RELEASE FACTOR ERF-1	Cytoplasma	AAC71899	cytosolic in other organisms
TRANSLATION ELONGATION FACTOR EEF1ALPHA	Cytoplasma		
TRANSLATION ELONGATION FACTOR EEF1BETA	Cytoplasma	AAF27524	cytosolic in other organisms
TUBULIN ALPHA	Cytoplasma	S07459	cytosolic in other organisms
TUBULIN BETA	Cytoplasma	A44949	cytosolic in other organisms
TUBULIN GAMMA	Cytoplasma	CAA44265	cytosolic in other organisms

UBIQUITIN CONJUGATING ENZYME E2	Cytoplasma	NP_473305	cytosolic in other organisms
VACUOLAR-TYPE H PUMPING PYROPHOSPHATASE-1	Cytoplasma	AAD17215	Mol. Biochem. Parasitol. 114 (2), 183-195 (2001)
VACUOLAR-TYPE H PUMPING PYROPHOSPHATASE-2	Cytoplasma	AAG21366	Mol. Biochem. Parasitol. 114 (2), 183-195 (2001)
VACUOLAR ATPASE SU A	Cytoplasma	Q03498	J Biol Chem 2000 Nov 3;275(44):34353-8
VACUOLAR ATPASE SU B	Cytoplasma	Q25691	J Biol Chem 2000 Nov 3;275(44):34353-8
ABRA	Extrazellulär	AAA29462	refs in Philos Trans R Soc Lond B Biol Sci. 2002 Jan 29;357(1417):25-33
AMA-1	Extrazellulär	P22621	refs in Philos Trans R Soc Lond B Biol Sci. 2002 Jan 29;357(1417):25-33
CIRCUMSPOROZOITE RELATED ANTIGEN	Extrazellulär	AAA21753	EMBO J 1987 Feb;6(2):485-91
CLAG	Extrazellulär	AAF97948	thought to be involved in cytoadherence on cell surface - no direct evidence of localisation; review in Mol Biochem Parasitol. 2001 Jul;115(2):129-43
EBA-175	Extrazellulär	AAF72186	refs in Philos Trans R Soc Lond B Biol Sci. 2002 Jan 29;357(1417):25-33
EMP3	Extrazellulär	AAF01135	Mol Biochem Parasitol 1993 May;59(1):59-72
ERD2	Extrazellulär	S39609	EMBO J 1993 Dec;12(12):4763-73
HSP70 BIP	Extrazellulär		
KAHRP	Extrazellulär	AAA29629	Cell. 1997 Apr 18;89(2):287-96; Proc Natl Acad Sci U S A 1987 Oct;84(20):7139-43; EMBO J 1987 May;6(5):1421-7
MAEBL	Extrazellulär	AAL10509	localised in P. yoelii; Proc Natl Acad Sci U S A 1998 Feb 3;95(3):1230-5
MSP-1	Extrazellulär	P13819	refs in Philos Trans R Soc Lond B

			Biol Sci. 2002 Jan 29;357(1417):25-33
MSP-3	Extrazellulär	AAC09377	refs in Philos Trans R Soc Lond B Biol Sci. 2002 Jan 29;357(1417):25-33
RAP1	Extrazellulär	S27833	refs in Philos Trans R Soc Lond B Biol Sci. 2002 Jan 29;357(1417):25-33
RAP2	Extrazellulär	AAF23400	import associated with RAP-1; refs in Philos Trans R Soc Lond B Biol Sci. 2002 Jan 29;357(1417):25-33
RBP2	Extrazellulär	AAK19245	Infect Immun 2001 Feb;69(2):1084-92
RESA	Extrazellulär	CAA00077	J Exp Med 1985 Aug 1;162(2):774-9
RHOPH3	Extrazellulär	A45554	Parasitology 1994 Apr;108 (Pt 3):269-80
RIFIN	Extrazellulär	NP_473134	review in Mol Biochem Parasitol. 2001 Jul;115(2):129-43
S-ANTIGEN	Extrazellulär	Q03400	refs in Philos Trans R Soc Lond B Biol Sci. 2002 Jan 29;357(1417):25-33
SUB-1	Extrazellulär	CAA05261	J Biol Chem 1998 Sep 4;273(36):23398-409
SUB-2	Extrazellulär	CAB43592	Proc Natl Acad Sci U S A 1999 May 25;96(11):6445-50
ACCASE	Apicoplast	N/A	localised in Toxo; Proc Natl Acad Sci U S A 2001 Feb; 27;98(5):2723-8
ACP	Apicoplast	AAC63959	EMBO J. 2000 Apr 17;19(8):1794-802.
ALAD	Apicoplast	AY064477	found in plastids of other organisms and has characteristic apicoplast leader
ALANYL-TRNA SYNTHETASE	Apicoplast		
ASPARTYL-TRNA SYNTHETASE	Apicoplast		
CHLOROPLAST (LEUCINE?) AMINOPEPTIDASE	Apicoplast		
CLPBV1	Apicoplast		

CPN60 APLAST	Apicoplast	P34940	branches with plastid enzymes on phylogenetic tree and has characteristic apicoplast leader
DIMETHYLADENOSINE TRANSFERASE	Apicoplast		
DNA TOPOISOMERASE (ATP-HYDROLYSING)	Apicoplast		
DOXP REDUCTOISOMERASE	Apicoplast	AAD03739	Plas leader targeted in Toxo; Science. 1999 Sep 3;285(5433):1573-6
DOXP SYNTHASE	Apicoplast	AAD03740	found in plastids of other organisms and has characteristic apicoplast leader
EF-TS APLAST	Apicoplast	NP_473178	branches with plastid EF-TSs and has characteristic apicoplast leader
EF-TS SANGER MP03010	Apicoplast		
FABD	Apicoplast	AAK83684	found in plastids of other organisms and has characteristic apicoplast leader
FABG	Apicoplast	AAK83686	found in plastids of other organisms and has characteristic apicoplast leader
FABH	Apicoplast	AAC63960	EMBO J. 2000 Apr 17;19(8):1794-802.
FABI	Apicoplast	AAK38273	found in plastids of other organisms and has characteristic apicoplast leader
FABI	Apicoplast	AAK38273	found in plastids of other organisms and has characteristic apicoplast leader
FABZ	Apicoplast	AAK83685	found in plastids of other organisms and has characteristic apicoplast leader
FERREDOXIN	Apicoplast		
GLUTAMYL-TRNA SYNTHETASE	Apicoplast		
GYRA	Apicoplast	N/A	found in plastids of other organisms and has characteristic apicoplast leader

GYRB	Apicoplast	N/A	found in plastids of other organisms and has characteristic apicoplast leader
HEMB	Apicoplast		
ISOPENTYL MONOP KINASE	Apicoplast	N/A	found in plastids of other organisms and has characteristic apicoplast leader
LIPOIC ACID SYNTHETASE	Apicoplast	N/A	found in plastids of other organisms and has characteristic apicoplast leader
LYSYL TRNA SYNTHASE	Apicoplast		
PBGD	Apicoplast	N/A	found in plastids of other organisms and has characteristic apicoplast leader
PETF	Apicoplast	N/A	found in plastids of other organisms and has characteristic apicoplast leader
PETH	Apicoplast	N/A	localised to toxo aplast but unpublished
PHENYLALANYL-TRNA SYNTHETASE ALPHA CHAIN	Apicoplast		
PHOSPHATE/PHOSPHOENOLPYRUVATE TRANSLOCATOR	Apicoplast		
PORPHOBILINOGEN DEAMINASE (HEMC)	Apicoplast		
PYRUVATE DEHYDROGENASE E1	Apicoplast	N/A	branches with plastid enzymes on phylogenetic tree and has characteristic apicoplast leader
PYRUVATE DEHYDROGENASE E1 BETA	Apicoplast	N/A	branches with plastid enzymes on phylogenetic tree and has characteristic apicoplast leader
PYRUVATE DH E2	Apicoplast	N/A	branches with plastid enzymes on phylogenetic tree and has characteristic apicoplast leader
PYRUVATE DH E3	Apicoplast	N/A	branches with plastid enzymes on phylogenetic tree and has characteristic apicoplast leader
PYRUVATE KINASE	Apicoplast	N/A	found in plastids of other organisms and has characteristic

			apicoplast leader
RIBOSOME RELEASE FACTOR FRR	Apicoplast		
RRNA METHYLASE	Apicoplast		
RTCB PROTEIN	Apicoplast		
SERYL-TRNA SYNTHETASE 1	Apicoplast		
SLR1419	Apicoplast		
TIC22	Apicoplast	N/A	localised but unpublished
TOC 34 (SLR1974)	Apicoplast		
TRNA-GUANINE TRANSGLYCOSYLASE (	Apicoplast		
TRNA ISOPENTENYLTRANSFERASE	Apicoplast	N/A	found in plastids of other organisms and has characteristic apicoplast leader
TRNA PSEUDOURIDINE SYNTHASE A	Apicoplast		
TRYPTOPHANYL-TRNA SYNTHETASE	Apicoplast		
VALYL-TRNA SYNTHETASE	Apicoplast		
YGBB	Apicoplast	N/A	found in plastids of other organisms and has characteristic apicoplast leader

7.3. Quellcode des fully connected feed-forward ANN

Am Beispiel des Netzes mit 13 Aminosäurehäufigkeiten als Inputvektoren, das auf den besten Mathews-Koeffizienten trainiert wurde.

```
/* ------------------------------------------------------------------------ */
/*
  Title: aas3N436ItMaxCC, Neural Network execution code.
  Automatically generated by SNN, Wed Jun 26 19:40:36 2002


  License Agreement:
  ------------------

  Copyright StatSoft Inc., 2000-2001, all rights reserved.
  This source code (Source Code Generated by STATISTICA Neural Networks,
  referred to as "CG" below) is owned by StatSoft Inc. and is protected
  by United States Copyright laws and international treaty provisions. You
  shall treat the CG like any copyrighted material.
  The CG may not be redistributed or used except in accordance with the
  conditions below.

  The licensee is granted a license to incorporate the CG as embedded
  software in their own hardware and software products, and to distribute an
  unlimited number of such embedded copies as part of this license subject to
  obtaining prior written consent from StatSoft Inc., and subject to
  the conditions listed below.

  Prior consent is required so that StatSoft Inc. can ensure that license
  conditions are not breached, and can track legitimate use of the CG. Consent
  shall not be refused unless StatSoft Inc. reasonably believes that a
  breach of license conditions will occur. Consent shall usually be granted
  within five working days of the request, providing that sufficient details of
  the intended use are given.

  Requests should be sent to SNN Project Director, StatSoft, Inc.,
  2300 East 14th Street, Tulsa OK 74104 USA, FAX: 918-749-2217,
  E-Mail: info@statsoft.com.

  The licensee may modify the CG as they see fit for embedded use, including
  recoding into alternative programming languages, altering the neural network
  architecture and weights, and otherwise modifying the CG, provided that they
  keep intact this copyright and license notice.

  The licensee may distribute products including the compiled version of CG.

  The licensee shall not:-
    Sublicense, rent, lease, or assign any portion of the CG to third parties.
    Allow compiled versions of the CG to be incorporated in products owned
    by third parties.
    Allow access to the CG to third parties.
    Use (implicitly or explicitly) any reference to StatSoft, Inc., STATISTICA,
    STATISTICA Neural Networks, or any trade names used by StatSoft, Inc.
    to describe, promote, or reference products in which CG is used, or which
    benefit from CG.

  Except as expressly stated herein, the CG is provided "AS IS." The licensee
  shall be entirely responsible for the selection of the CG and for the
```

/* standard includes. math.h needed for exp() function. */

```
#include <stdio.h>
#include <math.h>
#include <string.h>
#include <stdlib.h>

#ifndef FALSE
#define FALSE 0
#define TRUE 1
#endif

#define MENUCODE -999


static double aas3N436ItMaxCC21Thresholds[] =
{

/* layer 1 */
-0.49624930020517216, -0.72668947747500945, -1.317154709766758,

/* layer 2 */
-0.96202400482749961
```

```
};

static double aas3N436ItMaxCC21Weights[] =
{

/* layer 1 */
-1.6247470021818842, -0.87142723934368171, -2.7682147069725813, -
0.48081361543274681, -1.1269669476097366, 0.8765481142226631, 1.2727397065407147,
2.5553977374166728, 0.29189237463811124, -1.604352824472113, -0.25120900612038199,
-0.54871725466526988, 0.94486293689895895, 0.53781531498407975,
1.1969648829278581, -0.69004189969430429, -1.1522909732330191,
-0.36633306596620396, 2.5309010291160239, -3.4739921625554371, 3.9609323349429237,
-0.0010766222742467984, -3.1555624588035425, -0.32931344734860563, -
3.4769507299657545, 1.351815606931783, 0.94657337783621709, 1.1298743182826503,
1.3851631147309114, 1.6398768827302643, -0.14346164527293551, 0.57045584877762212,
0.25560302400233204, -0.44106212051209997, -0.2464572761185034,
0.78814476878137874, 0.14592008724276895, -0.18802097868874265,
0.11663796997845952,

/* layer 2 */
-4.0845249537761603, 5.4165153367367838, 4.3748712993678707


};

static double aas3N436ItMaxCC21Acts[34];

/* ------------------------------------------------------------ */
/*
  aas3N436ItMaxCC21Run - run neural network aas3N436ItMaxCC21

  inputs - the input variables of this network.
  The variable names are listed below, together with each
  variable's offset in the data set at the time code was
  generated (if the variable is then available).
  Variable (Offset)
  Var7
  Var9
  Var10
  NewVar1
  NewVar2
  NewVar4
  NewVar5
  NewVar6
  NewVar7
  NewVar8
  NewVar9
  NewVar11
  NewVar14

*/
/* ------------------------------------------------------------ */

void aas3N436ItMaxCC21Run( double inputs[], double outputs[], int outputType )
{
  int i, j, k, u;
  double *w = aas3N436ItMaxCC21Weights, *t = aas3N436ItMaxCC21Thresholds;

  /* Process inputs - apply pre-processing to each input in turn,
   * storing results in the neuron activations array.
   */
```

```
/* Input 0: standard numeric pre-processing: linear shift and scale. */
if ( inputs[0] == -9999 )
  aas3N436ItMaxCC21Acts[0] = 0.13142857142857176;
else
  aas3N436ItMaxCC21Acts[0] = inputs[0] * 4.8000000000000069 + 0;

/* Input 1: standard numeric pre-processing: linear shift and scale. */
if ( inputs[1] == -9999 )
  aas3N436ItMaxCC21Acts[1] = 0.1228571428571426;
else
  aas3N436ItMaxCC21Acts[1] = inputs[1] * 5.9999999999999876 + 0;

/* Input 2: standard numeric pre-processing: linear shift and scale. */
if ( inputs[2] == -9999 )
  aas3N436ItMaxCC21Acts[2] = 0.13333333333333341;
else
  aas3N436ItMaxCC21Acts[2] = inputs[2] * 4 + 0;

/* Input 3: standard numeric pre-processing: linear shift and scale. */
if ( inputs[3] == -9999 )
  aas3N436ItMaxCC21Acts[3] = 0.28816326530612218;
else
  aas3N436ItMaxCC21Acts[3] = inputs[3] * 3.4285714285714239 + 0;

/* Input 4: standard numeric pre-processing: linear shift and scale. */
if ( inputs[4] == -9999 )
  aas3N436ItMaxCC21Acts[4] = 0.17600000000000043;
else
  aas3N436ItMaxCC21Acts[4] = inputs[4] * 4.8000000000000069 + 0;

/* Input 5: standard numeric pre-processing: linear shift and scale. */
if ( inputs[5] == -9999 )
  aas3N436ItMaxCC21Acts[5] = 0.32714285714285773;
else
  aas3N436ItMaxCC21Acts[5] = inputs[5] * 3.0000000000000027 + 0;

/* Input 6: standard numeric pre-processing: linear shift and scale. */
if ( inputs[6] == -9999 )
  aas3N436ItMaxCC21Acts[6] = 0.40285714285714319;
else
  aas3N436ItMaxCC21Acts[6] = inputs[6] * 4 + 0;

/* Input 7: standard numeric pre-processing: linear shift and scale. */
if ( inputs[7] == -9999 )
  aas3N436ItMaxCC21Acts[7] = 0.400816326530612;
else
  aas3N436ItMaxCC21Acts[7] = inputs[7] * 3.4285714285714239 + 0;

/* Input 8: standard numeric pre-processing: linear shift and scale. */
if ( inputs[8] == -9999 )
  aas3N436ItMaxCC21Acts[8] = 0.12571428571428556;
else
  aas3N436ItMaxCC21Acts[8] = inputs[8] * 12.000000000000007 + -0.50000000000000089;

/* Input 9: standard numeric pre-processing: linear shift and scale. */
if ( inputs[9] == -9999 )
  aas3N436ItMaxCC21Acts[9] = 0.30190476190476201;
else
  aas3N436ItMaxCC21Acts[9] = inputs[9] * 4 + 0;

/* Input 10: standard numeric pre-processing: linear shift and scale. */
if ( inputs[10] == -9999 )
```

```
  aas3N436ItMaxCC21Acts[10] = 0.1014285714285712;
 else
  aas3N436ItMaxCC21Acts[10] = inputs[10] * 5.9999999999999876 + 0;

 /* Input 11: standard numeric pre-processing: linear shift and scale. */
 if ( inputs[11] == -9999 )
  aas3N436ItMaxCC21Acts[11] = 0.1942857142857147;
 else
  aas3N436ItMaxCC21Acts[11] = inputs[11] * 4.8000000000000069 + 0;

 /* Input 12: standard numeric pre-processing: linear shift and scale. */
 if ( inputs[12] == -9999 )
  aas3N436ItMaxCC21Acts[12] = 0.24190476190476204;
 else
  aas3N436ItMaxCC21Acts[12] = inputs[12] * 4 + 0;

 /*
  * Process layer 1.
  */

 /* For each unit in turn */
 for ( u=0; u < 3; ++u )
 {
  /*
   * First, calculate post-synaptic potentials, storing
   * these in the aas3N436ItMaxCC21Acts array.
   */

  /* Initialise hidden unit activation to zero */
  aas3N436ItMaxCC21Acts[13+u] = 0.0;

  /* Accumulate weighted sum from inputs */
  for ( i=0; i < 13; ++i )
   aas3N436ItMaxCC21Acts[13+u] += *w++ * aas3N436ItMaxCC21Acts[0+i];

  /* Subtract threshold */
  aas3N436ItMaxCC21Acts[13+u] -= *t++;

  /* Now apply the hyperbolic activation function, ( e^x - e^-x ) / ( e^x + e^-x ).
   * Deal with overflow and underflow
   */
  if ( aas3N436ItMaxCC21Acts[13+u] > 100.0 )
    aas3N436ItMaxCC21Acts[13+u] = 1.0;
  else if ( aas3N436ItMaxCC21Acts[13+u] < -100.0 )
   aas3N436ItMaxCC21Acts[13+u] = -1.0;
  else
  {
   double e1 = exp( aas3N436ItMaxCC21Acts[13+u] ), e2 = exp( -
aas3N436ItMaxCC21Acts[13+u] );
   aas3N436ItMaxCC21Acts[13+u] = ( e1 - e2 ) / ( e1 + e2 );
  }
 }

 /*
  * Process layer 2.
  */

 /* For each unit in turn */
 for ( u=0; u < 1; ++u )
 {
  /*
   * First, calculate post-synaptic potentials, storing
```

```
   * these in the aas3N436ItMaxCC21Acts array.
   */

  /* Initialise hidden unit activation to zero */
  aas3N436ItMaxCC21Acts[16+u] = 0.0;

  /* Accumulate weighted sum from inputs */
  for ( i=0; i < 3; ++i )
    aas3N436ItMaxCC21Acts[16+u] += *w++ * aas3N436ItMaxCC21Acts[13+i];

  /* Subtract threshold */
  aas3N436ItMaxCC21Acts[16+u] -= *t++;

  /* Now apply the logistic activation function, 1 / ( 1 + e^-x ).
   * Deal with overflow and underflow
   */
  if ( aas3N436ItMaxCC21Acts[16+u] > 100.0 )
     aas3N436ItMaxCC21Acts[16+u] = 1.0;
  else if ( aas3N436ItMaxCC21Acts[16+u] < -100.0 )
    aas3N436ItMaxCC21Acts[16+u] = 0.0;
  else
    aas3N436ItMaxCC21Acts[16+u] = 1.0 / ( 1.0 + exp( - aas3N436ItMaxCC21Acts[16+u] ) );
}

/* Type of output required - selected by outputType parameter */
switch ( outputType )
{
  /* The usual type is to generate the output variables */
  case 0:


    /* Post-process output 0, two-state nominal output */
    if ( aas3N436ItMaxCC21Acts[16] >= 0.49480152577427561 )
      outputs[0] = 2.0;
    else
      outputs[0] = 1.0;
    break;

  /* type 1 is activation of output neurons */
  case 1:
    for ( i=0; i < 1; ++i )
      outputs[i] = aas3N436ItMaxCC21Acts[16+i];
    break;

  /* type 2 is codebook vector of winning node (lowest actn) 1st hidden layer */
  case 2:
    {
      int winner=0;
      for ( i=1; i < 3; ++i )
        if ( aas3N436ItMaxCC21Acts[13+i] < aas3N436ItMaxCC21Acts[13+winner] )
          winner=i;

      for ( i=0; i < 13; ++i )
        outputs[i] = aas3N436ItMaxCC21Weights[13*winner+i];
    }
    break;

  /* type 3 indicates winning node (lowest actn) in 1st hidden layer */
  case 3:
    {
      int winner=0;
      for ( i=1; i < 3; ++i )
```

```
        if ( aas3N436ItMaxCC21Acts[13+i] < aas3N436ItMaxCC21Acts[13+winner] )
          winner=i;

       outputs[0] = winner;
     }
     break;
  }
}

/*
  Test harness. Compile including this main() procedure, as
  a windows console program or a DOS program, to interactively
  test that the software functions as expected.
*/


int main(int argc, char* argv[])
{
  int i;
  double inputs[13], outputs[13];
  double invalid[7];
  FILE *fp;
  FILE *OUTFILE;
  char *pnewline;
  int winner;
  double activation;
  char filename[100];
  pnewline = (char *) malloc(1000);

  sprintf(&filename,"%s.aa-7",argv[1]);

  if ( ( fp = fopen(filename,"r") ) ==0) {
       printf("Couldn't open file %s ! Aborting!\n\n",argv[1]);
       exit(1);
  }
  else {
       sprintf(&filename,"%s.out",argv[1]);

       OUTFILE = fopen(filename,"w");

       /*Reading line by line from file */

       while (fgets(pnewline,1000,fp)!=NULL) {
            sscanf(pnewline,"%lf %lf %lf %lf %lf %lf %lf %lf %lf %lf %lf %lf
%lf",&inputs[0],&inputs[1],&inputs[2],&inputs[3],&inputs[4],&inputs[5],
&inputs[6],&inputs[7],&inputs[8],&inputs[9],&inputs[10],&inputs[11],&inputs[12]);
            /* Get the input pattern */
            /* Run the neural network - Activation*/
            aas3N436ItMaxCC21Run( inputs, outputs, 1 );
            activation = outputs[0] ;
            /* Run network - winning neuron */
            aas3N436ItMaxCC21Run(inputs, outputs, 0 );
            winner = outputs[0];
            fprintf(OUTFILE,"%i [%.2lf]\n",winner,activation);

       }
   fclose(fp);
   fclose(OUTFILE);
   return(0);
   }
free(pnewline);
}
```

8. Lebenslauf

Name: Andreas Bender

Geboren: am 20. August 1976, in Berlin

Schulische Ausbildung:

08 / 1996	Abitur an der Humboldt-Oberschule Berlin-Tegel

Zivildienst:

10 / 1996 - 09 / 1997	Zivildienst im Virchow-Klinikum Berlin-Wedding

Studium:

WS 1997/98 – SS 1999	Technische Universität Berlin Vordiplom August 1999
WS 1999/00 - SS 2000	Auslandsaufenthalt am Trinity College Dublin, Irland
WS 2000/01- WS 2001/02	Technische Universität Berlin Ablegen der Diplomprüfungen
SS 2002	Diplomarbeit an der Goethe-Universität Frankfurt

9. Eidesstattliche Erklärung

Ich erkläre an Eides Statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form zu keiner anderen Prüfung vorgelegt und auch noch nicht veröffentlicht.

Frankfurt am Main, 20. September 2002

10. Referenzen

[1] Mathews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405, 442-451.

[2] Voet, D. and Voet, J.G. (1995) Biochemistry, Second Edition. John Wiley & Sons, Inc., New York.

[3] MITOMAP: A Human Mitochondrial Genome Database. Center for Molecular Medicine, Emory University, Atlanta, GA, USA. http://www.gen.emory.edu/mitomap.html, 2001.

[4] McFadden, G.I. (2001) Primary and secondary endosymbiosis and the origin of plastids. J. Phycol. 37, 951-959.

[5] Kurland, C.G. und Andersson, S.G.E. (2000) Origin and evolution of the mitochondrial genome. Mol. Biol. Rev. 64(4), 786-820.

[6] von Heijne, G. (Ed.) (1994) Signal peptidases, R.G. Landes Company, Austin.

[7] Emanuelsson O., von Heijne G. und Schneider G. (2001) Analysis and prediction of mitochondrial targeting peptides. Meth. Cell Biol. 65, 175-187.

[8] Claros M.G. und Vincens P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur. J. Biochem. 241, 779-786.

9 Nakai K. und Kanehisa M. (1992) A knowledge base for predicting protein localizeiton sites in eukaryotic cells. Genomics 14, 897-911.

[10] Waller R.F., Reed M.B., Cowman, A.F. und McFadden, G.I. (2000) Protein trafficking to the plastid of Plasmodium falciparum is via the secretory pathway. EMBO J. 19, 1794–1802.

[11] Etzold, T., Ulyanov A. Und Argos P. (1996) SRS: information retrieval system for molecular biology data banks. Methods Enzymol. 266, 114-128. SRS WWW Service at the European Bioinformatics Institute http://srs6.ebi.ac.uk/.

[12] Bairoch A. und Apweiler R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMPL in 2000. Nucleic Acids Res. 28, 45-48

[13] The Plasmodium Genome Database Collaborative (2001). PlasmoDB: An integrative database of the *P. falciparum* genome. Tools for accessing and analyzing finished and unfinished sequence data. Nucl. Acids Res. 29, 66-69.

[14] Salzberg S.L., Pertea M., Delcher A.L., Gardner M.J., und Tettelin H. (1999) Interpolated Markov models for eukaryotic gene finding. Genomics 59(1), 24-31

[15] Chen T. und Zhang M.Q. (1998) Pombe: a gene-finding and exon-intron structure prediction system for fission yeast. Yeast 14(8), 701-710

[16] ClustalW WWW Service at the European Bioinformatics Institute http://www2.ebi.ac.uk/clustalw.

[17] JalView WWW Service at the European Bioinformatics Institute http://www.ebi.ac.uk/jalview/.

[18] Zuegge J., Ralph S., Schmuker M., McFadden G.I. und Schneider G. (2001) Deciphering apicoplast targeting signals--feature extraction from nuclear-encoded precursors of Plasmodium falciparum apicoplast proteins. Gene, 280, 19-26.

[19] Kawashima, S., Ogata, H., und Kanehisa, M. (1999). AAindex: amino acid index database. Nucleic Acids Res. 27, 368-369 .

[20] Schneider, G. und Wrede, P. (1998). Artificial neural networks for computer-based molecular design. Prog. Biophys. Mol. Biol. 70(3),175-222.

[21] Stevens, J. (1986) Applied multivariate statistics for the social sciences. Hillsdale, NJ.

[22] Kaiser, H. F. (1958) The varimax criterion for analytic rotation in factor analysis. Pyrometrical 23, 187-200.

[23] Kohonen, T. (1982) Self-organized formation of topologically correct feature maps. Biol. Cybern. 43, 59-69.

[24] Kohonen, T. (1984). Self-Organization and Associative Memory. Springer Series in Information Sciences 8 (3rd edition 1989), Springer Verlag, Heidelberg.

[25] Kohonen, T. (1989) Learning vector quantization for pattern recognition. Report TKK-F-A601. University of Technology, Helsinki.

[26] Zupan, J., Gasteiger, J. (1999) Neural Networks for Chemists (2nd edition). VCH, Weinheim.

[27] StatSoft, Inc. (2001) STATISTICA (data analysis software system), version 6. www.statsoft.com.

[28] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems 2, 303-314.

[29] Andrea, T.A. und Kalayeh, H. (1991) Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. J. Med. Chem. 34, 2824-2836.

[30] Manallack, D.T. und Livingston, D.J. (1995) Neural networks and expert systems in molecular design. In van de Waterbeemd, H. (Ed.), Advanced Computer-Assisted Techniques in Drug Discovery. VCH, Weinheim.

[31] Rumelhart, D.E., Hinton, G.E. und Williams R.J. (1986) Learning Internal Representations by Error Propagation. In Rumelhart, D.E. und McClelland J.L. (Eds.), Parallel Distributed Processing, Volume 1, MIT Press, Cambridge, MA

[32] Jones, W.P. und Hoskins, J. (1987) Back-Propagation. A generalized delta-learning rule. Byte, Oktober 1987, 155-162.

[33] Werbos, P.J. (1974) Beyond regression: new tools for prediction and analysis in the behavioural sciences. Ph.D. thesis, Harvard University, Boston, MA.

[34] Parker, D.B. (1985) Learning logic. Technical Report TR-47, MIT Center for Research in Computational Economics and Management Science, Cambridge, MA.

[35] Emanuelsson et. al. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol Biol. 3000, 1005-1016.

[36] Claros, M.G. und Vincens, P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur. J. Biochem. 241, 779-786. Web interface at http://www.mips.biochem.mpg.de/cgi-bin/proj/medgen/mitofilter.

[37] Emanuelsson O., Nielsen H., Brunak S. und von Heijne G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J.Mol.Biol 300, 1005-1016. Web interface at http://www.cbs.dtu.dk/services/TargetP/.

[38] The Plasmodium Genome Database Collaborative. (2001) PlasmoDB: An integrative database of the *P. falciparum* genome. Tools for accessing and analyzing finished and unfinished sequence data. Nucl. Acids Res. 29, 66-69.

[39] Zuegge J., Ralph S., Schmuker M., McFadden G.I. und Schneider G. (2001) Deciphering apicoplast targeting signals--feature extraction from nuclear-encoded precursors of Plasmodium falciparum apicoplast proteins. Gene 280(1-2), 19-26.

[40] Kawashima, S., Ogata, H., und Kanehisa, M. (1999) AAindex: amino acid index database. Nucleic Acids Res. 27, 368-369.

[41] Bishop, C. (1995) Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

[42] Goldberg, D. E. (1989) Genetic Algorithms. Addison Wesley, Reading, MA

[43] Lobry J.R. (1997) Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene (205), 309-316.

[44] Shepherd, A. J. (1997). Second-Order Methods for Neural Networks. Springer, New York.

[45] Bishop, C. (1995). Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

[46] Florent I. et al. (1998) A Plasmodium falciparum aminopeptidase gene belonging to the M1 family of zinc-metallopeptidases is expressed in erythrocytic stages. Mol. Biochem. Parasitol. 97(1-2), 149-60.

[47]McIntosh M.T. et al. (2001) Two classes of plant-like vacuolar-type H^+-pyrophosphatases in malaria parasites. Mol. Biochem. Parasitol. 114 (2), 183-195.

[48] The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796-815.