Annotation der cDNA-Datenbank von *Verticillium longisporum* und Identifikation von adhäsions-relevanten Proteinen

Diplomarbeit

vorgelegt von
Ingrid Hartwig
aus
Hannover

angefertigt
im Institut für Mikobiologie und Genetik
an der biologischen Fakultät
der Georg-August-Universität zu Göttingen

2008

Referent : Prof. Dr. Burkhard Morgenstern

Koreferent : Prof. Dr. Gerhard H. Braus

Tag der Abgabe der Diplomarbeit : 31.10.2008

Inhaltsverzeichnis

Danksagung

Zunächst möchte ich mich an dieser Stelle bei allen Personen bedanken, die mich während dieser Arbeit unterstützt und angeleitet haben.

Ich danke Prof. Burkhard Morgenstern und Prof. Gerhard Braus dafür, dass sie mir diese Arbeit ermöglicht haben. Mein besonderer Dank gilt Maike Tech, die viel Zeit in die Betreuung und Anleitung sowie für die Korrektur dieser Arbeit investiert hat. Ebenfalls danke ich Susanna Braus-Stromeyer und ihrer Gruppe für die Bereitstellung der Sequenzen, die Einführung in das Thema, für Anregungen und Korrekturen.

Mein Dank gilt auch den hilfsbereiten Korrekturlesern Friederich Limbach und Meike Bruns.

Besonders bedanken möchte ich bei meinen Eltern, die mich von Anfang an unterstützt haben. Sie haben mich immmer dazu ermutigt weiter zu lernen und für meine Ziele zu arbeiten. Ich danke vor allem Jonas dafür, dass er mir während dieser Arbeit vielfach geholfen und mich zuverlässig mit Kaffee versorgt hat.

Abkürzungen und fremdsprachliche Ausdrücke

A	Adenin
Abb.	Abbildung
AS	Aminosäure
BLAST	Basic Local Alignment Search Tool
blastn	Unterprogramm von BLAST, zur Datenbanksuche mit Nukleotidsequenzen
bp	Basenpaar
C	Cytosin
C-Terminus	Carboxyl-Ende eines Proteins
ca.	Circa
cDNA	Komplementäre Desoxyribonukleinsäure
ddATP	Didesozyadenosintriphosphat
ddCTP	Didesoxycytidintriphosphat
ddGTP	Didesoxyguanosintriphosphat
ddNTP	Didesoxynukleosidtriphosphat
ddTTP	Didesoxythymidintriphosphat
DNA	Desoxyribonukleinsäure
downstream	In 3'-Richtung auf dem DNA- bzw. RNA-Strang gelegen
ELISA	Enzyme linked immunosorbent-assay
Enhancer	Abschnitt der DNA, der an der Verstärkung der Transkription beteiligt ist
ER	Endoplasmatisches Retikulum
EST	Expressed Sequence Tag
EU	Europäische Union
Exon	Teil der prä-mRNA, der nach dem Spleißen reifen mRNA
FASTA-Format	Format zur Darstellung der Sequenz von Proteinen und DNA
G	Guanin
gap	Lücke in der Sequenz bei einem Sequenzalignment
GPI	Glycosylphosphatidylinositol
ha	Hektar
HMM	Hidden Markov Model
ID	Identifikationsbezeichnung
IUPAC	International Union of Pure and Applied Chemistry
LR	Leserahmen
MCS	Multiple cloning site

mind.	Mindestens
Mio.	Millionen
mRNA	Messenger Ribonukleinsäure
N	Nicht identifiziertes Nukleotid
NCBI	National Center for Biotechnology Information
NR	Non Redundant
N-Terminus	Amino-Ende eines Proteins
NTP	Nukleosidtriphosphat
orf	Offener Leserahmen
Primer	Oligonukleotid, das als Ansatzpunkt für die DNA-Polymerase dient
Query	Abfrage einer Datenbank
RNA	Ribonukleinsäure
RTase	Revers-Transkriptase
S_{ppt}	Physical property term
$S_{profile}$	Bester Score beim Alignment mit bekannten GPI-Sequeunzen
Score	Punktzahl, die einem Treffer entsprechend seiner Güte zugeordnet wird
Ser	Serin
Silencer	Abschnitt der DNA, der an der Hemmung der Transkription beteiligt ist
Spleißen	Prozess, bei dem Introns aus der mRNA herausgeschnitten werden und Exons zu einer reifen mRNA verbunden werden
ssp.	Subspecies (bei nicht bekannter Unterart)
subsp.	Subspecies
T	Thymin
t	Tonnen
tblastx	Unterprogramm von BLAST zum Vergleich der sechs Leserahmen einer Nukleotidsequenz mit den sechs Leserahmen der Datenbanksequenzen
U	Uracil
UFOP	Union zur Förderung von Oel- und Proteinpflanzen e.V.
upstream	In 5'-Richtung auf einem DNA- oder RNA-Strang gelegen
UTR	Untranslated Regions
Thr	Threonin
Z	Zentriolwert

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

1.1 Ziele der Arbeit

In dieser Arbeit sollen exprimierte Gene des Pilzes *Verticillium longisporum*, einem Rapsschädling, mit bioinformatischen Methoden näher untersucht werden. Die Sequenzen einer cDNA-Bank werden zu diesem Zweck bereinigt und eine erste Annotation wird vorgeschlagen. Zusätzlich wird in der Datenbank nach Proteinen gesucht, die die Pathogenität des Pilzes *V. longisporum* begünstigen und Adhäsion, die Anhaftung des Organismus an die Wirtspflanze, ermöglichen. Besondere Aufmerksamkeit wird den Sequenzen gewidmet, die bereits experimentell untersucht wurden und bei denen Hinweise für einen Einfluss auf die Adhäsion vorhanden sind. Es ist zu klären, wodurch dieser Einfluss zustande kommt, ob es sich beispielsweise um Adhäsine handelt oder um Proteine, welche die Adhäsion regulieren.

Um die genannten Aufgaben zu erfüllen, werden verschiedene Ansätze verfolgt. Neben einer allgemeinen BLAST-Suche gegen verschiedene Datenbanken, die zur Annotation der cDNA-Sequenzen dient, wird zusätzlich nach Proteinen gesucht, die bestimmte Eigenschaften mit Adhäsinen teilen.

1.2 Rapsanbau in Deutschland

Der Rapsanbau in Deutschland gewinnt vor dem Hintergrund der steigenden Nachfrage nach Biodiesel und Rapsöl zunehmend an Bedeutung.

Zur Stärkung des Umweltschutzes und auch aus Gründen der Versorgungssicherheit, wird von den Mitgliedsstaaten der Europäischen Union (EU) eine Förderung des Einsatzes von Biokraftstoffen gefordert. Dadurch soll das Maßnahmenpaket zur Einhaltung des Kyoto-Protokolls

erfüllt und gleichzeitig die Abhängigkeit von Energieeinfuhren wie Erdöl gesenkt werden [2]. In Deutschland, das neben Frankreich der größte Rapsproduzent der EU ist [46], gelang eine Steigerung des Biokraftstoffanteils am Kraftstoffmarkt von 1,21% im Jahr 2003 auf 3,75% im Jahr 2005. Dabei stieg die Produktionskapazität von Biodiesel in Deutschland von 0,267 Millionen (Mio.) Tonnen (t) im Jahr 2000 auf 5,08 Mio. t im Jahr 2007 [47]. Ein erforderlicher Schritt, um bis 2020 in den EU-Ländern einen Biokraftstoffanteil von 10% am Kraftstoffverbrauch zu verwirklichen, wird auch in der weiteren Förderung des Rapsanbaus gesehen, aus dem Biodiesel in Europa vorrangig hergestellt wird [27].

Ebenso wie die Nachfrage nach Biodiesel steigt auch die Nachfrage nach Rapsöl. Betrug der Anteil von Rapsöl an den 2004 verkauften Speiseölen noch 7,2%, waren es 2006 bereits 10,7%. Somit handelt es sich beim Rapsöl um das einzige Öl, dessen Marktanteile nicht stagnieren oder rückläufig sind [48].

Durch die erhöhte Nachfrage und stärkere Förderung lässt sich der stark gestiegene Anbau von Winterraps (*Brassica napus*) erklären (Abb. 1.1). Während die Anbaufläche für Winterraps 2004 noch bei 1,267 Mio. Hektar (ha) lag, stieg sie 2005 zunächst um 3,7% auf 1,314 Mio. ha [7] und 2006 um weitere 6,2% auf 1,405 Mio. ha [8]. Nachdem die mit Winterraps angebaute Fläche 2007 einen Höchststand von 1,534 Mio. ha erreichte, sank die Aussaatfläche für die Ernte 2008 um 8,4% auf das Niveau von 2006.

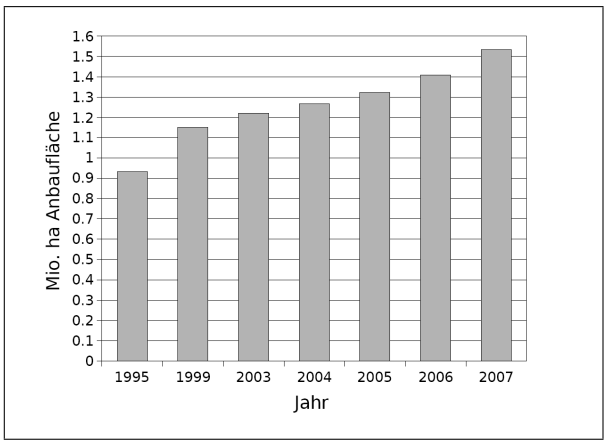

Abbildung 1.1: Diagramm der Anbaufläche von Raps in Deutschland

Für den Rückgang der Anbaufläche sind verschiedene Gründe ausschlaggebend. Neben ungünstigen Witterungsbedingungen für die Aussaat von Winterraps im Herbst 2007 kommen der gestiegene Getreidepreis und das Aussetzen der Flächenstilllegung hinzu [10]. Allein 2005 wurde 27,79% des Rapses in Deutschland auf Stilllegeflächen angebaut, wobei die mit der Stilllegung

verbundenen Direktzahlungen nicht wegfielen, wenn der Raps als Energiepflanzen und nicht für die Herstellung von Nahrungsmitteln verwendet wurde [47]. Nach der geringen Getreideproduktion von 2007 wurde der Stilllegungssatz für die Aussaat im Herbst 2007 und Frühjahr 2008 auf 0% gesetzt, wodurch auf der Fläche, auf der zuvor kein Anbau von Nahrungsmitteln möglich war, nun beispielsweise auch Getreide und andere Nahrungsmittel angebaut werden können [15].

Diese Entwicklungen zeigen die wachsende Bedeutung von Rapsanbau in Deutschland, wenngleich der starke Aufwärtstrend im Anbau 2008 nicht im gleichen Maße wie zuvor fortgesetzt werden kann.

1.3 *Verticillium longisporum* als Verursacher der Rapswelke

Der intensive Rapsanbau und die zunehmende Rapsdichte lassen Pflanzenpathogene wie *Verticillium longisporum* zu einem wachsenden Problem werden. So ist die seit 1985 in Deutschland bekannte Rapswelke (Verticilliumwelke), die durch diesen pathogenen Pilz verursacht wird, mittlerweile im gesamten Bundesgebiet verbreitet. Raps ist dabei der am häufigsten vertretene Wirt von *V. longisporum*, der auch andere Pflanzen, vor allem innerhalb der Gruppe der *Brassicaceae*, befällt. Zu den Wirtspflanzen zählen beispielsweise Pak Choi (*Brassica campestris chinensis*), Chinakohl (*Brassica rapa subsp. pekinensis*), Blumenkohl (*Brassica oleracea var. botrytis*), Senf (*Sinapis alba L*) und Ölrettich (*Raphanus sativus ssp. oleiformes*) [42, 24]. Es wird jedoch auch von Wirtspflanzen berichtet, die nicht zu den *Brassicaceae* gehören [23, 42].

Schleswig-Holstein und Mecklenburg-Vorpommern, die beiden Bundesländer mit den meisten Anbaugebieten von Raps, sind am stärksten von der Verticilliumwelke betroffen [46]. Ein großes Problem im Zusammenhang mit dieser Krankheit ist die Diagnose, die dadurch erschwert wird, dass die Symptome erst in einer späten Wachstumsphase auftreten und nicht eindeutig sind. Die meisten Symptome der Rapswelke können nur schwer von Alterungssymptomen unterschieden werden [42]. Zu den Symptomen der Rapswelke gehören: Verfrühte Abreife, gehemmtes Wachstum ab dem 21. Tag nach der Infektion [16] und eine hellgrüne bis gelbe Stängelfärbung ab beginnender Reife [42]. Die Welkesymptome treten bei Rapspflanzen deutlich stärker auf als bei anderen Wirtspflanzen, wo sie nur schwach ausgeprägt sind [57]. Die sicherste Diagnose ist das Auffinden von Mikrosklerotien an Ernteresten, was jedoch eine ausgesprochene Detailkenntnis erfordert und erst nach der Ernte festzustellen ist. Selbst nach dieser Diagnose wird nicht immer auf die richtige Ursache der frühen Abwelke geschlossen. Der ELISA-Test (Enzyme linked immunosorbent assay)[55], ein immunologisches Nachweisverfahren, ermöglicht dagegen auch schon das frühe Erkennen eines latenten Befalls mit dem Pathogen *Verticillium longisporum* [42]. In Ex-

perimenten führte der Befall einer Pflanze mit diesem Pathogen durch stark virolente Isolate bei 50% der Wirte zum Tod innerhalb von 42 Tagen [57].

Zunächst wurde die Rapswelke dem Erreger *Verticillium dahalie* zugeschrieben, der anfänglich nicht von *V. longisporum* unterschieden wurde. Jedoch unterscheiden sich die Pilze durch die Länge der Kondien, die bei *V. longisporum* deutlich größer sind. Nachdem *V. longisporum* zunächst als Variation von *V. dahliae* gesehen wurde, ist er inzwischen als eigene Art etabliert [24].

Neben morphologischen trugen auch genetische Unterschiede zu dem Schritt bei, eine eigenständige Art vorzuschlagen. Bei *V. longisporum* handelt es sich um einen nahezu diploiden Organismus, dessen DNA-Gehalt 1,78 mal so hoch ist wie der von *V. dahliae* [24]. Ebenfalls werden unterschiedliche Wirtspflanzen befallen, wobei *V. dahliae* das größere Wirtsspektrum hat, jedoch keine starke Pathogenität für Raps zeigt [57]. *V. longisporum* befällt vor allem Rapspflanzen und andere *Brassicaceae* [24].

Der Organismus *V. longisporum* kommt wahrscheinlich in kleinen Mengen in jedem Boden vor [23]. Ist eine Wirtspflanze vorhanden, wird durch die Wurzelexudate der Pflanze bereits nach elf Stunden das Keimen der Mikrosklerotien von *V. longisporum* stimuliert [16]. Die Infektion findet über die Wurzeln statt, wobei diese nicht verletzt sein muss, um das Eindringen des Pathogens zu ermöglichen. Die Hyphen erreichen die Wurzeloberfläche durch gerichtetes Wachstum entlang der Wurzelhaare und dringen direkt in epidermale Zellen ein oder greifen an Verzweigungen der epidermalen Zellen an. Dabei haften die *Verticillium*-Hyphen stark an den Wurzelhaaren an, jedoch nicht an deren Spitzen, die der einzige nicht bewachsene Teil der Wurzelhaare bleiben [16].

V. longisporum dringt ins Xylem der Pflanze ein und verbreitet sich dort zunächst in den Wurzeln, indem benachbarte Xylemzellen durch die Plasmodesmata befallen werden. Drei Wochen nach der Inokulation werden nach den Wurzelgefäßen auch die Gefäße der Sprossachse befallen [16]. Die Kondien breiten sich immer weiter aus und sind mit Beginn der Blüte in der gesamten Pflanze zu finden [58]. Sobald die Pflanze anfängt zu reifen, werden im vaskulären System Mikrosklerotien gebildet, die als Überdauerungsorgane dienen. Diese sind auch nach der Ernte in den Pflanzenresten auf dem Feld zu finden. Dort können sie bis zu fünfzehn Jahre überdauern, bis wieder Wirtspflanzen angebaut werden.

Dadurch, dass die Mikrosklerotien im Boden zurückbleiben, reichert sich der pathogene Pilz immer weiter an, je häufiger Wirtspflanzen angebaut werden. So ist das Auftreten der Krankheit davon abhängig, wie hoch die Anbaukonzentration der Wirtspflanzen auf einem Feld ist. Die Verringerung des Anbauabstandes von Raps von vier auf drei Jahre lässt den Befall der Pflanzen um 50% ansteigen und verdoppelt das Bodeninokulum. Ebenfalls von Bedeutung ist die Vegetationsdauer. Je länger die Pflanzen auf dem Feld verbleiben, desto höher ist die Konzentration von

Pathogenen im Boden. Eine um eine Woche verlängerte Anbauzeit führt zu einem verdoppelten Bodeninokulum [42]. Da weder Bodenart noch Fungizide einen Einfluss auf die Ausbreitung der Krankheit oder die Anreicherung des Pathogens im Boden haben, ist die Kontrolle der Fruchtfolge und das Einschränken der Wachstumsdauer derzeit die einzige Möglichkeit, der Etablierung der Krankheit in Norddeutschland entgegenzuwirken und die Verbreitung im Rest des Landes zu verhindern [42]. Dies jedoch würde eine Einschränkung des Rapsanbaus in Deutschland bedeuten.

1.4 Adhäsion

Adhäsion, das Haften an biotischen und abiotischen Oberflächen, ermöglicht Pilzen einen pathogenen Befall von Organismen [49]. Durch die Haftung sind die Pilze vor der Entfernung von günstigen Umgebung geschützt [51]. Die adhäsiven Eigenschaften können durch unterschiedliche Proteine vermittelt werden. Eine wichtige Rolle spielen dabei Adhäsine, aber auch Integrine und Hydrophobine können es Pilzen ermöglichen an verschiedenen Oberflächen zu haften.

Adhäsine werden von Pilzen für eine Vielzahl von Interaktionen mit ihrer Umgebung gebraucht. So können *Candida ssp.* durch Adhäsion an medizinischen Instrumenten und Geräten einen Biofilm bilden, der sich über Katheter, Prothesen oder auch Herzschrittmacher legt, und auf diesem Weg Infektionen im Menschen auslösen [11]. Ein anderes Beispiel ist *Metarhizium anisopliae*, ein Pilz der sowohl an der Oberfläche von Pflanzenwurzeln, als auch an Insekten und Plastik haften kann. Die Adhäsine *MAD1* und *MAD2* spielen dabei eine wichtige Rolle: *MAD1* für die Anhaftung an Insekten und *MAD2* für die an Pflanzen. Ohne diese beiden Proteine wird die Adhäsion um 90% schwächer und auch die Virulenz des Organismus wird reduziert, da der erste Schritt des Angriffs auf den Wirt nicht mehr möglich ist [53]. Ebenfalls möglich ist eine Adhäsion zwischen zwei Zellen der gleichen Art, wie es beispielsweise zwischen Hefezellen vorkommt. Diese Art der Interaktion, bei der die Zellen in wässriger Lösung durch Aneinanderhaften wie Flocken ausfallen, wird auch Flockulation genannt [51].

Vor dem Hintergrund dieser vielfältigen Interaktionen zwischen pathogenen Pilzen und biotischen und abiotischen Oberflächen liegt es nahe, auch für *V. longisporum* nach möglichen Adhäsinen zu suchen. Diese können für das Anhaften des Erregers an den Wurzeln von Rapspflanzen, sowie das Haften der Zellen innerhalb des vaskularen Systems, verantwortlich sein.

1.4.1 Adhäsine

Alle oben beschriebenen Interaktionen von Pilzen miteinander, mit anderen Organismen oder abiotischen Oberflächen werden durch Adhäsine ermöglicht (Abb. 1.2). Diese Zelloberflächen-

Proteine besitzen eine gemeinsame Grundstruktur, bestehend aus drei Domänen [20, 49]: Die N-terminale Domäne des Proteins ragt aus der Zelloberfläche hervor und ist für Ligandenbindung zuständig. Je nach Art der Bindung lassen sich zwei Gruppen von Adhäsinen unterscheiden: Zucker-sensitive und Zucker-insensitive [51]. Zucker-sensitive Adhäsine haben eine Lektin ähnliche Kohlenhydrat-Bindedomäne [25, 38], Zucker-insensitive binden an Peptide oder erhöhen die Hydrophobie der Zelle, sodass hydrophobe Wechselwirkungen mit Oberflächen ermöglicht werden [51].

Die zentrale Domäne ist circa 300 Aminosäuren (AS) lang, reich an Sequenzwiederholungen und enthält einen erhöhten Anteil an Serin (Ser) und Threonin (Thr) [51], der 35-55% beträgt [31]. Diese Domäne formt einen Stiel, durch den die N-terminale Domäne über die Zellwand hinauszuragen vermag, um mit anderen Oberflächen in Kontakt zu treten. Je länger dieser Sequenzteil ist, desto weiter können die Oberflächen, mit denen der Pilz interagiert, von der Zellwand entfernt sein [50]. Des Weiteren sind in dieser Region vermehrt hydrophobe Reste zu finden, sowie eine erhöhte Menge an Cysteinen in der C-terminalen Region dieser Domäne. Die Funktion dieser beiden Eigenschaften ist jedoch noch ungeklärt [12].

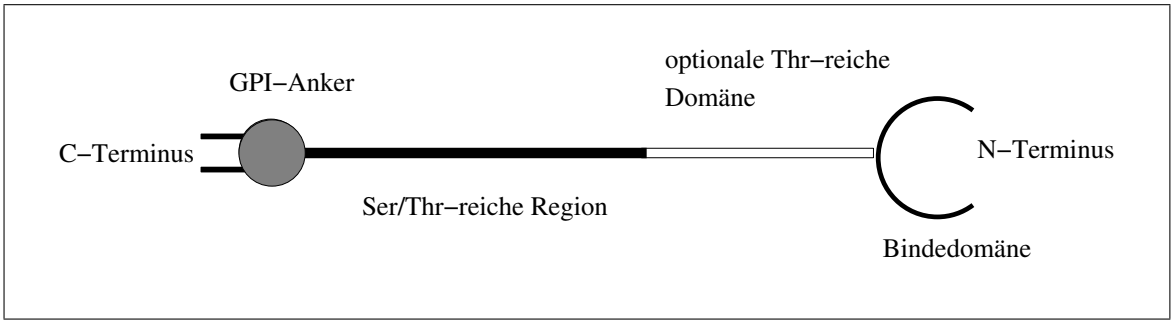

Abbildung 1.2: **Schema eines Adhäsins**Der GPI-Anker ermöglicht die Verankerung des Proteins in der Zellwand. Die Ser/Thr-reiche Domäne ragt über diese hinweg und ermöglich mit Hilfe der N-terminalen Bindedomäne die Interaktion mit Oberflächen

Am C-Terminus der Adhäsine wird posttranskriptional ein Glycosylphosphatidylinositol(GPI)-Anker angebracht [12]. Der GPI-Anker kann ein Protein sowohl in der Zellwand als auch an der Plasmamembran verankern. Dabei erfolgt nur die Anhaftung an der Zellwand durch eine kovalente Bindung. Diese bildet sich zwischen den Mannose-Resten des GPI-Ankers und den β-1,6-Glucanen der Zellwand [37].

In einigen Adhäsinen ist eine vierte, optionale Domäne zwischen der N-terminalen Bindedomäne und dem Ser- und Thr-reichen Stiel zu finden. Diese enthält Thr-reiche Sequenzwiederholungen sowie andere AS, die eine Faltung des Proteins zu einem β-Faltblatt ermöglichen. Diese

Region ist einige hundert AS lang und beeinflusst, ebenso wie die Ser/Thr-reiche Region, den Grad der Interaktion mit Oberflächen.

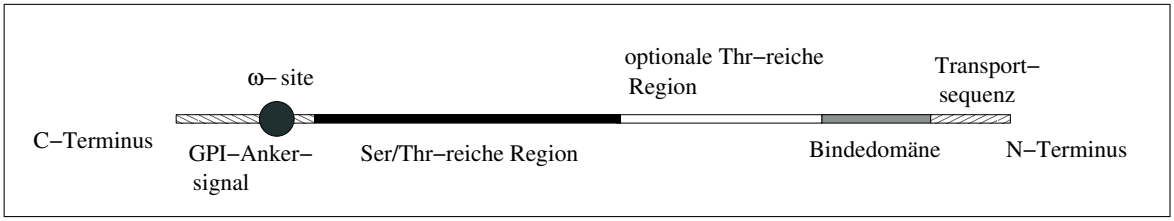

Abbildung 1.3: **Vorläuferprotein von Adhäsinen** Die vollständige Transportsequenz sowie der C-terminale Teil des GPI-Ankersignals werden im Lauf der posttranslationalen Modifikation entfernt. Der Teil des GPI-Ankersignals, der N-terminal der ω-site liegt, bleibt auch nach der Modifikation erhalten. An der ω-site wird im ER der GPI-Anker angefügt.

Vorläufer der Adhäsine haben zusätzliche Domänen, die im Laufe der posttranslationalen Modifikation entfernt werden (Abb. 1.3). Am N-terminalen Ende der Vorläuferproteine befindet sich eine 20-30 AS lange Signalsequenz, die für den Export des Proteins ins Endoplasmatische Retikulum (ER) nötig ist, wo der GPI-Anker durch die GPI-Transamidase angebracht wird [14]. Die Signale, die für das Durchlaufen des Sekretionsweges nötig sind, werden während der einzelnen Sekretionsschritte entfernt [20].

Damit der GPI-Anker angebracht werden kann, muss am C-terminalen Ende ebenfalls eine Signalsequenz mit einer Länge von ca. 40 AS vorhanden sein [14]. Diese Sequenz besitzt bestimmte Eigenschaften: An der Stelle, an der ein Teil der Signalsequenz abgespalten wird, der omega-site (ω *-site*), müssen AS mit kurzen Resten vorkommen, da dieser Teil des Proteins in das katalytische Zentrum der GPI-Transamidase passen muss. Ebenfalls ist eine hydrophobe Domäne am C-Terminus des Vorgängerproteins nötig, etwa 10-12 AS von der C-terminalen ω-site entfernt [34]. Dieser hydrophobe Bereich enthält viele Leucin-, Alanin- und Valin-Reste. Die 10-12 AS zwischen der ω-site und der hydrophoben Region sind nur moderat polar. In N-terminaler Richtung der ω-site ist ein weiterer polarer Bereich, der einen erhöhten Anteil an Serin und Threonin aufweist und in die Ser- und Thr-reiche Domäne des Adhäsins übergeht [14].

1.4.2 Weitere Proteine, die Adhäsion fördern

Neben Adhäsinen sind weitere Proteine bekannt, die die Anhaftung einer Zelle an andere Zellen oder abiotische Oberflächen fördern. Ein Beispiel sind Hydrophobine, kleine Proteine mit einer Länge von ca. 100 AS [56]. Sie ändern die Hydrophobie einer Zelle und ermöglichen so Interaktio-

nen zwischen Oberflächen und Zellen durch hydrophobe Wechselwirkung [35]. Dadurch sind sie in der Lage den Befall eines pflanzlichen Wirtes einzuleiten [13]. Unter den Hydrophobinen sind zwei unterschiedliche Klassen bekannt, die sich dadurch unterscheiden, dass Hydrophobine der zweiten Klasse einen erhöhten Anteil an geladenen AS besitzen [56]. Die Veränderung der Oberflächeneigenschaften wird allein durch die AS-Sequenz vermittelt, es müssen zu diesem Zweck keine Lipide am Protein gebunden sein [56]. In Hydrophobinen sind nur zwei Sequenzmotive bekannt: Das erste besteht aus acht Cysteinen, die in einem charakteristischen Muster zwischen beliebigen anderen AS verteilt liegen. Dabei liegen das zweite und dritte, sowie das sechste und siebte Cystein direkt nebeneinander.

Durch die Cystein-Reste werden intramolekulare Disulfidbindungen ausgebildet, wobei immer vier Cysteine in einer der zwei Domänen des Hydrophobins liegen. Der Abstand zwischen den Cystein-Resten ist gleichzeitig ein weiteres Unterscheidungsmerkmal zwischen Hydrophobinen der ersten und zweiten Klasse. Außerdem ist ein N-terminales Sekretionssignal vorhanden, mit dessen Hilfe das Protein aus der Zelle heraus transportiert wird. Nach dem Export wird von außen die Hydrophobie der Zelle verändert [13].

Die transmembranen Integrine fördern dagegen die Adhäsion an Zelloberflächen durch direktes Binden an Liganden. Sie bestehen aus den zwei Untereinheiten α und β, die nicht kovalent miteinander verknüpft, aber jeweils paarweise an der Zelloberfläche angeordnet sind. Jede von ihnen hat einen kurzen C-terminalen Bereich, der im Cytoplasma liegt und eine lange exrazelluläre, N-terminale Domäne. Die β-Untereinheit enthält im extrazellulären Bereich vier Sequenzwiederholungen, die jeweils 40 Cysteine enthalten. Diese bilden intramolekulare Disulfidbindungen [21]. Die Bindespezifität der beiden Untereinheiten unterscheidet sich voneinander. Als Liganden werden vor allem Proteine mit einer Sequenz, die ein Arginin-Glycin-Asparaginsäure Motiv aufweist, benutzt. Dies ist vor allem bei Proteinen der extrazellulären Matrix der Fall [45]. Da Integrine in der Membran eines Organismus verankert sind, aber gleichzeitig als Rezeptor für Proteine der Zelloberfläche dienen, ermöglichen sie die Anhaftung einer Zelle an der Oberfläche einer anderen.

1.5 Genetische Grundlagen

Das Erbgut eines Lebewesens liegt in Form der Desoxyribonukleinsäure (DNA) vor, die für den Aufbau von Proteinen kodiert [19]. Die in der DNA enthaltene Information wird in mehreren aufeinander folgenden Schritten verarbeitet und genutzt, um Proteine zu bilden. Im Zellkern findet die Transkription statt, die eigentliche Synthese der Proteine erfolgt im Cytosol.

Durch die spezifische Abfolge der Basen Adenin (A), Thymin (T), Guanin (G) und Cytosin (C)

innerhalb eines DNA-Stranges ist die Speicherung der unterschiedlichen genetischen Informationen möglich [54]. Während der Transkription wird durch eine DNA-abhängige RNA-Polymerase an einem DNA -Strang eine *mRNA* (messenger Ribonukleinsäure) gebildet [3]. Der Strang, der bei der Bildung der mRNA als Matrize dient, wird auch als kodogener Strang bezeichnet. An diesem DNA-Strang lagern sich bei der Transkription komplementäre Nukleotide an. Während in der DNA A und T, sowie G und C komplementär zueinander sind, wird in der RNA anstelle von Thymin Uracil (U) eingebaut.

Durch den Aufbau der DNA hat jeder Strang eine definierte Richtung. Die einzelnen Nukleotide sind über ihr 3'- und ihr 5'- Kohlenstoffatom mit Phosphatgruppen verknüpft. An den Enden des Stranges ist einmal ein 3'-Kohlenstoff und einmal ein 5'-Kohlenstoff ungebunden. Eines wird dementsprechend als 3'-Ende bezeichnet, das andere das 5'-Ende [30]. Bei der Transkription bewegt sich die RNA-Polymerase auf dem DNA-Strang immer nur von 3'- in 5'-Richtung, der RNA-Strang wird also von 5'- in 3'-Richtung synthetisiert.

Bei der Transkription entsteht zunächst ein *prä-mRNA*-Strang, der noch verschiedene Prozesse durchläuft, bevor er zur reifen mRNA wird. Zu den posttranskriptionalen Modifikationen gehört das Anfügen eines *Caps* und eines Poly-A-Schwanzes, sowie das *Spleißen*. Die Cap-Struktur, ein modifiziertes Guanin, wird am 5'-Ende der mRNA angefügt und ist für den Transport der mRNA ins Cytosol nötig. Am 3'-Ende der mRNA wird an einer Erkennungssequenz ein Poly-A-Schwanz angefügt. Wurde die Erkennungssequenz transkribiert, wird an dieser Stelle die RNA-Sequenz gespalten. Die Matrizen-unabhängige Poly(A)-RNA-Polymerase fügt 100-250 Adenine an und erhöht dadurch die Stabilität und Lebensdauer der RNA [40]. Weitere nicht-kodierende Bereiche der prä-mRNA (*untranslated regions - UTR*), sind die *Introns*, die beim Spleißen aus dem Strang herausgeschnitten werden. Sie verbleiben im Nukleus, wo sie abgebaut werden. Die Sequenzabschnitte, die ins Cytoplasma exportiert werden, sind die *Exons*. Sie können neben Bereichen, die für Proteine kodieren, auch nicht-translatierte Regionen aufweisen. Diese nicht-kodierenden Regionen treten dann aber ausschießlich an den Ende des Stranges und nicht zwischen kodierenden Bereichen auf.

Vom Zellkern aus wird die mRNA ins Cytoplasma transportiert, wo sich Ribosomen, katalytische Komplexe aus Proteinen und RNA, an den RNA-Strang anlagern. Im Verlauf der Translation werden durch die Ribosomen an der mRNA Proteine gebildet. Dabei wird die mRNA wird durch das Ribosom hindurch geschoben, wobei jeweils drei Basen gleichzeitig abgelesen werden und für eine AS kodieren [43]. Die Translation beginnt am Kodon AUG, das für die Aminosäure Methionin kodiert und endet mit einem der drei Stopp-Kodons UAA, UAG oder UGA [28].

Alle AS weisen eine gemeine Grundstruktur auf. Diese besteht aus einer an einem Kohlenstoff gebundenen Animogruppe, einer Carboxylgruppe, einem Wasserstoffatom und einer Restgrup-

pe, die sich von AS zu AS unterscheidet. Reagieren zwei AS unter Austritt eines Wassermoleküls zu einem Dipeptid, so wird der Kohlenstoff der Carboxylgruppe kovalent mit dem Stickstoff der Aminogruppe verknüpft. Eine der AS hat nun eine freie Carboxyl-, die andere eine freie Aminogruppe. Dem entsprechend werden die beiden Enden C-Terminus und N-Terminus genannt. Auch bei der Verlängerung des Dipeptides zu einem Polypeptid sind die zwei verschiedene Enden immer zu unterscheiden. Der N-Terminus des Proteins wird vom Startkodon, der C-Terminus vom 3'-Ende der mRNA kodiert [43].

1.6 Experimentelle Vorarbeiten an *V. longisporum*

In dieser Arbeit werden Sequenzdaten von *V. longisporum* analysiert. Diese wurden bereits im Vorfeld durch das Anlegen und Sequenzieren einer cDNA-Bank erhalten. Außerdem wurden erste Experimente durchgeführt, um Gene, die für die Adhäsion relevant sind, zu finden. Diese vorausgegangenen Arbeitsschritte werden im Nachfolgenden erläutert.

Eine cDNA-Bank ist eine Sammlung von DNA-Sequenzen, die aus der mRNA eines Organismus gewonnen wird. Sie enthält also nur transkribierte Bereiche des Genoms. Durch die Analyse der cDNA-Sequenzen lässt sich ermitteln, welche Gene aktiv sind. Dabei ist es jedoch von den Umweltbedingungen und Wachstumsphasen des jeweiligen Organismus abhängig, welcher Teil des Genoms zu dem Zeitpunkt des Erstellens der Bank exprimiert wird. Es wird immer nur eine Teilmenge der Gene erfasst.

1.6.1 Das Erstellen einer cDNA-Bank

Um cDNA-Sequenzen zu erstellen, muss zunächst die mRNA einer Zelle isoliert werden. Mit Hilfe einer RNA-abhängigen DNA-Polymerase, auch Reverse-Transkriptase (RTase) genannt, kann mit einem mRNA-Strang als Matrize ein DNA-Strang synthetisiert werden. Dieser komplementäre DNA-Strang wird auch *cDNA* (complementary DNA) genannt. Damit die RTase mit der Transkription beginnen kann, muss ein doppelsträngiger Bereich vorliegen. Dazu wird ein *Primer* hergestellt, ein Oligonukleotid, das komplementär zum 3'-Ende der mRNA ist. Er lagert sich an den Einzelstrang an und bildet so einen doppelsträngigen Bereich. Da bekannt ist, dass die mRNA einen Poly-A-Schwanz hat, kann ein Primer für diesen Bereich hergestellt werden. Er besteht dann ausschließlich aus Thymidinen. Dieser Primer lagert sich an den Poly-A-Schwanz an und die RTase beginnt am Primer einen zur RNA komplementären DNA-Strang zu bilden. Sobald die RTase vom DNA-Strang abfällt oder bis zum 5'-Ende der RNA durchgelaufen ist, liegt ein Doppelstrang aus RNA und cDNA vor. Die RNA wird durch alkalische Lyse abgebaut und so wird einzelsträngige cDNA gewonnen.

Um mit der cDNA als Matrize arbeiten zu können, ist wiederum ein Primer nötig. Da die Synthese der DNA aber immer von 5'- in 3'-Richtung läuft, die Matrize also vom 3'-Ende zum 5'-Ende hin abgelesen werden muss, kann der Poly-A-Schwanz bei diesem Schritt nicht als Primer benutzt werden. Die entstandene cDNA hat zwar einen Bereich, der komplementär zum Poly-A-Bereich ist, dieser jedoch liegt am 5'-Ende. Die Sequenz am 3'-Ende der cDNA ist unbekannt und unterscheidet sich bei verschiedenen Genen. Für die Sequenz des 3'-Endes lässt sich deshalb kein Primer definieren. Jedoch kann mit Hilfe einer terminalen Transferase ein Oligo-C-Schwanz am 3'-Ende angebracht werden. Für diesen kann nun wiederum ein Primer erstellt werden. Die cDNA-Sequenzen werden in das Plasmid pDONR222 kloniert und so gelagert.

1.6.2 DNA-Sequenzierung

Die Sequenzierung ist eine Methode um die Basenabfolge eines DNA-Stranges zu ermitteln. Das Prinzip der Sequenzierung soll hier anhand der Sanger-Coulson-Methode, auch als enzymatische, Didesoxy-Sequenzierung oder Kettenabbruchmethode bekannt, erläutert werden [39].
Der erste Schritt besteht darin einen Primer für eine einzelsträngige DNA, die sequenziert werden soll, zu erstellen. Da die Sequenz der cDNA vor der Sequenzierung nicht bekannt ist, wird der Primer für die Sequenz des Plasmides definiert, in das die cDNA integriert wurde. Eine DNA-Polymerase bildet an der einzelsträngigen Matrize einen zweiten Strang, wobei zu den benötigten Nukleotidtriphosphaten (NTP) ebenfalls eine geringe Menge an Didesoxynukleosidtriphosphaten (ddNTP) gegeben werden. Diese besitzen am 3'-Kohlenstoff keine Hydroxyl-, sondern eine Hydridgruppe. Wird ein ddNTP in die Sequenz eingebaut, ist eine weitere Strangverlängerung nicht mehr möglich und die Reaktion bricht ab. Es sind insgesamt vier Ansätze nötig, wobei in jedem der Ansätze eins der vier ddNTPs (ddATP, ddGTP, ddCTP, ddTTP) verwendet wird.

Die Längen der entstandenen Fragmente werden durch Elektrophorese ermittelt und verglichen. In die kürzeste Sequenz wurde zum frühesten Zeitpunkt der Sequenzierung ein ddNTP eingebaut, sodass der Strang nicht weiter verlängert werden konnte. Je nachdem in welchem der vier Ansätze dieser früheste Strangabbruch zu beobachen ist, lässt sich sagen, dass es sich bei dieser ersten Base um Adenin, Guanin, Thymin oder Cytosin handelt. Die nächstlängere Sequenz gibt Aufschluss über die Identität der folgenden Base. Es wird weiter so verfahren, bis auf diese Weise die Basenabfolge des gesamten Stranges ermittelt ist. Am Anfang und am Ende der Sequenz ist die Qualität der Sequenzierung am geringsten und die Basenabfolge kann in diesen Bereichen oftmals nicht zuverlässig bestimmt werden.

1.6.3 Suche nach adhäsiven Proteinen

Um zu ermitteln, ob einige der cDNA-Sequenzen für Proteine kodieren, die in *V. longisporum* Adhäsion an der Wurzeloberfläche ermöglichen, wurde in nicht-adhäsiven Hefe-Mutanten (*Saccharomyces cerevisiae*) die Wirkung von Genprodukten der cDNA-Bank getestet. Bei der verwendeten Hefe-Mutante handelt es sich um den Stamm BY4741, der eine Punktmutation im Gen von *Flo8* aufweist (Δ*flo8*). Bei Flo8 handelt es sich um einen Transkriptionsfaktor, der Adhäsion in Hefe reguliert [32]. Wenn dieses Protein nicht vorhanden ist, werden keine Adhäsine exprimiert und die Hefe kann nicht mehr an Oberflächen anhaften. Ebenfalls von Bedeutung ist *Flo11*, bei dem es sich um ein Adhäsin der Hefe handelt. Es ermöglicht Flockulation, sowie die Interaktion mit abiotischen Oberflächen, wie Agar oder Plastik [18]. Es wurde eine weitere Mutante des BY4741-Stammes verwendet, in dem außerdem das Gen Flo11 deletiert ist (Δ*flo8/11*). In beiden Fällen (Δflo8 und Δflo8/11) geht die Fähigkeit zur Adhäsion und Flockulation verloren [51].

Es soll geprüft werden, ob eine der nicht-adhäsiven Hefen die Fähigkeit zu Adhäsion oder Flockulation zurückgewinnen kann, wenn in ihr eine cDNA-Sequenz exprimiert wird (Abb. 1.4). Dies würde darauf hindeuten, dass die entsprechende Sequenz auch in *V. longisporum* mit der Adhäsion in Zusammenhang steht. Um diesen Versuch durchzuführen, wurden cDNA-Sequenzen in Δ *flo8*- und in Δ *flo 8/11*-Mutanten eingebracht. Die erstellten 456.584 Klone wurden auf Agar-Platten angezogen und anschließend gewaschen. Beim Waschen blieben nur die Mutanten haften, die cDNA-Sequenz enthalten, welche die Adhäsion in Hefe wieder herstellen. Beim ersten Durchgang (*Screen*) waren 59 Klone vorhanden, die am Agar anhafteten, beim zweiten Screen waren noch 43 Hefe-Mutanten dazu in der Lage. Diese 43 Hefe-Mutanten sind durch das Einbringen von 24 verschiedenen cDNA-Sequenzen in Hefe entstanden. Diese Sequenzen werden im Folgenden als Kandidaten aus dem Δ-flo8, bzw. aus dem Δflo8/11 Screen bezeichnet. Achtzehn der Kandidaten ermöglichen sowohl in Δ*flo8*-, als auch in Δ*flo8/11*-Stämmen eine wieder hergestellte Adhäsion. Die weiteren sechs Kandidaten vermitteln diese Fähigkeit nur in den Δ*flo8*-Stämmen der Hefe. Fünfzehn Kandidaten sind zusätzlich dazu in der Lage die Fähigkeit zur Flockulation in Hefezellen wiederherzustellen.

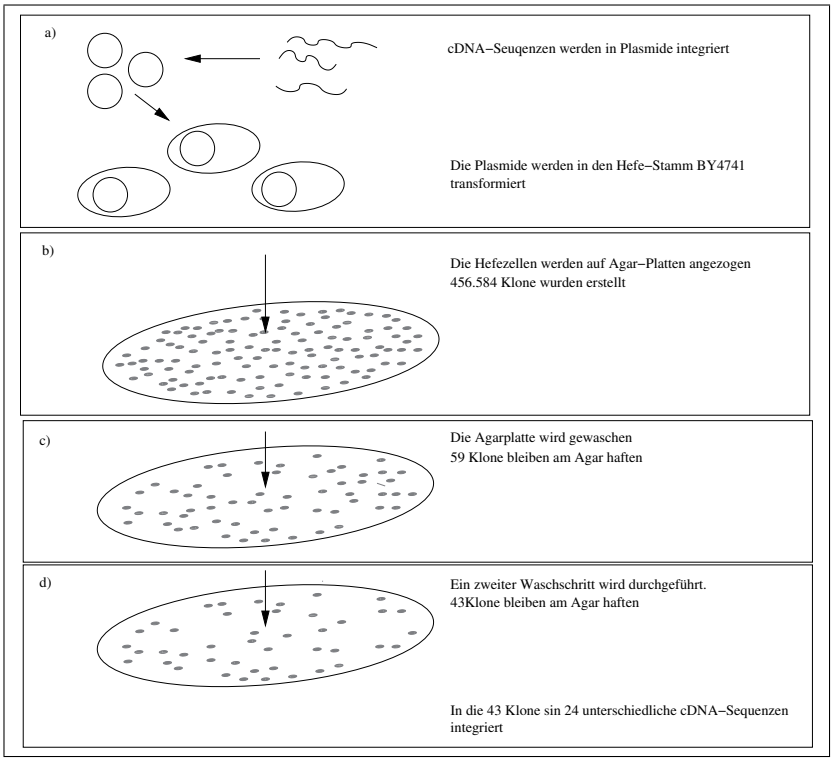

Abbildung 1.4: **Übersicht über die Suche nach cDNA-Sequenzen, die die Adhäsion in Hefe beeinflussen**

a) Die cDNA-Sequenzen werden in den Hefestamm BY4741 integriert, b) Die Hefezellen werden auf Agar angezogen, c-d) Die Agarplatte wird in zwei Schritten gewaschen, adhäsive Klone bleiben haften.

Kapitel 2

Methoden

2.1 Das Zusammenstellen der verwendeten Daten

Die in dieser Arbeit verwendeten cDNA-Sequenzen von *V. longisporum* wurden unter verschiedenen Wachstumsbedingungen gewonnen, um möglichst viele exprimierte Gene zu erhalten. Die Datensätze zweier Sequenzierungen wurden zum Erstellen der Datenbank vereint. Doppelt vorhandene Identifikationsnamen (ID) der Sequenzen wurden durch das Anfügen eines *.1*, bzw. *.2* eindeutig gemacht. Ebenfalls wurden Sequenzen, die kürzer als zwanzig Basenpaare (bp) sind, aus der Datenbank gelöscht.

2.2 Bereinigen der cDNA-Sequenzen

Die cDNA-Sequenzen liegen zur Aufbewahrung und Lagerung in dem Plasmid pDONR222 vor (Abb.2.1, 1). Der Primer, der zum Durchführen der Sequenzierung benötigt wird, setzt nicht direkt an der cDNA-Sequenz an, sondern an diesem Plasmid. Es wird ein Sequenzbereich in der Nähe der *Multiple Cloning Site* (MCS), an der die cDNA ins Plasmid integriert ist, benutzt. Dadurch soll möglichst wenig Information verloren gehen, da der erste sequenzierte Bereich oft eine schlechte Qualität hat, sodass er nicht verwendet werden kann. So wird nicht nur die cDNA selbst, sondern auch der jeweilige Bereich in 5′ und in 3′-Richtung (*up*- und *downstream*) aus dem Plasmid sequenziert. Die vorhandene Plasmidsequenz muss entfernt werden, damit sie die Ergebnisse von weiteren Untersuchungen der cDNA-Sequenzen nicht verfälscht.

Es kommt vermehrt am Anfang und am Ende des Sequenzierens zu Fehlern in der ermittelten Basenabfolge. Deshalb ist es wichtig, dass beim Vergleich der Plasmidsequenz mit der Sequenz in der Datenbank auch die Regionen erkannt werden, die sich an einigen Positionen von der Plas-

midsequenz unterscheiden. Das in dieser Arbeit verwendete Programm Basic Local Alignment and Search Tool (BLAST) [1] eignet sich gut dazu, lokale Ähnlichkeiten zwischen zwei Sequenz zu bestimmen (siehe Abschnitt 2.7.1). Wenn die Qualität der Sequenzierung zu gering ist und vermehrt Sequenzierfehler auftreten, werden jedoch auch durch eine BLAST-Suche nicht mehr alle Bereiche der Sequenz erkannt, die aus dem Plasmid stammen.

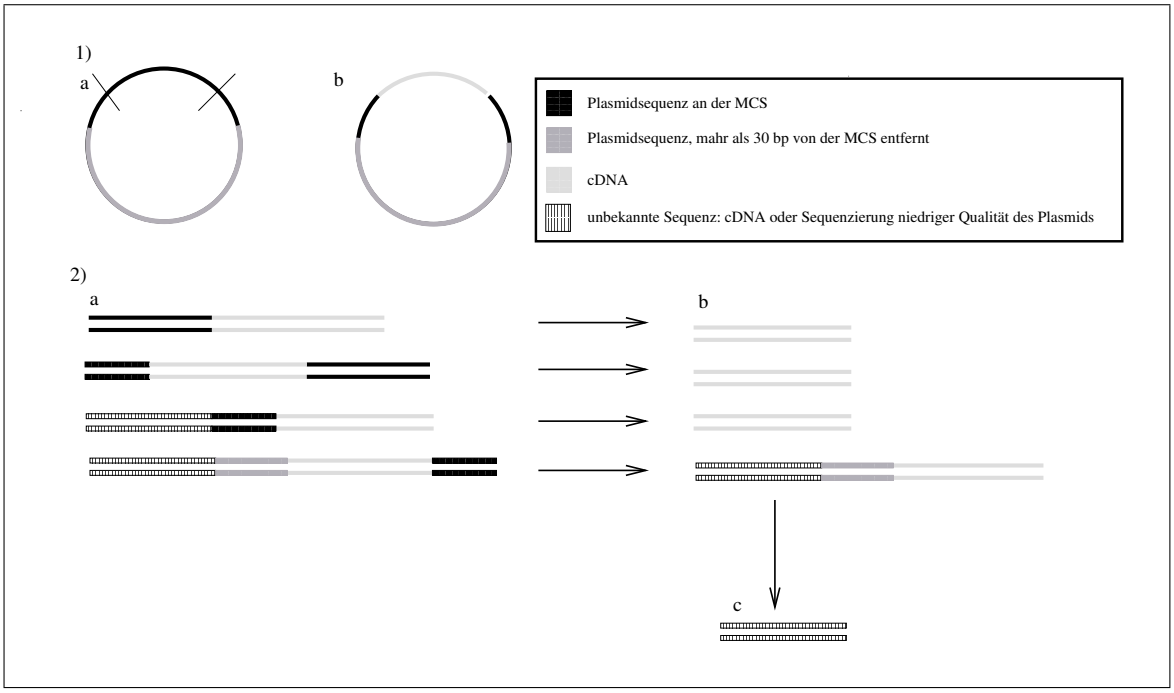

Abbildung 2.1: **Verschiedene Möglichkeiten des Auftretens von Plasmidsequenz in der cDNA:**

Die Abbildung zeigt auf welche Weise Teile der Plasmidsequenz in der cDNA integriert sein können und in welchen Schritten die Plasmidsequenz aus der cDNA-Sequenz entfernt wird. 1a) Das Plasmid enthält eine MCS, an der durch Restriktionsenzyme eine Spaltung erfolgt; b) Zwischen den Schnittstellen der MCS wird ein cDNA-Strang eingebracht; 2: a) Die Plasmidsequenz kommt an verschiedenen Stellen der cDNA-Sequenz vor: Am Anfang, am Ende oder an beiden Enden der cDNA. Ebenfalls kann in der Mitte der cDNA Plasmidsequenz vorhanden sein. Teilweise stammt die Plasmidsequenz aus von der MCS entfernten Plasmidregionen; b) Die Plasmidsequenz aus der Nähe der MCS wird aus der cDNA entfernt. Liegt upstream der an der MCS liegenden Plasmidsequenz ein Sequenzbereich, der nicht identifiziert werden kann, wird dieser ebenfalls entfernt. cDNA-Sequenzen mit Ähnlichkeiten zu Plasmidsequenzen, die nicht an der MCS liegen, werden zunächst nicht verändert; c) Die Plasmidsequenz, die nicht direkt an der MCS liegt, wird entfernt. Die cDNA-Sequenz downstream des Plasmidinserts wird ebenfalls entfernt.

Die meisten BLAST-Treffer werden in der unmittelbaren Nähe zur MCS erwartet, da an dieser Stelle die cDNA-Sequenz in das Plasmid integriert wurde. Diese Integration findet jedoch nicht immer an genau derselben Stelle statt, die Schnittstelle zum Einbringen der Sequenz kann sich verschieben, wenn das verwendete Restriktionsenzym die Plasmid-DNA nicht exakt an der MCS

schneidet. Aus diesem Grund kann nicht davon ausgegangen werden, dass immer die gleiche Plasmidsequenz up- und downstream der cDNA liegt. Es darf also nicht nur nach den Basenabfolgen gesucht werden, die direkt up-, beziehungsweise downstream der theoretischen Schnittstellen des Restriktionsenzyms liegen.

Der erste Schritt bestand darin, die Datenbank der cDNA-Sequenzen mit der Plasmidsequenz als Suchanfrage (*Query*) zu durchsuchen. Die Regionen mit großen Ähnlichkeiten zur Plasmidsequenz wurden aus der cDNA-Datenbank entfernt. Wenn es in der cDNA-Sequenz einen Treffer für einen Bereich upstream der MCS im Plasmid gibt, ist es wahrscheinlich, dass upstream des Treffers in der cDNA ebenfalls Plasmidsequenz liegt. In diesem Fall wird der upstream-Bereich nicht durch die BLAST-Suche erkannt, da zu viele Sequenzierfehler auftreten, muss aber zusätzlich zu dem eigentlichen Treffer aus der cDNA-Sequenz entfernt werden (Abb. 2.1, 2a+b). Umgekehrt wurde bei einem BLAST-Treffer in der cDNA, der eine lokale Ähnlichkeit zu downstream-Bereichen der MCS hat, auch die verbleibende downstream-Sequenz entfernt.

Aus diesem Grund ist es wichtig, dass bekannt ist, ob die Ähnlichkeit zwischen cDNA und Plasmidsequenz up- oder downstream der MCS liegt. Nur so können auch die Bereiche entfernt werden, die bei der BLAST-Suche nicht erkannt wurden.

Zusätzlich treten auch BLAST-Treffer für Regionen im Plasmid auf, die mehr als 30 bp von der MCS entfernt sind. Es ist aber sehr unwahrscheinlich, dass diese Bereiche in der cDNA noch auf eine ungenaue Integration in das Plasmid zurückzuführen sind. In diesem Fall wird die gesamte downstream des Treffers gelegene Sequenz aus der cDNA entfernt. Dies ist nötig, da die Herkunft dieser Sequenz nicht bekannt ist und sie somit nicht als cDNA-Sequenz verwendet werden kann (Abb. 2.1, 2c). Die Sequenz verbleibt nur in der cDNA-Bank, wenn ihre Länge nach den vorgenommenen Kürzungen noch über 21 bp liegt. Andernfalls wird sie entfernt.

2.2.1 Entfernen der MCS-nahen Plasmidsequenz aus der cDNA

Zur Bereinigung der cDNA-Sequenzen wurde im Rahmen dieser Arbeit das Perl-Programm `Remove_plasmid_sequence.pl` geschrieben. Es werden die BLAST-Ergebnisse der Suche von Plasmidsequenzen in der cDNA-Bank verwendet, um die cDNA-Bank so zu kürzen, dass nur noch cDNA-Sequenzen in ihr enthalten sind. Als Argument wird ein Verzeichnis, das die zu bearbeitende cDNA-Sequenzen im FASTA-Format enthält, erwartet. Die Ausgabe erfolgt ebenfalls im FASTA-Format, die bereinigten Sequenzen werden dabei in zwei separaten Verzeichnissen gespeichert. In eines der Verzeichnisse werden alle Sequenzen geschrieben, in denen keine Plasmidsequenz aus Regionen vorhanden ist, die mehr als 30 bp von der MCS entfernt liegt. Die übrigen Sequenzen werden in dem zweiten Verzeichnis gespeichert.

Bei allen Sequenzen, in denen Plasmidsequenz vorhanden ist, wird zunächst der Teil der Plas-

midsequenz entfernt, der in der Nähe der MCS liegt. Befindet sich upstream der Plasmidsequenz noch ein Sequenzteil, der bei der BLAST-Suche nicht als zum Plasmid gehörig erkannt wurde, wird dieser Bereich ebenfalls entfernt. In den Fällen, in denen die Plasmidsequenz downstream der cDNA-Sequenz gefunden wird, wird diese mit der unbekannten Sequenz downstream der Plasmidsequenz entfernt. cDNA-Sequenzen, die keine weitere Plasmidsequenz mehr aufweisen, werden als FASTA-Datei in einem der übergebenen Verzeichnis gespeichert. Wenn die Sequenz aber Bereiche aufweist, die der Plasmidsequenz zugewiesen werden können, die weiter als 30 bp von der MCS entfernt sind, wird diese Sequenz, ebenfalls im FASTA-Format, gespeichert und später von einem zweiten Programm, `Remove_plasmid_insert.pl` bearbeitet.

Bei den weiteren Argumenten handelt es sich um die Pfade der Dateien, die die Ergebnisse der BLAST-Suche der cDNA-Datenbank gegen die up- und downstream der MCS gelegenen Plasmidsequenzen, sowie jeweils die Ergebnisse für die Sequenzen des jeweils komplementären Stranges enthalten. Der vollständige Aufruf des Programmes `Remove_plasmid_sequence.pl` ist im Folgenden angegeben.

Programmaufruf:

```
Remove_plasmid_sequence.pl <Eingabe Fasta-Datei> <Ausgabe-
verzeichnis 1> <Ausgabeverzeichnis 2> <BLAST-Suche upstream,
+ Strang> <BLAST-Suche upstream, - Strang> <BLAST-Suche
downstream, - Strang> <Eingabe BLAST-Suche downstream, -Strang>
```

Die im FASTA-Format ausgegebenen Sequenzen enthalten in der Kopfzeile neben dem Namen der Sequenz, die neue Sequenzlänge, die Länge vor dem Kürzen und ein Doppelkreuz (#), gefolgt von der Anzahl der BLAST-Treffer für Plasmidsequenzen, die weiter als 30 bp von der MCS entfernt sind. Diese Werte sind jeweils durch drei Leerzeichen voneinander getrennt. In der Beispielausgabe (s.u.) ist eine Sequenz angegeben, deren ursprüngliche Länge 375 bp betrug. Nachdem die Plasmidsequenz entfernt wurde, beträgt sie 356 bp. Durch die Null hinter dem Doppelkreuz wird angegeben, dass keine Plasmidsequenz vorhanden war, die mehr als 30 bp von der MCS entfernt lag. Die Kürzungen wurden also vorgenommen, um Sequenzabschnitte zu entfernen, bei denen es sich wahrscheinlich um up- oder downstream-Bereiche der MCS handelt.

Beispielausgabe einer bereinigten Sequenz ohne MCS-ferne Plasmidsequenz:

```
>VL1080 356 375 #0
TGCCGTCAAGATGGTCGCCGCCCAGAAGCATGTTCCCATCGTGAAGAAGC
GCACCAAGCGCTTCGAGCGCCACCAGAGCGACCGCTTCATGCGTGTCGAC
CCCTCTTGGCGCAAGCCCAAGGGTATCGACAACCGCGTTCGCCGTCGGTT
CCGTGGTACCGCGCCCATGCCCTCGATCGGCTATGGCTCCAACAAGAAGA
CCAAGTACATGATGCCCTCCGGCCACAAGGCATTCCTCGTCAACAACGTT
TCCGACGTTGAGCTCCTCCTCATGCACAACCGCACCTTCGCCGCTGAGAT
CGCGCACGGCGTCTCCTCCAGGAAGCGCATCGACATTATCTCCCGCGCCA
AGCAAC
```

2.2.2 Entfernen von MCS-ferner Plasmidsequenz

Die cDNA-Sequenzen, die Plasmidsequenz enthalten, welche nicht in der Nähe der MCS liegt, werden durch das Programm `Remove_plasmid_insert.pl` weiter bearbeitet. Dabei wird nicht nur der Bereich entfernt, der homolog zur Plasmidsequenz ist, sondern auch die downstream davon gelegene Sequenz. Als Parameter werden drei Verzeichnisse übergeben: Das Verzeichnis, in dem die cDNA-Sequenzen im FASTA-Format gespeichert sind, das Verzeichnis, in das die Ausgabe der veränderten cDNA-Sequenzen erfolgt und das Verzeichnis, dass die BLAST-Suchen der Plasmidsequenzen upstream und downstream der MCS auf beiden DNA-Strängen enthält.

Programmaufruf:

```
Remove_plasmid_insert.pl <Verzeichnis der Fasta-Dateien> <Ausga-
beverzeichnis> <Verzeichnis der BLAST-Suchen up- und downstream
der MCS>
```

Für jede Sequenz wird der Startpunkt der Sequenzähnlichkeit zum Plasmid ermittelt. Dieser Sequenzteil, sowie die davon downstream gelegene Sequenz werden entfernt. Die Ausgabe ähnelt der von `Remove_plasmid_sequence.pl` im FASTA-Format. Jedoch ist hier nach dem Doppelkreuz eine Zahl aufgeführt, die größer als Null ist, da Sequenzabschnitte gefunden wurden, die im Plasmid mehr als 30 bp von der MCS entfernt sind.

Beispielausgabe einer von MCS-fernen Plasmidsequenzen bereinigten cDNA-Sequenz:

```
>VL0795  114   843   #1
CATGAGATCAACCTCCGCATCAGCAGCCAACTTGTCCGCGCAGCCCTGCC
CAAGATGCATCTCGTGCCGCAAGAGCTTGATAAGCTCGTCATTTCTCAGC
TGGGGTTCCTGGCG
```

2.3 Untersuchungen der Aminosäuresequenzen

Übersetzt man die Nukleotidsequenzen aus der cDNA-Datenbank in Aminosäuresequenzen, lassen sich aus dieser wertvolle Informationen gewinnen. Die spezifische Abfolge von AS legt die Eigenschaften eines Proteins fest. In dieser Arbeit wird in der AS-Sequenz nach Signalen gesucht, die für das Anbringen eines GPI-Ankers an ein Vorläuferprotein nötig sind. Auch wird nach der für Adhäsine typischen Ser/Thr-reichen Region gesucht. Um diese Untersuchungen für die cDNA-Sequenzen durchführen zu können, muss die DNA-Sequenz zunächst in eine AS-Sequenz übersetzt werden. Da nicht bekannt ist an welchem Strang und an welcher Position die Translation dieses Genes startet, werden alle sechs Leserahmen (LR) nacheinander übersetzt.

Die Übersetzung beginnt am ersten Nukleotid, sodass die drei ersten Nukleotide als erstes Kodon benutzt werden (1.LR). Dann werden das zweite, dritte und vierte Nukleotid verwendet, das erste wird ignoriert (2.LR). Für den dritten LR werden das dritte, vierte und fünfte Nukleotid der Sequenz als Kodon für die erste AS verwendet. Analog werden die drei LR im reversen Komplement übersetzt. In einigen Sequenzen kommen Nukleotide vor, deren Identität durch die Sequenzierung nicht eindeutig festgestellt werden konnte. Diese werden mit *N* bezeichnet. Kodons, die nicht-identifizierte Nukleotide enthalten, kann keine AS zugeordnet werden. Aus diesem Grund werden Sequenzen, die unbestimmte Nukleotide enthalten, in Bereiche unterteilt, in denen alle Nukleotide bestimmt werden konnten. Für diese Teilsequenzen werden dann mögliche AS-Sequenzen aller sechs LR bestimmt.

2.4 Ermitteln der kodierenden Regionen

Die cDNA-Sequenzen, die auch als Expressed Sequence Tags (*ESTs*) bezeichnet werden, können neben der für das Protein kodierenden Region auch Bereiche aufweisen, die nicht-kodierend sind. Diese werden nicht translatiert und enthalten somit keine Informationen zum Aufbau des Proteins. Bei Vorhersagen für die AS-Sequenz einer cDNA dürfen daher nur die kodierenden Bereiche

berücksichtigt werden. Um diese kodierenden Bereiche zu finden, kann das Programm ESTScan [22, 33] verwendet werden. Dieses Programm sucht nicht nur nach den kodierenden Sequenzen in der cDNA, es werden ebenfalls Insertionen und Deletionen, die auf Fehler beim Sequenzieren der ESTs zurückzuführen sind, erkannt und berichtigt.

2.4.1 ESTScan

Da ESTs oft mit Sequenzierfehlern behaftet sind und sich dadurch der Leserahmen der kodierenden Region verändern kann, müssen diese Fehler in der Sequenz bei der Vorhersage einer kodierenden Region berücksichtigt werden. ESTScan nutzt *Hidden Markov Modelle* (HMM) um kodierende Regionen in ESTs zu finden. Das Programm kann mit Sequenzierfehlern, die zu einem veränderten Leserahmen führen oder die zusätzliche Stopp-Kodons erzeugen und mit nicht identifizierten Nukleotiden arbeiten.

Das verwendete HMM wird mit RNA-Sequenzen trainiert, wobei das Training für Sequenzen, je nach GC-Gehalt, seperat erfolgt. Es werden vier Gruppen gebildet, die die Bereiche von weniger als 43 % GC-Gehalt, von 43 bis maximal 50 %, von 51% bis maximal 57 % und über 57 % abdecken. Bei der Eingabe einer Sequenz wird dann zuerst überprüft welcher GC-Gehalt vorliegt, damit die entsprechenden Einstellungen für diese Gruppe genutzt weden. Dieser Schritt ist wichtig, da die Häufigkeit, mit der einzelne Kodons von einem Organismus verwendet werden, stark von dem GC-Gehalt der jeweiligen Sequenz abhängen können. Die Sequenzen, mit denen das Training erfolgen soll, können individuell zusammengestellt werden.

Es wird ein fünf Buchstaben Alphabet benutzt, zu dem neben den Basen A, C, T und G auch N für nicht identifizierte Basen gehört. Alle anderen, mehrdeutigen Buchstaben der Nomenklatur von *International Union of Pure and Applied Chemistry* (IUPAC) werden in N umgewandelt. Diesen nicht-identifizierten Basen wird ein Durchschnittswert zugewiesen, der zwischen dem der bekannten vier liegt.

Das Programm kann fünf verschiedene Bereiche in einer Sequenz erkennen. Zu diesen Bereichen zählen die am 5'- und am 3'-Ende gelegenen nicht-kodierenden Regionen, die Profile der Start- und Stopp-Sites der Translation, sowie die kodierende Region. Da nicht in allen EST-Sequenzen alle diese Regionen vorkommen müssen, erkennt das Programm auch kodierende Regionen ohne Start- oder Stopp-Sites. Es ist ebenfalls nicht nötig, dass nicht-kodierende Sequenzen vorhanden sind, genauso wie nicht in allen Fällen eine kodierende Region vorliegen muss. ESTScan benutzt dabei das gleiche Modell wie GENESCAN [5] um kodierende Regionen einer Sequenz zu finden.

ESTScan liest die eingegebene EST-Sequenz und übersetzt dabei die mutmaßlich kodierende Region in eine AS-Sequenz. In der übersetzten Region werden nacheinander drei Zustände, F0,

F1 und F2, wiederholt. Sie entsprechen der Position des aktuellen Nukleotids im Kodon. Damit jedes eingelesene Nukleotid einmal jede Position durchläuft, werden zu Beginn des Einlesens jeweils einmal ein X, zwei und kein X der Nukleotidsequenz vorrangestellt. Dadurch wird jedes Nukleotid einmal als F0, einmal als F1 und einmal als F2 eingelesen.

Die Reihenfolge der drei Zustände kann auf zwei verschiedene Weisen durchbrochen werden. Dies geschieht, wenn entweder eine Insertion oder eine Deletion vorliegt. Handelt es sich bei dem eingelesenen Nukleotid um eines, dass als Insertion erkannt wurde, so wird diese Position zwar eingelesen, aber nicht als Bestandteil der möglichen kodierenden Region ausgegeben. Wird an einer Position eine Deletion vorhergesagt, wird an der entsprechenden Stelle ein zusätzliches »N« eingefügt. Auf diese Weise kann einer Verschiebung des Leserahmens entgegengewirkt werden.

Mit Hilfe des Viterbi-Algorithmus [52] wird der optimale Pfad durch die Zustände des HMM berechnet und die entsprechenden AS-Sequenzen werden ausgegeben. Bei der Ausgabe der vorhergesagten AS-Sequenz kann ein X am Anfang oder am Ende der Sequenz auftauchen, wenn am Anfang oder am Ende der Sequenz zusätzliche Positionen eingeschoben werden.

Bei der Übersetzung werden nur Insertionen und Deletionen berücksichtigt. Substitutionen werden ignoriert. Während die beiden Ersten den Leserahmen und somit die gesamte folgende AS-Sequenz ändern, wird durch eine Substitution nur ein Nukleotid durch ein anderes ausgetauscht. Die Folge wäre bei einer Übersetzung in AS-Sequenz eine falsche AS oder aber gar keine Änderung in der AS-Sequenz. Ein Sonderfall besteht jedoch in Substitutionen, durch die ein Stopp-Kodon innerhalb der kodierenden Sequenz auftaucht. Beim Akzeptieren eines falschen Stopp-Kodons als Ende der kodierenden Region ginge das Ende der Proteinsequenz verloren. Um dies zu verhindern, wird beim Auftreten eines Stopp-Kodons nicht automatisch die vorhergesagte kodierende Region beendet. So wird das Vorkommen von Stopp-Kodons im Leserahmen toleriert, jedoch werden alle anderen Kodons ihm vorgezogen.

In einer kodierenden Region kann eine lokale Anhäufung von Sequenzierfehlern vorkommen, die dazu führt, dass nicht die gesammte Region als kodierend erkannt wird. Aus diesem Grund sucht ESTScan nicht nur nach der ersten kodierenden Region. Die Suche wird an der Position, die sich der gefundenen Region anschließt, wieder neu gestartet. Die Ausgabe erfolgt nur, wenn die vorhergesagte Sequenz eine Länge von mindestens 50 Nukleotiden erreicht, andernfalls wird das Ergebnis herrausgefiltert und erscheint nicht in der Ausgabe.

In der Veröffentlichung [22]wird die Spezifität anhand der Vorhersage für 12.284 untranslatierte ESTs angegeben. Sie beträgt 69,2 %. Es wird eine Sensitivität von 94,5 % der erreicht. Zusätzlich konnten 64,2 % der Start- und 55,1 % der Stopp-Kodons exakt erkannt werden.

2.5 Suche nach Sequenzen mit Serin- und Threonin- reichen Regionen

Wie in Kapitel 1.4 beschrieben, weisen Adhäsine eine ca. 300 AS lange Region auf, die reich an Serin (Ser) und Threonin (Thr) ist. Der Anteil von Ser und Thr liegt hier bei 35 -55 % [31]. Dieser Bereich formt einen so genannten *Stalk*, einen Stiel, der es der Bindedomäne des Proteins ermöglicht aus der Zellwand herauszuragen und so Kontakt zu Oberflächen in der Umgebung der Zelle herzustellen. Da es beim Sequenzieren der zugehörigen cDNA-Sequenz jedoch häufig zu einem frühen Abfallen der Polymerase vom Matrizenstrang kommt, wird für Adhäsine in der cDNA-Datenbank keine vollständige Ser/Thr-reiche Region erwartet. Um Bereiche mit einem erhöhten Ser/Thr-Gehalt zu finden, werden alle sechs LR jeder cDNA-Sequenz untersucht.

Dem zu diesem Zweck erstellten Programm `find_stalk.pl` wird der Pfad zu einem Verzeichnis, das die zu untersuchenden AS-Sequenzen im FASTA-Format enthält übergeben. Außerdem müssen eine Datei, in die die gefundenen Ergebnisse ausgegeben werden, und die minimale Länge (L) der Ser/Thr-reiche Region übergeben werden.

Programmaufruf:

```
find_stalk.pl <Eingabeordner> <Ausgabeordner> <L Ser/Thr-Anteil>
```

Um die Lage dieser Ser/Thr-reichen Regionen zu ermitteln, werden immer zwanzig AS gleichzeitig betrachtet. Dieses zwanzig-AS-Fenster wird vom Anfang der Sequenz in Schritten von jeweils einer AS weiterbewegt. Bei jedem Schritt wird überprüft, ob die gesetzte Bedingung, ein Ser/Thr-Anteil von mindestens 35 %, erfüllt ist. Sobald dies der Fall ist, wird dieser Sequenzbereich markiert und das Fenster um ganze zwanzig AS weiter geschoben. Dadurch wird als nächstes das direkt angrenzende Fenster betrachtet. Solange die Bedingung erfüllt ist, wird der Bereich immer wieder um weitere zwanzig AS verlängert.

Ist die Bedingung nicht mehr erfüllt, wird das letzte Fenster schrittweise um jeweils eine AS zurückbewegt, bis wieder ein Bereich mit einem 35 %-igen Ser/Thr-Gehalt in dem Fenster liegt. Dabei kommt es zu Überlappungen mit der zuvor gefundenen Region. Für den ermittelten Gesamtbereich wird der Anteil von Serin und Threonin bestimmt. Dadurch, dass das letzte gefundene Fenster zurück verschoben wird, bis es die Bedingungen wieder erfüllt, kann der Ser/Thr-Anteil der gesamten Region unter 35 % fallen. In diesem Fall wird die Region abwechselnd an beiden Enden um je eine AS gekürzt, bis die Bedingung wieder erfüllt ist. Sollte es jedoch nötig sein, dass mehr AS entfernt werden müssen, als durch das letzte Fenster dazugewonnen würden,

die Gesamtlänge durch das Anfügen des letzten Fensters also abnähme, so wird das letzte Fenster ignoriert.

Das Fenster, mit dem die Suche nach weiteren Sequenzbereichen fortgesetzt wird, liegt direkt hinter dem zuletzt gefundenen. Wurde das letzte Fenster der Region jedoch nicht mehr verwendet, weil dadurch die Gesamtregion zu stark gekürzt worden wäre, wird eine AS übersprungen. So wird die Suche nach einer neuen Region begonnen und es wird verhindert, dass immer das gleiche Fenster gefunden und wieder gekürzt wird.

Nachdem die Untersuchung für eine Sequenz abgeschlossen ist, werden alle Sequenzbereiche, die nur zehn AS voneinander entfernt liegen, miteinander vereinigt. Dadurch ist es möglich, dass Regionen gefunden werden, deren Gesamtanteil an Serin und Threonin unter 35 % liegt.

Ausgegeben werden, jeweils durch Tabulatorzeichen voneinander getrennt, Name und Länge der Sequenz, sowie Informationen über jeden der ermittelten Bereiche mit einem erhöhten Ser/Thr-Gehalt. Für jeden einzelnen gefundenen Bereich werden Start- und Endpunkt, gefolgt von der in Klammern gesetzten Länge des Bereiches und, nach einem Doppelpunkt, der prozentuale Ser/Thr-Gehalt angegeben. Die Ausgabe der Ergebnisse erfolgt nur, wenn mindestens von einer der ermittelten Regionen die zuvor festgelegte, minimale Länge erreicht wurde.

Beispielausgabe der Ser/Thr-Suche:

		Region 1	Region 2
Name der Sequenz	Sequenzlänge	Start-Ende(Länge): Ser/Thr-Anteil	Start-Ende(Länge): Ser/Thr-Anteil
A57E02_rev_LR3.fna	243	54-73(19): 35.00	124-143(19): 35.00
A54D06_LR3.fna	218	0-36(36): 35.14	51-198(147): 41.89
A54D06_rev_LR3.fna	218	94-175(81): 35.37	
A54D04_LR1.fna	220	9-33(24): 36.00	77-215(138): 39.57
A54D03_LR1.fna	272	51-75(24): 36.00	100-205(105): 42.45

2.6 Suche nach Proteinen mit GPI-Ankersignal

Zur Bindung von Adhäsine an die Zellwand, muss an ihrem C-terminalen Ende ein GPI-Anker angebracht werden. Dazu muss eine Signalsequenz vorhanden sein, die bei der Prozessierung des Proteins durch einen GPI-Anker ersetzt wird, der dann kovalent an die Zellwand bindet. Das Auffinden eines solchen Ankersignals an der Proteinsequenz einer cDNA ist ein Hinweis darauf, dass es sich bei diesem Protein um ein Adhäsin handeln kann. Jedoch muss auch eine N-terminale Exportsequenz vorhanden sein, da nur

Proteine mit Exportsequenz in das ER transportiert werden, wo eine Ankerstruktur angefügt werden kann. Proteine, die nur eine Ankersignalsequenz aufweisen, nicht jedoch eine Exportsequenz, sind in vivo wahrscheinlich nicht verankert.

Um nach der Signalsequenz für das Anfügen eines GPI-Ankers zu suchen, werden verschiedene im Internet verfügbare Programme benutzt.

2.6.1 Vorhersage GPI-verankerter Proteine mit Big-Π

Big-Π ist ein Programm zur Vorhersage von GPI-Ankerstellen, das speziell für Pilze entwickelt wurde [14]. Für eine untersuchte Sequenz wird ein *Score*-Wert berechnet, der anzeigt, wie stark die vom Programm erfassten Eigenschaften der Sequenz einem Ankersignal entsprechen. Dazu werden sowohl Sequenzähnlichkeiten zu bekannten Vorläufern von GPI-verankerten Proteinen, als auch bestimmte Eigenschaften der Signalsequenz berücksichtigt (siehe Abschnitt 1.4).

Die Trainingsmenge, mit der das Programm trainiert wurde, enthält ausschließlich Sequenzen, die von Pilzen stammen und einen GPI-Anker aufweisen. Da es aber nicht viele experimentell bestätigte Sequenzen mit GPI-Anker gibt, werden außerdem 25 Sequenzen verwendet, die in der Literatur nur aufgrund von theoretischen Annahmen zu den GPI-Signalsequenzen gezählt werden, da sie sowohl die Signalsequenz, als auch die Exportsequenz aufweisen. Weitere 171 Sequenzen aus den Proteomen von fünf Pilzen werden dem Datensatz hinzugefügt, da sie sowohl einen hohen Score bei der Vorhersage von Signalsequenzen durch schon verfügbare Programme, als auch eine ebenfalls vorhergesagte N-terminale Exportsequenz aufweisen. Ein negatives Trainingsset wird nicht verwendet. Aus den C-terminalen Enden dieser Sequenzen wird ein *Alignment* ohne Lücken (*gaps*) erstellt. In diesem Alignment wird sowohl nach Zusammenhängen zwischen AS-Häufigkeiten und Position im Alignment, als auch nach Korrelationen an verschiedenen Positionen auftauchender AS gesucht.

Der Score setzt sich aus zwei Größen zusammen:

$$S = S_{profile} + S_{ppt}$$

mit:

S: Gesamtscore

$S_{profile}$: Bester Score für das Alignment mit bekannten GPI-Signalsequenzen

S_{ppt}: Physical property term

$S_{profile}$ wird für jede der 55 C-terminalen AS berechnet. Dabei wird jeweils angenommen, dass die ω-site an der aktuellen AS-Position liegt. Das Alignment der Trainings-Sequenzen liefert die Kriterien für hohe, bzw. niedrige Werte. Der $S_{profile}$- Wert für die gesamte Sequenz entspricht dem Score der am höchsten bewerteten AS. S_{ppt} ist der Score für den *physical property term*, der bestimmte Eigenschaften der Sequenz bewertet. Dabei wird davon ausgegangen, dass die AS, die den höchsten $S_{profile}$-Wert hat, die tatsächliche ω-site ist. Unter dieser Annahme wird die Sequenz, welche die ω-site umgibt, untersucht. Zu den beurteilten Eigenschaften gehören beispielsweise das Seitenkettenvolumen in verschiedenen Regionen der Signalsequenz und der Anteil geladener oder aromatischer Seitenketten. Insgesamt werden etwa 40 Größen überprüft. Je nachdem, ob beobachtete Eigenschaften denen von Ankersignalen entsprechen oder nicht, wird der Score erhöht.

Die Sensitivität, die bei der Vorhersage mit Big-Π erreicht wird, wird in der Veröffentlichung mit 90 % angegeben. Sie wurde neben einer Überprüfung der Selbstkonsistenz auch mittels einer Kreuzvalidierung bestimmt. Die Selektivität, die Rate der falsch positiven Ergebnissen, wird mit 0,1 % angegeben. Die Selektivität lässt sich nicht leicht über einen direkten Weg bestimmen, da experimentelle Nachweise über Funktionen, die eine Sequenz nicht hat, nur selten vorhanden sind. Allein über die Lokalisation eines Proteins in der Zelle lässt sich nicht ausschließen, dass es eine Signalsequenz besitzt, die in vivo zu einem Anbringen eines GPI-Ankers führen würde. Das Vorhandensein einer N-terminalen Exportsequenz, die in vivo nötig ist, soll durch dieses Programm nicht berücksichtigt werden, sondern nur die Eigenschaft des C-Terminus als Signalsequenz. Um dennoch die Rate von falsch positiven Ergebnissen angeben zu können, wurden große

Datenbanken, wie die aus Pilzen stammenden Sequenzen in SwissProt, für eine Vorhersage von GPI-Ankerstellen genutzt. Die Proteine mit einem positiven Score wurden auf ihre zelluläre Lokalisation und ihre Funktion hin untersucht. Von den 75 gefundenen Treffern wiedersprach keiner einer GPI-Verankerung des Proteins. Durch die hohe Genauigkeit der Vorhersage eignet sich dieses Programm gut zum Auffinden von Signalsequenzen für das Anbringen von GPI-Ankern in großen Datensätzen.

2.6.2 Vorhersage GPI-verankerter Proteine mit DGPI

Das Programm DGPI geht ähnlich vor, wie es schon für Big-II beschrieben wurde [29] (Veröffentlichung des Programmes nicht mehr verfügbar). Die Unterscheidung zwischen Sequenzen mit Ankersignalen und Sequenzen, die keine Ankersignale enthalten, wird mit Hilfe der durch die Literatur bekannten Eigenschaften der Signalsequenzen vorgenommen. Dabei werden die Proteinsequenzen nach drei voneinander getrennte Kriterien bewertet: Als erstes wird geprüft, ob eine N-terminale Exportsequenz vorhanden ist, die als Exportsignal ins ER die Vorraussetzung für eine GPI-Verankerung in vivo darstellt. Weiterhin werden hydrophile und hydrophobe AS-Reste am C-Terminus der Sequenz mit verschiedenen Scores bewertet, je nachdem, ob ihre Verteilung der von GPI-Ankersignalen entspricht oder nicht. Die dritte gesuchte Eigenschaft ist das Vorhandensein einer ω-site. Die Vorhersage, dass ein GPI-Anker angefügt werden kann, wird nur getroffen, wenn sowohl eine N-terminale Exportsequenz als auch ein C-terminales Signal vorliegen. Eine mögliche ω-site muss nicht vorhanden sein.

2.6.3 Vorhersage GPI-verankerter Proteine mit GPI-SOM

GPI-SOM ist ein Programm, das entwickelt wurde, um, mit Hilfe von Kohonen-Netzen, GPI-verankerte Proteine zu identifizieren [17]. Bei diesem Ansatz werden keine Alignments der Signalsequenzen benötigt, sondern ein neurales Netzwerk wird mit Positiv- und Negativbeispielen für GPI-Signalsequenzen trainiert. Das positive Trainingsset enthält Sequenzen von Proteinen, an deren C-terminalen Enden GPI-Anker angebracht werden. Verwendet wurden 110 Proteinsequenzen verschiedener Eukaryoten aus GenBank, bei denen die GPI-Verankerung experimentell nachgewiesen wurde, sowie 248 GPI-Proteinen aus *Arabidopsis thaliana*. Das negative Trainingsset umfasste 256 bekannte cytolytische

und 128 transmembrane Proteine eukaryotischer Organismen.

Um die Qualität der Vorhersagen zu überprüfen, wurde ein Testset zusammengestellt. Dieses enthielt zum Einen GPI-verankerte Proteine aus neuen Veröffentlichungen und zum Anderen das Testset, das auch schon für Big-ΠI benutzt wurde. Proteine ohne GPI-Anker werden für das Testset durch textbasierte Suchen in GenBank ermittelt. Darunter sind auch Sequenzen, die sowohl eine C-terminale transmembrane Domäne, als auch eine ER-Transportsequenz besitzen.

In einem Pilotexperiment wurde ein Perzeptron, ein einfaches neurales Netzwerk, mit verschiedenen Eingabearten der Trainingssequenzen trainiert. Dadurch sollte ermittelt werden, in welcher Art Informationen aus den Sequenzdaten verarbeitet werden sollten, um optimale Ergebnisse zu erhalten. Die besten Ergebnisse bei einer Vorhersage mit dem trainierten Netzwerk wurden mit einer Repräsentation der Sequenz durch Zentriolwerte (Z) erreicht. Diese repräsentieren die Durchschnittspositionen der AS A in einer Sequenz, gewichtet durch die jeweilige Nähe der AS A zum C-Terminus. Sie werden bei jeder Trainings- und Eingabesequenz für jede AS bestimmt und als Eingebewerte verwendet. Dabei werden alle Positionen p, an denen eine bestimmte AS auftaucht, mit einem Vielfachen von zwei multipliziert. Die Positionen werden zum C-Terminus hin gezählt. Es werden somit hohe Zentriolwerte für AS erhalten, die häufig vorkommen, aber auch für die, die nahe am C-Terminus liegen. AS, die nicht in der Sequenz erscheinen, haben einen Zentriolwert von Null.

Die Berechnung des Scores erfolgt nach:

$$Z(A) = 2^{-n} \sum_{i=1}^{n} 2^{i-1} p_{Ai}$$

Mit:

$Z(A)$: Zentriolwert der AS A

n: Anzahl der AS A

i: Index $\in$ {1,2,..,n}

p_{Ai}: Position, der i-ten AS A

Zusätzlich zum Zentriolwert werden mutmaßliche ω-sites mit einer Scoring-Matrix bewertet, die sich nach dem Aufbau der ω-site bekannter GPI-Proteine richtet. Die Positionen ω, $\omega + 1$ und $\omega + 2$ werden dabei berücksichtigt. Ein weiterer berücksichtigter Wert ist ein Score für die Hydrophobie der in der Sequenz auftauchenden AS.

Die optimale Länge zum Training und zur Vorhersage wurde ermittelt, indem ein Perzeptron wiederholt mit den gleichen Sequenzen, gekürzt auf unterschiedliche Längen, trainiert und anschließend getestet wurde. Die kürzeste Sequenz umfasst nur die fünf letzten AS vor dem C-Terminus. Die Häufigkeit, mit der falsch positive und falsch negative Treffer ermittelt werden, ist für solche kurzen Sequenzen sehr hoch. Die Genauigkeit der Vorhersage nimmt jedoch auch ab, wenn die Sequenz länger als 32 AS ist. Bei einer Länge von 29 bis 32 AS treten die wenigsten falschen Vorhersagen auf.

Die 32 C-terminalen AS wurden des Weiteren mit *in silico* Mutageneseexperimenten untersucht. Nacheinander wurden die zum Training benutzten Sequenzen an bestimmten Positionen maskiert. Dabei wurden für alle Sequenzen des Trainings- und Testsets die gleichen AS-Positionen durch ein X markiert, das für beliebige AS steht. Die maskierten Bereiche hatten in verschiedenen Durchläufen eine Länge von einer bis vier Positionen. So konnte ermittelt werden, dass vor allem die AS nah am C-Terminus wichtige Reste für die Erkennung von GPI-Signalsequenzen sind. Ein Eingabevektor, der nur Informationen zu den 22 wichtigsten der insgesamt verwendeten 32 AS-Positionen enthält, liefert beim Training mit einem Perzepton bessere Ergebnisse, als wenn alle 32 Residuen einbezogen werden.

Der Eingabevektor enthält Informationen zu den Zentriolwerten, der Hydrophobie und der ω-site. Diese werden durch 44 Komponenten an das Programm übermittelt: zwanzig Zentriolwerte, für jede AS einen; Hydrophobie der AS an der 22. betrachteten Position und je einen Werte für die Qualität und die Position der besten ω-site.

Das Inputlayer des Kohonen-Netzwerkes besteht aus 44 Neuronen, die die Informationen der Eingabe aufnehmen. Das Training erfolgt in 5000 Runden. Dabei wird in jeder Runde die Gewichtungen der Neuronen aktualisiert, die dem Eingabesignal am nächsten sind. Die Nachbarn dieser Neuronen werden ebenfalls höher gewichtet. Dies geschieht aber mit steigender Entfernung zum Eingabesignal in immer geringerem Ausmaß. Leere Einheiten, die nach dem vollständigen Ablauf des Trainings noch nicht getroffen wurden,

erhalten einen Wert, der sich nach dem ihrer benachbarten Neuronen richtet.

In der Veröffentlichung werden Sensitivität und Spezifität der Vorhersage mit Sequenzen der Positiv- und Negativtestsets ermittelt und mit Big-Π und DGPI verglichen. Für die Sequenzen, die an GPI-SOM und Big-Π übergeben werden, wird zuvor eine Suche nach dem N-terminalen Exportsignal durchgeführt. Es sollen nur Sequenzen einbezogen werden, bei denen eine GPI-Verankerung in vivo möglich wäre. Da DGPI selbst die Suche nach einer solchen Exportsequenz in die Vorhersage der GPI-Verankerung mit einbezieht, müssen hier die Eingabesequenzen nicht selektiert werden. Für GPI-SOM wird eine Sensitivität von 96,4% auf dem Testset ermittelt. Diese übersteigt im direkten Vergleich die von Big-Π um 28,9 Prozentpunkte. Die Selektivität erreicht nicht die guten Ergebnisse, die von Big-Π erzielt werden, und hängt von der Art der Sequenzen im Negativtestset ab. Vor allem bei den Suchen in Sequenzen von Proteinen mit C-terminaler transmembraner Domäne werden viele falsch-positive Vorhersagen gemacht. GPI-SOM kann ergänzend zu den bereits vorhandenen Programmen zur Vorhersage von GPI-Ankersignalen verwendet werden. Da mit der hohen Sensitivität auch eine hohe Rate an falsch positiven Ergebnissen verbunden ist, ist eine Bestätigung durch weitere Programme sinnvoll, sodass die Verwendung falsch positiver Ergebnisse verhindert wird.

2.7 Suche nach Sequenzhomologien mit BLAST

Mit dem Programm BLAST [1] werden verschiedene Sequenz-Datenbanken nach Ähnlichkeiten zur cDNA-Datenbank durchsucht. Für die Suche werden die Proteindatenbanken nr (*non redundant*) des National Center for Biotechnology Information (*NCBI*) und die SwissProt-Datenbank [4] verwendet. Diese werden mit dem Programm blastx durchsucht. Die Nukleotiddatenbanken Cogeme [41] und die Datenbank der offenen Leserahmen (*orf*) von Hefe (*Saccharomyces Genome Database*) werden sowohl mit dem Programm blastn, als auch mit tblastx durchsucht. Das Programm blastn vergleicht die Basenabfolge der Nukleotidsequenzen, tblastx ermöglicht die Suche nach Ähnlichkeiten zwischen den AS-Sequenzen der sechs Leserahmen von Nukleotiddatenbanken und cDNA-Sequenzen.

2.7.1 BLAST - Basic Local Alignment Search Tool

Das Programm BLAST dient dem Vergleich von Eingabe-Sequenzen (*Query-Sequenz*) mit den Sequenzen einer Datenbank und sucht nach Bereichen mit starker Ähnlichkeit zueinander. Dabei müssen sich die Sequenzen nicht über ihre gesamte Länge gleichen, auch lokale Ähnlichkeiten in Sequenzen werden miteinander aligniert. So lassen sich einzelne konservierte Sequenzabschnitte einander zuordnen und können über mögliche Funktionen einer unbekannten Sequenz Aufschluss geben.

BLAST ist durch ein heuristisches Verfahren dazu in der Lage, Vergleiche zwischen Sequenzen sehr schnell durchzuführen. Dies wird durch das Zerlegen der Eingabesequenz in kurze Teilsequenzen erreicht. Die Datenbank wird nicht nach Ähnlichkeiten zu der gesamten Eingabesequenz, sondern direkt nach den Teilsequenzen durchsucht. Für jede dieser Teilsequenzen wird zusätzlich eine Liste mit Segmenten erstellt, die bei einem Alignment mit der Teilsequenz einen festgelegten, minimalen Score überschreiten. Diese Segmente werden ebenfalls für die Datenbanksuche genutzt. Durch die vor der Suche erfolgte Unterteilung der Query-Sequenz in einzelne Segmente und das Zusammenstellen einer Liste mit ähnlichen Segmenten, die einen minimalen Score erreichen, kann die eigentliche Suche schneller erfolgen, als es ohne diese Unterteilung möglich wäre.

Wird ein Treffer für ein Segment ermittelt, so wird dieser in beide Richtungen verlängert. Für eine Ausgabe des Treffers ist es nötig, dass ein zweiter Score für diese verlängerte Region überschritten wird. Zusammen mit dem Namen und einer kurzen Beschreibung der Treffer-Sequenz sind der erreichte Score und der *e-Value* angegeben. Der e-Value beschreibt die Wahrscheinlichkeit einen Treffer mit dem gleichen Score in einer Datenbank zufälliger Sequenzen zu finden. Je kleiner also der e-Value ist, desto geringer ist die Wahrscheinlichkeit, dass der ermittelte Treffer auf eine zufällige Ähnlichkeit zurückzuführen ist. Da kleine Datenbanken weniger Sequenzen enthalten, in denen der Treffer gefunden werden kann, sind in dieser Arbeit für Treffer aus kleinen Datenbanken höhere e-Values zugelassen als für größere. Die Werte verschiedener Datenbanken lassen sich aus diesem Grund nicht direkt miteinander vergleichen.

In dieser Arbeit werden drei Arten der BLAST-Suche durchgeführt: 1. Die cDNA-Sequenzen werden mit Nukleotiddatenbanken verglichen 2. Die sechs verschiedenen

Leserahmen der cDNA-Sequenzen werden in Proteindatenbanken und 3. in sechs Leserahmen der Nukleotiddatenbanken gesucht. Durch die BLAST-Suche ist eine erste Annotation der Sequenzen möglich. Die Ergebnisse der am besten annotierten Datenbank SwissProt werden zuerst auf Treffer mit einem niedrigem e-Value durchsucht. Wird ein solcher Treffer gefunden, kann für die Annotation der cDNA-Sequenz verwendet werden. Besitzt kein Treffer der BLAST-Suche gegen SwissProt einen e-Value der niedrig genug ist, werden die BLAST-Ergebnisse anderer Datenbanken in folgender Reihenfolge hinzugezogen: SwissProt, Cogeme (zuerst die blastn, dann die tblastx-Suche), Saccharomyces Genome Database (ebenfalls zuerst die blastn, dann die tblastx Suche) und als letztes die nr Datenbank. Es werden dabei keine Treffer für die Annotation verwendet, bei denen es sich um Sequenzen mit unbekannter Funktion oder um hypothetische Proteine handelt. In diesen Fällen wird der nächst bessere BLAST-Treffer verwendet.

2.7.2 Sequenzähnlichkeiten zu Adhäsions-relevanten Proteinen

Von besonderem Interesse sind die BLAST-Ergebnisse mit Treffern, die auf eine mögliche Rolle der cDNA-Sequenz in Adhäsionsprozessen hindeuteten. Um diese Sequenzen zu finden, müssen alle Programm-Ausgaben der Suchen gegen Nukleotid- und Proteindatenbanken durchmustert werden. Dazu werden Treffer der BLAST-Suche ermittelt, in denen bestimmte Schlüsselwörter auftreten, die auf eine Beteiligung des entsprechenden Proteins an der Adhäsion hindeuten (Tabelle 2.1). Die Schlüsselwörter beziehen sich direkt auf adhäsive Proteine aber auch auf Prozesse, an denen diese Proteine beteiligt sind. Bei den benutzten Begriffen handelt es sich um: Adhäsion, adhäsiv, Adhäsin, Flockulation, Flockulin, flockulativ, Hydrophobin, Biofilm, Muzin, Schleimstoffe, Lektin, Integrin, Oberflächenanhaftung, Oberflächen-Glykoprotein, invasives Wachstum, Zellwandprotein, Oberflächenprotein und GPI.

Der Informationsgehalt dieser Schlüsselwörter unterscheidet sich voneinander. Während das Wort »Adhäsin« in einem BLAST-Ergebnis ein starker Hinweis darauf ist, dass das Protein, das von der entsprechenden Sequenz kodiert wird, einen Einfluss bei der Adhäsion hat, ist der Begriff »Zellwandprotein« nur ein schwacher Hinweis. Dieses Schlüsselwort ist allgemeiner und trifft auch auf viele andere Proteine zu. Um die Qualität der Schlüsselwörter zu unterscheiden, werden ihnen Gewichtungen zwischen null und eins

zugewiesen, wobei eins Schlüsselwörtern mit einem starken Bezug zur Adhäsion zugeordnet wird (Tabelle 2.1).

Tabelle 2.1: Schlüsselwörter und ihre Gewichtungen

Schlüsselwort	Gewichtung
Adhesin	1
Adhesive	1
Adhesion	1
Surface Attachment	1
Mucin	1
Mucilage	1
GPI	1
Surface glycoprotein	0,9
Lectin	0,8
Flocculin	0,7
Flocculent	0,7
Flocculation	0,7
Hydrophobin	0.7
Integrin	0,6
Invasive Growth	0,6
Cell wall protein	0,4
Surface protein	0,4
Biofilm	0,4

Mit dem Programm `get_best_Keywords.pl` werden die BLAST-Ergebnisse der Kandidaten durchsucht. Als Argumente werden zwei Verzeichnisse und eine Datei erwartet. Eines der Verzeichnisse enthält die BLAST-Ergebnisse, in das zweite werden die Ausgabedateien gespeichert. Die Datei muß die zu suchenden Schlüsselwörter enthalten. Die Schlüsselwörter müssen dabei jeweils in einer Zeile der Datei gespeichert sein. Zusätzlich muss der maximale e-Value übergeben werden, den ein BLAST-Ergebnis haben darf,

wenn es verwendet werden soll.

Programmaufruf:

```
get_best_Keywords.pl <Eingabeverzeichnis> <Ausgabedatei> <Datei mit
Schlüsselwörtern> max. e-Value
```

Die BLAST-Ergebnisse im Eingabeverzeichnis werden nacheinander durchsucht. Für jede Datei wird die Datenbank, in der gesucht wurde, sowie die Art der BLAST-Suche ermittelt. Zeilen, die keine Angaben über den Namen und die Funktion des BLAST-Treffers enthalten, werden ignoriert.

Die BLAST-Treffer der Kandidaten, deren e-Value den festgelegten Maximalwert nicht überschreitet, werden einzeln gemustert und mit allen übergebenen Schlüsselwörtern verglichen. Enthält einer der Treffer ein solches Schlüsselwort, wird es für die Ausgabe gespeichert. Für jede Sequenz werden Sequenzname, die Datenbank, in der gesucht wurde, die Art der BLAST-Suche und die gefundenen Schlüsselwörter gespeichert. Die Anzahl der aufgetretenen Schlüsselwörter wird hinter dem jeweiligen Schlüsselwort gespeichert. Tabulatorzeichen trennen die einzelnen Elemente der Ausgabe voneinander.

Beispielausgabe:

Seq.	DB	BLAST-Pr.	Schlüsselw.	Schlüsselw.
a12	yeast_orf_coding	TBLASTX	Flocculation = 1	GPI = 1
a11	yeast_orf_coding	TBLASTX	Flocculation = 3	
a15	yeast_orf_coding	TBASLTX	GPI = 1	

Durch ein weiteres Programm, `sort_Keywords.pl`, werden die Schlüsselwörter aus BLAST-Suchen in verschiedenen Datenbanken in einer Datei zusammengestellt. Beim Aufruf wird das Verzeichnis, das die Ausgabe des Programmes `get_Keywords.pl` für verschiedene Datenbanken enthält und ebenso wie eine Ausgabedatei übergeben.

Programmaufruf:

```
sort_Keywords.pl <Eingabeverzeichnis> <Ausgabedatei>
```

Für jede Sequenz werden alle vorhandenen Schlüsselwörter und die zugehörigen BLAST-Suchen ausgegeben.

Beispielausgabe:

```
a4 :
yeast orf coding, TBLASTX, Cell wall protein = 1, Flocculation
= 1
cogeme, TBLASTX, GPI = 1

a12 :
yeast orf coding, TBLASTX, Flocculation = 1, GPI = 1, Invasive
growth = 1, Surface glycoprotein = 1
cogeme, TBLASTX, Surface glycoprotein = 1

a15 :
yeast orf coding, TBLASTX, GPI = 1
```

2.7.3 Bewertung der vorhandenen Schlüsselwörter mit einem Score

Bei der Suche nach Schlüsselwörtern in den BLAST-Treffer, ist es nur sinvoll alle gefundenen Schlüsselwörter auszugeben, wenn die Zahl der cDNA-Sequenzen und die der BLAST-Treffer nicht zu hoch ist und sich die Ergebnisse manuell durchsehen lassen. Dies ist für die Suche der Schlüsselwörter in den BLAST-Ergebnissen der Kandidaten-Sequenzen der Fall (siehe Abschnitt 2.7.2).

Die BLAST-Treffer vieler Sequenzen, wie die der kompletten cDNA-Datenbank, lassen sich jedoch nicht mehr manuell bewerten. Daher wird für Anzahl und Qualität der BLAST-Treffer mit Schlüsselwörtern ein Score berechnet. So können die Ergebnisse verschiedener cDNA-Sequenzen miteinander verglichen werden, ohne sie manuell duchge-

hen zu müssen. Die Güte eines Treffers ist dabei davon abhängig, welche Schlüsselwörter in einer BLAST-Suche auftauchen und ob sie am Anfang oder am Ende der BLAST-Trefferliste stehen. Treffer, die in verschiedenen Datenbanken gefunden werden, werden dabei unabhängig voneinander betrachtet. Für jede Sequenz in jeder durchsuchten Datenbank wird ein Score ermittelt. Für die Auswertung der BLAST-Ergebnisse wurden die Programme `get_Matrix.pl` und `get_Score.pl` erstellt.

Suche nach Schlüsselwörtern in den BLAST-Ergebnissen

Im ersten Schritt werden, wie schon bei der ersten Suche nach den Schlüsselwörtern, die BLAST-Ergebnisse aller cDNA-Sequenzen eingelesen und nach Schlüsselwörtern durchsucht. Dazu werden dem Programm `get_Matrix.pl` das Verzeichnis, das die BLAST-Dateien enthält, und ein Verzeichnis, in das die Ausgabedateien geschrieben werden, übergeben. Des Weiteren muss der Pfad zu der Datei angegeben werden, die die Schlüsselwörter enthält. Aus dieser werden die Schlüsselwörter, die mit der BLAST-Ausgabe verglichen werden sollen, zeilenweise eingelesen. Der letzte übergebene Wert legt fest, wie hoch der e-Value eines BLAST-Ergebnisses maximal sein darf, damit es verwendet wird.

Programmaufruf:

```
get_Matrix.pl <Verzeichnis mit Eingabedateien> <Verzeichis für
Ausgabedateien> <Schlüsselwörter-Datei> max. e-Value
```

Die Abfrage der einzelnen Schlüsselwörter erfolgt in alphabetischer Reihenfolge. Für jede BLAST-Datei wird eine Ausgabedatei erstellt. Sie enthält eine Zeile für jeden BLAST-Treffer. In der Zeile ist, durch Tabulatorzeichen getrennt, jeweils eine `1` für ein gefundenes Schlüsselwort angegeben und eine `0`, wenn das entsprechende Schlüsselwort nicht gefunden wurde. Eine Zeile der Ausgabedatei enthält also die Information, welche der Schlüsselwörter in einem BLAST-Treffer vorhanden sind. In der Beispielausgabe enthalten die ersten beiden BLAST-Treffer der Sequenz keine Schlüsselwörter. Im dritten BLAST-Treffer ist jedoch ein Schlüsselwort vorhanden. Hier ist eine Eins angegeben. Die Eins ist die fünfte Ziffer der Zeile, die hier für den Begriff »Cell wall protein«, steht. Aus

der Matrix lässt sich also ablesen, dass der dritte BLAST-Treffer der Sequenz das Schlüsselwort »Cell wall protein« enthält und dass keine weiteren Schlüsselwörter in diesem BLAST-Ergebnis vorkommen. Nicht berücksichtigt sind dabei die BLAST-Ergebnisse, deren e-Value den festgelegten Maximalwert übersteigt.

Beispielausgabe der Matrixdatei für die Suche der cDNA-Sequenz VL0713 gegen Swiss-Prot:

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Berechnung des Scores

Für das BLAST-Ergebnis einer Sequenz setzt sich der Score aus der Summe aller einzelnen Treffer zusammen. Ein hoher Score wird also gebildet, wenn die besten Treffer der BLAST-Suche Schlüsselwörter enthalten, wenn viele Schlüsselwörter vorkommen und wenn es sich um Wörter mit einem engen Bezug zur Adhäsion handelt.

Das Programm `get_Score.pl` liest die erstellten Matrix-Dateien ein. Die Ausgabe des Programmes enthält die Score-Werte für jede Sequenz und erfolgt in eine gemeinsame Datei. Ebenfalls übergeben werden muss eine Datei, die die zuvor gesuchten Schlüsselwörter enthält. In jeder Zeile dieser Datei folgt dem Schlüsselwort ein Tabulatorzeichen und der jeweils zugehörige Gewichtungsfaktor.

Programmaufruf:

```
get_Score.pl <Verzeichnis der Matrixdateien> <Ausgabedatei> <Datei
mit Schlüsselwörtern und deren Gewichtung>
```

Die Matrix-Dateien werden zeilenweise eingelesen. Wenn in einer Zeile eine oder mehrere Einsen auftreten zeigt dies an, dass in dem entsprechenden BLAST-Treffer ein oder

mehrere Schlüsselwörter vorhanden sind. Um welches Schlüsselwörter es sich handelt, lässt sich aus der Position der Eins in der Matrix schließen. Die Reihenfolge der Zahlen richtet sich nach den alphabetisch sortierten Schlüsselwörtern, die vom Programm eingelesen wurden.

Für einen gefundenen Treffer in der Matrix-Datei wird nun die Gewichtung des jeweiligen Schlüsselwortes w mit einem Faktor r multipliziert. Dieser ergibt sich aus der Position des entsprechenden BLAST-Treffers in der BLAST-Trefferliste. Steht er an der ersten Position, so ist der erreichte BLAST-Score der höchste dieser Suche, daher ist auch r hoch. Entsprechend ist r bei den letzten Positionen niedrig.

Die Berechnung des Scores erfolgt nach:

$$S = \sum_{i=1}^{n} w_i \, A_i \, r_i$$

mit:

w = Gewichtung des Schlüsselwortes

A = Schlüsselwort, $\in$ {0,1}

n = Anzahl der BLAST-Treffer

i = Index der BLAST-Treffer $\in$ {1,2,...,n}

r = Rang des BLAST-Treffers, r= $\frac{n-(i-1)}{N}$

Der Score für die Suche nach Schlüsselwörtern wird in eine Datei ausgegeben. Die Ergebnisse in dieser Datei werden in einem weiteren Programm, `sort_score.pl` verarbeitet. Die Score-Werte der einzelnen BLAST-Suchen werden nach Größe geordnet, sodass der Name der Sequenzen, für die der höchste Score ermittelt wurde, an den ersten Positionen der Liste auftauchen.

Programmaufruf:

```
sort_score.pl <Verzeichnis der Dateien mit ungeordnetem Score>
<Ausgabeverzeichnis für die Dateien mit geordneten Scores>
```

Die Ausgabedatei enthält in jeder Zeile den Namen einer cDNA-Sequenz und den da-

zugehörigen Score.

Beispielausgabe der Score-Werte bei der Suche gegen die Cogeme-Datenbank:

```
VL3449 = 3.89375
VL0683 = 1.5
VL4342 = 1.5
VL1443 = 1.5
VL0761 = 1
VL3873 = 1
VL1067 = 1
VL1878 = 1
VL3009 = 1
VL4181 = 0.833333333333333
VL3796 = 0.825
VL4483 = 0.825
VL2301 = 0.818181818181818
```

2.8 Suche nach ähnlichen Sequenzen in der cDNA-Datenbank

Die Anzahl identischer Sequenzen in einer cDNA-Datenbank kann Aufschluss darüber geben, wie stark das entsprechende Gen in dem Organismus, aus dem die cDNA-Sequenz gewonnen wurde, exprimiert wird. So lassen sich die cDNA-Sequenzen gruppieren, die vollständig oder teilweise übereinstimmen. Weiterhin können die Sequenzen der Kandidaten bestimmten Sequenzen der kompletten cDNA-Bank zugeordnet werden.

2.8.1 Suche nach ähnlichen Sequenzen

Als Maß für die Übereinstimmung zweier Sequenzen wird das Ergebnis ihrer BLAST-Suche gegeneinander verwendet. Sequenzen, deren e-Value geringer als 10^{-180} ist und

bei denen eine der Sequenzen zu mindestens 50 % mit der anderen übereinstimmt, werden als identisch angesehen. Das Programm `same_cDNA.pl` wurde geschrieben, um Sequenzen der selben Gene zu suchen. Dazu muss das Verzeichnis übergeben werden, das die Ergebnisse der BLAST-Suche der cDNA-Sequenzen gegeneinander bzw. der Kandidatensequenzen gegen die cDNA, enthält. Ebenso muss der Pfad einer Ausgabedatei übergeben werden, sowie der e-Value, der nicht überschritten werden darf, wenn die Sequenzen dem selben Gen zugeordnet werden sollen.

Programmaufruf:

```
same_cDNA.pl <Verzeichnis mit BLAST-Ergebnissen> <Ausgabedatei> max.
e-Value
```

Die BLAST-Suche der Kandidaten- oder die der cDNA-Sequenzen gegen die cDNA-Datenbank wird eingelesen und nach Treffern mit einem e-Value, der kleiner als 10^{-180} ist, durchsucht. Wenn bei der BLAST-Suche eine Sequenz gefunden wird, die die gesetzte Bedingung erfüllt, werden die einzelnen Treffer der Sequenz nach Plus- und Minusstrang sortiert. Es wird entweder die Homologie zum Plus- oder die zum Minusstrang berücksichtigt, da nur die Ähnlichkeit zu einem der beiden Stränge nötig ist. Die Treffer für beide Stränge werden einzeln auf ihre Länge überprüft. Wenn mehr als 50 % der Basen der kürzeren Sequenz mit denen der längeren übereinstimmen, werden diese beiden Sequenzen als gleiche Sequenzen ausgegeben.

Beispielausgabe:

Q-Seq.	L	str	L	Hit-Seq	id	e-val
b7	719	plus	777	VL1651	441	0
b7	719	plus	518	VL3927	434	0
a12	1338	plus	807	VL1696	690	0
a2	1151	plus	701	VL0377	701	0
a2	1151	plus	674	VL4412	674	0
a2	1151	plus	591	VL1287	591	0
a2	1151	plus	700	VL2273	641	0

Die Ausgabe enthält den Namen der Sequenz, mit der die BLAST-Suche durchgeführt wurde (Q-Seq), die Länge dieser Sequenz (L), die Strangorientierung (str) der homologen Sequenz sowie deren Länge und Name (Hit-Seq). Als Information zum BLAST-Treffer wird die Zahl der identischen Basen (id) und und die e-Values (e-val) der einzelnen Treffer ausgegeben.

2.8.2 Suche nach Gruppen ähnlicher Sequenzen

Um die Sequenzpaare in größeren Gruppen zusammenzufassen, wird das Programm `Find_groups.pl` angewendet, das die Ausgabe des Programmes `same_cDNA.pl` weiter verarbeitet. Die einzelnen Sequenzpaare werden in Gruppen eingeordnet, welche sich zu größeren Gruppen zusammenfassen lassen, wenn sie übereinstimmende Gruppenmitglieder haben. Auf diese Weise werden Paare in größere Einheiten eingeordnet.

Programmaufruf:

```
Find_groups.pl <same_cDNA.pl Ausgabedatei> <Ausgabedatei>
```

Die Sequenzpaare, die in der Ausgabedatei des Programmes `same_cDNA.pl` gespeichert sind, werden nacheinander durchlaufen. Für Paare, bei denen keine der beiden Sequenzen schon einmal vorkam, wird eine neue Gruppe erstellt, der sie beide zugeordnet werden. Wird jedoch in einer der schon erstellten Gruppen einer der beiden Sequenznamen gefunden, so wird die zweite Sequenz dieser Gruppe zugefügt. Auf diese Weise werden alle Sequenzpaare durchlaufen.

Der zweite Schritt besteht darin, die Gruppen miteinander zu vergleichen. Durch die Art, mit der die Sequenzpaare zuvor gruppiert wurden, kann es zu Überschneidungen kommen. Dies geschieht, wenn beispielsweise für zwei Sequenzpaare jeweils eine Gruppe erstellt wurde, ein späteres Sequenzpaar jedoch eine Verknüpfung der Gruppen herstellt. Dann wird das neue Sequenzpaar einer der Gruppen zugeordnet, die zweite Gruppe bleibt aber bestehen. Aus diesem Grund ist nach der Gruppenbildung eine Vereinigung sich überschneidender Gruppen nötig.

Besteht beispielsweise die Eingabe aus vier Sequenz-Gruppen mit den Sequenzenpaaren A und B, C und D, E und F, sowie E und B, werden diese zunächst zu drei Gruppen zusammengefasst. Zunächst werden die Gruppen AB, CD und EF gebildet. Die Gruppe, die E und B enthält könnte beiden Gruppen, AB und EF zugefügt werden. In diesem Schritt wird sie aber nur einer der Grupen angefügt, sodass die drei Gruppen ABE, CD und EF entstehen. Die im zweiten Schritt würden die Gruppen ABE und EF vereinigt werden.

Da es durch das Zusammenfassen zweier Gruppen zu neuen Überschneidungen kommen kann, wird der Schritt des Zusammenfassens iterativ wiederholt bis es keine Veränderung mehr gibt. In die Ausgabedatei werden in jede Zeile die Mitglieder einer Gruppe geschrieben.

Beispielausgabe:

```
VL0644  VL0777
VL1231  VL0776  VL2753
VL3755  VL0774  VL2143
VL0984  VL0773  VL3763
VL4422  VL0772  VL0400
```

2.8.3 Suche nach Korrelationen zwischen den Kandidaten

Mit Hilfe von zwei verschiedenen Programmen wird nach Ähnlichkeiten zwischen den BLAST-Ergebnissen der Kandidaten gesucht. Ähnliche BLAST-Ergebnisse zweier Sequenzen können darauf hindeuten, dass die Sequenzen zu Genen gehören, die für ähnliche Proteine kodieren. Dadurch, dass die Ergebnisse der BLAST-Suche verwendet werden, lässt sich zeigen, wenn zwei Kandidaten-Sequenzen Sequenzähnlichkeiten zu unterschiedlichen Regionen eines Proteins besitzen. Beide erstellte Programme lesen die Ergebnisse der BLAST-Suche der Kandidaten gegen eine bestimmte Datenbank ein und vergleichen die gefundenen Treffer miteinander. Das Programm `Blast_Correlation.pl` vergleicht dabei nur die zehn besten Ergebnisse der Kandidaten, `Blast_score_correlation.pl` vergleicht alle BLAST-Treffer, deren e-Value einen zuvor festgelegten maximalen Wert nicht überschreitet.

Korrelation zwischen den zehn besten BLAST-Treffern

Bevor die zehn besten Ergebnisse der einzelnen BLAST-Suchen miteinander verglichen werden können, werden diese zunächst von dem Programm `Best_10_hits.pl` ermittelt. Dem Programm müssen ein Verzeichnis mit den BLAST-Ergebnissen der Kandidaten und eine Ausgabedatei übergeben werden.

Programmaufruf:

```
Best_10_hits.pl <Verzeichnis mit BLAST-Ergebnissen> <Ausgabedatei>
```

Die einzelnen Dateien mit den BLAST-Ergebnissen werden nacheinander eingelesen. Die Ausgabe für jede dieser Dateien enthält den Namen der Sequenz mit der die BLAST-Suche durchgeführt wurde, die verwendete Datenbank, die Art der BLAST-Suche und, für jeden der ersten zehn Treffer, die ID und den Namen des Treffers, sowie den e-Value. ID und Name jedes Treffers sind durch ein Leerzeichen voneinender getrennt, Name und e-Value mit einem Komma und einem Leerzeichen. Zwischen den weiteren Ausgaben, also dem Name der Sequenz, der Datenbank, der BLAST-Suche und den Informationen zu den einzelnen Treffern, stehen jeweils Tabulatorzeichen. Alle Ausgaben für eine Eingebesequenz werden jeweils in eine Zeile der Ausgabedatei geschrieben. Wenn nicht mehr als zehn BLAST-Treffer vorhanden sind, werden alle vorhandenen in die Liste aufgenommen.

Beispielausgabe für die besten BLAST-Treffer eines Kandidaten:

```
b21  All non-redundant GenBank CDS  BLASTX  ref|XP_381808.1|
hypothetical protein FG01632.1 [Gibberella zeae PH-1], Expect =
6e-04  ref|XP_363452.1| hypothetical protein MGG_01378 [Magnapor-
the grisea 70-15], Expect = 9.4
```

Die in dem Beispiel gezeigten Ergebnisse einer BLAST-Datei enthalten nur zwei BLAST-Treffer. Die erstellte Ausgabe wird an das Programm `BLAST_Correlation.pl` überge-

ben. Dieses ermittelt für jeden Kandidaten die ID und die Liste aller durch die BLAST-Suche gefundenen Treffer.

Programmaufruf:

```
BLAST_Correlation.pl <Eingabedatei mit den zehn besten BLAST-
Treffern jeder Sequenz> <Ausgabedatei>
```

Die einzelnen BLAST-Treffer jedes Kandidaten werden mit den Treffern jedes anderen Kandidaten verglichen. Dabei wird für jedes mögliche Sequenzpaar ein Score berechnet, der die Ähnlichkeit der BLAST-Ergebnisse ausdrückt. Für jeden Treffer einer BLAST-Suche, der für beide Sequenzen gefunden wurde, wird der Score erhöht. Taucht der Treffer an einer der ersten Positionen der BLAST-Ergebnisliste auf, wird der Score stärker erhöht als bei ähnlichen Treffern am Ende der BLAST-Liste. Diese unterschiedliche Gewichtung wird eingeführt, da die in der BLAST-Ausgabe zuerst aufgeführten Treffer einen besseren Score erreicht haben als die nachfolgenden und ihnen somit eine höhere Relevanz zugeschrieben wird. Der Score wird als prozentualer Anteil der höchst möglichen Punktzahl ausgegeben.

Die Ausgabe erfolgt als Matrix der Sequenzpaare (Tabelle 2.2), in der die berechneten Scorewerte aufgeführt sind. Der höchste Score wurde hier für das Sequenzpaar b20 und b13 berechnet. Folglich ähneln sich die BLAST-Ergebnisse dieser Sequenzen am stärksten.

Korrelation zwischen den Score-Werten der BLAST-Treffer

Bei der zuvor geschilderten Methode zur Bestimmung der Ähnlichkeit der BLAST-Ergebnisse zweier Kandidaten wird nur die Position der BLAST-Treffer in den Score mit einbezogen. Damit die Qualität der einzelnen BLAST-Treffer stärker berücksichtigt werden kann, wird von dem Programm `BLAST_score_correlation.pl` der BLAST-Score bei der Berechnung des Ähnlichkeits-Scores für die Ähnlichkeit der BLAST-Treffer verwendet.

Zuerst durchsucht das Progamm `Blast_hits_and_score.pl` ein Verzeichnis mit

Tabelle 2.2: Matrix-Beispiel mit Score-Werten für die Korrelation der zehn besten BLAST-Ergebnisse der Kandidaten bei einer BLAST-Suche gegen die SwissProt DB

-	a11	a12	a13	a17	a19	a2	a21	a22	a3	a4	a5	a7	b1	b11	b13	b16	b17	b20	b21	b5
a11	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9.09	0	0	0	15.45
a12	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10.00
a13	-	-	-	0	0	0	0	0	0	0	0	0	3.64	0	0	0	0	0	0	0
a17	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a19	-	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a2	-	-	-	-	-	-	0	0	0	10	0	0	0	0	0	0	0	0	0	0
a21	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0	11.82
a22	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0
a3	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0
a4	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	0
a5	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0
a7	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0
b1	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0
b11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0
b13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	37.27	0	0
b16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	11.82
b17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0
b20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
b21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
b5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

BLAST-Ausgaben und schreibt eine Ergebnisdatei, die die Namen der besten Treffer und deren Scores enthält. Dabei sind die besten Treffer nicht, wie zuvor in Abschnitt 2.8.3 beschrieben, alle Ergebnisse, die an einer der zehn ersten Positionen der BLAST-Liste stehen. Die BLAST-Ergebnisse werden nach ihren e-Value ausgewählt. Dazu muss dem Programm ein maximaler e-Value übergeben werden. BLAST-Treffer mit einem e-Value, der größer ist als dieser übergebene Wert, werden ignoriert.

Programmaufruf:

```
Blast_hits_and_score.pl <Verzeichnis von BLAST-Ausgaben verschiede-
ner Sequenzen> < Ausgabedatei> <max. e-Value>
```

Jede BLAST-Ausgabe wird nach Treffern durchsucht, deren e-Value unter dem übergebenen Wert liegt. Die Treffer werden dann mitsamt BLAST-Score und ID der Query-Sequenz in die Ausgabedatei geschrieben. In jeder Zeile der Ausgabedatei sind die Treffer für eine Sequenz zu finden. Angegeben ist zuerst die ID der Query-Sequenz. Es folgen, durch Tabulatorzeichen voneinander getrennt, die verwendete Datenbank, die Art der

BLAST-Suche und Informationen zu jedem Treffer, der unter dem angegebenen e-Value liegt. Zu jedem Treffer sind zuerst ID und Name angegeben, gefolgt von drei Gleichheitszeichen und der BLAST-Score, mit dem der Treffer gefunden wurde.

Programmausgabe:

<ID> <Datenbank> <BLAST-Prog> <BLAST-Treffer === Score>

```
b14 /gobics/scratch/ingrid/cogeme/cogemeAlle.fas  BLASTN
MagCon[5130] transcriptional regulator ===75.8
```

Das Programm `Blast_score_correlation.pl` liest diese Ausgabe ein und erstellt eine Tabelle, die Ähnlichkeits-Scores für jedes Kandidatenpaar enthält.

Tabelle 2.3: Scorewert-Beispiel für die Korrelation der BLAST-Ergebnisse mit niedrigem e-Value der SwissProt Datenbank

	a10	a11	a13	a16	a17	a22	a3	a4	a5	b1	b11	b13	b16	b17	b20	b5
a10	-	0	0	0	0	0	0	0	1194.2	0	0	0	0	0	0	0
a11	-	-	0	0	0	103.2	0	0	0	0	0	0	84.3	0	0	83.5
a13	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0
a16	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0
a17	-	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0
a22	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	0
a3	-	-	-	-	-	-	-	0	0	0	0	0	75.8	0	0	0
a4	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0
a5	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0
b1	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0
b11	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0
b13	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0
b16	-	-	-	-	-	-	-	-	-	-	-	-	-	0	79.3	80.8
b17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
b20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
b5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Dazu wird der BLAST-Score der jeweiligen Treffer verwendet. Die Scores von BLAST-Ergebnissen, die für zwei verschiedene Sequenzen gefunden werden, werden miteinander addiert. Die Summe aller Scores von identischen Treffern zweier Sequenzen bildet den Score für die Korrelation dieser Sequenzen.

Die Scores werden, wie auch bei der Berechnung des Scores durch die Position des Treffers in der Ergebnisliste, als Tabelle (Tabelle 2.3) aufgeführt.

Kapitel 3

Ergebnisse

3.1 Analyse der cDNA-Datenbank

Um die Sequenzen der cDNA-Datenbak zu analysieren, wurden die in Kapitel 2 beschriebene Methoden angewandt. Im Folgenden werden die Ergebnisse dargestellt, die durch das Verwenden verschiedener Programme erhalten wurden.

3.1.1 Bereinigung der cDNA von der Plasmidsequenz

Der erste Schritt bei der Arbeit mit der cDNA-Bank ist das Entfernen von Sequenzen, die zu kurz sind, um bei Sequenzanalysen berücksichtigt zu werden, so wie das Bereinigen der cDNA-Sequenzen von der Plasmidsequenz. Zunächst wurden von den ursprünglich 5.652 Sequenzen 776 entfernt, die weniger als zwanzig bp lang sind. Es verbleiben 4.876 Sequenzen, die nach Plasmidsequenzen durchsucht werden (siehe Abschnitt 2.2). Durch das Entfernen der Plasmidsequenzen werden einige cDNA-Sequenzen so weit gekürzt, dass ihre Länge geringer als 21 bp ist. Diese Sequenzen werden ebenfals aus der cDNA-Bank entfernt.

Tabelle 3.1 gibt eine Übersicht über die Zusammensetzung der Datenbank vor und nach der Kürzung der Sequenzen. Dabei sind die Sequenzen, die aus dem ursprünglichen Datensatz der Sequenzen entfernt wurden, da sie weniger als zwanzig bp lang sind, nicht mit aufgeführt.

In 73,79 % der cDNA-Sequenzen wurde Plasmidsequenz gefunden und entfernt. Dadurch sinkt die Anzahl langer Sequenzen. Es sind nach dem Kürzen keine Sequenzen

mehr vorhanden, die mehr als 1.000 bp lang sind und auch die Anzahl der 500-999 bp langen Sequenzen nimmt ab. Ebenfalls sinkt die Zahl der kurzen Sequenzen, die nur bis zu 50 bp lang sind. Das erklärt sich dadurch, dass diese nach der Kürzung oft unter die Grenze von 21 bp fallen und aus der Datenbank entfernt werden. So nimmt die Gesamtanzahl der in der Datenbank vorhandenen Sequenzen ab, 303 Sequenzen wurden nach der Bereinigung aus der Datenbank entfernt.

In 27 Sequenzen wurde Plasmidsequenz gefunden, die innerhalb des Plasmids mehr als 30 Nukleotide von der MCS entfernt liegt. In diesen Fällen wird mit der Plasmidsequenz zusammen auch die downstream gelegene cDNA-Sequenz entfernt.

Tabelle 3.1: **cDNA-DB vor und nach der Bereinigung von Plasmidsequenz**

a) Überblick überdie Sequenzen in der DB; b) Anzahl der Sequenzen bei einer bestimmten Länge vor und nach der Bereinigung der Datenbank.

* Sequenzen der Länge 0-20 wurden aus der bereinigten Datenbank entfernt und fließen nicht in die bestimmte Gesamt- oder Durchschnittslängemit ein. Die Anzahl dieser Sequenzen ist jedoch der Vollständigkeit halber mit angegeben

		Ursprüngliche cDNA-Bank	Sequenzen >20 bp in der cDNA-Bank	Bereinigte cDNA-Bank
a)	Anzahl der Sequenzen	5.652	4.601	4.573
	Gesamtlänge	2.706.725 bp	2.705.765 bp	2.659.599 bp
	Durchschnittslänge	479 bp	588 bp	582 bp
	Längste Sequenz	978 bp	978 bp	954 bp
	Kürzeste Sequenz	0	21 bp	21 bp
	Geänderte Sequenzen		-	3.598
	> 30 bp von MCS	-		27
b)	Länge: 0-20 bp	1.051	0	303*
	Länge: 21-49 bp	150	150	127
	Länge: 50-99 bp	86	86	93
	Länge: 100-299 bp	306	306	320
	Länge: 300-499 bp	615	615	663
	Länge: 500-999 bp	3.444	3.444	3.370

3.1.2 Ermittlung der Kopienzahl der cDNA-Sequenzen

Die cDNA-Datenbank kann mehrfach vorhandene Sequenzen enthalten. Damit kontrolliert werden kann, ob mehrere Sequenzen starke Ähnlichkeiten aufweisen und deshalb auch ähnlich gute Ergebnisse erzielen, werden diese Sequenzen mit Hilfe einer BLAST-Suche gruppiert. Zusätzlich kann mit Hilfe der Gruppierung nachvollzogen werden, ob eine Sequenz in hoher Kopienzahl in der Zelle vorlag, ob also das entsprechende Gen stark exprimiert ist. Insgesamt kommen von den 4.573 Sequenzen der cDNA-Datenbank 2.392 Sequenzen mehr als nur einmal vor (Tabelle 3.2). Diese mehrfach vorkommenden Sequenzen lassen sich in 541 verschiedene Gruppen einordnen. Die Größe dieser Gruppen bewegt sich zwischen zwei und 208 Sequenzen, durchschnittlich lassen sich 4,4 Sequenzen zusamenfassen. Da 2.181 Sequenzen nur einmal vorhanden sind und 541 mehrfach vorliegen, sind insgesamt 2.722 verschiedene Sequenzen in der Datenbank vorhanden.

Tabelle 3.2: **Kopiezahl einer Sequenz in der cDNA-Datenbank**
Die Tabelle zeigt, wie die Sequenzen der cDNA-Datenbank in Gruppen ähnlicher Sequenzen zusammengefasst weden.

cDNA-Sequenzen	Gesamtzahl	4573
	mit nur einer Kopie	2181
	mehrfache Kopien	2392
	verschiedene Sequenzen	2722
Gruppen	Anzahl	541
	Größten Kopiezahl	208
	kleinsten Kopiezahl	2
	Durchschnittliche Kopiezahl	4,4

Die mehrfach vorkommenden Sequenzen wurden nicht aus der Datenbank entfernt. Obwohl sie sich stark ähneln, können sie unterschiedlich lang sein und so verschiedene Re-

gionen einer Sequenz enthalten. Auch kann es vorkommen, dass sie in den homologen Regionen nicht vollkommen deckungsgleich sind. Die Unterschiede können auf verschiedene Kopien im Genom von *V. longisporum*, aber auch auf Sequenzierfehler hindeuten. Im ersten Fall ginge Information verloren, wenn nur eine Kopie in der Datenbank verbleibt, im zweiten könnte bei der Entscheidung für eine der Sequenzen die Variante ohne Sequenzierfehler verloren gehen, während nur eine fehlerhafte Version benutzt wird.

3.1.3 Kodierende Regionen der cDNA-Sequenzen

Mithilfe von ESTScan wurden für die Sequenzen der cDNA kodierende Regionen vorhergesagt (siehe Abschnitt 2.4.1). Die Ergebnisse, die man beim Anwenden von ESTScan auf eine cDNA-Sequenzen erhält, hängen stark davon ab, mit welchen Sequenzen das Programm trainiert wurde. Hier werden die Ergebnisse gegenübergestellt, die für die cDNA-Sequenzen erhalten werden, nachdem das Programm einerseits mit Sequenzen der Gruppe Sordariomycetes und andererseits Sequenzen aller Ascomycota trainiert wurde. Verticillium gehört zur Klasse der Sordariomycetes, die wiederum der Abteilung der Ascomycota, angehört. Sequenzen aus *V. longisporum* konnten jedoch nicht zum Training verwendet werden, da keine bekannt sind.

Die Tabelle 3.1.3 zeigt wie viele kodierende Regionen von ESTScan vorhergesagt werden, wenn man jeweils eines der Trainingssets verwendet. Auch angegeben ist, welche cDNA-Sequenzen dem Programm übergeben wurden. In den 4.573 Sequenzen, die dem Programm übergeben werden, werden nach dem Training von ESTScan mit Sordariomycetes 4.106 und nach dem mit Ascomycota 4.223 kodierende Sequenzen gefunden. In drei Fällen werden zwei kodierende Sequenzen, die den gleichen Score erreicht haben, für eine eingegebene cDNA-Sequenz vorgeschlagen.

Da nicht für alle cDNA-Sequenzen eine kodierende Region vorgeschlagen wird und die nicht-kodierenden Regionen entfernt werden, sinkt die Gesamtlänge der Sequenzen in der Datenbank von 2.659.599 bp auf 2.101.775 (Sordariomycetes), bzw. 2.165.871 bp (Ascomycota) ab. Die meisten kodierenden Sequenzen werden für Plusstrang vorhergesagt, auf dem Minusstrang wird nur in 195 (Sordariomycetes), bzw. 286 (Ascomycota))Fällen eine kodierende Region vorhergesagt. 2.432 (Sordariomyceten), bzw. 2.416 (As-

Tabelle 3.3: **Vorhergesagte kodierende Regionen in der cDNA-Datenbank** Die Daten der cDNA-Datenbank und der von ESTScan gefundenen kodierenden Regionen werden angegeben. Dabei werden die Ergebnisse von ESTScan, die mit unterschiedlichen Trainingsets erhalten werden, verglichen: Die, die nach dem Training der HMMs mit Sequenzen von Sordariomyceten, gefunden werden, sowie die, die nach dem Training mit Ascomyceten gefunden werden. Bei eimem X an erster Position der AS-Sequenz beginnt die kodierende Region vor der cDNA-Sequent (siehe Abschnitt 2.4.1

	cDNA	kodierende Seq (Sordariomycetes)	kodierende Seq (Ascomycota)
Anzahl der Sequenzen	4.573	4.106	4.223
2 Codierenden LR	-	3	3
Gesamtlänge	2.659.599 bp	2.101.775 bp	2.165.871 bp
Durchschnittslänge	582 bp	512 bp	513 bp
Gesamtlänge der AS-Sequenz	-	701.437	722.833
Durchschnittslänge der AS-Sequenzen	-	171	171
Längste Sequenz	954 bp	938 bp	954 bp
Kürzeste Sequenz	21 bp	97 bp	97 bp
Länge: 0-49 bp	127	0	0
Länge: 50-99 bp	93	7	5
Länge: 100-299 bp	320	637	652
Länge: 300-499 bp	663	1.135	1.158
Länge: 500-999 bp	3.370	2.327	2.408
Auf dem Minusstrang	-	195	286
1. Nukleotid ist kodierend	-	1.683	1.816
erste AS Methionin	-	2.432	2.416
erste AS X	-	1.123	1.241

comycota) der vorhergesagten kodierenden Regionen beginnt mit einem Methionin, was darauf hindeutet, dass der Anfang der kodierenden Region vollständig vorhanden ist. 1.123 (Sordariomycetes), bzw. 1.241 (Ascomycota) der vorhergesagten kodierenden Regionen beginnen mit einem X, da hier zusätzliche Nukleotide an die cDNA-Sequenzen angehängt wurden, um alle Leserahmen berücksichtigen zu können. In diesen Fällen beginnt die als kodierend vorhergesagte Region am Anfang der cDNA-Sequenz und weist kein Start-Kodon auf.

Obwohl die HMMs auf verschiedenen Sequenzen trainiert wurden, werden ähnliche kodierende Regionen vorhergesagt (Tabelle 3.4. Von den insgesamt 4.247 vorhergesagten kodierenden Regionen liegen 96,13 % auf dem selben Strang der cDNA-Sequenz. Dabei

beginnen 83,78 % und enden 91,10 % der als kodierend vorhergesagten Regionen mit dem selben Nukleotid. Durchschnittlich sind die Startpunkte 18,65 Nukleotide und die Endpunkte 9,28 Nukleotide voneinander entfernt. In 171 cDNA-Sequenzen wird nur von einem der unterschiedlich trainierten HMMs eine kodierende Region vorhergesagt, von dem anderen jedoch nicht.

Tabelle 3.4: **Vergleich der vorhergesagten kodierenden Sequenzen**
Vergleich der kodierenden Regionen, die mit ESTScan nach dem Training mit Sordariomycetes-Sequenzen vorhergesagt werden, mit den, die nach einem Training mit Ascomycote-Sequenzen erhalten werden. Dazu ist angegeben, wie viele der kodierenden Regionen bei den beiden Vorhersagen auf dem gleichen Strang liegen und die gleiche Start- oder Endposition haben. Ebenfalls angegeben sind die durchschnittliche und maximale Entfernung zwischen den Start- und Endpunkten, die für jeweils eine Sequenz bestimmt wurden, sowie die Zahl der Sequenzen, für die nur mit einem Trainingsset eine kodierende Region gefunden wird.

Seqenzen insgesamt	4573
Gleiche Orientierung Identischer Start Identisches Ende	4082 3558 3869
durchschnittliche Entfernung der Startpunkte	18,65
durchschnittliche Entfernung der Endpunkte	9,28
Maximaler Abstand zwischen den Startpunkten	691
Maximaler Abstand zwischen den Endpunkten	617
Nur in einer Vorhersage gefunden	171

3.1.4 Suche nach Sequenzen mit einer Serin- und Threonin-reichen Regionen

Die Proteinsequenz von Adhäsinen beinhaltet Bereiche, in denen die Aminosäuren Serin und Threonin gehäuft vorkommen (siehe Abschnitt 1.4.1). Werden in der Datenbank Proteine mit einer solchen Region gefunden, so ist dies ein Hinweis darauf, dass es sich um Adhäsine handeln könnte.

Die Ergebnisse der Suche nach Bereichen mit einem hohen Anteil der AS Serin (Ser) und Threonin (Thr) sind von verschiedenen Faktoren abhängig. Je nachdem, welcher Anteil von Ser und Thr in einer Sequenz mindestens erreicht werden muss, werden mehr oder weniger Treffer gefunden. Der aus der Literatur stammende Wert von 35 % [51] wird in dieser Arbeit übernommen. Dies ist jedoch für die Länge der Region nicht ohne weiteres möglich. Für sie ist in der Literatur ein Bereich von 300 AS angegeben. Im Fall der vorliegenden cDNA-Datenbank kann jedoch nicht davon ausgegangen werden, dass alle Sequenzen vollständig sequenziert wurden. So kann es vorkommen, dass nur ein Teil der Sequenz vorliegt. Die Mindestlänge, die zur Ausgabe der Ergebnisse verlangt wird, ist also auch vom vorliegenden Datensatz abhängig. Aus diesem Grund wurde die Suche im Rahmen dieser Arbeit mit verschiedenen Werten für die Mindestlänge durchgeführt.

Tabelle 3.5: **Serin- und Threonin-reiche Regionen in der AS-Sequenz**
Es wird jeweils die Zahl der Serin- und Threonin-reicher Regionen gezeigt, die für unterschiedliche Minimallängen gefunden werden. Da jeweils in allen sechs LR der cDNA-Sequenzen gesucht wurde, ist mit angegeben, in wie vielen unterschiedlichen cDNA-Sequenzen die gefundenen Regionenen liegen. Ebenfalls angegeben ist die Anzahl der gefundenen Serin- und Threonin-reichen Bereiche in den von ESTScan vorhergesagten kodierenden Regionen der cDNA. Die vorhergesagten kodierenden Regionen sind wiederum unterteilt in solche, die bei einer Vorhersage nach dem Training mit Sequenzen von Ascomycota und in solche, die nach dem Training mit Sordariomycetes gefunden werden.

Minimale Länge	LR der cDNA	Treffer in LR unterschiedlicher cDNA-Sequenzen	Ascomycota	Sordariomycetes
100 AS	1588	1586	24	23
200 AS	188	188	0	0
250 AS	15	15	0	0
300 AS	0	0	0	0

Eine Übersicht der gefundenen Bereiche mit einem erhöhten Serin- und Threoninanteil ist in Tabelle 3.5 gezeigt. In Tabelle 3.5 wird gezeigt, dass in der cDNA-Datenbank keine Ser/Thr-reichen Regionen mit einer Länge von 300 AS vorhanden ist. Es werden fünfzehn Sequenzen gefunden, die eine Ser/Thr-reiche Region besitzen, die länger als 250 AS ist. Die Zahl der gefundenen Sequenzen mit Ser/Thr-Bereichen, steigt, wenn die minimale Größe dieses Bereiches herruntergesetzt wird. Da die Suche für jede der möglichen AS-Sequenzen der sechs Leserahmen der cDNA-Sequenz durchgeführt wird, kann es vorkommen, dass in verschiedenen Leserahmen derselben cDNA-Nukleotidsequenz Treffer gefunden werden. Damit diese Treffer nicht doppelt gezählt werden, denn es ist unwahrscheinlich, dass beide Leserahmen kodierend sind, ist ebenfalls die Zahl der unterschiedlichen Sequenzen angegeben, in mindestens eine Ser/Thr-reiche Region gefunden wurde. Nur wenn die Minimallänge auf 100 As gesetzt wird, werden zwei Treffer gefunden, die in unterschiedlichen Leserahmen der selben Sequenz liegen.

Wenn nur die als kodierend vorhergesagten Bereiche der cDNA-Sequenz nach Ser/Thr-reichen Regionen durchsucht werden, werden deutlich weniger Treffer gefunden, als bei der Suche in den AS-Sequenzen zu allen Leserahmen. Es werden 24 (Ascomycota), bzw 23 (Sordariomycetes) Regionen gefunden, die Ser/Thr-reiche Bereiche mit einer Länge von mehr als 100 AS besitzen. Eine Länge von 200 AS für einen Bereich mit hohem Ser/Thr-Anteil wird dabei nicht überschritten. Die in den vorhergesagten kodierenden Regionen gefundenen Bereiche mit einem hohen Anteil von Serin und Threonin stammen jedoch nicht aus den cDNA-Sequenzen, in deren Leserahmen Ser/Thr-reiche Bereiche mit einer Länge von mehr als 200 AS liegen.

3.1.5 Suche nach GPI-Ankersignalen in der AS-Sequenz

Neben der Serin- und Threoni n-reichen Region weisen Adhäsine einen GPI-Anker an ihrem C-Terminus auf (siehe Abschnitt 1.4.1). In den Sequenzen der cDNA-Datenbank kann nach den Signalsequenzen, die das Anbringen eines GPI-Ankers ermöglichen, gesucht werden. Bei dieser Suche werden die drei Programme GPI-SOM, DGPI und Big-Π verwendet (siehe Abschnitt 2.6).

Da die Programme eine Eingabesequenz, bestehend aus den zwanzig kanonischen

AS erwarten, können nur Sequenzen verwendet werden, die weder Stopp-Kodons noch nicht identifizierte AS enthalten. Stopp-Kodons tauchen dann in der Eingabesequenz auf, wenn die vollständige cDNA-Sequenz in die sechs möglichen AS-Sequenzen übersetzt wird. Kommt in einem der LR ein Stopp-Kodon vor, wird die gesammte Sequenz nicht mehr verwendet. Die 4573 Sequenzen der cDNA-Datenbank enthalten 4228 AS-Sequenzen ohne Stopp-Kodon, wobei diese AS-Sequenzen aus den verschiedenen LR von 2612 cDNA-Sequenzen stammen. Tabelle 3.6 zeigt, wie viele der Sequenzen lang genug sind, um als Eingabesequenz genutzt werden zu können.

Die kodierenden Regionen, die von ESTScan vorhergesagt werden, können dagegen nicht identifizierte AS enthalten (siehe Abschnitt 2.4.1). In einem solchen Fall werden die nicht identifizierten AS am Anfang und am Ende der AS-Sequenz entfernt, bevor diese Sequenzen vom Programm eingelesen werden. Bei einer durch ESTScan vorhergesagten Insertion in der Sequenz ist es ebenfalls nicht möglich, die dem entsprechenden Kodon zugehörige AS eindeutig zu bestimmen. In einem solchen Fall taucht eine nicht identifizierte AS in der Mitte der AS-Sequenz auf. Eine solche Sequenz wird nicht mehr in die Vorhersage von Signalsequenzen einbezogen. Bei den als kodierend vorhergesagten Regionen ist das für 645 (Sordariomycetes), bzw. 717 (Ascomycota) Sequenzen der Fall. So können nicht alle zur Verfügung stehenden Sequenzen für eine Vorhersage genutzt werden.

Tabelle 3.6: **GPI-Signalsequenzen in der AS-Sequenz der cDNA**

Vergleich der Ergebnisse der GPI-Suche mit verschiedenen Programmen. Angegeben sind die Anzahl der verwendeten LR, die Anzahl der verschiedener Sequenzen, denen die LR zugeordnet ist, und die Zahl der Sequenzen für die Signalsequenz vorhergesagt werden. Für die vorgesagten kodierenden Regionen ist nur ein Wert angegeben, da für jede Sequenz nur eine Region vorrausgesagt wird.

	Anzahl der Eingabesequenzen						Anzahl der GPI-Bindestellen					
	GPI-SOM		DGPI		Big-II		GPI-SOM		DGPI		Big-II	
	LR	Seq	LR	Seq	LR	Seq	LR	Seq	LR	Seq	LR	Seq
LR der cDNA	3989	2529	4228	2612	3287	2397	304	296	115	113	27	27
vorherges. kod. (Sordario.)	3461		3461		3286		303		118		83	
vorherges. kod. (Ascom.)	3506		3506		3346		316		116		84	

In Tabelle 3.6 wird gegenübergestellt, wie viele Sequenzen von den verschiedenen Programmen durchsucht werden können und wie viele GPI-Signalsequenzen für sie vorhergesagt werden. Diese Zahlen unterscheidet sich voneinander, da die unterschiedlichen Programme verschiedene Eingabebedingungen an eine Sequenz stellen.

Die meisten cDNA-Sequenzen werden von DGPI mit 4228 AS-Sequenzen durchsucht, gefolgt von Big-II (3989) und GPI-SOM (3287). Wird nur die Zahl der verwendeten cDNA-Sequenzen berücksichtigt, die mind. ein LR ohne Stopp-Kodon besitzen, werden von DGPI die LR von 2612 cDNA-Sequenzen durchsucht. Die meisten Ankersignale werden von GPI-SOM vorhergesagt. Die angegebenen 304 Treffer liegen in 296 verschiedenen cDNA-Sequenzen. Es sind also nur für acht cDNA-Sequenzen GPI-Ankersignale in zwei LR der selben Sequenz gefunden worden. DGPI findet in 115 LR, die aus 133 cDNA-Sequenzen stammen, eine Signalsequenz. Die wenigsten Vorhersagen werden von Big-II getroffen, wo 27 Ankersequenzen gefunden werden.

Bei der Suche nach GPI-Ankersignalen in den als kodierend vorhergesagten Bereichen der cDNA, die durch ESTScan ermittelt wurden, ist für jede cDNA-Sequenz nur eine AS-Sequenz vorhanden. Die Werte für die verwendeten LR und den Sequezen, aus denen sie stammen, unterschieden sich hier deshalb nicht. Mit dem Programm GPI-SOM werden 303 (Sordariomycetes), bzw. 316 (Ascomycota) GPI-Ankersignale gefunden. Dabei ähneln sich die Ergebnisse für die vorhergesagten kodierenden Regionen, die unter Verwendung verschiedener Trainingssequenzen gefunden wurden. 281 der über 300 Sequenzen werden für beide kodierende Regionen gefunden. 57 Ankersignale werden für nur eine der als kodierend vorhergesagten Region vorrausgesagt. Das Gleiche trifft auf die Ergebisse des Programmes DGPI zu. Hier werden 104 Ankersignale in beiden als kodierend vorhergesagten Regionen gefunden. Auch die Zahl der Insgesamt vorhergesagten Signalsequenzen ist ähnlich. Es werden 118 (Sorariomycetes, bzw. 116 (Ascomycota) Sequenzen gefunden. Wird das Programm Big-II verwendet, werden 83 Signalsequenzen in zwei kodierenden Regionen gefunden. Dabei liegt die Gesamtzahl der vorhergesagten Signalsequenzen bei 83 (Sordariomycetes), bzw. 84 (Ascomycota) und übersteigt somit auch die Zahl der Ankersignale in den LR der cDNA-Sequenzen deutlich.

Die Zahl der Ankersignale, die von den verschiedenen Programmen gefunden werden, unterscheidet sich voneinander, abhängig davon, welche der beiden Programme miteinander verglichen werden (Tabelle 3.7). Die wenigsten Übereinstimmung gibt es zwischen den Vorhersagen von Big-ΠI und GPI-SOM. Nur für eine AS-Sequenz wird von beiden Programmen ein Ankersignal gefunden. Hier wird nur für eine Sequenz von beiden Programen eine GPI-Anker-Bindestelle gefunden. Die größte Übereinstimmung gibt es zwischen den Ergebnissen von Big-ΠI und DGPI, wenn die Vorhersage für die kodierenden Bereiche getroffen wird. Es werden 28 Ankersignale von beiden Programmen vorhergesagt. Das entspricht 23,73 % der Ergebnisse von DGPI und 33,73 % der Ergebnisse von Big-ΠI. Bei der Suche mit den AS-Sequenzen der verschiednen LR der cDNA-Sequenzen werden jedoch nur in sieben der Sequenzen Ankersignale von beiden Programmen gefunden. Die Ähnlichkeit zwischen den Ergebnissen der Programme GPI-SOM und DGPI ist nicht so stark. Hier werden für 5,93 % (Sordariomycetes), bzw. 6,80 % (Ascomycota) und 19,13 % (LR der cDNA) der mit DGPI gefundenen Ankersignale auch mit GPI-SOM Signalsequenzen gefunden. In keiner AS-Sequenz werden von allen Programmen Ankersignale gefunden.

Tabelle 3.7: **Vergleich der Vorhersage der GPI-Ankerstellen**
Es wird die Zahl der cDNA-Sequenzen angegeben, für die die Vorhersagen der Ankersignale von jeweils zwei Programmen übereinstimmen.

	LR der cDNAs		Ascomycota		Sordariomycetes	
	DGPI	Big-ΠI	DGPI	Big-ΠI	DGPI	Big-ΠI
GPI-SOM	22	1	8	1	7	1
DGPI	-	7	-	28	-	28

3.1.6 Ähnlichkeiten zwischen cDNA-Sequenzen und Sequenzen aus anderen Datenbanken

Bei der Suche nach Sequenzhomologien wurde BLAST verwendet. Für die durchsuchten Datenbanken Saccharomyces Genome Database, Cogeme, SwisProt und nr wurden verschiedene maximale e-Values gewählt, die die BLAST-Ergebnisse nicht überschreiten durften, wenn sie für weitere Analysen verwendet werden. Der maximale e-Value für die Suche gegen die Saccharomyces Genome Datebase wurde auf zehn gesetzt. Bei diesen Bedingungen gibt es mit blastn für zwanzig und mit tblastx für 78 cDNA-Sequenzen kein BLAST-Ergebnisse. Der maximale e-Value für die BLAST-Suche gegen die Cogeme- und die SwissProt-Datenbank wurde auf 0,1 gesetzt. In den BLAST-Ergebnissen der Suche gegen SwissProt gibt es 1.632 cDNA-Sequenzen, die für keine Ergebnisse vorliegen, deren e-Value niedrig genug ist. Bei der Suche mit blastn gegen Cogeme werden für 905, bei der Suche mit tblastx für 274 cDNA-Sequenzen keine BLAST-Ergebnisse gefunden, die unter diesen Bedingungen verwendet werden können. Der maximale e-Value für die BLAST-Ergebnisse der Suche gegen die bei weitem größten nr-Datenbank wurde der e-Value auf 10^{-4} gesetzt. Die BLAST-Ergebnisse von 836 cDNA-Sequenzen erfüllen diese Bedingung nicht und werden nicht verwendet.

Tabelle 3.8: **Ergebnisse der BLAST-Suche mit verschiedenen Datenbanken**
Es wird angegeben, für wie viele cDNA-Sequenzen keine BLAST-Treffer gefunden werden, deren e-Value kleiner als die festgelegten maximalen Werte (10; 0,1; 10^{-4}) ist.

Datenbank	BLAST-Suche	cDNA-Sequenzen ohne Treffer		
		e-Value ≤ 10	e-Value $\leq 0{,}1$	e-Value $\leq 10^{-4}$
Cogeme	blastn	12	905	1719
Cogeme	tblastx	60	274	901
Sacch. Gen. DB	blastn	20	3401	4259
Sacch. Gen. DB	tblastx	78	1953	2553
SwissProt	blastx	266	1632	2091
nr	blastx	0	658	836

3.1.7 Suche nach Schlüsselwörtern in den BLAST-Ergebnissen

Die BLAST-Ergebnisse der cDNA-Sequenzen weisen auf eine Vielzahl von Sequenzähnlichkeiten zu Datenbank-Sequenzen hin. Um in diesen Ergebnissen Ähnlichkeiten zu Nukleotid- und Proteinsequenzen, die an der Adhäsion von Zellen an Oberflächen beteiligt sind, zu finden, wurden die BLAST-Ergebnisse nach Schlüsselwörtern durchsucht (siehe Abschnitt 2.7.2). Jedem BLAST-Ergebnis jeder cDNA-Sequenz wird ein Score-Wert zugewiesen, der Qualität und Quantität, der in diesen Ergebnissen gefundenen Schlüsselwörter, widerspiegelt. In der Tabelle 3.9 ist angegeben, wie hoch der jeweils beste Score ist und für wie viele cDNA-Sequenzen ein Score-Wert erreicht wird, der großer als Null ist. Dabei werden immer zwei Werte gegenübergestellt. Die Werte, die erreicht werden, wenn alle BLAST-Ergebnisse mit einem e-Value, der kleiner als zehn ist akzeptiert werden und die, deren BLAST-Ergebnisse nur mit in die Berechnung einbezogen werden, wenn der e-Values unter dem für die entsprechende Datenbank gewählten Maximum liegt (siehe Abschnitt 3.1.6).

Für einem großen Teil der BLAST-Ergebnisse wird ein Score berechnet, der größer als

Tabelle 3.9: **Scorewerte der Schlüsselwort-Suche bei verschiedenen BLAST-Suchen**
Angegeben ist die Zahl der cDNA-Sequenzen, für deren BLAST-Ergebnisse ein Score errechnet werden kann, der größer als Null ist und der beste Score, der mit den BLAST-Ergebnissen einer cDNA-Sequenz für die Suche in einer bestimmten Datenbank erreicht wird.

DB	BLAST-Suche	Seq mit Score > 0	bester Score
Cogeme	blastn	72	3,89
Cogeme	tblastx	427	7,63
Sacchar. Gen. DB	blastn	766	2,53
Sacchar. Gen. DB	tblastx	1809	9,34
SWISSPROT	blastx	175	11,68
nr	blastx	168	32,66

Null ist. Die Zahl der cDNA-Sequenzen, deren Score diesen Wert überschreitet, hängt dabei von der Datenbank ab, gegen die die BLAST-Suche erfolgte. Die Datenbank-Suche, mit den meisten Ergebnissen, die Schlüsselwörter enthalten, ist die Saccharomyces Geno-

me Database bei einer Suche mit tblastx. In diesem Fall werden in den BLAST-Ergebnissen von 1809 cDNA-Sequenzen Schlüsselwörter gefunden.

Die höchsten Score-Ergebnisse, also die BLAST-Suchen, in denen viele Schlüsselwörter mit hoher Gewichtung und/oder an den ersten Positionen der BLAST-Trefferliste gefunden werden, sind bei Suchen in der nr- oder SwissProt-Datenbank zu finden. Mit einem Score von 32,66 wird für die cDNA-Sequenz VL2221 in den Ergebnissen der BLAST-Suche gegen die nr-Datenbank der höchste Wert erreicht. In den Ergebnissen der BLAST-Suchen gegen weitere Datenbanken sind jedoch keine Schlüsselwörter vorhanden, sodass es keinen weiteren Scorewert für diese Sequenz gibt, der über Null liegt. Das Schlüsselwort, das in den BLAST-Treffern der nr-Datenbank auftaucht ist GPI. Jedoch steht es in diesem Fall für die Glucose-6-Phosphat Isomerase, die ebenso wie der Glycosylphosphatidylinositol-Anker mit GPI abgekürzt wird. Dies ist auch der Grund, aus dem nur in einer der Datenbanken ein positiver Score erreicht wird. Nur in der nr-Datenbank wird diese Abkürzung verwendet.

Es gibt keine cDNA-Sequenz, die für alle sechs BLAST-Suchen einen positiven Schlüsselwort-Score erreicht. Nur neun Sequenzen haben für fünf der BLAST-Suchen einen Score, der größer als Null ist. 73 cDNA-Sequenzen haben vier, 86 Sequenzen haben drei und 500 Sequenzen haben zwei positive Scorewerte. Ein Großteil der Sequenzen besitzt nur einen Score, der größer als Null ist.

3.2 Analyse der Sequenzen der Adhäsionskandidaten

Es liegen 43 Kandidaten-Sequenzen vor, die im Vorfeld dieser Arbeit schon experimentell ermittelt wurden (siehe Abschnitt 1.6.3). Dabei handelt es sich um 24 verschiedene Sequenzen, einige sind mehrfach vertreten. Diese mehrfach vertretenen Sequenzen unterschieden sich jedoch durch ihre Länge und die Basenabfolge leicht voneinander. Aus diesem Grund kann es zu unterschiedlichen Ergebnissen kommen, weshalb die Ergebnisse für alle 43 Sequenzen dargestellt werden.

Da es sich nur um 43 Sequenzen handelt, können diese einzeln genauer untersucht werden, als dies für alle Sequenzen der cDNA-Datenbank möglich ist. Auch sind diese Sequenzen von besonderem Interesse, da für ihre Genprodukte bereits Hinweise auf ein

Mitwirken am Prozess oder der Regulation der Adhäsion vorliegen.

3.2.1 Bereinigung der cDNA von Plasmidsequenzen

Die Kandidaten-Sequenzen enthalten, wie auch die Sequenzen der cDNA-Datenbank, Plasmidsequenz. Aus diesem Grund ist es nötig, dass die Sequenzen bereinigt werden (siehe Abschnitt 2.2).

Tabelle 3.10: **Vergleich der bereinigten und unbereinigten Kandidaten-Sequenzen**
a) Daten zu den Kandidaten-Sequenzen vor und nach der Bereinigung von der Plasmidsequenz; b) Länge der einzelnen Sequenzen

		Kandidaten aus dem Δflo8-Screen		Kandidaten aus dem Δflo8/11-Screen	
		cDNA-Sequenzen	Bereinigte cDNA-Sequenzen	cDNA-Sequenzen	Bereinigte cDNA-Sequenzen
a)	Anzahl der Sequenzen	22	22	21	21
	Gesamtlänge (bp)	27.518	23.961	26.228	23.111
	Durchschnittslänge (bp)	1.251	1.089	1.249	1.101
	nicht identifizierte Nukleotide	0	0	5.486	4.204
	Längste Sequenz (bp)	1.376	1.376	1.149	1.309
	Kürzeste Sequenz (bp)	1.029	27	1.309	206
	Geänderte Sequenzen	-	11	-	5
	> 30 bp von MCS	-	4	-	5
b)	Länge: 0-20 bp	0	0	0	0
	Länge: 21-49 bp	0	1	0	0
	Länge: 50-99 bp	0	0	0	0
	Länge: 100-299 bp	0	0	0	1
	Länge: 300-499 bp	0	2	0	0
	Länge: 500-999 bp	0	1	0	4
	Länge: 1000-1499 bp	22	18	21	16

Tabelle 3.10 zeigt, wie stark die Kandidaten-Sequenzen dabei gekürzt werden. Während im unbereinigten Datensatz alle Sequenzen eine Länge zwischen 1.000 und 1.499 bp besitzen, erreichen von den Kandidaten aus dem Δflo8-Screen vier, und von den Kandidaten aus dem Δflo8/11-Screen fünf Sequenzen diese Länge nicht mehr. Die kürzeste berei-

nigte Sequenz aus dem Δflo8-Screen hat eine Länge von 27 bp. Die längste Sequenz ist 1.376 bp lang, sowohl vor als auch nach der Bereinigung. Auch bei den Kandidaten aus dem Δflo8/11-Screen muss die längste Sequenz mit 1.309 bp nicht bereinigt werden. Bei der Sequenzierung der Sequenzen aus dem Δflo8/11-Screen konnten einige Nukleotide nicht identifiziert werden. In den unbereinigten Sequenzen sind 5.486 von 26.228 Nukleotiden nicht identifiziert. Das entspricht durchschnittlich 261 nicht identifizierten Nukleotiden pro Sequenz. Die nicht identifizierten Nukleotide werden in der Sequenz mit »N« angegeben. Nach der Bereinigung von der Plasmidsequenz sind nur noch 4.204 nicht identifizierte Nukleotide bei einer Gesamtlänge von 23.111 bp vorhanden. Die durchschnittliche Häufigkeit von nicht identifizierten Nukleotiden pro Sequenz sinkt also auf 210 Nukleotide.

3.2.2 Vorhergesagte kodierende Bereiche in der Sequenz der Kandidaten

In den bereinigten cDNA-Sequenzen der Kandidaten werden mit dem Programm ESTScan kodierende Regionen vorhergesagt. Durch das Entfernen der nicht-kodierenden Sequenz nimmt die Länge der Sequenzen ab (Tabelle 3.11). Während die achtzehn (Δflo8), bzw. sechtzehn (Δflo8/11) der Kandidaten-Sequenzen länger als 1.000 bp sind, ist ein Großteil der vorhergesagten kodierenden Regionen zwischen 500 und 999 bp lang. Für Δflo8/11 ist das in fünfzehn (Sordariomycetes), bzw. zwölf (Ascomycota) mal der Fall. Bei den Δflo8-Sequenzen trifft es auf sechs (Sorsariomycetes), bzw. elf (Ascomycota) Sequezen zu. HMMs, die auf unterschiedlichen Sequenzen trainiert wurden, liefern bei der Vorhersage verschiedene Ergebnisse. So werden beispielsweise in den Sequenzen der Δflo8/11-Kandidaten einmal zwanzig und einmal neunzehn kodierende Regionen ermittelt. Unterschiede liegen auch in der Länge der vorhergesagten kodierenden Regionen.

3.2.3 Suche nach Sequenzen mit einer Serin- und Threonin-reichen Regionen

Die Tabelle 3.12 zeigt die Anzahl der AS-Sequenzen der Kandidaten, in denen Serin- und Threonin-reiche Bereiche gefunden werden. Angegeben sind sowohl Treffer, die in der AS-Sequenz der Kandidaten-cDNA gefunden werden, die vollständig und in allen sechs

Tabelle 3.11: **Vorhergesagte kodierende Regionen in den Kandidaten-Sequenzen** Die Daten der bereinigten Kandidaten-Sequenzen und der von ESTScan gefudenen kodierenden Regionen werden angegeben. Dabei werden die zwei Ergebnisse von ESTScan verglichen: Die, die nach dem Training der HMMs auf die Sequenzen von Sordariomycetes, gefunden werden, sowie die, die nach dem Training auf Ascomycota gefunden werden. Wird als erste AS ein X angegeben, konnte die Identität dieser AS nicht eindeutig festgestellt werden, da die vorhergesagte kodierende Region vor dem Beginn der cDNA anfängt.

	Δflo8	vorherges. kod. Seq Sordariomycetes	vorherges. kod. Seq Ascomycota	Δflo8/11	kod. Seq Sordariomycetes	kod. Seq Ascomycota
Anzahl der Sequenzen	22	21	21	21	20	19
Gesamtl. (bp)	23.961	19.058	18.909	23.111	15.080	16.229
Durchschnl. (bp)	1089	908	900	1101	754	854
Gesamtl. der AS-seq.	-	6.345	6.298	-	5.025	5.409
Durchschnl. der AS-seq.	-	302	300	-	251	285
längste Sequenz (bp)	1.376	1.338	1.338	1.309	1.338	1.244
kürzeste Sequenz (bp)	27	311	311	206	311	206
Länge: 0-49 bp	1	0	0	0	0	0
Länge: 50-99 bp	0	0	0	0	0	0
Länge: 100-299 bp	0	0	0	1	3	1
Länge: 300-499 bp	2	3	2	0	0	1
Länge: 500-999 bp	1	6	11	4	15	12
Länge: 1000-1499 bp	18	12	8	16	2	5
Auf dem Minusstrang	-	1	0	-	0	1
1. Nukleotid kodierend	-	8	10	-	11	15
erste AS Methionin	-	13	11	-	7	3
erste AS X	-	6	8	-	13	16

Leserahmen übersetzt wird, als auch Treffer, die in der durch ESTScan vorhergesagten kodierenden Region gefunden werden. Im Gegensatz zu den Ergebnissen, die bei der Suche in den Sequenzen der gesamten cDNA-Sequenz erhalten werden (siehe Abschnitt 3.1.4), wird kein Bereich gefunden, der länger als 200 AS ist. Die Länge von 100 AS wird nur vier mal überschritten. Werden dagegen die Ser/Thr-reichen Bereiche betrachtet, die eine Länge von 50 AS erreichen, wird für den Großteil der Sequenzen eine solche Region gefunden. Siebzehn der insgesamt 22 Δflo8-Kandidaten besitzen mindestens einen Leserahmen, der eine solche Region enthält. Von 21 Δflo8/11-Kandidaten trifft das auf fünfzehn Kandidaten zu.

Werden nur die als kodierend vorhergesagten Regionen in den Kandidaten-Sequenzen durchsucht, werden weniger Ser/Thr-reiche Regionen gefunden. Für Δflo8-Kandidaten werden nur acht Ser/Thr-reicher Bereiche, die länger als 50 AS sind gefunden. Von diesen acht Bereichen sind jedoch vier, bzw. drei, auch länger als 100 AS. Die vorhergesagten kodierenden Regionen der Δflo8/11-Kandidaten besitzen weniger Ser/Thr-reichen Be-

reiche mit einer Länge von mehr als 50 AS, als das für die AS-Sequenz der vollständigen Nukleotidsequenz der Fall ist. Sowohl für die vollständigen, als auch für die vorhergesagten kodierenden Regionen von Δflo8/11 werden vier Sequenzen gefunden, die einen Ser/Thr-reichen Bereich besitzen, der länger als 100 AS ist. Es wird jedoch von keiner das als kodierend vorhergesagten Region die Länge von 200 AS überschritten.

Tabelle 3.12: **Serin und Threonin-reiche Regionen in den Sequenzen der Kandidaten**
Dargestellt ist die Anzahl der Kandidaten mit einem Sequenz-Bereich, in dem vermehrt Kodons für Serin und Threonin auftreten. Die Anzahl der gefundenen Treffer wird in Abhängigkeit von der minimalen Länge, die diese Region hat, gezeigt. Es ist sowohl angegeben in wie vielen LR Treffer gefunden werden, als auch aus wievielen Sequenzen die LR mit den Treffern stammen. Zusätzlich sind die Treffer für die vorhergesagte kodierenden Regionen aufgeführt

Kandidat	Min. Länge	LR der Kandidaten-cDNA	versch. Seq.	kodierende Seq. Ascomycota	kodierende Seq. Sordariomycetes
Δflo8	50 AS	21	17	8	8
	100 AS	6	6	4	3
	200 AS	0	0	0	0
Δflo8/11	50 AS	20	15	9	8
	100 AS	4	4	4	4
	200 AS	0	0	0	0

In den als kodierend vorhergesagten Regionen, die von unterschiedlich trainierten HMMs vorhergesagt werden, werden die gleichen Ser/Thr-Bereiche gefunden. Insgesammt liegen in acht kodierenden Regionen der Kadidaten also Ser/Thr-Bereiche, die länger als 100 AS sind. Sieben dieser acht Kandidaten besitzen eine starke Sequenzähnlichkeit zueinander (Tabelle A.1), sodass letztlich nur zwei verschiedene Sequenzen einen Ser/Thr-Bereich besitzen, der länger als 100 AS ist. Bei diesen beiden Kandidaten handelt es sich um a2 und a6.

3.2.4 Suche nach GPI-Ankersignalen in der Aminosäure-Sequenz

Die drei Programme Big-II, DGPI und GPI-SOM werden zur Vorhersage von GPI-Ankersignalen verwendet. Dabei können, wie in Abschnitt 3.1.5 geschildert, nur die AS-Sequenzen verwendet werden, die ausschließlich die zwanzig kanonischen AS verwenden. Kandidaten-Sequenzen die Stopp-Kodons enthalten werden nicht zur Vorhersage genutzt. Nur fünf AS-Sequenzen der Δflo8-Kandidaten beinhalten kein Stopp-Kodon (Tabelle 3.13). Diese Sequenzen werden von DGPI durchsucht. Da die Sequenzen, die von den Programmen GPI-SOM und Big-II nach Ankersignalen durchsucht werden, eine Mindestlänge von 32, bzw. 53 AS erreichen müssen, sinkt hier die Zahl der verwendeten Sequenzen auf Eins.

Tabelle 3.13: **GPI-Signalsequenzen in der AS-Sequenz der Kandidaten**
Vergleich der Ergebnisse der GPI-Suche mit verschiedenen Programmen. Angegeben sind die Anzahl der durchsuchten Sequenzen und die Zahl der Sequenzen die eine Signalsequenz enthalten. Ebenso sind die Ergebnisse für die vorhergesagte kodierenden Regionen aufgeführt.

		Eingabesequenzen						GPI-Ankersignale					
Kandidaten	AS-Sequenz	GPI-SOM		DGPI		Big-II		GPI-SOM		DGPI		Big-II	
		LR	Seq	LR	Seq	LR	Seq	LR	Seq	LR	Seq	LR	Seq
Δflo8	LR der cDNA	1	1	5	2	1	1	0	0	0	0	0	0
	Sordariomycetes	17	17	17	17	17	17	2	2	1	1	1	1
	Ascomycota	14	14	14	14	14	14	2	2	1	1	1	1
Δflo8/11	LR der cDNA	24	19	745	21	23	19	1	1	2	2	0	0
	Sordariomycetes	4	4	4	4	4	4	0	0	0	0	0	0
	kod. Ascomycota	3	3	3	3	3	3	0	0	0	0	0	0

Nicht identifizierte Nukleotide in der cDNA-Sequenz der Kandidaten führen dazu, dass die zugehörige AS-Sequenz nicht mehr ermittelt werden kann. Dies kommt vermehrt bei Δflo8/11-Kandidaten vor (siehe Abschnitt 3.2.1). Diese Sequenzen werden in Teilsequenzen gegliedert, die jeweils ausschließlich identifizierte Nukleotide enthalten. Die verschiedenen Leserahmen werden für diese Teilsequenzen bestimmt. Dadurch kommt es für die Δflo8/11-Kandidaten zu der hohen Zahl von 879 durchsuchten LR. Davon

erreichen jedoch nur 24 die Mindestlänge von 32 AS, die benötigt wird, wenn das Programm GPI-SOM angewendet werden soll. Das Programm Big-Π kann nur 23 Sequenzen verwenden.

Die als kodierend vorhergesagten Regionen werden vor der Suche nach Ankersignalen nach nicht identifizierten AS durchsucht. Wie in Abschnitt 3.1.5 dargestellt, werden Sequenzen mit nicht identifizierten AS nicht zur Ankervorhersage genutzt. Siebzehn (Sordariomycetes), bzw. vierzehn (Ascomycota) der vorhergesagten kodierenden Regionen sind ausreichend lang und enthalten nur die kanonischen AS, sodass sie für die Suche nach Signalsequenzen genutzt werden können. Die Zahl der Δflo8/11-Sequenzen, die nach Ankersignalen durchsucht werden ist geringer. Da diese Sequenzen viele nicht identifizierte Nukleotide enthalten, werden nur vier (Sordariomycetes), bzw. drei (Ascomycota) der als kodierend vorhergesagten Bereiche gefunden, die nur identifizierte Nukleotide enthalten. Die Mindestlänge von 53 AS erreichen jedoch alle als kodierenden vorhergesagte Regionen.

Es werden nur wenige Ankersignale in den Kandidaten-Sequenzen gefunden. In den vorhergesagten kodierenden Regionen von Δflo8 werden zwei Signalsequenzen von GPI-SOM und eine von DGPI und Big-Π gefunden. Die Ergebnisse von Big-Π und DGPI sind dabei identisch. Von beiden Programmen wird eine Signalsequenz in dem Kandidaten a2 vorrausgesagt. GPI-SOM findet zusätzlich Ankersignale für a13 und a8. Für die kodierendenRegionen von Δflo8/11 und auf die LR der cDNA-Sequenzen von Δflo8 werden keine GPI-Anker-Bindestellen gefunden.

Die GPI-Ankersignale, die in den LR der Δflo8/11-Kandidaten gefunden werden stammen aus den drei unterschiedlichen cNDA-Sequenzen. Jedoch sind sich zwei der Sequenzen, b8 und b6, für die eine Ankersequenz gefunden wird, sehr ähnlich. In b8 wird von GPI-SOM ein Ankersignal vorhergesagt, in b6 von DGPI.

Es spielt keine Rolle,welches der Trainingssets benutzt wurde, um die mutmaßlichen kodierenden Regionen zu bestimmen. Für beide Regionen liegen jeweils die gleichen Ergebnisse vor. Die Ergebnisse, die für die vorhergesagten kodierenden Regionen und die LR der Kandidaten erhalten werden unterscheiden sich jedoch. Hier muss aber berücksichtigt werden, dass nur wenig LR für die Vorhersagen verwendet werden, da die meisten Sequenzen zu viele Stopp-Kodons enthalten.

3.2.5 Homologien zu Sequenzen verschiedener Datenbanken

Die Sequenzen der Kandidaten werden mit den Sequenzen verschiedener Datenbanken verglichen. Bei der Berücksichtigung aller BLAST-Treffer wird für eine Sequenz kein BLAST-Ergebnis erhalten (Tabelle 3.14). Bei einer Einschränkung der BLAST-Treffer durch einen maximalen e-Value von 0,1 steigt die Zahl der Kandidaten-Sequenzen, für die unter diesen Bedingungen kein Ergebnis erhalten wird, auf bis zu zwanzig an. Bei einem maximalen e-Value von 10^{-4} werden für die nr Datenbank, der einzigen, für die dieser Grenzwert gesetzt wurde, nur für drei Kandidaten-Sequenzen keine BLAST-Ergebnisse erhalten. Die mit Hilfe der BLAST-Ergebnisse erstellte Annotation der Kandidaten ist in Abschnitt 3.3.1 dargestellt.

Tabelle 3.14: **Ergebnisse der BLAST-Suche mit verschiedenen Datenbanken**
Gezeigt ist die Anzahl der Kandidaten-Sequenzen, für die keine BLAST-Ergebnisse gefunden werden. Dabei wird dargestellt, wie sich diese Zahl mit Veränderung eines maximalen e-Values verändert.

	Datenbank	BLAST-Suche	keine BLAST-Treffer		
			e-Value ≤ 10	e-Value $\leq 0{,}1$	e-Value $\leq 10^{-4}$
Δflo8/11	Cogeme	blastn	0	11	14
	Cogeme	tblastx	0	1	5
	Saccharomyces Genome Database	blastn	1	20	21
	Saccharomyces Genome Database	tblastx	1	12	16
	SWISSPROT	blastx	1	7	15
	nr	blastx	0	1	3
Δflo8	Cogeme	blastn	0	7	12
	Cogeme	tblastx	1	1	7
	Saccharomyces Genome Database	blastn	0	15	22
	Saccharomyces Genome Database	tblastx	0	9	13
	SWISSPROT	blastx	1	7	10
	nr	blastx	1	3	3

3.2.6 Suche nach Korrelationen zwischen den Kandidaten

Ein hoher Wert für die Korrelation zweier Sequenzen kann auf verschiedene Weisen zustande kommen. Beide Sequenzen können zu dem gleichen Bereich einer dritten Sequenz homolog sein. In diesem Fall könnten die Sequenzen für ähnliche Proteine kodieren. Die Sequenzen können dann dazu genutzt werden durch ein multiples Alignment ein Motiv dieser Proteinregion zu erstellen, das dann für weitere Suchen genutzt werden kann. Es kann jedoch auch vorkommen, dass die Sequenzen zu verschiedenen Bereichen des selben Proteins homolog sind. Es besteht dann die Möglichkeit, dass die beiden Sequenzen aus verschiedenen Teilen des gleichen Gens stammen, dass beim Sequenzieren aber beide Male nicht die vollständige Basenabfolge ermittelt werden konnte. Wird für zwei Sequenzen ein BLAST-Treffer gefunden, können die beiden Sequenzen aus unterschiedlichen Genen stammen, die für Domänen kodieren, die beide im gefundenen BLAST-Treffer vorkommen.

Der höchste Score, der von einem Sequenzpaar erreicht wird, liegt bei 98,18 für den Vergleich der zehn besten BLAST-Treffer und bei 21.583,5 für die Addition der BLAST-Scores identischer Treffer. Beide Male wird der Score von dem Sequenzpaar a5 und a10 erreicht. Die Übereinstimmung von 98,18 % der zehn besten Treffer wurde für die BLAST-Ergebnisse der SwissProt-Datenbank ermittelt. Beim Betrachten der BLAST-Ergebnissen ist zu beobachten, dass die Homologien zu den Treffern nicht in den gleichen Regionen der Proteine liegen. a10 ist zum vorderen Bereich des Proteins homolog, a5 zu einer Region im hinteren Bereich des Proteins. Dies trifft auf alle neun SwissProt-Treffer zu, die die beiden Sequenzen gemeinsam haben.

Beim direkten Vergleich der beiden Sequenzen mit BLAST (Abb.3.1) wird eine Homologie zwischen dem Ende von a10 und dem Anfang von a5 gefunden. In diesem Bereich von 260 bp sind 228 Positionen identisch. Es ist möglich, dass es sich bei a5 und a10 um verschiedene Teile des gleichen Proteins handelt. Das beste BLAST-Ergebnis bei der Suche gegen SwissProt ist für beide Sequenzen *sak1*, DNA-bindendes Protein, das in *Schizosaccharomyces pombe* den Austritt aus dem mitotischen Zellzyclus positiv reguliert.

Die Sequenzen b16 und b20 haben einen Ähnlichkeitsscore von 79,3 für den Vergleich der

```
a10:   627 ELEMRHLFGANRHRKLQEVAQELHGNERGPNSERTRQVFAMLWINSVCSKGK-GSVPRGR 803
           E E++ L   + H  L+ +A++L  +    NSE+ RQVF + W+   C + +  +V R +
sak1:   62 ETELKRLALEHEHYSLESLAEKLRMDHVSANSEKFRQVFGICWLKRACEEQQDAAVQRNQ 121

a10:   804 VYANYASRCATERITVLNPASFGKLVRVLFPGLKTRRLGVRGESKYHYVNFAL 962
           +YA+Y   C +  I  LN ASFGKLVR+LFP +KTRRLG+RG SKYHY    L
sak1:  122 IYAHYVEICNSLHIKPLNSASFGKLVRLLFPSIKTRRLGMRGHSKYHYCGIKL 174


a5:    335 AIALPRIEPFLPQGTDPDAAKS-LSALYRSHCTSLVECVRYCKEKTFFHLYTSFQGTLTM 511
           + +LP I+ +L    D   AKS L  +Y SHC +L+E VRY   K F    ++F  +L+
sak1:  370 SFSLPPIDYYLNGPYDNVEAKSALMNIYSSHCITLIESVRYMHLKQFLSEISNFPNSLSP 429

a5:    512 PVQKLLGNASVAPWIEECDFILYQRMMQIVSGLTLQVVPKPVLDTLRSISTRLVPHIREA 691
            +  LL +     WIE  D ++Y+ +++++  +TLQVVP PVL  LR ++  LV HI
sak1:  430 SLLALLSSPYFTKWIERSDTVMYREILKLLFPMTLQVVPPPVLVLLRHLAENLVNHISSI 489

a5:    692 FQGQPLHVIRAKEAPATVFAALLDRALKVNLTAHAAANMLSNPANRDQMYLDWITMVPVR 871
           +      +++ K   A +F+ LL R L+VN TAHAAA  L+NPA+R  +  DW   V  R
sak1:  490 YASHSSCLLQVKSETAAIFSNLLSRLLRVNDTAHAAARFLANPADRHLICNDWERFVSTR 549

a5:    872 KVA-------ESIPTRGMDDFVNLV--VSEIRDLLDPQSVPWEVECLTVYGDLASEKSDS 1024
            +        +      +D++ +++   S   +LLDP     E          AS+ S +
sak1:  550 FIVHRELMCNDKEAVAALDEWYSILSTCSNPSELLDPLKDKHE----------ASDTSMN 599

a5:   1025 S*QRWRVSGKNVLERWVTFW-SLLQRFP 1105
             +  ++ G  VL+R   F+  L  RFP
sak1:  600 RVELRQIDG--VLDRMADFFLELPSRFP 625
```

Abbildung 3.1: BLAST-Alignment der Sequenz von sak1 mit der von a5, bzw. a10.

Ergebnisse einer BLAST-Suche gegen die SwissProt-Datenbank. Die Bereiche des BLAST-Treffers, zu denen die beiden Sequenzen homolog sind, überschneiden sich. Während b16 zu den AS 448 bis 589 des Atrophin-1 der Ratte (*Rattus norvegicus*) homolog ist, hat b20 einen Treffer für den Bereich 371 bis 535 des gleichen Proteins. Bei einer BLAST-Suche der Sequenzen gegeneinander, wird kein Alignment gefunden. Die Ähnlichkeit der Sequenzen ist bei einem direkten Vergleich der Sequenzen nicht groß genug um ein Alignment zu erstellen, das als Motiv für weitere Suchen genutzt werden kann.

Die Ähnlichkeiten, die zwischen den BLAST-Treffern der Sequenzen gefunden wurden, reichen nicht aus, um Sequenzen in Grupppen einzuteilen, aus deren multiplen Alignments sich ein Motiv bilden lässt. Dieses Ergebnis wird durch BLAST-Suchen der Sequenzen gegeneinander bestätigt. Jedoch können die Ergebnisse der Sequenzen a10 und a5 miteinander verglichen werden, da die große Anzahl identischer BLAST-Treffer die Vermutung nahelegt, dass es sich um Sequenzen zweier ähnlicher oder des gleichen Gens

handelt. Da sich die Sequenzen bei einer BLAST-Suche aber nur teilweise überschneiden und auch die BLAST-Ergebnisse, die zwischen den beiden identische sind, nicht in den selben Bereichen der Datenbanksequenz gefunden werden, lässt sich kein aussagekräftiges Motiv bilden.

Suche nach Homologien zu Adhäsion-relevanten Proteinen

In den Ergebnissen der BLAST-Suche gegen verschiedenen Datenbanken sind Treffer vorhanden, die die gesuchten Schlüsselwörter beinhalten (Tabelle 3.15). Während in den BLAST-Ergebnissen der cDNA-Datenbank zu viele Schlüsselwörter gefunden werden, als dass eine manuelle Auswertung möglich wäre, lassen sich die Ergebnisse für die Kandidaten der Adhäsion einzeln durchgehen.

Bei den 43 cDNA-Sequenzen, die die Fähigkeit der Hefezellen zur Adhäsion wieder herstellen, handelt es sich nur um 24 verschiedene Sequenzen. Einige von ihnen kommen in leicht abgewandelter Form mehrfach vor. So kann es geschehen, dass sich die BLAST-Ergebnisse dieser sehr ähnlichen Sequenzen unterscheiden. Ein Beispiel sind die Sequenzen a11 und b2. b2 hat 918 von 1247 Basenpositionen mit a11 gemeinsam. Trotzdem werden verschiedene BLAST-Treffer gefunden. Während für a11 nur die Schlüsselwörter »Zellwand-Protein« und »Flockulation« in der BLAST-Suche mit der Saccharomyces Genome Database gefunden werden, ist in der BLAST-Suche der Sequenz b2 gegen die Cogeme-Datenbank ebenfalls ein Treffer mit einem Schlüsselwort enthalten. Bei diesem handelt es sich um »Muzin«, das in den BLAST-Ergebnissen von a11 nicht zu finden ist.

Für die Sequenzen a22, a12, a17, a4 und b13 werden Schlüsselwörter in den BLAST-Ergebnisse verschiedener Datenbanken gefunden. Die Treffer von a12 beinhalten dabei sowohl die Schlüsselwörter Flockulation, als auch GPI, Surface glycoprotein und Invasive growth. Die beiden Sequenzen a17 und a2 haben einen BLAST-Treffer, der das Schlüsselwort Adhäsion, bzw. Adhäsin enthält. Während der BLAST-Treffer von a2 direkt ein Adhäsin ist, handelt es sich bei dem von a17 um die Anker-Untereinheit des a-Agglutinin der Hefe. Dies bindet an die Adhäsions-Untereinheit Aga2p.

Tabelle 3.15: **Schlüsselwörter in den BLAST-Suchen der Kandidaten**
Dargestellt ist wie viele Schlüsselwörter in den BLAST-Ergebnissen der Kandidaten für verschiedene Datenbanken gefunden werden. Mit den Schlüsselwörtern zusammen ist angegeben, wie oft sie jeweils gefunden werden.

Sequenz	DB	BLAST	Schl. 1	Anzahl	Schl. 2	Azahl	Schl. 3	Anzahl
a16	Sacc. Gen. DB	TBLASTX	GPI	1				
a5	Sacc. Gen. DB	BLASTN	Surface glycoprotein	1	Invasive Growth	1	GPI	1
a22	Sacc. Gen. DB	TBLASTX	Invasive growth	1	GPI	1		
	Cogeme	TBLASTX	Lectin	1				
a12	Sacc. Gen. DB	TBLASTX	Invasive growth	1	Flocculation	1	GPI	1
	Cogeme	tBLASTX	Surface glycoprotein	1				
a10	Sacc. Gen. DB	TBLASTX	GPI	1				
a17	Sacc. Gen. DB	TBLASTX	Adhesion	1	GPI	2		
	Cogeme	TBLASTX	Cell wall protein	1				
b16	Sacc. Gen. DB	TBLASTX	GPI	1				
a4	Sacc. Gen. DB	TBLASTX	Cell wall protein	1	Flocculation	1		
	Cogeme	TBLASTX	GPI	1				
a2	Sacc. Gen. DB	TBLASTX	Adhesin	1				
a13	Sacc. Gen. DB	BLASTN	Invasive growth	2	Flocculation	1		
b13	Cogeme	TBLASTX	Mucin	1				
	Sacc. Gen. DB	TBLASTX	Invasive growth	1				
a19	yeast orf coding	BLASTN	Invasive growth	1				
b11	Sacc. Gen. DB	BLASTX	Flocculin	1	GPI	1		
b4	SwissProt	BLASTX	Mucin	1				
a11	Sacc. Gen. DB	TBLASTX	Cell wall protein	2	Flocculation	3		
b2 (a11)	Cogeme	TBLASTX	Mucin	1				
a7	Sacc. Gen. DB	TBLASTX	Flocculation	1	GPI	1		
b10 (a7)	Sacc. Gen. DB	TBLASTX	Invasive growth	1	GPI	1		

Suche nach Schlüsselwörtern in BLAST-Ergebnissen

Obwohl sich die BLAST-Ergebnisse der Kandidaten-Sequenzen, die Schlüsselwörter enthalten, überschaubar sind (siehe Abschnitt 3.2.6), ist auch das berechnen der zugehörigen Schlüsselwörter von Vorteil. So kann neben dem eigentlichen Schlüsselwort auch die Position der BLAST-Trefferliste berücksichtigt werden, an der dieses Schlüsselwort gefunden wird. Tabelle 3.16 zeigt die Scores der BLAST-Ergebnisse der Kandidaten.

Tabelle 3.16: **Scorewerte der Schlüsselwort-Suche bei verschiedenen BLAST-Suchen**
Berechnung des Scores siehe Abschnitt 2.7.3

	DB	BLAST-Suche	Seq mit Score > 0		bester Score		Durchschnittsscore	
			e ≤ 10	e ≤ max.	e ≤ 10	e ≤ max.	e ≤ 10	e ≤ max.
Δflo8/11	Cogeme	blastn	1	0	1,06	0	0,35	0
	Cogeme	tblastx	13	2	2,02	0,67	1,25	0,58
	Saccharomyces Gen. DB	blastn	0	0	0	0	0	0
	Saccharomyces Gen. DB	tblastx	11	11	1,33	1,33	0,76	0,76
	SwissProt	blastx	12	1	2,90	0,72	0,93	0,72
	nr	blastx	5	0	1,26	0	0,92	0
Δflo8	Cogeme	blastn	0	0	0	0	0	0
	Cogeme	tblastx	15	4	2,56	0,58	0,97	0,25
	Saccharomyces Gen. DB	blastn	4	4	2,5	2,5	1,48	1,48
	Saccharomyces Gen. DB	tblastx	13	13	2,7	2,7	0,93	0,93
	SwissProt	blastx	3	0	1,25	0	0,88	0
	nr	blastx	3	0	1,1	0	0,75	0

3.2.7 Suche nach Kandidaten-Sequenzen in der cDNA-Datenbank

Einige Sequenzen der Kandidaten für Adhäsion sind auch in der cDNA-Datenbak vorhanden (Tabelle 3.17). Durch eine BLAST-Suche der Kandidaten gegen die cDNA-Sequenzen werden diese Sequenzähnlichkeiten ermittelt (siehe Abschnitt 2.8). In allen Fällen werden die Sequenzpaare mit einem BLAST-Treffer gefunden, dessen e-Value gleich Null ist. Die Basen, die zwischen beiden Sequenzen identisch sind, entsprechen in den meisten Fällen nahezu der gesamten Länge der jeweils kürzeren Sequenz. Diese ist somit vollständig in der längeren enthalten. Die meisten homologen Sequenzen sind für den Kandidaten a16 vorhanden. Neunzehn Sequenzen in der cDNA-Datenbank weisen Ähnlichkeiten zu der Sequenz von a16 auf. Auch a2 ist mit neun Sequenzen in der cNDA-Datenbank vertreten. Das nicht alle Sequenzen der Kandidaten in der Datenbank gefunden werden können, liegt daran, dass beim Sequenzieren nicht jede in der cDNA-Bank vorhandene Sequenz auch abgelesen wird. Je höher jedoch die Anzahl der Stränge mit gleicher

Sequenz, desto wahrscheinlicher ist es, dass diese auch sequenziert wird und in der Datenbank auftaucht.

Tabelle 3.17: **Kandidaten-Sequenzen in der cDNA-Datenbank**

Die Kandidatensequenz und ihre Länge sind den cDNA-Sequenzen, zu denen sie eine starke Sequenzähnlichkeit besitzen gegenübergestellt. Für die cDNA-Sequenz ist ebenfalls die Länge angegeben, gefolgt von der Anzahl der Basen, die zwischen den beiden Sequenzen identisch sind. Unter »e-val.« ist der e-Value aufgeführt, den ein BLAST-Ergebnis der beiden Sequenzen erreicht.

Kand.	l	cDNA-Seqeunz	l	id. bp	e-Val.	Kand.	l	cDNA-Seq	l	id. bp	e-Val.
a2	1151	VL0377	701	701	0	a16	389	VL0622	762	381	0
a2	1151	VL4412	674	674	0	a16	389	VL3723	729	381	0
a2	1151	VL1287	591	591	0	a16	389	VL1058	703	381	0
a2	1151	VL2273	700	641	0	a16	389	VL3113	472	372	0
a2	1151	VL2949	498	497	0	a16	389	VL1959	570	366	0
a2	1151	VL1214	507	469	0	a16	389	VL3271	752	381	0
a2	1151	VL3778	380	379	0	a16	389	VL2240	541	347	0
a2	1151	VL3889	356	356	0	a16	389	VL0539	783	372	0
a2	1151	VL3191	477	444	0	a16	389	VL2549	561	372	0
a6	1200	VL3927	518	517	0	a16	389	VL4275	807	372	0
a6	1200	VL1651	777	479	0	a16	389	VL1539	745	372	0
a12	1338	VL1696	807	690	0	a16	389	VL1873	647	372	0
a16	389	VL0488	757	381	0	a16	389	VL1041	691	372	0
a16	389	VL3443	618	381	0	a16	389	VL2037	818	372	0
a16	389	VL0588	753	381	0	a16	389	VL2102	499	372	0
a16	389	VL4337	699	372	0						

3.3 Zusammenfassung der Ergebnisse

Die unterschiedlichen Ansätze, mit denen nach den für die Adhäsion von *V. longisporum* relevanten Proteinen gesucht wird, liefern verschiedene Ergebnisse. In einem Vergleich und einer Zusammenstellung dieser Ergebnisse sollen Sequenzen gezeigt werden, die mit mehreren Methoden gefunden werden. Wenn für eine Sequenz in der Datenbank mehrere Versionen und somit auch mehrere Ergebnisse vorhanden sind, werden diese zusammengefasst und einer der Sequenzen zugeordnet. Dabei wird immer das beste Er-

gebnis dargestellt.

3.3.1 Ergebnisse der Kandidaten-Sequenzen

Die Tabelle 3.18 zeigt eine Übersicht der Ergebnisse der einzelnen Kandidaten. Gezeigt sind die Annotation, die längste Ser/Thr-reiche Region, die über zwanzig AS hinausgeht und die Signalsequenzen für den Export ins ER und die GPI-Verankerung. Die Ergebnisse der BLAST-Suche und Annotation sind ausschließlich mit den vollständigen cDNA-Sequenzen erstellt, während die Signalsequenzen und Ser/Thr-reichen Bereiche auch mit den vorhergesagten kodierenden Regionen der cDNA-Sequenzen gesucht wurden. Dabei sind die Ser/Thr-Regionen und Signal-, bzw. Ankersequenzen, die in der AS-Sequenz eines Leserahmens der vollständigen Nukleotidsequenz gefunden wurden, mit einem Stern gekennzeichnet. Die Ergebnisse für die Export-Sequenzen stammen aus der DGPI-Vorhersage, in der sie mit der Suche nach GPI-Ankersignalen verbunden ist.

Während in den BLAST-Ergebnissen vieler Sequenzen Schlüsselwörter gefunden werden, ist die Zahl der vorhergesagten Anker- und Exportsignale gering. Nur für a2 wird sowohl ein GPI-Anker, als auch eine Exportsequenz vorhergesagt. Die Vorhersage stammt von den Programmen DGPI und Big-II und wird nur für die vorhergesagte kodierende Region getroffen. In der vorhergesagten kodierenden Region selbst wird ein 125 AS langer Bereich gefunden, der einen Ser/Thr-Gehalt von 34,92% besitzt. Die Ergebnisse der BLAST-Suche der a2-Sequenz gegen verschiedene Datenbanken enthalten das Schlüsselwort »Adhesin«. Bei der tblastx-Suche gegen die Saccharomyces Genome database wird ein Treffer für das Zellwand-Adhäsin FIG2 gefunden, das besonders bei der Paarung exprimiert wird und an der Intakthaltung der Zellwandstabilität beteiligt ist. Dieser BLAST-Treffer ist mit einem e-Value von 3,3 der vierte in der Trefferliste, jedoch der erste, der im selben Leserahmen liegt, wie die mit ESTScan vorhergesagte kodierende Region. Die gefundene Sequenzähnlichkeit tritt 48 bp vor dem Start der kodierenden Region auf und endet 101 bp hinter dem Methionin, das den Start der als kodierend vorhergesagten Region darstellt. Die Sequenzähnlichkeit besteht zu den letzten AS von FIG2. Derselbe Bereich, der Ähnlichkeiten zu FIG2 aufweist, kommt in einem zweiten Treffer der selben BLAST-Suche vor. Mit einem e-Value von 8,7 wird DEF1 gefunden, ein RNAPII-Degrationsfaktor. Dieser Abschnitt der AS-Sequenz besitzt außerdem eine Ähnlichkeit

Tabelle 3.18: **Überblick über die Ergebnisse, die für einzelne Kandidaten-Sequenzen erhalten werden**

Für jeden Kandidat ist angegeben welche Ergebnisse gefunden werden. Dabei werden Treffer für GPI-Ankersignale abgekürzt. Je nachdem, mit welchem Programm der Treffer gefunden wird, ist ein B (Big-II), ein D (DGPI) oder ein S (GPI-SOM) angegeben. Des weiteren aufgeführt sind: Von DGPI gefundene Exporsignale (Exp-S) und die längste Ser/Thr-reiche Region mit einer Länge von über zwanzig AS. Ergebnisse für Signalsequenzen und Ser/Thr-Region, die mit der vollständigen Sequenz, ohne Berücksichtigung der vorhergesagten kodierenden Region, erhalten werden, sind mit einem *-Zeichen gekenzeichnet.

Kandidat	GPI	Exp-S	Ser/Thr-Region	Annotation	Schlüsselwörter			
a2	DB	+	125	extracellular matrix protein	Adhesin			
a3		-	26/39*	R3H domain-containing protein 2				
a4		-	44/53*	A-agglutinin anchorage subunit precursor	Flocculation	Cell Wall Protein	GPI	Lectin
a5		-	67	Protein sak1	Invasive Growth	Surface Glycoprotein	GPI	
a8	S	-	39/75*	Protein sak1	GPI			
a6		+	107	Capsid protein p87	Flocculation	Invasive Growth	GPI	
a11		-	33*	AT-rich interactive domain-containing protein 1A	Floccilation	Cell Wall Protein	Mucin	
a12		-	60/73*	Cellulase	Flocculation	Surface Glycoprotein	GPI	Invasive Growth
a13	S	-	21/109*	Uncharacterized RNA-binding protein C3H8.09c	Flocculation	Invasive Growth		
a14		-	-	Zinc cluster protein PDR1				
a16		-	82*	Thiazole biosynthetic enzyme, mitochondrial precursor	GPI			
a17		-	59/73*	Salicylaldehyde dehydrogenase	Adhesion	Cell Wall Protein	GPI	
a19		-	98*	Ubiquinol-cytochrome c reductase iron-sulphur subunit	Invasive Growth			
a21		-	24/56*	Subunit of the SWI/SNF complex				
a22		-	53/151*	DNA-binding protein creA	Invasive Growth	Lectin	GPI	
b1		-	57*	Caskin-1				
b3	D*	-	41*	-				
b4		-	47/67*	RNA-binding protein FUS	Mucin			
b5		-	59*	AT-rich interactive domain-containing protein 1A				
b6	D*S*	-	89	Aflatoxin biosynthesis regulatory protein	Flocculin	GPI		
b13		-	69*	Peripheral benzodiazepine receptor-interacting protein	Invasive Growth	Mucin		
b16		-	24*	Synapsin-1	GPI			
b20		-	33*	Serine/arginine repetitive matrix protein 1				
b21		-	24/39*	DNAJ domain protein				

zu dem Protein Muzin-16 der SwissProt-Datenbank. Der entsprechende BLAST-Treffer ist der beste, der mit der Sequenz von a2 in der SwissProt-Datenbank gefunden wird, besitzt jedoch nur einen e-Value von 1,7. Bei der Suche mit blastn gegen die Cogeme-Datenbank wird ein Protein der extrazellulären Matrix, gefunden. Der Sequenzabschnitt, für den diese Ähnlichkeit besteht, ist 20 bp lang und liegt am Ende der als kodierend vorhergesagten Region.

Für den Kandidaten a12 werden mit vier gefundenen Schlüsselwörtern die meisten gefunden. Alle stammen aus zwei Ergebnissen der tblastx-Suche gegen die Saccharomyces Genome Database. Der erste BLAST-Treffer, der ein Schlüsselwort enthält ist für MUC1. Dies ist ein GPI-verankertes Zelloberflächenflockulin, das für haploides invasives Wachstum und diploide Formation von Pseudohyphen. Der e-Value dieses BLAST-Treffers beträgt 0,59, liegt jedoch nicht auf dem gleichen Strang wie die vorhergesagte kodierende Region. Der zweite Treffer, der ein Schlüsselwort enthält liegt zwar auf dem gleichen Strang wie die vorhergesagte kodierende Region, jedoch in einem anderen Leserahmen. Er ist für einen transkriptionalen Repressor, der an der Regulation der Flockulation beteiligt ist.

Neben vielen Ähnlichkeiten zu unbekannten Proteinen besteht auch eine Ähnlichkeit zu der Cogeme-Datenbanksequenz für Zellulase. Das BLAST-Ergebnis hat einen e-Value von 0,038 und liegt in der vorhergesagten kodierenden Region der Kandidatensequenz. Weitere BLAST-Ergebnisse für bekannte Proteine, die den maximalen e-Value nicht überschreiten liegen nur für die Saccharomyces Genome Database vor. Es handelt sich um Gene, deren Produkten an Genregulation, Spleicing und Replikation beteiligt sind.

In den BLAST-Ergebnisssen von a11 werden drei Schlüsselwörter gefunden. Das Wort »Muzin« wird in einem BLAST-Treffer für ein dem menschlichen Muzin ähnlichen Protein gefunden und liegt im LR der vorhergesagten kodierenden Sequenz. Die Schlüsselwörter »Flocculation« und »Cell wall protein« stammen aus BLAST-Ergebnissen der Suche gegen die Saccharomyces Genome Dateabase. Bei den Treffer handelt es sich um die Gene von Flo1, Flo5 und Flo9, die alle an der Flockulation von Hefe beteiligt sind.

Die Ergebnisse, die für die Kandidaten-Sequenzen erstellt wurden, enthalten nicht nur Hinweise auf mögliche Adhäsine. In einigen BLAST-Ergebnissen werden auch vermehrt Sequenzähnlichkeiten zu Transkriptionsfaktoren gefunden. Die Kandidaten, auf

die das zutrifft werden im Folgenen erwähnt.

Die BLAST-Ergebnisse von a22 enthalten, vor allem bei Suchen in SwissProt, nr, Cogeme und der Saccharomyces Genome Datebase eine große Zahl von Treffern für das DNA-bindende Protein creA, für Zinkfingerproteine und Regulationsfaktoren. Diese Treffer liegen im selben LR wie die von ESTScan vorhergesagte kodierende Region.

Es sind nur vier Treffer enthalten, in denen Schlüsselwörter vorkommen: Der erste Treffer stammt aus der tblastx-Suche in der Cogeme-Database und ist, mit einen e-Value von 0,015, für ein Zellulose bindendes Lektin-ähnliches Protein. Dieses Ergebnis steht jedoch erst an 43. Stelle. Für das Schlüsselwort »Invasive Growth« werden zwei BLAST-Ergebnisse, beide aus der tblastx-Suche in der Saccharomyces Genome Database, gefunden. Sie sind für die Proteine TEC1 und RIM101, die beide die Transkription regulieren und für invasives Wachstum nötig sind. Der letzte BLAST-Treffer, der ein Schlüsselwort enthält stammt ebenfalls aus der Saccharomyces Genome Database. Bei dem zugehörigen Protein handelt es sich um eine *Aspartic protease*, die mittels eines GPI-Ankers an Membranen verankert ist.

Die Kandidaten-Sequenz a5 besitzt, ebenso wie a12, Ähnlichkeiten zu MUC1. Diese Ähnlichkeit wird bei der blastn-Suche gegen die Saccharomyces Genome Database gefunden. Jedoch werden bei der BLAST-Suche gegen die SwissProt-Datenbank vor allem Treffer für DNA-bindende Proteine und Transkriptionsfaktoren gefundenen. Auch SAK1, das Protein, zu dem auch die Sequenz a10 eine große Ähnlichkeit aufweist, hat eine regulatorische Funktion (siehe Abschnitt 3.2.6).

Für die Kandidaten-Sequenz a13 wird ein GPI-Ankersignal gefunden, ebenso wie BLAST-Ergebnisse, die die Stichwörter »Flocculation« und »Invasive Growth« enthalten. Eine längere Ser/Thr-reiche Region ist zwar vorhanden, nicht jedoch in der als kodierend vorhergesagten Sequenz. Ein Vergleich der BLAST-Ergebnisse, die in verschiedenen Datenbanken gefunden werden, zeigt, dass vor allem RNA-bindende Proteine gefunden werden. In den Ergebnissen der SwissProt und der nr-Datenbank sind die einzigen Ergebnisse, die den maximalen e-Value nicht überschreiten, für die RNA-bindenden Proteine Nab3, Rbp und Q10145. Die Schlüssselwörter »Flocculation« und »Invasive Growth« werden in den Ergebnissen der BLAST-Suche gegen die Saccharomyces Genome Database gefunden. In beiden Fällen handelt es sich um einen Transkriptionsfaktor, der für die

Regulation des invasiven Wachstums nötig ist. Der eine ist Flo8, der in dem verwendeten Hefestamm BY4741 nicht funktionsfähig vorliegt, der andere MSS11. Beide liegen in der vorhergesagten kodierenden Region der Kandidaten-Sequenz, werden jedoch mit einem hohen e-Value gefunden. Dieser liegt für den BLAST-Treffer Flo8 bei 9,2 und für MSS11 bei 2,3.

Ein weiterer Kandidat, für den BLAST-Treffer vorliegen, die ein Schlüsselwort enthalten, ist b4. In diesem Fall handelt es sich um einen Treffer für ein Vorläuferprotein von Muzin-5B in der SwissProt-Datenbank. Dieser Treffer liegt jedoch nicht in der als kodierendend vorhergesagten Sequenz von b4. Der beste BLAST-Treffer in der SwissProt-Datenbank, der in der vorhergesagten kodierenden Region liegt ist für das RNA-bindende Protein FUS. Ebenfalls werden Sequenzähnlichkeiten zu dem Transkriptionsfaktor IF-2 gefunden. Auch bei den BLAST-Suchen gegen die Cogeme- und die nr-Datenbank werden Treffer für Transkriptionsfaktoren gefunden, die ebenfalls in der als kodierendend vorhergesagten Region von b4 liegen.

Für den Kandidaten b6 wird von dem Programm GPI-SOM ein GPI-Ankersignal gefunden. Dieses Signal ist jedoch nicht Teil der als kodierend vorrausgesagten Region. Die mit tblastx gefundene Sequenzähnlichkeit zu dem GPI-verankerten EGT2, einer Endoglucanase, wird jedoch in dem kodierenden LR gefunden. Der e-Value dieses BLAST-Treffers liegt bei 2,8. Ebenfalls einen e-Value von 2,8 hat der BLAST-Treffer für BSC1, einem Protein, das dem Zelloberflächenflocculin Muc1p ähnelt. Es werden aber wiederum Sequenzähnlichkeiten zu regulatorischen Proteinen und Transkriptionsfaktoren gefunden. Dies ist vor allem für die BLAST-Ergebnisse bei der Suche gegen SwissProt, Cogeme und nr der Fall. Die besten Ergebnisse für die Suche gegen SwissProt sind für die AFLR-Regulationsproteine, die in *Aspergillus* die Biosynthese von Aflatoxin, einem Mycotoxin, regulieren.

Der Kandidat a6 weist eine Ser/Thr-reiche Region mit einer Länge von 107 AS in der als kodierend vorhergesagten Region auf. Zusätzlich wird eine N-terminale Exportsequenz für diese Region vorrausgesagt. Die Schlüsselworte »GPI«, »Flocculation« und »Invasive Growth« werden in den Ergebnissen der BLAST-Suche gegen die Saccharomyces Genome Database gefunden. Bei den entsprechenden BLAST-Treffern handelt es sich um Flo10, einem Lectin-ähnlichem Protein, von dem eine Beeiligung an der Flockulation

vermutet wird. Das Schlüsselwort »GPI « wird in dem BLAST-Treffer für YPS3 gefunden, einer Protease, die mittels eines GPI-Ankers an der Plasmamembran verankert ist. Das einzige Schlüsselwort, das aus einem BLAST-Treffer stammt, der im gleichen LR wie die vorhergesagte kodierende Region liegt, ist »Invasive Growth«. Dieser BLAST-Treffer ist für den Transkriptionsrepressor RIM101, der invasives Wachstum als Antwort auf pH-Wert-Änderungen reguliert. Die Zahl der BLAST-Ergebnisse, in denen ein Schlüsselwort vorkommt, ist jedoch gering. Die meisten Treffer sind für Sequenzen, die in keinem direkten Zusammenhang zur Adhäsion stehen. Es werden vor allem Regulationsproteine und Transkriptionsfaktoren gefunden. Darunter auch der Transkriptionsfaktor AAX73422.1 aus *V. dahliae*, der sowohl in den Treffern der Cogeme-, als auch in denen der nr-Datenbank vorhanden ist.

3.3.2 Ergebnisse der cDNA-Sequenzen

In Tabelle 3.3.2 wird eine Übersicht über die besten Ergebnisse der cDNA-Sequenzen gegeben. Für die cDNA-Sequenzen, die einen Bereich mit hohem Ser/Thr-Gehalt haben, Sequenzen mit einem hohen Scorewert für die Suche nach Schlüsselwörtern in den BLAST-Ergebnissen und einem GPI-Ankersignal werden in der Tabelle die einzelnen Egebnisse zusammengetragen.

Die Sequenz VL3449 erreicht bei den Schlüsselwort-Suchen in den BLAST-Ergebissen für die Cogeme-, die nr- und die SwissProt-Datenbank einen hohen Score. Der Schlüsselwort-Score von VL3449, der für die Ergebnisse der blastn-Suche gegen Cogeme ermittelt wird, ist mit 3,89 der höchste, der für blastn-Suchen in der Cogeme-Datenbank gefunden wird. Die Schlüsselwort-Scores aus den BLAST-Ergebnissen der tblastx-Suche gegen Cogeme und der Suche in der nr-Datenbank sind jeweils die dritthöchsten. In den Schlüsselwort-Ergebnissen für die SwissProt-Datenbank beträgt der Score von VL3449 1,6.

Bei den gefundenen Schlüsselwörtern handelt es sich um die Begriffe »Hydrophobin« und »Zellwandprotein«. Alle BLAST-Ergebnisse der Suche gegen die SwissProt-Datenbank, deren e-Value unter 0,1 liegt, beziehen sich auf Hydrophobin-Vorläuferproteine. Fünf der sechs Treffer beinhalten das aus acht Cysteinen bestehende Muster (siehe Abschnitt 1.4.2), das auch in der durch ESTScan vorhergesagten kodierenden Region von

VL3449 vorhanden ist. Der beste Treffer hat dabei einen e-Value von $9e^{-20}$. Die BLAST-Treffer aus der nr-Datenbank sind nicht für Vorläuferproteine, sondern direkt für die Hydrophobine. Die achzehn besten Treffer, beginnend mit einem e-Value von $4e^{-20}$, sind für Hydrophobine. Bei der BLAST-Suche in Cogeme werden vor allem Treffer für Klasse II Hydrophobine gefunden. Für die als kodierend vorhergesagte Sequenz wird ein GPI-Ankersignal gefunden, das für die Funktion eines Proteins als Hydrophobin jedoch nicht nötig ist.

Für die cDNA-Sequenz VL4432 werden BLAST-Treffer gefunden, in denen das Schlüsselwort »Adhesin« vorkommt. Diese Treffer werden in den Datenbanken SwissProt, nr und Saccharomyces Genome Database gefunden. Weitere Schlüsselwörter, die in den BLAST-Ergebnissen gefunden werden sind: Flockulin, Zellwandprotein, Muzin, GPI und Adhäsion. Das SwissProt-Ergebniss mit dem niedrigsten e-Value ist für die Sequenz von einem Serin-reichen Adhäsin, das für Blutplättchenvorläufer benötigt wird. Dieser BLAST-Treffer liegt in der vorhergesagten kodierenden Region der Sequenz und hat einen e-Value von 70^{-7}. Auch die BLAST-Ergebnisse für die Vorläufer der Zellwandproteine TIR3, CCW14 und AWA1 liegen in der als kodierden vorausgesagten Region. Diese Treffer werden auch in der nr und in der Saccharomyces Genome Database wiedergefunden.

Zusätzlich wird für die Sequenz VL4432 ein Ankersignal von GPI-SOM gefunden. Die längste Ser/Thr-reiche Region ist 145 AS lang und liegt in der als kodierend vorhergesagten Region. Es ist somit eine der längsten Regionen, die für die als kodierend vorrausgesagten Bereiche der cDNA gefunden wird. Die gefundene Ser/Thr-reiche Region reicht bis zu der fünften AS vor dem Ende der Sequenz. Es besteht die Möglichleit, dass ein Teil der Sequenz bei der Sequenzierung nicht erfasst wurde und die kodierend Region somit länger sein könnte.

Die Sequenz, die den höchsten Score für Schlüsselwörter in den BLAST-Ergebnissen für die SwissProt-Datebank erreicht ist VL1789. Der Score liegt bei 11,68. Der BLAST-Treffer mit den niedrigsten e-Value enthält ein Schlüsselwort und ist für den »Sporozoite surface protein 2 precursor«. Eine N-terminale Signalsequenz, die für den Export des Proteins an die Zelloberfläche benötigt wird, wird vom Programm DGPI vorhergesagt. Die weiteren Treffer, die Schlüsselwörter enthalten und im gleichen LR liegen, wie die als kodierend vorhergesagte Region von VL1789, sind für Adhäsine und Vorläufer von

Adhäsinen, sowie für, Muzine, Zellwandproteine und Oberflächenproteine. Jedoch sind unter den 381 BLAST-Treffern, mit einem e-Value der kleiner als 0,1 ist, auch 335 Treffer, die in der vorhergesagten kodierenden Region von VL1789 liegen und kein Schlüsselwort enthalten. Ein Beispeil hierfür ist das »Neurofilament medium polypeptide«. Einige der Treffer besitzen, obwohl sie kein Schüsselwort enthalten, trotzdem einen Bezug zur Oberfächenbeschaffenheit der Zellwand. So wird ein Vorläufer des »Accumulation-associated protein« gefunden. Dieses Protein ist an der Zellwand lokalisiert, wie es auch für Adhäsine der Fall ist.

In der Cogeme-, nr-, und Saccharomyces-Datenabank werden ebenfalls BLAST-Treffer gefunden, die Schlüsselwörter enthalten. Es bestehen Sequenzähnlichkeiten zu dem Adhäsin MAD1 und dem Flockulin MUC1.

Die als kodierend vorhergesagte Region der cDNA-Sequenz VL3377 enthält eine 105 AS lange Ser/Thr-reiche Region. Der Anteil der beiden AS liegt bei 42,45 %. Zusätzlich werden von den beiden Programmen Big-II und DGPI Signalsequenzen für eine GPI-Ankerstelle gefunden. Auch eine N-terminale Exportsequenz ist vorhanden. Das einzige Schlüsselwort wird in den BLAST-Ergebnissen der Suche gegen die Saccharomyces-Datenbank gefunden. Der entsprechende Treffer ist für das Protein FIT3, ein Mannoprotein, das mittels eines GPI-Ankers in der Zellwand verankert ist. Das Protein ist als Adenosyl-Ribosylierungs-Faktor annotiert. Der BLAST-Treffer für dieses Protein stammt aus der Cogeme-Datenbank.

Eine der cDNA-Sequenzen, für die in fünf verschiedenen BLAST-Suchen Ergebnisse gefunden werden, die Schlüsselwörter enthalten, ist VL0203. Bei den Schlüsselwörtern handelt es sich um »Zellwandprotein«, »Oberflächen-Glykoprotein«, »Muzin« und »GPI«. In den BLAST-Ergebnissen, die in der SwissProt-Datenbank gefunden werden, enthält der Treffer mit dem niedrigsten e-Value das Schlüsselwort. Bei diesem Treffer handelt es sich um ein Vorläuferprotein von SED1 aus *S. cerevisiae*, ein Zellwandprotein. SED1 besitzt einen GPI-Anker und trägt zur Stabilisierung der Zellwand bei. Weitere Sequenzähnlichkeiten werden zum Clock-controlled Protein 6 gefunden, das wiederum SED1 ähnelt. In der Datenbank Cogeme wird ein BLAST-Treffer ermittelt, der zum Clock-controlled Protein 6 homolog ist und aus *V. dahliae* stammt. Dieser BLAST-Treffer hat einen e-Value von Null. Alle BLAST-Treffer, die die Schlüsselwörter »Zellwandprotein«,

»Oberfächen-Glycoprotein« und »GPI« enthalten, liegen im selben LR wie die als kodierend vorhergesagte Sequenz. Eine GPI-Ankerstelle wird nicht gefunden, jedoch eine N-terminale Exportsequenz.

Die cDNA-Sequenz VL0879 ähnelt, ebenso wie VL0203, der Datenbank-Sequenz von SED1 und Clock-controlled protein 6. Diese Sequenzähnlichkeit liegt ebenfalls in der vorhergesagten kodierenden Region. Die Proteinregionen, zu denen die cDNA-Sequenzen VL0879 und VL0203 ähnlich sind, überschneiden sich teilweise. Werden jedoch die beiden cDNA-Sequenzen in einer BLAST-Suche direkt miteinander verglichen, so werden keine Sequenzähnlichkeiten gefunden. In der Sequenz von VL0879 werden von den Programmen Big-II und DGPI GPI-Ankersignale gefunden. Da die Sequenzen VL0203 und VL0879 Sequenzähnlichkeiten zu den selben Datenbank-Sequenzen besitzen, ist es möglich, dass es sich um zwei Teile des selben Genes handelt.

Die cDNA-Sequenz VL2256 besitzt ein 176 AS langen Bereich in der vorrausgesagten kodierenden Region, in dem die AS Ser und Thr 40,68 % aller AS ausmachen. Es wird zusätzlich eine Signalsequenz von dem Programm Big-II vorhergesagt. Es werden in der BLAST-Suche keine Treffer ermittelt, die ein Schlüsselwort enthalten. In den Datenbanken SwissProt und nr werden keine BLAST-Treffer ermittelt. In der Cogeme-Datenbank werden zwei Treffer gefunden, deren e-Value bei 0,63 liegt. Es handelt sich um Treffer für die DNA-Sequenzen einer Glucosamine-Fructose-6-Phosphate Aminotransferase und eines Kinesins.

Die cDNA-Sequenz mit dem vierthöchsten Schlüsselwort-Score für die BLAST-Suche in der nr-Datenbank ist VL1444. Das einzige Schlüsselwort, das dabei gefunden wird ist GPI. Es werden in allen Datenbanken BLAST-Treffer gefunden, die dieses Schlüsselwort enthalten. Die einzelnen Treffer sind für GPI-Transamidasen oder Komponenten von GPI-Transamidasen. Diese Sequenz hat keine Ähnlichkeit zu Adhäsinen, jedoch ist die GPI-Transamidase ein Protein, das an der Prozessierung der Adhäsine beteiligt ist.

Die Sequenz VL3873 hat einen Schlüsselwort-Score für die Cogeme-Datenbank von Eins. Der BLAST-Treffer ist für ein Zell-Adhäsionsprotein. Alle BLAST-Treffer, die für diese cDNA-Sequenz in der Cogeme und nr-Datenbank gefunden werden und in der vorhergesagten kodierenden Region der cDNA-Sequenz liegen, sind Treffer für ein Adhäsionsprotein oder für Fasciclin, ein Adhäsin. In der SwissProt-Datenbank werden keine

Treffer gefunden, deren e-Values unter dem Wert von 0,1 liegen. In der Saccharomyces Genome Database werden weitere Treffer gefunden, die im gleichen LR liegen wie die vorhergesagte kodierende Region. Keiner der Treffer enthält ein Schlüsselwort. Es wird ein Exportsignal, jedoch kein Ankersignal und auch keine Ser/Thr-reiche Region gefunden.

Für die Sequenz VL3130 werden von den Programmen DGPI und GPI-SOM Ankersignale gefunden. Ebenfalls vorhanden ist ein Exportsignal. In fünf der BLAST-Suchen werden Treffer gefunden, die Schlüsselwörter enthalten. Der Treffer mit dem niedrigsten e-Value der BLAST-Suche in der SwissProt-Datenbank ist für einen 1,3-beta-Glucanosyltransferase Gel1-Vorläufer. Der e-Value liegt für diesen Treffer bei 80^{-29}. Den nächsthöheren e-Value von 20^{-24} erreicht der Treffer für einen Vorläufer eines Glycolipid-verankerten Oberflächenproteins. Treffer für diese Proteine werden ebenfalls bei der Suche in der nr-Datenbank gefunden. Es besteht ebenfalls eine Sequenzähnlichkeit zu der Sequenz der beta-1,3-Glucanosyltransferase aus *V. dahliae*. Dieser BLAST-Treffer wird bei der Suche in der Cogeme-Datenbank gefunden. Das Schlüsselwort »Oberflächen-Glykoprotein« wird in den BLAST-Ergebnissen nur im Zusammenhang mit der Glucanosyltransferase gefunden. Auch bei der BLAST-Suche in der Saccharomyces Genome Database ist der Treffer mit dem niedrigsten e-Value für die Glucanosyltransferase. Es werden die Schlüsselwörter Adhäsin und GPI gefunden, wobei die entsprechenden BLAST-Treffer für die Glucanosyltransferase, GPI2, ein Protein, das an der Synthese der GPI-Anker beteiligt ist, und für Fig2, ein Adhäsin sind. Der Treffer für Fig2 hat einen e-Value von 3,8.

Der zweitbeste Schlüsselwort-Score für die BLAST-Ergebnisse der SwissProt-Datenbank wird mit einem Wert von 9,8 von der Sequenz VL2977 erreicht. Der BLAST-Treffer mit dem niedrigsten e-Value ist für die Sequenz von SSP2, dem Vorläufer eines Sporozoiden Oberflächenproteins. Die Sequenzähnlichkeit liegt innerhalb der vorhergesagten kodierenden Region der cDNA-Sequenz. Es werden vier BLAST-Treffer gefunden, die das Schlüsselwort »Adhäsin« enthalten. Mit einem e-Value von 80^{-06} wird das Vorläuferprotein des Adhäsins PAc gefunden. Die weiteren Treffer sind für die Adhäsine CNA und Zonadhäsine aus Mensch und Maus. Es werden Ähnlichkeiten zu der Proteinsequenz von FP1, einem adhäsiven Matrixproteins aus den Muscheln *Mytilus galloprovincialis* und *Mytilus edulis* gefunden. Weitere Schlüsselwörter in den BLAST-Ergebnissen sind Mu-

zin, Zellwandprotein und Oberflächen-Glycoprotein. Dabei liegen nur die zwei Treffer für das Oberflächenprotein MSP1 liegen nicht im selben LR wie die vorhergesagte kodierende Sequenz. Der BLAST-Treffer mit dem niedrigsten e-Value, der bei der Suche in der Cogeme-Datenbank gefunden wird, ist für die »Mixed-linked Glucanase« aus *V. dahliae*. Dieser Treffer liegt in der als kodierend vorhergesagten Region von VL2977. Der e-Value dieses Treffers liegt bei 80^{-51}. Es werden jedoch auch Sequenzähnlichkeiten zu Oberflächen-Glycoproteinen gefunden. Für die als kodierend vorhergesagte Region von VL2977 werden weder Signalsequenzen, noch Bereiche mit einem hohen Anteil von Ser und Thr gefunden. Ein LR der cDNA-Sequenz, jedoch nicht der als kodierend vorhergesagte, weist eine höhere Menge an Ser und Thr auf. Da jedoch die BLAST-Treffer, die Schlüsselwörter enthalten, in der kodierenden Region ligen, lässt sich zwischen diesen und den Ser/Thr-reichen Bereich kein Zusammenhang herstellen.

VL0081 besitzt eine vorhergesagte kodierende Region, für die sowohl ein GPI-Ankersignal, als auch eine Ser/Thr-reiche Region gefunden wird. In den Ergebnissen der BLAST-Suche ist jedoch kein Schlüsselwort vorhanden. Es werden Treffer für eine Glycosidase gefunden, die im gleichen LR wie die als kodierend vorhergesagte Region liegen. Es existiert eine Sequenzähnlichkeit zu einer Glycosidase der Zellwand aus *V. dahliae*. Die vorhergesagte GPI-Verankerung dieses Proteines wird also durch das Ergebnis der BLAST-Suche unterstützt.

Tabelle 3.19: **Übersicht über die Ergebnisse der cDNA-Sequenzen**

Die Tabelle zeigt die Ergebnisse, die für einzelne cDNA-Sequenzen gefunden werden. Es sind angegeben: die ID der Sequenz, die vorhandenen GPI-Ankersignale, die vorhandenen Exportsignale, die längste Ser/Thr-reiche Region (bei einer Länge von mehr als 100 AS) und Schlüsselwörter, die in der BLAST-Suche vorkommen. Die BLAST-Ergebnisse wurden dabei für die vollständige cDNA-Sequenz ermittelt. Die Ergebnisse für die Signalsequenzen und für die Ser/Thr-reiche Region wurden sowohl für die vollständigen Sequenzen, als auch für die als kodierend vorhergesagten Regionen ermittet. Dabei sind die Ergebnisse, die nur für die vollständigen Sequenzen, nicht jedoch für die kodierenden Regionen gefunden werden, durch ein *-Symbol gekennzeichnet.

ID	GPI-Anker	Exp-S	Ser/Thr-Region	Schlüssewörter					
VL3449	D	+	-	Cell Wall Protein	Hydrophobin				
VL4432	S	-	145	Cell Wall Protein	Adhesion	GPI	Mucin	Adhesin	Flocculin
VL1789		+	-	Cell wall protein	Surface glycoprotein	GPI	Mucin	Adhesin	
				Invasive Growth	Surface protein	Lectin	Adhesive		
VL3377	BD	+	105	GPI					
VL0203	-	+	-	Cell Wall Protein	Surface Glycoprotein	GPI	Mucin		
VL0879	BD	+	-	Cell Wall Protein	Surface Glycoprotein	GPI			
VL2256	B	-	176	-					
VL1444	-	-	-	GPI					
VL3873	-	+	-	Adhesion					
VL3130	SD	+	-	Cell Wall Protein	Surface Glycoprotein	GPI	Adhesin	Surface Protein	
VL2977	-	-	117*	Cell Wall Protein	Surface Glycoprotein	Mucin	Surface Protein	Adhesin	Adhesive
VL0081	B	-	104	-					
VL1687	B	-	102	GPI					

Kapitel 4

Diskussion

Das Ziel der Arbeit war es, die cDNA-Sequenzen von *V. longisporum* zu bereinigen, eine erste Annotation vorzunehmen und die Sequenzen zu finden, die die Anhaftung des Pilzes an Wirtspflanzen fördern.

Die cDNA-Sequenzen enthielten zu Beginn der Arbeit Regionen, die aus dem Plasmid stammen, in dem die cDNA-Sequenzen gelagert wurden. Diese Bereiche, die nicht zum Genom von *V.longisporum* gehören, mussten aus den Sequenzen entfernt werden, damit die weiteren Ergebnisse nicht verfälscht werden konnten. Sequenzen, deren Länge zwanzig bp nicht überstieg, wurden verworfen.

Auch die Sequenzen, die in den Adhäsions-Screens gefunden wurden, wurden von der Plasmidsequenz bereinigt. Die danach kürzeste Sequenz, a14, ist nur 27 bp lang. In der ursprünglichen cDNA-Sequenz, die eine Länge von 1.358 bp hat, wurden zwei Regionen von MCS-ferner Plasmidsequenz gefunden. Da in einem solchen Fall alle downstream der Plasmidsequenz gelegene cDNA-Sequenz verworfen wird, kommt es zu der geringen Länge von 27 bp. Da nur dieser Teil der Sequenz untersucht wird, werden für diesen Kandidaten nicht mehr viele Ergebnisse erwartet.

Im Rahmen dieser Arbeit wurden die cDNA-Sequenzen mit Hilfe einer BLAST-Suche miteinander verglichen. Sequenzen mit starker Ähnlichkeit zueinander konnten in Gruppen zusammengefasst werden. Für Sequenzen, die aus einer Gruppe stammen, werden bei nachfolgenden Analysen ähnliche oder identische Ergebnisse erzielt. Durch das Entfernen der mehrfach vorhandenen Sequenzen würden auch die mehrfach vorhandenen

Ergebnisse beseitigt werden, sodass diese sich besser vergleichen ließen. Dieser Schritt wurde im Rahmen der Arbeit noch nicht vorgenommen, da durch das Entfernen eines Teiles der Sequenzen auch Informationen verloren gehen können. Dies würde beim Entfernen sich überschneidender Sequenzen geschehen. Auch bei sich überschneidenden, nicht vollständig identischen, cDNA-Sequenzen ist nicht bekannt, ob und welche dieser Sequenzen Sequenzierfehler enthalten. Auch möglich ist ein mehrfach vorhandenes Allel für dieses Gen in *V. longisporum*. In diesem Fall sind die Unterschiede in der cDNA-Sequenz auf sich leicht unterscheidende Gene zurückzuführen. Damit möglichst wenig Sequenzinformationen verloren gehen, wurden noch keine cDNA-Sequenzen aus der Datenbank entfernt.

Für die Analyse der AS-Sequenzen mussten die cDNA-Sequenzen in AS-Sequenz umgeschrieben werden. Dabei wurde zum Einen die gesamte cDNA-Sequenz in sechs verschiedene Leserahmen übersetzt, zum Anderen wurde nur die AS-Sequenz verwendet, die aus der vorhergesagten kodierenden Region der cDNA-Sequenz stammt. Bei keinem der Ansätze wurde berücksichtigt, dass im mitochondriellen Genom von *Verticillium ssp.* das Kodon UGA kein Stopp-Kodon ist, sondern als zusätzliches Kodon für Tryptophan verwendet wird [36]. Ob dies nur für das mitochondrielle Genom der Fall ist, oder ob sich die Ergebnisse auf das chromosomale Genom übertragen lassen, ist noch nicht geklärt. Die LR der cDNA-Sequenzen könnten unter Umständen also eine zu hohe Zahl an Stopp-Kodons enthalten. Die Trainingssequenzen, die für das Training von ESTScan verwendet wurden, enthalten Sequenzen aus einer Vielzahl von Organismen. Für den Großteil der Organismen wird UGA jedoch als Stopp-Kodon verwendet. Dadurch kann möglicherweise auch die Vorhersage der kodierenden Region mit Fehlern behaftet sein, wodurch die richtige kodierende Region nicht vorhergesagt werden kann. Die Ergebnisse, die für die AS-Sequenzen gefunden wurden, sind aus diesem Grund im Weiteren zu überprüfen um auszuschließen, dass Fehler in der AS-Sequenz die Ergebnisse beeinflussen.

Die in der Literatur [51] angegebene Länge der Ser/Thr-reichen Region von 300 AS wird in den cDNA-Sequenzen von *V. longisporum* nicht gefunden. In den meisten Fällen sind die Nukleotidsequenzen zu kurz, um für eine Region dieser Länge zu kodieren. Die längste cDNA-Sequenz mit 954 bp kann maximal für ein 318 AS langes Polypeptid

kodieren. Zusätzlich wurde das Kodon UGA als Stopp-Kodon übersetzt. Sollte es in *V. longisporum* jedoch für Tryptophan kodieren, ist die Länge der AS-Sequenzen dadurch zusätzlich verkürzt.

Es werden AS-Sequenzen mit Ser/Thr-reichen Bereichen gefunden, die kürzer als 300 AS sind. Dies kann ein Hinweis auf ein Adhäsin sein, da die cDNA-Sequenz nicht immer das ganze Gen enthält. Wird eine Mindestlänge von 250 AS erwartet, erfüllen fünfzehn Sequenzen diese Bedingung. In den als kodierend vorhergesagten Regionen werden nur Sequenzen mit einer Länge von weniger als 200 AS gefunden. 24 Sequenzen enthalten Ser/Thr-reiche Bereiche, die länger als 100 AS sind. Die fünfzehn Sequenzen, in deren LR ein Ser/Thr-reicher Bereich liegt, der länger als 250 AS ist, sind in diesen 24 Sequenzen jedoch nicht enthalten. Sie kommen also nicht in LR vor, die als kodierend vorhergesagt werden.

Bei der Suche nach GPI-Ankersignalen in den AS-Sequenzen werden, abhängig vom verwendeten Programm, unterschiedliche Ergebnisse erhalten. Bereits die Anzahl der gefundenen Ankersignale stimmt nicht miteinander überein. Es muss jedoch berücksichtigt werden, dass den Analysen jeweils eine andere Zahl an zu durchsuchenden Sequenzen zugrunde lag. Die größte Differenz bestand zwischen den Eingabesequenzen von Big-II und DGPI bei der Suche in den LR der cDNA-Sequenzen. DGPI durchsuchte 941 Sequenzen mehr als Big-II, da DGPI auch kürzere Eingabesequenzen akzeptiert. Des Weiteren sucht nur DGPI nach N-terminalen Exportsequenzen. Diese werden bei den Suchen mit Big-II und GPI-SOM nicht berücksichtigt und müssen in weiteren Analysen getrennt ermittelt werden. Die Sensitvität und Spezifität der Programme ist ebenfalls unterschiedlich. Da eine große Zahl von Sequenzen vorhanden ist, die durchsucht werden sollen, ist vor allem eine geringe Zahl von falsch positiven Ergebnissen wichtig. Dies wird von Big-II gewährleistet. Die Rate der falsch positiven Ergebnisse liegt bei 0,1 %. Zusätzlich sollte jedoch auch DGPI verwendet werden, um in den Ergebnissen von Big-II die Sequenzen zu finden, die zusätzlich eine N-terminale Transportsequenz besitzen. Die Ergebnisse von DGPI und GPI-SOM sind zu zahlreich, als das eine experimentelle Überprüfung jeder Vorhersage möglich wäre.

Obwohl cDNA-Sequenzen relativ kurz sind und dadurch nicht immer die vollständige kodierende Region enthalten, wurden in der Arbeit GPI-Ankersignale gefunden.

Die wenigsten Übereinstimmungen gibt es dabei zwischen den Ergebnissen von Big-ΠI und GPI-SOM. Nur für eine AS-Sequenz wird von beiden Programmen ein Ankersignal gefunden. Dies entspricht 0,32 % der GPI-SOM und 1,19 % der Big-ΠI-Ergebnisse (Ascomycetes). Die größte Ähnlichkeit besteht zwischen den Vorhersagen von DGPI und Big-ΠI. Es werden 23,73 % der von DGPI vorhergesagten GPI-Ankerstellen auch von Big-ΠI gefunden. Das entspricht 33,73 % der Ergebnisse von Big-ΠI (Sordariomycetes).

Da die Signalsequenz am C-terminalen Ende eines GPI-verankerten Proteins liegen muss, hat die Vollständigkeit der Sequenz einen großen Einfluss auf das Ergebnis der Vorhersage. Ist die cDNA-Sequenz unvollständig sequenziert, sodass das C-terminale Ende der AS-Sequenz der cDNA nicht der des tatsächlichen Proteins entspricht, wird der falsche Bereich der cDNA zur Vorhersage genutzt. Unter diesen Bedingungen würde in vivo kein GPI-Anker angefügt, da die Signalsequenz nicht am C-Terminus liegt. Für Sequenzen, die als Kandidaten für Adhäsine vorgeschlagen werden, sollten aus diesem Grund zu weiteren Untersuchungen möglichst verlässliche cDNA-Daten vorliegen. Eine gezielte Überprüfung der cDNA-Sequenzen, für die eine GPI-Anker-Bindestelle gefunden wurde, ist ebenfalls nötig.

Es existieren keine cDNA-Sequenzen in deren LR sowohl ein GPI-Anker, als auch eine mehr als 250 AS lange Ser/Thr-reiche Region gefunden wird. Werden nur die als kodierend vorhergesagten Regionen betrachtet, so werden für sechs Sequenzen ein Ser/Thr-reicher Bereich mit einer Länge von über 100 AS und ein GPI-Ankersignal von DGPI und Big-ΠI gefunden. Vier dieser cDNA-Sequenzen besitzen starke Sequenzähnlichkeiten zu der Kandidatensequenz a2. Die zwei weiteren cDNA-Sequenzen VL3377 und VL4201 sind wiederum ähnlich zueinander. Drei weitere Sequenzen besitzen neben der Ser/Thr-reichen Region ein von Big-ΠI vorhergesagtes Ankersignal. Es handelt sich um VL2256, VL1687 und VL0081. Nur eine vorhergesagte kodierende Region hat sowohl eine Ser/Thr-reiche Region, als auch ein von GPI-SOM gefundenes Ankersignal. Diese Vorhersage wird nur für die als kodierend vorhergesagte Region getroffen, die unter Verwendung des Ascomycota-Trainingssets gefunden wurde. Von 24 als kodierend vorhergesagten Regionen, die einen Ser/Thr-reichen Bereich enthalten, werden für sechs von zwei Programmen und für vier von einem Program auch GPI-Ankerstellen gefunden.

Die BLAST-Suche, die in verschiedenen Datenbanken durchgeführt wurde, lieferte

für einen großen Teil der cDNA-Sequenzen Ergebnisse (siehe Abschnitt 3.1.6). Bei den Suchen in der Saccharomyces Genome Database wurden für zwanzig (blastn), bzw. 78 (tblastx) cDNA-Sequenzen keine BLAST-Ergebnisse gefunden, deren e-Values unter dem festgelegten Maximalwert von zehn liegen. Die BLAST-Ergebnisse, die bei der Suche gegen SwissProt, nr und Cogeme erhalten wurden, sind durch den jeweils gewählten Maximalwert stärker eingeschränkt. Möglicherweise wurden so zu viele BLAST-Treffer der Suche gegen die Saccharomyces Genome Database mit einbezogen. In weiteren Untersuchungen der BLAST-Ergebnisse sollte ein größerer Teil der Treffer ausgeschlossen werden, indem der maximal akzeptierte e-Value für die BLAST-Suchen in der Saccharomyces Genome Database verringert wird. Dadurch könnten nur die besten BLAST-Treffer berücksichtigt werden.

In einer großen Zahl der BLAST-Ergebnisse wurden Schlüsselwörter gefunden. Dabei sind nicht alle BLAST-Treffer, die Schlüsselwörter enthalten, Adhäsine, Flockuline oder Hydrophobine. So handelt es sich bei den BLAST-Treffern für die cDNA-Sequenz VL222, für die der höchste Schlüsselwort-Score erreicht wird, um die Glucose-6-Phosphat Isomerase, die mit GPI abgekürzt wird. Ebenfalls werden Ähnlichkeiten zu Proteinen gefunden, die nicht direkt an der Adhäsion beteiligt sind, diese aber regulieren. Ein Beispiel dafür ist VL0761, das den fünfthöchsten Schlüsselwort-Score bei der BLAST-Suche in der nr-Datenbank erhält. Die BLAST-Treffer, die das Schlüsselwort GPI enthalten, sind in diesem Fall für *PIG-F*, ein GPI-Anker Biosynthesprotein. Durch den Einfluss dieses Proteins auf die Syntese von GPI-Ankern nimmt es gleichzeitig auch Einfluss auf Adhäsine. Ein Vergleich mit den Ergebnissen der Suche nach GPI-Ankersignalen, Exportsequenzen oder Ser/Thr-reichen Bereichen ist dann jedoch nicht nötig, da diese Eigenschaften der AS-Sequenz vor allem für Adhäsine erwartet werden. In weiteren Arbeiten könnten die BLAST-Ergebnisse, die ein Schlüsselwort enthalten, in verschiedene Gruppen geteilt werden, um gezielter die Sequenzen zu finden, die Adhäsinen ähneln. Dazu könnten BLAST-Treffer, die sowohl ein Schlüsselwort, als auch Begriffe wie »Biosynthesis« oder »Regulationsprotein« enthalten, in eine der Gruppen eingeordnet werden. Die zweite Gruppe würde dann nur die BLAST-Ergebnisse enthalten, in denen keine Begriffe vorkommen, die im Zusammenhang mit der Regulation stehen. Ein Vergleich der BLAST-Ergebnisse mit den Ergebnissen der Suche nach Ser/Thr-reichen Bereichen oder GPI-

Anker-Bindestellen müsste dann nur für diese Gruppe durchgeführt werden.

Durch den Vergleich der BLAST-Ergebnisse der Kandidaten konnte festgestellt werden, dass für die Kandidaten a5 und a10 identische BLAST-Treffer gefunden werden. Da sich die Sequenzen von a10 und a5 teilweise überlappen, ist es möglich, dass es sich um cDNA-Sequenzen der gleichen mRNA handelt, wobei unterschiedliche Regionen sequenziert wurden. Ein Beispiel dafür ist der BLAST-Treffer für sak1, der bei der Suche in der SwissProt-Datenbank gefunden wurde. Bei sak1 handelt es sich um einen transkriptionalen Repressor. a10 besitzt Sequenzähnlichkeiten zum vorderen Bereich des Proteines, a5 zum hinteren.

Zwischen den BLAST-Ergebnissen der anderen Kandidaten konnten keine starken Ähnlichkeiten gefunden werden. Auch bei dem direkten Vergleich der Kandidaten-Sequenzen mittels einer BLAST-Suche werden keine Ähnlichkeiten zwischen den Sequenzen gefunden, die größer sind als die zwischen a5 und a10. Eine Gruppierung der Kandidaten-Sequenzen ist aufgrund dieser großen Unterschiede nicht sinnvoll. Somit kann aus den Sequenzen der Kandidaten kein Sequenzmuster abgeleitet werden, mit dem eine Datenbank-Suche möglich ist.

Die Kandidaten-Sequenzen wurden durch zwei verschiedene Screens gefunden (siehe Abschnitt 1.6.3). Zum Einen wurde ein Δflo8-BY4741-Stamm benutzt, zum Anderen ein Δflo8/11-Stamm. Für die Kandidaten des Δflo8-Screens werden sowohl Transkriptionsfaktoren, also auch Adhäsionsproteine erwartet. Die Transkriptionsfaktoren könnten die Funktion von Flo8 erfüllen und die Expression von Flo11 wieder herstellen. Adhäsine können die Funktion von Flo11 übernehmen und die Adhäsion direkt wieder herstellen. Für die Kandidaten des Δflo8/11-Screens werden nur Sequenzen erwartet, deren Genprodukte die Adhäsion direkt wieder herstellen.

Obwohl für die Kandidaten des Δflo8/11-Screens keine Transkriptionsfaktoren erwartet werden, sind die Annotationen für b2, b4 und b6 für Proteine die dazu in der Lage sind DNA zu binden. Dabei wird der Kandidat b2 sowohl in den Kandidaten des Δflo8, als auch in den Kandidaten des Δflo8/11-Screens gefunden. Unter den Kandidaten des Δflo8/11-Screens befinden sich acht, die Sequenzähnlichkeiten zu Proteinen aufweisen, die an DNA binden können.

Kapitel 5

Zusammenfassung

Das Ziel, aus den cDNA-Sequenzen von *V. longisporum* eine bereinigte cDNA-Datenbank zusammenzustellen, wurde im Rahmen dieser Arbeit erreicht. Die Plasmidsequenz wurde von den cDNA-Sequenzen getrennt und entfernt. Während die cDNA-Sequenzen ursprünglich eine Länge von insgesamt 2.706.725 bp erreichten, waren es, nachdem die Plasmidsequenz entfernt und kurze cDNA-Sequenzen ausgeschlossen wurden, nur noch 2.659.599 bp.

Innerhalb der cDNA-Sequenzen konnten verschiedene Gruppen ähnlicher Sequenzen gefunden werden. Es wurden 2.722 verschiedene Gruppen gebildet, wobei 541 Gruppen mehrere Sequenzen enthalten, die sich stark ähneln. Das Entfernen von vollständig identischen Sequenzen muss noch erfolgen.

Bei der Analyse der AS-Sequenzen wurde ein Stopp-Kodon verwendet, das in *V. longisporum* unter Umständen nicht verwendet wird. Aus diesem Grund könne weitere Untersuchungen der AS-Sequenzen nötig werden. Die kodierenden Regionen, die auf Basis der verschiedener Trainingssequenzen der Sordariomycetes und Ascomycota gefunden wurden, ähneln sich stark. Dadurch gleichen sich auch die Ergebnisse, die mit den vorhergesagten kodierenden Regionen erzielt werden.

In einigen der AS-Sequenzen wurden GPI-Ankersignale gefunden. Dabei decken sich die Vorhersagen der drei verwendeten Programme jedoch nicht. Für weitere Untersuchungen sollte nur das Programm Big-II verwendet werden, da aufgrund der hohen Zahl an durchsuchten Sequenzen vor allem eine hohe Sensitivität nötig ist. Sequenzen, für die

GPI-Ankerstellen vorhergesagt wurden, müssen zusätzlich nach N-terminalen Exportsequenzen durchsucht werden.

Die vorhergesagten kodierenden Regionen, für die die längsten Ser/Thr-reichen Regionen gefunden werden, enthalten zu einem großen Teil auch Signalsequenzen für die GPI-Verankerung. Insgesamt werden jedoch nur wenige Ser/Thr-reiche Bereiche gefunden, die die Länge von 250 AS überschreiten. Werden nur die als kodierend vorhergesagten Regionen betrachtet, so ist keiner diese Bereiche länger als 200 AS. Da die Sequenzen, die durchsucht wurden, jedoch durchschnittlich nur 171 AS lang sind, wird auch keine große Menge an Sequenzen erwartet, deren Ser/Thr-reicher Bereich länger als 250 AS ist. Diese Methode ist für das Auffinden von möglichen Adhäsinen in einer cDNA-Datenbank nur bedingt geeignet, da cDNA-Sequenzen nicht die gesamte Länge eines Gens abdecken und lange Bereiche so meist nicht vollständig vorhanden sind.

Mit Hilfe der verschiedenen Ansätze, mit denen nach cDNA-Sequenzen gesucht wurde, die einen Einfluss auf die Adhäsion haben, konnten verschiedene Sequenzen gefunden werden. Neben anderen wurden auch Sequenzen gefunden, die homolog zum Kandidaten a2 sind. Dieser Kandidat würde also sowohl experimentell, durch die Adhäsions-Screens gefunden, als auch durch die Analyse der cDNA-Datenbank mit bioinformatischen Methoden.

Literaturverzeichnis

[1] ALTSCHUL S. F., GISH W., MILLER W., MYERS E.W., LIPMAN D. J. 1990. Basic local alignment search tool. *Mol.Biol.* **215**: 403-410.

[2] RICHTLINIE 2003/30/EG DES EUROPÄISCHEN PARLAMENTES UND DES RATES 2003. Förderung der Verwendung von Biokraftstoffen oder anderen erneuerbaren Kraftstoffen im Verkehrssektor. *http://209.85.135.104/search?q=cache:hTwWr2c7b3AJ:ec.europa.eu* **L 123/42 Amtsblatt der Europäischen Union**: 17.5.2003.

[3] ARCHAMBAULT J., FRIESEN J. D. 1993. Genetics of Eukaryotic RNA Polymerases I, II, and III. *Microbiological Reviews* **57(3)**: 703-724.

[4] BOIROCH A., BOECKMANN B. 1991. The SWISS-PROT protein sequence data bank. *Nucleic Acids Research* **19**: 2247-2249.

[5] BURGE C., KARLIN S. 1997. Prediction of complete gene structures in human genomic DNA. *Mol. Biol.* **268**: 78-94.

[6] STATISTISCHES BUNDESAMT DEUTSCHLAND 2004. Landwirtschaft in Zahlen 2003. *http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Presse/pk/2004/Landwirtschaft/Landwirtschaft__04*.

[7] STATISTISCHES BUNDESAMT DEUTSCHLAND 2005. Anbaufläche von Weizen und Raps steigt. *http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Presse/pm/2005/05/PD05__236__411* **Pressemitteilung Nr. 236**: 27.05.2005.

[8] STATISTISCHES BUNDESAMT DEUTSCHLAND 2006. Aussaatflächen 2006: Anbau von Raps und Wintergerste steigt. *http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Presse/pm/2006/05/PD06__212__411* **Pressemitteilung Nr. 212**: 26.05.2006.

[9] STATISTISCHES BUNDESAMT DEUTSCHLAND 2007. Aussaatflächen 2007: Weiterer Rückgang bei Sommerkulturen. *http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Presse/pm/2007/05/PD07__216__411* **Pressemitteilung Nr. 216**: 25.05.2007.

[10] STATISTISCHES BUNDESAMT DEUTSCHLAND 2008. Herbstaussaaten 2007: Mehr Wintergetreide, weniger Raps. *http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Presse/pm/2008/01/PD08__014__412* **Pressemitteilung Nr. 014**: 14.01.2008.

[11] DOUGLAS L. J. 2003. Candida biofilms and their role in infection. *Trends Microbiol.* **11(1)**: 30-36.

[12] DRANGINIS A. M., RAUCEO J. M., CORONADO J. E., LIPKE P. N. 2007. A Biochemical Guide to Yeast Adhesins: Glycoproteins for Social and Antisocial Occasions. *Microbiology and Molecular Biology Reviews* **71**: 282-294.

[13] EBBOLE D.J. 1997. Hydrophobins and fungal infrction of plants and animals. *Trends in Microbiology* **5**: 405-408.

[14] EISENHABER B., SCHNEIDER G., WILDPANER M., EISENHABER F. 2003. A Sensitive Predictor for Potential GPI Lipid Modification Sites in Fungal Protein Sequences and its Application to Genome-wide Studies for *Aspergillus nidulans*, *Candida albicans*, *Neurospora crassa*, *Saccharomyces cerevisiae* ans *Schizosaccharomyces pombe*. *Journal of Molecular Biology* **337**: 243-253.

[15] PRESSEMITTEILUNG DER EUROPÄISCHEN UNION 2007. Getreide: Rat genehmigt Stilllegungssatz von Null für die Aussaat von Herbst 2007 und Frühjahr 2008. *http://europa.eu/rapid/pressReleasesAction.do?reference=IP/07/1402* **Pressemitteilung IP/07/1402**: 26.09.2007.

[16] EYNCK C., KOOPMANN B., GRUNEWALDT-STOECKER G., KARLOVSKY P., VON TIEDEMANN A. 2006. Differential interactions of *Verticillium longisporum* and *V. dahliae* with *Brassica napus* detected with molecular and histological techniques. *European Journal of Plant Pathology* **doi:10.1007/s10658-007-9144-6**: (in press).

[17] FRANKHAUSER N., MÄSER P. 2005. Identification of GPI anchor attachment signals by a Kohoen self-organizing map. *Bioinfrmatics* **21(9)**: 1846-1852.

[18] GUO B, STYLES C. A., FENG Q. FINK G R. 2000. A *Saccharomyces* gene family involved in invasive growth, cell-cell adhesion, and mating. *PNAS* **97**: 12158-12163.

[19] HERSHEY A.D., CHASE M. 1952. Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. *The Gernal of General Physiology* **36**: 39-56.

[20] HOYER L. L., PAYNE T. L., HECHT J. E. 1998. Identification of Candida albicans ALS2 and ALS4 and Localization of Als Proteins to the Fungal Cell Surface. *Journal of Bacteriology* **180**: 5334-5343.

[21] HYNES R. O. 1987. Integrins: a family of cell surface receptors. *Cell* **48(4)**: 549-554.

[22] ISELI C., JOGENEEL V., BUCHER P. 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol.* 138-48.

[23] JOHANSSON A., GOUD J.-K. C.; DIXELIUS C. 2005. Plant host range of *Verticillium longisporum* and microsclerotia density in Swedish soils. *European Journal of Plant Pathology* **114**: 139-149

[24] KARAPAPA V.K., BAINBRIGE B.W., HEALE J.B. 1997. Morphological and molecular characterization of *Verticillium longisporum* comb. nov., pathogenic to oilseed rapey *Mycological Research* **101(11)**: 1281-1294.

[25] KOBAYASHI O., HAYASHI N., KUROKI R., SONE H. 1998. Region of Flo1 Proteins Responsible for Sugar Recognition. *Journal of Bacteriology* **180**: 6503-6510.

[26] KOHONEN T. 2001. Self-Organizing Maps. 3. Aufl. Berlin: Springer Series in Information Sciences.

[27] KOMMISSION DER EUROPÄISCHEN GEMEINSCHAFTEN 2007. Fortschrittsbericht Biokraftstoffe. *http://209.85.135.104/search?q=cache:wGwElB-NtE8J:ec.europa.eu* **Mitteilung der Kommission ab den Rat und das Europäische Parlament**: KOM(2006) 845.

[28] KOZAK M. 1989. The Scanning Model for Translation: An Update. *The Journal of Cell Biology* **108**: 229-241.

[29] KRONEGG J., BULOZ D. 1999. Detection/prediction of GPI cleavage site (GPI-anchor) in a protein (DGPI). *Retrieved [retrieve date] from http://129.194.185.165/dgpi/*.

[30] LEVENE P. A. 1919. The Structure of Yeast Nucleic Acid. *The Journal of Biological Chemistry* **40**: 415-424.

[31] LIPKE P. N., KURJAN J. 1992. Sexual agglutination in budding yeasts: structure, function, and regulation of adhesion glycoproteins. *Microbiol. Rev.* **56**: 180-194.

[32] LO W-S, DRANGINIS A. M. 1997. The cell surface Flocculin Flo11 is required for Pseudohyphae Formation and Invasion by *Saccharomyces cerevisiae*. *Molecular Biology of the Cell* **9**: 161-171.

[33] LOTTAZ C., ISELI C., JONGENEEL C.V., BUCHNER P. 2003. Modeling sequencing errors by computing Hidden Marcov models. *Bioinfrmatics* **19,2**: ii103-ii112.

[34] MORAN P., CARAS I. W. 1991. Fusion of sequence elements from non-anchored proteins to generate a fully functional signal for glycophosphatidylinositol membrane anchor attachment. *J. Cell Biol.* **115**: 1595-1600.

[35] NAKARI-SETÄLÄ T., AZEREDO J., HENRIQUES M., OLIVEIRA R., TEIXEITA J., LINDER M., PENTTILÄ M. 2002. Expression of a Fungal Hydrophobin in the *Saccharomyces cerevisiae* Cell Wall: Effect on Cell Surface Properties and Immobilization. *Applied and Enviromental Microbiology* **68**: 3385-3391.

[36] PANTOU M. P., KOUVELIS V. N., TYPAS M. A. 2006. The complete mitochondrial genome of the vascular wilt fungus *Verticillium dahliae*: a novel gene order for *Verticillium* and a diagnostic tool for species identification. *Curr. Genet.* **50**: 125-136.

[37] PITTET M., CONZELMANN A. 2006. Biosynthesis and function of GPI proteins in the yeast. *Saccharomyces cerevisiae Biochim. Biophys. Acta.* **1771**: 405-420.

[38] RIGDEN D. J., MELLO L. V., GALPERIN M. Y. 2004. The PA14 domain, a conserved all-β domain in bacterial toxins, enzymes, adhesins and signaling molecules. *Trends in Biochemical Sciences* **29**: 335-339.

[39] SANGER F., NICKLEN S., COULSON A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74**: 5463-5467.

[40] SHATKIN A.J., MANLEY J. L. 2000. The ends of the affair: capping and polyadenylation. *nature structural biology* **7(10)**: 838-842.

[41] SOANES D. M., SKINNER W., KEON J., HARGREAVES J., TALBOT N. J. 2002. Genomics of Pathogenic Fungi and the Development of Bioinformatic Resources. *The American Phytopathological Society* **15(5)**: 421-427.

[42] STEINBACH P. 2004. Standortbezogene Risikobewertung für den Erreger der Rapswelke Verticillium longisporum auf der Grundlage der Quantifizierung des Bodeninokulums. *http://www.ufop.de/downloads/Abschlussbericht_Rapswelke.pdf*.

[43] STRETTON A. O. W. 1965. The genetic code. *brit. med. Bull.* **21(3)**: 229-235.

[44] SZYMAŃSKI M., BARCISZEWSKI J. 2007. the genetic code - 40 years on. *Acta Biochimica Polonica* **54(1)**: 51-54.

[45] TUCKER S. L., TALBOT N. J. 2001. Surface Attachment and Pre-Penetration Stage Development by Plant Pathogenic Fungi. *Annu. Rev. Phytopathology* **39**: 385-417.

[46] UFOP UNION ZUR FÖRDERUNG VON OEL- UND PROTEINPFLANZEN E. V. 2003. Ein Vergleich der weltweit wichtigsten Anbauregionen fuer ölsaaten. *http://www.ufop.de/downloads/Weltoelsaaten_deutsch.pdf* **Endbericht für die Förderung von Öel und proteinpflanzen e. V.**.

[47] UFOP UNION ZUR FÖRDERUNG VON OEL- UND PROTEINPFLANZEN E. V. 2006. Biodiesel und pflanzliche Öle als Kraftstoffe. *http://209.85.135.104/search?q=cache:cFq3ENUjwP0J:www.ufop.de/*.

[48] UFOP UNION ZUR FÖRDERUNG VON OEL- UND PROTEINPFLANZEN E. V. 2007. UFOP Geschäftsbericht 2006/2007. *http://www.ufop.de/ufop_bericht.php* **Geschäftsbericht 2006/2007**.

[49] VERSTREPEN K. J., REYNOLDS T. B., FINK G. R. 2004. Origins of Variation in the Fungal Cell Surface *Nature Reviews Microbiology* **2**: 533-540.

[50] VERSTREPEN K. J., JANSEN A., LEWITTER F., FINK G. 2005. Intragenic tandem repeats generate functional variability. *nature genetics* **37**: 986-990.

[51] VERSTREPEN K. J., KLIS F. M. 2006. Flocculation, adhesion and biofilm formation in yeast. *Molecular Microbiology* **60**: 5-15.

[52] VITERBI A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* **13,2**: 260-269.

[53] WANG C., LEGER R. J. S. 2007. The MAD1 Adhesin of *Metarhizium anisopliae* Links Adhesion with Blastospore Production and Virulence to Insects and the MAD2 Adhesin Enables Attachment to Plasnts. *Eukaryotic Cell* **6**: 808-816.

[54] WATSON J. D., CRICK F. H. 1953. A structure for deoxyribose nucleic acid. *Nature* **171**: 137.

[55] WOLF G. A. 2003. Problemschläge früh erkennen. *DLG-Mitteilungen* **7**: 48-50.

[56] WÖRSTEN H. A. B., DE VOCHT M. L. 2000. Hydrophobins, the fungal coat unravelled. *Biochimica et Biophysica Acta* **1469**: 79-86.

[57] ZEISE K., VON TIEDEMANN A. 2002. Host Specializatoin among Vegetative Compability Groups of *Verticillium dahliae* in Relation to *Verticillium longisporum*. *J. Phytopathology* **150**: 112-119.

[58] ZHOU L., HU Q., JOHANSSON A., DIXELLIUS C. 2006. *Verticillium longisporum* and *V. dahliae*: infection and disease in *Brassica napus*. *Plant Pathology* **55**: 137-144

Anhang A

Überblick über die Kandidaten-Sequenzen

Tabelle A.1: **Ähnlichkeiten zwischen den Kadidaten-Sequenzen:**

Gezeigt wird, welche Kandidaten-Sequenzen starke Ähnlichkeit zueinander besitzen. In der Arbeit wird die linke ID genutzt.

a2	
a3	a1
a4	b18
a5	a9
a6	a7 a15 b7 b10 b12 b14
a8	a10 a18
a11	b2
a12	
a13	
a14	
a16	a20
a17	
a19	
a21	
a22	
b1	
b3	
b4	b9 b17
b5	
b6	b8 b11 b15 b19
b13	
b16	
b20	
b21	

Anhang B

Inhaltsverzeichnis der Daten-CD

Für die Dateien, die die Daten-CD enthält, werden folgende Bezeichnungen verwendet:

cDNA: Verzeichnisse oder Dateien, die Ergebnisse oder Sequenzen cDNAs enthalten

delta_flo8: Verzeichnisse oder Dateien, die Ergebnisse oder Sequenzen der Kandidaten enthalten, die mittels des Δflo8-Screens gefunden wurden.

delta_flo8_11: Verzeichnisse oder Dateien, die Ergebnisse oder Sequenzen der Kandidaten enthalten, die mittels des Δflo8/11-Screens gefunden wurden.

_a: Verzeichnisse oder Dateien, die Ergebnisse oder Sequenzen der kodierenden Bereiche enthalten, die gefunden werden, wenn ESTScan mit Sequenzen der Ascomycota trainiert wurde.

_s: Verzeichnisse oder Dateien, die Ergebnisse oder Sequenzen der kodierenden Bereiche enthalten, die gefunden werden, wenn ESTScan mit Sequenzen der Sordariomycetes trainiert wurde.

_rf: Verzeichnisse oder Dateien, die Ergebnisse oder Sequenzen enthalten, die erhalten werden, wenn die Nukleotid-Sequenz der cDNAs in sechs verschiedene Leserahmen übersetzt wird.

Die Namen der angelegten Verzeichnisse und Dateien sind fettgedruckt. Eingerückte Namen sind Unterverzeichnisse der weiter vorne stehenden.

Sequences: Im fasta-Format gespeicherte Sequenzen

all_seq: Alle unbearbeitete Sequenzen

cdna.fasta

delta_flo8.fasta

delta_flo8_11.fasta

long_cdna.fasta: Alle Sequenzen, die länger als 20 bp sind

seq: Sequenzen, die von der Plasmidsequenz bereinigt und länger als 20 bp sind.
(Die Benennung entspricht der der unbearbeiteten Sequenzen)

coding_sequences: Kodierende Bereiche in den Sequenzen

as: AS-Sequenz der kodierenden Bereiche

cdna_a.fasta

cdna_s.fasta

delta_flo8_a.fasta

delta_flo8_s.fasta

delta_flo8_11_a.fasta

delta_flo8_11_s.fasta

nucleotide: Nukleotid-Sequenz der kodierenden Bereiche
(Die Benennung entspricht der der kodierenden Bereiche, die als AS-Sequenz dargestellt sind)

BLAST: Ergebnisse der BLAST-Suchen, die mit den Sequenzen durchgeführt wurde

cdna

swissprot: Ergebnisse der blastx-Suche gegen die SwissProt-DB

cogeme_blastn: Ergebnisse der blastn-Suche gegen die Cogeme-DB

cogeme_tblastx: Ergebnisse der blastn-Suche gegen die Cogeme-DB

nr: Ergebnisse der blastx-Suche gegen die nr-DB

saccharomyces_blastn: Ergebnisse der blastn-Suche gegen die Saccharomyces-DB

saccharomyces_tblastx: Ergebnisse der tblastx-Suche gegen die Saccharomyces-DB

cdna: Ergebnisse der BLAST-Suche gegen die cDNA-DB

delta_flo8: (Die Benennung entspricht der der BLAST-Suchen, die für die cDNA-Sequenzen

durchgeführt wurden)

delta_flo8_11: (Die Benennung entspricht der der BLAST-Suchen, die für die cDNA-Sequenzen durchgeführt wurden)

GPI: Ergebnisse der Suche nach GPI-Anker-Signalsequenzen

cdna:

big_pi_rf.txt

big_pi_a.txt

big_pi_s.txt

dgpi_rf.txt

dgpi_a.txt

dgpi_s.txt

gpi_som_rf.txt

gpi_som_a.txt

gpi_som_s.txt

delta_flo8: (Die Bennenung entspricht der der Ergebnisse für die cDNA-Sequenzen)

delta_flo8_11: (Die Bennenung entspricht der der Ergebnisse für die cDNA-Sequenzen)

Ser_Thr: Ergebnisse der Suche nach Ser/Thr-reichen Regionen

cdna

100_rf.txt: Sequenzen mit Ser/Thr-reichen Bereichen, die länger als 100 AS sind.

100_a.txt

100_s.txt

200_rf.txt: Sequenzen mit Ser/Thr-reichen Bereichen, die länger als 200 AS sind.

200_a.txt

200_s.txt

250_rf.txt: Sequenzen mit Ser/Thr-reichen Bereichen, die länger als 250 AS sind.

250_a.txt

250_s.txt

delta_flo8:

50_lr.txt: Sequenzen mit Ser/Thr-reichen Bereichen, die länger als 50 AS sind.

50_a.txt

50_s.txt

100_lr.txt: Sequenzen mit Ser/Thr-reichen Bereichen, die länger als 100 AS sind.

100_a.txt

100_s.txt

200_lr.txt: Sequenzen mit Ser/Thr-reichen Bereichen, die länger als 200 AS sind.

200_a.txt

200_s.txt

delta_flo8_11: (Die Benennung entspricht der der Ser/Thr-reichen Regionen in den delta_flo8-Sequenzen)

Mapping: Zuordnung ähnlicher Sequenzen zueinander mittels einer BLAST-Suche

cdna_pairs.txt: Paare von cDNA-Sequenzen

delta_flo8_pairs.txt: Paare von Sequenzen des Δflo8-Screens

delta_flo8_11_pairs.txt: Paare von Sequenzen des Δflo8/11-Screens

cdna_groups.txt: Gruppen von cDNA-Sequenzen, die sich ähneln

Correlation: Ähnlichkeit zwischen den BLAST-Ergebnissen der Kandidaten-Sequenzen

delta_flo8

best_10: Für jede Sequenz werden jeweils die zehn besten BLAST-Treffer angezeigt

best_blast: Für jede Sequenz werden alle BLAST-Treffer mit ausreichend niedrigem e-Value angezeigt

correlation_best_10: Korrelationen der zehn besten-BLAST-Ergebnisse

correlation_best_blast: Korrelationen aller BLAST-Ergebnisse

delta_flo8_11 (Die Bennenung entspricht der der Ergebnisse für die Kandidaten des Δflo8-Screens)

IDs.txt: Gegenüberstellung der ursprünglichen und neu vergebenen IDs