Methodenentwicklung und Anwendungsbeispiele für Protein- und Nukleotidsequenzanalysen

Diplomarbeit

im Fachbereich Biologie

der Johann Wolfgang Goethe-Universität

Frankfurt am Main

von

Tobias Doerks

Frankfurt am Main, Mai 1998

Inhaltsverzeichnis

Inhaltsverzeichnis

1. Einleitung

1.1. Bioinformatik

Die Bioinformatik dient der Lösung biologischer Probleme und Erkenntnisgewinnung mit Hilfe informatischer Methodik. und stellt das Bindeglied zwischen der Informationswissenschaft und der Lehre des Lebens dar. Eine festgeschriebene Definition des Begriffes „Bioinformatik" existiert nicht: Sie umfaßt ein weites Feld, beginnend bei der automatischen Sequenzierung ganzer Genome, über Funktionsanalysen durch Homologiesuchen in Datenbanken, Strukturvorhersagen und Modelling, bis hin zur chipgesteuerten Prothetik.

Unterstützende Arbeit in der Molekularbiologie leistet die Bioinformatik bei der Aufnahme und Verwaltung von Nukleotid– und Aminosäuresequenzen in Datenbanken. Der rasante Fortschritt bei der Sequenzierung ganzer Genome führt zu explosionsartig ansteigender Datenfülle und damit zu stetig wachsenden bioinformatischen Anwendungsmöglichkeiten, die ihrerseits notwendig sind, um diese Datenfülle zu bewältigen.

Die Genomprojekte erbrachten bislang die vollständige Sequenz der Genome von Species aus allen drei Überreichen:

Eubakterien: *Haemophilus influenza* [17], *Mycoplasma genitalium* [18], *Synechocystis sp* [29]. u.a.

Archaebakterien: *Methanococcus jannaschii* [16] u.a.

Eukaryonten: *Saccharomyces cerevisiae* u.a.

Ende 1998 soll auch die Sequenzierung des Nemathelminten (Rundwurm) *Caenorhabditis elegans* und den Menschen abgeschlossen sein.

Die Bedeutung der immer umfangreicher werden Datenbanken und der damit in Verbindung stehenden Homolgiesuchen erklärt sich aus der Annahme, daß die Primärstruktur eines Proteins den wesentlichen Informationsgehalt zur Deutung der dreidimensionalen Struktur zur Verfügung stellt und damit Rückschluß auf die Funktion erlaubt.

Bis auf wenige Ausnahmen in einigen Prokaryonten kodieren Nukleotidtripletts für die 21 proteinogenen Aminosäuren (Zahl einschließlich Selenocystein). So kann eine bekannte cDNA – Sequenz, EST (Expressed Sequence Tag) oder ein „Offener Leserahmen" in die zugehörige Aminosäuresequenz übersetzt werden. Nimmt man an, daß in Abhängigkeit von exogenen Faktoren wie Temperatur, das Vorhandensein spezifischer Faltungsproteine (Chaperone) und anderes die Primärstruktur ursächlich für Sekundär – und Tertiärstruktur ist, bedeutet das Wissen um ein in seiner Funktion beschriebenes homologes Protein die Möglichkeit der Funktionsvorhersage (siehe Kapitel 1.3.). Da es bei der Fülle von Sequenzen unmöglich ist, die Funktion ihrer Proteinprodukte molekularbiologisch im Experiment

aufzuklären, sind vergleichende Sequenzanalysen eine wissenschaftliche Notwendigkeit.

War die Bedeutung der Bioinformatik vor zehn Jahren noch nicht abzusehen, ist sie heute als wesentlicher Bestandteil der biologischen Forschung nicht mehr wegzudenken und kann gegenwärtig als eine der am stärksten expandierenden Wissenschaften betrachtet werden. Die Sequenzierung von Genen bis hin zu vollständigen Genomen ist nur der Anfang. Zukünftig werden die Interpretation von Einzelsequenzen und multiplen Sequenzalignments, die zur Herleitung von Sekundär – und Tertiärstrukturen führen, verstärkt in den Vordergrund treten. Im Rahmen der Sequenzanalysen ist die Unterscheidung zwischen Sequenzähnlichkeiten, die auf Bereichen gleicher intrinsischer Eigenschaft beruhen und tatsächlicher Homolgie von essentieller Wichtigkeit (siehe Kapitel 1.2, Kapitel 1.4 und Kapitel 3.1). Durch das Aufdecken von Homologien hat man nicht nur die Möglichkeit 3D - Strukturen zu identifizieren, sondern kann auch die Funktion eines bezüglich seiner Eigenschaften bisher unbekannten Proteins (siehe Kapitel 1.3, Kapitel 1.4, Kapitel 3.2 und Anhang) vorhersagen; dies soll in absehbarer Zeit zur Vorhersage kompletter Stoffwechselwege führen. Die bioinformatische Methodik wird in Zukunft die molekularbiologische Labortätigkeit verstärkt unterstützen, viele zeitintensive Arbeiten überflüssig machen und ihr so zu maximaler Effizienz verhelfen.

1.2. Sekundärstrukturen und intrinsische Eigenschaften von Proteinen

Bedeutsamer als die Kenntnis der Aminosäuresequenz eines Proteins ist das Wissen um die nicht selten daraus ableitbare und konserviertere Sekundärstruktur. Das Zusammenspiel von Wasserstoffbrückenbindungen, Schwefelbrücken, hydrophoben oder polaren Wechselwirkungen und Ionenbindungen faltet das Protein in alpha-Helix ,beta-Faltblatt oder eine andere charakteristische Struktur.
Diese sekundären Elemente gebendem Protein die Möglichkeit, seine ihm eigene dreidimensionale Struktur anzunehmen und somit funktionell als Enzym oder Strukturprotein zu wirken, intrinsische Eigenschaften zu erlangen.

Neben anderen sind für die Vorhersage von Funktion und Lokalisation in der Zelle drei intrinsche Eigenschaften von Bedeutung:

N-terminale Signalpeptide ermöglichen einem Protein mitunter noch während seines Translationsprozesses das Eindringen ins rauhe Endoplasmatische Reticulum, in dem die charakteristische Sequenz meist abgetrennt wird; Proteine mit diesem Signal sind membranständig oder zur Sezernation bestimmt. Die Sequenz beginnt N-terminal mit einer oder wenigen polaren oder geladenen Aminosäuren, setzt sich durch hydrophobe Aminosäuren bis zu einer Länge zwischen 20 und 35 Einheiten fort und endet eine bzw. drei Aminosäuren vor der Abspaltstelle mit kleinen ungeladenen Aminosäuren.

Abbildung nicht vorhanden

Abb. 1 Schematische Darstellung der Synthese von Sekretproteinen am Endoplasmatischen Reticulum [Abbildung aus Darnell, Molekulare Zellbiologie, 1. Auflage]

Transmembranregionen sind meist alpha-helikal und ueberwiegend aus hydrophoben Aminosäuren zusammengesetzt. Eine oder mehere Helices durchziehen die Lipiddoppelschicht über eine Länge von etwa 15 bis 35 Aminosäuren. Sie fixieren das Protein in der Membran, wo es seine Aufgabe z.B als Rezeptor oder Transporter erfüllen kann. In selteneren Fällen wie z.B bei Porinen durchdringen auch

beta-Faltblattstrukturen die Membran und bilden faßartig angeordnet Tunnel mit Transportereigenschaft.

Abbildung nicht vorhanden

Abb. 2 Beispiele für integrale alpha-helikale Membranproteine hinsichtlich ihrer Anordnung in der Membran des Endoplasmatischen Reticulums [Abbildung aus Darnell, Molekulare Zellbiologie, 1. Auflage]

Coiled Coils haben die Form gestauchter alpha-Helices, die in der Lage sind, sich paarweise oder zu mehreren umeinanderzuwinden. Man findet diese Strukturen u.a. in fibrillaeren Proteinen wie Myosinen, Tropomyosinen oder Keratinen, in denen dieser Bereich mehr als tausend Aminosäuren umspannen kann.

In globulären Proteinen erstreckt sich die Coiled Coil – Region über deutlich kürzere Bereiche; in DNA – bindenden Proteinen (Zink-Finger) ist sie kaum größer als zwanzig Aminosäuren. Die Coiled Coil – Helix besteht aus Heptad-Repeats

(Aminosäurekennzeichnung: a bis g).Sie setzt sich überwiegend aus polaren oder geladenen Aminosäuren zusammen, deren Reste aus der Helix herausragen. An den Positionen a und d befinden sich hydrophobe Aminosäuren; sie bilden den hydrophoben Helixkern, der das Zusammenlagern der Helices herbeiführen kann.

Abbildung nicht vorhanden

Abb. 3 Hier handelt es sich um die schematische Darstellung eines Myosinmoleküls bestehend aus zwei Paaren ungleicher leichter Ketten und zwei identischen schweren Ketten. Die schweren Ketten tragen an ihren N - terminalen globulären Bereichen die leichten Ketten. Der C - terminale fibrilläre Rest weist Coiled Coil – Struktur auf und erlaubt den schweren Ketten so das Sichumeinanderwinden [Abbildung aus Darnell, Molekulare Zellbiologie, 1. Auflage]

1.3. Funktionsbestimmung uncharakterisierter Proteinfamilien

In den Sequenzdatenbanken existieren viele Proteine, über deren Funktion oder Lokalisation in der Zelle die Annotation keine Auskunft gibt. Diese Proteine werden gewöhnlich als hypothetisch („hypothetical") oder einfach als Offener Leserahmen („ORF – open reading frame") bezeichnet und können im Falle signifikanter Homologie untereinander in Familien mit vermutlich gleicher oder ähnlicher Funktion zusammengefaßt werden. In der SWISSPROT-Datenbank sind 58 derartige Familien („UPFs – uncharacterized protein families") definiert (httpd://www.expasy.ch/cgi-bin/lists?upflist.txt), die mehr als 400 SWISSPROT-Sequenzen enthalten. Grundlage dieser Definition sind signifikante Sequenzähnlichkeiten, die durch Datenbanksuchen mit gapped Blast [3] erkannt wurden.

Datenbanksuchen stellen ein Standardverfahren dar, um aus Sequenzähnlichkeit einen Rückschluß auf Struktur und Funktion des Proteins ziehen zu können (siehe Kapitel 2.2.2.).

Ist die Startsequenz homolog zu einem annotierten Protein mit bekannter Funktion, so ist es in einigen Fällen möglich, die Funktion des annotierten Proteins auf das mit bisher unbekannter Funktion zu übertragen. Diese Vorgehensweise erweist sich als nicht problemlos. Da mit zunehmender Datenfülle immer mehr Suchmaschinen diese Arbeit automatisieren [11], steigt die Gefahr von Fehlannotationen [9],[12]. Ursächlich hierfür sind bereits falsch annotierte Sequenzen oder Homologien zu nicht katalytischen Teilbereichen des annotierten Proteins sein. Fragwürdig bleibt oft auch die Spezifität der Funktionsvorhersage. So kann in einem Fall eine funktionelle Tendenz bestehen, so daß die Funktion teilweise übertragbar ist, nicht aber mit der Spezifität, die der beste Treffer suggeriert. In einem anderen Fall bleibt die Multifunktionalität eines Multidomänenproteins unentdeckt und seine Funktion nur partiell erkannt, weil der erste Treffer automatisch als alleiniger Vorhersagekandidat dient. Zudem besteht die Möglichkeit, daß die zuerst gefundene Sequenz zwar eine deutliche Ähnlichkeit über die gesamte Sequenz aufzeigt, im Bereich der katalytischen Reste, der metallbindenden Aminosäuren oder der aktiven Zentren jedoch nicht übereinstimmt. Solchen Fehlannotationen vorzubeugen obliegt dem Wissenschaftler, verlangt nach seinem geübten Auge.

1.4. Intention

Im Rahmen dieser Diplomarbeit sollte ein Programm entwickelt werden, das die intrinsischen Eigenschaften eines Proteins vorhersagt. Es soll die Ausgabe der Ergebnisse von drei Programmen (Coils2, TopPred2 und SignalP) analysieren und interpretieren, eine Prognose über die Präsenz von Coiled Coils, Transmembranregionen und Signalpeptiden erstellen und die betroffenen Bereiche der Aminosäuresequenz angeben. Die Nutzung dieses Hilfsmittels ist über das World-Wide-Web möglich.

In einem weiteren Teil soll die Funktion von Proteinen, deren funktionelle Eigenschaften unbekannt sind,

über Homologiesuchen und Deutung der Ähnlichkeiten zu Proteinen mit bekannter Funktion aufgeklärt und beschrieben werden.

Es handelt sich hierbei um Proteine, die aufgrund von Sequenzähnlichkeit in 58 Familien, sogenannten UPFs (uncharacterized protein families), zusammengefaßt sind.

2. Material und Methoden

2.1. Material

2.1.1. Computer

Prozessoren und Speichergrößen der Unix – Computer, die die Grundlage zur Anfertigung der Diplomarbeit darstellen, sind nachstehend aufgelistet.

1 100 MHZ IP22 Processor

FPU: MIPS R4610 Floating Point Chip Revision: 2.0

CPU: MIPS R4600 Processor Chip Revision: 2.0

Hauptspeichergröße: 32 Mbytes

1 133 MHZ IP22 Processor

FPU: MIPS R4610 Floating Point Chip Revision: 2.0

CPU: MIPS R4600 Processor Chip Revision: 2.0

Hauptspeichergröße: 96 Mbytes

8 40 MHZ IP7 Processors

FPU: MIPS R2010A/R3010 VLSI Floating Point Chip Revision: 4.0

CPU: MIPS R2000A/R3000 Processor Chip Revision: 3.0

Hauptspeichergröße: 256 Mbytes

16 75 MHZ IP21 Processors

FPU: MIPS R8010 Floating Point Chip Revision: 2.2

CPU: MIPS R8000 Processor Chip Revision: 0.1

Hauptspeichergröße: 2048 Mbytes

4 250 MHZ IP19 Processors

FPU: MIPS R4000 Floating Point Coprocessor Revision: 0.0

CPU: MIPS R4400 Processor Chip Revision: 6.0

Hauptspeichergröße: 320 Mbytes

2.1.2. Datenbanken

EMBL

DNA - und RNA - Sequenzen werden in der EMBL – Nukleotidsequenz – Datenbank verwaltet. Sie werden der Literatur entnommen oder direkt von Wissenschaftlern und Sequenziergruppen zugesandt. Sie enthält über 210000 Einträge.

TREMBL

Die TREMBL – Datenbank enthält Aminosäuresequenzen, die aus der Translation der EMBL – Nukleotidsequenzen gewonnen werden. Übersetzungen mit internen Stopcodons werden nicht berücksichtigt.

SWISS-PROT [6]

Die Datenbank SWISS-PROT enthält fast siebzigtausend 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 und zu PDB (Protein Data Base).

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.

PROSITE / PROSITEDOC [5]

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.

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. Derzeit sind mehr als 7000 Einträge in PDB enthalten.

2.1.3. Programme

2.1.3.1. Vorhersage von Sekundärstrukturen und intrinsischen Eigenschaften

In diesem Abschnitt sind die Programme aufgeführt, die zur Vorhersage von sekundären Strukturen in Proteinen und intrinsischen Eigenschaften verwendet wurden.

Coils2 [35] – Programm zur Vorhersage von Coiled Coil Strukturen

Das Programm vergleicht eine Proteinsequenz mit Sequenzen, die bekanntermaßen in der Lage sind, parallele doppelsträngige coiled coils zu bilden. Die Datenbank enthält Sequenzen von Myosinen, Tropomyosinen und Keratinen. Aus einem Ähnlichkeitswert wird dann die Wahrscheinlichkeit der Bildung eines coiled coil berechnet.

TopPred2 [24] - Programm zur Vorhersage von Transmembranregionen

Die Vorhersage von a-helicalen Transmembranbereichen beruht auf der Analyse der Hydrophobizitdt des Proteins. Der Algorithmus beurteilt die Lage hydrophober Aminosäuren zueinander und berechnet hieraus Wahrscheinkeitswerte.

SignalP [25] – Programm zur Vorhersage von N-terminalen Signalsequenzen

SignalP ermöglicht den Vergleich einer Suchsequenz mit Datensätzen aus Signalpeptiden. 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.

PHD [42] – Programm zur Vorhersage von alpha-Helices oder beta-Faltblattstruktur

Dieses Programm sagt die Sekundärstruktur (Helixstruktur, ß-Faltblattstruktur, Knicks und Schleifen) unter Verwendung eines neuronalen Netzes voraus.

2.1.3.2. Homologiesuche

Im folgenden werden Programme beschrieben, die mittels vergleichender Sequenzanalyse im Falle von Sequenzähnlichkeiten zwischen Proteinen Rückschlüsse auf Domänenstruktur und Funktion zulassen.

Blast

Der Sequenzvergleich zwischen Suchsequenz und Datenbank beruht auf dem Blast (basic local alignment search tool) – Algorithmus [4]. 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 Ähnlichkeit 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 Ähnlichkeitsbeurteilung

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 [26] und die Gonnet – Serie [8], während die ältere Dayhoff – Pam –Matrix als weniger effektiv gilt.

Der Einsatz von Filtern wie SEG und XNU erlaubt das selektive Nicht – Berücksichtigen (maskieren) von Sequenzabschnitten, die für die Funktions – oder Verwandtschaftsvorhersage nicht relevant sind (z.B serin – oder prolinreiche Regionen). Die Suche wird so für die wesentlichen Bereiche sensibilisiert.

Gapped Blast [3]

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 [3]

Ergänzend zum herkömmlichen BLAST erstellt PSI-BLAST ein Profil (siehe Kapitel 2.2.2.) aus den für signifikant ähnlich befundenen Sequenzen und erlaubt iterative Suchen mit dem jeweiligen Profil.

MoST [47]

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 verglichen wird.

Sequenzen, die signifikant ähnlich zur Matrix sind, werden zur Bildung einer neuen Matrix verwendet, die erneut mit Datenbanksequenzen verglichen werden kann.

Swise [10]

Der zugrunde liegende Algorithmus erlaubt den Vergleich eines Profiles gegen alle sechs Leseraster einer DNA-Sequenz. Es können somit auch Sequenzen, die Verschiebungen des Rasters aufweisen, gefunden werden.

2.1.3.3. Sonstiges

Profilberechnung

PairWise [10]

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.

Aligment-Berechnung und - editierung

ClustalW [48]

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

ClustalX [49]

Die Funktion von ClustalX entspricht der von ClustalW mit zusätzlicher graphischer Benutzeroberfläche.

SeaView

Neben der Möglichkeit zur Erstellung von multiplen Sequenz-Alignments gestattet SeaView ergänzend die Editierung von Sequenzen.

Graphische Darstellung von 3D-Strukturen

Rasmol

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

Darstellung phylogenetischer Bäume

Njplot und TreeTool

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

2.2 Methoden

2.2.1. Programmierung

Die Erstellung des Programms (Intrinsic Features) zur Interpretation der Ausgabedaten der Sekundärstrukturvorhersageprogramme Coils2, TopPred2 und SignalP erfolgte in Perl [52] (Parctical Extraction and Report Language). Hierbei handelt es sich um eine interpretierte Sprache, die optimiert ist für das Durchsuchen von Textdateien, die Extraktion von Informationen aus solchen und das Zusammenfassen von gefundenen Daten.

HTML (HyperText Markup Language) dient zur Programmierung der Web-Oberfläche.

2.2.2. Konzeption der Homologiesuche

Eine Homologiesuche erfolgt in der Regel mit der Absicht, zu einer Amino – oder Nukleotidsequenz eines Proteins, dessen Funktion teilweise oder vollständig unbekannt ist, Proteine aufzuspüren, die zu der Startsequenz abschnittsweise oder über die Gesamtsequenz Homologie aufzeigen. Informationen über die homologen Sequenzen lassen sich mit Einschränkungen (siehe Kapitel 1.3.) auf die Startsequenz übertragen.

Das Ergebnis von Suchprogrammen wie z.B BLAST wird durch den Einsatz verschiedener Aminosäure – Austausch – Matrizen und Filter beeinflußt (siehe Kapitel 2.1.3.2.). Als Ausgangspunkt empfiehlt sich die Matrix BLOSUM45, die mäßig stark divergiert. Da die Divergenz das sogenannte Rauschen begünstigt, sollten bei kürzeren Sequenzen eher stringentere Matrizen wie BLOSUM62 oder Gonnet PAM120/PAM160 verwendet werden, während bei langen Sequenzen oft stark divergente Matrizen (Gonnet PAM250 bis PAM350) zu einem besseren Suchergebnis führen. Wie in Kapitel 2.1.3.2 erwähnt geben Filter dem Benutzer die Möglichkeit, Bereiche, die Ähnlichkeiten zu Proteinen aufdecken, die zur näheren Charakterisierung nicht hilfreich sind, durch X zu ersetzen, d.h zu maskieren. Sie werden bei der BLAST – Suche also nicht berücksichtigt.

Nach vollendeter Suche ist eine reziproke Suche mit gefundenen Homologen sinnvoll. Sie kann die Signifikanz der Erstsuche untermauern und ergänzend weitere entferntere Homologe zur Startsequenz aufdecken.

Anschließend erleichtert die Erstellung eines Alignments die Interpretation der Ergebnissse. Konservierte Bereiche treten hervor; Domänen, also Strukturen die für eine funktionell bedeutsame Faltung typisch sind, oder Motive, kurze Sequenzabschnitte mit hochkonservierten Einzelaminosäuren können erkennbar werden. Das Alignment bietet die Möglichkeit zur Erstellung einer positionsspezifischen Bewertungsmatrix, eines Profils. Dieses kann nun zur erneuten Homologiesuche herangezogen werden. Man unterscheidet hierbei zwischen Motivsuchen, die auf schnellen Wort-Such Algorithmen basieren und Profilsuchen, die auf der gründlicheren aber langsameren dynamic programming" – Programmiertechnik beruhen. Dieser Algorithmus erstellt ein optimales Alignment aus allen möglichen Kombinationen von Sequenzen. Die Anwendung von Profilsuchen ermöglicht ein iteratives Vorgehen; in Profilsuchen gefundene Proteine werden mit in das Profil aufgenommen, prägen es durch ihren Einfluß und gestatten so weitere Suchen mit immer neuem Profil.

3. Ergebnisse und Diskussion

3.1. Entwicklung eines Web-Tools zur Vorhersage intrinsischer Eigenschaften von Proteinen

Das Programm „Intrinsic Features" unterstützt die Vorhersage von intrinsischen Eigenschaften eines Proteins. Es ist in der Lage, die Ausgabe von drei Programmen (Coils2, TopPred2 und SignalP) auszuwerten, die ihrerseits mittels spezifischer Algorithmen (siehe Kapitel 2.1.3.) die Primärstruktur von Proteinen analysieren und Wahrscheinlichkeitsangaben über das Vorhandensein von Coiled Coils, Transmembranregionen und Signalpeptiden machen. „Intrinsic Features" verbessert die Vorhersagegüte, indem es Parameterwahl optimiert und die Ausgabe der drei Programme bestmöglich interpretiert . Das Tool wird über das world – wide – web zur Verfügung gestellt.

3.1.1. Wahl der Parameter

Das Programm SignalP berechnet über Ähnlickeitswerte die Wahrscheinkeit für die N-terminalen Aminosäuren als Bestandteil einer Signalsequenz, das Vorhandensein einer Abspaltstelle und die Kombination aus beiden. Die drei Wahrscheinlichkeiten werden separat mit einem „ja" oder „nein" kommentiert, und zusätzlich wird durch ein zusammenfassendes Urteil („ja" oder „nein") die Präsenz einer Signalsequenz bestätigt oder verworfen. Fallen drei der vier Urteile zugunsten eines Signalpeptides aus, so schließt „Intrinsic Features programm – intern ebenfalls auf das Vorhandensein eines Signalpeptides. SignalP berücksichtigt bei der Vorhersage den Organsimus, der das Protein mit der potentiellen Signalsequenz hervorbringt; durchschnittliche Länge der Signalsequenz und die Aminosäuren an der Abspaltstelle variieren interspezifisch. „Intrinsic Features" gibt die Möglichkeit zwischen gram – positiven und gram – negativen Prokaryonten und Eukaryonten auszuwählen. Erfolgt keine ausdrückliche Wahl, wird die Prognose für alle drei Gruppen durchgeführt.

Coils2 gibt Wahrscheinlichkeitswerte für jede einzelne Aminosäure der Sequenz aus.

Die Bewertung liegt zwischen null (sehr wahrscheinlich nicht Bestandteil einer Coiled Coil – Helix) bis eins (sehr wahrscheinlich Bestandteil einer Coiled Coil – Helix). Coils2 läßt drei Fenster, also Coiled Coil – Sequenzen von unterschiedlicher Größe (14 Aminosäuren, 21

Aminosäuren und 28 Aminosäuren) mit der Sequenz vergleichen. In der Ergebnisausgabe finden sich somit drei Wahrscheinlichkeitswerte für jede Aminosäure. Liegt über einen Bereich von mindestens 15 Aminosäuren der Wahrscheinlichkeitswert des Fensters 28 über 0.8, der des Fensters 21 über 0.4 und der des Fenster 14 in diesem Bereich wenigstens zeitweilig über 0.8, interpretiert „Intrinsic Features" die Region als Coiled Coil. Sind die Lücken zwischen zwei so als Coiled Coil – Region erkannten Abschnitten größer als 30 Aminosäuren, wird dieser Abschnitt auch als Coiled Coil gedeutet. Der Benutzer hat die Möglichkeit, unterschiedliche Matrizen zu wählen und die Gewichtung des Vorhandenseins der hydrophoben Aminosäuren an den Positionen a und g (hydrophober Helixkern), festzulegen.

TopPred2 beurteilt in seiner Ausgabe die Transmembranregionen mit einem Wahrscheinlichkeitswert zwischen null und drei. Ist die Anzahl der gefundenen Transmembranregionen größer als 3 oder die Proteinsequenz nicht länger als 120 Aminosäuren, behält „Intrinsic Features" die Cutoff – Werte (Grundeinstellung, Lower Cutoff: 0,6 und Upper Cutoff: 1) für wahrscheinlich und sehr sicher bei, erklärt aber nur solche zu Transmembranregionen, deren Startaminosäure jenseits der fünften liegt. Dies soll die Verwechslung mit Signalsequenzen ausschließen. Liegt die Anzahl der von TopPred2 gefundenen Transmembranregionen unter 4, und die Proteinsequenz ist länger als 120 Aminosäuren, werden nur Sequenzabschnitte, deren Wahrscheinlichkeitswert größer 2 ist, als transmembran interpretiert. „Intrinsic Features bietet Möglichkeit den „Lower Cutoff", den Wert, ab dem die Richtigkeit der Aussage wahrscheinlich ist und den „Upper Cutoff, ab dem die Richtigkeit der Aussage annähernd sicher ist, auszuwählen. Der Benutzer kann von den Standardeinstellungen abweichend die Suche sensibler oder stringenter gestalten. Mit der Fenstergröße wird über die erlaubte Länge der Transmembranregion entschieden (window size top : Minimumlänge, window size bottom : Maximumlänge).

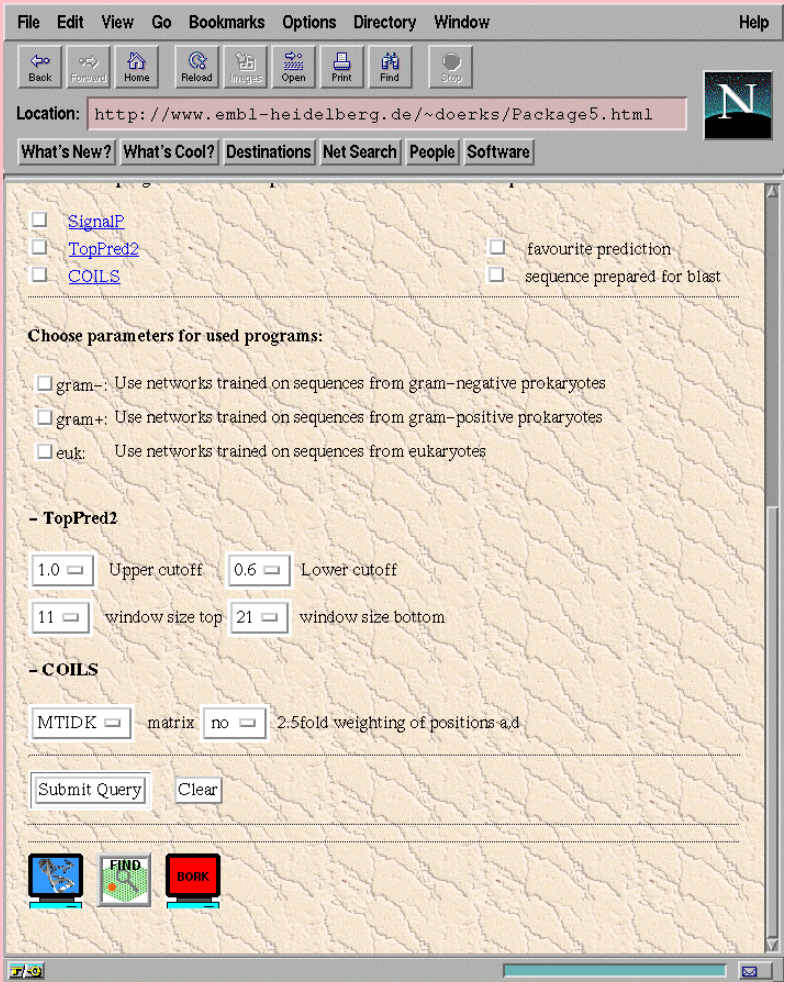

Abb. 4 Parameterauswahl

3.1.2. Darstellung des Eingabe – und Ausgabeformulars

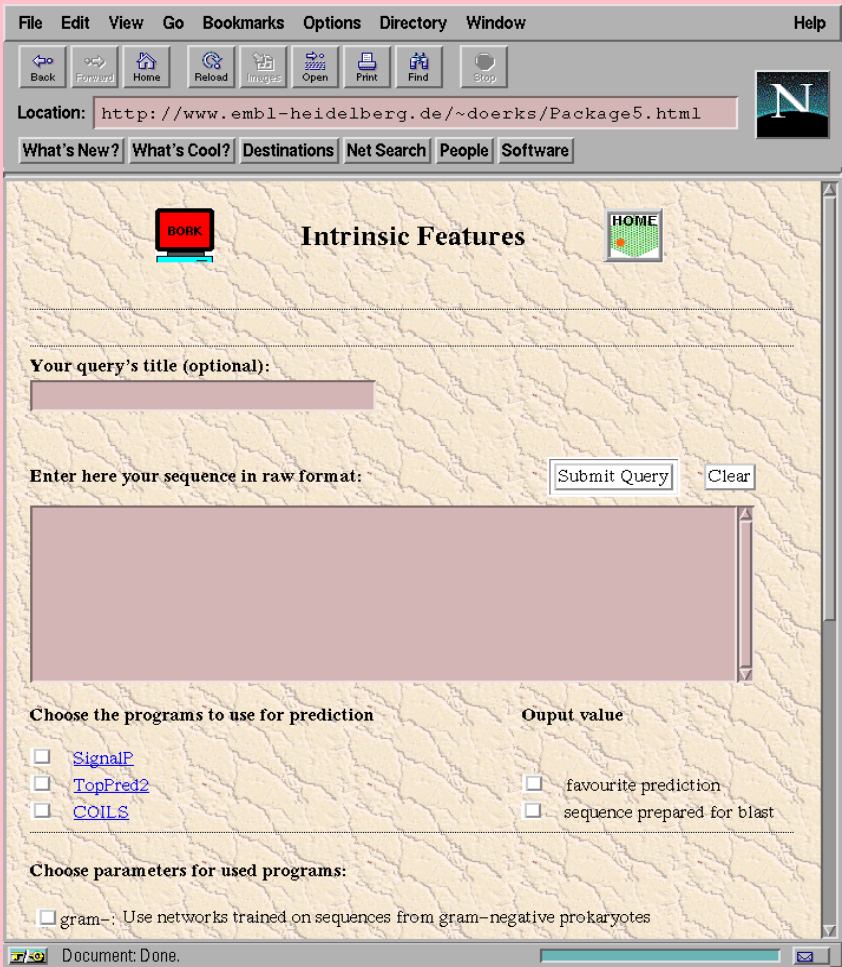

Abb.5 Web-Seite „Intrinsic Features – Eingabeformular

Abbildung 5 zeigt die Eingabeaufforderung des Programms. Der Benutzer gibt an dieser Stelle die zu analysierende Aminosäuresequenz in das Textfeld ein und wählt das gewünschte Analyseprogramm und die Ausgabeform. Es besteht die Möglichkeit, sich die Interpretation, beruhend auf der programm - internen Auswertung des jeweiligen Basisprogramms, sowie die für eine BLAST – Suche vorbereitete Sequenz ausgebenzulassen. Im zweiten Fall sind wie bei der Filterung in BLAST (siehe Kapitel 2.1.3.) die Regionen mit intrinsischen Eigenschaften maskiert (Abb. 6). Die Maskierung ermöglicht dem Benutzer sensiblere BLAST – Suchen. Signalsequenzen, Transmembranregionen und Coiled Coils weisen aufgrund ihrer Eigenschaften Sequenzähnlichkeit auf, sind aber kein Homologiekriterium und nicht hilfreich bei einer spezifischeren Funktionsvorhersage.

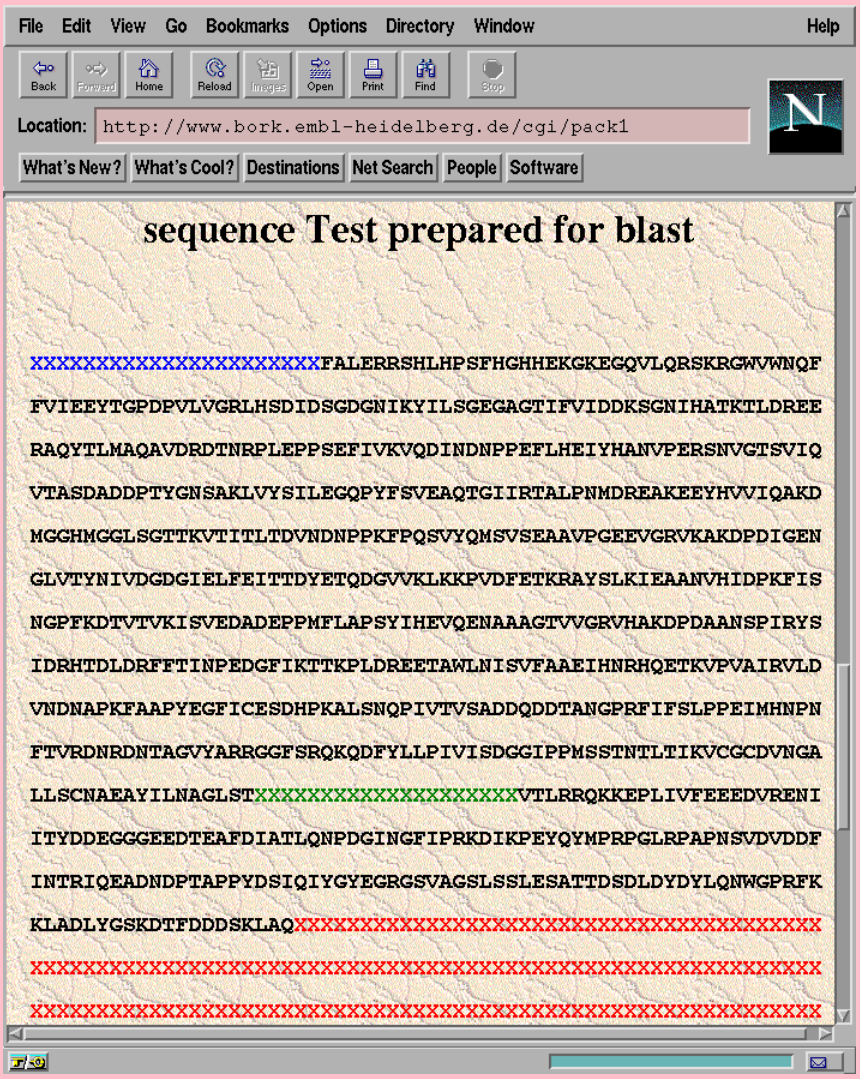

Abb. 6 Ausgabe der Sequenz – Signalsequenz (blau), Transmembranregion (grün) und Coiled Coils (rot) maskiert für BLAST - Suche

3.2. Funktionsbestimmung uncharakterisierter Proteinfamilien (UPFs)

3.2.1. Funtionsanalyse der Einzelfamilien

In den folgenden Unterkapiteln des Kapitels 3.2.1. werden die Analyseergebnisse der Familien näher beschrieben, über deren Funktion eine Vorhersage möglich ist. Familien mit Transmembranproteinen sind in den Unterkapiteln nicht explizit aufgeführt, da in keinem Fall genauere Charakterisierung möglich war. Es handelt sich hierbei um die UPFs 3, 5, 13, 14, 16, 18, 28, 32, 43, 48, 52, 56 und 58. Die Tabelle bietet eine Übersicht über die vorhergesagte Funktion der jeweiligen Familie.

Tabelle 1. Vorhergesagte Funktion für 25 UPFs

UPF –Nummer	Anzahl der Famileinmitglieder	vorhergesagte Funktion
02	70	Pseudouridylate – Synthase
04	60	Methyltransferase
07	15	Cytidyltransferase
08	30	ATPase
09	40	GTPase
10	10	Aldose 1-epimerase
11	10	Methyltransferase
12	25	Nitrilase
17	30	Hydrolase
19	15	Phosphat – bindendes Protein (TIM BARREL – Struktur)

20	40	N6-adenin-spezifische Methylase
21	50	ATPase
26	30	Zweidomänen – Protein : Eisen / Schwefel – bindend und Amidotransferase
30	10	Amidotransferase
31	30	Kinase
34	20	Pyrimidin – bindende Oxidoreduktase (TIM BARREL – Struktur)
35	20	Mutator mutt protein (7,8–Dihydro-8-Oxoguanintriphosphatase)
36	70	Hydrolase
37	10	Oxygenase
38	35	ATPase**
42	10	ATPase
46	15	Phosphatase
49	50	N6 – adenin – spezifische Methylase
53	40	CBS – Domänen – Protein
55	10	Glutaredoxin – ähnliches Protein

* Bei der Mitgliederanzahl der Familien handelt es sich um eine ungefähre Angabe, da sich die Datenbankeinträge täglich ändern und .der Begriff Familie und damit die Zugehörigkeit nicht exakt definiert sind.

3.2.1.1. UPF0002

Für UPF0002 konnte nach der ersten iterativen Suche mit PSI – BLAST die Funktion der Familie bestimmt werden. An Position 40 der Suche konnte über die gesamte Sequenz (hypothetisches Protein: YPUL_BACSU) eine signifikante Ähnlichkeit (Wahrscheinlichkeitswert : 2e-48) zu einer Pseudouridylat – Sytnhase [27] festgestellt werden

Tabelle 2. Auszug einer PSI-BLAST – Ergebnisausgabe für UPF0002 nach der 1.Iteration

Position	Annotation	Wahrscheinlichkeit
1	Gnl|PID|e332795 (Z98268) hypothetical protein MTCI125.33 [Mycobacterium tuberculosis] . .	(2e-75)
4 5	Sp|P33643|SFHB_ECOLI SFHB PROTEIN Gnl|PID|e1185138 (Z99112) alternate gene name: ylmL; similar to hypothetical proteins [Bacillus subtilis] . .	(1e-67) (3e-65)
		(7e-50)
		(2e-48)
		(7e-48)

37	Sp|Q12362|RIB2_YEAST DRAP DEAMINASE >gi|1078332|pir||S50972 RIB2 protein – yeast (Saccharomyces cerevisiae) >gi|642221 (Z21618) DRAP deaminase [Saccharomyces cerevisiae] >gi|1419887|gnl|PID|e252279 (Z74808) ORF YOL066c [Saccharomyces cerevisiae] . .
40	Sp|P33918|RSUA_ECOLI 16S PSEUDOURIDYLATE 516 SYNTHASE (16S PSEUDOURIDINE 516 SYNTHASE) (URACIL HYDROLYASE)
41	Sp|Q47417|YQCB_ERWCA EXOENZYME REGULATION REGULON ORF1 >gi|628643|pir||S45107 hypothetical protein 1 – Erwinia carotovora >gi|496598 (X79474) ORF1 [Erwinia carotovora] . .

Die Ergebnisausgabe zeigt anschaulich die Problematik bei der Interpretation von Homolgiesuchen. So findet sich bereits an Position 4 eine homologe Sequenz zu einem sogenannten SFHB – Protein.SFHB ist der Name eines Gens, dessen Produkt Supressoreigenschfaft bezüglich einer Temperatursensitivitätsmutation des Genes ftsh1 hat. Dieses Wissen bedeutet keinen Informationsgewinn über die tatsächliche Funktion des Proteins.

Aus Position 37 wird Homologie zu einer Deaminase deutlich. Die Sequenzähnlichkeit zu dem hypothetischen Proteinen der UPF erstreckt sich aber nicht über den katalytischen Bereich. Die Deaminase ist somit zur Funktionsableitung ungeeignet. Irreführend ist auch die Homolgie zu dem Protein auf Position 41; seine Benennung beschreibt die Lage auf dem Bakterienchromosom, beinhaltet keine Information über die Funktion.

3.2.1.2. UPF0004

Nach der ersten PSI – BLAST – Suche fand sich zu den Mitgliedern der UPF0004 ein über die gesamte Sequenz homologes Protein mit bekannter Funktion. Das hypothetische Protein (Y865_METJA) ist zu 24 % identisch und zu 39% ähnlich einer Methyltransferase (Phosphonoacetaldehyd Methylase (gi|1061002|)). Die Ähnlichkeit von Aminosäuren wird nach Größe, Ladung, Hydrophobizität und dem Umstand, ob die Reste aromatisch oder aliphatisch sind, beurteilt.

3.2.1.3. UPF0007

Noch vor der ersten iterativen Suche zeigte ein PSI – BLAST – Ergebnis, daß die Mitglieder der UPF0007 signifikante Ähnlichkeit zu einer Cytidyltransferase haben [23]. Das nachstehende Baumdiagramm verdeutlicht die ausgeprägtere Verwandtschaftsnähe zu Cytidyltransferasen als zu anderen Nukleotidyltransferasen.

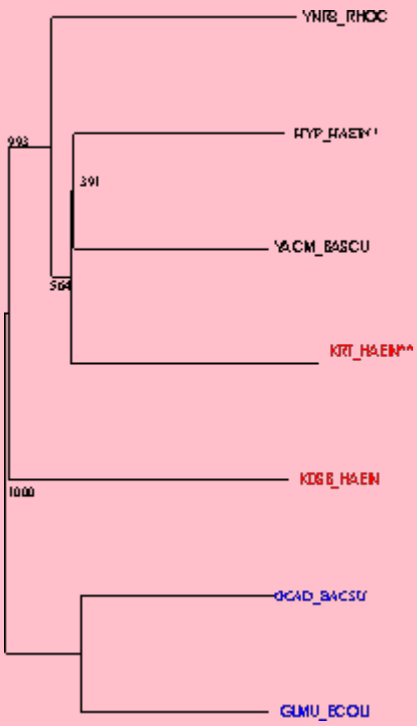

Abbildung 7 Phylgenetischer Baum aus UPF0007 – Mitgliedern (schwarz) Cytidyltransferasen (rot) und Uridilyltransferasen (blau). Der Baum wurde mit CLUSTALX berechnet. Die Werte neben den Ästen sind Indikatoren für die phylogenetische Distanz (Werte größer als 800 bedeuten signifikante Aufspaltung der Äste)

*pir - Datenbankeintrag, pir|g64156

**pir - Datenbankeintrag, pir|s49238.

3.2.1.4. UPF0008

Aus der zweiten iterativen Suche mit PSI – BLAST ging hervor, daß sich bei den hypothetischen Proteinen der UPF0008 um ATPasen der PP – Familie handelt. Sie katalysieren die Reaktion von ATP zu AMP und PP_i . Das charakteristische Motiv [14] mit dem hochkonservierten Glycin ist in allen Fällen vorhanden.

```
Query 190 KVLCLISDGIDSPVAAFMM 209
          KVL ++S G+DS V A ++
Sbjct 236 KVLVMVSGGVDSAVCAALL 255
```

Abbildung 8 Paarweises Alignment zwischen hypothetischen UPF0008 - Protein (sp|Q58341|Y931_METJA) (Query) und einer GMP SYNTHASE (sp|Q09580|GUAA_CAEEL PROBABLE) (Sbjct) im Bereich der ATP – Bindungsstelle (kursiv).

3.2.1.5. UPF0009

Nach der ersten PSI – BLAST – Suche fand sich zu den Mitgliedern der UPF0009 ein homologes Protein mit GTPase – Funktion [1]. Die typischen GTP – Bindungsmotive sind konserviert.

```
Query 27 VIVVGRSNVGKSTFVR 73 VDMPGFGY
            +GRSNVGKS+        VD+PG+G+
Sbjct 23 ICFMGRSNVGKSSLIN 68 VDLPGYGF
```

Abbildung 9 Paarweises Alignment zwischen hypothetischen UPF0009 - Protein (Y320_METJA) und GTP – Bindungsprotein (Y335_MYCGE) in den Bereichen der GTP – Bindungsstellen (kursiv).

3.2.1.6. UPF0010

Nach der dritten iterativen PSI – BLAST – Suche wurde signifikante Homologie der UPF0010 – Mitglieder zu Aldose 1 –Epimerasen [19] deutlich.

Das hypothetische Protein LAXP_LACLA ist zu 20 % identisch und zu 34% ähnlich einer ALDOSE 1-EPIMERASE (GALM_HAEIN) und die für die enzymatische Aktivität bedeutsamen drei Histidine sind in allen Fällen konserviert.

3.2.1.7. UPF0011

Eine PSI – BLAST – Suche führte nach der ersten Iteration zu einer signifikanten Ähnlichkeit der UPF0011 – Proteine über die gesamte Sequenz zu einer Methyltransferase (Uroporphyrin - II C – Methyltransferase) [20].

Die Aminosäuren der Sequenzen des hypothetischen Proteins Y056_MYCPN und der Uroporphyrin - II C – Methyltransferase HEM4_CLOJO sind zu 15% identisch und zu 39% ähnlich.

Eine spezifischere Funktionsvorhersage als die Methylaseaktivität war nicht möglich.

3.2.1.8. UPF0012

Nach der ersten PSI – BLAST – Suche fand sich zu den Mitgliedern der UPF0012 ein über die gesamte Sequenz homologes Protein mit bekannter Funktion. Das hypothetische Protein ist zu 27 % identisch und bezüglich der Eigenschaften von Aminosäuren zu 42% ähnlich einer Nitrilase (NRL4_ARATH

NITRILASE) [39].

Zu dieser Nitrilase – Familie existiert in der Prosite – Datenbank ein Eintrag. Ein konserviertes Motiv in der UPF0012 – Familie weicht nur geringfügig von dem eingetragenen Nitrilase – Motiv ab.

3.2.1.9. UPF0017

Nach der ersten iterativen Suche mit PSI – BLAST wird Homologie zwischen den hypothetischen Proteinen der UPF0017 – Familie und Hydrolasen mit der charakteristischen a / b - Hydrolase – Faltung [41] (SCOP [38]) deutlich. Eine spezifischere funktionelle Zuordnung innerhalb dieser Superfamilie ist nicht möglich.

Die erstgefundene Hydrolase ist eine Peroxidase (gnl|PID|d1011335) mit 13% Sequenzübereinstimmung und 29% Ähnlichkeit im Vergleich mit dem hypothetischen Protein YYC5_CAEEL. Das folgende Baumdiagramm berechnet aus einem Alignment verschiedener Hydrolasesequenzen und den hypothetischen Proteinen zeigt, daß die UPF0017 - Familie keiner Hydrolasefamilie signifikant näher steht als der anderen.

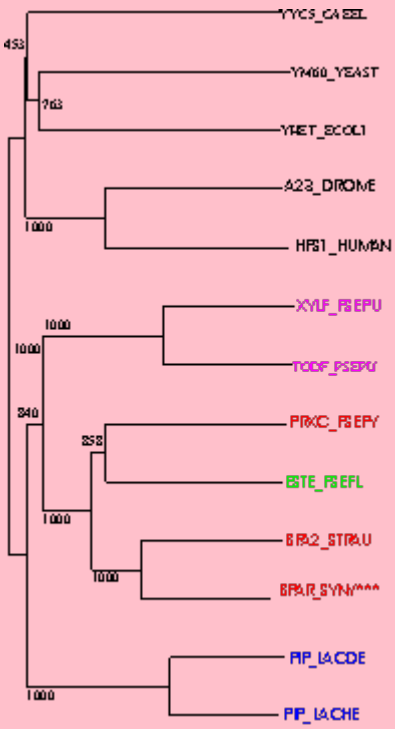

Abbildung 10 Phylogenetischer Baum der UPF017 – Mitglieder (schwarz), Peroxidasen (rot), Esterasen (grün), Peptidasen (blau) und andere Hydrolasen (pink). Der Baum wurde mit CLUSTALX berechnet.

*** genbank - Eintrag gi|1001804

3.2.1.10. UPF0019

Die erste iterative Suche mit PSI – BLAST führte zu dem Auffinden einer Homologie zwischen den Mitgliedern der UPF0019 – Familie und den C – terminalen 100 Aminosäuren einer Indol – 3 – Glycerol – Phosphatsynthase (TRPC_PHYPR). In diesem Bereich sind 22% der Aminosäuren identisch und 45% ähnlich zu dem hypothetischen Protein H47_STELP. Die katalytischen Reste sind nicht übereinstimmend vorhanden, aber die Sequenzähnlichkeit spricht für charakteristische TIM (Triose – Phosphat – Isomerase) - Barrelstruktur eines phosphatbindenden Proteins [38, 41].

3.2.1.11. UPF0020

Nach der ersten iterativen PSI – BLAST – Suche fand sich zu den Mitgliedern der UPF0020 – Familie ein über die gesamte Sequenz homologes Protein mit bekannter Funktion. Das hypothetische Protein YPSC_BACSU ist zu 17 % identisch und bezüglich der Eigenschaften von Aminosäuren zu 35% ähnlich einer

N6 – adeninspezifischen Methylase (MTV1_VIBS3) [39].

Zu dieser Methylase – Familie existiert in der Prosite – Datenbank ein Eintrag. Ein konserviertes Motiv in der UPF0020 – Familie weicht von dem eingetragenen

N6 – adeninspezifischen Methylase – Motiv nur geringfügig ab.

3.2.1.12. UPF0021

Aus der ersten iterativen Suche mit PSI – BLAST ging hervor, daß sich bei den hypothetischen Proteinen der UPF0021 um ATPasen der PP – Familie handelt. Das charakteristische Motiv [14], das schon in der UPF0008 – Familie auftaucht, ist auch hier in allen Fällen konserviert.

3.2.1.13. UPF0026

Eine Suche mit PSI – BLAST und eine ergänzende Suche mit dem Programm Most führten zu dem Ergebnis, daß die Mitglieder der UPF0026 – Familie Proteine mit zwei Domänen sind. Der N –terminale Bereich trägt das charakteristische Eisen / Schwefel – bindungsmotiv eines Molybdän – Cofaktor – Biosyntheseprotein [X], und die c –terminale Region (über 200 Aminosäuren) hat signifikante Ähnlichkeit mit einer Amidoptransferase [X]. Das hypothetische Protein Y117_HELPY ist zu 25% identisch und zu 45% ähnlich der Amidotransferasedomäne einer Glutamatsynthase (GLSF_ANTSP).

```
Query 26 LSPSKKQCNYNCIYCELGK
                +CN   C YC +
Sbjct 44 RLSLTDKCNLRCTYCMPAE
```

Abbildung 11 Paarweises Alignment zwischen dem hypothetischen UPF0026 - Protein (Y117_HELPY) und dem Molybdän – Cofaktor – Biosyntheseprotein (pir||S57490) in dem Bereich der Eisen / Schwefel – Bindungsstelle (drei Cysteine).

3.2.1.14. UPF0030

Bereits die erste PSI – BLAST – Suche offenbarte eine signifikante Sequenzhomologie zwischen den Mitgliedern der UPF0030 – Familie und der

hisH Amidotransferase – Familie [51]. Das hypothetische Protein (YMY5_YEAST) erweist sich als zu 24% identisch und zu 40% ähnlich im Vergleich mit der

Glutamin – Amidotransferase hisH (HIS5_CYAPA). In allen Fällen sind die katalytischen Reste konserviert.

3.2.1.15. UPF0031

Zwei iterative PSI – BLAST – Suchen waren notwendig, um die Proteine als Kinasen zu identifizieren. Das hypothetische Protein (YKP1_YEAST) ist signifikant homolog (17% Identität, 29% Ähnlichkeit) zu einer Phosphomethylpyrimidin – Kinase [45] und zu weiteren Kinasen.

3.2.1.16. UPF0034

Nach der ersten iterativen Suche mit PSI – BLAST wird Homologie zwischen den hypothetischen Proteinen der UPF0034 – Familie und Pyrimidin - bindenden Oxidoreduktasen [38,54] (TIM - Barrel – Struktur) deutlich. Das hypothetische Protein () erweist sich als zu 19% identisch und zu 30% ähnlich im Vergleich mit der Dihydropyrimidin – Dehydrogenase (DPYD_HUMAN). Die Pyrimidin – Bindungsstelle ist konserviert.

3.2.1.17.UPF0035

Eine Homologiesuche unter Verwendung der WiseTools ließ signifikante Ähnlichkeit der UPF0035 Proteine mit Mutator mutt proteinen (7,8–Dihydro–8-Oxoguanintriphosphatasen erkennen. Das hypothetische Protein (YEAB_ECOLI) ist signifikant homolog (26% Identität, 42% Ähnlichkeit) zu einem mutt protein (gi|1500003) [2].

3.2.1.18.UPF0036

Nach der ersten iterativen Suche mit PSI – BLAST wird Homologie zwischen den hypothetischen

Proteinen der UPF0036 – Familie und Hydrolasen mit der charakteristischen a / b - Hydrolase – Faltung [41] (SCOP [38]) deutlich. Eine spezifischere funktionelle Zuordnung innerhalb dieser Superfamilie ist nicht möglich, wie der das nachfolgende Baumdiagramm verdeutlicht.

Die erstgefundene Hydrolase ist eine Hydroxyacylgluthation - Hydrolase (GLO2_HAEIN) mit 27% Sequenzübereinstimmung und 41% Sequenzähnlichkeit zu dem hypothetischen Protein (Y139_MYCPN).

3.2.1.19.UPF0037

Eine PSI – BLAST – Suche führte zu dem Auffinden einer signifikanten Ähnlichkeit der UPF0037 – Proteine über die gesamte Sequenz zu Oxidasen, Oxigenasen und anderen Oxidoreduktasen [36].

Die Aminosäuren der Sequenzen der Oxidase (GOX_RAT) und dem hypothetischen Protein (MJ0862) sind zu 23% identisch und zu 37% ähnlich.

3.2.1.20.UPF0038

Aus der ersten noch nicht iterativen Suche mit PSI – BLAST ging hervor, daß sich bei den hypothetischen Proteinen der UPF0038 ebenfalls um ATPasen der PP – Familie handelt. Das charakteristische Motiv [14], das schon in der UPF0008 – und UPF0021 – Familie auftaucht, ist auch hier in allen Fällen konserviert.

3.2.1.21. UPF0042

Nach der ersten PSI – BLAST – Suche fand sich zu den Mitgliedern der UPF0042 ein homologes Protein mit ATPase – Funktion [14]. Das typische ATP – Bindungsmotiv, daß schon in den anderen ATPasen der UPFs auftaucht, ist auch hier konserviert.

3.2.1.22. UPF0046

Nach der fünften iterativen Suche mit PSI – BLAST wurde zu den Mitgliedern der UPF0046 – Familie ein über die gesamte Sequenz homologes Protein mit Proteasefunktion . Das hypothetische Protein C25E10.12 gene product (gi|1226307) ist zu 13% identisch und zu 25% ähnlich mit einer Protease (THER_BACST) [54].

Die katalytischen Reste sind allerdings in keinem Fall konserviert. Dieser Umstand führt zu dem Schluß, daß es sich zwar um Enzyme mit ähnlicher dreidimensionaler Struktur aber dennoch unterschiedlicher Funktion handelt. Tatsächlich erscheint nach der neunten iterativen PSI – BLAST – Suche ein homologes Protein mit Phosphatasefunktion. Das hypothetische Protein ist zu 14% identisch und zu 29% ähnlich einer Phosphatase (gi|2464946|) [31]. Das nachstehende Alignment verdeutlicht die Präsenz der metallbindenden und für die katalytische Aktivität relevanten Aminosäuren. Ursächlich für das ungewöhnlich späte Auftauchen der Phophatase ist die hohe Mitgliederzahl der Proteasefamilie. Die Proteasen beeinflussen entsprechend ihrer Anzahl stark das Profil, so daß mit jeder weiteren Iteration, immer mehr Proteasen in der Ausgabe erscheinen.

```
                          * *                          *

YBPT_CAEEL  DTPIFENKVRFVCISDTHEKLHEILP------YIPDGDVLIHSGDFTNCG-DIGEVIKFN
239E_HUMAN  DTPKPAGHTRFVCISDTHSRTDGIQ--------MPYGDILLHTGDFTELG-LPSEVKKFN
YW12_CAEEL  DTPVKPDHVRFVCIGCTHGEQFDIS-------KLPPGDVLLVAGDFTSCG-LPNEVHNFN
PPAF_PHAVU  DVPYT-----FGLIGDLGQSFDSNTTLSHYELSPKKGQTVLFVGDLSYADRYPNHDNVRW
PPAF_ARATH  DVPYT-----FGLIGDLGQTYDSNSTLSHYEMNPGKGQAVLFVGDLSYADRYPNHDNNRW
PPB_LYSEN   DICDTSGNA-CQGTSDLIVSINPT------------AVFTAGDNAYNSGTLSEYNSRY


                                 *

YBPT_CAEEL  A-EIGSLPHKHK----IVIAGNHELGFE----------DGEEMSERQ-LAGLNMLG---
239E_HUMAN  D-WLGNLPYEYK----IVIAGNHELTFDKEFM-----ADLVKQDYYR-FPSVSKLKPED
YW12_CAEEL  K-LLGKLKYSYK----VVIGGNHECTFDDTFLKLKQESEPKEMALKQALLSAIHSDSKG
PPAF_PHAVU  D-TWGRFTERSVAYQPWIWTAGNHEIEFAPEIN----ETEPFKPFSYR-YHVPYEASQST
PPAF_ARATH  D-TWGRFVERSVAYQPWIWTAGNHEIDFVPDIG----EIEPFKPFMNR-YHTPHKASGSI
PPB_LYSEN  APTWGRFKALTSP------SPGNHDYSTTGAKG----YFDYFNGSGNQ--TGPAGDR---



YBPT_CAEEL  INKAYELLSNCTYLCDKS-YEAYGLKIYGAPWH SMP-GYSFFRQRGQKILHKWNQIPAKI
239E_HUMAN  FDNVQSLLTNSIYLQDSE-VTVKGFRIYGAPWT PWFNGWGFNLPRGQSLLDKWNLIPEGI
YW12_CAEEL  GISAKDLLSNAIYLEDNA-TKSRQLGIQSLSWT TTI---------GQMEPD----PGGV
PPAF_PHAVU  SPFWYSIKRASAHIIVLSSHIAYGRGTPQYTWL KKE---------LRKVKR---SETPW
PPAF_ARATH  SPLWYSIKRASAYIIVMSCYSSYGIYTPQYKWL EKE---------LQGVNR---TETPW
PPB_LYSEN    SKGYYSWDVGDWHFVSLNTMSGGTVAQAQIDWL KAD---------LAANTK------PC
```

Abbildung 12 Multiples Alignment von Mitgliedern der UPF046 – Familie (kursiv) und Phosphatasen (PPAF_ARATH, PPAF_PHAVU, PPB_LYSEN). Hochkonservierte Aminosäuren sind blau und fett, metallbindende Aminosäuren sind rot, fett und mit Sternen markiert.

3.2.1.23. UPF0049

Nach der zweiten iterativen PSI – BLAST – Suche fand sich zu den Mitgliedern der UPF0049 – Familie ein über die gesamte Sequenz homologes Protein mit bekannter Funktion. Das hypothetische Protein YS02_CAEEL ist zu 16 % identisch und bezüglich der Eigenschaften von Aminosäuren zu 27% ähnlich einer

N6 – adeninspezifischen Methylase (MTV1_VIBS3) [39].

Zu dieser Methylase – Familie existiert in der Prosite – Datenbank ein Eintrag. Ein konserviertes Motiv in der UPF0049 – Familie weicht von dem eingetragenen

N6 – adeninspezifischen Methylase – Motiv nur geringfügig ab.

3.2.1.24. UPF0053

Eine PSI – BLAST – Suche führte zu dem Auffinden einer signifikanten Ähnlichkeit der UPF0053 – Proteine zu CBS – Domänen - Proteinen [40].

Die Mitglieder dieser Familie setzen sich ausschließlich aus zwei dieser Domänen zusammen.

3.2.1.25. UPF0055

Nach der ersten noch nicht iterativen Suche mit PSI – BLAST wird Homologie zwischen den hypothetischen Proteinen der UPF0055 – Familie und Glutaredoxinen deutlich. Das hypothetische Protein (YDHD_ECOLI) erweist sich als zu 25% identisch und zu 50% ähnlich im Vergleich mit einem

Glutaredoxin (gi|2708324) [28]. Die für den Elektronentransport verantwortliche Region ist konserviert; das zweite an der Schwefelbrücke beteiligte Cystein ist in allen Fällen durch ein Serin ersetzt.

3.2.2. Zusammenfassung - graphische Darstellung der Ergebnisse

Von den 58 uncharakterisierten Proteinfamilien ist in 34,5% der Fälle keine Funktionsvorhersage möglich, 22,4% sind Transmembranproteine, deren Funktion aus der Homologie zu Proteinen mit bekannter Funktion nicht ableitbar ist, und in 43,1% der Fälle können funktionelle Eigenschaften der Familie genauer vorhergesagt werden.

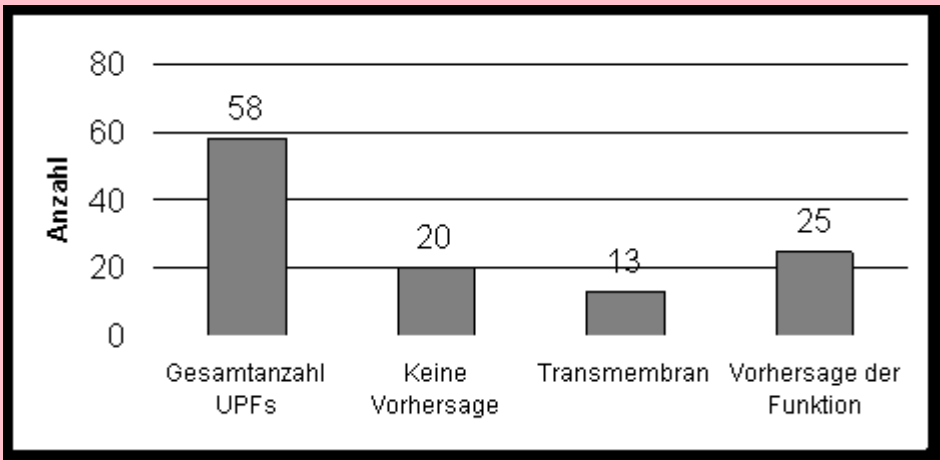

Abbildung 13 graphische Darstellung der Möglichkeiten zur Charakterisierung von UPFs

Die Güte der Funktionsvorhersage ist unterschiedlich; in allen Fällen kann enzymatische Aktivität angenommen werden. Bei 44% der funktionell charakterisierten Familien kann garnicht oder nur geringfügig auf mehr als die enzymatische Hauptklasse**** geschlossen werden, oder es ist das Vorhandensein einer Domäne bekannt, die in unterschiedlichen Proteinfamilien auftaucht Bei den Transmembranproteinen handelt es sich in allen Fällen um mehrfach transmembrangängige Proteine. Eine genauere Vorhersage ist nicht möglich, da die hydrophoben Helices den überwiegenden Teil des Proteins einnehmen und bei Ähnlichkeitssuchen beliebige Transmembranproteine auffinden lassen, die keinen Homologieschluß zulassen. (CBS – Domänen – Protein). 20% sind ATPasen oder GTPasen, und bei 36% kann aus der Homologie zu Sequenzen mit bekannter Funktion die Familienzugehörigkeit abgeleitet werden.

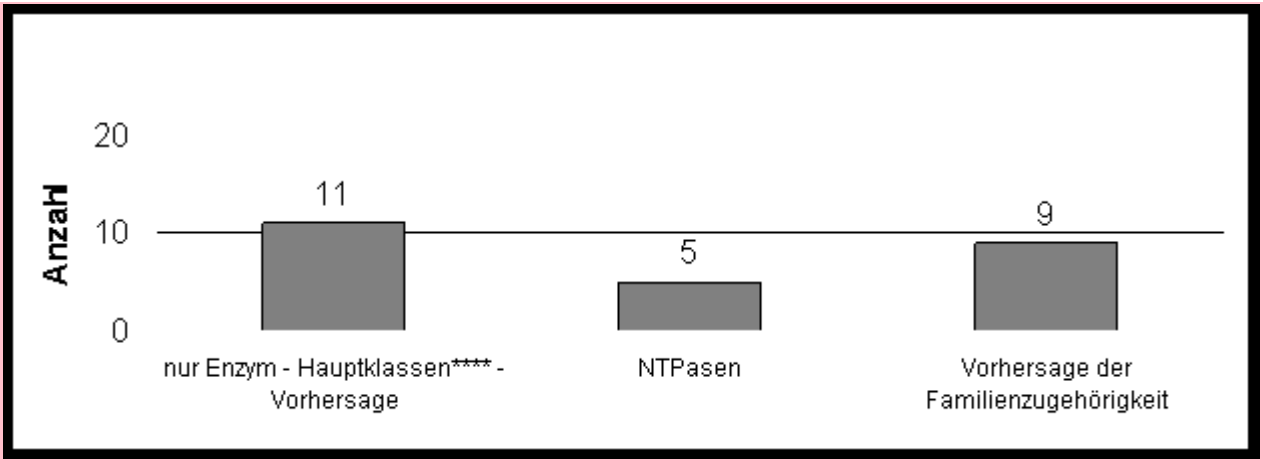

Abbildung 14 graphische Darstellung der Vorhersagespezifität, der Balken "nur Enzym – Hauptklassen – Vorhersage" beinhaltet auch das

CBS – Domänen - Protein

**** die sechs Hauptklassen der Enzyme: Oxidoreduktasen, Transferasen, Hydrolasen, Lyasen, Isomerasen und Ligasen

Die Tatsache, daß häufig nur oder kaum spezifischer als die enzymatische Hauptklasse vorhergesagt werden kann, veranschaulicht die Schwierigkeit bei der Funktionsvorhersage. Der Vergleich der hypothetischen Familie mit verschiedenen Familien der enzymatischen Hauptklassen läßt keine spezifischere Zuordnung zu.

Bei den NTPasen ist jeweils nur das charakteristische Bindungsmotiv konserviert, sodaß nähergehende Funktionsbestimmung nicht möglich ist. Die Familien UPF0026, UPF0046 und UPF0053 verdeutlichen im besonderen die Problematik bei der Funktionsbestimmung von Proteinen. Die Mitglieder der UPF0053 bestehen aus zwei CBS – Domänen. Proteine, die diese Domäne tragen sind in unterschiedliche enzymatische Prozesse involviert. Die Charakterisierung der Funktion der UPF0046 war erst nach der neunten Iteration möglich. Auffällige Sequenzähnlichkeit zu Proteasen bei früheren Iterationen hätte zu einer Fehlinterpretation führen können. Die Mitglieder der UPF0026 sind Beispiele für Zweidomänenproteine und vermitteln einen Eindruck von der notwendigen Sorgfalt bei der Interpretation von Homologiesuchen.

Abbildung 15 weist auf die Methode hin, die zum Auffinden des homologen Proteins mit bekannter Funktion nötig war. Fast jede zweite Familie konnte noch vor der ersten iterativen Suche mit PSI – BLAST funktionell charakterisiert werden. In einem drittel der Fälle war eine bei drei Familien eine zweite iterative Suche notwendig. Eine UPF konnte erst nach der neuten Iteration identifiziert werden. Die Charakterisierung von zwei Familien erfolgte mit Swise und Most.

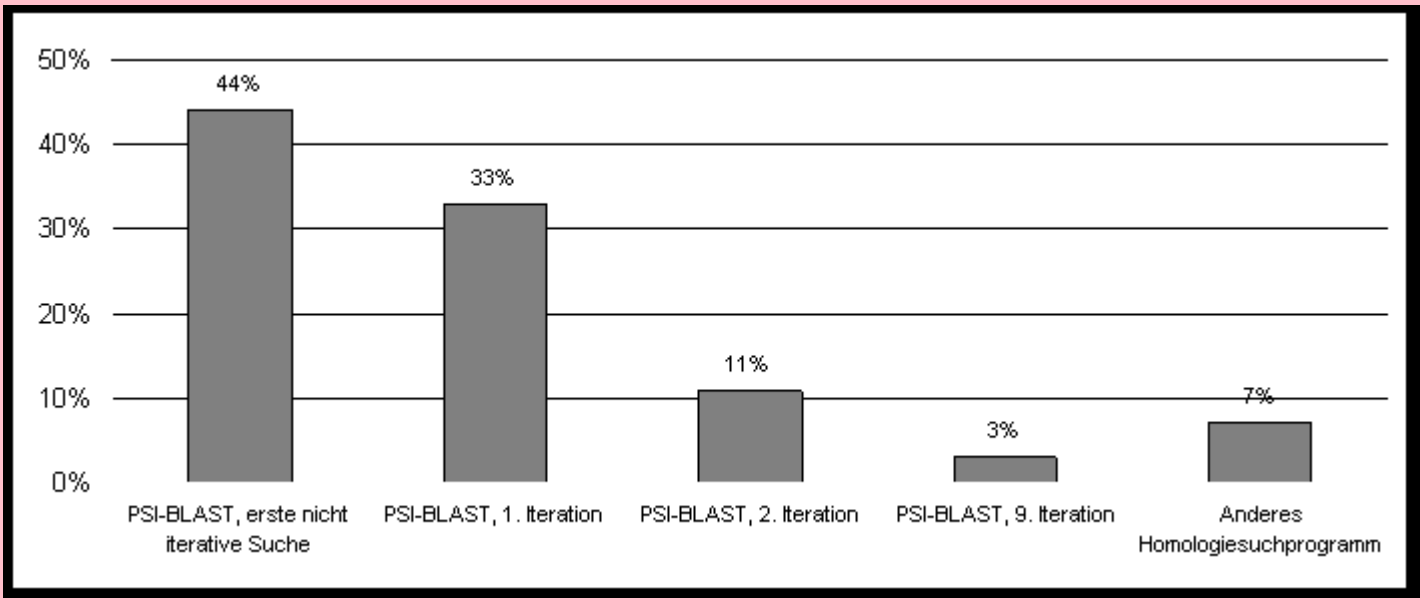

Abbildung 15 graphische Darstellung des Verfahrens, das zur funktionellen Charakterisierung der UPFs führte

4.0 Literaturverzeichnis

1. J. Ahnn, P. E. March, H. E. Takiff, M. Inouye (1986)

   A GTP-binding protein of Escherichia coli has homology to yeast RAS proteins.

   Proc Natl Acad Sci U S A 83(23):8849-8853

2. M. Akiyama, T. Horiuchi, .M. Sekiguchi (1987)

   Molecular cloning and nucleotide sequence of the mutT mutator of Escherichia coli that causes A:T to C:G transversion.

   Mol Gen Genet 206(1):9-16

3. S. F. Altschul, et al (1997)

   Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.

   Nucleic Acids Res. 25(17): 3389-3402 Review.

4. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. (1990)

   Basic local alignment search tool.

   J. Mol. Biol., 215:403—410

5. A. Bairoch, et al. (1997)

   The PROSITE database, its status in 1997.

   Nucleic Acids Res. 25(1): 217-221

6. A. Bairoch, et al. (1997)

   The SWISS-PROT protein sequence data bank and its supplement TREMBL.

   Nucleic Acids Res. 25(1): 31-36.

7. A. Bateman (1997)

The structure of a domain common to archaebacteria and the homocystinuria disease protein.

Trends Biochem Sci 22(1):12-13

8. S. A. Benner ans M. A. Cohen and G. H. Gonnet. (1994)

Amino acid substitution during functionally constrained divergent evolution of protein sequences.

Protein Eng, 7:1323—1332

9. U. Bhatia, K. Robinson & W. Gibert (1997)

Dealing with database explosion: a cautionary note.

Science 276, 1724-1725

10. E. Birney, et al. (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(14): 2730-2739

11. M. Boguski, et al. (1994)

I think therefore I publish.

Trends Biochem Sci. (2): 71

12. P. Bork, et al. (1996)

Go hunting in sequence databases but watch out for the traps.

Trends Genet. (10): 425-42.

13. P. Bork, et al. (1996)

Applying motif and profile searches.

Methods Enzymol.; 266: 162-184.

14. P. Bork and E. V. Koonin (1994)

A P- Loop-Like Motif in a Widespread ATP Pyrophosphatase Domain: Implications

for the Evolution of Sequence Motifs and Enzyme Activity

Proteins 20: 347-355

15. C. Boursaux-Eude, D. Margarita, A. M. Gilles, O. Barzu, I. Saint Girons (1997)

Borrelia burgdorferi uridine kinase: an enzyme of the pyrimidine salvage pathway

for endogenous use of nucleotides.

FEMS Microbiol Lett 151(2):257-261

16. C. J. Bult, O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton,

J. A. Blake, L. M. FitzGerald, and et al. (1996)

Complete genome sequence of the methanogic archaeon,

Methanococcus jannaschii.

Science, 273:1058—73

17. R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R.

Kerlavage, C. J. Bult, J.~F. Tomb, B. A. Dougherty, J. M. Merrick, and et al. (1995)

Whole-genome random sequencing and assembly of Haemophilus influenzae

Science, 269:496—512

18. C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. (1995)

Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, and et al.
The minimal gene complement of Mycoplasma genitalium.
Science, 270:397—403

19. P. A. Frey (1996)
The Leloir pathway: a mechanistic imperative for three enzymes to change the stereochemical configuration of a single carbon in galactose
FASEB 10(4):461-470

20. E. Fujino, T. Fujino, S. Karita, K. Sakka, K. Ohmiya (1995)
Cloning and sequencing of some genes responsible for porphyrin biosynthesis from the anaerobic bacterium Clostridium josui.
J Bacteriol 77(17):5169-5175

21. F. K. Gleason., A. Holmgren (1988)
Reduction of mutant phage T4 glutaredoxins by Escherichia coli thioredoxin reductase
FEMS Microbiol. Rev. 54:271-298

22. M. Gribskov, A. D. McLachlan, and D. Eisenberg. (1987)
Profile analysis: detection of distantly related proteins.
Proc Natl Acad Sci, 84:4355—8

23. R. C. Goldman, T. J. Bolling, W. E. Kohlbrenner, Y. Kim, J. L. Fox (***1986)***
Primary structure of CTP:CMP-3-deoxy-D-manno-octulosonate cytidylyltransferase (CMP-KDO synthetase) from Escherichia coli.
J Biol Chem 261(34):15831-15835

24. Gunnar von Heijne (1992)

Membrane Protein Structure Prediction, Hydrophobicity Analysis and the

Positive-inside Rule"

J. Mol. Biol. 225, 487-494

25 Henrik Nielsen, Jacob Engelbrecht, Søren Brunak and Gunnar von Heijne: (1997)

Identification of prokaryotic and eukaryotic signal peptides and prediction of their

cleavage sites.

Protein Engineering 10, 1-6

26. S. Henikoff and J. G. Henikoff. (1992)

Amino acid substitution matrices from protein blocks.

Proc Natl Acad Sci U S A, 89:10915—9

27. T. Hidaka, M. Goda, T. Kuzuyama, N. Takei, M. Hidaka, H. Seto (1995)

Cloning and nucleotide sequence of fosfomycin biosynthetic genes of Streptomyces

wedmorensis.

Mol Gen Genet 249(3):274-280

28. A. Holmgren (1988)

Thioredoxin and glutaredoxin: small multi-functional redox proteins with active-site

disulphide bonds.

Biochem. Soc. Trans. 16:95-96

29. T. Kaneko, A. Tanaka, S. Sato, H. Kotani, T. Sazuka, N. Miyajima, M. Sugiura,

and S. Tabata. (1995)

Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis

sp. strain PCC6803. I. sequence features in the 1 Mb region from map positions 64 to 92 of the genome.
DNA Res, 2:191—198

30. P. Karlovsky, H. H. Prell (1991)
The TRP1 gene of Phytophthora parasitica encoding indole-3-glycerolphosphate synthase-N-(5'-phosphoribosyl)anthranilate isomerase: structure and evolutionary distance from homologous fungal genes.
Gene 109(1):161-165

31 Klabunde, T. et al. (1996)
Mechanism of Fe(III)-Zn(II) purple acid phosphatase based on crystal structures
J. Mol. Biol. 259, 737-748

32. M. Kobayashi, H. Komeda, N. Yanaka, T. Nagasawa, H. Yamada (1992)
Nitrilase from Rhodococcus rhodochrous J1. Sequencing and overexpression of the gene and identification of an essential cysteine residue
J. Biol. Chem. 267:20746-20751

33. M. Kobayashi, H. Izui, T. Nagasawa, H. Yamada (1993).
Nitrilase in biosynthesis of the plant hormone indole-3-acetic acid from indole-3-acetonitrile: cloning of the Alcaligenes gene and site-directed mutagenesis of cysteine residues.
Proc. Natl. Acad. Sci. U.S.A. 90:247-251

34. Z. H. Lu, R. Zhang, R. B. Diasio (1992)
Purification and characterization of dihydropyrimidine dehydrogenase from human liver.

J Biol Chem 267(24):17102-17109

35. A. Lupas, M. Van Dyke and J. Stock, (1991)
"Predicting Coiled Coils from Protein Sequences."
Science,. 252: p. 1162-1164

36. K. Maeda-Yorita, K. Aki, H. Sagai, H. Misaki, V. Massey (1995)
L-lactate oxidase and L-lactate monooxygenase:
mechanistic variations on a common structural theme.
Biochimie;77(7-8):631-642

37 C. Menendez C, G. Igloi, H. Henninger, R. Brandsch (1995)
A pAO1-encoded molybdopterin cofactor gene (moaA) of
Arthrobacter nicotinovorans: characterization and
site-directed mutagenesis of the encoded protein
Arch. Microbiol. 164:142-151

38 Murzin, A.G., Brenner, S.E., Hubbard, T. and Chotia, C. (1995)
Mol. Biol. 247, 536-540

39. K. E. Narva, J. L. van Etten, B. E. Slatko, J. S. Benner (1988)
The amino acid sequence of the eukaryotic DNA [N6-adenine]methyltransferase,
M.CviBIII, has regions of similarity with the prokaryotic isoschizomer M.TaqI and other
DNA [N6-adenine] methyltransferases.

Gene 74:253-259

40. C. P. Ponting (1997)

CBS domains in CIC chloride channels implicated in myotonia and nephrolithiasis

(kidney stones).

J Mol Med 75(3):160-163

41 F. Rentier-Delrue, SC. Mande, S. Moyens, P. Terpstra, V. Mainfroid, K. Goraj,

M. Lion, WG. Hol, JA. Martial (1993)

Cloning and overexpression of the triosephosphate isomerase

genes from psychrophilic and thermophilic bacteria. Structural

comparison of the predicted protein sequences.

J Mol Biol 229(1):85-93

42 B. Rost, C. Sander and R. Schneider (1994)

PHD – an automatic mail server for protein secondary structure prediction.

Comput Appl Biosci, 10:53-60

43 H. Sakakibara, M. Watanabe, T. Hase, T. Sugiyama (1991)

Molecular cloning and characterization of complementary DNA

encoding for ferredoxin-dependent glutamate synthase in maize

leaf.

J Biol Chem 266(4):2028-2035

44. W. Sidler, E. Niederer, F. Suter, H. Zuber (1986)

The primary structure of Bacillus cereus neutral proteinase and comparison with

thermolysin and Bacillus subtilis neutral proteinase.

Biol Chem Hoppe Seyler 367(7):643-657

45. J. A. Sigrell, A. D. Cameron, T. A. Jones, S. L. Mowbray (1997)

Purification, characterization, and crystallization of Escherichia coli ribokinase.

Department of Molecular Biology, Uppsala University, Sweden.

Protein Sci (11):2474-2476

46. M. D. Sintchak, M. A. Fleming, O. Futer, S. A. Raybuck, S. P. Chambers, P. R. Caron, M. A. Murcko, K. P. Wilson (1996)

Structure and mechanism of inosine monophosphate dehydrogenase in complex with the immunosuppressant mycophenolic acid.

Cell 85(6):921-930

47 R. L. Tatusov, S. F. Altschul, and E. V. Koonin (1994)

Detection of conserved segments in proteins: iterative scanning ofsequence databases with alignment blocks.

PNAS, 91:12091-12095

48. J. D. Thompson, D. G. Higgins, and T. J. Gibson. (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

49 Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997)

Nucleic Acids Res 1997 Dec 15;25(24):4876-4882

The CLUSTAL_X windows interface: flexible strategies for multiple

sequence alignment aided by quality analysis tools.

Nucleic Acids Res 1997 Dec 15;25(24):4876-4882

50. L. Wall, T. Christiansen, and R. L. Schwartz (1996)

Programming Perl.

O'Reilly & Associates Inc., 2nd edition,.

51. M. Weng, H. Zalkin (1987)

Structural role for a conserved region in the CTP synthetase glutamine amide

transfer domain.

J. Bacteriol. 169:3023-3028

52. J. Wrzesinski, A. Bakin, K. Nurse, B. G. Lane, J. Ofengand (1995)

Purification, cloning, and properties of the 16S RNA pseudouridine 516 synthase

from Escherichia coli.

Biochemistry 34(27):8904-8913

53. H. Yokota, P. Fernandez-Salguero, H. Furuya, K. Lin, O. W. McBride, B. Podschun,

K.D. Schnackerz, F. J. Gonzalez (1994)

cDNA cloning and chromosome mapping of human dihydropyrimidine dehydrogenase,

an enzyme associated with 5-fluorouracil toxicity and congenital thymine uraciluria.

J Biol Chem 269(37):23192-23196

Weitere Quellen :

J. Darnell, H. Lodish, D. Baltimore

Molekulare Zellbiologie

1. Auflage, 1993

J. Schultz

Evolution von Proteinkinasen :

Modulidentifizierung mittels Sequenzanalyse

Diplomarbeit, Universität Konstanz, 1996

Anhang - Veröffentlichungen