Domänen-Architektur von langen Signalpeptiden

- *in silico* und *in vitro* -

Dissertation

zur Erlangung des Doktorgrades

der Naturwissenschaften

vorgelegt beim Fachbereich Biowissenschaften

der Johann Wolfgang Goethe - Universität

in Frankfurt am Main

von

Jan Alexander Hiß

aus Groß-Gerau

Frankfurt 2008

(D 30)

vom Fachbereich Biowissenschaften der

Johann Wolfgang Goethe – Universität als Dissertation angenommen.

Dekan : …….

Gutachter :

Prof. Dr. Gisbert Schneider (JWG-Universität Frankfurt am Main)
Prof. Dr. Paul Wrede (Charité-Universitätsmedizin Berlin)

Datum der Disputation : .....

Inhaltsverzeichnis

Abkürzungsverzeichnis

ATP	Adenosintriphosphat
ATPase	Adenosintriphosphat hydrolysierendes Enzym
AS	Aminosäuren
bp	*base pair*
cTP	Chloroplasten-Transit-Peptide
DCBD2	Discoidin, CUB and LCCL domain-containing protein 2
DNA	*Desoxyribonukleinacid*
E. coli	*Escherichia coli*
EPG	e*pidermal growth factor*
ER	Endoplasmatisches Reticulum
GTP/GDP	Guanosintriphosphat / Guanosindiphosphat
gram-, gram+	Gram-negativ, Gram-positiv
GTPAse	Guanosintriphosphat hydrolisierendes Enzym
HEK	*human embryonic kidney*
HIV	*human immunodeficiency virus*
HHV	Humaner Herpesvirus
HLA	Histokompatibilitätsantigene
HMM	Hidden-Markov-Modell
HMMSTR	*Hidden-Markov-Model for local sequence-structure correlation*
KNN	Künstliches Neuronales Netz
LCMV	*lymphocytic choriomeningitis virus*
mcc	Matthews Korrelations-Koeffizient
MCMV	*murine cytomegalovirus*
MCF7Zellen	*human breast adenocarcinoma cell line (SUCHEN MAMA Cariz*
MCS	*multiple cloning site*
MHC	*major histocompatibility complex*
MPP	*mitochondrial processing peptidase*
MMTV	*mouse mammary tumor virus*
MIP	*mitochondrial intermediate peptidae*
mRNA	*messenger ribonucleic acid*
mTP	*mitochondrial targeting peptide*
NAC	*nascent polypeptide-associated complex*
NtraC-Modell	Modell zur Domänen-Architektur langer Signalpeptide
NKC	*natural killer cell*
PEXEL	*Plasmodium* export element Aminosäure-Motiv
PCR	*polymerase chain reaction*
RC	*Reliability Class*, TargetP Ausgabewert
RGMA	*repulsive guidance molecule* A
RNA	*ribonucleic acid*

rRNA	*ribosomal ribonucleic acid*
SA	Signalanker
SEAP	*secreted alkaline phosphatase*
7SL-RNA	*ribonucleic acid*, Komponente des SRP
shrew-1	*adherens junction-associated protein 1*
SP	Signalpeptid
SRP	s*ignal recognition particle*
SPS	Signalpeptidase-Schnittstelle
SRS	*sequence retrival system*
SVM	*support vector machine*
TMD	*transmembrane domain*
TMS	*transmembrane sequence*
tra	Übergangsbereich, engl. *transition area*
vSP	virale Signalpeptide

Einleitung

Die Mehrzahl aller Proteine einer eukaryotischen Zelle ist im Zellkern kodiert und wird im Cytosol synthetisiert (Boeckmann *et al.*, 2005). Innerhalb der eukaryotischen Zellen existieren jedoch eine Vielzahl von Kompartimenten mit unterschiedlicher Protein-Ausstattung (Dacks *et al.*, 2008). Man unterscheidet dabei vom endoplasmatischen Reticulum (ER) abstammende und von einer einzelnen Membran umschlossene Kompartimente und von mehreren Membranen umgebene semiautonome Kompartimente, die vermutlich durch Endosymbiose entstanden sind (de Duve, 2007; Tabak *et al.*, 2008). Eine Übersicht der wichtigsten eukaryotischen Zellkompartimente ist in Abbildung 1 gegeben. Um die individuelle Protein-Ausstattung dieser verschiedenen Kompartimente zu gewährleisten, werden diese post- und co-translational von Proteinen gezielt angesteuert (Abb. 1). Die vom ER abstammenden Kompartimente, mit Ausnahme der Peroxisomen (Tabak *et al.*, 2008), erhalten ihre Proteinausstattung hauptsächlich über Vesikel vom ER (Dacks *et al.,* 2008). Proteine für das ER selber, den endoplasmatischen Raum, für die von mehrfachen Membranen umgebenen Kompartimente (Plastide, Mitochondrien, Chloroplasten) sowie Proteine für die Peroxisomen enthalten Signalsequenzen innerhalb der Proteinsequenz (Blobel und Dobberstein, 1975a und 1975b). Signalsequenzen können z.B. C-terminal (Peroxisomen; Gould *et al.*, 1987 und 1988), N-terminal (Mitochondrien, Chloroplasten, endoplasmatisches Retikulum; Walter und Johnson, 1994; Patron und Waller, 2007) oder innerhalb der Sequenz (nicht-klassische Sekretion; Bendtsen *et al.*, 2004b) liegen.

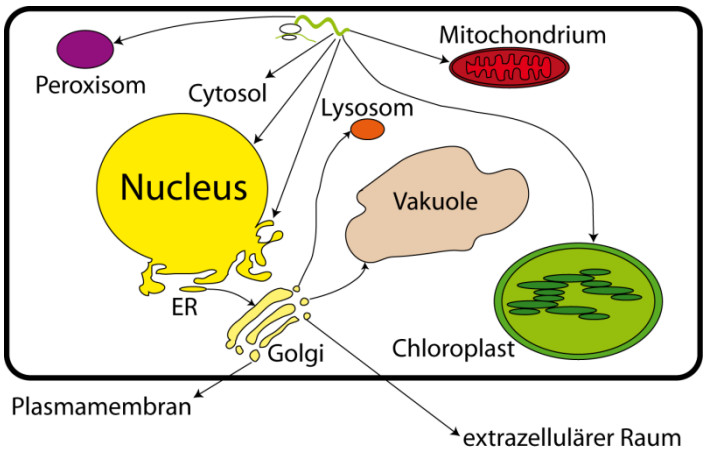


Abbildung 1: Auswahl an eukaryotischen Kompartimenten und deren Targeting-Wegen. **Grüne Linie**: im Cytosol synthetisiertes Protein. **Pfeile**: Mögliche Zielkompartimente neusynthetisierter Proteine. **ER**: Endoplasmatisches Retikulum. **Golgi**: Golgi-Apparat.

Cytosolische Proteine besitzen keine Signalsequenz. Eine grundlegende Unterscheidung, die auch bei Bakterien anzutreffen ist, ist somit die zwischen intra- und extrazellulärer Lokalisation, Protein mit Signalpepitd und Protein ohne ein solches (Wiech *et al*., 1991). Es wurde gezeigt, dass bestimmte Signalpeptide zwischen Prokaryoten und Eukaryoten ausgetauscht werden können (Hegde und Bernstein, 2006), was zur Sekretion von vorher cytosolischen Proteinen führt. Von Heijne, demonstrierte bereits 1984, dass sich eukaryotische und prokaryotische Signalpeptide in ihrem N-Terminus nicht in ihrer Netto-Ladung unterscheiden. Dies wird auf einen vergleichbaren selektiven Druck für diese Signale zurückgeführt (von Heijne, 1984a). In eukaryotischen Zellen ist neben der Unterscheidung intra- oder extrazellulär, Signalpeptid oder nicht, noch die Unterscheidung zwischen den zusätzlichen Kompartimenten (z.B. Mitochondrien, Chloroplasten, Peroxisomen, Zellkern) und deren Signalpeptiden zu berücksichtigen.

Welchen Vorteil hat die Zelle von dieser Unterteilung, die einen komplexen selektiven Sortierungsapparat benötigt? Funktional betrachtet erlaubt die Kompartimentierung der eukaryotischen Zelle, unterschiedliche physiologische Umgebungen zu schaffen. Dies kann z.B. bestimmten biochemischen Abläufen dienen (Fettsäure-Abbau in Peroxisomen; Tabak *et al*., 2008), der Erhaltung einer vom Cytosol abweichenden Ca^{2+} Konzentration (ER; Meldolesi und Pozzan, 1998), eines abweichenden pH-Wertes (Lysosomen; Nohl und Gille, 2005) oder eines Protonengradienten (Mitochondrien; Nicholls, 1974). Diese unterschiedlichen Bedingungen erlauben der Zelle, z.B. im Falle der Lysosomen, die für sie ansonsten schädlichen pH-Werte in einem vom Cytosol durch eine selektive Membran getrennten Kompartiment bereit zu halten. Im Falle von Peroxisomen können Proteine für den einen biochemischen Ablauf, den Fettsäure-Abbau, selektiv zu den Peroxisomen sortiert werden. Eine Auswahl an Proteinen, die für die Peroxisomen bestimmt sind, enthalten eine C-terminale Signalsequenz und werden post-translational und gefaltet importiert (Goldman und Blobel, 1978). Im Unterschied dazu besitzt die Mehrzahl an Proteine für das endoplasmatische Retikulum (ER) und somit für die Plasmamembran oder den extrazellulären Raum meist N-terminal liegende Signalsequenzen (Tabelle 1) und werden co-translational und ungefaltet ins Lumen des ER transloziert (Helenius *et al*., 1992; Braakman *et al*., 1992). Ein Protein inklusive Signalsequenz wird im Unterschied zum nativen Protein als „Präprotein“ oder Vorstufen-Protein bezeichnet.

Tabelle 1: Auswahl eukaryotischer Signalsequenzen für die Proteinsortierung. Angepasst nach Alberts *et al*., 2004, S. 772. **C:** C-Terminus, **N**: N-Terminus.

Funktion der Signalsequenz	**Beispiel für Signalsequenz**
Import in den Zellkern	...PPKKKRKV...
Export aus dem Zellkern	...LALKLAGLD...
Import in Mitochondrien	N–MLSLRGIRFFKPATRTLCSSRYLL
Import in ein Plastid	…ASLGSSMSSLSLSSNS…LSPITLS..
Import in ein Peroxisom	...SKL–C
Import in das ER	N–MMSFVSLLLVGILFWATEAEQLTKCE
Rückkehr (Verbleib) im ER	–…KDEL…–C

Wie funktioniert die Unterscheidung zwischen extra- und intrazellulären Proteinen? Dies wird erreicht über einen selektiven Transport zum ER. Hierfür und für die Translokation in das Lumen des ER sind in Säugerzellen drei Komponenten notwendig und *in vitro* hinreichend (Görlich und Rapoport, 1993):

- Das SRP (*signal recognition particle*) ist ein Ribonukleoprotein-Komplex aus sechs Proteinen (SRP9, SRP14, SRP19, SRP54, SRP68, SRP72) und einer 7SL-RNA (Walter und Blobel, 1980; Keenan *et al*., 2001). Die 7SL-RNA ist eine von der RNA-Polymerase-III transkribierte aber nicht translatierte RNA. Das SRP erkennt die Signalsequenz des neu synthetisierten Proteins und dirigiert den Komplex aus mRNA, Ribosom und Signalsequenz zum SRP-Rezeptor in der ER-Membran (Lipp *et al*., 1987).

- Der Sec61 heterotrimere Komplex aus den Untereinheiten Sec61α, Sec61β und Sec61γ) stellt den Translokationskanal in der ER-Membran dar (Görlich und Rapoport, 1993; Deshaies *et al*., 1991, Prinz *et al*., 2000a). In Kalies *et al*. (2008) wird postuliert, dass ein einzelner Sec61-Komplex für die Protein-Translokation hinreichend ist.

- Das TRAM-Protein (*translocating chain-associated membrane protein*) dient als ER-membranständiger Co-Faktor (Görlich *et al.*, 1992; Walter, 1992). Das TRAM-Protein ist ein Bestandteil des Translocons der ER-Membran. In Görlich *et al*. (1992) wurde gezeigt, dass es Gruppen von Proteinen gibt, für die TRAM notwendig und Gruppen, für die TRAM nicht notwendig ist, um eine Translokation in das Lumen des ER zu ermöglichen.

In Abb. 2 A-C ist der Weg eines Proteins vom Synthesestart im Cytosol bis zur co-translationalen Translokation in das Lumen des ER entsprechend der Lehrmeinung (Stryer, 1996; Alberts *et al*., 2004) dargestellt.

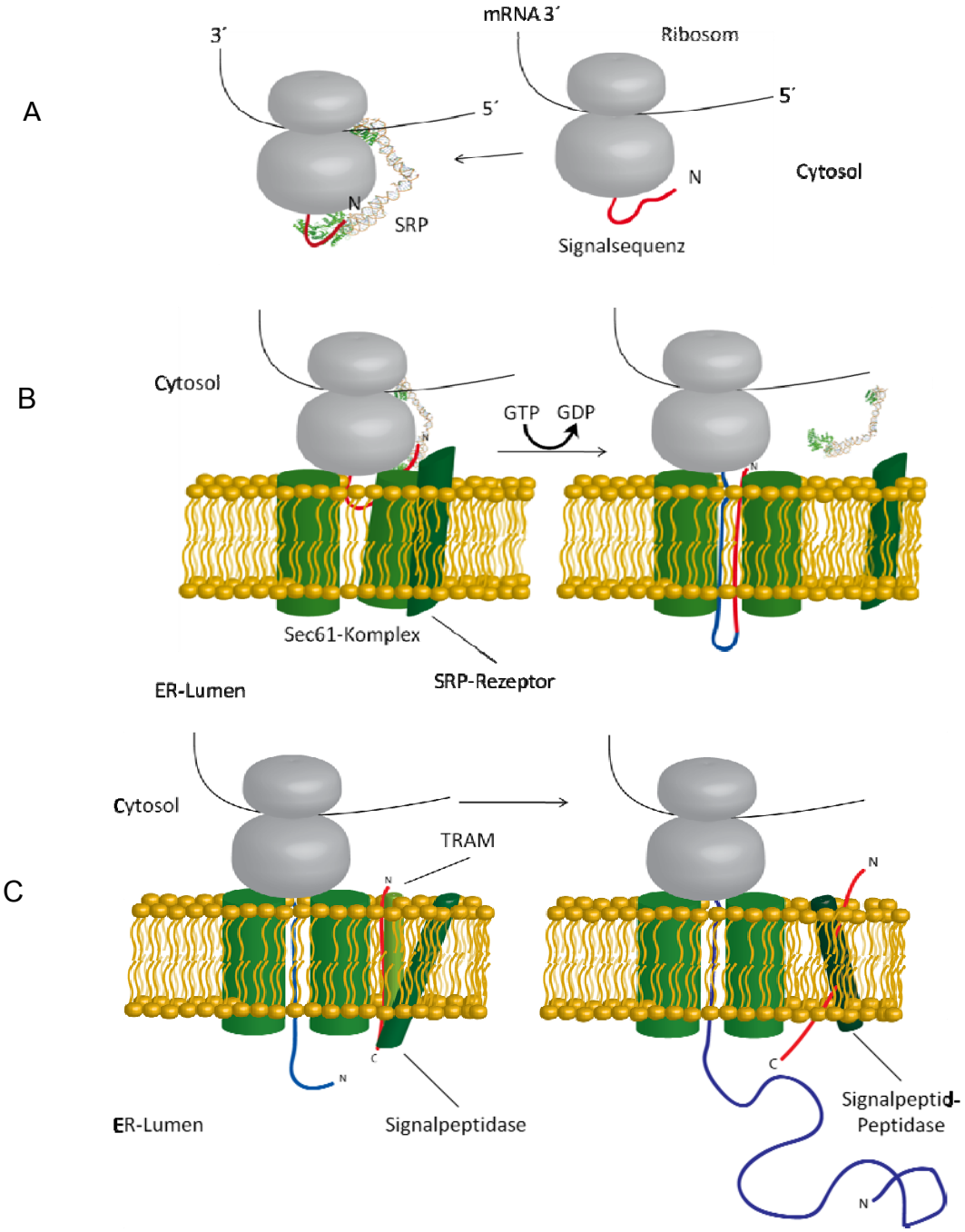


Abbildung 2: A: Translationsbeginn im Cytosol. Translationsarrest durch SRP-Interaktion mit dem Ribosom und der Signalsequenz. **B:** Erkennung des Ribosom/SRP-Komplexes an der ER-Membran durch den SRP-Rezeptor und den Sec61-Komplex. Freisetzung des SRP unter Hydrolisierung von GTP zu GDP. **C:** Übergabe der Signalsequenz (**rot**) an das TRAM-Protein und Abtrennung der Signalsequenz durch die Signalpeptidase. Fortsetzung der Translation des Proteins in das Lumen des ER **(blau)** und Abbau des Signalpeptides durch die Signalpeptid-Peptidase. **SRP**: *signal recognition particle* (Struktur nach Halic *et al*., 2004). **N**: N-Terminus. **C**: C-Terminus.

Die Signalsequenz wird während der Translation am Ribosom durch das SRP erkannt, es kommt zu einem Arrest der Translation (Lipp *et al*., 1987; Abb. 2a). Etwa 40 Aminosäuren der neu translatierten Aminosäure-Kette sind im Inneren des Ribosoms verborgen (Wolin und Walter, 1993). Nach der Translation von circa 50 Aminosäure-Resten wird das Signalpeptid vom SRP erkannt und es kommt zum Translationsarrest (Siegel und Walter, 1988; Wolin und Walter, 1993). Der Komplex aus Ribosom, SRP, mRNA und bereits translatierter Sequenz wird zur ER-Membran dirigiert (Abb. 2b). Hier findet die Erkennung des Komplexes durch den SRP-Rezeptor in der ER-Membran (Gilmore *et al.*, 1982) und eine Interaktion zwischen dem Sec61-Komplex und der 28S rRNA - große Untereinheit des Ribosoms - statt (Prinz *et al*., 2000b). Das SRP löst sich unter GTP-Hydrolyse ab (Bange *et al*., 2007; Zhang *et al*., 2008) und „übergibt" die Sequenz an den Translokationsapparat (sec61-Komplex und TRAM-Protein) in der ER-Membran (Abb. 2b). Die Signalsequenz liegt nun innerhalb der ER-Membran. Sie wird an der Signalpeptidase-Schnittstelle von der ER-membranständigen Signalpeptidase (Evans *et al*., 1986) abgetrennt und lateral in die Membran des ER freigesetzt (Abb. 2c). Die Translation des Proteins wird in das Lumen des ER fortgesetzt. Das abgespaltene Signalpeptid wird durch die Signalpeptid-Peptidase abgebaut (Weihofen *et al*., 2002; Abb. 2c). Die primäre Funktion des Signalpeptides ist somit die Direktion des Prä-Proteins zum ER.

In Prokaryoten existiert ein zu SRP54 homologes Protein (Ffh oder P48). Ffh fungiert entsprechend zu SRP54 als GTPase, die zusammen mit einer 4,5 S-RNA (so bennant in *Escherichia coli*) die Funktion des SRP übernimmt (Bernstein *et al.*, 1989, Keenan *et al*., 2001). Ein zur eukaryotischen SRP-Rezeptor Komponente SRα homologes Protein, das ebenfalls GTPAse Aktivität besitzt, existiert in Bakterien (FtsY) und liegt löslich oder mit der inneren Membran assoziiert vor (Keenan *et al*., 2001). In der inneren Membran wurde ein zu Sec61 homolger Protein-Kanal (SecYEF) gefunden, der zusammen mit dem ATP-abhängigen Motorprotein SecA für die Transloaktion von Proteinen über die innere Membran hinreichend ist (Hartl *et al.,* 1990; Driessen und Nouwen, 2008).

Eine Signalpeptid- und ER-unabhängige Methode zur Sezernierung von Proteinen in Eukaryoten stellen *signal patches* dar. *Signal patches* sind Sortierungssignale, die über die Primär-Struktur verteilt vorliegen. Durch die Faltung des Proteins (Tertiär-Struktur) kommen diese Bereiche in räumliche Nähe und werden dann als Sortierungssignal erkannt. *Signal patches* dienen zur Sekretion von Proteinen über den sogenannten „nicht-klassischen"- oder

leaderless-Sekretionsweg (Bendtsen *et al*., 2004b). Dabei werden z.B. die Proteine FGF-1, FGF-2, UGT1A6 und IL-1 unabhängig vom ER und nicht-glykosyliert direkt sezerniert (Rubartelli und Sitia, 1997; Ouzzine *et al.*, 1999; Hughes, 1999; Cooper, 2002). *Signal patches* können ebenfalls bei Proteinen mit Signalpeptid, nach Translokation in das Lumen des ER als Sortierungssignal innerhalb des Golgi-Apparates erkannt werden und zur selektiven Ansteuuerung von von Vesikel versorgten Kompartimenten dienen (Pfeffer und Rothman, 1987).

Können Signalsequenzen durch maschinelle Lernverfahren *in silico* erfasst werden? Die Signalsequenzen für den sekretorischen Weg (im Folgenden ER-Targeting) weisen nach Gierasch (1989) und Pugsley (1990) keine strenge Sequenz-Homologie auf. Im Mittel sind Signalsequenzen 22 Aminosäuren lang; zum Vergleich: Das C-terminale Signal für den Transport zum Peroxisom ist drei Aminosäuren lang. Signalsequenzen besitzen jedoch einen einheitlichen Aufbau (von Heijne, 1985). Die ER-Signalsequenz besitzt nach von Heijne, 1985 einen dreigeteilten Aufbau: *n*-, *h*- und *c*-Region (Abb. 3). Die N-terminal gelegene *n*-Region ist nach von Heijne zwischen einer und 17 Aminosäuren lang. Die *n*-Region ist somit stark variabel in ihrer Länge und enthält oft geladene Reste. Darauf folgen 7-16 Aminosäuren innerhalb einer aus vorwiegend hydrophoben Resten zusammengesetzten Kern- oder *h*-Region. Die Signalsequenz endet mit der 4-5 Aminosäuren langen, klar definierten Signalpeptidase-Schnittstelle, der *c*-Region (Abb. 2). Die Signalpeptidase-Schnittstelle folgt dabei der -3 -1-Regel (von Heijne, 1983; von Heijne, 1984b, von Heijne, 1986b): In Position -1 und -3 in Relation zur Signalpeptidase-Schnittstelle sind kleine, neutrale Aminosäuren häufig und aromatische, geladene und große Aminosäuren selten, mit der Ausnahme von Glutamin in Position -1. Dies ermöglicht die Vorhersage von Signalpeptiden und deren Schnittstelle durch maschinelle Lernverfahren (Schneider *et al*., 1993; Schneider und Wrede, 1994; Schneider und Fechner, 2004).

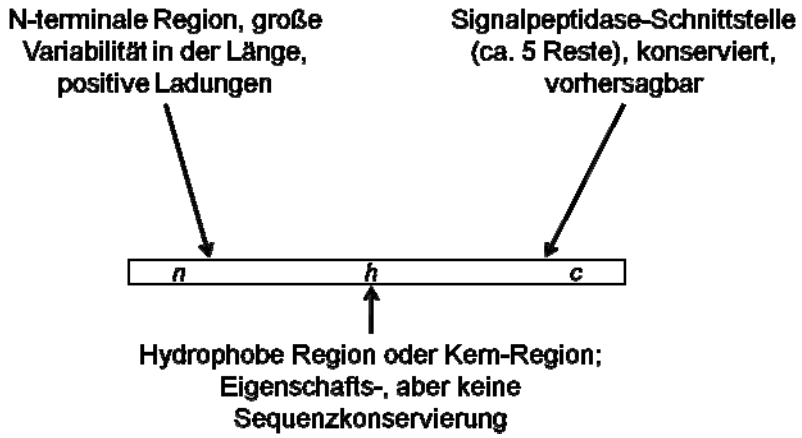


Abbildung 3: Allgemeiner Aufbau eukaryotischer N-terminaler Signalpeptide für den sekretorischen Pathway nach von Heijne, 1985.

Haben Signalpeptide mehr Funktionen als das Targeting zum ER?

Signalsequenzen können neben ihrer Funktion für die Sortierung von Proteinen, auch selber nach Abspaltung vom nativen Protein Funktionen übernehmen (Hegde, 2002). Im Folgenden ist eine Auswahl an Proteinen aufgelistet, deren Signalpeptide nach der Abspaltung durch die Signalpeptidase eine Post-Targeting-Funktion besitzen:

- Li *et al.*, 1996: Die ineffiziente Abspaltung der Signalsequenz des HIV-1 Glycoprotein gp120 hat Einfluß auf die Faltung, die intrazelluläre Lokalisation und die Assoziation des Proteins mit Calnexin.

- HIV Glycoprotein gp160 (Martoglio *et al*., 1997)

- Chen *et al*., 2001: Die Signalsequenz von Prepro-αfactor und Preinvertase muss abgeschnitten werden, damit eine Glykosylierung des Proteins stattfinden kann.

- Kim *et al*., 2002: Eine Substrat-Spezifität für den Translokations-Apparat ist innerhalb der Signalsequenzen kodiert und in homologen Proteine konserviert.

- Braud *et al*., 1997 und 1998; Long, 1998, Ulbrecht *et al.*, 2000: Die Signalsequenzen von klassischen MHC-I Proteinen (*major histo compatibility complexes,* HLA-A, -B, -C, und –G) werden nach Abspaltung durch die Signalpeptidase durch die Signalpeptid-Peptidase prozessiert und auf dem nicht-klassischen MHC-I Protein HLA-E präsentiert. Diese Präsentation führt zur Inhibierung von NKC (*natural killer cells)*.

- *Lassa virus* Glycoprotein-C (Eichler *et al*., 2003a)
- *Lymphocytic choriomeningitis virus* Glycoprotein-C (Schrempf *et al*., 2007)
- Preprolactin (Martoglio *et al*., 1997)

Dabei werden die Glycoproteine des *Lassa virus* und des *lymphocytic choriomeningitis virus* nach der Abspaltung durch die Signalpeptidase vermutlich nicht weiter prozessiert. Im Fall von HIV-1 GLycoprotein gp120 findet eine ineffiziente Abspaltung durch die Signalpeptidase statt (Li *et al*., 1996). Im Fall von HLA-A ist eine Prozessierung des Signalpeptides durch die Signalpeptid-Peptidase für die Post-Targeting-Funktion als HLA-E Epitop notwendig (Lemberg *et al*., 2001). Es wurde gezeigt, dass für einen Abbau durch die Signalpeptid-Peptidase eine vorhergehende Abspaltung des Signalpeptides sowie die Anwesenheit Helix-brechender Aminosäuren innerhalb der Transmembran-Segmentes (TMS) des Signalpeptides notwendig sind, der Abbau durch die Signalpeptid-Peptidase somit regulierbar ist (Lemberg und Martoglio, 2002). Der Abbau durch die Signalpeptid-Peptidase kann zur Freisetzung von Fragmenten des Signalpeptides aus der Membran des ER führen. Diese Signalpeptid-Fragmente können entweder im Lumen des ER weitere Funktionen übernehmen (HLA-A; Ulbrecht *et al.*, 2000) oder im Cytosol (Preprolactin und HIV-1 p-gp160 interagieren mit Calmodulin; Martoglio *et al*., 1997)

Welche Bedeutung hat die korrekte Sortierung von Proteinen für die Zelle? Die korrekte Sortierung der Proteine ist für die Zelle von Bedeutung, da eine fehlerhafte Sortierung die Ursache von verschiedenen Krankheiten ist (Tabelle 2). Das krankheitsverursachende Protein ist dabei nicht inaktiv oder abwesend, sondern wird intrazellulär inkorrekt sortiert.

Tabelle 2: Auswahl von Krankheiten, die auf fehlerhafter Proteinsortierung beruhen.

Krankheit / Pathogen	Ursache	Krankhafte Folge	Referenz
Mukoviszidose	Diverse Punktmutationen; häufig Position 508 (Phenlyalanin).	Retention eines Chloridkanals (CTFR) im ER, daher zu wenige Chloridkanäle in der Plasmamembran.	Chen *et al.,* 2000; Chen *et al*., 2004.

Muriner Cytomegalovirus (MCMV)	Fragmente des pathogenen gp48 werden nicht als Epitop erkannt. Sie werden im Komplex mit MHC-I in die Lysosomen umgeleitet.	Der MCMV-Virus entgeht der Detektion durch cytotoxische T-Zellen.	Bubeck *et al*., 2002; Loch und Tampé, 2005 (Review).
Humaner Herpesvirus (HHV-7)	Fragmente des pathogenen U21 werden nicht als Epitop erkannt.Sie werden im Komplex mit MHC-I in die Lysosomen umgeleitet.	Der HHV-7-Virus entgeht der Detektion durch cytotoxische T-Zellen.	Loch und Tampé, 2005 (Review); Hudson *et al.,* 2003.
Varicella-zoster Virus (VZV)	Fragmente des ORF66 Protein werden im Komplex mit MHC-I im Golgi zurückgehalten.	Der VZV-Virus entgeht der Detektion durch cytotoxische T-Zellen.	Abendroth *et al*., 2001; *Loch* und Tampé, 2005 (Review).
Adenovirus	Das E19 bzw. dessen Fragmente werden im Komplex mit MHC-I im ER zurückgehalten.	Der Adenovirus entgeht der Detektion durch cytotoxische T-Zellen.	Cox *et al*., 1991; Bennett *et* al., 1999; Loch und Tampé, 2005.
Thrombose	Eine Mutation an Position -3 des Signalpeptides in Antithrombin (V→E) führt zu einer neuen Signalpeptidase-Schnittstelle.	Antithrombin ist um zwei Aminosäuren verkürzt und daher inaktiv, es kommt zu Thrombosen.	Daly *et al*., 1990.
Hypoparathyroidism/ Hypocalcaemia	Eine Punktmutation (C18R) im hydrophoben Kern des Signalpeptides des Präproproteins des Hormons Preproparathyroid führt zu einem Rückhalt des Proteins im ER und somit zu keiner Sekretion des Proteins.	Die Abwesenheit des Hormons Preproparathyroid führt zu einer reduzierten Menge an freiem Calcium im Blut. Die nicht sekretierten Proteine im ER führen zu erhöhtem Zellstress und zur Apoptose der Zellen.	Arnold *et al*., 1990; Datta *et al*., 2007.

Reduzierte Blutgerinnung (Faktor $X_{Santo\ Domingo}$)	Eine Punktmutation an Position 3 im Signalpeptid des humanen Gerinnungsfaktors X (G→A) führt zu einer Verhinderung der Sezernierung des Proteins.	Die geringe Konzentration von Gerinnungsfaktor X im Blut führt zu einer reduzierten Blutgerinnung.	Watzke *et al*., 1991
Alzheimer	Eine Transmembran Variante des Prion Proteins PrP(Sc) – (Ctm)PrP enthält ein nicht geschnittenes Signalpeptid und wird im ER zurückgehalten.	Akkumulation des ansonsten membranständigen Prion Proteins führt zum Zelltod von Neuronen.	Stewart *et al*., 2001

Wie in Tabelle 2 zu erkennen ist, ist die subzelluläre Lokalisation von Proteinen sowohl für die Entstehung von Krankheiten basierend auf körpereigenen Fehlfunktionen als auch für Pathogene von Bedeutung. Diverse Viren haben Methoden entwickelt, um durch ein verändertes Targeting von Proteinen, z.B. zum Lysosom, der Detektion durch das Immunsystem zu entgehen. Auch für eukaryotische Pathogene ist eine Interaktion mit der Wirtszelle von entscheidender Bedeutung. *Plasmodium falciparum* als intrazellulärer Parasit sezerniert für seine Pathogenität entscheidende Proteine in das Lumen infizierter Erythrozyten und integriert diese Proteine in die Plasmamembran des Erythrozyten (Cooke *et al*., 2004). Der Parasit lebt dabei in einem durch das Eindringen in den Erythrozyten entstandenen Kompartiment, der parasitären Vakuole. Es wurde gezeigt, dass diese Proteine, die zusätzlich die Membran dieser parasitären Vakuole überwinden müssen, ein bestimmtes Sequenzmotiv enthalten: VTS bzw. PEXEL (Hiller *et al*., 2004; Marti *et al*., 2004). Die Identifikation solcher Protein-Targeting-Sequenzen und Sequenzmotive mit Hilfe von maschinellen Lernverfahren ist bereits mehrfach erfolgreich angewendet worden (Schneider *et al*., 1994; Schneider und Wrede, 1994; Schneider und Wrede, 1998). Das entsprechend identifizierte PEXEL-Motiv tritt dabei bei sezernierten Proteinen *mit* und *ohne* Signalsequenz auf. Von uns wurde kürzlich gezeigt, dass nach *in silico*-Abspaltung der Signalsequenzen die Position des PEXEL-Motivs in Proteinen mit und ohne Signalsequenz zur Deckung kommt

(Hiss *et al.*, 2008a). Dies ist ein Hinweis auf eine Erweiterung des klassischen Sekretionsweges mit Hilfe von Signalsequenzen. Die Sortierung von Proteinen in Zellen durch Signalpeptide stellt somit ein für den reibungslosen Ablauf in Zellen entscheidendes Kriterium da. Das Vorhersagen der korrekten oder fehlerhaften subzellulären Lokalisation von Proteinen ist daher bei der Analyse von Krankheiten von Bedeutung.

Motiviert durch die Erkenntnis, dass Signalpeptide Post-Targeting-Funktionen haben können Hedge (2002) und dass dies potentiell in Korrelation zu deren Länge steht, beschäftigt sich diese Arbeit mit der Analyse langer (≥40 Aminosäuren) Vertebrata-Signalpeptide. Es wurde analysiert:

- Welchen Einfluß haben die Länge der Signalpeptide und deren Aufbau auf die Existenz von Post-Targeting-Funktionen?
- Lassen sich Signalpeptide aufgrund ihrer Länge funktional gruppieren?
- Gibt es Unterschiede in den Signalpeptiden, die für verschiedene eukaryotische Kompartimente kodieren?
- Haben lange Signalpeptide einen von kurzen Signalpeptiden abweichenden Aufbau bzw. besitzen sie eine interne Organisation?

Material und Methoden

Der Material- und Methodenteil ist in einen bioinformatischen und einen biologisch-experimentellen Abschnitt untergliedert.

Bioinformatische Methoden

Online verfügbare bioinformatische Programme

Im Rahmen der Arbeit wurden die im Folgenden aufgelisteten, frei verfügbaren bioinformatischen Programme („Tools“) verwendet. Soweit nicht anders vermerkt, wurden die jeweils hier angegebenen Parametereinstellungen verwendet.

SignalP (http://www.cbs.dtu.dk/services/SignalP/) in der Version 3.0 (Bendtsen *et al*., 2004a).

SignalP 3.0 ist eine auf künstlichen neuronalen Netzen (KNN) und Hidden-Markov-Modellen (Nielsen und Krogh, 1998) basierende Software zur Vorhersage von Signalsequenzen in Eukaryoten und Bakterien. SignalP 3.0 wurde ebenfalls in einer für akademische Nutzung freien lokalen Version in das im Rahmen dieser Arbeit entwickelte Programm „NtraC“ integriert. Der Nutzungsvertrag ist im Anhang A1 zu finden.

Verwendete Einstellungen in der englischen Originalbezeichnung der webbasierten Benutzeroberfläche:

Output format:	Standard
Method:	both (neural networks + Hidden Markov models)
Truncation:	120 residues
Graphics:	GIF (inline)

Bei Verwendung von SignalP 3.0 im Rahmen der NtraC-Vorhersage werden die folgenden abweichenden Einstellungen verwendet:

Output format:	short (no graphics)
Method:	both (neural networks + Hidden Markov models)
Truncation:	100 residues
Graphics:	no graphics

Nach Angaben aus Bendtsen *et al.*, 2004a wurde für SignalP 3.0 auf einem positiven Beispielsatz mit 1192, 334 und 153 Sequenzen für Eukaryoten, Gram-negative und Gram-positive Bakterien trainiert. Der Matthews Korrelations-Koeffizient (*mcc*; Gleichung 1, Matthews, 1975) für die drei Datensätze ist im Folgenden angegeben:

- SignalP 3.0 Neuronales Netzwerk
  - Eukaryota *mcc* = 0,98
  - Gram-negative Bakterien *mcc* = 0,95
  - Gram-positive Bakterien *mcc* = 0,98

- SignalP 3.0 Hidden- Markov-Modell
  - Eukaryota *mcc* = 0,94
  - Gram-negative Bakterien *mcc* = 0,94
  - Gram-positive Bakterien *mcc* = 0,98
  -

$$mcc = \frac{tp \times tn - fp \times fn}{\sqrt{(tp+fn)(tp+fp)(tn+fp)(tn+fn)}}, \qquad \text{(Gleichung 1)}$$

wobei:

tp = true positive, Anzahl positiv-korrekte Vorhersagen.

fp = false positive, Anzahl falsch als positiv vorhergesagter Instanzen.

tn = true negative, Anzahl negativ-korrekte Vorhersagen.

fn = false negative, Anzahl falsch als negativ vorhergesagte Instanzen.

Die Ergebnisse wurden jeweils fünffach kreuzvalidiert. Nach Nielsen *et al*., 1997 wurden für das Training der künstlichen neuronalen Netze bei eukaryotischen Sequenzen für den C-*Score* zwei und für den Y-*Score* vier Hidden-Neuronen verwendet.

Die Trainingsdaten wurden dabei mithilfe des unterschiedlichen isoelektrischen Punktes von Signalpeptiden und deren Proteinen von falsch-positiven „preproteins“ gereinigt. Diese “preproteins” konnten auch durch die unterschiedliche Aminosäure-Zusammensetzung zwischen nativem Protein und Signalpeptid identifiziert werden (Cedano *et al*., 1997; Chou, 2001). Des Weiteren wurde festgestellt, dass bestimmte Aminosäuren selten bei Signalpeptidase-Schnittstellen auftreten (Bendtsen *et al*., 2004b). Beispiele, die im eukaryotischen Trainingssatz Lysin (Anzahl: 7) oder Arginin (Anzahl: 12) an der -1 Position relativ zur Signalpeptidase-Schnittstelle besaßen, wurden entfernt. Alle positiven

Trainingssequenzen besaßen bei den Eukaryoten eine der folgenden Aminosäuren an Position -1: Alanin, Cystein, Glycin, Leucin, Prolin, Glutamin, Serin oder Threonin.

Alle Phagen- und Virussequenzen wurden aus dem Trainingssatz entfernt.

Die SignalP Neuronale-Netz-Vorhersage erfolgt jeweils für ein Fenster, das über die Sequenz gleitet und eine bestimmte Zahl an Aminosäuren umfasst. Die Vorhersage für diese Fenster wird dann zu einem Punkt in der Ausgabe zusammengefasst. Für die Signalpeptid-Erkennung wurde bei Eukaryoten ein symmetrisches Fenster mit 27 Aminosäuren verwendet. Für die Vorhersage der Signalpeptidase-Schnittstelle wurde ein asymmetrisches Fenster mit 20 Aminosäuren *upstream* und vier Aminosäuren *downstream* der aktuellen Position verwendet.

Die entscheidende Neuerung in SignalP 3.0 im Vergleich zum Vorgänger ist die Einführung des D-*Score* für die Signalpeptidase-Schnittstellen-Vorhersage. Dieser D-*Score* stellt eine Kombination aus dem Durchschnitt des mittleren S-*Score* und dem maximalen Y-*Score* dar (Gleichung 2). Er erweitert den S-*Score*, der in SignalP 2.0 zur Unterscheidung zwischen sezernierten und nicht sezernierten Proteinen verwendet wurde.

$$Y_i = \sqrt{C_i \Delta_d S_i}\,, \qquad \text{(Gleichung 2)}$$

wobei $\Delta_d S_i$ die Differenz zwischen mittlerem S-*Score* von *d* Positionen vor und nach der Position *i* ist (Gleichung 3).

$$\Delta_d S_i = \frac{1}{D}\left(\sum_{j=1}^{d} S_{i-j} - \sum_{j=0}^{d-1} S_{i+j}\right)\,. \qquad \text{(Gleichung 3)}$$

Zusätzlich wird die Länge der vorhergesagten Sequenz in den endgültigen *Score* mit einbezogen. Die durchschnittliche Länge von Signalpeptiden beträgt bei Eukaryoten 22 und bei Gram-negativen und Gram-positiven Bakterien 24 Aminosäuren. Das neuronale Netzwerk bestraft beim Verschieben des Analyse-Fensters über die Sequenz Positionen, die diesen Durchschnitt über- oder unterschreiten. Damit werden extrem lange bzw. extrem kurze Signalpeptide teilweise nicht erkannt. Das Hidden-Markov-Modell (HMM) bestraft lange Signalpeptide implizit durch seine Struktur, d.h. wiederholtes Durchlaufen Aminosäure

emittierender Zustände wird sukzessive unwahrscheinlicher. Daher ist auch das HMM nicht in der Lage, lange Signalpeptide in bestimmten Fällen vorherzusagen.

Beispiel: NUC_STAAU (Thermonuclease aus *Staphylococcus aureus*) ist in Swiss-Prot mit einem Signalpeptid mit 26 Aminosäuren Länge annotiert. In der Original-Publikation (Miller *et al*., 1987) werden zwischen 60 und 79 Aminosäuren beschrieben, Bendtsen *et al*. 2004a spricht von 63 Aminosäuren Länge, SignalP 3.0 schlägt 28 vor.

TargetP (http://www.cbs.dtu.dk/services/TargetP/) in der Version 1.1 (Emanuelsson *et al*., 2000; Emanuelsson *et al*., 2007).

TargetP 1.1 ist eine auf neuronalen Netzen (KNN) basierende Software zur Vorhersage von mitochondrialen Targeting-Peptiden. TargetP wurde in einer für akademische Nutzung freien lokalen Version in das im Rahmen dieser Arbeit entwickelte Programm „NtraC“ integriert. Der Nutzungsvertrag ist im Anhang A1 zu finden.

Verwendete Einstellungen in der englischen Originalbezeichnung der webbasierten Benutzeroberfläche:

Organism Group:	non-plant
Prediction scope:	deselected
Cutoffs:	no cutoffs

Wenn nicht anders vermerkt, wurden alle Vorhersagen bezüglich der mTP (mitochondrial targeting peptide) Targeting-Kapazität mit TargetP 1.1 durchgeführt.

TargetP besitzt drei unabhängig voneinander trainierte neuronale Netze für die Vorhersage von Signalpeptiden, Chloroplasten-Transit-Peptiden und mitochondrialen Targeting-Peptiden. Die KNN enthielten jeweils vier Neuronen in der „versteckten Schicht“. Die drei Netze wurden in einem neuen Netz zusammengefasst, dieses wurde ohne eine versteckte Schicht trainiert (Abb. 4). Als Eingabe für dieses „integrative Netzwerk“ dienten jeweils die Werte der ersten 100 aminoterminalen Reste (Emanuelsson *et al*., 2000).

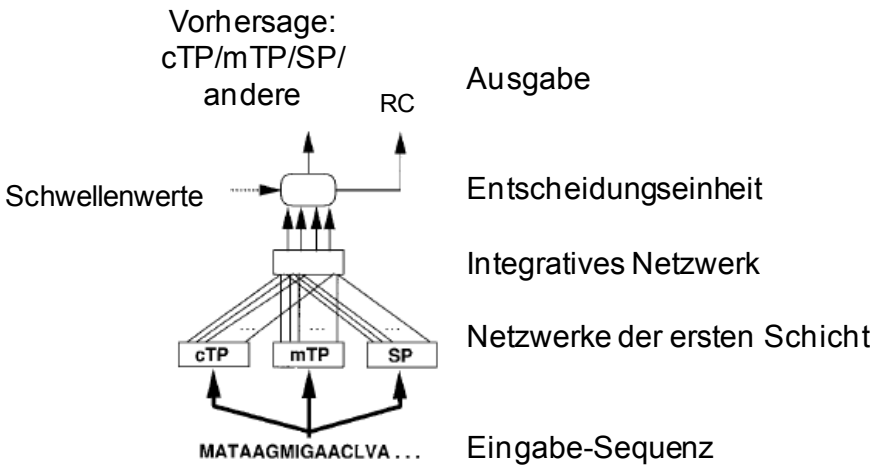


Abbildung 4: Aufbau der in TargetP verwendeten KNN (übernommen, angepasst und übersetzt aus aus Emanuelsson *et al*., 2000). **Netzwerke der ersten Schicht** treffen eine voneinander unabhängige Aussage, ob die Eingabe-Sequenz ein cTP, mTP oder SP enthält. Das **integrative Netzwerk** fasst die Ausgaben dieser Netze zusammen. Die **Entscheidungseinheit** trifft, basierend auf den Schwellenwerten des Benutzers, die finale Entscheidung, ob die Sequenz ein cTP, mTP, SP, ein anderes oder kein Signal enthält. **RC**: *reliability class*, der Unterschied zwischen dem höchsten und dem zweithöchsten Ausgabewert. Je höher der Unterschied, desto niedriger die RC Klasse, und umso verlässlicher die Vorhersage.

Die Fenstergröße, mit der der *Score* für jede Sequenzposition errechnet wird, beträgt bei mTP 35 Reste. Der Matthews Korrelations-Koeffizient (*mcc*) für mTP-Vorhersagen war 0,77 (Gl. 1). In der Arbeit von Emanuelsson *et al*., 2000 wird gezeigt, dass das Programm PSORT bei vergleichbarer Aufgabenstellung einen *mcc* von 0,64 hat. Das Programm PSORT wurde in der vorliegenden Arbeit daher nicht verwendet.

SVMTurn (http://gecco.org.chemie.uni-frankfurt.de/SVMTurn/SVMTurn.html) in der Preliminary Trial Version (Meissner *et al*., 2008).

SVMTurn ist ein auf Support-Vektor-Maschinen (Cai *et al*., 2003; Zhang *et al*., 2005) basierender Algorithmus zur Vorhersage von β-Turn-Strukturen in Proteinen. Ergebnisse der β-Turn-Vorhersage (SVM-*Score*) wurden in das im Rahmen dieser Arbeit entwickelte Programm „NtraC“ integriert.

Verwendete Einstellungen in der englischen Originalbezeichnung der webbasierten Benutzeroberfläche:

Prediction of turns with length of:	4 residues
Threshold value:	0.9995

Die SVM wurden mit Hilfe der Software SVMlight trainiert (Joachims, 1999). Die Vorhersage „Turn/Nicht-Turn“ wurde einer 10-fachen 50/50 Kreuzvalidierung unterzogen. Der *mcc* lag bei 0,55. Bei einem Training auf einem zweiten Datensatz, das nur durch Wasserstoffbrücken stabilisierte β-Turns enthielt, lag der *mcc* bei 0,58. Die auf diesem Trainingssatz erzeugten SVM-Scores für vier Aminosäuren-β-Turns wurden im Rahmen des NtraC-Modells verwendet.

Prosite ScanProsite (http://www.expasy.org/tools/scanprosite/?Q2PFL5) Release 20.36.

Prosite ScanProsite ist ein Vorhersage-Programm für Domänen in Proteinsequenzen (Sigrist *et al*., 2002; Hulo *et al*., 2006; de Castro *et al*., 2006). Es stellt ein regelbasierendes Suchprogramm dar (Sigrist *et al*., 2002), das mit Hilfe eine webbasierten Benutzeroberfläche (de Castro *et al*., 2006) eine gegebene Protein-Sequenz durchsucht. Die Aussage des Programms stellt somit keine Wahrscheinlichkeit dar, sondern ist binär: Das Muster wird gefunden oder nicht. In der aktuellen Version (20.36) enthält PROSITE 1449 Dokumentationen über Domänen. Diese Domänen können über 1331 Muster gesucht werden. Dabei werden die 1331 Muster mit 737 Regeln kombiniert bzw. erweitert. Wenn über ein Muster eine Protein-Domäne identifiziert wurde, kann eine zusätzliche Regel greifen, die z.B. an die Existenz bestimmter Reste in strukturell kritischen Bereichen geknüpft ist. Auch innerhalb einer Protein-Familie auftretende Kombinationen aus mehreren Mustern können in einer Regel erfasst werden. Damit kann neben der Zuordnung der Domänen durch die Mustererkennung auch eine Zuordnung zu einer Protein-Familie mithilfe einer Regel getroffen werden.

Sequence retrieval system (SRS) (http://srs6.ebi.ac.uk) Release 7.1.3 (Zdobnov *et al.,* 2002a and 2002b; Etzold *et al.,* 2003; Harte *et al*., 2004).

Das SRS stellt eine Suchmaske zur Recherche in diversen Protein- und Nukleotidsequenz Datenbanken zur Verfügung. Mithilfe des SRS wurden in dieser Arbeit Suchanfragen in der Proteinsequenz UniProtKB-Datenbank (Kapitel „Datenbanken“) durchgeführt.

Verwendete Einstellungen, die bei allen Suchen gleich waren, bzw. Suchfelder, die unterschiedlich belegt bei allen Suchen verwendet wurden, sind im Folgenden angegeben (Angaben in der englischen Originalbezeichnung entsprechend der webbasierten Benutzeroberfläche):

Library Page
Available Databanks: UniProtKB

Extended Query Form
Taxonomy: (entsprechend der Fragestellung) z.B. “Eukaryota“ oder “*virus“

ProteinExistence: *AND* : “evidence at protein level”

Feature subentry fields:	FtKey	*AND* “Signal”
	FtLength	nach Fragestellung z.B. “≥ 40“
	FtDescription	nach Fragestellung z.B. “≤200“

Die Ausgabe wurde jeweils im FASTA-Format (Abschnitt „Datenformate“) als Textdatei gespeichert.

„*“ bedeutet eine beliebige Anzahl und Auswahl an Buchstaben. Alle Suchfelder wurden mit „AND“ verknüpft. Bei dem Feld „ProteinExistence“ und „FtKey“ ist die Auswahl zusätzlich unabhängig von der globalen Selektion zu treffen. In allen Feldern werden vordefinierte Schlüsselbegriffe verwendet. Eine in diesem Sinne falsche Eingabe, z.B. *Eukaryoten* statt *Eukaryota,* führt zu keinem Ergebnis.

ClustalW2 (http://www.ebi.ac.uk/Tools/clustalw2/index.html) Version 2.0 (Larkin *et al*., 2007).

ClustalW2 ist eine in C++ geschriebene Software für das multiple Alignment von Proteinsequenzen.

Verwendete Einstellungen in der englischen Originalbezeichnung der webbasierten Benutzeroberfläche:

Gap Open Penalty	=	10
Gap Extension Penalty	=	1
matrix	=	Gonnet 250
ENDGAP	=	-1
GAPDIST	=	4

HMMSTR/Rosetta Prediction Server (http://www.bioinfo.rpi.edu/~bystrc/hmmstr/) Version vom 27.08.2007 (Bystroff *et al*., 2000; Bystroff und Shao, 2002).

HMMSTR basiert auf Hidden-Markov-Modellen (Rabiner, 1989). Ein Hidden-Markov-Modell ist eine Kette von verbundenen Zuständen. Jeder Zustand kann mit einer gewissen

Wahrscheinlichkeit z.B. eine Aminosäure emittieren oder zu einem anderen Zustand wechseln. Das hier verwendete HMM kann Sekundärstruktur-Motive mit einer Genauigkeit von 74% vorhersagen. Für die Kreuzvalidierung wurden homologe Proteine mit Hilfe von PSIBLAST (Altschul *et al*., 1997) gesucht und aus dem Trainingssatz entfernt.

Datenformate

FASTA-Format: Format zum Speichern von Sequenzdaten (Pearson und Lipman, 1988). Eine FASTA-Datei enthält eine Beschreibungszeile, gefolgt von beliebig vielen Sequenzzeilen. Die Beschreibungszeile beginnt mit einem „>“. Es werden 80 Buchstaben pro Zeile empfohlen. Aminosäuren werden im Standard Ein-Buchstaben-Code beschrieben (Tabelle 3). Es ist möglich, mehrere Einträge in einer FASTA-Datei zusammenzufassen. Jeder Eintrag muss mit einem „>“ beginnen. Ein Beispieleintrag der Sequenz der ATPase Untereinheit 8 aus *Podospora anserina* ist im Folgenden dargestellt:

```
>UniProtKB|Q02653|ATP8_PODAN ATP synthase protein 8;
MPQLVPFYFVNEITFTFIILAITVYILSKYILPRFVRLFLSRTFISKLLG
```

Tabelle 3: Endogene Aminosäuren im Ein- und Drei-Buchstaben-Code nach Stryer, 1996, S.23.

Aminosäure	Ein-Buchstaben-Code	Drei-Buchstaben-Code
Alanin	A	Ala
Cystein	C	Cys
Aspartat	D	Asp
Glutamat	E	Glu
Phenylalanin	F	Phe
Glycin	G	Gly
Histidin	H	His
Isoleucin	I	Ile
Lysin	K	Lys
Leucin	L	Leu
Methionin	M	Met
Asparagin	N	Asn
Prolin	P	Pro
Glutamin	Q	Gln
Arginin	R	Arg
Serin	S	Ser
Threonin	T	Thr
Valin	V	Val
Tryptophan	W	Trp
Tyrosin	Y	Tyr

Datenbanken

Sequenzinformationen zu Proteinen wurden aus den folgenden öffentlich zugänglichen Datenbanken entnommen:

UniProtKB, Version 13.6 und 14.0 (Verwendete Version jeweils angegeben; Wu *et al*., 2007). Zugriff erfolgte über das SRS. Am 30.07.2008 waren 6.462.751 Protein-Einträge indiziert (http://www.uniprot.org/).

Die UniProtKB Protein-Datenbank ist im Jahr 2002 aus der Fusion der

- Swiss-Prot (Swiss Institute of Bioinformatics (SIB); Geneva, Switzerland)
- TrEMBL (European Bioinformatics Institute (EBI); Hinxton, United Kingdom)
- und PIR (Protein Information Resource (PIR); Washington DC, USA)

Datenbank entstanden. Sie enthält Sequenzdaten, Publikationsverweise und Annotationen, die die Funktion der Proteine betreffen.

NCBI Entrez Protein (National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/ ; Referenz für alle NCBI Datenbanken: Wheeler *et al*., 2003).

Es handelt sich bei NCBI Entrez Protein-Datenbank um eine Protein-Sequenz-Datenbank, die aus den folgenden Datenbanken zusammengestellt wurde: SwissProt, PIR, PRF, PDB und Übersetzung von annotierten kodierenden Regionen in GenBank und RefSeq.

NCBI Entrez Gene (National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov/; Referenz für alle NCBI Datenbanken: Maglott *et al*., 2007).

Die NCBI Entrez Gene-Datenbank ist eine Gen-Sequenz-Datenbank. Sie enthält ebenfalls Informationen zu mRNA-Sequenzen und ist in sechs Organismen-Gruppen unterteilt: Archaea, Bacteria, Eukaryota, Viruse, Viroide und Plasmide.

Software

Im Rahmen dieser Arbeit wurden die folgenden im Arbeitskreis von Prof. Dr. Anna Starzinski-Powitz oder im Arbeitskreis von Prof. Dr. Gisbert Schneider verfügbaren Programme verwendet:

Clone Manager Professional Suite

Version 8, Scientific & Educational Software, Cary NC 27513 USA. Mithilfe dieser Software wurden alle in dieser Arbeit verwendeten Primer/Oligonucleotide entworfen.

Imaris

Version 5.0.3 von Bitplane, 8048 Zürich, Schweiz. Alle Bilder des konfokalen Laser-Mikroskops wurden mithilfe dieser Software bearbeitet.

Matlab (The Mathworks, Natick, Massachusetts) Version 7.6, Release 2008a mit folgenden zusätzlichen Modulen:

Bioinformatics Toolbox	Version 3.1
Fixed-Point Toolbox	Version 2.2
Fuzzy Logic Toolbox	Version 2.2.7
Genetic Algorithm and Direct Search Toolbox	Version 2.3
MATLAB Compiler	Version 4.8
Neural Network Toolbox	Version 6.0
Optimization Toolbox	Version 4.0
Signal Processing Blockset	Version 6.7
Signal Processing Toolbox	Version 6.9
Simulink Fixed Point	Version 5.6
Statistics Toolbox	Version 6.2

Verwendet wurde der zweiseitige Kolmogorov-Smirnov-Test, wie er in Matlab implementiert ist. Der Kolmogorov-Smirnov ist ein kumulativer, von der Verteilung unabhängiger Test, der feststellt, ob zwei Verteilungen der gleichen Grundgesamtheit entstammen (Rassokhin und Agrafiotis, 2000). Der im Rahmen dieser Arbeit entwickelte NtraC-Algorithmus (Kapitel „Ergebnisse“) ist in Matlab implementiert worden. Für die webbasierte Benutzeroberfläche wurde mithilfe des Matlab Compilers eine Standalone-Version erzeugt.

Pymol Version 1.1r1

DeLano, W.L. The PyMOL Molecular Graphics System (2002) DeLano Scientific, Palo Alto, CA, USA. Alle in dieser Arbeit gezeigten Protein-Struktur-Bilder wurden mithilfe von Pymol erzeugt. Des Weiteren wurde Pymol zur Betrachtung aller PDB Einträge verwendet.

Adobe©Illustrator CS2

Version 12.0 , Adobe Systems GmbH, München, Deutschland. Adobe Illustrator wurde zur Erstellung der selbstgestalteten Abbildungen dieser Arbeit verwendet.

Biologische Materialien und Methoden

Polymerase Chain Reaction

Die Polymerase Chain Reaction PCR wurde in PCR-Cyclern der Firma Biometra (Personal Cycler) und MJ-Research (PTC-100) durchgeführt.

Die PCR dient zur selektiven Amplifikation von Nukleotid-Sequenzen. Problemspezifisch entworfene Primer-Sequenzen binden dabei an gegebene Template. Eine hitzebeständige Polymerase baut unter der Verwendung zugesetzter Nucleotide den Gegenstrang der gewünschten Sequenz. Eine PCR wurde in fünf Schritten durchgeführt:

1) Denaturierung der DNA bei 94^{o}C für 10 min
2) Denaturierung der DNA bei 94^{o}C für 30 sek
3) Anlagerung der Primer bei 52^{o}C für 45 sek
4) Elongation bei 68^{o}C (pfx) bzw. 72^{o} C (pfu)
   für 2 min (pfx) bzw. 3,5 min (pfu)
5) 35-fache Wiederholung der Schritte 2-4 und abschließend eine 10-minütige Elongationsphase bei 68^{o}C (pfx) bzw. 72^{o}C (pfu).

Zell-Linien

Alle Experimente zur Lokalisation (SEAP-Assay) wurden in HEK 293T Zell-Linien durchgeführt. HEK-Zellen sind humane fötale Nierenzellen (LGC Promochem, http://www.lgcpromochem-atcc.com/ →CRL-11268™).

Die Zellen wurden mithilfe von MATra (Magnet Assisted Transfection, IBA) transfiziert.

Wenn entsprechend vermerkt, wurden für die Immunfluoreszenz MCF7-Zellen verwendet. MCF7-Zellen sind humane Brustdrüsenkrebszellen (LGC Promochem, http://www.lgcpromochem-atcc.com/ →HTB-22™).

Die Zellen wurden mithilfe von jetPEI^{TM} (poly plus transfection) transfiziert.

Die biochemische Aufreinigung von Mitochondrien erfolgte aus HEK 293T Zellen.

Zellmedium

Als Medium für die Zell-Linien wurde DMEM (Dulbecco´s Modified Eagle´s Medium) verwendet, mit 10% FCS (Fötales Kälberserum) und 1% Penicillin/Streptomycin-Lösung.

Bakterienstamm

Alle Plasmide wurden in *Escherichia coli* DH5α Zellen amplifiziert und kloniert (Bachmann, 1983).

Plasmide

In der Arbeit wurden die folgenden Plasmide eingesetzt:

- pcDNA3.1(-) Leervektor (Invitrogen)
- pGEM©-T-Easy-Vektor (Promega)

(-) steht für die Orientierung der multiplen Klonierungsstelle (MCS, multiple cloning site). pcDNA3.1(-) dient zur Expression der Konstrukte in Säugerzellen. Der T-Easy-Vektor dient in einem Vorschritt zur Transformation kompetenter Bakterien. Er ermöglicht durch ein *Blau-Weiß-Screening* gewachsener Kolonien die Auswahl erfolgreich transformierter Bakterienklone. Das *Blau-Weiß-Screening* wird ermöglicht, da auf dem T-Easy-Vektor das lacZ-Gen liegt, welches für die β-Galactosidase kodiert. Die β-Galactosidase ist in der Lage, Lactose zu Glucose und Galactose sowie den Farbstoff Xgal (5-Brom-4-Chlor-3-Indolyl-β-D-Galactopyranosid, Roth) zu hydrolisieren. Das hydrolisierte Xgal wird dabei zu einem schwer löslichen und blauen Indigofarbstoff umgesetzt. Bei dem T-Easy-Vektor liegt die MCS innerhalb des β-Galactosidase-Gens. Wird das untersuchte Gen somit korrekt in den Vektor integriert, wird das β-Galaktosidase-Gen inaktiviert, die Kolonien verbleiben weiß. Ist der T-Easy-Vektor ohne das zu untersuchende Gen religiert, ist die β-Galactosidase aktiv, die Kolonien erscheinen blau. Zusätzlich wird IPTG (Isopropyl-ß-D-thiogalactopyranosid, Roth) zur Verfügung gestellt, das den aktiven Repressor des lacZ-Gens (bzw. ursprünglich des gesamten lacZ-Operons) inaktiviert.

Oligonucleotide

Alle in dieser Arbeit verwendeten Oligonucleotide wurden von der Firma MWG-Biotech (Ebersberg, Deutschland) bezogen und sind in Tabelle 4 aufgeführt. Der Entwurf der Sequenzen erfolgte mithilfe der Software „Clone Manager Professional Suite“ (Version 8, Scientific & Educational Software).

Tabelle 4: Oligonukleotid /Primer-Sequenzen für die Konstrukte von DCBD2 und RGMA. C-Dom: C-Domäne. N-Dom: N-Domäne.

Bezeichnung	Oligonukleotid/Primer-Sequenz	Orientierung	Restriktionsschnittstelle
5´NheI_C_Dom_DCBD_ SEAP	5′-AATTGCTAGCATGTCCTCCTTCTCCATGCCTCTGTTCCTCCTG CTCTTACTTGTCCTGCTCCTGCTGCTCGAGGACGCTGGAGCCC AGCAAGGTGAATTCATCATCCCAGTTGAGG-3′	vorwärts	NheI
N_Dom_1_C_Dom_DCBD	5′-CAAGTCCGGGCCGCGGCCGCCGCCCCCGCCTGGGCCGCGCTCC CCCTCTCCCGCTCCCTCCCTCCCTGCTCCAACTCCTCCTCCTT CTCCATGCCTCTGTTCC-3′	vorwärts	------
5´-NheI_N_Dom_2_DCBD	5′-AATTGCTAGCATGGCGAGCCGGGCGGTGGTGAGAGCCAGGCGC TGCCCGCAGTGTCCCCAAGTCCGGGCCGCGGCCGCC -3′	vorwärts	NheI
N_Dom_1_DCBD_ SEAP	5′-CAAGTCCGGGCCGCGGCCGCCGCCCCCGCCTGGGCCGCGCTCC CCCTCTCCCGCTCCCTCCCTCCCTGCTCCAACTCCGAATTCAT CATCCCAGTTGAGG-3′	vorwärts	------
3´-SEAP Myc	5′-TTGGTACCTTACAGATCCTCTTCTGAGATGAGTTTTTGTTCAC CCGGGTGCGCGGCGTCG-3′	rückwärts	ACC65I
5´NotI_DOM_RGMA_ SEAP	5′-TTGCGGCCGCATGGCCCTGGGATTCTGGCCGACCCTGCGCTTC CTTCTCTGCAGCTTCCCCGCAGCCACCTCCCCGTGCAAGATCC TCGAATTCATCATCCCAGTTGAG-3′	vorwärts	NotI
5´NotI_N_DOM_RGMA_ SEAP	5′-TTGCGGCCGCATGCAGCCGCCAAGGGAGAGGCTAGTGGTAACA GGCCGAGCTGGATGGATGGGTATGGGGAGAGGGGCAGGACGTT CAGAATTCATCATCCCAGTTGAGG-3′	vorwärts	NotI

Anm.: Die Primer der Konstrukte von shrew-1 sind im Anhang A2 aufgeführt.

Restriktionsenzyme

Für die Klonierung sowie den analytischen Verdau der DCBD2 und RGMA-Konstrukte wurden folgende Enzyme verwendet:

- NheI (Fermentas)
- NotI (Fermentas)
- ACC65I (Fermentas)

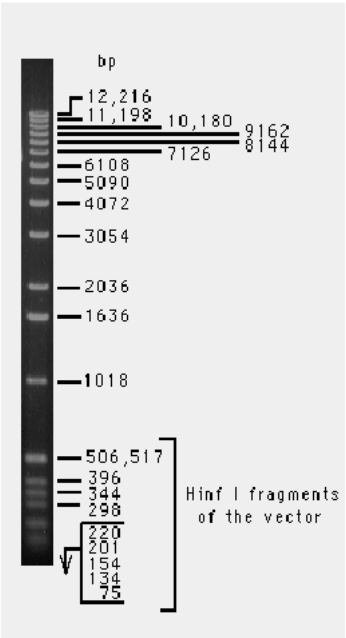
Für den Verdau wurde jeweils der mitgelieferte „yellow"- (NheI/ACC65I) oder „orange"-(NotI/ACC65I) Puffer verwendet. Für die PCR wurden die folgenden Enzyme verwendet:

- *pfx*-DNA Polymerase (Invitrogen)
- *pfu*-DNA Polymerase (Promega)

Für die Transformation kompetenter Bakterien wurden die folgenden Enzyme verwendet:

- Lysozym (Roth, Muramidase)
- T4-DNA Ligase (Fermentas, Ligase)
- T4-DNA Ligase für T-Easy Vektor (Promega)
- CIAP (Fermentas, Phosphatase)

Alle im Rahmen dieser Arbeit hergestellten Agarosegele zur Auftrennung von DNA-Fragmenten wurden mit einer 1Kb-DNA-Leiter (Abb. 5; Invitrogen) zum Größenvergleich versehen.



Verwendete 1Kb-DNA-Leiter Zusammensetzung:

5 µl Marker
50 µl DNA/Ladepuffer
195 µl H_20

Abbildung 5: 1Kb-DNA—Leiter. Angaben des Herstellers. Größenangaben in Basenpaaren. Aufgetragen: 0,5 µg DNA-Größenstandard in 0,9% Agarosegel. Färbung mit Ethidiumbromid.

Primäre und sekundäre Antikörper

Für die Immunfluoreszenz sowie die Detektion in Western-Blots wurden die in Tabelle 5 aufgeführten primären und die in Tabelle 6 aufgeführten sekundären Antikörper verwendet.

Tabelle 5: Primäre Antikörper für Western-Blot und Immunfluoreszenz. WB: Western-Blot. IF: Immunfluoreszenz.

Bezeichnung	monoklonal /polyklonal	Verdünnung	Spezies	verwendet in
Anti-GAPDH	monoklonal	1:10.000	Maus	WB
Anti-Cytochrom-C	monoklonal	1:500	Maus	WB
Anti-grp94	monoklonal	1:500	Ratte	WB
Anti-Myc	polyklonal	1:1.000(WB), 1:100(IF)	Kaninchen	WB und IF

Tabelle 6: Sekundäre Antikörper für Western-Blot und Immunfluoreszenz. WB: Western-Blot. IF: Immunfluoreszenz

Bezeichnung	monoklonal / polyklonal	Verdünnung	Spezies	verwendet in
Anti-GAPDH *wird erkannt durch:* Anti-Maus HRP-konjugiert	monoklonal	1:30.000	Maus	WB
Anti-Cytochrom C *wird erkannt durch:* Anti-Maus HRP-konjugiert	polyklonal	1:15.000	Ziege	WB
Anti-grp94 *wird erkannt durch:* Anti-Ratte HRP-konjugiert	monoklonal	1:15.000	Ziege	WB
Anti-Myc *wird erkannt durch:* Anti-Kaninchen HRP-konjugiert	monoklonal	1:15.000(WB), 1:200(IF)	Ziege	WB und IF
Jeweils konjugiert mit Alexa Fluor® 594 goat anti-mouse IgG (H+L)	monoklonal	1:200	Ziege	IF

Experimental-Kits

Im Rahmen der PCR, der Transformation kompetenter Bakterien und der Transfektion von HEK- bzw. MCF7-Zellen kamen die folgenden Kits zum Einsatz:

- Rapid Gel Extraction System (Marligen Biosciences)
- NucleoSpin Extract II (Machery-Nagel)
- Rapid Plasmid Miniprep System (Marligen Biosciences)
- High Purity Plasmid Midiprep Kit (Marligen Biosciences)

Immunfluoreszenz

Alle Immunfluoreszenz-Bilder wurden mithilfe des konfokalen Laser-Scanning-Mikroskops (CLSM, Leica TCS SP5) unter Anleitung von Dr. Alexander Schreiner (JWG-Universität, Frankfurt/Main) erstellt. Die Bearbeitung der Bilder erfolgte mithilfe der Software „Imaris“ (Version 5.0.3) von Bitplane.

- Die Zellkernfärbung erfolgte mit DAPI-Stammlösung (Hoechst 33258) 0,5 mg/ml.
- Die Färbung der Mitochondrien erfolgte mit Mitotracker (Invitrogen)
    - 0,1 mM Mitotracker-Stammlösung auf 1 ml,
    - 0,05 mg Mitotracker Red CMXRos in 100 µl DMSO lösen,
    - Zugabe von 900 µl DMEM.
- Alternativ wurden Mitochondrien mit mito-GFP (GFP mit Cytochrom-C-Signal) gefärbt.

SEAP-Assay

Der *Secreted Alkaline Phosphatase*-Assay (SEAP-Assay) dient als indirekter Nachweis für einen Import des untersuchten Konstrukts in das Lumen des ER.

Für eine Aktivität der SEAP ist eine N-Glykosylierung im Lumen des ER notwendig. Ein Transport der SEAP ohne ihr natives Signalpeptid zum ER erfolgt nicht. Im Rahmen des Tests wird das native SEAP-Signalpeptid durch die zu untersuchende Sequenz ersetzt. Ist diese Sequenz in der Lage, die SEAP zum ER zu transportieren, wird dies durch eine Aktivierung der SEAP angezeigt.

Die Aktivierung der SEAP durch Glykosylierung im ER-Lumen wird dabei indirekt durch die Fähigkeit festgestellt, *para*-Nitrophenylphosphat zu *para*-Nitrophenolat und Phosphat umzusetzen (Berger *et al*., 1988; Abb. 6).

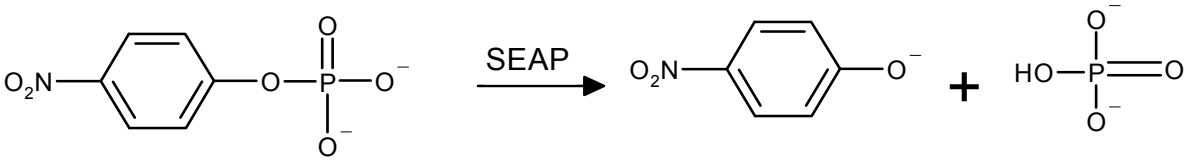


Abbildung 6: SEAP-Substratumsetzung. *Para*-Nitrophenylphosphat wird durch die SEAP zu *para*-Nitrophenolat und Phosphat durch die SEAP umgesetzt.

Die Substratumsetzung ruft eine messbare Veränderung des Absorptionsspektrums bei 405nm hervor. Die Geschwindigkeit der Substratumsetzung ist dabei mit der Menge an aktiver SEAP korreliert. Dadurch ist sowohl ein qualitativer als auch ein quantitativer Rückschluss auf die Fähigkeit der zu untersuchenden Sequenz, die SEAP zum ER zu dirigieren, möglich.

Bei zwei durchgeführten Messungen wurde der Mittelwert ohne Standardabweichung angegeben. Bei mehr als zwei durchgeführten Messungen wurde der Standardfehler (Quotient aus Standardabweichung und Wurzel des Stichprobenumfanges) angegeben.

Die Angaben zur Durchführung (Kapitel „Messung der SEAP-Aktivität im Zell- und Proteinextrakt“) sind an Berger *et al.,* 1988 und Dipl.-Biologe Eduard Resch (JWG-Universität, Frankfurt/Main, persönliche Kommunikation) angelehnt und wurden vom Autor den Gegebenheiten des DCBD2- und RGMA-Kontextes angepasst.

Messung der SEAP-Aktivität im Zell- und Proteinextrakt

Alle Absorptionsspektren wurden an einem Spektrophotometer des Typs Ultrospec 2100pro der Firma Amersham Biosciences gemessen.

Die Messung wurde mit einem Volumen von 6 µg Proteinextrakt durchgeführt.

Vorgehen:

1) Proteinextrakt wird in 1,5 µl Reaktionsgefäß in 50 µl ddH_2O verdünnt
2) Inkubation 10 min bei 65°C
3) Überführung in Messküvette und Vermischung mit 100 µl SEAP-Puffer
4) Inkubation 10 min bei 37°C

5) Zugabe von 20 µl SEAP-Substratlösung, Temperatur der Lösung 37°C, durchmischen
6) Inkubation auf 37°C für 30 min
7) Abnahme der SEAP-Substratlösung jeweils nach 5 min und nach 30 min
8) Messung der Absorptionsveränderung bei 405 nm

Als Referenz für das Spektrophotometer dient ein entsprechend der Beschreibung (*vide supra* Punkte 1-8) behandelter Ansatz aus Zellen, die mit Leervektor transfiziert wurden. Die Messung und Kalibrierung des Gerätes auf den Referenzwert erfolgt unmittelbar nach Zugabe der 20 µl Substratlösung.

Zusammensetzung der verwendeten Komponenten:

SEAP-Substratlösung:
120 mM *p*-Nitrophenphosphat in SEAP-Puffer gelöst.

SEAP-Puffer

1 M	Diethanolamin
10 mM	L-Homoargininhydrochlorid
0,5 mM	$MgCl_2$

SEAP-Aktivität im Überstand

Der zu messende Überstand wurde wie folgt erhalten:

Die transfizierten Zellen wurden in *six-well*-Schalen 48 h kultiviert und 1 ml des Überstandes wurde entnommen.

Weiteres Vorgehen:

1) Zentrifugation, 2 min, 13000 U/min
2) Entnahme und Überführung der oberen 800 µl des Überstandes in 1,5 µl Reaktionsgefäß
3) Entnahme von 50 µl für den Assay, 750 µl bei -80°C gelagert
4) Inkubation, 10 min, 65°C
5) Zentrifugation, 2 min, 13000 U/min
6) Überführung von 10 µl des Überstandes (oberer Teil) in Messküvette

7) Vermischung mit 90 µl ddH_2O und 100 µl SEAP-Puffer
8) Inkubation, 10 min, 37°C
9) Zugabe von 20 µl SEAP-Substratlösung, mischen
10) Inkubation, 30 min, 37°C
11) Abnahme der SEAP-Substratlösung jeweils nach 5 min und nach 30 min
12) Messung der Absorptionsveränderung bei 405 nm

Als Referenz für das Spektrophotometer diente ein entsprechend den oben beschriebenen Punkten 1-8 behandelter Ansatz aus Zellen, die mit Leervektor transfiziert wurden. Die Messung und Kalibrierung des Gerätes auf den Referenzwert erfolgte unmittelbar nach Zugabe der 20 µl Substratlösung.

PNGase-F Verdau

Durch Zugabe von PNGase-F werden N-Glykosylierungen entfernt (Lottspeich und Zorbas, 1998; New England Biolabs). Ein Protein, dass im ER glykosyliert wurde, zeigt im Western-Blot eine Größenveränderung in Abhängigkeit der Zugabe von PNGase-F. Ein Protein, das nicht glykosyliert wurde, ist unempfindlich gegenüber einer Zugabe von PNGase-F.

Ergebnisse und Diskussion

Lange Signalpeptiden (≥40 Aminosäuren) wurden hinsichtlich ihres Aufbaus, ihrer potentiellen Post-Targeting-Funktionen und Diskriminierung zu kurzen (<40 Aminosäuren) Signalpeptiden untersucht.

Zuerst wurde die Relevanz des Aspektes „Länge“ bei Signalsequenzen durch eine statistische Analyse der Längenverteilung verschiedener ER-Targeting-Signale (lange und kurze Signalpeptide, mitochondriale Targeting-Peptide, Chloroplasten-Targeting-Peptide, virale Signalpeptide) untersucht (Kapitel „Statistiken zu Signalpeptiden“). Darauf aufbauend wurde ein Modell für die Domänen-Architektur von langen Signalpeptiden (NtraC) entwickelt und zur Vorhersage *in silico* angewendet (Kapitel „NtraC-Modell *in silico“*). Das entwickelte NtraC-Modell wurde im Anschluss am Beispiel von drei Proteinen (shrew-1, DCBD2, RGMA) *in vitro* überprüft (Kapitel „NtraC-Modell *in vitro*“). Alle drei Proteine zeigten *in vitro* eine Verhaltensweise entsprechend der *in silico* gemachten Vorhersagen (Hiss *et al*., 2008b).

Die Begriffe „lange Signalpeptide“ und „kurze Signalpeptide“ werden hier wie folgt definiert und im Weiteren verwendet:

- Signalpeptide mit ≥40 Aminosäuren → „lange Signalpeptide“
- Signalpeptide mit <40 Aminosäuren → „kurze Signalpeptide“

Die Definition ist wie folgt begründet: Signalpeptide weisen eine durchschnittliche Länge von 22 Aminosäuren auf (von Heijne, 1985). Für eine potentielle zweite Targeting-Funktion innerhalb eines Signalpeptides, die dem gleichen Aufbau eines kurzen Signalpeptides folgt, wurde ebenfalls die Länge von 22 Aminosäuren berücksichtigt. Die Grenze für lange Signalpeptide wurde somit auf die doppelte Länge durchschnittlicher Signalpeptide und unter Berücksichtigung der Signalpeptidase-Schnittstelle (Abzug von vier Aminosäuren) auf 40 Aminosäuren festgelegt.

Statistiken zu Signalpeptiden

Signalpeptide sind unabhängig von ihrer Länge in der Lage, Proteine zum ER zu dirigieren (von Heijne, 1985; Hiss *et al.,* 2008b). Eine Unterscheidung aufgrund des Zielkompartimentes von langen und kurzen Signalpeptiden ist daher nicht möglich und in diesem Kontext auch nicht sinnvoll. Eine Diskriminierung langer und kurzer Signalpeptide kann aber *in silico* durch statistische Analysen der Verteilung der Aminosäuren innerhalb der Signalpeptide und der

Länge der Signalpeptide erfolgen. Des Weiteren wird untersucht, ob ein Zusammenhang zwischen der Länge der Targeting-Signale und dem von ihnen kodierten Kompartiment besteht. Zur Diskriminierung zwischen langen und kurzen Signalpeptiden wurden daher die in den Kapiteln „Längenverteilung von Signalpeptiden“ und „Aminosäure-Häufigkeiten in Signalpeptiden“ beschriebenen Analysen durchgeführt.

Längenverteilung von Signalpeptiden

In Kapitel „Längenverteilung von Signalpeptiden“ wird untersucht, ob sich Targeting-Signale für unterschiedliche Kompartimente durch ihre Länge (gemessen in Anzahl der Aminosäuren) unterscheiden. Hierzu wurden alle eukaryotischen Signalpeptide „nicht-putativer“ Proteine aus der UniProtKB (Version 13.6) extrahiert und nach ihrer Länge unterteilt (Abb. 7).

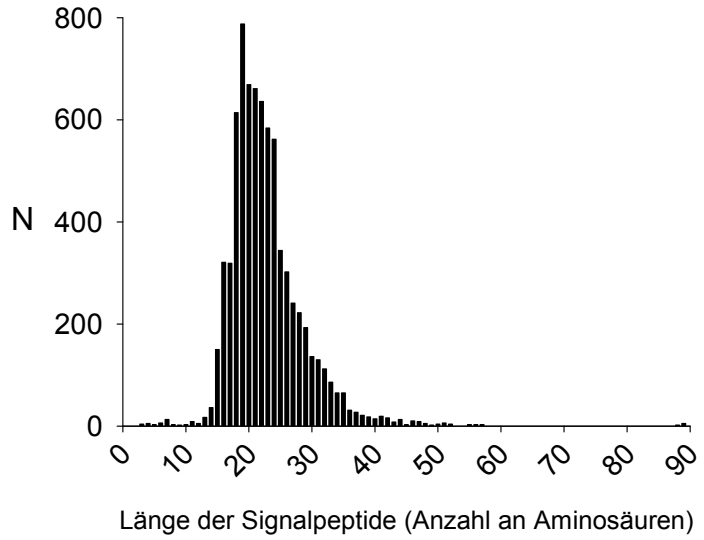


Abbildung 7: Längenverteilung **eukaryotischer** Signalpeptide (*N* = 7.539). Nur nicht-putative Proteindaten wurden verwendet. Sequenzen aus UniProtKB (Version 13.6).

Die mittlere Länge von eukaryotischen Signalpeptiden nach Abb. 7 beträgt 23 Aminosäuren mit einer Standardabweichung von ±6. Dies ist vergleichbar mit den Beobachtungen in von Heijne (1985), der eine mittlere Länge von 22 Aminosäuren beschreibt. Diese mittlere Länge wurde auch für das Training von SignalP 3.0 verwendet (Bendtsen *et al*., 2004a). In der SRS-Suchanfrage im Rahmen dieser Arbeit sind in UniProtKB (Version 13.6) 136 Sequenzen aus nicht-putativen Proteinen mit einem Signalpeptid mit ≥40 Resten gefunden worden (Anhang Tabelle A3). Diese 136 Signalpeptide (2% aller eukaryotischen Signalpeptide) bilden die Gruppe der langen Signalpeptide. Ihre Längenverteilung wird gesondert betrachtet (Abb. 8).

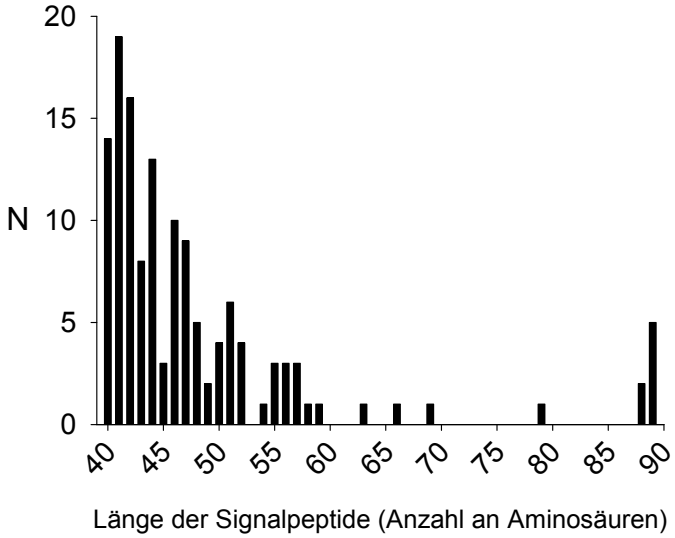


Abbildung 8: Längenverteilung **eukaryotischer** Signalpeptide mit ≥ 40 Resten (*N* = 136) nicht-putativer Proteine. Sequenzen aus UniProtKB (Version 13.6).

Das längste in UniProtKB (Version 14.0) annotierte, nicht-potentielle eukaryotische Signalpeptid ist das Crumbs-Protein aus *Drosophila melanogaster* mit 88 Aminosäuren (Tepass *et al*., 1990). Nach Aussage der UniProtKB besitzen die als eukaryotisch annotierten Proteine ENK1, ENK2, ENK3,ENK4, ENK5 und ENK6 Signalpeptide ≥ 88 Aminosäuren. Hierbei handelt es sich aber um virale Proteine, die im Laufe der Evolution ins Wirtsgenom integriert wurden. Diese haben Signalpeptide mit 88 (ENK1) bzw. 89 (ENK2 bis ENK6) Aminosäuren Länge (Loewer *et al.,* 1995; Toenjes *et al*., 1999; Barbulescu *et al*., 1999; Turner *et al*., 2001; de Parseval *et al.*, 2003). Diese sechs Proteine haben mit Ausnahme von ENK1 (88 Aminosäuren Länge) ein zu 100% konserviertes langes Signalpeptid mit 89 Aminosäuren (Abb. 8). Obwohl diese Proteine als „eukaryotisch“ annotiert sind, werden sie hier aufgrund ihres Ursprungs als retrovirale *beta type-B* Hüll-Proteine, nicht als eukaryotische SP betrachtet. Die Proteine haben ihren ursprünglichen fusogenen Charakter, also die Vermittlung der Fusion zwischen Wirtsmembran und Virus, verloren. Die Signalpeptide sind jeweils als „potentiell“ annotiert. Die Proteine ENK1-6 sind als „Membranprotein“ annotiert. Die Annotation erfolgte durch Ähnlichkeit zu anderen Proteinen (*by similarity)*.

Das Auftreten von 136 in der Datenbank annotierten eukaryotischen Signalpeptiden ≥ 40 Aminosäuren wird als Hinweis gewertet, dass lange Signalpeptide kein Einzelphänomen sind. Im Rahmen der automatisierten Datenbank-Suche können nur lange Signalpeptide erfasst werden, die in der Datenbank auch als solche annotiert sind. Vorhergesagte lange Signalpeptide werden aber in bestimmten Fällen als Artefakt betrachtet bzw., wenn eine

experimentelle Validierung stattfindet, wird diese nicht in die Datenbank übernommen. Ein Beispiel hierfür ist shrew-1, welches ohne Signalpeptid in UniProtKB (Version 14.0) annotiert ist, aber ein experimentell bestätigtes Signalpeptid von 43 Aminosäuren Länge besitzt (Resch *et al*., 2008). Die manuelle Erweiterung des Datensatzes um vergleichbare Sequenzen wird daher fortlaufend durchgeführt.

Bei der Betrachtung langer eukaryotischer Signalpeptide fällt die Präsenz ins Genom integrierter ehemaliger viraler Signalpeptide (Enk1-6, Enk 17) auf. Sechs weitere Proteine (FCG2B, C163A, ITA5, MCP, Sema7, Tyro 3) sind Zelloberflächenproteine, die für die Erkennung der Wirtszelle durch den Virus bzw. dessen Lebenszyklus von Bedeutung sind (Anhang Tabelle A4). Das Auftreten langer Signalpeptide bei viralen Proteinen ist in der Literatur bekannt (Henderson *et al*., 1983; Lindemann *et al*., 2001; Liu und Miller, 2005).

In Abbildung 9 ist die Verteilung viraler Signalpeptide dargestellt (*N* = 213). Diese Gruppe beinhaltet nicht ehemalige ins Genom integrierte virale Proteine, sondern nur als viral annotierte Proteine. Virale Signalpeptide zeigen mit einer mittleren Länge von 15 (±16) Aminosäuren (Abb. 9) eine um 7 Aminosäuren kürzere durchschnittliche Länge im Vergleich zu eukaryotischen Signalpeptiden (23 Aminosäuren, Abb. 7). Die höhere Standardabweichung (±16) im Vergleich zu eukaryotischen Signalpeptiden (±6) ist auf die im Verhältnis höhere Anzahl an Signalpeptiden mit ≥40 Aminosäuren zurückzuführen. 11 virale Signalpeptide haben mehr als 40 Aminosäuren. Dies entspricht 5%. Bei eukaryotischen Signalpeptiden haben 2% eine Länge von ≥40 Aminosäuren. Die gefundenen 11 viralen Sequenzen ≥40 sind im Anhang, Tabelle A4, aufgeführt.

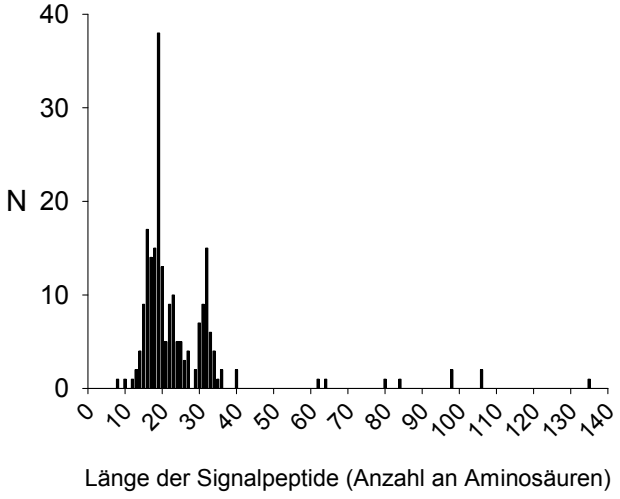


Abbildung 9: Längenverteilung **viraler** Signalpeptide (*N* = 213). Nur nicht-putative Proteindaten wurden verwendet. Sequenzen aus UniProtKB (Version 14.0) entnommen.

Die Längenverteilung viraler Signalpeptide ähnelt der Längenverteilung eukaryotischer Signalpeptide (Mittelwert, Schwerpunkt der Verteilung). Das im Vergleich zu eukaryotischen Signalpeptiden vermehrte Auftreten von langen Signalpeptiden bei viralen Proteinen kann an die Existenz zusätzlicher Funktionen geknüpft sein (Kapitel „Post-Targeting-Funktionen von Signalpeptiden“). Die experimentelle Bestätigung dieser zusätzlichen Funktionen bei langen viralen Signalpeptiden hat eine bessere Annotation in den Datenbanken zur Folge. Vorhergesagte lange Signalpeptide bei Viren werden nicht direkt als Artefakt betrachtet.

Virale Signalpeptide stellen somit eine eigene Gruppe dar, deren Längenverteilung sich an der eukaryotischer Signalpeptide orientiert, aber funktionsabhängige Abweichungen der Länge toleriert.

Bei mitochondrialen Targeting-Peptiden und Chloroplasten-Transit-Peptiden ist ebenfalls ein gehäuftes Auftreten langer Sequenzen zu beobachten (Abb. 10 und 11). Die Signalsequenzen von Mitochondrien und Chloroplasten unterscheiden sich vor allem in ihrer Schnittstelle (Schneider *et al*., 1993).

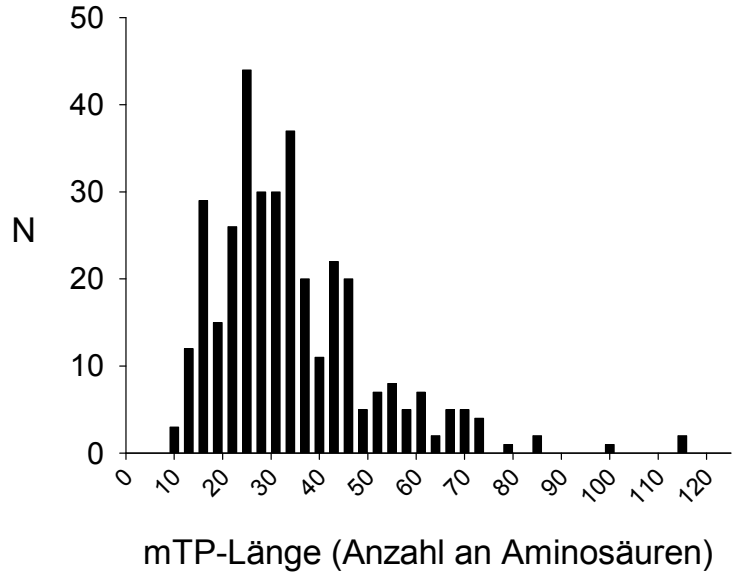


Abbildung 10: Längenverteilung **mitochondrialer** Targeting-Peptide (mTP, *N* = 1.101). Sequenzen aus UniProtKB (Version 13.6) entnommen; nur nicht-putative Proteine wurden berücksichtigt. Vereinfachte Darstellung mit jedem 3. Balken. Vollständige Darstellung im Anhang A5.

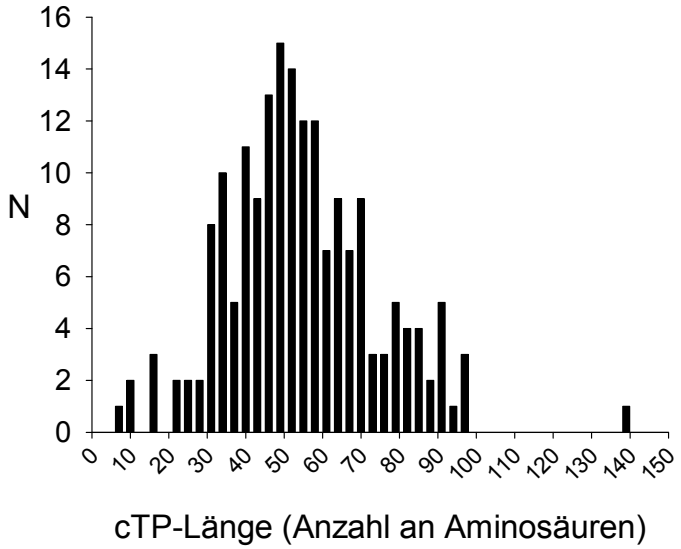


Abbildung 11: Längenverteilung von **Chloroplasten-Transit-Peptiden** (cTP, *N* = 514). Sequenzen aus UniProtKB (Version 13.6) entnommen; nur nicht-putative Proteine wurden berücksichtigt. Vereinfachte Darstellung mit jedem 3. Balken. Vollständige Darstellung im Anhang A6.

Bei mTP beträgt die mittlere Länge 36 Aminosäuren mit einer Standardabweichung von ± 18 (Abb. 10). Von 1.101 betrachteten mTP-Sequenzen haben 343 (31%) eine Länge von mehr als 40 Aminosäuren. Bei cTP beträgt die mittlere Länge 56 Aminosäuren mit einer Standardabweichung von ± 19. Von 514 untersuchten cTP-Sequenzen haben 428 (81%) eine Länge von mehr als 40 Aminosäuren (Abb. 11).

In den Gruppen cTP und mTP treten lange Signalpeptide somit im Vergleich zu SP um das 41-fache (cTP) bzw. 16-fache (mTP) häufiger auf. Dies könnte darauf zurückzuführen sein, dass Proteine, die für die Mitochondrien oder die Chloroplasten bestimmt sind, mehrere Membranen überwinden müssen. Im Falle von mitochondrialen Proteinen sind dies die innere und äußere Mitochondrien-Membran, im Falle von Chloroplasten höherer Pflanzen sowie Rot- und Grünalgen die innere und äußere Chloroplasten-Membran sowie gegebenenfalls noch die Thylakoid-Membran (Cavalier-Smith, 1999). Im Falle von Braunalgen und Euglenophyta können die Chloroplasten von 3-4 äußeren Membranen umschlossen sein (Cavalier-Smith, 1999). Mitochondriale Targeting-Signale unterliegen entsprechend dem von ihnen kodierten mitochondrialen Subkompartiment (Matrix, innere und äußere Membran, Intermembranraum) einer multiplen Prozessierung (Schneider *et al*., 1998). Diese zusätzliche Prozessierung erfolgt durch die *mitochondrial processing peptidase* (MPP) in der Matrix der Mitochondrien (Gakh *et al*., 2002). Eine Subgruppe von Signalen wird im Anschluss durch *mitochondrial intermediate peptidase* (MIP) prozessiert (Isaya und Kalousek, 1994). Für die

statistische Auswertung der Targeting-Signale wurde keine Unterscheidung entsprechend der Subkompartimente von Chloroplasten und Mitochondrien vorgenommen. Die Targeting-Signale von Mitochondrien und Chloroplasten wurden unabhängig von dem von ihnen kodierten Subkompartiment als Mitochondrien- bzw. Chloroplasten-Signal in die Untersuchung aufgenommen. Die hier beobachtete mittlere Länge von 56 Aminosäuren für Chloroplasten-Transit-Peptide stimmt mit dem von Emanuelsson *et al*., 2000 ermittelten Wert von 56 Aminosäuren überein.

Zusammenfassend wird beobachtet, dass Signalpeptide bei gleichem Zielkompartiment variable Längen aufweisen können (Abb. 7, 10 und 11). Targeting-Signale für unterschiedliche Kompartimente weisen jedoch eine unterschiedliche durchschnittliche Länge auf (SP: 23, mTP: 36, cTP: 56). Die unterschiedliche mittlere Länge ist als Hinweis zu werten, dass die Länge von Signalpeptiden mit den Zielkompartimenten und der Anzahl der das Kompartiment umgebenden Membranen verknüpft ist. Die hohen Varianzen der Längenverteilung von Targeting-Signalen könnte auf die Toleranz von Post-Targeting-Funktionen hindeuten.

Aminosäure-Häufigkeiten in Signalpeptiden

Als weitere Unterscheidungsmöglichkeit für Targeting-Signale wurde ihre Aminosäure-Zusammensetzung untersucht. Bekannt ist, dass Proteine entsprechend ihres Zielkompartimentes (z.B. integrale Membranproteine, membran-verankerte Proteine, extrazelluläre Proteine, intrazelluläre Proteine und kern-lokalisierte Proteine) eine unterschiedliche Zusammensetzung an Aminosäuren aufweisen (Cedano *et al*., 1997; Chou, 2001). Von Nakei und Horton (1999) wurde der Unterschied in der Aminosäure-Zusammensetzung der ersten 20 N-terminalen Reste zur Erkennung von mitochondrialen Targeting-Peptiden (mTP) herangezogen. Dabei ist Arginin nach Schneider und Broger (1999) die Aminosäure, die am besten geeignet ist um zwischen mTP und dem maturen mitochondrialen Protein zu unterscheiden. Diese und andere Eigenschaften werden bereits von *in silico*-Vorhersage-Methoden zum Erkennen von Protein-Targeting-Signalen verwendet (Schneider und Fechner, 2004). Motiviert durch diese Beobachtungen wurde eine Untersuchung der Aminosäure-Häufigkeit in den annotierten Targeting-Signalen der

aktuellen UniProtKB (Version 14.0) vorgenommen. In Tabelle 7 ist die Verteilung der 20 genetisch kodierten Aminosäuren innerhalb der verschiedenen Targeting-Signale (SP, mTP, cTP, vSP) aufgeführt.

In Tabelle 7 stellt jede Zeile die Verteilung der 20 genetisch kodierten Aminosäuren in den Targeting-Signalen für das jeweilige Kompartiment dar, gefolgt von einer Zeile mit der zugehörigen Standardabweichung. Bei den Signalpeptiden wurde zusätzlich zwischen langen und kurzen Signalpeptiden und zwischen Gruppen nach dem NtraC-Modell organisierter putativer Signalpeptide unterschieden. Eine speziesspezifische Aminosäure-Hintergrundverteilung kann hier als Bezugspunkt nicht berücksichtigt werden, da die verwendeten Sequenzen aus unterschiedlichen Spezies stammen.

Zuerst wurde ein Vergleich der gesamten Verteilung der Aminosäure-Häufigkeiten innerhalb eines Targeting-Signals (z.B. mTP) mit der gesamten Verteilung eines anderen Targeting-Signals (z.B. SP) durchgeführt. Hierzu wurde ein zweiseitiger Kolmogorov-Smirnov-Test (KS-Test, Signifikanz-Niveau 5%; Rassokhin und Agrafiotis, 2000) für jede Kombination der Aminosäure-Häufigkeitsverteilungen zweier ER-Targeting-Signale durchgeführt.
Ein signifikanter Unterschied der gesamten Verteilungen zwischen beliebigen Paaren aus Aminosäure-Häufigkeitsverteilungen zweier ER-Targeting-Signale aus Tabelle 7 ist nicht zu erkennen.
Exemplarisch bedeutet das: Wird die Verteilung der Aminosäure-Häufigkeit als Unterscheidungskriterium herangezogen, ist statistisch kein Unterschied zwischen z.B. Signalpeptiden und mitochondrialen Targeting-Peptiden zu erkennen.
Fokussiert man auf einzelne Reste, wird jedoch ein Unterschied zwischen SP und mTP bei Arginin erkennbar. Arginine sind mit ihrer positiven Ladung für die Erkennung duch Tom 20 und 22 (Mokranjac und Neupert, 2007) für das mitochondriale Targeting von Bedeutung und in mTP mit 13% (±5) im Vergleich zu kurzen SP und viralen SP mit jeweils 3% (±4) überrepräsentiert (Tabelle 7). Es wurde für mTP gezeigt, dass geladene Reste einen Einfluss auf die Funktionalität der mTP-Schnittstelle haben (Schneider *et al*., 1995). Bei langen SP nehmen Arginine im Vergleich zu kurzen SP von 3% (±4) auf 7% (±4) zu. Betrachtet man die NtraC-organisierten langen Signalpeptide (SP ≥40 NtraC), bei denen wir gezeigt haben, dass sie mTP-Funktionen enthalten können (Hiss *et al.,* 2008), steigt der Anteil an Argininresten auf 8% (±5).

Noch deutlicher wird die Zunahme von Arginin bei der Gruppe der langen Signalpeptide für die nach dem NtraC-Modell eine mTP-Funktion vorhergesagt wurde. Diese zeigen mit einer Auftrittshäufigkeit von 10% (±3) eine zu mTP (13% ±5) vergleichbare und zu kurzen SP (3% ±4) klar unterscheidbare Auftrittshäufigkeit (Tabelle7, Rahmen). Dies deutet auf eine zusätzliche Funktion der NtraC-organisierten Signalpeptide hin, die an Arginine gekoppelt ist, z.B. ein mitochondriales Targeting.

Es ist weiterhin zu beobachten, dass Alanin und Leucin in allen Targeting-Signalen bezogen auf eine Gleichverteilung (5%) überrepräsentiert ist, während die geladenen Reste Histidin, Asparagin, Glutamin, Aspartat und Glutamat in allen Targeting-Signalen unterrepräsentiert sind. Gleiches glit für die aromatischen Aminosäuren Tyrosin und Tryptophan. Dies könnte im Zusammenhang mit der *h*-Region (von Heijne, 1985) stehen. Hier werden vermehrt hydrophobe Reste erwartet und nur einzelne geladene Aminosäuren toleriert. Des Weiteren argumentiert dies generell für einen Zusammenhang zwischen dem Aspekt Ladung und dem Targeting von Proteinen.

Die beobachtete Unterrepräsentation von geladenen und aromatischen Aminosäuren könnte zur Vorhersage herangezogen werden, ob es sich bei einer Sequenz um eine Targeting-Sequenz handelt oder nicht. Es wurde von Schneider *et al*. (1997) gezeigt, dass sich mitochondriale und Chloroplasten-Signale mithilfe künstlicher neuronaler Netze trennen lassen. Von Schneider *et al*. (1997) wurden jedoch die Aminosäuren nicht im 1-Buchstabencode dargestellt, sondern durch eine Kodierung der Aminosäure gemäß ihrer Volumen-, Refraktivitäts- und Hydrophobizitätswerte sowie der Berücksichtigung von Mustern innerhalb der Sequenz. Hierbei wurden auftretende Muster z.B. geladener Reste innerhalb von mTP-Sequenzen im Unterschied zu cTP-Sequenzen berücksichtigt. Das in Schneider *et al*. 1997 beschriebene komplette Fehlen von geladenen Resten in cTP kann hier nicht bestätigt werden (Tabelle 7). Der Stichprobenumfang für cTP bei Schneider *et al*. (1997) ist N =43, in unserer Untersuchung N = 514. Dies und sowie den Beschreibung von bekannten dualen Targeting-Signalen, die sowohl von der Mitochondrien- als auch der Chloroplasten-Importmaschinerie erkannt werden (Silva-Filho, 2003) sind weiter Argumente für die Existenz geladener Reste in cTPs.

Tabelle 7: Auftrittshäufigkeit von Aminosäuren in Signalpeptiden (Angaben in Prozent). **SP <40**: Eukaryotische Signalpeptide mit <40 Aminosäuren (*N* = 7.539). **SP ≥ 40**: Eukaryotische Signalpeptide mit ≥ 40 Aminosäure (*N* = 145). **SP≥40 NtraC**: Eukaryotische Signalpeptide mit ≥ 40 Aminosäure, die eine NtraC-Organisation aufweisen (*N* = 185). **SP≥40 NtraC(mTP):** Eukaryotische Signalpeptide mit ≥ 40 Aminosäure, die eine NtraC-Organisation aufweisen und deren N-Domäne als mTP vorhergesagt wurde (*N* = 32). **mTP**: Mitochondriale Targeting-Peptide (*N* = 1.101). **cTP**: Chloroplasten-Transit-Peptide (*N* = 514). **vSP**: Virale Signalpeptide (*N* = 213). Sequenzen aus UniProtKB (Version 14.0) entnommen. Putative Proteine wurden jeweils von der Betrachtung ausgeschlossen.

Aminosäuren	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
SP < 40	13 ±9	3 ±4	1 ±2	1 ±2	5 ±6	6 ±5	1 ±2	5 ±5	2 ±3	23 ±9	6 ±3	1 ±2	4 ±5	2 ±3	3 ±4	8 ±6	5 ±5	8 ±6	2 ±3	1 ±2
SP ≥40	12 ±6	3 ±3	1 ±2	2 ±2	3 ±3	8 ±6	2 ±2	2 ±3	2 ±3	18 ±6	4 ±2	2 ±3	9 ±7	3 ±3	7 ±4	8 ±4	4 ±4	5 ±4	2 ±3	1 ±1
SP≥40 NtraC	12 ±6	3 ±3	1 ±1	2 ±3	2 ±3	10 ±6	1 ±2	2 ±3	2 ±3	19 ±6	4 ±2	1 ±2	10 ±6	3 ±3	8 ±5	7 ±4	3 ±3	5 ±3	3 ±3	1 ±1
SP≥NtraC(mTP)	15 ±7	2 ±2	1 ±1	1 ±1	3 ±3	12 ±5	1 ±2	1 ±2	1 ±1	19 ±6	3 ±2	1 ±1	11 ±7	3 ±2	10 ±3	8 ±3	3 ±3	3 ±3	3 ±2	0 ±1
mTP	12 ±8	2 ±3	1 ±1	1 ±2	4 ±4	6 ±5	2 ±2	3 ±4	3 ±4	13 ±6	5 ±3	2 ±3	5 ±4	3 ±3	13 ±5	11 ±7	5 ±4	6 ±4	2 ±2	2 ±2
cTP	11 ±8	2 ±2	1 ±2	1 ±2	4 ±3	4 ±4	2 ±2	3 ±3	5 ±3	9 ±4	3 ±2	4 ±3	7 ±5	3 ±3	8 ±5	18 ±7	7 ±4	6 ±4	0 ±1	1 ±1
vSP	9 ±8	4 ±4	1 ±2	1 ±3	6 ±6	5 ±5	1 ±2	7 ±7	4 ±4	19 ±8	7 ±4	2 ±3	4 ±5	2 ±3	3 ±4	7 ±5	5 ±5	9 ±6	2 ±3	2 ±3

Zur Unterscheidung zwischen unterschiedlichen Targeting-Signalen (SP, mTP, cTP) ist die Auftrittshäufigkeit von Aminosäuren innerhalb der Targeting-Sequenz ohne Berücksichtigung zusätzlicher Information, z.B. der Position innerhalb der Sequenz, nach den vorliegenden Ergebnissen nicht geeignet. Die Aussage der Unterscheidbarkeit zwischen Signalpeptid und nativem Protein von Cedano *et al*. (1997) und Chou (2001) bleibt davon unberührt.

Unter viralen Signalpeptiden (vSP) sind alle Signalpeptide zusammengefasst, deren zugehörige Proteine in der UniProtKB (Version 14.0) als virale annotiert sind. Virale Signalpeptide zeigen eine Verteilung der Aminosäure-Häufigkeiten, die kurzen Signalpeptiden entspricht. Eine Zuordnung, basierend auf der Aminosäure-

Häufigkeitsverteilung, zu langen oder kurzen Signalpeptiden, ist daher nicht möglich. Dies entspricht auch der Beobachtung beim Vergleich der Längenverteilung von viralen Signalpeptiden und eukaryotischen Signalpeptiden (Kapitel „Längenverteilung von Signalpeptiden“). Fasst man die Daten aus der Längenverteilung und der Aminosäure-Häufigkeitsverteilung von viralen und langen Signalpeptiden zusammen, stellen sowohl virale als auch lange Signalpeptide eine Variation kurzer eukaryotischer Signalpeptide dar. Aufgrund der vorliegenden Ergebnisse wird festgehalten, dass

- Targeting-Signale anhand ihrer Aminosäure-Häufigkeitsverteilung nicht zu unterscheiden sind,
- von kurzen SP über lange SP hin zu NtraC-organisierten langen SP und mTP der Anteil an Arginin in den Targeting-Sequenzen zunimmt.

Dies ist als Hinweis zu werten, dass lange Signalpeptide eine eigene Untergruppe kurzer Signalpeptide darstellen, deren Unterscheidbarkeit mit der zwischen mTP und SP zu vergleichen ist. Lange Signalpeptide könnten, entsprechend viralen Signalpeptiden, eine Variation kurzer Signalpeptide darstellen, aber einem zusätzlichen selektiven Druck ausgesetzt sein.

Post-Targeting-Funktionen von Signalpeptiden

Virale Signalpeptide besitzen bekannte Post-Targeting-Funktionen. Die zusätzliche Länge, die sich besonders bei den viralen Signalpeptiden zeigt (Kapitel „Längenverteilung von Signalpeptiden“), bietet „Sequenzraum“ für die Kodierung dieser zusätzlichen Funktionen. Dieser Sequenzraum steht auch bei eukaryotischen langen Signalpeptiden zur Verfügung und wird auch hier für die Kodierung zusätzlicher Funktionen verwendet (Kurys *et al*., 2000; Hiss *et al*., 2008b). Im Folgenden sind Beispiele publizierter Post-Targeting-Funktionen und zusätzlicher Funktionen von langen Signalpeptiden fallweise aufgeführt.

1) Das humane *Cytomegalo-Virus* codiert in dem langen Signalpeptid des UL40 Virus-Proteins (37 Aminosäuren) ein Epitop für HLA-E. Es ist eine Kopie des Epitops, das natürlicherweise in HLA-C vorkommt und von der Zelle zur Überwachung des HLA-C Niveaus verwendet wird. Durch Kodierung des Epitops in einem Virus-Protein kann der Virus die HLA-C Expression abschwächen und gleichzeitig der Detektion von

natürlichen Killer-Zellen durch ein gesenktes HLA-E Niveau entkommen (Ulbrecht *et al.*, 2000).

2) Das *Lassa virus* Glycoprotein-C besitzt ein 58 Aminosäure langes Signalpeptid. Dieses Signalpeptid ist für die Reifung des funktionalen Glycoprotein-C notwendig. Ein Ersatz durch ein kurzes Signalpeptid führte zwar zu einem ER-Targeting, nicht aber zu einem funktionalen Glycoprotein-C. Eine Co-Expression des Wildtyp-Signalpeptides führte wieder zu einem reifen Glycoprotein-C. Dies zeigt, dass das lange Signalpeptid neben dem ER-Targeting eine weitere für die Reifung des Glykoprotein-> notwendig Funktion besitzt und nach der Abspaltung stabil bleibt (Eichler *et al*., 2003a; Eichler *et al*., 2003b).

3) Ein zu 2) identisches Verhalten konnte für das lange Signalpeptid (58 Aminosäuren) des Glycoprotein-C des *lymphocytischen choriomeningitis virus* (LCMV) gezeigt werden (Schrempf *et al*., 2007). Der Lassa-Virus und LCMV gehören beide zur Familie der Arena-Viren.

4) Das lange Signalpeptid (98 Aminosäuren) des REM Proteins aus *mouse mammary tumor virus* (MMTV) ist nach Abspaltung stabil. Es enthält ein Kern-Lokalisationssignal und akkumuliert nach Abspaltung in den Nucleoli (Dultz *et al*., 2008).

5) Das humane Interleukin 15 besitzt zwei Isoformen. Eine mit einem 29, eine mit einem 48 Aminosäuren langen Signalpeptid. Die Signalpeptide liegen auf unterschiedlichen Exons. Das kurze Signalpeptid wird durch Exon 5 kodiert, das lange Signalpeptid durch Exon 3-5. Die Länge des Signalpeptides hat Einfluss auf die ER-Targeting-Effizienz, die Glykosylierung und das Sezernieren (Kurys *et al*., 2000).

6) Das humane shrew-1 besitzt ein Signalpeptid von 43 Aminosäuren Länge. Das Signalpeptid enthält eine kryptische Domäne, die *in vitro* als mTP fungieren kann. Der Domänen-Aufbau entspricht dem in dieser Arbeit vorgestellten NtraC-Modell (Hiss *et al*., 2008b).

7) Die humanen Proteine DCBD2 und RGMA besitzen jeweils ein langes Signalpeptid. Beide Signalpeptide enthalten (entsprechend zu 6) eine Domäne, die *in vitro* als mTP fungieren kann (Resch und Hiss, unpublizierte Daten). Die Experimente und

Ergebnisse zu beiden Proteinen werden in dieser Arbeit (Kapitel „Das NtraC-Modell *in vitro“*) vorgestellt.

8) Das murine C4b-bindende Protein ist Teil des Komplement-Systems des Immunsystems. Es besitzt ein experimentell validiertes 56 Aminosäuren langes Signalpeptid. In Ogata *et al*., 1993 wird vorgeschlagen, dass das Signalpeptid eine Rolle bei der Faltung von C4b spielen könnte.

Diese dokumentierten Beispiele von Post-Targeting-Funktionen bei langen Signalpeptiden und weitere in Hegde, 2002 beschriebene waren Teil der Motivation dieser Arbeit, eine maschinelle Vorhersage für die Existenz und Lokalisation zusätzlicher Funktionen innerhalb langer Signalpeptide zu entwerfen und zu implementieren.

Das NtraC-Modell *in silico*

Im Rahmen dieser Arbeit wurde ein Konzept zur Identifizierung unabhängiger funktioneller Domänen in langen eukaryotischen Signalpeptiden (NtraC) entwickelt und publiziert (Hiss *et al.,* 2008b). Das Modell identifiziert dabei zunächst einen Übergangsbereich (engl. *transition area*, *tra*), basierend auf Sekundärstruktur-Vorhersagen für β-Turns. Die Detektion potentieller β-Turns erfolgt dabei unter Verwendung eines nach Meissner (Meissner *et al*. 2008) errechneten SVM-*Score* (Kapitel „Bioinformatische Online Programme“, SVMTurn). Vorhergesagte β-Turns werden zur Unterteilung des langen Signalpeptides in einen N-terminalen („N-Domäne“) und einen C-terminalen („C-Domäne“) Teil (Abb. 12) herangezogen. Die Lage des Übergangsbereiches („tra“) ist dabei zusätzlich motiviert durch die Konservierung der potentiellen β-Turns in homologen Sequenzen der zu überprüfenden Proteine. In den homologen Sequenzen ist nicht die exakte Sequenz, sondern die Fähigkeit, einen β-Turn zu formieren, konserviert. β-Turns sind für die Faltung von Proteinen entscheidende Sekundärstruktur-Elemente (Marcelino und Gierasch, 2008).

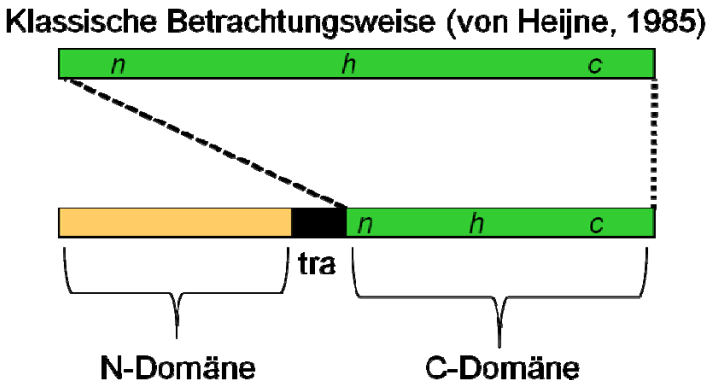


Abbildung 12: Das NtraC-Modell. **Orange**: N-Domäne (‚N'traC). **Schwarz**: Übergangsbereich (N„tra"C, engl. *transition area*). **Grün**: C-Domäne (Ntra,C'). ***n*, *h* und *c***: *n-, h-* und *c*-Region klassische Betrachtungsweise langer Signalpeptide entsprechend von Heijne, 1985.

Das Modell ermöglicht durch die Unterteilung in zwei Domänen, individuelle Targeting-Funktionen der einzelnen Domänen zu erkennen (Abb. 12). Eine Domäne hat dabei die ursprüngliche, meist die SP-Funktion, entsprechend des Targeting-Verhaltens des gesamten Signalpeptides. Die zweite Domäne ist für das Targeting nicht notwendig. Sie enthält potentielle zusätzliche Funktionen oder alternative Targeting-Signale. Die durch das NtraC-Modell beschriebenen funktionellen SP-Domänen weisen für sich eine *n-, h-* und *c*-Einteilung und somit die Charakteristika eines funktionellen Signals auf (Abb. 12). Der typische in von Heijne, 1985 beschriebene Gesamtaufbau von Signalpeptiden mit einer *n-*, *h-* und *c*-Region bleibt davon unberührt (Abb. 12). Die Signalsequenz als Ganzes betrachtet hat somit eine *n-, h-* und *c*-Einteilung, eine SP-Domäne für sich genommen hat ebenfalls eine *n-, h-* und *c*-Einteilung. Die Domänen können aber auch zu ER-Targeting-Signalen abweichende Targeting-Funktionen aufweisen, obwohl der Gesamtaufbau des Signalpeptides einem ER-Targeting-Signal entspricht. Im Rahmen dieser Arbeit wurden die im Folgenden aufgeführten potentiellen Funktionen jeweils für die N- und C-Domäne untersucht:

- Signalpeptid (SP, endoplasmatisches Retikulum)
- Signalanker (SA, Plasmamembran-Anker)
- Mitochondriales Targeting-Peptid (mTP, Mitochondrien)
- Chloroplasten-Transit-Peptid (cTP, Chloroplasten)
- Signalpeptid Gram-positive Bakterien (gram+, Bakterienmembran)
- Signalpeptid Gram-negative Bakterien (gram-, Bakterienmembran)
- Nicht spezifiziert (Die Domäne enthält keine spezifizierbare Funktion, wobei die zweite Domäne eine identifizierbare Funktion enthalten muss.)

Dabei ist zu beachten, dass die Reihenfolge und die Kombinationen der potentiellen Targeting-Funktionen der einzelnen Domänen keinen Einfluss auf die Zugehörigkeit zum NtraC-Modell haben. Eine Zugehörigkeit zum Modell wird postuliert, wenn das Signalpeptid in funktionale Domänen unterteilt werden kann. Eine Liste der im Rahmen dieser Arbeit betrachteten Kombinationen aus Targeting-Funktionen der Domänen ist in Tabelle 8 angegeben. Eine Erweiterung des NtraC-Modells um zusätzliche Targeting-Funktionen ist möglich.

Es müssen somit die folgenden Voraussetzungen für eine Organisation entsprechend dem NtraC-Modell bei einem langen Signalpeptid gegeben sein:

1) Hinreichende Gesamtlänge des Signalpeptides, um mindestens zwei Domänen enthalten zu können.
2) Existenz eines Übergangsbereiches der potentielle β-Turns enthält zur Unterteilung des Signalpeptides in zwei Domänen. Zur Festlegung können homologen Proteine, soweit bekannt, herangezogen werden.
3) Hinreichende Länge mindestens einer der resultierenden Domänen, um ein funktionelles Targeting-Signal zu enthalten.
4) Existenz einer durch SignalP oder TargetP erfassbaren Targeting-Funktion (entsprechend Tabelle 8).

Tabelle 8: Übersicht über die im Rahmen dieser Arbeit betrachteten Kombinationen aus Targeting-Funktionen. **SP**: Signalpeptid. **SA**: Signalanker. **mTP**: Mitochondriales Targeting-Peptid. **Gram-positiv**: Bakterielles Signalpeptid in Gram-positiven Bakterien. **Gram-negativ**: Bakterielles Signalpeptid in Gram-negativen Bakterien.

N-Domäne	C-Domäne
mTP	SP
SP	mTP
Nicht spezifiziert	SP
SP	Nicht spezifiziert
SA	Nicht spezifiziert
Nicht spezifiziert	SA
mTP	Nicht spezifiziert
Nicht spezifiziert	mTP
Nicht spezifiziert	Gram-positiv
Gram-positiv	Nicht spezifiziert
Nicht spezifiziert	Gram-negativ
Gram-negativ	Nicht spezifiziert
SP	Gram-positiv
Gram-positiv	SP
Gram-negativ	SP
SP	Gram-negativ
SP	SP
mTP	SA
SA	mTP

NtraC-Algorithmus

Der Ablauf des NtraC-Algorithmus ist im Folgenden in Pseudocode dargestellt:

1) *Wenn* Signalpeptidase-Schnittstelle gegeben, *dann* gehe zu **3),** sonst zu **2).**
2) *Schneide* gegebene Sequenz nach 100 Aminosäuren *ab*.
   Sage Position der Signalpeptidase-Schnittstelle unter Verwendung von SignalP 3.0 *voraus*.
3) *Schneide* gegebene Sequenz an Signalpeptidase-Schnittstelle *ab*.
4) *Suche nach* potentiell β-Turn-formierenden Bereichen (Übergangsbereich) innerhalb der verbleibenden Sequenz (SVM-*Score* nach Meissner *et al*., 2008)
5) *Zerlege* die verbleibende Sequenz an jedem gefundenen potentiellen Übergangsbereich in eine N- und eine C-Domäne.
6) *Sage* die potentielle Targeting-Funktion jeder gefundenen Domäne unter Verwendung von SignalP 3.0 und TargetP 1.1 *voraus*.
7) *Überprüfe* jede mögliche mögliche Kombinatione aus N- und C-Domänen.
8) **Ausgabe:** Aussage über die Zugehörigkeit zum NtraC-Modell, ggf. mit gefundenen möglichen Kombinationen aus N- und C-Domänen.

Webbasierte NtraC-Benutzeroberfläche

Für den beschriebenen NtraC-Algorithmus wurde im Rahmen dieser Arbeit eine öffentlich zugängliche webbasierte Benutzeroberfläche entwickelt. Diese Benutzeroberfläche wurde in Hiss *et al*., 2008b publiziert und ist unter folgender Adresse zu erreichen:

www.modlab.de → Software → NtraC

(http://gecco.org.chemie.uni-frankfurt.de/NtraC/NtraC.html)

Ziel der webbasierten Benutzeroberfläche ist es, einen öffentlich zugänglichen Algorithmus zur Analyse von langen Signalpeptiden hinsichtlich ihrer NtraC-Organisation zu bieten.

Die webbasierte Benutzeroberfläche ist in vier Schritte unterteilt, die in den Abb. 13-14 dargestellt und erläutert werden. Die Abbildungen sind Screenshots der originalen Weboberfläche. Eine Übersetzung wird jeweils in der Abbildungsbeschreibung zur Verfügung gestellt.

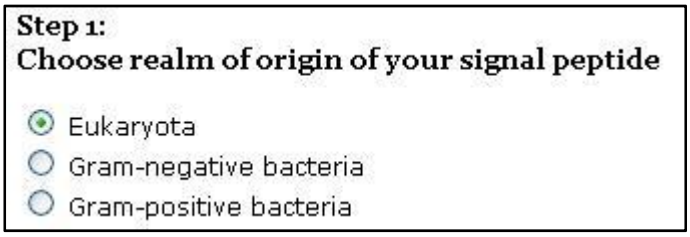


Abbildung 13: Schritt 1 der NtraC-Benutzeroberfläche. Übersetzung: Wählen Sie das Ursprungsreich ihres Signalpeptides: Eukaryota, Gram-negative Bakterien, Gram-positive Bakterien.

Im ersten Schritt (Abb. 13) erfolgt die Auswahl des Ursprungsreiches des zu untersuchenden Signalpeptides. Dies ist notwendig, falls die Signalpeptidase-Schnittstelle des zu untersuchenden Signalpeptides nicht bekannt ist und vorhergesagt werden soll. Die Signalpeptidase-Schnittstelle ist in den verschiedenen Reichen (Eukaryota, Bakterien) sowie in den verschiedenen Gruppen von Bakterien (Gram-negativ, Gram-positiv) unterschiedlich aufgebaut. Für eine korrekte Vorhersage ist daher die Information des Ursprungsreiches der Sequenz notwendig.

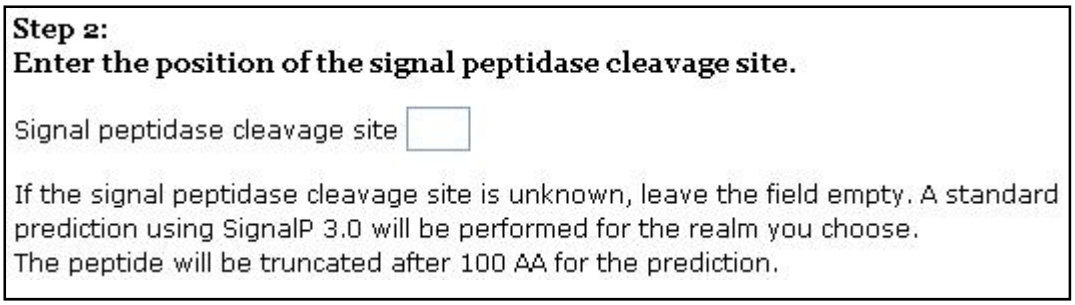


Abbildung 14: Schritt 2 der NtraC-Benutzeroberfläche. Übersetzung: Geben Sie die Position der Signalpeptidase-Schnittstelle ein. Falls die Signalpeptidase-Schnittstelle unbekannt ist, lassen Sie das Eingabefeld bitte frei. Eine Standard-Vorhersage unter der Verwendung von SignalP 3.0 wird in diesem Fall entsprechend dem von Ihnen gewählten Reich durchgeführt. Die Sequenz wird für die Vorhersage nach 100 Aminosäuren abgeschnitten.

Im zweiten Schritt (Abb. 14) ist die manuelle Eingabe der Signalpeptidase-Schnittstelle möglich. Dies erlaubt eine größere Flexibilität für den Anwender, da verschiedene z.B. auf experimentellen Erfahrungen oder eigenen Beobachtungen basierende Schnittstellen berücksichtigt werden können. Ist über die Sequenz nichts bekannt, kann das Feld freigelassen werden. Es erfolgt dann eine automatische Vorhersage der Position der Signalpeptidase-Schnittstelle, basierend auf dem in Schritt 1 gewählten Reich. Die Vorhersage erfolgt mit der lokalen Version von SignalP 3.0 (Bendtsen *et al*., 2004a) für die akademische Verwendung. Der Verwendung von SignalP 3.0 und TargetP 1.1 im Rahmen der NtraC-Vorhersage wurde vertraglich zugestimmt (Anhang A1).

Das Festlegen der Signalpeptidase-Schnittstelle hat einen starken Einfluss auf die Qualität der Vorhersage. Bei einzelnen Sequenzen mit langen Signalpeptiden aus Eukaryoten wird die Signalpeptidase-Schnittstelle bei einer automatischen Vorhersage mit SignalP nicht korrekt identifiziert (Tabelle 9). Dies ist darauf zurückzuführen, dass die Sequenzen eine degenerierte, näher am N-Terminus liegende Signalpeptidase-Schnittstelle besitzen. Diese wird gegenüber der korrekten, aber weiter *upstream* liegenden Schnittstelle bevorzugt (Tabelle 9, Nr 1, 3 bis 5). Ursache dafür ist die Berechnung und das Training von SignalP 3.0. Im Training wurden Signalpeptide mit <15 und >45 Aminosäuren nicht berücksichtigt (Bendtsen *et al*., 2004a). Des Weiteren werden Signalpeptidase-Schnittstellen mit einem Straf-Term belegt, die weiter vom N-Terminus entfernt liegen als die durchschnittliche Länge eines Signalpeptides von 22 Aminosäuren (Bendtsen *et al*., 2004a). Es ist nicht möglich, diese Parameter manuell in SignalP 3.0 zu ändern. Des Weiteren werden Signalpeptidase-

Schnittstellen, die weit vom N-Terminus entfernt liegen, in bestimmten Fällen (Tabelle 9, Nr.2 und 6) erkannt, aber der davor liegende Bereich wird in der automatischen Vorhersage nicht als Signalpeptid, sondern als Signalanker identifiziert. In beiden Fällen führt eine automatische Vorhersage durch SignalP somit zu einem fälschlich verkürzten Signalpeptid oder einer Nicht-Identifizierung des Signalpeptides. Entsprechend werden die auf den Vorhersagen aufbauenden NtraC-Kalkulationen fehlerhaft. Bei bestehendem Wissen oder auch bei Vermutungen über die Signalpeptidase-Schnittstelle empfiehlt der Autor daher ausdrücklich die manuelle Eingabe der Schnittstellenposition. Ein N-terminal liegendes Transmembran-Segment kann ebenfalls aufgrund des hydrophoben Charakters mit einem Signalpeptid verwechselt werden (Lohmann *et al*., 1994).

Tabelle 9: Durch automatische Vorhersage (SignalP 3.0) falsch identifizierte Signalpeptidase-Schnittstellen. SP: Signalpeptid. SA: Signalanker.

Nr.	Proteinname	Organismus	Signalpeptidase -Schnittstelle SignalP-Vorhersage	Signalpeptidase -Schnittstelle tatsächliche Position	SignalP Aussage
1	Glycoprotein	*Lassa* virus	34	58	SP
2	Beta-fructofuranosidase, lösliches Isoenzym I	*Daucus carota*	0	57	SA
3	Ring-infiziertes Erythrozyt Oberflächen Antigen	*Plasmodium falciparum*	18	65	Nicht-sekretiertes Protein
4	Sporulations-spezifisches Protein 2	*Saccharomyces cerevisiae*	23	56	Nicht-sekretiertes Protein
5	Verzweigt-kettige alpha-keto Säure Dehydrogenase E1-beta Untereinheit	*Gallus gallus*	30	50	Nicht-sekretiertes Protein
6	ENK1-ENK7	*Homo sapiens*	88 bzw.89	88 bzw. 89	SA

Im dritten Schritt (Abb. 15) wird die Art der NtraC-Organisation ausgewählt. Die Auswahlmöglichkeiten kommen zustande, da das NtraC-Modell zwar die Existenz, nicht aber die Reihenfolge der Targeting-Funktionen festlegt. Die in dieser Arbeit betrachteten möglichen N/C-Domäne-Kombinationen sind in Tabelle 8 aufgeführt. Der limitierende Faktor stellt die Vorhersagemöglichkeiten von SignalP und TargetP dar. Eine Erweiterung auf zusätzliche potentielle Funktionen, soweit diese durch etablierte automatisierte Vorhersageprogramme erfasst werden, ist möglich. Die hier zur Verfügung gestellten optionalen Kombinationen ermöglichen eine Erweiterung des Modells auf unterschiedliche Kombinationen aus Targeting-Funktionen (Tabelle 8).

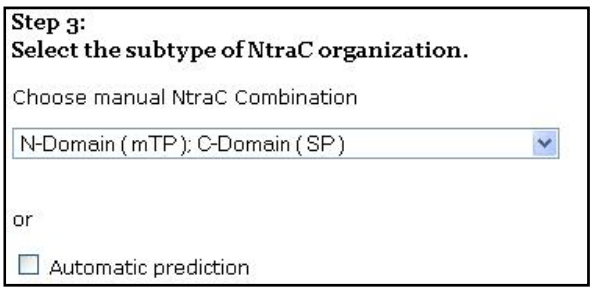


Abbildung 15: Schritt 3 der NtraC-Benutzeroberfläche. Übersetzung: Wählen Sie den Subtyp der NtraC-Organisation. Wählen Sie manuell eine NtraC-Kombination. (Es folgt ein Aufklappmenü mit verschiedenen Auswahlmöglichkeiten, hier zu sehen N-Domäne (mTP = mitochondriales Targeting-Peptid), C-Domäne (SP = Signalpeptid)). Alternativ zur manuellen Auswahl wählen Sie die „Automatische Vorhersage“.

Zur Steigerung der Benutzerfreundlichkeit wurde zusätzlich eine automatische Vorhersage implementiert. Diese kann durch Setzen des entsprechenden Auswahlhakens in der Benutzeroberfläche „Schritt 3“ aktiviert werden. Diese nicht triviale Funktionalität untersucht die gegebene Sequenz auf jede mögliche Kombination aus Domänen. Dem Anwender wird so ermöglicht, ohne Vorwissen über die Kombination aus Domänen eine Vorhersage durchzuführen. Des Weiteren erlaubt es auch, bei bekannter Kombination der Domänen zusätzliche Informationen über die Verhaltensweisen in anderen Reichen zu erhalten. Dies ist von Bedeutung, wenn z.B. eukaryotische Sequenzen für die experimentelle Überprüfung in bakterielle Zellen eingebracht werden, aber zusätzlich potentielle bakterielle Signale enthalten. Die automatische Vorhersage stellt daher die vom Autor empfohlene Auswahl dar.

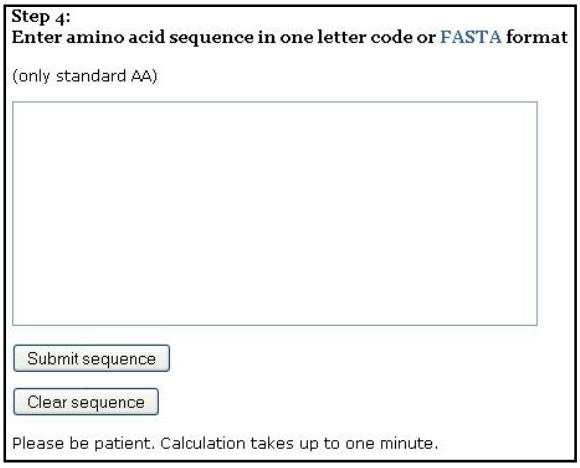


Abbildung 16: Schritt 4 der NtraC-Benutzeroberfläche. Übersetzung: Geben Sie die Aminosäure-Sequenz im Ein-Buchstaben-Code oder im FASTA-Format ein. Nur Standard-Aminosäuren (endogene). Submit Sequence: Start der Kalkulation; Clear Sequence: Löschen der Sequenz. Hinweis: Bitte haben Sie etwas Geduld, die Kalkulation kann bis zu einer Minute dauern.

Im vierten und letzten Schritt (Abb. 16) erfolgt die Eingabe der zu untersuchenden Aminosäure-Sequenz. Dies kann im Ein-Buchstaben-Code (Tabelle 6) oder im FASTA-Format erfolgen. Für das FASTA-Format wird die Beschreibung des NCBI als Link (http://www.ncbi.nlm.nih.gov/blast/FASTA.shtml) zur Verfügung gestellt. In der Aminosäure-Sequenz dürfen nur die 20 endogenen Aminosäuren (Tabelle 6) enthalten sein. Der Buchstabe X für eine beliebige Aminosäure kann nicht verwendet werden.

Durch Drücken des Knopfes „Submit sequence“ wird die Analyse der Sequenz gestartet. Diverse Fehler, wie z.B. „keine Sequenz“, „zu kurze Sequenz“ oder die „Verwendung unbekannter Aminosäuren“ werden abgefangen und dem Anwender mitgeteilt. Das Ergebnis („Ausgabe“) wird dem Anwender unmittelbar im Anschluss an die Berechnung auf einer für jede Anfrage neu erstellten Ergebnisse-Seite angezeigt.

Ausgabe der webbasierten Benutzeroberfläche

Sollte die gegebene Sequenz keine NtraC-Organisation aufweisen, erhält der Anwender einen entsprechenden Hinweis. Dieser Hinweis enthält gegebenenfalls Erläuterungen für die negative Aussage des Programms, wie z.B., dass kein Übergangsbereich (tra) gefunden werden konnte oder dass es keine N/C-Domänen-Kombination entsprechend der vom Anwender gewählten Vorgabe gibt. Dies ermöglicht dem Anwender eine erneute Eingabe der Sequenz mit veränderten Parametern.

Im Falle einer NtraC-Organisation der Sequenz und der Auswahl „automatische Vorhersage“ erhält der Anwender die in Abbildung 17 und 18 exemplarisch beschriebene Ausgabe. Die Ausgabe ist unterteilt in zwei Abschnitte. Im ersten Abschnitt (Abb. 17) werden zunächst die gewählten Parameter „Ursprungsreich der Sequenz“ und „Signalpeptidase-Schnittstelle“ dargestellt. Darauf folgt eine Liste gefundener N- und C-Domänen, deren vorhergesagte Targeting-Funktion sowie deren Position innerhalb der Sequenz. Der Wert am Ende der Zeile gibt die Ausgabewerte von SignalP bei SP, SA, gram-, gram+ und TargetP bei mTP wieder. Im Falle von SignalP 3.0 handelt es sich um eine Wahrscheinlichkeit mit Werten im Bereich [0,1]. Im Falle von TargetP 1.1 handelt es sich um die Ausgabewerte der KNN und damit um keine Wahrscheinlichkeiten (Emanuelsson *et al*., 2000). Ein hoher Wert drückt nach Emanuelsson *et al*. (2000) dennoch die wahrscheinlichste Lokalisation aus. Der Abstand in den Sequenzen der C-Domänen stellt die verwendete Signalpeptidase-Schnittstelle dar.

```
Setup:

Realm:     Eukaryotic
Signal Peptidase Cleavage site: 43
Predicted domains:

gi|308420,16, SP, prob. C-domain(20-43): 0.920
 PLGSHAWILIAMFQLAVDLPACEA LGPGP
gi|308420,20, SP, prob. C-domain(24-43): 0.856
 HAWILIAMFQLAVDLPACEA LGPGP
gi|308420,21, SP, prob. C-domain(25-43): 0.715
 AWILIAMFQLAVDLPACEA LGPGP
gi|308420,20, mTP, prob. N-domain(1-23): 0.422
 MWIQQLLGLSSMSIRWPGRPLGS
gi|308420,21, mTP, prob. N-domain(1-24): 0.441
 MWIQQLLGLSSMSIRWPGRPLGSH
-------------------------------------------------------
```

Abbildung 17: Erster Abschnitt der automatischen Ausgabe, beispielhaft für die Sequenz von shrew-1 (AAP35025).

```
Possible N- and C-Domain combination
gi|308420,20, mTP, prob. N-domain(1-23): 0.422
 MWIQQLLGLSSMSIRWPGRPLGS
gi|308420,20, SP, prob. C-domain(24-43): 0.856
 HAWILIAMFQLAVDLPACEA LGPGP
gi|308420,21, mTP, prob. N-domain(1-24): 0.441
 MWIQQLLGLSSMSIRWPGRPLGSH
gi|308420,21, SP, prob. C-domain(25-43): 0.715
 AWILIAMFQLAVDLPACEA LGPGP
```

Abbildung 18: Zweiter Abschnitt der automatischen Ausgabe, beispielhaft für die Sequenz von shrew-1 (AAP35025).

Im zweiten Abschnitt (Abb. 18) der Ausgabe werden mögliche Kombinationen aus den gefundenen N-und C-Domänen aufgeführt. Wurde ein Übergangsbereich gefunden, der die Sequenz so unterteilt, dass für beide resultierenden Domänen eine Targeting-Vorhersage gemacht werden kann, stellt dies ein gegenüber einer Einzeldomäne hervorzuhebendes Ereignis dar. Daher werden diese Kombinationen, soweit vorhanden, in einem zweiten Abschnitt der Ausgabe gesondert aufgeführt. Es werden im zweiten Abschnitt keine zusätzlichen Domänen aufgeführt, sondern mögliche Kombinationen aus den im ersten Abschnitt (Abb. 17) gefundenen Domänen aufgezeigt.

Der Autor empfiehlt für die Verwendung der webbasierten NtraC-Benutzeroberfläche die folgenden Einstellungen:

- manuelle Angabe einer Signalpeptidase-Schnittstelle (Schritt 2),
- Auswahl „Automatic prediction“ (Schritt 3).

Die webbasierte NtraC-Benutzeroberfläche stellt eine Möglichkeit für andere Wissenschaftler dar, Sequenzen hinsichtlich ihrer Zugehörigkeit zum NtraC-Modell zu untersuchen. Diese Zugänglichkeit und somit die Entwicklung der webbasierten Benutzeroberfläche war ein erklärtes Ziel dieser Arbeit.

Datenbankrecherche mit dem NtraC-Algorithmus

Mit dem in Kapitel „NtraC-Algorithmus“ beschriebenen Algorithmus wurde eine systematische Suche nach langen Signalpeptiden durchgeführt, die dem NtraC-Modell entsprechen. Dazu wurden mithilfe des Sequence-Retrival-System (SRS, Version 53.2) aus der UniProtKB-Datenbank (Version 13.6) alle Vertebrata-Proteine mit Signalpeptiden mit mindestens 40 Aminosäuren extrahiert (*N* = 296; Anhang, Tabelle A7).

Die Einteilung der gefundenen 296 Signalpeptide entsprechend der Zugehörigkeit zum NtraC-Modell ist in Abbildung 19 dargestellt. Die im Vergleich zu den insgesamt annotierten Vertebrata-Signalpeptiden (5.245) geringe Anzahl an langen Signalpeptide (296 ≙ 6%) ist auf die Annotation innerhalb der Datenbank zurückzuführen. Im Unterschied zu den in Kapitel „Längenverteilung von Signalpeptiden“ aufgeführten Signalpeptiden mit mehr als 40 Aminosäuren (*N* = 136) sind bei dieser Suche (*N* = 296) auch putative Proteine berücksichtigt worden. Dies ist motiviert durch die Tatsache, dass für die statistische Analyse der

Längenverteilung die exakte Länge der Signalpeptide von Bedeutung ist, für die Analyse nach dem NtraC-Modell primär deren Existenz. Mithilfe des SRS wurden nur Sequenzen gefunden, die ein *FeatureKey* „Signal“ mit einem *FeatureKey value* „≥ 40“ als Eintrag besitzen. Diverse experimentell erkannte lange Signalpeptide (z.B. shrew-1; Resch *et al.*, 2008) sind in der Datenbank nicht aufgeführt. Dies ist darauf zurückzuführen, dass die entsprechende publizierte Information der Datenbank noch nicht mitgeteilt wurde oder im Fall von langen Signalpeptiden auch einfach nicht vorhanden ist. *In silico* vorhergesagte Signalpeptide mit ≥40 Aminosäuren Länge werden oft als Artefakt betrachtet und nicht in die Datenbank übernommen bzw. experimentell nicht berücksichtigt. Des Weiteren werden lange Signalpeptide in bestimmten Fällen bereits *in silico* nicht erkannt (Tabelle 9). Der Vorteil dieser quantitativ betrachtet schlechten Datenlage ist die Qualität der Daten. Die Annotation langer Signalpeptide erfolgt in der Regel nur nach experimenteller Validierung oder persönlicher Kommunikation, selten automatisch.

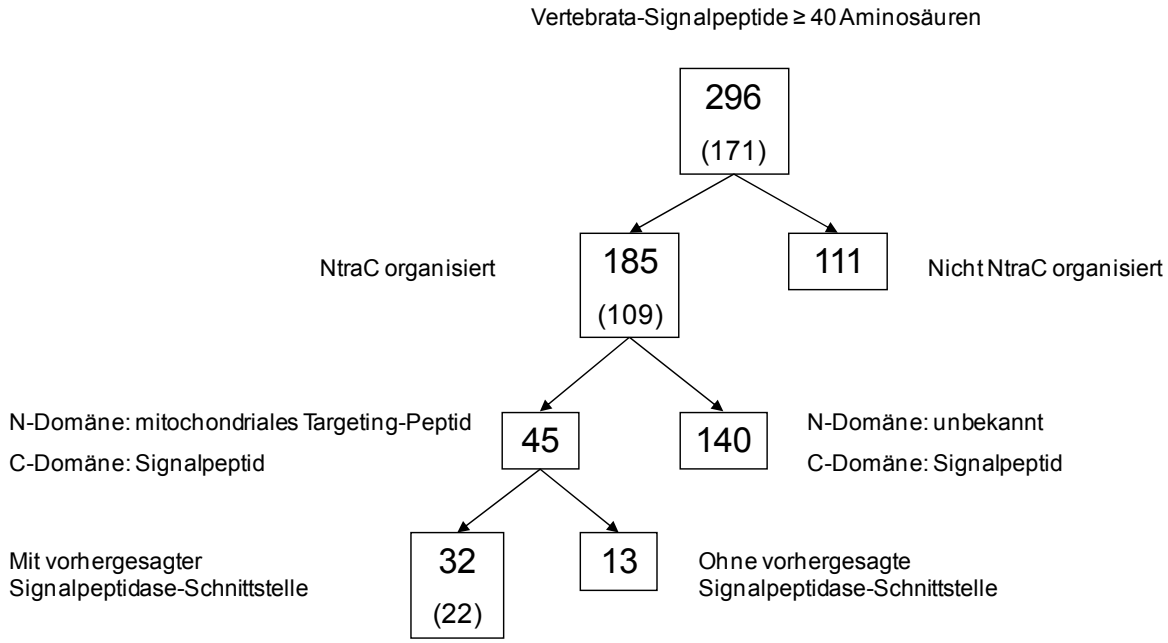


Abbildung 19: Aufteilung der langen Vertebrata-Signalpeptide entsprechend dem NtraC-Modell. **Zahlen in Klammern** geben die Werte ohne orthologe Proteine wieder.

Von den mithilfe von SRS gefundenen 296 Vertebrata-Signalpeptiden mit ≥40 Resten weisen insgesamt 185 (63%) eine NtraC-Organisation auf (Abb. 19). Dabei liegt eine NtraC-Organisation vor, wenn ein Übergangsbereich vorhergesagt wird und mindestens eine Domäne eine vorhersagbare Targeting-Funktion besitzt. 32 Signalsequenzen (11%; Anhang Tabelle A8) weisen die folgende Kombination von N- und C-Domäne auf:

- N-Domäne: mitochondriales Targeting-Peptid,
- C-Domäne: Signalpeptid.

Diese Kombination entspricht der Domänen-Architektur, der in dieser Arbeit experimentell untersuchten langen Signalsequenzen von shrew-1, „Discoidin, CUB and LCCL domain-containing protein 2“ (DCBD2) und *Repulsive guidance molecule* A (RGMA; Kapitel „Das NtraC-Modell *in vitro*). DCBD2 und RGMA sind in den 32 aufgeführten Sequenzen in Abb. 19 enthalten, shrew-1 nicht (entsprechend der Annotation in der UniProtKB Datenbank).
Die vorliegenden Ergebnisse werden vom Autor dahingehend gewertet, dass es sich bei dem vorgeschlagenen NtraC-Modell um ein grundlegendes Konzept für die Architektur langer Signalpeptide handelt. Das Auftreten einer NtraC-Organisation in 63% aller annotierten Vertebrata-Signalpeptide weist auf eine funktionelle Bedeutung dieser Signalpeptid-Architektur hin. Zur Validierung dieser These wurde die Gruppe der 185 NtraC-organisierten Proteine hinsichtlich weiterer Gemeinsamkeiten untersucht (Kapitel „Semantische Wolke“).

Semantische Wolke

Es stellt sich die Frage, ob die Proteine, die durch die NtraC-Organisation ihrer Signalpeptide eine Gruppe bilden, weitere funktionelle Gemeinsamkeiten aufweisen. Zu diesem Zweck wurde eine semantische Wolke (s*emantic cloud*, Kim *et al*., 2008) für die Proteine erstellt. Die semantische Wolke stellt hier eine Liste mit Schlüsselbegriffen dar, die aus den UniProtKB-Einträgen der Proteine stammen und mit dem Protein in Zusammenhang stehen. Bei der manuellen Durchsicht der Proteine wurde das gehäufte Auftreten der folgenden feststehenden Schlüsselbegriffe aus der UniProtKB festgestellt:

- *glycoprotein*
- *alternative splicing*
- *epidermal growth factor (EGF)*
- *cell adhesion*
- *immunglobulin*
- *metal-binding*
- *lysosome*
- *single-pass transmembrane domain (single-pass TMD)*

Das Auftreten dieser Schlüsselbegriffe wurde in einem zweiten Schritt automatisiert erfasst. Die Ergebnisse sind in Abbildung 18 aufgeführt. Die blauen Balken stellen dabei jeweils den Hintergrund in Form des Auftretens bei Signalpeptiden der Länge 1-40 dar.

Die Auswahl der in Abbildung 20 betrachteten Schlüsselwörter erfolgte zunächst manuell durch den Autor nach Sichtung aller Datenbankeinträge. Im Anschluss wurde zusätzlich eine qualitative maschinelle Auswertung vorgenommen, bei der alle Schlüsselwörter, die über einen bestimmten Schwellenwert an Datenbankeinträgen (≥10%, 20%, 50%) vorhanden waren, herausgefiltert wurden. Die Auswertung erfolgte getrennt für jede Gruppe (Signalpeptide der Länge 1-40, der Länge 40-100, NtraC-organisiert, NtraC-organisiert N-Domäne: mTP, C-Domäne: SP) und ist in den Tabellen 10-13 nach diesen Gruppen getrennt aufgeführt.
Der Ansatz in Abbildung 18 ermöglicht einen Vergleich der Zu- oder Abnahme der Häufigkeit des Auftretens der ausgewählten Schlüsselbegriffe.

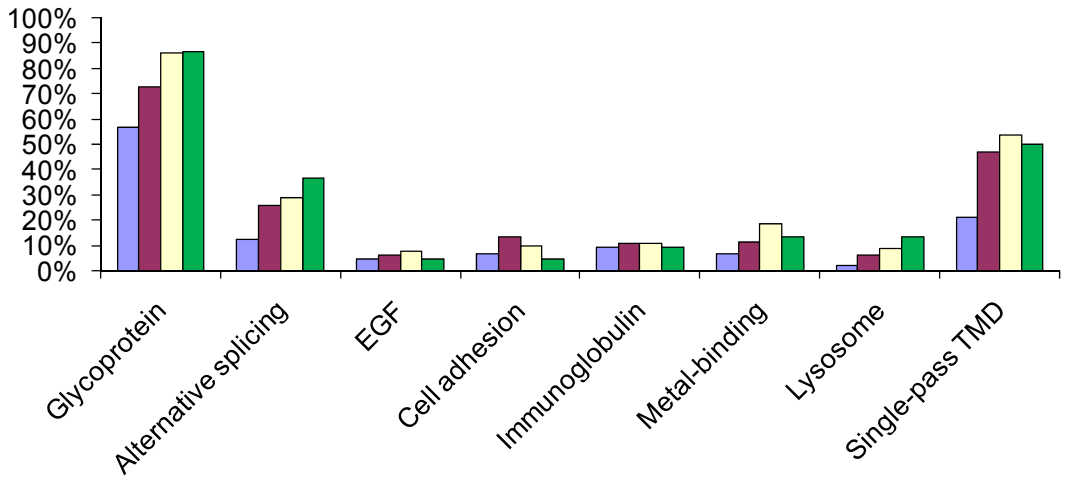


Abbildung 20: Verteilung von geschützten Schlüsselbegriffen in UniProtKB-Datenbankeinträgen Signalpeptid tragender Vertebrata-Proteine. **Blau**: Vertebrata-Proteine mit annotierter ER-Signalsequenz der Länge 1-39 (*N* = 14.153, putative eingeschlossen). **Rot**: Vertebrata-Proteine mit annotierter ER-Signalsequenz der Länge 40-100 (*N* = 296). **Gelb**: NtraC-organisierte Signalsequenzen der Länge 40-100, ohne orthologe Proteine, nur Signalpeptide, deren N- und C-Domäne mindestens eine Targeting-Funktion enthält (*N* = 109). **Grün**: NtraC-organisierte Signalsequenzen der Länge 40-100, ohne orthologe Proteine, mit N-Domäne: mTP, C-Domäne: SP (*N* = 32).

Der systematische Ansatz in Tabelle 10-13 stellt sicher, dass keine Schlüsselbegriffe, die in mehr als 50% der Einträge auftreten, bei der manuellen Durchsicht übersehen wurden. In den Tabellen 10-13 ist zu erkennen, dass zwei Begriffe in mehr als 50% der

Proteinbeschreibungen jeder Gruppe (kurze SP, lange SP, lange SP mit NtraC, lange SP mit NtraC und mTP+SP) vorkommen. Dies sind die Begriffe „Glycoprotein“ und „Signal“. Der Begriff „Signal“ dient dabei als positive Kontrolle der Methoden, da er in jedem Eintrag vorkommen muss, da alle Proteine ein Signalpeptid enthalten. Betrachtet man bei dem Begriff Glycoprotein nur die qualitativen Aussagen der Tabellen 10-13, ist kein Unterschied zwischen den Gruppen erkennbar.

Tabelle 10: Verteilung von Schlüsselbegriffen in der Gruppe „ Vertebrata-Signalpeptide 1-40“ (*N* = 14.353).

Schlüsselbegriff tritt in ≥ 10% der Einträge auf	**Schlüsselbegriff tritt in ≥ 20% der Einträge auf**	**Schlüsselbegriff tritt in ≥ 50% der Einträge auf**
alternative splicing	*glycoprotein*	*glycoprotein*
direct protein sequencing	*membrane*	*signal*
glycoprotein	*repeat*	
hydrolase	*signal*	
membrane	*transmembrane*	
polymorphism		
receptor		
repeat		
signal		
transmembrane		

Tabelle 11: Verteilung von Schlüsselbegriffen in der Gruppe „Vertebrata-Signalpeptide 40-100“ (*N* = 296).

Schlüsselbegriff tritt in ≥ 10% der Einträge auf	**Schlüsselbegriff tritt in ≥ 20% der Einträge auf**	**Schlüsselbegriff tritt in ≥ 50% der Einträge auf**
alternative splicing	*alternative splicing*	*glycoprotein*
cell adhesion	*glycoprotein*	*membrane*
glycoprotein	*membrane*	*signal*
hydrolase	*repeat*	*transmembrane*
membrane	*signal*	
metal-binding	*transmembrane*	
phosphorylation		
polymorphism		
receptor		
repeat		
signal		
transmembrane		

Tabelle 12: Verteilung von Schlüsselbegriffen in der Gruppe „NtraC-organisierte Signalpeptide, ohne orthologe Proteine“ (*N* = 109).

Schlüsselbegriff tritt in ≥ 10% der Einträge auf	**Schlüsselbegriff tritt in ≥ 20% der Einträge auf**	**Schlüsselbegriff tritt in ≥ 50% der Einträge auf**
alternative splicing	*alternative splicing*	*glycoprotein*
calcium	*glycoprotein*	*membrane*
extracellular matrix	*hydrolase*	*signal*
glycoprotein	*membrane*	*transmembrane*
hydrolase	*repeat*	
membrane	*signal*	
metal-binding	*transmembrane*	
protease		
receptor		
repeat		
signal		
transmembrane		
zinc		

Tabelle 13: Verteilung von Schlüsselbegriffen in der Gruppe „NtraC-organisierte Signalpeptide mit N-Domäne: mTP, C-Domäne: SP“ (*N* = 32).

Schlüsselbegriff tritt in ≥ 10% der Einträge auf	**Schlüsselbegriff tritt in ≥ 20% der Einträge auf**	**Schlüsselbegriff tritt in ≥ 50% der Einträge auf**
alternative splicing	*alternative splicing*	*glycoprotein*
calcium	*glycoprotein*	*membrane*
glycoprotein	*hydrolase*	*signal*
GPI-anchor	*membrane*	*transmembrane*
hydrolase	*repeat*	
lipoprotein	*signal*	
lysosome	*transmembrane*	
membrane		
metal-binding		
protease		
repeat		
signal		
transmembrane		

Unter Berücksichtigung der quantitativen Auswertung aus Abbildung 18 ist aber zu erkennen, dass das Auftreten des Begriffes Glycoprotein in der Gruppe der langen Signalpeptide um 15% und in der Gruppe der NtraC-organisierten langen Signalpeptide um 25% im Vergleich zur Gruppe kurzer Signalpeptide zunimmt. Dies ist als ein erster Hinweis zu werten, dass Proteine mit NtraC-organisierten Signalpeptiden neben dem Signalpeptid weitere Gemeinsamkeiten aufweisen. Die *in vitro* untersuchten Proteine DCBD2 und RGMA

(Kapitel „Das NtraC-Modell *in vitro*) sind Glycoproteine. Glycoproteine spielen insbesondere bei der viralen Erkennung von Wirtszellen eine entscheidende Rolle (Eichler *et al*., 2003; Isaacson *et al*, 2008).

Des Weiteren treten die Begriffe „Membran“ und „Transmembran“ in mehr als 50% der Proteine den Gruppen, Vertebrata-SP 40-100, NtraC-organisierte SP und NtraC-organisierte SP mit N-Domäne: mTP, C-domäne: SP auf (Tabelle 11,12 und 13). Da es sich hierbei um die Analyse von Kommentaren und Beschreibungen zu den Proteinen handelt, ist anzunehmen, dass beide Begriffe jeweils auf ein Membranprotein hindeuten. Dies ist konform zu der Überlegung, dass die Proteine, die ein ER-Targeting-Signal haben, vornehmlich extrazelluläre Proteine oder Membranproteine sind. Eine quantitative Analyse (Abb. 20) des Begriffes „*single-pass transmembrane domain*“ zeigt, dass 50% der Proteine mit NtraC-organisiertem Signalpeptid auch *single-pass* Transmembranproteine sind. Im Vergleich dazu sind Proteine mit kurzem Signalpeptid nur in 26% der Fälle auch als einzelspännige Transmembranproteinen annotiert. Auch die in dieser Arbeit experimentell überprüften Proteine mit einem NtraC-organisierten Signalpeptid (shrew-1 und DCBD2, RGMA) sind einzelspännige Transmembranproteine. Für RGMA ist lediglich „Membranprotein“ annotiert.

Bei den Begriffen „alternatives Spleißen“ und „Lysosom“ ist ebenfalls eine Zunahme der entsprechend annotierten Proteine in der Gruppe mit NtraC-organisierten Signalpeptiden im Vergleich zu kurzen und langen Signalpeptiden ohne NtraC zu erkennen (Abb. 20). Die Proteine DCBD2 und RGMA sind ebenfalls mit dem Vermerk „alternatives Spleißen“ annotiert.

Zusammenfassend wird aufgrund der vorliegenden *in silico* Schlüsselwort-Analyse festgestellt, dass die Gruppe an Proteinen, die lange NtraC-organisierte Signalpeptide besitzen, zusätzlich eine Anreicherung an Proteinen mit folgenden Eigenschaften aufweisen:

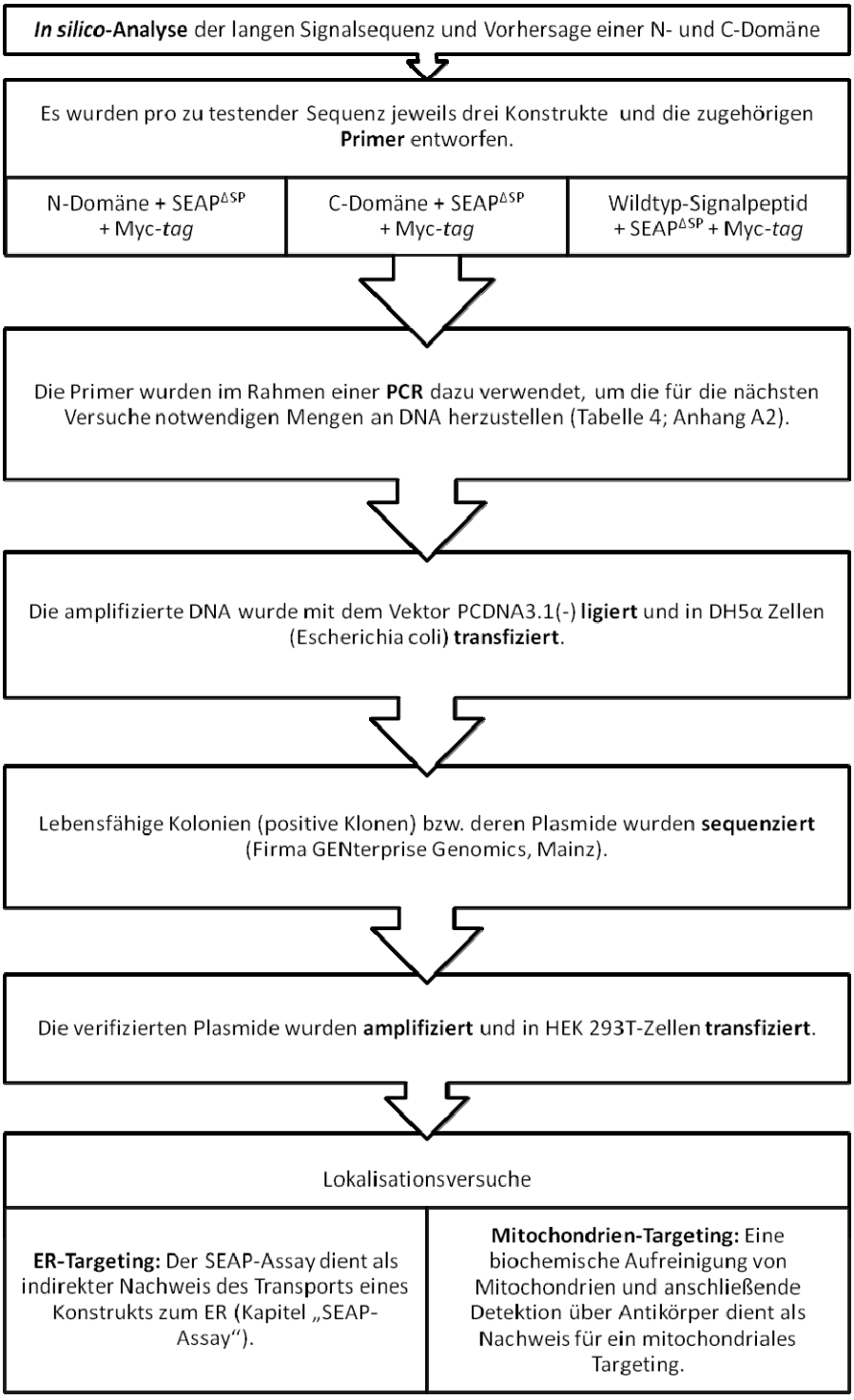
- Glycoproteine
- einzelspännige Transmembranproteine
- alternative gespleißte Proteine.

Ein funktioneller Zusammenhang zwischen dem NtraC-organisierten Signalpeptid und den genannten Aspekten der Proteine muss in weiteren Experimenten überprüft werden.

Das NtraC-Modell *in vitro*

Zur weiteren Überprüfung des NtraC-Modells wurden zusätzlich zu den in Kapitel „Das NtraC-Modell *in silico*“ beschriebenen *in silico*-Experimenten im Rahmen dieser Arbeit *in vitro*-Experimente in Kooperation mit dem Labor von Prof. Dr. Anna Starzinski-Powitz (JWG-Universität, Frankfurt/Main) durchgeführt. Die Experimente wurden sowohl von Dipl.-Biologe Eduard Resch als auch unter dessen und Dr. Alexander Schreiners Betreuung durch den Autor selbst ausgeführt.

Ziel der Experimente war, durch Kombination der N- und C-Domäne mit einem Reporter-Gen bzw. durch Antikörper-Detektion die vorhergesagte zelluläre Lokalisation der Konstrukte zu überprüfen. Für die drei hier untersuchten langen Signalsequenzen von shrew-1, DCBD2 und RGMA wurde jeweils für die N-Domäne eine mTP-Funktion und für die C-Domäne eine SP-Funktion vorhergesagt und experimentell bestätigt.
Die experimentellen Details der Assays zur Überprüfung der zellulären Lokalisation sind im Kapitel „Biologische Methoden“ aufgeführt. Im Folgenden werden die allgemeine Durchführung als Ablauf-Protokoll sowie das Arbeits-Konzept beschrieben. Das Vorgehen war für alle getesteten Konstrukte identisch und wird daher getrennt von den sequenzspezifischen experimentellen Ergebnissen und Konstrukten hier im Vorfeld erläutert.
Das folgende Schema gibt einen Überblick über die *in vitro*-Untersuchungen:



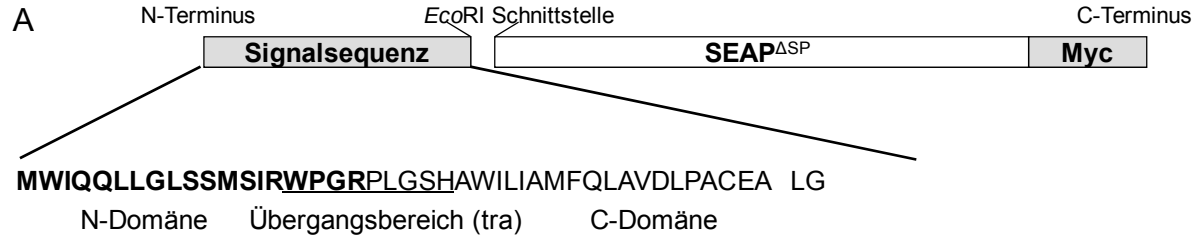
Shrew-1-Protein

Shrew-1 wurde ursprünglich bei einer Biopsie aus einer Epithel-Zelllinie isoliert (Bharti *et al*., 2004). Shrew-1 besitzt eine N-terminale lange Signalsequenz von 43 Aminosäuren (Aminosäure 1-43; Resch *et al*., 2008). Es ist ein Typ-I Transmembranprotein mit einem Transmembran-Segment an Position 283-303, gefolgt von einer cytoplasmatischen Domäne (Aminosäuren 304-411). Es wurde gezeigt, dass shrew-1 innerhalb polarisierter Zellen

basolateral lokalisiert und dort mit den E-Cadherin vermittelten Adhärenz-Verbindungen interagiert (Bharti *et al*., 2004; Jakob *et al*., 2006). In nicht polarisierten Zellen zeigt shrew-1 eine gleichmäßige Plasmamembran-Lokalisation. Shrew-1 spielt des Weiteren eine Rolle bei der Regulation von Zellinvasivität und Zellmotilität über eine Interaktion mit CD147 (Schreiner *et al*., 2007).

Die Konstrukte für shrew-1 wurden im Rahmen dieser Arbeit vom Autor *in silico* entworfen, die *in vitro*-Experimente wurden von Dipl. Biologe Eduard Resch durchgeführt. Die Angaben entstammen der gemeinsamen Publikation (Hiss *et al*., 2008b, Abb. 21). Alle Konstrukte wurden N-terminal mit der SEAP ohne ihr endogenes Signalpeptid fusioniert. C-terminal wurde an die SEAP jeweils der Myc-*tag* (Aminosäure-Sequenz: EQKLISEEDL) fusioniert.

A



B

Konstrukte	Signalsequenzen
SH-SEAP$^{\Delta SP}$	**MWIQQLLGLSSMSIR<u>WPGR</u>**PLGSHAWILIAMFQLAVDLPACEA LG
SH^{G18I}-SEAP$^{\Delta SP}$	**MWIQQLLGLSSMSIR<u>WPIR</u>**PLGSHAWILIAMFQLAVDLPACEA▼LG
SH$^{\Delta WPGR}$-SEAP$^{\Delta SP}$	**MWIQQLLGLSSMSIR**PLGSHAWILIAMFQLAVDLPACEA▼LG
SH$^{\Delta WPGR/mut}$-SEAP$^{\Delta SP}$	**MWIQQLLGLSSMSI**APSD–HAWILIAMFQLAVDLPACEA▼LG
SH^{C}-SEAP$^{\Delta SP}$	M<u>PLGSH</u>AWILIAMFQLAVDLPACEA▼LG
SH^{N}-SEAP$^{\Delta SP}$	**MWIQQLLGLSSMSIR<u>WPGR</u>**
SEAP$^{\Delta SP}$	Sekretierte Alkaline Phosphatase ohne endogenes Signalpeptid

Abbildung 21: Übersicht shrew-1-Konstrukte. **A**: Aufbau der shrew-1 Signalsequenz. **Fett**: N-Domäne (Reste 1-19). **Unterstrichen**: Übergangsbereich (Reste 16-24). **Unveränderte Schrift**: C-Domäne (Reste 20-43). ▼: Signalpeptidase-Schnittstelle. „LG" shrew-1 Reste 44 und 45 zur Vervollständigung der Signalpeptidase-Schnittstelle. **B**: Übersicht der entworfenen Signalsequenz-Konstrukte. Alle Konstrukte wurden N-terminal an die SEAP ohne ihr endogenes Signal (SEAP$^{\Delta SP}$) fusioniert. C-terminal sind alle Konstrukte mit dem Myc-*tag* fusioniert. Entnommen und übersetzt aus Hiss *et al.*, 2008b.

Entsprechend dem NtraC-Modell wurde im Signalpeptid von shrew-1 ein Übergangsbereich (tra) von Position 16-24 (WPGRPLGSH) detektiert (Abb. 21). Der Übergangsbereich besteht hier aus drei in der Sequenz überlappenden potentiellen β-Turns. Dabei stellt WPGR nach Vorhersage-Score den Turn mit dem höchsten SVM-Score dar. Zur Unterstützung der Vorhersage hinsichtlich der Position des Übergangsbereiches wurden die Signalpeptide orthologer shrew-1-Sequenzen aus *Macaca mulatta*, *Canis familaris*, *Bos taurus*, *Mus*

musculus, *Rattus norvegicus* und *Danio rerio* betrachtet. Hierzu wurde mit Hilfe von CLUSTALW (Larkin *et al*., 2007) ein multiples Alignment durchgeführt (Abb. 22).

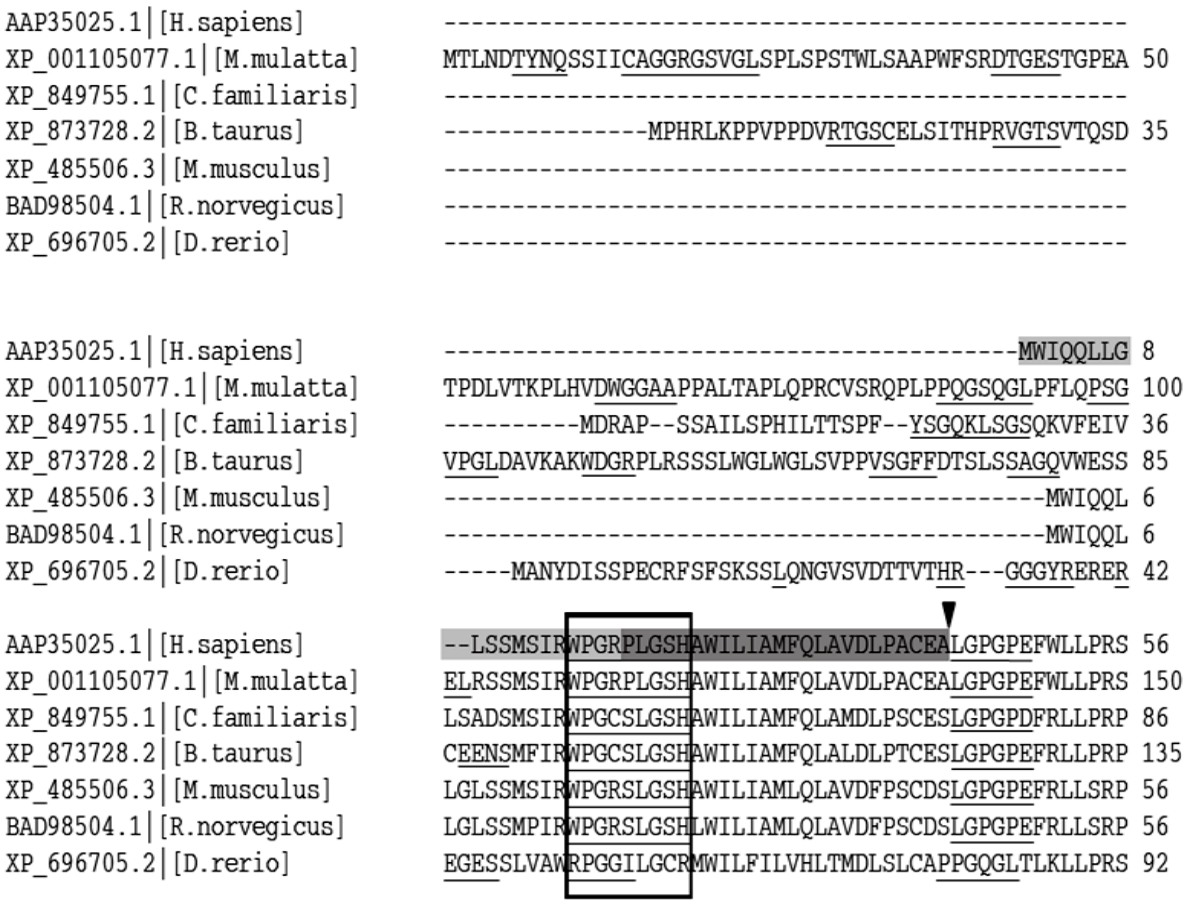


Abbildung 22: Multiples Alignment der Signalpeptide von shrew-1 aus den Spezies *Homo Sapiens*, *Macaca mulatta*, *Canis familaris*, *Bos taurus*, *Mus musculus*, *Rattus norvegicus* und *Danio rerio,* durchgeführt mit ClustallW Version 1.83. Entnommen aus Hiss *et al*., 2008b. **Unterstrichen**: potentiell β-Turn formierende Bereiche. **Schwarzer Kasten**: Übergangsbereich nach NtraC-Modell. ▼: Signalpeptidase-Schnittstelle. **Hellgrau hinterlegt**: N-Domäne. **Dunkelgrau hinterlegt**: C-Domäne nach NtraC-Modell.

In Abb. 22 sind die vorhergesagten potentiellen β-Turns „WPGRPLGSH“ durch einen schwarzen Kasten markiert. Der Übergangsbereich ist hier der β-Turn mit dem höchsten Score „WPGR“. Für die Zerlegung in N-und C-Domäne wurde nur dieser erste Turn „WPGR“ verwendet. In Abb. 22 ist zu erkennen, dass der erste dominante β-Turn in allen Spezies in seiner Funktion als „β-Turn“ konserviert ist. Nur in diesem Bereich ist in allen sieben Spezies an gleicher Position innerhalb der Signalpeptid-Sequenz ein potentieller β-Turn zu erkennen.

Dies ist ein starker Hinweis auf einen evolutionären Druck, dieses potentiell Sekundärstruktur-Element zu erhalten und unterstützt die Vorhersage für einen β-Turn in diesem Bereich. Basierend auf diesen Befunden wurde in der Signalpeptid-Sequenz von shrew-1 in *H.sapiens* der Bereich „WPGR“ (16-19) als Übergangsbereich festgelegt.
Der N-terminal gelegene Bereich inklusive des Übergangsbereiches bildet die N-Domäne (Position 1-24; Abb. 22, hellgrau hinterlegt). Die N-Domäne wird von TargetP 1.1 als potentielles mitochondriales Targeting-Peptid (*Score* = 0,3) vorhergesagt.
Der C-terminal gelegene Teil bis zur Signalpeptidase-Schnittstelle bildet die C-Domäne (Position 25-43; Abb. 22, dunkelgrau hinterlegt). Die C-Domäne wurde von SignalP 3.0 als Signalpeptid (Wahrscheinlichkeit = 0,99) vorhergesagt.

Entsprechend der Vorhersage des NtraC-Modells wurden die Konstrukte für shrew-1 (Abb. 21, B) entworfen und *in vitro* hinsichtlich ihrer Targeting-Kapazität getestet.

Experimentelle Ergebnisse von shrew-1

Die ER-Targeting-Kapazität der N- und C-Domäne von shrew-1 wurde mithilfe eines SEAP-Assays (Kapitel „SEAP-Assay“) erfasst. Die Ergebnisse sind in Abb. 23 und 24 dargestellt. Im SEAP-Assay, ausgehend vom Zell-Lysat (Abb. 23), ist zu erkennen, dass das Absorptions-Niveau der C-Domäne nach 5 min ist geringer als das der beiden Wildtyp-Signalpeptide nach 5 min. Das Absorptions-Niveau der C-Domäne des Signalpeptides nach 30 min ist vergleichbar mit dem des SEAP-Wildtyp-Signalpeptid und dem des shrew-1-Wildtyp-Signalpeptid nach 30 min. Das Absorptions-Niveau der N-Domäne nach 5 min und 30 min ist mit dem des Leervektor vergleichbar.
Somit verhalten sich beide Konstrukte (N-Domäne und C-Domäne) entsprechend der Vorhersage des NtraC-Modells: Die N-Domäne fungiert nicht als ER-Targeting-Signal, die C-Domäne fungiert als ER-Targeting-Signal.

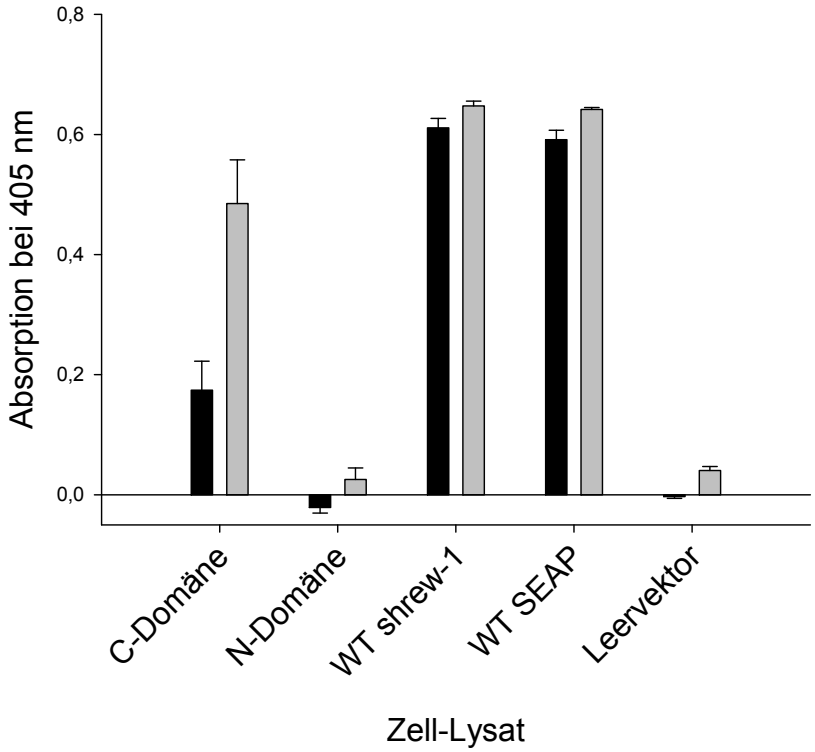


Abbildung 23: Ergebnisse des SEAP-Assays für das Zell-Lysat (*N* = 4). **Schwarz**: Absorption nach 5 min Inkubation mit SEAP-Substrat. **Grau**: Absorption nach 30 min Inkubation mit SEAP-Substrat. Fehlerbalken repräsentieren den Standardfehler.

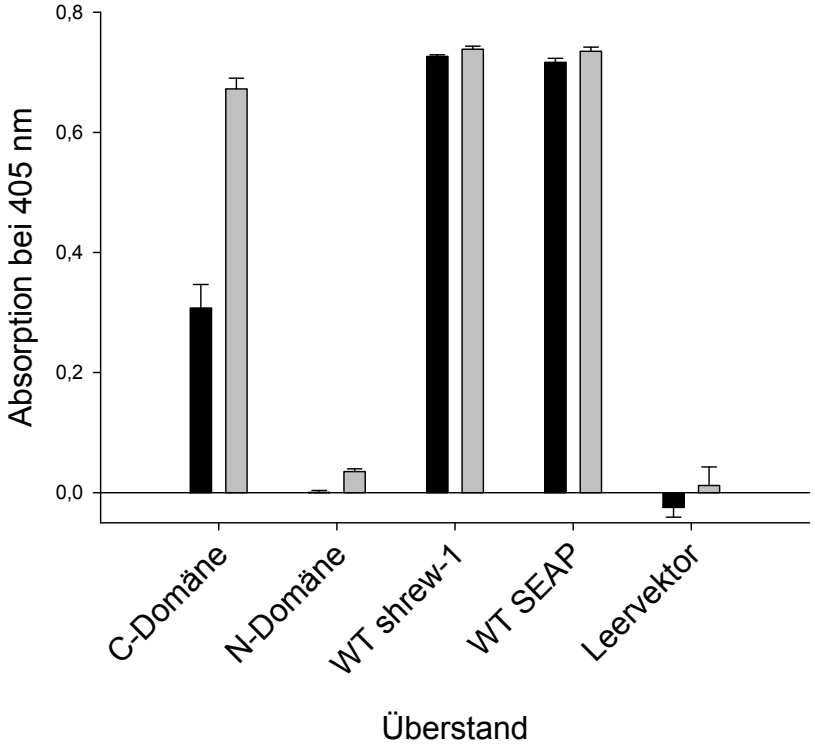


Abbildung 24: Ergebnisse des SEAP-Assays für den Überstand (*N* = 4). **Schwarz**: Absorption nach 5 min Inkubation mit SEAP-Substrat. **Grau**: Absorption nach 30 min Inkubation mit SEAP-Substrat. Fehlerbalken repräsentieren den Standardfehler.

Der SEAP-Assay, basierend auf dem Überstand (Abb. 24), ist konform zum SEAP-Assay, basierend auf dem Zell-Lysat (Abb. 23). Die C-Domäne ist in der Lage, die SEAP zum ER zu dirigieren, ist aber im Vergleich zum Wildtyp-Signalpeptid von shrew-1 quantitativ schwächer. Die geringere ER-Targeting-Kapazität der C-Domäne alleine im Vergleich zum Wildtyp-shrew-1-Signalpeptid schlägt sich auch in einer geringeren Sekretion, messbar im Überstand, nieder. Der einzige Unterschied zwischen dem Konstrukt mit der C-Domäne und dem shrew-1-Wildtyp-Signalpeptid ist die Abwesenheit der Aminosäuren 1-23. Die schwächere Sekretion ist daher auf eine schwächere ER-Targeting-Kapazität der C-Domäne ohne N-Domäne zurückzuführen. Der SEAP-Assay des Überstands bestätigt des Weiteren die Ergebnisse des Zell-Lysates für die N-Domäne. Die N-Domäne ist offensichtlich nicht in der Lage, die SEAP zum ER zu dirigieren; das Konstrukt mit der N-Domäne wird nicht sezerniert.

Die vorliegenden experimentellen Daten werden wie folgt interpretiert:

1) Die alleinige Anwesenheit der C-Domäne ist hinreichend, um das SEAP-Konstrukt zum ER zu dirigieren.

2) Die C-Domäne stellt ein - quantitativ - ausgedrückt schwächeres ER-Targeting-Signal als das Wildtyp-Signalpeptid von shrew-1 und des Wildtyp-Signalpeptid der SEAP dar.

3) Die N-Domäne ist *nicht* in der Lage, das SEAP-Konstrukt zum ER zu dirigieren.

Die Vorhersagen für die SP-Targeting-Kapazität der shrew-1 Signalpeptid-Domänen, basierend auf dem NtraC-Modell, wurden damit durch das Experiment eindeutig bestätigt.

Die experimentellen Ergebnisse zu Vorhersage über die mTP-Targeting-Kapazität der Domänen folgen im Abschnitt „Mitochondriales Targeting“.

Mitochondriales Targeting

Die Targeting-Kapazität der N-Domäne wurde *in silico* als mitochondriales Targeting-Peptid vorhergesagt. Die C-Domäne wird nicht als mTP vorhergesagt. Die Überprüfung dieser Hypothese erfolgte durch eine biochemische Aufreinigung von Mitochondrien (Abb. 25). Die Konstrukte werden dabei in einem Western-Blot durch einen Anti-Myc-Antikörper (Material und Methoden) detektiert.

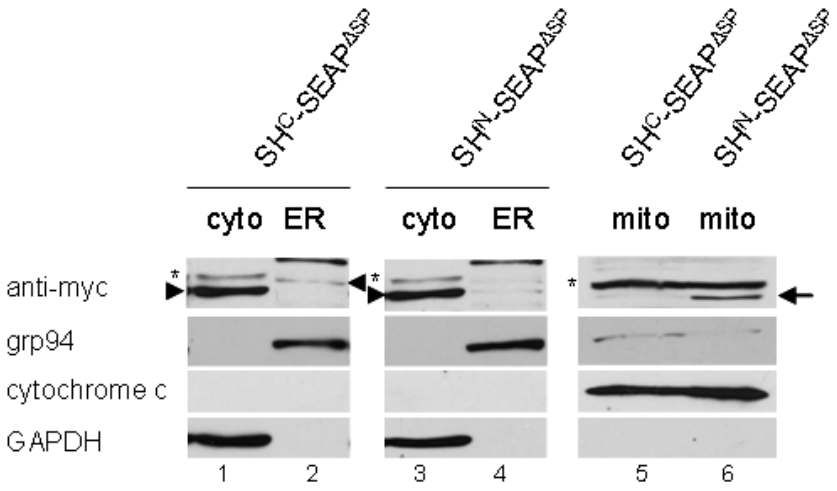


Abbildung 25: Western-Blot der biochemischen Aufreinigung von Mitochondrien. **$^{\Delta SP}$:** Alle Konstrukte sind jeweils mit der SEAP ohne ihr endogenes Signal und mit einem Myc-*tag* zu verstehen. **SH^{C}**: C-Domäne von shrew-1. **SH^{N}**: N-Domäne von shrew-1. **anti-Myc**: Antikörper gegen den Myc-*tag*. **grp94**: ER-Lumen Protein Marker. **GAPDH**: Cytosol Marker. Cytochrom **C**: Mitochondrialer Marker. *: Unspezifische Bande. ► bzw. ➔ : markieren jeweils die Bande des Konstruktes.

Die Cytochrom-C Zeile (Abb. 25, Spalte 1-4) zeigt, dass keine Verunreinigung durch Mitochondrien vorliegt. Die unterschiedliche Größe der detektierten C-Domänen-Konstrukte (Abb. 25, Spalte 1 kleiner als Spalte 2) ist auf die Glykosylierung der SEAP im ER zurückzuführen. Trotz Abspaltung des Signalpeptides wird durch die Glykosylierung eine Größenzunahme des Konstruktes erreicht.

In Abbildung 25 ist in Spalte 1 und 2 zu erkennen, dass die C-Domäne im Cytosol stark und im ER schwach zu detektieren ist. Dies ist in Einklang mit dem SEAP-Assay (Abb. 23 und Abb. 24), der die C-Domäne als schwaches ER-Targeting-Signal zeigt. Spalte 5 zeigt ein sehr schwaches mitochondriales Signal für die C-Domäne, was im Vergleich zur Kontrolle grp94 folgenden Schluss zulässt: Die mitochondriale Fraktion war, entsprechend der Bandenstärke von grp94 (Abb. 25, Spalte 5 und 6), mit ER verunreinigt. Die sehr schwache Bande der C-Domäne in der mitochondrialen Fraktion (Abb. 25, Spalte 5) ist mit der sehr schwachen Bande von grp94 in der mitochondrialen Fraktion zu vergleichen. Das Auftreten der C-Domäne in der mitochondrialen Fraktion wird daher als Verunreinigung durch die ER-Fraktion aufgefasst. Eine mitochondriales Targeting für die C-Domäne wird nicht postuliert.

Eine Bande der N-Domäne ist im Cytosol (Abb. 25, Spalte 3), aber nicht im ER (Spalte 4) zu finden. Dies ist konform zum SEAP-Assay (Abb. 23 und Abb. 24), der für die N-Domäne keine

ER-Targeting-Kapazität zeigt. Entscheidend ist, dass die N-Domäne in der mitochondrialen Fraktion (Abb. 25, Spalte 6) zu detektieren ist. Dies ist als deutlicher Hinweis darauf zu werten, dass die N-Domäne eine mitochondriale Targeting-Kapazität besitzt.

Die experimentellen Ergebnisse der mitochondrialen Targeting-Kapazität für die N- und C-Domäne entsprechen somit der Vorhersage durch das NtraC-Modell.

Immunfluoreszenz-Aufnahmen

Zur zusätzlichen Analyse der mitochondrialen Lokalisation der N-Domäne von shrew-1 wurden Immunfluoreszenz-Aufnahmen in Hek 293T-Zellen hergestellt (Abb. 26).

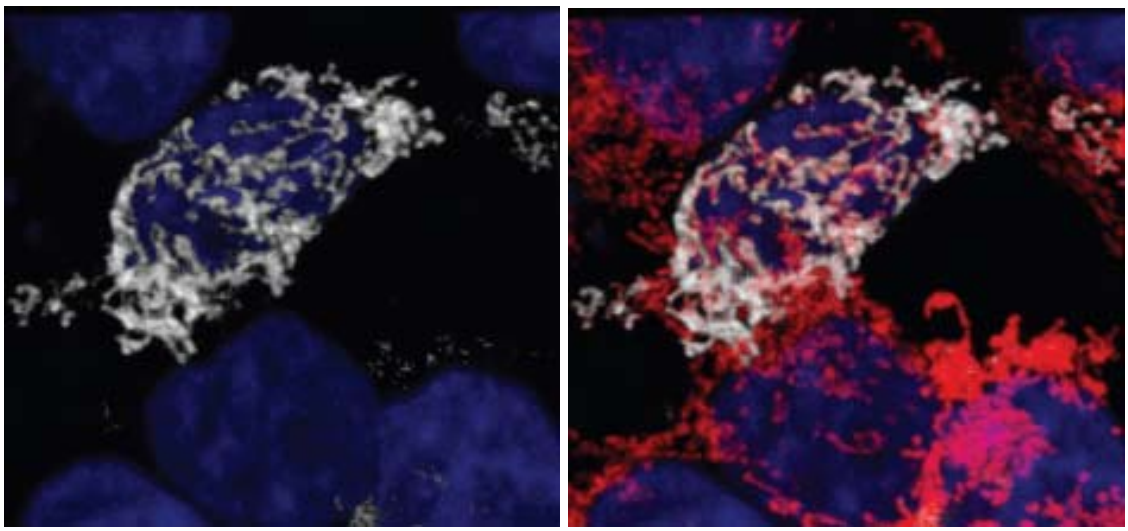


Abbildung 26: Immunfluoreszenz-Aufnahmen der N-Domäne von shrew-1 in Hek 293T-Zellen. Dargestellt ist jeweils dieselbe Zelle links und rechts. **Blau**: Zellkernfärbung mit DAPI. **Weiß**: Co-Lokalisation zwischen der N-Domäne von shrew-1 und Mitotracker. Rechts zusätzlich in **Rot**: Mitochondrien-Färbung mit Mitotracker angezeigt (Inkubation 1 Stunde).

In Abbildung 26 ist eine Co-Lokalisation zwischen dem Konstrukt mit der N-Domäne von shrew-1 und den durch Mitotracker gefärbten Mitochondrien erkennbar (Abb. 26, weiß). Zum Vergleich ist in Abb. 26 rechts in rot zusätzlich zur Co-Lokalisation die Färbung mit Mitotracker angezeigt. Die zu erkennende Co-Lokalisation zwischen der N-Domäne und dem Mitotracker ist als weiterer Hinweis zu werten, dass die N-Domäne von shrew-1 in der Lage ist, als mitochondriales Targeting-Peptid zu fungieren.

Von George *et al*. (1998) wurde beschrieben, dass die Entscheidung für einen Transport zu den Mitochondrien in Hefe früh in der Translation getroffen wird und abhängig ist von der Menge des anwesenden *nascent polypeptide-associated complex* (NAC). Dieser Aspekt

wurde im Versuchsaufbau nicht berücksichtigt und erlaubt daher nur eine qualitative, keine quantitative Aussage über die mitochondriale Targeting-Kapazität der N-Domäne.

Als weiteres Argument für die grundsätzliche mTP-Targeting-Kapazität der N-Domäne ist ihr Aufbau zu werten. Der Aufbau der N-Domäne folgt der in Schneider *et al.* (1998) beschriebenen Verteilung von Arginin-Resten. In dieser Arbeit wird ein loses Muster von Arginin in der -10, -3 und -2 Position von mTP beschrieben. In der erweiterten N-Domäne von shrew-1 (Positionen 1-24) ist ein Arginin an Position -6 und -10 zu finden, welches zur Erkennung als mTP beitragen könnte.

Zusammenfassend ist für das lange Signalpeptid von shrew-1 festzuhalten, dass die experimentellen Ergebnisse die *in silico* gemachten Vorhersagen, basierend auf dem NtraC-Modell, eindeutig bestätigt wurden:

1) Die N-Domäne des shrew-1-Signalpeptides (Positionen 1-19) kann *in vitro* als mitochondriales Targeting-Signal fungieren.
2) Die C-Domäne (Positionen 20-43) kann *in vitro* als ER-Targeting-Signal und als Signal für den sekretorischen Pathway fungieren.
3) Die ER-Targeting-Kapazität des Wildtyp-Signalpeptides bleibt von der NtraC-Domänen-Architektur unberührt.
4) Die C-Domäne alleine ist entsprechend der Vorhersage ein funktionales, aber schwächeres Signalpeptid als das Wildtyp-shrew-1-Signalpeptid.

Die Ergebnisse wurden publiziert (Hiss *et al*., 2008b). Das lange Signalpeptid von shrew-1 (43 Aminosäuren) ist damit das erste experimentell validierte Beispiel für ein dem NtraC-Modell entsprechend organisiertes Signalpeptid.

Weiterführende Überlegungen

Die hier gezeigte Existenz zweier Domänen mit unterschiedlicher Targeting-Kapazität ist als Erweiterung zu dem bisher berichteten alleinigen ER-Targeting von shrew-1 zu sehen. Dies könnte ein Hinweis auf eine Regulation sein, in deren Rahmen die alternative Targeting-Funktion nur unter bestimmten Bedingungen angesprochen wird. Eine Erkennung der N-Domäne (mTP) oder der C-Domäne (SP) könnte unter bestimmten Bedingungen favorisiert werden. Alternativ könnte die Existenz der N-Domäne auch eine Regulation der ER-Targeting-Effizienz darstellen. Ein Hinweis hierauf ist die schwächere ER-Targeting-Effizienz der C-Domäne alleine im Vergleich zum Wildtyp-shrew-1 Signalpeptid.

Des Weiteren ist die 3D-Strukturuntersuchung des Signalpeptides von shrew-1 geplant. Eine erste *in silico*-Vorhersage mit Hilfe des Rosetta-Algorithmus ist in Abb. 27 dargestellt. Es ist zu erkennen, dass der Übergangsbereich (Abb. 27, rot) die N-Domäne und die C-Domäne nicht nur räumlich, sondern auch hinsichtlich der Sekundärstruktur-Eigenschaft trennt. In der N-Domäne werden zwei anti-parallele β-Stränge vorhergesagt. Typischerweise haben mitochondriale Signale eine alpha-helikale Struktur (von Heijne, 1986a; Roise *et al*., 1988; Emanuelsson *et al*., 2001), welche die geladenen Reste auf eine Seite für die Interaktion mit Tom20 und Tom22 bringt (Mokranjac und Neupert, 2007). Dies könnte hier durch die β-Faltblätter erreicht werden. Die geordnete Struktur der N-Domäne hebt sich von der ungeordneten Struktur der C-Domäne ab (Abb. 27). Der Bereich mit Sekundärstruktur-Elementen (N-Domäne) ist durch den Übergangsbereich von dem Bereich ohne Sekundärstruktur-Elemente (C-Domäne) getrennt. Eine experimentelle Strukturauflösung wird in Kooperation mit Dipl.-Chemikerin Kerstin Schmöe aus dem Arbeitskreis von Prof. Dr. Volker Dötsch (JWG –Universität, Frankfurt/Main) bearbeitet.

Die NtraC-Organisation könnte somit über den Sekundärstruktur-Aspekt eines β-Turns im Übergangsbereich strukturell unterschiedliche Bereiche innerhalb des Signalpeptides voneinander trennen.

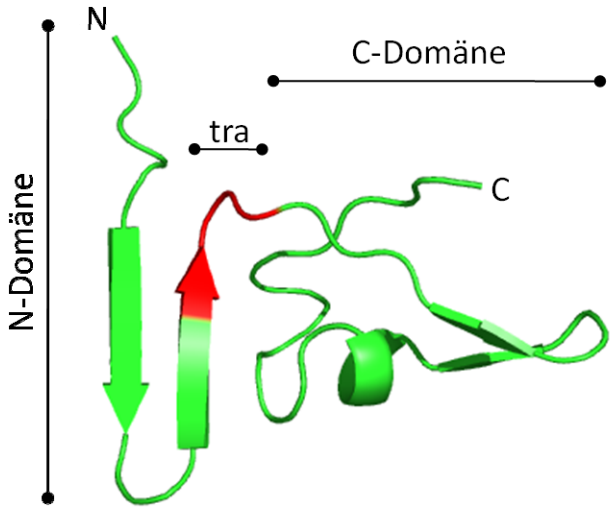


Abbildung 27: Mit dem Rosetta-Algorithmus vorhergesagte 3D-Struktur des shrew-1-Wildtyp-Signalpeptides. Gezeigt sind die Aminosäuren 1-43. **Rot**: Übergangsbereich Aminosäuren „WPGR“. **N**: N-Terminus. **C**: C-Terminus.**N-Domäne, tra, C-Domäne**: Unterteilung nach NtraC-Modell.

DCBD2-Protein

Die Discoidin-Proteine sind eine Protein-Familie aus Phospholipid-bindenden Lectinen und Proteinen, die in Zell-Zell-Adhäsion involviert sind (Baumgartner *et al*., 1998; Kane und Davie, 1988). DCBD2 steht für „Discoidin, CUB and LCCL domain-containing protein 2“.

Alle im Folgenden beschriebenen Konstrukte des Signalpeptides von DCBD2 wurden im Rahmen dieser Arbeit vom Autor *in silico* entworfen. Die zellbiologischen Untersuchungen hinsichtlich der Lokalisation wurden unter Anleitung von Eduard Resch und Dr. Alexander Schreiner in dem Labor von Prof. Dr. Anna Starzinski-Powitz vom Autor selbst durchgeführt.

Die CUB-Domäne ist eine extrazelluläre Domäne von ca. 110 Resten Länge, die vornehmlich bei in der Entwicklung regulierten Proteinen auftritt. Die Struktur wird als *beta-barrel*-ähnlich zu Immunglobulinen vorhergesagt (Bork und Beckmann, 1993; Bork, 1991). Die LCCL-Domäne ist nach den am besten charakterisierten Proteinen benannt, in denen sie gefunden wurde: Limulus Faktor C, Coch-5b2 und Lgl1 (Trexler *et al*., 2000). Die Domäne umfasst ein ca. 100 Reste langes Motiv, das ein konserviertes Histidin (YxxxSxxCxAAVHxGVI) enthält. Die LCCL-Domäne wird als selbständig faltende Domäne angesehen, die bei den genannten Domäne-organisierten Proteinen durch Exon-Neukombination integriert wurde. Sie besteht nach Sekundärstruktur-Vorhersagen aus sechs β-Strängen und zwei α-Helices (Trexler *et al*., 2000). Es wird vermutet, dass die LCCL-Domäne in das Binden von Lipopolysacchariden involviert ist (Trexler *et al*., 2000; Robertson *et al*., 1998). Die LCCL-Domäne tritt ebenfalls bei extrazellulären Proteinen von Parasiten des Stamms Apicomplexa gehäuft auf, z.B. *Plasmodium falciparum* (Dessens *et al*., 2004). Der Name des DCBD2-Gens ist dcbld2, synonym verwendet wird clcp1 und esdn (Koshikawa *et al*., 2002; Kobuke *et al*., 2001). Das DCBD2-Protein stammt aus Endothel- und glattem Muskelgewebe und ist als neuropin-ähnlich klassifiziert. Es ist ein einzelspänniges Typ-I Transmembranprotein, das nach Gefäßverletzungen verstärkt synthetisiert wird. Es wurde ebenfalls im Zusammenhang mit metastasierendem Lungenkrebs beobachtet und besitzt eine experimentell bestätigte Signalsequenz mit 66 Aminosäuren (Kobuke *et al*., 2001). Im Jahr der Publikation der Signalsequenz (2001) wurde die DCBD2-Signalsequenz (66 Aminosäuren) als die längste bekannte eukaryotische sekretorische Signalsequenz bezeichnet. Die Sequenz von DCBD2 ist in Abbildung 28 gegeben, und die Signalsequenz (66 Aminosäuren) ist unterstrichen.

>gi|54792129|ref|NP_563615.3| discoidin, CUB and LCCL domain containing 2 [Homo sapiens]
MASRAVVRARRCPQCPQVRAAAAAPAWAALPLSRSLPPCSNSSSFSMPLFLLLLLVLLLLLEDAGAQQGDGCGHT
VLGPESGTLTSINYPQTYPNSTVCEWEIRVKMGERVRIKFGDFDIEDSDSCHFNYLRIYNGIGVSRTEIGKYCGL
GLQMNHSIESKGNEITLLFMSGIHVSGRGFLASYSVIDKQDLITCLDTASNFLEPEFSKYCPAGCLLPFAEISGT
IPHGYRDSSPLCMAGVHAGVVSNTLGGQISVVISKGIPYYESSLANNVTSVVGHLSTSLFTFKTSGCYGTLGMES
GVIADPQITASSVLEWTDHTGQENSWKPKKARLKKPGPPWAAFATDEYQWLQIDLNKEKKITGIITTGSTMVEHN
YYVSAYRILYSDDGQKWTVYREPGVEQDKIFQGNKDYHQDVRNNFLPPIIARFIRVNPTQWQQKIAMKMELLGCQ
FIPKGRPPKLTQPPPPRNSNDLKNTTAPPKIAKGRAPKFTQPLQPRSSNEFPAQTEQTTASPDIRNTTVTPNVTK
DVALAAVLVPVLVMVLTTLILILVCAWHWRNRKKKTEGTYDLPYWDRAGWWKGMKQFLPAKAVDHEETPVRYSSS
EVNHLSPREVTTVLQADSAEYAQPLVGGIVGTLHQRSTFKPEEGKEAGYADLDPYNSPGQEVYHAYAEPLPITGP
EYATPIIMDMSGHPTTSVGQPSTSTFKATGNQPPPLVGTYNTLLSRTDSCSSAQAQYDTPKAGKPGLPAPDELVY
QVPQSTQEVSGAGRDGECDVFKEIL

Abbildung 28: Aminosäure-Sequenz von DCBD2 im FASTA-Format. Unterstrichen: Signalpeptid. Entnommen aus der NCBI Datenbank (NP_563615).

Im Folgenden ist die Signalsequenz einzeln dargestellt. Vorhergesagte β-Turns (Länge: 4 Aminosäuren) sind unterstrichen, der Übergangsbereich „CSNS“ und die Domänen nach NtraC-Modell sind angezeigt:

MASRAVVRARRCPQCPQVRAAAAAPAWAALPLSRSLPPCSNSSSFSMPLFLLLLLVLLLLLEDAGAQQ

N-Domäne — tra — C-Domäne

In Abbildung 29 ist ein Alignment mit orthologen DCBD2-Sequenzen von *H. sapiens*, *M.musculus* und *R.norvegicus* dargestellt. Zusätzlich wurden zwei durch NCBI automatisch vorhergesagte Sequenzen von *Pan troglodytes* und *Bos taurus* mit einbezogen. Für diese Sequenzen ist ein annotierter Gen-Eintrag vorhanden; das Protein und dessen funktionelle Ähnlichkeit zu DCBD2 wurde vom NCBI vorhergesagt.

Basierend auf der NtraC-Vorhersage und dem multiplen Alignment (Abb. 29) wurde in DCBD2 der Übergangsbereich (tra) auf die Aminosäuren „CSNS“ (Position 39-42) festgelegt. In Abbildung 29 ist zu erkennen, dass der Übergangsbereich (tra) in allen Sequenzen funktionell konserviert ist. Dies ist ein Hinweis für die funktionelle Bedeutung dieses Bereichs.

```
CLUSTAL 2.0.8 multiple sequence alignment

SeqA Name                          Len(aa)  SeqB Name                          Len(aa)  Score
================================================================================================
1    UniProtKB|Q96PD2|DCBD2_HUMAN   775     2    UniProtKB|Q91ZV3|DCBD2_MOUSE 769      84
1    UniProtKB|Q96PD2|DCBD2_HUMAN   775     3    UniProtKB|Q91ZV2|DCBD2_RAT   769      85
1    UniProtKB|Q96PD2|DCBD2_HUMAN   775     4    gi|114588144|P.troglodytes   775      99
1    UniProtKB|Q96PD2|DCBD2_HUMAN   775     5    gi|119879152|B.taurus        770      91
2    UniProtKB|Q91ZV3|DCBD2_MOUSE   769     3    UniProtKB|Q91ZV2|DCBD2_RAT   769      92
2    UniProtKB|Q91ZV3|DCBD2_MOUSE   769     4    gi|114588144|P.troglodytes   775      84
2    UniProtKB|Q91ZV3|DCBD2_MOUSE   769     5    gi|119879152|B.taurus        770      83
3    UniProtKB|Q91ZV2|DCBD2_RAT     769     4    gi|114588144|P.troglodytes   775      85
3    UniProtKB|Q91ZV2|DCBD2_RAT     769     5    gi|119879152|B.taurus        770      83
4    gi|114588144|P.troglodytes     775     5    gi|119879152|B.taurus        770      90
================================================================================================

UniProtKB|Q96PD2|DCBD2_HUMAN      MASRAVVRARRCPQCPQVRAAAAAPAWAALPLSRSLPP----CSNSSSFS   46
UniProtKB|Q91ZV3|DCBD2_MOUSE      MASRAPLRAARSPQGPGGPAAPAATGRAALPSAGCCPLP--PGRNSSSRP   48
UniProtKB|Q91ZV2|DCBD2_RAT        MASRAPLRAARSPQDPGGRAAPAATGRAPLPSAGWCPLP--PGRNSSSRP   48
XP_001141399|P.troglodytes        MASRAVVRARRCPQCPQVRAAAAAPAWAALPLSRSLPP----CSNSSSFS   46
XP_602937|B.taurus                MASRAVVRAGHSPQRFLVRAAVAAPARAAFPLSRSYPLPFRSNSSSTSFP   50
                                  *****  **   **     ** **   *  *     *        * *

UniProtKB|Q96PD2|DCBD2_HUMAN      MPLFLLLLLVLLLLLEDAGA▼QQGDGCGHTVLGPESGTLTSINYPQTYPNS   96
UniProtKB|Q91ZV3|DCBD2_MOUSE      R-----LLLLLLLLLQDAGG▼QQGDGCGHTVLGPESGTLTSINYPHTYPNS   93
UniProtKB|Q91ZV2|DCBD2_RAT        R-----LLLLLLLLLPDAGA▼QKGDGCGHTVLGPESGTLTSINYPHTYPNS   93
XP_001141399|P.troglodytes        MPLFLLLLLVLLLLLEDAGA▼QQGDGCGHTVLGPESGTLTSINYPQTYPNS   96
XP_602937|B.taurus                RPLFLLLLLILLLLLEDAGA▼QQGDGCGHTVLGPESGTLTSINYPHTYPNS  100
                                   *** ***** *** * *** ******************** *****
```

Abbildung 29: Multiples Alignment der DCBD2-Sequenzen von *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Pan troglodytes* und *Bos taurus.* Dargestellt sind jeweils die ersten 96 (*H.sapiens, P.troglodytes*) bzw. 93 (*M.musculus, R.norvegicus*) bzw. 100 (*B.taurus*) Aminosäuren, die jeweils die Signalsequenz enthalten. **Unterstrichen**: Übergangsbereich (tra) nach NtraC-Modell. ▼: Signalpeptidase-Schnittstelle (bei *P.troglodytes* und *B.taurus* durch SignalP vorhergesagt). *: Identische Aminosäuren.

Die Einteilung des Signalpeptides von DCBD2 entsprechend dem NtraC-Modell in eine N- und eine C-Domäne ist im Folgenden dargestellt:

N-Domäne (1-42) :
MASRAVVRARRCPQCPQVRAAAAAPAWAALPLSRSLPPCSNS

C-Domäne (43-66/68), inklusive Signalpeptidase-Schnittstelle GA QQ :
SSFSMPLFLLLLLVLLLLLEDAGA QQ

Die N-Domäne wird dabei als mitochondriales Targeting-Peptid durch TargetP vorhergesagt (*Score* 0,86). Die C-Domäne alleine wird von SignalP als Signalpeptid mit einer Wahrscheinlichkeit von 0,72 vorhergesagt. Macht man eine Vorhersage für die C-Domäne in ihrem nativen Proteinkontext, also mit den Aminosäuren 43-733, erhält man eine Signalpeptid-Wahrscheinlichkeit von 1.

Bei dem orthologen DCBD2-Protein in *B. taurus* zerlegt der Übergangsbereich, abweichend zum humanen DCBD2, die Signalsequenz in zwei ER-Targeting-Signale. Sowohl die in der N-Domäne als auch die in der C-Domäne liegende Signalsequenz besitzt eine eigene Signalpeptidase-Schnittstelle. Aufbauend auf dieser Vorhersage wurden die folgenden Konstrukte entworfen, um die vorhergesagte Targeting-Funktion der Domänen zu untersuchen:

N-Domäne :
MASRAVVRARRCPQCPQVRAAAAAPAWAALPLSRSLPPCSNS + SEAP$^{\Delta SP}$ + Myc-tag

C-Domäne :
SSFSMPLFLLLLLVLLLLLEDAGAQQ + SEAP$^{\Delta SP}$ + Myc-tag

Wildtyp-DCBD2-Signalpeptid :
MASRAVVRARRCPQCPQVRAAAAAPAWAALPLSRSLPP<u>CSNS</u>SSFSMPLFLLLLLVLLLLLED
AGAQQ + SEAP$^{\Delta SP}$ + Myc-tag

wobei:
+ SEAP$^{\Delta SP}$: *secreated alkaline phosphatase* ohne deren endogenes Signalpeptid.
+ Myc-*tag* : Myc-*tag* für die Antikörper-Detektion.

Die entworfenen Konstrukte wurden mithilfe eines SEAP-Assays (Material und Methoden, SEAP-Assay) hinsichtlich ihrer ER-Targeting-Kapazität getestet. Die mitochondriale Targeting-Kapazität wurde durch eine biochemische Aufreinigung von Mitochondrien und die Detektion des Konstrukts mithilfe von Antikörpern in den entsprechenden Fraktionen erreicht.

Experimentelle Ergebnisse für DCBD2

Die ER-Targeting-Kapazität der N- und C-Domäne von DCBD2 und des Wildtyp-Signalpeptides wurde mithilfe eines SEAP-Assays erfasst. Die Ergebnisse sind in Abbildung 30 und 31 dargestellt.

Das Absorptionsniveau der C-Domäne bei 405 nM ist sowohl nach 5 min als auch nach 30 min im Zell-Lysat mit dem Niveau des Wildtyp-SEAP-Signalpeptid (Positiv-Kontrolle) und dem des DCBD2 Wildtyp-Signalpeptid vergleichbar (Abb. 30). Dies ist als klarer Hinweis zu werten, dass die vorgeschlagene C-Domäne von DCBD2 entsprechend der *in silico*-Vorhersage die Funktion einer ER-Targeting-Sequenz erfüllen kann.

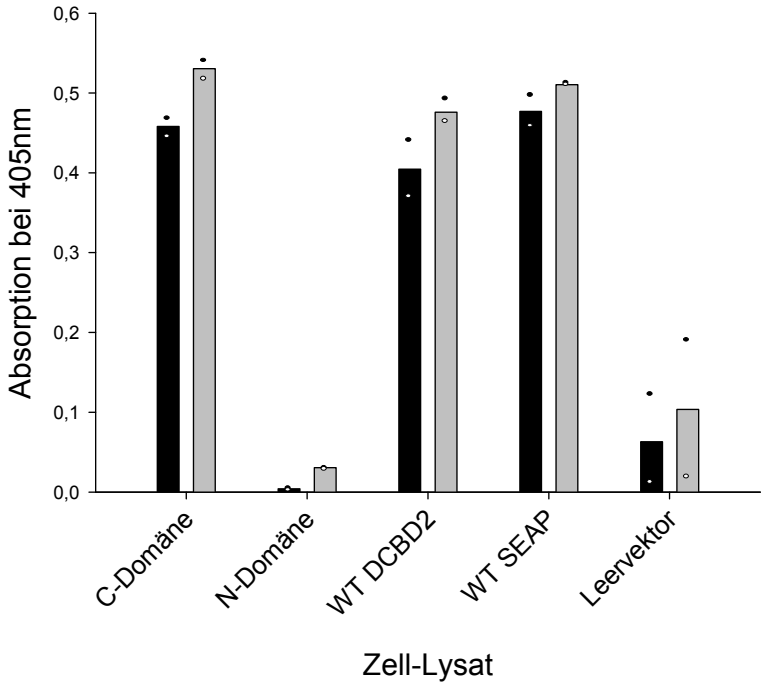


Abbildung 30: Ergebnisse des SEAP-Assays für das Zell-Lysat (*N* = 2). **Schwarz**: Absorption nach 5 min Inkubation mit SEAP-Substrat. **Grau**: Absorption nach 30 min Inkubation mit SEAP-Substrat. **Punkt**: Einzelner Messwert.

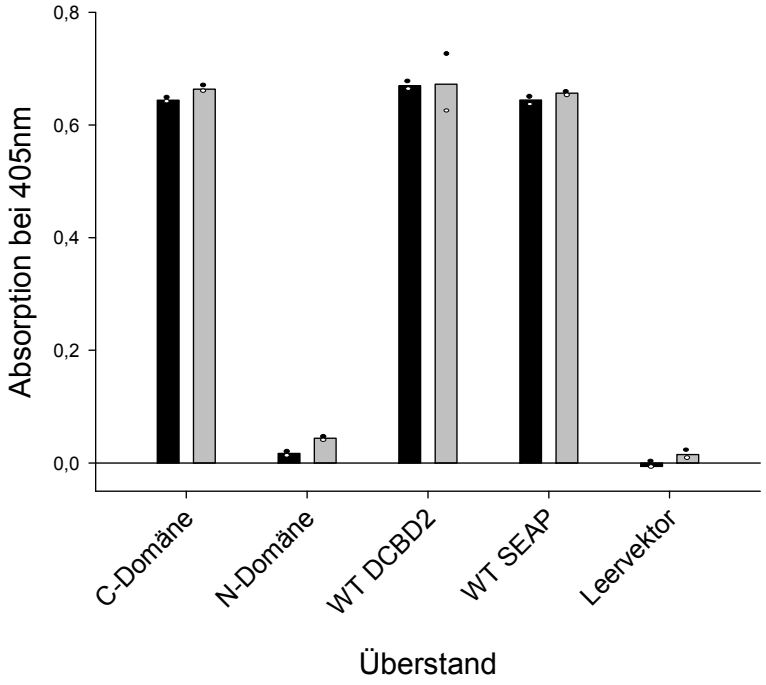


Abbildung 31: Ergebnisse des SEAP-Assays für den Überstand (N = 2). **Schwarz**: Absorption nach 5 min Inkubation mit SEAP-Substrat. **Grau**: Absorption nach 30 min Inkubation mit SEAP-Substrat. **Punkt**: Einzelner Messwert.

Die N-Domäne zeigt ein Absorptionsniveau bei 405 nM von 0,005, schwächer als der Leervektor mit 0,06. Eine ER-Targeting-Funktion wird für die N-Domäne von DCBD2 basierend auf diesen Ergebnissen verneint (Abb. 30 und 31). Die DCBD2 Wildtyp-Signalpeptid ER-Targeting-Kapazität ist mit der des Wildtyp-SEAP-Signalpeptides (Positiv-Kontrolle) vergleichbar. Im Unterschied zu shrew-1 ist die DCBD2 C-Domäne alleine ein effizientes ER-Targeting-Signal, vergleichbar mit dem DCBD2 Wildtyp-Signalpeptid. Die shrew-1 C-Domäne hingegen war alleine ein weniger effizientes ER-Targeting-Signal als das shrew-1 Wildtyp-Signalpeptid (Abb. 23 und 24). Der Grund für die unterschiedliche ER-Targeting-Effizienz kann aufgrund der hier durchgeführten Versuche nicht hergeleitet werden. Der Unterschied in der ER-Targeting-Kapazität ist nicht im *Score* der SignalP Vorhersage erkennbar: shrew-1 C-Domäne (*Score* = 0,9), DCBD2 C-Domäne (*Score* = 0,7).
Die Absorptionsniveaus im Überstand entsprechen insgesamt (Schwankungen maximal ± 0,2) denen im Zell-Lysat. Die C-Domäne von DCBD2 besitzt eine ER-Targeting-Kapazität vergleichbar mit dem Wildtyp-Signalpeptid von DCBD2 und dem Wildtyp-Signalpeptid der SEAP.
Die experimentellen Ergebnisse entsprechen und bestätigen somit die Vorhersagen hinsichtlich der ER-Targeting-Kapazitäten der N- und C-Domäne von DCBD2.

Mitochondriales Targeting

Die mitochondriale Targeting-Fähigkeit der N- und C-Domäne wurde mithilfe einer biochemischen Aufreinigung der Mitochondrien überprüft. Die Ergebnisse sind als Western-Blot in Abbildung 32 dargestellt. Der Western-Blot zeigt, dass die C-Domäne sowohl im Cytosol (Abb. 32, Spalte 2) als auch im Verhältnis dazu stark angereichert im ER zu finden ist (Abb. 32, Spalte 1, Pfeilspitze). Dies ist konform mit den Ergebnissen des SEAP-Assays, dass die C-Domäne ER-Targeting-Kapazitäten besitzt. Die N-Domäne ist im Vergleich dazu im ER minimal (Abb. 32, Spalte 4) und im Cytosol schwach zu finden (Abb. 32, Spalte 3, Pfeilspitze). Dies ist ebenfalls konform zur Aussage des SEAP-Assays, dass die N-Domäne keine ER-Targeting-Kapazitäten besitzt. Die schwache cytosolische Fraktion entsteht durch die translation der Konstrukte im Cytosol. Die mitochondriale Targeting-Kapazität der N-Domäne zeigt sich in Abb. 32, Spalte 6.

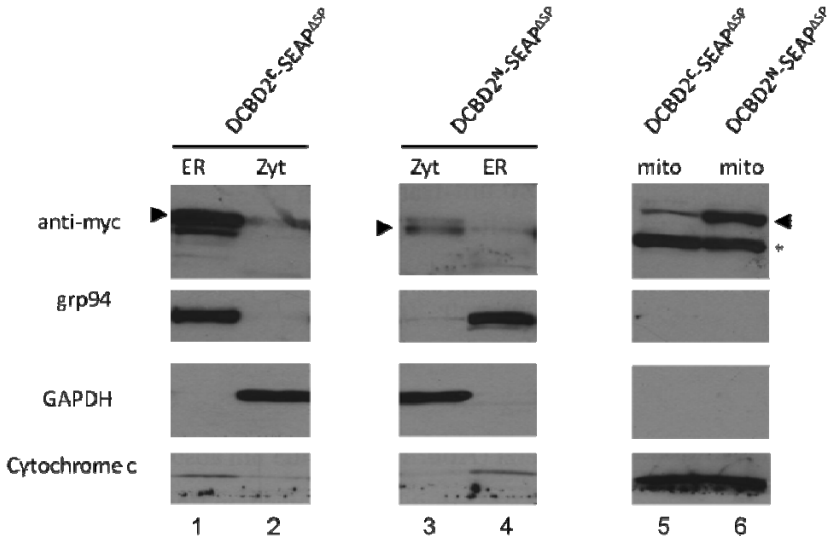


Abbildung 32: Western-Blot der biochemischen Aufreinigung von Mitochondrien. Alle Konstrukte sind jeweils mit der SEAP ohne ihr endogenes Signal ($^{\Delta SP}$) und mit einem Myc-*tag* zu verstehen. **DCDBD2^{C}**: C-Domäne von DCBD2. **DCBD2^{N}**: N-Domäne von DCBD2. **anti-Myc**: Antikörper gegen den Myc-*tag*. **grp94**: ER-Lumen Protein-Marker. **GAPDH**: Cytosol Marker. **Cytochrom-C**: Mitochondrialer Marker. *: unspezifische Bande. ►: markiert jeweils die Bande des Konstrukts. **ER**: Endoplasmatisches Retikulum. **Zyt**: Cytosol.

Die durch die Pfeilspitze markierte N-Domänen-Bande (Abb. 32, Spalte 6) ist in ihrer Stärke vergleichbar zur Bande von Cytochrom-C (natives mitochondriales Protein). Sie ist im Vergleich zur C-Domäne deutlich angereichert. Das sehr deutliche mitochondriale Signal könnte auf eine Besonderheit in der Sequenz der N-Domäne zurückzuführen sein: In Anlehnung zu dem in Schneider *et al*., 1998 beschriebenen losen Arginin-Muster in mTP (-10,-3,-2) hat die N-Domäne von DCBD2 ein Arginin an Position -9. Ein zu Schneider *et al*. (1998) passendes Muster ist *upstream* des Übergangsbereiches zu finden. Wird Position 20 (Alanin) innerhalb der N-Domäne als „-1“ Position betrachtet, befinden sich drei Arginine in -2, -10 und -11 Position. Wird Position 12 (Cystein) als „-1“ Position betrachtet befinden sich Arginine in -2,-3,-5 und -9 Position. Dieses für das mitochondriale Targeting wichtige Muster (-10, -3 , -2 Position mit Argininen besetzt) wird also je nach Definition der „-1“ Position mehrfach in der langen N-Domäne von DCBD2 erfüllt. Dies könnte ein Grund für die sehr starke mitochondriale Lokalisation der N-Domäne in den Assays darstellen. Wir vermuten daher der Tom20/Tom22-Rezeptor in der äußeren mitochondrialen Membran hat die Möglichkeit, mehrere Muster innerhalb der Sequenz alternativ zu erkennen.

Des Weiteren liegen zwei Arginine (Position 11 und 19) im Abstand von acht Aminosäuren in der N-Domäne von DCBD2 vor. Diese ergeben, wird Position 20 als „-1“ Position betrachtet,

ein -2 und -10 Muster von Argininen. Dieses -2, -10 Muster ist nach Isaya *et al.,* 1991 und Isaya und Kalousek (1995) ein Hinweis auf die sequentielle Prozessierung durch die MPP (*matrix processing peptidase*) und die MIP (*mitochondrial intermediate peptidase*). Die N-Domäne von DCBD2 könnte somit ein mitochondriales Matrix-Signal darstellen. Die experimentelle Überprüfung der Vorhersage für die Lokalisation in der mitochondrialen Matrix ist aus den vorliegenden Versuchen nicht ableitbar. Aufgrund der vorliegenden Ergebnisse wird postuliert, dass die N-Domäne von DCBD2 ein generelles mitochondriales Targeting-Signal ist. Die Kapazität ist mit der eines nativen mitochondrialen Signals (Cytochrom-C) vergleichbar. Die durch den Autor vorhergesagte mitochondriale Targeting-Kapazität der N-Domäne wurde somit *in vitro* bestätigt.

Für das Glycoprotein gp120 des HI-Virus wurde in früheren Studien gezeigt, dass die geladenen Reste in der Signalsequenz von gp120 (30 Aminosäuren) einen negativen Effekt auf die ER-Targeting-Effizienz haben (Li *et al*., 1996). Dies wurde auf eine reduzierte Abspaltung des Signalpeptides durch die Existenz der geladenen Reste zurückgeführt. Im Fall von DCDBD2 haben die positiven Ladungen der N-Domäne jedoch keinen Einfluss auf die ER-Targeting-Effizienz, was sich im Vergleich der Targeting-Effizienz der C-Domäne alleine und dem Wildtyp-Signalpeptid zeigt (Abb. 30 und 31). Ein PNGase-F Verdau (Lottspeich und Zorbas, 1998; Kapitel „PNGase-F Verdau), der N-Glykosylierung entfernt, zeigte jedoch eine cytosolische Spezies des Wildtyp-Signalpeptides, die nicht N-glykosyliert (Abb. 33, Spalte 8, Sternchen) ist und die bei dem C-Domänen-Konstrukt nicht vorhanden ist (Abb. 33, Spalte 4). Zwei mögliche Ursachen für die Existenz der cytosolischen Spezies sind:

- eine reduzierte ER-Targeting-Kapazität, die aber durch die erreichte Sättigung im SEAP-Assay nicht aufgelöst werden konnte und nur im Western-Blot erkennbar ist.

- Es handelt sich um eine mitochondriale Spezies des Wildtyp-Signalpeptides von DCBD2.

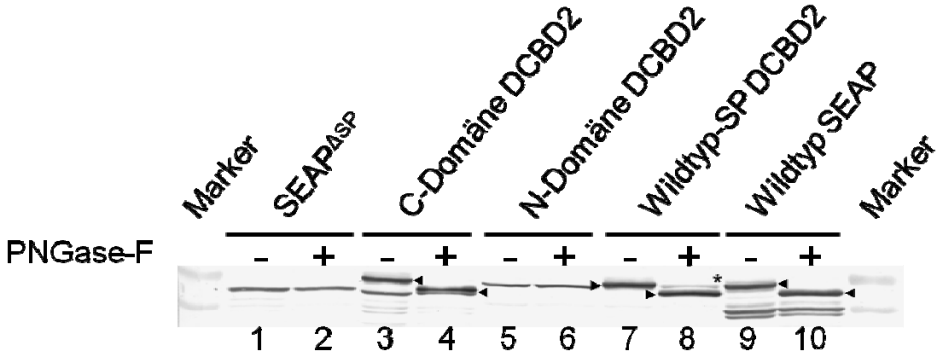


Abbildung 33: PNGase-F Verdau der DCBD2-Konstrukte. **-/+**: ohne bzw. mit PNGase-F. **Marker**: Fermentas PageRuler. **Pfeilspitzen**: Konstrukte ohne und mit PNGase-F Verdau. ***: N**icht glykosylierte Spezies des Wildtyp-SP von DCBD2.

Zur zusätzlichen experimentellen Untermauerung der vorhergesagten Lokalisationen der Domänen wurden Immunfluoreszenzbilder in MCF7-Zellen angefertigt (Abb. 34a-d). In Abb. 34a ist zu erkennen, dass eine der beiden im Fokus liegenden Zellen erfolgreich mit dem Konstrukt (DCBD2 N-Domäne) transfiziert werden konnte. Im Vergleich dazu ist in Abb. 34b zu erkennen, dass beide Zellen erfolgreich mit mito-gfp transfiziert wurden. In Abb. 34c und 34d ist eine Co-Lokalisation zwischen der mito-gfp-Färbung und der Anti-Myc-Färbung des Konstrukts in weiß zu erkennen. Die Immunfluoreszenz-Aufnahmen stimmen somit mit den in der biochemischen Aufreinigung erhaltenen Daten überein und werden als ein weiterer Hinweis für die mitochondriale Lokalisation der DCBD2 N-Domäne gewertet.

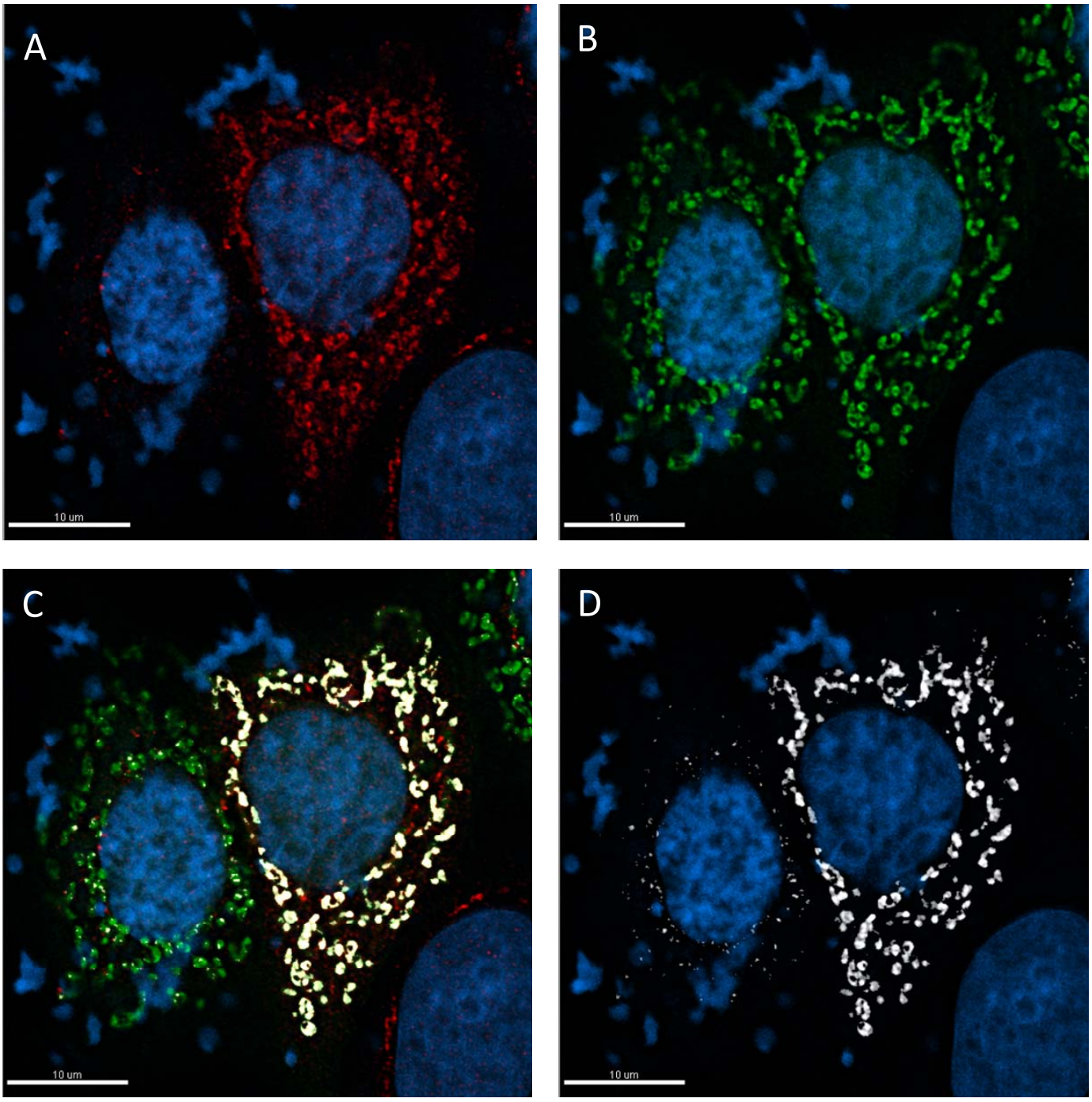


Abbildung 34: Immunfluoreszenz-Aufnahmen in MCF7-Zellen. Maßstab jeweils 10 µm. Jeweils in **Blau** Zellkernfärbung mit DAPI. **A:** Färbung mit Alexa 594, Myc-*tag* der DCBD2 N-Domäne (**Rot**). **B:** Mitochondrien mit mito-gfp gefärbt (**Grün**). **C:** Überlagerung von A und B und Co-Lokalisation zwischen mito-gfp und Alexa 594 (**Weiß**). **D:** Co-Lokalisation zwischen mito-gfp und Alexa 594 (**Weiß**).

Zusammenfassend ist für das lange Signalpeptid von DCBD2 festzuhalten, dass die experimentellen Ergebnisse die *in silico* gemachten Vorhersagen, basierend auf dem NtraC-Modell, eindeutig bestätigen:

1) Die N-Domäne des DCBD2-Signalpeptides (Positionen 1-42) kann *in vitro* als mitochondriales Targeting-Signal fungieren.
2) Die N-Domäne von DCBD2 ist hinsichtlich ihrer mitochondrialen Targeting-Effizienz mit dem bekannten mTP von Cytochrom-C vergleichbar.
3) Die C-Domäne (Positionen 43-66) kann *in vitro* als ER-Targeting-Signal und als Signal für den sekretorischen Pathway fungieren.
4) Die ER-Targeting-Kapazität des Wildtyp-Signalpeptides bleibt von der NtraC-Domänen-Architektur unberührt.
5) Die C-Domäne alleine ist entsprechend der Vorhersage ein funktionales und hinsichtlich seiner Effizienz ein mit dem Wildtyp-DCBD2-Signalpeptid vergleichbares ER-Targeting-Signal.

DCBD2 stellt damit ein zweites *in vitro* validiertes Beispiel für die NtraC-Organisation langer Vertebrata-Signalpeptide dar.

Weiterführende Betrachtungen zu DCBD2

Es wurden Vorhersagen der Sekundärstruktur des humanen und des DCBD2-Signalpeptides aus *R.norvegicus* unter Verwendung des Programmes Rosetta (Bystroff *et al.*, 2000; Bystroff und Shao, 2002) durchgeführt (Abb. 35 a und b).

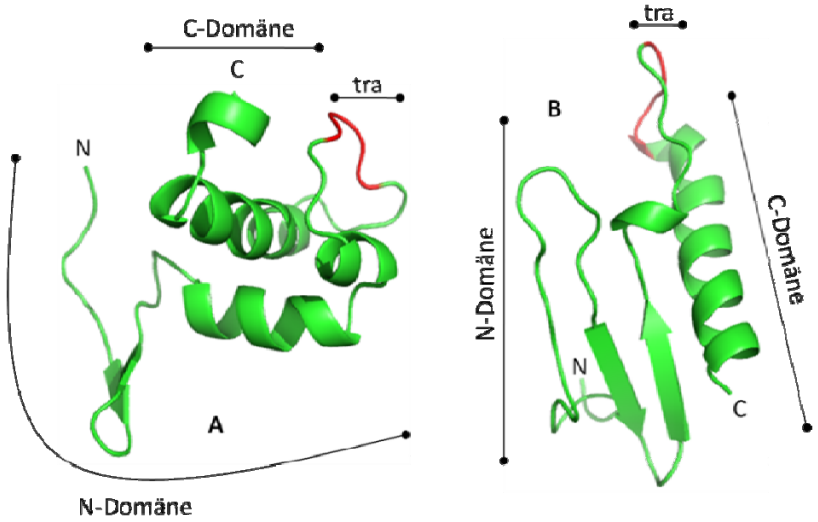


Abbildung 35: Sekundärstruktur-Vorhersagen mit dem Rosetta Algorithmus. **A**: DCBD2 Signalpeptid aus *H.sapiens.* **B**: DCBD2 Signalpeptid aus *R.norvegicus*; in **rot** jeweils die Übergangsbereiche. **N**: N-Terminus. **C**: C-Terminus. N-Domäne, C-Domäne, tra: Einteilung nach NtraC-Modell.

Die C-Domäne wird sowohl im DCBD2-Signalpeptid von *R.norvegicus* als auch im humanen DCBD2 als α-Helix vorhergesagt. Dies ist im Unterschied zu shrew-1 zu sehen, wo die C-Domäne als ungeordnet vorhergesagt wurde. Eine Sekundärstruktur-Ähnlichkeit zwischen DCBD2 und shrew-1 ist in der N-Domäne zu erkennen. Die N-Domänen von DCBD2 in Mensch und Ratte sowie die N-Domäne in shrew-1 enthalten zwei anti-parallele β-Stränge. Hier könnte ein Zusammenhang zwischen Struktur (anti-parallele β-Stränge) und Funktion der N-Domäne (mitochondriales Targeting-Peptid) bestehen.

RGMA-Protein

Die RGM-Familie (*repulsive guidance molecule*) stellt eine Familie membranständiger Proteine dar, die bei dem Wachstum und der Ausrichtung von Axonen eine Rolle spielen. RGM,A' wurde als Überlebensfaktor für Axone in *G.gallus*-Embryonen identifiziert und hat Einfluss auf die Differenzierung sich entwickelnder Neuronen (Monnier *et al.,* 2002; Matsunaga *et al*., 2004; Matsunaga *et al*., 2006). RGMA besitzt ein 45 Aminosäuren, RGMB ein 47 Aminosäuren und RGMC ein 35 Aminosäuren langes Signalpeptid (Camus und Lambert, 2007). Die Signalpeptide sind jeweils als putativ annotiert. Ein multiples Alignment der Signalpeptide ist in Abbildung 36 gezeigt.

```
CLUSTAL 2.0.8 multiple sequence alignment

SeqA Name                       Len(aa)  SeqB Name                       Len(aa)  Score
================================================================================
1    UniProtKB|Q96B86|RGMA_HUMAN   450      2    UniProtKB|Q6NW40|RGMB_HUMAN   437      47
1    UniProtKB|Q96B86|RGMA_HUMAN   450      3    UniProtKB|Q6ZVN8|RGMC_HUMAN   426      47
2    UniProtKB|Q6NW40|RGMB_HUMAN   437      3    UniProtKB|Q6ZVN8|RGMC_HUMAN   426      41
================================================================================


UniProtKB|Q96B86|RGMA_HUMAN      MQPPRERLVVTGRAGWMGMGRGAGRS--ALGFWPTLAFLLCSFPA ----A 44
UniProtKB|Q6NW40|RGMB_HUMAN      MGLRAAPSSAAAAAAEVEQRRRPGLCPPPLELLLLLLFSLGLLHA▼GDCQQ 50
UniProtKB|Q6ZVN8|RGMC_HUMAN      ----------MGEPGQSPSPRSSHGSPPTLSTLTLLLLLCGHAHS▼----- 35
                                             *       *     *

UniProtKB|Q96B86|RGMA_HUMAN      TSP▼CKILKCNSEFWSATSGSHAPASD---------------DTPEFCAAL 79
UniProtKB|Q6NW40|RGMB_HUMAN      PAQ CRIQKCTTDFVSLTSHLNSAVDG---------------FDSEFCKAL 85
UniProtKB|Q6ZVN8|RGMC_HUMAN      --Q CKILRCNAEYVSSTLSLRGGGSSGALRGGGGGGRGGGVGSGGLCRAL 83
                                     * *  *     * *                            * **
```

Abbildung 36: Multiples Alignment der Sequenzen der RGM Familie von *Homo sapiens* (RGMA, RGMB, RGMC). Dargestellt sind jeweils die ersten 79, 85 bzw. 83 Aminosäuren, die jeweils die Signalsequenz enthalten. **Unterstrichen**: Übergangsbereich (tra) nach NtraC-Modell. ▼: Signalpeptidase-Schnittstelle nach UniProtKB. *: Identische Aminosäuren.

Die unterstrichenen Bereiche stellen jeweils den vorhergesagten Übergangsbereich dar (Abb. 36). Es ist zu erkennen, dass in allen drei Vertretern der Familie an vergleichbarer Position ein Übergangsbereich zu finden ist. Die C-Domäne des Signalpeptides aller drei Proteine (RGMA, RGMB, RGMC) wird als ER-Targeting-Signal (*Score* = 0,99, 0,7 und 1,0) vorhergesagt. Im Falle von RGMB und RGMC wird für die N-Domäne ein mTP mit einem *Score* von 0,39 bzw. 0,38 durch TargetP vorhergesagt. Dieser mTP-*Score* ist mit dem der N-Domäne von shrew-1 (*Score* = 0,3) vergleichbar. Da RGMA mit 0,42 den höchsten mTP-*Score* erhält, wurde es für den Test im Rahmen dieser Arbeit ausgewählt.
Entsprechend dem NtraC-Modell wurden für das Signalpeptid von RGMA die folgenden Konstrukte vom Autor *in silico* entworfen und in Kooperation mit Dipl.-Biologe Eduard Resch *in vitro* getestet:

N-Domäne (1-26):
MQPPRERLVVTGRAGWMGMGRGAGRS + SEAP$^{\Delta SP}$ + Myc-*tag*

C-Domäne (27-47):
ALGFWPTLAFLLCSFPAATSP + SEAP$^{\Delta SP}$ + Myc-*tag*

Wildtyp-RGMA-Signalpeptid (1-47) :
MQPPRERLVVTGRAGWMGMGRGAGRSALGFWPTLAFLLCSFPAATSP + SEAP$^{\Delta SP}$ + Myc-*tag*

wobei:
+ SEAP$^{\Delta SP}$: *secreated alkaline phosphatase* ohne deren endogenes Signalpeptid
+ Myc-*tag* : Myc-*tag* für die Antikörper Detektion

Die ER-Targeting-Kapazität der drei Konstrukte wurde mithilfe eines SEAP-Assay, angelehnt an Berger *et al*. (1988), überprüft. Die Ergebnisse sind in Abbildung 37 für das Zell-Lysat und in Abbildung 38 für den Überstand dargestellt.
Die C-Domäne von RGMA zeigt im Zell-Lysat im Vergleich zum Wildtpy-SEAP-Signalpeptid ein schwaches, aber im Vergleich zum Leervektor eindeutiges Signal (Abb. 37, C-Domäne). Die ER-Targeting- Kapazität der C-Domäne zeigt sich auch im Überstand, wo die C-Domäne 72% der Signalstärke des Wildtyp-SEAP-Signalpeptides nach 30 min erreicht. Die stärkere Zunahme der Absorptionsrate im Überstand im Vergleich zum Zell-Lysat spricht für eine starke Sekretion des Konstruktes. Die N-Domäne zeigt dabei auch im Überstand ein Verhalten vergleichbar zum Leervektor (Negativ-Kontrolle). Eine ER-Targeting-Kapazität der N-Domäne kann nicht festgestellt werden.

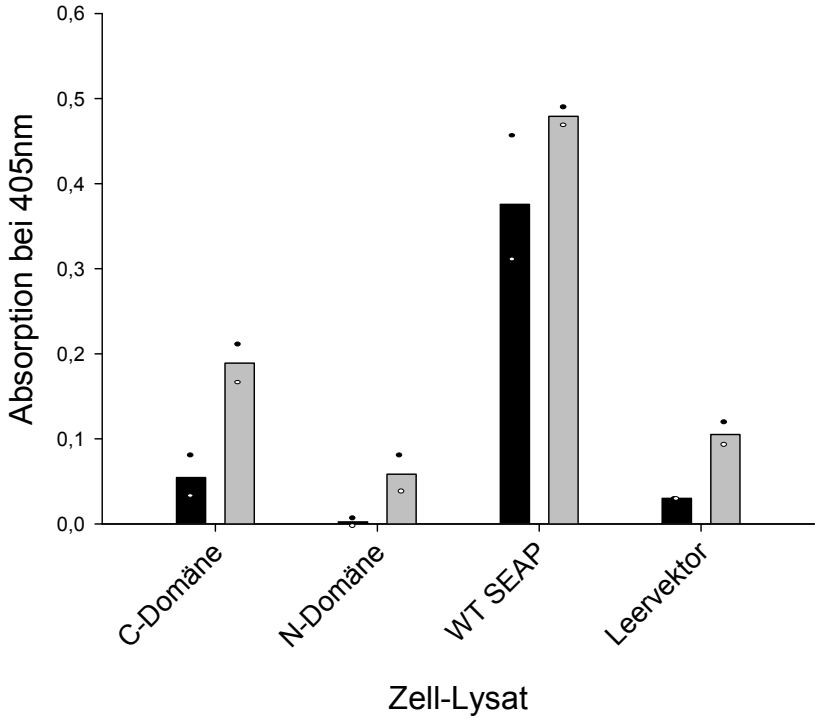


Abbildung 37:. Ergebnisse des SEAP-Assays für das Zell-Lysat der RGMA Konstrukte (*N* = 2). **Schwarz**: Absorption nach 5 min Inkubation mit SEAP-Substrat. **Grau**: Absorption nach 30 min Inkubation mit SEAP-Substrat.

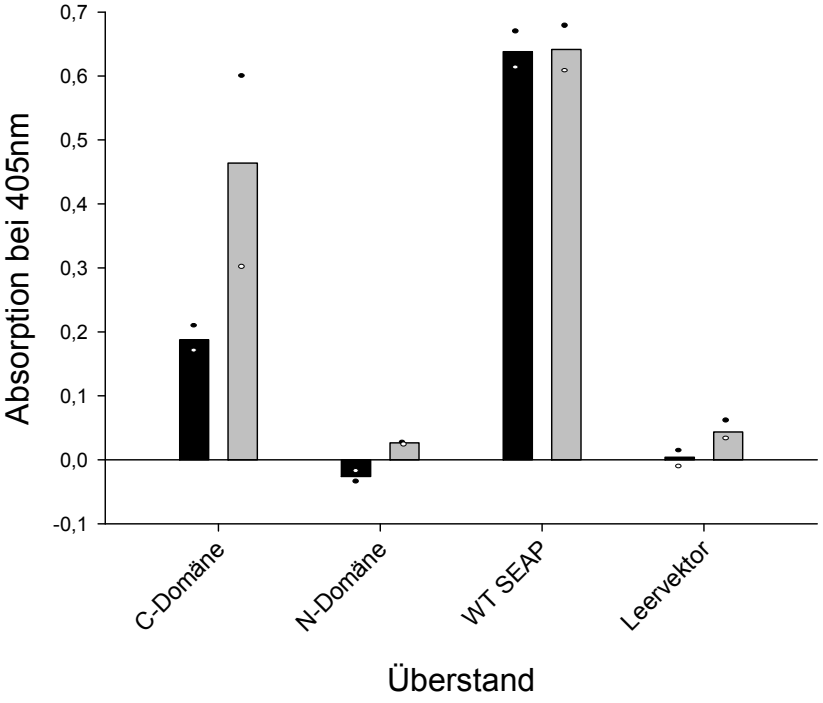


Abbildung 38: Ergebnisse des SEAP-Assays für den Überstand RGMA Konstrukte (*N* = 2). **Schwarz**: Absorption nach 5 min Inkubation mit SEAP-Substrat. **Grau**: Absorption nach 30 min Inkubation mit SEAP-Substrat.

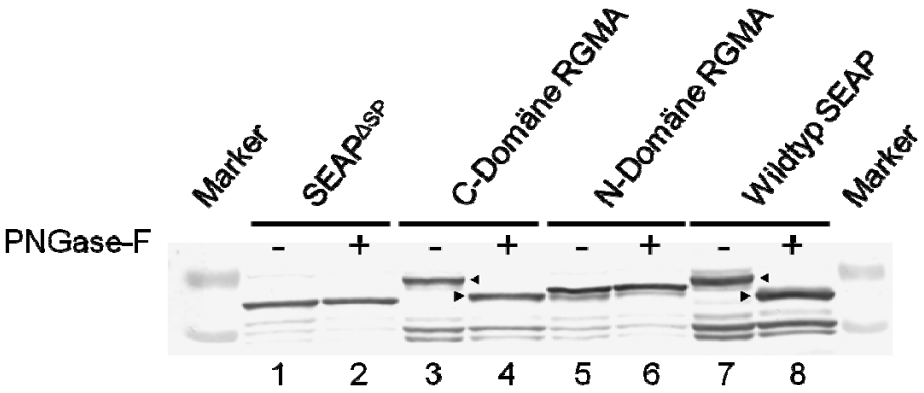


Abbildung 39: PNGase-F Verdau der RGMA Konstrukte. **-/+**: ohne bzw. mit PNGase-F. **Marker**: Fermentas PageRuler. **Pfeilspitzen**: Konstrukte ohne und mit PNGase-F Verdau.

Die postulierte ER-Targeting-Kapazität wird auch durch einen Western-Blot eines PNGase-F Verdau zur Entfernung von N-Glykosilierungen unterstützt (Abb. 39). In Abbildung 39 ist zu erkennen, dass die C-Domäne von RGMA eine Größenveränderung nach Zugabe von PNGase-F zeigt (Abb. 39, Spalte 3 und 4, Pfeilspitzen). Diese Größenveränderung ist auch bei dem Wildtyp-SEAP-Protein mit Signalpeptid (Positiv-Kontrolle) zu erkennen (Abb. 39, Spalte 7 und 8, Pfeilspitzen). Die SEAP ohne ihr Signalpeptid (Abb. 39, Spalte 1 und 2) sowie die N-Domäne von RGMA (Abb. 39, Spalte 5 und 6) zeigen keine Größenveränderung nach Zugabe von PNGase-F. Diese Ergebnisse werden als weiterer Hinweis gewertet, dass die C-Domäne von RGMA als ER-Targeting-Signal fungieren kann, die N-Domäne nicht.

Die *in vitro*-Ergebnisse bestätigen somit eindeutig die *in silico* gemachten Vorhersagen, basierend auf dem NtraC-Modell. Für das lange Signalpeptid von RGMA wird daher zusammenfassend festgehalten:

1) Die N-Domäne des RGMA-Signalpeptides (Positionen 1-26) besitzt keine ER-Targeting-Kapazität.
2) Über die mTP-Targeting-Kapazität der N-Domäne kann ohne eine biochemische Aufreinigung von Mitochondrien zum gegenwärtigen Zeitpunkt keine Aussage getroffen werden.
3) Die C-Domäne (Positionen 27-47) kann *in vitro* als ER-Targeting-Signal und als Signal für den sekretorischen Pathway fungieren.
4) Das Signal im Überstand nimmt stärker zu als im Zell-Lysat, der Grund ist aus dem Versuchsaufbau nicht abzuleiten.
5) Die C-Domäne alleine ist entsprechend der Vorhersage ein funktionales, im Vergleich zum Wildtyp SEAP-Signalpeptid aber schwächeres ER-Targeting-Signal.

RGMA stellt damit ein drittes *in vitro* validiertes Beispiel für ein NtraC-organisiertes langes Vertebrata-Signalpeptid dar.

Abschließende Betrachtungen zum NtraC-Modell

Die funktionale Bedeutung dieser wiederkehrenden NtraC-Organisation muss durch weitere Versuche untersucht werden. Die vorliegenden Daten legen einen Bezug zur Translokations- und Sekretionseffizienz nahe. Von Kim *et al*. (2002) wird ein Zusammenhang zwischen Signalpeptid-Sequenz und Translokations-Effizienz für Signalpeptide postuliert. Ein Zusammenhang zwischen Hydrophobizität und Translokations-Effizienz wurde bereits mehrfach gezeigt (Gierasch, 1989; Pugsley, 1989; Laforet und Kendall, 1991; Dalbey und von Heijne, 1992; Schneider *et al*., 1994). Eine Überprüfung im Zusammenhang mit der Länge der Signalpeptide findet in diesen Studien jedoch nicht statt und könnte weitere Zusammenhänge aufzeigen. Die Translokations-Effizienz der langen Signalpeptide könnte durch die Assoziation mit zusätzlichen Proteinen des Translokationsapparates (TRAM) und der N-Domäne erreicht werden. Hier würden Cross-Linking-Versuche z.B. zwischen dem TRAM-Protein und der Signalsequenz weiterführende Daten zur Verfügung stellen.

Ein weiterer Ansatz für die Untersuchung der funktionalen Bedeutung der NtraC-Organisation ist der Austausch NtraC-organisierter Signalpeptide im Wildtyp-Protein-Kontext. Sind Unterschiede in der Lokalisation und/oder Funktion des nativen Proteins zu beobachten, wenn ein NtraC-organisiertes langes Signalpeptid durch ein kurzes Signalpeptid ersetzt wird? Kann die Funktion durch ein alternatives NtraC-organisiertes Signalpeptid gerettet werden? Die RGM-Familie stellt ein bevorzugtes Beispiel für einen solchen Austausch dar, da der Austausch von Signalpeptiden zwischen RGMA-orthologen (*M.musculus*, *G.gallus*) und RGMA-paralogen (RGMB, RGMC) untersucht werden kann.

Die vorliegenden *in silico*-Daten in Kombination mit den durchgeführten mehrfachen *in vitro*-Experimenten sind als Hinweis zu werten, dass lange Signalpeptide in Erweiterung zu kurzen Signalpeptiden eine domänenartige Organisation aufweisen können. Diese NtraC-Domänen-Architektur ist in homologen und paralogen Sequenzen konserviert, was auf einen selektiven Druck zur Erhaltung hindeutet. Das NtraC-Modell bietet somit eine neue Betrachtungsweise für lange Signalpeptide. Diese ermöglicht es, ansonsten „kryptische“ Signale innerhalb der Signalsequenz zu identifizieren und der experimentellen Untersuchung zugänglich zu machen.

Ausblick

Alternative Leseraster in DCBD2 und RGMA

Weiterführend wurde untersucht, ob der domänenartige Aufbau von Signalpeptide in Zusammenhang mit alternativen Leserastern im Bereich der Signalsequenz steht.

Hierzu wurden die codierenden genomischen Nucleotid-Sequenzen der humanen Proteine DCBD2 und RGMA über Querverweise aus der UniProtKB extrahiert. Die Existenz aller zwei Proteine ist auf Protein-Level nachgewiesen.

Alternative Leseraster in RGMA

Bei RGMA handelt es sich um ein einzelspänniges Typ-I Transmembranprotein, das eine NtraC-Organisation in seinem Signalpeptid aufweist. Das multiple Alignment mit ClustalW (Abb. 40) zeigt eine Konservierung des Übergangsbereichs in den Spezies *Homo sapiens, Macaca fascicularis, Mus musculus* und *Gallus gallus* (Abb. 40, unterstrichen). Bei dem humanen RGMA liegen die Nukleotide für die Aminosäuren „MQPPRE“ in dem *open reading frame* 3, für die folgenden Aminosäuren bis „LCSFPA“ (Position -5 zur Signalpeptidas-Schnittstelle) im *open reading frame* 2 und die verbleibenden Aminosäuren bis zum Ende des Proteins wieder im *open reading frame* 3. Damit liegt der Kernbereich des Signalpeptides ohne die Signalpeptidase-Schnittstelle in einem alternativen Leseraster. Für die RGMA-Sequenz von *M. musculus* ist eine zweite Isoform, beginnend mit „MGM“, ohne experimentellen Beweis annotiert (VSP_011316; Abb.40, hellgrau). Für die Sequenz von *H. sapiens* ist ebenfalls eine zweite Isoform annotiert (VSP_022294), die in einem verschobenen Leseraster liegt. Dort ist „MQPP“ ersetzt durch „ARGRGGRSLPARCSRRRSEAALSSQRDLFSSPFFL NSSS“. Für die Existenz dieser Sequenz liegt kein experimenteller Beweis vor.
Des Weiteren enthalten alle N-Domänen der RGMA-Homologen diverse zusätzliche potentielle Turns (aus Gründen der Übersichtlichkeit nicht dargestellt), die C-Domänen hingegen nicht. Besonders zu beachten ist die Sequenz von *G. gallus*. Der Datenbankeintrag beginnt hier mit den Aminosäuren „MGM“. Eine Suche in der Originalpublikation der Sequenz (Monnier *et al.*, 2002) ermöglichte eine manuelle Translation des *upstream* gelegenen Bereichs (Abb. 40, dunkelgrau).

Die resultierende Sequenz ist in 9 von 16 Resten mit der humanen und der murinen Sequenz identisch. Verschiedene Erklärungen sind plausibel:

- In den Spezies wird jeweils ein alternativer Transkriptionsstart verwendet.
- Es findet ein alternatives Spleiß-Ereignis statt und/oder es findet eine Leserasterverschiebung statt.

Der Datenbankeintrag für RGMA aus *G. gallus* (RGMA_chick) stellt nach der Überzeugung des Autors eine Isoform dar, die durch ein verschobenes Leseraster entstanden ist. Die manuell translatierte Sequenz bei RGMA von *G.gallus* endet mit „NSSS“. Die annotierte zweite Isoform des RGMA aus *H.sapiens* ersetzt „MQPP“ mit einer auf „NSSS“ endenden Sequenz. Dies und die hohe Sequenzkonservierung des manuell translatierten, nicht annotierten Bereichs von RGMA bei *G.gallus* (Abb. 40, dunkelgrau) argumentiert für die Existenz einer zweiten Isoform bei *G.gallus,* vergleichbar mit der annotierten Sequenz der orthologen RGMA-Proteine. Das Signalpeptid von RGMA setzt sich somit aus Fragmenten auf verschiedenen Leserastern zusammen, die zu verschiedenen Isoformen führen. Die An- oder Abwesenheit dieser Signalpeptid-Fragmente bzw. deren möglicherweise durch alternatives Splicing entstandenen Isoformen unterliegen gegebenfalls einer Kontrolle und haben damit potentiell Einfluss auf die Targeting-Funktion.

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	RGMA_HUMAN	450	2	RGMA_MACFA	458	97
1	RGMA_HUMAN	450	3	RGMA_MOUSE	454	92
1	RGMA_HUMAN	450	4	RGMA_CHICK	432	79
2	RGMA_MACFA	458	3	RGMA_MOUSE	454	88
2	RGMA_MACFA	458	4	RGMA_CHICK	432	78
3	RGMA_MOUSE	454	4	RGMA_CHICK	432	79

```
RGMA_HUMAN MQPP--------RERLVVTGRAGWMGMGRGAGRSALGFWPTLAFLLCSFPAATSP▼CKILK 52
RGMA_MACFA MGGPGPRRAGTSRERLVVTGRAGWMGMGRGAGRSALGFWPTLAFLLCSFPAAT▼SPCKILK 60
RGMA_MOUSE MQPP--------RERLVVTGRAGWMGMGRGAGRSALGLWPTLAFLLCSFPAAISP▼CKILK 52
RGMA_CHICK NSSS--------RERIVVKARAGWMGMGRGAGSTALGLFQILPVFLCIFPPVTSP▼CKILK 34
                     *** **  ************  ***    *   ** **   *******

RGMA_HUMAN CNSEFWSATS-GSHAPASDDTPEFCAALRSYALCTRRTARTCRGDLAYHSAVHGIEDLMS 111
RGMA_MACFA CNSEFWSATS-GSHAPASDDTPEFCAALRSYALCTRRTARTCRGDLAYHSAVHGIEDLMS 119
RGMA_MOUSE CNSEFWSATSSGSHAPASDDVPEFCAALRTYALCTRRTARTCRGDLAYHSAVHGIEDLMS 112
RGMA_CHICK CNSEFWAATS-GSHHLGAEETPEFCTALRAYAHCTRRTARTCRGDLAYHSAVHGIDDLMV 93
           ****** *** ***       **** *** ** ********************** ***
```

Abbildung 40: Multiples Alignment durchgeführt mit ClustalW2 der RGMA-Sequenzen von *Homo sapiens, Macaca fascicularis, Mus musculus* und *Gallus gallus*. **Hellgrau**: Annotierter Start der alternativen Isoform bei *M.musculus*. **Dunkelgrau**: Nicht im Datenbankeintrag enthaltene manuelle, translatiert *upstream* des Starcodons liegende Sequenz bei *G.gallus*. **Unterstrichen**: Übergangsbereich (tra). ▼: Signalpeptidase-Schnittstelle. „*“: Konservierter Rest in allen Sequenzen.

Alternative Leseraster in DCBD2

Die Signalsequenz des humanen DCBD2 ist 66 Aminosäuren lang (Kobuke *et al*., 2001; Signalpeptidase-Schnittstelle: ▼).

MASRAVVRARRCPQCPQVRAAAAAPAWAALPLSRSLPPCSNSSSFSMPLFLLLLLVLL
LLLEDAGA▼QQG…

Bei DCBD2 liegt das Signalpeptid inklusive der Signalpeptidase-Schnittstelle „AGA QQG“ im *open reading frame* 1. Die verbleibenden Aminosäuren des Proteins liegen > 1000 bp entfernt auf dem *open reading frame* 2. Bei den gemachten *in silico*-Vorhersagen sowie bei dem Entwurf und der Testung der Konstrukte *in vitro* zeigte sich, dass die Signalpeptidase-Schnittstelle mit circa zwei bis drei Aminosäuren *downstream* umschrieben werden muss, um erkannt zu werden. Dies könnte im Zusammenhang mit DCBD2 darauf hindeuten, dass das Signalpeptid inklusive Signalpeptidase-Schnittstelle und drei Resten *downstream* der Signalpeptidase-Schnittstelle durch die Lage in einem alternativen Leseraster ein Modul darstellt. Dieses Signalpeptid-Modul könnte in einer regulierten Weise mit abgelesen werden, ist aber möglicherweise evolutionär gesehen nicht immer Teil des DCBD2-Proteins gewesen.

Ein weiterer Hinweis hierauf ist das Signalpeptid des hypothetischen Protein BAE73105 aus *Macaca fascicularis*. Das hypothetische Protein stellt eine Variante von DCBD2 dar, das N- und C-terminal verkürzt ist (Abb. 41). BAE73105 enthält C-terminal nur eine verkürzte Signalsequenz, die der C-Domäne bei DCBD2 (Abb. 41, grün) inklusive der Signalpeptidase-Schnittstelle entspricht (Abb. 41, Kasten). Des Weiteren ist BAE73105 analog zu DCBD2 aus einer CUB-Domäne (Abb. 41, hellgrau), einer LCCL-Domäne (Abb. 41, dunkelgrau) und einer FA58C-Domäne (Abb. 41, gelb) aufgebaut. Namensgebend für die FA58C-Domäne ist ihr Auftreten in den Gerinnungsfaktoren V und VIII, wo sie zweimal wiederholt C-terminal vorliegt. Im Falle von BAE73105 ist sie im Vergleich zu DCBD2 C-terminal halbiert (Abb. 41, gelb). Das Signalpeptid von BAE73105 ab den Aminosäuren „MPLF“ bis drei Reste *downstream* der Signalpeptidase-Schnittstelle „QQG“ liegt im Leseraster 1, der Rest des Proteins in Leseraster 2. Dies entspricht dem Ende des Signalpeptides von DCBD2. Weiterhin fällt auf, dass sowohl bei DCBD2 als auch bei BAE73105 nach „QQG“ in der genomischen Sequenz „E“ statt „D“ kommt. Die Asparaginsäure (D), die die korrekte Fortsetzung der Protein-Sequenz darstellt, liegt in einem anderen Leseraster. BAE73105 würde somit eine

Domäne des Signalpeptides auf einem alternativen Leseraster enthalten, DCBD2 beide. Der domänenartige Aufbau von DCBD2 und BAE73105 spiegelt sich somit auch in den Leserastern und in der NtraC-Domänen-Architektur des Signalpeptides wider.

Um eine Aussage über den Zusammenhang zwischen dem Verwenden alternativer Leseraster und dem Auftreten solcher Signalpeptid-Domänen-Fragmente bei homologen Proteinen treffen zu können, sind weiterführende systematische Analysen notwendig. Die vorliegenden Daten werden als Hinweis darauf gewertet, dass Signalpeptide vom nativen Protein unabhängige Instanzen bereits auf genomischer Ebene darstellen können.

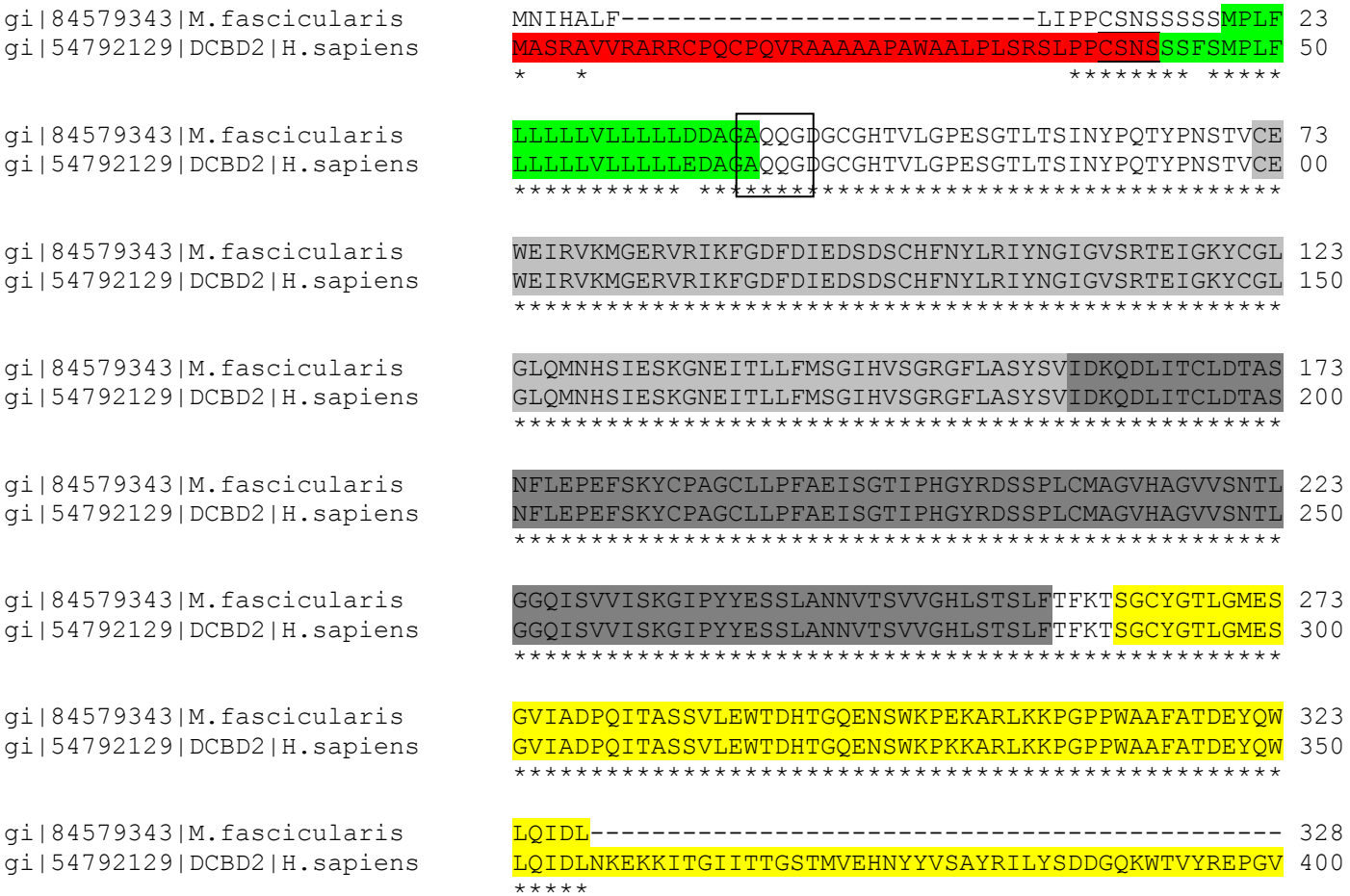


Abbildung 41: Sequenz-Alignment zwischen humanen DCBD2 und dem hypothetischen Protein BAE73105 (CLUSTAL 2.0.8). **Unterstrichen**: Übergangsbereich (tra). **Kasten**: Signalpeptidase-Schnittstelle GA-QQG. Das Alignment ist nur für die 328 Reste von BAE73105 gezeigt. *****: Konservierter Rest in allen Sequenzen. **Rot**: N-Domäne (NtraC). **Grün**: C-Domäne (NtraC). **Hellgrau**: CUB-Domäne. **Dunkelgrau**: LCCL-Domäne. **Gelb**: FA58C-3-Domäne.

β-Turns und Signalpeptidase-Schnittstellen

Im Rahmen des NtraC-Modells erfolgte eine Untersuchung von Signalsequenzen, um Bereiche zu identifizieren, die potentiel β-Turns, bestehend aus vier Resten, bilden können. Diese β-Turns sind für die Positionierung des Übergangsbereiches des NtraC-Modells entscheidend. Bei der Untersuchung der Signalsequenzen ist aufgefallen, dass häufig „falsch-positive“ β-Turns im Bereich der Signalpeptidase-Schnittstelle auftauchen. Falsch-positiv sind diese aber nur in Bezug auf das NtraC-Modell, da direkt an der Signalpeptidase-Schnittstelle kein Übergangsbereich liegen kann. Sie stellen gleichwertig vorhergesagte β-Turns im Vergleich zu den „Übergangsbereich-β-Turns“ dar, nur außerhalb des NtraC-Modell-Kontextes. Das gehäufte Auftreten von β-Turns in unmittelbarer Nähe der Signalpeptidase-Schnittstelle könnte aber unabhängig vom NtraC-Modell von Bedeutung sein und wurde daher quantitativ untersucht. Hierzu wurden von uns aus der UniProtKB (Version 14.0) unter Verwendung des SRS (Version 7.1.3) alle Eukaryota-Proteine mit einer Signalsequenz extrahiert (*N* = 14.696). Hierbei sind sowohl putative Proteine als auch putative Signalsequenzen enthalten. Aus den gefundenen Signalsequenzen wurden 578 Sequenzen mit einer unbekannten (1-x) oder keiner annotierten Länge entfernt. Die verbleibenden Sequenzen (*N* = 14.118) wurden hinsichtlich des Auftretens eines β-Turns in der Umgebung der Signalpeptidase-Schnittstelle untersucht (Abb. 42). Dabei wurde eine Unterteilung in Proteine mit kurzer Signalsequenz (1-40 Aminosäuren), langer Signalsequenz (40-100 Aminosäuren) und NtraC-organisierten langen Signalpeptiden vorgenommen. Betrachtet wurden in allen drei Gruppen Bereiche zwischen 10 Aminosäuren *upstream* und 15 Aminosäuren *downstream* der Signalpeptidase-Schnittstelle (Abb. 42). In Abbildung 42 ist Folgendes zu erkennen:

1) Es ist kein Unterschied zwischen den Gruppen kurzer, langer und langer NtraC-organisierter Signalpeptide zu erkennen. Die Existenz von β-Turns steht nicht im Zusammenhang mit der Länge oder der Organisation des Signalpeptides.

2) 50% aller Sequenzen weisen einen β-Turn in der Nähe (-5/+10 Positionen) zur Signalpeptidase-Schnittstelle (SPS) auf.

3) Bei einer Erweiterung des betrachteten Bereiches auf die Positionen -10 bis +15, ausgehend von der SPS, besitzen 80% der Sequenzen einen β-Turn im Bereich der SPS.

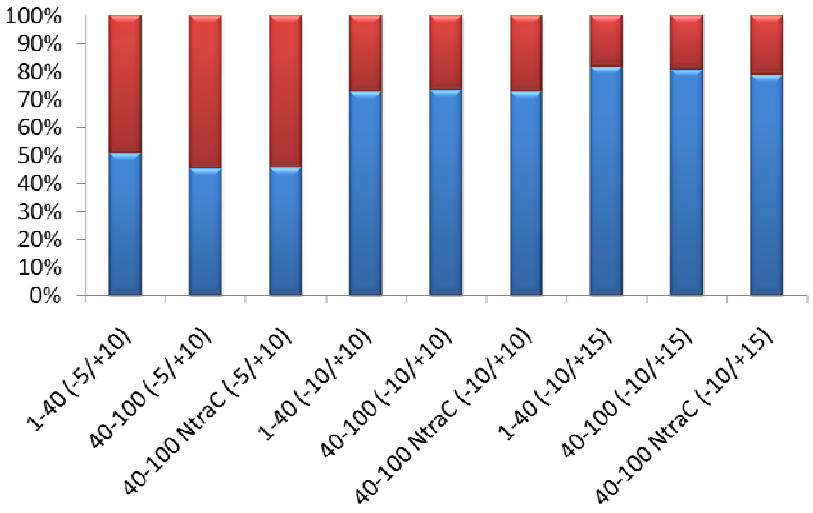


Abbildung 42: Auftreten von β-Turns im Bereich der Signalpeptidase-Schnittstelle von Vertebrata-Signalpeptiden. SP-Länge 1-40 Aminosäuren (*N* = 14.118). SP-Länge 40-100 Aminosäuren (*N* = 296). SP-Länge 40-100 und NtraC-organisiert (*N* = 185). **Blau**: Sequenzen mit β-Turn im Bereich der SP-Schnittstelle. **Rot**: Sequenzen ohne β-Turn im Bereich der SP-Schnittstelle. **Zahlen in Klammern**: Berücksichtigte Reste *upstream* (-) und *downstream* (+) der SP-Schnittstelle.

Das gehäufte Auftreten von β-Turns im Bereich der SPS (Abb. 42) könnte in funktionalem Zusammenhang mit der Signalpeptidase stehen. Hier werden weitere Untersuchungen angestrebt, ob

- SP ohne β-Turn evtl. nicht oder ineffizient geschnitten werden,
- eine Mutation des β-Turns im Bereich der SPS die Abspaltung des Signalpeptides verhindert,
- zur Signalpeptidase homologe Peptidasen bzw. deren Substrate ebenfalls β-Turns im Bereich der Schnittstelle aufweisen.

Sollten die vorliegenden Ergebnisse auf andere Peptidasen übertragbar sein, ermöglicht dies evtl. eine neue Art der Vorhersage von Peptidase-Schnittstellen: Bei unbekannter Peptidase-Schnittstelle wie z.B. bei Apicoplasten-Transit-Peptiden (Hempel *et al.*, 2007) könnte eine Suche nach β-Turns als Grundlage für Mutationsstudien zur Identifizierung der Peptidase-Schnittstelle dienen.

Bakterielle Autotransporter

Im Fokus dieser Arbeit stand die Untersuchung von eukaryotischen Signalpeptiden, speziell der Vertebrata, mit einer Länge von mehr als 40 Aminosäuren. Erweiternd wurde auch eine erste Untersuchung bakterieller Autotransporter des Typ-V durchgeführt, die ebenfalls lange Signalpeptide aufweisen (Henderson *et al*., 2004). Gram-positive und Gram-negative Bakterien sezernieren Proteine bzw. transportieren Proteine zur Plasmamembran über den sogenannten Sec-abhängigen *Pathway* (Clérico *et al*., 2008). Dieser besteht zentral aus SecY, einem Homologen des Sec61-Komplexes in Eukaryoten (Wiech *et al*., 1991). Ihre Signalpeptidase-Schnittstelle folgt ebenfalls der -3 -1 Regel von von Heijne (1983). In Schneider und Wrede (1994) wurde gezeigt, dass an Position -2 eine große Aminosäure wie Tryptophan liegen muss. Prokaryoten besitzen entsprechend auch eine Signalpeptidase, die in der inneren Plasmamembran lokalisiert ist (von Heijne, 1992). Der hydrophobe Kern-Bereich liegt bei Position -6 (Schneider *et al.,* 1993). Basierend auf diesen Regeln konnten artifizielle Signalpeptidase-Schnittstellen für *E.coli in silico* entworfen und erfolgreich *in vivo* getestet werden (Wrede *et al.,* 1998). Bei Gram-negativen Bakterien treten zusätzlich bakterielle Autotransporter auf (Dautin und Bernstein, 2007; Wells *et al*., 2007). Bakterielle Autotransporter sind Proteine, die in Kombination oder alternativ zu dem Sec-Sekretionsweg eigene Komplexe bilden, um den periplasmatischen und/oder den extrazellulären Raum zu erreichen. Typ-V-Autotransporter überwinden die innere Membran mit Hilfe des Sec-Komplexes (Henderson *et al.,* 2004; Yen und Stathopoulos, 2007). Für die Überwindung oder die Integration in die äußere Membran bilden sie eine eigene Pore. Sie besitzen einen typischen Aufbau, bestehend aus einer N-terminalen Signalsequenz, gefolgt von der zu transportierenden Domäne, einer Linker-Region und C-terminal einer β-*barrel*-formierenden Domäne (Yen und Stathopoulos, 2007). Die Signalsequenz ist dabei, vergleichbar mit der Signalsequenz in Eukaryoten, für den Transport zu und die Interaktion mit dem Sec-Komplex entscheidend (Yen und Stathopoulos, 2007). Nach Translokation des Proteins über die innere Membran wird die Signalsequenz abgespalten. Die β-Domäne insertiert in die äußere Membran und bildet ein β-*barrel*. Durch die so formierte Pore kann nun die zu transportierende Domäne den extrazellulären Raum erreichen oder in die äußere Membran integriert werden (Henderson *et al*., 2004).

Von Henderson *et al*. (2004) wird berichtet, dass mindestens 80 Autotransporter, darunter AIDA-1, Pet und Hbp, eine ungewöhnlich lange Signalsequenz besitzen. Die Signalsequenz

wird in verschiedene Bereiche, entsprechend der Ladung und der Hydrophobizität der Aminosäuren, unterteilt (Abb. 43, unverändert entnommen aus Henderson *et al*., 2004). Hierbei wird beobachtet, dass der C-terminale Bereich der Signalsequenz einer normal kurzen Signalsequenz ähnelt und der N-terminale Bereich maßgeblich für die Variation in der Länge verantwortlich ist. Dies ist konform mit der Idee des NtraC-Modells. Beim NtraC-Modell enthält eine Domäne die notwendige und hinreichende Signalsequenz (z.B. die C-Domäne), während die zweite Domäne eine potentielle zusätzliche Funktion besitzt, die mit dem primären Targeting nicht interferiert. Eine Untersuchung der von Henderson *et al.* (2004) angegebenen 46 Autotransporter-Signalsequenzen hinsichtlich einer NtraC-Organisation führte zu folgendem Ergebnis:

Von den 46 gegebenen Autotransporter-Signalsequenzen weisen 42 eine NtraC-Unterteilung auf (Abb. 44).

Abbildung 44 ist an den Aufbau der von Henderson publizierten Abbildung 43 angelehnt, um den Vergleich der Ergebnisse zu erleichtern. Die Unterteilung der Region n1, n2, h1 und h2 ist in Abbildung 43 und 44 identisch zu sehen. Diese NtraC-Architektur ist mit einer Verschiebung von ±5 Aminosäuren mit der in der Henderson *et al*., 2004 gemachten Unterteilung identisch. Dies argumentiert für die von Henderson *et al*. (2004) vorgenommenen Unterteilung, da sie ebenfalls von dem NtraC-Modell vorhergesagt wird. Zu betonen ist, dass das NtraC-Modell für die Vorhersage eine grundlegend andere Methode verwendet als Henderson *et al.* Bei Henderson *et al.* erfolgt die Unterteilung aufgrund der Hydrophobizitäts- und der Ladungsverteilung, im NtraC-Modell erfolgt die Unterteilung aufgrund der Sekundärstruktur. Die gefundene Übereinstimmung des NtraC-Modells mit den publizierten Ergebnissen ist als Hinweis zu werten, dass das NtraC-Modell auf bakterielle lange Signalsequenzen erweitert werden kann.

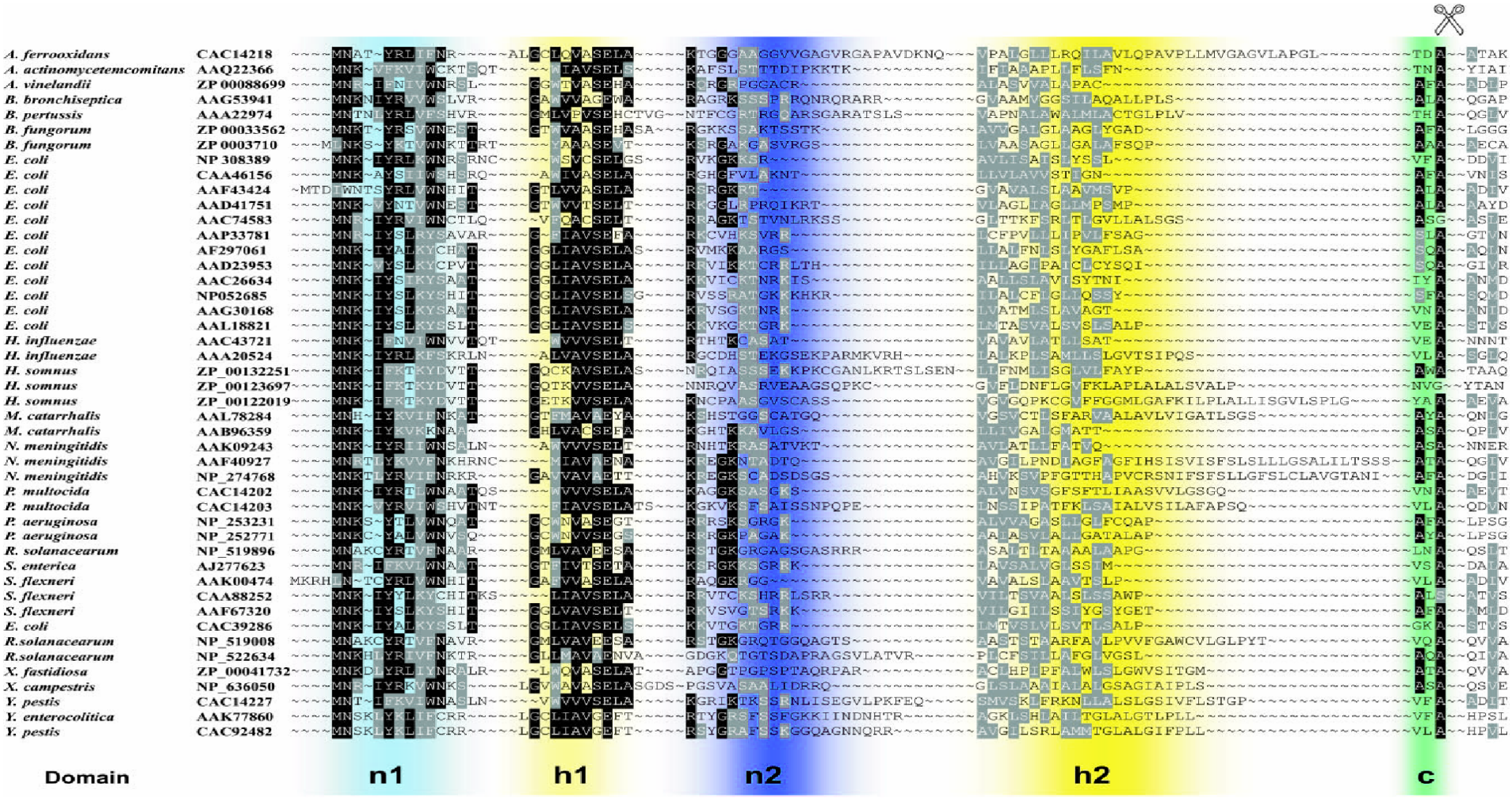


Abbildung 43: Unverändert entnommene Abbildung aus Henderson *et al*. (2004). Übersetzung der originalen Bildunterschrift: Struktur und Alignment der langen Signalsequenzen. Darstellung der langen Autotransporter-Signalsequenzen aus einem weiten Bereich der Gram-negativen Bakterien. **Blau**: Positiv geladene n1 und n2 Domänen. **Gelb**: Hydrophobe Domänen h1 und h2. **Grün**: Signalpeptidase-Schnittstellen. Die n2, h2 und C-Domänen sind charakteristisch für Signalsequenzen, die über den Sec-abhängigen Sekretionsweg in Kombination mit dem SecB-Chaperon sekretiert werden. Die *in silico* vorhergesagten und/oder empirisch ermittelten Schnittstellen zwischen der Signalsequenz und der zu transportierenden Domäne sind markiert. Konservierte Reste sind hervorgehoben.

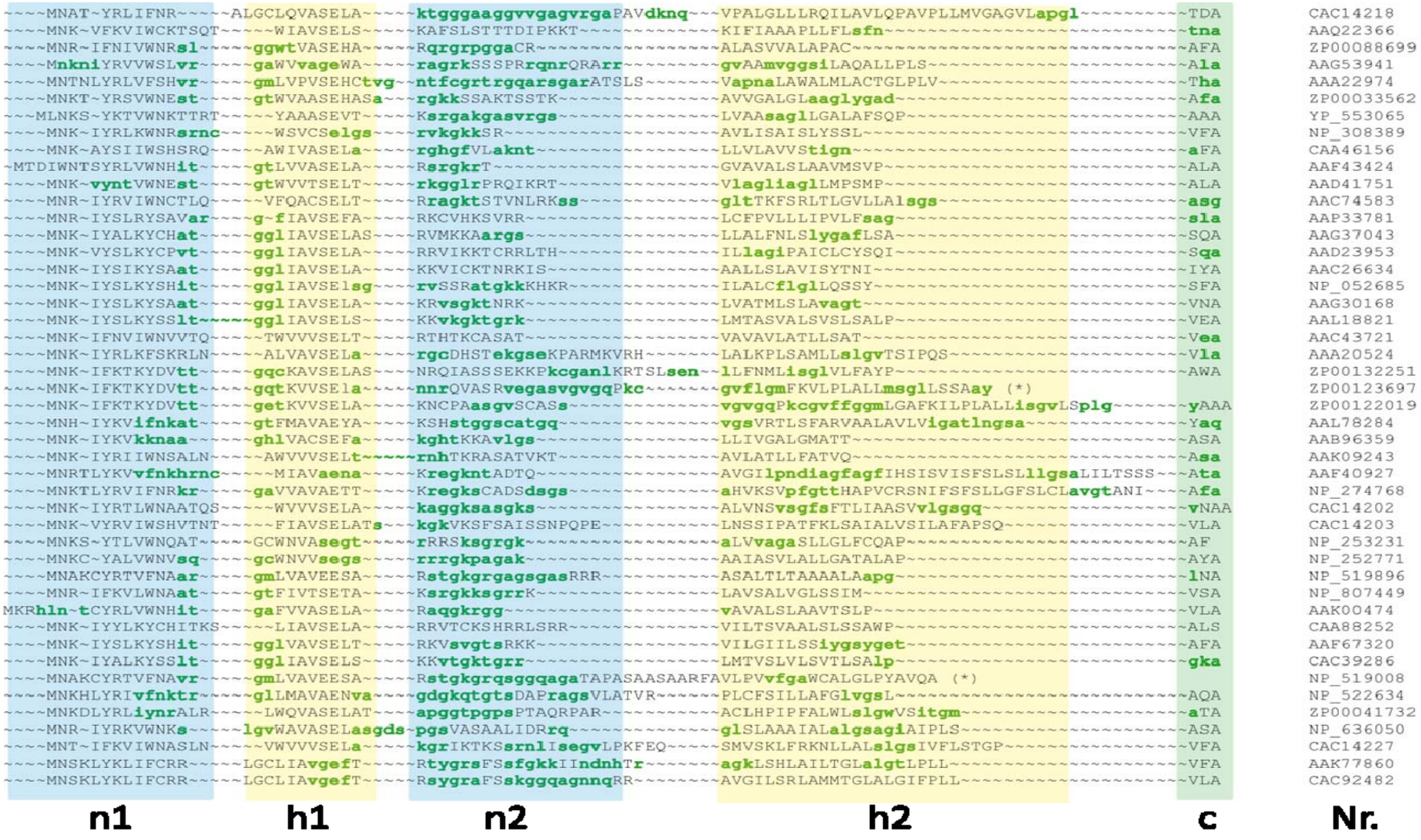


Abbildung 44: Liste von 46 Autotransporter-Signalsequenzen, angepasst nach Henderson *et al*. 2004, vgl. Abb. 43. **(*)**: Zu Henderson *et al*. abweichende aktuelle Sequenzen. **Blau**: Positiv geladene n1 und n2 Domänen. **Gelb**: Hydrophobe Domänen h1 und h2. **Grün**: Signalpeptidase-Schnittstellen. **Nr.**: UniProtKB/SRS Zugriffsnummer. **Dunkelgrün/hellgrün hervorgehobene Kleinbuchstaben**: Vorhergesagte 4-Reste-β-Turns, potentielle Übergangsbereiche nach NtraC-Modell.

Virale Signalpeptide

Bei viralen Signalpeptiden sind 5% der annotierten Beispiele länger als 40 Aminosäuren (Abb. 9, Kapitel „Längenverteilung von Signalpeptiden“). Als hervorgehobenes Beispiel sei hier das Glycoprotein-C des *lymphocytic choriomeningitis virus* (LCMV) genannt (Beyer *et al*., 2001). Das Protein hat eine lange Signalsequenz von 58 Aminosäuren (▼= Signalpeptidase-Schnittstelle):

MGQIVTMFEALPHIIDEVINIVIIVLIIITSIKAVYNFATCGILALVSFLFLAGRSCG▼MY…

Es wurde gezeigt, dass das aus dieser langen Signalsequenz resultierende Signalpeptid nach der Abspaltung stabil ist (Froeschke *et al*., 2003) und in neu entstehende Viren verpackt wird. Das Signalpeptid ist direkt oder indirekt notwendig, um Glycoprotein-C in seine funktionale Form zu überführen - und damit für die Infektiosität des Virus (Schrempf *et al*., 2007). Ein elegantes Experiment zeigt dies durch den Ersatz der Wildtyp-Signalsequenz mit einer normal kurzen. Dies führte zwar zu einem ER-Targeting des Glykopoteins-C, dieses war aber inaktiv. Das Wildtyp-Signalpeptid des Glycoproteins-C weist eine NtraC-Organisation auf. Die C-Domäne weist dabei ein ER-Targeting-Signal auf; die N-Domäne hat keine vorhergesagte Targeting-Funktion und ist damit für zusätzliche Funktionen frei. Die N-Domäne könnte als Signalanker fungieren. Die NtraC-Einteilung ist wie folgt (Übergangsbereich unterstrichen):

N-Domäne:
MGQIVTMFEALPHIIDEVINIVIIVLIIITSIKAVYNF

C-Domäne:
ATCGILALVSFLFLAGRSCG▼MY…

Von Schrempf *et al*. (2007) wurde ebenfalls eine Unterteilung des Signalpeptides vorgenommen. Diese ist motiviert durch das geladene Lysin an Position 33, das von zwei hydrophoben Bereichen flankiert wird. Die in Schrempf *et al*. (2007) vorgeschlagenen Orientierungen des Signalpeptides über die Membran sind in Abbildung 45 dargestellt.

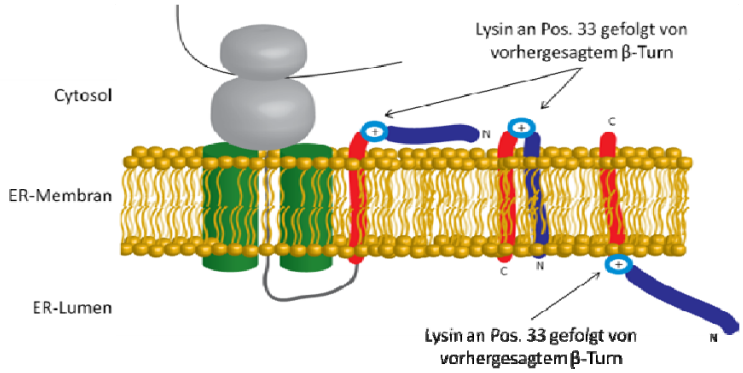


Abbildung 45: Mögliche Orientierungen des Signalpeptides von Glycoprotein-C von LCMV in der ER-Membran, vorgeschlagen in Schrempf *et al*. (2007). **Hellgrau**: Ribosom mit mRNA. **Grün**: Sec61-Komplex. **Rot**: Erster hydrophober Bereich im Signalpeptid *und* N-Domäne nach NtraC-Modell**. Dunkelblau**: Zweiter hydrophober Bereich im Signalpeptid *und* C-Domäne nach NtraC-Modell. **Pfeile**: Lysin an Position 33 dient zur Unterteilung des Signalpeptides in zwei hydrophobe Bereiche nach Schrempf *et al. und* Position eines Übergangsbereiches nach NtraC-Modell.

In Schrempf *et al*. (2007) werden verschiedene Orientierungen des Signalpeptides in der ER-Membran vorgeschlagen (Abb. 45). Die Lage des N-Terminus im ER-Lumen wird unterstützt durch artifizielles Einbringen einer Glykosylierungsstelle, die glykosyliert wurde (Schrempf *et al.,* 2007). Das positiv geladene Lysin an Position 33 (Abb. 45, blauer Kreis) wird dabei von Schrempf *et al*. zur Unterteilung des Signalpeptides in zwei hydrophobe Bereiche herangezogen (Abb. 45, rot und dunkelblau). Dieses Lysin liegt zwei Aminosäuren *upstream* zu einem von uns vorhergesagten β-Turn „VYNF“. Dieser potentielle β-Turn stellt für das Signalpeptid des Glycoproteins-C einen Übergangsbereich des NtraC-Modells dar (Abb. 45, Pfeile). Die vorgeschlagenen Orientierungen des langen Signalpeptids von Schrempf *et al*. (2007) sind somit konform zu einer Unterteilung basierend auf dem NtraC-Modell. Als Ergänzung zu Schrempf *et al.* ermöglicht das NtraC-Modell, einen funktionalen Aspekt einzuführen. Die vorhergesagte C-Domäne des LCMV-Glycoproteins-C (Position 39-58) wird von uns als hinreichend für das ER-Targeting vorgeschlagen. Die vorhergesagte N-Domäne (Position 1-38) wird von uns als Signalanker oder Interaktionspartner für andere Proteine vorgeschlagen.

Ein weiteres Beispiel für ein virales NtraC-organisiertes Signalpeptid ist das lange Signalpeptid des Hüll-Glycoproteins des humanen *Foamy virus*. In Lindemann *et al*. (2001) wird eine Länge von 146 Aminosäuren für das Signalpeptid gezeigt. Das Signalpeptid hat eine Post-Targeting-Funktion, es reguliert die Freisetzung von Viren-Partikeln (Stanke *et al*.,

2005). In Lindemann *et al*. (2001) wird weiterhin durch Konstrukte des Signalpeptides mit unterschiedlicher Länge gezeigt, dass die N-terminalen 15 Aminosäuren nicht für das Targeting zum ER notwendig sind. Die Aminosäuren 1-15 sind jedoch notwendig für den *budding*-Prozess des Virus.
Eine automatisierte Analyse des langen Signalpeptides mit Hilfe der webbasierten Benutzeroberfläche des NtraC-Algorithmus brachte folgende Ergebnisse:

- Das lange Signalpeptid dieses Glycoproteins weist eine NtraC-Organisation auf.
- Die N-Domäne umfasst die Aminosäuren 1-18 (inklusive Übergangsbereich 15-18). Eine Targeting-Funktion für die N-Domäne wird nicht vorhergesagt.
- Die C-Domäne umfasst die Reste 20-146 und wird als Signalanker oder bakterielles Sekretionssignal vorhergesagt.

Die Ergebnisse des NtraC-Modells stimmen somit mit den experimentellen Ergebnissen aus Lindemann *et al.* (2001) überein. Der NtraC-Algorithmus erkennt eine C-Domäne mit Signal-Charakter und eine N-Domäne mit unbekannter Funktion, deren Lage mit den experimentell gezeigten Positionen zur Deckung kommt.

Die Übereinstimmung der hier gezeigten *in silico*-Analyse viraler Signalpeptide mit existierenden experimentellen Befunden argumentieren dafür, dass die NtraC-Domänen-Architektur auf lange virale Signalpeptide übertragbar ist.

Zusammenfassung

Ziel der Arbeit war die Analyse von langen eukaryotischen Signalpeptiden, mit einer Länge von mindestens 40 Aminosäuren, und ihre Diskriminierung zu kurzen SP. Signalpeptide sind notwendig, um die im Cytosol translatierten Proteine zum Ort ihrer Funktion zu dirigieren. Sie spielen dadurch eine fundamentale Rolle bei der Entwicklung von Zellen. Signalpeptide weisen keine Sequenzhomologie, aber einen typischen, in drei Regionen gegliederten Aufbau (*n-, h-, c*-Region) auf. In den letzten Jahren wurden zunehmend Beispiele von Signalpeptiden gefunden, die neben dem Targeting zum endoplasmatischen Retikulum weitere Post-Targeting-Funktionen aufweisen. Auffällig ist hier die besondere Länge der Signalpeptide. Für die Analyse dieser langen Signalpeptide standen bis jetzt keine gezielt entwickelten Vorhersageprogramme zur Verfügung. Im Rahmen dieser Arbeit wurde diese Gruppe langer Signalpeptide untersucht und ein Modell zu deren interner Organisation entwickelt. Das entwickelte „NtraC"-Modell erweitert etablierte sequenzbasierte Ansätze für kurze SP um eine Sekundärstruktur-motivierte Perspektive für lange Sinalpeptide. Zuerst wird dabei ein Übergangsbereich (*transition area*, N„tra"C), der potentiell β-Turn bildende Aminosäuren enthält, identifiziert. Dieser dient im Modell zur Zerlegung des SP in zwei hinsichtlich ihrer Funktion unabhängige Domänen: eine N-terminale N-Domäne (‚N'traC) und eine C-terminale C-Domäne (Ntra‚C'). Diese mit bekannten Vorhersageprogrammen nicht identifizierbaren „kryptischen" Domänen innerhalb der Signalpeptid-Sequenz können unterschiedliche Targeting-Kapazitäten aufweisen und entsprechen für sich genommen eigenständigen Protein-Targeting-Signalen. Im Fall einer ER-Targeting Kapazität z.B. weist eine Domäne für sich genommen eine *n-, h-,* und *c*-Region auf. 63% aller Vertebrata-Signalpeptide entsprechen der in dieser Arbeit vorgeschlagenen NtraC-Organisation. Eine basierend auf dem NtraC-Modell vorgeschlagene Architektur für die langen Signalpeptide von shrew-1 (43 Aminosäuren), DCBD2 (66 Aminosäuren) und RGMA (47 Aminosäuren) wurde vom Autor selbst *in vitro* überprüft. Für alle drei Proteine wurden eine N-Domäne mit mitochondrialer Targeting-Funktion und eine C-Domäne mit Signalpeptid-Funktion vorhergesagt. Die langen Signalpeptide der Proteine wurden bisher als reine ER-Targeting-Signale betrachtet. Die vorliegende Studie zeigt jedoch, dass in diesen langen Signalpeptiden multiple Targetingsignale kodiert sind. Die ER-Targeting-Kapazität der C-Domänen wurde durch SEAP-Assays überprüft, die mTP-Funktion der N-Domäne durch biochemische Aufreinigung von Mitochondrien. Die *in silico*-Vorhersagen konnten in vollem Umfang für

alle drei Proteine *in vitro* bestätigt werden. Eine Untersuchung der semantischen Wolke aller Proteine mit NtraC-organisiertem Signalpeptid zeigte, dass eine NtraC-Organisation in mehr als 50% der Fälle im Zusammenhang mit Typ-I Transmembranproteinen auftritt. Auch die Proteine der hier experimentell untersuchten Signalpeptide von shrew-1, DCBD2, RGMA sind Typ-I Transmembranproteine. Des Weiteren weisen 15% aller langen Vertebrata-Signalpeptide eine Domänen-Kombination analog zu shrew-1, DCBD2 und RGMA auf. Der gefundene analoge Aufbau der langen Signalpeptide könnte somit funktionelle Gruppen von Proteinen zusammenführen, die bisher anderweitig nicht gruppiert werden konnten.

Es konnte weiterhin gezeigt werden, dass bakterielle Autotransporter Gram-negativer Bakterien in Variation ebenfalls eine NtraC-Organisation in ihren Signalpeptiden aufweisen. Gleiches konnte für Gruppen langer viraler Signalpeptide gezeigt werden. Das NtraC-Modell ist somit nicht auf Vertebrata-Signalpeptide beschränkt.

In der vorliegenden Arbeit wurde ein Modell zur Domänen-Architektur langer Signalpeptide entwickelt und erfolgreich angewendet: das NtraC-Modell. Ein Vorhersage-Algorithmus zur *in silico*-Untersuchung langer Signalpeptide wurde implementiert und in einer webbasierten Benutzeroberfläche öffentlich zugänglich gemacht. Das Modell trifft auf 63% der annotierten langen Vertebrata-Signalpeptide zu. Des Weiteren wurden, basierend auf dem NtraC-Modell, für die langen Signalpeptide von drei Proteinen (shrew-1, DCBD2, RGMA) *in vitro*-Versuche durchgeführt. Die erhaltenen *in vitro*-Ergebnisse unterstützen klar die These, dass lange Signalpeptide eine aus definierten Domänen bestehende Organisation aufweisen können.

Literatur

Abendroth, A., Lin, I., Slobedman, B., Ploegh, H. und Arvin, A.M. (2001) Varicella-zoster virus retains major histocompatibility complex class I proteins in the Golgi compartment of infected cells. *J. Virol.* 75, 4878–4888.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. und Walter, P. *Molekularbiologie der Zelle.* WILEY-VCH Verlag, Weinheim (2004).

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. und Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.

Arnold, A., Horst, S.A., Gardella, T.J., Baba, H., Levine, M.A. und Kronenberg, H.M. (1990) Mutation of the signal peptide-encoding region of the preproparathyroid hormone gene in familial isolated hypoparathyroidism. *J. Clin. Invest.* 86, 1084–1087.

Bachmann, B.J. (1983) Linkage maps of Escherichia coli K-12, edition 7. *Microbiol. Rev.* 47, 180-230.

Bange, G., Wild, K. und Sinning, I. (2007) Protein translocation: checkpoint role for SRP GTPase activation. *Curr. Biol.* 17, R980-982.

Barbulescu, M., Turner, G., Seaman, M.I., Deinard, A.S., Kidd, K.K. und Lenz, J. (1999) Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Curr. Biol.* 9, 861-868.

Baumgartner, S., Hofmann, K., Chiquet-Ehrismann, R. und Bucher P. (1998) The discoidin domain family revisited: new members from prokaryotes and a homology-based fold prediction. *Protein Sci.* 7, 1626-1631.

Bendtsen, J.D., Nielsen H., von Heijne, G. und Brunak, S. (2004a) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 340, 783-795.

Bendtsen, J.D., Jensen, L.J., Blom, N., von Heijne, G. und Brunak, S. (2004b) Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.* 17, 349-356.

Bennett, E.M., Bennink, J.R., Yewdell, J.W. und Brodsky, F.M. (1999) Cutting edge: adenovirus E19 has two mechanisms for affecting class I MHC expression. *J. Immunol.* 162, 5049–5052.

Berger, J., Hauber, J., Hauber, R., Geiger, R. und Cullen, B.R. (1988) Secreted placental alkaline phosphatase: a powerful new quantitative indicator of gene expression in eukaryotic cells. *Gene* 66, 1-10.

Bernstein, H.D., Poritz, M.A., Strub, K., Hoben, P.J., Brenner, S. und Walter, P. (1989) Model for signal sequence recognition from amino-acid sequence of 54K subunit of signal recognition particle. *Nature* 340, 482-486.

Beyer, W.R., Miletic, H., Ostertag, W. und von Laer, D. (2001) Recombinant expression of lymphocytic choriomeningitis virus strain WE glycoproteins: a single amino acid makes the difference. *J. Virol.* 75, 1061-1064.

Bharti, S., Handrow-Metzmacher, H., Zickenheiner, S., Zeitvogel, A., Baumann, R. und Starzinski-Powitz, A. (2004) Novel membrane protein shrew-1 targets to cadherin-mediated junctions in polarized epithelial cells. *Mol. Biol. Cell* 15, 397-406.

Blobel, G. und Dobberstein, B. (1975a) Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. *J. Cell Biol.* 67, 835-851.

Blobel, G. und Dobberstein, B. (1975b) Transfer of proteins across membranes. II. Reconstitution of functional rough microsomes from heterologous components. *J. Cell Biol.* 67, 852-862.

Boeckmann, B., Blatter, M.C., Famiglietti, L., Hinz, U., Lane, L., Roechert, B. und Bairoch, A. (2005) Protein variety and functional diversity: Swiss-Prot annotation in its biological context. *C. R. Biol.* 328, 882-899.

Bork, P. (1991) Complement components C1r/C1s, bone morphogenic protein 1 and Xenopus laevis developmentally regulated protein UVS.2 share common repeats. *FEBS Lett.* 282, 9-12.

Bork, P. und Beckmann, G. (1993) The CUB domain. A widespread module in developmentally regulated proteins. *J. Mol. Biol.* 231, 539-545.

Braakman, I., Helenius, J. und Helenius, A. (1992) Manipulating disulfide bond formation and protein folding in the endoplasmic reticulum. *EMBO J.* 11, 1717-1722.

Braud, V., Jones, E.Y. und McMichael, A. (1997) The human major histocompatibility complex class Ib molecule HLA-E binds signal sequence-derived peptides with primary anchor residues at positions 2 and 9. *Eur. J. Immunol.* 27, 1164-1169.

Braud, V.M., Allan, D.S., O'Callaghan, C.A., Söderström, K., D'Andrea, A., Ogg, G.S., Lazetic, S., Young, N.T., Bell, J.I., Phillips, J.H., Lanier, L.L. und McMichael, A.J. (1998) HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. *Nature* 391,795-799.

Bubeck, A., Reusch, U., Wagner, M., Ruppert, T., Muranyi, W., Kloetzel, P.M. und Koszinowski, U.H. (2002) The glycoprotein gp48 of murine cytomegalovirus proteasome-dependent cytosolic dislocation and degradation. *J. Biol. Chem.* 277, 2216-2224.

Bystroff, C., Thorsson, V. und Baker, D. (2000) HMMSTR: A hidden markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.* 301, 173-190.

Bystroff, C. und Shao, Y. (2002) Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics* 18, 54-61.

Cai, Y-D., Liu, X.-J., Li, Y.-X., Xu, X.-B und Chou, K-C. (2003) Prediction of beta-turns with learning machines. *Peptides* 24, 665-669.

Camus, L.M. und Lambert, L.A. (2007) Molecular evolution of hemojuvelin and the repulsive guidance molecule family. *J. Mol. Evol.* 65, 68-81.

Cavalier-Smith, T. (1999) Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J. Eukaryot. Microbiol.* 46, 347-366.

Cedano, J., Aloy, P., Pérez-Pons, J.A. und Querol, E. (1997) Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.* 266, 594-600.

Chen, E.Y., Bartlett, M.C. und Clarke, D.M. (2000) Cystic fibrosis transmembrane conductance regulator has an altered structure when its maturation is inhibited. *Biochemistry* 39, 3797-3803.

Chen, E.Y., Bartlett, M.C., Loo, T.W. und Clarke, D.M. (2004) The DeltaF508 mutation disrupts packing of the transmembrane segments of the cystic fibrosis transmembrane conductance regulator. *J. Biol. Chem.* 279, 39620-39627.

Chen, X., VanValkenburgh, C., Liang, H., Fang, H. und Green, N. (2001) Signal Peptidase and Oligosaccharyltransferase Interact in a Sequential and Dependent Manner within the Endoplasmic Reticulum. *J. Biol. Chem.* 276, 2411-2416.

Chou, K.C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43, 246–255.

Clérico, E.M., Maki, J.L. und Gierasch, L.M. (2008) Use of synthetic signal sequences to explore the protein export machinery. *Biopolymers* 90, 307-319.

Cooke, B.M., Lingelbach, K., Bannister, L.H. und Tilley, L. (2004) Protein trafficking in Plasmodium falciparum-infected red blood cells. *Trends Parasitol.* 20, 581–589.

Cooper, D.N. (2002) Galectinomics: finding themes in complexity. *Biochim. Biophys. Acta* 1572, 209-231.

Cox, J., Bennink, J. und Yewdell, J. (1991) Retention of adenovirus E19 glycoprotein in the endoplasmic reticulum is essential to its ability to block antigen presentation. *J. Exp. Med.* 174, 1629–1637.

Dacks, J.B., Peden, A.A. and Field, M.C. (2008) Evolution of specificity in the eukaryotic endomembrane system. *Int. J. Biochem. Cell Biol.* doi:10.1016/j.biocel.2008.08.041

Dalbey, R.E. und von Heijne, G. (1992) Signal peptidases in prokaryotes and eukaryotes--a new protease family. *Trends Biochem. Sci.* 17, 474-478.

Daly, M., Bruce, D., Perry, D.J., Price, J., Harper, P.L., O'Meara, A. und Carrell, R.W. (1990) Antithrombin Dublin (-3 Val→Glu): an N-terminal variant which has an aberrant signal peptidase cleavage site. *FEBS Lett.* 273, 87-90.

Datta, R., Waheed, A., Shah, G.N. und Sly, W.S. (2007) Signal sequence mutation in autosomal dominant form of hypoparathyroidism induces apoptosis that is corrected by a chemical chaperone. *Proc. Nat. Acad. Sci. U.S.A.* 104, 19989-19994.

Dautin, N. und Bernstein, H.D. (2007) Protein secretion in gram-negative bacteria via the autotransporter pathway. *Annu. Rev. Microbiol.* 61, 89-112.

de Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V., Petra, S., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A. und Hulo, N. (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* 34, W362-W365.

de Duve, C. (2007) The origin of eukaryotes: a reappraisal. *Nat. Rev. Genet.* 8, 395-403.

de Parseval, N., Lazar, V., Casella, J.-F., Benit, L. und Heidmann, T. (2003) Survey of human genes of retroviral origin: identification and transcriptome of the genes with coding capacity for complete envelope proteins. *J. Virol.* 77, 10414-10422.

Dessens, J.T., Sinden, R.E. und Claudianos, C. (2004) LCCL proteins of apicomplexan parasites. *Trends Parasitol.* 20, 102-108.

Deshaies, R.J., Sanders, S.L., Feldheim, D.A. und Schekman, R. (1991) Assembly of yeast Sec proteins involved in translocation into the endoplasmic reticulum into a membrane-bound multisubunit complex. *Nature* 349, 806-808.

Driessen, A.J. und Nouwen, N. (2008) Protein translocation across the bacterial cytoplasmic membrane. *Annu. Rev. Biochem.* 77, 643-667.

Dultz, E., Hildenbeutel, M., Martoglio, B., Hochman, J., Dobberstein, B. und Kapp, K. (2008) The signal peptide of the mouse mammary tumor virus Rem protein is released from the endoplasmic reticulum membrane and accumulates in nucleoli. *J. Biol. Chem.* 283, 9966-9976.

Eichler, R., Lenz, O., Strecker, T., Eickmann, M., Klenk, H.D. und Garten, W. (2003a) Identification of Lassa virus glycoprotein signal peptide as a trans-acting maturation factor. *EMBO Rep.* 4, 1084-1088.

Eichler, R., Lenz, O., Strecker, T. und Garten, W. (2003b) Signal peptide of Lassa virus glycoprotein GP-C exhibits an unusual length. *FEBS Lett.* 538, 203-206.

Emanuelsson, O., Nielsen, H., Brunak, B. und von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005-1016.

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

Emanuelsson, O., Brunak, S., von Heijne, G. und Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* 2, 953-971.

Etzold, T., Harris, H. and Beulah, S. *Bioinformatics: Managing Scientific Data*. *Chapter: SRS, An integration platform for databanks and analysis tools in bioinformatics.* Elsevier Science (USA) (2003).

Evans, E.A., Gilmore, R. und Blobel, G. (1986) Purification of microsomal signal peptidase as a complex. *Proc. Nat. Acad. Sci. U.S.A.* 83, 581-585.

Froeschke, M., Basler, M., Groettrup, M. und Dobberstein, B. (2003) Long-lived signal peptide of lymphocytic choriomeningitis virus glycoprotein pGP-C. *J. Biol. Chem.* 278, 41914-41920.

Gakh, O., Cavadini, P. und Isaya, G. (2002) Mitochondrial processing peptidases. *Biochim. Biophys. Acta* 1592, 63-77.

George, R., Beddoe, T., Landl, K. und Lithgow, T. (1998) The yeast nascent polypeptide-associated complex initiates protein targeting to mitochondria in vivo. *Proc. Nat. Acad. Sci. U.S.A.* 95, 2296-2301.

Gierasch, L.M. (1989) Signal Sequences. *Biochemistry* 28, 923-930.

Gilmore, R., Walter, P. und Blobel, G. (1982) Protein translocation across the endoplasmic reticulum. II. Isolation and characterization of the signal recognition particle receptor. *J. Cell Biol.* 95, 470-477.

Görlich, D., Hartmann, E., Prehn, S. und Rapoport, T.A. (1992) A protein of the endoplasmic reticulum involved early in polypeptide translocation. *Nature* 357, 47-52.

Görlich, D. und Rapoport, T.A. (1993) Protein translocation into proteoliposomes reconstituted from purified components of the endoplasmic reticulum membrane. *Cell* 75, 615–630.

Goldman, B.M. und Blobel, G. (1978) Biogenesis of peroxisomes: intracellular site of synthesis of catalase and uricase. *Proc. Nat. Acad. Sci. U.S.A.* 75, 5066-5070.

Gould, S.J., Keller, G.A. und Subramani, S. (1987) Identification of a peroxisomal targeting signal at the carboxy terminus of firefly luciferase. *J. Cell Biol.* 105, 2923-2931.

Gould, S.J., Keller, G.A. und Subramani, S. (1988) Identification of peroxisomal targeting signals located at the carboxy terminus of four peroxisomal proteins. *J. Cell Biol.* 107, 897-905.

Halic, M., Becker, T., Pool, M.R., Spahn, C.M., Grassucci, R.A., Frank, J. und Beckmann, R. (2004) Structure of the signal recognition particle interacting with the elongation-arrested ribosome. *Nature* 427, 808-814.

Harte, N., Silventoinen, V., Quevillon, E., Robinson, S., Kallio, K., Fustero, X., Patel, P., Jokinen, P. und Lopez, R. (2004) Public web-based services from the European Bioinformatics Institute. *Nucleic Acids Res.* 32, W3-9.

Hartl, F.U., Lecker, S., Schiebel, E., Hendrick, J.P. und Wickner, W. (1990) The binding cascade of SecB to SecA to SecY/E mediates preprotein targeting to the E. coli plasma membrane. *Cell* 63, 269-279.

Hegde, R.S. (2002) Targeting and beyond: new roles for old signal sequences. *Mol. Cell* 10, 697-698.

Hegde, R.S. und Bernstein, H.D. (2006) The surprising complexity of signal sequences. *Trends Biochem. Sci.* 31, 563-571.

Helenius, A., Marquardt, T. und Braakman, I. (1992) The endoplasmic reticulum as a protein-folding compartment. *Trends Cell Biol.* 2, 227-231.

Hempel, F., Bozarth, A., Sommer, M.S., Zauner, S., Przyborski, J.M. und Maier, U.-G. (2007) Transport of nuclear-encoded proteins into secondarily evolved plastids. *Biol. Chem.* 388, 899-906.

Henderson, L.E., Sowder, R., Smythers, G. und Oroszlan, S. (1983) Terminal amino acid sequences and proteolytic cleavage sites of mouse mammary tumor virus env gene products. *J. Virol.* 48, 314-319.

Henderson, I.R., Navarro-Garcia, F., Desvaux, M., Fernandez, R.C. und Ala'Aldeen, D. (2004) Type V Protein Secretion Pathway: the Autotransporter Story. *Microbiol. Mol. Biol. Rev.* 68, 692-744.

Hiller, N.L., Bhattacharjee, S., van Ooij, C., Liolios, K., Harrison, T., Lopez-Estrano, C. und Haldar, K. (2004) A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science* 306, 1934–1937.

Hiss, J.A., Przyborski, J.M., Schwarte, F., Lingelbach, K. und Schneider, G. (2008a) The Plasmodium export element revisited. *PLoS ONE* 3, e1560.

Hiss, J.A., Resch, E., Schreiner, A., Meissner, M., Starzinski-Powitz, A. und Schneider, G. (2008b) Domain organization of long signal peptides of single-pass integral membrane proteins reveals multiple functional capacity. *PLoS ONE* 3, e2767.

Hudson, A.W., Blom, D., Howley, P.M. und Ploegh, H.L. (2003) The ER-lumenal domain of the HHV-7 immunoevasin U21 directs class I MHC molecules to lysosomes. *Traffic* 4, 824–837.

Hughes, R.C. (1999) Secretion of the galectin family of mammalian carbohydrate-binding proteins. *Biochim. Biophys. Acta* 1473, 172-185.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M. und Sigrist, C.J. (2006) The PROSITE database. *Nucleic Acids Res.* 34, D227-D230.

Isaacson, M.K., Juckem, L.K. und Compton, T. (2008) Virus entry and innate immune activation. *Curr. Top. Microbiol. Immunol.* 325, 85-100.

Isaya, G., Kalousek, F., Fenton, W.A. und Rosenberg, L.E. (1991) Cleavage of precursors by the mitochondrial processing peptidase requires a compatible mature protein or an intermediate octapeptide. *J. Cell. Biol.* 113, 65-76.

Isaya, G. und Kalousek, F. (1995) Mitochondrial intermediate peptidase. *Meth. Enzymol.* 248, 556-67.

Jakob, V., Schreiner, A., Tikkanen, R. und Starzinski-Powitz, A. (2006) Targeting of transmembrane protein shrew-1 to adherens junctions is controlled by cytoplasmic sorting motifs. *Mol. Biol. Cell* 17, 3397-3408.

Joachims, T. *Making large-scale SVM learning practical. Adv kernel methods - support vector learning.* MIT-Press. Cambridge, MA (1999).

Kane, W.H. und Davie, E.W. (1988) Blood coagulation factors V and VIII: structural and functional similarities and their relationship to hemorrhagic and thrombotic disorders. *Blood* 71, 539-555.

Kalies, K.U., Stokes, V. und Hartmann, E. (2008) A single Sec61-complex functions as a protein-conducting channel. *Biochim. Biophys. Acta* Epub ahead of print.

Keenan, R.J., Freymann, D.M., Stroud, R.M. und Walter, P. (2001) The signal recognition particle. *Annu. Rev. Biochem.* 70,755-775.

Kim, S.J., Mitra, D., Salerno, J.R. und Hegde, R.S. (2002) Signal sequences control gating of the protein translocation channel in a substrate-specific manner. *Dev. Cell* 2, 207-217.

Kim, H-L., Passant, A, Breslin, J. Scerri, S. und Decker, S. (2008) Review and alignment of tag ontologies for semantically-linked data in collaborative taggin spaces. *IEEE International Conference on Semantic Computing.* in press.

Kobuke, K., Furukawa, Y., Sugai, M., Tanigaki, K., Ohashi, N., Matsumori, A., Sasayama, S., Honjo, T. und Tashiro, K. (2001) ESDN, a novel neuropilin-like membrane protein cloned from vascular cells with the longest secretory signal sequence among eukaryotes, is up-regulated after vascular injury. *J. Biol. Chem.* 276, 34105-34114.

Koshikawa, K., Osada, H., Kozaki, K., Konishi, H., Masuda, A., Tatematsu, Y., Mitsudomi, T., Nakao, A. und Takahashi, T. (2002) Significant up-regulation of a novel gene, CLCP1, in a highly metastatic lung cancer subline as well as in lung cancers in vivo. *Oncogene* 21, 2822-2828.

Kurys, G., Tagaya, Y., Bamford, R., Hanover, J.A., und Waldmann, T.A. (2000) The Long Signal Peptide Isoform and Its Alternative Processing Direct the Intracellular Trafficking of Interleukin-15. *J. Biol. Chem.* 275, 30653-30659.

Laforet, G.A. und Kendall, D.A. (1991) Functional limits of conformation, hydrophobicity, and steric constraints in prokaryotic signal peptide cleavage regions. Wild type transport by a simple polymeric signal sequence. *J. Biol. Chem.* 266, 1326-1334.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. und Higgins, D.G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947-2948.

Lemberg, M.K., Bland, F.A., Weihofen, A., Braud, V.M. und Martoglio, B. (2001) Intramembrane proteolysis of signal peptides: an essential step in the generation of HLA-E epitopes. *J. Immunol.* 167, 6441-6446.

Lemberg, M.K. und Martoglio, B. (2002) Requirements for signal peptide peptidase-catalyzed intramembrane proteolysis. *Mol. Cell* 10, 735-744.

Li, Y., Bergeron, J.J., Luo, L., Ou, W-J., Thomas, D.Y. and Kang, C.Y. (1996) Effects of inefficient cleavage of the signal sequence of HIV-1 gp120 on its association with calnexin, folding, and intracellular transport. *Proc. Nat. Acad. Sci. U.S.A.* 93, 9606-9611.

Lindemann, D., Pietschmann, T., Picard-Maureau, M., Berg, A., Heinkelein, M., Thurow, J., Knaus, P., Zentgraf, H. und Rethwilm, A. (2001) A particle-associated glycoprotein signal peptide essential for virus maturation and infectivity. *J. Virol.* 75, 5762-5771.

Lipp, J., Dobberstein, B. und Haeuptle, M-T. (1987) Signal recognition particle arrests elongation of nascent secretory and membrane proteins at multiple sites in a transient manner. *J. Biol. Chem.* 262, 1680-1684.

Liu, S.-L. und Miller, A.D. (2005) Transformation of madin-darby canine kidney epithelial cells by sheep retrovirus envelope proteins. *J. Virol.* 79, 927-933.

Loch, S. und Tampé, R. (2005) Viral evasion of the MHC class I antigen-processing machinery. *Pflügers Arch. – Eur. J. Physiol.* 451, 409–417.

Loewer, R., Toenjes, R.R., Korbmacher, C., Kurth, R. und Loewer, J. (1995) Identification of a Rev-related protein by analysis of spliced transcripts of the human endogenous retroviruses HTDV/HERV-K. *J. Virol.* 69, 141-149.

Lohmann, R., Schneider, G., Behrens, D. und Wrede, P. (1994) A neural network model for the prediction of membrane-spanning amino acid sequences. *Protein Sci.* 3, 1597-1601.

Long, E.O. (1998) Signal sequences stop killer cells. *Nature* 391,740-741, 743.

Lottspeich, F. und Zorbas, H. *Bioanalytik.* Spektrum Akademischer Verlag GmbH, Heidelberg, Berlin. (1998).

Maglott, D., Ostell, J., Pruitt, K.D. und Tatusova, T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 35, D26-D31.

Marcelino, A.M. und Gierasch, L.M. (2008) Roles of beta-Turns in protein folding: from peptide models to protein engineering. *Biopolymers* 89, 380-391.

Marti, M., Good, R.T., Rug, M., Knuepfer, E. und Cowman, A.F. (2004) Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* 306, 1930-1933.

Martoglio, B., Graf, R. und Dobberstein, B. (1997) Signal peptide fragments of preprolactin and HIV-1 p-gp160 interact with calmodulin. *EMBO J.* 16, 6636-6645.

Matsunaga, E., Tauszig-Delamasure, S., Monnier, P.P., Mueller, B.K., Strittmatter, S.M., Mehlen, P. und Chédotal, A. (2004) RGM and its receptor neogenin regulate neuronal survival. *Nat. Cell Biol.* 6, 749-755.

Matsunaga, E., Nakamura, H. und Chédotal, A. (2006) Repulsive guidance molecule plays multiple roles in neuronal differentiation and axon guidance. *J. Neurosci.* 26, 6082-6088.

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

Meissner, M., Koch, O., Klebe, G. und Schneider, G. (2008) Prediction of turn types in protein structure by machine-learning classifiers. *Proteins* Epub ahead of print.

Meldolesi, J. and Pozzan, T. (1998) The endoplasmic reticulum Ca2+ store: a view from the lumen. *Trends Biochem. Sci.* 23, 10-14.

Miller, J.R., Kovacevic, S. und Veal, L.E. (1987) Secretion and processing of staphylococcal nuclease by Bacillus subtilis. *J. Bacteriol.* 169, 3508-3514.

Mokranjac, D. und Neupert, W. (2007) Protein import into isolated mitochondria. *Meth. Mol. Biol.* 372, 277-286.

Monnier, P.P., Sierra, A., Macchi, P., Deitinghoff, L., Andersen, J.S., Mann, M., Flad, M., Hornberger, M.R., Stahl, B., Bonhoeffer, F. und Mueller, B.K. (2002) RGM is a repulsive guidance molecule for retinal axons. *Nature* 419, 392-395.

Nakai, K. und Horton, P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24, 34-36.

Nicholls, D.G. (1974) The influence of respiration and ATP hydrolysis on the proton-electrochemical gradient across the inner membrane of rat-liver mitochondria as determined by ion distribution. *Eur. J. Biochem.* 50, 305-315.

Nielsen, H., Engelbrecht, J., Brunak, S. und von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10, 1-6.

Nielsen, H. und Krogh, A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 6, 122-130.

Nohl, H. und Gille, L. (2005) Lysosomal ROS formation. *Redox Rep.* 10, 199-205.

Ogata, R.T., Mathias, P., Bradt, B.M. und Cooper, N.R. (1993) Murine C4b-binding protein. Mapping of the ligand binding site and the N-terminus of the pre-protein. *J. Immunol.* 150, 2273-2280.

Ouzzine, M., Magdalou, J., Burchell, B. und Fournel-Gigleux, S. (1999) An internal signal sequence mediates the targeting and retention of the human UDP-glucuronosyltransferase 1A6 to the endoplasmic reticulum. *J. Biol. Chem.* 274, 31401-31409.

Patron, N.J. und Waller, R.F. (2007) Transit peptide diversity and divergence: A global analysis of plastid targeting signals. *Bioessays* 29, 1048-1058.

Pearson, W.R. und Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.U.S.A.* 85, 2444-2448.

Pfeffer, S.R. und Rothman, J.E. (1987) Biosynthetic protein transport and sorting by the endoplasmic reticulum and Golgi. *Annu. Rev. Biochem.* 56, 829-852.

Prinz, A., Hartmann, E. und Kalies, K.U. (2000a) Sec61p is the main ribosome receptor in the endoplasmic reticulum of Saccharomyces cerevisiae. *Biol. Chem.* 381, 1025-1029.

Prinz, A., Behrens, C., Rapoport, T.A., Hartmann, E. und Kalies, K.U. (2000b) Evolutionarily conserved binding of ribosomes to the translocation channel via the large ribosomal RNA. *EMBO J.* 19, 1900-1906.

Pugsley, A.P. (1990) Translocation of proteins with signal sequences across membranes. *Curr. Opin. Cell Biol.* 2, 609-616.

Rabiner, L. R. (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE* 77, 257-286.
Rassokhin, D.N. und Agrafiotis, D.K. (2000) Kolmogorov-Smirnov statistic and its application in library design. *J. Mol. Graphics Modell.* 18, 368–382.

Resch, E., Quaiser, S., Quaiser, T., Schneider, G., Starzinski-Powitz, A. und Schreiner, A. (2008) Synergism of shrew-1's signal peptide and transmembrane segment required for plasma membrane localization. *Traffic* 9, 1344-1353.

Robertson, N.G., Lu, L., Heller, S., Merchant, S.N., Eavey, R.D., McKenna, M., Nadol, J.B. Jr., Miyamoto, R.T., Linthicum, F.H. Jr., Lubianca Neto, J.F., Hudspeth, A.J., Seidman, C.E., Morton, C.C. und Seidman, J.G. (1998) Mutations in a novel cochlear gene cause DFNA9, a human nonsyndromic deafness with vestibular dysfunction. *Nat. Genet.* 20, 299-303.

Roise, D., Theiler, F., Horvath, S.J., Tomich, J.M., Richards, J.H., Allison, D.S. und Schatz G. (1988) Amphiphilicity is essential for mitochondrial presequence function. *EMBO J.* 7, 649-653.

Rubartelli, A. und Sitia, R. , in Kuchler,K., Rubartelli A. and Holland B.I. (eds), *Unusual Secretory Pathways: from Bacteria to Man: Secretion of Mammalian Proteins that Lack a Signal Sequence*. Landes, Austin, TX, 87–114. (1997).

Stanke, N., Stange, A., Lüftenegger, D., Zentgraf, H. und Lindemann D. (2005) Ubiquitination of the prototype foamy virus envelope glycoprotein leader peptide regulates subviral particle release. *J. Virol.* 79, 15074-15083.

Schneider, G. und Broger, C. (1999) Visualizing sequence space by self-organizing feature maps: Classification of protein targeting signals. *Endocytobiology* 7, 589-602.

Schneider, G., Röhlk, S. und Wrede, P. (1993) Analysis of cleavage-site patterns in protein precursor sequences with a perceptron-type neural network. *Biochem. Biophys. Res. Commun.* 194, 951-959.

Schneider, G. und Wrede, P. (1994) The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. *Biophys. J.* 66, 335-344.

Schneider, G., Schuchhardt, J. und Wrede P. (1995) Peptide design in machina: development of artificial mitochondrial protein precursor cleavage sites by simulated molecular evolution. *Biophys. J.* 68, 434-447.

Schneider, G., Sjöling, S., Wallin, E., Wrede, P., Glaser, E. und von Heijne, G. (1998) Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides. *Proteins* 30, 49-60.

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

Schneider, G. und Fechner, U. (2004) Advances in the prediction of protein targeting signals. *Proteomics* 4, 1571-1580.

Schreiner, A., Ruonala, M., Jakob, V., Suthaus, J., Boles, E., Wouters, F. und Starzinski-Powitz, A. (2007) Junction protein shrew-1 influences cell invasion and interacts with invasion-promoting protein CD147. *Mol. Biol. Cell* 18, 1272-1281.

Schrempf, S., Froeschke, M., Giroglou, T., von Laer, D. und Dobberstein, B. (2007) Signal peptide requirements for lymphocytic choriomeningitis virus glycoprotein C maturation and virus infectivity. *J. Virol.* 81, 12515-12524.

Siegel, V. und Walter, P. (1988) Each of the activities of signal recognition particle (SRP) is contained within a distinct domain: analysis of biochemical mutants of SRP. *Cell* 52, 39-49.

Sigrist, C.J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. und Bucher, P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* 3, 265-274.

Silva-Filho, M.C. (2003) One ticket for multiple destinations: dual targeting of proteins to distinct subcellular locations. *Curr. Opin. Plant. Biol.* 6, 589-595.

Stewart, R.S., Drisaldi, B. and Harris, D.A. (2001) A transmembrane form of the prion protein contains an uncleaved signal peptide and is retained in the endoplasmic Reticulum. *Mol. Biol. Cell* 12, 881-889.

Stryer, L., *Biochemie. 4. Auflage.* Spektrum Akademischer Verlag, Heidelberg-Berlin (1996).

Tabak, H.F., van der Zand, A. und Braakman, I. (2008) Peroxisomes: minted by the ER. *Curr. Opin. Cell Biol.* 20, 393-400.

Tepass, U., Theres, C. und Knust, E. (1990) Crumbs encodes an EGF-like protein expressed on apical membranes of Drosophila epithelial cells and required for organization of epithelia. *Cell* 61, 787-799.

Toenjes, R.R., Czauderna, F. und Kurth, R. (1999) Genome wide screening, cloning, chromosomal assignment and expression of full-length human endogenous retrovirus type K (HERV-K) *J. Virol.* 73, 9187-9195.

Trexler, M., Bányai, L. und Patthy, L. (2000) The LCCL module. *Eur. J. Biochem.* 267, 5751-5757.

Turner, G., Barbulescu, M., Su, M., Jensen-Seaman, M.I., Kidd, K.K. und Lenz, J. (2001) Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr. Biol.* 11, 1531-1535.

Ulbrecht, M., Martinozzi, S., Grzeschik, M., Hengel, H., Ellwart, J.W., Pla, M. und Weiss, E.H. (2000) Cutting edge: the human cytomegalovirus UL40 gene product contains a ligand for HLA-E and prevents NK cell-mediated lysis. *J. Immunol.* 164, 5019-5022.

von Heijne G. (1983) Patterns of amino acids near signal-sequence cleavage sites. *Eur. J. Biochem.* 133, 17-21.

von Heijne, G. (1984a) Analysis of the distribution of charged residues in the N-terminal region of signal sequences: implications for protein export in prokaryotic and eukaryotic cells. *EMBO J.* 3, 2315–2318.

von Heijne G. (1984b) How signal sequences maintain cleavage specificity. *J. Mol. Biol.* 173, 243-251.

von Heijne, G. (1985) Signal sequences. The limits of variation. *J. Mol. Biol.* 184, 99-105.

von Heijne, G. (1986a) Mitochondrial targeting sequences may form amphiphilic helices. *EMBO J.* 5, 1335-1342.

von Heijne, G. (1986b) A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.* 14, 4683-4690.

Walter, P. und Blobel, G. (1980) Purification of a membrane-associated protein complex required for protein translocation across the endoplasmic reticulum. *Proc. Nat. Acad. Sci. U.S.A.* 77, 7112-7116.

Walter, P. (1992) Protein translocation. Travelling by TRAM. *Nature* 357, 22-23.

Walter, P. und Johnson, A.E. (1994) Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. *Annu. Rev. Cell Biol.* 10, 87-119.

Watzke, H.H., Wallmark, A., Hamaguchi, N., Giardina, P., Stafford, D.W. und High, K.A. (1991) Factor XSanto Domingo. Evidence that the severe clinical phenotype arises from a mutation blocking secretion. *J. Clin. Invest.* 88, 1685-1689.

Weihofen, A., Binns, K., Lemberg, M.K., Ashman, K. und Martoglio, B. (2002) Identification of signal peptide peptidase, a presenilin-type aspartic protease. *Science* 296, 2215-2218.

Wells, T.J., Tree, J.J., Ulett, G.C. und Schembri, M.A. (2007) Autotransporter proteins: novel targets at the bacterial cell surface. *FEMS Microbiol. Lett.* 274, 163-172.

Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A. und Wagner, L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* 31, 28-33.

Wiech, H., KIappa, P. und Zimmermann, R. (1991) Protein Export in Prokaryotes and Eukaryotes. *FEBS Lett.* 285, 182-188.

Wiemann, S., Weil, B., Wellenreuther, R., Gassenhuber, J., Glassl, S., Ansorge, W., Böcher, M., Blöcker *et al*. (2001) Towards a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.* 11, 422-435.

Wolin, S.L. und Walter, P. (1993) Discrete nascent chain lengths are required for the insertion of presecretory proteins into microsomal membranes. *J. Cell Biol.* 121, 1211-1219.

Wrede, P., Landt, O., Klages, S., Fatemi, A., Hahn, U. und Schneider, G. (1998) Peptide design aided by neural networks: biological activity of artificial signal peptidase I cleavage sites. *Biochemistry* 37, 3588-3593.

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

Yen, Y.T. und Stathopoulos, C. (2007) Identification of autotransporter proteins secreted by type V secretion systems in gram-negative bacteria. *Methods Mol. Biol.* 390, 33-46.

Zdobnov, E.M., Lopez, R., Apweiler, R. und Etzold, T. (2002a) The EBI SRS server--recent developments. *Bioinformatics* 18, 368-373.

Zdobnov, E.M., Lopez, R., Apweiler, R. und Etzold T. (2002b) The EBI SRS server-new features. *Bioinformatics* 18, 1149-1150.

Zhang, Q., Yoon, S. und Welsh, W.J. (2005) Improved method for predicting beta-turn using support vector machine. *Bioinformatics* 21, 2370-2374.

Zhang, X., Kung, S. und Shan, S.O. (2008) Demonstration of a multistep mechanism for assembly of the SRP x SRP receptor complex: implications for the catalytic role of SRP RNA. *J. Mol. Biol.* 381, 581-593.

Anhang

Anhang A1: Vertrag über die Nutzung von SignalP und TargetP als Programmteile der webbasierten NtraC-Benutzeroberfläche.

Agreement on the use of SignalP and TargetP in a public web interface

between

Prof. Dr. H. Nielsen, K. Rapacki, Prof. Dr. S. Brunak
Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark,
DK-2800 Kongens Lyngby, Denmark

and

Jan A. Hiss, Prof. Dr. G. Schneider
Chem- & Bioinformatics, Beilstein Endowed Chair for Cheminformatics
Johann Wolfgang Goethe-University, Siesmayerstr. 70, D-60323 Frankfurt am Main, Germany

Description: The sole aim of the web interface is to provide free public access to a prediction method for long signal sequences. Long signal sequences often show a standard overall NHC partitioning but they are also known to posses additional functions besides ER targeting. Hiss/Schneider developed a method to decrypt these long signal peptides by dissecting the signal sequence in a special way. Typically, this yields two fragments or "domains". Usually, one domain resembles a standard signal peptide and a second domain impedes ER targeting. A web-interface will be developed for making predictions (that is, long signal fragmentation and domain analysis). For the analysis part, it is intended to use the software SignalP and TargetP (developed by H. Nielsen, S. Brunak, K. Rapacki).

- The web interface will be hosted on the university web server of the Schneider group in Frankfurt, Germany. The method will only be available as a web interface, and not distributed as a stand-alone package, neither for commercial nor non-commercial uses.
- SignalP and TargetP will be run by a call from wrapper software developed by Hiss/Schneider. There will be no possibility to influence the parameters of SignalP and TargetP for the user of the web interface.
- The results of SignalP and TargetP will be provided to the user "as is" without modification.
- If a user provides a sequence he/she may enter the position of the known signal peptidase cleavage site. If the site is unknown, automatic prediction using SignalP for the first 100 residues is performed (signalp -t euk -trunc 100 -short -q -d output.txt SignalP_Input.txt). The same adapted call is performed if the user chooses gram positive or gram negative as realm of origin of his sequence.
- After dissection of the long signal sequence the fragments are checked for their potential targeting capacity using SignalP and TargetP (same system call as above but without the 100 residue truncation). If the sequences are predicted as mTP, SP, SA, gram+ or gram- signal (a value of 0.4 must be reached) the user gets a feedback as to the domain structure of the long signal sequence.
- SignalP and TargetP publications will be properly acknowledged and quoted on the web site and a link will be placed pointing to the official SignalP and TargetP prediction sites.

Frankfurt, 4 January 2008

J. A. Hiss

G. Schneider

Kongens Lyngby, 22 January 2008

H. Nielsen

S. Brunak

K. Rapacki

Anhang A2: Oligonucleotid-Sequenzen für die SEAP Konstrukte. Entnommen aus Hiss *et al*., 2008b. Die Restriktionsendonucleasen-Schnittstellen NotI, Acc65I, EcoRI, XhoI und HindIII wurden für die Klonierung in den pcDNA3.1(-) (Invitrogen, Karlsruhe) Vektor eingebracht und sind jeweils unterstrichen.

NotI SH^{C}-$SEAP^{\Delta SP}$:
5'-TTGCGGCCGCATGCCCCTCGGAAGCCATGCCTGG-3'.

NotI SH^{N}-$SEAP^{\Delta SP}$:
5'-TGCGGCCGCATGTGGATTCAACAGCTTTTAGGACTCAGCTCCATGT
CCATCCGCTGGCCGGGCCGCGAATTCATCATCCCAGTTGAG-3'.

NotI SH^{G18I}-$SEAP^{\Delta SP}$:
5'-GCGGCCGCATGTGGATTCAACAGCTTTTAGGACTCAGCTCCATGTC
CATCCGCTGGCCGATCCGCCCCCTCGGAAGCC-3'.

NotI $SH^{\Delta WPGR}$-$SEAP^{\Delta SP}$:
5'-TTGCGGCCGCATGTGGATTCAACAGCTTTTAGGACTCAGCTCCATGT
CCATCCGCCCCCTCGGAAGCCATGCCTGG-3'.

NotI $SH^{\Delta WPGR/mut}$-$SEAP^{\Delta SP}$:
5'-TTGCGGCCGCATGTGGATTCAACAGCTTTTAGGACTCAGCTCCATGT
CCATCGCCCCCTCGGACCATGCCTGG-3'.

Acc65I SEAP Myc-tagging:
5'-TTGGTACCTTACAGATCCTCTTCTGAGATGAGTTTTTGTTCACCCGG
GTGCGCGGCGTCG-3'.

SH-$SEAP^{\Delta SP}$ diente für alle Konstrukte (auch DCBD2 und RGMA) als Vorlage. Es wurde durch Fusion mit dem shrew-1 Signalpeptid und der Amplifizierung durch PCR erzeugt.

NotI shrew-1 SP 5'-TTTGCGGCCGCATGTGGATTCAACAGCTT-3',
und EcoRI shrew-1 SP 5'-TTGAATTCGCCCAGGGCCTCGCAGGC-3' und SEAP, ohne endogenes Signalpeptid, wurden durch PCR mithilfe der folgenden Primer amplifiziert:

EcoRI 5'-TTGAATTCATCATCCCAGTTGAGGAG-3', und HindIII
5'-TTTTAAGCTTTTAACCCGGGTGCGCGGC-3' an der EcoRI- Schnittstelle.

Der Myc-*tag* des Vorlagen-Konstruktes wurde durch den folgenden Primer (Acc65I SEAP Myc-tag) erreicht:
5'-TTGGTACCTTACAGATCCTCTTCTGAGATGAGTTTTTGTTCACCCGGG
TGCGCGGCGTCG-3'.

$SEAP^{\Delta SP}$ wurde durch den folgenden Xhol Start-Codon Primer generiert.
$SEAP^{\Delta SP}$ 5'–TTTCTCGAGATGATCATCCCAGTTGAGGAG-3' und
HindIII 5'-TTTTAAGCTTTTAACCCGGGTGCGCGGC-3'.

Tabelle A3: Mit SRS aus UniProtKB (Version 13.6) entnommene eukaryotische Signalsequenzen ≥40 aus nicht-putativen Proteinen (N = 136).

SRS Nummer	Signalsequenz Annotation	Signalsequenz
ADA23_HUMAN	potentiell	MKPPGSSSRQPPLAGCSLAGASCGPQRGPAGSVPASAPARTPPCRLLLVLLLLPPLAAS
AGAL_ORYSJ	nicht-potentiell	MARASSSSSPPSPRLLLLLLVAVAATLLPEAAALGNFTAESRGARWRSRRARRRA
ASM_HUMAN	nicht-potentiell	MPRYGASLRQSCPRSGREQGQDGTAGAPGLLWMGLVLALALALALA
ATS1_HUMAN	potentiell	MQRAVPEGFGRRKLGSDMGNAERAPGSRSFGPVPTLLLLAAALLAVSDA
ATS1_MOUSE	potentiell	MQPKVPLGSRKQKPCSDMGDVQRAARSRGSLSAHMLLLLLASITMLLC
ATS4_BOVIN	potentiell	MSHMDSHPGRGLADGWLWGIQPRLLLPTVPVSGSRLVWLLLLASLLPSAWP
ATS4_HUMAN	potentiell	MSQTGSHPGRGLAGRWLWGAQPCLLLPIVPLSWLVWLLLLLLASLLPSARL
C163A_HUMAN	potentiell	MSKLRMVLLEDSGSADFRRHFVNLSPFTITVVLLLSACFVT
C163B_HUMAN	nicht-potentiell	MMLPQNSWHIDFGRCCCHQNLFSAVVTCILLLNSCFLISS
C4BPA_HUMAN	nicht-potentiell	MHPPKTPSGALHRKRKMAAWPFSRLWKVSDPILFQMTLIAALLPAVLG
C4BPA_MOUSE	nicht-potentiell	MCAKQQQTLLPTRAAHGRLHRNRDVVAWPFSTLCRVSGPTLFQMTFTAALWVAVFG
CADE_DROME	potentiell	MSTSVQRMSRSYHCINMSATPQAGHLNPAQQQTHQQHKRKCRDLGRRLIPARLLLGVIVAISLLSPALA
CADM1_HUMAN	potentiell	MASVVLPSGSQCAAAAAAAAPPGLRLRLLLLLFSAAALIPTGDG
CADM1_MOUSE	potentiell	MASAVLPSGSQCAAAAAVAAAAAPPGLRLRLLLLLLSAAALIPTGDG
CATL_DROME	potentiell	MNHLGVFETRFRPRTRHKSQRAQLIPEQITMRTAVLLPLLALLAVAQA
CLN5_HUMAN	potentiell	MAQEVDTAQGAEMRRGAGAARGRASWCWALALLWLAVVPGWSRVSG
CR1_HUMAN	nicht-potentiell	MGASSPRSPEPVGPPAPGLPFCCGGSLLAVVVLLALPVAWG
CRB_DROME	nicht-potentiell	MAKIANASLSQQQKQRQAETATTTTTTVAASVETATTTARSRDRTKSAAQITSHLLKRAISVYSSPQWIPLFILIYLATDVASVAVPT
CRRY_MOUSE	potentiell	MEVSSRSSEPLDPVWLLVAFGRGGVKLEVLLLFLLPFTLG
CXCL5_MOUSE	nicht-potentiell	MSLQLRSSAHIPSGSSSPFMRMAPLAFLLLFTLPQHLAEA
DCBD2_HUMAN	potentiell	MASRAVVRARRCPQCPQVRAAAAAPAWAALPLSRSLPPCSNSSSFSMPLFLLLLLVLLLLLEDAGA
DPY10_CAEEL	potentiell	MKNNAKEDYRTFSLTTNYSRQMIYRCVTGLQIGFSLFSFIIVCVA
E134_MAIZE	nicht-potentiell	MPSSAQVLLCLAAVLAAAAATTAEAHSQCLDNPPDRSIHGRQL
EMP46_YEAST	nicht-potentiell	MTTRKTASSLQLLGKITGTKAGTKQKKMNFINGLIWLYMCVWMVHG
ENK1_HUMAN	potentiell	MHPSEMQRKAPPRRRRHRNRAPLTHKMNKMVTSEQMKLPSTKKAEPPTWAQLKKLTQLATKYLENTKVTQTPESMLLAALMIVSMVVS
ENK2_HUMAN	potentiell	MNPSEMQRKAPPRRRRHRNRAPLTHKMNKMVTSEEQMKLPSTKKAEPPTWAQLKKLTQLATKYLENTKVTQTPESMLLAALMIVSMVVS
ENK3_HUMAN	potentiell	MNPSEMQRKAPPRRRRHRNRAPLTHKMNKMVTSEEQMKLPSTKKAEPPTWAQLKKLTQLATKYLENTKVTQTPESMLLAALMIVSMVVS
ENK4_HUMAN	potentiell	MNPSEMQRKAPPRRRRHRNRAPLTHKMNKMVTSEEQMKLPSTKKAEPPTWAQLKKLTQLATKYLENTKVTQTPESMLLAALMIVSMVVS

ENK5_HUMAN	potentiell	MNPSEMQRKAPPRRRRHRNRAPLTHKMNKMVTSEEQMKLPSTKKAEPPTWAQLKKLTQLATKYLENTKVTQTPESMLLAALMIVSMVVS
ENK6_HUMAN	potentiell	MNPSEMQRKAPPRRRRHRNRAPLTHKMNKMVTSEEQMKLPSTKKAEPPTWAQLKKLTQLATKYLENTKVTQTPESMLLAALMIVSMVVS
EXG1_COCCA	nicht-potentiell	MRFSSLLACLGAVGIQAAAIPFQRRVDNTTDSGSLDAAQAAA
FCG2B_HUMAN	potentiell	MGILSFLPVLATESDWADCKSPQPWGHMLLWTAVLFLAPVAG
FCG2C_HUMAN	potentiell	MGILSFLPVLATESDWADCKSPQPWGHMLLWTAVLFLAPVAG
FREM2_HUMAN	potentiell	MHSAGTPGLSSRRTGNSTSFQPGPPPPPRLLLLLLLLLSLVSRVPA
FZD1_HUMAN	potentiell	MAEEEAPKKSRAAGGGASWELCAGALSARLAEEGSGDAGGRRRPPVDPRRLARQLLLLLWLLEAPLLLG
GABR2_HUMAN	potentiell	MASPRSSGQPGPPPPPPPPPARLLLLLLLPLLLPLAPGAWG
GABR2_RAT	potentiell	MASPPSSGQPRPPPPPPPPPARLLLPLLLSLLLWLAPGAWG
GBRAL_DROME	potentiell	MCTMPATRDASGSGDASTDLIAARSLSSHQGQRSNLRIFKLLISCCLLMLCIYPNAWP
GRP78_YEAST	nicht-potentiell	MFFNRLSAGKLLVPLSVVLYALFVVILPLQNSFHSSNVLVRG
GUC2D_BOVIN	nicht-potentiell	MTACTFLAGGLRDPGLCGPTRWAPSPPGLPPIPPRPRLRLRPPLLLLLLLPRSVLS
GUC2D_HUMAN	potentiell	MTACARRAGGLPDPGLCGPAWWAPSLPRLPRALPRLPLLLLLLLLLQPPALS
GUC2E_RAT	potentiell	MSAWLLPAGGFPGAGFCIPAWQSRSSLSRVLRWPGPGLPGLLLLLLLPSPSAFS
GUC2G_MOUSE	potentiell	MASRTRSESPLEPRLYAGAGSRADHPSLVLMLSVVMLVTCLEA
HEXB_HUMAN	potentiell	MELCGLGLPRPPMLLALLLATLLAAMLALLTQVALVVQVAEA
HYAL1_MOUSE	potentiell	MLGLTQHAQKVWRMKPFSPEVSPGSSPATAGHLLRISTLFLTLLELAQVCRG
ICOSL_MOUSE	durch Vergleich	MQLKCPCFVSLGTRQPVWKKLHVSSGFFSGLGLFLLLLSSLCAASA
IL24_HUMAN	nicht-potentiell	MNFQQRLQSLWTLARPFCPPLLATASQMQMVVLPCLGFTLLLWSQVSGAQG
IL9R_HUMAN	potentiell	MGLGRCIWEGWTLESEALRRDMGTWLLACICICTCVCLGV
INVB_DAUCA	potentiell	MDTYHFLPSRDLEHASSYTPRPDSPETRHEPDPDRSKTNRRPIKIVSSVLLSTLILS
ITA4_MOUSE	potentiell	MFSTKSAWLRNGGADQGPRGIALREAVMLLLYFGVPTGPS
ITA5_HUMAN	nicht-potentiell	MGSRTPESPLHAVQLRWGPRRRPPLLPLLLLLLPPPPRVGG
ITB8_HUMAN	potentiell	MCGSALAFFTAAFVCLQNDRRGPASFLWAAWVFSLVLGLGQG
KIRR1_MOUSE	potentiell	MTLESPSTRLMTCQSSLLPEKPRFLSQKMWAPHLVVAYLIFVTLALA
KLK11_HUMAN	potentiell	MQRLRWLRDWKSSGRGLTAAKEPGARSSPLQAMRILQLILLALATGLVGG
LAMA5_MOUSE	nicht-potentiell	MAKRGGQLCAGSAPGALGPRSPAPRPLLLLLAGLALVGEA
LAR_CAEEL	potentiell	MIQFRNKNNSMNRIARHLRNVARRKGSSLLLFLMLSTVLVAA
LIFR_HUMAN	potentiell	MMDIYVCLKRPSWMVDNKRMRTASNFQWLLSTFILLYLMNQVNS
LIFR_MOUSE	potentiell	MAAYSWWRQPSWMVDNKRSRMTPNLPWLLSALTLLHLTMHANG
LMA2L_HUMAN	potentiell	MAATLGPLGSWQQWRRCLSARDGSRMLLLLLLLGSGQGPQQVGA
LMAN2_CANFA	nicht-potentiell	MAAEGWIWRWGWGRRCLGRPGLPGPGPGPATPLFLLLLLGPVVA

LMAN2_HUMAN	potentiell	MAAEGWIWRWGWGRRCLGRPGLLGPGPGPTTPLFLLLLLGSVTA
LRC55_HUMAN	nicht-potentiell	MGSLQHCCCLLPKMGDTWAQLPWPGPPHPAMLLISLLLAAGLMHSDA
LRIG2_HUMAN	potentiell	MAPAPLGVPEEQLLGCRSRVLSRLLFIAQTALLLLPAAGA
LTBP3_HUMAN	potentiell	MPGPRGAAGGLAPEMRGAGAAGLLALLLLLLLLLLGLGGRVEG
LY9_MOUSE	nicht-potentiell	MADLKRYWCDWALGPLSENPRMSQQQIFSPILWIPLLFLLMGLGASG
LYAM3_HUMAN	nicht-potentiell	MANCQIAILYQRFQRVVFGISQLLCFSALISELTNQKEVAA
LYAM3_MOUSE	potentiell	MAGCPKGSWTPRLRSVILGGAQLIWFSALISELVNQKEVAA
MA2B1_BOVIN	nicht-potentiell	MVGDARPSGVRAGGCRGAVGSRTSSRALRPPLPPLSSLFVLFLAAPCAWA
MA2B1_CAVPO	potentiell	MGASVLPLGLGAGDCQSSSGRRMSACLPRTALSFLLSLLLATPGARA
MA2B1_HUMAN	nicht-potentiell	MGAYARASGVCARGCLDSAGPWTMSRALRPPLPPLCFFLLLLAAAGARA
MCP_BOVIN	potentiell	MRASCTPLKAPLRRPERLASSGRFAWVLLLAPLLLLPTSSDA
MCP_MOUSE	potentiell	MTAAPLMPDSTHPCRRRKSYTFFWCSLGVYAEALLFLLSHLSDA
MCP_PIG	nicht-potentiell	MMAFCALRKALPCRPENPFSSRCFVEILWVSLALVFLLPMPSDA
MCP_RAT	potentiell	MTAAPLTPDPTHPRRRRKSYTFFSLGIYAEALLFLLSSLSDA
MMP15_HUMAN	potentiell	MGSDPSAPGRPGWTGSLLGDREEAARPRLLPLLLVLLGCLG
MMP24_HUMAN	potentiell	MPRSRGGRAAPGPPPPPPPPGQAPRWSRWRVPGRLLLLLLPALCCLPGAARA
MMP24_MOUSE	potentiell	MPRSRGGRAAPGQASRWSGWRAPGRLLPLLPALCCLAAAAG
MPRI_BOVIN	potentiell	MEAAAGRSSHLGPAPAGRPPRCPLLLQLQLLLLLLLLPPGWVPG
MPRI_HUMAN	nicht-potentiell	MGAAAGRSPHLGPAPARRPQRSLLLLQLLLLVAAPGSTQA
MS57C_DROME	potentiell	MPINDFISCYLKQLQRISIVSIHQVVKMHGTHFLILLLLCGVLG
NCLN_HUMAN	potentiell	MLEEAGEVLENMLKASCLPLGFIVFLPAVLLLVAPPLPAADA
NEUR1_HUMAN	nicht-potentiell	MTGERPSTALPDRRWGPRILGFWGGCRVWVFAAIFLLLSLAASWSKA
NFAM1_HUMAN	potentiell	MENQPVRWRALPGLPRPPGLPAAPWLLLGVLLLPGTLRLAGG
NGL1_HUMAN	potentiell	MLNKMTLHPQQIMIGPRFNRALFDPLLVVLLALQLLVVAGLVRA
NGL1_MOUSE	potentiell	MLNKMTLHPQQIMIGPRFNRALFDPLLVVLLALQLLVVAGLVRA
NLGN1_HUMAN	potentiell	MALPRCTWPNYVWRAVMACLVHRGLGAPLTLCMLGCLLQAGHVLS
NLGN1_MOUSE	potentiell	MALPRCMWPNYVWRAMMACVVHRGSGAPLTLCLLGCLLQTFHVLS
NLGN1_RAT	potentiell	MALPRCMWPNYVWRAMMACVVHRGSGAPLTLCLLGCLLQTFHVLS
NLGNX_HUMAN	nicht-potentiell	MSRPQGLLWLPLLFTPVCVMLNSNVLLWLTALAIKFTLIDS
NOTCH_DROME	potentiell	MQSQRSRRRSRAPNTWICFWINKMHAVASLPASLPLLLLTLAFANLPNTVRG
NPW_RAT	potentiell	MDLSALASSREVRGPGPGAPVNRPLLPLLLLLLLLPLPASA
NRX1B_RAT	nicht-potentiell	MYQRMLRCGAELGSPGGGSSGGAGGRLALLWIVPLTLSGLLGVAWG

NRX2B_HUMAN	durch Vergleich	MPPGGSGPGGCPRRPPALAGPLPPPPPPPPPLLPLLPLLLLLLLGAAEG
NRX2B_RAT	nicht-potentiell	MPPGGSGQGGCPRRPPALAGPLPPPPPPPPLPLLLGLLLLLGAAEG
NUDEL_DROME	potentiell	MNYNMDEMEATRLLRHPRRWWSIGFGKRIVAISILVIIVLLFS
P30_TOXGO	potentiell	MSVSLHHFIISSGFLTSMFPKAVRRAVTAGVFAAPTLMSFLRCGVMA
PCSK6_HUMAN	potentiell	MPPRAPPAPGPRPPPRAAAATDTAAGAGGAGGAGGAGGPGFRPLAPRPWRWLLLLALPAACSA
PERM_HUMAN	nicht-potentiell	MGVPFFSSLRCMVDLGPCWAGGLTAEMKLLLALAGLLAILATPQPSEG
PLBL2_CAEEL	potentiell	MTRLIRSKKQFLIRSLHSVFYYLGSLLHSTFEMNVFIGLLLA
PLBL2_HUMAN	nicht-potentiell	MVGQMYCYPGSHLARALTRALALALVLALLVGPFLSGLAGA
PLBL2_MOUSE	nicht-potentiell	MAAPVDGSSGGWAARALRRALALTSLTTLALLASLTGLLLSGPAGA
PLXB3_HUMAN	potentiell	MCHAAQETPLLHHFMAPVMARWPPFGLCLLLLLLSPPPLPLTGA
PLXD1_HUMAN	potentiell	MAPRAAGGAPLSARAAAASPPPFQTPPRCPVPLLLLLLLGAARAGA
PME3_CITSI	potentiell	MTRIKEFFTKLSESSTNQNISNIPKKKKKLFLALFATLLVVAAVIGIVAG
PTP10_DROME	potentiell	MLYQLSKATTRIRLKRQKAVPQHRWLWSLAFLAAFTLKDVRC
PTPRN_RAT	durch Vergleich	MRRPRRPGGPAGCGGSEGSGGLRLLVCLLLLSGRPGGCSA
PVRL3_HUMAN	potentiell	MARTLRPSPLCPGGGKAQLSSASLLGAGLLLQPPTPPPLLLLLFPLLLFSRLCGALA
PVRL3_MOUSE	potentiell	MARTPGPAPLCPGGGKAQLSSAFPPAAGLLLPAPTPPPLLLLLIPLLLFSRLCGALA
QSOX1_CHICK	potentiell	MWRRRARSGGGGGGGGGAAPRCRWWPAVLALLAAALPAARS
R4RL2_HUMAN	potentiell	MLPGLRRLLQAPASACLLLMLLALPLAAPSCPMLCTCYSSPPTVSC
RGMA_HUMAN	potentiell	MQPPRERLVVTGRAGWMGMGRGAGRSALGFWPTLAFLLCSFPAATSP
ROR2_DROME	potentiell	MAAGQWVGVVERVLRGMVLKWGANLAVLGLCVFLFASATHA
SAS_DROME	potentiell	MQTCRRRKASGGQSTIKWSRMCLATLCGLLLLGIQIERAAS
SCA_DROME	potentiell	MRDWQTFPDLQKKKVSRDHLNCPATMAGSNVLWPILLAVVLLQISVAFVSG
SDK_DROME	potentiell	MLKSAASSLRRRRPKTTITATLAIEMPSQPKLASLLAVLVLLCYCDS
SELN_HUMAN	potentiell	MGRARPGQRGPPSPGPAAQPPAPPRRRARSLALLGALLAAAAA
SEM7A_HUMAN	potentiell	MTPPPPGRAAPSAPRARVPGPPARLGLPLRLRLLLLLWAAAASA
SERR_DROME	potentiell	MFRKHFRRKPATSSSLESTIESADSLGMSKKTATKRQRPRHRVPKIATLPSTIRDCRSLKSACNLIALILILLVHKISA
SLIK5_HUMAN	potentiell	MHTCCPPVTLEQDLHRKMHSWMLQTLAFAVTSLVLSCAET
SODE_ONCVO	potentiell	MINSFIVIFLSFLIFINYANLVCVEATHVYGRRSHSNGMHGN
SORC2_HUMAN	potentiell	MAHRGPSRASKGPGPTARAPSPGAPPPPRSPRSRPLLLLLLLLGACGAAG
SPS2_YEAST	potentiell	MPIWKTQTFFTSISVIQIVNKETKVSTKKEKDSMLNQLNTILRFLFLFLQLIKSSA
TEFF2_HUMAN	nicht-potentiell	MVLWESPRQCSSWTLCEGFCWLLLLPVMLLIVARPVKLAA
THS7A_HUMAN	potentiell	MGLQARRWASGSRGAAGPRRGVLQLLPLPLPLPLLLLLLLRPGAGRA

TNR21_HUMAN	potentiell	MGTSPSSSTALASCSRIARRATATMIAGSLLLLGFLSTTTA
TR10B_HUMAN	nicht-potentiell	MEQRGQNAPAASGARKRHGPGPREARGARPGPRVPKTLVLVVAAVLLLVSAESAL
TR10D_HUMAN	nicht-potentiell	MGLWGQSVPTASSARAGRYPGARTASGTRPWLLDPKILKFVVFIVAVLLPVRVDS
TYRO3_HUMAN	potentiell	MALRRSMGRPGLPPLPLPPPPRLGLLLAALASLLLPESAA
UNC5C_MOUSE	potentiell	MRKGLRATAARCGLGLGYLLQMLVLPALALLSASGTGSAA
VAS1_HUMAN	potentiell	MMAAMATARVRMGPRCAQALWRMPWLPVFLSLAAAAAAAAA
VIT1_AEDAE	nicht-potentiell	MATDGITSRFGFNERRRTHNRNSCRILEDKMLAKLLLLALAGLTAA
VLDLR_CHICK	potentiell	MRSSRQRGDRSAATGGGCGARRWALPRCGALCLLLALGCLRTA
XYN1_TRIRE	potentiell	MVAFSSLICALTSIASTLAMPTGLEPESSVNVTERGMYDFVLGAHNDHRRR
YF1M_CAEEL	potentiell	MNSFLFGFLNLINVLKINYLQLMRRGCRHHLAAVVLLIATFPPLAYN
YP003_HUMAN	potentiell	MGAQGAQESIKAMWRVPGTTRRPVTGESPGMHRPEAMLLLLTLALLGGPTWA
Q41038_PEA	potentiell	MASACASSAIAAVAISTPSSQKNGSPSGTSKAFLGRKLKVNSSTASPS

Tabelle A4: Virale Signalpeptide ≥ 40 Aminosäuren (N = 11). Nur nicht-putative Proteindaten wurden verwendet. Sequenzen aus UniProtKB (Version 14.0) entnommen. Keine identischen Sequenzen aufgeführt.

SRS Nummer	Signalsequenz Annotation	Sequenz
ARC3_CBCP	nicht-potentiell	MKGIRKSILCLVLSAGVIAPVTTSIVQSPQKCYACTVDKG
YP_398578	nicht-potentiell	MKGLRKSILCLVLSAGVIAPVTSGMIQSPQKCYAYSINQK
ENV_CAEVG	nicht-potentiell	MDAGASYMRLTGEENWVEVTMDEEKERKGKDVQQGKYRPQVSKPIINRDTNTSFAYKGIFLWGIQITMWILLWTNMCVRA
ENV_JSRV	potentiell	MPKRRAGFRKGWYARQRNSLTHQMQRMTLSEPTSELPTQRQIEALMPYAWNEAHVQPPVTPTNILIMLLLLLQRVQNGAAAAFW
ENV_MMTVC	nicht-potentiell	MPNHQSGSPTGSSDLLLDGKKQRAHLALRRKRRREMRKINRKVRRMNLAPIKEKTAWQHLQALIFEAEEVLKTSQTPQTSLTLF LALLSVLGPPPVSG
ENV_MMTVG	durch Vergleich	MPNHQSGSPTGSSDLLLSGKKQRPHLALRRKRRREMRKINRKVRRMNLAPIKEKTAWQHLQALISEAEEVLKTSQTPQNSLTLF LALLSVLGPPPVTG
ENV_RSVP	potentiell	MRRALFLQAFLTGYPGKTSKKDSKEKPLATSKKDPEKTPLLPTRVNYILIIGVLVLCEVTGVRA
ENV_RSVSA	potentiell	MEAVIKAFLTGYPGKTSKKDSKEKPLATSKKDPEKTPLLPTRVNYILIIGVLVLCEVTGVRA
ENV_VILV	potentiell	MASKESKPSRTTWRDMEPPLRETWNQVLQELVKRQQQEEEEQQGLVSGKKKSWVSIDLLGTEGKDIKKVNIWEPCEKWFAQVVWG VLWVLQIVLWGCLMWEVRKGN
ENV_VILVK	Potentiell	MASKESKPSRTTRRGMEPPLRETWNQVLQELVKRQQQEEEEQQGLVSGKKKSWVSIDLLGTEGKDIKKVNIWEPCEKWFAQVVWG VLWVLQIVLWGCLMWEVRKGN
FUS_CDVO	Potentiell	MHRGIPKSSKTQTHTQQDRPPQPSTELEETRTSRARHSTTSAQRSTHYDPRTSDRPVSYTMNRTRSRKQTSHRLKNIPVHGNHEA TIQHIPESVSKGARSQIERRQPNAINSGSHCTWLVLWCLGMASLFLCSKA

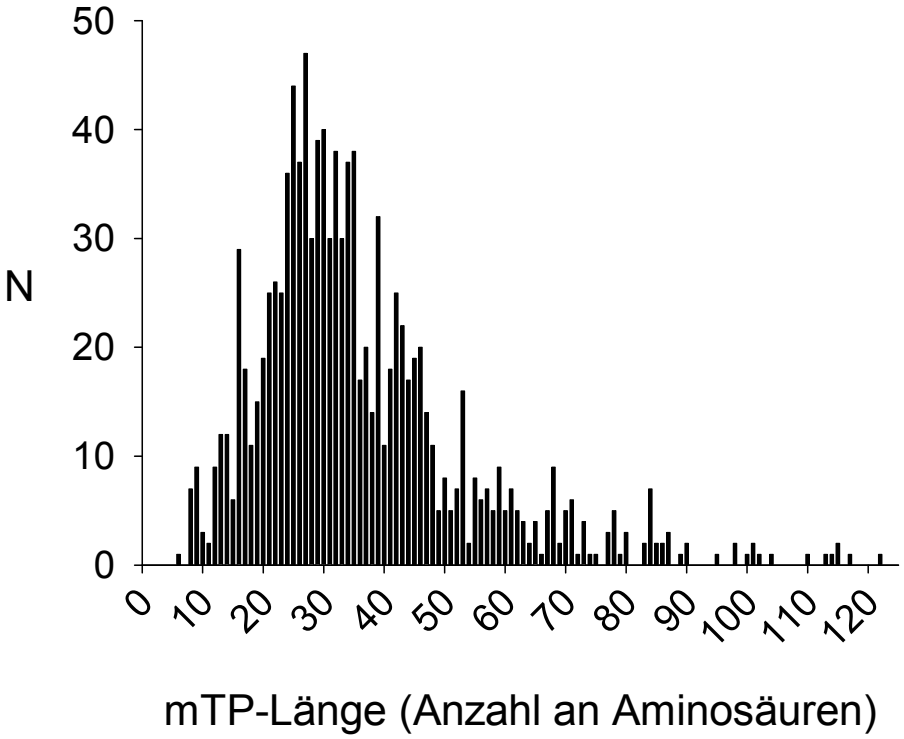


Abbildung A5: Längenverteilung **mitochondrialer** Targeting-Peptide (mTP, *N* = 1.101) nicht-putativer Proteine. Sequenzen aus UniProtKB-Version 13.6 entnommen.

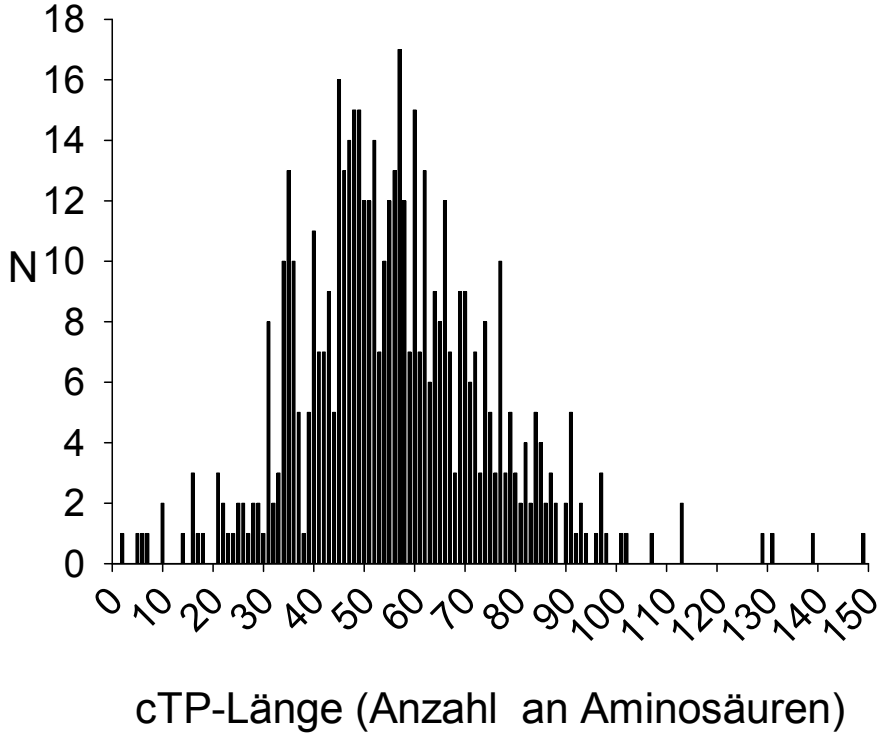


Abbildung A6: Längenverteilung von **Chloroplasten-Transit-Peptiden** (*N* = 514) nicht-putativer Proteine. Sequenzen aus UniProtKB-Version 13.6 entnommen.

Tabelle A7: Vertebrata-Signalpeptide mit ≥ 40 Aminosäuren, die eine NtraC-Organisation abweichend zu shrew-1 und DCBD2 aufweisen. Entnommen und angepasst aus Hiss *et al*., 2008b. Unterstrichene Aminosäuren stellen potentielle Übergangsbereiche nach dem NtraC-Modell dar.

NCBI Accession Number	Signal peptide sequence
Q60813	MSVAAAGRGFASSLSSPQIRRIALKEAKLTPHIWAALHWNLGLRLVPSVRVGILVLLIFLPSTFC
P08607	MCAKQQQTLLPTRAAHGRLHRNRDVVAWPFSTLCRVSGPTLFQMTFTAALWVAVFG
P61569	MNPSEMQRKAPPRRRRHRNRAPSSHKMNKMMMSEEQMKLPSTNKAEPLTWAQLNKLTQLATKCLENTKMTQTPESMLLAALMIVSTVVS
P61570	MNPSEMQRKAPPRRRRHRNRAPLTHKMNKMVTSEEQMKLPSTKKAEPPTWAQLKKLTQLATKYLENTKVTQTPESMLLAALMIVSMVVS
Q6ZVY0	MNPSEMQGKAPPQRQRTRNRTSLTRRVNKMVISEEQMKLPSTKKAGPPTWAQLKKLTQLAEKSLENTRVTQTPENKLLAALMIVSTVVS
P61565	MHPSEMQRKAPPRRRRHRNRAPLTHKMNKMVTSEQMKLPSTKKAEPPTWAQLKKLTQLATKYLENTKVTQTPESMLLAALMIVSMVVS
Q69384	MNPSEMQRKAPPRRRRHRNRAPLTHKMNKMVTSEEQMKLPSTKKAEPPTWAQLKKLTQLATKYLENTKVTQTPESMLLAALMIVSMVVS
O71037	MNPSEMQRKAPPRRRRHRNRAPLTHKMNKMVTSEEQMKLPSTKKAEPPTWAQLKKLTQLATKYLENTKVTQTPESMLLAALMIVSMVVS
Q9UKH3	MNPSEMQRKAPPRRRRHRNRAPLTHKMNKMVTSEEQMKLPSTKKAEPPTWAQLKKLTQLATKYLENTKVTQTPESMLLAALMIVSMVVS
Q902F9	MNPSEMQRKAPPRRRRHRNRAPLTHKMNKMVTSEEQMKLPSTKKAEPPTWAQLKKLTQLATKYLENTKVTQTPESMLLAALMIVSMVVS
Q902F8	MNPSEMQRKAPPRRRRHRNRAPLTHKMNKMVTSEEQMKLPSTKKAEPPTWAQLKKLTQLATKYLENTKVTQTPESMLLAALMIVSMVVS
Q9UKH7	MNPSEMQRKAPPRRRRHCNRAPLTHKMNKMVTSEEEMKLPSTKKAEPLTWAQLKKLTQLATKYLENTKVTQTPESMLLAALMIVSMVVS
Q9UBN6	MGLWGQSVPTASSARAGRYPGARTASGTRPWLLDPKILKFVVFIVAVLLPVRVDS
O75077	MKPPGSSSRQPPLAGCSLAGASCGPGRGPAGSVPASAPARTPPCRLLLVLLLLPPLAAS
Q9R1V7	MKPPGSISRRPTLTGCSLPGASCGPGRCPAGPVPARAPPCRLLLVLLLLPALATS
P33727	MGRRGAASLPRPGSPRRPLLPGVLPLLLRLLLLPSRPGAGA
P50429	MGKLSPCTGRSRPGGPGPQLPLLLLLLQLLLLLLSPARASG
Q5FYB0	MAPRGCAGHPPPPSPQACVCPGKMLAMGALAGFWILCLLTYGYLSWGQA
Q8BM89	MAPRDSAEPLPPLSPQAWAWSGKFLAMGALAGFSVLSLLTYGYLCWG
Q8TE60	MECALLLACAFPAAGSGPPRGLAGLGRVAKALQLCCLCCASVAAALA
Q4VC17	MECALLCLCALRAAGPGPPWGPAGLGRLAKALQLCCFCCASVAVALA
Q9UHI8	MQRAVPEGFGRRKLGSDMGNAERAPGSRSFGPVPTLLLLAAALLAVSDA
P97857	MQPKVPLGSRKQKPCSDMGDVQRAARSRGSLSAHMLLLLLASITMLLC
Q9WUQ1	MQPEVPLGSGKLKPCSDMGDIQRAAKFRSSQSAHMLLLLLASITMLLCVRGAHG
Q9TT93	MSHMDSHPGRGLADGWLWGIQPRLLLPTVPVSGSRLVWLLLLASLLPSAWP
O75173	MSQTGSHPGRGLAGRWLWGAQPCLLLPIVPLSWLVWLLLLLLASLLPSARL
Q8BNJ2	MSQMGLHPRRGLTGHWLQRFQPCLPLHTVQWRRLLLLAFLLSLAWPASP
Q5RFQ8	MSQTGSHPGRGLAGRWLWGAQPCLLLPIVPLSWLVWLLLLLLASLLPSARL

Q2VLG4	MSKLRMVLLEDSGSADVRRHFVNLSPFTIAVVLLLRACFVTSSLG
Q2VL90	MDKLRMVLHENSGSADFRRCSAHLSSFTFAVVAVLSACLVTSSLGG
Q9NR16	MMLPQNSWHIDFGRCCCHQNLFSAVVTCILLLNSCFLISS
Q9BY67	MASVVLPSGSQCAAAAAAAAPPGLRLRLLLLLFSAAALIPTGDG
Q8R5M8	MASAVLPSGSQCAAAAAVAAAAAPPGLRLRLLLLLLSAAALIPTGDG
Q5RBP6	MGVKAAQTGIWASQQQSIRVVGFQAQTAHRAICLLGFVVLVLLQCCSA
P17927	MGASSPRSPEPVGPPAPGLPFCCGGSLLAVVVLLALPVAWG
Q8AWW5	MYLAAVSAGRRRPGGDGGGGGGGWHLAAAGWLLLLALLLGQPGTRA
Q9Y426	MAMARLGSWLGEAQWLALVSLFVAALATVGLYLAQWALARA
Q05754	MQPHLSHQPCWSLPSPSVREAASMYGTAVAIFLVILVAALQG
Q9H6D8	MPSGCHSSPPSGLRGDMASLVPLSPYLSPTVLLLVSCDLGFVRA
Q5SZK8	MHSAGTPGLSSRRTGNSTSFQPGPPPPPRLLLLLLLLLSLVSRVPA
O57328	MAERRGPAGGGSGEVGGRRAGGDRCPRRPPALPLLLLLWAAALPAGG
Q9UP38	MAEEEAPKKSRAAGGGASWELCAGALSARLAEEGSGDAGGRRRPPVDPRRLARQLLLLLWLLEAPLLLG
O70421	MAEEAAPSESRAAGRLSLELCAEALPGRREEVGHEDTASHRRPRADPRRWASGLLLLLWLLEAPLLLG
Q08463	MAEEAVPSESRaagrPSLELCAVAlPGRREEVGHQDTAGHRRPRAHSRCWARGLLLLLWLLEAPLLLG
Q5ZLR1	MAGAIIENMSTRKLCIVGGILLVFQVIAFLVGGLIAPSPTTA
Q5T9L3	MAGAIIENMSTKKLCIVGGILLVFQIIAFLVGGLIAPGPTTA
Q6DID7	MAGAIIENMSTKKLCIVGGILLVFQIVAFLVGGLIAPAPTTA
Q5R9R3	MAGAIIENMGTKKLCIVGGILLVFQIIAFLVGGLIAPGPTTA
Q6P689	MAGAIIENMSTKKLCIVGGILLVFQIVAFLVGGLIAPAPTTA
P55203	MTACTFLAGGLRDPGLCGPTRWAPSPPGLPPIPPRPRLRLRPPLLLLLLLPRSVLS
O19179	MSACALLAGGLPDPRLCAPARWARSPPGVPGAPPWPQPRLRLLLLLLLLPPSALS
Q02846	MTACARRAGGLPDPGLCGPAWWAPSLPRLPRALPRLPLLLLLLLLQPPALS
P51839	MAGLQQGCHPEGQDWTAPHWKTCRALPGPRGLTVRHLRTVSSISVFSVVFWGVLLWADSLSLPAWA
O02740	MFLAPWPFSHLMLWFVTLGRQRGQHGLASFKLLWCLWLLVLMSLPL
Q5SDA5	MFLGPWPFSRLLSWFAISSRLSGQHGLPSSKFLRCLCLLALLPLLRWGQA
P51842	MFLGPWPFSRLLSWFAISSRLSGQHGLTSSKFLRYLCLLALLPLIWWGQA
Q6TL19	MASRTRSESPLEPRLYAGAGSRADHPSLVLMLSVVMLVTCLEA
P55205	MASRARSEPPLEHRFYGGAESHAGHSSLVLTLFVVMLMTCLEA
Q91ZJ9	MLGLTQHAQKVWRMKPFSPEVSPGSSPATAGHLLRISTLFLTLLELAQVCRG
Q9JHJ8	MQLKCPCFVSLGTRQPVWKKLHVSSGFFSGLGLFLLLLSSLCAASA

Q01113	MGLGRCIWEGWTLESEALRRDMGTWLLACICICTCVCLGV
Q00651	MFSTKSAWLRNGGADQGPRGIALREAVMLLLYFGVPTGPS
Q6UXG2	MAEPGHSHHLSARVRGRTERRIPRLWRLLLWAGTAFQVTQG
Q8NC54	MAAAVPKRMRGPAQAKLLPGSAIQALVGLARPLVLALLLVSAALSSVVS
Q9QYN3	MRRLKSDWKLSTETREPGARPALLQARMILRLIALALVTGHVGG
Q61001	MAKRGGQLCAGSAPGALGPRSPAPRPLLLLLAGLALVGEA
Q2KIY5	MVAPMYGSPGGRLARAVTRALALALVLALLVGLFLSGLTGA
P42703	MAAYSWWRQPSWMVDNKRSRMTPNLPWLLSALTLLHLTMHANG
O70535	MGAFSWWRQPSWMADNKRGRMTPLPWLLSALTLLHLMMHVNG
P49256	MAAEGWIWRWGWGRRCLGRPGLPGPGPGPATPLFLLLLLGPVVA
Q12907	MAAEGWIWRWGWGRRCLQRPGLLGPGPGPTTPLFLLLLLGSVTA
Q9DBH5	MAAEAWLWRWGWGWGQRCPGRPGLPGPGPSPTTFLHLLLLLGPVAA
Q6ZSA7	MGSLQHCCCLLPKMGDTWAQLPWPGPPHPAMLLISLLLAAGLMHSDA
Q3UY51	MGSLQHCCCQLPKMGDTWAQLPWPGPPHSALLLVFFLLAAGVMHSDA
Q4KLL3	MGSLQYCCCQLPKMGDTWAQLPWPGPPHSALLLVFFLLAAGVMHSDA
Q9NS15	MPGPRGAAGGLAPEMRGAGAAGLLALLLLLLLLLLGLGGRVEG
Q29451	MVGDARPSGVRAGGCRGAVGSRTSSRALRPPLPPLSSLFVLFLAAPCAWA
Q8VHC8	MGASVLPLGLGAGDCQSSSGRRMSACLPRTALSFLLSLLLATPGARA
O46432	MGADARPLGVRAGGGGRGAARPGTSSRALPPPLPPLSFLLLLLAAPGARA
O00754	MGAYARASGVCARGCLDSAGPWTMSRALRPPLPPLCFFLLLLAAAGARA
O09159	MGTGPLTSGVRAGGGNTGWLWMSSCNLGSPVLPISFLFWLLLAAPGARA
Q641Q3	MRGAARAAWGRAGQPWPRPPAPGPPPPPLPLLLLLLAGLLGGAGA
Q9Y5R2	MPRSRGGRAAPGPPPPPPPPGQAPRWSRWRVPGRLLLLLLPALCCLPGAARA
P08169	MEAAAGRSSHLGPAPAGRPPRCPLLLQLQLLLLLLLLPPGWVPG
Q9IAL7	MALCKKTVGSVLEEWCLNEPLFGCKRHQNVRKKLRLIRIIGLLVSVVAISTFSLSISA
Q9UI40	MDLQQSTTITSLEKWCLDESLSGCRRHYSVKKKLKLIRVLGLFMGLVAISTVSFSISA
O54701	MDLHQSATVRLLQEWCSHESPSGCRRHYNTRKKLKLIRVIGLVMGLVAVSTVPFSISA
Q969V3	MLEEAGEVLENMLKASCLPLGFIVFLPAVLLLVAPPLPAADA
Q8VCM8	MLEEAGEVLENVLKASCLPLGFIVFLPAVLLLVAPPLPAADA
Q5XIA1	MLEEAGEVLENVLKASCLPLGFIVFLPAVLLLVAPPLPAADA
Q99519	MTGERPSTALPDRRWGPRILGFWGGCRVWVFAAIFLLLSLAASWSKA
O35657	MVGADPTRPRGPLSYWAGRRGQGLAAIFLLLVSAAESEARA

Q8NET5	MENQPVRWRALPGLPRPPGLPAAPWLLLGVLLLPGTLRLAGG
Q9HCJ2	MLNKMTLHPQQIMIGPRFNRALFDPLLVVLLALQLLVVAGLVRA
Q8C031	MLNKMTLHPQQIMIGPRFNRALFDPLLVVLLALQLLVVAGLVRA
Q8N2Q7	MALPRCTWPNYVWRAVMACLVHRGLGAPLTLCMLGCLLQAGHVLS
Q99K10	MALPRCMWPNYVWRAMMACVVHRGSGAPLTLCLLGCLLQTFHVLS
Q62765	MALPRCMWPNYVWRAMMACVVHRGSGAPLTLCLLGCLLQTFHVLS
Q8K1M5	MDLSALASSREVRGPGPGAPVNRPLLPLLLLLLLLPLPASA
Q28142	MYQRMLRCGAELGSPGGGGGGGGRLALLWIVPLTLSGLLGVAWG
P58400	MYQRMLRCGAELGSPGGGGGGGGGGGAGGRLALLWIVPLTLSGLLGVAWG
Q63373	MYQRMLRCGAELGSPGGGSSGGAGGRLALLWIVPLTLSGLLGVAWG
Q96JQ0	MQKELGIVPSCPGMKSPRPHLLLPLLLLLLLLLGAGVPGAWG
Q63415	MPPRAPPAPGPRPPPRAAGRHGLSPLAPRPWRWLLLLALPAVCSA
P05164	MGVPFFSSLRCMVDLGPCWAGGLTAEMKLLLALAGLLAILATPQPSEG
Q63259	MRRPRRPGGPAGCGGSEGSGGLRLLVCLLLLSGRPGGCSA
Q8BQC3	MAEPRTASPRRLPALRRPGFLPPLLPPPPPLLLLLLLLPLPAPSLG
Q9JLB9	MARTPGPAPLCPGGGKAQLSSAFPPAAGLLLPAPTPPPLLLLLIPLLLFSRLCGALA
Q8JGM4	MWRRRARSGGGGGGGGGGAAPRCRWWPAVLALLAAALPAARS
O60895	MASLRVERAGGPRLPRTRVGRPAALRLLLLLGAVLNPHEALA
Q9WUP0	MAPLRVERAPGGSRLGVTRAQRPTALCLPPLLLLLLLLLGAVSA
Q9JHJ1	MAPLRVERAPGGSQLAVTSAQRPAALRLPPLLLLLLLLLLGAVST
Q6NW40	MGLRAAPSSAAAAAAEVEQRRRPGLCPPPLELLLLLLFSLGLLHA
Q7TQ33	MGVRAAPSCAAAPAAAGAEQSRRPGLWPPSPPPPLLLLLLLSLGLLHA
P50228	MSLQLRSSAHIPSGSSSPFMRMAPLAFLLLFTLPQHLAEA
Q64519	MKPGPPRRGTAQGQRVDTATHAPGARGLLLPPLLLLLLAGRAAG
P33671	MKPGPPRRGTAQGQRVDTATHGPGARGLLLPPLLLLLLAGRAAG
Q8AV58	MVGRKVDREIIARRNSRRDGMMMKLNFCFFFCRRWWAFLLLQLHMLQALA
Q9W6G6	MKTAGEPDRRRQRRQVRTGRFSCAWWSTSVMLFFSLPEGNC
Q9Z123	MLARAERPRPGPRPPPVSLFPPPSSLLLLLLAMLSAPVCG
Q96PQ0	MAHRGPSRASKGPGPTARAPSPGAPPPPRSPRSRPLLLLLLLLGACGAAG
Q9EPR5	MAHRGPPSAPKRPGPTAPDRSFQALLPPCWPRSWPLLLLLLVLVAACGA
Q96GP6	MEGAGPRGAGPARRRGAGGPPSPLLPSLLLLLLLWMLPDTVAP
Q9UIK5	MVLWESPRQCSSWTLCEGFCWLLLLPVMLLIVARPVKLAA

Q9QYM9	MVLWESPRQCSSWTLCEGFCWLLLLPVTLLIIARPVKLAA
Q6R5N8	MSGLYRILVQLEQSPYVKTVPLNMRRDFFFLVVTWMPKTVKMNGSSFVPSLQLLLMLVGFSLPPVAET
Q6ZP80	MRLNIAIFFGALFGALGVLLFLVAFGSDYWLLATEVGRCSG
O14763	MEQRGQNAPAASGARKRHGPGPREARGARPGPRVPKTLVLVVAAVLLLVSAESAL
Q9QZM4	MEPPGPSTPTASAAARADHYTPGLRPLPKRRLLYSFALLLAMLQAVFVPVTA
P49744	MTMITPSSKLTLTKGNKSWSSTRCGAFLLLHLVLQPWQRAGA
O95185	MRKGLRATAARCGLGLGYLLQMLVLPALALLSASGTGSAA
O08747	MRKGLRATAARCGLGLGYLLQMLVLPALALLSASGTGSAA
Q761X5	MRKGLRATAARCGLGLGYLLQMLVLPALALLSASGTGSAA
Q15904	MMAAMATARVRMGPRCAQALWRMPWLPVFLSLAAAAAAAAA
P98165	MRSSRQRGDRSAATGGGCGARRWALPRCGALCLLLALGCLRTA
P31286	MRKNLWTFQFGGSGLVGSAMVSQHFVVLLMSLYCLTQS
Q9ULT6	MRPRSGGRPGATGRRRRRLRRRPRGLRCSRLPPPPPLPLLLGLLLAAAGPGAARA
Q5SSZ7	MRPRSGGRPGAPGRRRRRLRRGPRGRRLPPPPPLPLLLGLLLAAAGPGAARA
Q9BS86	MEAFALGPARRGRRRTRAAGSLLSRAAILLFISAFLVRVPSSVG
Q62522	MEALAPGRAPRGRRRAGASGSVLSPLSLAAVLLCALLRAPPAVG
Q2YHT5	MEISQQAGWCKKPASPMNTRAALEAVRNTAWTIVLLTSAAVMGAS
Q90YI0	MNFTEGCEATGRRPGSAGSRRRRAPRPGPVALLPLLLPLLLPPAAAV
Q0VD19	MPRHGVSPGQGLPRSGREQASDRSLGAPCLRLLWLGLALA
Q04519	MPHHRASSGQDHLRAGWEQRLERSLPAPRVGLLWMGLGLALVLA
Q86VB7	MSKLRMVLLEDSGSADFRRHFVNLSPFTITVVLLLSACFVT
Q28065	MKHQRVPVMILHSKGTMASWPFSRLWSISDPILFQVTLVATLLATVLG
P60509	MDPLHTIEKVPARRNIHDRGHQGHRMGDGTPGRPKISVQQMTRFSLIIFFLSAPFVVNA
P42702	MMDIYVCLKRPSWMVDNKRMRTASNFQWLLSTFILLYLMNQVNS
P51511	MGSDPSAPGRPGWTGSLLGDREEAARPRLLPLLLVLLGCLG
Q5VX71	MYHGMNPSNGDGFLEQQQQQQQPQSPQRLLAVILWFQLALC
Q8BH32	MYHGMNPSNGDGFLEQQLQQQQPQSPQRLLAVILWFQLALC
Q5R8M2	MYHGMNPSNGDGFLEQQQQQQPQSPQRLLAVILWFQLALC
Q8WTU2	MHKEAEMLIGPQLDEKRWGWRLGGDSAAPPFLPQALSFLLLLPL

Tabelle A8: 32 Vertebrata-Signalpeptide mit ≥ 40 Aminosäuren, die eine NtraC-Organisation entsprechend shrew-1, DCBD2 und RGMA aufweisen (N-Domäne: mTP, C-Domäne: SP). Entnommen und angepasst aus Hiss *et al*., 2008b. Unterstrichene Aminosäuren stellen potentielle Übergangsbereiche nach dem NtraC-Modell dar.

NCBI Accession Nummer	Sequenz
P70505	MSVAASASRSASTLCSPQIQQGALKEAKVPPHIWAARHWNLGLRLVPGHASVRAGILVLLIFLPSTLC
P17405	MPRYGASLRQSCPRSGREQGQDGTAGAPGLLWMGLVLALALALALA
Q96PD2	MASRAVVRARRCPQCPQVRAAAAAPAWAALPLSRSLPPCSNSSSFSMPLFLLLLLVLLLLLEDAGA
Q91ZV3	MASRAPLRAARSPQGPGGPAAPAATGRAALPSAGCCPLPPGRNSSSRPRLLLLLLLLLQDAGG
Q91ZV2	MASRAPLRAARSPQDPGGRAAPAATGRAPLPSAGWCPLPPGRNSSSRPRLLLLLLLLLPDAGA
Q28110	MGIPSFLAFPAARRNRAHCTPWHPWGHMLLWTALLFLAPVSG
Q1LZH9	MRLLSLAPDRPRRGGPRHLTSGSPALPPPPPLLLLLLLGGCLGVSGA
P50426	MRFLSLAPDRPRRGGPRHLPSGSPAPPPPPPLLLLLLGGCLGVSGA
P52785	MSAWLLPAGGLPGARFCVPARQSPSSFSRVLRWPRPGLPGLLLLLLLPSPSALS
P51840	MSAWLLPAGGFPGAGFCIPAWQSRSSLSRVLRWPGPGLPGLLLLLLLPSPSAFS
P51841	MFLGLGRFSRLVLWFAAFRKLLGHHGLASAKFLWCLCLLSVMSLPQQVWT
Q8K201	MAASALGRMCGAAREKLSPGPGARGLGALARSLVLALLLVPVLC
Q5R5B8	MAAAALKRMRGPAQAKLLPGSAIQALVGLARPLVLALLLVSAALSSVVS
Q9UBX7	MQRLRWLRDWKSSGRGLTAAKEPGARSSPLQAMRILQLILLALATGLVGG
Q5XNR9	MMNISLRLRRPPWMVDSNGRRMTSHFQWLLLTFILLYLMNQVTS
Q9H0V9	MAATLGPLGSWQQWRRCLSARDGSRMLLLLLLLGSGQGPQQVGA
P59481	MAAASRPSWWQRWRRRAWARDGAKLLLFLLLLGSGPGPRHVRA
Q6VE48	MRASCTPLKAPLRRPERLASSGRFAWVLLLAPLLLLPTSSDA
Q8VE43	MRGAVWAARRRAGQQWPRSPGPGPGPPPPPPLLLLLLLLLGGASA
Q5RJL6	MRGVVWAARRRAGQQWPRSPGPGPGPPPPPPLLLLLLLLLGGASA
Q9R0S2	MPRSRGGRAAPGQASRWSGWRAPGRLLPLLPALCCLAAAAG
Q99PW6	MPRSRGGRAAPGQAARWSGWRAPGRLLPLLPALCCLAAAAG
P29122	MPPRAPPAPGPRPPPRAAAATDTAAGAGGAGGAGGAGGPGFRPLAPRPWRWLLLLALPAACSA
Q9NQS3	MARTLRPSPLCPGGGKAQLSSASLLGAGLLLQPPTPPPLLLLLFPLLLFSRLCGALA
Q96B86	MQPPRERLVVTGRAGWMGMGRGAGRSALGFWPTLAFLLCSFPAATSP
Q9N0A6	MGGPGPRRAGTSRERLVVTGRAGWMGMGRGAGRSALGFWPTLAFLLCSFPAAT
Q6PCX7	MQPPRERLVVTGRAGWMGMGRGAGRSALGLWPTLAFLLCSFPAAISP
Q9QUR8	MTPPPPGRAAPSAPRARVLSLPARFGLPLRLRLLLVFWVAAASA
Q9UPZ6	MGLQARRWASGSRGAAGPRRGVLQLLPLPLPLPLLLLLLLRPGAGRA
Q9EPU5	MGTRASSITALASCSRTAGQVGATMVAGSLLLLGFLSTITA
Q8IZC6	MGAGSARGARGTAAAAAARGGGFLFSWILVSFACHLASTQG
Q91443	MGRHSALGLSGNRQVSPCTGTRPFKVVGSRSPVQPLCILLALTVCIGTS

Publikationen

The *Plasmodium* Export Element Revisited

J. A. Hiss, J. M. Przyborski, F. Schwarte, K. Lingelbach, G. Schneider

PLoS ONE 3, e1560.

Signalsequenzen von *Plasmodium falciparum*, die das *Plasmodium Export Element* (PEXEL) enthalten und aus verschiedenen Proteinfamilien stammen, können basierend auf den das PEXEL-Motiv flankierenden Aminosäuren nach Protein-Familien unterschieden werden. Es zeigen sich in dieser Region für die Protein-Familien charakteristische Hydrophobizitäts-muster. Es wird gezeigt, dass es bei Proteinen mit PEXEL-Motiv und Signalpeptid durch die *in silico*-Abspaltung des Signalpeptides zu einer positionellen Überlagerung der PEXEL-Motive mit Proteinen mit PEXEL-Motiv, aber ohne natives Signalpeptid kommt.

OPEN ACCESS Freely available online PLoS one

The *Plasmodium* Export Element Revisited

Jan Alexander Hiss [1]*, Jude Marek Przyborski [2], Florian Schwarte [1], Klaus Lingelbach [2], Gisbert Schneider [1]

1 Johann Wolfgang Goethe-University, Institute of Cell Biology and Neuroscience, Centre for Membrane Proteomics, Frankfurt am Main, Germany, 2 Faculty of Biology, Philipps-University Marburg, Marburg, Germany

Abstract

We performed a bioinformatical analysis of protein export elements (PEXEL) in the putative proteome of the malaria parasite Plasmodium falciparum. A protein family-specific conservation of physicochemical residue profiles was found for PEXEL-flanking sequence regions. We demonstrate that the family members can be clustered based on the flanking regions only and display characteristic hydrophobicity patterns. This raises the possibility that the flanking regions may contain additional information for a family-specific role of PEXEL. We further show that signal peptide cleavage results in a positional alignment of PEXEL from both proteins with, and without, a signal peptide.

Citation: Hiss JA, Przyborski JM, Schwarte F, Lingelbach K, Schneider G (2008) The Plasmodium Export Element Revisited. PLoS ONE 3(2): e1560. doi:10.1371/journal.pone.0001560

Editor: Per Westermark, Uppsala University, Sweden

Received September 5, 2007; Accepted January 15, 2008; Published February 6, 2008

Funding: This research was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften, and the Centre for Membrane Proteomics at Goethe-University, Frankfurt am Main. J.M.P is supported by the DFG priority program (SPP) 1131.

Competing Interests: The authors have declared that no competing interests exist.

*E-mail: hiss@bioinformatik.uni-frankfurt.de

Introduction

Plasmodium falciparum (*P. falciparum*) is an intracellular parasite of the human red blood cell and the cause of the most virulent form of malaria. The severe pathology is at least partly a result of a modification of the host cell plasma membrane by proteins synthesized and exported by the parasite [1]. From a cell biological point of view this is unusual, as the non-infected human erythrocyte lacks a machinery to facilitate directed protein transport. An additional obstacle is the location of the parasite within a so-called "parasitophorous vacuole" (PV), which separates the parasite from the host cell cytosol, the vacuolar membrane thereby forming a further barrier for proteins destined for the host cell. In two elegant studies, Hiller *et al.* [2] and Marti *et al.* [3] identified a short peptide sequence, referred to as the vacuolar transport signal (VTS) or Plasmodium export element (PEXEL), respectively. This motif is frequently found in parasite proteins that are transported beyond the confines of the vacuolar membrane. Although VTS and PEXEL differ slightly in their structure, they share the conserved five-residue motif Rx(L,I)x(-D,E,Q).

Dominant protein families of the *P. falciparum* "exportome" are parasite-encoded surface proteins such as the erythrocyte membrane protein 1 (PfEMP1)–the major *P. falciparum* virulence factor [4–6]–and the RIFIN and STEVOR surface antigen families [7,8]. We found that 28% of the putative *P. falciparum* proteome contain the PEXEL/VTS pattern. This is a large number of proteins, and raises the question whether the presence of the motif is the sole defining criterion for exported parasite proteins. In fact, residues surrounding the PEXEL motif were found to be important in correct trafficking or folding of exported proteins [9,10], and a recent study suggests that the short pentameric core motif alone is insufficient to cause protein traffic across the PV membrane [11]. Apparently, additional factors need to be taken into account when predicting the size and members of the Plasmodium exportome as well as antigens at the surface of the infected erythrocyte. This hypothesis is substantiated by the observation that members of the RIFIN protein family locate in different cellular compartments despite the fact that all members of the RIFIN protein family contain a PEXEL sequence: A-type RIFINs are transported to the surface of infected erythrocytes via Maurer's clefts, whereas B-type RIFINs remain inside the parasite [12]. Wahlgren and coworkers already speculated that residue positions in the PEXEL motif and additional family-specific conserved stretches of amino acids are required for differential protein targeting [12], which is in agreement with the studies by Przyborski *et al.* [10]. One question arising from these preliminary findings is whether the PEXEL-flanking sequence regions contain family-specific information.

Results and Discussion

So motivated, we analyzed residue positions surrounding the PEXEL motif. We compiled a set of 5,571 unique proteins from *P. falciparum* extracted from PlasmoDB [13], TIGR/NCBI clone 3D7 [14], and EMBL-EBI [15]. Pattern matching with SEED-TOP [16] retrieved 1,557 (28%) sequences containing the PEXEL motif. 412 (7.4%) hits were found by a generalized Hidden-Markov-Model (false-positive rate: 5%), which requires, in addition to the PEXEL motif, a preceding hydrophobic region for prediction of exported proteins [17]. For further analysis, we extracted stretches of 25 amino acids from these 412 predicted proteins containing the central five-residue PEXEL motif and ten additional residues on both sides (data available as supplementary material). When multiple PEXEL motifs existed in one protein sequence, only the most N-terminal occurrence was extracted.

We performed all-against-all pair-wise alignment of the 25-residue fragments using BLAST ([16]; Gapped BLAST was run with the BLOSUM62 matrix, [18], and gap-open cost = 11, gap-elongation = 1). Only 6% of the sequences aligned to proteins

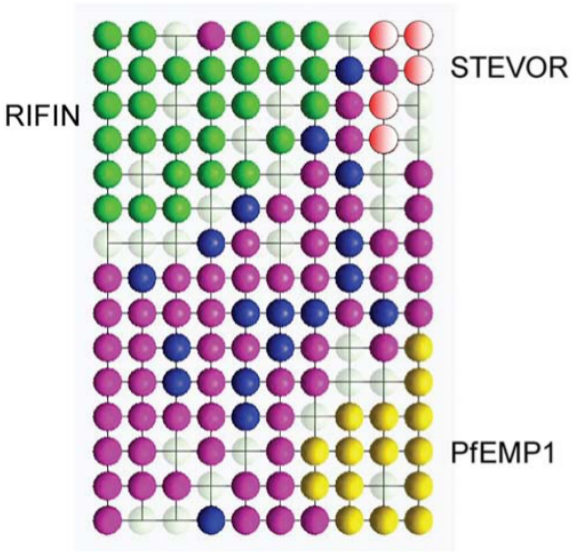


Figure 1. SOM projection of the PEXEL containing sequence fragments from *P.falciparum* proteins. The SOM contains 106 15 topologically ordered data clusters ("neurons"). Locations of RIFIN (green), STEVOR (red), and PfEMP1 (yellow) proteins are highlighted. The location of hypothetical proteins is shown in magenta. Blue color indicates "other" PEXEL-containing fragments. A neuron is assigned one particular class if more than 50% of its clustered proteins belong to one family. White neurons do not contain any proteins ("empty sequence space"). Map generated with the software SOMMER [27].
doi:10.1371/journal.pone.0001560.g001

outside their family, and 78% of all fragments aligned to sequences of the corresponding protein family with average E-values up to 0.1. These results indicate that the residues flanking the PEXEL motif contain family-specific information. It is evident that the shortness of the sequences used (25 residues), and the failure to align 22% of the sequence fragments limit this approach for general prediction of potentially exported proteins and protein family assignment. It has been argued before that straightforward sequence alignment may not be appropriate to find all members of the P.falciparum exportome because individual protein families are particularly deviating in their primary sequences, for example beta-barrel proteins from outer bacterial and organelle membranes [9].

In a complementary approach, we encoded the sequence fragments by seven physicochemical amino acid properties [19]: hydrophilicity [20] and hydrophobicity [21] scales, volume [22], surface [23], bulkiness, refractivity, and polarity [24]. This led to a 256 7 = 175-dimensional vectorial sequence representation. We employed Kohonen's self organizing map (SOM) technique [25] for visualizing the data distribution by nonlinear projection of this high-dimensional sequence space [26]. As a result of SOM training, the topology of the data distribution is shown on a two-dimensional map, and cluster formation of RIFIN, STEVOR, and PfEMP1 sequences is observed (Figure 1). The physicochemical sequence representation led to a reasonable grouping of the three dominant PEXEL-containing protein families.

Noteworthy, based on relative amino acid frequency only (calculated from full-length sequences), the three protein families cannot be distinguished (Kolmogorov-Smirnov test significance at the 5% level).

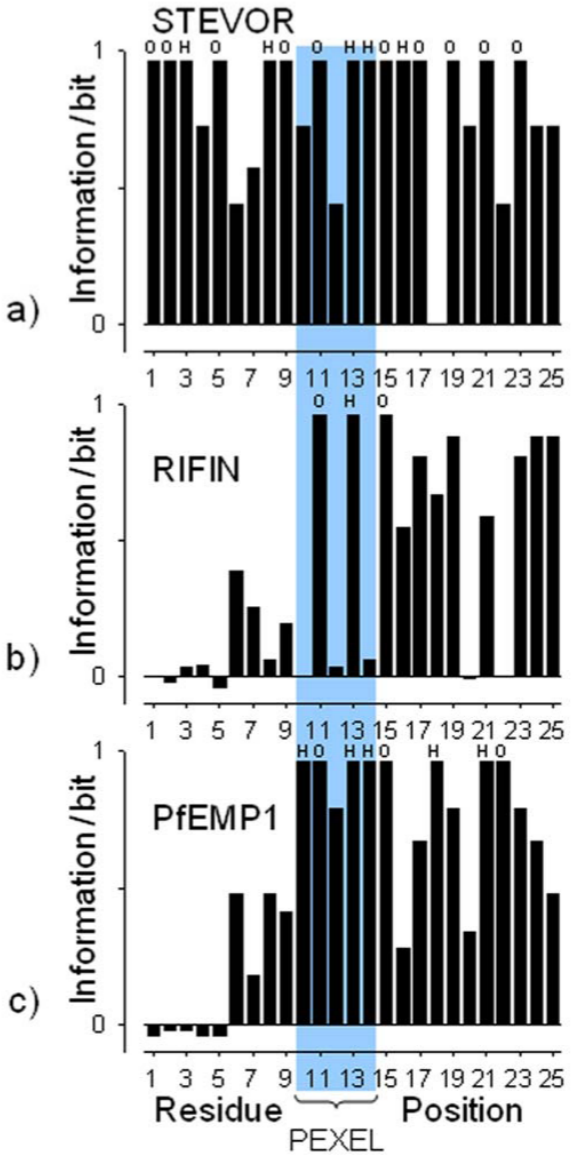


Figure 2. Occurrence of hydrophobic and hydrophilic residues up and downstream the PEXEL motif . Information plot of the PEXEL motif and surrounding residue positions in the protein families STEVOR (N = 30) (a), RIFIN (N = 125) (b) and PfEMP1 (N = 58) (c). Large values indicate sequence positions with conserved hydrophobic (H) or hydrophilic (O) residues (see text for residue classification). The position of the PEXEL motif is highlighted.
doi:10.1371/journal.pone.0001560.g002

The SOM is grounded on a non-deterministic process. Thus the projections slightly differ in repeated runs. We selected a SOM projection with a small mean quantization error. Clustering strength was evaluated by calculating the topological intrafamily distance of proteins on the SOM. We found that the PEXEL containing families RIFIN (average distance: 4.17), PfEMP1 (average distance: 2.99), and STEVOR (average distance: 2.20) have smaller pair-wise distances than the remaining "hypotheti-

cal" proteins (average distance: 5.99). This supports our notion that RIFIN, PfEMP1, and STEVOR form local distributions.

The RIFIN cluster contains two major sub-families, A-RIFIN and B-RIFIN proteins [28]. The average topological distance of SOM neurons containing B-RIFIN is 2.5, and for A-RIFIN 3.5, indicating that B-RIFIN proteins are more similar to each other than the A-RIFINs, based on the sequence fragments analyzed.

The SOM projection was then used to predict the family membership of the remaining 180 PEXEL-containing hypothetical P.falciparum proteins. A conservative prediction was performed as we focused only on neurons containing at least 50% members from one protein family. Among the candidate proteins, one co-localizes with the RIFIN, five with the STEVOR, and one with the PfEMP1 family on the SOM (see supplementary material). Noteworthy, these suggested assignments are based on the similarity of the PEXEL motif flanking regions only.

The hypothetical sequences that do not co-cluster with known members of the RIFIN, PfEMP1, and STEVOR families are not necessarily false-positives. They might belong to other PEXEL containing protein families. In our study, we focused only on the three dominant PEXEL-containing protein families from P. falciparum. For determination whether they represent actual false-positives with regard to intracellular localization, biological experiments are required. This is beyond the scope of the present study.

The formation of clusters of protein families on the SOM corroborates the hypothesis that family-related information exists in the flanking areas of the PEXEL motif. This would not be without precedent, as precursor proteins targeted to cellular compartments such as the mitochondria and chloroplasts often contain essential protein targeting information at their N-terminus, sometimes encoded on an extra exon. A similar situation can be found in proteins targeted to the apicoplast of P. falciparum, and, e.g. in exported P. falciparum homologues of the HSP40 chaperone family [29].

The apparent positional conservation of the PEXEL motif (approximately 20 amino acids C-terminal to the hydrophobic sequence, and situated 15-20 amino acids N-terminal to the beginning of the mature protein) has been suggested to be required for correct recognition by the transport machinery [30]. As of today, there is no experimental evidence to suggest that the PEXEL containing region is actually cleaved. N-terminal protein sequencing of exported proteins has been attempted, but so far without success [31]. Additionally, Western blot analysis shows no size difference between proteins within the parasite's secretory pathway and those that have reached the erythrocyte cytosol, although this size shift should be able to be detected [32,33].

In the present study, we show an apparent family-specific conservation of physicochemical residue profiles for short PEXEL-flanking regions (vide infra). This raises the possibility that this region may be more than just a "simple transport signal", e.g. playing a role in alternative transport mechanisms, or in regulation of protein transport. To this end, it is noteworthy that a PEXEL containing RESA-GFP chimera was only correctly transported to its correct sub-cellular location when expressed under control of its endogenous promoter. Expression of the same protein under control of a heterologous promotor led to retention of the reporter within the lumen of the PV [34]. We speculate that the PEXEL-flanking regions might therefore influence regulated secretion of proteins, either temporally, or even in response to external stimuli. In other systems, evidence is also accumulating to suggest that

Table 1. Pearson correlation between family-derived hydrophilicity profiles and 25-residue sequence fragments containing the central PEXEL motif.

Profile from	*Fragments from*		
	PfEMP1	RIFIN	STEVOR
PfEMP1	0.75 (0.09)	0.15 (0.09)	0.07 (0.11)
RIFIN	0.15 (0.07)	0.73 (0.08)	0.46 (0.10)
STEVOR	0.07 (0.12)	0.43 (0.09)	0.80 (0.16)

Standard deviation in brackets.
doi:10.1371/journal.pone.0001560.t001

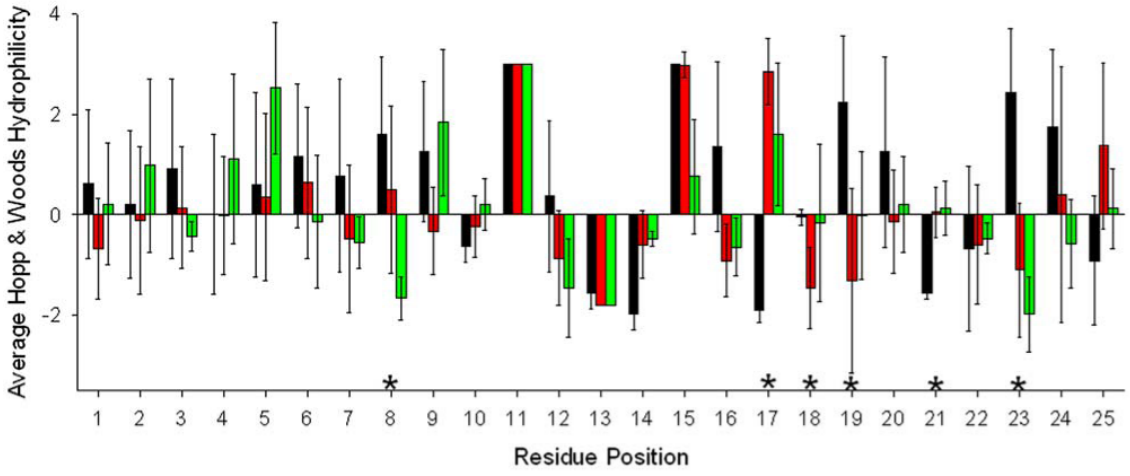


Figure 3. Average position-specific hydrophobicity in 25-residue fragments containing the central PEXEL motif (at positions 11-15). Color indicates the protein family (black: PfEMP1, gray: RIFIN, light gray: STEVOR). Error bars give standard deviations. Asterisks indicate positions characteristic for one of the families. Note that the Hopp & Woods scale [20] is a "hydrophilicity" scale with negative values for hydrophobic residues.
doi:10.1371/journal.pone.0001560.g003

targeting signals, such as endoplasmic reticulum (ER) signals, far from being "just greasy peptides", can contain important regulatory information [35,36].

In all three Plasmodium protein families studied, the downstream flanking regions show high information content with regard to hydrophobic and hydrophilic residues. Noteworthy, the upstream flanking region of the STEVOR examples exhibits additional conserved patterns not present in the known PfEMP1 and RIFIN proteins (Figure 2).

For calculation of the position-specific information content the software H-BloX was used [37] (Eq. 1). The 25-residue fragments were encoded by a two-letter alphabet containing "hydrophobic" (A,C,F,G,I,L,M,T,V,W) and "hydrophilic" residues (D,E,H,K, N,P,Q,R,S,Y).

$$I \sim H_{background} - H_{observed}, \text{ where} \qquad (Eq:1)$$

$$H_{observed} \sim -\left(p_{hydrophobic} \log_2 p_{hydrophobic} + p_{hydrophilic} \log_2 p_{hydrophilic}\right):$$

The expected distribution $H_{background}$ of hydrophobic and hydrophilic residues was calculated from the amino acid distribution found in the predicted *P.falciparum* proteome (in percent: A = 1.9, C = 1.8, D = 6.5, E = 7.2, F = 4.4, G = 2.8, H = 2.4, I = 9.2, K = 11.7, L = 7.5, M = 2.2, N = 14.5, P = 2.0, Q = 2.7, R = 2.6, S = 6.4, T = 4.1, V = 3.9, W = 0.5, Y = 5.7).

Site-directed mutagenesis of charged residues within this region has previously been shown to cause an accumulation of chimeric reporter proteins within the parasite's endoplasmic reticulum [10]. This region is predicted to contain several putative chaperone binding sites, suggesting that disruption of chaperone binding sites may interfere with chaperone mediated protein folding and quality control, leading to an aggregation of incorrectly folded protein, and a corresponding reduction in protein export. Mutation of residues "downstream" of the PEXEL motif had minimal or no effect on the localization of a STEVOR protein [10], highlighting the relative importance of its PEXEL preceding sequence.

We then computed averaged hydrophobicity profiles of PEXEL plus flanking residues for each of the three protein families, using the hydrophilicity scale according to Hopp and Woods [20]. Table 1 gives the correlation coefficients for matching the family-specific profiles against the fragments from the three families. We observe that there is only low cross-family correspondence of the property patterns, again suggesting family specificity of the flanking regions.

More detailed analysis of the position-specific preference of hydrophobic or hydrophilic residues indicate that position 8 is important for discrimination of STEVOR proteins, whereas positions 17–19, 21, 23 are characteristic of PfEMP1 proteins (Figure 3). Position 18 is dominated by glycine in PfEMP, resulting in high information content (Figure 2) yet a hydrophobicity value of close to zero (Figure 3).

A further hint towards a family-specific function of the N-terminal flanking region is that, according to our analysis, only 24% of the proteins with a PEXEL motif actually possess a standard signal sequence. It has been reported that PEXEL is preferably located 15–20 amino acids downstream of an N-terminal hydrophobic signal sequence [2,3]. In Figure 4a, the PEXEL motif distribution in our set of 412 proteins is shown. We observe three groups of sequences with preferences around positions 20, 43, and 85. All PfEMP1 proteins lack a standard

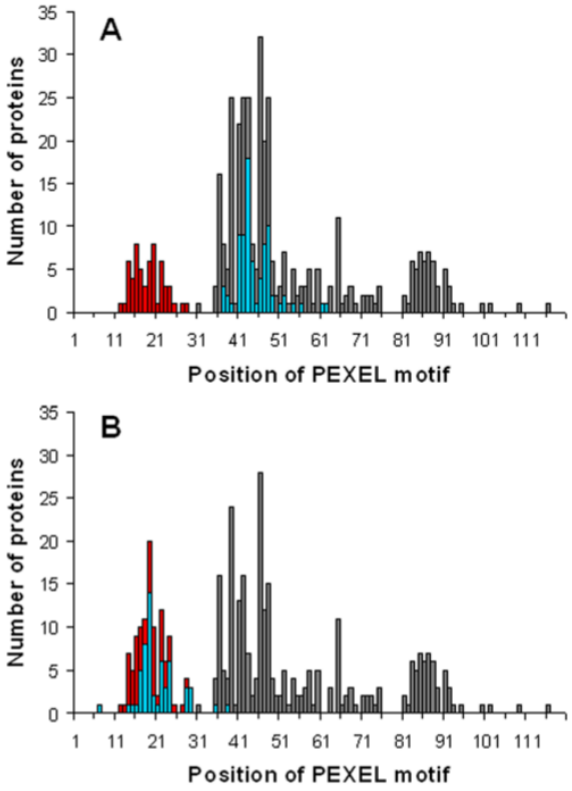


Figure 4. Distribution of the first position of the PEXEL motif. A) position of the PEXEL motif in sequences of the PfEMP1 protein family (red), exported proteins with a predicted signal peptide (blue), and exported proteins lacking a predicted signal peptide (gray). In B) the blue bars show the positions of the PEXEL motif after cleaving off the predicted signal peptide. Only sequences with a predicted signal peptidase cleavage site (score . 0.5 according to SignalP [38,39]) are included. Gray bars in B) represent the unchanged distribution of PEXEL in proteins lacking a predicted signal sequence. Note that all bars are displayed on top of each other.
doi:10.1371/journal.pone.0001560.g004

signal peptide, and the PEXEL location is near the protein N-terminus between residue positions 12 and 28. In contrast, in proteins containing a predicted signal peptide we find the PEXEL motif in a range of approximately 30 residues, between positions 37 and 63. We then artificially cleaved off the signal peptide in precursors with a predicted cleavage site and analyzed the resulting mature proteins: PEXEL motifs shift to positions 13–29, which is now comparable to the position of the PEXEL motif in PfEMP1 proteins lacking a signal peptide (Figure 4b).

The group of proteins with a PEXEL preference around position 85 does not contain a canonical signal peptide, but rather a recessed N-terminal hydrophobic segment, which has previously been shown to function as an ER targeting signal [2,33,39]. The proteins with a PEXEL preference around positions 35–50 are predicted to be exported as they, in addition to a PEXEL sequence, possess a hydrophobic N-terminal segment (Figure 4b, gray bars). Many of these proteins may actually contain an export signal which is not recognized by SignalP. As no standard algorithm predicts cleavage of these sequences and it is unclear whether these sequences are actually cleaved at all, no shift in the position of PEXEL is predicted in the analyses shown in Figure 4b.

These analyses support the hypothesis that, although different mechanisms may exist for initial entry of PEXEL containing proteins to the secretory pathway, mediated either by an N-terminal signal sequence, or another, as yet uncharacterized mechanism, certain positional constraints are exerted on the PEXEL motif, potentially related to the nature of the protein translocation machinery. As a consequence, recessed signal sequences such as those present in glycophorin-binding protein 130 (GBP130) and the ring-infected erythrocyte surface antigen (RESA) 2 might be actually cleaved to bring the PEXEL motif into the correct positional preferences required for further transport.

On this note, the strong conservation of the initial arginine residue in the PEXEL motif is of interest. Arginine residues can often be found in protein targeting motifs such as the TAT (twin arginine translocation) signal peptide [41], and arginine based ER retention signals. It is possible that the arginine residue in the PEXEL motif associates the exported protein with the membrane of the parasitophorous vacuole prior to passage through the putative translocon. Such membrane binding properties have recently been shown for arginine residues in the TAT signal [42].

Summarizing, we found conserved hydrophobicity profiles rather than conserved residue patterns in the PEXEL-flanking regions. This hints toward potential recognition of the PEXEL motif and flanking regions by an interacting macromolecule and supports earlier experimental findings [12]. Any conserved property profile most likely is a result from gene duplication and other evolutionary events leading to the formation of different protein families. Bioinformatical analysis alone will not be able to undoubtedly determine whether these patterns are part of a PEXEL-related targeting signal or responsible for a completely different function. Still, our study provides a well-motivated basis for the necessary biochemical experiments. Although we may use the PEXEL motif to speculate about the nature of the *P. falciparum* ''exportome'', we are only now beginning to understand the processes governed by this sequence, their biological importance, and how such processes are regulated, possibly by residues directly abutting the PEXEL sequence itself.

Acknowledgments

We thank N. Joannin for helpful discussion and access to data pre-publication.

Author Contributions

Conceived and designed the experiments: JH GS JP KL. Performed the experiments: JH FS. Analyzed the data: JH GS FS JP KL. Wrote the paper: JH GS JP KL.

References

1. Cooke BM, Lingelbach K, Bannister LH, Tilley L (2004) Protein trafficking in Plasmodium falciparum-infected red blood cells. Trends Parasitol. 20: 581–589.
2. Hiller NL, Bhattacharjee S, van Ooij C, Liolios K, Harrison T, et al. (2004) A host-targeting signal in virulence proteins reveals a secretome in malarial infection. Science 306: 1934–1937.
3. Marti M, Good RT, Rug M, Knuepfer E, Cowman AF (2004) Targeting malaria virulence and remodeling proteins to the host erythrocyte. Science 306: 1930–1933.
4. Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, et al. (1995) Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. Cell 82: 101–110.
5. Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC, et al. (1995) Cloning the P. falciparum gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. Cell 82: 77–87.
6. Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, et al. (1995) The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. Cell 82: 89–100.
7. Kyes SA, Rowe JA, Kriek N, Newbold CI (1999) RIFINs: a second family of clonally variant proteins expressed on the surface of red cells infected with Plasmodium falciparum. Proc. Natl. Acad. Sci. U. S. A. 96: 9333–9338.
8. Cheng Q, Cloonan N, Fischer K, Thompson J, Waine G, et al. (1998) Stevor and rif are Plasmodium falciparum multicopy gene families which potentially encode variant antigens. Mol. Biochem. Parasitol. 97: 161–176.
9. Marti M, Baum J, Rug M, Tilley L, Cowman AF (2005) Signal-mediated export of proteins from the malaria parasite to the host erythrocyte. J. Cell. Biol. 171: 587–592.
10. Przyborski JM, Miller SK, Pfahler JM, Henrich PP, Rohrbach P, et al. (2005) Trafficking of STEVOR to the Maurer's clefts in Plasmodium falciparum-infected erythrocytes. EMBO J. 24: 2306–2317.
11. Nunes MC, Goldring JP, Doerig C, Scherf A (2007) A novel protein kinase family in Plasmodium falciparum is differentially transcribed and secreted to various cellular compartments of the host cell. Mol. Microbiol. 63: 391–403.
12. Petter M, Haeggström M, Khattab A, Fernandez V, Klinkert MQ, et al. (2007) Variant proteins of the Plasmodium falciparum RIFIN family show distinct subcellular localization and developmental expression patterns. Mol. Biochem. Parasitol. 156: 51–61.
13. Stoeckert CJ Jr, Fischerm S, Kissinger JC, Heiges M, Aurrecoechea C, et al. (2006) PlasmoDB v5: new looks, new genomes. Trends Parasitol. 22: 543–546. (www.plasmodb.org/plasmo/home.jsp, version of 19 Feb. 2006).
14. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419: 498–511. (www.tigr.org, version of 8 June 06).
15. Cochrane G, Aldebert P, Althorpe N, Andersson M, Baker W, et al. (2006) EMBL Nucleotide Sequence Database: developments in 2005. Nucl. Acids Res. 34: D10–D15. (www.ebi.ac.uk/embl/, version of 18 June 2006).
16. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucl. Acids Res. 25: 3389–3402.
17. Sargeant TJ, Marti M, Caler E, Carlton JM, Simpson K, et al. (2006) Lineage-specific expansion of proteins exported to Erythrocytes in malaria parasites. Genome Biol. 7: R12.
18. HenikoffS, HenikoffJG (1993) Performance evaluation of amino acid substitution matrices. Proteins 17: 49–61.
19. Schneider G, Wrede P (1993) Development of artificial neural filters for pattern recognition in protein sequences. J. Mol. Evol. 36: 586–595.
20. Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. U. S. A. 78: 3824–3828.
21. Engelmann DM, Steitz TA, Goldman A (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu. Rev. Biophys. Chem. 15: 321–353.
22. Zamyatnin AA (1972) Protein volume in solution. Prog. Biophys. Mol. Biol. 24: 107–123.
23. Chothia C (1975) Structural invariants in protein folding. Nature 254: 304–308.
24. Jones DD (1975) Amino acid properties and side-chain orientation in proteins: a cross correlation approach. J. Theor. Biol. 50: 167–183.
25. Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol. Cybern. 42: 59–69.
26. Schneider G (1999) How many potentially secreted proteins are contained in a bacterial genome? Gene 237: 113–121.
27. Schmuker M, Schwarte F, Brück A, Proschak E, Tanrikulu Y, et al. (2007) SOMMER: self-organising maps for education and research. J. Mol. Model. 13: 225–228.
28. Joannin N, Abhiman S, Sonnhammer EL, Wahlgren M (2008) Sub-grouping and sub-functionalization of the RIFIN multi-copy protein family. BMC Genomics 9: 19, in press.
29. Ralph SA, Foth BJ, Hall N, McFadden GI (2007) Evolutionary pressures on apicoplast transit peptides. Mol. Biol. Evol. 21: 2183–2194.
30. Knuepfer E, Rug M, Cowman AF (2005) Function of the plasmodium export element can be blocked by green fluorescent protein. Mol. Biochem. Parasitol. 142: 258–262.
31. Baumeister S, Burgwedel A, Maier UG, Lingelbach K (1999) Reconstitution of protein transport across the vacuolar membrane in Plasmodium falciparum-infected permeabilized erythrocytes. Novartis Found Symp. 226: 145–154; discussion 154–156.
32. Benting J, Mattei D, Lingelbach K (1994) Brefeldin A inhibits transport of the glycophorin-binding protein from Plasmodium falciparum into the host erythrocyte. Biochem. J. 300: 821–826.
33. Ansorge I, Benting J, Bhakdi S, Lingelbach K (1996) Protein sorting in Plasmodium falciparum-infected red blood cells permeabilized with the pore-forming protein streptolysin O. Biochem. J. 315: 307–314.
34. Rug M, Wickham ME, Foley M, Cowman AF, Tilley L (2004) Correct promoter control is needed for trafficking of the ring-infected erythrocyte surface antigen to the host cytosol in transfected malaria parasites. Infect. Immun. 72: 6095–6105.

35. Hegde RS, Bernstein HD (2006) The surprising complexity of signal sequences. Trends Biochem. Sci. 31: 563–571.
36. Martoglio B, Dobberstein B (1998) Signal sequences: more than just greasy peptides. Trends Cell Biol. 8: 410–415.
37. Zuegge J, Ebeling M, Schneider G (2001) H-BloX: visualizing alignment block entropies. J. Mol. Graph. Model. 19: 304–306.
38. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340: 783–795.
39. Nielsen H, Krogh A (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. In: Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB 6: AAAI Press, Menlo Park). pp 122–130.
40. Wickham ME, Rug M, Ralph SA, Klonis N, McFadden GI, et al. (2001) Trafficking and assembly of the cytoadherence complex in Plasmodium falciparum infected human erythrocytes. EMBO J. 20: 5636–5649.
41. Sargent F, Berks BC, Palmer T (2006) Pathfinders and trailblazers: a prokaryotic targeting system for transport off olded proteins. FEMS Microbiol. Lett. 254: 198–207.
42. Shanmugham A, Wong Fong Sang HW, Bollen YJ, Lill H (2006) Membrane binding of twin arginine preproteins as an early step in translocation. Biochemistry 45: 2243–2249.

Domain Organization of Long Signal Peptides of Single-Pass Integral Membrane Proteins Reveals Multiple Functional Capacity

J. A. Hiss, E. Resch, A. Schreiner, M. Meissner, A. Starzinski-Powitz, G. Schneider

PLoS ONE 3, e2767

Wir beschreiben ein neues Modell (NtraC) für eine bipartite Domänen-Architektur langer Vertebrata-Signalpeptide, das in 63% aller annotierten Vertebrata-Signalpeptide mit mehr als 40 Aminosäuren Länge auftritt. Die Domänen werden als voneinander unabhängige, vollständige Targeting-Signale für unterschiedliche Kompartimente vorhergesagt. Die für das lange Signalpeptid von shrew-1 *in silico* vorgeschlagene NtraC-Architektur wird *in vitro* bestätigt.

OPEN ACCESS Freely available online PLoS one

Domain Organization of Long Signal Peptides of Single-Pass Integral Membrane Proteins Reveals Multiple Functional Capacity

Jan A. Hiss ·, Eduard Resch ·, Alexander Schreiner, Michael Meissner, Anna Starzinski-Powitz ", Gisbert Schneider "*

Centre for Membrane Proteomics, Institute of Cell Biology and Neuroscience, Goethe-University, Frankfurt am Main, Germany

Abstract

Targeting signals direct proteins to their extra - or intracellular destination such as the plasma membrane or cellular organelles. Here we investigated the structure and function of exceptionally long signal peptides encompassing at least 40 amino acid residues. We discovered a two-domain organization ("NtraC model") in many long signals from vertebrate precursor proteins. Accordingly, long signal peptides may contain an N-terminal domain (N-domain) and a C-terminal domain (C-domain) with different signal or targeting capabilities, separable by a presumably turn-rich transition area (tra). Individual domain functions were probed by cellular targeting experiments with fusion proteins containing parts of the long signal peptide of human membrane protein shrew-1 and secreted alkaline phosphatase as a reporter protein. As predicted, the N-domain of the fusion protein alone was shown to act as a mitochondrial targeting signal, whereas the C-domain alone functions as an export signal. Selective disruption of the transition area in the signal peptide impairs the export efficiency of the reporter protein. Altogether, the results of cellular targeting studies provide a proof-of-principle for our NtraC model and highlight the particular functional importance of the predicted transition area, which critically affects the rate of protein export. In conclusion, the NtraC approach enables the systematic detection and prediction of cryptic targeting signals present in one coherent sequence, and provides a structurally motivated basis for decoding the functional complexity of long protein targeting signals.

Citation: Hiss JA, Resch E, Schreiner A, Meissner M, Starzinski-Powitz A, et al. (2008) Domain Organization of Long Signal Peptides of Single-Pass Integral Membrane Proteins Reveals Multiple Functional Capacity. PLoS ONE 3(7): e2767. doi:10.1371/journal.pone.0002767

Editor: Janet Kelso, Max Planck Institute for Evolutionary Anthropology, Germany

Received March 10, 2008; Accepted June 25, 2008; Published July 23, 2008

Funding: This research was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften, the Centre for Membrane Proteomics at Goethe-University, Frankfurt am Main, the Sonderforschungsbereich 579 ("RNA-Ligand Interactions", project A11.2), the Sonderforschungsbereich 628 ("Functional Membrane Proteomics", project p7), and the Deutsche Forschungsgemeinschaft Sta 187/16-1.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: g.schneider@modlab.de

. These authors contributed equally to this work.

" These authors also contributed equally to this work.

Introduction

Targeting signals are contiguous stretches of amino acids that direct proteins to their sub-cellular destinations or the extracellular space [1]. With few exceptions, the vast majority of extracellular proteins are exported from mammalian cells via the endoplasmic reticulum (ER) secretory pathway [2]. While most signal sequences are N-terminally located, deviant examples have been reported with internal signals like in human UDP-glucuronosyltransferase [3], or bacterial C-terminal secretion signals like in virulence factor from Mycobacterium tuberculosis [4], and Escherichia coli (E. coli) haemolysin [5].

Canonical N-terminal signals are processed by signal peptidases [6]. The sequence similarity among these cleavable "signal peptides" coding for the ER and subsequent protein export is low as they do not share common residue motifs but rather possess common physicochemical features coding for the appropriate cellular compartment [7,8]. Signal recognition by the cellular decoding machinery may include multiple recognition events [9,10]. This renders perfect in silico prediction of subcellular locations and the detection of targeting signals still impossible although many encouraging attempts have been made [11–16]. For example, to counter the dissimilarity in signal peptides for prediction processes, the amino acid composition has been taken into account resulting in improved accuracy [8,17,18]. Despite their dissimilarity, N-terminally located targeting sequences are sometimes interchangeable between proteins in eukaryotes and even between different kingdoms. One such example is Escherichia coli (E. coli) beta-lactamase, which can be exported by Xenopus oocytes [19]. Still, general signal interchangeability cannot be postulated [20,21]. Public web servers are available for predicting the subcellular localization of proteins in various organisms, for example Cell-PLoc (http://chou.med.harvard.edu/bioinf/Cell-PLoc/) [22] or the SignalP suite (http://www.cbs.dtu.dk/services/SignalP/) [14].

In eukaryotes, a canonical N-terminally located protein export signal typically contains three distinguishable parts: a positively charged N-terminal section (n-region), a hydrophobic core (h-region), and a signal peptidase recognition site (c-region) [8,11]. The approximate average length of such signal peptides is 22 amino acid residues [23]. While the c-region typically consists of five residues, both the h- and the n-region show more variability in length.

This variability has been suggested to enable alternative functions [10,24]. In fact, much longer examples of signal peptides are known to exhibit additional functions besides precursor targeting [10,25,26], for example regulation of the protein export rate as described for interleukin-15 [27], or signal peptide accumulation in the nucleoli in the case of mouse mammary tumor virus Rem protein after release from the endoplasmic reticulum [28].

In the present study, we introduce a structurally motivated modularization of long signal peptides into separate functional modules, and demonstrate the actual functional relevance of this concept for the long signal peptide of the integral membrane protein shrew-1 (SH) as an example. Shrew-1 was originally isolated from an epithelial-like cell line obtained from an endometriosis biopsy [29]. It contains a cleavable N-terminal signal peptide of 43 residues [30], an extracellular domain (residues 44–282), a transmembrane segment (residues 283–303) and a cytoplasmic domain (residues 304–411). Shrew-1 is transported to the basolateral part of the plasma membrane in polarized epithelial cells and interacts with the E-cadherin mediated adherens junction complex [29,31]. In non-polarised cells, like transformed epithelial cells, shrew-1 also displays plasma membrane localization, though apparently less polarized. Shrew-1 appears to be involved in the regulation of cell invasion and motility and, in line with this, interacts with protein CD147, a known promoter of invasiveness [32].

Based on proteome analysis by machine-learning systems, we propose a bipartite domain model (''NtraC'' model) of long signal peptides from single-pass integral membrane proteins. According to this model, such long signal peptides may contain two separate functional domains: an N-terminal domain (''N-domain'') and a C-terminal domain (''C-domain'') traceable by a turn-rich linker area connecting both. We denote this linker element ''transition area'' (tra). Proof-of-principle for the validity of the NtraC domain model is provided by in vitro targeting experiments with shrew-1.

Results

Many single-spanning integral membrane proteins possess long signal peptides with a bipartite domain organization

Analysis of long signal peptides was performed in two steps: First, potential domains were predicted using a novel machine-learning technique for turn prediction [33]. Potential turn-containing regions were found to be predominantly located in the central portion of these long signals. Based on the location of this ''transition area'', long signal peptides were dissected into two parts, an N-terminal ('N') and a C-terminal ('C') fragment. Then, the resulting sequence fragments were scrutinized for potential targeting functions. The concept of this NtraC model of signal peptide organization is based on the hypothesis that the two functional modules in a long signal peptide may exhibit individually distinct tasks in the context of protein targeting. This requires a minimal peptide length, and for the present study we decided to focus only on signal peptide domains containing conventional signals with an expected average length of approximately 20 residues each. This choice is motivated by the observed average length of targeting signals coding for a single compartment [23]. Certainly, we cannot exclude the existence of other targeting signals of hitherto unknown structure e.g. unusually short signals) within long signal peptides.

Searching for long signal peptides $40 residues) in the UniProtKB database (release 53.2) [34] yielded 296 vertebrate proteins, including homologues. All sequences were analyzed with regard to their potential NtraC organization. Within our NtraC analysis software, predictions for potential targeting signals were done using the software SignalP 3.0 [23] (signals coding for protein transport into the ER, signal peptide and signal anchor prediction) and TargetP [35] (signals coding for mitochondrial import). Potential turn-forming elements were detected using our software tool SVMTurn (www.modlab.de/Software/SVMTurn) [33]. SVMTurn uses Support Vector Machine classifiers for recognition of various turn types in amino acid sequences. Turns with intramolecular hydrogen bonds encompassing four, five, and six residues are predicted with approximately 80% accuracy.

According to NtraC (www.modlab.de/Software/NtraC) analysis, 185 of 296 (62%) long signal peptides obey the NtraC domain organization with a C-domain coding for an ER targeting signal (Suppl. Table S1). We found no strict conservation of turn residues in all 185 sequences. As expected for beta-turns, Gly is overrepresented at residue position 3 of a regular beta turn [36]. 45 of thee 185 candidate proteins possess both an N-domain coding for a putative mitochondrial transit peptide and a C-domain coding for an endoplasmic reticulum (ER) targeting signal (Figure 1). For 13 of these sequences, signal peptidase cleavage sites were not predicted. Thus, they might act as signal anchors. All 32 remaining candidates, which show a predicted domain combination analogous to shrew-1 (N-Domain: mTP, C-domain: SP) and posses a predicted signal peptidase cleavage site, are listed in Table 1. The C-domains of the remaining 140 NtraC-organized sequences code for ER targeting. In contrast to shrew-1, however, their N-domains may contain an additional feature or targeting function that is different from conventional mitochondrial targeting signals.

To check the influence of a potential bias in these results due to clusters of homologues in the set of 296 candidate genes, we manually eliminated all orthologues. This procedure did not affect the ratio of NtraC-organized vs. non-NtraC-organized samples (Figure 1, values in brackets). In the human genome alone, we found 105 signal peptides with $40 residues overall, among which 71 (68% of 105) are NtraC-organized.

We provide a public web service for NtraC analysis of amino acid sequences (www.modlab.de/Software/NtraC) and invite the scientific community to scrutinize our NtraC domain model using this prediction server.

Proteins with NtraC-organized signal sequences apparently have common features. 19 of the 32 candidate sequences are annotated in UniProt as type-I membrane proteins containing a single potential transmembrane segment (TMS). Among these, the only experimentally validated TMS is the one of shrew-1 [29], which was a clear motivation for us to use this protein for the cellular proof-of-principle study. We then performed TMS predictions for the 13 remaining sequences using the software tools Phobius [37] and SVMtm [38], which in all cases gave rise to the same results: Two proteins yielded strong positive scores indicating the likely presence of a TMS, three received weaker scores favoring TMS presence, and eight are seemingly devoid of a TMS. These results increase the number of candidate proteins from 19 to 24 out of 32, corresponding to 75% as a conservative estimation.

Summarizing, we identified a class of long signal peptides distinguished by the NtraC domain architecture. This structural and functional organization is present in signal peptides of many single-pass membrane proteins. For further study, we selected one of these proteins, human shrew-1 as an example.

Experimental system for assessment of prediction results: Shrew-1 signal peptide and SEAP reporter protein

Based on the theoretical analysis described in the previous paragraph, we used secreted alkaline phosphatase (SEAP) as a reporter protein in order to probe the targeting capacity of the predicted domains of shrew-1's signal peptide. The SEAP reporter

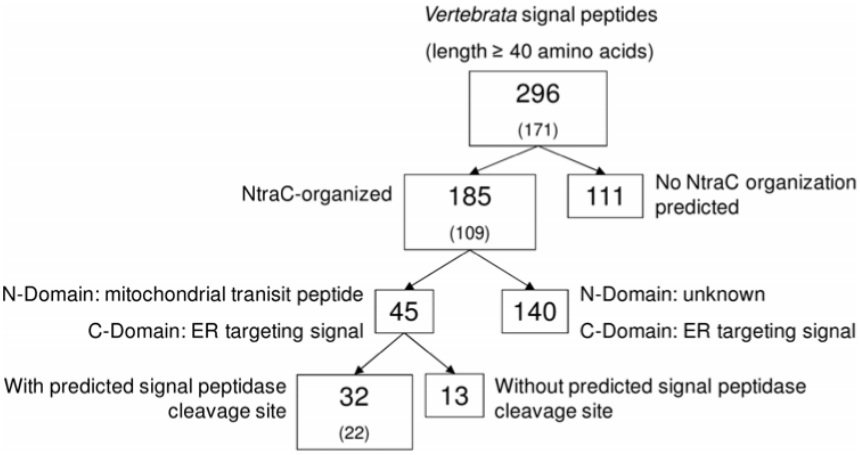


Figure 1. Overview of NtraC-organized sequences among long signal sequences found in vertebrate proteins. Set sizes without orthologues are given in brackets. The numbers represent conservative estimates based on validated prediction tools for targeting signal recognition and turn structure prediction.
doi:10.1371/journal.pone.0002767.g001

system allows for the exchange of the intrinsic signal peptide by other potential signal peptide sequences, which can then be tested for biological activity [39]. SEAP is a glycoprotein which becomes N-glycosylated by oligosaccharyl transferase located in the ER [40]. Therefore, its N-glycosylation status is an indication of translocation into the ER lumen, which in turn is a prerequisite for SEAP secretion into the supernatant.

The C-domain acts as a secretion signal. According to the NtraC model, the shrew-1 signal peptide (residues 1–43 [30], SignalP 3.0probability = 0.95) is divided into three domains: It contains an N-domain (residues 1–19) and a C-domain (residues 20–43) connected by the transition area (residues 16–24). The C-domain is predicted as a standard secretion signal containing an h-, and c-region (SignalP 3.0probability = 0.9), whereas the N-domain receives a prediction as a mitochondrial transit peptide (TargetPprobability = 0.3).

Within the transition area, three adjacent and partly overlapping b-turns were predicted (positions 16–24). Interestingly, no further turns were found in the remainder of the signal peptide. The position of the turns appears to be evolutionary conserved among different species, as shown by a multiple sequence alignment of seven vertebrate shrew-1 homologues, suggesting a fundamental functional importance of this region (Suppl. Figure S1).

To functionally test the predicted signal peptide domains, six constructs coding for different SEAP fusion proteins were devised (Figure 2). They were transfected into HEK 293T cells, and SEAP activity was determined in both the supernatants and in whole cell lysates.

As shown in Figure 3A, the C-domain (SH^{C}-SEAP ^{DSP}) alone is able to direct SEAP fusion protein to the supernatant. The N-domain (SH^{N}-SEAP ^{DSP}) alone does not have this targeting capacity. The same holds for the whole cell lysates (Figure 3A, white bars).

Compared to full length shrew-1 signal peptide (SH-SEAP^{DSP}), SEAP activity in both the supernatant and whole cell lysates of SH ^{C}-SEAP ^{DSP} transfected cells was decreased to about one third. This implies that the full-length signal peptide is required for full export efficiency, but basic targeting information is encoded in the C-domain of the long signal peptide.

Notably, both fusion proteins were detectable by Western blotting (Figure 3B). This raises the question for the reason of inactivity of the N-domain containing protein. One explanation would be impaired translocation from the cytosol into the ER, which in turn should have resulted in lacking N-glycosylation of SEAP. To check this hypothesis, we subjected the lysates to PNGase F treatment, which removes N-linked glycans that are selectively found on ER-translocated active protein. Figure 3B shows that the SH^{N}-SEAP ^{DSP} protein is not N-glycosylated (lanes 7 and 8), whereas SH^{C}-SEAP ^{DSP} and SH-SESP ^{DSP} contain an N-glycosylated SEAP population (lanes 3 and 5, band marked by an asterisk). We conclude that SH^{N}-SEAP ^{DSP} was not transported into the ER. It is noteworthy that SH ^{N}-SEAP ^{DSP} was found in two non-glycosylated bands (lanes 7 and 8), indicating the existence of two populations with different molecular mass. The position of the bands is in line with the idea that the upper band contains the N-domain of the signal peptide, which might have been cleaved offin the faster migrating protein (lower band) by some non-ER protease activity.

The N-domain directs the reporter protein to mitochondria. The observation of two non-glycosylated bands in the Western blot analysis raised the question, whether the SH^{N}-SEAP ^{DSP} fusion protein is able to target to mitochondria, as predicted by our sequence analysis (vide supra). Therefore, we analyzed mitochondrial localization of SH^{N}-SEAP ^{DSP}. HEK 293T cells were transfected with either SH^{N}-SEAP ^{DSP} or SH ^{C}-SEAP ^{DSP}, and mitochondria were isolated by differential centrifugation followed by density gradient centrifugation. Cytosolic (cyto) and ER fractions obtained by differential centrifugation were positive for GAPDH as a cytosolic marker protein, or grp94 as an ER marker, and negative for cytochrome C as a mitochondrial marker (Figure 4, lanes 1–4). Mitochondria obtained by density centrifugation were completely negative for GAPDH, only a weak band corresponding to grp94 was detectable, and cytochrome C was prominently detected, indicating efficient purification of mitochondria (Figure 4, lanes 5 and 6).

SH ^{C}-SEAP ^{DSP} was detectable in an unglycosylated state in the cytosolic fraction (Figure 4, lane 1) and in an N-glycosylated state in the ER fraction (Figure 4, lane 2). In contrast, it was barely detectable in the mitochondrial fraction (Figure 4, lane 5). A different distribution was found for SH^{N}-SEAP ^{DSP}, which was present in the cytosolic fraction, but not in the ER fraction (Figure 4, lanes 3 and 4). This observation is in line with the

Table 1. 32 Vertebrate signal peptides . 40 amino acids, which are predicted to be NtraC organized and are similar in their domain capacity to shrew-1.

ID	NCBI Accession Number	Signal peptide sequence
1	P70505	MSVAASASRSASTLCSPQIQQGALKEAKVPPHIWAARHWNLGLRLVPGHASVRAGILVLLIFLPSTLC
2	P17405	MPRYGASLRQSCPRSGREQGQDGTAGAPGLLWMGLVLALALALALA
3	Q96PD2	MASRAVVRARRCPQCPQVRAAAAAPAWAALPLSRSLPPCSNSSSFSMPLFLLLLLVLLLLLEDAGA
4	Q91ZV3	MASRAPLRAARSPQGPGGPAAPAATGRAALPSAGCCPLPPGRNSSSRPRLLLLLLLLLQDAGG
5	Q91ZV2	MASRAPLRAARSPQDPGGRAAPAATGRAPLPSAGWCPLPPGRNSSSRPRLLLLLLLLLPDAGA
6	Q28110	MGIPSFLAFPAARRNRAHCTPWHPWGHMLLWTALLFLAPVSG
7	Q1LZH9	MRLLSLAPDRPRRGGPRHLTSGSPALPPPPPLLLLLLLLGGCLGVSGA
8	P50426	MRFLSLAPDRPRRGGPRHLPSGSPAPPPPPPLLLLLLLGGCLGVSGA
9	P52785	MSAWLLPAGGLPGARFCVPARQSPSSFSRVLRWPRPGLPGLLLLLLLPSPSALS
10	P51840	MSAWLLPAGGFPGAGFCIPAWQSRSSLSRVLRWPGPGLPGLLLLLLLPSPSAFS
11	P51841	MFLGLGRFSRLVLWFAAFRKLLGHHGLASAKFLWCLCLLSVMSLPQQVWT
12	Q8K201	MAASALGRMCGAAREKLSPGPGARGLGALARSLVLALLLVPVLC
13	Q5R5B8	MAAAALKRMRGPAQAKLLPGSAIQALVGLARPLVLALLLVSAALSSVVS
14	Q9UBX7	MQRLRWLRDWKSSGRGLTAAKEPGARSSPLQAMRILQLILLALATGLVGG
15	Q5XNR9	MMNISLRLRRPPWMVDSNGRRMTSHFQWLLLTFILLYLMNQVTS
16	Q9H0V9	MAATLGPLGSWQQWRRCLSARDGSRMLLLLLLLGSGQGPQQVGA
17	P59481	MAAASRPSWWQRWRRRAWARDGAKLLLFLLLLGSGPGPRHVRA
18	Q6VE48	MRASCTPLKAPLRRPERLASSGRFAWVLLLAPLLLLPTSSDA
19	Q8VE43	MRGAVWAARRRAGQQWPRSPGPGPGPPPPPPLLLLLLLLLGGASA
20	Q5RJL6	MRGVVWAARRRAGQQWPRSPGPGPGPPPPPPLLLLLLLLLGGASA
21	Q9R0S2	MPRSRGGRAAPGQASRWSGWRAPGRLLPLLPALCCLAAAAG
22	Q99PW6	MPRSRGGRAAPGQAARWSGWRAPGRLLPLLPALCCLAAAAG
23	P29122	MPPRAPPAPGPRPPPRAAAATDTAAGAGGAGGAGGAGGAGGPGFRPLAPRPWRWLLLLALPAACSA
24	Q9NQS3	MARTLRPSPLCPGGGKAQLSSASLLGAGLLLQPPTPPPLLLLLFPLLLFSRLCGALA
25	Q96B86	MQPPRERLVVTGRAGWMGMGRGAGRSALGFWPTLAFLLCSFPAATSP
26	Q9N0A6	MGGPGPRRAGTSRERLVVTGRAGWMGMGRGAGRSALGFWPTLAFLLCSFPAAT
27	Q6PCX7	MQPPRERLVVTGRAGWMGMGRGAGRSALGLWPTLAFLLCSFPAAISP
28	Q9QUR8	MTPPPPGRAAPSAPRARVLSLPARFGLPLRLRLLLVFWVAAASA
29	Q9UPZ6	MGLQARRWASGSRGAAGPRRGVLQLLPLPLPLPLLLLLLLRPGAGRA
30	Q9EPU5	MGTRASSITALASCSRTAGQVGATMVAGSLLLLGFLSTITA
31	Q8IZC6	MGAGSARGARGTAAAAAARGGGFLFSWILVSFACHLASTQG
32	Q91443	MGRHSALGLSGNRQVSPCTGTRPFKVVGSRSPVQPLCILLALTVCIGTS

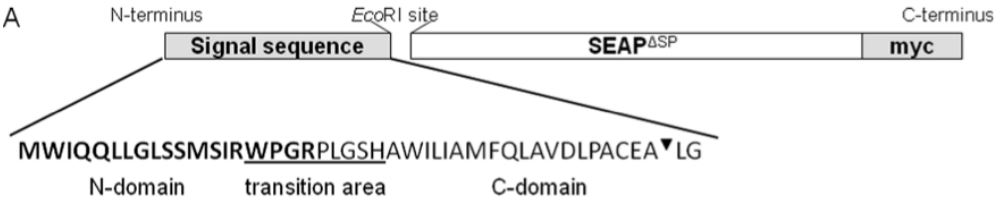
Underlined residues are predicted turns belonging to the transition area.
doi:10.1371/journal.pone.0002767.t001

absence of SEAP activity in the supernatant and whole cell lysates extracted from cells transfected with this fusion protein (Figure 3). Most importantly, SH^{N}-$SEAP^{DSP}$ was prominently detected in the mitochondrial fraction, which received further confirmation by immunofluorescence studies in HEK 293T cells (not shown). This experimental observation is in perfect agreement with the computational prediction.

Deletion of the transition area decreases secretion. The results presented so far show that the C-domain is sufficient for secretion of SEAP fusion protein, whereas the N-domain has no ER translocation capacity, but rather accommodates a mitochondrial targeting activity. However, when compared to the full length signal sequence the C-domain exhibits a decreased secretion activity. This observation gave rise to the question whether the transition area (residues 16–24) influences the efficiency of ER translocation.

To test this hypothesis, we generated constructs coding for three different SEAP fusion proteins, containing mutations and deletions in the transition area of the otherwise wild-type shrew-1 signal peptide. One contains a Gly^{R} Ile substitution at position 18 (SH^{G18I}-$SEAP^{DSP}$) which was predicted to prevent the formation of the first turn in the transition domain. In the second construct, we deleted the first four amino acids with the highest turn forming potential (SH^{DWPGR}-$SEAP^{DSP}$) of the predicted transition domain. In the third construct, we deleted the first four amino acids of the transition area and introduced additional substitutions in the remaining four amino acids in order to completely disrupt the transition area ($SH^{DWPGR/mut}$-$SEAP^{DSP}$) (for a schematic of all constructs, see Figure 2 B).

Each of these constructs was transfected into HEK 293T cells, and again SEAP activity was determined in the supernatants as well as in whole cell lysates. As shown in Figure 5A, SEAP activity decreases with increasing disruption of the transition area. $SH^{DWPGR/mut}$-$SEAP^{DSP}$ showed the lowest activity which is similar to the activity of SH^{C}-$SEAP^{DSP}$. This is consistent with



B

Constructs	Signal sequences
SH-SEAP$^{\Delta SP}$	MWIQQLLGLSSMSIR**WPGR**PLGSHAWILIAMFQLAVDLPACEA▼LG
SH^{G18L}-SEAP$^{\Delta SP}$	MWIQQLLGLSSMSIR**WPIR**PLGSHAWILIAMFQLAVDLPACEA▼LG
SH$^{\Delta WPGR}$-SEAP$^{\Delta SP}$	MWIQQLLGLSSMSIRPLGSHAWILIAMFQLAVDLPACEA▼LG
SH$^{\Delta WPGR/mut}$-SEAP$^{\Delta SP}$	MWIQQLLGLSSMSIAPSD—HAWILIAMFQLAVDLPACEA▼LG
SH^{C}-SEAP$^{\Delta SP}$	MPLGSHAWILIAMFQLAVDLPACEA▼LG
SH^{N}-SEAP$^{\Delta SP}$	MWIQQLLGLSSMSIR**WPGR**
SEAP$^{\Delta SP}$	Secreted alkaline phosphatase lacking a targeting signal

Figure 2. Shrew-1 (SH) signal sequence and the construction of the SEAP fusion proteins. (A) Organization of shrew-1 signal sequence. Bold: N-domain (shrew-1 residues 1–19). Standard type: C-domain (shrew-1 residues 20–43). Underlined: transition area (shrew-1 residues 16–24). ▼: signal sequence cleavage site. LG: shrew-1 residues 44 and 45. (B) Diagrams of SEAP constructs with assigned shrew-1 signal sequences. Signal sequences are N-terminally fused to the SEAP protein lacking the endogenous signal peptide (SEAP$^{\Delta SP}$). C-terminally, all fusion proteins are tagged with myc (EQKLISEEDL). For cleavage site recognition (PACEA▼LG) shrew-1 residues 44 and 45 (LG) are included in the constructs.
doi:10.1371/journal.pone.0002767.g002

the assumption that the transition area may be needed for the overall secretion activity of the shrew-1 signal sequence.

The dependency of secretion efficiency on the integrity of the transition area should be mirrored in the presence of N-glycosylated SEAP. This was tested by Western blotting (Figure 5B). With increasing impairment of the transition area the ratio of N-glycosylated (upper band;) to non-glycosylated SEAP fusion protein (lower band, arrow) species decreased by one order of magnitude from 1.94 to 0.17 (Figure 5B). We conclude that protein export efficiency appears to be correlated with the existence and integrity of the transition area separating N- and C-domains of the shrew-1 signal peptide.

Discussion

Here we report the first systematic approach for predicting structure and function ofl ong signal peptides of single-pass integral membrane proteins. Sequence analysis tools suggest a general organization model for these sequences, which was validated in a proof-of-principle study using the type I membrane protein shrew-1. Most importantly, according to our NtraC model a structural feature of the transition area is a crucial determinant ofl ong signal peptide modularization: A potentially turn- or loop-forming central element (transition area) acts as some kind of separation unit between two sequence domains with different targeting capacity. Results of cellular targeting studies highlight the functional importance of the transition area. A minimal interpretation is that it affects ER translocation of the reporter protein.

The N-domain (residues 1–19) was able to act as a mitochondrial targeting signal in our experiments. Similar observations have been made for other proteins containing consecutive ''tandem'' signals rather than ''cryptic'' signals as described by the NtraC model. The transmembrane glycoprotein nicastrin, which is an essential component of gamma-secretase [41], is such an example. Gamma-secretase was found to translocate into mitochondria in Alzheimer patients, potentially inducing apoptosis [42]. Transport into the organelle is mediated by a mitochondrial transit signal following the N-terminal cleavable signal peptide of nicastrin. Notably, in contrast to the shrew-1 example and the NtracC domain model, the sequential order of the targeting signals is inverted in nicastrin and other proteins containing such a ''tandem'' signal, e.g. microsomal CYP2E1 [43]. This demonstrates that the prediction and discovery of proteins with multiplex locations is important for an understanding of the regulation of cell process such as apoptosis.

Mitochondrial targeting of shrew-1 and other proteins containing NtraC-organized long signals may not occur constitutively but in a regulated manner or only under cellular stress, and our results indicate that the mitochondrial targeting signal (N-domain) and the ER targeting signal (C-domain) are not sequentially processed. The N-domain of shrew-1 harbors no ER translocation activity, but is able to mediate mitochondrial targeting. We wish to stress that this activity has been proven for the isolated N-domain in the context of the experimental setup used in the present study, and it needs further investigation to determine the conditions under which this activity is found in the context of the full-length signal peptide. Possibly this cryptic activity is revealed under certain physiological situations only.

As an extension to the already known tandem signals like in the nicastrin or CYP2E1 precursors [41,43], our NtraC model provides a framework for cryptic signals. The domain model is of general relevance, as at least 62% of the known vertebrata proteins with a signal peptide exceeding 40 residues show an NtraC-organization. Although it remains unclear if and under which conditions or regulatory control mitochondrial targeting of these proteins occurs, we were able show that NtraC-organized signal peptides can exhibit additional functions besides ER targeting or protein export. Prediction of such important structural elements has now become feasible.

Due to its amphipathic nature, we further speculate that the N-domain might be involved in dimerization or stabilization of shrew-1 in the plasma membrane or interaction with other proteins [29,32]. Positively charged arginine residues in the N-domain could help the signal peptide to adopt its native conformation in the plasma membrane. It would thereby follow the ''positive inside rule'' [44]

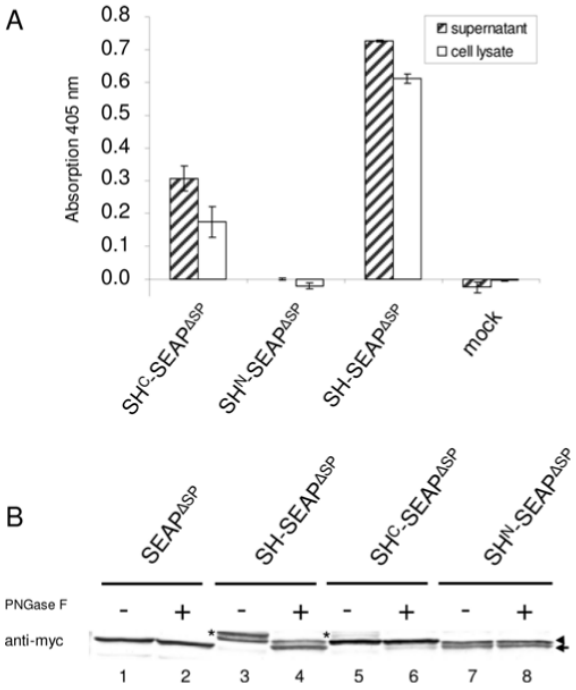


Figure 3. Influence of the isolated N- and C-domain on the expression, the activity and secretion of the SEAP fusion proteins. (A) SEAP activity was recorded in the supernatant (hatched bars) and whole cell lysate (white bars) of transfected HEK 293T cells after 5 minutes of substrate incubation. Cells transfected with the empty vector were used as negative control (mock). Error bars show s.e.m. (N = 4). (B) Cell lysates of HEK 293T cells expressing either SEAP^{DSP}, SH-SEAP^{DSP}, SH^{C}-SEAP^{DSP} or SH^{N}-SEAP^{DSP} were treated with PNGase F (+) or were left untreated (-) and Western blots were prepared. Fusion proteins were probed with anti-myc antibody. SH-SEAP^{DSP}, SH^{C}-SEAP^{DSP} or SH^{N}-SEAP^{DSP} fusion proteins show double bands while SEAP^{DSP} reveals a single band which lacks N-glycosylation. SH-SEAP^{DSP} and SH^{C}-SEAP^{DSP} possess an N-glycosylated protein population (*) that shifts to the position of SEAP^{DSP} (arrow) after PNGase F treatment. The protein population that exhibits no PNGase F sensitivity (b) is not N-glycosylated and not N-terminally processed. SH^{N}-SEAP^{DSP} shows no PNGase F sensitivity at all, but is also characterized by a doublet. The lower band (arrow) corresponds to the position of SEAP^{DSP} indicating N-terminal processing, whereas the upper band (b) complies with the non processed protein population.
doi:10.1371/journal.pone.0002767.g003

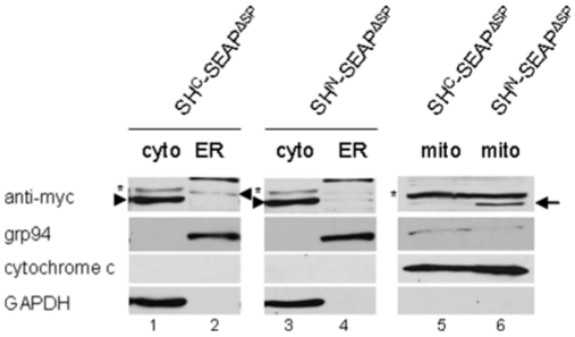


Figure 4. Detection of SH ^{N}-SEAP ^{DSP} in mitochondria. Mitochondria were isolated from HEK 293T transfected with either SH^{N}-SEAP^{DSP} or SH^{C}-SEAP^{DSP}, and Western blots were prepared with cytosolic (cyto), ER and mitochondrial (mito) fractions. SEAP fusion proteins were probed with antibody against the myc-tag (anti-myc). Marker proteins were grp94 for ER, cytochrome c for mitochondrial, and GAPDH for cytosolic fractions. Only SEAP fusion protein containing the N-domain of shrew-1's SP (SH^{N}-SEAP^{DSP}) was clearly detectable in the mitochondrial fraction (lane 6, arrow). Asterisks on the left indicate unspecific bands. Arrowheads mark the positions of SEAP fusion proteins in the cytosolic and ER fractions.
doi:10.1371/journal.pone.0002767.g004

and arrest the C-terminal part inside the membrane while being available for protein-protein interactions on the cytoplasmic side.

The C-domain is sufficient for protein export via the ER, but not as effective as the full-length signal peptide. Most strikingly, the transition area which was first predicted to only link the N- to C-domain, turned out to be essential for the full ER translocation activity of the C-domain. It is noteworthy that the transition area is the only part of the long signal peptides predicted to predominantly contain b-turns. Thus, turn formation seems to be not only a structural element separating the N- and C-domains, but a decisive feature of long signal peptides supporting the ER translocation activity of the C-domain. The NtraC model thereby explains earlier observations made for interleukin-15, which is subjected to different export rates depending on the length of its signal peptide [27].

Our model also provides a rational explanation for membrane targeting of bacterial autotransporters, which possess long signal peptides: These are in accordance with our NtraC model, where the C-domain alone is sufficient for transport to the inner membrane but for proper processing the complete signal peptide is required [45]. In the present study, we restricted our analysis to single-spanning integral membrane proteins with signals that have a similar organization as the long signal peptide of shrew-1. The role of the transition area besides making the N- and C-domain distinguishable is subject to further research.

Materials and Methods

Oligonucleotides used for cloning of SEAP fusion constructs

Constructs were generated by PCR (Suppl. Text S1).

Cell lines, cell culture and transfection

HEK 293T (CRL-11268; ATCC, Manassas, USA) were cultured in Dulbecco's Modified Eagle Medium (DMEM; (Invitrogen GmbH, Karlsruhe, Germany) with 10% fetal calf serum (FCS; PAA LABORATORIES, Co ¨lbe, Germany) and 1% penicillin/streptomycin (Invitrogen GmbH, Karlsruhe, Germany). 66 10^{5} cells were seeded per 12 cm^{2} of culture dish and transfected with 3mg DNA 24 h later by using Magnet Assisted Transfection (MATra, IBA GmbH, Go ¨ttingen, Germany) according to the manufacturer's instructions.

SEAP activity assays

SEAP activity assays were performed according to [39] using 10 ml of the supernatants or 6mg of protein from cleared whole cell lysates.

Immunoblotting and antibodies

After collection of supernatant for SEAP assays, cells were washed with PBS and lysed with 100ml RIPA buffer (150 mM NaCl, 50 mM Tris-HCl, pH 7.5, 0.5% sodium deoxycholate, 1% Nonidet P-40, 0.1% SDS) containing proteinase inhibitor cocktail Complete (Roche Diagnostics GmbH, Mannheim, Germany) at 4uC for 30 min. Lysates were cleared by centrifugation in a

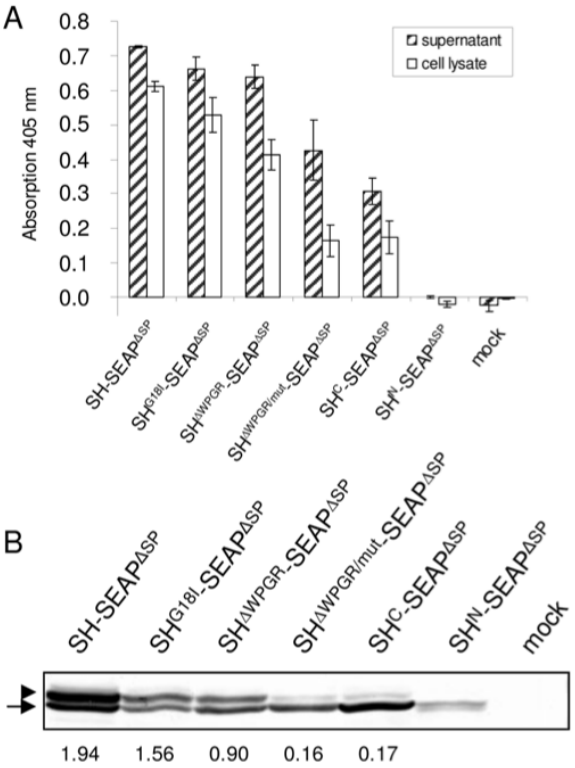


Figure 5. Mutation of the transition area impairs secretory activity of SEAP fusion proteins. (A) SEAP activity was measured in the supernatant (hatched bars) and whole cell lysate (white bars) of transfected HEK 293T cells after 5 min substrate incubation. Cells transfected with the empty vector were used as negative control (mock). Error bars show s.e.m. (N = 4). The data for cells with constructs SH^{N}-$SEAP^{DSP}$, SH^{C}-$SEAP^{DSP}$, SH-$SEAP^{DSP}$ and mock are adopted from Figure 2A. (B) Western blots were prepared from whole cell lysates of transfected HEK 293T cells, and SEAP fusion proteins were detected with anti-myc antibody. The upper bands of the fusion proteins, except of that from SH^{N}-$SEAP^{DSP}$, represent the N-glycosylated and N-terminally processed protein population (c), the lower band the non processed population (arrow). The values below the lanes show the density ratio of the upper band to the lower band for each fusion protein which decreases the more the transition area is impaired.
doi:10.1371/journal.pone.0002767.g005

microcentrifuge at 4C for 5 min. Where indicated, cell lysates were treated with PNGase F which removes N-glycans according to the manufacturer's instructions (New England Biolabs, Frankfurt, Germany). For immunoblotting, 20mg of protein from each cell lysate was separated in a 6% SDS PAA-gel. Protein blots were incubated with rabbit polyclonal anti-myc antibody (0.5g/ml; Sigma-Aldrich Chemie GmbH, Mu¨nchen, Germany) diluted in TBST (10 mmol/L Tris-HCl, pH 7.4, 150 mmol/L NaCl; 0.05% Tween 20). Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was probed with mouse monoclonal anti-GAPDH antibody (1mg/ml, Ambion/Applied Biosystems, Darmstadt, Germany), cytochrome c with mouse monoclonal anti-cytochrome c antibody (0.4mg/ml; medac, Wedel, Germany) and Grp94 with rat monoclonal anti-Grp94 antibody (2mg/ml; medac, Wedel, Germany). Secondary alkaline phosphatase-conjugated goat anti-rabbit, horseradish peroxidase conjugated goat anti-mouse antibody and horseradish peroxidase conjugated goat anti-rat antibody (all Jackson ImmunoResearch, Dianova GmbH, Hamburg, Germany) were used for detection offi rst antibodies. Enzyme substrates were NBT/BCIP (Roche Diagnostics GmbH, Mannheim, Germany) for alkaline phosphatase or a solution of luminol (2.5 mM), p-coumaric acid (0.4 mM), Tris-HCl, pH 8.5 (100 mM) and 0.009% H_2O_2 for horseradish peroxidase.

Densitometric analysis

The densitometric analysis of the Western blots was performed with Image J (Scion). The densities of the corresponding bands on the blot were measured and the ratio of the upper band to the lower band of each construct was calculated.

Isolation of mitochondria

24 hours after transfection of HEK 293T cells mitochondria were isolated with the Qproteome Mitochondria Isolation Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Briefly, after removal of nuclei, cell debris, cytosolic and microsomal cell fractions, the mitochondria pellet was resuspended in 0.5 M sucrose buffer (1 mM EDTA, 0.1% BSA, 10 mM Tris-HCl, pH 7.5), layered on a 1–2 M sucrose gradient (1 mM EDTA, 0.1% BSA, 10 mM Tris-HCl, pH 7.5) and centrifuged for 2 h at 25,000 rpm. The mitochondrial band was collected, diluted with 2 volumes of 1 mM EDTA, 10 mM Tris-HCl, pH 7.4 buffer and pelleted by centrifugation at 20,000 g for 15 min. 20 mg of protein of each fraction was loaded on a 10% PAA-gel and separated by SDS-PAGE.

Supporting Information

Table S1 Vertebrate signal peptides. 40 amino acids, which are predicted to be NtraC organized but differ in their domain capacity from shrew-1. Underlined residues are predicted turns belonging to the T-domain
Found at: doi:10.1371/journal.pone.0002767.s001 (0.17 MB DOC)

Figure S1 Multiple sequence alignment of the signal peptides of shrew-1 homologues
Found at: doi:10.1371/journal.pone.0002767.s002 (0.37 MB DOC)

Text S1 Oligonucleotides used for cloning of SEAP fusion constructs.
Found at: doi:10.1371/journal.pone.0002767.s003 (0.06 MB DOC)

Acknowledgments

We thank Matthias Schmidt and Monika Kamprad for technical support, and Bernhard Dobberstein, Katja Kapp, and Paul Wrede for fruitful discussion. Norbert Dichter helped us set up the web interface.

Author Contributions

Conceived and designed the experiments: JAH ER AS ASP. Performed the experiments: JAH ER. Analyzed the data: JAH ER AS ASP GS. Contributed reagents/materials/analysis tools: MM. Wrote the paper: JAH GS. Performed the biological experiments: ER. Designed and supervised the biological experiments and analyzed the biological data: AS-P AS. Analyzed the bioinformatical data and developed the NtraC model: JH.

References

1. Blobel G (2000) Protein targeting. Biosci Rep 20: 303–344.
2. Nickel W (2005) Unconventional secretory routes: direct protein export across the plasma membrane of mammalian cells. Traffic 6: 607–614.
3. Ouzzine M, Magdalou J, Burchell B, Fournel-Gigleux S (1999) An internal signal sequence mediates the targeting and retention of the human UDP-glucuronosyltransferase 1A6 to the endoplasmic reticulum. J Biol Chem 274: 31401–31409.
4. Champion PA, Stanley SA, Champion MM, Brown EJ, Cox JS (2006) C-terminal signal sequence promotes virulence factor secretion in Mycobacterium tuberculosis. Science 313: 1632–1636.
5. Gray L, et al. (1989) A novel C-terminal signal sequence targets Escherichia coli haemolysin directly to the medium. J Cell Sci Suppl 11: 45–57.
6. von Heijne G, ed (1994) Signal Peptidases. Austin: R.G. Landes.
7. Watson ME (1984) Compilation of published signal sequences. Nucl Acids Res 12: 5145–5164.
8. Izard JW, Kendall DA (1994) Signal peptides: exquisitely designed transport promoters. Mol Microbiol 13: 765–773.
9. Jungnickel B, Rapoport TA (1996) A posttargeting signal sequence recognition event in the endoplasmic reticulum membrane. Cell 82: 261–270.
10. Martoglio B (2003) Intramembrane proteolysis and post-targeting functions of signal peptides. Biochem Soc Trans 31: 1243–1247.
11. von Heijne G (1990) The signal peptide. J Membr Biol 115: 195–201.
12. Feng ZP (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 58: 491–499.
13. Schneider G, Fechner U (2004) Advances in the prediction of protein targeting signals. Proteomics 4: 1571–1580.
14. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2: 953–971.
15. Shen H-B, Chou K-C (2007) Signal-3L: A 3-layer approach for predicting signal peptides. Biochem Biophys Res Comm 363: 297–303.
16. Chou K-C, Shen H-B (2007) Signal-CF: A subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Comm 357: 633–640.
17. Horton P, et al. (2007) WoLF PSORT: protein localization predictor. Nucl Acids Res 35: W585–W587.
18. Tamura T, Akutsu T (2007) Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition. BMC Bioinformatics 8: 466.
19. Wiedmann M, Huth A, Rapoport TA (1984) Xenopus oocytes can secrete bacterial beta-lactamase. Nature 309: 637–639.
20. Al-Qahtani A, Teilhet M, Mensa-Wilmot K (1998) Species-specificity in endoplasmic reticulum signal peptide utilization revealed by proteins from Trypanosoma brucei and Leishmania. Biochem J 331: 521–529.
21. Hegde RS, Bernstein HD (2006) The surprising complexity of signal sequences. Trends Biochem Sci 31: 563–571.
22. Chou K-C, Shen H-B (2008) Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. Nature Protocols 3: 153–162.
23. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783–795.
24. Froeschke M, Basler M, Groettrup M, Dobberstein B (2003) Long-lived signal peptide ofl ymphocytic choriomeningitis virus glycoprotein pGP-C. J Biol Chem 278: 41914–41920.
25. Ramanujan S, Bernstein HD (2006) The surprising complexity of signal sequences. Biochem Sci 31: 563–571.
26. Martoglio B, Dobberstein B (1998) Signal sequences: more than just greasy peptides. Trends Cell Biol 8: 410–415.
27. Kurys G, Tagaya Y, Bamford R, Hanover JA, Waldmann TA (2000) The long signal peptide isoform and its alternative processing direct the intracellular trafficking ofi nterleukin-15. J Biol Chem 275: 30653–30659.
28. Dultz E, Hildenbeutel M, Martoglio B, Hochman J, Dobberstein B, et al. (2008) The signal peptide of the mouse mammary tumor virus Rem protein is released from the endoplasmic reticulum membrane and accumulates in nucleoli. J Biol Chem 283: 9966–9976.
29. Bharti S, Handrow-Metzmacher H, Zickenheiner S, Zeitvogel A, Baumann R, et al. (2004) Novel membrane protein shrew-1 targets to cadherin-mediated junctions in polarized epithelial cells. Mol Biol Cell 15: 397–406.
30. Resch E, Quaiser S, Quaiser T, Schneider G, Starzinski-Powitz A, et al. (2008) Synergism of shrew-1's signal peptide and transmembrane segment required for plasma membrane localization. Traffic, in press.
31. Jakob V, Schreiner A, Tikkanen R, Starzinski-Powitz A (2006) Targeting of transmembrane protein shrew-1 to adherens junctions is controlled by cytoplasmic sorting motifs. Mol Biol Cell 17: 3397–3408.
32. Schreiner A, Ruonala M, Jakob V, Suthaus J, Boles E, et al. (2007) Junction protein shrew-1 influences cell invasion and interacts with invasion-promoting protein CD147. Mol Biol Cell 18: 1272–1281.
33. Meissner M, Koch O, Klebe G, Schneider G (2008) Prediction of turns types in protein structure by machine-learning classifiers. Proteins, in press.
34. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, et al. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucl Acids Res 34: D187–191.
35. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005–1016.
36. Hutchinson EG, Thornton EG (1994) A revised set of potentials for beta-turn formation in proteins. Protein Sci. 3: 2207–2213.
37. Kä̈ll L, Krogh A, Sonnhammer E (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics 21: 251–257.
38. Yuan Z, Mattick JS, Teasdale RD (2004) SVMtm: Support vector machines to predict transmembrane segments. J Comput Chem 25: 632–636.
39. Berger J, Hauber J, Hauber R, Geiger R, Cullen BR (1988) Secreted placental alkaline phosphatase: a powerful new quantitative indicator of gene expression in eukaryotic cells. Gene 66: 1–10.
40. Foulquier F, Harduin-Lepers A, Duvet S, Marchal I, Mir AM, et al. (2002) The unfolded protein response in a dolichyl phosphate mannose deficient Chinese hamster ovary cell line points out the key role of a demannosylation step in the quality-control mechanism of N-glycoproteins. Biochem J 362: 491–498.
41. Takasugi N, Tomita T, Hayashi I, Tsuruoka M, Niimura M, et al. (2003) The role of presenilin cofactors in the gamma-secretase complex. Nature 422: 438–441.
42. Hansson CA, Frykman S, Farmery MR, Tjernberg LO, Nilsberth C, et al. (2004) Nicastrin, presenilin, APH-1, and PEN-2 form active gamma-secretase complexes in mitochondria. J Biol Chem 279: 51654–51660.
43. Robin MA, Anandatheerthavarada HK, Biswas G, Sepuri NB, Gordon DM, et al. (2002) Bimodal targeting of microsomal CYP2E1 to mitochondria through activation of an N-terminal chimeric signal by cAMP-mediated phosphorylation. J Biol Chem 277: 40583–40593.
44. von Heijne G (1986) The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO J 5: 3021–3027.
45. Szabady RL, Peterson HP, Skillman KM, Bernstein HD (2005) An unusual signal peptide facilitates late steps in the biogenesis of a bacterial autotransporter. Proc Natl Acad Sci USA 102: 221–226.

Danksagung

Ich danke Prof. Dr. Gisbert Schneider für die Betreuung der Arbeit, für die fortwährende Motivation und Unterstützung und die wissenschaftliche Ausbildung.

Ich danke Prof. Dr. Paul Wrede für die Betreuung der Arbeit, die zahlreichen kreativen und hilfreichen Diskussionen und den Anstoß, immer die Perspektive zu wechseln.

Ich danke Prof. Dr. Anna Starzinski-Powitz für die Möglichkeit, meine Thesen *in vitro* überprüfen zu können und für die Unterstützung bei allen biologischen Aspekten der Arbeit.

Ich danke Dr. Alexander Schreiner und Dipl.-Biologe Eduard Resch für die Unterstützung bei allen experimentellen Fragen - und für die Freundschaft.

Ich danke den *üblichen Verdächtigen* aus meinem Jahrgang für die Begleitung in allen Aufs und Abs des Studiums, für die Freundschaft und die vielen wundervollen Reisen.

Ich danke dem modlab Team für die schöne Zeit, die immer freundliche Atmosphäre und die Kuchen.

Ich danke den Mitgliedern des Labors von Frau Prof. Dr. Anna Starzinski-Powitz für die tolle Arbeits-Atmosphäre und die fröhliche Stimmung.

Ich danke Norbert Dichter und Monika Kamprad für die Bereitstellung aller technischen und experimentellen Mittel, ohne die diese Arbeit nicht möglich gewesen wäre.

Ich danke meinen Eltern für die lebenslange Unterstützung meiner Person und Arbeit und den unerschütterlichen Glaube an meine Fähigkeiten, meine Zukunft und mich.

Lebenslauf

Persönliche Daten

Name, Vorname	Hiß, Jan Alexander
Geburtsdatum /-ort	2. Oktober 1979 / Groß-Gerau
Nationalität	deutsch

Schulbildung

06/1999	Allgemeine Hochschulreife (Note: 1,0)

Hochschulstudium

10/2000- 08/2005

Bioinformatik (Diplom)

Johann Wolfgang von Goethe-Universität, Frankfurt/Main

Diplomarbeit:

- Interdisziplinäre Diplomarbeit in Frankfurt/Main und an der FU/ Charité Berlin
- Titel: „Peptiddesign unter Zuhilfenahme eines Ameisenalgorithmus in einer virtuellen Fitnesslandschaft“
- Betreuer: Prof. Dr. G. Schneider
  (J.W.G.- Universität Frankfurt/Main)
  Prof. Dr. P. Wrede
  (Freie Universität Berlin)
  (Note Diplomarbeit: sehr gut 1,1)
- Abschluss: Diplom-Bioinformatiker
- Note Diplom: sehr gut (1,3)

ab 10/2005

Promotion in Biologie

Johann Wolfgang Goethe-Universität, Frankfurt/Main, FB15 Biowissenschaften

Titel der Dissertation:

Domänen-Architektur von langen Signalpeptiden
–*in silico* und *in vitro* –

Publikationen

Hiss, J.A., Resch, E., Schreiner, A., Meissner, M., Starzinski-Powitz, A. und Schneider, G. (2008) Domain Organization of Long Signal Peptides of Single-Pass Integral Membrane Proteins Reveals Multiple Functional Capacity. *PLoS ONE* 3, e2767.

Hiss, J.A., Przyborski, J.M., Schwarte, F., Lingelbach, K. und Schneider, G. (2008) The Plasmodium Export Element Revisited. *PLoS ONE 3*, e1560.

Hiss, J.A., Bredenbeck, A., Losch, F.O., Wrede, P., Walden, P. und Schneider, G. (2007) Design of MHC I stabilizing peptides by agent-based exploration of sequence space. *Protein Eng. Des. Sel.* 20, 99-108.

Eidesstattliche Versicherung

Eidesstattliche Versicherung

Ich erkläre hiermit an Eides Statt, dass ich die vorgelegte Dissertation über

Domänen-Architektur von langen Signalpeptide

- *in silico* und *in vitro* -

selbständig angefertigt und mich anderer Hilfsmittel als der in ihr angegebenen nicht bedient habe, insbesondere, dass aus Schriften Entlehnungen, soweit sie in der Dissertation nicht ausdrücklich als solche mit Angabe der betreffenden Schrift bezeichnet sind, nicht stattgefunden haben.

Frankfurt am Main, den..... .....

(Unterschrift)