Annotation und Klassifikation nuklearer Domänen

Dissertation
zur Erlangung des akademischen Grades des
Doktors der Naturwissenschaften

vorgelegt beim Fachbereich Biologie
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

Tobias Doerks
aus Heidelberg

Heidelberg 2001

Dank gilt meinem Chef Dr. Peer Bork, der mich an seinem Wissen teilhaben ließ, mir alles über Sequenzanalyse vermittelt hat und stets ein hilfreicher Betreuer war.
Besonders bedanke ich mich auch bei Prof. Dr. Anna Starzinski-Powitz, ohne deren Engagement ein Kontak zum EMBL nie stattgefunden hätte, und die meine Arbeit vor der Universität Frankfurt am Main vertritt.
Darüber hinaus bin ich meinen Kollegen nachhaltig dankbar für produktiven wissenschaftlichen Austausch und andere soziale Aktivitäten.

Inhaltsverzeichnis

Teil 1

Einleitung

1.1. Bioinformatik-Grundlagen der Genomanalyse

Die Bioinformatik kann als relativ junge Disziplin der Naturwissenschaften verstanden werden, die die Informatik mit der Biologie verbindet. Sie bedient sich computergestützter Methoden zur Lösung biologischer Fragestellungen, Verarbeitung komplexer laborbiologischer Resultate und Erstellung wissenschaftlicher Analysen belebter Systeme. Eine präzise Definition des Begriffes „Bioinformatik“ existiert nicht. Sie umfaßt ein weites Feld, beginnend bei der automatischen Sequenzierung und Annotation pro- und eukaryontischer Genome, über homologiebasierende Protein- und Nukleotidsequenzanalysen, Verwaltung und Erstellung biologischer Datenbanken, bis hin zu Struktur- und Funktionsvorhersagen und "modelling".

In der Molekularbiologie leistet die Bioinformatik bei der Aufnahme und Verwaltung der Nukleotid- und Aminosäuresequenzen in Datenbanken wertvolle unterstützende Arbeit. Im Rahmen der Genomprojekte ist es in den letzten Jahren zu einem explosionsartigen Anstieg der Datenfülle und damit zu stetig wachsenden bioinformatischen Anwendungsmöglichkeiten gekommen, die ihrerseits notwendig sind, um diese Datenfülle zu bewältigen.
Beginnend mit *Haemophilus influenzae* sind bis heute mehr als 40 eubakterielle und 9 Archeagenome vollständig sequenziert (siehe detalierte Liste unter http://www.TIGR.ORG/tdb/mdb/mdbcomplete.html) Desweiteren sind die eukaryontischen Genome von *Saccharomyces cerevisia, Caenorhabditis elegans, Drosophila melanogaster*, *Arabidopsis thaliana* und als "Jahrhunderprojekt" der Mensch (International Human Genome Sequencing Consortium 2001) bis heute entschlüsselt und öffentlich zugänglich. Weitere folgen in kürze.
Konsequenz der rapide zunehmenden Datenmenge an Sequenzen ist eine relative Abnahme der experimentell untersuchten und charakterisierten Proteine (z.B. 7% der circa 20000 *C.elegans* Proteine sind Teil laborbiologischer Experimente). Eine noch gravierendere Diskrepanz kann für Spezies angenommen werden, die nicht generell als experimentelle Organismen Verwendung finden. Dies veranschaulicht die Bedeutung homologiebasierender Sequenzanalysen, die in struktureller und speziell

funktioneller Hinsicht die Lücke zwischen wenigen experimentell untersuchten und neuen unbekannten Proteinen schließen können.

Die Funktion eines Proteins wird auf unterschiedlichen Ebenen zu beschreiben versucht. Da ist sind z.B. wenige spezifsiche Aminosäuren, die für die katalytische Aktivität eines Enzyms verantwortlich sind. Diese funktionell spezifischen Aminosäuren sind Ihrerseits eingebunden in die dreidimensionale Struktur, die notwendig ist für die zweckmäßige Bereitstellung der katalytischen Reste im räumlichen Kontext. Die Funktion des Proteins in der Zelle hängt nicht allein von seiner Beschaffenheit ab; es interagiert mit anderen Proteinen und Zellkomponenten gewebsspezifisch und abhängig von der Lokalisation (z.B. nuklear, cyotoplasmatisch oder membrangebunden). Diese und andere Aspekte müssen bei der Übertragung der Funktion eines experimentell untersuchten Proteins auf unbekannte Homologe berücksichtigt werden.

Standardmethoden (siehe Methoden) bestimmen die Signifikanz der Sequenzähnlichkeit und damit die Homologie. Diese erlaubt nun einen Rückschluß auf die Struktur und bedingt auf die Funktion des Proteins. Dieser Analyse liegt die Idee zugrunde, daß beide Proteine aus einem gemeinsamen Vorläufer enstanden sind. In unterschiedlichen Spezies kann es sich bei entsprechender Homologie über die gesamte Sequenz um Orthologe, also dem funktionell vergleichbaren Gegenstück handeln, in der gleichen Spezies sind es Paraloge mit ähnlicher Funktion in anderem Kontext.

Die Entscheidung über die funktionelle Übertragung, über Orthologie und Paralogie erfordert ein hohes Maß an Erfahrung und Berücksichtigung der Verwandtschaft aller Homologer.

Ein durchschnittliches Protein besteht aus separaten strukturellen Einheiten, sogenannten Domänen (siehe 1.3.1.). Jede Domäne bedeutet unterschiedliche Funktionalität innerhalb des Proteins (z.B. DNA-Bindung, katalytische Aktivität oder Proteininteraktion). Bei jeder Funktionsvorhersage ist die Analyse der Modularität (partielle Homologie) daher eine unbedingte Notwendigkeit, um Fehlannotationen auszuschließen.

Bei Domänen handelt es sich nicht nur um funktionell sondern auch evolutiv bedeutsame Einheiten. Ihre phylogenetische Verteilung und Ihr korreliertes Auftreten kann Erkenntnisse über evolutive Prozesse in Proteinen und Organismen bedeuten.

1.2. Modulare Architektur von Proteinen

1.2.1. Definiton des Domänenbegriffes

Die umfassende Berücksichtigung von Informationen über Proteinsequenzen bildet das Rückgrat zur Aufstellung struktureller, funktioneller oder evolutiver Hypothesen und ist vor dem Hintergrund vorliegender bereits sequenzierter Genome eine der bedeutendsten Herausforderungen des 21. Jahrhunderts. Wesentlich für das Verständnis der Evolution von Proteinfamilien und ihrer funktionellen Aufgaben in Organismen ist das Wissen um das Domänenkonzept. Eine Domäne ist eine im Genom bewegliche, strukturell konservierte Einheit. Auf Proteinebene stellt sie eine dreidimensionale globuläre Struktur dar, die die Fähigkeit besitzt, sich eigenständig um einen hydrophoben Kern zu falten (Janin et al. 1985). Domänen bilden innerhalb der modularen Archtitektur eines Proteins fundamentale nicht weiter teilbare Einheiten.

Experimentelle Strukturbestimmungen der letzten 30 Jahre haben gezeigt, daß Bereiche ähnlicher Sequenz auch strukturelle Ähnlichkeit besitzen und somit homologe Mitglieder derselben Familie sind (Doolittle et al. 1995, Henikoff et al. 1997). Die strukturelle Homologie kann über die Sequenzähnlichkeit erhalten sein, auf deren Ebene sie bei manchen Domänen nicht mehr feststellbar ist (Murzin 1998).

Trotz der enormen Anzahl von Proteinen wird basierend auf umfangreichen Sequenzanalysen nur von circa 1000 strukturell unterschiedlichen Domänen ausgegangen (Chothia 1992, Green et al. 1993). Dies erlaubt den Schluß, daß funktionelle Neuentstehungen im Rahmen evolutiver Prozesse vermehrt auf Genduplikationen, Mutationen in vorhandenem Sequenzenmaterial und Domänenassoziation als auf Schaffung neuer Gene aus nicht kodierender DNA beruhen. Die Analyse von Domänenarchitekturen ermöglicht damit wissenschaftliche Einblicke in die evolutive Entwicklung von Proteinfamilien.

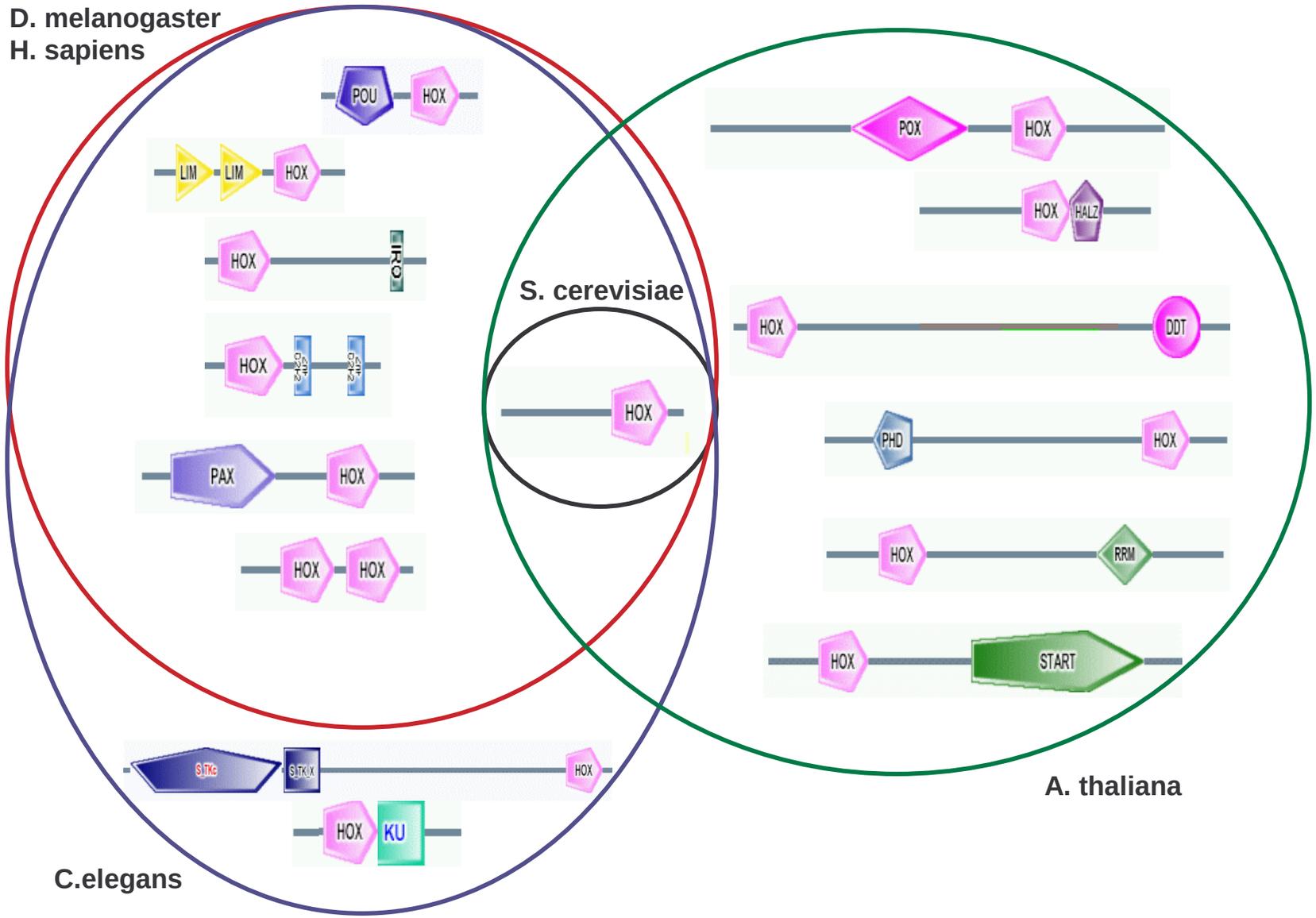
1.2.2. Beispiel Homeobox

Die Homeobox (HOX)-Domäne ist für die sequenz-spezifische DNA-Bindung in Transkriptionsfaktoren verantwortlich, die eine wesentliche Rolle bei der Embryonalentwicklung und Furchunsprozessen spielen (Gehring et al. 1994). Die Bindung an die DNA erfolgt über eine "helix-turn-helix" Struktur; die zweite α-Helix lagert sich in die große Grube der DNA ein, während die erste α-Helix die Bindungsstruktur stabilisiert.

Die Domäne wurde in den 80er Jahren in Drosophila entdeckt und beschrieben (Gehring 1985) und seitdem in Hefe, Pflanzen und Tieren einschließlich Vertebraten identifiziert. Bis heute sind mehr als 2500 Homeobox-Proteine bekannt, deren modulare Architektur die Evolution von Domänen veranschaulicht (siehe Abbildung 1).

Während in *Saccharomyces cerevisiae* nur zehn Homeobox-Proteine vorkommen, die ausschließlich allein auftreten, hat sowohl in Pflanzen als auch Tieren eine evolutive Radiation stattgefunden. Die Assoziation mit anderen funktionellen Domänen weist auf die stammesgeschichtlich frühe Trennung von Pflanzen und Tieren hin. In *Arabidopsis thaliana* findet sich die HOX-Domäne allein und gemeinsam mit Pflanzen-spezifischen Domänen wie der funktionell nicht charakterisierte POX- Domäne oder einem atypischen Leucin-Zipper (HALZ) oder mit anderen über die Spezies hinaus verbreiteten Domänen. Sie ist assoziiert mit Protein-, RNA-, DNA- oder Lipid-bindenden Domänen (PHD, RRM, DDT oder START (diese Domäne ist in *C. elegans* und *Drosophila* aber nicht in menschlichen Proteinen vorhanden)). Keine Mehrdomänen-Architektur der fast 150 pflanzlichen Proteine ist in Tieren anzutreffen.

In Drosophila, C. elegans und Mensch tritt die Homeoboxdomäne allein, intern dupliziert oder mit unterschiedlichen Nukleinsäure- oder Proteinbindungsdomänen auf (siehe Abbildung 2). Hinsichtlich der Architektur stimmen die 171 Proteine in Drosophila mit den 328 menschlichen überein, während in Nematoden im Laufe der phylogenetischen Entwicklung unter den 136 Proteinen zwei weitere C.elegans-spezifische Architekturen entstanden sind; die Homeobox-Domäne tritt in einer Proteinkinase und mit einer in Proteinkinase-Inhibitoren vorhandenen Kunitz-Domäne assoziiert auf.


D. melanogaster
H. sapiens
POU
HOX
LIM
LIM
HOX
HOX
IRO
HOX
PAX
HOX
HOX
HOX
S. cerevisiae
HOX
C.elegans
STKc
STKX
HOX
HOX
KU
POX
HOX
HOX
HALZ
HOX
DDT
PHD
HOX
HOX
RRM
HOX
START
A. thaliana

Beschreibung Abbildung 2
Graphische Darstellung der modularen Architektur von Homeobox-Proteinen und ihre Verteilung in verschiedenen Spezies. Schwarzer Kreis: Protein-Architektur der Homeobox-Proteine in *Saccharomyces cerevisiae*, grüner Kreis: Protein-Architekturen der Homeobox-Proteine in *Arabidopsis thaliana*, roter Kreis: Protein-Architekturen der Homeobox-Proteine in *Drosophila melanogaster* und *Homo sapiens*, blauer Kreis: Protein-Architekturen der Homeobox-Proteine in *Caenorrhabditis elegans*.

Zusammenfassend verdeutlicht die Verteilung der Homeobox-Domäne in den fünf Spezies drei wesentliche stammesgeschichtliche Aspekte, die architektonisch und somit im funktionellen Kontext einfache Präsenz in der einzelligen Hefe, die evolutive Radiation in mehrzelligen Organismen und die Phylum-spezifische Konservierung der modularen Architektur.

1.3. Annotation von nuklearen Proteinen

1.3.1. Definiton nuklearer Lokalisation

Biochemische Stoffwechselwege, Signaltransduktionskaskaden und Zellreplikations- und Aufbauprozesse sind nicht kompartimentgebunden, sondern miteinander über die gesamte Zelle oder darüber hinaus verknüpft. Als Folge dieser intrazellulären Wechselwirkungen sind Proteine nicht zwanglsläufig auf ein Zellkompartiment beschränkt. So kann ein Transkriptionsfaktor bereits im Cytosol Liganden binden, bevor er in den Zellkern eindringt, zu translatierende mRNA passiert begleitet von Transportproteinen die Kernmembran, Enzyme werden in zellfreien Raum sezerniert und vieles mehr. Der aus der evolutiven Entwicklung hervorgegangene komplexe modulare Aufbau von Proteinen und die damit verbundene Präsenz gleicher Domänen in unterschiedlichen Proteinfamilien, bedeutet eine oft vielfältige Verteilung einer Domänenfamilie auf unterschiedliche Zellbereiche, assoziiert mit den verschiedensten funktionellen Komplexen.
Dieser Sachverhalt erschwert eine eindeutige lokalisationsspezifische Zuordnung für Proteine und insbesondere Domänen. Im Rahmen dieser Arbeit wurden neben

anderen schwerpunktmäßig nukleare Proteine und Domänen annotiert, klassifiziert und vertiefend untersucht. Die Definiton nukleare Domäne erhebt nicht den unerfüllbaren Anspruch der Ausschließlichkeit, sondern stellt die sinnvolle Abgrenzung zur klassischen cytosolischen signalling-Domäne dar und weißt auf Ihre Relevanz bei Nukleus- oder Nukleoid-lokalisierten Prozessen hin.

1.3.2. Historische Entwicklung nuklearer Domänen

Auch wenn sich die molekulare Analyse nuklearer Proteine bereits über einige Jahrzehnte erstreckt, ist die systematische Identifikation nuklearer Domänen ein relativ junges Phänomen.

Als vor sechzehn Jahren der C2H2-Zingfinger als erste nukleare Domäne beschrieben wurde (Miller et al. 1985), war die Domänenentdeckungsperiode der extrazellulären Domänen schon weit fortgeschritten (siehe Abbildung 2). Ihre Identifikation begann Ende der siebziger Jahre (Sottrup-Jensen et al. 1975). Während die intensive Analyse der nuklearen Domänen mehr als fünf Jahre später einsetzte, ist die regelmäßige Domänenidentifikation in "signalling"-Proteinen nur um ein bis zwei weitere Jahre verschoben. Den Höhepunkt 1995 erreichend stieg die Anzahl der Entdeckungen nuklearer Domänen stetig an, extrazelluläre Domänen wurden mit den beginnenden 90'ger Jahren jährlich weniger identifiziert. Zeitversetzt zur Erstbeschreibung der Domänen wurden und werden die dreidimensionalen Strukturen bestimmt (Abbildung 2, dünne Linien). Während zwischen 1980 und 1985 noch durchschnittlich 8,3 Jahre und zwischen 1985 und 1990 noch 7,5 Jahre zwischen der Entdeckung extrazellulärer Domänen und Auflösung ihrer Struktur vergingen, waren es zwischen 1990 und 1995 und für nukleare Proteine über die gesamten 15 Jahre durchschnittlich nur 4,5 Jahren. In den letzten fünf Jahren zeigte sich tendenziell eine geringfügige Abnahme neu entdeckter nuklearer Domänen; diese Tendenz wurde für das Jahr 2001 im Rahmen dieser Arbeit mit einem 26%igen Zuwachs neuer nuklearer Domänen durchbrochen (siehe Kapitel 3.3.).

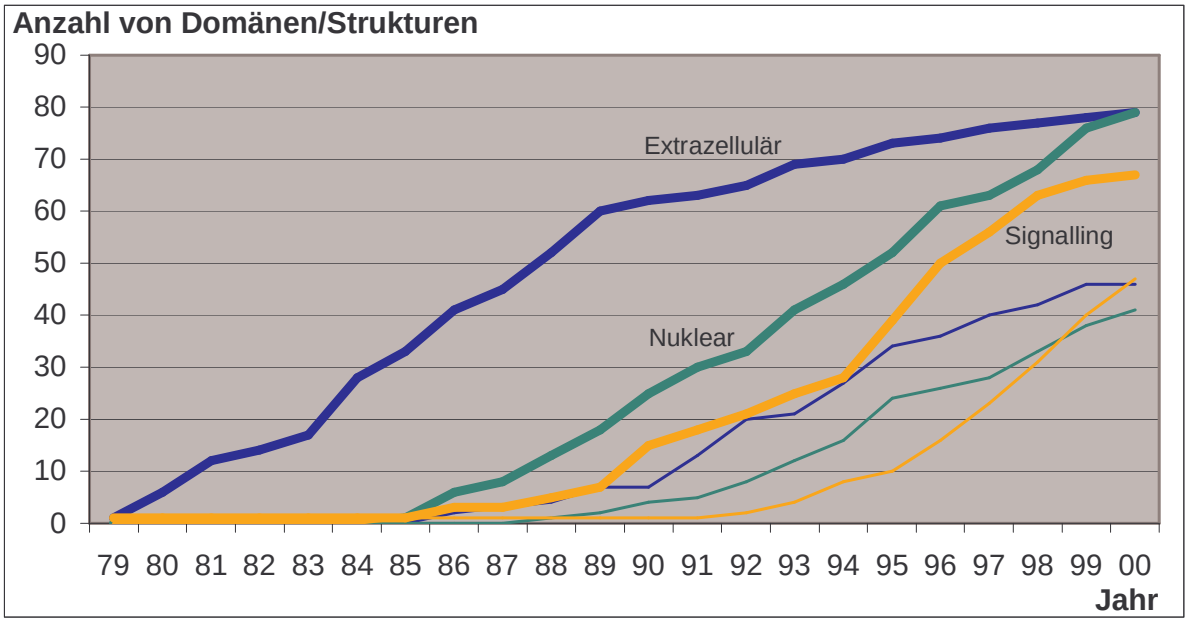


Abbildung 1. (aus Copley et al. in press). Kurvendiagramm der Domänenentdeckung und Bestimmung der Tertärstruktur. Anzahl der entdeckten Domänen oder experimentell bestimmten Strukturen (X-Achse) pro Jahr (Y-Achse). Dicke Linien: entdeckte Domänen, dünne Linien: farblich äquivalent zugehörige Strukturen (blau: extrazelluläre Domänen/Strukturen, grün: nukleare Domänen/Strukturen, orange: "signalling"-Domänen/Strukturen.

1.3.3. Zukunft der Domänenanalyse

Es ist naheliegend, dass die Anzahl unterschiedlicher Domänen in Proteinen begrenzt ist (Cothia 1992, Green et al. 1993). Abbildung 2 in Kapitel 1.3.2 zeigt neben der geringfügigen Abnahme der Entdeckung neuer nuklearer Domänen, eine deutliche Sättigung besonders bei extrazellulären und "signalling"-Domänen. Auch wenn intensive automatisierte Analysen gepaart mit aufwendigen Einzeluntersuchungen wie im Falle der nuklearen Domänen (siehe Kapitel 3.3) einen kurzfristigen sprunghaften Anstieg neuer Entdeckungen bedeuten können, führt die begrenzte Anzahl vorhandener Domänen zukünftig zu einer Schwerpunktverlagerung. Neben der Domänenentdeckung werden vermehrt Subfamilien-spezifische Funktionsanalysen im Mittelpunkt stehen. Motiv-orientierte Untersuchungen können dann detaillierte Aussagen über z.B. sequenzspezifische DNA-Bindung oder substratspezifische Katalyse ermöglichen.

Eine weitere Aufgabe wird es sein, sich mit dem Zusammenführen von Domänenfamilien sorgfältig zu beschäftigen. Aufgrund der rasant wachsenden Anzahl von Proteinsequenzen treten gehäuft "linker"-Sequenzen auf, die die Homologie zwischen zwei Familien deutlich werden lassen. Desweiteren führen verbesserte sensitivere Suchalgorithmen (z.B. PSI-BLAST oder Hmms) zur Offenlegung neuer interdomänalen Homologien, und im Rahmen der Auflösung von Strukturen werden Homologien erkennbar, die auf Sequenzebene nicht mehr detektierbar sind.

1.4. Zielsetzung

Ziel dieser Doktorarbeit war die umfassende Klassifikation und Annotation nuklearer Domänen, die systematische Analyse der modularen Architektur nuklearer Proteine, die daraus resultierende Entdeckung und funktionelle Charakterisierung unbekannter Domänen und darüber hinaus ihre Einordnung in einen phylogenetischen und evolutiven Kontext.

In einem ersten Schritt sollten bekannte nukleare Domänen aus Proteinsequenz- und Domänendatenbanken, sowie aus der Literatur ermittelt werden; unter Berücksichtigung korrekter Domänengrenzen sollten multiple Sequenzalignments und Hidden Markov Models erstellt und in das Simple Modular Architecture Research Tool implementiert werden. Dieses wird ergänzt durch umfangreiche Literaturinformationen hinsichtlich Entdeckung, Struktur und experimentellen Hintergrund.

Im zweiten Schritt werden ausgesuchte nukleare und kontextabhängig andere funktionell bedeutsame Proteine homologiebasierenden Sequenzanalyseverfahren unterzogen, um neue Domänen zu identifizieren und zu charaktisieren, die das Wissen um Funktion und Evolution mehren und den Weg für laborbiologische Experimente ebnen.

Abschließend sollte die Gesamtheit nuklearer Proteine automatisch untersucht und einzeln ausgewertet werden. Es gilt in einer Komplettananlyse aller unbekannten Regionen möglichst umfassend annähernd alle unentdeckten nuklearen Domänen aufzuspüren und zu beschreiben.

Teil 2

Methoden

2.1. Datenbanken, Analysemethoden und Programme

2.1.1. Sequenz- und Strukturdatenbanken

EMBL (Baker et al. 2000)
DNA- und RNA-Sequenzen werden in der EMBL–Nukleotidsequenz–Datenbank verwaltet. Sie werden der Literatur entnommen oder direkt von Wissenschaftlern und Sequenziergruppen zugesandt.

TREMBL (Bairoch et al 2000)
Die TREMBL–Datenbank enthält Aminosäuresequenzen, die aus der Translation der EMBL–Nukleotidsequenzen gewonnen werden. Übersetzungen mit internen Stopcodons werden nicht berücksichtigt.

SWISSPROT (Bairoch et 2000)
Die Datenbank SWISSPROT enthält Proteinsequenzen aus Übersetzungen der EMBL-Datenbank, aus der PIR (Protein Identification Resource)–Datenbank, aus der Literatur oder aus direkt eingesandten Sequenzen. Sie bietet unter anderem Referenzen zur Prosite–Datenbank, zur EMBL–Datenbank, zu PDB (Protein Data Base) und zu Domänendatenbanken wie Pfam und Smart.

GenBank (Benson et al. 1999)
Datenbank aller bekannten Nukleotid- und Proteinsequenzen inklusive biologischer Information und Referenzen.

nrdb
In NRDB (Non Redundant Data Base) sind Aminosäuresequenzen zusammengestellt, die aus anderen Banken wie TREMBL, SWISSPROT, PIR oder Genbank gewonnen werden. Identische Sequenzen werden zu einem NRDB–Eintrag zusammengefaßt. Die Datenbank enthält 756743 Einträge (Stand: 16.10.2001)

PDB

PDB (Protein Data Base) enthält Informationen über die 3D–Struktur von Proteinen; sie gibt die Koordinaten der einzelnen Atome eines Proteins und damit ihre Orientierung im Raum an. Am 19 September 2001 enhielt die Datenbank 15729 Einträge.

2.1.2. Datenbank-Suchalgorithmen

Blast (Altschul et al. 1990)

Der Sequenzvergleich zwischen Suchsequenz und Datenbank beruht auf dem Blast (basic local alignment search tool) – Algorithmus. Grundlage bilden HSPs (Highscoring Segment Pairs), Fragmente willkürlicher aber gleicher Länge aus Such– und Datenbanksequenz, die aligniert ein lokales Maximum bilden und einen Schwellenwert überschreiten können.

Die Signifikanz der gefundenen Homologie zwischen zwei Sequenzen wird durch den sogenannten p-Wert repräsentiert. Er ist ein Maß für die Wahrscheinlichkeit, daß die Sequenzähnlichkeit zufällig ist. Dem Benutzer stehen fünf unterschiedliche Suchmethoden zur Verfügung:

- blastn: Suche von Nukleotidsequenz gegen Nukleotidsequenzdatenbank
- blastp: Suche von Proteinsequenz gegen Proteinsequenzdatenbank
- blastx: Suche von sechs Leserastern übersetzter Nukleotidsequenz gegen Proteinsequenzdatenbank
- tbblastn: Suche von Proteinsequenz gegen sechs Leseraster übersetzter Nukleotidsequenzdatenbank
- tblastx: Suche von sechs Leserastern übersetzter Nukleotidsequenz gegen sechs Leseraster übersetzter Nukleotidsequenzdatenbank

Mit der Auswahl von unterschiedlichen Matrizen kann die Homologiebeurteilung beeinflußt werden. Es handelt sich hierbei um Aminosäure–Austausch–Matrizen, die die evolutive Signifikanz einer Mutation bewerten. Sie unterscheiden sich in der Art ihrer Berechnung und in dem Datensatz, der ihre Grundlage darstellt.

Gebräuchlich sind Blosum und die Gonnet–Serie, während die ältere Dayhoff–Pam–Matrix als weniger effektiv gilt.
Der Einsatz von Filtern wie SEG und XNU erlaubt das selektive nicht Berücksichtigen von Sequenzabschnitten, die für die Funktions– oder Verwandtschaftsvorhersage als nicht relevant gelten (z.B. serin– oder prolinreiche Regionen). Die Suche wird so für die wesentlichen Bereiche sensibilisiert.

Gapped Blast (Altschul et al. 1990)
Zusätzlich zu den Funktionen von Blast erstellt Gapped Blast auch Alignments, die Lücken enthalten. Das heißt, Alignments homologer Bereiche, die durch nicht homologe Bereiche unterbrochen sind, werden als ein Gesamtalignment dargestellt.

PSI-BLAST (Altschul et al 1997, Altschul et al 1998)
Ergänzend zum herkömmlichen BLAST gestattet PSI-BLAST die Erstellung eines Profils aus den für signifikant homolog befundenen Sequenzen und iterativ weitere Suchen mit dem jeweiligen Profil.

MoST
MoST (Motif-Search-Tool) gestattet iterative Homologiesuchen mit einer positionsabhängigen Wichtungsmatrix. Ein Alignment-Block ohne Insertionen oder Deletionen wird zur Bildung einer Matrix herangezogen, die dann mit Sequenzen aus einer Datenbank hinsichtlich ihrer Ähnlichkeit verglichen wird.
Sequenzen, die signfikant homolog zur Matrix sind, werden zur Bildung einer neuen Matrix verwand, die erneut mit Datenbanksequenzen verglichen werden können.

Searchwise (Birney et al. 1996)
Der zugrunde liegende Algorithmus erlaubt den Vergleich eines Profils gegen alle sechs Leseraster einer DNA-Sequenz, kann somit auch Sequenzen, die Verschiebungen des Rasters aufweisen, auffinden.

PairWise (Birney et al. 1996)
PairWise dient zur Berechnung eines Profils aus einem Alignment.
Die Berechnung erfolgt vor dem Hintergrung, daß stark unterschiedliche Sequenzen in einem Alignment stärkeren Einfluß auf das Profil nehmen als sehr ähnliche.

HMMer (Eddy 1998)
Profil hidden Markov models (profile HMMs) dienen zur sensitiven Datenbanksuche unter Verwendung statistischer Beschreibungen der Konsensus-Sequenz einer Domänenfamilie. Die HMM-Sequenzanalyse-software ist frei erhältlich (siehe http://hmmer.wustl.edu/)

2.1.3. Domänen - und Motivdatenbanken

PROSITE / PROSITEDOC (Hofmann et al. 1999)
Die Prosite–Datenbank enthält Aminosäuresequenzprofile und Muster, die für Proteinfamilen oder Domänen charakteristisch sind. Die vergleichende Suche einer Sequenz gegen die Datenbank ermöglicht das Auffinden bekannter Domänen oder die Zuordnung zu einer definierten Familie.
Neben signifikanten Mustern sind in der Datenbank Referenzen zu SWISSPROT-Sequenzen, die das entsprechende Profil aufweisen, enthalten.
Zusätzlich finden sich Verweise auf die PROSITEDOC-Datenbank, die ihrerseits Einträge über Eigenschaften und Funktionen von Proteinfamilien und Literaturreferenzen beinhaltet.

Pfam (Bateman et al. 2000)
Pfam ist eine umfangreiche Zusammenstellung aus multiplen Sequenz-alignments und hidden Markov models, die 2216 Protein-Domänen abdecken (Stand Oktober 2001)

Smart (Schultz et al. 1998, Schultz et al. 2000)
Smart, das **S**imple **M**odular **A**rchitecture **R**esearch **T**ool ist eine Zusammenstellung aus handgefertigten Alignments und den zugehörigen hidden Markov models für extrazelluläre, Signaltransduktions- , nukleare und anderen Proteindomänen.
Geprüfte "thresholds" erlauben eine präzise Vorhersage der modularen Architektur von Proteinen. Desweiteren werden Informationen zur Struktur. Funktion, Lokalisation und anderes und Links zu Datenbanken wie Pfam und PBD bereitgestellt.

Interpro (Apweiler et al. 2001)
Interpro ist eine umfangreiche Zusammenstellung aus Informationen von Domänen-, Motiv- und anderen Proteinsequenzdatenbanken über Domänen- und Proteinfamilien.

2.1.4. Vorhersage intrinsicher Eigenschaften von Proteinen

Coils2 (Lupas et al. 1991) **–Programm zur Vorhersage von Coiled Coil Strukturen**
Das Programm vergleicht eine Proteinsequenz mit Sequenzen, die bekanntermaßen in der Lage sind, parallele dopppelsträngige coiled coils zu bilden. Die Datenbank enthält Sequenzen von Myosinen, Tropomyosinen und Keratinen. Aus einem Ähnlichkeitswert wird dann eine Coiled coil-Ausbildewahrscheinlichkeit berechnet.

TopPred2 (von Hejne 1992) **- Programm zur Vorhersage von Transmembranregionen**
Die Vorhersage von α-helicalen Transmembranbereichen beruht auf der Analyse des Proteins hinsichtlich seiner Hydrophobizität. Der Algorithmus beurteilt die Lage hydrophober Aminosäuren zueinander und bildet so die Grundlage zur Bildung von Wahrscheinlichkeitswerten.

SignalP (Nielsen et al. 1997) **- Programm zur Vorhersage von N-terminalen Signalsequenzen**
SignalP ermöglicht den Vergleich einer Suchsequenz mit Datensätzen aus Signalpeptid-tragenden Proteinen. Es stehen drei unterschiedliche Datensätze für gram-positive Prokaryonten, gram-negative Prokaryonten und Eukaryonten zur Verfügung.
Ähnlichkeitswerte führen zur Berechnung von Wahrscheinlichkeiten für die N-terminalen Aminosäuren als Bestandteil einer Signalsequenz, für das Vorhandensein einer Abspaltstelle und der Kombination aus beiden. Daraus resultiert abschließend die Prognose.

SEG (Wootton et al. 1996)
SEG detektiert Regionen von "low compositional complexity" in Proteinsequenzen.

2.1.5. Programme zur Anfertigung und Bearbeitung von Alignments

ClustalW (Thompson 1994)
Das Programm dient zur Erstellung von multiplen Sequenz-Alignments. Nach paarweiser Alignierung der einzelnen Sequenzen wird eine Distanzmatrix berechnet, die die Basis für den Entwurf eines Stammbaumes ist. Gemäß ihrer Verwandtschaft werden die Sequenzen zu einem Gesamtalignment zusammengefügt.

ClustalX (Thompson 1997)
Die Funktion von ClustalX entspricht der von ClustalW mit zusätzlicher graphischer Benutzeroberfläche.

SeaView (Galtier et al. 1996)
Neben der Möglichkeit zur Erstellung von multiplen Sequenz-Alignments gestattet SeaView ergänzend die Editierung von Sequenzen.

MACAW (Schuler et al. 1991)
Die "Multiple Alignment Construction and Analysis Workbench" erlaubt die Analyse von Segmenten eines Alignments und prüft die Signifikanz der Homologie.

consensus
Kalkuliert die Konsensus-Sequenz für Alignments (N. Brown und J. Lai, unpublished; see http://www.bork.embl-heidelberg.de/Alignment/consensus.html)

2.1.6. Vorhersage der Sekundär- und Tertiärstruktur von Proteinen

PHD (Rost et al.1994)
Unter Verwendung eines neuronalen Netzes bietet dieses Programm die Möglichkeit der Vorhersage von Sekundärstrukturen (α-Helixstruktur, β-Faltblattstruktur, turns und loops).

3D-PSSM (Kelley et al. 1999)
Beruhend auf der Erkennung schwacher Homologien berechnet 3D-PSSM eine positionsspezifische Matrix zur Vorhersage der Tertiärstruktur.

2.1.7. Sonstiges

Sequenzvergleich

Dotter (Sonnhammer et al. 1995)
Dieses Dotplot-Programm erlaubt den detailierten Vergleich zweier Sequenzen. Sequenzabschnitte werden eins zu eins auf 100% Aminosäure- oder Nukleotidübereinstimmung geprüft und in einem Koordinatensystem graphisch dargestellt.

Graphische Darstellung von 3D-Strukturen

Rasmol
Rasmol ist geeignet, pdb-Datenbank-Einträge nutzend Proteine in ihrer 3D-Struktur abzubilden.

Darstellung phylogenetsicher Bäume

Njplot

Diese Programme erlauben die graphische Darstellung und Bearbeitung von phylogenetischen Bäumen, wie sie von ClustalX oder ClustalW erstellt werden.

Meta_Annotator (Eisenhaber et al. 1998)

Das Programm sagt die zelluläre Lokalisation basierend auf der Annotation in SWISSPROT vorher.

SEALS (Walker et al. 1997)

Das "**S**ystems for **E**asy **A**nalysis of **L**ots of **S**equences" ist ein Programmpaket, das automatische und einfache Proteinsequenzanalysen großer Datenmengen ermöglicht.

Grouper

Ein Programm aus dem "SEALS - package", das in einem Blast-ähnlichen Verfahren homologe Sequenzen automatisch in Gruppen zusammenfaßt.

SRS (Etzold et al. 1996)

"**S**equence **R**etrieval **S**ystem"

RC-Methode (Doerks et al. accepted)

Die RC-Methode dient zum automatischen Aufspüren möglicher neuer Domänen. Proteinsequenzabschnitte (von 30 Aminosäuren und länger), die sich zwischen in der Smart-Datenbank abgelegten Domänen befinden, werden aus nrdb extrahiert und unter Zuhilfenahme des Programms Grouper aus dem SEALS-package in homologen Gruppen zusammengeführt.

2.2.2. Grundlagen der Sequenzanalyse

Die Definition homologer Bereiche (Domänen) in Proteinen oder die Charakterisierung von orthologen und paralogen Sequenzen erfolgt über das Aufspüren signifikanter Sequenzähnlichkeiten.
Standard-Suchmethoden wie BLAST und FASTA entdecken nach Schätzungen ein Drittel aller Homologen in Datenbanken (Park et al. 1998). Die Sensitivität kann durch reziproke Suchen mit gefundenen Sequenzen (Park et al. 1997) und durch iterative Suchen verbessert werden (Salamov et al. 1999).
PSI-BLAST (Altschul et al. 1997, 1998) nutzt die Ergebnisse einer BLAST-Suche, um ein internes Alignment aus der Startsequenz und den in der Datenbank gefundenen Sequenzen zu erstellen. Ein Profil (positions-spezifische Bewertungsmatrix) beruhend auf diesem Alignment kann dann verwendet werden, um erneut gegen die Datenbank zu suchen. Dieser Vorgang läßt sich wiederholen, bis keine weiteren Homologen detektierbar sind. Die einfache und schnelle iterative Methode ermöglicht das Auffinden von erheblich mehr Sequenzen als die herkömmliche BLAST-Variante.
Um die Divergenz der Sequenzen und konservierte Eigenschaften hervortreten zu lassen, werden die homologen Bereiche zu einem multiplen Alignment mit z.B. Clustalx (Thompson et al. 1997) zusammengefaßt .
Konservierte Abschnitte können farbig hervorgehoben und dadurch Domänen, also Strukturen die für eine funktionell bedeutsame Faltung typisch sind, oder Motive, kurze Sequenzabschnitte mit hochkonservierten Einzelaminosäuren erkennbar werden.
Eine weitere systematische und effiziente Nutzung der Informationen, die das Aligment bereitstellt, beruht auf der Erstellung von Hidden Markov Models (HMMs) (Eddy 1998). Diese konstruieren Wahrscheinlichkeitsannahmen unter Berücksichtigung statistischer Gewichtungen der Typs der Aminosäure in Bezug auf seine Position und eines möglichen Austauschs im Alignment.
HMMs bieten verglichen mit PSI-BLAST-Suchen den Vorteil, daß sie nicht auf automatisch konstruierte lokale Alignments beschränkt sind, sondern auf einem handgefertigten Alignment beruhend nicht automatisch detektierbare Sequenzähnlichkeiten berücksichtigen können.

Neben Homologie können auch intrinsische Eigenschaften für Sequenzähnlichkeiten verantwortlich sein. Strukturelle und funktionelle Ansprüche an die Proteinsequenz erzwingen eine Aminosäurekomposition, die ähnlich aber nicht in evolutiver Hinsicht homolog ist. Hierbei handelt es sich z. B. um hydrophobe Transmembranhelices, N-terminale Signalpeptide oder gestaucht α-helicale CoiledCoil-Strukturen, die aus polaren und geladene Aminosäuren und einem hydrophoben Kern bestehen.

Um Fehler bei der Interpretation von Sequenzähnlichkeiten zu minimieren, müssen die mit intrinsischen Eigenschaften verbundenen Aminosäurekompositionen berücksichtigt werden, was durch Ihre Vorhersage ermöglicht wird (van Hejne 1992, Nielsen et al. 1997, Lupas et al. 1991).

Teil 3

Ergebnisse und Diskussion

3.1. Systematische Annotation nuklearer Domänen

Grundlage der systematischen Analyse nuklearer Domänen bildet eine Zusammenstellung des Großteils aller bekannten überwiegend nuklearen oder nukleoiden Domänen.

3.1.1. Zusammenstellung, Alignment- und Profilanfertigung nuklearer Domänen

Die systematische Suche in Sequenz-Datenbanken und Literaturrecherche ermöglichte die Zusammenstellung einer umfangreichen Liste überwiegend nuklearer und nukleoider Proteine und ihrer Domänen.
Mehr als 35000 Proteine, die von 164 Domänen abgedeckt werden, sind zur Zeit aus nrdb (non-redundant database) extrahierbar.
Für alle Domänen wurden wenn möglich unter Zuhilfename der Tertiärstruktur die exakten Grenzen bestimmt und multiple Alignments angefertigt (Thompson et al. 1994, 1997, Galtier et al. 1996). Diese dienten zur Erstellung von Hidden Markov Models (Eddy 1998).
Siehe Tabelle im Anhang.

3.1.2. Implementation der Domänenkollektion in die Domänendatenbank Smart

Die nuklearen Domänen wurden in das Domänenvorhersage-Programm Smart (Schultz et al. 1998, Schultz et al. 2000) (http://smart.embl-heidelberg.de/) implementiert. Mit 164 von 635 Domänen insgesammt stellen sie einen wesentlichen Anteil von mehr als 25% des Vorhersagepotenzials dar (siehe Tabelle im Anhang).
Die Aufteilung der Domänen kann in Smart aufgrund funktioneller Überschneidungen (Signaltransduktion / nukleare Lokalisation) von der Tabelle im Anhang abweichen.

3.2. Entdeckung und funktionelle Analyse neuer uncharakterisierter Domänen in nuklearen und anderen Proteinen

Die individuelle Analyse ausgesuchter nuklearer und nicht-nuklearer Proteinfamilien ermöglichte die Idenfikation unbekannter Domänen in unterschiedlichen biologischen Zusammenhängen. Strukturelle und funktionelle Untersuchungen geben tiefere Einblicke in die biologischen Aufgaben der neuentdeckten Domänen, nukleare, Membran-assoziierte oder cytosolische Prozesse.

3.2.1. L27, eine neue Hetero-Dimer-bildende Domäne in den Rezeptor-Targeting-Proteinen Lin-2 and Lin-7 (Doerks et al. 2000)

Membran-assoziierte Guanylat-Kinasen (MAGUKs) sind essentiell fuer die Organisation von Zelloberflächenproteinen und ihrer Interaktion mit dem Cytoskelett (Anderson et al. 1996). Sie beeinflussen die Ausbildung von Zell-Zell-Kontakten und sind bei der Tumorsupression wesentlich. Die *C.elegans*-Guanylat-Kinase Lin-2 bildet einen Komplex mit den Rezeptor-targeting-Proteinen Lin-7 und Lin-10; dieser Komplex ist verantwortlich für den korrekten Transport des Let-23 Wachstumsfaktor zur basolateralen Membran von Epithelzellen (Hoskins et al. 1995, Kaech et al. 1998). Ähnliche Heterotrimerkomplexe sind auch aus Vertebraten bekannt (Butz et al. 1998, Borg et al. 1998).

Lin-2 und Lin-7 sowie orthologe Proteine tragen eine PDZ-Domäne, von der bekannt ist, dass sie bei der Komplexbildung keine Rolle spielt (Kaech et al. 1998, Butz et al. 1998) (siehe Abbildung 2).

Eine PSI-BLAST-Suche (Altschul et al. 1997, Altschul et al. 1998) mit der Region N-terminal zur PDZ-Domäne (aus dem Lin-2-ähnlichen dlg2-Protein) findet diese Domäne in allen Lin-2-Orthologen Maguks.

Dotplot-Analysen (Sonnhammer et al. 1995) dieser Proteine detektieren eine interne Duplikation, die durch Macaw (Schuler et al. 1991) - Analysen (P-value 10^{-50}) bestätigt wird (siehe Abbildung 3).

Eine zusätzliche BLASTP-Suche (Altschul et al. 1990) mit der duplizierten Region in dlg2 gegen die Datenbank wormpep18 zeigt Sequenzähnlichkeit (E-value 0.1) zu dem Protein Lin-7.

Der entdeckte Bereich ist ebenfalls im N-terminus lokalisiert und wird von einer PDZ-Domäne gefolgt. Nicht nur der biologische Kontext stützt die Signifikanz der Sequenzähnlichkeit sondern auch die weiterführenden Analysen des multiplen Alignments (siehe Beschreibung von Abbildung 3).

Die einfache Präsenz in Lin-7-Proteinen und das doppelte Auftreten in Lin-2 und Ortholgen weist auf eine unabängige Domäne hin. Die neue Domäne erhielt den Namen L27 nach der besser charakterisierten MAGUK-Subfamilie (Lin-2) und ihren Bindungspartnern, die Lin-7-Proteinfamilie.

Die L27-Domäne erstreckt sich über ungefähr 50 Aminosäuren; wesentliche Komponenten sind konservierte negativ geladene und eine aromatische Aminosäure (siehe Abbildung 3). Auffällig und von vermutlich funktioneller Bedeutung ist der Aspekt, dass die zweite Kopie der Domäne in Lin-2-Proteinen ein konserviertes Histidin enthält (ausgenommen Lin-2 in *C. elegans*).

Es wurde bereits experimentell bestätigt, dass die Region N-terminal zu der PDZ-Domäne in Lin-2-Ortholgen die Lin-7-Bindung vermittelt (Kaech et al. 1998, Butz et al. 1998). Punktmutationen konservierter Aminosäuren in der zweiten Helix der L27-Domäne verhindern die Bindung von Lin-2 an Lin-7 in der Maus (nicht publizierte Daten).

Die neu gewonnenen Erkenntnisse stützen die Hypothese einer Heterodimerbildung von Lin-2 und Lin-7 im Bereich der L27-Domäne.

Obwohl weitere BLAST-Suchen gegen gebräuchliche Datenbanken keine weiteren homologen Proteine entdeckten, ähnelt die Anordnung konservierter geladener und hydrophober Aminosäuren in Verbindung mit Sekundärstrukturvorhersagen dem Kernbereich einer Histidin-tragendenden Phosphortransfer-HPt-Domäne (Kato et al. 1997). Die HPt-Domäne vermittelt die Phosphortransfer-Reaktion in Zwei-Komponent-Signaltransduktionssystemen. Macaw-Analysen erhärten diese Entdeckung.

Sekundärstrukturvorhersagen (Rost et al.1994) und die Ähnlichkeit zur HPt-Domäne ließen fuer die L27-Domäne eine "three helical bundle"-Struktur annehmen.

Zusammenfassend bedeutet die Entdeckung der L27-Domäne eine genauere Charakterisierung von Rezeptor-targeting-Proteinen, sowie ein besseres Verständnis um Ihre modulare Architektur und damit zusammenhängende strukturelle und funktionelle Vorhersagen, die schlüssig deren Dimerbildung erklären.

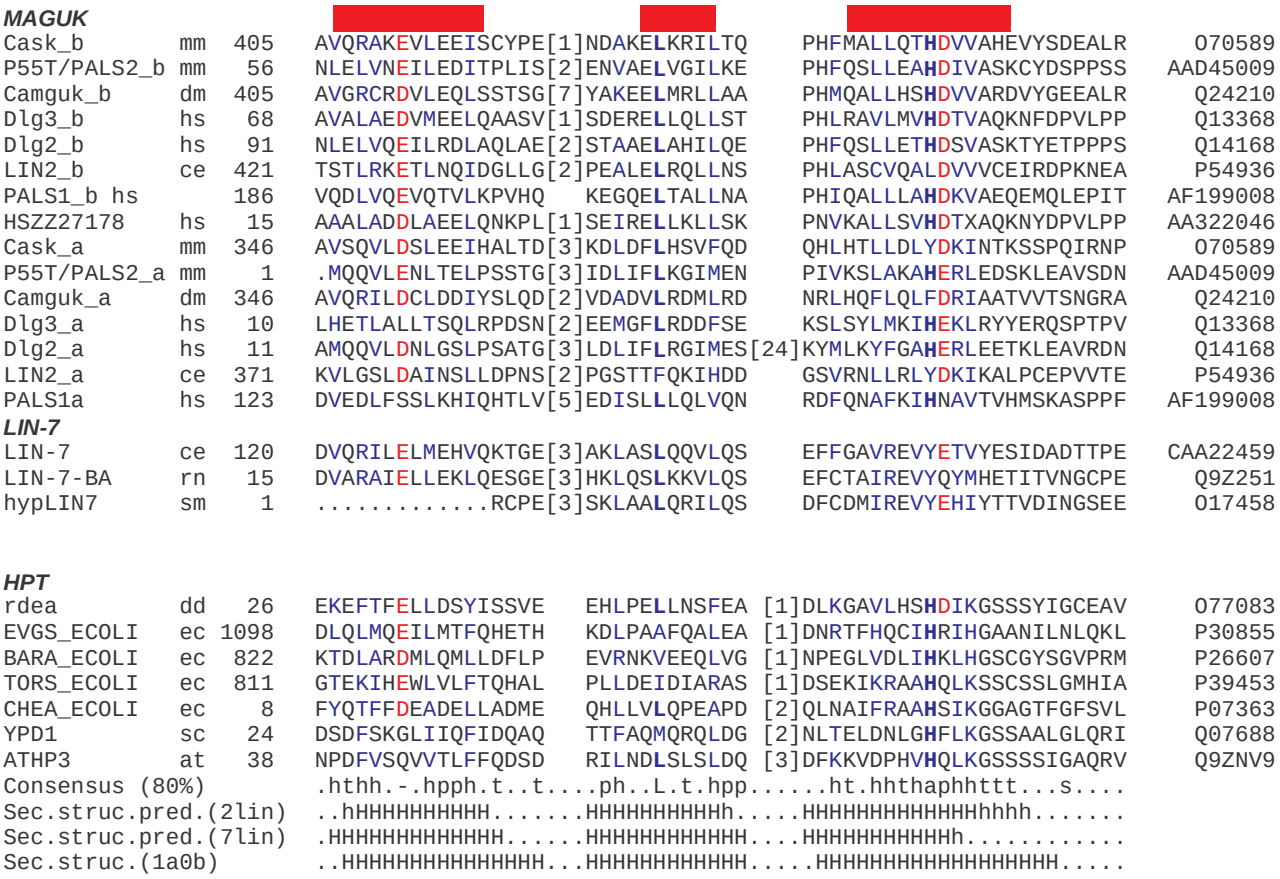


Abbildung 3. Multiples Alignment der L27-Domäne von MAGUK Lin-2-homologen Proteinen, Lin-7-homologen Proteinen und ausgesuchten Mitgliedern HPt-Domänen tragender Proteine. Erste Spalte: Protein-Namen (wiederholt auftretende Domänen sind mit a und b gekennzeichnet); zweite Spalte: Spezies-Bezeichnung (at: *Arabidopsis thaliana*; Ce: *Caenorhabditis elegans*; dd: *Dictyostelium discoideum*; dm: *Drosophila melanogaster*; ec: *Echerichia coli*; hs: *Homo sapiens*; mm: *Mus musculus*; rn: *Rattus norvegicus*; sc: *Saccharomyces cerevisae*; sm: *Schistosoma*

mansoni); dritte Spalte: erste Aminosäure der Domäne im jeweiligen Protein; letzte Spalte: Datenbank Accession Nummer. Konservierte negativ geladene Aminosäuren sind rot markiert; konservierte hydrophobe Aminosäuren sind blau; zusätzliche konservierte Aminosäuren sind fett gedruckt. Die Konsensus-Sequenz (konservierte Aminosäuren in 80% aller Sequenzen) befindet sich unter dem Alignment; h, p, a, t, s und - stehen für hydrophobe, polare, aromatische, turn-artige, kleine (s =small) und negativ geladene Aminosäuren. Konservierte Aminosäuren, hydrophobe Bereiche und die Sekundärstruktur ähneln dem Kern Histidin-tragender Phosphotransfer-HPt-Domänen; ein Alignment ausgewählter Sequenzen befindet sich unter dem MAGUK-Alignment. Die vorhergesagte Sekundärstruktur basierend auf dem Aligment der Lin-2- und Lin-7-Homologen und die bekannte Sekundärstruktur der HPt-Domäne (1a0b) stehen in den letzten beiden Zeilen (H, Helix bekannt oder vorhergesagt mit einer durchschnittlichen Genauigkeit > 82%); h, Helix bekannt oder vorhergesagt mit einer durchschnittlichen Genauigkeit < 82%)) (Rost et al. 1994). Das multiple Alignment ist in drei separate Blöcke unterteilt (hervorgehoben durch Zahlen in Klammern). MACAW Alignment-Analysen (Schuler et al. 1991) zeigen signifikante Sequenzähnlichkeit für die beiden Kopien der Domäne in Lin-2 und für die Domäne in Lin-7 (markiert mit einer roten Linie über dem Alignment). Für den ersten Block mit einem P value von $7.7*10^{-13}$ über 12 Aminosäuren, für den zweiten Block mit einem P value von $4,9*10^{-8}$ über 6 Aminosäuren und für den dritten Block mit einem P value von 10^{-50} über 13 Aminosäuren.

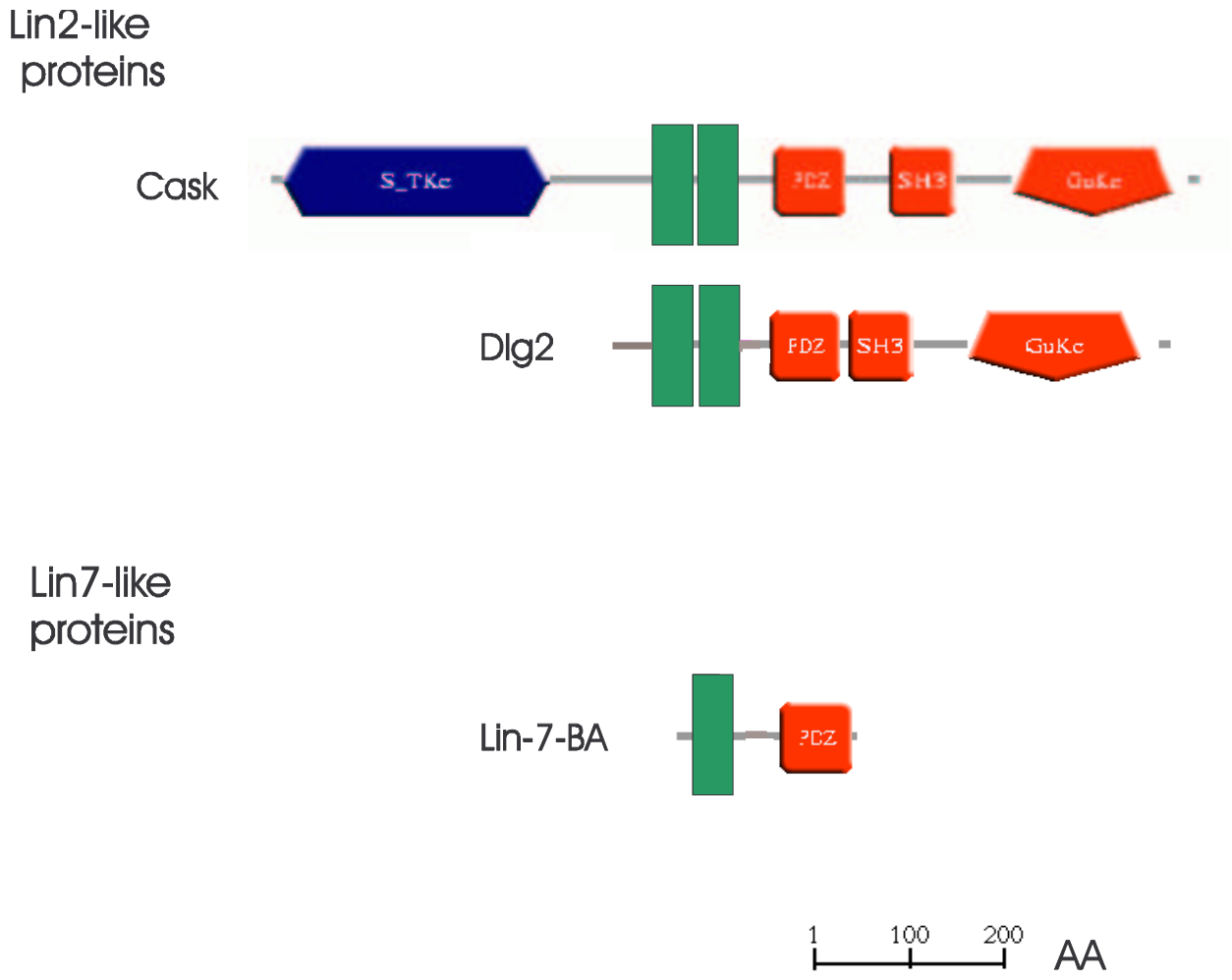


Abbildung 4. Domänen-Architektur der L27-Domänen-tragenden Proteine (grün).
Die Abbildung beinhaltet nur Proteine unterschiedlichen modularen Aufbaus. Die Domännamen sind dem Simple Modular Architecture Research Tool (Schulz et al. 1998, Schulz et al. 2000) (http://smart.embl-heidelberg.de) entlehnt. GuKc: Guanylat-Kinase-Domäne; S_TKc: Serin/Threonin Protein-Kinase-Domäne; SH3: src Homologie 3 Domäne; PDZ: Domäne in PSD-95, dlg und ZO-1/2 - Proteinen.

3.2.2. GRAM, eine neue Domäne in Glucosyltransferasen, Myotubularinen und anderen Membran-assoziierten Proteinen (Doerks et al. 2000)

UGT51/52 Glucosyltransferasen sind wesentlich für die Biosynthese von Sterolglucosiden, Membran-assoziierten Lipiden, die in vielen Eukaryoten anzutreffen sind (Warnecke et al. 1999). Der katalytischen Domäne geht in Hefe und *Dictyostelium* eine N-terminale Ausdehnung voraus, die homologen Proteinen in Pflanzen fehlt (Warnecke et al. 1999).

Eine PSI-BLAST-Suche (Altschul et al. 1997, Altschul et al. 1998) gegen nrdb mit einer konservierten Region (Aminosäuren 207 bis 274) der N-terminalen Verlängerung (siehe Abbildung 5) von *Dictyostelium discoideum* findet den Bereich mit signifikanter Ähnlichkeit ($E=10^{-4}$) in anderen Glucosyltransferasen in Hefe. Nach der zweiten Iteration wird die konservierte Region in TBC-Domänen-tragenden Proteinen ($E=2*10^{-8}$), die die GTPase-Aktivität in Rab-ähnlichen GTPasen regulieren (Neuwald et al. 1997) und in hypothetischen Proteinen ($E=3*10^{-9}$) detektiert (siehe Abbildung 6). Weitere iterative Suchen spüren eine Kopie der Domäne in den Rab-ähnlichen GTPase Aktivatoren ($E=2*10^{-7}$) und als verkürztes Duplikat in den Glucosyltransferasen in Hefe ($E=7*10^{-4}$) auf. Dieses verkürzte Duplikat geht C-terminal direkt in eine PH-Domäne über.

Nach der fünften Iteration wird die neue Domäne in C2-Domänen-tragenden Proteinen ($E=2*10^{-5}$) und in einigen kleinen Formin-bindenden ABA (abscisic acid-responsive-element-binding) Proteinen ($E=1*10^{-4}$) gefunden. Diese Proteine spielen bei der Streßantwort in höheren Pflanzen eine Rolle (Giraudat et al. 1994, Choi et al. 2000). Die Ergebnisse werden durch Hidden Markov Model-Suchen (Eddy 1998) gestützt.

Weitere PSI-BLAST-Suchen zeigten Ähnlichkeit unmittelbar unterhalb des Signifikanz-Schwellenwertes zu MTM1/MTMR1 Myotubularinen (Laporte et al. 1996, 1998) (E=0.064), einer Familie dual-spezifischer Phosphatasen und zu Sbf (SET (Suvar3-9, Enhancer-of-zeste, Trithorax)-domain-binding factor) Proteinen (E=1.4); Sbf-Familien-Mitglieder sind Myotubularin-ähnliche Proteine, die die Phosphatase-Aktivität verloren haben (Cui et al. 1998). Mutationen in der neuen Domäne und anderen Regionen des mtm1-Gens sind verantwortlich für eine angeborene

Myopathie, die durch Hypotonie und Atmungsinsuffizienz gekennzeichnet ist (de Guyon et al. 1997, Laporte et al.1996).
Die schwache Homologie wurde durch MACAW-Alignment-Analysen bestätigt (P values zwischen 10^{-11} and 10^{-50}).
Die neue Domäne erhielt den Namen GRAM nach den besser charakterisierten Glucosyltransferasen, Rab-ähnlichen GTPase-Aktivatoren und Myotubularinen.
Die Gram-Domäne ist nur in einer von sechs als biochemische GTPase-Aktivatoren identifizierten Familien vorhanden (Albert et al. 1999), was eine Funktion hinsichtlich der Aktivationsregulierung unwahrscheinlich macht.
Alle nicht-katalytsichen Domänen (PH-Domäne (Gibson et al. 1994, Zhang et al. 1995), TBC-Domäne (Neuwald et al. 1997), C2-Domäne (Rizo et al. 1998, Davletov et al. 1993) , FYVE-Domäne (Gaullier et al. 1998), C1-Domäne (Hurley et al. 1998) und andere), mit denen die GRAM-Domäne in Proteinen gemeinsam auftritt, sind erstlinig in Membran-assoziierte Prozesse involviert.

Die GRAM-Domäne hat eine durchschnittliche Länge von 70 Aminosäuren (50 in der verkürzten Variante). Sekundärstrukturvorhersagen mit PHD (Rost et al. 1994) zeigen vier β-Stränge, die für den Kern der Domäne eine β-Faltblattstruktur annehmen lassen. Jeder Strang besitzt eine konservierte aromatische Aminosäure; weitere Charakteristika sind konservierte geladene Aminosäuren und ein invariantes Glycin (Abbildung 5). Die C-terminale α-Helix fehlt in der verkürzten PH-Domänen-assoziierten GRAM-Domäne. Vergleichbare Verkürzungen sind kein ungewöhnliches Phänomen (siehe z.B. Ubiquitin-konjugiernde Enzym-Familie (Ponting et al. 1997, Koonin et al. 1997)).

Zusammenfassend scheint die neu entdeckte Domäne intrazelluläre Protein- oder Lipidbindungseigenschaften im Rahmen der Signaltransduktion zu erfüllen und in wichtige Membran-assoziierte Prozesse involviert zu sein.
Sie ist sowohl Spezies-übergreifend in Proteinfamilien vertreten (z.B. Rab-ähnliche GTPase-Aktivatoren, Myotubularine (MTM1/MTMR1)) als auch in Spezies-spezifische Stoffwechselvorgänge eingebunden (z.B. Glucosyltransferasen).

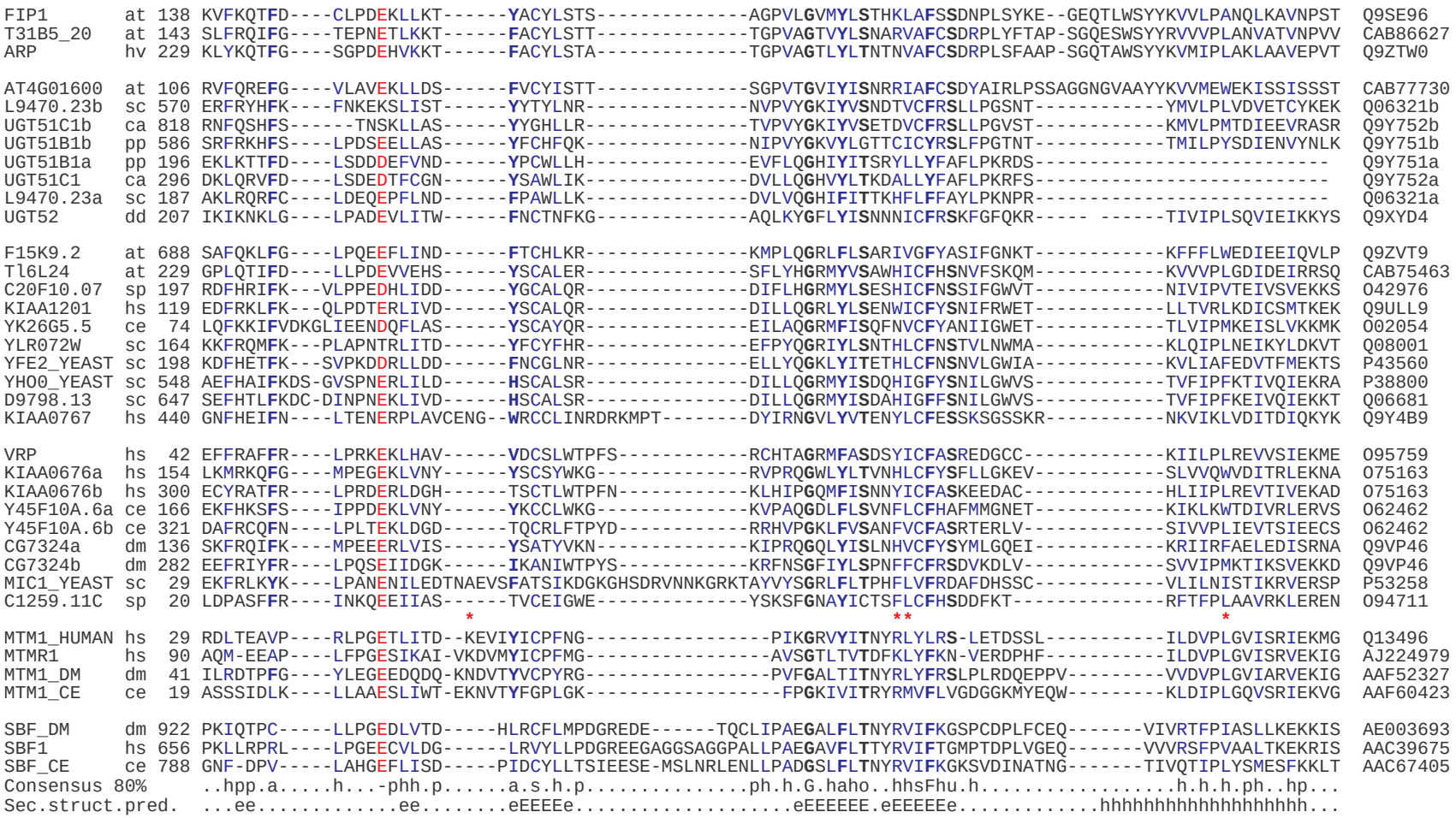
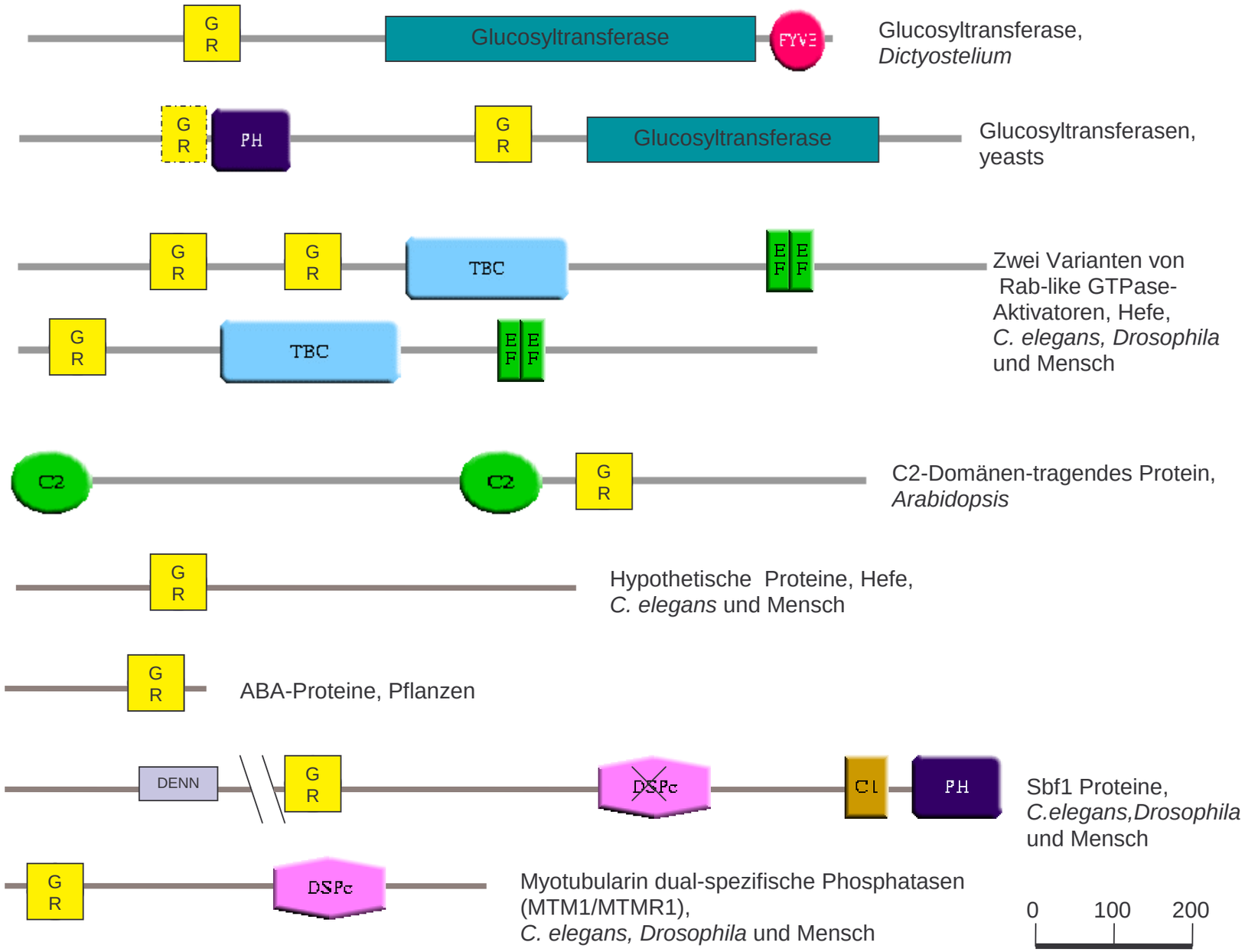

FIP1 at 138 KVFKQTFD----CLPDEKLLKT------YACYLSTS--------------AGPVLGVMYLSTHKLAFSSDNPLSYKE--GEQTLWSYYKVVLPANQLKAVNPST Q9SE96
T31B5_20 at 143 SLFRQIFG----TEPNETLKKT------FACYLSTT--------------TGPVAGTVYLSNARVAFCSDRPLYFTAP-SGQESWSYYRVVVPLANVATVNPVV CAB86627
ARP hv 229 KLYKQTFG----SGPDEHVKKT------FACYLSTA--------------TGPVAGTLYLTNTNVAFCSDRPLSFAAP-SGQTAWSYYKVMIPLAKLAAVEPVT Q9ZTW0
AT4G01600 at 106 RVFQREFG----VLAVEKLLDS------FVCYISTT--------------SGPVTGVIYISNRRIAFCSDYAIRLPSSAGGNGVAAYYKVVMEWEKISSISSST CAB77730
L9470.23b sc 570 ERFRYHFK----FNKEKSLIST------YYTYLNR---------------NVPVYGKIYVSNDTVCFRSLLPGSNT------------YMVLPLVDVETCYKEK Q06321b
UGT51C1b ca 818 RNFQSHFS------TNSKLLAS------YYGHLLR---------------TVPVYGKIYVSETDVCFRSLLPGVST------------KMVLPMTDIEEVRASR Q9Y752b
UGT51B1b pp 586 SRFRKHFS----LPDSEELLAS------YFCHFQK---------------NIPVYGKVYLGTTCICYRSLFPGTNT------------TMILPYSDIENVYNLK Q9Y751b
UGT51B1a pp 196 EKLKTTFD----LSDDDEFVND------YPCWLLH---------------EVFLQGHIYITSRYLLYFAFLPKRDS--------------------------- Q9Y751a
UGT51C1 ca 296 DKLQRVFD----LSDEDTFCGN------YSAWLIK---------------DVLLQGHVYLTKDALLYFAFLPKRFS--------------------------- Q9Y752a
L9470.23a sc 187 AKLRQRFC----LDEQEPFLND------FPAWLLK---------------DVLVQGHIFITTKHFLFFAYLPKNPR------------------------- Q06321a
UGT52 dd 207 IKIKNKLG----LPADEVLITW------FNCTNFKG--------------AQLKYGFLYISNNNICFRSKFGFQKR----- ------TIVIPLSQVIEIKKYS Q9XYD4
F15K9.2 at 688 SAFQKLFG----LPQEEFLIND------FTCHLKR---------------KMPLQGRLFLSARIVGFYASIFGNKT------------KFFFLWEDIEEIQVLP Q9ZVT9
Tl6L24 at 229 GPLQTIFD----LLPDEVVEHS------YSCALER---------------SFLYHGRMYVSAWHICFHSNVFSKQM------------KVVVPLGDIDEIRRSQ CAB75463
C20F10.07 sp 197 RDFHRIFK---VLPPEDHLIDD------YGCALQR---------------DIFLHGRMYLSESHICFNSSIFGWVT------------NIVIPVTEIVSVEKKS O42976
KIAA1201 hs 119 EDFRKLFK---QLPDTERLIVD------YSCALQR---------------DILLQGRLYLSENWICFYSNIFRWET------------LLTVRLKDICSMTKEK Q9ULL9
YK26G5.5 ce 74 LQFKKIFVDKGLIEENDQFLAS------YSCAYQR---------------EILAQGRMFISQFNVCFYANIIGWET------------TLVIPMKEISLVKKMK O02054
YLR072W sc 164 KKFRQMFK---PLAPNTRLITD------YFCYFHR---------------EFPYQGRIYLSNTHLCFNSTVLNWMA------------KLQIPLNEIKYLDKVT Q08001
YFE2_YEAST sc 198 KDFHETFK---SVPKDDRLLDD------FNCGLNR---------------ELLYQGKLYITETHLCFNSNVLGWIA------------KVLIAFEDVTFMEKTS P43560
YHO0_YEAST sc 548 AEFHAIFKDS-GVSPNERLILD------HSCALSR---------------DILLQGRMYISDQHIGFYSNILGWVS------------TVFIPFKTIVQIEKRA P38800
D9798.13 sc 647 SEFHTLFKDC-DINPNEKLIVD------HSCALSR---------------DILLQGRMYISDAHIGFFSNILGWVS------------TVFIPFKEIVQIEKKT Q06681
KIAA0767 hs 440 GNFHEIFN----LTENERPLAVCENG--WRCCLINRDRKMPT--------DYIRNGVLYVTENYLCFESSKSGSSKR-----------NKVIKLVDITDIQKYK Q9Y4B9
VRP hs 42 EFFRAFFR----LPRKEKLHAV------VDCSLWTPFS------------RCHTAGRMFASDSYICFASREDGCC------------KIILPLREVVSIEKME O95759
KIAA0676a hs 154 LKMRKQFG----MPEGEKLVNY------YSCSYWKG--------------RVPRQGWLYLTVNHLCFYSFLLGKEV------------SLVVQWVDITRLEKNA O75163
KIAA0676b hs 300 ECYRATFR----LPRDERLDGH------TSCTLWTPFN------------KLHIPGQMFISNNYICFASKEEDAC------------HLIIPLREVTIVEKAD O75163
Y45F10A.6a ce 166 EKFHKSFS----IPPDEKLVNY------YKCCLWKG--------------KVPAQGDLFLSVNFLCFHAFMMGNET------------KIKLKWTDIVRLERVS O62462
Y45F10A.6b ce 321 DAFRCQFN----LPLTEKLDGD------TQCRLFTPYD------------RRHVPGKLFVSANFVCFASRTERLV------------SIVVPLIEVTSIEECS O62462
CG7324a dm 136 SKFRQIFK----MPEEERLVIS------YSATYVKN--------------KIPRQGQLYISLNHVCFYSYMLGQEI------------KRIIRFAELEDISRNA Q9VP46
CG7324b dm 282 EEFRIYFR----LPQSEIIDGK------IKANIWTPYS------------KRFNSGFIYLSPNFFCFRSDVKDLV------------SVVIPMKTIKSVEKKD Q9VP46
MIC1_YEAST sc 29 EKFRLKYK----LPANENILEDTNAEVSFATSIKDGKGHSDRVNNKGRKTAYVYSGRLFLTPHFLVFRDAFDHSSC------------VLILNISTIKRVERSP P53258
C1259.11C sp 20 LDPASFFR----INKQEEIIAS------TVCEIGWE--------------YSKSFGNAYICTSFLCFHSDDFKT--------------RFTFPLAAVRKLEREN O94711
MTM1_HUMAN hs 29 RDLTEAVP----RLPGETLITD--KEVIYICPFNG----------------PIKGRVYITNYRLYLRS-LETDSSL-----------ILDVPLGVISRIEKMG Q13496
MTMR1 hs 90 AQM-EEAP----LFPGESIKAI-VKDVMYICPFMG----------------AVSGTLTVTDFKLYFKN-VERDPHF-----------ILDVPLGVISRVEKIG AJ224979
MTM1_DM dm 41 ILRDTPFG----YLEGEEDQDQ-KNDVTYVCPYRG----------------PVFGALTITNYRLYFRSLPLRDQEPPV---------VVDVPLGVIARVEKIG AAF52327
MTM1_CE ce 19 ASSSIDLK----LLAAESLIWT-EKNVTYFGPLGK----------------FPGKIVITRYRMVFLVGDGGKMYEQW---------KLDIPLGQVSRIEKVG AAF60423
SBF_DM dm 922 PKIQTPC-----LLPGEDLVTD-----HLRCFLMPDGREDE------TQCLIPAEGALFLTNYRVIFKGSPCDPLFCEQ-------VIVRTFPIASLLKEKKIS AE003693
SBF1 hs 656 PKLLRPRL----LPGEECVLDG------LRVYLLPDGREEGAGGSAGGPALLPAEGAVFLTTYRVIFTGMPTDPLVGEQ-------VVVRSFPVAALTKEKRIS AAC39675
SBF_CE ce 788 GNF-DPV-----LAHGEFLISD-----PIDCYLLTSIEESE-MSLNRLENLLPADGSLFLTNYRVIFKGKSVDINATNG-------TIVQTIPLYSMESFKKLT AAC67405
Consensus 80% ..hpp.a.....h...-phh.p......a.s.h.p.............ph.h.G.haho..hhsFhu.h.................h.h.h.ph..hp...
Sec.struct.pred. ...ee.............ee........eEEEEe.................eEEEEEE.eEEEEEe.............hhhhhhhhhhhhhhhhhhh...

Abbildung 5

Beschreibung von Abbildung 5. Multiples Alignment der GRAM-Domäne für (ABA-responsive-element binding) Proteins (FIP1, T31B5_20, ARP, AT4G01600), PH-Domänen-tragende Glucosyltransferasen (L9470.23, UGT51B1, UGT51C1), FYVE-Domänen-tragende Glucosyltransferase (UGT52), hypothetische Proteine (F15K9.2, T16L24, C20F10.07, KIAA1201, YK26G5.5, YLR072W, YFE2_YEAST, YHO0_YEAST, D9798.13, KIAA0767), TBC-Domänen-tragende Rab-ähnliche GTPase–Aktivatoren (VRP, KIAA0676, Y45F10A.6, CG7324, MIC1_YEAST, C1259.11C), Myotubularin dual-spezifische Phosphatasen (MTM1_HUMAN, MTMR1, MTM_DM, MTM_CE), Sbf1-ähnliche dual-spezifische Phosphatasen (SBF_DM, SBF1, SBF_CE). Erste Spalte: Protein-Namen (wiederholt auftretende Domänen sind mit a und b gekennzeichnet); zweite Spalte: Spezies-Bezeichnung (at: *Arabidopsis thaliana*; ca: *Candida albicans;* ce: *Caenorhabditis elegans*; dd: *Dictyostelium discoideum*; dm: *Drosophila melanogaster*; hs: *Homo sapiens*; hv: *Hordeum vulgare*; pp: *Pichia pastoris* sc: *Saccharomyces cerevisae;* sp: *Schizosaccharomyces pombe*); dritte Spalte: erste Aminosäure der Domäne im jeweiligen Protein; letzte Spalte: Datenbank Accession Nummer. Konservierte geladene Aminosäuren sind rot markiert; konservierte hydrophobe Aminosäuren sind blau; zusätzliche konservierte Aminosäuren sind fett gedruckt. Mutationen in der GRAM-Domäne im Gen MTM1 sind mit Sternen über der Sequenz MTM1_HUMAN gennzeichnet.

Die Konsensus-Sequenz (konservierte Aminosäuren in 80 % aller Sequenzen) befindet sich unter dem Alignment; h, p, a, u, s, o und - stehen für hydrophobe, polare, aromatische, winzige, kleine (s =small), alkoholische und negativ geladene Aminosäuren. Die vorhergesagte Sekundärstruktur steht in der letzten Zeile (H, Helix bekannt oder vorhergesagt mit einer durchschnittlichen Genauigkeit > 82%; h, Helix bekannt oder vorhergesagt mit einer durchschnittlichen Genauigkeit < 82%; B, β-Strang vorhergesagt mit einer durchschnittlichen Genauigkeit > 82%; b, β-Strang vorhergesagt mit einer durchschnittlichen Genauigkeit < 82%) (Rost et al. 1994).


G R
Glucosyltransferase
FYVE
Glucosyltransferase, *Dictyostelium*
G R
PH
G R
Glucosyltransferase
Glucosyltransferasen, yeasts
G R
G R
TBC
E F
E F
G R
TBC
E F
E F
Zwei Varianten von Rab-like GTPase-Aktivatoren, Hefe, *C. elegans, Drosophila* und Mensch
C2
C2
G R
C2-Domänen-tragendes Protein, *Arabidopsis*
G R
Hypothetische Proteine, Hefe, *C. elegans* und Mensch
G R
ABA-Proteine, Pflanzen
DENN
G R
DSPc
C1
PH
Sbf1 Proteine, *C.elegans,Drosophila* und Mensch
G R
DSPc
Myotubularin dual-spezifische Phosphatasen (MTM1/MTMR1), *C. elegans, Drosophila* und Mensch
0
100
200

Beschreibung von Abbildung 6. Domänen-Architektur von GRAM-Domänen-tragenden Proteinen.

Die Abbildung beinhaltet nur Proteine unterschiedlichen modularen Aufbaus. Die Domännamen sind dem Simple Modular Architecture Research Tool (Schultz et al. 1998, Schultz et al. 2000) (http://smart.embl-heidelberg.de) entlehnt.

Abkürzungen: C1, Protein-Kinase C konservierte Region 1; C2, Protein-Kinase C konservierte Region 2 (CalB); EF, EF-hand, Calcium-bindendes Motiv; FYVEE, Domäne in Fab1, YOTB, Vac1 und EEA1; PH, Pleckstrin-homologe Domäne; DSPc, katalytische Domäne dual-spezifischer Phosphatasen; TBC, Domäne in Tre-2, BUB2p und Cdc16p. Die DSPc-Domäne in Sbf1-Proteinen ist nicht katalytisch aktiv, gekennzeichnet durch ein X. Die C1-Domäne ist nur im Sbf1-Protein von *Drosophila melanogaster* vorhanden.

Die DENN-Domäne ist definiert von Pfam (Bateman et al. 2000), die Domänengrenzen der Glucosyltransferase wurden der Literatur entnommen (Warnecke et al. 1999).

3.2.3. DDT, eine neue DNA-bindende Domäne in unterschiedlichen Transkriptionsfaktoren, Chromosom-assoziierten und anderen nuklearen Proteinen (Doerks et al. 2001)

Chromatin-umbildende Komplexe sind notwendig für die Zerstörung und Umgestaltung der Nuleosomenanordnung und beeinflussen die transkriptionelle Aktivität dieser Regionen (Kornberg et al. 1999). Unterschiedliche Bromodomänen-tragende Proteine sind mit diesen Komplexen assoziiert. Sie sind verantwortlich für Proteinbindung und in regulative Prozesse involviert (Barlev et al. 1998, Dhalliun et al. 1999).

Alle bekannten Domänen im BPTF Transkriptionsfaktor (Bromodomäne- und PHD-Finger) (Jones et al. 2000) sind als Proteinbindungsdomänen charakterisiert (Barlev et al. 1998, Aaasland et al. 1995). Eine Analyse unterschiedlicher Regionen des Transkriptionsfaktors sollte das Auffinden einer potentiell vorhandenen DNA-bindenden Domäne ermöglichen.

PSI-BLAST-Suchen (Altschul et al 1997, Altschul et al 1998) gegen NRDB (non-redundant database) mit dem N-terminalen Bereich (konservierte Aminosäuren 102 bis 162, Abbildung 7) von BPTF entdeckten signifikante Ähnlickeit zu Proteinen mit N-terminalen AT-Hooks (Aravind et al. 1998) in *C.elegans* ($E=7x10^{-11}$) und

Drosophila ($E=2x10^{-15}$) und annähernd 100% Identität zu einem C-terminal verkürzten menschlichen Transkriptionfaktor (FALZ or FAC1, 810 Aminosäuren Länge), von dem angenommen wird, dass er bei der Alzheimer Erkrankung eine Rolle spielt (Bowser et al. 1995).

Nach weiteren iterativen PSI-BLAST-Suchen wurde Homologie zu einem vier PHD-Domänen-tragenden (Aasland et al. 1995) Protein MOI20 unbekannter Funktion ($E=10^{-3}$), einem Homeobox-tragenden (Gehring et al. 1994), vermutlich DNA-bindenden Protein ($E=10^{-4}$), zu hypothetischen Proteinen in *Arabidopsis thaliana* ($E=4x10^{-6}$) und (mit schwacher Signifikanz, E=~0.04) zu hypothetischen Hefeproteinen gefunden.

Zusätzliche Hidden Markov Model-Suchen (Eddy 1998) zeigen Sequenzähnlichkeit kaum unterhalb des Grenzwertes zur BAZ-Familie chromatin-umbildender Faktoren (E=0.15); BAZ-Proteine (Jones et al. 2000) ähneln in ihrer modularen Architektur den BPTF-Transkriptionsfaktoren. Diese Proteine spielen eine Rolle beim Williams Syndrom, einer komplexen Entwicklungsstörung mit multisystemischen Defekten (Morris et al. 1988, Lu et al. 1998, Bochar et al. 2000). Die schwache Homologie wurde durch MACAW-Alignment-Analysen (Schuler et al. 1991) bestätigt (P-values between 10^{-12} and 10^{-50}). Die neu entdeckte Domäne wurde DDT nach den besser charakterisierten DNA-bindenden Homeobox-tragenden Proteinen, den unterschiedlichen Transkriptions- und Chromatinumbildungs-Faktoren benannt.

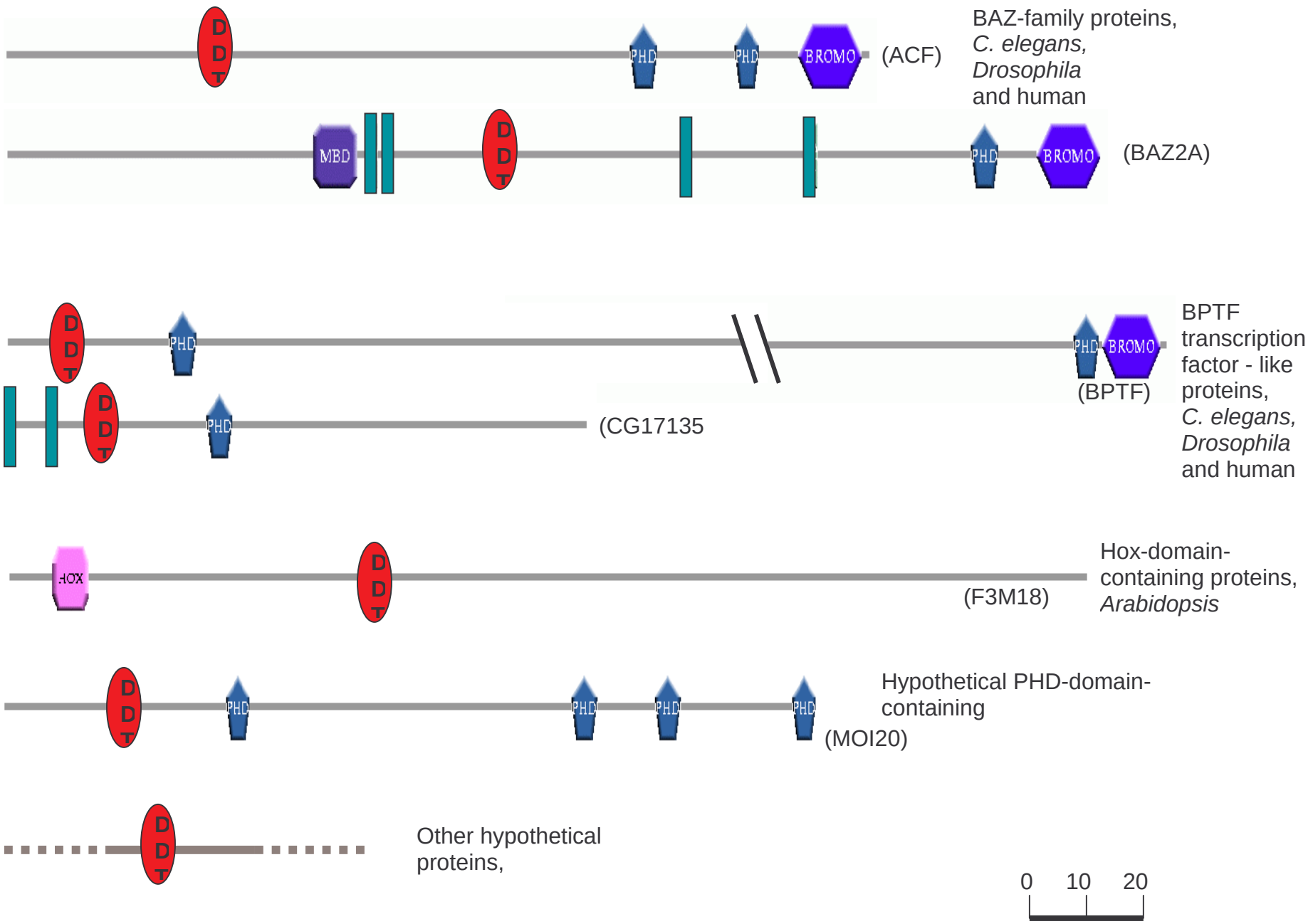
Die durchschnittlich 60 Aminosäuren lange DDT-Domäne ist ausschliesslich mit nuklearen Domänen assoziiert (siehe Abbildung 8); die Analyse des multiplen Alignments zeigt charakteristische konservierte geladene Aminosäuren, N-terminale Phenylalanine und C-terminale Leucine (siehe Abbildung 7). Eine frühere Veröffentlichung (Jones et al. 2000) über das angebliche Vorhandenseins eines LXXLL (Heery et al. 1997) Protein-Protein-Interaktionsmotives ist nicht nachvollziehbar; das Motiv ist nur in der BAZ-Subfamilie konserviert und damit nicht von funktioneller Relevanz.

Eine DNA-bindende Funktion der DDT-Domäne ist wahrscheinlich. Von einigen der identifizierten Proteinen ist bekannt, dass sie DNA binden, ohne daß eine DNA-bindende Domäne charakterisiert ist (z.B. ACF1 (Ito et al. 1999) und BPTF (Jones et al. 2000).

Weiterhin wurde bereits experimentell nachgewiesen, dass FAC1 (die verkürzte Variante von BPTF) in der N-terminalen Region (Aminosäuren 1 bis 398), die die DDT-Domäne beinhaltet, essentiell für DNA-Bindung ist (Jordan-Sciutto et al 1999). PHD (Rost et al. 1994) sagt drei Helices als Sekundärstruktur für die DDT-Domäne voraus; vielen DNA-Bindungsdomänen liegt ein Faltungstyp aus drei Helices zugrunde.

```
BPTF     hs 102 NEHIMNVIAIYEVLRNFGTVLR-LSPFR------FEDFCAALVSQ-EQCTLMAEMHVVLLKAVLREEDT Q9UIG2
CG17135 dm 189 NTHVLRALSIYEVLRRFRHMVR-LSPFR------FEDLCAALACE-EQSALLTEVHIMLLKAILREEDA Q9W0T0
F26H11   ce 253 TASIMDAVEIYEILRSYHRTLR-ITPFT------FEDFCAALISH-NNSCIMAEVHMALLRNCLKSDDE O45409
BAZ1A    hs 573 PEIFGDALMVLEFLNAFGELFD-LQDEFPDG-VTLEVLEEALVGN-DSEGPLCELLFFFLTAIFQAIAE Q9UIG1
BAZ1B    hs 605 NTLFGDVAMVVEFLSCYSGLL--LPDAQYP--ITAVSLMEALSADKGGFLYLNRVLVILLQTLLQDEIA Q9UIG0
ACF1     dm 347 EHLLGDAFMMREFMHTYTGLLSGIEVFRQN--LSFYEMTRALTAR-EIAGPLSDILLVLLGTVFDLQKE Q9Y0W1
ZK783    ce 525 SQGFADALMVHEFVQNFGHVLG-IDLEIA---PKLESLCAGLDGDANHAEQTLQLTRQLLRLALEFPGM Q23590
H20J04   ce 473 NAEFEDYLFIFQFFNSFKQLLP-LKEIRGSDEVQFSDIIIAIKCNDPQNSSFADLLRVLLSIRTDIADE AAF39888
F3M18    at 658 DETVGNLLMVWRFLISFSDVLD-LWPFT------LDEFIQAFHDY-DSR-LLGEIHVTLLRSIIRDVED Q9SGP0
MLN1     at 515 DENVANLLMVWRFLITFADVLG-LWPFT------LDEFAQAFHDY-DPR-LMGEIHIVLLKTIIKDIEG BAB10985
MOI20    at 193 EEAVAHLLSVYGFLRSFSFQLY-ICPFE------LNDFVGALYFS-GPNSLLDAVHVALLRALKGHLER BAA98208
MAH20    at 298 MDCVGDLLMVWDFCTSFGRQLH-LWRFS------LEDFENAVCHKESNLVLIMEVHASLFRFLINERGD BAB10012
Ypl216   sc 376 QFPTERLLVVYQFLSFFGRFIG-LSHFN------FDQFLTTIKCT-SPEALVDEYVKINFLKTYNSKGS Q08964
F1N21    at 217 TEEAGNVCQLFEFCSAFGKALA-LKEGHAET-IVRELFICGRNTRRQQYCSITQMMIQLLDLISKDREM O49273
YGN3     sc 424 FDSFGKLLQAYQFLNTFGSKIC-LSHFS------LDQFITSLKCT-DPYELKGEVVLVNIRTQTSKEQE P53125
hypAT1   at 187 EEAVVYLLSVYGFLRSFSVQLY-ICPFG------LDDFVGALNFL-GPNSLLDAVHVALMRALKGHLER BAB11682
hypAT2   at 258 PEDAGNVFQFLEFCSAFGKALD-LRKGQAE--CVIREMLSGRSKRRQQYSTLTQMIIQLLTVILEDRGE BAB10159
Consensus(80%) ...h.phh.lhpFhpsFuphL..lp.hp......hpph..ul.sp.p....h.plhhhLLphhhpp...
sec.struc.pred ...hhHHHHHHHHHHHHHHhh.............HHHHHHHH.......hHHHHHHHHHHHHHHHh....
```

Abbildung 7. Multiples Alignment der DDT-Domäne für BPTF (bromodomain, PHD finger transcription factor) Transkriptionsfaktor - ähnliche Proteine (BPTF, CG17135, F26H11), BAZ (bromodomain adjacent to zinc finger) - Proteine (BAZ1A, BAZ1B, BAZ2A, BAZ2B, ACF1, ZK783, H20J04), HOX-Domänen-tragende Proteine (F3M18, MLN1), PHD-Domäne-tragendes Protein (MOI20) und andere hypothetische Proteine (MAH20 Ypl216 F1N21 YGN3 hypAT1 hypAT2). Erste Spalte: Protein-Namen; zweite Spalte: Spezies-Bezeichnung (at: *Arabidopsis thaliana*; ce: *Caenorhabditis elegans*; dm: *Drosophila melanogaster*; hs: *Homo sapiens*; sc: *Saccharomyces cerevisae*); dritte Spalte: erste Aminosäure der Domäne im jeweiligen Protein; letzte Spalte: Datenbank Accession Nummer. Konservierte geladene Aminosäuren sind rot markiert; konservierte hydrophobe Aminosäuren sind blau; zusätzliche konservierte Aminosäuren sind fett gedruckt. Die Konsensus-Sequenz (konservierte Aminosäuren in 80 % aller Sequenzen) befindet sich unter dem Alignment; h, p, u, s, l und - stehen für hydrophobe, polare, winzige, kleine (s =small), aliphatische und negativ geladene Aminosäuren. Die vorhergesagte Sekundärstruktur steht in der letzten Zeile (H, Helix vorhergesagt mit einer durchschnittlichen Genauigkeit > 82%); h, Helix vorhergesagt mit einer durchschnittlichen Genauigkeit < 82%)) (Rost et a. 1994).


DDT
PHD
PHD
BROMO
(ACF)
BAZ-family proteins,
C. elegans,
Drosophila
and human
MBD
DDT
PHD
BROMO
(BAZ2A)
DDT
PHD
PHD
BROMO
(BPTF)
BPTF
transcription
factor - like
proteins,
C. elegans,
Drosophila
and human
DDT
PHD
(CG17135
HOX
DDT
(F3M18)
Hox-domain-
containing proteins,
Arabidopsis
DDT
PHD
PHD
PHD
PHD
Hypothetical PHD-domain-
containing
(MOI20)
DDT
Other hypothetical
proteins,
0
10
20

Beschreibung von Abbildung 8. Domänen-Architektur von DDT-Domänen-tragenden Proteinen.
Die Abbildung beinhaltet nur Proteine unterschiedlichen modularen Aufbaus. Die Domännamen sind dem Simple Modular Architecture Research Tool (Schultz et al. 1998, Schultz et al. 2000) (http://smart.embl-heidelberg.de) entlehnt.
Abkürzungen: AT, AT_Hook, DNA-bindende Domäne mit Präferenz für A/T-reiche Regionen; BROMO, Bromodomäne; HOX, DNA-bindende Homeodomäne; MBD, Methyl-CpG-bindende Domäne; PHD, PHD C4HC3 Zinkfinger.

3.2.4. BSD, eine neue putativ DNA-bindende Domäne in Transkriptionsfaktoren, synapsen-assozierten und anderen hypothetischen Proteinen (Doerks et al. 2001)

Die RNA-Polymerase B benötigt zur Transkriptionsinitiation verschiedene akzessorische Proteine. Für die Transkritionsfaktoren das Säugerprotein BTF2 und das Hefehomologe TFB1, essentielle Komponenten dieses Initiationskomplexes sind keine funktionellen Bereiche näher charakterisiert (Fischer et al. 1992, Gileadi et al. 1992). Untersuchungen unterschiedlicher Regionen von BTF2 führten zur Entdeckung eines zentralen konservierten Abschnittes (Aminosäuren 180-232, siehe Abbildung 9). PSI-BLAST-Suchen (Altschul et al. 1997, Altschul et al. 1998) mit dieser Region zeigten schwache Sequenzähnlichkeit (E-value=0.14) zu DOS2-ähnlichen Proteinen und zu einer Proteinfamilie mit Homologie zu einem Synapsen-assoziierten Protein in *Drosophila* (E-value=2.4). Die Funktion der DOS2-Proteinfamilie ist unbekannt; Synapsen-assoziierte Proteine werden Neuron-spezifisch exprimiert und sind wichtige molekulare Bestandteile des Nervensystems (Reisch et al. 1995).
Reziproke Hidden Markov Model-Suchen (Eddy 1998) beginnend mit den Synapsen-assoziierten Proteinen und verwandten Homologen in Pflanzen (inklusive eines Ubox-Domänen-tragenden Proteins) wurden durchgeführt, um die Homologie zwischen diesen und den anderen Familien zu belegen.
Die erste HMMer-Suche zeigte signifikante Ähnlichkeit zu DOS2-ähnlichen Proteinen (E-value=7.2 x 10^{-7}) und in weiteren HMMer-Iterationen die erwartete Homologie zu

den BTF2-Transkriptionsfaktoren (E-value=$2.3x10^{-3}$), einem BTB/POZ-Domänen-tragenden (Zollman et al. 1994) Protein und anderen hypothetischen Proteinen in Protozoen (Fig.2). Die neu entdeckte Domäne wurde BSD nach den besser charakterisierten BTF2-Transkriptionsfaktoren, Synapsen-assoziierten und DOS2-ähnlichen Proteinen benannt.
Ein multiples Sequenz-Alignment wurde angefertigt (Thompson et al. 1994), um für alle Proteine die Domänengrenzen korrekt zu definieren (siehe Abbildung 9). Eine modulare zusammenhänge Sekundärstruktur ist auf den definierten Bereich beschränkt, die Domäne mündet N- und C-terminal in "low compositional complexity" Regionen (z.B. N-Terminus von T16K5.150 (185-199) und C-Terminus von DOS2 (240-265).

Die BSD-Domäne hat eine durchschnittliche Länge von 60 Aminosäuren; Sekundärstrukturvorhersagen mit PHD (Rost et al. 1994) gehen von drei Helices aus, die sich in kleinen Domänen häufig zu einem "three helical bundle" falten. In der dritten vorhergesagten Helix sind ein Phenylalanin und ein Tryptophan in jeder BSD - Domänen-Sequenz konserviert.
Die BSD-Domäne ist in vielen Spezies von einfachen Protozoen bis zum Menschen verbreitet, was eine grundlegende und wichtige Funktion annehmen lässt.
Das Vorkommen der Domäne in BTF2-Transkriptionsfaktoren deutet auf eine Rolle in Chromatin-assoziierten Prozessen hin. Diese Vermutung wird durch die modulare Architektur hypothetischer Homologer gestützt (siehe Abbildung 10). Die Ubox-Domäne ist eine RING-Finger ähnliche Domäne, die in Ubiquitinierungsvorgänge involviert ist (Aravind et al. 2000); eine nennenswerte Anzahl von Ubiquitinierungsproteinen steht in Verbindung mit Chromatin-assoziierten Prozessen (Hershko et al. 1998, Ciechanover et al. 1998). In einem anderen Fall geht der BSD-Domäne eine BTB-Domäne voraus; eine Protein-Protein-Interaktionsdomäne, die häufig in Transkriptionsfaktoren gemeinsam mit DNA-bindenden C2H2-Zinkfingern vorkommt (Zollman et al. 1994, Bardwell et al. 1994).
Zusammenfassend rechtfertigen die Resultate die Annahme, daß die BSD-Domäne DNA-bindende Funktion hat.

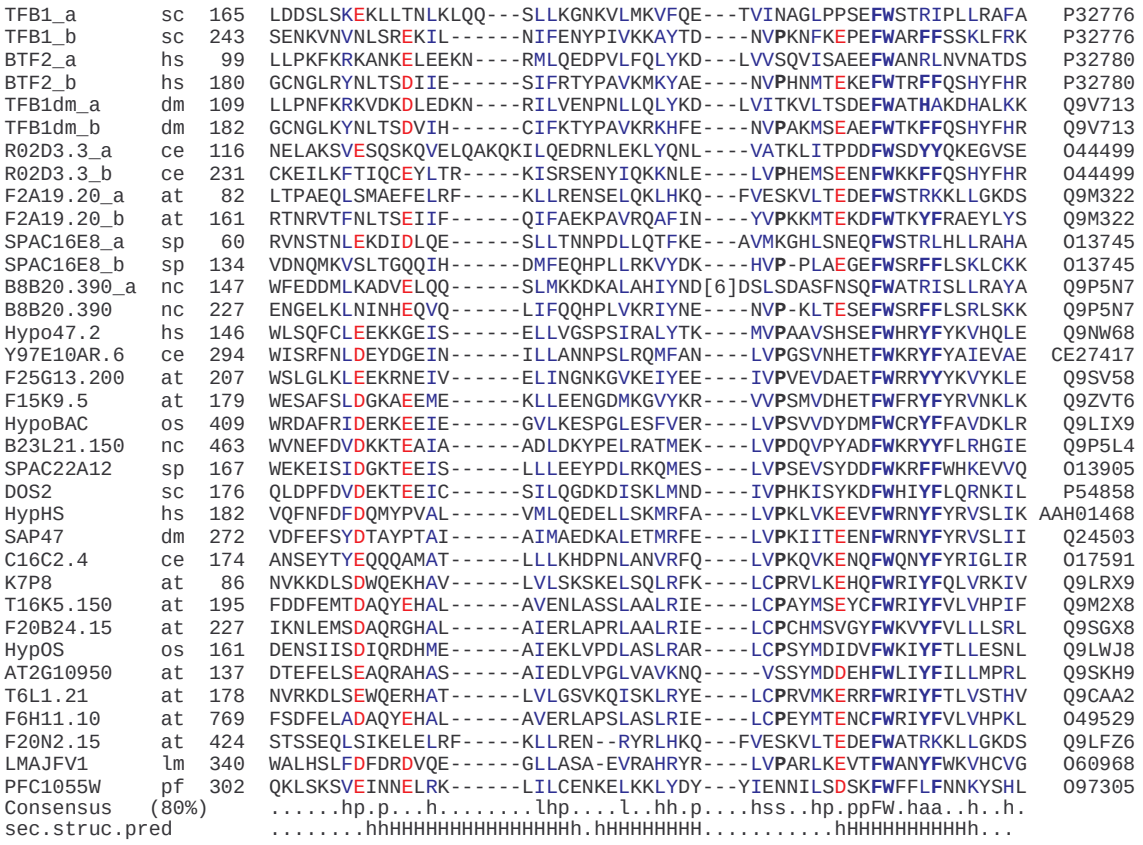


Abbildung 9. Multiples Alignment der BSD-Domäne für BTF2-Transkriptionsfaktoren (TFB1, BTF2, TFB1dm, R02D3.3, F2A19.20, SPAC16E8, B8B20.390), DOS2-ähnliche Proteine (Hypo47.2, Y97E10AR.6, F25G13.200, F15K9.5, HypoBAC, B23L21.150, SPAC22A12, DOS1), Proteine, homolog zu einem synapsen-assoziierten Protein (HypHS, SAP47, C16C2.4, K7P8, T16K5.150, F20B24.15, HypOS, AT2G10950, T6L1.21 und mit einer N-terminalen UBOX (F6H11.10), einem BTB-Domäne-tragenden Protein (F20N2.15) und andere hypothetische Proteine (LMAJFV1, PFC1055W). Erste Spalte: Protein-Namen (wiederholt auftretende Domänen sind mit a und b gekennzeichnet); zweite Spalte: Spezies-Bezeichnung (at: *Arabidopsis thaliana*; ce: *Caenorhabditis elegans*; dm: *Drosophila melanogaster*; hs: *Homo sapiens*; lm: *Leishmania major*, nc: *Neurospora crassa*, os: *Oryza sativa*, pf: *Paramecium falciparum;* sc: *Saccharomyces cerevisae;* sp: *Saccharomyces pombe*); dritte Spalte: erste Aminosäure der Domäne im jeweiligen Protein; letzte Spalte: Datenbank Accession Nummer. Konservierte negativ geladene Aminosäuren sind rot markiert; konservierte hydrophobe Aminosäuren sind blau; zusätzliche konservierte Aminosäuren sind fett gedruckt. Die Konsensus-Sequenz (konservierte Aminosäuren in 80 % aller Sequenzen) befindet sich unter dem Alignment; h, p, s, l und a stehen für hydrophobe, polare, kleine (s =small), aliphatische und aromatische Aminosäuren. Die vorhergesagte

Sekundärstruktur steht in der letzten Zeile (H, Helix vorhergesagt mit einer durchschnittlichen Genauigkeit > 82%); h, Helix vorhergesagt mit einer durchschnittlichen Genauigkeit < 82%)) (Rost et al. 1994).

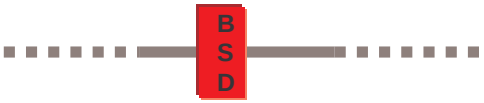
BTF2-like transcription factors
Yeast, Arabidopsis, C. elegans, Drosophila and human

Arabidopsis,
C. elegans, Drosophila and human

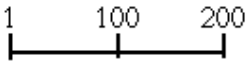
Proteins with homology to
Synapse-associated proteins of
Drosophila

Arabidopsis,

BTB-domain-containing proteins
Arabidopsis,



Dos2-like proteins in Yeast, *Arabidopsis, C. elegans* and human and other hypothetical proteins *Paramecium* and *Leishmania*



Beschreibung von Abbildung 10. Domänen-Architektur von BSD-Domänen-tragenden Proteinen.
Die Abbildung beinhaltet nur Proteine unterschiedlichen modularen Aufbaus. Die Domänennamen sind dem Simple Modular Architecture Research Tool (Schultz et al. 1998, Schultz et al. 2000) (http://smart.embl-heidelberg.de) entlehnt.
Abkürzungen: BTB, Broad-complex, Tamtrack and Bric a Brac; Ubox, eine modifizierte Ring Finger-Domäne assoziiert mit Ubiquitination.
Es sei angemerkt, daß die Bezeichung synapsen-assoziiertes Protein irreführend sein kann. Die Lokalisation wurde nur für *Drosophila* (Reisch et al. 1995) experimentell nachgewiesen; verwandte Homologe wurden auch in Spezies ohne Synapsen gefunden.

3.3. Automatische Analyse unbekannter Regionen in nuklearen Proteinen - Detektion 28 neuer Domänen-Familien (Doerks et al. accepted)

107 Domänen in Smart sind als überwiegend nuklear definiert (Stand Oktober 2000). Hierfür wurde das Computer-Programm Meta-A(nnotator) (Eisenhaber et al. 1998) verwendet, welches die Lokalisation aus der Swiss-Prot-Annotation herausliest. Domänenfamilien, deren Proteine in mehr als 80% der Swiss-Prot-Einträge nuklear annotiert sind, wurden in dieser Analyse berücksichtigt. Aus 86 Domänen wurden 11 falsch positive eliminiert und weitere 32 überwiegend nukleare Domänen aus der Literatur entnommen. Mit der RC-Methode (Doerks et al. accepted) wurden alle unbekannten Regionen, die nicht von einer bekannten Smart-Domäne abgedeckt werden, aus 24000 überwiegend nuklearen Proteinen in nrdb extrahiert.
Sequenzabschnitte von weniger als 30 Aminosäuren Länge wurden nicht berücksichtigt, da Faltunsgvermögen zu einer eigenständigen Domäne wenig wahrscheinlich ist.
Die extrahierten Zwischen-Domänen-Bereiche wurden mit dem Programm grouper des SEALS Programmpaketes (Walker et al. 1997) mit einem "single linkage clustering"-Grenzwert von 50 Bits in homologen Gruppen zusammengeführt. Der

jeweils längste Vertreter einer Gruppe wurde auf das Vorhandensein von "coiled-coil" (Lupas et al. 1991) und "low complexity" (Wootton et al. 1996) überprüft und zu einer

PSI-BLAST-Suche (Altschul et al. 1997, Altschul et al. 1998) herangezogen; es wurden acht iterative Suchen für Sequenzen mit einem Grenzwert von E< 0.001 durchgeführt.

Regionen, die in unterschiedlichem Domänenkontext in mehr als 5 Proteinen auftraten, wurden zur weiteren Analyse bereitgestellt (siehe Abbildung 11).

Um sicherzustellen, dass es sich bei den detektierten Bereichen um bisher unbekannte Module handelt, wurde nachgeprüft, ob sie in der Pfam-Domänendatenbank (Bateman et al. 2000) oder der Literatur beschrieben sind; wenn möglich wurden zusätzliche Genomanalysen durchgeführt, um künstliche Genfusionierungen auszuschliessen (z.B. nukleare Domäne nur in einem Fall mit einer extrazellulären Domäne assoziiert).

Von den neu entdeckten Domänen wurden multiple Alignments angefertigt (Thompson et al. 1994), die genauen Domänengrenzen bestimmt und zusätzliche Hidden markov Model-Suchen (Eddy 1998) durchgeführt, um mögliche weitere Familienmitglieder zu identifizieren

Ergebnis ohne iterative PSI-Blast-Suche:	Ergebnis mit iterativen PSI-Blast - Suchen:
aus ~15000 Sequenzabschnitten	aus ~15000 Sequenzabschnitten
↓	↓
~ 4000 cluster	~10000 cluster
↓	↓
~ 150 in unterschiedlichem Domänen-Kontext	~ 400 in unterschiedlichem Domänen-Kontext
↓	↓
~ 8 neue Domänen	~ 20 neue Domänen

Abbildung 11. Ergebnis der automatischen und semi-automatischen Analyse der Domänensuche mit und ohne PSI-BLAST-Iterationen.

3.3.1 Klassifikation

Es werden neue Domänen in verschiedenen Proteinfamilien in unterschiedlichen Spezies, Spezies-spezifische Domänen, Domänen mit entfernter, nicht signifikanter Ähnlichkeit zu bereits annotierten Domänen, und Extensionen, die stets mit einer spezifischen Domäne assoziiert sind, unterschieden.
Die automatische Analyse nuklearer Proteine führte zum Auffinden folgender Domänen:

i) 15 neue Domänen in verschiedenen Proteinfamilien in unterschiedlichen Spezies und molekularen Kontexten (siehe 3.3.2, Tabelle 1). Drei dieser Domänen wurden während Erstellung dieser Arbeit publiziert (Doerks et al. 2001, Clissold et al. 2001, Callebaut et al. 2001)
ii) 3 Domänen wurden in einer oder sehr eng verwandten Spezies gefunden (siehe 3.3.3, Tabelle 2)
iii) 7 weitere Domänen-Familien ähneln nicht signifikant bereits beschriebenen Modulen (eine Domäne wurde während Erstellung dieser Arbeit publiziert (Aravind 2000)) (siehe 3.3.4, Tabelle 3)
iv) 3 Domänen sind mit einer spezifischen Domäne assoziiert und somit N- oder C-terminale Extensionen (siehe Tabelle 4)

Damit führte die Analyse zu der Entdeckung von 28 neuen Domänen, in mehr als 1200 Proteinen.

3.3.2. 15 neue Domänen in verschiedenen Proteinfamilien in unterschiedlichen Spezies

Die abschliessende Analyse führte zur Entdeckung von 15 neuen Domänen, die keinerlei Ähnlichkeit zu bekannten Domänen zeigen und in unterschiedlichen

Spezies auftreten. Die vorhergesagten Funktionen erstrecken sich von Metall-Protein- oder Nukleotidbindung bis zu katalytischer Aktivität.

Domäne	Beschreibung	Länge in AS	Sek. Strukt Vorh.	Vorherges. Funktion	Anzahl proteine	Assoziierte Domänen	Spezies	Acc.Nr. einer repräsentativen Sequenz (Domänen-Grenzen)
JmjC*	Jumonji related family	100	β	Metallo-Enzym	140	BRIGHT, jmjN PHD, FBOX, LRR, C2, TPR PLAc, CXXC†, ZnF_C2H2	Eu, y, a, c, d, h	O14607§ (1042-1205)
CSZ	Domain in chromatin remodeling S1 domain containing and Zinc finger proteins	750	α / β	DNA-Bindung, Chromatin-Modulation	35	S1, SH2, C2HC, HhH	Eu, y, a, c, d, h	P34703 (389-1120)
RPR	Proteins involved in regulation of nuclear pre-mRNA	120	α	Protein-Interaktion	40	RRM, PWWP, SURP†, G-Patch	y, a, c, d, h	Q9SJQ7 (88-225)
DDT*	Different transcription and chromosome remodeling factors	60	α	DNA-Bindung	30	AT_Hook, PHD, HOX, BROMO, MBD	y, a, c, d, h	Q9UIG2§ (102-161)

TLDc	TBC, LysM and other proteins	220	α / β + β	enzymatisch	30	TBC, LysM, R3H, FBOX	y, a, c, d, h	Q9VNA1§ (1163-1325)
PUG	Protein kinases, UBA or UBX domain containing proteins and glycanases	60	α /β	RNA-Bindung	25	C2H2, UBA, TGc, UBX, S_TKc, STYKc	y, a, c, d, h	Q9MAT3 (323-386)
HSA	Helicases and SANT domains	70	α	DNA-Bindung	20	SANT, BROMO DEXDc, HELIc	y, a, c, d, h	P25439§ (501-573)
PSP	Proline-rich, in spliceosome associated proteins	60	α	RNA - oder snRNP-Bindung	15	SAP, C2HC	y, a, c, d, h	O16997 (299-357)
FYRN	Trithorax and X-chromosome inactivating proteins	40	α / β	unbekannt	25	PHD, SET, PWWP	a, c, d, h	Q24742§ (1869-1914)
FYRC	Trithorax and X-chromosome inactivating proteins	90	α / β	unbekannt	25	PHD, SET, PWWP	a, c, d, h	Q24742§ (3495-3583)
RUN*	TBC, PH, FYVE and other proteins	65	α	GTPase signalling	40	DENN†, TBC, PLAT, PH, C1, FYVE, GST, SH3	c, d, h	BAB14033 (115-178)
BRK	Transcription factors and CHROMO domain helicases	50	α / β	unbekannt	20	CHROMO, PHD, SANT, TFSM2, DEXDc, BROMO	c, d, h	O15025§ (882-931)
DZF	DSRM or ZnF_C2H2 domain containing proteins	250	α / β	unbekannt	40	C2H2, DSRM	c, d, h	O88531 (762-1016)

NEUZ	Domain in neuralized-like proteins	120	β	unbekannt	10	SOCS, RING, SPRY, SH2	c, d, h	Q19299 (199-321)
ZnF_TTF	Domain in transposases and transcription factors	100	α + β	Metall-Bindung	20	KRAB, BTB	a, d, h	Q9ZWT4 (100-199)

Beschreibung von Tabelle 1. Tabelle der neuen Domänen in unterschiedlichen Spezies. Erste Spalte: Domänen-Name; zweite Spalte: Beschreibung der Domäne (z.B. assoziierte Domänen oder genauer charakterisierte Proteine); dritte Spalte: ungefähre Länge der Domäne (Anzahl der Aminosäuren); vierte Spalte: Sekundärstrukturvorhersage (Roste et al. 1994) (α: Domäne besteht aus α-Helices; β: Domäne besteht aus β-Faltblättern; α/β: Domäne besteht aus α-Helices und β-Faltblättern; fünfte Spalte: vorhergesagte Funktion der neuen Domäne; sechste Spalte: Anzahl der Proteine, die die neue Domäne tragen; siebte Spalte: Namen der assoziierten Domänen (die Domänen-Namen leiten sich aus dem Simple Modular Architecture Research Tool (http://smart.embl-heidelberg.de) (Schultz et al. 1998, 2000) ab oder sind definiert in Pfam (Bateman et al. 2000)†). Achte Spalte: Spezies, in denen die neue Domäne auftritt; eu: Eubakterien; virus: Viren; y: Hefe (yeast); a: *Arabidopsis thaliana*; c: *Caenorhabditis elegans*; d: *Drosophila melanogaster*; h: *Homo sapiens*); neunte Spalte: Accession Nummer eines repräsentativen Proteins und die Region, in der sich die neue Domäne befindet.

* neue Domäne wurde bereits publiziert oder ist "in press".

§ zusätzliche HMMer-Suchen (Eddy 1998) waren notwendig, um alle Proteine, die die neue Domäne tragen, zu identifizieren

3.3.2.1 Die PUG-Domäne, befindlich in N-Glykanasen und anderen nuklearen Proteinen

Die PUG-Domäne repräsensiert beispielhaft die neuen Domänen, die in vielen Spezies in unterschiedlichen molekularen Kontexten auftreten.

Ein hypothetisches *Arabidopsis* Protein (Acc. Nr.: Q9MAT3) trägt zwei Zink-Finger-Motive (ZnF_C2H2), gefolgt von einer UBA-Domäne (Hofmann et al. 1996). Im C-terminalen Bereich nach einer vorhergesagten "coliled coil"-Region sind keine weiteren Domänen beschrieben. PSI-Blast-Suchen mit diesem C-terminalen Sequenzabschnitt (siehe Abbildung 12) zeigen signifikante Ähnlichkeit (E-value<10^{-5}) zu UBX-Domänen-tragenden Proteinen und zu N-Glykanasen (PNGasen) in Metazoen. Bei orthologen Glykanasen in *S. cerevisiae*, *S. pombe* und *Arabidopsis* fehlt diese Domäne, was eine bedeutsame Rolle in vielzelligen *Animalia* annehmen

läßt. Nur die *S. cerevisiae* PNGase ist experimentell untersucht; sie ist im Nukleus lokalisiert, tritt aber auch in geringer Menge im Cytosol auf (Suzuki et al. 2000).
Die neu entdeckte Domäne wurde PUG nach den besser charakterisierten Peptid:N-Glykanasen und anderen möglichweise nuklearen UBA- oder UBX-Domänen-tragenden Proteinen benannt. PNGasen sind in die Antwort auf ungefaltete Proteine (Unfolded Protein Response (UPR) involviert (Suzuki et al. 2000); die UPR führt zu einer Erhöhung der Transkription von Proteinen des Endoplasmatischen Retikulums und zu einer Anreicherung ungefalteter Proteine im ER. Die PUG-Domäne tritt mit zwei Domänen (UBA und UBX) assoziiert auf, die eine zentrale Rolle in der Ubiquitin-vermittelten Proteolyse spielen. Die PUG-Domäne scheint somit ein Bindeglied zwischen der ER-assoziierten Reaktion auf ungefaltete Proteine und dem Proteinabbau durch Ubiquitination zu sein.
Ergänzende HMMer-Suchen zeigten schwache Sequenzähnlichkeit zu IRE1p-ähnlichen Kinasen (Acc. Nr.: Q9SHL6) (*E*-value: 0.21) in einer Region homolog zu dem C-terminalen Schwanz von 2'-5'-oligo(A)-abhängigen Ribonukleasen (Zhou et al. 1993) (Siehe Abbildung 13). Obwohl die Ähnlichkeit nicht signifikant ist, ist die Übereinstmmung beachtlich, berücksichtigt man den Aspekt, dass die IREp-Kinasen bei der UPR eine wesentliche Rolle spielen (Shamu et al. 1996). Der C-terminale Schwanz von IRE1p wird für die Induktion der UPR benötigt (Shamu et al. 1996) und zeigt Endoribonuklease-Aktivität (Sidrauski et al. 1997). Diese Aktivität stimmt überein mit der C-terminalen Lokalisierung der RNase-Aktivität in homologen 2'-5'-oligo(A)-abhängigen Ribonukleasen (Bork et al. 1993). Dieses führt zu dem Schluss, dass eine divergente PUG-Domäne im C-terminalen Bereich von IREp vorhanden ist.

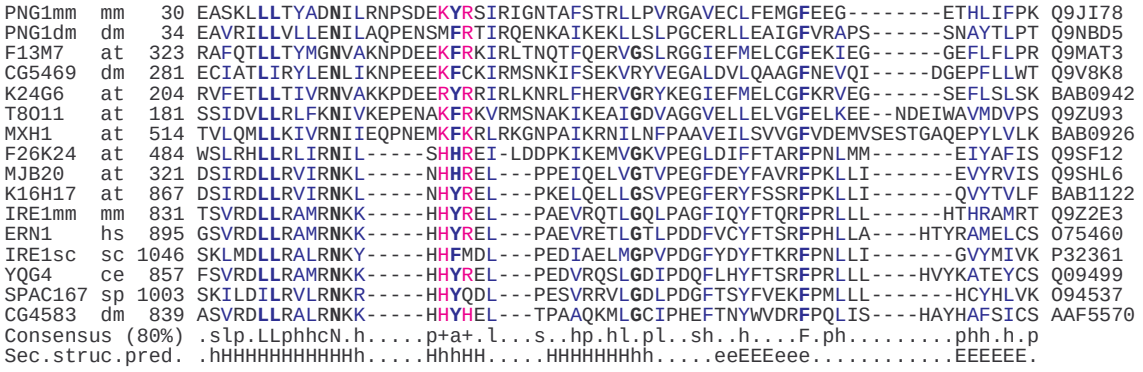
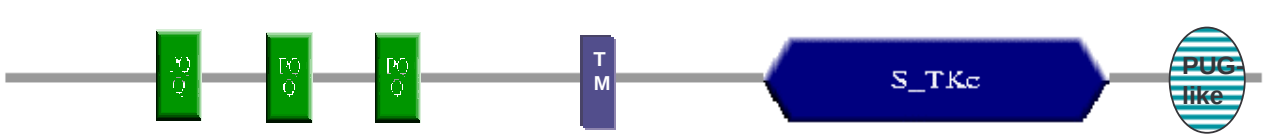
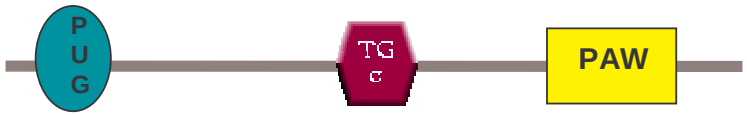


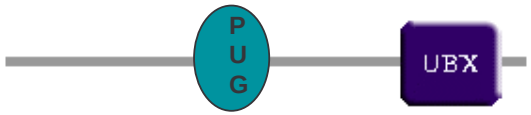
Abbildung 12. Multiples Alignment der PGN-Domänen in N-Glykanasen (PGN1mm, PGN1gm), UBX-Domänen-tragende Proteine (F13M7, CG5469), HOX-Domänen-tragende Proteine (F3M18, MLN1), UBA/Zink-Finger-Domänen-tragende Proteine (K24G6, T8011), hypothetische Zink-Metalloproteinase (MXH1) und Serin/Threonin-Protein-Kinasen (F26K24, MJB20, K16H17, IRE1mm, ERN1, IRE1sc, YQG4, SPAC167, CG4583). Erste Spalte: Protein-Namen; zweite Spalte: Spezies-Bezeichnung (at: *Arabidopsis thaliana*; Ce: *Caenorhabditis elegans*; dm: *Drosophila melanogaster*; hs: *Homo sapiens*; mm: *Mus musculus*; rn: *Rattus norvegicus*; sc: *Saccharomyces cerevisae*; sp: *Saccharomyces pombe*); dritte Spalte: erste Aminosäure der Domäne im jeweiligen Protein; letzte Spalte: Datenbank Accession Nummer. Konservierte positiv geladene Aminosäuren sind pink markiert; konservierte hydrophobe Aminosäuren sind blau; zusätzliche konservierte Aminosäuren sind fett gedruckt. Die Konsensus-Sequenz (konservierte Aminosäuren in 80% aller Sequenzen) befindet sich unter dem Alignment; h, p, u, s, l und + stehen für hydrophobe, polare, winzige, kleine (s =small), aliphatische und positiv geladene Aminosäuren. Die vorhergesagte Sekundärstruktur steht in der letzten Zeile (H, Helix bekannt oder vorhergesagt mit einer durchschnittlichen Genauigkeit > 82%); h, Helix bekannt oder vorhergesagt mit einer durchschnittlichen Genauigkeit < 82%)) (Rost et al. 1994).



Serine/Threonine protein kinases,
Yeast, *Arabidopsis, C. elegans, Drosophila* and human

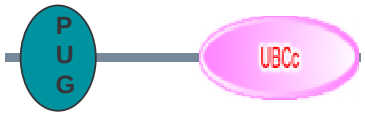


N-glycanases ,
Drosophila and mouse



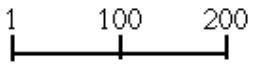
UBX-domain-containing proteins,
Arabidopsis and *Drosophila*

UBA/Zinc-Finger-domain-containing proteins,
Arabidopsis, Plasmodium



UBCc-domain-containing protein,
Plasmodium

Hypothetical zinc metalloproteinase ,
Arabidopsis



Beschreibung von Abbildung 13. Domänen-Architektur von PUG-Domänen-tragenden Proteinen.

Die Abbildung beinhaltet nur Proteine unterschiedlichen modularen Aufbaus. Die Domänennamen sind dem Simple Modular Architecture Research Tool (Schultz et al. 1998, 2000) (http://smart.embl-heidelberg.de) entlehnt.

Abkürzungen: C2H2, Zink-Finger C2H2 DNA-bindende Domäne; PQQ, "beta-Propeller repeat"; S_TKc, Serin/Threonin Protein-Kinase katalytische Domäne; TGc, Transglutaminase/Protease katalytische Domäne; UBA, Ubiquitin assoziierte Domäne; UBX, Domäne in Ubiquitination regulierenden Proteinen.

3.3.3. Drei neue Spezies-spezifische Domänen

Die Spezies-spezifische Radiation einer Domänen-Familie, also die auffällige häufige Verbreitung einer Domäne in einem Genom verglichen mit anderen Genomen ist ein weit verbreitetes Phänomen (International Human Genome Consortium 2001). In extremen Fällen ist es nicht möglich, Ähnlichkeit zu einer Domäne in einer Spezies über diese hinaus zu entdecken. Dies kann auf eine Genom-spezifische Neuentstehung oder wahrscheinlicher auf eine gesteigerte Geschwindigkeit der molekularen Evolution (keine Sequenzähnlichkeit ist mehr detektierbar) zurückzuführen sein. Ein alternatives Szenario des enormen Verlustes der Domäne in vielen Spezies scheint weniger wahrscheinlich.

3 (~11%) Domänen sind phylogenetsich auf eine Art bzw. Gattung beschränkt (ausgenommen Extensionen, siehe Tabelle 2).

Domäne	Beschreibung	Länge in AS	Sek. Strukt Vorh.	Vorherges. Funktion	Anzahl proteine	Assoziierte Domänen	Spezies	Acc.Nr. einer repräsentativen Sequenz (Domänen-Grenzen)
FBD	Domain in FBOX and other domain containing plant proteins	80	α / β	unbekannt	160	FBOX, LRRcap, BRCT, AAA	a	Q9LXJ7 (304-382)
ZnF_PMZ	Plant mutator transposase zinc finger domain	27	α / β	Metall-Bindung	125	AT_Hook, ZnF_C2HC, PHD	a	Q9SH73 (3212-3239)
SPK	SET and PHD domain containing proteins and protein kinases	120	α / β	Protein-interaktion	40	SET, ICE_p10†, ICE_p20†, ZnF_C2HC, PHD, STYKc	c	Q9XU06 (139-250)

Tabelle 2 Tabelle der neuen Domänen in unterschiedlichen Spezies. Erste Spalte: Domänen-Name; zweite Spalte: Beschreibung der Domäne (z.B. assoziierte Domänen oder genauer charakterisierte Proteine); dritte Spalte: ungefähre Länge der Domäne (Anzahl der Aminosäuren); vierte Spalte: Sekundärstrukturvorhersage (Rost et al. 1994) (α: Domäne besteht aus α-Helices; β: Domäne besteht aus β-Faltblättern; α/β: Domäne besteht aus α-Helices und β-Faltblättern; fünfte Spalte: vorhergesagte Funktion der neuen Domäne; sechste Spalte: Anzahl der Proteine, die die neue Domäne tragen; siebte Spalte: Namen der assoziierten Domänen (die Domänen-Namen leiten sich aus dem Simple Modular Architecture Research Tool (http://smart.embl-heidelberg.de) (Schultz et al. 1998, 2000) ab oder sind definiert in Pfam (Bateman et al. 2000)†). Achte Spalte: Spezies, in denen die neue Domäne auftritt; a: *Arabidopsis thaliana*; c: *Caenorhabditis elegans*; neunte Spalte: Accession Nummer eines repräsentativen Proteins und die Region, in der sich die neue Domäne befindet.

3.3.3.1 Die SPK-Domäne in Nematoden-spezifischen Proteinen

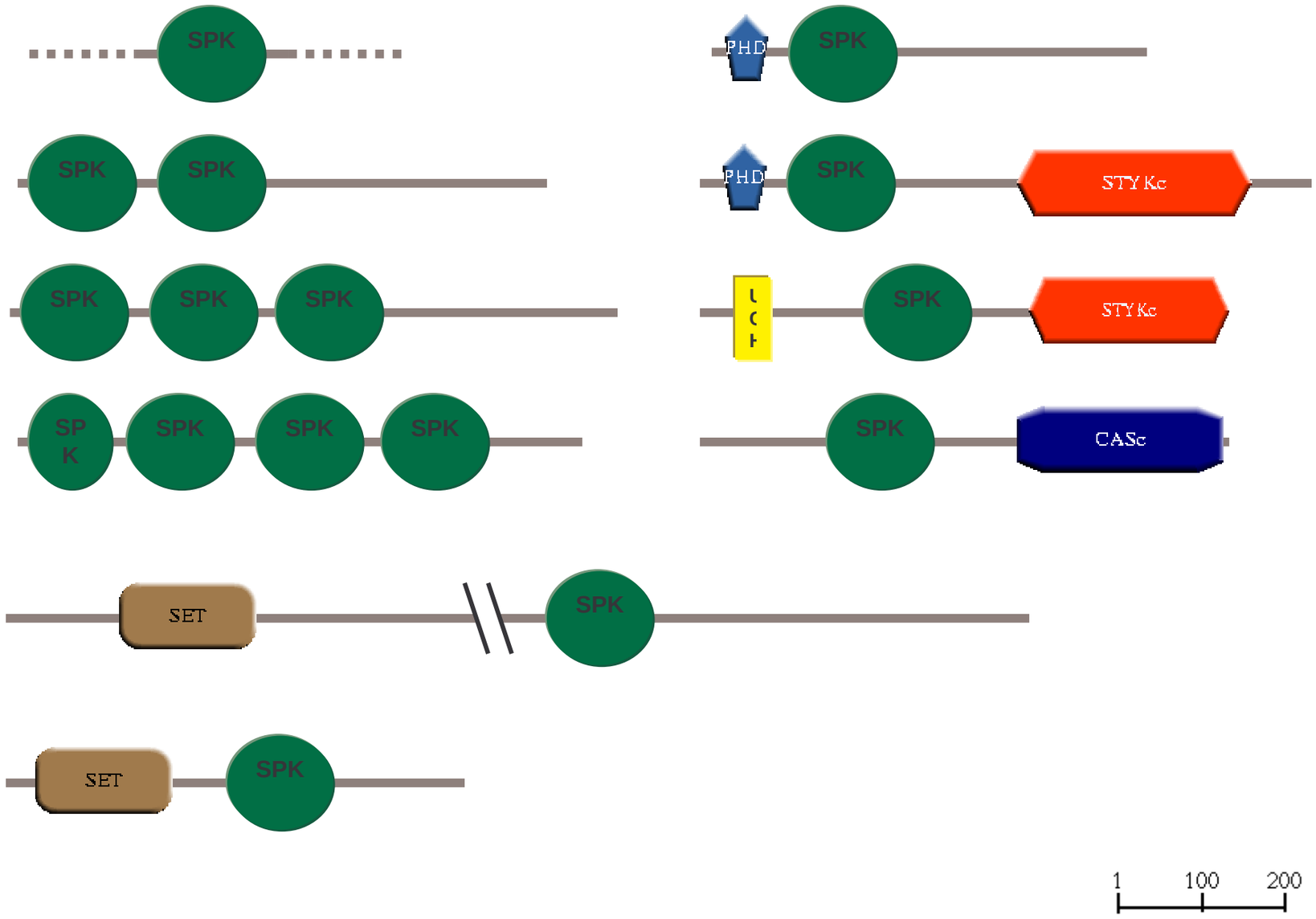
PSI-BLAST-Suchen mit einer Region C-terminal zu einer SET-Domäne (Cui et al. 1998) (siehe Abbildung 14) eines hypothetischen Proteins Y43F11A.5 (Acc. Nr.: Q9U2G8) führten zur Entdeckung einer neue Domäne in diversen *C. elegans*-Proteinen jedoch in keiner anderen Spezies. Die Domäne ist ungefähr 120 Aminosäuren lang und tritt assoziiert mit der katalytischen Domäne von Caspasen (CASc), einer unspezifischen Proteinkinase-Domäne (STYKc), als auch mit einer SET Methyltransferase-Domäne auf. Es finden sich häufig multiple Tandem-Kopien der Domäne in der gleichen Sequenz (siehe Abbildung 15). Die neu entdeckte Domäne erhielt den Namen SPK (asssoziiert mit SET, PHD (Aasland et al. 1995) und Proteinkinasen).

```
Y14H12D.2     23  DMSHFLDFLAKKSKN---ISRPLML-KDLFSAYKEEAGYPGTVATLRLKLRWDLAVKIPLAANFEDDEKAQMLFATSTSAK-EEFLKRLREK-
C08F11.7      46  IPSNFLFFLKQQTKD---IQKPLEL-RSLFRGYIAYAKSRKQMETMMRLIIPQLSKTVEETTDFSDQEKVQMIFGARIRMN-QSFWERFKPS-
Y57A10A.1    511  EMEKCMDFLVEIIDN---FTDPVIKTDIWKLYYSRMKP-DVSEKVINNRFQSKLAPIIHRLDNYSIETRVRIMFVMGVPVE-AGFLKNLRKT-
Y49E10.7      56  EFVRFVQFLVNKTEK---SNEPIDQ-RRVFGKFRALEHGVLDVTTYRNRFHGVVARNMILLDNLSIDLSIRMIFALSGKME-TNFLSEVEKH-
ZK402.5       78  DDVRLMTFLVEKTKE---ANEPLVA-TKVFMEFGKKENARCSDGAYYRKFHKKLAPNMDQLDNYSIESRLRVMLGLVGEVS-DDFLTQVQTE-
T28A8.3_a      7  QLERLMSFLVEQTKD---SIEPLVV-LKVFTEYSNRENDGLSYRFYYDRFRTSVALNMAKLTEYTIEDRIRVMFGFAGEVS-DDFLTKIETI-
T28A8.3_b    139  DLVRFMDFLVEKTKD---TIVPVAA-CKVLTEYSKRENDGLSYSVYYRRFRRSVAPNMAKFENYSIEERVRIMFGFAGKVA-DDFLEQIKTE-
Y57A10A.7_a    4  PLGRLINFLVEWTKN---VVEPLVG-GRIFNRFAELDGAGLSHIYYYIRFHEHIAPNIAKWDNFNVTTRIRMMFALSGEVP-DDFLALIKLT-
Y57A10A.7_b  133  DLPRLMNFLVEKTKD---ATEPMVS-IAELSEFRRRERSELTEQAYYGKFHRQLAPKMGQIVNYSIEERVRVMFGFAGEVT-DDFLKQIQTI-
Y57A10A.7_c  265  EITRFMNFLVEKAKV---STDPMFA-SVVFTEFRKSEEDGLAYSTYCRKFYNYLAPRMDQFVNYSIEERVRLMFGLGGEVT-EDFLKLCRKE-
B0205.1_a      6  DLKIFMDFLVEKTKD---AFQPMIA----------------------------------------VQARIRLMYALGGKVE-SDFLERIETH-
B0205.1_b     82  HFTRFMDFLIQKTKY---AVEPMNF-NQVLEEFCRLEPDCRHYGVYYVRFHHKLAPNMDALNNYNIYDRIRLMFVLNGKVS-GDFFKTIKTH-
B0205.1_c    197  SDTRLMDFLIQKTRH---SVEPMQATNLIFKEFSERPS-RILDKMYQLRFYQKLAPNMNEWNRYSIEQRIRLMFVLKGKVA-DDFLKTIPILS[13]
B0205.1_d    332  TNIRFMDFLIQNTKD---GAESIRL---IYNEFALLEGNVLSASTYSSRFCKKLAPKMSQWSNYRIEDRIRMMFALKGKVE-SDFLAQMIQS-
C47E8.8     1118  ELERMIDFLVEITEK---ISVPAIK-TDIWKLYQARMKQNVKEECIRQRFMSKLAPIIHRLDNYSIEARVRIIFVLSVKVD-ADFLKELRKN-
Y43F11A.5    242  ADVDFLQFLAQKSSSD--EIRPLFR-FQICEEYIDTYENPSSENYLNNRFLRVLAMRIPLLRGFDLETKARMMFVAGIPLE-ERFLIELRRY-
C16A11.4     113  NLEEFLEFLSEASEN---ISRPLAL-TTLFEDYKEHINYPQSVRILRKQLESNLLETISMSPNYYADRKAQMLFAMGLRVK-GEFLERLERN-
C16A11.3     106  YIQRFLDFLSDKSEN---IKSPLPL-TKFFQDYKTKSNCSQSLATAKKKLLANLFEHIVMSAKYNDDRKAEMLFAIGLRVE-DEFLERLRGR-
Y48E1B.13    154  EMERFLKFLKDKTKQVEKRKEPFSQ-KEIYAVFQRRIKSELCIETVKKKFQPLLPN-AIQTCEFDEETMIRMIYGAGIRIDSVDFWNRFTSK-
H12I13.1     216  YLNQCIKFLHTKSEN---VGKPFVR-TTFYANFKNVVY-YKGTATTLGRGLQPLLHDAIISSEYKDFRKCQMLFVMSIQITNQSFLYRLAKNS
Consensus (80%)   ....hhpFLhppo.p......Ph.h...hh..a..h....hp...h..+hh..ls.ph...spap.pp+h+hhFshs..lp..-FLpplppp.
Sec.struct.pred.  .hHHHHHHHHHHHHh.......hhHHHHHHHHHHHHHh.....HHHHHHHHHHHHHHHHHHHHh...HHHHHHHHHHh.hhhhHHHHHHHHHh..

Y14H12D.2         ATVDVDNLQRIT--YYKSTK-------LEFKG   Q9TYP8
C08F11.7          AHLTLDKFSRLA--FFKSDT-------LLLIA   Q9U3Q7
Y57A10A.1         AVVRVDENQMIT--KYVACA---DDGGLKLEG   Q9U209
Y49E10.7          GIVVLNDGKRIC--EYASRD-----GKLKLEA   Q23477
ZK402.5           GIVKLDEKKRIC--KFTSHD-----GKLKLEA   Q9XU06
T28A8.3_a         GVVELDDGKRIC--KYASHD-----GKLKLEG   Q9XU06
T28A8.3_b         GTVQLDAKRRIS--KYASND-----GNLKLEG   Q9NA85
Y57A10A.7_a       GTVHLDEKKRIC--KYATHD-----GKLKLEG   Q9NA85
Y57A10A.7_b       GVVELDKNNRIC--EYISYD-----GTLKLGA   Q9NA85
Y57A10A.7_c       GTVELNKKRKIT--KYTSND-----GILSLQD   O61746
B0205.1_a         GTVELDRKKKIT--KYVSND-----GKLKLGG   O61746
B0205.1_b         GTVELDEKRRIR--KYTSND-----GKLSLKK   O61746
B0205.1_c         GAVQLDENYRIV--RYTSNN-----GKVELSE   O61746
B0205.1_d         GIVRLDNQNRII--EYVAND-----GKMKLKG   Q9XTT7
C47E8.8           AVVRVDDEQLIT--KYVSCN----DGGLKLEG   Q18690
Y43F11A.5         GSVELDEKHRIT--NFKANN-----GDFTLKG   Q9U2G8
C16A11.4          AVVEVDGYHRIT--FYKSKN-------LEFRG   O76580
C16A11.3          ATVEVDDCQRIT--FYKSKTSTSSPETVSFVL   O76581
Y48E1B.13         ATISLDCYSRLI--SYSSDS-------LTLSG   O18203
H12I13.1          FSITLDRFYRIEVCEYQSKK---FSGVHKCPG   Q9N5M9
Consensus (80%)   uslplDp.pRIs..pYhSps.......lpl.u
Sec.struct.pred.  .eEEe..hhHhhhhhhhhh........ee...
```

Abbildung 14

Beschreibung von Abbildung 14. Multiples Alignment der SPK-Domänen in ausgesuchten hypothetischen Proteinen (Y14H12D.2, C08F11.7, Y57A10A.1, Y49E10.7, ZK402.5, T28A8.3, Y57A10A.7, B0205.1), SET-Domänen-tragende Proteine (C47E8.8, Y43F11A.5), PHD-Domänen-tragendes Protein (C16A11.4), PHD-Domänen-tragende Proteinkinase (C16A11.3), Caspase (Y48E1B.13), UCH-2-Domänen-tragende Proteinkinase (H12I13.1). Erste Spalte: Protein-Namen (mehrfach auftretende Domäne im gleichen Protein ist mit a,b,c oder d gekennzeichnet); zweite Spalte: erste Aminosäure der Domäne im jeweiligen Protein; letzte Spalte: Datenbank Accession Nummer. Konservierte positiv geladene Aminosäuren sind pink; konservierte negativ geladene Aminosäuren sind rot konservierte hydrophobe Aminosäuren sind blau; zusätzliche konservierte Aminosäuren sind fett gedruckt. Die Konsensus-Sequenz (konservierte Aminosäuren in 80% aller Sequenzen) befindet sich unter dem Alignment; h, p, u, s, l, a, o und + stehen für hydrophobe, polare, winzige, kleine (s =small), aliphatische, aromatisch, alkoholisch und positiv geladene Aminosäuren. Die vorhergesagte Sekundärstruktur steht in der letzten Zeile (H, Helix bekannt oder vorhergesagt mit einer durchschnittlichen Genauigkeit > 82%); h, Helix bekannt oder vorhergesagt mit einer durchschnittlichen Genauigkeit < 82%)) (Rost et al. 1994).


SPK
PHD
SPK
SPK
SPK
PHD
SPK
STYKc
SPK
SPK
SPK
LCI
SPK
STYKc
SP
K
SPK
SPK
SPK
SPK
CASc
SET
SPK
SET
SPK
1
100
200

Beschreibung von Abbildung 15. Domänen-Architektur von SPK-Domänen-tragenden Proteinen. Die Abbildung beinhaltet nur Proteine unterschiedlichen modularen Aufbaus. Die Domänennamen sind dem Simple Modular Architecture Research Tool (Schultz et al. 1998, 2000) (http://smart.embl-heidelberg.de) entlehnt.
Abkürzungen: CASc, katalytische Domäne von Caspasen; PHD, PHD C4HC3-Zinkfinger; SET, (Su(var)3-9, Enhancer-of-zeste, Trithorax)-Domäne; STYKc, katalytische Domäne von Proteinkinasen. Die UCH-2 (Ubiquitin carboxy-terminale Hydrolase Familie 2)-Domäne wird von Pfam (Eddy 1998) vorhergesagt.

3.3.4. Sieben neue Domänen mit nicht signifikanter Ähnlichkeit zu bereits annotierten Domänen

Sieben neu identifizierte Domänen zeigten entfernte Ähnlichkeit zu bereits charakterisierten, traten aber in einem neuen molekularen Kontext auf. Es handelte sich möglicherweise um Mitglieder derselben Superfamilie; ihre Entdeckung lieferte funktionelle Hinweise für Proteine, die diese Domänen tragen. (siehe Tabelle 3)

Domäne	Beschreibung	Länge in AS	Sek. Strukt Vorh.	Vorherges. Funktion	Anzahl proteine	Assoziierte Domänen	Spezies	Acc.Nr. einer repräsentativen Sequenz (Domänen-Grenzen)
ZnF_BED*	BED zinc finger, Related to C2H2 /C2HC zinc fingers (based on pattern similarity)	60	β	Metall-Bindung	50	AT_Hook, PTPc_DSPc	y, a, c, d, h	Q9LWM2 (169-224)
CPDc	Catalytic domain of ctd-like phosphatases, related to phosphatase superfamily (based on pattern similarity)	120	α / β	Phosphatase	70	BRCT, DSRM, UBQ	y, a, c, d, h	Q9PTJ8 (93-236)
RWD	RING finger and WD repeat containing proteins and DEXDc helicases, related to the UBCc domain (revealed by hmm searches)	110	α / β	Protein-interaktion	60	S_TKc, RING, WD, UPF29†, DEXDc, HELIc	y, a, c, d, h	Q9QZ05§ (25-137)
BTP	Bromodomain transcription factors and PHD domain containing Proteins, related to archaeal histone-like transcription factors,	90	α	DNA-Bindung	25	AT_Hook, BROMO, PHD	y, a, c, d, h	Q9S7R9 (41-131)

	defined by PFAM (revealed by PSI-Blast results with less significance (E=0.041))							
ZPW	Zinc finger, PHD domain and WD repeats containing proteins, related to SANT domain (after the second iteration Q9SR68 bridges to SANT domains (E=0.002))	90	α	DNA- oder Protein-Bindung	60	C2H2, PHD, WD	Virus, a, c, d	Q9V5Y9 (22-110)
Znf_DBF	Zinc finger in DBF-like proteins, related to C2H2 zinc fingers (revealed by pattern similarityn and hmm searches, E value = 1.4)	50	α	Metall-Bindung	10	BRCT, AT_Hook	y, d, h	O93843 (590-638)
CHK	C4-zinc finger and HLH domain containing kinase subfamily of choline kinases (after the second iteration P35790 bridges to choline kinases, defined by PFAM) (E=0.003))	200	α / β	Enzym	70	ZnF_C4, HLH, i.c†	Eu, c, d	Q9VBT6 (129-321)

Beschreibung von Tabelle 3. Tabelle der neuen Domänen in unterschiedlichen Spezies. Erste Spalte: Domänen-Name; zweite Spalte: Beschreibung der Domäne (z.B. assoziierte Domänen oder genauer charakterisierte Proteine); dritte Spalte: ungefähre Länge der Domäne (Anzahl der Aminosäuren); vierte Spalte: Sekundärstrukturvorhersage (Rost et al. 1994) (α: Domäne besteht aus α-Helices; β: Domäne besteht aus β-Faltblättern; α/β: Domäne besteht aus α-Helices und β-Faltblättern; fünfte Spalte: vorhergesagte Funktion der neuen Domäne; sechste Spalte: Anzahl der Proteine, die die neue Domäne tragen; siebte Spalte: Namen der assoziierten Domänen (die Domänen-Namen leiten sich aus dem Simple Modular Architecture Research Tool (http://smart.embl-heidelberg.de) (Schultz et al. 1998, 2000) ab oder sind definiert in Pfam (Bateman et al. 2000)†). Achte Spalte: Spezies, in denen die neue Domäne auftritt; eu: Eubakterien; virus: Viren; y: Hefe (yeast); a: *Arabidopsis thaliana*; c: *Caenorhabditis elegans*; d: *Drosophila melanogaster*; h: *Homo sapiens*); neunte Spalte: Accession Nummer eines repräsentativen Proteins und die Region, in der sich die neue Domäne befindet.

* neue Domäne wurde bereits publiziert oder ist "in press".

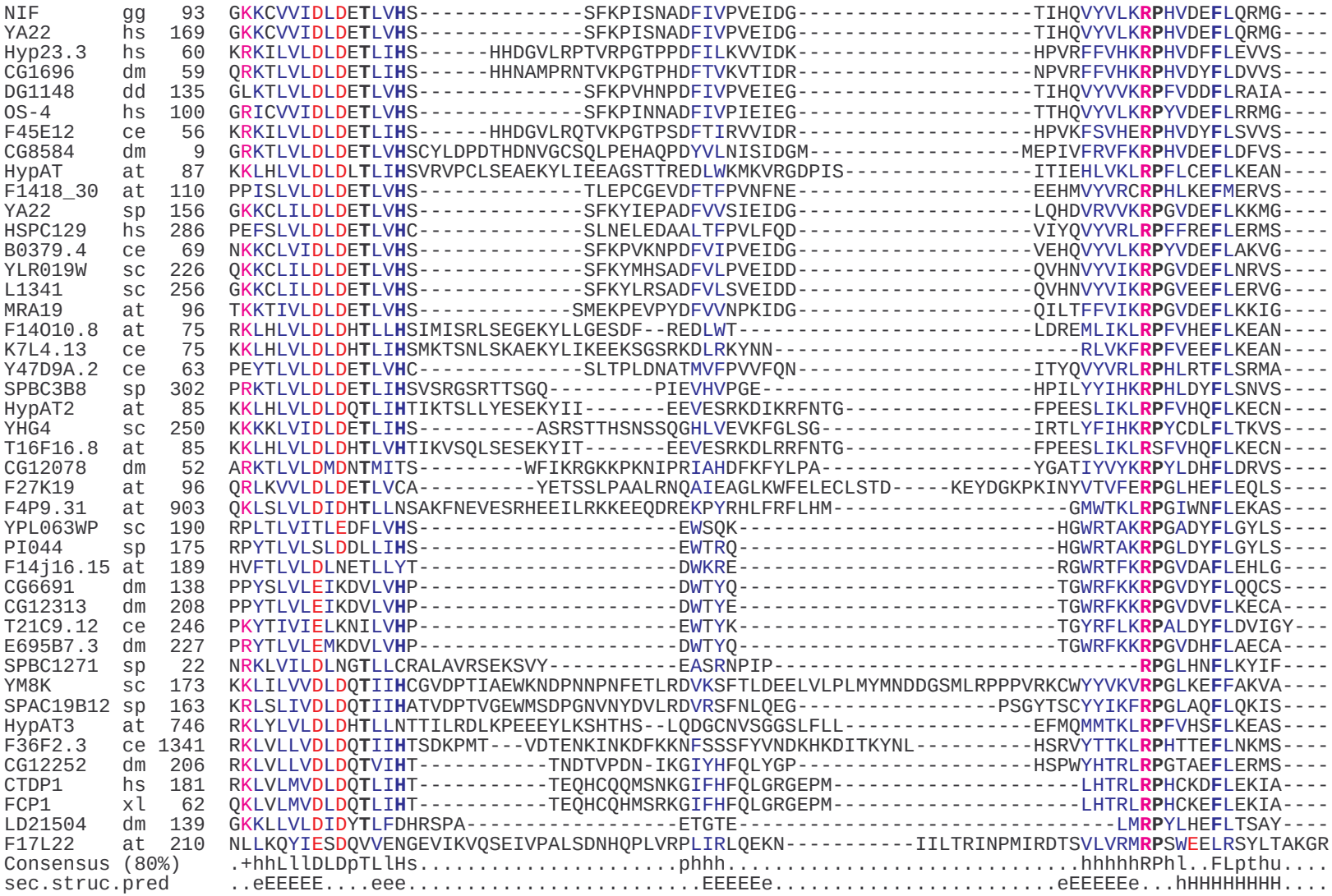
§ zusätzliche HMMer (Eddy 1998)- Suchen waren notwendig, um alle Proteine, die die neue Domäne tragen, zu identifizieren

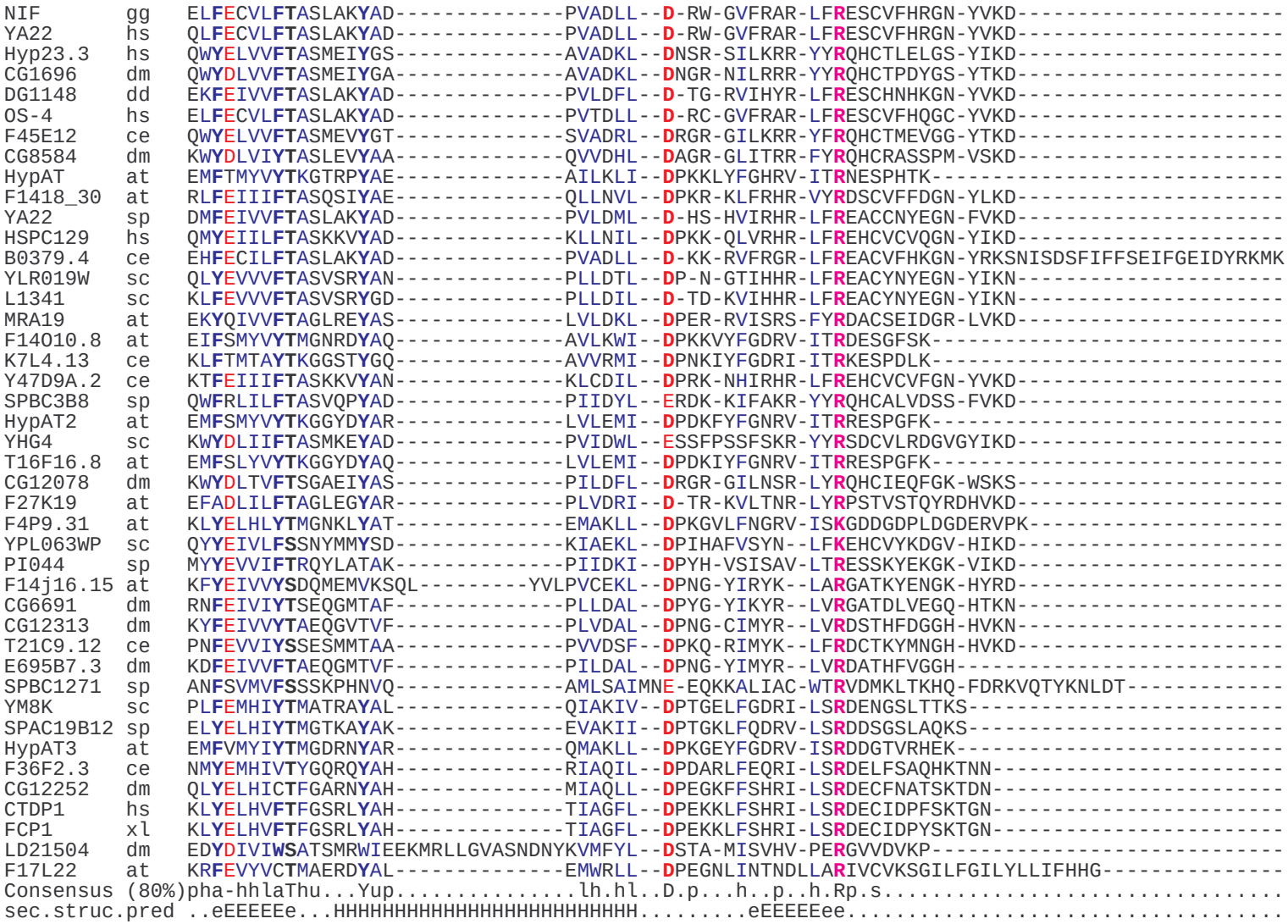
3.3.4.1 Die RWD-Domäne, nicht katalytische Subfamilie ubiquitin-assoziierter Enzyme

Die RWD-Domäne repräsensiert beispielhaft die neuen Domänen, die nicht signifikante Ähnlichkeit zu charakterisierten Domänen haben.

Die GCN2 eIF2alpha-Kinase und Histidyl-tRNA-Synthetase (SpTREMBL accession: Q9QZ05) ist eine wesentliche Komponente der Translationskontrolle (Jentsch et al. 1991, Sattleger et al. 1998). Eine PSI-Blast-Suche mit einem N-terminalen vor der inaktiven Proteinkinase-Domäne gelegenen Sequenzabschnitt (siehe Abbildung 4b) des GCN2 Proteins der Maus detektiert diverse Orthologe in unterschiedlichen Spezies von Hefe bis Mensch. Weitere PSI-BLAST-Iterationen und zusätzliche Hidden Markov Model-Suchen zeigten signifikante Ähnlichkeit zu "WD-repeat"-tragenden Proteinen, DEAD-Helikasen in Hefe, einem Protein der UPF0029-Familie (Uncharakterisierte Proteinfamilie Nummer 29), vielen hypothetischen Proteinen und

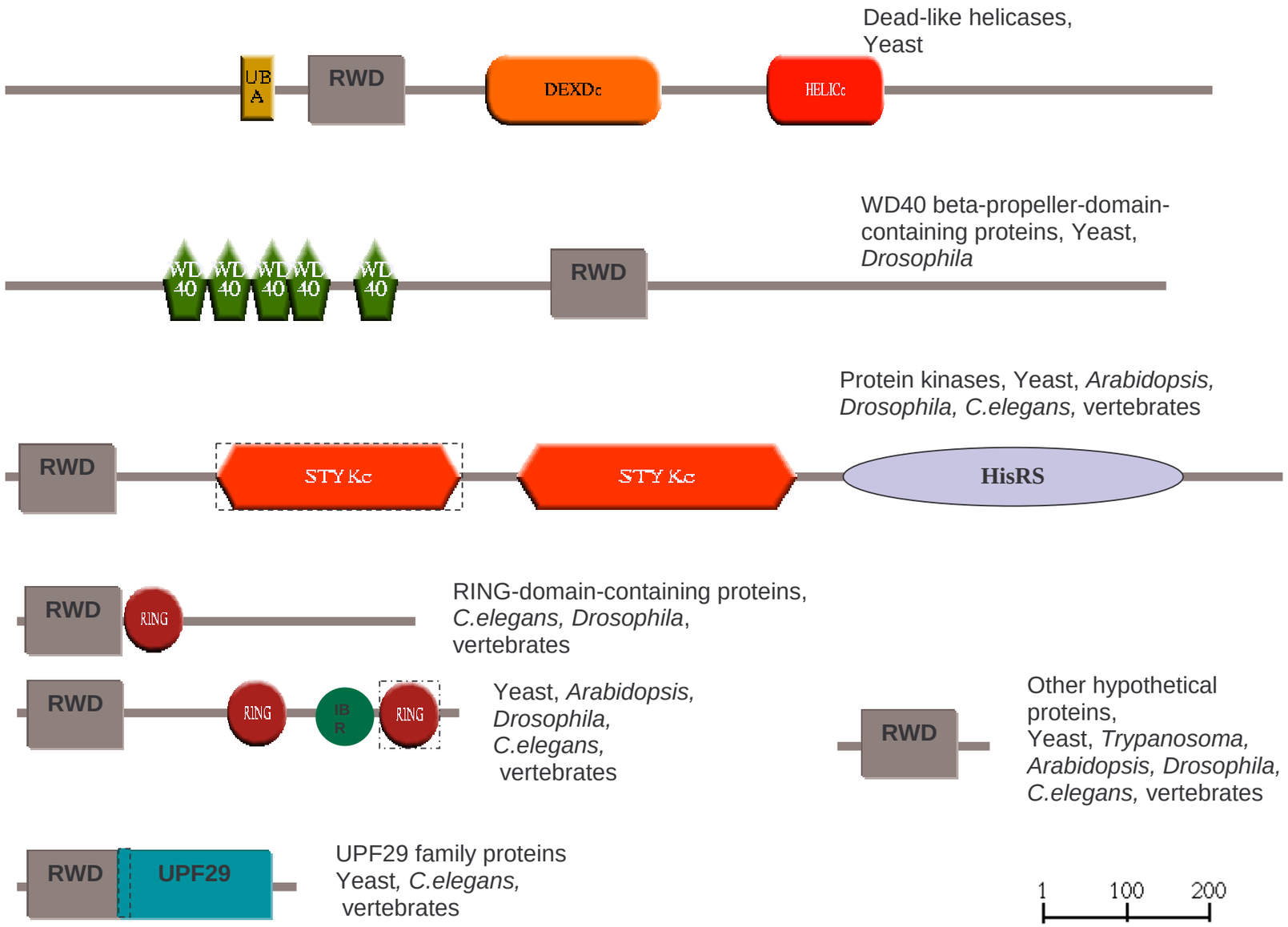
einigen RING-Finger-tragenden Proteinen. Die neu entdeckte Domäne erhielt den Namen RWD nach den besser charakterisierten RING-Finger und WD-Domänen-tragenden Proteinen und DEAD-Helikasen. PSIBLAST-Suchen beginnend mit verschiedenen RWD-Sequenzen zeigen Homologie zu der UBCc-Domäne ubiquitin-assoziierter Enzyme, (z.B. Acc. Nr: Q94721 detektiert Q9SDY5 nach der dritten Iteration, *E*-value = 9×10^{-4}). Das für katalytsiche Aktivität notwendige Cystein ist in den meisten Mitgliedern der neuen Subfamilie nicht konserviert (siehe Abbildung 4a). Die Entdeckung ist vor dem Hintergrund früher experimenteller Studien über das Protein A07 (Acc. Nr.: Q9QZR0) von besonderem Interesse. Dieses RWD- und RING-Finger-tragende Protein besitzt einen Sequenzabschnitt (Aminosäure 85 bis 363), der Ubiquitin-assoziierte Enzyme (E2) bindet und als Substrat für die Ubiquitinierung dient (Lorich et al. 1999).


NIF gg 93 GKKCVVIDLDETLVHS----------SFKPISNADFIVPVEIDG----------TIHQVYVLKRPHVDEFLQRMG----
YA22 hs 169 GKKCVVIDLDETLVHS----------SFKPISNADFIVPVEIDG----------TIHQVYVLKRPHVDEFLQRMG----
Hyp23.3 hs 60 KRKILVLDLDETLIHS------HHDGVLRPTVRPGTPPDFILKVVIDK----------HPVRFFVHKRPHVDFFLEVVS----
CG1696 dm 59 QRKTLVLDLDETLIHS------HHNAMPRNTVKPGTPHDFTVKVTIDR----------NPVRFFVHKRPHVDYFLDVVS----
DG1148 dd 135 GLKTLVLDLDETLVHS----------SFKPVHNPDFIVPVEIEG----------TIHQVYVVKRPFVDDFLRAIA----
OS-4 hs 100 GRICVVIDLDETLVHS----------SFKPINNADFIVPIEIEG----------TTHQVYVLKRPYVDEFLRRMG----
F45E12 ce 56 KRKILVLDLDETLIHS------HHDGVLRQTVKPGTPSDFTIRVVIDR----------HPVKFSVHERPHVDYFLSVVS----
CG8584 dm 9 GRKTLVLDLDETLVHSCYLDPDTHDNVGCSQLPEHAQPDYVLNISIDGM----------MEPIVFRVFKRPHVDEFLDFVS----
HypAT at 87 KKLHLVLDLDLTLIHSVRVPCLSEAEKYLIEEAGSTTREDLWKMKVRGDPIS----------ITIEHLVKLRPFLCEFLKEAN----
F1418_30 at 110 PPISLVLDLDETLVHS----------TLEPCGEVDFTFPVNFNE----------EEHMVYVRCRPHLKEFMERVS----
YA22 sp 156 GKKCLILDLDETLVHS----------SFKYIEPADFVVSIEIDG----------LQHDVRVVKRPGVDEFLKKMG----
HSPC129 hs 286 PEFSLVLDLDETLVHC----------SLNELEDAALTFPVLFQD----------VIYQVYVRLRPFFREFLERMS----
B0379.4 ce 69 NKKCLVIDLDETLVHS----------SFKPVKNPDFVIPVEIDG----------VEHQVYVLKRPYVDEFLAKVG----
YLR019W sc 226 QKKCLILDLDETLVHS----------SFKYMHSADFVLPVEIDD----------QVHNVYVIKRPGVDEFLNRVS----
L1341 sc 256 GKKCLILDLDETLVHS----------SFKYLRSADFVLSVEIDD----------QVHNVYVIKRPGVEEFLERVG----
MRA19 at 96 TKKTIVLDLDETLVHS----------SMEKPEVPYDFVVNPKIDG----------QILTFFVIKRPGVDEFLKKIG----
F14O10.8 at 75 RKLHLVLDLDHTLLHSIMISRLSEGEKYLLGESDF--REDLWT----------LDREMLIKLRPFVHEFLKEAN----
K7L4.13 ce 75 KKLHLVLDLDHTLIHSMKTSNLSKAEKYLIKEEKSGSRKDLRKYNN----------RLVKFRPFVEEFLKEAN----
Y47D9A.2 ce 63 PEYTLVLDLDETLVHC----------SLTPLDNATMVFPVVFQN----------ITYQVYVRLRPHLRTFLSRMA----
SPBC3B8 sp 302 PRKTLVLDLDETLIHSVSRGSRTTSGQ---------PIEVHVPGE----------HPILYYIHKRPHLDYFLSNVS----
HypAT2 at 85 KKLHLVLDLDQTLIHTIKTSLLYESEKYII-------EEVESRKDIKRFNTG----------FPEESLIKLRPFVHQFLKECN----
YHG4 sc 250 KKKKLVIDLDETLIHS----------ASRSTTHSNSSQGHLVEVKFGLSG----------IRTLYFIHKRPYCDLFLTKVS----
T16F16.8 at 85 KKLHLVLDLDHTLVHTIKVSQLSESEKYIT-------EEVESRKDLRRFNTG----------FPEESLIKLRSFVHQFLKECN----
CG12078 dm 52 ARKTLVLDMDNTMITS---------WFIKRGKKPKNIPRIAHDFKFYLPA----------YGATIYVYKRPYLDHFLDRVS----
F27K19 at 96 QRLKVVLDLDETLVCA----------YETSSLPAALRNQAIEAGLKWFELECLSTD-----KEYDGKPKINYVTVFERPGLHEFLEQLS----
F4P9.31 at 903 QKLSLVLDIDHTLLNSAKFNEVESRHEEILRKKEEQDREKPYRHLFRFLHM----------GMWTKLRPGIWNFLEKAS----
YPL063WP sc 190 RPLTLVITLEDFLVHS----------EWSQK----------HGWRTAKRPGADYFLGYLS----
PI044 sp 175 RPYTLVLSLDDLLIHS----------EWTRQ----------HGWRTAKRPGLDYFLGYLS----
F14j16.15 at 189 HVFTLVLDLNETLLYT----------DWKRE----------RGWRTFKRPGVDAFLEHLG----
CG6691 dm 138 PPYSLVLEIKDVLVHP----------DWTYQ----------TGWRFKKRPGVDYFLQQCS----
CG12313 dm 208 PPYTLVLEIKDVLVHP----------DWTYE----------TGWRFKKRPGVDVFLKECA----
T21C9.12 ce 246 PKYTIVIELKNILVHP----------EWTYK----------TGYRFLKRPALDYFLDVIGY---
E695B7.3 dm 227 PRYTLVLEMKDVLVHP----------DWTYQ----------TGWRFKKRPGVDHFLAECA----
SPBC1271 sp 22 NRKLVILDLNGTLLCRALAVRSEKSVY----------EASRNPIP----------RPGLHNFLKYIF----
YM8K sc 173 KKLILVVDLDQTIIHCGVDPTIAEWKNDPNNPNFETLRDVKSFTLDEELVLPLMYMNDDGSMLRPPPVRKCWYYVKVRPGLKEFFAKVA----
SPAC19B12 sp 163 KRLSLIVDLDQTIIHATVDPTVGEWMSDPGNVNYDVLRDVRSFNLQEG----------PSGYTSCYYIKFRPGLAQFLQKIS----
HypAT3 at 746 RKLYLVLDLDHTLLNTTILRDLKPEEEYLKSHTHS--LQDGCNVSGGSLFLL----------EFMQMMTKLRPFVHSFLKEAS----
F36F2.3 ce 1341 RKLVLLVDLDQTIIHTSDKPMT---VDTENKINKDFKKNFSSSFYVNDKHKDITKYNL----------HSRVYTTKLRPHTTEFLNKMS----
CG12252 dm 206 RKLVLLVDLDQTVIHT----------TNDTVPDN-IKGIYHFQLYGP----------HSPWYHTRLRPGTAEFLERMS----
CTDP1 hs 181 RKLVLMVDLDQTLIHT----------TEQHCQQMSNKGIFHFQLGRGEPM----------LHTRLRPHCKDFLEKIA----
FCP1 xl 62 QKLVLMVDLDQTLIHT----------TEQHCQHMSRKGIFHFQLGRGEPM----------LHTRLRPHCKEFLEKIA----
LD21504 dm 139 GKKLLVLDIDYTLFDHRSPA----------ETGTE----------LMRPYLHEFLTSAY----
F17L22 at 210 NLLKQYIESDQVVENGEVIKVQSEIVPALSDNHQPLVRPLIRLQEKN----------IILTRINPMIRDTSVLVRMRPSWEELRSYLTAKGR
Consensus (80%) .+hhLllDLDpTLlHs.....phhh.....hhhhhRPhl..FLpthu....
sec.struc.pred ..eEEEEE....eee.....EEEEEe.....eEEEEEe...hHHHHHHHH....


NIF gg ELFECVLFTASLAKYAD----------PVADLL--D-RW-GVFRAR-LFRESCVFHRGN-YVKD----------
YA22 hs QLFECVLFTASLAKYAD----------PVADLL--D-RW-GVFRAR-LFRESCVFHRGN-YVKD----------
Hyp23.3 hs QWYELVVFTASMEIYGS----------AVADKL--DNSR-SILKRR-YYRQHCTLELGS-YIKD----------
CG1696 dm QWYDLVVFTASMEIYGA----------AVADKL--DNGR-NILRRR-YYRQHCTPDYGS-YTKD----------
DG1148 dd EKFEIVVFTASLAKYAD----------PVLDFL--D-TG-RVIHYR-LFRESCHNHKGN-YVKD----------
OS-4 hs ELFECVLFTASLAKYAD----------PVTDLL--D-RC-GVFRAR-LFRESCVFHQGC-YVKD----------
F45E12 ce QWYELVVFTASMEVYGT----------SVADRL--DRGR-GILKRR-YFRQHCTMEVGG-YTKD----------
CG8584 dm KWYDLVIYTASLEVYAA----------QVVDHL--DAGR-GLITRR-FYRQHCRASSPM-VSKD----------
HypAT at EMFTMYVYTKGTRPYAE----------AILKLI--DPKKLYFGHRV-ITRNESPHTK----------
F1418_30 at RLFEIIIFTASQSIYAE----------QLLNVL--DPKR-KLFRHR-VYRDSCVFFDGN-YLKD----------
YA22 sp DMFEIVVFTASLAKYAD----------PVLDML--D-HS-HVIRHR-LFREACCNYEGN-FVKD----------
HSPC129 hs QMYEIILFTASKKVYAD----------KLLNIL--DPKK-QLVRHR-LFREHCVCVQGN-YIKD----------
B0379.4 ce EHFECILFTASLAKYAD----------PVADLL--D-KK-RVFRGR-LFREACVFHKGN-YRKSNISDSFIFFSEIFGEIDYRKMK
YLR019W sc QLYEVVVFTASVSRYAN----------PLLDTL--DP-N-GTIHHR-LFREACYNYEGN-YIKN----------
L1341 sc KLFEVVVFTASVSRYGD----------PLLDIL--D-TD-KVIHHR-LFREACYNYEGN-YIKN----------
MRA19 at EKYQIVVFTAGLREYAS----------LVLDKL--DPER-RVISRS-FYRDACSEIDGR-LVKD----------
F14O10.8 at EIFSMYVYTMGNRDYAQ----------AVLKWI--DPKKVYFGDRV-ITRDESGFSK----------
K7L4.13 ce KLFTMTAYTKGGSTYGQ----------AVVRMI--DPNKIYFGDRI-ITRKESPDLK----------
Y47D9A.2 ce KTFEIIIFTASKKVYAN----------KLCDIL--DPRK-NHIRHR-LFREHCVCVFGN-YVKD----------
SPBC3B8 sp QWFRLILFTASVQPYAD----------PIIDYL--ERDK-KIFAKR-YYRQHCALVDSS-FVKD----------
HypAT2 at EMFSMYVYTKGGYDYAR----------LVLEMI--DPDKFYFGNRV-ITRRESPGFK----------
YHG4 sc KWYDLIIFTASMKEYAD----------PVIDWL--ESSFPSSFSKR-YYRSDCVLRDGVGYIKD----------
T16F16.8 at EMFSLYVYTKGGYDYAQ----------LVLEMI--DPDKIYFGNRV-ITRRESPGFK----------
CG12078 dm KWYDLTVFTSGAEIYAS----------PILDFL--DRGR-GILNSR-LYRQHCIEQFGK-WSKS----------
F27K19 at EFADLILFTAGLEGYAR----------PLVDRI--D-TR-KVLTNR-LYRPSTVSTQYRDHVKD----------
F4P9.31 at KLYELHLYTMGNKLYAT----------EMAKLL--DPKGVLFNGRV-ISKGDDGDPLDGDERVPK----------
YPL063WP sc QYYEIVLFSSNYMMYSD----------KIAEKL--DPIHAFVSYN--LFKEHCVYKDGV-HIKD----------
PI044 sp MYYEVVIFTRQYLATAK----------PIIDKI--DPYH-VSISAV-LTRESSKYEKGK-VIKD----------
F14j16.15 at KFYEIVVYSDQMEMVKSQL---------YVLPVCEKL--DPNG-YIRYK--LARGATKYENGK-HYRD----------
CG6691 dm RNFEIVIYTSEQGMTAF----------PLLDAL--DPYG-YIKYR--LVRGATDLVEGQ-HTKN----------
CG12313 dm KYFEIVVYTAEQGVTVF----------PLVDAL--DPNG-CIMYR--LVRDSTHFDGGH-HVKN----------
T21C9.12 ce PNFEVVIYSSESMMTAA----------PVVDSF--DPKQ-RIMYK--LFRDCTKYMNGH-HVKD----------
E695B7.3 dm KDFEIVVFTAEQGMTVF----------PILDAL--DPNG-YIMYR--LVRDATHFVGGH----------
SPBC1271 sp ANFSVMVFSSSKPHNVQ----------AMLSAIMNE-EQKKALIAC-WTRVDMKLTKHQ-FDRKVQTYKNLDT----------
YM8K sc PLFEMHIYTMATRAYAL----------QIAKIV--DPTGELFGDRI-LSRDENGSLTTKS----------
SPAC19B12 sp ELYELHIYTMGTKAYAK----------EVAKII--DPTGKLFQDRV-LSRDDSGSLAQKS----------
HypAT3 at EMFVMYIYTMGDRNYAR----------QMAKLL--DPKGEYFGDRV-ISRDDGTVRHEK----------
F36F2.3 ce NMYEMHIVTYGQRQYAH----------RIAQIL--DPDARLFEQRI-LSRDELFSAQHKTNN----------
CG12252 dm QLYELHICTFGARNYAH----------MIAQLL--DPEGKFFSHRI-LSRDECFNATSKTDN----------
CTDP1 hs KLYELHVFTFGSRLYAH----------TIAGFL--DPEKKLFSHRI-LSRDECIDPFSKTGN----------
FCP1 xl KLYELHVFTFGSRLYAH----------TIAGFL--DPEKKLFSHRI-LSRDECIDPYSKTGN----------
LD21504 dm EDYDIVIWSATSMRWIEEKMRLLGVASNDNYKVMFYL--DSTA-MISVHV-PERGVVDVKP----------
F17L22 at KRFEVYVCTMAERDYAL----------EMWRLL--DPEGNLINTNDLLARIVCVKSGILFGILYLLIFHHG----------
Consensus (80%) pha-hhlaThu...Yup.....lh.hl..D.p...h..p..h.Rp.s.....
sec.struc.pred ..eEEEEEe...HHHHHHHHHHHHHHHHHHHHHHHH.....eEEEEEee.....

```
NIF          gg   ----------------------LSRLGR----------ELSKVIIVDNSPASYIFHPEN-AVPVQSWFDDMTDT  Q9PTJ8
YA22         hs   ----------------------LSRLGR----------ELSKVIIVDNSPASYIFHPEN-AVPVQSWFDDMTDT  O15194
Hyp23.3      hs   ----------------------LSVVH----------SDLSSIVILDNSPGAYRSHPDN-AIPIKSWFSDPSDT  O95476
CG1696       dm   ----------------------LSAIC----------SDLNRIFIIDNSPGAYRCFPNN-AIPIKSWFSDPMDT  Q9VRG7
DG1148       dd   ----------------------LSRLGR----------DLKSTIIVDNSPSSYLFHPEN-AIPIDSWFDDKDDR  Q9XYL0
OS-4         hs   ----------------------LSRLGR----------DLRKTLILDNSPASYIFHPEN-AVPVQSWFDDMADT  O14595
F45E12       ce   ----------------------LSAIH----------PDLSSICILDNSPGAYRKFPHN-AIPIPSWFSDPNDT  Q20432
CG8584       dm   ----------------------LTLVT----------PDMSGVLIIDNSPYAYRDFPDN-AIPIKTFIYDPDDT  Q9V4W8
HypAT        at   ----------------------TLDMVLA----------DERGVVIVDDTRKAWPNNKSN-LVLIGRYNYFRSQS  BAB02545
F1418_30     at   ----------------------LSVLGR----------DLSRVIIVDNSPQAFGFQVEN-GVPIESWFNDPSDK  CAB87659
YA22         sp   ----------------------LSQLGR----------NLEDSIIIDNSPSSYIFHPSH-AVPISSWFNDMHDM  Q09695
HSPC129      hs   ----------------------LNILGR----------DLSKTIIIDNSPQAFAYQLSN-GIPIESWFMDKNDN  AAF29093
B0379.4      ce   DYRKMKNLHFSSYFPEKIEKYYLSRLGR----------NLNQTLIIDNSPASYAFHPEN-AVPVTTWFDDPSDT  O02204
YLR019W      sc   ----------------------LSQIGR----------PLSETIILDNSPASYIFHPQH-AVPISSWFSDTHDN  Q07949
L1341        sc   ----------------------LSQIGR----------PLSDIIILDNSPASYIFHPQH-AIPISSWFSDTHDN  Q07800
MRA19        at   ----------------------LGFVMR----------DLRRVVIVDDNPNSYALQPEN-AFPIKPFSDDLEDV  BAB09212
F14O10.8     at   ----------------------TLDLVLA----------DECGVVIVDDTRHVWPDHERN-LLQITKYSYFRDYS  AAF88157
K7L4.13      ce   ----------------------TLDLVLA----------DERGIVIVDNTPNVWPHHKRN-LLEITSYFYFKNDG  AAF34842
Y47D9A.2     ce   ----------------------LTILGR----------DPSKTMILDNAVQSFAYQLDN-GIPIESWFHDRNDT  AAF60646
SPBC3B8      sp   ----------------------ISICN----------IHLSRIMIIDNSPASYNAHKEN-AIPIEGWISDPSDV  O59718
HypAT2       at   ----------------------TLDLVLA----------DERGIVIVDDTSSVWPHDKKN-LLQIARYKYFGDKS  BAB09563
YHG4         sc   ----------------------LSIVKDSEENGKGSSSSLDDVIIIDNSPVSYAMNVDN-AIQVEGWISDPTDT  P38757
T16F16.8     at   ----------------------TLDLVLA----------DERGIVVVDDKSSVWPHDKKN-LLQIARYKYFGDQS  Q9ZVR2
CG12078      dm   ----------------------VLLAC----------PDLSNVVLLDNSSTECSFNAEN-AILIKSYEIGCRDE  Q9VZS0
F27K19       at   ----------------------LLSTSK----------NMCRTVIVDNNPFSFLLQPSN-GIPCIAFSAGQPND  CAB87850
F4P9.31      at   -------------------SKDLEGVMG----------MESSVVIIDDSVRVWPQHKMN-LIAVERYHL-SHYG  O22804
YPL063WP     sc   ----------------------LSKLNR----------DLSKVIIIDTDPNSYKLQPEN-AIPMEPWNGEADDK  Q02776
PI044        sp   ----------------------LSYLNR----------DLSRVIMIDTNPESWSKQPDN-AIAMAPWTGNPKDK  O13636
F14j16.15    at   ----------------------LSKLNR----------DPKKILFVSANAFESTLQPEN-SVPIKPYKLEADDT  AAF79316
CG6691       dm   ----------------------LDYLNR----------DLSRVIVVDCDPYTTPLHPDN-SLVLTKWLGNDDDV  Q9V9P3
CG12313      dm   ----------------------LDNLNR----------DLKRVVVVDWDRNSTKFHPSN-SFSIPRWSGNDNDT  Q9W0S3
T21C9.12     ce   ----------------------LSKLNR----------DLSKVIYIDFDAKSGQLNPEN-MLRVPEWKGNMDDT  Q22647
E695B7.3     dm   ----------------------------------------HVIVVDWDANATKMHPDN-TFGLARWHGNDDDG  O76907
SPBC1271     sp   -----------VWEKIHHDSTGKPVSW----------SQYNTIIVDDSKTKCAAHPYN-HIAVSDFVAKSHSN  O94336
YM8K         sc   ----------------------LAKLFPT----------DQSMVVVIDDRGDVWNWCP-N-LIKVVPYNFFVGVG  Q03254
SPAC19B12    sp   ----------------------LRRLFPC----------DTSMVVVIDDRGDVWDWNP-N-LIKVVPYEFFVGIG  CAC00553
HypAT3       at   ----------------------SLDVVLG----------QESAVLILDDTENAWPKHKDN-LIVIERYHFFSSSC  BAB08870
F36F2.3      ce   ----------------------LKALFPC----------GDNLVVIIDDRSDVWMYSE-A-LIQIKPYRFFKEVG  O62235
CG12252      dm   ----------------------LKALFPN----------GDSMVCIIDDREDVWNMAS-N-LIQVKPYHFFQHTG  Q9W147
CTDP1        hs   ----------------------LRNLFPC----------GDSMVCIIDDRKDVWKFAP-N-LITVKKYVYFQGTG  Q9Y5B0
FCP1         xl   ----------------------LRNLFPC----------GDSMVCIIDDREDVWKFAP-N-LITVKKMCIFQGTG  Q9PT70
LD21504      dm   ----------------------LGVIWALY------KQYNSSNTIMFDDIRRNFLMNPKS-GLKIRPFRQAHLNR  Q9XZ16
F17L22       at   -----------------FKKSLFNVFLDG-------TCHPKMALVIDDRLKVWDEKDQPRVHVVPAFAPYYSPQ  Q9SVT0
Consensus (80%).....................hshh............s.p.slllDsp..sa..p.pN.hl.l..a....pss
sec.struc.pred .....................hhh.............eEEEE..............eEEe..ee.....
```

Beschreibung von Abbildung 16. Multiples Alignment der RWD-Domänen. Erste Spalte: Protein-Namen; zweite Spalte: Spezies-Bezeichnung (at: *Arabidopsis thaliana*; Ce: *Caenorhabditis elegans*; dm: *Drosophila melanogaster*; hs: *Homo sapiens*; mm: *Mus musculus*; rn: *Rattus norvegicus*; sc: *Saccharomyces cerevisae*; sp: *Saccharomyces pombe*); dritte Spalte: erste Aminosäure der Domäne im jeweiligen Protein; letzte Spalte: Datenbank Accession Nummer. Konservierte positiv geladene Aminosäuren sind pink markiert; konservierte hydrophobe Aminosäuren sind blau; zusätzliche konservierte Aminosäuren sind fett gedruckt. Die Konsensus-Sequenz (konservierte Aminosäuren in 80 % aller Sequenzen) befindet sich unter dem Alignment; h, p, u, s, l und + stehen für hydrophobe, polare, winzige, kleine (s =small), aliphatische und positiv geladene Aminosäuren. Die vorhergesagte Sekundärstruktur steht in der letzten Zeile (H, Helix bekannt oder vorhergesagt mit einer durchschnittlichen Genauigkeit > 82%); h, Helix bekannt oder vorhergesagt mit einer durchschnittlichen Genauigkeit < 82%)) (Rost et al. 1994).


Dead-like helicases,
Yeast
UB
A
RWD
DEXDc
HELICc
WD40 beta-propeller-domain-
containing proteins, Yeast,
Drosophila
WD
40
RWD
Protein kinases, Yeast, Arabidopsis,
Drosophila, C.elegans, vertebrates
RWD
STY Kc
STY Kc
HisRS
RWD
RING
RING-domain-containing proteins,
C.elegans, Drosophila,
vertebrates
RWD
RING
IB
R
RING
Yeast, Arabidopsis,
Drosophila,
C.elegans,
vertebrates
RWD
Other hypothetical
proteins,
Yeast, Trypanosoma,
Arabidopsis, Drosophila,
C.elegans, vertebrates
RWD
UPF29
UPF29 family proteins
Yeast, C.elegans,
vertebrates
1
100
200

Beschreibung von Abbildung 17. Domänen-Architektur von RWD-Domänen-tragenden Proteinen.
Die Abbildung beinhaltet nur Proteine unterschiedlichen modularen Aufbaus. Die Domänennamen sind dem Simple Modular Architecture Research Tool (Schultz et al. 1998, 2000) (http://smart.embl-heidelberg.de) entlehnt.
Abkürzungen: Superfamilie DEAD-ähnlicher Helikasen (N-terminale Domäne); HELICc, Superfamilie von Helikasen (C-terminale Domäne); RING, RING-Finger-Domäne; STYKc, Proteinkinasen (Substratspezifität nicht klassifiziert); UBA, Ubiquitin-assoziierte Domäne; WD40, WD40-"repeats". Die RING-Finger-Domäne in der gestrichelten Box wird von SMART oder Pfam nicht erkannt. Die STYKc-Domäne in der gestrichelten Box ist degeneriert (zerstückelt und nicht katalytisch). The IBR ("In between Ring fingers")-Domäne ist in Pfam (Bateman et al. 2000) beschrieben. Die HisRS (Histidyl-tRNA-Sythetase)-Domäne ist in der Literatur (Sattleger et al. 1998) beschrieben.

3.3.5. Drei neue Domänen-spezifische Extensionen

Drei neu identifizierte homologe Sequenzabschnitte in Proteinen sind in allen Fällen mit bereits charakterisierten Domänen assoziiert. Es handelt sich um vermutlich funktionell bedeutsame Extensionen. (siehe Tabelle 4)

Domäne	Beschreibung	Länge in AS	Sek. Strukt Vorh.	Vorherges. Funktion	Anzahl proteine	Assoziierte Domänen	Spezies	Acc.Nr. einer repräsentativen Sequenz (Domänen-Grenzen)
AWS	Associated with SET domain, subdomain of PRESET‡ (hmm searches, E value = 0.52)	50	α / β	Nukleosom aufbau	25	SET, PWWP, AT_Hook, WW, PHD, POSTSET, BAH	y, a, c, d, h	P46995 (63-119)
POX	Domain associated with HOX-domains	50	α	unbekannt	20	HOX	a	Q38897 (199-337)
PRE_C2HC	Associated with zinc fingers	70	α / β	unbekannt	15	ZnF_C2HC	d	O44939 (546-616)

Beschreibung von Tabelle 4 Tabelle der drei neuen domänen-spezifischen Extensionen. Erste Spalte: Domänen-Name; zweite Spalte: Beschreibung der Domäne (z.B. assoziierte Domänen oder genauer charakterisierte Proteine); dritte Spalte: ungefähre Länge der Domäne (Anzahl der Aminosäuren); vierte Spalte: Sekundärstrukturvorhersage (Rost et al. 1994) (α: Domäne besteht aus α-Helices; β: Domäne besteht aus β-Faltblättern; α/β: Domäne besteht aus α-Helices und β-Faltblättern; fünfte Spalte: vorhergesagte Funktion der neuen Domäne; sechste Spalte: Anzahl der Proteine, die die neue Domäne tragen; siebte Spalte: Namen der assoziierten Domänen (die Domänen-Namen leiten sich aus dem Simple Modular Architecture Research Tool (http://smart.embl-heidelberg.de) (Schultz et al. 1998, 2000). Achte Spalte: Spezies, in denen die neue Domäne auftritt; eu: Eubakterien; virus: Viren; y: Hefe (yeast); a: *Arabidopsis thaliana*; c: *Caenorhabditis elegans*; d: *Drosophila melanogaster*; h: *Homo sapiens*); neunte Spalte: Accession Nummer eines repräsentativen Proteins und die Region, in der sich die neue Domäne befindet.

3.3.6. Funktionsvorhersagen

Auf der Grundlage umfassender Literaturrecherche und des gemeinsamen Auftretens der neu entdeckten Domänen mit bereits funktionell charakterisierten in gleichen Proteinen sind Funktionsvorhersagen für 76,6% der 28 neuen Domänen-Familien möglich. Dies bedeutet eine Zunahme der funktionellen Beschreibung für 700 Proteine (Proteine, die von den neuen Domänen abgedeckt werden
(Tabelle 1 bis 4)). Die vorhergesagten Funktionen betreffen vielfältige zelluläre Prozesse und Stoffwechselwege, wie DNA/RNA-, Metallionen- oder Proteinbindung bis hin zu katalytischen Eigenschaften.
Kurzbeschreibungen der Funktionen sind in den Tabellen 1 bis 4 aufgeführt, fünf charakteristische Beispiele sind (siehe unten) konkret beschrieben.

Chromatin-bindende Domänen

Das CSZ-Domänen-tragende Protein SPT6 und seine Orthologen regulieren die Transkription durch Stabilisierung und Formung der Chromatinstruktur (Chiang et al. 1996, Winston 2001). Die Fähigkeit von SPT6, Histone zu binden, ist experimentell belegt (Bortvin et al. 1996). Die CSZ-Domäne tritt assoziiert mit S1- und zwei SH2-Domänen auf, die weder für Histon- noch Chromatinbindung verantwortlich sind. Aus dem Kontext dieser Informationen ist eine Histon- oder Chromatin- bindende Funktion für die neue CSZ-Domäne als wahrscheinlich ableitbar.

Im Rahmen des Projektes kam es zur Identifikation einer Domäne als "tandem-repeat" in vielen hypothetischen Proteinen und als einzeln assoziiert mit PHD- und TFS2M-Domänen in *Drosophila* (CG6525), so wie in *Drosophila* brahma und kismet. Für das kismet Protein in *Drosophila* und seine Orthologen wurde gezeigt, dass es sich um "chromatin-remodelling"- Faktoren handelt. Die neu entdeckte Domäne schliesst eine in der Literatur beschrieben Region (BRK) mit ein, für die Chromatin-Bindung angenommen wird (Daubresse et al. 1999). Eine Chromatin-bindende Eigenschaft der neuen Domäne ist deshalb wahrscheinlich.

Protein-Interaktions-Domänen

Frühere Studien belegen eine Interaktion der RPR-Domäne im Protein pcfl 1 mit der Carboxyl-terminalen Domäne der größten Untereinheit der RNA-Polymerase II

(Yuryev et al. 1996). Dies führt zu dem Schluss, dass RPR generell die Polymerase bindet oder weniger spezifisch eine Protein-Interaktionsfunktion hat.

PSP-Domänen scheinen Proteine zu binden. Das PSP-Domänen-tragende Protein Cus1p ist Bestandteil eines spliceosomalen Komplexes, assoziiert mit U2 snRNA (Gozani et al. 1996). Cus1p bindet direkt an das snRNP Hsh155p über eine Region, die mit der PSP-Domäne überlappt (Pauling et al. 2000).

"The nuclear factor 90" (NF90) ist Substrat und Regulator des eukaryotischen "initiation factor 2 kinase double-stranded RNA-activated protein kinase" -Komplex. Die neue DZF-Domäne in NF90 überlappt mit einer Region bekannt als "NF45 homology domain", die verantwortlich ist für die Konformationsausbildung von NF90 im Komplex mit NF45; hier bindet sie NF45 oder andere Proteine (Parker et al. 2001).

3.3.7. Mögliche Krankeitsrelevanz neuer Domänen

Vier (14%) der neuen Domänen und eine Extension kommen in Proteinen vor, die mit schwerwiegenden menschlichen Erbkrankheiten in Verbindung gebracht werden können. Die betreffenden beschädigten Gene oder chromosomalen Bereiche sind verantwortlich für Krebs, neurodegenerative Erkrankungen und Cromosomenaberrationen (siehe Tabelle 5).
Es ist nicht bekannt, ob Mutationen speziell in den neuen Domänen für die phänotypischen Effekte ursächlich sind.

Domäne	Prot. Acc.Nr.	Erkrankung	OMIM Acc. Nr.
AWS	O96028	Wolf - Hirschhorn - Syndrom (Stec et al. 1998)	602952
RWD	CAB88085	Monosomie 21 (Orti et al. 2000)	- - -
DNP	O70656	Malignes Astrocytom (Nakamura et al. 1998)	- - -
FYRN/FYRC	Q03164	Akute Leukaemie (Djabali et al. 1992)	159555

Tabelle 5. Tabelle der neuen Domänen und Domänen-spezifischen Extensionen, die möglicherweise bei der Entstehung von Erbkrankheiten eine Rolle spielen.
Erste Spalte: Domänenname; zweite Spalte: Accession-Nr. eines krankheitsrelevanten Proteins; dritte Spalte: Bezeichnung der Erkrankung; vierte Spalte: Accession-Nr. der Krankheit in der OMIM-Datenbank

3.3.8. Zusammenfassung

Einige ausführlich beschriebene und gut charakterisierte Signaltransduktionsdomänen wie z.B. SH2 oder PH sind in einer Vielzahl von Proteinen vorhanden, assoziiert mit einer grossen Anzahl von Domänen. Die weite Verbreitung erklärt ihre frühe Entdeckung und Charakterisierung. So ist es nicht überraschend, dass die Mehrheit der hier identifizierten Domänen häufig in weniger Proteinen erscheint; im Durchschnitt treten sie in vier verschiedenen Architekturen in circa 30 Proteinen auf. Aber auch verbreitetere Domänen wie BRK (mit sieben unterschiedlichen Architekturen) sind Ergebnis der Analyse.
Nur drei (11 %) der neu entdeckten Domänen sind Spezies-spezifisch; zwei sind auf Pflanzen beschränkt und eine ist Nematoden-spezifisch (siehe Tabelle 2)
Dieses kann bedeuten, dass selbst wenn Spezies-spezifische Soffwechselwege existieren, diese bereits vorhandene Komponenten nutzen oder dass Spezies-spezifische Domänen erheblich seltener mit taxonomisch verbreiteten Domänen assoziiert sind; solche können von dieser Analyse nicht berücksichtigt werden.

Zusammenfassend hat die Analyse zur Entdeckung von 28 neuen Domänen geführt, von denen ein Großteil bevorzugt im Nukleus lokalisiert zu sein scheint (10 sind nukleusspezifisch, 18 im Cytosol und Nukleus lokalisiert). Die systematische Suche nach neuen Domänen hat zu einem 26%igen Anstieg der in den 15 Jahren entdeckten nuklearen Domänen geführt. Die vorhergesagten Funktionen bewegen sich von enzymatischer Aktivität bis zur Protein- oder Nukleotidbindung.
Obwohl die neu entdeckten Domänen durchschnittlich in weniger Proteinen auftreten, als dies für viele früher charakterisierte Domänen gilt, spricht ihre überwiegend weit Spezies-übergreifende Verbreitung für eine evolutiv bedeutsame biologisch grundlegend relevante Rolle.

3.4. Perspektiven - Weiterführende Analysen nuklearer Domänen

Die Zusammenstellung bekannter nuklearer Domänen kombiniert mit der umfassenden Analyse aller nuklearen Proteine vervollständigt das Wissen um die modulare Architektur im Kern befindlicher Proteine in ihrer Gesamtheit und bildet eine zuverlässige Grundlage für weitergehende Untersuchungen.

Die Ergebnisse dieser Arbeit können für Korrelationsananlysen herangezogen werden, die Aussagen über funktionelle Wechselwirkungen zwischen Domänen, phylogenetische Relevanz und molekulare Evolution gestatten. Des weiteren ebnet sich der Weg für Subfamilien-orientierte Funktionsanalysen anhand von Motivunterscheidungen und für vergleichende Untersuchungen von nuklearen Domänen, um über die Sequenzähnlichkeit hinaus z.B. auf struktureller Ebene Homologien zu entdecken, die vertiefende Einblicke in evolutive Mechanismen ermöglichen.

Dies bedeutet für die Zukunft die Bereitstellung vielversprechender wissenschaftlicher Ansätze, die als vordringliche Herausforderung verstanden werden können.

Teil 4

Zusammenfassung

Im Rahmen dieser Doktorarbeit wurde der Großteil aller nuklearen Proteine annotiert und klassifiziert. Aus Literatur, Proteinsequenz- und Domänendatenbanken wurden bekannte nukleare Domänen ermittelt, ihre Grenzen unter Zuhilfenahme von Tertiärstrukturen oder Sekundärstrukturvorhersagen bestimmt und multiple Sequenzalignments erstellt. Die handgerfertigten Aligments wurden zur Anfertigung von Hidden Markov Models herangezogen und in das Domänenvorhersageprogramm Simple Modular Architecture Research Tool (Schultz et al. 1998, Schultz et al. 2000) (http://smart.embl-heidelberg.de/) implementiert. Hier sind umfassend Informationen über Literatur, phylogentische Verteilung, Anzahl beteiligter Proteine und Funktion für 164 Domänen (118 entstammen dieser Arbeit) mehr als 35000 Proteine abdeckend zusammengefasst.

Aufbauend auf der vollständigen Kollektion nuklearer Proteine wurden ausgewählte nukleare und nicht-nukleare Proteine auf der Grundlage homologiebasierender Sequenzanalyseverfahren untersucht. Die Arbeit führte zur Entdeckung von vier biologisch relevanten neuen Domänen:

- L27, eine neue Hetero-Dimer bildende Domäne in den Rezeptor-Targeting-Proteins Lin-2 and Lin-7 (Doerks et al. 2000)
- GRAM, eine neue Domäne in Glucosyltransferasen, Myotubularinen und anderen Membran-assoziierten Proteinen (Doerks et al. 2000)
- DDT, eine neue DNA-bindende Domäne in unterschiedlichen Transkriptionsfaktoren, Chromosom-assoziierten und anderen nuklearen Proteinen (Doerks et al. 2001)
- BSD, eine neue putativ DNA-bindende Domäne in Transkriptionsfaktoren, Synapsen-assozierten und anderen hypothetischen Proteinen (Doerks et al. submitted)

Abschliessend erfolgte die automatische Analyse von 24000 nuklearen Proteinen, aus denen 550 hypothetisch neue Domänen hervorgingen. Die intensive Aufarbeitung dieser 550 konservierten Sequenzbereiche erbrachte die Entdeckung von 28 neuen nuklearen oder teilweise nuklearen Domänen unterschiedlicher Speziesverbreitung, Funktion und biologischer Relevanz (Doerks et al. accepted).

Teil 4
Literatur

Aasland, R., Gibson, T. J. and Stewart, A. F. 1995. The PHD finger: implications for chromatin-mediated transcriptional regulation. *Trends Biochem Sci* **20***,* 56-59

Albert, S., Gallwitz, D. 1999 Two new members of a Ypt/Rab GTPase activating proteins. Promiscuity of substrate recognition. *J Biol Chem* **274**, 33186-33189

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. 1990 Basic local alignment search tool *J Mol Biol* **215**, 403-410

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

Altschul, S.F., Koonin, E.V. 1998 Iterated profile searches with PSI-BLAST - a tool for discovery in protein databases. *Trends Biochem Sci* **23**, 444-447

Anderson, A. M. 1996 *Curr. Biol.* ***6****,* 382-384

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D.R., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J.A., Zdobnov, E.M. 2001
The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29**, 37-40

Aravind, L. 2000. The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. *Trends Biochem Sci* **25**, 421-423.

Aravind, L. and Koonin, E. V. 2000 The U box is a modified RING finger - a common domain in ubiquitination. *Curr Biol* **10**, 1324-1324

Aravind, L. Landsman, D. 1998 AT-Hook motifs identified in wide variety of DNA-binding proteins *Nucleic Acids Res* **26***,* 4413-21

Bairoch, A., Apweiler, R. 2000 The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000 *Nucleic Acids Res* **28***,* 45-48

Baker, W., van Broek, den., Camon, A. E., Hingamp, P., Sterk, P., Stoesser, G. Tuli, M. A. 2000 The EMBL nucleotid sequence database. *Nucleic Acids Res* **28**, 19-23

Bardwell, V. J. and Treisman, R. 1994 The POZ domain: a conserved protein-protein interaction motif. *Genes Dev* **15**, 1664-1677

Barlev, N. A., Poltoratsky, V., Owen-Hughes, T., Ying, C., Liu, L., Workman, J. L., Berger, S. L. 1998 Repression of GCN5 histone acetyltransferase activity via bromodomain- mediated binding and phosphorylation by KU-DNA-dependent protein kinase complex *Mol Cell Biol* **18***,* 1349-1358

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. and Sonnhammer, E. L. 2000. The Pfam protein families database. *Nucleic Acids Res* **28***,* 263-266

Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., Ouellette, B. F., Rapp, B. A., Wheeler, D. L. 1999 GenBank *Nucleic Acids Res* **27***,* 263-266

Birney, E., Thompson, J.D., Gibson, T.J. 1996 PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res.* **24**, 2730-2739

Bochar, D. A., Savard, J., Wang, W., Lafleur, D. W., Moore, P., Cote, J., Shiekhattar, R. 2000 BRCA1 is associated with a human SWI/SNF-related complex: linking chromatin remodeling to breast cancer *PNAS* **97***,* 1038-43

Borg, J.-P., Straight, S. W., Kaech, S. M., de Taddeo-Borg, M., Kroon, D. E., Karnak, D., Turner, R. S., Kim, S. K. and Margolis, B. 1998 Identification of an evolutionarily conserved heterotrimeric protein complex involved in protein targeting. *J. Biol. Chem.* **273**, 31633-31636

Bork, P. & Sander C. 1993. A hybrid protein kinase-RNase in an interferon-induced pathway? *FEBS Lett* **334**, 149-152

Bowser, R., Giambrone, A., Davies, P. 1995 FAC1, a novel gene identified with the monoclonal antibody Alz50, is developmentally regulated in human brain
Dev Neurosci **17***,* 20-37

Brandt, P., Ramlow, S., Otto, B. and Bloecker, H. 1995 Nucleotide sequence analysis of a 32500 bp region of the right arm of *Saccharomyces cerevisiae* Chromosome IV *Yeast* **12**, 85-90

Buchberger, A., Howard, M. J., Proctor, M. and Bycroft, M. 2001. The UBX domain: a widespread ubiquitin-like module. *J Mol Biol* **307**, 17-24

Butz, S., Okamoto, M., Suedhof, T. C. 1998 A tripartite protein complex with the potential to couple synaptic vesicle exocytosis to cell adhesion in brain. *Cell* **94**, 773-782

Callebaut, I., de Gunzburg, J., Goud, B. and Mornon, J. 2001. RUN domains: a new family of domains involved in Ras-like GTPase signaling. *Trends Biochem Sci.* **26**, 79-83

Choi, H., Hong, J., Ha, J., Kang, J., Kim, S. Y. 2000 ABFs, a family of ABA-responsive element binding factors. *J Biol Chem* **275**, 1723-1730

Chothia, C. 1992 Proteins. One thousand families for the molecular biologist.
Nature **357**, 543-544

Clissold., P. M. and Ponting, C. P. 2001. JmjC: Cupin metalloenzyme-like domains in Jumonji, Hairless nand Phospholipase A2-beta. *Trends Biochem. Sci*, **26**, 7-9

Ciechanover, A. 1998 The ubiquitin-proteasome pathway: on protein death and cell life *Embo J* **17***,* 7151-7160

Cui, X., De Vivo, I., Slany, R., Miyamoto, A., Firestein, R. and Cleary, M. L. 1998. Association of SET domain and myotubularin-related proteins modulates growth control. *Nat. Genet.* **18**, 331-337

Davletov, B. A., Suedhof, T. C. 1993 A single C2 domain from synaptotagmin I is sufficient for high affinity Ca^{2+}/phospholipid binding. *J Biol Chem* **268**, 26386-26390

de Gouyon, B. M., Zhao, W., Laporte, J., Mandel, J. L., Metzenberg, A., Herman, G. E. 1997 Characterization of mutations in the myotubularin gene in twenty six patients with X-linked myotubular myopathy. *Hum Mol Genet* **6**, 1499-504

Dhalliun, C., Carlson, J. E., Zeng, L., He, C., Aggarwal, A. K., Zhou, M. M. 1999 Structure and ligand of a histone acetyltransferase bromodomain *Nature* **399***,* 491-496

Djabali, M., Selleri, L., Parry, P., Bower, M., Young, B. D. and Evans,G. A. 1992. A trithorax-like gene is interrupted by chromosome 11q23 translocations in acute leukaemias *Nat. Genet.* **2**, 113-118

Doerks, T., Bork, P., Kamberov, E., Makarova, O., Muecke, S.
and Margolis, B. 2000 L27, a novel heterodimerization domain in receptor targeting proteins Lin-2 and Lin-7 *Trends Biochem Sci* **25**, 317-318

Doerks, T., Strauss, M., Brendel, M. and Bork, P. 2000
GRAM, a novel domain in glucosyltransferases, myotubularins and other putative membrane-associated proteins *Trends Biochem Sci* **25**, 483-485

Doerks, T., Copley, R. R. and Bork, P. 2001. DDT, a novel domain in differrent transcription and chromosome remodeling factors. *Trends Biochem Sci*, **26**, 145-146

Doerks*, T., Copley*, R. R., Schultz, J., Ponting, C. P. and Bork P. 2001
Systematic identification of novel protein domain families associated with nuclear functions *Genome Res* accepted

Doolittle, R. F. 1995. The multiplicity of domains in proteins. *Annu Rev Biochem.* **64**, 287-314

Eddy, S. R. 1998. Profile hidden Markov models. *Bioinformatics* **14***,* 755-763

Eisenhaber, F. and Bork, P. 1998. Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol* **8**, 169-170

Etzold, T., Ulyanov, A., Argos, P.
SRS: information retrieval system for molecular biology data banks. 1996
Methods Enzymol **266**, 114-128

Fischer, L., Gerard, M., Chalut, C., Lutz, Y., Humbert, S., Kanno, M., Chambon, P. and Egly, J. M. 1992 Cloning of the 62-kilodalton component of basic transcription factor BTF2. *Science* **257**, 1392-1395.

Gaullier, J., Simonsen, A., D'Arrigo, A., Bremnes, B., Stenmark, H. 1998 FYVE fingers bind PtdIns(3)P. *Nature* **394**, 432-433

Gehring, W. J. 1985 Homeotic genes, the homeobox, and the spatial organization of the embryo. *Harvey Lect.* **81**, 153-172

Gehring, W. J., Affolter, M., Burglin, T. 1994 Homeodomain proteins *Annu Rev Biochem* **63***,* 487-526

Gibson, T. J., Hyvonen, M., Musacchio, A., Saraste, M., Birney, E. 1994 PH domain: the first anniversary. *Trends Biochem Sci* **19**, 349-353

Gileadi, O., Feaver, W. J. and Kornberg R. D. 1992 Cloning of a subunit of yeast RNA polymerase II transcription factor b and CTD kinase *Science* **257**, 1389-1392.

Giraudat, J., Parcy, F., Bertauche, N., Gosti, F., Leung, J., Morris, P. C., Bouvier-Durand, M., Vartanian, N. 1994 Current advances in abscisic acid action and signalling. *Plant Mol Biol* **26**, 1557-1577

Green, P., Lipman, D., Hillier, L., Waterston, R., States, D. and Claverie, J.M. 1993 Ancient conserved regions in new gene sequences and the protein database. *Science* **259**, 1711-1776

Heery, D. M., Kalkhoven, E., Hoare, S., Parker, M.G. 1997 A signature motif in transcriptional co-activators mediates binding to nuclear receptors *Nature* **387***,* 733-6

von Heijne, G. 1992 Membrane Protein Structure Prediction, Hydrophobicity Analysis and the Positive-inside Rule *J. Mol. Biol*. **225**, 487-494

Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K. and Hood, L. 1997 Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**, 609-614

Hershko, A. and Ciechanover, A. 1998 The ubiquitin system *Annu Rev Biochem* **67**, 425-479

Hofmann K. and Bucher, P. 1996. The UBA domain: a sequence motif present in multiple enzyme classes of the ubiquitination pathway. *Trends Biochem. Sci.* **21**, 172-173

Hofmann K. and Bucher, P., Falquet, L., Bairoch, A. 1999 The PROSITE database, its status in 1999 *Nucleic Acids Res* **27***,* 215-219

Hoskins, R., Hajnal, A., Harp, S., Kim, S. K. 1995 *Development* **122***,* 97-111

Hurley, J. H., Newton, A. C., Parker, P. J., Blumberg, P. M., Nishizuka, Y. 1997 Taxonomy and function of C1 protein kinase C homology domains. *Protein Sci* **6**, 477-480

International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921

Ito T., Levenstein, M. E., Fyodorov, D. V., Kutach, A. K., Kobayashi, R., Kadonaga, J. T. 1999 ACF consists of two subunits, Acf1 and ISWI, that function cooperatively in the ATP-dependent catalysis of chromatin assembly *Genes Dev* **13***,* 1529-39

Janin, J. and Chothia, C. 1985. Domains in proteins: definitions, location, and structural principles. *Methods Enzymol.* **115**, 420-430.

Jentsch, S., Seufert, W. and Hauser, H.-P. 1991. Genetic analysis of the ubiquitin system. *Biochim. Biophys. Acta* **1089**, 127-139

Jones, H. J., Hamana, N., Shimane, M. 2000 Identification and characterization of BPTF, a novel bromodomain transcription factor *Genomics* **63**, 35-39

Jones, H. J., Hamana, N., Nezu, J., Shimane, M. 2000 A novel family of bromodomains *Genomics* **63**, 40-45

Jordan-Sciutto, K. L., Dragich, M. J., Rhodes, J. L., Bowser, R. 1999 Fetal Alz-50 clone 1, a novel zinc finger protein, binds a specific DNA sequence and acts as a transcriptional regulator *J Biol Chem* **274***,* 35262-8

Kaech, S. M., Whitfield, C. W., Kim, S. K. 1998 The LIN-2/LIN-7/LIN-10 complex mediates basolateral membrane localization of the C. elegans EGF receptor LET-23 in vulval epithelial cells.*Cell 94*, 761-771

Kato, M., Mizuno, T., Shimizu, T., Hakoshima, T. 1997 Insights into multistep phosphorelay from the crystal structure of the C-terminal HPt domain of ArcB. *Cell* **88***,* 717-723

Kelley, L.A., Maccallum, R., Sternberg, M.J.E.
Recognition of Remote Protein Homologies Using Three-Dimensional Information to Generate a Position Specific Scoring Matrix in the program 3D-PSSM
RECOMB 99, Proceedings of the Third Annual Conference on Computational Molecular Biology
Pages 218-225
Editors: Sorin Istrail, Pavel Pevzner, Michael Waterman
Publisher: The Association for Computing Machinery, New York, New York 10036
April 1999

Koonin, E. V., Abagyan R. A. 1997 TSG101 my be the prototype of a class of dominant negative ubiquitin regulators. *Nat Genet* **16**, 330-331

Kornberg, R. D., Lorch, Y. 1999 Twenty-five years of the nucleosome, fundamental particle of the eukaryotic chromosome *Cell* **98***,* 285-294

Laporte, J., Blondeau, F., Buj-Bello, A., Tentler, D., Kretz, C., Dahl, N., Mandel, J. L. 1998 Characterization of the myotubularin dual specificity phosphatase gene family from yeast to human. *Hum Mol Genet* **7**, 1703-1712

Laporte, J., Guiraud-Chaumeil, C., Vincent, M. C., Mandel, J.L., Tanner, S. M., Liechti-Gallati, S., Wallgren-Pettersson, C., Dahl, N., Kress, W., Bolhuis, P.A., Fardeau, M., Samson, F., Bertini, E. 1997 Mutations in the MTM1 gene implicated in X-linked myotubular myopathy. ENMC International Consortium on Myotubular Myopathy. European Neuro-Muscular Center. *Hum Mol Genet* **6**, 1505-11

Laporte, J., Hu, L. J., Kretz, C., Mandel, J. L., Kioschis, P., Coy, J. F., Klauck, S. M., Poustka, A., Dahl, N. 1996 A gene mutated in X-linked myotubular myopathy defines a new putative tyrosine phosphatase family conserved in yeast. *Nat Genet* **13**, 175-82

Lorick, K. L., Jensen, J. P., Fang, S., Ong, A. M., Hatakeyama, S. & Weissmann, A. M. 1999. RING fingers mediate ubiquitin-conjugating enzyme (E2)-dependent ubiquitination. *Proc. Natl. Acad. Sci.* **96**, 11364-11369

Lu, X., Meng, X., Morris, C. A., Keating, M. T. 1998 A novel human gene, WSTF, is deleted in Williams Syndrome *Genomics* **54***,* 241-249

Lupas, A., Van Dyke, M. and Stock, J. 1991. Predicting coiled coils from protein sequences. *Science* **252**, 1162-1164

Miller, J., McLachlan, A. D. and Klug, A. 1985. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J.* **4**, 1609-1614

Morris, C. A., Leonard, C. O., Dilates, C. 1988
Natural history of William's syndrome: Physical characteristics
J. Pediatric **113***,* 318-326

Murzin, A.G. 1998 How far divergent evolution goes in proteins
Curr Opin Struc Biol **8**, 380-387

Nakamura, H., Yoshida, M., Tsuiki, H., Ito, K., Ueno, M., Nakao, M.,Oka, K., Tada, M., Kochi, M., Kuratsu, J., Ushio, Y. and Saya, H. 1998. Identification of a human homolog of the Drosophila neuralized gene within the 10q25.1 malignant astrocytoma deletion region *Oncogene* **16**, 1009-1019

Neuwald, A. F. 1997 A shared domain between a spindle assembly checkpoint protein and Ypt/Rab-specific GTPase-activators. *Trends Biochem Sci* **22**, 243-244

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

Orti, R., Rachidi, M., Vialard, F., Toyama, K., Lopes, C., Taudien, S., Rosenthal, A., Yaspo, M.-L., Sinet, P. M. and Delabar, J. M. 2000. Characterization of a novel gene, C21orf6, mapping to a critical region of chromosome 21q22.1 involved in the monosomy 21 phenotype and of its murine ortholog, orf5. *Genomics* **64**, 203-210

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. and Chothia, C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods.
J Mol Biol **284**, 1201-1210**21**, 172-173

Park, J., Teichmann, S.A., Hubbard, T. and Chothia, C. 1997 Intermediate sequences increase the detection of homology between sequences. *J Mol Biol* **273**, 349-354

Ponting C. P., Yu-Dong, C., Bork, P. 1997 The breast cancer gene product TSG101: a regulator of ubiquitination? *J Mol Med* **75**, 467-469

Rizo, J., Suedhof, T. C. 1998 C_2-domains, structure and function of a universal Ca^{2+}-binding domain. *J Biol Chem* **273**, 15879-15882

Reisch, D., Heimbeck, G., Hofbauer, A., Klagges, B., Pflugfelder, G. O. and Buchner, E. 1995 The sap47 gene of Drosophila melanogaster codes for a novel conserved neuronal protein associated with synaptic terminals. *Brain Res Mol Brain Res* **32**, 45-54.

Rost, B, Sander, C. & Schneider, R. 1994. PHD--an automatic mail server for protein secondary structure prediction. *Cabios* **10**, 53-60

Sattleger, E., Hinnebusch, A. G. & Barthelmess, I. B. 1998. cpc-3, the Neurospora crassa homologue of yeast GCN2, encodes a polypeptide with juxtaposed eIF2alpha kinase and histidyl-tRNA synthetase-related domains required for general amino acid control. *J. Biol. Chem.* **273**, 20404-20416

Schuler, G. D., Altschul, S. F., Lipman, D. J. 1991 A workbench for multiple alignment construction and analysis. *Proteins* **9**, 180-190

Schultz, J., Copley, R., Doerks, T., Ponting, C. and Bork, P. 2000. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res* **28**, 231-234

Schultz, J., Milpetz, F., Bork, P. and Ponting, C. P. 1998. SMART, a simple modular architecture research tool: Identification of signalling domains *Proc. Natl. Acad. Sci.* **95**, 5857-5864

Shamu, C. E. and Walter, P. 1996. Oligomerization and phosphorylation of the Ire1p kinase during intracellular signaling from the endoplasmic reticulum to the nucleus. *EMBO J.* **15**, 3028-3039

Sidrauski, C. and Walter, P. 1997. The transmembrane kinase Ire1p is a site-specific endonuclease that initiates mRNA splicing in the unfolded protein response.*Cell* **90**, 1031-1039

Singer, J. D., Manning, B. E. and Formosa, T. 1995 Control of single copy DNA replication requires genes that act in ubiquitin metabolism *EMBL Data Library*, U19857

Sonnhammer, L. L. and Durbin, R. 1995 A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis.
Gene **167**, GC1-10

Sottrup-Jensen, L. et al. 1975 Amino-acid sequence of activation cleavage site in plasminogen: homology with "pro" part of prothrombin. *Proc Natl Acad Sci* **72**, 2577-2581

Stec, I, Wright, T. J., van Ommen, G. J. B., de Boer, P. A. J., van Haeringen, A. Moorman, A. F. M, Altherr, M. R. and den Dunnen, J. T. 1998. WHSC1, a 90 kb SET domain-containing gene, expressed in early development and homologous to a Drosophila dysmorphy gene maps in the Wolf-Hirschhorn syndrome critical region and is fused to IgH in t(4;14) multiple myeloma *Hum. Mol. Genet.* **7**, 1071-1082

Suzuki, T., Park, H., Hollingsworth, N. M., Sternglanz R. and Lennarz W. J. 2000. PNG1, a yeast gene encoding a highly conserved peptide:N-glycanase. *J. Cell Biol.* **149**, 1039-1052

Thompson, J. D., Higgins, D. G. and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.*Nucleic Acids Res.* **22**, 4673-4680

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG 1997
The CLUSTAL_X windows interface: flexible strategies for multiple
sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **15**,
4876-4882

Walker, D. R. and Koonin E. V. 1997. SEALS: a system for easy analysis of lots of sequences.*Ismb* **5**, 333-339

Warnecke, D., Erdmann, R., Fahl, A., Hube, B., Mueller, F., Zank, T., Zaehringer, U., Heinz, E. 1999 Cloning and functional expression of UGT genes encoding sterol glucosyltransferases from *Saccharomyces cerevisiae*, *Candida albicans*, *Pichia pastoris*, and *Dictyostelium discoideum*. *J Biol Chem* **274**, 13048-13059

Wootton, J. C. and Federhen, S. 1996. Analysis of compositionally. biased regions in sequence databases. *Methods Enzymol* **266**, 554-571

Zhang, P., Talluri, S., Deng, H., Branton, D., Wagner, G. 1995 Solution structure of the pleckstrin homology domain of Drosophila beta-spectrin. *Structure* **3**, 1185-1195

Zhou, A., Hassel, B. A. and Silverman R. H. 1993. Expression cloning of 2-5A-dependent RNAase: a uniquely regulated mediator of interferon action. *Cell* **72**, 753-765

Zollman, S., Godt, D., Prive, G. G., Couderc, J. L. and Laski, F. A. 1994
The BTB domain, found primarily in zinc finger proteins, defines an evolutionarily conserved family that includes several developmentally regulated genes in Drosophila. *Proc Natl Acad Sci U S A* **91***,* 10717-10721.

Weitere Quellen:

Jörg Schultz 2001 Dissertation: SMART, a Simple Modular Architecture Research Tool - Development and Applications

Teil 6
Publikationen

Publikationen

1. Bork, P., Doerks, T., Springer, T. A. and Snel, B. 1999 Domains in plexins: links to integrins and transcription factors *Trends Biochem Sci* **24**, 261-263

2. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. and Bork, P. 2000 SMART: a web-based tool for the study of genetically mobile domains *Nucleic Acids Res* **28**, 231-234

3. Stutz, F., Bachi, A., Doerks, T., Braun, I. C., Seraphin, B., Wilm, M., Bork, P. and Izaurralde, E 2000 REF, an evolutionary conserved family of hnRNP-like proteins, interacts with TAP/Mex67p and participates in mRNA nuclear export *RNA* **6**, 638-650

4. Schultz, J., Doerks, T., Ponting, C. P., Copley, R. R. and Bork P. 2000 More than 1,000 putative new human signalling proteins revealed by EST data mining *Nat Genet* **25**, 201-204

5. Doerks, T., Bork, P., Kamberov, E., Makarova, O., Muecke, S. and Margolis, B. 2000 L27, a novel heterodimerization domain in receptor targeting proteins Lin-2 and Lin-7 *Trends Biochem Sci* **25**, 317-318

6. Suyama, M., Doerks, T., Braun, I. C., Sattler, M., Izauralde, E. and Bork, P. 2000 Prediction of structural domains of TAP reveals details of its interaction with p15 and nucleoporins *EMBO Reports* **1**, 53-58

7. Dandekar, T., Huynen, M., Regula, J. T., Ueberle, B., Zimmermann, C. U., Andrade, M. A., Doerks, T., Sanchez-Pulido, L., Snel, B., Suyama, M., Yuan, Y. P., Herrmann, R. and Bork, P. 2000 Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames *Nucleic Acids Res* **28**, 3278-3288

8. Doerks, T., Strauss, M., Brendel, M. and Bork, P. 2000 GRAM, a novel domain in glucosyltransferases, myotubularins and other putative membrane-associated proteins *Trends Biochem Sci* **25**, 483-485

9. Doerks, T., Copley, R. R. and Bork, P. 2001
DDT, a novel domain in differrent transcription and chromosome remodeling factors *Trends Biochem Sci* **26**, 145-146

10. International Human Genome Sequencing Consortium 2001
Initial sequencing and analysis of the human genome
Nature **15**, 860-921

11. Ciccarelli, F. D., Copley, R. R., Doerks, T., Russell, R. B. and Bork P. 2001
CASH - a beta-helix domain widespread among carbohydrate-binding proteins
Trends Biochem Sci submitted

12. Doerks*, T., Copley*, R. R., Schultz, J., Ponting, C. P. and Bork P. 2001
Systematic identification of novel protein domain families associated with nuclear functions *Genome Res* accepted

13. Doerks, T. Huber, S., Buchner, E. and Bork P. 2001
BSD, a novel domain in transcription factors, synapse-associated and other hypothetical proteins *Trends Biochem Sci* submitted

14. Letunic, I., Goodstadt, L., Dickens, N. J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R. R., Ponting, C. P. and Bork, P. 2002
Recent improvements to the Smart domain-based sequence annotation resource *Nucleic Acids Res* **30**, 1-3

15. Copley, R. R., Doerks, T., Letunic, I. and Bork, P. 2001
Protein Domain Analysis in the Era of Complete Genoms *FEBS lett* submitted

16. Nicodeme, P. Doerks, T. and Vingron, M. 2001
Motif Statistics Apllied to Proteome Analysis submitted

17. Kerkhoff, E., Simpson, J. C., Leberfinger, C. B., Otto, I. M., Doerks, T., Bork, P., Rapp, U. R., Raabe, T. and Pepperkok, R. 2001
The Spir actin organizers are involved in vesicle transport processes
Curr Biol accepted

Teil 7
Anhang

Domänen Name	Beschreibung	Anzahl der Proteine in nrdb*	Achaea	Bacteria	Homo sapiens	Drosophila melanogaster	Caenorhabditis elegans	Arabidopsis thaliana	Saccharomyces cerevisae
35EXOc	katalytische Domäne der 3'-5' Exonuklease	139	--	76	5	6	5	21	1
53EXOc	katalytische Domäne der 5'-3' Exonuklease	98	--	83	--	--	--	2	--
A1pp+	Appr-1-p "processing enzyme"	172	6	26	17	2	1	5	2
ADEAMc	tRNA-spezifische und Doppelstrang RNA-Adenosin-Deaminase	51	--	--	12	5	5	1	1
AP2+	DNA-bindende Domäne in Pflanzen	248	--	--	--	--	--	185	--
AP2Ec+	Ap Endonuklease-Familie 2	44	10	21	--	--	1	--	1
AT_hook+	bevorzugt A/T reiche DNA-Regionen bindende Domäne	257	2	11	53	35	17	46	6
AWS	Domäne, assoziiert mit der SET Domäne	30	--	--	12	5	3	6	1
AXH	Domäne in Ataxinen und HMG-Domänen-tragenden Proteinen	8	--	--	3	1	1	--	--
BAH+	Bromo- Domänen homologe Domäne	115	--	--	19	9	5	33	5

BASIC+	Basische Domäne in HLH-Proteinen der MYOD-Familie	59	--	--	7	1	--	--	--
BBC+	Domäne C-terminal zur BBOX	85	--	--	44	6	10	--	--
BCL	(B-Zell-Lymphoma), mit den Regionen BH1 und BH2	121	--	--	30	3	1	--	--
BRCT+	"brest cancer carboxy-terminal domain	257	--	55	36	17	31	19	11
BRIGHT	"ARID (A/T-rich region interaction) domain	78	--	--	28	8	6	9	2
BRLZ	Basische Region und Leuzin Zipper	682	--	--	117	43	29	107	16
BROMO+	Bromo Domäne	225	--	--	79	27	21	31	9
BTP	Domäne in Bromodomänen-tragenden Transkriptions-faktoren	25	--	--	4	6	2	6	1
CBF	CCAAT-bindende Transkriptions-faktoren	35	--	--	1	1	2	14	1
CHK	ZnF_C4 - und HLH- Domänen-tragende Kinase-Domäne	72	--	2	--	44	26	--	--
CHROMO+	Chromatin-organisierende Domäne	237	--	--	55	24	27	30	4
ChSh+	"Chromo Shadow Domain"	39	--	--	9	5	4	1	--
CYCLIN+	Domäne in Cyclinen und TFIIB	588	33	--	66	34	29	74	18
CPDc	katalytische	85	--	1	14	10	5	25	5

	Domäne von CTD-ähnlichen Phosphatasen								
CSZ	Domäne in chromatin-organisierenden S!-Domänen-tragenden Proteinen	47	--	25	4	4	3	2	1
Cu_FIST	"Copper-Fist"	8	--	--	--	--	--	--	3
DSRM	Doppelstrang-RNA-bindende Domäne	262	--	43	62	29	15	20	2
DWA	Domäne A in "dwarfin family proteins"	133	--	--	19	11	12	--	--
DWB	Domäne B in "dwarfin family proteins"	85	--	--	12	12	15	--	--
DZF	Domäne in DSRM oder ZnF_C2H2 - Domänen-tragenden Proteinen	46	--	--	30	2	2	--	--
ENDO3c	katalytische Domäne der Endonuklease III	209	31	125	13	2	1	14	4
ETS	"erythroblast transformation specific domain	207	--	2	54	11	13	--	--
EXOIII	katalytische Domäne verschiedener Exonukleasen	219	2	113	21	9	12	14	5
FBD	pflanzen-spezifische Domäne in FBOX- und BRCT-Domänen-	162	--	--	--	--	--	159	--

	tragenden Proteinen								
FES	Eisen/Schwefel-bindende Domäne in DNA-Lyasen	117	17	69	11	1	1	7	1
FH	"Forkhead"-Domäne	282	--	--	59	21	27	--	5
FYRC^{+}	"FY-reiche"-Domäne (C-terminale Region)	28	--	--	9	5	1	7	--
FYRN^{+}	"FY-reiche"-Domäne (N-terminale Region)	26	--	--	8	5	1	7	--
G_patch^{+}	Glyzin-reiche nukleinsäure-bindende Domäne	152	--	--	46	18	16	17	5
H15	Domäne in Histonfamilien 1 und 5	244	--	--	21	3	13	25	1
H2A	Domäne in Histonfamilie 2A	208	--	--	34	2	18	16	3
H2B	Domäne in Histonfamilie 2B	192	--	--	31	1	10	14	2
H3	Domäne in Histonfamilie 3	421	1	--	28	3	14	20	2
H4	Domäne in Histonfamilie 4	127	--	--	15	1	15	10	1
HALZ	Homeobox-assoziierter Leuzin Zipper	35	--	--	--	--	--	13	--
HELICc	Helikase Superfamilie, C-terminale Domäne	1870	83	411	200	92	106	196	74
HhH1^{+}	"helix-hairpin-	513	64	332	20	5	2	13	2

	heilx"-DNA-Bindungsmotiv, Klasse 1								
HhH2	"helix-hairpin-heilx"-DNA-Bindungsmotiv, Klasse 2	166	10	85	11	6	4	10	4
HLH	"helix-loop-helix"-Domäne	1220	--	--	185	84	51	156	10
HMG	"high mobility group"	710	--	--	113	32	24	22	10
HMG17	"high mobility group" 14 und 17	28	--	--	10	--	--	--	--
$HNHc^{+}$	HNH-Nukleasen	248	2	122	1	--	1	4	4
HNS^{+}	Domäne in histon-ähnlichen Proteinen der HNS-Familie	35							
HOLI	Ligandenbindungs-domäne in nuklearen Hormonrezeptoren	990	--	--	103	31	383	--	--
HOX	Homeodomäne	2778	--	--	321	167	140	135	10
$HRDC^{+}$	Domäne in Helikasen und RNasen, c-terminal	78	--	40	6	3	3	6	2
HSA	Domäne in Heliasen und assoziiert mit SANT Domänen	24	--	--	5	5	3	3	2
HSF	"heat shock factor"	92	--	--	8	1	2	30	5
HTH_ARAC	"helix_turn_helix, arabinose control proteins"	605	---	564	--	--	--	--	--
HTH_ARSR	"helix_turn_helix, arsenical resistance operon repressor"	57	178	---	---	---	---	---	---

HTH_ASNC	"helix_turn_helix, ASNC type"	180	50	125	---	---	---	---	---
HTH_CRP	"helix_turn_helix, cAMP Regulatory proteins"	164	10	147	---	---	---	---	---
HTH_DEOR	"helix_turn_helix, deoxyribose operon repressor"	156	3	142	---	---	---	---	---
HTH_DTXR	"helix_turn_helix, diphteria toxin regulatory element"	56	19	34	---	---	---	---	---
HTH_GNTR	"helix_turn_helix, gluconate operon transcriptional repressor"	438	2	405	---	---	---	---	---
HTH_ICLR	"helix_turn_helix, isocitrate lyase regulation"	118	1	107	---	---	---	---	---
HTH_LACI	"helix_turn_helix, lactose operon repressor"	316	--	298	---	---	---	---	---
HTH_LUXR	"helix_turn_helix, Lux Regulon"	524	---	493	---	---	---	---	---
HTH_MARR	"helix_turn_helix, multiple antibiotic resistance protein"	297	7	275	---	---	---	---	---
HTH_MERR	"helix_turn_helix, mercury resistance"	262	2	244	---	---	---	---	---
HTH_XRE	"helix_turn_helix, XRE-family like proteins"	737	51	532	2	1	1	5	1
IPT	ig-ähnliche Domäne in Plexinen und Transkriptions-faktoren	147	---	2	46	16	4	---	2
IRF	"interferon	66	--	--	16	--	--	--	--

	regulatory factor"								
IRO^{+}	Motiv in Iroquois-Klasse Homeodomänen-tragenden Proteine	29	---	---	2	6	1	---	---
JmjC^{+}	Domäne in Cupin Metalloenzym-Superfamilie	159	---	10	54	12	15	21	5
JmjN^{+}	Domäne in der Jumonji-Familie von Transkriptions-faktoren	52	---	---	15	4	2	9	3
KH	K-Homologie RNA-Bindungsdomäne	596							
KRAB^{+}	"krueppel associated box"	397	---	---	270	---	---	---	---
LEM^{+}	Domäne in kernmembran-assoziierten Proteinen	28	---	---	7	6	3	---	---
LER^{+}	"leucin-rich region"	107	---	---	73	---	---	---	---
LIGANc	katalytische Domäne der Ligase N	71	---	63	1	---	---	---	---
MADS	MADS-Box DNA-Bindungsdomäne	610	---	---	5	5	3	138	4
MBD	Methyl-CpG-Bindungsdomäne	62	---	---	26	6	3	10	---
MCM	"minichromosome maintenance proteins"	118	13	---	21	12	7	10	6
MUTSac	ATPase-Domäne in "DNA mismatch repair"-Proteinen der MUTS-Familie	217	8	139	10	2	8	9	6
MUTSd	DNA-Bindungsdomäne in "DNA mismatch	131	3	69	9	2	7	8	6

	repair"-Proteinen der MUTS-Familie								
NEUZ	Domäne in "neuralized proteins"	20	---	---	6	6	3	---	---
ORANGE+	Domäne in Transkriptions-faktoren	82	---	---	17	14	---	---	---
ParBc+	ParB-ähnliche Nukleasedomäne	170	13	138	1	1	---	1	1
PAX+	"Paired Box domain"	203	---	---	17	16	11	---	---
PAC+	Domäne c-terminal zu Pas Domänen	577	36	228	49	20	12	21	---
PAS+	teilweise nukleare Signaltransduktions domäne	962	63	480	65	24	16	21	4
PHD+	PHD Zinkfinger	641	---	---	188	58	53	192	16
POL3Bc	DNA-Polymerase III, beta subunit	88	---	85	---	---	---	---	---
POLAc	DNA-Polymearase, Domäne A	118	---	75	4	3	2	3	1
POLBc	DNA-Polymerase (Typ B)	369	40	5	11	9	5	6	4
POLIIIAc	DNA-Polymerase, alpha subunit	140	15	113	---	---	---	1	---
POLXc	DNA-Polymerase X Familie	46	3	6	9	---	---	2	2
POP4+	Domäne in menschlicher Untereinheit der RNase MRP und P und in archaeal Ribonukleo-proteinen	19	11	---	2	1	1	---	1
PostSET+	Cystein-reiche Domäne c-terminal zu SET Domänen	108	---	---	34	14	13	26	2

POU^{+}	Domäne in Pit-Oct-Unc Transkriptions-faktoren	195	1	1	28	8	6	---	---
POX	Domäne assoziiert mit HOX Domänen	28	---	---	---	---	---	21	---
PRE_C2HC	Domäne n-terminal zu C2HC Zinkfingern	14	---	---	---	8	---	---	---
PreSET	Domäne n-terminal zu einigen SET Domänen	56	---	---	14	5	3	19	---
PRY	Domäne assoziert mit SPRY Domänen	155	---	---	83	---	---	---	---
PSP	Prolin-reiche Domäne in spliceosomalen Proteinen	19	---	---	7	2	3	3	1
PUA	angenommen RNA-bindende Domäne in Pseudouridin-Synthasen	129	46	34	7	7	3	5	5
PUG	Domäne in Proteinkinasen, N-Glykanasen und anderen nuklearen Proteinen	31	---	---	9	4	1	8	1
RIBOc	katalytische Domäne der Ribonuklease III	102	---	55	10	4	5	13	3
RPOL4c	RNA-Polymerase II Untereinheit	21	---	---	2	2	3	3	2
RPOL8c	RNA-Polymerase Untereinheit 8	14	---	---	3	1	1	2	1
RPOL9	RNA-Polymerase Untereinheit 9	38	14	---	4	3	3	2	4
RPOLA_N	RNA-Polymerase I	230	15	62	3	4	5	7	3

	Untereinheit (N-terminus)								
RPOLCX	RNA-Polymerase Untereinheit CX	19	4	---	3	---	1	3	1
RPOLD	RNA-Polymerase Untereinheit D	145	13	53	6	2	2	7	2
RPR	Domäne in Proteinen, die in pre-mRNA-Regulation involviert sind	52	---	1	12	5	5	14	3
RRM	RNA-bindende Domäne (Motive RNP1 und RNP2)	1841	---	23	423	163	143	321	54
RWD	Domäne in Ring-Finger und WD-repeat Domänen tragenden Proteinen	73	---	---	17	8	7	4	5
SAND^{+}	DNA-bindende Domäne	59	---	---	27	1	4	---	---
SANT^{+}	DNA-bindende Domäne in Transkriptions-faktoren	953	---	---	78	37	28	440	20
SAP^{+}	DNA-bindende Domäne, involviert in Chromosomen-Organisation	122	---	---	43	11	7	10	5
SET^{+}	Domäne in (Su(var3-9, Enhancer-of-zeste. Trithorax)-Proteinen	287	---	---	72	36	42	56	8
SFM^{+}	Domäne in Splicing-Faktor-Proteinen	17	---	---	5	2	2	2	1
Skp1^{+}	Domäne in Skp1-	104	---	---	6	11	28	26	3

	Proteinen								
Sm	Domäne in snRNP-Proteinen	221	12	---	49	17	18	26	16
SMR	Domäne in kleinen MutS-ähnlichen Proteinen	73	---	36	1	2	6	15	2
SPK	Domäne in SET- und PHD-Domänen-tragenden Proteinen	45	---	---	---	---	45	---	---
SRA	Domäne in Deinococcus radiodurans	32	---	2	2	---	---	24	---
STE	Domäne in STE-Transkriptions-faktoren	10	---	---	---	---	---	---	1
SWAP	Domäne in Splicing-Regulations-proteinen	50	---	---	11	7	4	17	1
TBOX	DNA-bindende Domäne, Erstfund iim T locus des Brachyury-Proteins	164	---	---	20	12	24	---	---
TCH	Domäne in Transkriptionsfaktoren und CHROMO-Domänen Helikasen	21	---	---	10	6	2	---	---
TEA	Domäne in "transcriptional enhancer aktivators"	24	---	---	9	1	3	---	1
TFIIE^{+}	Domäne in "Transcription initiation factor IIE"	21	7	1	2	1	1	3	1
TFS2M	Domäne in	44	---	---	9	3	1	6	2

	"transcription elongation factor S-II"								
TFS2N	Domäne in "transcription elongation factor S-II" (N-terminus)	54	---	---	7	3	2	14	1
TOP1Ac	Bakterielle DNA-Bindungsdomäne der DNA-Topoisomerase I	123	20	81	3	2	4	3	1
TOP1Bc	Bakterielle ATP-Bindungsdomäne der DNA-Topoisomerase I	121	20	79	3	2	4	3	1
TOP2c	katalytische Domäne der Topoisomerase II	678	5	499	7	1	6	6	1
TOP4c	katalytische Domäne der Topoisomerase IV	337	5	258	5	1	5	4	2
TOPEUc	katalytische Domäne der eukaryotischen Topoisomerase	35	---	---	3	1	3	3	1
TOPRIM	Domäne in DNA-Primasen, Topoisomerasen und anderen Enzymen	287	38	201	3	2	4	5	1
TUDOR^{+}	Domäne in Proteinen nukleinsäure-assoziierter Komplexe	121	---	---	42	20	11	6	---
UAS	Domäne in UBA-Domänen-tragenden Proteine	29	---	---	8	4	2	8	2

Ubox^{+}	Modifizierte Ring Finger Domäne	132							
XPGI	I-Region in Xeroderma pigmentosum-Proteinen	74	10	---	10	6	4	8	7
XPGN	N-Region in Xeroderma pigmentosum-Proteinen	76	10	---	14	6	4	10	6
ZnF_BED	DNA-Bindungsdomäne in Transposasen	70	---	---	7	14	4	19	1
ZnF_A20^{+}	A20-ähnliche Zinkfinger	37	---	---	11	2	2	10	---
ZnF_AN1^{+}	AN1-ähnliche Zinkfinger	55	4	---	8	3	3	12	2
ZnF_C2C2	DNA-Bindungsdomäne in Transkriptions-faktoren	98	14	1	9	6	5	5	5
ZnF_C2H2	Zinkfinger DNA-Bindungsdomäne	3388	14	14	1006	432	253	203	51
ZnF_C2HC	DNA-Bindungsdomäne in erstlinig viralen Nukleocapsid-Proteinen	2765	---	5	71	53	57	270	15
ZnF_C3H1	Zinkfinger, DNA-Bindungsdomäne	308	---	---	61	22	39	67	7
ZnF_C4	Zinkfinger in nuklearen Rezeptoren	1083	---	---	112	35	378	---	---
ZnF_DBF	Zinkfinger in DBF-ähnlichen Proteinen	13	---	---	2	4	---	---	1
ZnF_GATA^{+}	DNA-Bindungsdomäne	191	---	---	23	12	19	31	12

	(sequenz-spefisch an GATA-Box)								
ZnF_NFX	Domäne in Transkriptions-repressoren NK-X1	19	---	---	6	3	2	2	1
ZnF_PMZ	Zinkfinger in pflanzlichen Transposasen	138	---	---	3	---	---	100	---
ZnF_RBZ	Zinkfinger in Ran-bindende Proteine	168	---	2	43	19	15	33	2
ZnF_TTF	Zinkfinger in Transposasen und Transkriptions-faktoren	17	---	---	1	1	---	7	---
ZnF_U1[+]	U1-ähnliche Zinkfinger	113	---	---	36	16	8	13	5
ZPW		65	---	---	---	48	12	3	---

Tabelle (im Anhang): Beschreibung von Tabelle im Anhang. Liste der in Smart implemtierten überwiegend nuklearen Domänen. Erste Spalte: Domänen-Name; zweite Spalte: Beschreibung der Domäne (z.B. assoziierte Domänen oder genauer charakterisierte Proteine); dritte Spalte: Anzahl der Proteine, die die Domäne tragen in nrdb; vierte Spalte: Anzahl der Proteine in Archebakterien; fünfte Spalte: Anzahl der Proteine in Eubakterien; sechste Spalte: Anzahl der Proteine in *Homo sapiens;* sechste Spalte: Anzahl der Proteine in *Drosophila melanogaster*; siebte Spalte: Anzahl der Proteine in *Caenorhabditis elegans*; achte Spalte: Anzahl der Proteine in *Arabidopsis thaliana*; ; achte Spalte: Anzahl der Proteine in *Saccharomyces cerevisae.*

* Stand der Datenbank nrdb 30.09.01

+ Alignment und Hidden Markov Model wurden von Chris Ponting bereitgestellt.

Lebenslauf

Persönliche Daten

Name:	Tobias Doerks
Anschrift:	Marktstr. 47 69123 Heidelberg
Geburtsdatum:	01.11.1971
Familienstand:	verheiratet
Staatsangehörigkeit:	deutsch

Schulausbildung / Studium

1978-1982	Grundschule Quickborn
1982-1991	Dietrich-Bonhoeffer-Gymnasium Quickborn
Juni 1991	Allgemeine Hochschulreife
07/1991-07/1992	Wehrdienst
10/1992-10/1994	Grundstudium der Biologie an der Johann Wolfgang Goethe-Universität Frankfurt a. M.
10/1994-09/1998	Hauptstudium der Biologie an der Johann Wolfgang Goethe-Universität Frankfurt a. M. Hauptfächer: Humangenetik und Anthropolgie, Nebenfächer: Mikrobiologie und Pharmakologie
09/1997-05/1998	Erstellung der Diplomarbeit mit dem Thema "Methodenentwicklung und Anwendungsbeispiele für Protein- und Nukleotidsequenzanalysen" unter Betreuunug von Prof. Dr. Anna Starzinski-Powitz im Fachbereich Anthropologie und Humangenetik der Johann Wolfgang Goethe-Universität in Frankfurt am Main und von Dr. P. Bork als Besucher am Europäischen Laboratorium für Molekularbiologie (EMBL)
06/1998-11/2001	Anfertigung einer Promotionsarbeit als PhD am am Europäischen Laboratorium für Molekularbiologie (EMBL) mit dem Thema: "Annotation und Klassifikation nuklearer Domänen" unter Betreuung von Dr. P Bork in der Abteilung Strukturbiologie/Bioinformatik