Phylogenetische Untersuchung von vier RNA-bindenden Proteindomänen

Axel Wintsche

13. März 2009

Zusammenfassung

RNA ist neben der Proteinsynthese an vielen regulatorischen Prozessen in der Zelle beteiligt. Dabei ist die RNA meist mit Proteinen gebunden. Dies verhindert ihre Degradation und kann dem RNA-Proteinkomplex neue Funktionalität geben. Die Bindung zwischen RNA-Molekül und Protein kann durch spezielle Proteindomänen etabliert werden. Auf RNA-Seite kann zudem eine Spezifische Sequenz oder Struktur notwendig sein. In diesem Praktikum wurde die evolutionäre Konservierung von vier RNA-bindenen Proteindomänen untersucht (RRM, SAM, PUF und KH). Von Interesse war dabei die Verteilung im phylogenetischen Baum. Für die Proteindomänen wurden phylogenetische Bäume und statistische Auswertungen erstellt. Domänenkombinationen in Proteinen wurden auch auf Homologie untersucht.

1 Material und Methoden

1.1 Proteindaten

Die Basis bildet die Superfamily Datebank in der Version vom 18.01.2009. In der Datenbank wurde nach annotierten Family IDs (siehe Tabelle 1) der hier betrachteten Proteindomänen gesucht. Die Proteine in der SuperFamily werden durch HMM annotiert. Verwendet werden dabei mehrere Modelle für die Familie und die Superfamilie. Domänen werden annotiert wenn mindestens ein HMM einer Familie oder einer Superfamilie einen *evalue* unter einer bestimmten Grenze haben (siehe SuperFamily Webseite für genauere Informationen). Dadurch können Proteine fälschlicherweise annotiert sein, sog. “false positives”. Aus diesem Grund wurde für jede Proteindomäne ein Subset der gefundenen Proteine erstellt. In dem Subset sind alle Proteine welche mit einem $evalue < 1e^{-4}$ von mindestens einem Family HMM erkannt werden. Für diese signifakanten Subsets sind die Bäume und statistischen Auswertungen verfügbar aber sie werden nachfolgend nicht betrachtet.

Domäne	Family ID
RRM	54929
SAM	47773
PUF	63611
KH	54792

Tabelle 1: Proteindomänen und ihre Family ID (ScopID)

1.2 Statistiken

Um zu sehen wie weit die Proteindomänen in bestimmten Taxa verbreitet sind ist es notwendig zu berechnen wie viele Organismen eines Taxon überhaupt in der Superfam bekannt sind. Bei dieser Berechnung, der Anzahl aller Spezies pro Taxon, wurden alle in der Superfamily vorhandenen Spezies auf ihre NCBI ID gemappt. Aus dieser Liste wurden alle mehrfachen Einträge entfernt (da Superfam → NCBI ID nicht injektiv!) da diese im Baum als ein Organismus gezählt werden (bei unterschiedlichen Superfam Organismen mit gleicher NCBI ID wurde derjenige Organismus mit den meisten annotierten Domänen gewählt). Die Organismen wurden dann in dem NCBI-Baum von den Blättern (Organismus) bis zur Wurzel propagiert und in allen durchlaufenen Knoten (Taxa) die Anzahl um eins erhöht. Genauer wurde in einer Hashtabelle jeder Organismus (NCBI ID) mit dem Wert 1 initialisiert und danach für alle Organismen (= Blätter) folgende rekursion angewandt:

Solange ID > 1 (da NCBI ID 1 die Wurzel ist)

1. ID = NCBI ID des Vaterknotens(ID)
2. falls Hasheintrag von ID existiert: Hash(ID) += 1
   sonst neuer Hasheintrag: Hash(ID) = 1
3. gehe zu 1

Die so ermittelten Werte repräsentieren also die Anzahl der zugehörigen Organismen eines Taxon. Nun kann die Anzahl annotierter Organismen für ein Taxon mit diesen Werten verglichen werden. Dabei viel auf das bei keiner Domäne alle Organismen annotiert waren. Da alle Domänen in den Metazoa annotiert waren wurde untersucht wie viele Organismen nicht annotiert waren, immer genau 64 von den bekannten 69 Organismen. Nun wurde ermittelt welche Organismen jeweils fehlten. Es stellte sich heraus das immer die fünf gleichen Organismen nicht annotiert waren.

Die Anzahl annotierter Proteine wurde aus der Datenbank Superfamily gewonnen. Erstaunlich war das die Anzahl annotierter Proteindomänen auch in nahe verwandten Organismen z.T. sehr unterschiedlich war. So wurde z.B. die RRM Domäne in Drosophila melanogaster 335 mal annotiert, in den restlichen Drosophiliden im Schnitt aber viel seltener. Weitere Unterscheide sind in Tabelle 3 zu sehen. Ursache dieser Unterschiede kann die in Abschnitt 1.2 beschriebene Methode der Annotation von Proteindomänen sein. Sind die HMM auf Grundlage dieser Spezies erstellt worden werden in diesen eventuell mehr Treffer gefunden. Aber auch unvollstaendig sequenzierte Genome können eine Rolle spielen

Spezies	Assembly / Release
Homo sapiens (Ensembl)	NCBI36
Pan troglodytes (Ensembl)	CHIMP2.1
Pongo pygmaeus (Ensembl)	PPYG2
Mus musculus (Ensembl)	NCBIM37
Rattus norvgegicus (Ensembl)	RGSC3.4
Drosophila melanogaster (Flybase)	5.15
Drosophila yakuba (Flybase)	1.2
Caenorhabditis elegans (Wormbase)	Ce147
Caenorhabditis briggsae (Wormbase)	Cb3

Tabelle 2: Genomversionen im Vergleich

	#Domänen			
Spezies	RRM	SAM	PUF	KH
Homo sapiens	825	220	13	111
Pan troglodytes	578	166	7	73
Pongo pygmaeus	409	104	5	48
Mus musculus	786	189	12	84
Rattus norvgegicus	674	163	11	80
Spermophilus tridecemlineatus	198	58	4	32
Drosophila melanogaster	510	90	7	77
Drosophila yakuba	183	41	4	24
Caenorhabditis elegans	353	46	15	58
Caenorhabditis briggsae	171	23	14	29

Tabelle 3: Unterschiede in der Anzahl annotierter Domänen naher verwandter Spezies.

(Tabelle 2). Vergleicht man die Annotationen aus dem Subset der signifikanten Proteine wird dieser Unterscheid noch deutlicher (Tabelle 4).

1.3 Dateien

Es wurden für die Domänen statistische Daten auf den phylogenetischen Baum der NCBI Taxonomie projeziert. Die entstandenen Dateien sind auf der Webseite verfügbar. Die Datei ass_18-Jan-2009.tab.gz kann von der Superfam Webseite bezogen werden.

1.4 Scripte & Pipelines

In Tabelle 5 sind die Skripts, welche im Rahmen des Pratkikums entstanden sind gelistet und kurz erklärt.

Die Bäume und Dateien wurden wie folgt erstellt.
Für einzelne Domänen (Bsp PUF):

1. grep -P "\t63611\t" 964.ass.tab >ass.tab

Spezies	#Domänen			
	RRM	SAM	PUF	KH
Homo sapiens	85	16	8	44
Pan troglodytes	61	11	6	24
Pongo pygmaeus	44	4	3	18
Mus musculus	119	14	9	39
Rattus norvgegicus	97	9	6	29
Spermophilus tridecemlineatus	20	4	2	8
Drosophila melanogaster	51	8	5	1
Drosophila yakuba	13	3	1	-
Caenorhabditis elegans	31	-	12	7
Caenorhabditis briggsae	12	-	10	4

Tabelle 4: Unterschiede in der Anzahl signifikant annotierter Domänen naher verwandter Spezies.

2. ./SIGNI
3. ./GETALL (braucht `specnames.pl`, `motifpergene.pl`, superfam2ncbi.tab)
4. ./GETSIG (braucht `specnames.pl`, `motifpergene.pl`, superfam2ncbi.tab)
5. `tax -f idall.list -n` >all.tree
6. `tax -f idsig.list -n` >sig.tree
7. `addStats.pl statsall.list all.tree` (siehe auch Quellcode)
8. `addStats.pl statssig.list sig.tree` (siehe auch Quellcode)

Für kombinierte Domänen (Bsp. RRM-KH):

1. grep -P "\t54929\t" 964.ass.tab >ass1.tab
2. grep -P "\t54792\t" 964.ass.tab >ass2.tab
3. ./JOINIT (braucht `specnames.pl`, `motifpergene.pl`, superfam2ncbi.tab)
4. `tax -f idall.list -n` >all.tree
5. `addStats.pl statsall.list all.tree` (siehe auch Quellcode)

Name	Beschreibung
addStats.pl	Fügt in einen Newick-Baum Daten der statistischen Auswertung ein. Es werden an den Blättern die Werte und an den inneren Knoten jeweils Minimum, Durchschnitt und Maximum dieser Werte des Unterbaumes und die Anzahl darin enthaltener Organismen hinzugefügt.
getNCBIname.pl	Dieses Skript wurde verwendet um die Superfamily DB Organismen auf die NCBI Namen zu projezieren. Als Eingabe wird eine Liste mit den Spalten "Bezeichner"\t"NCBI ID" benötigt. Mit diesem Skript wurde die superfam2ncbi.tab erstellt.
HowMuch.pl	Aus einer Liste von NCBI IDs wird anhand des taxonomischen NCBI Baumes eine Liste aus Knoten und zugehöriger Anzahl im Unterbaum enthaltener Organismen (nur diese aus der Eingabeliste) erstellt. Aus der Liste aller in der Superfamily bekannter Organismen wurde mit diesem Skript die Datei viechcount erstellt.
kindel.pl	Zu einem vom Nutzer gewählten Taxon der NCBI (über Name oder ID) werden alle Kind-Knoten des taxonomischen NCBI Baumes und die darin enthaltene Anzahl aus der Superfamily bekannten Organismen ausgegeben. Mit der Option **-a** werden zusätzlich alle diese Organismen ausgegeben. So kann z.B. die Liste annotierter Organismen in den Metazoa mit den aus der Superfamily bekannten Metazoa verglichen werden.
motifpergene.pl	Join von zwei Listen. Beide Listen haben in der ersten Spalte den "scientific name" der NCBI als Schlüssel. In der zweiten Spalte steht entweder die zugehörige Anzahl annotierter Domänen oder annotierter Gene. Diese Werte werden um die Zahl Domänen pro Gen erweitert und in eine neue Liste geschrieben. Das Skript ist Bestandteil der Pipelines.
specnames.pl	Bestandteil der Pipelines. Es joint zwei Tabellen anhand der ersten Spalte.
tree2org.pl	Extrahiert aus einem Newick Baum alle Blattknoten. Wurde verwendet um aus den all.tree Dateien z.B. alle gefundenen Metazoa zu extrahieren (siehe auch kindel.pl und Option -a). Dazu wurde all.tree mit dem Programm "Dendroscope" geöffnet, der zu untersuchende Unterbaum komplett markiert und extrahiert (in extra Datei). Diese wurde gespeichert und tree2org.pl auf dem STDIN übergeben. Die so erstellte Liste kann dann für weitere Auswertungen herangezogen werden.

Tabelle 5: Liste von im Praktikum erstellten Perl-Skripts

2 RNA Recognition Motif (RRM)

Die RRM-Domäne ist eine der meist verbreitetsten Proteindomänen unter den Eukaryoten [4]. Andere Bezeichnungen sind RNA-bindende Domäne (RBD) oder Ribonucleoprotein Domäne (RNP). Identifiziert wurde sie in den '80 Jahren bei Untersuchungen von Proteinkomplexen mit "mRNA-precursors" und heterogener Kern-RNA [2]. Weitere Arbeiten [4] konnten zeigen das Proteine mit RRM-Domänen an vielen unterschiedlichen Funktionen in der Zelle beteiligt sind. Die RRM-Domäne besteht allgemein aus 90 Aminosäuren und faltet sich in eine $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$ Struktur [1, 4].

2.1 Evolution

Die RRM-Domäne war in Bakterien, Archaea als auch Eukaryoten annotiert. Vergleicht man die Anzahl der Proteine, welche die RRM-Domäne annotiert haben unter den Organismen, ergibt sich eine Partitionierung entsprechend der taxonomischen Einordnung dieser. In den Archaea und Bakterien finden sich wenige Proteine mit RRM verglichen mit der Anzahl eukaryotischer Proteine mit RRM (siehe Tabelle 6).

Taxon	#Proteine: min/avg/max	(annotiert/in superfam)
Bacteria	1/2.07/8	(159/668)
Archaea	1/2.5/4	(2/52)
Eukaryota	14/129.95/562	(193/193)
Protostomia	110/151.18/335	(22/22)
Deuterostomia	125/269.73/562	(37/37)
Sarcopterygii	125/260.79/540	(28/28)
Teleostei	231/355/562	(5/5)

Tabelle 6: Übersicht der RRM-Domänenverteilung in den Taxa.

Die geringe Abdeckung bei den Bakterien ist bei genauerer Betrachtung ein Wegfall kompletter Unterbäme. Es wurden von allen 668 in der Datenbank der `Superfam` gelisteten Bakterien nur 159 RRM-annotiert. Betrachtet man die Untergruppen (siehe Tabelle 7) stellt sich heraus, dass die RRM-Domänen nicht über die gesammten Bakterien verteilt gefunden wurden, sondern in bestimmten Untergruppen fast vollständig vorhanden sind, in anderen dagegen garnicht. Allerdings könnten sich die Organismen in bestimmten Untergruppen im Genom sehr ähnlich sein, was das vollständige Vorhandensein der Domäne erklärt. Mit dem heutigen Verständis der Evolution würde sich die Verteilung der RRM-Domäne nicht ohne zusätzliche Betrachtung der Lebensräume der Organismen und der Funktion RRM-enthaltender Proteine erklären [5].

In den Eukaryoten ist die RRM-Domäne in allen Organismen vorhanden. Anstiege in der Anzahl RRM-enthaltender Proteine sind vor der Abspaltung der Eukaryoten, der Streptophyta (siehe Tabelle 6) und der Teleostei festzustellen, sowie ein generell leichter Anstieg mit zunehmender Komplexität des Organismus.

Taxon	annotiert/in superfam
Bacteria	159/668
Cyanobacteria	33/33
Proteobacteria	83/354
Gammaproteobacteria	49/172
Alphaproteobacteria	0/86
Betaproteobacteria	11/57
delta/epsilon subdivisions	23/39
Firmicutes	7/128
Bacilli	0/94
Clostridia	7/34
Deinococcus-Thermus	0/4
unclassified Bacteria	0/1
Fusobacteria	0/1
Chlamydiae/Verrucomicrobia group	4/16
Verrucomicrobia	3/3
Chlamydiae	1/13
Bacteroidetes/Chlorobi group	20/24
Bacteroidetes	10/14
Chlorobi	10/10
Fibrobacteres/Acidobacteria group	2/2
Aquificae	0/3
Chloroflexi	1/7
Thermotogae	0/7
Actinobacteria	0/53
Planctomycetes	1/1
Spirochaetes	8/13
Spirochaetaceae	2/7
Leptospiraceae	6/6
Tenericutes	0/21

Tabelle 7: Übersicht der RRM-Domänenverteilung in dem Taxon Bacteria.

3 Sterile Alpha Motive (SAM)

Die SAM-Domäne findet sich in vielen funktionell unterschiedlichen Proteinen wie etwa Transkriptions- und Translationsregulatoren und Kinasen [3, 9, 6]. Dabei ist die Domäne selber auch an verschiedenen Funktionen beteiligt, meist an Protein-Protein Interaktionen. Aber auch posttranskriptionelle Regulation durch RNA-Bindung wurde entdeckt [8]. Die Sekundärstruktur der SAM-Domäne faltet sich in vier bis fünf α-Helices, welche ein globuläres Bündel mit hydrophobem Kern bilden.

3.1 Evolution

Taxon	#Gene: min/avg/max	(annotiert/in superfam)
Archaea	-	(0/52)
Bacteria	1/1.4/2	(5/668)
Eukaryota	1/28.84/415	(188/193)
Fungi	1/5.12/14	(78/78)
Metazoa	15/72.73/415	(64/64)
Protostomia	20/33.05/66	(22/22)
Deuterostomia	44/102.59/415	(37/37)
Sarcopterygii	44/91.79/189	(28/28)
Teleostei	114/184.6/415	(5/5)

Tabelle 8: Übersicht der SAM-Domänenverteilung in den Taxa.

Die SAM-Domäne ist fast ausschließlich in Eukaryoten vorhanden (siehe Tabelle 8). In diesen ist sie in den Untergruppen (bis auf die Alveolata) vollständig zu finden. Dies läßt ihren Ursprung in der Entstehung der Eukaryoten vermuten. Bei den fünf Funden in den Bakterien gibt es keinen experimentellen Beleg über ihre Funktionalität. Nach [7] könnte die große Sequenzdivergenz der SAM Domäne das Finden von Homologen durch Sequenzähnlichkeit in Bakterien verhindern.

Die Anzahl der Proteine mit SAM in den Organismen (Tabelle 8) im phylogenetischen Baum unterstützt die gängige Ansicht über die Zunahme der Genomgröße durch z.B. Genomduplikationen. Innerhalb der Deuterostomia (sogar der Mammalia) ist im Gegensatz zu dem restlichen Baum eine große Varianz der Genanzahl einzelner Organismen beobachtbar. Auch hier könnten die in 1.2 beschriebenen Probleme der Grund sein. Das Vorkommen von Proteinen mit mehr als einer SAM-Domäne (Tabelle 9) ist spezifisch für Metazoa.

Taxon	avg SAM pro Protein
Fungi	1.00
Metazoa	1.21
Protostomia	1.27
Deuterostomia	1.17

Tabelle 9: Hier ist die durchschnittliche Anzahl an SAM-Domänen pro Protein für bestimmte Taxa gelistet. Der Durchschnitt pro Organismus berechnet sich durch die Anzahl seiner annotierten SAM-Domänen dividiert mit der Anzahl seiner annotierten SAM-enthaltenden Proteine.

4 Pumilio family (PUF)

Die PUF-Domäne, auch PUM oder PUM-HD genannt, besteht aus mehreren (meist acht) Pumilio-repeats welche wiederum aus zwei Helices bestehen [11]. Funktionell handelt es sich um eine RNA bindende Domäne [12].

4.1 Evolution

Homologe Sequenzen der PUF-Domäne sind in allen Eukaryoten zu finden (Tabelle 10) und in der Bakterie Wolbachia endosymbiont of Drosophila melanogaster. Auch die Anzahl PUF-enthaltender Proteine in einem Organismus ist erstaunlich homogen. Ausnahmen sind die Protostomia mit weniger PUF-Domänen und die Streptophyta mit deutlich mehr PUF-Domänen (17.6) als der eukaryotische Durchschnitt. Das die Teleosten durchschnittlich nur genauso viele PUF-Proteine wie die Deuterostomia besitzen ist für die hier betrachteten Domänen besonders. Allen gemeinsam ist das alle Proteine die PUF-Domäne maximal einmal besitzen.

Taxon	#Gene: min/avg/max	(annotiert/in superfam)
Archaea	-	(0/52)
Bacteria	1/1/1	(1/668)
Eukaryota	2/6.82/42	(193/193)
Fungi	3/6.45/14	(78/78)
Metazoa	2/5.44/17	(64/64)
Protostomia	2/3.59/7	(22/22)
Deuterostomia	3/6.14/17	(37/37)
Teleostei	4/6.6/10	(5/5)

Tabelle 10: Übersicht der PUF-Domänenverteilung in den Taxa.

5 K Homology Domain (KH)

Die KH-Domäne wurde erstmals in K-Proteinen entdeckt (daher K homolog). Wie sich jedoch zeigte ist die KH-Domäne ein sehr allgemeines und weitverbreitetes Motiv [10]. Ein gemeinsames Merkmal der KH-enthaltenden Proteine ist eine physische oder funktionelle Assoziation mit RNA-Molekülen – wie sich zeigte durch die KH-Domäne induziert.

5.1 Evolution

Taxon	#Gene: min/avg/max	(annotiert/in superfam)
Archaea	1/1.73/3	(52/52)
Bacteria	1/1.43/5	(646/668)
Eukaryota	3/21.13/139	(193/193)
Fungi	5/8.38/16	(78/78)
Metazoa	18/41.19/139	(64/64)
Protostomia	19/28.5/77	(22/22)
Deuterostomia	25/49.97/139	(37/37)
Teleostei	51/75.2/139	(5/5)

Tabelle 11: Übersicht der KH-Domänenverteilung in den Taxa.

Die KH-Domäne läßt sich in den drei großen Reichen Archaea, Bakterien und Eukaryoten finden (Tabelle 11). Damit scheint sie die evolutiv älteste hier betrachtete Domäne zu sein. Während sie in den Archaea und Bakterien nur ein paar mal pro Organismen zu finden ist besitzen Eukaryoten viele Proteine mit dieser Domäne. Innerhalb der Eukaryoten gibt es quantitative Unterschiede, gut sichtbar zwischen Protostomia, Deuterostomia und Teleostei. Bei den Protostomia ist im Gegensatz zu den Deuterostomia die Anzahl KH-enthaltender Proteine recht homogen. Durch in 1.2 beschriebene Probleme sind auch hier wieder ungewöhnlich viele Proteine in z.B. C. elegans, D. melanogaster, H. sapiens und anderen festzustellen. In den Bakterien gibt es im Gegensatz zu den Archaea und Eukaryota keine Proteine mit mehr als einer KH-Domäne (Tabelle 12).

Taxon	avg KH pro Protein
Bacteria	1.00
Archaea	1.49
Eukaryota	2.14
Fungi	2.47
Metazoa	2.22
Protostomia	2.09
Deuterostomia	2.30

Tabelle 12: Hier ist die durchschnittliche Anzahl an KH-Domänen pro Protein für bestimmte Taxa gelistet. Der Durchschnitt pro Organismus berechnet sich aus die Anzahl seiner annotierten KH-Domänen dividiert mit der Anzahl seiner annotierten KH-enthaltenden Proteine.

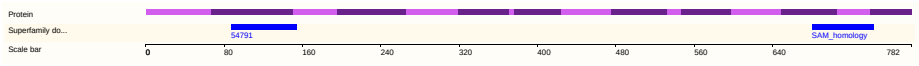

Abbildung 1: SAM und KH im Danio rerio ENSDARP00000094302-Protein

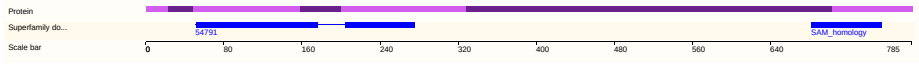

Abbildung 2: SAM und KH im Drosophila melanogaster FBpp0080363-Protein

6 Mehrere RNA-bindende Domänen

6.1 SAM & KH

Proteine mit SAM- und KH-Domäne finden sich in fast allen Metazoa, in zwei Euglenozoa, Monosiga brevicollis (Choanoflagellida) und Batrachochytrium dendrobatidis (Fungi). Alle Proteine in den Metazoa scheinen homolog zu sein. Die Domänen befinden sich an den Enden der Aminosäurekette (Bild 1, Bild 2, Bild 3). Teleosten besitzen zwei oder mehr dieser Proteine, welche paralog sind [Ensembl-Daten]. Der Rest hat durchschnittlich ein "Gen" welches aber mehrere Transkripte besitzen kann (so z.B. Canis lupus familiaris oder D. melanogaster).

6.2 RRM & PUF

Die Kombination von RRM- und PUF-Domäne in einem Protein ist nur den Vertretern der Fungi eigen. In den Fungi gibt es 67 von 81 Organismen mit annotierten RRM-PUF Proteinen. Die meisten davon besitzen nur ein einziges solches Protein. Die Ausnahme bilden hier die Saccharomyces und Schizosaccharomyces. Die Domänen liegen bei Saccharomyces cerevisiae mittig der Aminosäurekette (Bild 4). Vom N-terminalen Ende gesehen kommt erst die RRM-Domäne gefolgt von PUF (Pumilo_RNA).

6.3 RRM & KH

Proteine, welche die RRM- und die KH-Domäne enthalten finden sich nur in eukaryotischen Lebewesen. Sie sind vollständig in den Deuterostomia vorhanden. Bei den Protostomia konnte diese Domänenkombination bei den Drosophilinae, Daphnia pulex und Lottia gigantea gefunden werden. RRM-KH Proteine finden sich auch noch in wenigen Alveolata und Vividiplantae (Tabelle 13). Ob es sich hier um funktionelle Proteine handelt oder um "false positives" muss überprüft werden.

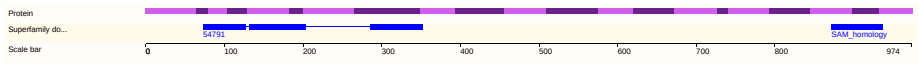

Abbildung 3: SAM und KH im Homo sapiens ENSP00000362993-Protein

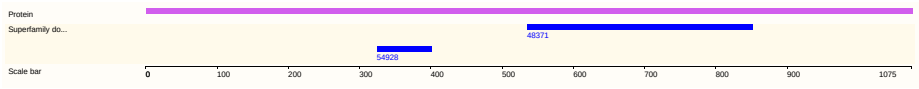

Abbildung 4: RRM und PUF in Saccharomyces cerevisiae YPR042C-Protein

Taxon	annotiert/in superfam
Bacteria	0/668
Archaea	0/52
Eukaryota	61/193
Rhodophyta	0/1
Haptophyceae	1/1
Parabasalidea	0/1
Heterolobosea	0/1
Viridiplantae	9/18
Fungi/Metazoa group	47/143
Fungi	0/78
Choanoflagellida	1/1
Metazoa	46/64
Alveolata	4/13
stramenopiles	0/6
Euglenozoa	0/5
Diplomonadida group	0/1
Amoebozoa	0/3

Tabelle 13: Übersicht der RRM-KH-Domänenverteilung.

Im Schnitt hat jeder Organismus der Deuterostomia 3 RRM-KH Proteine, die restlichen Organismen nur eines. Hier noch ungeklärt ist die Frage ob die Proteine der Metazoa homolog sind. Die Domänenverteilung in den Proteinen (Bild 5 Bild 6 Bild 7) spricht aber dafür. Auch bei den Proteinen eines Organismus scheint es sich um homologe zu handeln (Bild 5, 8, 9, 10).

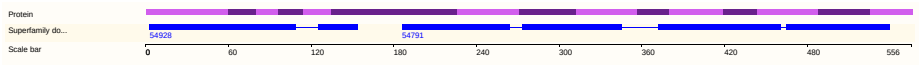

Abbildung 5: RRM und KH in Homo sapiens ENSP00000320204-Protein

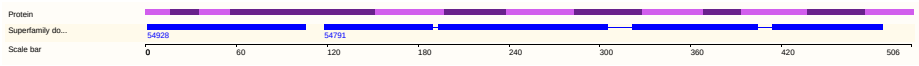

Abbildung 6: RRM und KH in Danio rerio ENSDARP00000040323-Protein

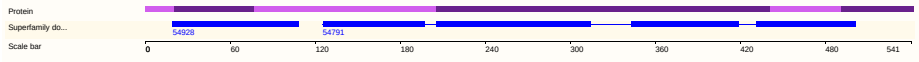

Abbildung 7: RRM und KH in Aedes aegypti AAEL006876-PA-Protein

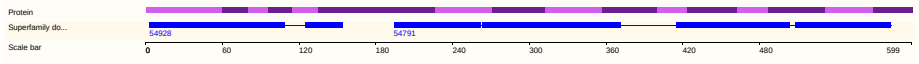

Abbildung 8: RRM und KH in Homo sapiens ENSP00000371634-Protein

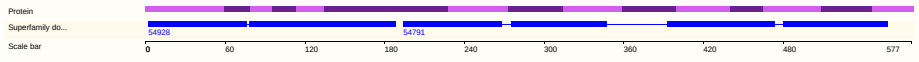

Abbildung 9: RRM und KH in Homo sapiens ENSP00000290341-Protein

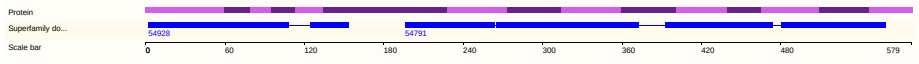

Abbildung 10: RRM und KH in Homo sapiens ENSP00000258729-Protein

Literatur

[1] A. Cléry, M. Blatter, and F.H.T. Allain. RNA recognition motifs: boring? Not quite. *Current Opinion in Structural Biology*, 18(3):290–298, 2008.

[2] G. Dreyfuss, MS Swanson, and S. Pinol-Roma. Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. *Trends Biochem Sci*, 13(3):86–91, 1988.

[3] C.A. Kim and J.U. Bowie. SAM domains: uniform structure, diversity of function. *Trends in Biochemical Sciences*, 28(12):625–628, 2003.

[4] C. Maris, C. Dominguez, and F.H.T. Allain. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS Journal*, 272(9):2118–2131, 2005.

[5] K. Maruyama, N. Sato, and N. Ohta. Conservation of structure and cold-regulation of RNA-binding proteins in cyanobacteria: probable convergent evolution with eukaryotic glycine-rich RNA-binding proteins. *Nucleic Acids Research*, 27(9):2029–2036.

[6] CP PONTING. SAM: A novel motif in yeast sterile and Drosophila polyhomeotic proteins. *Protein Science*, 4(9):1928, 1995.

[7] CP Ponting, L. Aravind, J. Schultz, P. Bork, and EV Koonin. Eukaryotic Signalling Domain Homologues in Archaea and Bacteria. Ancient Ancestry and Horizontal Gene Transfer. *Journal of Molecular Biology*, 289(4):729–745, 1999.

[8] I.C.P.C.M. Processing. The RNA-binding SAM domain of Smaug defines a new family of post-transcriptional regulators. *Nature Structural Biology*, 10:614–621, 2003.

[9] J. SCHULTZ, CP PONTING, K. HOFMANN, and P. BORK. SAM as a protein interaction domain involved in developmental regulation. *Protein Science*, 6(1):249, 1997.

[10] H. Siomi, M.J. Matunis, W.M. Michael, and G. Dreyfuss. The pre-mRNA binding K protein contains a novel evolutionarily conserved motif. *NUCLEIC ACIDS RESEARCH*, 21:1193–1193, 1993.

[11] X. Wang, P.D. Zamore, and T.M.T. Hall. Crystal Structure of a Pumilio Homology Domain. *Molecular Cell*, 7(4):855–865, 2001.

[12] PD Zamore, JR Williamson, and R. Lehmann. The Pumilio protein binds RNA through a conserved domain that defines a new class of RNA-binding proteins, 1997.