Die Entdeckung neuer Proteindomänen sowie zwei Anwendungen ihrer genomweiten Identifizierung

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr.rer.nat.)

vorgelegt der

Mathematisch-Naturwissenschaftlich-Technischen Fakultät
(mathematisch-naturwissenschaftlicher Bereich)
der Martin-Luther-Universität Halle-Wittenberg

von Dipl. Biotechnol. Eike Staub
geboren am 12.7.1972 in Osnabrück

Gutachter:

1. Prof. Dr. Dr. Thomas Braun

2. Prof. Dr. Andre Rosenthal

3. Prof. Dr. Stefan Posch

Verteidigt am 29.04.2004 in Halle (Saale)

urn:nbn:de:gbv:3-000006642
[http://nbn-resolving.de/urn/resolver.pl?urn=nbn%3Ade%3Agbv%3A3-000006642]

Diese Arbeit entstand während meiner Zeit bei der Firma metaGen Pharmaceuticals GmbH. Allen Mitarbeitern sei herzlich für die kollegiale Arbeitsatmosphäre gedankt. Besonderer Dank gilt meinen Kollegen Dr. Bernd Hinzmann, Dr. Detlev Mennerich und Dr. Stefan Röpcke, die stets engagierte Diskussionspartner für die Erörterung fachlicher Fragen gewesen sind. Ich danke Prof. Dr. André Rosenthal für die Initiierung vieler herausfordernder Bioinformatik-Projekte innerhalb metaGens, in denen ich die für diese Dissertation notwendige Erfahrung sammeln konnte. Prof. Dr. André Rosenthal und Prof. Dr. Dr. Thomas Braun seien gedankt für fruchtbaren Ideenaustauch und Ihre konstruktive Kritik während der Entstehung der Texte dieser Dissertation. Prof. Dr. Gottfried Otting sei dafür gedankt, dass er mir Gelegenheit gegeben hat, mich in die Interpretation seiner Resultate einzubringen. Letztlich danke ich besonders meiner Frau Claudia für Ihre Geduld mit mir, wenn ich die Aufmerksamkeit für alltägliche Dinge vermissen ließ, weil ich mich in Gedanken mit den Problemen dieser Arbeit beschäftigte.

Inhaltsverzeichnis

1 Einleitung

Während des letzten Jahrzehnts erlebten die Biowissenschaften eine explosionsartige Vermehrung des Wissen über die genetischen Baupläne verschiedenster Organismen. Grundlage für diesen Wissenszuwachs war der technologische Fortschritt in der Entzifferung der Information des Erbmaterials, die Sequenzierung von DNA. Beginnend mit der Entschlüsselung der Erbinformation von einfachen Bakterien, die nur aus einigen Millionen Basenbausteinen besteht [1-3], hatte man in nur wenigen Jahren die Möglichkeit geschaffen, auch größere Genome, wie das der Bäckerhefe *Saccharomyces cerevisiae* und der ersten multizellulären Modellorganismen, des Wurms *Caenorhabditis elegans* und der Fruchtfliege *Drosophila melanogaster* aufzuklären [4-6]. Zum Jahrtausendwechsel gelang es dann, die komplette DNA Sequenz des ersten humanen Chromosoms zu ermitteln [7], bevor ein Jahr später die etwa drei Milliarden Basen lange Sequenz des menschlichen Erbguts entschlüsselt werden konnte [8,9].

Die Entschlüsselung von Genomen ging mit der Entstehung eines neuen Wissenschaftszweigs innerhalb der Genetik einher, der Genomik, die sich mit der genomweiten Analyse von Zellen oder Organismen befaßt. Die parallele experimentelle Untersuchung einer Vielzahl von Genprodukten mit Hilfe neuer Techniken, wie etwa der Expressionsanalyse von mehreren tausend Genen auf mRNA Ebene in nur einem Experiment mittels DNA Chips [10-12] oder der Aufklärung von Interaktionen zwischen Proteinen mittels moderner massenspektrometrischer Verfahren [13,14], erzeugt ungeheure Datenmengen. Die Interpretation dieser Daten erfordert den massiven Einsatz rechnergestützter Methoden.

Die Enthüllung der Sequenzen von Genen eilt ihrer experimentellen Charakterisierung weit voraus. Nur für einen Bruchteil der Gene eines Genoms ist auch auf Experimenten basierendes Wissen vorhanden. Die Funktionsvorhersage von Genen und Genprodukten allein auf Grundlage ihrer Sequenzinformation ist daher schnell zu einem der wichtigsten Teilbereiche der Genomik geworden. Sie ist das übergeordnete Thema dieser Arbeit.

1.1 Klassische Funktionsvorhersage von Proteinen

Die klassische Methode, die Funktion eines neuen Gens vorherzusagen, basiert auf der Ähnlichkeit der abgeleiteten Proteinsequenz zu den Sequenzen bekannter Proteine [15-17]. Eine signifikante Ähnlichkeit zweier Proteinsequenzen deutet auf einen gemeinsamen Sequenzvorfahren hin, also auf die Homologie zweier Proteine. Homologe Proteine weisen fast immer eine ähnliche dreidimensionale Struktur auf.

Oftmals ist die Struktur sogar noch konserviert, wenn die Homologie auf Sequenzebene nicht mehr nachzuweisen ist [18]. Auf Sequenzebene zeigt sich die Konservierung der Struktur besonders in der Konservierung jener Aminosäuren, die zur Ausbildung einer Struktur essentiell sind. In extrazellulären Bereichen von Proteinen sind daher häufig Paare von Cysteinen konserviert, die Disulfidbrücken bilden. In Sekundärstrukturelementen ist oftmals die regelmäßige Abfolge hydrophober und polarer Aminosäuren charakteristisch. So sind Gruppen von hydrophoben Aminosäuren, die auf einer Seite einer α-Helix liegen, häufig zum Kern des Proteins hin ausgerichtet [18]. Dies offenbart sich auf Sequenzebene in der periodischen Wiederkehr von hydrophoben Aminosäuren im Abstand von drei bis vier Sequenzpositionen. Es sind gerade diese strukturell wichtigen Ähnlichkeiten zwischen homologen Proteinsequenzen, die zur Detektion von Homologie zwischen entfernt verwandten Proteinen ausgenutzt werden.

Signifikante Sequenzähnlichkeit deutet also in erster Linie auf eine ähnliche Struktur hin, aber nicht zwingend auf eine gleiche Funktion. Für ein neues Protein lässt sich oftmals die Zugehörigkeit zu einer Proteinfamilie mit ähnlicher 3D-Struktur durch Sequenzanalyse vorhersagen. Auch ist innerhalb einer Proteinfamilie häufig ein allgemeiner biochemischer Wirkmechanismus konserviert, zum Beispiel ein enzymatischer Reaktionsmechanismus, die Rolle als Bindungspartner für einen Co-Faktor oder die Funktion als Protein-Protein-Adapter [19]. Dagegen lässt sich die präzise Funktion eines Proteins, wie zum Beispiel die Substratspezifität eines Enzyms, aufgrund von Sequenzähnlichkeiten bislang nicht vorhersagen.

1.2 Proteindomänen als Bausteine modularer Proteinarchitektur

Um die Probleme der Proteinsequenzanalyse zu verstehen, ist ein Verständnis des Aufbaus und der Evolutionsmechanismen von Proteinen notwendig. Die dreidimensionale Struktur eines Proteins ist häufig in Fragmente unterteilt, die scheinbar autonome Faltungseinheiten darstellen [20,21]. Solche Faltungseinheiten, deren globuläre Struktur maßgeblich durch innere Kräfte stabilisiert wird, während äußere Wechselwirkungen, etwa mit anderen Bereichen des Proteins, nur an der Oberfläche dieser Einheiten stattfinden, nennt man auch Domänen. Domänen stellen oftmals auch funktionelle Einheiten dar, wie etwa katalytisch aktive Untereinheiten von Enzymen oder wie Adapterdomänen, welche Interaktionen zwischen Proteinen vermitteln.

Manche Proteindomänen können als evolutionär mobile Module angesehen werden, da sie in verschiedenen Proteinen in unterschiedlichem Domänenkontext auf-

treten [20,22]. Die Wiederverwendung von Domänen hat vor allem in der Evolution von Proteinen multizellulärer Organismen eine wichtige Rolle gespielt. Das als „domain shuffling“ bekannte Phänomen wird erleichtert durch die Mosaikstruktur von Genen in Exons und Introns. So liegen vor allem extrazelluläre Domänen in Vertebratenproteinen häufig in mehreren Kopien pro Protein vor, die jeweils auf einzelnen Exons kodiert sind. Intron-vermittelte Rekombination der genomischen DNA erleichtert die Entstehung neuer Gene durch Neukombination von Exons, auch „exon shuffling“ genannt [23]. Auf diese Weise können neuartige Proteine entstehen, die sich aus mehreren evolutionär mobilen Domänen zusammensetzen. Während anerkannt ist, dass die Rekombination von Proteinmodulen in neue Proteine durch „exon shuffling“ in der Entwicklung von Eukaryonten eine entscheidende Rolle gespielt hat [24,25], ist weiterhin umstritten, ob auch die „urtümlichen“ Gene („ancient genes“) der ersten Organismen in Exons und Introns organisiert waren und ob schon während der Entwicklung der ersten Lebewesen die Entstehung neuer Gene und Proteine durch die Neukombination von Proteinmodulen geprägt war. Diese Streitfrage ist mit dem Alter von Introns eng verknüpft. Sie führte zu dem noch heute andauernden wissenschaftlichen Disput zwischen Verfechtern der „introns early“ und „introns late“ Theorien [26-28].

Die Domänenstruktur von Proteinen stellt ein wesentliches Erschwernis für die Analyse neuer Proteinsequenzen dar. Führt man sich ein aus unbekannten evolutionär mobilen Domänen zusammengesetztes Protein vor Augen, so wird dieses in unterschiedlichen Bereichen zu den verschiedenen homologen Domänen anderer Proteine Ähnlichkeiten aufweisen. Oftmals können Methoden des Sequenzvergleichs nur eine partielle Ähnlichkeit zwischen den tatsächlichen homologen Sequenzbereichen zweier Proteine feststellen. Die Ähnlichkeit von Domänenkopien aus Proteinfamilien unterschiedlicher Domänenarchitektur ist zudem häufig nur marginal signifikant, so dass es schwierig ist, bei Datenbanksuchen zwischen wahren und zufälligen Treffern zu unterscheiden. Aus diesen Gründen sind die Grenzen zwischen Domänen oftmals nur schwer zu definieren. Erleichtert wird die Entdeckung eines neuen Proteinmoduls, wenn dieses in mehreren Kopien pro Protein vorkommt, da man durch die Untersuchung der intramolekularen Sequenzwiederholungen besser Start, Ende und Länge einer Domäne ableiten kann. Solche intramolekularen repetitiven Einheiten nennt man auch „Repeats“.

Zahlreiche Arbeitsgruppen versuchten die Entdeckung neuer Proteindomänen zu automatisieren [29-31]. Entscheidend für die Qualität der Resultate scheint die Sensitivität der Sequenzvergleiche, der Umgang mit niedrig-komplexen

Subsequenzen und die korrekte Fragmentierung von Proteinsequenzen zur Definition der Domänengrenzen. Die trotz automatischer Ansätze noch immer andauernde Entdeckung von neuen mobilen Proteindomänen wird u.a. dokumentiert durch die regelmäßigen Publikationen neuer Domänen in der „Protein Sequence Motif" Sektion der Fachzeitschrift „Trends in Biochemical Sciences". Dies macht deutlich, dass die korrekte umfassende Beschreibung einer neuen Domäne immer noch mit einer aufwendigen manuellen Sequenzanalyse verbunden ist, die nicht durch vollautomatische Verfahren ersetzt werden kann.

1.3 Vorhersage der Proteinstruktur

Sequenzvergleichende Methoden werden im Bereich der 3D-Strukturvorhersage von Proteinen extensiv angewendet. Ein Ziel der strukturellen Genomik ist es, alle in der Natur vorkommenden Faltungsmotive zu ermitteln. Nach Schätzungen besteht das gesamte Proteinuniversum nur aus etwa 1.000 bis 10.000 verschiedenen Faltungstypen [21,32-36]. Um neue Faltungstypen zu entdecken, wird systematisch nach denjenigen Proteinen gesucht, die keine Homologie zu Proteinen mit bekannter Struktur zeigen. Die Strukturen solcher Proteine werden vorrangig aufgeklärt, um möglichst schnell einen Großteil der existierenden Faltungsmotive zu erfassen [37]. Die erfolgreichsten Methoden zur Vorhersage der dreidimensionalen Struktur eines uncharakterisierten Proteins beruhen auf der Anwendung mathematischer Modelle von Proteinfamilien oder Proteindomänen bekannter Struktur [38]. Es werden sogenannte „multiple Alignments" der Sequenzen einer Proteinfamilie erstellt, das sind Ausrichtungen der Sequenzen, in denen sich entsprechende homologe Aminosäuren untereinander stehen. Multiple Alignments von Proteinsequenzfamilien werden durch sogenannte „Positionsspezifische Score Matrizen" (PSSMs) oder durch „Hidden Markov Modelle" (HMM) modelliert [39-41]. Es ist bekannt, dass Alignment-basierte Methoden eine vielfach höhere Sensitivität und Spezifität als paarweise Methoden des Sequenzvergleichs haben [42]. Die höchste Spezifität und Sensitivität lässt sich derzeit mit HMM-basierten Algorithmen erreichen [43]. Allein durch die Anwendung von HMMs bereits bekannter Faltungsmotive lassen sich derzeit etwa 50% der Proteine eines neu sequenzierten Genoms einer Strukturfamilie zuordnen [44].

1.4 Neue Methoden zur Vorhersage der Proteinfunktion

Die Verfügbarkeit von kompletten Genomsequenzen vieler Organismen hat in den letzten Jahren die Entwicklung alternativer Verfahren der Funktionsvorhersage von Genen und Proteinen ermöglicht. Sie sind wichtige Ergänzungen der traditionellen

Methoden und sollen deshalb kurz dargestellt werden. Genannt seien im Besonderen drei Verfahren, die nur auf Sequenzinformation beruhen [45], sowie zwei weitere, die sich die Verfügbarkeit von Massendaten aus modernen experimentellen Methoden der Genomik zunutze machen.

Die Nutzung sogenannter „phylogenetischer Profile“ zur Funktionsvorhersage beruht auf der Annahme, dass zwei Proteine, die in einem funktionellen Zusammenhang stehen, häufig beide gemeinsam in verschiedenen Genomen vorkommen oder fehlen [46,47]. Das Muster der Präsenz von orthologen Proteinen über die Genome mehrerer Spezies hinweg nennt man phylogenetisches Profil. Man leitet dann für zwei Proteine einen funktionellen Zusammenhang ab, wenn sie ähnliche phylogenetische Profile besitzen. Hat eins der Proteine eine bekannte Funktion, so lässt sich die Funktion des zweiten vorhersagen.

Auch die Konservierung der Nachbarschaft zweier Genen in den Genomen mehrerer Organismen deutet häufig auf eine funktionelle Interaktion hin [48,49]. Schon lange ist bekannt, dass funktionell interagierende Proteine von Bakterien vielfach auf sogenannten Operons kodiert sind. Diese genomischen Abschnitte werden in eine einzelne mRNA transkribiert, von der aus mehrere verschiedene Proteine translatiert werden können. In Operons organisierte Gene sind oftmals nur in ihrer Gesamtheit für ein Bakterium von Nutzen. Paradebeispiel ist das gemeinsame Vorkommen von sequenzspezifischen Restriktionsendonukleasen und DNA Methyltransferasen. Die Nachbarschaft interagierender Gene im Genom ist bei Vertebraten weit seltener zu beobachten als bei Bakterien. Dennoch kann für ein unbekanntes Vertebratenprotein über eine Homologiebeziehung zu einem bakteriellen Protein mit konservierter Genomnachbarschaft indirekt eine Funktion hergeleitet werden.

Weiterhin kann die Fusion von Genen auf ihre gemeinsame Funktion hindeuten [47,50]. Die Fusion zweier Gene ist besonders dann von unmittelbarem Vorteil für einen Organismus, wenn deren Proteinprodukte in aufeinanderfolgenden Schritten einer biochemischen Reaktion zusammenwirken oder sogar in einem Proteinkomplex direkt miteinander interagieren. Solche Fusionen können zum Beispiel bewirken, dass der Weg eines Reaktionsprodukts zum nächsten Enzym einer biochemischen Reaktionskette verkürzt wird. Wenn man also die Fusion eines unbekannten Gens mit einem Gen bekannter Funktion zu einem Hybrid-Gen - auch „Rosetta Stone Sequence“ genannt - feststellt, liegt es nahe, dass diese Fusion aufgrund funktioneller Abhängigkeit der zwei Proteinprodukte in einem Organismus fixiert worden ist. Durch die extensive Analyse von Genfusionen über mehrere Genome hinweg haben verschiedene Arbeitsgruppen die Funktion zahlreicher

unbekannter Gene vorhergesagt [45,47,50].

Der massive technologische Fortschritt im spezifischen Nachweis von Proteinen mittels moderner massenspektrometrischer Verfahren [13,14] und die Parallelisierung genetischer Screens wie der Yeast-Two-Hybrid Methode [51] ermöglichten die Aufklärung von Proteininteraktionen im Hochdurchsatzverfahren. Die Analyse der Position eines unbekannten Proteins innerhalb eines Protein-Protein-Interaktionsnetzwerks kann dann Hypothesen über seine Funktion liefern, wenn es mit bereits funktionell charakterisierten Proteinen interagiert. Derzeit sind Datensätze zu Protein-Protein-Interaktionen vor allem für die Bäckerhefe *Saccharomyces cerevisiae* verfügbar. Vergleiche der Daten mit bekannten Proteinkomplexen zeigen, dass diese noch sehr fehlerbehaftet sind [52]. Indirekte Schlussfolgerungen für Proteininteraktionen in anderen Organismen basieren auf den klassischen Methoden der Feststellung von Homologie, oder besser Orthologie, durch Sequenzanalysen.

Durch die parallelisierte Analyse der mRNA-Expression mittels hochdichter DNA-Chips stehen inzwischen für viele Organismen umfangreiche Datensätze über die Expression tausender Gene in unterschiedlichen Zelltypen oder physiologischen Situationen zur Verfügung [53]. Gene, die in einem funktionellen Zusammenhang zueinander stehen, werden bei einer Veränderung des Zellzustandes oftmals koordiniert reguliert. Ein Beispiel ist die Anpassung des Stoffwechsels einer Zelle an eine neue Nährstoffsituation. In Bakterien oder Hefen wird diese veränderte Lebensbedingung von der Anpassung der Expression des geeigneten enzymatischen Apparats begleitet, um die neue Nahrungsquelle optimal zu nutzen. Tatsächlich scheint die Expression von Genen, deren Genprodukte Proteinkomplexe bilden, oftmals koordiniert reguliert zu werden [54]. Die Co-Regulation der Expression von Genen wurde daher ebenfalls für die Vorhersage von funktionellen Zusammenhängen genutzt. Diese Art der Vorhersage umfasst auch transiente und mittelbare Interaktionen zwischen Genprodukten [53].

1.5 Hintergründe der Entdeckungen neuer Proteindomänen in dieser Arbeit

Die Entdeckung und funktionelle Beschreibung neuer Proteindomänen ist das zentrale Thema dieser Arbeit. Dies wird dokumentiert durch fünf Manuskripte, die von einzelnen Entdeckungen und Charakterisierungen neuer Proteindomänen handeln. Die Domänen wurden durch verschiedene Ansätze identifiziert. Letztlich handelt es sich um fünf erfolgreiche Fälle, durch detaillierte Sequenzanalyse neue Domänen zu entdecken, die von einer wesentlich höheren Anzahl erfolgloser

Versuche begleitet wurden. Drei der fünf entdeckten Proteindomänen sind evolutionär mobile Module, tauchen also in unterschiedlichem Domänenkontext in unterschiedlichen Proteinen auf.

Den Anstoß zum ersten Manuskript [55] lieferte ein experimenteller Befund innerhalb der Firma metaGen. Das Transkript des bislang uncharakterisierten Proteins „Apoptotic Speck-like protein containing a Caspase recruitment domain" (ASC) war in EST-Datenbanken von Brusttumorgeweben weit häufiger repräsentiert als in EST-Datenbanken von normalen Brustgeweben. Die Überexpression der ASC mRNA in Brusttumoren wurde im Labor mittels Dot Blots, RT-PCR und in-situ Hybridisierung bestätigt. Aufgrund dieser Befunde war es wünschenswert, durch eine detaillierte Untersuchung der ASC Proteinsequenz so viel wie möglich über die mutmaßliche Funktion des ASC Proteins zu erfahren. Dies führte schließlich zur Entdeckung der „Domain in Apoptosis and Interferon Response" (DAPIN) als gemeinsames Motiv einer bislang unentdeckten Familie von Wirbeltierproteinen [55]. Diese Proteinfamilie erlangte aufgrund ihrer Verbindung zu inflammatorischen Prozessen und Erbkrankheiten in nur kurzer Zeit eine hohe Aufmerksamkeit [56,57].

Die intensive Suche nach neuen Proteinfamilien durch kontinuierliches Literaturstudium führte zum Spindlin Protein [58]. Spindlin ist während der Meiose mit dem Spindelapparat assoziiert. Während der Oogenese wird Spindlin im Zuge der MAP/Mos-Kinase-Signaltransduktion phosphoryliert, was auf eine wichtige Funktion in der Regulation der Chromosomensortierung während der meiotischen Zellteilung hindeutet [59]. In der vorliegenden Arbeit wurde neben der Vorhersage der Proteinstruktur die Entdeckung diverser neuer Spindlin-ähnlicher Genprodukte in Vertebraten beschrieben [60]. Die kombinierte Analyse der Genstrukturen und Proteinsequenzen gibt Aufschlüsse über die Evolution der Spin/Ssty-Genfamilie. Zudem stellt diese Arbeit die bislang umfangreichste Beschreibung dieser neuen Vertebraten-spezifischen Proteinfamilie dar und kann somit als Referenz dienen.

Im Zuge der Anwendung eines automatischen Sequenzanalyseprotokolls auf alle humanen Proteine mit vorhergesagten Transmembranhelices fiel eine repetitive Struktur in der Ektodomäne des humanen NG2 Proteins auf. Eine nähere Analyse führte zur Entdeckung des „Chondroitinsulfat Proteoglycan" (CSPG) Repeats [61]. Diese repetitive Einheit existiert in Proteinen mit unterschiedlichen Domänenarrangements und kann daher als evolutionär mobiles Modul bezeichnet werden. Für einige bisher uncharakterisierte oder hypothetische Proteine lieferte die Entdeckung von CSPG-Repeats in ihren Sequenzen einen ersten Hinweis auf ihre mögliche zelluläre Funktion. Die Entdeckung des CSPG-Repeats erlaubte zudem die Interpretation bereits publizierter experimenteller Befunde. So konnte mit dem

neuen Wissen über die Domänensubstruktur der NG2-Ektodomäne das bisherige Strukturmodel von NG2 verfeinert werden.

Meine Mitarbeit im Bereich Sequenzanalyse im Rahmen des Europäischen Konsortiums für das Studium von autosomal dominanter lateraler temporaler Epilepsie (ADLTE) führte zur Entdeckung der vierten Proteindomäne, dem Epitempin (EPTP) Repeat [62]. Durch das Konsortium wurden in zwei unabhängigen Familien mit hereditärer Epilepsie Mutationen des humanen Gens LGI1 (Leucine-rich Glioma Inactivated 1) gefunden. Die Mutationen verändern vor allem den bislang uncharakterisierten C-Terminus des LGI1 Proteins, indem sie zu einem verfrühten Abbruch der Proteinsynthese führen. Durch die Analyse der intramolekularen repetitiven Struktur der LGI1-Sequenz und die Entdeckung entfernt verwandter Proteine konnte ein hypothetisches Modell der Struktur des C-Terminus erstellt werden, welches eine Ähnlichkeit zu einer bereits bekannten repetitiven Domänenstruktur aufweist. Der besondere Wert dieser Entdeckung entstand durch die systematische Analyse aller Genloci der Proteine, welche den EPTP-Repeat enthalten. Die chromosomalen Regionen fast aller Genprodukte mit EPTP-Repeats sind mit anderen Epilepsiesubtypen oder neurologischen Krankheiten assoziiert.

Die fünfte Studie zur Entdeckung einer neuen Proteindomäne nahm ihren Anfang in einem gemeinsamen Projekt mit Prof. Dr. Braun über die Zwei-Komponenten-Signaltransduktion durch Histidinkinaserezeptoren in Bakterien. Obwohl man weiß, dass phosphorylierte Histidine einen erheblichen Teil aller phosphorylierten Aminosäuren in Proteinen von Eukaryonten ausmachen, sind die Mechanismen oder Moleküle, die diese Histidin-Phosphorylierungen bewirken, bisher weitgehend unbekannt. Meine initialen Versuche, verschiedene Proteinsequenzdatenbanken von Säugetieren nach bekannten charakteristischen Proteinmodulen aus der bakteriellen Zwei-Komponenten-Signaltransduktion zu durchsuchen, blieben erfolglos. Einige Zeit später führte ich eine Analyse der Proteinsequenz des humanen Gens HIG (hypoxia-inducible gene) durch. Dabei zeigte sich, dass die Familie der HIG-ähnlichen eukaryotischen Proteine eine schwache Ähnlichkeit zu einer Proteinen der NtrY-Subfamilie bakterieller Histidinkinasen aufweist [63]. Diese Studie schildert die nähere Untersuchung einer potentiellen Homologie der beiden Familien, die eine besondere Bedeutung für die Suche nach den Mechanismen der Histidin-Phosphorylierung in Eukaryonten hätte.

1.6 Anwendungen der genomweiten Identifizierung von Proteindomänen in dieser Arbeit

Die Suche nach kurzen Motiven in Proteinsequenzen steht im Mittelpunkt des

sechsten Manuskripts [64]. Ziel dieser Arbeit war die Identifizierung von Immunorezeptor Tyrosin-basierten inhibitorischen Motiven (ITIMs) in einer humanen Proteindatenbank. Das Problem bei der Suche von kurzen Sequenzmotiven in Sequenzdatenbanken ist die Signifikanz eines Treffers. Wegen des geringen Informationsgehalts von kurzen Proteinmotiven ist die Zahl der falsch-positiven Treffer bei Datenbanksuchen sehr hoch. Die Zuverlässigkeit der Vorhersage eines ITIMs sollte erhöht werden, indem zusätzlich die Vorhersage von extrazellulären Domänen, Signalpeptiden und Transmembranhelices, also ein zum ITIM passender Sequenzkontext, gefordert wurde. Den resultierenden humanen ITIM-Proteinen konnten orthologe Proteine der Maus und mRNA-Expressionswerte in humanen Geweben zugeordnet werden, was neue Hypothesen über die Rolle von ITIM-vermittelter Signaltransduktion erlaubte.

Die Aufklärung des Proteoms des humanen Nukleolus mittels moderner massenspektrometrischer Verfahren [14] war die Anregung, eine umfassende Beschreibung der Proteindomänen eines definierten funktionellen Netzwerks von Proteinen, hier des Nukleolus, zu erstellen. Dadurch konnte die wohl umfangreichste Beschreibung des Proteindomänenrepertoires eines bestimmten zellulären Kompartments durchgeführt werden [65]. Die Präsenz der einzelnen Proteindomänen des Nukleolus in den Proteomen von verschiedenen Archaebakterien, Eubakterien und Eukaryonten in Zusammenhang mit ihrer biochemischen Funktion lieferte neue Hinweise, wie sich die Evolution des Nukleolus gestaltet haben könnte.

1.7 Referenzen der Einleitung

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM and others. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269(5223):496-512.
2. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM and others. The minimal gene complement of Mycoplasma genitalium. Science 1995;270(5235):397-403.
3. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD and others. Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. Science 1996;273(5278):1058-1073.
4. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M and others. Life with 6000 genes. Science 1996;274(5287):546, 563-547.

5. Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium. Science 1998;282(5396):2012-2018.

6. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF and others. The genome sequence of Drosophila melanogaster. Science 2000;287(5461):2185-2195.

7. Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, Toyoda A, Ishii K, Totoki Y, Choi DK and others. The DNA sequence of human chromosome 21. Nature 2000;405(6784):311-319.

8. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W and others. Initial sequencing and analysis of the human genome. Nature 2001;409(6822):860-921.

9. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA and others. The sequence of the human genome. Science 2001;291(5507):1304-1351.

10. Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat Biotechnol 1997;15(13):1359-1367.

11. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS. A gene expression map for Caenorhabditis elegans. Science 2001;293(5537):2087-2092.

12. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A and others. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A 2002;99(7):4465-4470.

13. Rappsilber J, Ryder U, Lamond AI, Mann M. Large-scale proteomic analysis of the human spliceosome. Genome Res 2002;12(8):1231-1245.

14. Andersen JS, Lyon CE, Fox AH, Leung AK, Lam YW, Steen H, Mann M, Lamond AI. Directed proteomic analysis of the human nucleolus. Curr Biol 2002;12(1):1-11.

15. Doolittle RF. On the trail of protein sequences. Bioinformatics 2000;16(1):24-33.

16. Doolittle RF. Some reflections on the early days of sequence searching. J Mol Med 1997;75(4):239-241.

17. Doolittle RF. Do you dig my groove? Nat Genet 1999;23(1):6-8.

18. Hill EE, Morea V, Chothia C. Sequence conservation in families whose members have little or no sequence similarity: the four-helical cytokines and cytochromes. J Mol Biol 2002;322(1):205-233.

19. Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C. The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli. J Mol Biol 2001;311(4):693-708.

20. Chothia C, Gough J, Vogel C, Teichmann SA. Evolution of the protein repertoire. Science 2003;300(5626):1701-1703.

21. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature 2002;420(6912):218-223.

22. Doolittle RF, Bork P. Evolutionarily mobile modules in proteins. Sci Am 1993;269(4):50-56.

23. Gilbert W. Why genes in pieces? Nature 1978;271(5645):501.

24. Patthy L. Exon shuffling and other ways of module exchange. Matrix Biol 1996;15(5):301-310; discussion 311-302.

25. Patthy L. Protein Evolution.: Blackwell Science Ltd.; 1999.

26. Roy SW, Lewis BP, Fedorov A, Gilbert W. Footprints of primordial introns on the eukaryotic genome. Trends Genet 2001;17(9):496-501.

27. Wolf YI, Kondrashov FA, Koonin EV. No footprints of primordial introns in a eukaryotic genome. Trends Genet 2000;16(8):333-334.

28. Wolf YI, Kondrashov FA, Koonin EV. Footprints of primordial introns on the eukaryotic genome: still no clear traces. Trends Genet 2001;17(9):499-501.

29. Gracy J, Argos P. Automated protein sequence database classification. II. Delineation Of domain boundaries from sequence similarities. Bioinformatics 1998;14(2):174-187.

30. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D. ProDom: automated clustering of homologous domains. Brief Bioinform 2002;3(3):246-251.

31. Heger A, Holm L. Picasso: generating a covering set of protein family profiles. Bioinformatics 2001;17(3):272-279.

32. Chothia C. Proteins. One thousand families for the molecular biologist. Nature 1992;357(6379):543-544.

33. Zhang C, DeLisi C. Estimating the number of protein folds. J Mol Biol 1998;284(5):1301-1305.

34. Wang ZX. A re-estimation for the total numbers of protein folds and superfamilies. Protein Eng 1998;11(8):621-626.

35. Govindarajan S, Recabarren R, Goldstein RA. Estimating the total number of protein folds. Proteins 1999;35(4):408-414.

36. Wolf YI, Grishin NV, Koonin EV. Estimating the number of protein folds and families from complete genome data. J Mol Biol 2000;299(4):897-905.

37. Goh CS, Lan N, Echols N, Douglas SM, Milburn D, Bertone P, Xiao R, Ma LC, Zheng D, Wunderlich Z and others. SPINE 2: a system for collaborative structural proteomics within a federated database framework. Nucleic Acids Res 2003;31(11):2833-2838.

38. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins 2001;Suppl 5:2-7.

39. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299(2):499-520.

40. Gough J, Chothia C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res 2002;30(1):268-272.

41. Karplus K, Barrett C, Cline M, Diekhans M, Grate L, Hughey R. Predicting protein structure using only sequence information. Proteins 1999;Suppl 3:121-125.

42. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 1998;284(4):1201-1210.

43. Madera M, Gough J. A comparison of profile hidden Markov model procedures for remote homology detection. Nucleic Acids Res 2002;30(19):4321-4328.

44. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 2001;313(4):903-919.

45. Huynen M, Snel B, Lathe W, 3rd, Bork P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 2000;10(8):1204-1210.

46. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 1999;96(8):4285-4288.

47. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. Nature 1999;402(6757):83-86.

48. Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 1998;23(9):324-328.

49. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 1999;96(6):2896-2901.

50. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature 1999;402(6757):86-90.

51. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P and others. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000;403(6770):623-627.

52. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M. Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet 2002;18(10):529-536.

53. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD and others. Functional discovery via a compendium of expression profiles. Cell 2000;102(1):109-126.

54. Jansen R, Lan N, Qian J, Gerstein M. Integration of genomic datasets to predict protein complexes in yeast. J Struct Funct Genomics 2002;2(2):71-81.

55. Staub E, Dahl E, Rosenthal A. The DAPIN family: a novel domain links apoptotic and interferon response proteins. Trends Biochem Sci 2001;26(2):83-85.

56. Martinon F, Burns K, Tschopp J. The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proIL-beta. Mol Cell 2002;10(2):417-426.

57. Mariathasan S, Vucic D. POPing the fire into the pyrin? Biochem J 2003;373(Pt 1):1-2.

58. Oh B, Hwang SY, Solter D, Knowles BB. Spindlin, a major maternal transcript expressed in the mouse during the transition from oocyte to embryo. Development 1997;124(2):493-503.

59. Oh B, Hampl A, Eppig JJ, Solter D, Knowles BB. SPIN, a substrate in the MAP kinase pathway in mouse oocytes. Mol Reprod Dev 1998;50(2):240-249.

60. Staub E, Mennerich D, Rosenthal A. The Spin/Ssty repeat: a new motif identified in proteins involved in vertebrate development from gamete to embryo. Genome Biol 2002;3(1):RESEARCH0003.

61. Staub E, Hinzmann B, Rosenthal A. A novel repeat in the melanoma-associated chondroitin sulfate proteoglycan defines a new protein family. FEBS Lett 2002;527(1-3):114-118.

62. Staub E, Perez-Tur J, Siebert R, Nobile C, Moschonas NK, Deloukas P, Hinzmann B. The novel EPTP repeat defines a superfamily of proteins implicated in epileptic disorders. Trends Biochem Sci 2002;27(9):441-444.

63. Staub E, Braun T. Sequence similarity between prokaryotic nitrogen-sensing histidine kinases and vertebrate hypoxia-inducible proteins. Cellular Signalling. Submitted. 2003.

64. Staub E, Rosenthal A, Hinzmann B. Systematic identification of immunoreceptor tyrosine-based inhibitory motifs (ITIMs) in the human proteome. Cellular Signalling. In Press. Online publication since Oct, 30th 2003 2003.

65. Staub E, Fiziev P, Rosenthal A, Hinzmann B. Insights into the evolution of the nucleolus by an analysis of its protein domain repertoire. BioEssays. Accepted for publication. 2003.

2 The DAPIN family: a novel domain links apoptotic and interferon response proteins

Die folgende Erklärung legt die individuellen Beiträge der Co-Autoren zum nachfolgenden Manuskript dar. Von den Co-Autoren unterschriebene Fassungen dieser Erklärung liegen dieser Dissertation bei. Die Erklärungen sollen belegen, dass die erzielten Resultate im wesentlichen auf meiner Arbeit beruhen und ihre Verwendung innerhalb dieser Dissertation gerechtfertigt ist.

The DAPIN manuscript: declaration about contributions of authors

Hereby I, co-author of the manuscript *„The DAPIN family: a novel domain links apoptotic and interferon response proteins“* which appeared in *Trends in Biochemical Sciences (2001) vol.26, no.2, pp.83-85* declare my contributions to the results and to the preparation of the manuscript. Furthermore, I agree that the statements below describe the contributions of the other co-authors correctly.

1) Eike Staub
- conducted the protein sequence analysis,
- discovered and characterised the domain,
- predicted the death domain-like structure,
- inferred the function of the domain,
- wrote the text and prepared the figures for the manuscript,
- served as the corresponding author during the review process.

2) Edgar Dahl
- raised the interest in the sequence analysis of the ASC protein by his initial observation of the differential expression of the ASC transcript in breast tumors,
- contributed to final stages of manuscript preparation by helpful comments on the style of the manuscript.

3) André Rosenthal
- was the supervisor of the project,
- contributed to final stages of manuscript preparation by helpful comments on the style of the manuscript.

The DAPIN family: a novel domain links apoptotic and interferon response proteins

Eike Staub, Edgar Dahl and André Rosenthal

We report the discovery of a protein domain, hereafter referred to as DAPIN, in diverse vertebrate and viral proteins that is associated with tumor biology, apoptosis and inflammation. Based on a secondary structure prediction, we suggest an all-α fold for DAPIN, which is also adopted by apoptotic protein domains of the CARD, death domain and death effector domain type.

Using an EST data-mining approach to search for genes that are differentially expressed between normal and cancerous breast tissue[1], we identified a gene which was recently named ASC (after apoptosis-associated speck-like protein containing a CARD). This name was given because the protein precipitates with monoclonal antibodies that target apoptotic speck-like bodies[2]. The ASC protein sequence contains a C-terminal caspase recruitment domain (CARD), which is an adapter domain in apoptotic proteins (e.g. RAIDD, ICH-1, Ced-3, ICE) that binds to other CARDs via homophilic interactions[3]. The presence of a CARD suggests that ASC is involved in apoptosis.

The N-terminal sequence of ASC is similar to that of the Mediterranean fever protein pyrin. Mutations in the pyrin-coding *MEFV* gene are the cause for familial Mediterranean fever (FMF), an autosomal recessive disease characterized mainly by recurrent attacks of fever and serositis[4,5]. Eleven of 16 mutations that contribute to the disease phenotype are located in the B30.2 domain[6], with none in the N-terminal region[7]. Pyrin localizes to the perinuclear cytoplasm and interacts with a putative Golgi transport protein[8].

Using the similarity between ASC and pyrin at the N-terminus as a starting point, we identified a further 11 proteins with similar N-termini by an iterative process that involved construction of multiple alignments with CLUSTAL X, building of hidden Markov models (HMM) and protein database scanning[9,10]. The final manually edited alignment (Fig. 1) resulted in a HMM that detected each family member with an expectation value E of $<1 \times 10^{-38}$. We named this new protein domain DAPIN (domain in apoptosis and interferon response).

The DAPIN seems to be unique to vertebrates or vertebrate-specific viruses, as no similarity to any known or predicted protein from fly, worm, yeast or bacteria could be detected. The 3′ ends of the first coding exons of the pyrin, IFI 16 (interferon inducible gene 16)[4,11] and *MNDA* genes (www.ensembl.org; Gene ID, ENSG00000073840) are consistent with the C-terminal boundaries of the alignment. Secondary structure prediction with JNet (Ref. 12) suggests an all-α structure with five α-helices. Thus, DAPIN resembles apoptotic adapter domains, the CARD, the death domain and the death effector domain, which all fold into six α-helices[13–15]. This similarity might indicate a common evolutionary origin for the four domains. The combination of DAPIN and CARD in two different proteins (Fig. 2) might be the product of domain duplication and divergent evolution.

The best-characterized DAPIN family members originate from gene clusters of interferon-inducible genes on human chromosome 1q22 and the syntenic mouse region. They code for the HIN-200 family of hematopoetic interferon-inducible nuclear proteins[16,17]. All proteins translocate to the nucleus after induction. Single family members differ in the dependence of their induction on distinct interferon subtypes and in their expression pattern among the different hematopoetic cell lineages. This family includes AIM2 (absent in melanoma 2), a possible tumor suppressor that was discovered in a screen for differentially expressed genes between malignant and benign melanoma cells[18].

The structural feature common to all members of the HIN-200 family is the presence of one or two copies of a 200-amino-acid domain[16,17,19] (HIN-200-aa in Fig. 2). Furthermore, a high degree of amino acid similarity in the N-terminal DAPIN region was described for the HIN-200 family members, except for interferon-inducible protein 202 (IFI 202). Attention was drawn to an 'imperfect' Leu zipper motif[19] (compare the alignment position 72–94 of proteins 1–6, Fig. 1) with Leu residues being partly replaced by Val, Ile and Met. Our secondary structure prediction suggests two α-helices in this sequence region and therefore contradicts the assumption of a Leu-zipper

http://tibs.trends.com 0968-0004/01/$ – see front matter PII: S0968-0004(00)01717-5

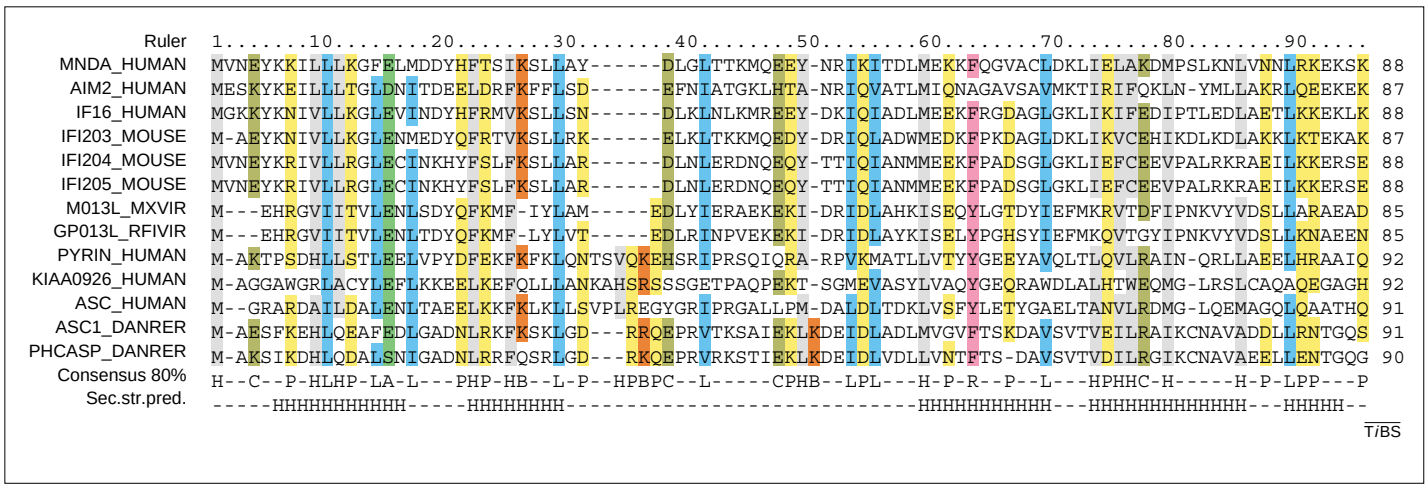

Fig. 1. Multiple alignment of DAPIN (domain in apoptosis and interferon response) regions in 13 family members. The alignment was constructed using the CLUSTAL X program with subsequent manual editing. The amino terminal Met was not used for hidden Markov model (HMM) construction. The numbers on the far right-hand side indicate the position of the DAPIN residue in the sequence that is closest to the C-terminus. The amino acid residues are colored according to an 80% consensus: P, polar (DEHKNQRST; yellow); C, charged (DEHKR; brown); A, negative (ED; green); B, positive (HRK; red); R, aromatic (FHWY; purple); L, aliphatic (ILV; blue); H, hydrophobic (ACFGHILMVWY; grey). Secondary structure prediction (sec. str. pred.): H, α-helix; E, β-strand. The protein names consist of gene and organism identifier (DANRER, danio rerio; MXVIR, myxoma virus; MNDA, myeloid nuclear differentiation antigen; RFIVIR, rabbit fibroma virus; PHCASP, pyrin-homolog caspase). GenBank identifiers: MNDA_HUMAN, 730038; AIM2_HUMAN, 2558942; IFI16_HUMAN, 184569; IFI203_MOUSE, 6016336; IFI204_MOUSE, 124489; IFI205_MOUSE, 2833215; M013L_MXVIR, 6523868; GP013L_RFIVIR, 6578691; PYRIN_HUMAN, 4557743; KIAA0926_HUMAN, 4589484; ASC_HUMAN, 6482372; ASC1_DANRER, 7673624; PHCASP_DANRER, 7673640.

conformation. Deletion mutant studies have shown that homodimerization of MNDA depends on amino acids 52–82 in the DAPIN region[20]. IFI 16 is a transcriptional repressor and its first 159 amino acids are sufficient to bind DNA, either directly or indirectly[21]. The abundance of basic amino acids in the DAPIN region seems to be a special feature of the HIN-200 subfamily and might facilitate direct or indirect DNA binding. However, we do not assume a DNA-binding function, but a protein–protein interaction function for the DAPIN domain based on the predicted secondary structure similarity to other apoptotic adapter domains and on the MNDA dimerization studies.

Another DAPIN protein, the large predicted human KIAA0926 protein, has been presented as a member of the novel NACHT (after NAIP, CIIA, HET-E and TP1 proteins) NTPase family[22], which suggests a function in apoptosis or activation of major histocompatibility complex (MHC)-class II transcription. In addition, KIAA0926 contains a CARD on its C-terminus. The domain architecture and size of this protein support the assumption that KIAA0926 plays an important role in the control of apoptosis, possibly as a docking platform for other proteins.

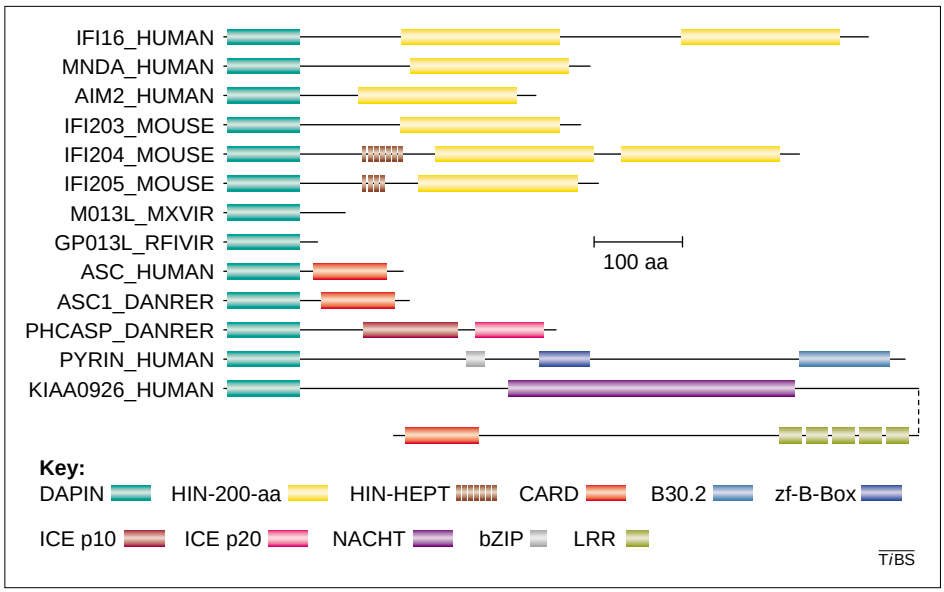

Fig. 2. Domain architecture of the DAPIN proteins (scale is approximate). Myeloid nuclear differentiation antigen (MNDA), AIM2 (absent in melanoma 2), IFI 16, IFI 203, IFI 204 and IFI 205 (where IFI is the interferon inducible protein) belong to the HIN-200 family. The domain key is given underneath. Abbreviations: ASC, apoptosis-associated speck-like protein containing a CARD; B30.2, named after the exon B30.2, identified in a search for coding sequences in the major histocompatibility complex (MHC)-class I region; bZIP, basic leucine zipper domain; CARD, caspase recruitment domain; DAPIN, domain in apoptosis and interferon response; HIN-200aa, characteristic 200-amino-acid domain for HIN-200 family (hematopoetic interferon-inducible proteins with a 200-amino-acid repeat); HIN-HEPT, heptamer repeats from HIN-200 proteins IFI 204 and IFI 205; ICE p10/ICE p20, interleukin-1β-converting enzyme p10 or p20 domain; LRR, Leu-rich repeats; NACHT, after NAIP, CIIA, HET-E and TP1 proteins; zf-B-Box, B-box zinc finger domain.

Two open-reading frames, named M013L and GP013L from the genomes of rabbit viruses, myxoma and fibroma virus[23,24], encode highly similar proteins that consist almost entirely of the DAPIN domain. This supports the idea that the DAPIN exon is a functionally independent module. Myxoma virus infection causes a rapid systemic infection in European rabbits and is almost always lethal, whereas fibroma virus causes a benign fibroma at the site of invasion. In adult hosts an adaptive immune response to fibroma virus is able to cause tumor shrinking and virus elimination. M013L and GP013L are novel candidates for viral proteins that tackle the host immune response. We hypothesize that this is achieved by the formation of nonfunctional heterodimers with cellular DAPIN proteins, thereby interfering with programmed cell death in response to virus infection.

Recently, two zebrafish genes that encode other DAPIN members have been discovered[25]. One is a predicted

ortholog of the human ASC protein; the other is an ICE (interleukin-1β-converting enzyme)-like protease. The combination of the DAPIN with ICE-like protease p10 and p20 domains is consistent with the assumption of an adapter function for DAPIN that complements the protease effector function in this novel caspase.

The identification of a common DAPIN domain links a well-characterized family of nuclear interferon-inducible proteins to other proteins with putative functions in apoptosis and tumor biology, viral infection and inflammation. Future research on DAPIN proteins should reveal the physiological and biochemical function of this domain and the small viral DAPIN proteins might be especially helpful for this task.

Acknowledgements

We would like to thank an anonymous reviewer for his / her helpful suggestions.

References

1 Schmitt, A.O. *et al.* (1999) Exhaustive mining of EST libraries for genes differentially expressed in normal and tumor tissues. *Nucleic Acids Res.* 27, 4251–4260
2 Masumoto, J. *et al.* (1999) ASC, a novel 22-kDa protein, aggregates during apoptosis of human promyelocytic leukemia HL-60 cells. *J. Biol. Chem.* 274, 33835–33838
3 Hofmann, K. *et al.* (1997) The CARD domain: a new apoptotic signalling motif. *Trends Biochem. Sci.* 22, 155–156
4 The French FMF Consortium (1997) A candidate gene for familial Mediterranean fever. *Nat. Genet.* 17, 25–31
5 The International FMF Consortium (1997) Ancient missense mutations in a new member of the RoRet gene family are likely to cause familial Mediterranean fever. *Cell* 90, 797–807
6 Vernet, C. *et al.* (1993) Evolutionary study of multigenic families mapping close to the human MHC class I region. *J. Mol. Evol.* 37, 600–612
7 Mulley, J.C. (1999) The genetic basis for periodic fever. *Am. J. Hum. Genet* 64, 939–942
8 Chen, X. *et al.* (2000) The familial mediterranean fever protein interacts and colocalizes with a putative golgi transporter. *Proc. Soc. Exp. Biol. Med.* 224, 32–40
9 Jeanmougin, F. *et al.* (1998) Multiple sequence alignment with CLUSTAL X. *Trends Biochem. Sci.* 23, 403–405
10 Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics* 14, 755–763
11 Trapani, J.A. *et al.* (1994) Genomic organization of IFI16, an interferon-inducible gene whose expression is associated with human myeloid cell differentiation: correlation of predicted protein domains with exon organization. *Immunogenetics* 40, 415–424
12 Cuff, J.A. *et al.* (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 40, 502–511
13 Chou, J.J. *et al.* (1998) Solution structure of the RAIDD CARD and model for CARD/CARD interaction in caspase-2 and caspase-9 recruitment. *Cell* 94, 171–180
14 Eberstadt, M. *et al.* (1998) NMR structure and mutagenesis of the FADD (Mort1) death-effector domain. *Nature* 392, 941–945
15 Huang, B. *et al.* (1996) NMR structure and mutagenesis of the Fas (APO-1/CD95) death domain. *Nature* 384, 638–641
16 Dawson, M.J. *et al.* (1996) HIN-200: a novel family of IFN-inducible nuclear proteins expressed in leukocytes. *J. Leukoc. Biol.* 60, 310–316
17 Landolfo, S. *et al.* (1998) The Ifi 200 genes: an emerging family of IFN-inducible genes. *Biochimie* 80, 721–728
18 DeYoung, K.L. *et al.* (1997) Cloning a novel member of the human interferon-inducible gene family associated with control of tumorigenicity in a model of human melanoma. *Oncogene* 15, 453–457
19 Johnstone, R.W. *et al.* (1999) Transcription and growth regulatory functions of the HIN-200 family of proteins. *Mol. Cell Biol.* 19, 5833–5838
20 Xie, J. *et al.* (1997) MNDA dimerizes through a complex motif involving an N-terminal basic region. *FEBS Lett.* 408, 151–155
21 Johnstone, R.W. *et al.* (1998) The human interferon-inducible protein, IFI 16, is a repressor of transcription. *J. Biol. Chem.* 273, 17172–17177
22 Koonin, E.V. *et al.* (2000) The NACHT family – a new group of predicted NTPases implicated in apoptosis and MHC transcription activation. *Trends Biochem. Sci.* 25, 223–224
23 Cameron, C. *et al.* (1999) The complete DNA sequence of myxoma virus. *Virology* 264, 298–318
24 Willer, D.O. *et al.* (1999) The complete genome sequence of shope (rabbit) fibroma virus. *Virology* 264, 319–343
25 Inohara, N. *et al.* (2000) Genes with homology to mammalian apoptosis regulators identified in zebrafish. *Cell Death Differ.* 7, 509–510

E. Staub*
E. Dahl
A. Rosenthal
metaGen Gesellschaft für Genomforschung mbH, 14195 Berlin, Germany.
*e-mail: eike.staub@metagen.de

3 The Spin/Ssty repeat: a new motif in proteins involved in vertebrate development from gamete to embryo

Die folgende Erklärung legt die individuellen Beiträge der Co-Autoren zum nachfolgenden Manuskript dar. Von den Co-Autoren unterschriebene Fassungen dieser Erklärung liegen dieser Dissertation bei. Die Erklärungen sollen belegen, dass die erzielten Resultate im wesentlichen auf meiner Arbeit beruhen und ihre Verwendung innerhalb dieser Dissertation gerechtfertigt ist.

The Spin/Ssty manuscript: declaration about contributions of authors

Hereby I, co-author of the manuscript *„The Spin/Ssty repeat: a new motif identified in proteins involved in vertebrate development from gamete to embryo“* which appeared in *Genome Biology (2002) vol.3, no.1, pp.RESEARCH0003.1* declare my contributions to the results and to the preparation of the manuscript. Furthermore, I agree that the statements below describe the contributions of the other co-authors correctly.

1) Eike Staub
- conducted the protein sequence analysis,
- defined the gene models of the genes that were discovered from genome sequence data,
- discovered and characterised the repeat,
- predicted the β-sheet structure of the repeat,
- performed the phylogenetic analysis,
- wrote the text and prepared the figures for the manuscript,
- served as the corresponding author during the review process.

2) Detlev Mennerich
- raised the interest in the sequence analysis of the SPIN-like proteins,
- contributed to the final stages of manuscript preparation by helpful comments on the style of the manuscript.

3) André Rosenthal
- was the supervisor of the project,
- contributed to the final stages of manuscript preparation by helpful comments on the style of the manuscript.

Research

The Spin/Ssty repeat: a new motif identified in proteins involved in vertebrate development from gamete to embryo

Eike Staub, Detlev Mennerich and André Rosenthal

Address: metaGen Pharmaceuticals GmbH, Oudenarderstrasse 16, D-13347 Berlin, Germany.

Correspondence: Eike Staub. E-mail: eike.staub@metagen.de

Published: 7 December 2001

Genome ***Biology*** 2001, **3(1)**:research0003.1–0003.6

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/3/1/research/0003

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 19 September 2001
Revised: 10 October 2001
Accepted: 23 October 2001

Abstract

Background: The homologous genes *Spin* (*spindlin*) and *Ssty* were first identified as genes involved in gametogenesis and seem to occur in multiple copies in vertebrate genomes. The mouse spindlin (Spin) protein was reported to interact with the spindle apparatus during oogenesis and to be a target for cell-cycle-dependent phosphorylation. The transcript of the mouse *Ssty* gene is specific to sperm cells. In the chicken, spindlin was found to co-localize with SUMO-1 to nuclear dots during interphase in fibroblasts, but to co-localize with chromosomes during mitosis. Thus, *Spin*/*Ssty* genes might be important in the transition from sperm cells and oocytes to the early embryo, as well as in mitosis.

Results: Here we report the discovery of a new protein motif of around 50 amino acids in length, the Spin/Ssty repeat, in proteins of the Spin/Ssty (spindlin) family. We found that in one member of this family, the human *SPIN* gene, each repeat resides in its own exon, supporting our view that Spin/Ssty repeats are independent functional units. On the basis of different secondary-structure prediction methods, we propose a four-stranded β-structure for the Spin/Ssty repeat.

Conclusions: The discovery of the Spin/Ssty repeat might contribute to the further elucidation of the structure and function of spindlin-family proteins. We predict that the tertiary structure of spindlin-like proteins is composed of three modules of Spin/Ssty repeats.

Background

During early oocyte development, the transcription of maternal genes ceases with the onset of meiosis. After fertilization and zygote formation, transcription of the embryonic genome starts at the two-cell stage or later, depending on the organism [1-3]. Thus, the amount of maternal mRNAs must be sufficient to drive the gamete through meiosis, fertilization and through the first zygotic cell division - a time span of almost 2 days in mice [1]. During this period the activation of translation from many different deadenylated, and thus dormant, mRNAs is controlled by their cytoplasmic polyadenylation [1,4].

In these early phases of mouse development, one of the most frequent transcripts regulated in this manner is that of the *spindlin* (*Spin*) gene [1,5]. The protein encoded by *Spin* is a meiotic-spindle-associated protein specific to the oocyte [1,5], that is phosphorylated during meiosis [6,7]. Oh *et al.* showed that phosphorylation modulates the ability of the Spin protein to interact with the spindle apparatus during oogenesis [6]. Phosphorylation is dependent on the Mos/MAP kinase pathway, which is controlled by meiotic-checkpoint proteins cyclin B and Cdc2 in *Xenopus laevis* oocytes [6,8]. Sequence similarity and mRNA expression suggest that a complementary role in sperm development

seems to be fulfilled by the gene *Ssty* (Y-linked spermiogenesis specific transcript), a multicopy testis-specific spermatogenesis gene on the mouse Y chromosome long arm [9]. In contrast to the oocyte-specific expression of *Spin*, the *Ssty* mRNA is specifically expressed in sperm cells [9]. Dosage reduction by partial deletion of *Ssty* genes was suggested to cause deformed sperm heads and infertility [10,11]. However, reports on *Ssty* expression on the protein level are still lacking. Recently, two *Spin*-type genes from the chicken, *Gallus gallus*, have been cloned - *Spin-W* and *Spin-Z*, located on the W and Z sex chromosomes, respectively [12]. They are nearly identical to each other in their coding regions, and both were reported to be expressed in early embryos, but *Spin-Z* is also expressed in various adult tissues. Transfection of fibroblasts with DNA expressing fluorescent protein-tagged chSpin-W and the small ubiquitin-related modifier SUMO-1 showed the co-localization of these proteins in nuclear dots during interphase. Localization was shown to depend on the carboxy-terminal 30 amino acids of chSpin-W, especially on the presence of two phenylalanines in positions 244 and 247. However, SUMO-1 and chSpin-W could not be shown to interact directly. In contrast to its interphase localization, the red fluorescent protein-chSpinW fusion associated with chromosomes during mitosis. Although experimental results indicate that the spindlin protein family includes important players in meiosis and early embryogenesis, as well as in mitosis, their biochemical function is largely unknown.

Results and discussion

Repeat identification and analysis

At the beginning of our analysis, pairwise sequence similarity among proteins of the spindlin family was already public knowledge, with the reported average sequence identity between members being approximately 70% (entry PF02513 (Spin/Ssty protein family) in the Pfam 6.2 protein database). When we tried to identify additional family members of this protein family by scanning the NCBI nonredundant protein database (nr) using BLASTP and the human Spin protein sequence (GenBank RefSeq identifier NP_006708) as a query, we noticed a second high-scoring segment pair in the hit of the human Spin sequence with itself. Therefore we scanned the human Spin sequence for internal repeats with the program dotter and found a triple repeat spanning nearly the complete protein sequence. We aligned the repeats using CLUSTALX and corrected the alignment manually for subsequent construction of a hidden Markov model (HMM). By scanning the nr database with this model we identified the repeat in open reading frames (ORFs) of other known members of the Spin/Ssty gene family with expectation (E) values below 1e-9. Among these, we detected three repeats of typical length of 53 amino acids in the ORF of mouse *Ssty*, encompassing the two smaller 71 base-pair (bp) repeats that were previously noticed at the cDNA level [9]. Spindlin-family protein sequences in the nr database are from human, mouse and chicken. Among the human and mouse sequences, many were hypothetical protein sequences translated from genomic or cDNA sequences. These sequences were too similar at the protein level to conclude that they derive from different genes. To determine the number of Spin/Ssty-like genes for *Mus musculus* and *Homo sapiens*, we decided to isolate an initial redundant set of possible transcripts on the basis of the human and mouse RefSeq and UniGene databases and the database of confirmed peptides of the Ensembl human genome annotation project (Version 1.1.3), and finally to reduce the redundancy of identified transcripts by thorough sequence comparison. We identified the initial set of Spin/Ssty-like transcripts in these databases by TBLASTN searches using known spindlin-family protein sequences as queries.

For *H. sapiens,* we detected four different genes of the Spin/Ssty family. According to Ensembl, the chromosomal region Xp11.1 contains two *SPIN*-like genes: one coding for a *spindlin*-like transcript (Ensembl: ENST00000218159; RefSeq: NM_019003.1; UniGene: Hs.2294334; GenBank-Clone: Z82211) and a second in close proximity, which was named *spindlin-like 2* (Ensembl: ENST00000252781; GenBankClone: Z82211). These transcripts are 99.7% identical to each other at the nucleotide level in their protein-coding regions and were first described by Laval *et al.* as members of the human X-linked DXF34 sequence family [13]. Another SPIN-family gene resides on chromosome Xq12 (Ensembl: ENST00000253399). The best characterized family member, the human *SPIN* gene (Ensembl: ENST00000223559; RefSeq: NM_006717.1; UniGene: Hs.3335321; GenBank-Clone: AL353748) is located on chromosome 9q22.2 and comprises three exons.

For *M. musculus*, scanning the RefSeq and UniGene resources revealed three Spin/Ssty-like transcripts with complete coding regions. The known *Spin* gene (RefSeq: NM_011462.1; UniGene: Mm.S939555) and the *Ssty* gene (also called *Smy*; RefSeq: NM_009220.1; UniGene: Mm.S936711) are around 70% identical on the protein level. A novel 1,056 bp cDNA (RefSeq: NM_023546.1; UniGene: Mm.S1997937) seems to encode a complete spindlin family protein with around 80% protein sequence identity to Ssty. Other mouse transcripts that could potentially encode complete proteins of the spindlin family seem to exist, as there are 11 additional independent cDNA assemblies in UniGene (Mm.S1975038, Mm.S1922195, Mm.S499811, Mm.S227336, Mm.S1973836, Mm.S707442, Mm.S781768, Mm.S502745, Mm.S782972, Mm.S778767, Mm.S787945). Their ORFs are interrupted or incomplete, however. Increased expressed sequence tag (EST) coverage and quality of these assemblies might reveal more functional spindlin family members. The high number of SPIN-like transcripts in mice is in agreement with previous reports [11,13] that presented evidence for the existence of a multi-copy *Ssty*-like gene family on the mouse Y chromosome. As three of four human Spin/Ssty-like genes

consist of a single exon, and alternative transcripts of the human triple-exon gene *SPIN* have not yet been reported, alternative splicing is unlikely to contribute to the diversity of *Spin/Ssty* transcripts in mouse.

To identify Spin/Ssty-family genes from other organisms, we scanned the dbEST database using the TBLASTN program and known spindlin-family proteins as queries. We found additional ESTs in several organisms. We assembled ESTs from *Bos taurus* (GenBank AV588979, AV588980, BE667003, BF045945), determined the full coding region by alignment with the human Spin protein sequence and added the Spin/Ssty repeat regions to the repeat alignment (Figure 1). Furthermore, we detected Spin/Ssty repeats in several single ESTs that represent fragments of putative Spin/Ssty-family genes. However, we were not able to obtain full coding regions by assembling these ESTs. Among them were several ESTs from *Rattus norvegicus*, an EST from *X. laevis* (GenBank BG018656), two ESTs from *Oryzias latipes* (GenBank AU169984, AU178597) and one EST from *Danio rerio* (GenBank AW077586), indicating the existence of Spin/Ssty repeats in fish and frog proteins. We did not detect Spin/Ssty repeats in the proteomes of *Drosophila melanogaster* or *Caenorhabditis elegans*. Thus, Spin/Ssty repeats are currently restricted to vertebrate proteins.

The subsequent analysis of Spin/Ssty repeats is exclusively based on repeats from known proteins or complete ORFs, in order to exclude low-quality sequences from the analysis. To include Spin/Ssty repeats from a fish protein, an exception is made for the *O. latipes* EST AU169984, which contains an incomplete ORF comprising two complete Spin/Ssty repeats without interruption by frameshift errors.

Using our initial HMM we identified three repeats per protein (two for the incomplete *O. latipes* protein) with E values below 1e-15. We aligned the repeats (Figure 1) and constructed three HMMs: two by using only repeats with less than 75 and 90% pairwise sequence identity, another by using all repeats in the seed alignment. All HMMs re-identified the repeats with E values below 1e-22. However, scanning the nr database with these new models did not identify further Spin/Ssty repeats. We submitted a description and an alignment of the Spin/Ssty repeat to Pfam (Pfam 6.6: PF02513), which replaced the previous Spin/Ssty protein family entry.

For single combinations of Spin/Ssty repeats, the pairwise sequence identity drops below 15%. To test the significance of the similarity among the repeat subtypes (amino-terminal, central, carboxy-terminal) and to exclude HMM training artifacts, we carried out a cross-validation test. We constructed HMMs for each repeat subtype and tried to detect the repeats of the remaining subtypes. For this approach we used five nonredundant proteins (gg_SPINZ, bt_SPINH, hs_SPINX2, mm_SSTY, mm_SPINL; Figure 1). We could identify the complete set of repeats from the five proteins with E values below 5e-3 and thus confirmed that the subgroups are evolutionarily related.

Phylogenetic analysis of the Spin/Ssty repeats from the five nonredundant proteins with the neighbor-joining method after removal of gapped alignment columns confirmed the existence of three subtypes of repeats, the amino-terminal, central and carboxy-terminal subtype (Figure 2). In the genomic structure of the human *SPIN* gene on chromosome 9 each Spin/Ssty repeat resides in its own exon, supporting our view that the Spin/Ssty repeats represent structural or functional units. In summary, the phylogenetic analysis and the gene structure of the *SPIN* gene suggest that the first spindlin-family protein evolved by subsequent duplications of an ancient exon and that these duplications preceded the speciation events leading to birds and mammals.

Structure prediction

We made secondary-structure predictions using several programs via the Jpred[2] server with the alignment of the whole family and the alignments of each of the amino-terminal, central and carboxy-terminal repeat subfamilies as a query. The consensus prediction for the whole alignment suggests four β strands for the Spin/Ssty repeat. Although the isolated central Spin/Ssty repeat is predicted to form an α helix in exchange for the second β strand, the single predictions for the amino- and carboxy-terminal repeat subtypes are in agreement with the prediction based on the whole family. Because in most cases the accuracy of secondary-structure predictions is higher when alignments of more diverse protein-family members are used, we believe that the predictions based on the whole family are the most reliable, and we suggest an all-β structure with four β strands for all Spin/Ssty repeats. Attempts to assign a known protein fold to the Spin/Ssty repeat using different fold-prediction methods via the Structure Prediction Meta Server did not lead to significant predictions.

Conclusions

Our findings might serve as a basis for future work on this new class of repeats. The Spin/Ssty repeat alignment will assist in detecting further family members in other species and in the search for an evolutionary origin of the spindlin family of proteins. The detection of Spin/Ssty repeats in proteins with other domain architectures might provide a clue to the function of the spindlin family. Knowledge of the repeat structure of spindlin-like proteins can support further experimental work. Once interaction partners or biochemical functions are identified for the spindlin-like proteins, hypotheses based on the repeat architecture can be generated for further experiments: site-directed mutagenesis studies, that are targeted on conserved residues, are most likely to disrupt the structure or destroy the function of a protein; attempts to delete certain regions of spindlin-family proteins or to swap regions between family members in order to explore their function, can now be guided by the repeat architecture in

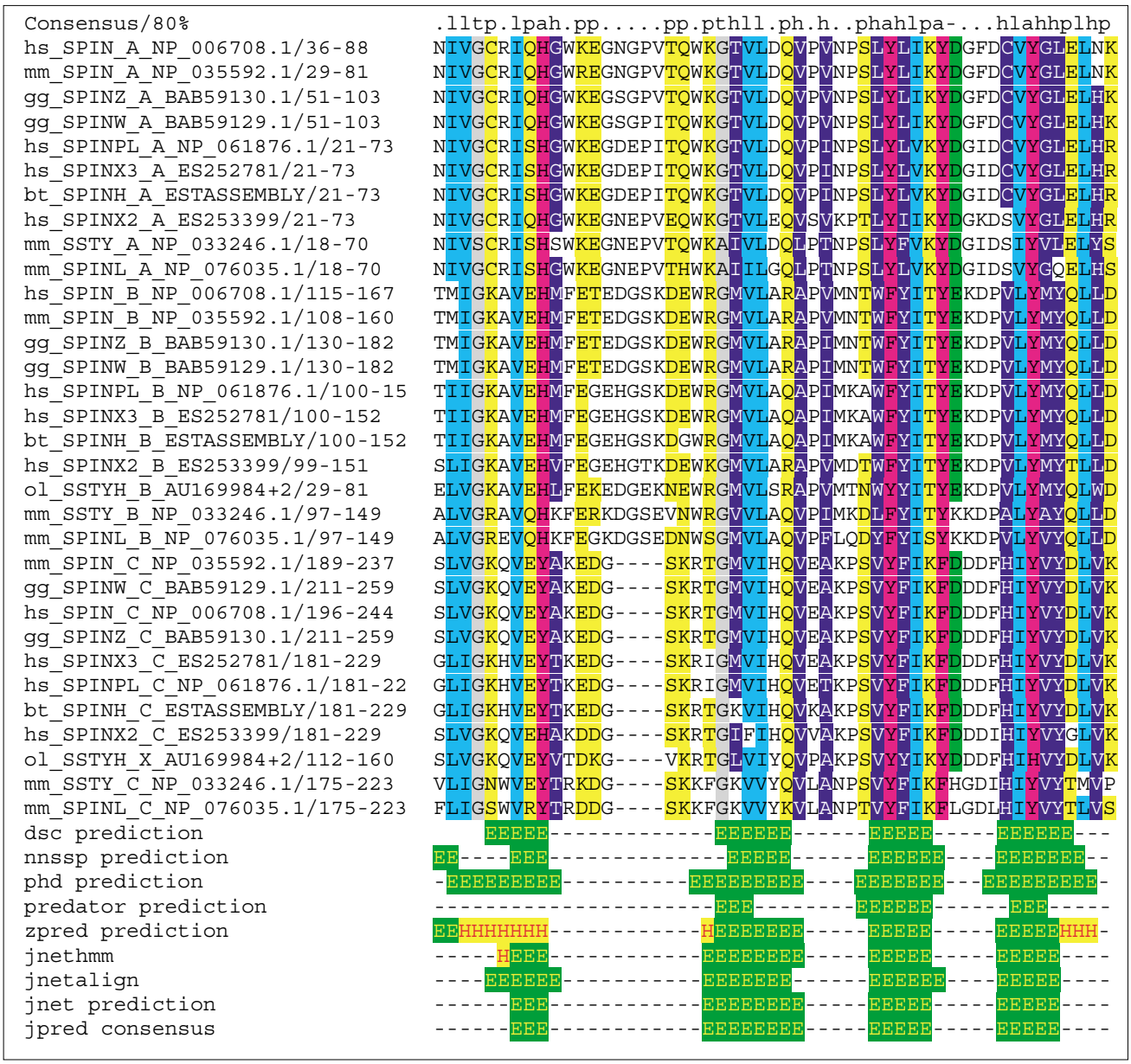

Figure 1
Alignment, consensus and secondary structure of Spin/Ssty repeats. The upper part shows the alignment of Spin/Ssty repeats. A two-letter organism-specific code (mm, *Mus musculus*; hs, *Homo sapiens*; ol, *Oryzias latipes*; bt, *Bos taurus*; gg, *Gallus gallus*) appears on the far left of each line, followed by a protein identifier, the repeat subtype (type A, amino-terminal; type B, central; type C, carboxy-terminal), the database identifier, the start and end residue of the Spin/Ssty repeat in the protein and the protein sequence. Amino acids are colored according to an 80% consensus. h, hydrophobic (ACFGILMVWP, white letters on dark blue); l, aliphatic (ILV, cyan); p, polar (NQSTY, yellow); a, aromatic (FHWY, purple); -, acidic (ED, green); +, basic (HKR, red letters on yellow); t, tiny (GAS, gray). The secondary-structure predictions of various programs run by the Jpred[2] server and the Jpred[2] consensus prediction are presented below. E, β strand; H, α helix.

these sequences to choose more reasonable borders. Finally, we hope that our findings will support the exploration of the tertiary structure of spindlin-like protein, as the Spin/Ssty sequence repeat is probably reflected by a repeated structural element with four β strands, which currently cannot be assigned to a known type of protein fold.

Materials and methods

Searching sequence databases

We scanned several databases to identify ESTs, ORFs, known protein sequences or gene structures of the Spin/Ssty gene family. We used the following databases, which can all be downloaded from the NCBI ftp server [14] (database

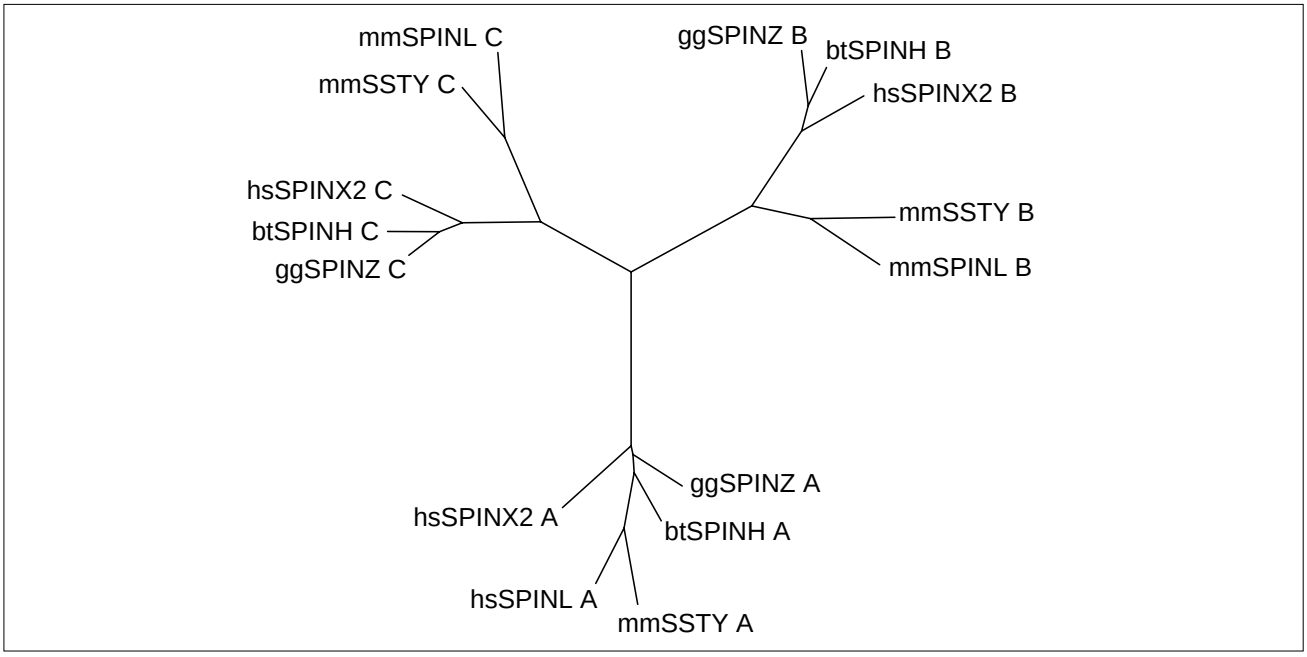

Figure 2
Phylogenetic tree of Spin/Ssty repeats. The tree was built from 15 repeats of five sequences. The labels stand for the repeats in the five proteins and consist of three fields: a two-letter code for the organism, an identifier for the protein sequence and the repeat subtype (see Figure 1 for terminology). Note the three groups of repeats: the amino-terminal repeats form one subtree, the central repeats form a second, and the carboxy-terminal repeats form a third. Thus, the phylogenetic classification of repeats matches the classification of the repeats by their positions in the proteins.

filenames are given in brackets) or the ENSEMBL ftp server [15]: the Non-redundant Protein Sequence Database (nr.Z), dbEST (est.Z), the mouse and human RefSeq mRNA and peptide sequences (hs.fna.gz, hs.faa.gz, mouse.fna.gz, mouse.faa.gz), the mouse and human UniGene databases (Hs.seq.uniq, Build #141; Mm.seq.uniq, Build #95) and the ENSEMBL set of confirmed human peptides and corresponding transcripts (ensembl.pep.gz, ensembl.cdna.gz, Ver.1.1.3). Pairwise sequence-similarity searches in these databases were carried out using the gapped versions of the programs of the BLAST program package version 2.1.2 with default scoring schemes [16].

Repeat analysis

The aim of the program dotter [17] is to visualize local sequence similarity between two sequences by allowing the user to view the dot matrix of the sequence comparisons and the alignment of the sequences in parallel. Here, dotter was used to compare sequences with themselves to examine them for repeats. Finally, it was used to refine the borders of repeat regions before their selection for the alignment.

Multiple alignment and phylogenetic tree construction

Multiple alignments were carried out with CLUSTALX version 1.8.1 [18] using the BLOSUM62 substitution matrix. The neighbor-joining algorithm [19] of CLUSTALX was used to build phylogenetic trees after gaps were removed from the alignments. The drawtree program of the PHYLIP package version 3.5 was used to visualize the tree [20].

Protein-sequence profile searches

For sensitive detection of repeats we built profile HMMs from the diverse alignments using the HMMER program suite [21] with default options for model building with hmmbuild (hmmls/domain alignment) and calibration with hmmcalibrate (sampled sequences: 5,000; mean length 350). Protein database searches with these HMMs were carried out using the hmmsearch program.

EST assembly

Having identified ESTs of a putative novel SPIN-family gene, we used the program Gap version 4.4 [22] for their assembly to derive a consensus representation of the complete mRNA sequence.

Secondary-structure prediction

Secondary-structure predictions were performed with the consensus method of the Jpred^2 server [23]. This method is built on several other well-known secondary-structure prediction algorithms such as DSC [24], Jnet [25], NNSSP [26], PHD [27] and Zpred [28]. According to the authors, the Jpred^2 consensus method reaches a level of 75% accuracy in secondary-structure prediction and outperforms the single methods.

Tertiary-structure prediction

We tried to assign known protein folds to the identified repeats by four widely used methods: 3D-PSSM [29], FUGUE [30], SUPERFAMILY [31] and SAM-T99 [32].

References

1. Oh B, Hwang S, McLaughlin J, Solter D, Knowles BB: **Timely translation during the mouse oocyte-to-embryo transition.** *Development* 2000, **127:**3795-3803.
2. Schultz RM: **Regulation of zygotic gene activation in the mouse.** *BioEssays* 1993, **15:**531-538.
3. Telford NA, Watson AJ, Schultz GA: **Transition from maternal to embryonic control in early mammalian development: a comparison of several species.** *Mol Reprod Dev* 1990, **26:**90-100.
4. Huarte J, Stutz A, O'Connell ML, Gubler P, Belin D, Darrow AL, Strickland S, Vassalli JD: **Transient translational silencing by reversible mRNA deadenylation.** *Cell* 1992, **69:**1021-1030.
5. Oh B, Hwang SY, Solter D, Knowles BB: **Spindlin, a major maternal transcript expressed in the mouse during the transition from oocyte to embryo.** *Development* 1997, **124:**493-503.
6. Oh B, Hampl A, Eppig JJ, Solter D, Knowles BB: **SPIN, a substrate in the MAP kinase pathway in mouse oocytes.** *Mol Reprod Dev* 1998, **50:**240-249.
7. Howlett SK: **A set of proteins showing cell cycle dependent modification in the early embryo.** *Cell* 1986, **45:**387-396.
8. Frank-Vaillant M, Haccard O, Ozon R, Jessus C: **Interplay between Cdc2 kinase and the c-Mos/MAPK pathway between metaphase I and metaphase II in *Xenopus* oocytes.** *Dev Biol* 2001, **231:**279-288.
9. Bishop CE, Hatat D: **Molecular cloning and sequence analysis of a mouse Y chromosome RNA transcript expressed in the testis.** *Nucleic Acids Res* 1987, **15:**2959-2969.
10. Burgoyne PS, Mahadevaiah SK, Sutcliffe MJ, Palmer SJ: **Fertility in mice requires X-Y pairing and a Y-chromosomal "spermiogenesis" gene mapping to the long arm.** *Cell* 1992, **71:**391-398.
11. Conway SJ, Mahadevaiah SK, Darling SM, Capel B, Rattigan AM, Burgoyne PS: **Y353/B: a candidate multiple-copy spermiogenesis gene on the mouse Y chromosome.** *Mamm Genome* 1994, **5:**203-210.
12. Itoh Y, Hori T, Saitoh H, Mizuno S: **Chicken *spindlin* genes on W and Z chromosomes: transcriptional expression of both genes and dynamic behavior of spindlin in interphase and mitotic cells.** *Chromosome Res* 2001, **9:**283-299.
13. Laval SH, Reed V, Blair HJ, Boyd Y: **The structure of DXF34, a human X-linked sequence family with homology to a transcribed mouse Y-linked repeat.** *Mamm Genome* 1997, **8:**689-691.
14. **National Center for Biotechnology Information ftp server** [ftp://ncbi.nlm.nih.gov]
15. **ENSEMBL ftp server** [ftp://ftp.ensembl.org]
16. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
17. Sonnhammer ELL, Durbin R: **A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis.** *Gene* 1995, **167:**GC1-GC10.
18. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ: **Multiple sequence alignment with CLUSTALX.** *Trends Biochem Sci* 1998, **23:**403-405.
19. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4:**406-425.
20. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164-166.
21. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
22. Bonfield JK, Smith KF, Staden R: **A new DNA sequence assembly program.** *Nucleic Acids Res* 1995, **23:**4992-4999.
23. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ: **Jpred: A consensus secondary structure prediction server.** *Bioinformatics* 1998, **14:**892-893.
24. King RD, Sternberg MJE: **Machine learning approach for the prediction of secondary structure.** *J Mol Biol* 1990, **216:**441-457.
25. Cuff JA, Barton GJ: **Application of multiple sequence alignment profiles to improve protein secondary structure prediction.** *Proteins* 2000, **40:**502-511.
26. Salamov AA, Solovyev VV: **Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments.** *J Mol Biol* 1995, **247:**11-15.
27. Rost B, Sander C: **Combining evolutionary information and neural networks to predict protein secondary structure.** *Proteins* 1994, **19:**55-72.
28. Zvelebil MJJM, Barton GJ, Taylor WR, Sternberg MJE: **Prediction of protein secondary structure and active sites using the alignment of homologous sequences.** *J Mol Biol* 1987, **195:**957-961.
29. Kelley LA, MacCallum RM, Sternberg MJE: **Enhanced genome annotation using structural profiles in the program 3D-PSSM.** *J Mol Biol* 2000, **299:**501-522.
30. Shi J, Blundell TL, Mizuguchi K: **FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties.** *J Mol Biol* 2001, **310:** 243-257.
31. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.** *J Mol Biol* 2001, **313:** 903-919.
32. Karplus K, Barrett C, Hughey R: **Hidden Markov models for detecting remote protein homologies.** *Bioinformatics* 1998, **14:**846-856.

4 A novel repeat in the melanoma-associated chondroitin sulfate proteoglycan defines a new protein family

Die folgende Erklärung legt die individuellen Beiträge der Co-Autoren zum nachfolgenden Manuskript dar. Von den Co-Autoren unterschriebene Fassungen dieser Erklärung liegen dieser Dissertation bei. Die Erklärungen sollen belegen, dass die erzielten Resultate im wesentlichen auf meiner Arbeit beruhen und ihre Verwendung innerhalb dieser Dissertation gerechtfertigt ist.

The CSPG manuscript: declaration about contributions of authors

Hereby I, co-author of the manuscript *„A novel repeat in the melanoma-associated chondroitin sulfate proteoglycan defines a new protein family“* which appeared in *FEBS letters (2002) vol.527, pp.114-118* declare my contributions to the results and to the preparation of the manuscript. Furthermore, I agree that the statements below describe the contributions of the other co-authors correctly.

1) Eike Staub
- conducted the protein sequence analysis,
- defined the gene models of the genes that were discovered from genome sequence data,
- discovered and characterised the repeat,
- predicted the secondary structure of the repeat,
- predicted the sequence similarity to domains adopting a cadherin fold,
- wrote the text and prepared the figures for the manuscript,
- served as the corresponding author during the review process.

2) Bernd Hinzmann
- confirmed the gene models of hypothetical genes in this study by EST assembly,
- contributed to the final stages of manuscript preparation by helpful comments on the style of the manuscript.

3) André Rosenthal
- was the supervisor of the project,
- contributed to the final stages of manuscript preparation by helpful comments on the style of the manuscript.

FEBS Letters 527 (2002) 114–118 FEBS 26438

A novel repeat in the melanoma-associated chondroitin sulfate proteoglycan defines a new protein family

Eike Staub*, Bernd Hinzmann, André Rosenthal

metaGen Pharmaceuticals GmbH, Oudenarder Str. 16, D-13347 Berlin, Germany

Received 18 June 2002; revised 26 July 2002; accepted 1 August 2002

First published online 14 August 2002

Edited by Gunnar von Heijne

Abstract The human melanoma-associated chondroitin sulfate proteoglycan (MCSP) and its rat ortholog NG2 are thought to play important roles in angiogenesis-dependent processes like wound healing and tumor growth. Based on electron microscopy studies, the highly glycosylated ectodomain of NG2 has been subdivided into the globular N-terminus, a flexible rod-like central region and a C-terminal portion in globular conformation. We identified a novel repeat named CSPG in the central ectodomain of NG2, MCSP and other proteins from fly, worm, human, sea urchin and a cyanobacterium which shows similarity to cadherin repeats. As earlier electron microscopy studies indicate, the folding of the tandem repeats compresses the length of the proposed repeat region by a factor of ~10 compared to the fully extended peptide chain. We identified two conserved negatively charged residues which might govern the binding properties of CSPG repeats. The phyletic distribution of CSPG repeats suggests that horizontal gene transfer contributed to their evolutionary history.

Key words: Chondroitin sulfate proteoglycan; Kringle domain; Protein repeat; Cadherin repeat; β-Sandwich; CSPG repeat

1. Introduction

In multicellular organisms, interactions of cells with the extracellular matrix (ECM) are of fundamental importance, ensuring the remodelling and maintenance of tissue architecture of multicellular organisms. Different families of membrane proteins mediate these interactions with varying degrees of specificity to their binding partners in the ECM [1]. The most prominent members are the heterodimeric receptors of the integrin family [2], but in recent years several membrane glycoproteins have been identified as ECM-binding components, such as the syndecans which are important for tissue homeostasis and cancer development [3–5]. The functionality of the syndecans is governed by the attached heparan sulfate chains which can interact with a wide range of ligands, albeit with low ligand specificity [6]. Another family of membrane-bound heparan sulfate proteoglycans, the glypicans, has also been implicated in tumor formation as mutant glypican-3 causes the Simpson–Golabi–Behmel overgrowth syndrome [7,8]. The same molecule was recently shown to be a negative regulator of breast cancer [7,8].

Here we focus on the melanoma-associated chondroitin sulfate proteoglycan (MCSP) and its putative rat ortholog NG2. Human MCSP was first identified by its function as a high molecular weight melanoma-associated antigen [9]. Even before the gene was known, a monoclonal antibody directed against an anti-MCSP antibody and thus mimicking the unavailable natural MCSP protein proved to be an effective suppressor of anchorage-independent tumor growth [10]. NG2 was identified as a developmentally regulated membrane protein in various developing tissues [11]. The rat NG2 protein comprises 2326 amino acids and has a signal peptide, followed by a large extracellular domain, a transmembrane domain, and a 76 residue cytoplasmic tail. The ectodomain was subdivided into the D1, D2 and D3 domains based on sequence features. Four internal repeats of ~200 amino acids and two of ~30 amino acids length were described for the ectodomain. Apart from a 12 residue segment which was noted to resemble a Ca^{2+}-binding fragment in the second chicken *N*-cadherin repeat, no similarities to other proteins were noted. In electron microscopy images of NG2, the ectodomain appeared to be subdivided into three parts: the globular N-terminus, the globular C-terminus and the rod-like central region [12].

NG2 can be proteolytically processed, resulting in the release of almost the entire ectodomain [13]. Some biochemical studies on NG2/MCSP concentrated on their ligand-binding properties. In electron microscope images, Tillet et al. observed that collagen fibers aligned with the central flexible rod-like D2 domain and collagen V and VI were shown to bind specifically to the D2 domain in ligand-binding assays [12]. In addition, NG2 binds plasminogen and its fragments like angiostatin, as long as they harbor positively charged kringle domains. It was proposed that multiple kringle binding sites in NG2 exist, that the interaction does not depend on chondroitin sulfate (CS) chains and that positively charged residues on kringle domains bind acidic clusters in NG2, which leads to sequestering of angiostatin in gliomas [14,15]. The same mode of binding was also suggested to explain the interaction of NG2 with the PDGF-α receptor in the developing rat brain [16]. In adherent cell lines, NG2 was shown to be organized in arrays and to co-localize with actin and myosin-containing stress fibers [17]. Although the exact function of NG2 and MCSP is still unknown, they may be implicated in angiogenesis, tissue invasion and cell spreading [18–21].

*Corresponding author. Fax: (49)-30-45082 101.
E-mail address: eike.staub@metagen.de (E. Staub).

0014-5793/02/$22.00
PII: S0014-5793(02)03195-2

2. Materials and methods

The non-redundant protein database from the NCBI was used as the basic pool of protein sequences in this study. Furthermore, we searched for sequence similarities in smaller databases of different proteomes from yeast, fly and worm (yeast.aa, drosoph.aa, wormpep) available from the FTP servers of the NCBI or the Sanger Institute. We used the BLASTP program [22] of the BLAST package with standard parameters to detect pairwise sequence similarities in these databases. The iterative PSIBLAST method was used to construct sequence profiles starting from single sequence fragments. During the iterations the inclusion of sequences, which are in the twilight zone of sequence similarity, into the profile was carefully checked. Inclusion thresholds were adjusted to include only true homologs, but were never raised above expectation (E) values of 0.008. Recursive searches using identified sequences were applied to confirm the detected similarities. As an independent and more sensitive method Hidden Markov Models (HMMs) of sequence alignments were applied to search protein sequence databases for additional homologous sequence fragments. To build, calibrate and apply profile HMMs we used the programs of the HMMER package [23]. Intramolecular repeats were visualized by a dot plot analysis using the program DOTTER [24]. The significance of the similarity between putative intramolecular repeats was confirmed by cross-comparisons using two additional methods, the PRSS program [25] from the FASTA package and the PROSPERO program [26]. Multiple alignments were created using CLUSTALX [27] and edited using JALVIEW (Clamp, M., unpublished). For the coloring of the alignment according to consensus rules we used the CHROMA program [28]. Signal peptides were predicted using the SIGNALP program in version 2.0 [29]. We predicted transmembrane helices using TMHMM version 2.0 [30]. The secondary structure of proteins was predicted on the basis of protein sequence alignments using the PHD prediction server [31].

3. Results and discussion

3.1. Identification of the CSPG repeat

As the knowledge about protein domains increased dramatically during the last years, we rescanned the NG2 sequence using the Smart and Pfam domain databases [32,33]. We discovered two laminin-G domains in the N-terminus of NG2. These domains occupy a large part of the formerly defined D1 region in the ectodomain of NG2. Though laminin-G domains are widespread among many extracellular proteins, their general function is not known. In laminin-G, this domain is implicated in heparin binding. It shows sequence similarity to pentraxins and thrombospondin-like molecules and a common fold was predicted for members of this superfamily [34].

Dot plot analysis [24] of the NG2 sequence (GenPep accession CAA39884.2) revealed extensive repeat structures between residues 420 and 2135 in the NG2 ectodomain. Most diagonals were separated by approximately 100 amino acids, which is an indicator for the repeat size. We chose the subsequence from 1124–1226 as a prototype of the putative repeat because it appeared to be similar to the highest number of other subsequences. When we compared this fragment with the whole NG2 ectodomain sequence using PROSPERO [26], four copies of the repeat were detected with expectation (E) values below 1e−5. When we used this subsequence as a seed in a PSIBLAST [22] query of the non-redundant protein database at the NCBI (nr) with an inclusion threshold E value of 0.008, the search converged after four rounds having identified 63 repeat copies in 10 different proteins from human, rat worm, fly, sea urchin and a cyanobacterium. We detected 10 copies of the repeat in the NG2 sequence. We confirmed this finding by extensive reciprocal PSIBLAST searches using different repeat copies as queries. The PSIBLAST searches usually converged after three to five rounds, having identified largely overlapping sets of subsequences. In addition, we were able to proof the significance of the similarity between repeats by extensive pairwise comparisons using the PROSPERO algorithm and the PRSS algorithm [25]. Hereafter, we refer to the repeat as CSPG repeat.

To achieve maximum sensitivity in database searches we aligned the CSPG repeat sequences and constructed a HMM using the HMMer program package [23]. A search in the nr database using the HMM revealed eight non-redundant protein sequences with a total of 74 repeat copies ($E < 1e{-}3$) (see Fig. 1). For each of the rat NG2 and human MCSP proteins 15 repeat copies were now found covering the whole region that was expected to contain repeats after dot plot analysis. We predicted the secondary structure on the basis of the manually edited alignment using the PHD prediction server (Fig. 1). The CSPG repeat is likely to obtain an all-β fold, possibly comprising eight β-strands. The sixth β-strand starts with a conserved aromatic residue, which is followed by a conserved serine or threonine. Conserved acidic residues are present in the subsequent loop regions between strands 6 and 7 as well as between 7 and 8. For most β-strands one can observe a typical alternating pattern of hydrophobic and non-hydrophobic residues. Hydrophobic side chains that point to the same side of the β-sheet are probably buried in the interior of the CSPG domain. We applied several fold recognition methods using single sequences or the whole alignment as queries, but no significant predictions were obtained.

We also determined the domain architectures of the CSPG repeat proteins (Fig. 2). The CSPG repeat occurs in 1–15 tandem copies per protein. In some proteins the CSPG repeat is combined with laminin-G domains and EGF-like domains. These domains are common in extracellular proteins and are known to mediate interactions between cell surface molecules and extracellular ligands or matrix components. We found 12 copies of the CSPG repeat in the embryonic blastocoelar matrix protein ECM3 (GenPep AAG00570.1) from the sea urchin *Lytechinus variegatus* next to a five-fold tandem repeat of Calx-β motifs. These motifs were originally found to occur in cytoplasmic regulatory regions of Na^{+}–Ca^{2+}-exchange proteins and integrin-β4. In ECM3 they reside in a putative extracellular region between a predicted signal peptide and a single transmembrane domain. The function of these, presumably extracellular, Calx-β motifs is not clear, although it is assumed that they bind calcium [35]. Support for the hypothesis that extracellular Calx-β motifs bind calcium comes from the sponge MAFp4 ECM protein which requires calcium for self-association.

One CSPG repeat prediction in ECM3 (residues 1145–1240) overlapped with a weak prediction of a cadherin repeat (residues 1169–1260; $E = 0.54$) detected by Smart [33]. This was an indicator of similarity between cadherin-like repeats and CSPG repeats. PSIBLAST searches starting with this sea urchin sequence fragment converged on a set of CSPG repeats without identification of cadherin repeats, although marginal similarity to cadherins was detected in weak hits with E values below the profile inclusion threshold. To gain sensitivity we built a profile HMM from an alignment of CSPG repeats with less than 70% pairwise identity. When we applied the HMM to the sequences of cadherins we obtained nine matches with E values in the range between 0.1 and 0.0095 which can be

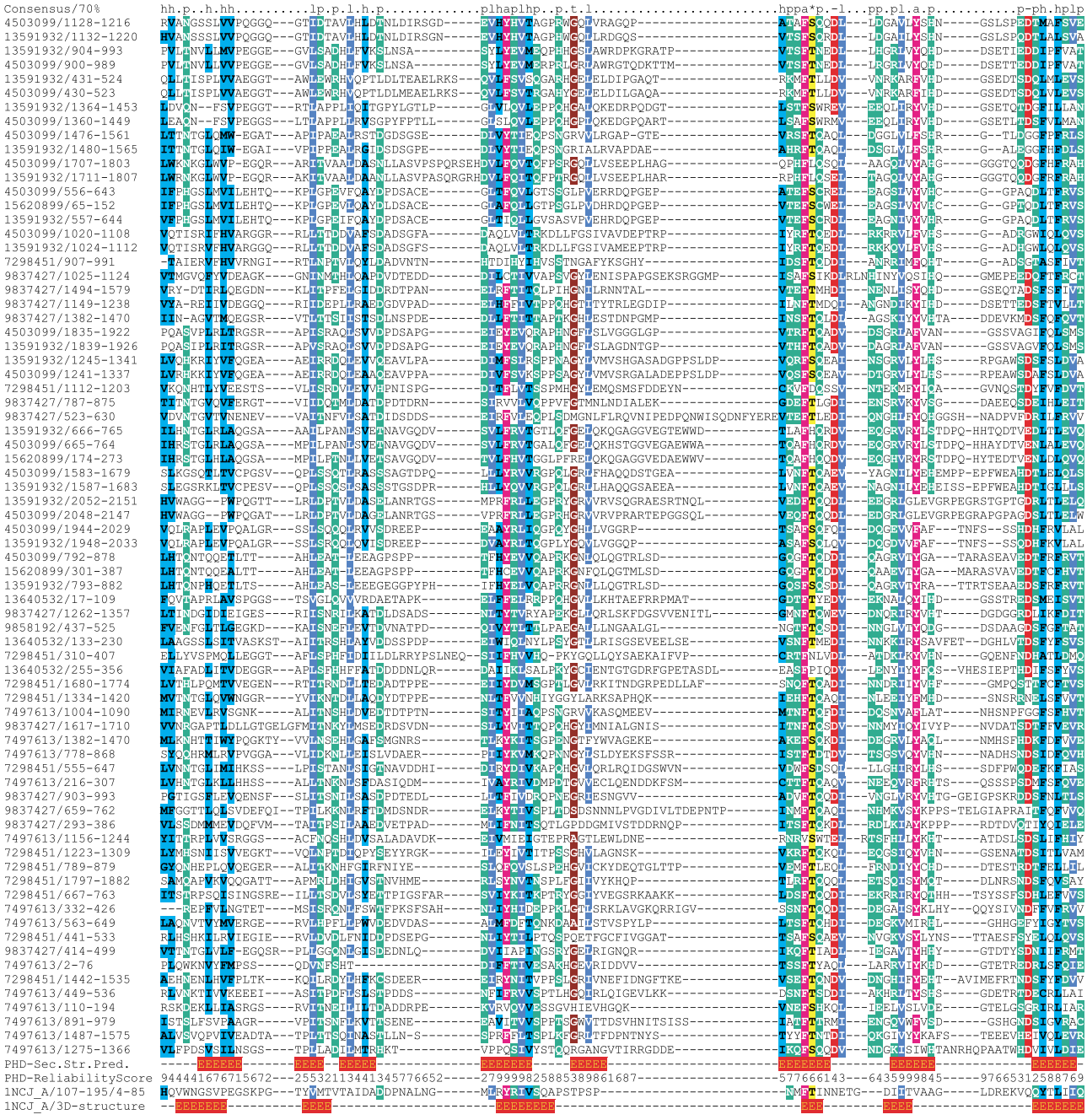

Fig. 1. Alignment of selected CSPG repeats. The identifier of each sequence in the non-redundant database (NCBI) is followed by the position of each repeat in the sequence. The protein names and species can be taken from the legend of Fig. 2. The alignment was colored according to a 70% consensus using the following amino acid classification: negatively charged: white on red, DE (-); hydroxylic: black on yellow, ST (*); aliphatic: white on dark blue, ILV (l); positively charged: black on green, HKR (+); tiny: white on brown, AGS (t); aromatic: white on purple, FHWY (a); polar: white on green, CDEHKNQRST (p); hydrophobic: black on light blue, ACFGHILMTVWY (h). Below the alignment of CSPG repeats the predicted secondary structure as determined using the PHD prediction server and the corresponding reliability values are printed (0–9; 9 is most reliable) [31]. The predicted secondary structure can easily be compared to the secondary structure of a cadherin repeat in the solved 3D structure of an *N*-cadherin fragment (PDB code 1NCJ) which is given in the last two lines together with its protein sequence. Secondary structure code: *E* stands for β-strand, *H* for α-helix.

considered significant. Therefore, we hypothesize that CSPG and cadherin repeats are distantly related. Apart from this similarity, the CSPG repeat predictions did not overlap with any predictions of known domains from Pfam and Smart.

3.2. Structural and functional implications

The identification of laminin-G domains and the novel CSPG repeats permitted a finer partitioning of the ectodomains of the NG2 and MCSP oncoproteins compared to the originally proposed D1/D2/D3 division. The presence of CSPG repeats in proteins from worm, fly and sea urchin shed light on the phyletic distribution of the CSPG repeats. The presence of a single CSPG repeat in a cyanobacterium may suggest that the CSPG repeat is an ancient protein module that was preserved during evolution. Alternatively, it may be an example of domain accretion by horizontal gene transfer

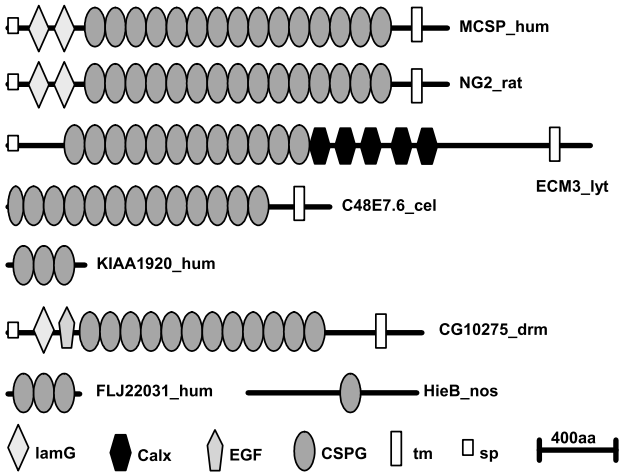

Fig. 2. Domain architecture of CSPG repeat proteins. The following domain abbreviations were used: lamG, laminin-G domain; EGF, EGF-like domain; Calx, Calx-β domain, CSPG, CS proteoglycan repeat; sp, signal peptide; tm, transmembrane helix. The protein identifiers consist of a protein and a species abbreviation. The protein names and their GenBank identifiers are: MCSP, gi:4503099; NG2, gi:13591932; embryonic blastocoelar matrix protein (ECM3), gi:9837427; KIAA1920, gi:15620899; hypothetical protein FLJ-22031, gi:13640532; CG10275, gi:7298451; C48E7.6, gi:7497613; HieB, gi:9858192. The species abbreviations are hum for *Homo sapiens*, rat for *Rattus norvegicus*, lyt for *L. variegatus*, drm for *Drosophila melanogaster*, cel for *Caenorhabditis elegans*, nos for *Nostoc* species PCC9229. The domain positions on the sequences are drawn approximately to scale.

from an unknown multicellular eukaryote to a cyanobacterium. We noticed the similarity of the phyletic distributions of CSPG repeats and the Calx-β motif which was also found in higher eukaryotes and cyanobacterial proteins. Both motifs are not present in other eukaryotic organisms like yeast, fly and worm. We think that a scenario in which these motifs are deleted from yeast, fly and worm genomes is less likely than the horizontal transfer of genes or gene fragments with CSPG repeats and Calx-β motifs between a marine eukaryote and a cyanobacterium.

Evidence that CSPG repeats fold into structural modules comes from the reconsideration of earlier electron microscopy studies in the light of the repeat discovery. Assuming a maximum extension of the polypeptide chain and a per-residue distance of 0.36 nm, the ~1700 residues of the CSPG repeat region would result in an extended polypeptide chain of 612 nm. The length of the rod-like central domain of the NG2 ectodomain was estimated to be in the range of 30–110 nm in electron microscopy images [12]. This implies that the folding of the repeat region results in a significant (~10-fold) compression of the length of the polypeptide chain. It is likely that this protein shrinking is conferred by the folding of CSPG repeats into structural units.

Further evidence for the relatedness between cadherin and CSPG repeats came from a comparison of the secondary structures of cadherin repeats with known 3D structures and the predicted secondary structures of CSPG repeats. The cadherin repeat in the second domain of an *N*-cadherin fragment folds into a β-sandwich (PDB code 1NCJ) [36]. We aligned the 1NCJ sequence to the CSPG repeat alignment (see Fig. 1). The six β-strands of the second cadherin domain aligned to a large extent with the predicted β-strands in CSPG repeats. Furthermore, cadherin-like and CSPG repeats are both thought to obtain a rod-like structure and the size of the repeat units from both families is ~100 residues. Therefore, we hypothesize that CSPG repeats and cadherin repeats share a common ancestor and structural fold. Do they also have similar biochemical properties? Compared to the many negatively charged residues involved in calcium binding of cadherin repeats, the CSPG repeats contain only two negatively charged positions in their C-terminal half (see Fig. 1). A calcium-binding capacity has not been reported for CSPG repeat proteins yet. It cannot be inferred from sequence analysis alone whether CSPG repeats bind calcium by their two acidic residues.

Insights into the biochemical function of CSPG repeats can be gained by reviewing the literature on the MCSP/NG2 proteins. One function of CSPG repeats may be the binding and presentation of the CS chains which then determine the functional properties of the molecule. However, the binding of the D2 and D3 regions of NG2 to positively charged kringle domains of the plasmin(ogen)/angiostatin system seemed to be independent of the presence of CS chains. As these regions comprise most of the CSPG repeats and multiple binding sites seem to exist, the binding of kringle domains is possibly facilitated by negatively charged conserved residues in the CSPG repeats (see Fig. 1). Another function of CSPG repeats is the binding of collagen. The D2 region of NG2 was shown to bind collagen and almost completely consists of CSPG repeats.

We conclude that the CSPG repeat is a novel cadherin-like and tumor-relevant protein module which we expect to mediate interactions between cells and the ECM in species as divergent as cyanobacteria, fly, worm, sea urchin and human. Furthermore, we propose that horizontal gene transfer contributed to the evolutionary history of genes which encode CSPG repeats.

References

[1] Pluschke, G., Vanek, M., Evans, A., Dittmar, T., Schmid, P., Itin, P., Filardo, E.J. and Reisfeld, R.A. (1996) Proc. Natl. Acad. Sci. USA 93, 9710–9715.
[2] Hynes, R.O. (1999) Trends Cell Biol. 9, M33–M37.
[3] Woods, A. and Couchman, J.R. (1998) Trends Cell Biol. 8, 189–192.
[4] Liu, W., Litwack, E.D., Stanley, M.J., Langford, J.K., Lander, A.D. and Sanderson, R.D. (1998) J. Biol. Chem. 273, 22825–22832.
[5] Alexander, C.M., Reichsman, F., Hinkes, M.T., Lincecum, J., Becker, K.A., Cumberledge, S. and Bernfield, M. (2000) Nat. Genet. 25, 329–332.
[6] Carey, D.J. (1997) Biochem. J. 327, 1–16.
[7] Pilia, G. et al. (1996) Nat. Genet. 12, 241–247.
[8] Xiang, Y.Y., Ladeda, V. and Filmus, J. (2001) Oncogene 20, 7408–7412.
[9] Bumol, T.F., Wang, Q.C., Reisfeld, R.A. and Kaplan, N.O. (1983) Proc. Natl. Acad. Sci. USA 80, 529–533.
[10] Harper, J.R. and Reisfeld, R.A. (1983) J. Natl. Cancer Inst. 71, 259–263.
[11] Nishiyama, A., Dahlin, K.J., Prince, J.T., Johnstone, S.R. and Stallcup, W.B. (1991) J. Cell Biol. 114, 359–371.
[12] Tillet, E., Ruggiero, F., Nishiyama, A. and Stallcup, W.B. (1997) J. Biol. Chem. 272, 10769–10776.
[13] Nishiyama, A., Lin, X.H. and Stallcup, W.B. (1995) Mol. Biol. Cell 6, 1819–1832.
[14] Goretzki, L., Lombardo, C.R. and Stallcup, W.B. (2000) J. Biol. Chem. 275, 28625–28633.
[15] Chekenya, M. et al. (2002) FASEB J. 12, 12.

[16] Nishiyama, A., Lin, X.H., Giese, N., Heldin, C.H. and Stallcup, W.B. (1996) J. Neurosci. Res. 43, 315–330.
[17] Lin, X.H., Dahlin-Huppe, K. and Stallcup, W.B. (1996) J. Cell Biochem. 63, 463–477.
[18] Burg, M.A., Pasqualini, R., Arap, W., Ruoslahti, E. and Stallcup, W.B. (1999) Cancer Res. 59, 2869–2874.
[19] Iida, J., Pei, D., Kang, T., Simpson, M.A., Herlyn, M., Furcht, L.T. and McCarthy, J.B. (2001) J. Biol. Chem. 276, 18786–18794.
[20] Eisenmann, K.M. et al. (1999) Nat. Cell Biol. 1, 507–513.
[21] Iida, J., Meijne, A.M., Spiro, R.C., Roos, E., Furcht, L.T. and McCarthy, J.B. (1995) Cancer Res. 55, 2177–2185.
[22] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[23] Eddy, S.R. (1998) Bioinformatics 14, 755–763.
[24] Sonnhammer, E.L. and Durbin, R. (1995) Gene 167, GC1–GC10.
[25] Pearson, W.R. (2000) Methods Mol. Biol. 132, 185–219.
[26] Mott, R. (2000) J. Mol. Biol. 300, 649–659.
[27] Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G. and Gibson, T.J. (1998) Trends Biochem. Sci. 23, 403–405.
[28] Goodstadt, L. and Ponting, C.P. (2001) Bioinformatics 17, 845–846.
[29] Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Int. J. Neural Syst. 8, 581–599.
[30] Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) J. Mol. Biol. 305, 567–580.
[31] Rost, B. and Sander, C. (1994) Proteins 19, 55–72.
[32] Bateman, A. et al. (2002) Nucleic Acids Res. 30, 276–280.
[33] Letunic, I. et al. (2002) Nucleic Acids Res. 30, 242–244.
[34] Beckmann, G., Hanke, J., Bork, P. and Reich, J.G. (1998) J. Mol. Biol. 275, 725–730.
[35] Hodor, P.G., Illies, M.R., Broadley, S. and Ettensohn, C.A. (2000) Dev. Biol. 222, 181–194.
[36] Tamura, K., Shan, W.S., Hendrickson, W.A., Colman, D.R. and Shapiro, L. (1998) Neuron 20, 1153–1163.

5 The novel EPTP repeat defines a superfamily of proteins implicated in epileptic disorders

Die folgende Erklärung legt die individuellen Beiträge der Co-Autoren zum nachfolgenden Manuskript dar. Von den Co-Autoren unterschriebene Fassungen dieser Erklärung liegen dieser Dissertation bei. Die Erklärungen sollen belegen, dass die erzielten Resultate im wesentlichen auf meiner Arbeit beruhen und ihre Verwendung innerhalb dieser Dissertation gerechtfertigt ist.

The EPTP manuscript: declaration about contributions of authors

Hereby I, co-author of the manuscript *„„The novel EPTP repeat defines a superfamily of proteins implicated in epileptic disorders"* which appeared in *Trends in Biochemical Sciences (2002) vol.27, no.9, pp.441-444* declare my contributions to the results and to the preparation of the manuscript. Furthermore, I agree that the statements below describe the contributions of the other co-authors correctly.

1) Eike Staub
- conducted the protein sequence analysis,
- discovered and characterised the domain including the prediction of structure and function,
- contributed to the definition of gene structures for previously unknown genes of the family (together with Bernd Hinzmann),
- identified the link from sequence similarity to function between LGI1 and VLGR1 proteins,
- contributed to the discussion about the significance of the transmembrane helix prediction for the LGI1 protein,
- discovered the links to epileptic disorders mapping in chromosomal regions which comprise further EPTP homologs,
- wrote the text and prepared the figures for the manuscript.

2) Jordi Pérez-Tur
- raised the interest in the detailed sequence analysis of LGI1 by his previous work on the role of this gene in autosomal dominant lateral temporal epilepsy (ADLTE),
- contributed to the discussion about the significance of the transmembrane helix prediction for the LGI1 protein,
- contributed useful comments on the clinical aspects of the discovery.

3) Reiner Siebert
- raised the interest in the detailed sequence analysis of LGI1 by his previous work on the role of this gene in autosomal dominant lateral temporal epilepsy (ADLTE),
- contributed to the discussion about the significance of the transmembrane helix prediction for the LGI1 protein,
- contributed useful comments on the clinical aspects of the discovery.

4) Nicholas K. Moschonas
- raised the interest in the detailed sequence analysis of LGI1 by his previous work on the role of this gene in autosomal dominant lateral temporal epilepsy (ADLTE),
- contributed useful comments on the molecular genetic aspects of the discovery.

5) Carlo Nobile
- raised the interest in the detailed sequence analysis of LGI1 by his previous work on the role of this gene in autosomal dominant lateral temporal epilepsy (ADLTE),
- contributed to the discussion about the significance of the transmembrane helix prediction for the LGI1 protein,
- contributed useful comments on the molecular genetic aspects of the discovery.

6) Panagiotis Deloukas
- raised the interest in the detailed sequence analysis of LGI1 by his previous work on the role of this gene in autosomal dominant lateral temporal epilepsy (ADLTE),
- contributed to the discussion about the significance of the transmembrane helix prediction for the LGI1 protein,
- contributed useful comments on the clinical aspects of the discovery.

7) Bernd Hinzmann
- was the supervisor of the study,
- contributed to the definition of gene structures for previously unknown genes of the family (together with Eike Staub),
- contributed to the discussion about the significance of the transmembrane helix prediction for the LGI1 protein.

The novel EPTP repeat defines a superfamily of proteins implicated in epileptic disorders

Eike Staub, Jordi Pérez-Tur, Reiner Siebert, Carlo Nobile, Nicholas K. Moschonas, Panagiotis Deloukas and Bernd Hinzmann*

Recent studies suggest that mutations in the *LGI1*/Epitempin gene cause autosomal dominant lateral temporal epilepsy. This gene encodes a protein of unknown function, which we postulate is secreted. The LGI1 protein has leucine-rich repeats in the N-terminal sequence and a tandem repeat (which we named EPTP) in its C-terminal region. A redefinition of the C-terminal repeat and the application of sensitive sequence analysis methods enabled us to define a new superfamily of proteins carrying varying numbers of the novel EPTP repeats in combination with various extracellular domains. Genes encoding proteins of this family are located in genomic regions associated with epilepsy and other neurological disorders.

The human *LGI1*/Epitempin gene was first discovered in the T98G glioblastoma cell line, in which it is rearranged as a result of a t(10;19)(q24;q13) balanced translocation [1]. The gene is predominantly expressed in neural tissues, especially in the brain. It has been localized to the human chromosomal region 10q24, which made it a strong candidate for positional cloning of genes leading to autosomal dominant lateral temporal epilepsy. The *LGI1* gene was recently found to be mutated in families with this type of epilepsy [2,3]. As well as the presence of a signal peptide and leucine-rich repeats (LRRs), a protein-specific tandemly repeated domain of about 130 residues was discovered in the C-terminal sequence of the LGI1 protein. We now present a deeper analysis of this repeat in which advanced sequence analysis techniques were used to find novel homologs of the LGI1 protein that could serve as starting points for the investigation of other epileptic disorders.

*On behalf of the European Collaborative Consortium for the study of Autosomal Dominant Lateral Temporal Epilepsy. A complete list of all members of the consortium is given in the Acknowledgements section.

Repeat identification

The LGI1 protein contains three LRRs, which are flanked by the typical cysteine-rich N-terminal and C-terminal sequences. LRRs are widespread among different classes of intracellular and extracellular proteins [4]. The repeat units are sequentially ordered in a curved structure. The concave side is formed by a parallel β sheet, which mediates interactions with other proteins. Furthermore, we identified an unknown tandem repeat of about 130 residues in the C-terminus of the human LGI1 protein by use of the Prospero program [3]. Alignments of the C-terminal repeat, which we named Epitempin (EPTP) repeat, were used in PSI-BLAST [5] and HMMER [6] searches; these identified additional copies in other members of the protein family (Pfam 7.2: PF03736) [7]. Recently, we noted a shorter internal repeat of about 50 residues in the alignment of these original EPTP repeats. The significance of this finding was supported by extensive pairwise comparisons of copies of the short repeat by means of the programs PRSS [8] and Prospero [9], which predominantly yielded expectation (E) values below 1×10^{-2}. We redefined the term EPTP repeat to refer to the shorter 50-residue repeat. On the basis of an alignment of these short repeats, we carried out iterative hidden Markov model (HMM) searches (inclusion threshold E value <0.1) to identify novel EPTP repeats in the non-redundant protein database of the US National Center for Biotechnology Information. The searches converged after two rounds on a redundant set of protein sequences with up to seven copies of the EPTP repeat per protein. Additional repeat copies were identified by PSI-TBLASTN searches in EST and genome databases.

The final alignment of all identified EPTP repeats (Fig. 1) was used for structure prediction. The PHD method [10] predicted an all-β secondary structure consisting of four β strands. We also applied the 3D-PSSM and the SUPERFAMILY fold-recognition methods to the alignment and to single repeats but we could not achieve significant predictions. However, with two copies of the repeat as a query, β-sandwich folds were commonly among the top hits; this finding supports our hypothesis of a β-sheet structure of the EPTP repeat. Because the EPTP repeats tend to be present in seven copies per protein, they could constitute another class of seven-fold β-sheet repeats, which fold into a β-propeller structure.

We found seven tandem repeats in each C-terminus of the human and mouse LGI1 proteins. These repeats covered the entire sequence C-terminal to the LRR region. Seven EPTP repeats were also detected in novel members of the LGI family – the LGI2 and LGI4 proteins from mouse and human (chromosomal regions 4p15.2 and 19q13.12), for which we assembled the complete coding regions by using sequence and raw trace data from public EST and genome databases. A common characteristic of members of this LGI subfamily of EPTP-repeat proteins is the presence of LRRs in the N-terminus. The existence of the *LGI* genes mentioned above has since been confirmed [11]. In addition to the LGI family members mentioned, Gu *et al.* identified the *LGI3* gene (8p21.3) by mining of the Celera human genome database [11]. We rediscovered *LGI3* in a PSI-TBLASTN search of dbEST and confirmed its genomic location on chromosome 8p21.3 by reordering the underlying public genomic data (GenBank 18693501) with the *LGI1* sequence as a template. In contrast to other recent studies on the LGI proteins [1,11], we did not predict a transmembrane helix in any LGI protein with the currently most reliable program for prediction of transmembrane helices, TMHMM [12]. The presence of a transmembrane helix in the C-terminus of LGI1 contradicts our hypothesis that all seven EPTP repeats fold into a

http://tibs.trends.com 0968-0004/02/$ – see front matter PII: S0968-0004(02)02163-1

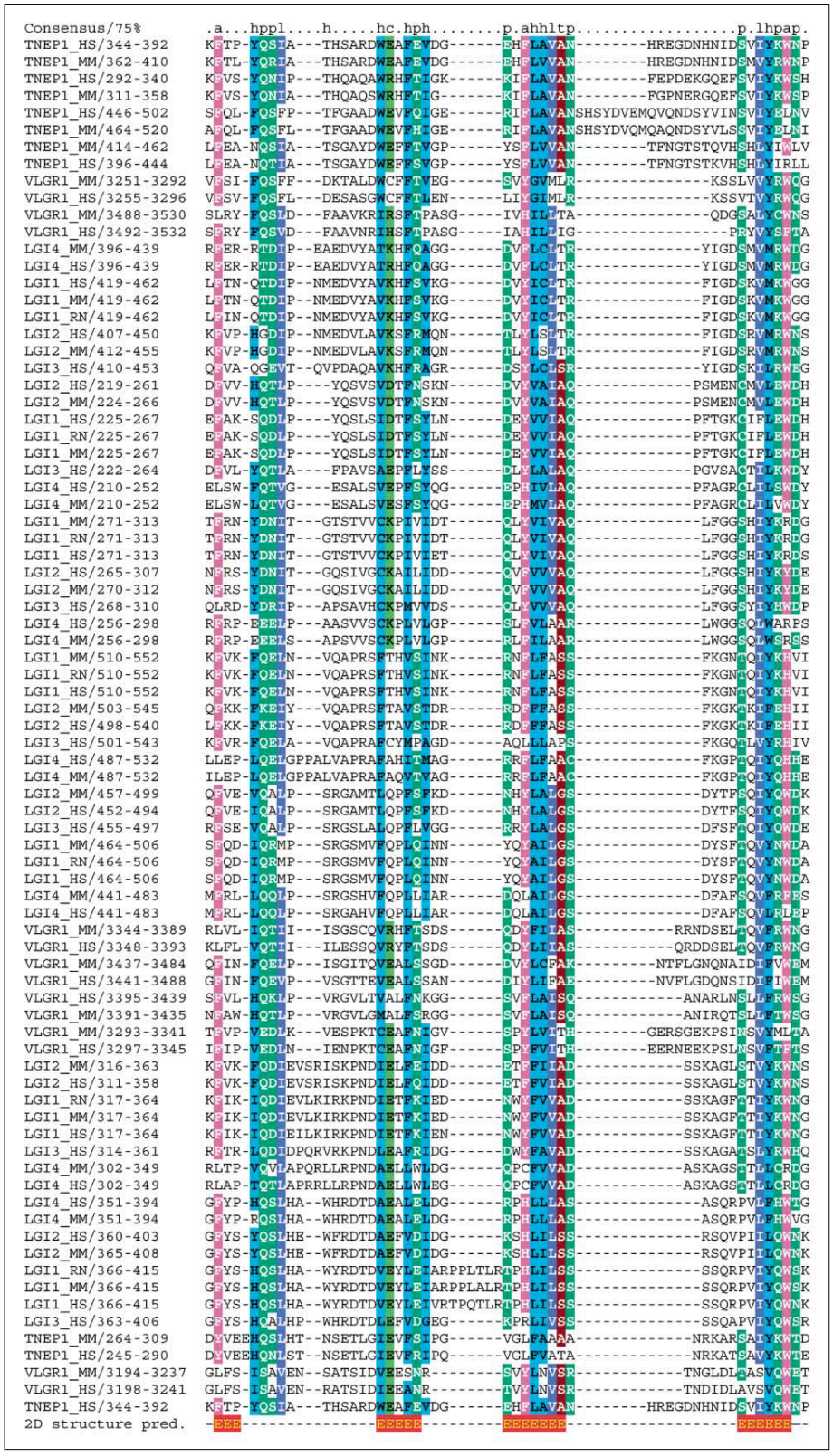

Fig. 1. Alignment of EPTP repeats. The name of each protein is followed by an organism-specific two-letter code (HS, *Homo sapiens*; MM, *Mus musculus*; RN, *Rattus norvegicus*) and by the position of each repeat in the sequence. The names of the proteins stand for sequences derived from the following entries in the public databases: TNEP1_HS, Unigene: Hs.352217; TNEP1_MM, EMBL: AJ487520; VLGR1_MM, GenBank: 16904210; VLGR1_HS, GenBank: 19882213; LGI4_MM, EMBL: AJ487521; LGI4_HS, EMBL: AJ487419; LGI1_HS, GenBank: 4826816; LGI1_MM, GenBank: 9938002; LGI1_RN, EMBL: AJ487517; LGI2_HS, EMBL: AJ487516; LGI2_MM, EMBL: AJ487515; LGI3_HS, EMBL: AJ487518. The alignment was manually refined and colored according to a 75% consensus by use of CHROMA [26] with the following classification of amino acids: aliphatic (l), white on dark blue, ILV; tiny (t), white on brown, AGS; aromatic (a), white on purple, FHWY; charged (c), black on light green, DEHKR; polar (p), white on green, CDEHKNQRST; hydrophobic (h), black on light blue, ACFGHILMTVWY. Predicted secondary structure: E stands for β strand. An alignment of the redefined EPTP repeat will replace the previous entry in Pfam (PF03736). This multiple sequence alignment (alignment number ALIGN_000431) has been deposited with the European Bioinformatics Institute (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/ALIGN_000431.dat).

uniform structure. Therefore, we propose that the LGI proteins are secreted and do not span the cell membrane.

Notably, we discovered proteins other than those having the typical LGI-like domain organization. The human and mouse very large G-protein-coupled receptors (VLGR) [13] each have seven EPTP repeats in their extracellular domains. In the VLGR1 protein, the EPTP repeats are located between tandem arrays of Calx-β repeats (Fig. 2). In the VLGR1 C-terminus, a G-protein-coupled receptor proteolytic site (GPS) domain [14] is followed by seven putative transmembrane helices resembling the transmembrane region of the flamingo and latrophilin proteins, which are involved in dendrite formation and synaptogenesis [15,16]. Calx-β repeats were originally found in intracellular parts of Na^{+}–Ca^{2+} exchange proteins and in integrins [17], but they are also present in extracellular parts of other proteins, such as the sea urchin protein ECM3, where they are thought to bind calcium [18]. Another novel human protein sequence was detected in the non-redundant protein database and by a PSI-TBLASTN search in the Unigene database (Hs.352217). We named the novel gene *TNEP1*, because the hypothetical coding region of the cDNA has a thrombospondin N-terminal domain and five EPTP repeats. The gene is localized on chromosome 21q22.3. A predicted gene model for this locus (GenBank:XM_092850) alternatively suggests a larger protein with seven EPTP repeats, which has to be confirmed by cDNA sequencing. Sequence similarity of VLGR1 or TNEP1 proteins to the LGI family has not been noted previously. We also found EPTP repeats by TBLASTN searches in the unassembled genomic sequences of *Tetraodon nigroviridans* and *Danio rerio*. This finding suggests a

minimum evolutionary age of 400 million years for the EPTP repeat.

Functional implications

Our study revealed several new genes that are potential candidates for epileptic disorders of unknown origin and other neurological diseases. These include several new members of the LGI subfamily and representatives of two further subfamilies, TNEP1 and VLGR1. Some of the novel candidate genes are located in chromosomal regions already associated with epileptic disease.

The *VLGR1* gene (OMIM:602851) is located in the human chromosomal region 5q14.1, a region associated with familial febrile convulsions of type 4 (OMIM:604352) [19,20]. The *VLGR1* gene was reported to be homologous to the mouse *Mass1* gene, which is mutated in the Frings mouse, a model of audiogenic seizures [21]. In combination, this is strong evidence that *VLGR1* is implicated in human epileptic disease. Another neurological syndrome mapping to human chromosome 5q13.1–15.1 is the Usher syndrome type II (OMIM:605472), which is characterized by retinitis pigmentosa and hearing loss [22].

The *LGI4* gene in human chromosomal region 19q13.12 is located in direct proximity to the *SCN1B* gene, which encodes a voltage-gated sodium ion channel β1 subunit. Mutations in *SCN1B* lead to a variable epilepsy subtype called generalized epilepsy with febrile seizures plus [23]. Another type of epilepsy, benign familial infantile convulsions (OMIM:601764), also maps to chromosome 19q13 [24] and is apparently not caused by mutations in *SCN1B* [25]. We consider the dysfunction of the LGI4 protein to be a possible mechanism leading to benign familial infantile convulsions.

The *TNEP1* gene is located on chromosome 21q22.3 near the Down syndrome critical region. The neurological phenotype of most individuals with Down syndrome involves mental retardation (OMIM:190685). The detailed involvement of genes from the Down syndrome critical region is far from clear, but *TNEP1* is a promising novel candidate to contribute to the neurological phenotype in this disorder.

We have presented a novel type of repeat that is present in vertebrate proteins involved in neurological disorders. The EPTP repeat is the only sequence motif that links the three subfamilies represented by *TNEP1*, *VLGR1* and *LGI1*. Two of these genes have been studied in detail and both were found to be relevant to epileptic disorders by independent approaches. Therefore, we hypothesize that EPTP repeats in the known and novel gene products of the described superfamily have a prominent role in the development of epileptic disorders.

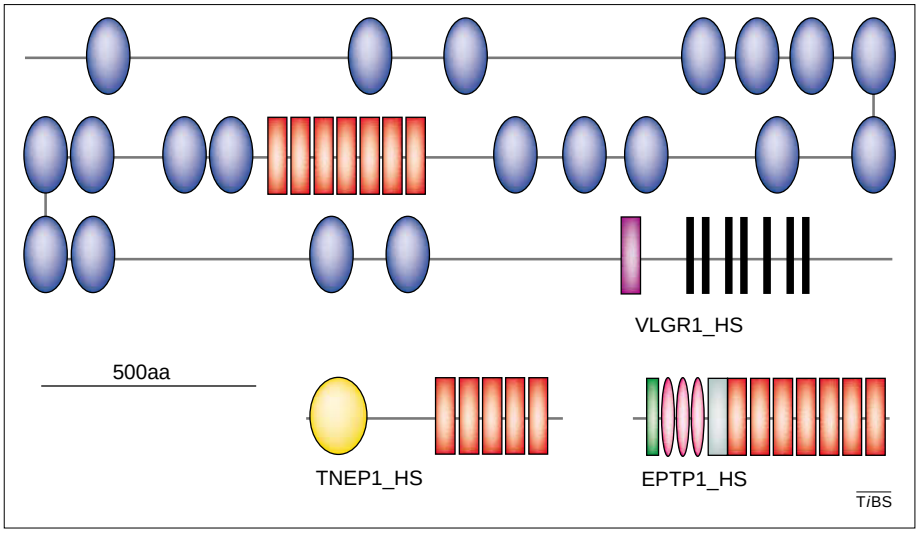

Fig. 2. Domain architecture of representative EPTP-repeat proteins. The name of each protein is followed by an organism-specific two-letter code (HS, *Homo sapiens*; MM, *Mus musculus*; RN, *Rattus norvegicus*). The domains and proteins are drawn approximately to scale. The domains are: transmembrane helix (black); Calx-β, a domain in Na^{+}-Ca^{2+}-exchange proteins (blue ovals); G-protein-coupled receptor proteolytic site (purple block); thrombospondin N-terminal domain (yellow oval); Epitempin repeat (red block); leucine-rich repeats (pink oval); leucine-rich repeats N-terminal domain (dark green); leucine-rich repeats C-terminal domain (light green).

Acknowledgements

Other members of the European Collaborative Consortium for the Study of Autosomal Dominant Lateral Temporal Epilepsy are: L. French, J. Galan, S. Gesk, A. Gorostidi, L. Kluwe, A. López de Munain, J.F. Martí Massó, V.F. Mautner, R. Michelucci, J.M. Morante-Redolat, J.J. Poza, J.F. Prud'homme, A. Rosenthal, A. Sáenz, T. Sarafidou, U. Stephani, and C.A. Tassinari. The study was partly supported by the Hensel Stiftung (Kiel). Work at J.P-T's laboratory was funded through grants from the Generalitat Valenciana (AE01/072) and Fondo de Investigaciones Sanitarias (00/0900). C.N. was supported by the Italian League Against Epilepsy. We thank an anonymous reviewer for helpful comments.

Note in proof

After revision of this manuscript, Scheel *et al.* published the discovery of the epilepsy-associated repeat (EAR), which is identical to the EPTP repeat [27]. These authors also predict a β-propeller fold for the novel domain.

References

1 Chernova, O.B. *et al.* (1998) A novel gene, LGI1, from 10q24 is rearranged and downregulated in malignant brain tumors. *Oncogene* 17, 2873–2881
2 Kalachikov, S. *et al.* (2002) Mutations in LGI1 cause autosomal-dominant partial epilepsy with auditory features. *Nat. Genet.* 30, 335–341
3 Morante-Redolat, J.M. *et al.* (2002) Mutations in the LGI1/Epitempin gene on 10q24 cause autosomal dominant lateral temporal epilepsy. *Hum. Mol. Genet.* 11, 1119–1128
4 Kobe, B. and Kajava, A.V. (2001) The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* 11, 725–732
5 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
6 Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics* 14, 755–763
7 Bateman, A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.* 30, 276–280
8 Pearson, W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* 132, 185–219
9 Mott, R. (2000) Accurate formula for P-values of gapped local sequence and profile alignments. *J. Mol. Biol.* 300, 649–659
10 Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232, 584–599
11 Gu, W. *et al.* (2002) The LGI1 gene involved in lateral temporal lobe epilepsy belongs to a new subfamily of leucine-rich repeat proteins. *FEBS Lett.* 519, 71–76

12 Möller, S. *et al*. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* 17, 646–653
13 McMillan, D.R. *et al*. (2002) Very large G protein-coupled receptor-1, the largest known cell surface protein, is highly expressed in the developing central nervous system. *J. Biol. Chem*. 277, 785–792
14 Ponting, C.P. *et al*. (1999) A latrophilin/CL-1-like GPS domain in polycystin-1. *Curr. Biol*. 9, R585–R588
15 Gao, F.B. *et al*. (2000) Control of dendritic field formation in *Drosophila*: the roles of flamingo and competition between homologous neurons. *Neuron* 28, 91–101
16 Sudhof, T.C. (2001) α-Latrotoxin and its receptors: neurexins and CIRL/latrophilins. *Annu. Rev. Neurosci*. 24, 933–962
17 Schwarz, E.M. and Benzer, S. (1997) Calx, a Na–Ca exchanger gene of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U. S. A*. 94, 10249–10254
18 Hodor, P.G. *et al*. (2000) Cell-substrate interactions during sea urchin gastrulation: migrating primary mesenchyme cells interact with and align extracellular matrix fibers that contain ECM3, a molecule with NG2-like and multiple calcium-binding domains. *Dev. Biol*. 222, 181–194
19 Nakayama, J. *et al*. (2000) Significant evidence for linkage of febrile seizures to chromosome 5q14-q15. *Hum. Mol. Genet*. 9, 87–91
20 Durner, M. *et al*. (2001) Genome scan of idiopathic generalized epilepsy: evidence for major susceptibility gene and modifying genes influencing the seizure type. *Ann. Neurol*. 49, 328–335
21 Skradski, S.L. *et al*. (2001) A novel gene causing a mendelian audiogenic mouse epilepsy. *Neuron* 31, 537–544
22 Pieke-Dahl, S. *et al*. (2000) Genetic heterogeneity of Usher syndrome type II: localisation to chromosome 5q. *J Med. Genet*. 37, 256–262
23 Wallace, R.H. *et al*. (1998) Febrile seizures and generalized epilepsy associated with a mutation in the Na^+-channel beta1 subunit gene SCN1B. *Nat. Genet*. 19, 366–370
24 Guipponi, M. *et al*. (1997) Linkage mapping of benign familial infantile convulsions (BFIC) to chromosome 19q. *Hum. Mol. Genet*. 6, 473–477
25 Moulard, B. *et al*. (2000) Study of the voltage-gated sodium channel beta 1 subunit gene (SCN1B) in the benign familial infantile convulsions syndrome (BFIC). *Hum. Mutat*. 16, 139–142
26 Goodstadt, L. and Ponting, C.P. (2001) CHROMA: consensus-based colouring of multiple alignments for publication. *Bioinformatics* 17, 845–846
27 Scheel, H. *et al*. (2002) A common protein interaction domain links two recently identified epilepsy genes. *Hum Mol Genet*. 11, 1757–1762

Eike Staub*
Bernd Hinzmann
metaGen Pharmaceuticals GmbH, Oudenarder Str. 16, d-13347 Berlin, Germany.
*e-mail: eike.staub@metagen.de

Jordi Pérez-Tur
Unitat de Genètica Molecular, Institut de Biomedicina de València-CSIC, València, Spain.

Reiner Siebert
Institute of Human Genetics, University Hospital Kiel, Kiel, Germany.

Carlo Nobile
CNR-Istituto di Neuroscienze, Sezione di Biologia e Fisiopatologia Neuromuscolare, Padova, Italy.

Nicholas K. Moschonas
Institute of Molecular Biology and Biotechnology, Foundation of Research and Technology and Dept of Biology, University of Crete, Heraklion, Crete, Greece.

Panagiotis Deloukas
The Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK.

6 Sequence similarity between prokaryotic nitrogen-sensing histidine kinases and vertebrate hypoxia-inducible proteins

Die folgende Erklärung legt die individuellen Beiträge der Co-Autoren zum nachfolgenden Manuskript dar. Von den Co-Autoren unterschriebene Fassungen dieser Erklärung liegen dieser Dissertation bei. Die Erklärungen sollen belegen, dass die erzielten Resultate im wesentlichen auf meiner Arbeit beruhen und ihre Verwendung innerhalb dieser Dissertation gerechtfertigt ist.

The NtrY/HIG manuscript: declaration about contributions of authors

Hereby I, co-author of the manuscript *„Sequence similarity between prokaryotic nitrogen-sensing histidine kinases and vertebrate hypoxia-inducible proteins“* which is submitted for publication in *Cellular Signalling*, declare my contributions to the results and to the preparation of the manuscript. Furthermore, I agree that the statements below describe the contributions of the other co-authors correctly.

1) Eike Staub
- conducted the protein sequence analysis,
- discovered the similarity between NtrY and HIG protein families,
- wrote the text and prepared the figures for the manuscript,
- serves as the corresponding author during the review process.

2) Thomas Braun
- raised the interest in the search for vertebrate homologs of bacterial proteins that function in histidine phosphorylation-dependent signalling,
- contributed to the manuscript preparation by hints to relevant literature and by comments on the style of the manuscript.

Sequence similarity between prokaryotic nitrogen-sensing histidine kinases and vertebrate hypoxia-inducible proteins

Eike Staub§* and Thomas Braun#

§ metaGen Pharmaceuticals GmbH, Oudenarder Str. 16, D-13347 Berlin, Germany
current address: Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology, Ihnestr. 73, D-14195 Berlin, Germany
University of Halle-Wittenberg, Institute of Physiological Chemistry, Hollystr. 1, D-06097 Halle, Germany

* author to whom correspondence should be addressed
email: staub@molgen.mpg.de
tel ++49-(0)30-8413-1157
fax ++49-(0)30-8413-1152

Abstract

Whereas in eukaryotic signalling pathways protein kinases mainly act on serine, threonine or tyrosine residues, in bacterial signal transduction predominantly histidine residues become phosphorylated. Although it is known that histidine phosphorylation also occurs in eukaryotes, the players and mechanisms of metazoan histidine phosphorylation remain elusive. Here we demonstrate that proteins encoded by the hypoxia-inducible gene (HIG) family, a group of mammalian proteins of unknown biochemical function, have significant sequence similarity to a subfamily of prokaryotic histidine kinases, which function in the response to nitrogen. The similarity region comprises the typical membrane-proximal sensory domains of eubacterial NtrY-like histidine kinases and the membrane-proximal regions of plant, fungal and metazoan HIG proteins. To our knowledge, this is the first report of significant sequence similarity which links a sensory domain of a prokaryotic two-component signal transduction pathway to a family of vertebrate proteins. Based on this sequence similarity and on a potential functional link of both families, the involvement in cellular processes dependent on the levels of gaseous compounds, we hypothesise that HIG and NtrY protein families are homologous. We suggest that HIG proteins are top candidates for future experimental studies that try to link bacterial phosphohistidine-dependent signal transduction to metazoan cellular signalling.

Introduction

Protein phosphorylation is a common signal transduction mechanism in all organisms. In eukaryotes the most widely known protein kinases catalyse the phosphorylation of hydroxyamino acids, i.e. serine/threonine protein kinases and tyrosine kinases. In contrast, the phosphorylation of histidine residues, which is the predominant type of protein phosphorylation in bacterial signal transduction, is only poorly characterised in mammals. Protein kinases specific for histidines have not yet been described in vertebrates, only in plants, fungi and prokaryotes [1,2]. Only rough estimates about the degree of histidine phosphorylation in mammals exists [3-5]. In lower eukaryotes, however, 6% of the phosphoamino acids of basic nuclear proteins are phosphohistidines, which is about two orders of magnitudes greater than the abundance of phosphotyrosine in this subset of proteins [6,7]. This suggests that the importance of histidine phosphorylation in eukaryotic signal transduction has not been fully recognised yet.

In prokaryotes, histidine phosphorylation is the key mechanism of so called 'two-component' signalling pathways, which link extracellular stimuli such as changing osmolarity, oxygen, or nitrogen levels to gene regulation, but also affect other functions. In two-component pathways, signals are sensed by cell surface-located receptors, the sensors. This results in their dimerisation and in auto-phosphorylation of cytoplasmic histidines. The high energy histidine-bound phosphates are transferred to aspartates in the receiver domains of response regulators. These proteins are often transcriptional regulators, which are activated upon phosphorylation-mediated conformational changes [8]. Typically, receptor histidine kinases have an extracellular sensory loop flanked by transmembrane regions, a dimerisation domain and a kinase domain [8].

One bacterial subfamily of histidine kinases is composed of the NtrY receptors, which are involved in the control of nitrogen-related environmental stimuli [9]. Recently, the hypoxia-inducible gene (HIG) family has been identified in vertebrates. Protein products of this family have not yet been characterised biochemically. The founding gene HIG is upregulated in response to hypoxia in the hypoxia-tolerant fish *Gillichthys mirabilis* [10]. Both families are predicted to comprise α-helical transmembrane proteins of largely uncharacterised biochemical function. During this study the sequence similarity between these two families is investigated. The results are discussed in the light of their potential meaning for cellular signalling in metazoa.

Materials and Methods

For all types of sequence similarity searches in databases we used the non-redundant protein database (nr) at the NCBI (http://www.ncbi.nlm.nih.gov/Database/) and the EBI set of bacterial protein sequence databases derived from completely sequences bacterial genomes (http://www.ebi.ac.uk/proteomes/). An initial HMM of the HIG protein family (HIG_1_N) was obtained from the Pfam database (version 7.0) of protein families [11]. Searches for local sequence similarity between query protein sequences and database proteins were carried out using BLASTP and PSIBLAST using three substitution matrices (PAM250, BLOSUM62, BLOSUM45) in combination with various E value cut-offs and profile inclusion thresholds [12]. Protein alignments were constructed using CLUSTALX [13]. HMM models of protein alignments were built and calibrated using the HMMER package [14]. The E value statistic of each HMM was calibrated using the default options of hmmcalibrate; the scores of 5000 random sequences, each of 325 residues length, were fitted to an extreme value distribution that was subsequently used for the calculation of E values for query scores. The PRSS program of the FASTA package was used to align two pairs of sequences and to assign a P value as an estimate for the significance of the alignments. The P value is estimated on the distribution of scores from alignments of one sequence with a randomly shuffled version of the other sequence [15]. To obtain an estimate for the similarity between two alignments, we used the LAMA web server with default options (minimal alignment length of 4 residues, minimum reported Z score of 5.6, calculation of the expectation (E) value on the basis of 5000 blocks (1700 more than in version 9.1 of the BLOCKS database) [16]. Using the COMPASS program we identified local similarities between alignment profiles using a Smith-Waterman-like algorithm allowing for the insertion of gapped columns during profile-to-profile alignment [17]. The calculation of E values for the resulting profile-to-profile alignments is based on the number of aligned columns in a profile database that can be specified explicitly.

Results and Discussion

The experimental evidence for a function of HIG proteins in hypoxia and the presence of two transmembrane domains in the N-terminal region of HIG proteins, like in histidine kinases, encouraged us to investigate the role of these proteins by sequence analysis. Using a Hidden Markov Model (HMM) of the HIG family from the Pfam database we scanned the protein databases of several eubacteria for HIG homologues. Indeed, we found a marginal similarity of HIG proteins to a number of bacterial proteins, among these the NtrY protein of *Caulobacter crescentus*. Using

this sequence fragment as a query, we performed PSIBLAST searches in the non-redundant (nr) protein database of the NCBI using various cut-offs. Applying an E value inclusion threshold of 0.01 we identified a family of 19 NtrY-like proteins. However, we were not able to detect significant similarity to HIG proteins by reciprocal PSIBLAST searches.

To gain sensitivity in database searches, we built an alignment of the putative sensory transmembrane region of NtrY proteins and derived a profile HMM by the use of the hmmbuild program of the HMMER package. The application of such an HMM on the HIG proteins resulted in alignments with the NtrY model consensus that hardly showed gapped regions. Two hits had E values of 0.033 and 0.05. For reciprocal HMM analysis, we built a HMM from a subsection of the Pfam HIG alignment (name HIG_1_N) which comprised the putative NtrY homology region. The application of the HIG HMM to NtrY-like proteins resulted in four hits with E values less than 0.06. However, because in searches of the large nr database using these HMMs the E values were no longer significant, these findings cannot be regarded as a proof of significant similarity.

To confirm the marginal similarity found in reciprocal HMM searches by an independent method, we applied a complementary approach based on the extensive pairwise cross-comparison of single sequences from one subfamily with single sequences from the other subfamily using the PRSS program. For the resulting 154 Smith-Waterman alignments, the median P value of all comparisons was 0.25, the 25^{th} percentile was 0.09, and the minimum P value was 0.0022. By this analysis we showed that not only the composition, but also the order of amino acid residues in sequences of both families contributes to the similarity. Because multiple sequence pairs of the two families can be aligned with significant P values below 0.05 we argue that this is further evidence for the significance of inter-family similarity.

To further investigate the hypothesis of an ancestral relation between the two families we applied two profile-to-profile alignment comparison methods. The LAMA method identifies ungapped homologous BLOCKS between two protein sequence alignments and estimates E values for the findings. The alignments of the HIG and NtrY proteins both passed the check for biased composition of BLOCKS. LAMA found a common BLOCK of 50 residues length with a score of 24. Assuming a database size of 5000 Blocks, this results in a significant Z score of 7.1 and an E value of 1.5e-2. This is further evidence that the sequence similarity between the two protein families is significant.

Because LAMA does not allow for gaps in the identified common BLOCK of two alignments, it might lose sensitivity. Therefore, we applied a second profile-to-profile comparison method allowing for gaps. The COMPASS method identifies local similarities between alignment profiles using a Smith-Waterman-like algorithm. The

calculation of E values for the resulting profile-to-profile alignments is based on the number of aligned columns in a profile database. When we compared the two automatic CLUSTALX alignments of the complete sequences of the NtrY and HIG protein sets, COMPASS identified a common region that is nearly identical to the aligned region shown in figure 1 with a score of 108 and an E value of 1.25e-10 when the database size is set to the length of the larger alignment. Explicitly specifying a database size of 1.000.000, which is greater than the number of aligned columns with gap fraction < 0.5 in the Pfam database, resulted in an E value of 3.08e-6. The COMPASS results clearly indicate significant sequence similarity in the common region of NtrY and HIG proteins. This is further evidence that both families are homologous. We argue that this is not due to a bias in amino acid composition that could arise from the in corporation of transmembrane helix residues, because (a) the PRSS analysis showed that the order of amino acids in HIG an NtrY sequences contributes significantly to the quality of pairwise sequence alignments and (b) the alignments of individual families passed the check for composition-biased alignments as implemented in the LAMA method. In the further course of the analysis we assume that NtrY and HIG proteins are homologous.

We constructed a combined alignment of the common region of the two subfamilies. Using the profile-to-profile alignment mode of CLUSTALX we obtained a combined NtrY/HIG superfamily alignment. We iteratively constructed a HMM, scanned the nr database for further homologues, and built a new HMM. The iterative searches converged in the 3rd round, resulting in the identification of 62 sequences in the nr database. After removal of redundancy at the 95% identity level, 48 sequences remained (see alignment figure 1). Among those were sequences from yeast, a filamentous fungus, plants, fly, mosquito and worm. This means that members of the postulated NtrY/HIG superfamily of proteins are present in all phyla, with the notable exception of the archaebacterial lineage. We hypothesise that the eukaryotic NtrY/HIG-like proteins are a eubacterial invention.

Furthermore, we argue that members of the HIG protein family, of which one member is upregulated in response to hypoxia at the transcript level, are probably involved in the sensing of small molecules. This might include sensing of local concentrations of oxygen, nitrogen, or NO near the surface of an animal cell. It is not clear from our work which specific member of the superfamily might detect what small molecule. However, it is tempting to speculate that HIG proteins sense oxygen and NtrY proteins sense nitrogen or nitrogen-related gaseous compounds taking into consideration the context in which these proteins were discovered. The alignment predicts a special role for basic residues in the central region of the domain (see residues 30 and 31 in figure 1). Their negative charges could be important for the binding of anions or gaseous compounds with partially negative

charge. The hydrophobicity in the transmembrane region is highly conserved, especially the aliphatic residues 13, 39 and 52 and a single tiny hydrophobic amino acid in residue 54, suggesting that these residues are important for the structure of the receptors.

A throughout evaluation of the domain architecture of NtrY and HIG proteins revealed that the vast majority of NtrY-like proteins have typical domains of histidine kinase receptors like HAMP domains, PAS domains, phosphoacceptor domains and the kinase domains themselves. Of the two NtrY-like proteins that lack these domains, one is described as a fragmentary sequence and the other is a hypothetical protein. The HIG-like proteins are devoid of such domains and have much shorter cytoplasmic tails. Two hypothetical HIG-like proteins with unusual domain composition stood out. The predicted rat protein XP_228571.1 most likely resulted from an erroneous gene prediction, leading to a fusion of a ribosomal L18ae-like protein with a HIG-like domain. In the hypothetical *Arabidopsis thaliana* T17F15.100 protein a RING finger domain was fused to the C-terminal part. RING fingers are known to be the catalytic domains of E3 ubiquitin ligases that confer the target specificity to ubiquitin-dependent protein degradation. Like the sequence itself, this functional link to proteasomal protein degradation is hypothetical. It remains an open question whether metazoan proteins have lost their additional histidine kinase domains or the bacterial sequences gained them. Since almost all bacterial NtrY-like proteins share the typical histidine kinase domain, we reason that the postulated common ancestor of HIG/NtrY proteins was a histidine kinase and that early domain loss in the metazoan lineage shaped the contemporary HIG proteins.

In conclusion, we have identified a eukaryotic protein domain in a broad range of species including vertebrates that has significant similarity to bacterial sensory domains of histidine kinase receptors. There is only limited knowledge about the biochemical functions of both families. Both seem to be involved in processes in which the concentration of gases, here nitrogen and oxygen, play a role. This weak functional link is in accordance with our assumption of homology. A further proof that the similarity between both families is not due to convergent evolution of unrelated sequences, but instead is due to evolution by descent from a common ancestral sequence, has to come from future experimental studies. Independent of the question whether the observed similarity is due to analogy or homology, we expect that the biochemical analysis of mammalian HIG proteins will lead to new insights into the mechanisms which allow animal cells to sense small compounds like gases in their local environment. Based on the assumption of homology we suggest that the HIG family of proteins is a good starting point for studies that try to discover the source of phosphohistidine-dependent signal transduction in

mammalian cells. However, even if HIG proteins will be confirmed as the first vertebrate homologs of histidine kinases, the source of histidine phosphorylation in eukaryotes would still remain unknown, because HIG proteins do not have a cytoplasmic histidine kinase domain. Nevertheless, HIG proteins as possible sensory receptors could then facilitate the search for downstream phosphohistidine signalling proteins by biochemical means. Therefore, we consider the HIG-like proteins to be important molecules for the elucidation of mechanisms of histidine phosphorylation in eukaryotic proteins.

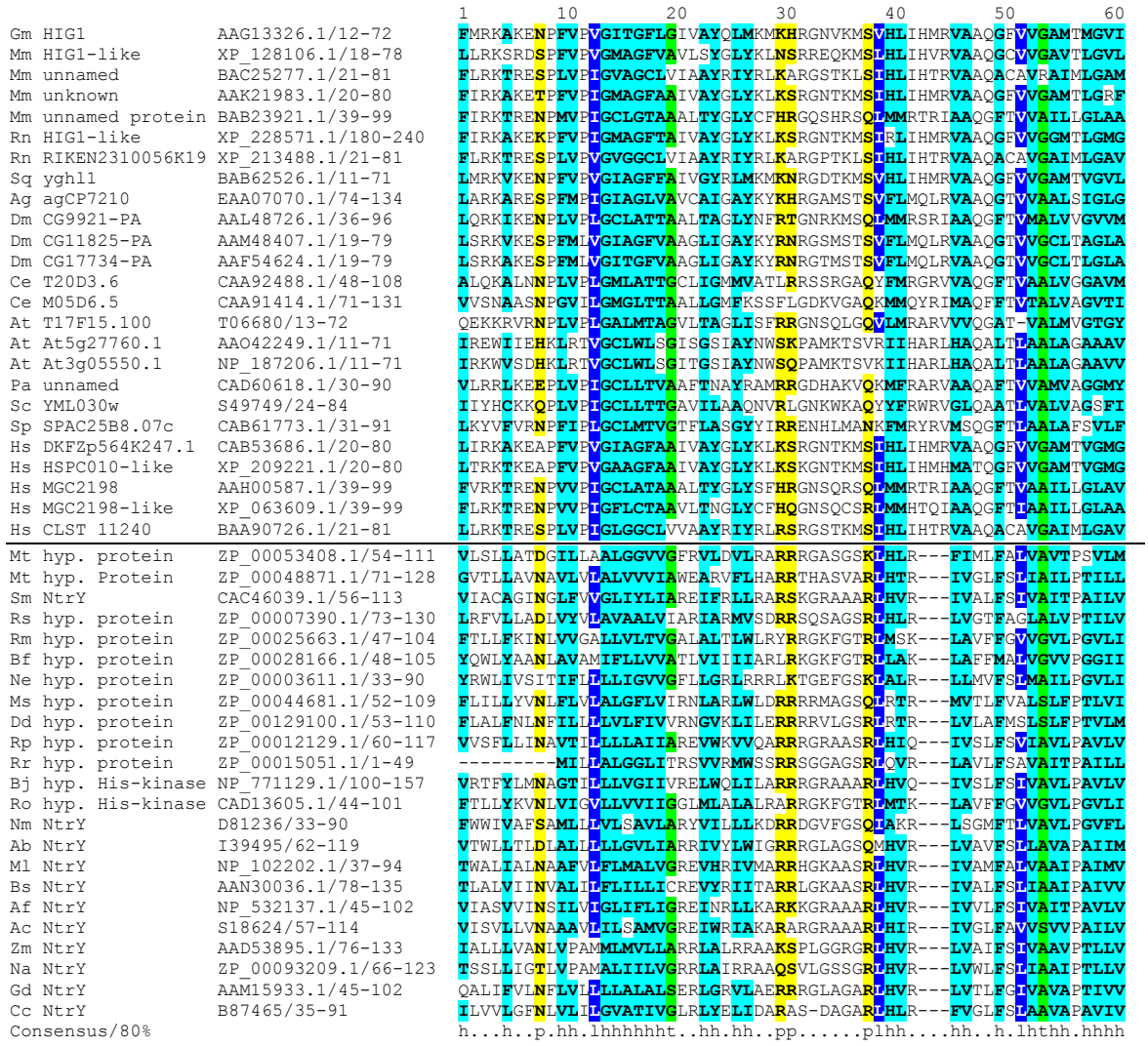

Figure 1

Sequence alignment of the common domain of NtrY and HIG proteins. Each line comprises a two-letter organism code, the protein name, a sequence database identifier (NCBI non-redundant protein database) followed by the location of the displayed sequence fragment and the domain sequence itself. Organism code: Hs *Homo sapiens*, Mm *Mus musculus*, Rn *Rattus norvegicus*, Sq *Seriola quinqueradiata*, Ag *Anopheles gambiae*, Dm *Drosophila melanogaster*, Ce *Caenorhabditis elegans*, At *Arabidopsis thaliana*, Pa *Podospora anserine*, Sc *Saccharomyces cerevisiae*, Sp *Schizosaccharomyces pombe*, Rs *Rhodobacter sphaeroides*, Rm *Ralstonia metallidurans*, Ro *Ralstonia solanacearum*, Bf *Burkholderia fungorum*, Ne *Nitrosomonas europaea*, Nm *Neisseria meningitidis* MC58, Mt *Magnetospirillum magnetotacticum*, Ms *Magnetococcus sp. MC-1*, Dd *Desulfovibrio desulfuricans G20*, Ab *Azospirillum brasilense*, Ml *Mesorhizobium loti*, Bs *Brucella suis 1330*, Sm *Sinorhizobium meliloti*, Af *Agrobacterium tumefaciens*, Bj *Bradyrhizobium japonicum*, Rp *Rhodopseudomonas palustris*, Ac *Azorhizobium caulinodans*, Rr *Rhodospirillum rubrum*, Zm *Zymomonas mobilis*, Na *Novosphingobium aromaticivorans*, Gd *Gluconacetobacter diazotrophicus*, Cc *Caulobacter crescentus CB15*. The amino acids are coloured according an 80% consensus rule and the following classification: DE negative (-) yellow, ST hydroxy (*) brown, ILV aliphatic (|) dark blue, HKR positive (+) red, AGS tiny (t) green, FHWY aromatic (a) purple, DEHKR charged (c) dark green, CDEHKNQRST polar (p) light orange, ACFGHILMTVWY hydrophobic (h) light blue. The horizontal bar separates the HIG subfamily and the NtrY subfamily.

References

1. Hwang I, Chen HC, Sheen J. Two-component signal transduction pathways in Arabidopsis. Plant Physiol 2002;129(2):500-515.
2. Santos JL, Shiozaki K. Fungal histidine kinases. Sci STKE 2001;2001(98):RE1.
3. Chen CC, Bruegger BB, Kern CW, Lin YC, Halpern RM, Smith RA. Phosphorylation of nuclear proteins in rat regenerating liver. Biochemistry 1977;16(22):4852-4855.
4. Chen CC, Smith DL, Bruegger BB, Halpern RM, Smith RA. Occurrence and distribution of acid-labile histone phosphates in regenerating rat liver. Biochemistry 1974;13(18):3785-3789.
5. Smith DL, Bruegger BB, Halpern RM, Smith RA. New histone kinases in nuclei of rat tissues. Nature 1973;246(5428):103-104.
6. Matthews HR, Huebner VD. Nuclear protein kinases. Mol Cell Biochem 1984;59(1-2):81-99.
7. Matthews HR. Protein kinases and phosphatases that act on histidine, lysine, or arginine residues in eukaryotic proteins: a possible regulator of the mitogen-activated protein kinase cascade. Pharmacol Ther 1995;67(3):323-350.
8. Wolanin PM, Thomason PA, Stock JB. Histidine protein kinases: key signal transducers outside the animal kingdom. Genome Biol 2002;3(10):REVIEWS3013.
9. Pawlowski K, Klosse U, de Bruijn FJ. Characterization of a novel Azorhizobium caulinodans ORS571 two- component regulatory system, NtrY/NtrX, involved in nitrogen fixation and metabolism. Mol Gen Genet 1991;231(1):124-138.
10. Gracey AY, Troll JV, Somero GN. Hypoxia-induced gene expression profiling in the euryoxic fish Gillichthys mirabilis. Proc Natl Acad Sci U S A 2001;98(4):1993-1998.
11. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2002;30(1):276-280.
12. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29(14):2994-3005.
13. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. Multiple sequence alignment with Clustal X. Trends Biochem Sci 1998;23(10):403-405.
14. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14(9):755-763.
15. Pearson WR. Effective protein sequence comparison. Methods Enzymol 1996;266:227-258.
16. Pietrokovski S. Searching databases of conserved sequence regions by aligning protein multiple-alignments. Nucleic Acids Res 1996;24(19):3836-3845.
17. Sadreyev R, Grishin N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 2003;326(1):317-336.

7 Systematic identification of immunoreceptor tyrosine-based motifs in the human proteome

Die folgende Erklärung legt die individuellen Beiträge der Co-Autoren zum nachfolgenden Manuskript dar. Von den Co-Autoren unterschriebene Fassungen dieser Erklärung liegen dieser Dissertation bei. Die Erklärungen sollen belegen, dass die erzielten Resultate im wesentlichen auf meiner Arbeit beruhen und ihre Verwendung innerhalb dieser Dissertation gerechtfertigt ist.

The ITIM manuscript: declaration about contributions of authors

Hereby I, co-author of the manuscript *„Systematic identification of immunoreceptor tyrosine- based inibitory motifs (ITIMs) in the human proteome“* which is accepted for publication in *Cellular Signalling*, declare my contributions to the results and to the preparation of the manuscript. Furthermore, I agree that the statements below describe the contributions of the other co-authors correctly.

1) Eike Staub
- planned and conducted the whole motif search strategy,
- conducted the sequence searches for the identification of orthologous mouse genes,
- determined the RNA expression levels of selected genes from public expression data,
- interpreted the results regarding the meaning of ITIM signalling in diverse mammalian tissues,
- wrote the text and prepared the figures for the manuscript,
- served as the corresponding author during the review process.

2) André Rosenthal
- served as a discussion partner in the strategic planning of manuscript preparation,
- contributed to the final stages of manuscript preparation by helpful comments on the style of the manuscript.

3) Bernd Hinzmann
- was the supervisor of the project,
- served as a discussion partner in the strategic planning of manuscript preparation,
- contributed to the final stages of manuscript preparation by helpful comments on the style of the manuscript.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Cellular Signalling xx (2003) xxx–xxx
CELLULAR SIGNALLING
www.elsevier.com/locate/cellsig

Systematic identification of immunoreceptor tyrosine-based inhibitory motifs in the human proteome

Eike Staub*, André Rosenthal, Bernd Hinzmann

MetaGen Pharmaceuticals GmbH, Oudenarderstr. 16, 13347 Berlin, Germany

Received 25 July 2003; received in revised form 11 September 2003; accepted 11 September 2003

Abstract

Immunoreceptor tyrosine-based inhibitory motifs (ITIMs) are short sequences of the consensus {ILV}-x-x-Y-x-{LV} in the cytoplasmic tail of immune receptors. The phosphorylation of tyrosines in ITIMs is known to be an important signalling mechanism regulating the activation of immune cells. The shortness of the motif makes it difficult to predict ITIMs in large protein databases. Simple pattern searches find ITIMs in ~30% of the protein sequences in the RefSeq database. The majority are false positive predictions. We propose a new database search strategy for ITIM-bearing transmembrane receptors based on the use of sequence context, i.e. the predictions of signal peptides, transmembrane helices (TMs) and protein domains. Our new protocol allowed us to narrow down the number of potential human ITIM receptors to 109 proteins (0.7% of RefPep). Of these, 36 have been described as ITIM receptors in the literature before. Many ITIMs are conserved between orthologous human and mouse proteins which represent novel ITIM receptor candidates. Publicly available DNA array expression data revealed that ITIM receptors are not exclusively expressed in blood cells. We hypothesise that ITIM signalling is not restricted to immune cells, but also functions in diverse solid organs of mouse and man.

Keywords: ITIM; NK cell activation; KIR; LIR; SIGLEC

1. Introduction

In recent years, much progress has been made towards the elucidation of molecular mechanisms governing the activation of immune cells. The activation of natural killer (NK) cells results in the lysis of abnormal cells of a host. Mechanisms by which NK cells distinguish healthy cells from infected or aberrant cells begin to emerge. According to a current model, NK cells sense the level of class-I MHC-like molecules on potential target cells. If potential target cells expose abnormally low numbers of MHC-I-like molecules on their surface, as it is the case in virally infected or transformed cells, then NK cells do no longer recognise these cells as "self". The missing "self" signal then triggers the activation of NK cells. This model, known as the "missing self" concept [1–3], is supported by a number of molecular studies. It seems that signals from activating and inhibitory receptors on NK cells sense the expression of "self" molecules on target cells which are proteins of the MHC class I family. It turned out that activation is the default signal for NK cells which has to be overruled by an inhibitory signal. Inhibitory receptors recognise MHC class I-like "self" proteins on target cells. If a potential aberrant target cell of a NK cell exposes abnormally low numbers of MHC I-like proteins on its surface, activation of the NK cell may be triggered by the relaxation of inhibition from inhibitory receptors, a mode of negative regulation that is also used by other cell types [2–4].

The activation signal is transmitted by several types of receptors. The most prominent group of activators is a family of Ig-like receptors on NK cells comprising NKp46,

Abbreviations: NK, natural killer; ITIM, immunoreceptor tyrosine-based inhibitory motif; ITAM, immunoreceptor tyrosine-based activation motif; Ig, immunoglobulin; KIR, killer cell Ig-like receptor; LIR, leukocyte Ig-like receptor; SH2 domain, Src-homology domain 2; SH3, Src-homology domain 3; NCR, natural cytotoxicity receptors; SHP, SH2 domain-containing phosphatase; LAIR, leukocyte cellular adhesion inhibitory receptor; IRTA, immunoglobulin superfamily receptor translocation associated; EpCAM, epithelial cell adhesion molecule; MHC, major histocompatibility complex; FN3, fibronectin 3; EGF, epidermal growth factor.

* Corresponding author. Present address: Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany. Tel.: +49-30-49910136, mob. +49-179-2439761; fax: +49-1212-5-11701401.
E-mail address: eike_staub@web.de (E. Staub).

0898-6568/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.cellsig.2003.08.013

NKp44 and NKp30, originally termed natural cytotoxicity receptors (NCR) [5]. A second group of activators is formed by some members of the killer cell Ig-like receptor (KIR) family. Whereas many KIRs exert inhibitory functions, those KIRs that have short cytoplasmic tails seem to activate NK or T cells. Activation via KIRs with short cytoplasmic tails involves the binding to adapters like DAP10/DAP12 which have ITAMs in their membrane region [6,7]. The ITAMs are thought to be phosphorylated by src-family kinases, thus generating binding sites for Syk kinase which transmits downstream signals for activation into the cytoplasm [8,9].

The inhibitory KIRs are characterised by the presence of small sequence motifs, the immune receptor tyrosine-based inhibitory motifs (ITIMs), that reside in their cytoplasmic tail. ITIM-bearing receptors form signalling complexes with activating receptors, leading to the phosphorylation of their ITIMs [10]. The inhibitory mechanism was shown to be dominant and molecule-specific, suggesting a local effect on the cell surface [3,11]. This is consistent with the observation that NK cells can specifically kill one contacting cell while leaving another contacting cell intact. The activation signals are turned off when the phosphorylated ITIMs recruit SH2-domains of phosphatases like SHP-1 or SHP-2 to the signalling complexes [3]. In turn, the SHP phosphatases dephosphorylate the activating receptors, thereby shutting down the signal. The molecular targets of SHP-like phosphatases have been reviewed recently [12].

On NK cells, three major subfamilies of inhibitory receptors have been defined. The best characterised subfamily comprises the inhibitory members of the KIR family. A second family of inhibitory receptors has been identified which is not restricted to NK cells but also binds MHC I-like ligands via its Ig-like domains. The gene family was first named Ig-like transcripts (ILT), later LIR for leukocyte Ig-like receptors and MIR for macrophage Ig-like receptors. Similar to KIRs, the ILTs can be subdivided in members with short or long cytoplasmic tails. The third group of inhibitory receptors on NK cells is formed by C-type lectin-like inhibitory receptors (CLIR) [3]. In contrast to KIRs and ILTs, the CLIR receptors are type II transmembrane proteins with an amino-terminal cytoplasmic tail. All three subfamilies are thought to bind MHC I-like molecules, so-called markers of "self", which are surface-exposed or secreted by potential target cells of NK cells.

Although the general concept of ITIM-mediated inhibitory effects was developed with NK cells as a model system, ITIMs seem to be of general importance for inhibitory effects in blood cells. Members of the SIGLEC family of ITIM-comprising type I transmembrane proteins function in the regulation of diverse types of leukocytes [13–15]. In the process of B-cell activation, signalling through the antigen binding B-cell receptor (BCR) usually activates extracellular signal-related kinase (ERK) and induces calcium mobilisation. The negative modulation of this signal cascade depends on an ITIM of the CD72 receptor [16,17].

With this extension of inhibitory ITIM signalling concept to diverse types of blood cells in mind, it is intriguing that a discrepancy exists in the tissue expression of the known ITIM signalling receptors and of the downstream phosphatases. Whereas the expression of known ITIM receptors is often restricted to blood cells, the phosphatases SHP-1 and SHP-2 are expressed in a variety of other organs [18,19]. This raises the question about the functional context of SHP-1/SHP-2-like phosphatases in these tissues. At present, we do not know to which other signal transduction pathways SHP-1/SHP-2-like phosphatases belong to and to which other receptors they may bind. Similarly, it is currently unknown how large the pool of interaction partners for the known ITIMs really is. How many other SH2-domain-containing proteins are present in the cytoplasm that are also able to interact with ITIMs? And do these also modulate activation signals?

A prerequisite for a broader view on ITIM signalling is the identification of an exhaustive set of potential ITIM receptors from the human proteome. The amino acid consensus pattern of the ITIMs in the cytoplasmic tails of inhibitory immune receptors was recently defined [20,21]. To have an estimate of the complexity of ITIM-dependent signalling at the inner surface of human cells, we developed a protocol for the identification of signalling receptors with ITIM motifs from protein sequence databases. We restricted the search to type I transmembrane proteins because they comprise the vast majority of known ITIM receptors and because many known signalling receptors belong to this class. Detection of potential type II ITIM receptors like the NKG2A-like receptors cannot be achieved using our new method. This seems to be reasonable to us, because the amino-terminal transmembrane helices of type II ITIM receptors are often hard to distinguish from signal peptides, making this subclass of ITIM receptors less suitable for an automatic analysis. The protocol was applied to a set of well-annotated human protein sequences. Furthermore, we analysed the conservation of the identified ITIM motifs in orthologous receptors of the mouse. We determined the expression pattern of a large fraction of ITIM receptors in human tissues by using publicly available data from DNA array experiments and discussed the implications of our findings.

2. Materials and methods

2.1. Protein sequence analysis pipeline

We used the sequences of human proteins from the NCBI RefSeq project (16886 sequences on the 20th of August 2002) as a starting point for our analysis [22]. Transmembrane helices (TMs) in the proteins were predicted using the software TMHMM version 2.0 [23]. Signal peptides were predicted using SIGFIND, a new signal peptide prediction program based on recurrent neural networks (http://

www.stepc.gr/~synaptic/sigfind.html) that was recently shown to be among the best signal peptide prediction programs available [24]. Extracellular protein domains were predicted using the extracellular subsection of the SMART domain database [25]. For database searches, we used profile Hidden Markov Models (HMMs) of the SMART alignments and applied the hmm search program of the HMMER package [26]. We used domain-specific expectation (E) value thresholds that are available for all SMART HMMs. For the detection of ITIM motifs, we applied our own regular expression searches using scripts written in the Perl programming language. Such scripts were also generated for other tasks like the removal of signal peptides and the extraction of putative C-terminal cytoplasmic tails from the protein sequences. The complete protocol of the extraction of ITIMs is depicted in Fig. 1.

2.2. Identification of orthologous mouse proteins

For all human transmembrane proteins with ITIMs in their putative cytoplasmic portion, we applied an automatic protocol to identify the orthologous mouse proteins [27]. A requirement for the correct assignment of orthology is the use of a complete proteome of an organism, because one has to be sure not to miss the real ortholog. Therefore, we used the ENSEMBL protein databases of *Homo sapiens* and *Mus musculus* to generate a set of orthologous sequence pairs from mouse and human [28]. We used each protein sequence from one ENSEMBL database and searched it in the other ENSEMBL database and vice versa. Sequence similarities with expectation (E) values below 1×10^{-6} were regarded as positive indicators of homology. Two proteins were regarded as a most probable pair of orthologs when both identified each other as the best hit in the reciprocal BLASTP searches. Additionally, the longest region of local sequence similarity, the BLAST high-scoring segment pair, had to cover at least 85% of each sequence and the sequence identity in this region had to exceed 70%. A weakness of this reciprocal BLAST best-hit method of orthology assignment is the detection of so-called in-paralogs that have arisen by duplications of one of the orthologs after the divergence of the species under consideration. However, it is a quick and straightforward method which allows one to identify the closest related genes between two species as measured by pairwise sequence similarity. It is therefore frequently applied in large-scale sequence analysis projects [27,29–31].

We matched the identified ITIM-containing receptors represented as RefSeq protein sequences to the orthologous pairs of human and mouse proteins from ENSEMBL by BLASTP searches. RefSeq and ENSEMBL proteins with alignments equal or above the 98% threshold were regarded as identical. For the orthologous mouse proteins, we predicted sequence features like protein domains, TM helices, signal peptides and ITIMs as described above.

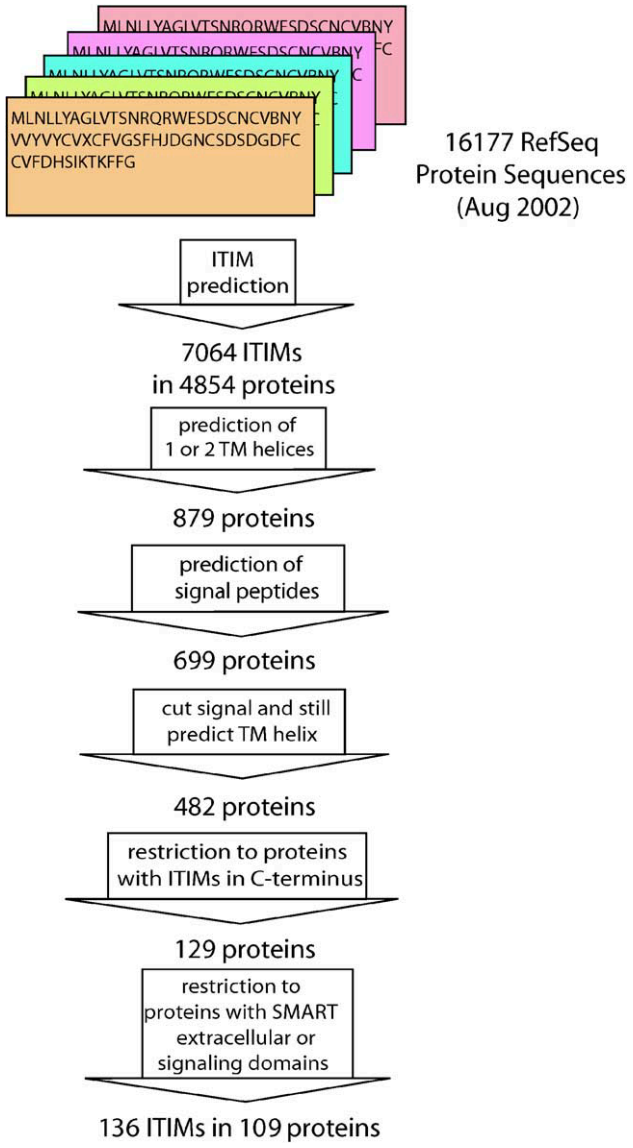

Fig. 1. Sequence analysis protocol for the detection of type I receptors with ITIMs in their cytoplasmic tail. Starting with 16177 human protein sequences of the RefSeq database, we identified 4854 proteins comprising sequence stretches which fit the ITIM consensus. Eight hundred seventy-nine of these proteins were predicted to contain one or two transmembrane helices. Initially, we included proteins with predictions of two helices because TM helix prediction programs often have problems in distinguishing signal peptides from transmembrane segments due to their similar biochemical composition. Six hundred ninety-nine of these proteins are predicted to comprise a leading signal peptide, a prerequisite for transmembrane proteins with N-termini located at the extracellular side of the cell. After removal of these peptides from the sequences, we still obtained 482 proteins that are predicted to have a TM helix and thus can be considered as type-I receptors. When we further restricted this set only to proteins which have ITIMs in regions that are C-terminal to the transmembrane helix, we arrived at a set of 129 proteins. One hundred nine of these contain known extracellular or signalling protein domains that are represented by alignments and HMMs in the SMART database (see Table 1).

Table 1
ITIM-bearing type I transmembrane proteins of *H. sapiens* which are present in the RefSeq protein database

Human RefPep protein ID	Human GENPEP GI	HUGO gene symbol	Known	Description	TM helices	ITIMs	Smart domain assignments
NP_004293.1	4757720	ACVR1B		activin A type IB receptor precursor; serine(threonine) protein kinase	(127–149)	(LPYYDL:428–433)	GS(177–207), S_TKc(207–497)
NP_064732.1	10862692	ACVR1B		activin A type IB receptor, isoform b precursor	(127–149)	(LPYYDL:428–433)	GS(177–207), S_TKc(207–469)
NP_068713.2	21536466	AXL		AXL receptor tyrosine kinase isoform 1; AXL transforming sequence/gene; oncogene AXL	(450–472)	(LLYSRL:632–637)	IG(41–136), IG(145–224), FN3(225–318), FN3(334–415), S_TKc(536–810), TyrKc(536–803)
NP_001690.2	21536468	AXL		AXL receptor tyrosine kinase isoform 2; AXL transforming sequence/gene; oncogene AXL	(441–463)	(LLYSRL:623–628)	IG(41–136), IG(145–224), FN3(225–318), FN3(334–415), S_TKc(527–801), TyrKc(527–794)
NP_006698.1	5729748	BTLN3		butyrophilin-like 3; butyrophilin-like receptor	(167–189)	(LIYTLL:394–399)	IG(25–132), PRY(253–303), SPRY(304–430)
NP_612116.1	19913375	C6orf25	yes	G6B protein isoform G6b-B precursor; G6B protein; immunoglobulin receptor	(143–165)	(LLYADL:209–214)	IG(20–123)
NP_001762.1	4502651	CD22		CD22 antigen; Siglec 2	(684–706)	(ISYTTL:760–765), (VTYSAL:794–799), (IHYSEL:820–825), (VDYVIL:840–845)	IG(29–137), IG(146–239), IG(250–328), IG(338–416), IG(427–502), IG(514–590), IG(601–678)
NP_001763.1	4502655	CD33		CD33 antigen; Siglec 3; gp67	(260–282)	(LHYASL:338–343)	IG(26–139), IG(148–232)
NP_001788.2	16306532	CDH11		cadherin 11, type 2, isoform 1 preproprotein; cadherin-11; OB-cadherin; osteoblast cadherin	(618–640)	(LDYDYL:767–772)	CA(76–157), CA(181–266), CA(290–382), CA(405–486), CA(513–600)
NP_001786.1	4502727	CDH5		cadherin 5, type 2 preproprotein; 7B4 antigen; VE-cadherin; vascular endothelial cadherin;endothelial-specific cadherin; cd144 antigen	(598–620)	(VDYDFL:755–760)	CA(68–149), CA(173–256), CA(280–371), CA(393–477), CA(500–583)
NP_001703.2	19923195	CEACAM1	yes	carcinoembryonic antigen-related cell adhesion molecule 1 (biliary glycoprotein)	(433–455)	(VTYSTL:491–496), (IIYSEV:518–523)	IG(40–141), IG(152–234), IG(244–318), IG(333–415)
NP_000386.1	4559408	CSF2RB		colony stimulating factor 2 receptor, beta, low-affinity (granulocyte–macrophage); interleukin 3 receptor/granulocyte–macrophage colony stimulating factor 3 receptor, beta (high affinity)	(443–465)	(LEYLCL:626–631)	FN3(133–221), FN3(339–422)
NP_000751.1	4503081	CSF3R		Colony-stimulating factor-3 receptor (granulocyte)	(626–648)	(VLYGQL:750–755)	FN3(123–212), FN3(237–316), FN3(333–417), FN3(429–515), FN3(527–609)
NP_005206.1	4885175	DCC		deleted in colorectal carcinoma	(1099–1121)	(VPYTPL:1361–1366)	IG(46–137), IG(146–231), IG(246–328), IG(337–418), IG(425–520), FN3(429–511), FN3(528–607), FN3(622–705), FN3(726–805), FN3(844–929), FN3(945–1031)

NP_054699.1	7669483	DDR1		discoidin receptor tyrosine kinase isoform a; PTK3A protein tyrosine kinase 3A; cell adhesion kinase; epithelial discoidin domain receptor 1; neurotrophic tyrosine kinase, receptor, type 4	(417–439)	(ISYPML:738–743)	FA58C(30–185), TyrKc(610–905)
NP_001945.2	7669487	DDR1		discoidin receptor tyrosine kinase isoform b	(417–439)	(ISYPML:700–705)	FA58C(30–185), S_TKc(572–867), TyrKc(572–867)
NP_054700.1	7669485	DDR1		discoidin receptor tyrosine kinase isoform c	(417–439)	(ISYPML:744–749)	FA58C(30–185), S_TKc(610–911), TyrKc(610–911)
NP_006173.1	5453814	DDR2		discoidin domain receptor family, member 2 precursor; neurotrophic tyrosine kinase, receptor-related 3	(399–421)	(VSYTNL:682–687)	FA58C(29–185), S_TKc(563–849), TyrKc(563–849)
NP_060893.1	8922179	DKFZp761P1010		hypothetical protein DKFZp761P1010	(26–48)	(LLYEMV:315–320)	S_TKc(114–384), TyrKc(114–380)
NP_005609.2	10518497	DLL1		delta-like 1; delta-like 1 (mouse) homolog; delta-like 1 protein; delta (Drosophila)-like 1	(545–567)	(VDYNLV:639–644)	DSL(159–221) , EGF(256–286), EGF(291–326), EGF(331–364), EGF(369–403), EGF(408–441), EGF(446–479), EGF(484–517), EGF(225–255)
NP_115740.3	21361589	DRIP78		dopamine receptor interacting protein; LYST-interacting protein LIP6	(37–59)	(ITYFAL:294–299)	DnaJ(152–209)
NP_001380.2	20127422	DSCAM		Down syndrome cell adhesion molecule; human CHD2-52 down syndrome cell adhesion molecule	(1595–1617)	(VHYQSV:1706–1711)	IGc2(37–109), IG(130–218), IG(231–311), IG(320–403), IG(413–502), IG(510–594), I IG(602–687), IG(696–785), IG(794–883), FN3(885–969), FN3(985–1073), FN3(1088–1174), FN3(1189–1270), IG(1292–1377), FN3(1380–1460), FN3(1477–1557)
NP_001934.1	4503403	DSG2		desmoglein 2 preproprotein; HDGC, included	(612–634)	(VPYVMV:1010–1015)	CA(69–156), CA(180–269) , CA(292–386), CA(412–496)
NP_005223.1	4885209	EPHA1		EphA1; eph tyrosine kinase 1 (erythropoietin-producing hepatoma amplified sequence; oncogene EPH; ephrin receptor EphA1)	(548–570)	(IPYRTV:921–926)	EPH_lbd(27–204), FN3(333–429), FN3(447–525), S_TKc(632–892), TyrKc(632–888), SAM(918–984)
NP_004422.1	4758278	EPHA2		EphA2; ephrin receptor EphA2; epithelial cell receptor protein tyrosine kinase	(536–558)	(IAYSLL:958–963)	EPH_lbd(28–201), FN3(436–516), FN3(329–419), S_TKc(613–875), TyrKc(613–871), SAM(901–968)
NP_065387.1	10140845	EPHA8		ephrin receptor EphA8 precursor; eph- and elk-related tyrosine kinase; tyrosine-protein kinase receptor eek; hydroxyaryl-protein kinase	(541–563)	(VCYGRL:649–654)	EPH_lbd(31–204), FN3(440–521), FN3(329–419), S_TKc(635–896), TyrKc(635–892), SAM(927–994)
NP_000112.1	4503591	EPOR		erythropoietin receptor precursor	(251–273)	(LKYLYL:452–457)	FN3(145–228)
NP_061008.2	19923536	ERMAP		erythroblast membrane-associated protein; erythroid membrane-associated protein	(155–177)	(LLYEHV:186–191)	IG(35–144), PRY(237–289), SPRY(290–415)
NP_003992.2	21361191	FCGR2B	yes	Fc fragment of IgG, low affinity IIb, receptor for (CD32)	(217–239)	(ITYSLL:283–288)	IG(49–122), IG(130–208)

(continued on next page)

Table 1 (*continued*)

Human RefPep protein ID	Human GENPEP GI	HUGO gene symbol	Known	Description	TM helices	ITIMs	Smart domain assignments
NP_443171.2	21314764	FCRH3	yes	Fc receptor-like protein 3; SH2 domain-containing phosphatase anchor protein 2	(572–594)	(VLYSEL:690–695)	IG(29–96), IG(105–186), IG(196–281), IG(294–375), IG(389–470), IG(482–563)
NP_075052.1	12597641	FLJ21302		hypothetical protein FLJ21302	(269–291)	(IWYNIL:295–300)	LRR_TYP(75–95), LRR_TYP(96–119), LRR_TYP(120–143), LRR_TYP(144–168), LRRCT(176–227)
NP_002560.1	4505579	FURIN		furin preproprotein; proprotein convertase subtilisin/kexin type 3; FES upstream region (FUR); dibasic processing enzyme	(716–738)	(ISYKGL:757–762)	FU(577–620), FU(638–681)
NP_004954.1	4826752	GUCYC2C		guanylate cyclase 2C (heat stable enterotoxin receptor)	(432–454)	(IDYYNL:539–544), (LEYLQL:1058–1063)	S_TKc(500–752), TyrKc(494–745), CYCc(788–983)
NP_005525.1	5031783	IFNGR2		interferon gamma receptor 2 (interferon gamma transducer 1); interferon gamma receptor accessory actor-1; interferon-gamma receptor beta chain precursor	(248–270)	(LKYRGL:271–276)	FN3(140–221)
NP_000866.1	4557665	IGF1R		insulin-like growth factor 1 receptor precursor	(936–958)	(VLYASV:971–976)	FU(227–270), FN3(489–592), FN3(611–814), FN3(832–914), S_TKc(999–1267), TyrKc(999–1266)
NP_065840.1	21357327	IGSF9		immunoglobulin superfamily, member 9	(722–744)	(LQYLSL:908–913)	IG(26–131), IG(143–224), IG(233–320), IG(408–488), FN3(492–577), FN3(608–687)
NP_003846.1	4504655	IL18R1		interleukin 18 receptor 1	(331–353)	(LFYRHL:357–362)	IG(25–112), IG(125–207), IG(222–316), TIR(374–523)
NP_002173.1	4504661	IL1RAP		interleukin 1 receptor accessory protein, membrane form	(360–382)	(VQYKAV:501–506), (LSYSSL:562–567)	IG(32–132), IG(145–232), IG(251–350), TIR(404–549)
NP_059112.1	11225607	IL1RAPL2		interleukin 1 receptor accessory protein-like 2	(357–379)	(LSYTKV:406–411)	IG(38–134), IG(149–232), IG(250–349), TIR(401–559)
NP_002175.1	4504675	IL6ST		interleukin 6 signal transducer (gp130, oncostatin M receptor)	(620–642)	(VQYSTV:757–762)	FN3(222–308), FN3(327–409) , FN3(424–505) , FN3(519–602)
NP_057331.1	7706717	IMPG2		interphotoreceptor matrix proteoglycan 200; Spacrcan protein	(1102–1124)	(VKYNPV:1157–1162)	SEA(239–349), SEA(892–1015)
NP_112572.1	14550416	IRTA1	yes	immunoglobulin superfamily receptor translocation associated 1	(386–408)	(VVYSEV:491–496)	IG(29–101), IG(108–186), IG(197–282), IG(295–378)
NP_112571.1	14550414	IRTA2	yes	immunoglobulin superfamily receptor translocation associated 2	(851–873)	(VVYSEV:922–927), (IIYSEV:952–957)	IG(29–101), IG(108–186), IG(196–281), IG(293–374), IG(386–467), IG(479–560), IG(572–653), IG(665–746), IG(758–838)
NP_443163.1	16418407	KALI	yes	activating NK receptor precursor; natural killer-, T- and B-cell antigen; NTBA receptor	(226–248)	(LEYVSV:271–276)	IG(27–127)
NP_055573.1	7662026	KIAA0254		KIAA0254 gene product	(53–75)	(LAYHTV:574–579), (VQYLRL:874–879)	PXA(95–272), PX(527–659)
NP_055033.1	7657271	KIR2DL1	yes	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 1; 47.11; cl-42; p58.1	(243–265)	(VTYTQL:300–305), (IVYTEL:330–335)	IG(34–114), IG(134–221)

NP_055034.1	7657273	KIR2DL2	yes	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 2; cl-43; nkat6; p58.2	(243–265)	(VTYTQL:300–305), (IVYAEL:330–335)	IG(34–114), IG(134–221)
NP_055326.1	7657275	KIR2DL3	yes	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 3; cl-6; nkat2; nkat2a; nkat2b; p58 natural killer cell receptor family; p58.2	(243–265)	(VTYAQL:301–306), (IVYTEL:331–336)	IG(34–114), IG(134–221)
NP_056952.1	7705568	KIR2DL3	yes	NK-receptor	(243–265)	(VTYAQL:301–306), (IVYTEL:331–336)	IG(34–114), IG(134–221)
NP_002246.2	20149517	KIR2DL4	yes	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 4; 15.212; 103AS	(243–265)	(VTYAQL:298–303)	IG(36–120), IG(131–218)
NP_065396.1	11968154	KIR2DL5	yes	killer cell immunoglobulin-like receptor, two domains, long cyto	(241–263)	(VTYAQL:296–301)	IG(34–116), IG(129–216)
NP_037421.1	7019441	KIR3DL1	yes	killer cell immunoglobulin-like receptor, three domains, long cytoplasmic tail, 1; cl-11; cl-2; nkat3; p70	(341–363)	(VTYAQL:396–401), (ILYTEL:426–431)	IG(34–118), IG(129–209), IG(229–316)
NP_006728.1	5803052	KIR3DL2	yes	killer cell immunoglobulin-like receptor, three domains, long cytoplasmic tail, 2; p140; cl-5; nkat4; nkat4a; nkat4b	(339–361)	(VTYAQL:396–401)	IG(34–118), IG(129–209), IG(229–316)
NP_002278.1	4504943	LAIR1	yes	leukocyte-associated Ig-like receptor 1, isoform a precursor; leukocyte-associated immunoglobulin-like receptor 1	(164–186)	(VTYAQL:249–254), (ITYAAV:279–284)	IG(34–121)
NP_068352.1	11231177	LAIR1	yes	leukocyte-associated Ig-like receptor 1, isoform b precursor; leukocyte-associated immunoglobulin-like receptor 1	(147–169)	(VTYAQL:232–237), (ITYAAV:262–267)	IG(34–121)
NP_002301.1	4504993	LIFR		leukemia inhibitory factor receptor precursor	(835–857)	(VIYIDV:972–977)	FN3(47–121), FN3(333–419), FN3(433–520), FN3(535–616), FN3(724–820)
NP_006660.1	5729927	LILRB1	yes	leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 1; leukocyte immunoglobulin-like receptor 1; CD85 antigen	(461–483)	(VTYAEV:560–565), (VTYAQL:612–617)	IG(34–119), IG(130–220), IG(231–319), IG(331–420)
NP_005865.1	5031911	LILRB2	yes	leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 2; leukocyte immunoglobulin-like receptor 2	(461–483)	(VTYAQL:560–565)	IG(34–117), IG(230–318), IG(330–419)
NP_006855.1	5803060	LILRB3	yes	leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 3; leukocyte immunoglobulin-like receptor 3	(442–464)	(VTYAPV:541–546), (VTYAQL:593–598)	IG(34–118), IG(230–316), IG(330–419)
NP_006838.2	21314641	LILRB4	yes	leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 4; leukocyte immunoglobulin-like receptor 5	(258–280)	(VTYAKV:358–363), (VTYARL:410–415)	IG(34–118), IG(129–218)

(continued on next page)

Table 1 (*continued*)

Human RefPep protein ID	Human GENPEP GI	HUGO gene symbol	Known	Description	TM helices	ITIMs	Smart domain assignments
NP_006831.1	5803070	LILRB5	yes	leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 5	(457–479)	(VTYAQL:552–557)	IG(34–118), IG(129–215), IG(229–317), IG(329–418)
NP_004819.1	4758694	NCR2	yes	natural cytotoxicity triggering receptor 2; lymphocyte antigen 95; activating NK-receptor; NK-p44	(193–215)	(ILYHTV:257–262)	IG(25–130)
NP_006334.1	5453738	MERTK		c-mer proto-oncogene tyrosine kinase	(502–524)	(LLYSRL:683–688)	IG(100–194), IG(203–281), FN3(284–368), FN3(384–470), S_TKc(587–861), TyrKc(587–854)
NP_000521.1	4505243	MPZ		myelin protein zero (Charcot-Marie-Tooth neuropathy 1B)	(167–189)	(VLYAML:228–233)	IG(45–157)
NP_003944.1	4506357	MPZL1		myelin protein zero-like 1; protein zero related	(169–191)	(VIYAQL:239–244)	IG(43–155)
NP_666017.1	22095335	NFAM1		NFAT activation molecule 1	(164–186)	(LVYENL:265-270)	IG(50–150)
NP_620587.1	20502982	NKIR		NK inhibitory receptor precursor	(156–178)	(LCYADL:203–208), (ISYASL:247–252)	IG(25–126)
NP_006363.2	21361312	NSAP1		NS1-associated protein 1	(21–43)	(LTYVLL:856–861)	PXA(108–253), RGS(285–417), PX(507–626)
NP_056428.1	14149694	PAL		retina specific protein PAL	(530–552)	(VTYVNL:569–574)	LRRNT(22–63), LRR_TYP(106–129), LRR_TYP(82–105), LRR_TYP(131–153), LRR_TYP(154–177), LRRCT(201–252), IG(260–346), FN3(425–505)
NP_116755.1	14589946	PCDH11Y		protocadherin 11 Y-linked isoform c; protocadherin Y; protocadherin 22	(34–56)	(VRYSIV:743–748)	CA(193–279), CA(303–385), CA(415–496), CA(520–600), CA(624–703), CA(730–812)
NP_002579.2	14277675	PCDHGC3		protocadherin gamma subfamily C, 3, isoform 1 precursor; protocadherin 2; protocadherin 43; cadherin-like 2	(693–715)	(VFYRQV:795–800)	CA(51–133), CA(157–242), CA(266–350), CA(374–455), CA(479–565), CA(593–672),
NP_115778.1	14277677	PCDHGC3		protocadherin gamma subfamily C, 3, isoform 2 precursor; protocadherin 2; protocadherin 43; cadherin-like 2	(693–715)	(VFYRQV:795–800)	CA(51–133), CA(157–242), CA(266–350), CA(374–455), CA(479–565), CA(593–672),
NP_061752.1	11128023	PCDHGC5		protocadherin gamma subfamily C, 5, isoform 1 precursor	(690–712)	(LKYMEV:757–762)	CA(33–131), CA(264–348), CA(372–452), CA(476–562), CA(592–674),
NP_115783.1	14277685	PCDHGC5		protocadherin gamma subfamily C, 5, isoform 2 precursor	(690–712)	(LKYMEV:757–762)	CA(33–131), CA(155–240), CA(264–348), CA(372–452), CA(476–562), CA(592–674),
NP_005009.1	4826890	PDCD1		programmed cell death 1 precursor	(168–190)	(VDYGEL:221–226)	IG(39–145)
NP_002600.1	4505683	PDGFRB		platelet-derived growth factor receptor beta precursor; beta	(534–556)	(LSYMDL:798–803), (VLYTAV:1007–1012)	IG(39–121), IG(220–311), IG(325–416), S_TKc(600–962), TyrKc(600–958)
NP_000433.2	21314617	PECAM1	yes	platelet/endothelial cell adhesion molecule (CD31 antigen); platelet/endothelial cell adhesion molecule	(603–625)	(VQYTEV:688–693)	IG(42–128), IG(332–403), IG(241–322), IGc2(338–393), IG(416–497), IG(508–595)
NP_038467.1	7305385	PILR	yes	paired immunoglobulin-like receptor alpha	(196–218)	(IVYASL:267–272)	IG(38–150)
NP_000940.1	4506107	PRLR		prolactin receptor	(236–258)	(VEYLEV:312–317)	FN3(28–112), FN3(127–215)

NP_002812.2	15826840	PTK7		PTK7 protein tyrosine kinase 7; TK7 protein tyrosine kinase	(704–726)	(LEYVDL:875–880)	IG(38–122), IG(231–320), IG(135–220), IG(328–409), IG(418–499), IG(508–589), IG(598–682), TyrKc(796–1061)
NP_542970.1	18426911	PTPNS1	yes	protein tyrosine phosphatase, non-receptor type substrate 1; myd-1 antigen; signal regulatory protein, alpha type 2; SHP substrate-1; signal regulatory protein, alpha type 1; tyrosine phosphatase SHP substrate 1; macrophage fusion receptor	(372–394)	(ITYADL:427–432), (LTYADL:468–473)	IG(40–145), IGc1(165–238), IG(258–350)
NP_004639.1	4758978	PTPNS1	yes	protein tyrosine phosphatase, non-receptor type substrate 1	(371–393)	(ITYADL:426–431), (LTYADL:467–472)	IG(40–144), IGc1(164–237), IG(257–349)
NP_109592.1	13677214	PTPRO		receptor-type protein tyrosine phosphatase O, isoform a precursor; glomerular epithelial protein-1; protein tyrosine phosphatase PTP-U2	(879–901)	(VIYENV:1208–1213)	FN3(433–518), FN3(530–616), FN3(632–712), FN3(723–802), PTPc(937–1197), PTPc_motif(1091–1194)
NP_002839.1	4506323	PTPRO		receptor-type protein tyrosine phosphatase O, isoform b precursor	(821–843)	(VIYENV:1180–1185)	FN3(723–802), FN3(530–616), FN3(433–518), FN3(632–712), PTPc(909–1169), PTPc_motif(1063–1166)
NP_109596.1	13677222	PTPRO		receptor-type protein tyrosine phosphatase O, isoform c precursor	(68–90)	(VIYENV:397–402)	PTPc(126–386), PTPc_motif(280–383)
NP_109594.1	13677218	PTPRO		receptor-type protein tyrosine phosphatase O, isoform c precursor; glomerular epithelial protein-1; protein tyrosine phosphatase PTP-U2; phosphotyrosine phosphatase U2; PTPase U2; PTPROt	(68–90)	(VIYENV:397–402)	PTPc(126–386), PTPc_motif(280–383)
NP_002842.1	4506329	PTPRZ1		protein tyrosine phosphatase, receptor-type, zeta polypeptide 1	(1639–1661)	(LAYYTV:1861–1866)	FN3(312–398), PTPc(1722–1993), PTPc_motif(1889–1990), PTPc(2021–2283), PTPc_motif(2181–2280)
NP_006496.2	19923372	PVR		Polio virus receptor	(345–367)	(VSYSAV:396–401)	IG(34–142), IG(253–330)
NP_004551.2	19743898	ROR2		receptor tyrosine kinase-like orphan eceptor 2 precursor; neurotrophic tyrosine kinase receptor-related 2	(403–425)	(LVYDKL:622–627)	IG(68–153), KR(314–396), TyrKc(473–746)
NP_002935.2	19924165	ROS1		proto-oncogene c-ros-1 protein precursor; v-ros avian UR2 sarcoma virus oncogene homolog 1	(1860–1882)	(LNYMVL:2272–2277), (LNYACL:2332–2337)	FN3(99–177), FN3(195–272), FN3(558–658), FN3(948–1027), FN3(1044–1137), FN3(1469–1539), FN3(1558–1642), FN3(1659–1738), FN3(1753–1839), TyrKc(1945–2215)
NP_149121.1	15055513	SIGLEC10	yes	sialic acid binding Ig-like lectin 10	(548–570)	(LDYINV:595–600), (LHYATL:665–670)	IG(26–140), IG(149–235), IG(261–341), IG(365–443)
NP_443116.1	16418393	SIGLEC11	yes	sialic acid binding Ig-like lectin 11	(550–572)	(LHYASL:630–635)	IG(27–141), IG(150–236), IG(260–340), IG(364–442)
NP_003821.1	4502659	SIGLEC5	yes	sialic acid binding Ig-like lectin 5; OB binding protein-2; CD33 antigen-like 2	(440–462)	(LHYASL:518–523)	IG(26–140), IG(149–233), IG(254–332)
NP_001236.1	4502657	SIGLEC6	yes	sialic acid binding Ig-like lectin 6; CD33 antigen-like 1	(333–355)	(LHYAVL:411–416)	IG(25–131), IG(142–222), IG(246–324)

(continued on next page)

Table 1 (*continued*)

Human RefPep protein ID	Human GENPEP GI	HUGO gene symbol	Known	Description	TM helices	ITIMs	Smart domain assignments
NP_055200.1	7657570	SIGLEC7	yes	sialic acid binding Ig-like lectin 7; adhesion inhibitory receptor molecule 1; D-siglec precursor	(354–376)	(IQYAPL:435–440)	IG(31–144), IG(153–237), IG(261–338)
NP_201586.1	16506826	SIGLECL1	yes	SIGLEC-like 1; sialic acid-binding immunoglobulin-like lectin-like gene	(362–384)	(IQYASL:445–450)	IG(38–151), IG(160–244), IG(268–346)
NP_443729.1	16506828	SIGLECL1	yes	SIGLEC-like 1; sialic acid-binding mmunoglobulin-like lectin-like gene	(480–502)	(IQYASL:563–568)	IG(29–142), IG(156–269), IG(278–362), IG(386–464)
NP_055947.1	17864088	SNX13		sorting nexin 13	(34–51)	(LLYLLL:231–236)	PXA(97–284), RGS(373–513), PX(563–676)
NP_620075.1	20270263	SPAP1	yes	SH2 domain-containing phosphatase anchor protein 1 (SPAP1) isoform a; Fc receptor-like protein 2	(147–169)	(VVYSQV:219–224), (VIYSSV:247–252)	IG(53–134)
NP_110391.2	19923629	SPAP1	yes	SH2 domain-containing phosphatase anchor protein 1 (SPAP1); Fc receptor-like protein 2	(400–422)	(VVYSQV:472–477), (VIYSSV:500–505)	IG(23–103), IG(113–198), IG(211–294), IG(306–387)
NP_004090.3	21361755	STOM		erythrocyte membrane protein band 7.2 (stomatin)	(32–54)	(VVYYRV:121–126)	PHB(52–211)
NP_003171.1	4507337	SYT5		synaptotagmin 5	(29–51)	(VPYVEL:181–186), (LDYDKL:329–334)	C2(124–226), C2(255–369)
NP_004603.1	4759226	TGFBR1		transforming growth factor, beta receptor I	(126–148)	(LPYYDL:426–431)	GS(175–205), S_TKc(205–471), TyrKc(205–492)
NP_005415.1	4885631	TIE		tyrosine kinase with immunoglobulin and epidermal growth factor homology domains	(764–786)	(LSYPVL:829–834)	IG(131–213), EGF(223–256), EGF(267–303), EGF(314–345), IG(357–444), FN3(446–530), FN3(543–631), FN3(644–726), TyrKc(839–1107)

NP_057646.1	7706093	TLR7	toll-like receptor 7	(843–865)	(VAYSQV:1039–1044)	LRRNT(33–69), LRR_TYP(126–149), LRR_TYP(203–226), LRR_TYP(227–247), LRR_TYP(289–312), LRR_SD22(313–334), LRR_SD22(420–440), LRR_TYP(396–419), LRR_TYP(516–540), LRR_TYP(541–564), LRR_TYP(595–618), LRR_TYP(649–672), LRR_TYP(698–721), LRR_TYP(723–745), LRR_TYP(748–769), LRRCT(783–834), TIR(890–1036)
NP_057694.2	20302166	TLR8	toll-like receptor 8, isoform 1	(844–866)	(LFYWDV:869–874)	LRR_TYP(84–103), LRR_TYP(142–165), LRR_TYP(218–241), LRR_TYP(242–262), LRR_TYP(304–327), LRR_TYP(328–352), LRR_TYP(354–378), LRR_TYP(384–407), LRR_TYP(411–434), LRR_TYP(435–458), LRR_TYP(601–623), LRR_TYP(656–679), LRR_TYP(705–728), LRR_TYP(729–752), LRR_TYP(753–776), LRRCT(790–841), TIR(897–1043)
NP_619542.1	20302168	TLR8	toll-like receptor 8, isoform 2	(826–848)	(LFYWDV:851–856)	LRR_TYP(66–85), LRR_TYP(200–223), LRR_TYP(124–147), LRR_TYP(224–244), LRR_TYP(286–309), LRR_TYP(310–334), LRR_TYP(336–360), LRR_TYP(366–389), LRR_TYP(393–416), LRR_TYP(417–440), LRR_TYP(583–605), LRR_TYP(638–661), LRR_TYP(687–710), LRR_TYP(711–734), LRR_TYP(735–758), LRRCT(772–823), TIR(879–1025)
NP_000585.1	10835155	TNF	tumor necrosis factor, alpha (cachectin)	(35–57)	(LIYSQV:133–138)	TNF(88–233)
NP_009199.1	6005958	Z39IG	Ig superfamily protein	(284–306)	(LDYEFL:386–391)	IG(26–136), IG(150–231)

The RefSeq and GenBank database identifiers for the human protein sequences are given in the first two columns, followed by the HUGO symbols of the genes which encode the proteins. The column "Known" indicates whether a protein was known before to comprise an ITIM. It is followed by the full-length name of the protein as given in the RefSeq header line. The last three columns contain the positions of the transmembrane helices, the ITIMs and the extracellular or signalling protein domains of each protein. The ITIM motifs are displayed in combination with their positions in the sequences. The protein domain abbreviations are the same as in the SMART database.

2.3. Expression profiles for transcripts of ITIM receptor-encoding genes

For the assignment of mRNA expression profiles to ITIM receptor genes, we used the large collection of expression profiles from the GeneAtlas experiments [32]. The Affymetrix U95 DNA chip with >12,000 probesets was used for this large-scale-expression analysis. We searched all sequences that are represented on the U95 DNA chip in the RefSeq set of mRNA sequences by BLASTN using an E-value threshold of 1e−30 and an identity threshold of 96%. Finally, we were able to link Affymetrix U95 gene-specific probesets to corresponding RefSeq proteins via their mRNA sequences.

The expression intensities in the original data set do not support easy interpretations as they are not based on an intuitive intensity scale. Therefore, we re-scaled the expression values of each experiment in the following manner. The Null-level of expression intensities was defined on the basis of the assumption that a maximum of 40% of genes is expressed in a distinct cell type. We also found that no more than 1% of all expression values per experiment reach the level of saturation. An intuitive and reasonable scaling was obtained by the construction of an ordinal scale setting the expression values of the 60th percentile and the 98.5th percentile after ranking all expression values of a single experiment to 0 and 10, respectively. This implies that only the top 1.5% of all expression values exceed the value 10 and that each expression value above 10 indicates that a gene is among the 1.5% highest expressed genes. The 0-level of expression in this normalisation method was found be in agreement with the experimentally supported minimum value of 200 (measured as Affymetrix Average Difference) to call a gene expressed or not in the original data. Our normalisation method is a general alternative to other expression data normalisation methods. In contrast to others, it is not largely based on the background noise level and avoids to consider saturated signals for normalisation.

3. Results and discussion

3.1. The sequence analysis pipeline reveals 94 human genes encoding ITIM-bearing type I receptors

Database search methods for larger protein domains representing independent structural units are advanced and accompanied by rigorous significance measures [26,33]. Detection of short motifs by pattern searches in protein sequence databases is complicated by a strong background noise which makes it difficult to distinguish a true positive motif from the large number of false positive signals. False positive signals can often be easily identified, when a motif only makes sense in a distinct cellular context. In the case of ITIMs, the presence of the amino acid consensus pattern in a metabolic enzyme or a transcription factor is obviously not indicative of a classical functional ITIM, as it would hardly be able to mediate signal transduction at the cellular surface. To overcome the problem of low specificity of pattern searches, we developed a search protocol aiming at the systematic identification of ITIMs in sequence databases. Our protocol uses information of pattern searches, protein domain context, predicted membranous localisation and conservation in orthologs. The main goal of our search strategy was to reduce the number of hits compared to usual pattern searches by using extra information that can be inferred from protein sequences.

When we applied pattern searches using the ITIM consensus sequences to the complete set of human protein sequences from the RefPep project [22], we identified 7064 ITIMs in 4854 out of the total number of 16,177 proteins (30%, see Fig. 1). The majority of these pattern-based predictions highlight proteins which obviously have no link to cytoplasmic ITIM signalling, e.g. metabolic enzymes or transcription factors. Many ITIMs were also identified in regions of proteins where they have no functional importance and thus represent false positive predictions, e.g. in extracellular protein domains. To narrow down the number of ITIMs, we decided to restrict the set of candidates to type I receptors because the large majority of currently known ITIMs resides in proteins of this family. Typical sequence features of type I receptors are the presence of a leading signal peptide which is cleaved during membrane insertion and an additional helical transmembrane domain. Because it is a weakness of current transmembrane helix prediction programs to distinguish between signal peptides and TM helices [34], we first filtered out all ITIM proteins with predictions of one or two TM helices. We then retained all proteins with an additional signal peptide prediction. Finally, we kept only those proteins in which the prediction of a single TM helix was retained after the removal of the signal sequence. This reduced the number of candidate ITIM receptors by 90% to 482 proteins. However, three quarters of these receptors comprises ITIM-like patterns in their extracellular domains. Only 129 type I transmembrane proteins have ITIMs in their cytoplasmic tail where they can function in phosphorylation-dependent signalling. We searched these 129 proteins for known protein domains by profile Hidden Markov Models from the SMART and PFAM databases. Only 21 proteins did not match known extracellular or signalling domains of the SMART database. In total, we identified 109 type I receptor proteins with known signalling domains and ITIMs in their cytoplasmic C-termini, representing 0.7% of all proteins from the start set. Compared to regular pattern searches alone, the additional use of sequence context resulted in a 45-fold reduction in the number of hits. The 109 transmembrane proteins are encoded by 94 different genes. A total of 27 proteins are products of 12 alternatively spliced mRNAs. All 109 proteins and their 94 HUGO gene symbols are displayed in Table 1.

3.2. The re-identification of known ITIM receptors shows the sensitivity of our algorithm

In numerous previous studies, 38 of the investigated human proteins have already been proposed to modulate immune cell behaviour via ITIM-dependent signalling. In the following, we discuss the results for each known family of ITIM receptors.

The classification of KIR proteins is based on the number and subtypes of Ig-like domains, the length of the cytoplasmic region (long, short, absent) and on the divergence in common regions of the sequence (2% threshold). Based on sequence similarity, seven KIR subfamilies were distinguished in the literature. In larger nucleotide databases, which are less well curated and therefore more redundant than RefSeq, more than hundred human KIR-like sequences exist. This is a result of the enormous variation between haplotypes [35,36]. We identified KIR protein variants with ITIMs encoded by seven genes, KIR2DL1, KIR2DL2, KIR2DL3, KIR2DL4, KIR2DL5, KIR3DL1, KIR3DL2. They represent all human genes that are regarded as KIR genes. This indicates that our protocol is performing well on known ITIM-bearing KIRs.

The second family of type I ITIM-bearing receptors is the ILT/LIR/MIR family which is encoded by a genomic locus called leukocyte receptor complex (LRC) on 19q13.4 proximal to the KIR cluster [37–41]. The locus shows a high number of different haplotypes and the differences between haplotypes are large, even in the number of functional KIR and ILT genes [36,38,42]. Haplotype analysis of the ILT/LIR/MIR cluster indicates that it is more stable with regard to the number of ILT/LIR/MIR genes than the KIR cluster. However, ILT6 is not necessarily a functional gene in all haplotypes [38]. Five out of thirteen reported ILT proteins are known to contain ITIMs [43]. During our screen, we detected all of them, LIR1/LILRB1, LIR2/LILRB2, LIR3/LILRB3, LIR5/LILRB4, and LIR8/LILRB5.

We also identified Ig-like receptors with ITIMs which do not clearly belong to the prominent classes of Ig-like receptors that were mentioned. The leukocyte-associated inhibitory receptor (LAIR-1), also encoded by the LCR on 19q13.4, comprises only one Ig-like domain and is expressed in various mononuclear leukocytes [44]. LAIR-1 inhibits NK cell and T-cell activation by ITIM-dependent SHP-1 and SHP-2 recruitment [45]. LAIR-1 does not bind MHC I-like ligands, but the colon carcinoma-associated epithelial cell adhesion molecule (EpCAM), which seems to be the natural ligand for LAIR-1, suggesting a role of LAIR-1 in the protection of colon mucosa from destruction by immune cells [46]. Additionally, we identified ITIMs in both, the Ig superfamily receptor genes translocation associated proteins IRTA1 and IRTA2, which are located in the human genomic region 1q21 near a putative chromosomal hotspot for translocations in B-cell malignancies [47]. Subsequently, three further Ig-like genes were found in the same region and two of these are ITIM receptors [48]. The SH2 domain-containing phosphatase anchor protein 1 (SPAP1), also-called Fc receptor-like protein 2 (FcRH2), occurs in isoforms with one or four Ig domains, but both have two predicted cytoplasmic ITIMs. The SH2 domain-containing phosphatase anchor protein 2 (SPAP2), alternatively called Fc receptor-like protein 3 (FcRH3), has six Ig domains and only a single ITIM.

The paired immunoglobulin-like receptor α (PILR) has been discovered by its ability to bind the phosphatase SHP-1 in an ITIM-dependent manner [49]. The NTBA receptor (HGNC symbol KALI) was described as an activating and an inhibitory receptor on NK cells and to bind the SH2 domain protein SH2D1A. In X-linked proliferative disease (XLP), the SH2D1A gene is mutated and NTBA contributes to the inability of NK cells to kill Epstein-Barr virus-infected cells [50]. The ITIM in the NK cell cytotoxicity triggering receptor 2 (NCR2), also called NKp44 or Ly95, has also been described before. The functionality of the ITIM has not been clarified yet. In contrast, NCR2 was shown to be an activating receptor [51].

Furthermore, we re-identified several receptors of the SIGLEC family which were proposed to have ITIMs in their sequences, namely SIGLECL1 [52], SIGLEC2/CD22 [53], SIGLEC3/CD33 [54], SIGLEC5 [55], SIGLEC6 [56,57], SIGLEC7 [58], SIGLEC10 [59], SIGLEC11 [15]. For SIGLEC4, SIGLEC8 and SIGLEC9, we did not obtain predictions of ITIMs. For SIGLEC4 and SIGLEC8, alternative splice variants exist which either lack or have ITIMs [13,60]. The ITIM-bearing variant A of SIGLEC4 (NP_002352.1) was not detected by our approach. It has numerous cytoplasmic tyrosine-based motifs (VLYSPE, DKYESE, LSYSHS, DSYTLT and AEYAEI) which do not fit the ITIM consensus exactly. The absence of concrete reports about the functionality of these ITIMs, e.g. the demonstration of SHP phosphatase binding, does not allow to consider the computational classification as either correct or wrong. For SIGLEC8, only a single protein variant without ITIMs was present in the database. The SIGLEC9 protein was not integrated into RefSeq at the time of the analysis, but investigation of its sequence revealed that its ITIM is detectable. We conclude that we classified all available sequences of SIGLEC receptors correctly.

Also non-classical examples of proteins which were either shown to bind SHP-like phosphatases or even shown to contain functional ITIMs were redetected: the platelet/endothelial cell adhesion molecule PECAM-1 (NP_000433.2) which inhibits signalling from the collagen glycoprotein VI (GPVI) receptor on human platelets and the antigen receptor on B cells (BCR) via ITIMs [61–64]. The initial discovery of an ITIM was based on the FcγRIIB/CD32 receptor. It was shown to inhibit mast cell proliferation in response to high affinity IgE receptors [65] and to inhibit B-cell proliferation in response to antigen that is induced through the Ras pathway [66,67], both via ITIM-mediated inhibition. A similar role for FcγRIIB/CD32 was suggested in the inhibition of phagocytosis of monocytes/macrophages when it is

up-regulated in response to Interleukin 4 (IL4) [68]. The ITIM in the C terminus of the carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1, also called biliary glycoprotein or CD66a) is required for the inhibition of tumour growth. It mediates the binding of CEACAM1 to SHP phosphatases [69,70]. The G6B protein (C6orf25) is another Ig-like receptor which was shown to bind SHP phosphatases through phosphorylated ITIMs [71]. The signal regulatory protein alpha 2 (SIRPα2), also called protein tyrosine phosphatase, non-receptor type, substrate 1 (PTPNS1), SHP substrate-1 (SHPS1) or macrophage fusion receptor (MFR), is an Ig-like receptor with two ITIMs that mediates binding to SHP phosphatases [72].

We missed the ITIM in the SHP-2 interacting transmembrane adapter protein (SIT) which inhibits the TCR signal in T cells, binds the SHP-2 phosphatase via an ITIM [73]. However, it was among the 129 proteins for which a signalling domain or an extracellular domain of SMART was not obligatory (see Fig. 1). We also missed the ITIM in the IL4 receptor α (IL4R). It was shown to inhibit proliferation of IL4-stimulated cells and to recruit SHP-1, SHP-2 and the SH2-domain inositol phosphatase SHIP [21]. The transmembrane helix of IL4R was not detected by the TMHMM2.0 program.

To our knowledge, these 38 type I transmembrane receptors are all known type I ITIM receptors with extracellular or signalling domains that were present in the RefSeq database at the time of the analysis. We are aware of the fact that we missed 2 out of 38 real ITIM-bearing transmembrane protein using our approach. One case was due to the stringent criterion of a required extracellular or signalling domain. The SHP2 interacting transmembrane adaptor (SIT) does not have such a domain but has a functional ITIM that is conserved between mouse and human. The other ITIM receptor was missed due to a misprediction of the TMMM2.0 program, which failed to predict the transmembrane helix in the IL4R protein.

Based on the high rate of redetection of known type I ITIM receptors (94,6%), we conclude that our analysis pipeline is able to detect ITIM-bearing type I transmembrane receptors in protein sequence databases with high sensitivity.

3.3. Many ITIMs are conserved in orthologous pairs of human and mouse proteins

We found 29 pairs of ITIM-bearing type I transmembrane proteins of mouse and human to be each others' best match in genome-wide protein sequence searches. This "reciprocal best hit" criterion is frequently used as an operational criterion for orthology in comparative genomics studies. We note that our thresholds for the conclusion of orthology are very strict and that they only allow the identification of orthologous pairs of sequences with rather strong evidence (Table 2).

The fact that we were not able to find orthologous mouse counterparts for all of the identified human ITIM receptors has several reasons. First, the orthologous protein in mouse might not have an ITIM in its sequence. Second, two orthologous sequences might miss each other in sequence searches due to a too high divergence of the sequences after separation of the two species, thus making their common origin undetectable. Given the close evolutionary relationship between mouse and human, it is unlikely that this effect had a large influence on our results. Third, the orthologous sequences are partly incomplete. Those sequences were rigorously discarded in our automatic procedure because of the requirement to show sequence similarity over a stretch of 85% of the shorter sequence. Fourth, the current protein annotations still comprise erroneous regions, which let the sequence identity drop below 70% in the aligned regions, below the minimum threshold in our analysis. Fifth, it is possible that there is no pair of orthologous sequences present in the mouse and human ENSEMBL protein databases or that the genes have not been discovered yet. Sixth, the orthologous human or mouse ENSEMBL protein sequences might not have a counterpart in the well-curated, but incomplete RefSeq sequence.

Only two human ITIM receptors of the well-known KIR, LIR and SIGLEC families led to the identification of orthologous sequences in the mouse RefSeq database: CD22/SIGLEC2 and SIGLEC7. This finding probably reflects the late and divergent evolution of the main ITIM immune receptor classes in mouse and man. Additionally, the annotation of KIR and ILT gene clusters in the mouse genomes might not be complete. Both aspects could make it difficult to detect clear orthologs in the canonical ITIM receptor families.

The major finding of the orthology analysis is that there are conserved ITIMs in orthologous pairs of well-characterised type I receptors which were not known to comprise ITIMs before. Among these are receptors with kinase activity in their cytoplasmic tail like the ephrin receptors A2 and A8, the AXL oncogene, the TGF β receptor I, the Activin A receptor type 1B, the PDGF receptor β chain and the mouse tyrosine kinase receptor 1 (here called TK1, NP_005415.1). Orthologous receptor pairs with Ig-like domains comprise the interleukin 18 receptor 1 (IL18R1/NP_003846.1), the interleukin 1 receptor accessory protein (IL1RAP) and the IL1RAP-like protein 2 (IL1RAPL2) which all have three extracellular Ig-like domains, a transmembrane region and an intracellular TIR domain. Single ITIMs are located between their transmembrane regions and the TIR domains in all human and mouse orthologs. The erythroblast membrane-associated protein (ERMAP) orthologs have only a single Ig-like domain in their extracellular N-termini, but have two SPRY domains in the C-termini [74].

Furthermore, several predicted orthologous ITIM receptors have combinations of multiple FN3 and Ig-like modules in their extracellular N-terminus and a fully conserved ITIM. The mentioned AXL and TK1 kinases also have multiple Ig-like and FN3 domains. The down syndrome cell

Table 2
ITIM-bearing type I transmembrane proteins of *H. sapiens* which have orthologs in *M. musculus*

Human RefPep ID	HUGO	Human protein description	Human ITIMs	Mouse RefPep ID	Mouse protein description	Mouse ITIMs
NP_064732.1	ACVR1B	activin A type IB receptor, isoform b precursor	(LPYYDL:428–433)	NP_031421.1	activin A receptor, type 1B	(LPYYDL:428–433)
NP_068713.2	AXL	AXL receptor tyrosine kinase isoform 1	(LLYSRL:632–637)	NP_033491.1	AXL receptor tyrosine kinase	(LLYSRL:626–631)
NP_001762.1	CD22	CD22 antigen; siglec-2	(ISYTTL:760–765), (IHYSEL:820–825), (VDYVIL:840–845)	NP_033975.1	CD22 antigen	(VSYAIL:775–780), (IHYSEL:835–840), (VDYVTL:855–860)
NP_001788.2	CDH11	cadherin 11, type 2, isoform 1	(LDYDYL:767–772)	NP_033996.1	cadherin 11; osteoblast–cadherin	(LDYDYL:767–772)
NP_001786.1	CDH5	cadherin 5, type 2 preproprotein; VE-cadherin	(VDYDFL:755–760)	NP_033998.1	cadherin 5; VE-cadherin	(IDYDFL:754–759)
NP_000386.1	CSF2R	colony stimulating factor 2 receptor, beta	(LEYLCL:626–631)	NP_031806.1	colony stimulating factor 2 receptor, beta 1	(LEYMCL:628–633)
NP_000751.1	CSF3R	colony stimulating factor 3 receptor (granulocyte)	(VLYGQL:750–755)	NP_031808.1	colony stimulating factor 3 receptor (granulocyte)	(VLYGQV:751–756)
NP_006173.1	DDR2	discoidin domain receptor family, member 2	(VSYTNL:682–687)	NP_072075.1	discoidin domain receptor family, member 2	(VSYANL:646–651)
NP_005609.2	DLL1	delta-like 1 protein	(VDYNLV:639–644)	NP_031891.1	delta-like 1	(VRYPTV:634–639)
NP_001380.2	DSCAM	CHD2–52 down syndrome cell adhesion molecule	(VHYQSV:1706–1711)	NP_112451.1	down syndrome cell adhesion molecule	(VHYQSV:1706–1711)
NP_004422.1	EPHA2	ephrin receptor EphA2	(IAYSLL:958–963)	NP_034269.1	Eph receptor A2	(IAYSLL:957–962)
NP_065387.1	EPHA8	ephrin receptor EphA8 precursor	(VCYGRL:649–654)	NP_031965.1	Eph receptor A8	(VCYGRL:648–653)
NP_000112.1	EPOR	erythropoietin receptor	(LKYLYL:452–457)	NP_034279.1	erythropoietin receptor	(LKYLYL:451–456)
NP_002560.1	FURIN	furin; proprotein convertase subtilisin/kexin type 3	(ISYKGL:757–762)	NP_035176.1	proprotein convertase subtilisin/kexin type 3	(ISYKGL:756–761)
NP_065840.1	IGSF9	immunoglobulin superfamily, member 9	(LQYLSL:908–913)	NP_291086.1	immunoglobulin superfamily, member 9	(LQYLSL:924–929)
NP_002173.1	IL1RAP	interleukin 1 receptor accessory protein	(VQYKAV:501–506), (LSYSSL:562–567)	NP_032390.1	interleukin 1 receptor accessory protein	(VQYKAV:501–506), (LSYSSL:562–567)
NP_059112.1	IL1RAPL2	interleukin 1 receptor accessory protein-like 2	(LSYTKV:406–411)	NP_109613.1	Interkeukin 1 receptor accessory protein-like 2	(LSYTKV:406–411)
NP_002175.1	IL6ST	interleukin 6 signal transducer	(VQYSTV:757–762)	NP_034690.1	interleukin 6 signal transducer	(VEYSTV:755–760)
NP_002301.1	LIFR	leukemia inhibitory factor receptor precursor	(VIYIDV:972–977)	NP_038612.1	leukemia inhibitory factor receptor	(VVYIDV:967–972)
NP_002579.2	PCDHGC3	protocadherin gamma subfamily C, 3, isoform 1	(VFYRQV:795–800)	NP_291059.1	protocadherin gamma subfamily C, 3	(VFYRQV:795–800)
NP_061752.1	PCDHGC5	protocadherin gamma subfamily C, 5, isoform 1	(LKYMEV:757–762)	NP_291061.1	protocadherin gamma subfamily C, 5	(LKYMEV:757–762)
NP_002600.1	PDGFRB	platelet-derived growth factor (PDGF) receptor beta	(LSYMDL:798–803), (VLYTAV:1007–1012)	NP_032835.1	platelet-derived growth factor (PDGF) receptor beta	(LSYTDL:797–802), (VLYTAV:1006–1011)
NP_109596.1	PTPRO	receptor-type protein tyrosine phosphatase O	(VIYENV:397–402)	NP_035346.1	protein tyrosine phosphatase, receptor type, O	(VIYENV:397–402)
NP_055200.1	SIGLEC7	sialic acid binding Ig-like lectin 7; siglec-7	(IQYAPL:435–440)	NP_112458.1	sialic acid-binding immunoglobulin-like lectin E	(IHYATL:430–435)
NP_004090.3	STOM	erythrocyte membrane protein band 7.2 (stomatin)	(VVYYRV:121–126)	NP_038543.1	erythrocyte protein band 7.2; protein 7.2b	(VVYYRV:121–126)
NP_003171.1	SYT5	synaptotagmin 5	(VPYVEL:181–186), (LDYDKL:329–334)	NP_058604.1	synaptotagmin 9; synaptotagmin IX	(VPYVEL:181–186), (LDYDKL:329–334)

(continued on next page)

Table 2 (*continued*)

Human RefPep ID	HUGO	Human protein description	Human ITIMs	Mouse RefPep ID	Mouse protein description	Mouse ITIMs
NP_004603.1	TGFBR1	transforming growth factor, beta receptor I	(LPYYDL:426–431)	NP_033396.1	transforming growth factor, beta receptor I	(LPYYDL:426–431)
NP_005415.1	TIE	tyrosine kinase with Ig and EGF homology domains	(LSYPVL:829–834)	NP_035717.1	tyrosine kinase receptor 1	(LSYPVL:825–830)
NP_000585.1	TNF	tumor necrosis factor, alpha (cachectin)	(LIYSQV:133–138)	NP_038721.1	tumor necrosis factor–alpha; TNF alpha	(LVYSQV:136–141)

The HUGO gene symbols of human proteins, the descriptions and identifiers of human and mouse protein sequences in the RefSeq protein database are given. For each orthologous pair of proteins, the sequences of the ITIM motifs and the position of the ITIMs in the sequences are displayed.

adhesion protein (DSCAM) orthologs represent the largest proteins of the FN3-like subfamily of ITIM receptors, each having nine Ig-like and six FN3 domains. DSCAM is located in the down syndrome critical region on human chromosome 21 and plays a role in the development of the brain [75], suggesting a possible role of ITIMs in brain development. The immunoglobulin superfamily member 9 (IGSF9) is orthologous to the mouse neural cell adhesion molecule-like protein NR1 and has several FN3 and Ig-like domains in its N-terminus.

Seven orthologous pairs of ITIM proteins have solely FN3 domains in their extracellular part. Among these are proteins like the colony stimulating factor 3 receptor (CSF3R) and the leukaemia inhibitory factor (LIFR). We propose that LIFR has a cytoplasmic intracellular ITIM and five extracellular fibronectin 3-like domains instead of only three as proposed previously [76]. The interleukin 6 signal transducer is related to LIFR and has only four FN3 domains. The mouse and human colony stimulating factor 2 receptors (CSF2R) have only two FN3 domains. Single FN3 domains are found in the erythropoetin receptor and in the receptor type protein tyrosine phosphatase ζ in which the ITIM resides in the first intracellular phosphatase domain. In contrast, the related receptor type protein tyrosine phosphatase O lacks any extracellular domain but also has an ITIM in its phosphatase domain.

Another popular class of proteins, which unexpectedly were found to comprise ITIMs, are cadherin-like receptors. The vascular endothelial (VE) cadherin, also known as cadherin 5, 7B4 antigen or CD144 antigen, and the osteoblast cadherin, also known as cadherin 11, each have five cadherin-like domains and a single ITIMs in their cytoplasmic tail. Two members of the protocadherin gamma subfamily C, numbers 3 and 5, have six CA domains and conserved ITIMs between mouse and man.

Several further orthologous ITIM proteins were found which do not fall into larger families of membrane receptors. There is the delta-like protein 1 with multiple EGF-like domains as well as C2 domains, the furin protein, the toll-like receptors 7 and 8 with their abundant leucine-rich repeats, the erythrocyte membrane protein band 7.2. We also detected an ITIM in the tumour necrosis factor α which is initially expressed as type I receptor and subsequently shed from the cell surface [77]. The detection of an ITIM in the cytoplasmic part of TNF-α might influence further work on the original transmembrane form of TNF-α.

The conservation of the short ITIMs in the mentioned proteins through the last ~70 million years since the speciation of mouse and human is indicative of a conserved function of these motifs. Because the function of ITIMs as phosphorylation-dependent attachment sites in immune receptor signalling is well characterised, we suggest that typical ITIM-related signalling molecules, like the broadly expressed SHP phosphatases, might also be recruited to the previously undetected ITIM receptors, a hypothesis which is to be tested by experiment.

During our analysis, we also noted the conservation of an ITIM in a less well-characterised molecule, the popeye-2 protein, which lacks any known protein domains. It was therefore not included in the final set of 109 proteins. Popeye-2 was proposed to function in cardiac muscle development [78]. We note that also the well-characterised ITIM protein SIT does not contain any known extracellular or signalling domain. As popeye-2 does not show similarity to other proteins and the biochemical mechanism by which it functions is largely unknown, it is reasonable to direct popeye-2 research towards ITIM-related signal transduction.

3.4. Uncharacterised ITIMs are present in receptors of diverse families

Because past ITIM signalling research is based on studies of Ig-like receptors, it is reasonable to search for more uncharacterised ITIMs in other Ig-like receptors. Consequently, in addition to the already mentioned orthologs in human and mouse, the majority of additional ITIM receptors have Ig-like domains. The NK cell inhibitory receptor (NKIR, NP_620587.1) has a single Ig-like domain and two ITIMs in the cytoplasmic tail. Similarly, the natural cytotoxicity receptor NK-p44 was found to have a single Ig-like domain and an ITIM, although its main function seems to be NK cell activation mediated by its membrane-proximal ITAM [51,79]. Further proteins with single Ig-like domains are the programmed cell death (PCD) 1 protein, the myelin protein zero (Charcot-Marie-Tooth disease 1B) and the homologous myelin protein zero-like 1 receptor, the NFAT activation molecule 1. The Ig superfamily protein encoded by sequence NP_009199.1 has only two Ig like domains

Table 3
Human type I transmembrane proteins with ITIMs for which data about their mRNA expression in human tissues is available

RefPep ID	HUGO	Description	U95 ID	SG	TR	TG	HE	LU	TM	PA	LI	KI	SP	AG	PR	TE	OV	UT	CE	BR	CX	CN	AM	TA	CC	SC	DR	PG	T−	T+	BL
NP_068713.2	AXL	AXL receptor tyrosine kinase isoform 1	38433_at	0	1	0	0	1	1	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0
NP_006698.1	BTLN3	butyrophilin-like 3; butyrophilin-like receptor	31645_at	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1	0	0	0	0	0
NP_001762.1	CD22	CD22 antigen	38522_s_at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
NP_001763.1	CD33	CD33 antigen (gp67)	36802_at	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
NP_001786.1	CDH5	cadherin 5, type 2 preproprotein	37196_at	0	1	1	1	2	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0
NP_001703.2	CEACAM1	carcinoembryonic ag-related cell adhesion molecule 1	36082_at	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NP_054700.1	DDR1	Discoidin receptor tyrosine kinase isoform c	36643_at	2	4	4	0	1	1	1	0	1	0	1	2	0	1	0	2	2	3	3	3	3	6	6	1	2	0	0	0
NP_001380.2	DSCAM	Down syndrome cell adhesion molecule	36699_at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
NP_000112.1	EPOR	erythropoietin receptor precursor	37986_at	1	0	1	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0
NP_003992.2	FCGR2B	Fc fragment of IgG, low affinity IIb	34665_g_at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
NP_002560.1	FURIN	furin preproprotein;	35338_at	4	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NP_004954.1	GUCYC2C	guanylate cyclase 2C	34450_at	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NP_005525.1	IFNGR2	interferon gamma receptor 2;	41140_at	2	2	2	1	2	2	1	1	1	1	2	2	3	1	2	1	2	1	2	1	1	1	2	2	2	1	1	4
NP_000866.1	IGF1R	insulin-like growth factor 1 receptor precursor	34718_at	1	1	1	0	0	1	0	0	1	0	1	2	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1
NP_002173.1	IL1RAP	interleukin 1 receptor accessory protein	38546_at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
NP_055573.1	KIAA0254	KIAA0254 gene product	37045_at	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0

(continued on next page)

Table 3 (*continued*)

RefPep ID	HUGO	Description	U95 ID	SG	TR	TG	HE	LU	TM	PA	LI	KI	SP	AG	PR	TE	OV	UT	CE	BR	CX	CN	AM	TA	CC	SC	DR	PG	T−	T+	BL
NP_055326.1	KIR2DL3	KIR, two domains, long cytoplasmic tail, 3	36886_f_at	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NP_037421.1	KIR3DL1	KIR, three domains, long cytoplasmic tail, 1;	36887_f_at	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NP_006728.1	KIR3DL2	KIR, three domains, long cytoplasmic tail, 2	36735_f_at	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
NP_068352.1	LAIR1	Leukocyte-associated Ig-like receptor 1, isoform b	37470_at	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	3
NP_002301.1	LIFR	Leukemia inhibitory factor receptor precursor	39617_at	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NP_006660.1	LILRB1	Leukocyte immunoglobulin-like receptor, B1	35927_r_at	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
NP_005865.1	LILRB2	Leukocyte immunoglobulin-like receptor, B2	39221_at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
NP_006855.1	LILRB3	Leukocyte immunoglobulin-like receptor, B3	37148_at	0	0	0	0	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12
NP_006838.2	LILRB4	Leukocyte immunoglobulin-like receptor, B4	36753_at	1	1	0	1	1	2	1	0	0	0	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1
NP_006831.1	LILRB5	Leukocyte immunoglobulin-like receptor, B5	36789_f_at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
NP_004819.1	LY95	Lymphocyte antigen 95 (activating NK-receptorp44)	34064_s_at	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NP_006334.1	MERTK	c-mer proto-oncogene tyrosine kinase	40648_at	0	0	0	0	0	1	0	0	0	0	2	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
NP_000521.1	MPZ	Myelin protein zero	37070_at	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	5	3	0	1	1	1

NP_003944.1	MPZL1	Myelin protein zero-like 1	33769_at	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
NP_056428.1	PAL	Retina specific protein PAL	31659_at	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	1	1	1
NP_002600.1	PDGFRB	Platelet-derived growth factor receptor beta	36993_at	0	1	1	1	2	1	0	0	0	1	1	1	1	2	3	0	0	1	0	0	0	0	1	2	0	0	0	0
NP_000433.2	PECAM1	Platelet/endothelial cell adhesion molecule	37397_at	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2
NP_002812.2	PTK7	PTK7 protein tyrosine kinase 7	38819_at	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
NP_542970.1	PTPNS1	Protein tyrosine phosphatase, non-receptor type substrate 1	40144_at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
NP_109592.1	PTPRO	Receptor-type protein tyrosine phosphatase O, isof. A	32199_at	1	1	0	0	0	0	0	0	2	0	1	0	1	0	0	0	1	1	1	2	1	0	0	0	1	0	0	0
NP_006496.2	PVR	Polio virus receptor	32699_s_at	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
NP_055200.1	SIGLEC7	Sialic acid binding Ig-like lectin 7	34978_g_at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
NP_004090.3	STOM	Erythrocyte membrane protein band 7.2 (stomatin)	40419_at	2	5	4	3	10	5	2	3	7	2	5	5	3	4	6	3	3	4	6	4	5	5	6	8	4	5	5	3
NP_003171.1	SYT5	Synaptotagmin 5	39242_at	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	4	3	1	0	0	0	1	0	0	0
NP_004603.1	TGFBR1	Transforming growth factor, beta receptor I	32903_at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
NP_009199.1	Z39IG	Ig superfamily protein	37976_at	2	1	1	1	4	3	4	2	2	1	2	1	1	1	1	1	0	1	1	1	1	1	2	1	1	1	1	2

The protein sequences are represented by their identifiers in the RefSeq database, the HUGO symbols of their genes and a short description. For each gene, the probeset identifier of the U95 Affymetrix gene chip is displayed. The mRNA expression level of the genes that encode these proteins range from 0 (low) to 14 (high). The expression values were obtained from a publicly available data set and were re-scaled for easier interpretation (see Materials and methods). The last columns contain the expression levels in diverse human tissues: SG salivary gland, TR trachea, TG thyroid gland, HE heart, LU kung, TM thymus, PA pancreas, LI liver, KI kidney, SP spleen, AG adrenal gland, PR prostate, TE testis, OV ovary, UT uterus, CE cerebellum, BR whole brain, CX cortex, CN caudate nucleus, AM amygdala, TA thalamus, CC corpus callosum, SC spinal cord, DR dorsal root ganglion, PG pituitary gland, CD34+ Thy-progenitor cells from GCSF treated patients, CD34+ Thy+ hematopoietic stem cells from GCSF treated patients, BL whole blood.

and one ITIM. The butyrophilin-like 3 (BTLN3) receptor is a homolog of the orthologous ERMAP receptors. According to a previous study, it is expressed in small intestine, colon, testis, and in leukocytes [80]. In ERMAP-like proteins, a single ITIM is located between the TM helix and the cytoplasmic SPRY domains. The deleted in colorectal cancer (DCC) protein has multiple FN3 and Ig-like domains. We hypothesise that it exerts an anti-proliferative role on tumour cells by the function of its cytoplasmic ITIM.

Ig-like domains are often combined with FN3 domains in cell surface receptors by extensive protein domain shuffling. The presence of ITIMs in the C-termini of many of such proteins suggests a general role of ITIM-related intracellular signalling in all Ig-like and FN3-like receptors. In addition to the mentioned mouse and human receptors, the IFN-γ receptor 2 and the retina specific protein PAL are representatives of the family of FN3-like ITIM receptors. Notably, the PAL protein additionally has a leucine-rich repeat region similar to the toll-like receptors which act in innate immunity [81].

A special subclass of Ig-like and FN3-like ITIM receptors is defined by the presence of protein kinase domains in their cytoplasmic parts. In addition to the mentioned pairs of orthologs, we identified five other protein kinases with ITIMs. The protein tyrosine kinase 7 (PTK7) has multiple Ig-like domains. The receptor tyrosine kinase-like protein 2 has an Ig-like domain and an additional Kringle domain. The c-mer oncoprotein has two FN3 and two Ig-like domains. The IGF1 receptor has several FN3 domains and a Furin-like domain. The c-ros-1 protein has multiple FN3-like domains. A hypothetical receptor kinase, the DKFZp761p1010 protein, also has an ITIM but no known extracellular domains. In each of these proteins, the ITIM is located in the kinase domain. It is therefore possible that ITIM phosphorylation and SH2 domain binding directly alter the catalytic activity of these proteins.

The remaining novel ITIM receptors fall into different classes. The discovery of ITIMs in the cadherin-like proteins desmoglein 2 and protocadherin 22 increases the number of ITIM receptors in this family to six. Three protein with extracellular PX, PXA were found, the NS1-associated protein 1, the KIAA0254 protein, and the sorting nexin 13. The hypothetical FLJ21302 protein has a characteristic leucine-rich repeat region and therefore possibly belongs to the toll-like receptor family of ITIM receptors. The interphotoreceptor matrix proteoglycan 200 has two extracellular SEA modules but does not share similarity with other ITIM receptors.

3.5. mRNA expression profiles of ITIM transmembrane proteins suggest a role for ITIM signalling in diverse tissues of solid organs

We were able to retrieve expression profiles across human tissues for 42 of our ITIM receptors from a public expression database (Table 3). Many ITIM receptors showed an expression pattern which is typical for a blood cell-related receptor. Expression of these proteins was almost exclusively found either in pure blood cell populations or in tissues related to the life cycle of blood cells like thymus, spleen, or kidney. Examples for these typical blood cell ITIM receptors include the leukocyte inhibitory receptor genes LIR1, LIR2, LIR3, and LIR5, the genes for the SIGLEC proteins CD33 and SIGLEC7, the poliovirus receptor or the IL1 accessory protein.

In contrast, we also found many other ITIM receptors to be expressed in diverse solid organs. The IFN-γ 2 receptor is ubiquitously expressed in all examined cell types. The leukocyte immunoglobulin-like receptor 5 is expressed in blood cells but also in trachea and the lung, in secretory tissues like the salivary and adrenal gland, the testis and ovary, in nervous tissues and in blood cells. The insulin-like growth factor receptor 1 is expressed in the salivary gland, the trachea and the lung, the thymus, kidney and adrenal gland. The highest expression values of synaptotagmin 5 (SYT5) can be measured in the brain, more precisely in the amygdala, the cortex and the pituitary gland, but no expression was found in blood cells. The platelet-derived growth factor β chain is expressed in the trachea, heart, thyroid gland, lung, kidney, spleen, adrenal gland, prostate testis, ovary, uterus, cortex, dorsal root ganglia, and pituitary gland. The AXL oncogene is expressed in trachea, lung, thymus, adrenal gland and reproductive tissues. The protein tyrosine kinase 3A mRNA is detected in diverse solid organs but not in blood cells. The vascular endothelial (VE) cadherin is expressed in many organs but not in the blood cells themselves. The down syndrome cell adhesion molecule (DSCAM) seems to be exclusively expressed in the brain, more precisely the amygdala.

Based on the heterogeneous expression patterns of the mentioned ITIM-bearing transmembrane proteins, we postulate that ITIM signalling is not a blood cell-specific signal transduction mechanism. Instead, we believe that ITIM-dependent signalling will be revealed as a widespread mechanism for the control of extracellular signals in diverse tissues and in various cellular contexts by future experimental studies.

4. Conclusions

We developed a systematic search for ITIM-bearing type I receptors in human proteins based on a combination of pattern searches and protein sequence context. The performance of the approach was tested on a set of type I ITIM receptors that is known from the literature. It was shown that the algorithm is highly sensitive, as it re-identified 36, but missed only 2 known human type I ITIM receptors. Additionally, it is far more specific than simple ITIM pattern searches alone. The total number of predictions in a set of 16177 proteins was lowered from 7064 using pattern searches alone to 109 using our approach. Thus, we propose

that our strategy is useful in mammalian sequence analysis on a genome-wide scale. The combination of pattern searches for short motifs with domain context information may easily be applied to a variety of other short motifs which otherwise could only be detected with high false positive rates.

We identified and described several previously undetected ITIMs in known and unknown protein sequences. Some of these were even conserved in human and mouse orthologous proteins. The analysis of the extracellular parts of all known and possible ITIM receptors revealed that a small set of protein domains was shuffled during evolution to construct the vast majority of ITIM receptors. As expected, the Ig-like domain is the most frequent module in ITIM receptors, but also FN3, CA, LRR and EGF domains are used frequently. Sometimes protein kinase modules, which have internal conserved ITIMs, are fused to these receptors. It is not clear yet whether these ITIMs can influence the catalytic function of these globular domains. However, we propose that the conservation of ITIMs in uncharacterised cytoplasmic tails of orthologous and homologous receptors indicates possible roles of ITIM signalling in many well known but also in novel proteins.

The analysis of mRNA expression patterns of the ITIM receptors confirmed the specificic blood cell expression of some receptors, but also revealed a significant number of ITIM receptor genes which are expressed in solid organs. Therefore, we predict ITIM signalling to be important in cell types different from immune cells.

We would like to comment that our search for new ITIM-containing receptors may not be exhaustive and that additional undetected ITIM receptors may be hidden in the human genome. This may be due to the fact that we restricted our search to the RefSeq set of proteins. Second, our analysis was focused exclusively on receptors containing a signal peptide at their N-terminus, a single transmembrane helix and a known extracellular or signalling domain. This strict approach enabled us to reach a high specificity. However, ITIMs are not only present in type I receptors. By such an approach, we excluded the C-type lectin like ITIM receptors (CLIRs) from the analysis. Notably, the Bradykinin B2 receptor, which is a seven TM helix GPCR, was also found to have an ITIM and binds SHP-2 [82]. These shortcomings of our analysis have to be addressed by new ITIM search protocols in the future.

References

[1] Ljunggren HG, Karre K. Immunol Today 1990;11(7):237–44.
[2] Lanier LL. Annu Rev Immunol 1998;16:359–393.
[3] Long EO. Annu Rev Immunol 1999;17:875–904.
[4] Moretta L, Bottino C, Pende D, Mingari MC, Biassoni R, Moretta A. Eur J Immunol 2002;32:1205–11.
[5] Moretta L, Biassoni R, Bottino C, Mingari MC, Moretta A. Immunol Today 2000;21(9):420–22.
[6] Lanier LL, Corliss BC, Wu J, Leong C, Phillips JH. Nature 1998; 391(6668):703–07.
[7] Lanier LL, Corliss B, Wu J, Phillips JH. Immunity 1998;8(6): 693–701.
[8] Leibson PJ. Immunity 1997;6(6):655–61.
[9] Brumbaugh KM, Binstadt BA, Billadeau DD, Schoon RA, Dick CJ, Ten RM, et al. J Exp Med 1997;186(12):1965–74.
[10] Long EO, Barber DF, Burshtyn DN, Faure M, Peterson M, Rajagopalan S, et al. Immunol Rev 2001;181:223–33.
[11] Watzl C, Stebbins CC, Long EO. J Immunol 2000;165(7): 3545–48.
[12] Blery M, Olcese L, Vivier E. Hum Immunol 2000;61(1):51–64.
[13] Crocker PR, Varki A. Immunology 2001;103(2):137–45.
[14] Moyron-Quiroz JE, Partida-Sanchez S, Donis-Hernandez R, Sandoval-Montes C, Santos-Argumedo L. Scand J Immunol 2002;55(4): 343–51.
[15] Angata T, Kerr SC, Greaves DR, Varki NM, Crocker PR, Varki A. J Biol Chem 2002;277(27):24466–74.
[16] Adachi T, Flaswinkel H, Yakura H, Reth M, Tsubata T. J Immunol 1998;160(10):4662–5.
[17] Adachi T, Wakabayashi C, Nakayama T, Yakura H, Tsubata T. J Immunol 2000;164(3):1223–1229.
[18] Li L, Dixon JE. Semin Immunol 2000;12(1):75–84.
[19] Plutzky J, Neel BG, Rosenberg RD. Proc Natl Acad Sci U S A 1992;89(3):1123–7.
[20] Burshtyn DN, Yang W, Yi T, Long EO. J Biol Chem 1997;272(20): 13066–72.
[21] Kashiwada M, Giallourakis CC, Pan PY, Rothman PB. J Immunol 2001;167(11):6382–7.
[22] Pruitt KD, Tatusova T, Maglott DR. Nucleic Acids Res 2003;31(1): 34–7.
[23] Krogh A, Larsson B, von Heijne G, Sonnhammer EL. J Mol Biol 2001;305(3):567–80.
[24] Reczko MFP, Staub E, Hatzigeorgiou A. LNCS 2002;2452:60ff.
[25] Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, et al. Nucleic Acids Res 2002;30(1):242–44.
[26] Eddy SR. Bioinformatics 1998;14(9):755–63.
[27] Snel B, Bork P, Huynen MA. Proc Natl Acad Sci U S A 2002; 99(9):5890–5.
[28] Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, et al. Nucleic Acids Res 2002;30(1):38–41.
[29] Daubin V, Gouy M, Perriere G. Genome Res 2002;12(7):1080–90.
[30] Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV. BMC Evol Biol 2001;1(1):8.
[31] Remm M, Storm CE, Sonnhammer EL. J Mol Biol 2001;314(5): 1041–52.
[32] Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, et al. Proc Natl Acad Sci U S A 2002;99(7):4465–70.
[33] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Nucleic Acids Res 1997;25(17):3389–02.
[34] Moller S, Croning MD, Apweiler R. Bioinformatics 2001;17(7): 646–53.
[35] Uhrberg M, Valiante NM, Shum BP, Shilling HG, Lienert-Weidenbach K, Corliss B, et al. Immunity 1997;7(6):753–63.
[36] Rajalingam R, Gardiner CM, Canavez F, Vilches C, Parham P. Tissue Antigens 2001;57(1):22–31.
[37] Martin AM, Kulski JK, Witt C, Pontarotti P, Christiansen FT. Trends Immunol 2002;23(2):81–8.
[38] Wilson MJ, Torkar M, Haude A, Milne S, Jones T, Sheer D, et al. Proc Natl Acad Sci U S A 2000;97(9):4778–83.
[39] Borrego F, Kabat J, Kim DK, Lieto L, Maasho K, Pena J, et al. Mol Immunol 2002;38(9):637–60.
[40] Wagtmann N, Rojo S, Eichler E, Mohrenweiser H, Long EO. Curr Biol 1997;7(8):615–8.
[41] Samaridis J, Colonna M. Eur J Immunol 1997;27(3):660–5.
[42] Gomez-Lozano N, Gardiner CM, Parham P, Vilches C. Immunogenetics 2002;54(5):314–9.
[43] Natarajan K, Dimasi N, Wang J, Mariuzza RA, Margulies DH. Annu Rev Immunol 2002;20:853–85.

[44] Meyaard L, Adema GJ, Chang C, Woollatt E, Sutherland GR, Lanier LL, et al. Immunity 1997;7(2):283–90.
[45] Meyaard L, Hurenkamp J, Clevers H, Lanier LL, Phillips JH. J Immunol 1999;162(10):5800–4.
[46] Meyaard L, van der Vuurst de Vries AR, de Ruiter T, Lanier LL, Phillips JH, Clevers H. J Exp Med 2001;194(1):107–12.
[47] Miller I, Hatzivassiliou G, Cattoretti G, Mendelsohn C, Dalla-Favera R. Blood 2002;99(8):2662–9.
[48] Davis RS, Wang YH, Kubagawa H, Cooper MD. Proc Natl Acad Sci U S A 2001;98(17):9772–7.
[49] Mousseau DD, Banville D, L'Abbe D, Bouchard P, Shen SH. J Biol Chem 2000;275(6):4467–74.
[50] Bottino C, Falco M, Parolini S, Marcenaro E, Augugliaro R, Sivori S, et al. J Exp Med 2001;194(3):235–46.
[51] Cantoni C, Bottino C, Vitale M, Pessino A, Augugliaro R, Malaspina A, et al. J Exp Med 1999;189(5):787–96.
[52] van den Berg TK, Nath D, Ziltener HJ, Vestweber D, Fukuda M, van Die I, et al. J Immunol 2001;166(6):3637–40.
[53] Yousef GM, Ordon MH, Foussias G, Diamandis EP. Biochem Biophys Res Commun 2001;284(4):900–10.
[54] Angata T, Hingorani R, Varki NM, Varki A. J Biol Chem 2001; 276(48):45128–36.
[55] Connolly NP, Jones M, Watt SM. Br J Haematol 2002;119(1): 221–38.
[56] Patel N, Brinkman-Van der Linden EC, Altmann SW, Gish K, Balasubramanian S, Timans JC, et al. J Biol Chem 1999;274(32): 22729–38.
[57] Richard M, Veilleux P, Rouleau M, Paquin R, Beaulieu AD. J Leukoc Biol 2002;71(5):871–80.
[58] Ito A, Handa K, Withers DA, Satoh M, Hakomori S. FEBS Lett 2001;498(1):116–20.
[59] Whitney G, Wang S, Chang H, Cheng KY, Lu P, Zhou XD, et al. Eur J Biochem 2001;268(23):6083–96.
[60] Foussias G, Yousef GM, Diamandis EP. Biochem Biophys Res Commun 2000;278(3):775–81.
[61] Gibbins J. Trends Cardiovasc Med 2002;12(5):213.
[62] Newman DK, Hamilton C, Newman PJ. Blood 2001;97(8): 2351–57.
[63] Wilkinson R, Lyons AB, Roberts D, Wong MX, Bartley PA, Jackson DE. Blood 2002;100(1):184–93.
[64] Jackson DE, Gully LM, Henshall TL, Mardell CE, Macardle PJ. Tissue Antigens 2000;56(2):105–16.
[65] Malbec O, Attal J, Fridman W, Daeron M. Mol Immunol 2002; 38(16–18):1295.
[66] Tridandapani S, Phee H, Shivakumar L, Kelley TW, Coggeshall KM. Mol Immunol 1998;35(17):1135–46.
[67] Koncz G, Pecht I, Gergely J, Sarmay G. Eur J Immunol 1999;29(6): 1980–89.
[68] Tridandapani S, Siefker K, Teillaud JL, Carter JE, Wewers MD, Anderson CL. J Biol Chem 2002;277(7):5082–89.
[69] Izzi L, Turbide C, Houde C, Kunath T, Beauchemin N. Oncogene 1999;18(40):5563–72.
[70] Chen T, Zimmermann W, Parker J, Chen I, Maeda A, Bolland S. J Leukoc Biol 2001;70(2):335–40.
[71] de Vet EC, Aguado B, Campbell RD. J Biol Chem 2001;276(45): 42070–76.
[72] Cant CA, Ullrich A. Cell Mol Life Sci 2001;58(1):117–24.
[73] Pfrepper KI, Marie-Cardine A, Simeoni L, Kuramitsu Y, Leo A, Spicka J, et al. Eur J Immunol 2001;31(6):1825–36.
[74] Ye TZ, Gordon CT, Lai YH, Fujiwara Y, Peters LL, et al. Gene 2000;242(1–2):337–45.
[75] Yamakawa K, Huot YK, Haendelt MA, Hubert R, Chen XN, Lyons GE, et al. Hum Mol Genet 1998;7(2):227–37.
[76] Tomida M, Gotoh O. J Biochem (Tokyo) 1996;120(1):201–5.
[77] Maskos K, Fernandez-Catalan C, Huber R, Bourenkov GP, Bartunik H, Ellestad GA, et al. Proc Natl Acad Sci U S A 1998;95(7): 3408–12.
[78] Andree B, Hillemann T, Kessler-Icekson G, Schmitt-John T, Jockusch H, Arnold HH, et al. Dev Biol 2000;223(2):371–82.
[79] Vitale M, Bottino C, Sivori S, Sanseverino L, Castriconi R, Marcenaro E, et al. J Exp Med 1998;187(12):2065–72.
[80] Shibui A, Tsunoda T, Seki N, Suzuki Y, Sugano S, Sugane K. J Hum Genet 1999;44(4):249–52.
[81] Medzhitov R. Nat Rev Immunol 2001;1(2):135–45.
[82] Duchene J, Schanstra JP, Pecher C, Pizard A, Susini C, Esteve JP, et al. J Biol Chem 2002;277(43):40375–83.

8 Insights into the evolution of the nucleolus by an analysis of its protein domain repertoire

Die folgende Erklärung legt die individuellen Beiträge der Co-Autoren zum nachfolgenden Manuskript dar. Von den Co-Autoren unterschriebene Fassungen dieser Erklärung liegen dieser Dissertation bei. Die Erklärungen sollen belegen, dass die erzielten Resultate im wesentlichen auf meiner Arbeit beruhen und ihre Verwendung innerhalb dieser Dissertation gerechtfertigt ist.

The NUCLEOLUS manuscript: declaration about contributions of authors

Hereby I, co-author of the manuscript *„Insights into the evolution of the nucleolus by an analysis of its protein domain repertoire“* which is accepted for publication in *BioEssays*, declare my contributions to the results and to the preparation of the manuscript. Furthermore, I agree that the statements below describe the contributions of the other co-authors correctly.

1) Eike Staub
- conducted the whole sequence analysis process,
- determined the strategy for the discovery of new domains,
- characterised the new domains by exhaustive literature searches,
- determined the distributions of domains across phyla,
- noted the link between biochemical function of the domains and their differing patterns of occurrence in species from different phyla,
- outlined the strategy how to represent the complete data in the WWW,
- interpreted the results with regard to nucleus and nucleolus evolution,
- wrote the text and prepared the figures for the manuscript,
- serves as the corresponding author during the review process.

2) Petko Fiziev
- was responsible for the visualisation of the complete data set on nucleolar protein domains and nucleolar protein domain architectures on the WWW.

2) André Rosenthal
- contributed to the final stages of manuscript preparation by helpful comments on the style of the manuscript.

3) Bernd Hinzmann
- was the supervisor of the project,

- served as a partner in discussions about the meaning of the results for the reconstruction of early eukaryotic evolution,
- contributed to the final stages of manuscript preparation by helpful comments on the style of the manuscript.

Insights into the evolution of the nucleolus by an analysis of its protein domain repertoire

Eike Staub*[,1,2], Petko Fiziev[1,2], André Rosenthal[1] and Bernd Hinzmann[1]

[1] metaGen Pharmaceuticals GmbH, Oudenarder Str. 16, D-13347 Berlin, Germany

[2] Max-Planck Institute for Molecular Genetics, Dept. of Computational Molecular Biology, Ihnestr. 73, D-14195 Berlin, Germany

*author to whom correspondence should be addressed

tel +49-(0)30-8413 1157

fax +49-(0)30 8413 1152

eike.staub@molgen.mpg.de

keywords: nucleolus, protein domains, proteomics repeats, ribosome, RNA metabolism, nucleolar organiser region

Abstract

Recently, the first investigation of nucleoli using mass spectrometry led to the identification of 271 proteins. This represents a rich resource for a comprehensive investigation of nucleolus evolution. We applied a protocol for the identification of known and novel conserved protein domains of the nucleolus, resulting in the identification of 115 known and 91 novel domain profiles. The phyletic distribution of nucleolar protein domains in a collection of complete proteomes of selected organisms from all domains of life confirms the archaebacterial origin of the core machinery for ribosome maturation and assembly, but also reveals substantial eubacterial and eukaryotic contributions to nucleolus evolution. We predict that in different phases of nucleolus evolution, protein domains with different biochemical functions were recruited to the nucleolus. We suggest a model for the late and continous evolution of the nucleolus in early eukaryotes and argue against an endosymbiotic origin of the nucleolus and the nucleus.

Supplementary information

We present alignments of the novel motifs and sketches of domain compositions of hundreds of already known or novel homologues of nucleolar proteins on our website (http://www.nucleolus.net/nucleolus/).

Introduction

Nucleoli are membrane-less dense compartments in the nuclei of eukaryotic cells [1]. They are associated with regions on chromosomes that comprise arrays of ribosomal RNA (rRNA) genes, so called nucleolar organiser regions (NOR). Nucleoli are thought to be the ribosome factories of the cell. Consequently, numerous building blocks of ribosomes can be found in nucleoli, both rRNAs and proteins. However, because many steps are required to build a ribosome, all those proteins are present in nucleoli which contribute to its biogenesis. In recent years, evidence emerged that nucleoli also have other functions than the assembly of ribosomes. They were proposed to function in the assembly of the signal recognition particle, in the processing of certain mRNAs, tRNAs and small nuclear RNAs, in the maturation of telomerase, nuclear export, sequestering of gene silencers, and in the regulation of the cell cycle [2-7]. This illustrates that the knowledge about the biological function of nucleoli is still fragmentary, despite the fact that the first nucleoli were already purified 40 years ago.

A breakthrough in nucleolus research was recently presented by Andersen and co-workers [8]. They presented the first proteomic analysis of purified human nucleoli using mass spectrometry. One half of the 271 identified peptides were known proteins. Only ten percent were known to be nucleolar before. For the other known proteins their association with the nucleolus was shown for the first time. The study also revealed the nucleolar localisation of many previously uncharacterised hypothetical proteins. Some of these were confirmed to be part of the nucleolus by fluorescence microscopy after expression with a YFP tag. Given that the dynamic change in the localisation of some nucleolar proteins depends on the status of a cell, the authors did not claim that they captured all nucleolar proteins by their approach. Nevertheless, their study is a big step towards a complete inventory of the human nucleolar proteome. It provides the seeds for the identification of homologous proteins in other species and in the human proteome itself, possibly leading to the discovery of additional building blocks of the nucleolus. Moreover, it facilitates the comprehensive investigation of the evolutionary past of the nucleolus by the analysis of nucleolar sequences.

In this manuscript, we describe the results of a search for known and novel conserved motifs of nucleolar proteins using sensitive sequence analysis techniques. After the identification and analysis of known protein domains and sequence features, we isolated novel repeats and domains in previously

uncharacterised sequence fragments of nucleolar proteins. We identified homologous protein domains in the proteomes of human, mouse, fly, worm, yeast, ear cress and of diverse eubacteria and archaebacteria which allowed us to determine the distribution of a comprehensive set of conserved nucleolar protein domains across phyla. The implications of these results for the evolution of the nucleolus and the nucleus are discussed.

Results and Discussion

Identification of 115 known and 91 novel protein domains in nucleolar proteins

On the basis of the results of Andersen et al. [8] we extracted a set of 235 nucleolar protein sequences from public databases. Our approach for the discovery of new motifs is similar to that successfully applied by Doerks et al. to identify novel protein domains in nuclear proteins [9]. Our set of proteins was searched for low-complexity regions, transmembrane helices and coiled-coil regions to exclude these regions from the subsequent analysis. We localized 115 different known protein domains from the Pfam database (version 7.3) in 177 (75%) proteins of our set (see figure 1). Subsequences of known domains were cut out in the nucleolar sequences. We identified intra-molecular repeats in 21 masked protein sequences which were also excluded, but were kept for manual evaluation of the repeats. Finally, approximately 55% of the original sequence remained unmasked. Because fragments of less than 30 amino acids length are not suitable for the detection of novel domains, we discarded another 7% of the sequence. 513 protein fragments remained, representing 48% of the sequence. To reduce the redundancy in this set of sequence fragments we performed pair-wise sequence similarity searches using *BLASTP* of all fragments versus each other. Detected similarities were used as relations in a single linkage clustering of the fragments. As only one fragment per cluster was selected for the subsequent analysis, a set of 488 fragments remained which had the potential to comprise novel conserved domains.

For the detection of sequences which are homologous to our set of 488 sequence fragments we performed iterative *PSIBLAST* searches in the NCBI non-redundant protein database (nr) using an expectation (E) value of 0.001 as a threshold to include a detected database sequence into the sequence profile of the next iteration. The number of maximum iterations was restricted to eight. Further iterations are unlikely to provide new information as usually a search either converges or

becomes unspecific due to an incorporation of false-positive sequences or sequences of low complexity into the profile. Therefore, we discarded *PSIBLAST* results which did not converge within 8 rounds. Using *PSIBLAST* profiles of each fragment we searched a database containing all copies of the 177 Pfam domains in the pfamseq database. When a profile detected one of these known domain copies, we also excluded the fragment from the analysis because it is likely to be distantly related to a known domain. We automatically built alignments from successful *PSIBLAST* results. To correct misalignments, each alignment was trimmed manually and reduced to regions of sufficient sequence conservation. Alignments that only presented trivial sequence similarities were discarded. For each of the remaining 213 alignments we built profile Hidden Markov Models (HMM) which allowed us to search for the conserved domains with high sensitivity in the nrdb90 database. The visualisation of known and new domains in all identified proteins in nrdb90 facilitated the exclusion of less interesting alignments from the analysis. We excluded those motifs that exclusively occurred in direct proximity to known domains. These can simply be regarded as domain extensions. Conflicts between overlapping novel domains were resolved. By picking only those domains that were characteristic of a set of sequences, we ended up with a set of 91 new domain signatures and repeats from 89 of the 235 proteins in the original set. Using our HMMs together in combination with Pfam HMMs we redetected 210 out of the 235 original proteins compared to 177 of 235 proteins using only the Pfam HMMs. The coverage of the total sequence space of the 235 proteins with domains was raised by 10.5%. We conclude that our set of HMMs is a tool which will enhance the detection of nucleolar protein domains in uncharacterised protein sequences. For all novel domains and repeats we provided a basic annotation based on the available annotations of single family members that were already characterised (see Table 2).

The distribution of nucleolar protein domains across the kingdoms of life

To elucidate the evolutionary history of the nucleolus, we decided to search several protein sets from completely sequenced genomes for occurrences of all known and novel conserved domains that are present in the investigated nucleolar proteins. We analysed species from different branches of the tree of life: human and mouse as mammals, the worm *Caenorhabditis elegans*, the insect *Drosophila melanogaster*, the baker's yeast *Saccharomyces cerevisiae* as a unicellular eukaryote, as well as multiple archaea and eubacteria representing the major bacterial lineages. Of the

conserved protein domains that can be found in human nucleolar proteins, the largest fraction of 59 domains (58 known and one novel) can be found in a minimum of one protein in all three domains of life. 25 domains were found in archaea and eukaryotes to the exclusion of eubacteria. 13 protein domains are present in eubacteria and eukaryotes to the exclusion of archaea. The vast majority, here 109, can be detected only in eukaryotes.

These results are only meaningful if we can exclude extensive lateral gene transfer (LGT) between phyla as an explanation for the distribution patterns of these domains. A hint for such a late spread of the protein domains across phyla would be the occurrence of a single domain in only a small subset of organisms from one phylum, either eubacteria or archaea. Therefore, we applied a more stringent rule to conclude that a distinct protein domain is present in a certain domain of life. In the following, only those protein domains were discussed which allowed a clearer statement about their presence or absence in each of the phyla: archaebacteria, eubacteria and eukaryotes. To be considered, a protein domain had to be present in a minimum of 4 different proteins from one phylum.

Each proteins domain was classified according to its phyletic distribution, thus providing information about its putative evolutionary age. Subsequently, the cellular functions of the domains that fit a particular phyletic pattern were analysed. The interpretation of the evolutionary age of the domains in combination with their cellular function allowed conclusions about the timely order in which the pre-nucleolar protein machinery could, at the earliest, have acquired certain domains and their associated cellular functions (see also box 1).

Ancient nucleolar protein domains mainly stem from ribosomal proteins or ribosome maturation factors

The fact that 59 human nucleolar protein domains can be found in all kingdoms of life indicates that a large fraction of the building blocks for nucleolar proteins was already present in the last universal common ancestor (LUCA). Ignoring the domains with less than four hits in a distinct kingdom, we yielded a set of 54 domains which can be regarded as ancient nucleolar domains. The proteins comprising these domains form the ancient core of the nucleolar protein machinery. They include the large group of protein domains from ribosomal proteins, reflecting the well recognised role of the nucleolus as the ribosome assembly factory. DEAD/DEAH box helicases are among the most abundant nucleolar proteins.

These RNA helicases are thought to unwind RNA during the assembly of diverse nucleoprotein complexes [10].

Diverse ancient RNA binding protein domains can be found in nucleoli. Many of these occur in ribosomal proteins, but are also present in non-ribosomal proteins. The S1 domain, named after its occurrence in the ribosomal protein S1, is a widespread RNA binding domain that can be found in a large number of other RNA-associated proteins [11]. The S4 domain is a putative RNA binding domain of diverse bacterial and eukaryotic ribosomal proteins and of RNA modifying enzymes like pseudouridine synthases and deaminases, RNA methylases, and tyrosyl-tRNA synthetases [12]. The transcription antitermination protein NusG of bacteria and various ribosomal proteins like L24 have a common RNA associated domain which is named KOW after its discoverers (Kyprides, Ouzounis, Woese) [13]. The PUA domain is a further putative RNA binding domain. Its name reflects its occurrence in pseudouridine synthase and archaeosine transglycosylase, but it is also present in other RNA modifying proteins like archaeosine synthases, rRNA methylases, and other families related to RNA metabolism [12]. The K homology domain (KH) is defined by its similarity to the human heterogeneous nuclear ribonucleoprotein (hnRNP) K. It is an RNA binding module that is present in a wide variety of quite diverse nucleic acid-binding proteins, e.g. the prokaryotic ribosomal protein S3 [14].

Apart from RNA binding domains, there are other ubiquitous protein domains which are diagnostic of RNA modification functions in proteins from the nucleolus. The RTC domain is named after its presence in RNA 3'-terminal phosphate cyclases which catalyse the ATP-dependent conversion of the 3'-phosphate to the 2',3'-cyclic phosphodiester in RNA [15]. The ribonuclease PH family signature is specific for 3'-5' exoribonucleases. Among these are ribonuclease PH which removes nucleotides from the CCA terminus of tRNA, polyribonucleotide nucleotidyltransferase (PNPase) that degrades messenger RNA starting from the 3’ end, and diverse proteins of the exosome which is responsible for 3' processing of the 5.8S rRNA [16,17]. The detection of the TruB domain reveals the base modification function of pseudouridylate synthases in nucleoli. Named after the prototype TruB which converts uracil to pseudouridine in many tRNAs, this family also comprises Cbf5p that modifies uracil in rRNA [18]. It is reasonable to assume that these RNA binding domains, which are either enzymatic themselves or associated with other catalytic RNA modifying domains, are relicts from an ancient RNA world and constitute the oldest part of the nucleolus.

Other ancient protein domains in the nucleolus are involved in the folding of proteins, they are so called chaperones. DnaJ domains (J-domains) are associated with the hsp70 heat-shock system, a ubiquitous protein folding system [19]. The cpn60-like proteins are proteins with homology to components of the bacterial GroEL protein folding system and which is essential for the correct folding and assembly of polypeptides into oligomeric structures [20,21]. The presence of DnaJ-like and TCP-1/cpn60-like proteins in the nucleolus and all domains of life suggests that the original function of these chaperones was ribosome assembly.

Several other ancient protein families with diverse functions can be found in the nucleolus. A few are related to the modification of DNA structure. We detected domain signatures of subunits of topoisomerase II, including those of DNA gyrase A and of DNA gyrase B [22,23]. Another ancient domain is the forkhead-associated domain (FHA), a phosphopeptide recognition domain found in many regulatory proteins like kinases, phosphatases, kinesins, transcription factors, RNA-binding proteins and metabolic enzymes [24]. To our knowledge, the emergence of FHA domains in genomes of archaea has not been described before. The GTP binding domain of elongation factor Tu is an ancient part of the translation machinery which is functionally linked to the nucleolus via its ribosomal association [25]. A second type of GTPase domain with prototypes in mouse MMR1 and human HSR1 is also found in the nucleolus [26]. The Metallophosphoesterase family comprises enzymes of different substrate specificity like nucleases (yeast MRE11, bacterial SbcD), phosphoserine phosphatases, nucleotidases, sphingomyelin phosphodiesterases and 2'-3' cAMP phosphodiesterases [27]. The Nol1_Nop2_Sun domain is the characteristic central domain of the proliferating cell nuclear antigen (PCNA) p120, which is encoded by the NOL1 gene and is thought to function as a RNA methylase in the nucleolus. The structural maintenance of chromosomes (SMC) proteins have recently been shown to act in processes like DNA repair, epigenetic silencing, and sister chromatid cohesion where they form ring-like structures around the chromatids. Their N- and C-terminal domains are ATPase domains which are linked by two coiled-coil hinge regions and a central globular domain. The central globular domain is the only ancient domain of this screen that has not been integrated into Pfam before [28]. The function of SMC proteins in the nucleolus is yet unknown. It is reasonable to assume that either they regulate DNA structure or transcription in the nucleolar organiser region (NOR). A similar function can be anticipated for SNF2_N domains which occur in proteins involved in transcription regulation (e.g., SNF2, STH1, brahma or MOT1) and chromatin

unwinding (e.g. ISWI) and other processes related to DNA structure regulation [29]. Thioredoxin domains are ancient domains that catalyse the oxidation and reduction of disulfide bonds, thereby facilitating protein folding [30]. The A1pp domain is a modular domain that is present in proteins like the rat macro-H2A histone protein, proteins from single strand RNA viruses and a third largely uncharacterised protein family with members in all kingdoms of life. A function in an ubiquitous cellular process was proposed. The detection of the A1pp domain in our analysis suggests that its role is associated with the nucleolus [31]. ABC transporters are responsible for the active transport of small molecules across cellular membranes. Their two ATP binding subunits can either be joined to the two transmembrane domains in one protein or exist as a separate protein. We also found ABC transporter ATP binding domains [32] and a Band 7 domain [33] in nucleolar proteins. Their occurrences among nucleolar proteins are hard to explain. Band 7 proteins are integral membrane proteins which should not co-purify with nucleoli, suggesting that their identification during mass spectrometry is possibly an artefact.

The distribution of functional classes of ancient nucleolar protein domains shows a strong bias towards ribosomal domains and domains acting in RNA modification and binding. Recently, Anantharaman et al. provided an excellent analytical review about the enzymes of RNA metabolism, many of which can be found in the set of ancient nucleolar domains presented here [34]. The various other ancient nucleolar protein domains mostly function in the regulation of DNA structure or in protein folding, probably regulating the accessibility and transcription of ribosomal genes in nucleolar organiser regions or supporting ribosome assembly. The structural core of the ribosome, the enzymes modifying the rRNA, and those supporting the correct assembly of the ribosome apparently represent the oldest part of the human nucleolus.

Nucleolar protein domains of archaebacterial origin function in ribosome maturation and translation

Another large fraction of nucleolar protein domains, 25 in this study, can be detected only in archaebacteria, but not in eubacteria. In the light of the previously proposed chimeric origin of the eukaryotic genome this finding is not surprising [35]. As the nucleolus is spatially linked to the rRNA genes and therefore adapted to them, it is reasonable to assume that a considerable fraction of the ribosome factory has the same archaebacterial origin as the eukaryotic rRNA genes. Evidence that core parts of the required RNA modification machinery were derived from an

archaebacterial ancestor comes from several earlier studies. Recently, Omer et al. have shown that small RNAs (sRNAs) exist in archaea and that they are homologous to eukaryotic small nucleolar RNAs (snoRNAs) [36]. Single examples of archaeal homologues of nucleolar proteins have also been noted before and were confirmed by this study. Those of fibrillarin (yeast Nop1p), NOP56/NOP58 or Imp4 [36,37]are already known and were proposed as indicators of an archaeal origin of eukaryotic RNA processing, and even as indicators of the archaeal origin of the nucleus [37]. We found that several other protein families or domains of the nucleolus are only present in archaea to the exclusion of eubacteria. Among these are four ribosomal protein domains, characteristic extensions of the small subunit proteins S3A and S4 and the large subunit proteins L15 and L31 (Ribosomal_L15e, Ribosomal_L31e, Ribosomal_S4e, Ribosomal_S3Ae) [38,39]. Several motifs of proteins which function in the process of translation in eukaryotes have also been found in archaea, but not in eubacteria (eIF-5a, EIF-5a_N, eIF6, eRF1_1, eRF1_2, eRF1_3) [40-43]. The eIF-5a proteins are linked to the nucleolus via their functional relation to translation. The roles of these proteins in the nucleolus are not clear yet.

We are aware of the fact that archaeal and eubacterial eIF-5a proteins are likely to be homologous to eubacterial EFP proteins (alignment not shown). The rather close relationship of archaebacterial and eukaryotic eIF-5a proteins and the distant relationship of both subfamilies to eubacterial EFPs has prevented the detection of this homology. This case illustrates the limited sensitivity of sequence searches, even when using HMMs. However, it also shows that limited sensitivity is not rendering our evolutionary interpretation invalid: Classifying the common domain of the eIF-5a and EFP families as “ancient” would have made these families uniformative with regard to the question about the origin of the eukarotic eIF-5a proteins. However, the closer relationship of eukaryotic eIF-5a proteins with the archaeal ones can clearly be deduced from the sequences. Therefore we expect that also other hypothetical cases of undetected homology will not change the general tendency of our results.

In combination with the results for ancient protein domains, these findings on archaebacterial sequence families support the hypothesis that the ribosome itself, many domains from the functionally related translation machinery, and the core human nucleolar machinery which includes RNA modification enzymes, stem from an archaebacterial ancestor.

The CBFD_NFYB_HMF family of proteins is characterised by a common domain between mammalian transcription factors of the CCAAT-binding factor (CBF) family

and archaeal histone proteins. It is probably involved in the regulation of ribosomal gene regulation in the nucleolar organiser regions [44]. Sm proteins are found in small nuclear ribonucleoprotein particles (snRNPs) like the spliceosomal U1, U2, U4/U6 and U5 and are also found in archaebacteria which do not have a splicing apparatus. The detection of a human Sm protein in the nucleolus points to an original role in ribosome maturation for Sm proteins [45]. Homologues of the archaebacterial subunit H of DNA-dependent RNA polymerases can be found in all eukaryotic RNA polymerases [46]. Their appearance in this analysis simply documents the transcriptional activity of ribosomal genes in NORs. This finally stresses the attractiveness of a model for the evolution of the nucleolus, in which a continuity of all aspects of ribosome generation is proposed; namely that ribosomal genes, the transcription machinery of ribosomal genes, and the machinery for maturation and assembly of the ribosome all stem from the genome of a single archaebacterial ancestor.

Nucleolar domains of eubacterial origin fulfil rather modern cellular functions

A considerable number of nucleolar protein domains are found in eubacteria and eukaryotes but not in archaea. However, they are much less abundant than the archaea-only protein domains: only 8 out of 13 of these domains fulfil our stringent criteria. Among these are again several proven or hypothetical RNA binding domains like the widespread RNA recognition motif (RRM) [47], the helicase and RNase D carboxy-terminal domain (HRDC) [48], the double stranded RNA binding (DsRBD/DSRM) domain [14] and the R3H domain, named after its conserved arginine and histidines [49]. In contrast to the archaeal RNA binding domains, these eubacterial RNA binding domains can not be found in enzymes which modify the bases of rRNA. Instead, they seem to be involved in more modern cellular functions related to RNA, e.g. like the regulation of splicing, the regulation of translocation of mRNAs or the control of the cell cycle. The functions of most of these eubacterial RNA binding domains in the nucleolus are not completely understood.

The 3'-5' exonuclease domain is the only eubacterial domain which is known to be catalytic. Prototypes of this domain are defined by the proofreading domain of E.coli DNA polymerase I, RNase D and Werner syndrome helicase [50,51]. RNase D is involved in the processing of tRNA, suggesting a similar function for its nucleolar counterpart. WD40 domains are β-propeller-like protein-protein interaction domains that are present in a wide range of proteins with various roles and diverse

cellular functions [52]. The frequent occurrence of WD40 domains among eukaryotic nucleolar proteins might be a sign of an increasing tendency towards compaction that is mediated or facilitated by protein-protein interaction domains. The BRCT domain is characteristic of proteins with functional relations to eukaryotic cell cycle control [53]. They might provide a link to the timely regulation of nucleolus disassembly and reassembly during the cell cycle.

Based on the function of the mentioned eubacterial domains as interaction-mediating and regulatory components and based on the fact that they rarely are present in the key rRNA mediating enzymes, we hypothesise that these domains did not take part in the key function of the early nucleolus. Because of the limited sensitivity of sequence searches it could be possible that some of these domains are actually hidden ancient domains. Nevertheless, the eubacterial sequences would then be more closely related to their eukaryotic homologs than to their undetected archaebacterial counterparts. We conclude that these protein domains are eubacterial contributions to nucleolus evolution that were acquired relatively late. We think that these domains have been recruited to a kind of ancestral nucleolar structure, probably of lower density than todays nucleoli, after the core rRNA modification enzymes and the core ribosome assembly machinery had evolved.

A large fraction of nucleolar protein domains evolved in eukaryotes

Most of the domains which were newly characterised in this study, precisely 80, are specific for eukaryotes. The new domains do not neccessarily represent new protein folds: they are rather lineage-specific conserved sequence regions of unknown structure. Their co-occurrence with other well-characterised domains, which in many cases define large protein families, means that most of them are subfamily-specific extensions. This can be observed for the ancient family of DEAD box RNA helicases which have various different C-terminal extensions and are the most widespread class of proteins in the nucleolus. In this type of families, it is unlikely that all the different extensions represent new domain folds. The homology between different types of extensions may simply remain undetected because of a too high degree of sequence divergence. Thus, different types of extsensions may well have a common structural fold or function. However, they certainly can be regarded as specific adaptions to the developing structure of the nucleolus, probably playing novel roles that had to be fulfilled during the formation of the nucleolus as a compartment and during its subsequent evolution.

Among the 115 known protein domains that occurred in the set of 235 nucleolar proteins, 29 domains were only found in eukaryotes (Tab 2). As these domains were not found in prokaryotes, and some even not in yeast, they probably represent those types of nucleolar protein domains that have evolved most recently and that are the latest acquisitions of the nucleolus. Limited sensitivity in sequence searches could have prevented the detection of some of these domains in bacteria. However, for these cases the degree of sequence divergence of eukaryotic and prokaryotic relatives must have been so high that it is disputable whether structure and function of the yet undetected relatives are still simlar. With regard to the question whether the evolution of the nucleolus was dominated by archaebacterial or eubacterial influences the known and novel eukaryote-specific domains are not informative. However, the abundance of eukaryote-specific domains that occur in all eukaryotic phyla considered here suggests that large sequence parts of todays nucleoli evolved early or at least changed fast during early eukaryotic evolution.

Some of the eukaryotic domain families have undergone a dramatic increase in the number of copies per genome. For example the exclusively eukaryotic high mobility group (HMG) box [54] can be found in seven yeast proteins, whereas the human genomes already encodes 124 proteins with this motif.

Only four of the eukaryotic-only domains stem from the ribosome (Ribosomal_L6e, Ribosomal_L14e, Ribosomal_L22e, Ribosomal_L27e) [55,56], thus reflecting the ancient origin of the ribosomal proteins. There are two other eukaryotic domains which are thought to function in RNA binding. The SRP14 protein is a part of the signal recognition particle (SRP) which targets secretory proteins to the membrane of the rough endoplasmatic reticulum. SRP14 is essential for RNA binding in the SRP [57]. Recently, the assembly of the SRP has been linked to the nucleolus [4]. The D111/G-patch domain occurs in diverse eukaryotic proteins related to RNA processing. Based on associated sequence features the G-patch domain was predicted to function in mRNA splicing or polyadenylation [58].

A well represented class of eukaryotic-only nucleolar domains is involved in the regulation of the compactness of DNA and in the assembly of complexes of nucleic acids and protein. Among them are the HMG box domains which are typically found in proteins that preferentially bind to distorted DNA structures [54,59]. They function in diverse eukaryotic nucleoprotein assemblies like the signal recognition particle, the nucleolus, or the transcription initiation complex [59]. Poly-ADP ribose polymerases and their PARP domains, PARP-like zinc fingers and PARP regulatory regions are not present in the yeast proteome, but are abundant in multicellular

eukaryotes with ten PARP family members in humans. PARPs catalyse the DNA-dependent transfer of ADP-ribose to some DNA-binding proteins, thereby decreasing their affinity to DNA, e.g. in response to DNA damage [60]. In the nucleolus, PARP activity might regulate the condensation of nucleolar matter and the accessibility of nucleosomal DNA. The CHROMO (CHRromatin Organization MOdifier) domain [61] and the CHROMO shadow domain [62] are present in proteins that function in the regulation of chromatin condensation and gene silencing. Another identified putative chromatin regulating domain is the AT-rich interaction domain (ARID) [63]. The SAP motif (after SAF-A/B, Acinus and PIAS) motif is a possible DNA binding domain which also seems to be implicated in the organisation of chromatin [64]. The histone superfamily comprises the proteins of the nucleosomal core, the histones, as well as other DNA binding proteins [65]. The histone fold seems to be a general motif regulating the compactness of complexes between DNA and proteins. Proteins of the nucleoplasmin family are chromatin decondensation proteins and directly interact with histones, thereby regulating the structure of nucleosomes [66].

The domains involved in chromatin organisation are the most abundant functional class within the group of eukaryotes-only nucleolar domains. How can this be explained? It is obvious that the emergence of chromatin during eukaryote evolution must have been a challenge for the correct assembly of ribosomes. Parallel to the evolution of gene-deactivating chromatin, the accessibility of rRNA and protein genes must have been maintained. Only then, the assembly of ribosomes, and thus protein synthesis in general, could be ensured. We hypothesise that the same machinery which regulated DNA structure in early eukaryotes, also was required for the evolution of compactness of the nucleolus. This assumption would also explain where the nucleolus life cycle has its origin and how the nucleolar structure depends on the cell cycle. A consequence of this hypothesis is that the compactness of present day nucleoli is made possible by proteins with chromatin-related functions.

Several eukaryotic protein domains with other functions can be found in the nucleolus. For some domains a role in nucleolus biology can be assumed, for others a function in the nucleolus is hard to imagine. The zinc knuckle is a zinc binding motif of the CCHC type. Besides its frequent occurrence in retroviral nucleocapsid proteins and plant transposases, it is found in a family 5'-3'-exoribonucleases of which some act as DNA strand transferases and others in nucleocytoplasmic transport of RNA [67-70]. Based on its ssDNA and RNA-related function, a role for this domain in the nucleolus seems to be plausible. We also detected a signature of the

N-terminal domain of DNA Topoisomerase I, an enzyme that relaxes positive and negative supercoils [71]. It is generally necessary for replication, recombination, and for transcription, here probably of ribosomal genes in the NOR. The armadillo repeat mediates protein-protein interactions and was first discovered in the Drosophila segment polarity gene armadillo, a homologue of the human nucleocytoplasmic signalling protein β-catenin. Both regulate transcription and cell division via HMG box transcription factors of the TCF/LEF family [72]. Armadillo repeats are as well present in the yeast nucleolar protein Srp1p, which is essential for the crescent shape of yeast nucleoli [73]. The IBB domain mediates the assembly of the importin complex which is required for the nuclear localisation signal-dependent import of proteins into the nucleus [74]. The detection of proteins acting in nuclear import is not surprising when one considers the enormous amount of rRNA and protein that has to be imported into the eukaryotic nucleus [75]. The C2 domain is thought to be involved in calcium-dependent phospholipid binding of protein kinase C (PKC) [76]. The FAT and FATC domains were first characterised by their presence in a family of large proteins with partial similarity to phosphatidyl inositol kinases (PIK). Although they were called PIK-related kinases, none of them was shown to posses PIK activity, but some were shown to function as Ser/Thr kinases. Members of the FAT/FATC family include such prominent members as the Ataxia telangiectasia mutant (ATM) protein or the RAD3 protein, regulators of DNA damage response and the cell cycle [77]. The annexins are a protein family which is involved in cytoskeletal interactions and in the inhibition of phospholipases. They bind to phospholipids in a calcium-dependent manner [78]. Given the identification of different signatures of phospholipid signalling-related proteins among nucleolar proteins, it is reasonable to assume a special function of these modules in the regulation of nucleolar function or structure. The cellular function of another interesting protein family, the translationally controlled tumor proteins (TCTP), is largely unknown, although it was shown to bind tubulin and calcium. The TCTP is expressed in normal mammalian cells, but preferably in growing tumours [79,80] and its 3D structure shows similarity to the human chaperone protein Mss4 [81]. Finally, we detected the C subunit of V-type ATP synthases. For their occurrence among nucleolar proteins there is no plausible explanation. An artefact in the purification process of nucleoli for mass spectrometry can not be excluded.

Conclusions

The core proteins of the eukaryotic nucleolus stem from an archaebacterial ancestor

Nucleoli can be observed in eukaryotes but not in bacteria. On the other hand, the key function of nucleoli, ribosome biogenesis, is crucial for all living species. Their importance is stressed by the estimation that 60% of transcription in a rapidly growing yeast cell is devoted to rRNA synthesis. Generally, the process of ribosome maturation involves molecules which are not parts of the ribosome itself, as for example rRNA base modification enzymes or small guide RNAs. Because ribosome maturation seems to be essential, the core parts of the eukaryotic nucleolar machinery already must have been present in the first eukaryote and also in the last universal common ancestor (LUCA) of all presently living organisms. This requirement is reflected by the huge number of ancient protein domains in nucleolar proteins which function in the ribosome itself, in ribosome assembly or in ribosome maturation.

Some younger RNA-associated protein domains seem to have evolved after the split of archaea and eubacteria in an archaebacterial ancestor of contemporary eukaryotes. It is widely accepted that this ancestor carried rRNA genes of an archaebacterial type in its genome. Also the presence of homologous small nucleolar RNAs (snoRNAs) in archaeal and eukaryotic genomes has been reported[36]. In this study, we found that far more homologues of human nucleolar protein domains occur in archaea and not in eubacteria than vice versa. This supports a theory which proposes an archaebacterial origin of the nucleolus. In this theory, the archaebacterial domains were already present in the pre-nucleolar proteins of the first eukaryote, whereas the eubacterial domains were added subsequently. The cellular functions of most archaeal domains are directly related to the ribosome or to protein translation, others to gene regulation and transcription. This suggests a common archaeal origin of the ribosomal genes, their transcription machinery, and the apparatus for maturation as well as assembly of the ribosome.

Eubacterial nucleolar protein domains were added lately in nucleolus evolution

In later phases of nucleolus evolution, some eubacterial protein domains with other RNA-related functions or with capabilities to mediate protein-protein interactions

appeared. We assume that their genes have been transferred to the nucleolus from a eubacterial genome and that they have contributed new functions in the early evolution of eukaryotes. Furthermore, the requirement to keep the ribosome assembly process efficient in a large eukaryotic cell must have been important, finally leading to a dense sub-nuclear organelle without membranous borders. In a large eukaryotic cell, all components of the ribosome assembly process had to be brought or kept in close proximity to each other. A dilution of the key components of ribosome biogenesis would have meant to generate ribosomes less efficiently. Having in mind that eukaryotic cells have become much larger than their prokaryotic ancestors, we believe that this anti-dilution effect was the major driving force in the evolution of the nucleolar machinery towards a dense sub-nuclear compartment. The nucleolar machinery had to develop the capability to retain their function in a densely-packed environment of DNA, RNA and proteins. To achieve this goal, certainly many novel functions had to be invented to fine-tune the nucleolar system. This is reflected in our results by the huge amount of known and novel eukaryotic protein domains which mediate protein-protein interactions (e.g. WD40 or armadillo repeats) or function in the packing of nucleic acids and proteins in chromatin. In parallel to the evolution of a nuclear membrane, an efficient transport system had to be invented to transport ribosomal proteins in and fabricated subunits out of the nucleus. In a growing yeast cell, each minute ~1000 ribosomal proteins have to be imported and ~25 subunits have to be exported through nuclear pores [75]. This would explain the detection of protein domains among nucleolar proteins that are related to the transport of proteins and RNA through nuclear pore.

The chimeric nature of the nucleolar protein domain repertoire does not support an endosymbiotic origin of the nucleus

It is currently under debate whether the nucleus has an endosymbiotic origin or has evolved gradually around the genomic DNA of an archaeal precursor cell [82-86]. Our findings show the chimeric nature of an essential part of the nucleus, the nucleolus. It also revealed that not only the ribosome itself, but also the core nucleolar components involved in ribosomal RNA maturation and ribosome assembly are of archaebacterial origin. These findings support and extend the view that those parts of the first eukaryote which relate to the processing of genomic information stem from an archaebacterial ancestor of early eukaryotes [87].

What does this mean for a hypothetical scenario in which a eubacterial endosymbiont becomes the nucleus of the first eukaryote? According to such a model, an archaeal origin of the nucleolar ribosome biogenesis machinery would mean that the ribosomal and nucleolar genes were transferred from an archaeal host genome to the eubacterial symbiont nucleus to replace the endogenous genes. Given the importance of a durable integrity of the ribosome synthesis machinery to maintain effective protein synthesis which is reflected by the enormous energy cost of ribosome synthesis [75], we consider such a scenario to be highly unlikely.

Other models for nucleus evolution aim to explain the chimeric nature of the eubacterial nucleus [84,88]. Using endosymbiosis as an explanation, either an archaebacterial symbiont could have invaded a eubacterial host or an archaebacterium could have invaded another archaebacterium. Alternatively, a fusion event between an archaebacterium and a eubacterium could have led to the chimeric nucleus. In all these models a subsequent step has to be integrated in which endosymbiosis of another eubacterium finally lead to the evolution of mitochondria. Although such models can not be fully excluded by the data of this study, several points argue against them. These models predict the existence, or eventually co-existence, of three different genomes and protein synthesis machineries in the early eukaryotes, a redundancy which hardly is an effective evolutionary strategy. Probably, successful endosymbiosis between prokaryotes depends on a favourable energy constitution of the resulting cell-hybrid, e.g. the exchange and use of each others waste metabolites to produce energy. Energetic advantages are not explained by theories that propose fusion or endosymbiosis as mechanisms leading to chimeric eukaryotic nuclear genomes. In addition, an endosymbiotic origin of the nucleus fails to explain other features of nucleus biology, e.g. the nature of the nucleus membrane (no free-living prokaryote is separated from the environment in the same manner in which the nucleus is separated from the cytoplasm) or the mode of nucleus replication (no organism is known which disintegrates its cell membrane during cell division) [86,89].

Recently, Martin and Müller proposed the 'hydrogen hypothesis', a more parsimonious model of early eukaryotic evolution regarding events like endosymbiosis or fusion [90] (see figure 1). According to these authors, mitochondria evolved by endosymbiosis of an anaerobic hydrogen-producing heterotrophic α-proteobacterium in an autotrophic hydrogen-dependent archaebacterium. The chimeric origin of nuclear genes could be explained by stepwise gene transfer from the symbiont to the host genome. The nuclear membrane and nucleus

substructures like nucleoli could have evolved slowly: the origin of intracellular membrane systems like the endoplasmatic reticulum and the nucleus could have been a result of an excess of membrane synthesis enzymes [82]. With regard to nucleolus evolution, the hydrogen hypothesis is consistent with an archaeal origin of the ribosome as well as an archaeal origin of the core nucleolar machinery. It can explain subsequent eubacterial contributions of nucleolar protein domains to the nucleus by gene transfer from the hydrogenosome (=mitochondrial) genome. It is compatible with the findings, that a substantial amount of nucleolus protein domains were invented after the common ancestor of eukaryotes emerged. It is not in conflict with the structure of the nuclear membrane or its disintegration during mitosis. It avoids critical steps that are energetically not favourable in a theory proposing nucleus endosymbiosis, like the maintenance of three genomes and protein synthesis machineries without a compensating advantage in energy metabolism for each cell. Thus, it seems to be a parsimonious and elegant model that is able to explain the chimeric nature of the nucleolus proposed in this study.

Text Box: remarks on the interpretation of phylogenetic profiles of protein domains with regard to cellular evolution

Consider a contemporary protein that comprises a certain domain. It is clear that the emergence of this domain in an ancestral organism is a prerequisite for the emergence of the protein during evolution: the time-point of the emergence of the domain during evolution must have preceded, or at least conicided with the time-point of the emergence of the protein. Often the protein domain is older than the protein architecture in which it is used today. The reason for this is the frequent reuse of protein domains as functional modules during evolution. Additionally, the protein domain could in principle work in completely unrelated functional contexts in those proteins where it is detected.

So which conclusions can be drawn from a single phylogenetic domain profile? One can conclude from a phylogenetic domain profile that the protein domain was already available as a potential building block of cellular structure in those ancestral organism whose descendants have the domain. One can not conclude that this protein domain was actually already used in the context of a particular cellular structure or function in that ancestral organism.

The same rules hold for the interpretation of a whole collection of protein domain profiles. Therefore, domain profiles are valuable tools to exclude that a certain cellular structure (like a biochemical pathway or the nucleous) could have already existed at a certain timepoint during evolution. As such, they are helpful to deduce the earliest possible timepoints at which particular modules of a cellular unit, here the nucleolus, could have evolved.

We point out that similar guidelines also apply to the interpretation of phylogenetic profiles determined by other measures than protein domain absence or presence. One example is the interpretation of the presence/absence patterns of orthologous proteins from signal transduction pathways in metazoan species. Here, the uncertainty of the assignment of a clear function of a particular domain is analogous to the uncertainty about the functional meaning of the detection of an ortholog. The reasons are a) that orthologous proteins are reused during various developmental stages in a single organism and b) that different organism use the same sets of orthologous genes for the control of different developmental programs. This is illustrated by the diverse functions of the *wingless* gene of the fruit fly and its numerous metazoan orthologs: nobody would expect that the common ancestor of all organisms which have *wingless* orthologs had wings.

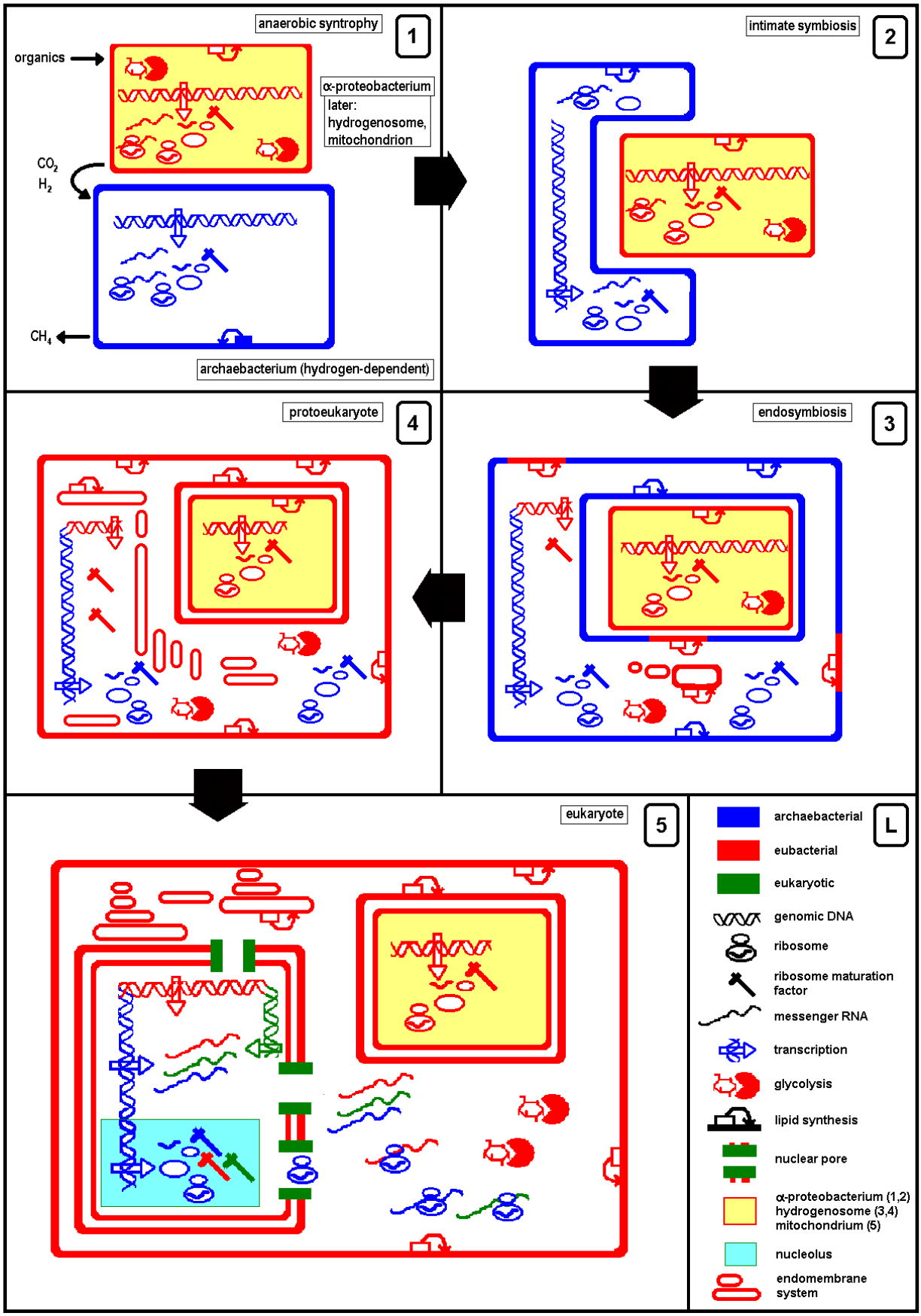

Figure 1. A possible scenario of nucleolus evolution according to the "hydrogen hypothesis for the first eukaryote".

Figure 1.

Here we illustrate how the evolution of the nucleolus can be incorporated into the "hydrogen hypothesis for the first eukaryote" [90]. This figure is adapted from Martin and Russel [91]. Our version of the figure accomodates our view of a late and continous evolution of the chimeric eukaryotic nucleolus. The "hydrogen hypothesis" predicts a late emergence of the nucleus (subsequent to the emergence of the mitochondrial precursor) and is therefore well suited to explain our results. We describe the evolution of nucleolar components in five key phases leading to the evolution of the first eukaryotic cell according to Martin and Russel [91]. For completeness substantial parts of their argumentations are repeated here. (1) Anaerobic syntrophy. Because it is energetically favourable, an α-proteobacterium and an archaebacterium share the same anaerobic environment: The (possibly facultative) anaerobic eubacterium is chemoheterotoph, can use organic molecules as energy and carbon sources and generates hydrogen as a waste product. This is willingly used by an obligate anaerobic hydrogen-dependent archaebacterium, possibly a methanogen. At this stage both prokaryotes have their own types of ribosomes and ribosome maturation factors. No characteristic compartments for ribosome maturation in these cells exist. (2) Intimate and stable symbiosis. The hydrogen-dependent archaebacterium tries to maximise its hydrogen consumption. Therefore it maximises its interacting surface while at the same time it has to ensure that the flow of carbon sources to the eubacterium continues, not to let the eubacterial fermentative production of hydrogen run dry. (3) Endosymbiosis. As soon as the archaebacterial host has found a way to feed the α-proteobacterium with carbohydrates (e.g. by symbiont-to-host lateral transfer of carbohydrate transporter genes), endosymbiosis can complete. The former external symbiont becomes a hydrogenosome. It is possible that also copies of other eubacterial genes, e.g. for glycolytic enzymes, membrane synthesis or RNA metabolism were already transferred to the archaebacterial host at this timepoint. Glycolysis probably has worked in both, the host and the symbiont cytoplasm. The enzymatic production of lipids of the eubacterial type in the archaebacterial cytoplasm could have resulted in the production of host-incompatible lipid vesicles: the beginning of an endomembrane system that will later evolve into the endoplasmatic reticulum and the nuclear membrane. The eubacterial contributions to the future eukaryotic nucleolus could have entered the host in this phase of evolution, although it is not clear whether they were used in the context of ribosome maturation so soon after

the symbiont-to-host transfer. (4) The protoeukaryote. Symbiont-to-host gene transfer has continued. Proteins synthesised in the host cytosol can now be transferred back to the hydrogenosome, allowing for a reduction of the symbiont genome. The hydrogenosome's ability to metabolise sugars is lost on its way to become a specialised organelle. The endomembrane system has extended and lipids of the eubacterial type have replaced their archaebacterial counterparts in all cellular membranes. The protoeukaryote cell is already substantially larger than its precursors. It still lacks a nuclear membrane and a nucleolus. For ribosome assembly the situation is suboptimal, because the components are diluted in the cytoplasm of the large protoeukaryotic cell. Therefore, the protoeukaryote is under pressure to form a "genome compartment" which serves to concentrate components that act in the assembly and regulation of large information-processing machineries (like the nucleolus or the transcription initiation complex). (5) The eukaryote. A facultative anaerobic heterotrophic cell with a mitochondrial precursor of endosymbiotic origin. Now a nuclear membrane is established, separating genome information management from the cytosol. Transcription and translation are uncoupled. Many new eukaryotic genes were already invented to regulate nuclear structure, e.g. proteins for nuclear import/export, the nuclear matrix and the reorganisation of the nuclear membrane during cell division. The genomic site of ribosomal gene transcription has now evolved into a dense subnuclear compartment by the reuse of eubacterial protein domains, the invention of new eukaryotic proteins and many new eukaryotic extensions of old proteins. It is not clear whether a dense pre-nucleolar structure evolved before or after the nuclear membrane. We suggest that the main driving force for the evolution of a densely-packed pre-nucleolar compartment was the compensation for the dilution of nucleolar components in a cell of larger volume. This dilution-effect would have been even larger in cells lacking nuclei. Thus, the start of the evolution of the ribosome assembly machinery towards a densely-packed compartment could have coincided with or even preceeded the start of nucleus evolution.

Materials and Methods

Sequence databases

During this study we used the following databases: the non-redundant protein database (nr) at the NCBI, the pfamseq database version 7, the nrdb90 database, the NCBI pdbaa database of protein sequences with solved 3D structures, the International Protein Index (IPI) databases of *Homo sapiens* and *Mus musculus* proteins, the wormpep database version 79 of *Caenorhabditis elegans* proteins, the NCBI databases yeast.aa and drosoph.aa of *Saccharomyces cerevisiae* and *Drosophila melanogaster*, the *Arabidopsis thaliana* protein set from the EBI, and protein sets from completely sequenced bacterial genomes provided by the EBI, namely those of the eubacteria *Bacillus subtilis*, *Borrelia burgdorferi*, *Brucella melitensis*, *Campylobacter jejuni*, *Caulobacter crescentus*, *Chlamydia trachomatis*, *Clostridium acetobutylicum*, *Deinococcus radiodurans*, *Escherichia coli* K12, *Haemophilus influenzae*, *Lactococcus lactis*, *Pseudomonas aeroguinosa*, *Rhizobium meliloti*, *Rickettsia prowazekii*, *Salmonella typhimurium*, *Synechocystis* sp. PCC6803, *Thermotoga maritima*, *Treponema pallidum* and of the archaebacteria *Archaeoglobus fulgidus*, *Halobacterium* sp. strain NRC-1, *Methanobacterium thermoautotrophicum*, *Methanococcus jannaschii*, *Pyrobaculum aerophilum*, *Pyrococcus abyssi*, *Pyrococcus horikoshi*, *Sulfolobus solfataricus*, *Sulfolobus tokodaii*, *Thermoplasma acidophilum*, *Thermoplasma volcanicum*.

Detection of known protein domains and other sequence features

Protein sequences were scanned for known domains and repeats using the Pfam database (version 7.3) [92]. Transmembrane helices were predicted using *TMHMM* version 2.0 [93]. For the prediction of signal peptides we used *SIGNALP V2.0* [94]. Sequences were investigated for the presence of coiled coils using the *COILS* algorithm [95]. Low-complexity regions were detected using the *SEG* program [96].

Repeat analysis

The program *DOTTER* [97] was used to visualise local sequence similarity when we compared sequences with themselves in order to examine them for repeats. Additionally, we refined the borders of repeat regions prior to their selection for the alignment with the help of *DOTTER*. The programs *PROSPERO* [98] and *PRSS* [99] from the *FASTA* program package were used to evaluate the significance of the repeats.

Sequence similarity searches, multiple alignments and phylogenetic trees

Pairwise sequence similarity searches were carried out using the gapped versions of the programs of the *BLAST* program package version 2.1.2 with default scoring schemes [100]. The *PSIBLAST* program was used to identify profiles and alignments based on single sequence queries. *PSIBLAST* profiles were stored using the -C option and applied using the -R option. Alignments were generated using *CLUSTALX* [101] and edited using *JALVIEW* by written by M. Clamp. The *hmmbuild* and *hmmcalibrate* programs of the *HMMER* package were used to construct HMMs from alignments with default options for model building with *hmmbuild* (hmmls/domain alignment) and calibration (sampled sequences: 5000; mean length 350) [102]. Database searches using these HMMs were carried out using the *hmmsearch* program of the same package.

Table 1

Distribution of known protein domains of the nucleolus across phyla.

The abbreviations in columns stand for *Homo sapiens* (hs), *Mus musculus* (mm), *Caenorhabditis elegans* (ce), *Drosophila melanogaster* (dm), *Arabidopsis thaliana* (at), *Saccharomyces cerevisiae* (sc), Archaeabacterial species (ar), Eubacterial species (eu). For each domain, the number of domain copies per eukaryotic genome or per bacterial lineage is given. The domains are ordered according to their distribution in eukaryotic, archaeal and eubacterial lineages. In summary lines for each classification (e.g. "eukaryotic only" or "archaeabacterial plus eukaryotic"), the numbers of domains per class are given: the left number considers all domains for which a minimum of one copy has been found in a bacterial lineage or eukaryotic genome; the right number counts only those domains which fulfil a more stringent threshold for the conclusion that a domain occurs in a distinct lineage (see results section). These latter domains can be recognised by the use of brackets which mark distinct counts. Those counts are considered less significant for the conclusion whether a domain is present in a certain lineage.

Pfam Name	hs	mm	ce	dm	at	sc	ar	eu
Domains detected in eubacteria and eukaryotes (10/7)								
3_5_exonuclease	6	4	8	5	11	1	0	30
BRCT	39	27	29	12	14	10	0	21
GTP_CDC§	33	18	2	7	2	7	0	(1)
HEAT§	16	11	6	2	7	3	0	(1)
HRDC	6	3	3	2	3	2	0	19
Topoisomerase_I§	2	2	2	1	2	1	0	(3)
rrm	475	320	128	141	237	55	0	12
WD40	424	291	130	161	221	87	0	34
dsrm	43	23	14	16	18	2	0	21
R3H	10	11	3	5	2	2	0	11
Domains detected in archaea and eukaryotes (18/16)								
CBFD_NFYB_HMF	32	22	38	6	32	8	15	0
Fibrillarin	2	4	1	2	3	1	12	0
IF_tail§	7	10	12	2	0	0	(1)	0
IMP4	2	4	2	2	2	2	8	0
LIM§	107	95	40	37	11	4	(1)	0
Sm	31	26	17	16	25	16	21	0
eIF-5a	5	1	2	1	3	2	11	0
EIF-5a_N	7	2	2	1	3	2	12	0
eIF6	2	1	1	1	2	1	11	0
eRF1_1	6	2	3	3	5	2	23	0
eRF1_2	4	2	3	3	5	2	22	0
eRF1_3	4	2	3	3	5	2	23	0
RNA_pol_H	3	1	1	1	6	1	12	0
Nop	6	3	3	3	7	3	12	0
Ribosomal_L15e	14	5	1	1	2	2	12	0

Pfam Name	hs	mm	ce	dm	at	sc	ar	eu
Ribosomal_L31e	27	13	2	1	3	2	12	0
Ribosomal_S4e	16	4	1	2	3	2	12	0
Ribosomal_S3Ae	49	5	1	1	2	2	12	0
Domains detected only in eukaryotes (29)								
ARID	30	15	4	6	8	3	0	0
Armadillo_seg	79	49	9	15	70	4	0	0
C2	211	188	58	42	105	10	0	0
Chromo_shadow	12	7	4	4	0	0	0	0
FAT	11	8	4	4	4	5	0	0
FATC	13	11	7	6	4	5	0	0
G-patch	37	31	17	17	14	4	0	0
HMG_box	124	82	17	23	15	7	0	0
IBB	14	11	3	4	7	1	0	0
Nucleoplasmin	28	9	0	2	1	0	0	0
PARP	10	6	4	2	3	0	0	0
PARP_reg	5	3	4	1	3	0	0	0
PI3_PI4_kinase	35	29	13	11	9	8	0	0
Ribosomal_L6e	14	3	1	2	3	2	0	0
Ribosomal_L14e	3	3	1	1	2	2	0	0
Ribosomal_L22e	10	3	1	2	2	2	0	0
Ribosomal_L27e	8	4	1	1	3	2	0	0
SAP	32	24	7	8	8	5	0	0
SRP14	4	1	1	1	1	1	0	0
TCTP	11	2	1	1	2	1	0	0
Topoisomer_I_N	3	1	2	1	2	1	0	0
V-ATPase_C	2	4	1	3	1	1	0	0
actin	64	35	12	14	19	8	0	0
annexin	31	19	4	4	7	0	0	0
chromo	39	29	19	15	13	2	0	0
histone	73	50	74	5	46	10	0	0
ubiquitin	91	75	27	26	68	12	0	0
zf-CCHC	53	54	34	22	173	11	0	0
zf-PARP	6	2	3	1	2	0	0	0
Ancient domains detected in eu- and archaebacteria and in eukaryotes (58/53§)								
Pfam Name	hs	mm	ce	dm	at	sc	ar	eu
A1pp	9	7	1	1	4	0	9	7
ABC_tran	141	122	70	67	132	36	386	1229
ATP-synt_ab	12	7	5	10	8	4	24	70
ATP-synt_ab_C	10	7	5	10	8	4	24	38
ATP-synt_ab_N	10	7	5	10	8	4	24	52
Band_7	26	17	13	13	15	2	19	57
DEAD	147	123	78	70	116	80	126	207
DNA_gyraseB	4	2	4	1	4	1	4	32
DNA_topoisoIV	3	2	4	1	3	1	4	34
DnaJ	74	62	36	37	98	21	5	74
Exonuclease§	25	13	15	7	13	6	(2)	46
FHA	41	27	10	18	15	15	4	32
GTP_EFTU	75	53	29	32	37	27	85	182
GTP_EFTU_D2	50	26	19	20	26	15	57	130
GTP_EFTU_D3	53	10	8	10	8	5	17	21
HATPase_c	33	26	11	7	32	8	58	613
KH-domain	74	47	30	22	25	7	59	69

Pfam Name	hs	mm	ce	dm	at	sc	ar	eu
KOW	39	20	10	9	19	12	49	38
MMR_HSR1	18	13	10	8	28	12	29	67
Metallophos	43	36	65	34	67	22	94	145
Mov34§	24	15	8	10	14	4	6	(1)
Nol1_Nop2_Sun	10	6	5	6	7	3	26	21
PHD§	166	135	63	60	212	17	(2)	(1)
PUA	3	3	2	2	2	4	49	14
RNA_pol_A	5	6	3	3	7	3	13	21
RNA_pol_A2	4	6	3	3	5	3	12	18
RNase_PH	8	7	7	6	9	6	19	29
RTC	4	3	1	2	0	1	10	4
Ribosomal_L10	11	3	2	2	6	3	12	19
Ribosomal_L13	19	3	2	2	6	3	12	18
Ribosomal_L2	3	2	2	3	6	3	12	19
Ribosomal_L22	29	24	3	2	5	3	12	18
Ribosomal_L3	13	6	2	5	4	2	12	19
Ribosomal_L30	33	7	1	2	5	4	11	13
Ribosomal_L4	16	4	2	2	4	3	12	18
Ribosomal_L5	2	5	2	1	6	3	12	19
Ribosomal_L5_C	2	7	2	1	4	3	12	19
Ribosomal_L6	11	7	1	2	5	3	12	19
Ribosomal_L7Ae	40	33	6	6	11	6	20	6
Ribosomal_S13	7	6	1	1	5	3	12	18
Ribosomal_S15	3	3	2	2	5	2	12	19
Ribosomal_S17	7	5	1	1	6	3	12	19
Ribosomal_S2	55	7	2	2	5	3	12	19
Ribosomal_S4	7	2	1	3	3	3	4	21
Ribosomal_S7	8	2	2	3	5	2	12	19
Ribosomal_S9	8	4	2	1	5	3	12	19
S1	8	9	5	6	14	4	35	119
S4	12	3	3	4	13	6	23	125
SMC_C	13	7	9	6	8	6	20	18
SMC_N	10	5	10	6	10	7	29	52
SNF2_N	48	42	29	19	40	21	19	25
TruB_N	3	4	1	1	2	2	12	19
cpn60_TCP1	34	18	11	14	19	11	23	26
helicase_C	181	165	93	75	144	77	107	197
ku§	4	3	2	2	3	2	(1)	6
pro_isomerase§	96	30	20	19	30	8	(1)	28
thiored	45	32	32	30	66	10	19	57
tubulin	66	21	17	14	17	4	21	19

Table 2

Phyletic distribution and descriptions for novel protein domains of the nucleolus.

The basic organisation of the table and the abbreviations in column headings are the same as table 1. Additionally, we proposed names for each novel domain. We also provided accession numbers (ACC) which can be used to retrieve information about the alignment of a domain and the domain architecture of all proteins of a domain family from our website (see supplement). Short descriptions of each domain are given as an initial annotation for each domain.

Acc	Proposed Name	Comment	size (aa)	representative	Pfam-seq	hs	mm	ce	dm	at	sc	ar	eu
Domains detected in archaebacteria and eukaryotes (7/7)													
NUC001	NOSIC	central domain in Nop56/SIK1-like proteins	53	SIK1_YEAST	49	5	3	3	3	8	3	8	0
NUC002	GAR1L	characteristic domain in GAR1-like snoRNPs	61	GAR1_YEAST	20	4	3	2	2	3	2	7	0
NUC011	DKCLD	TruB_N/PUA domain associated; N-terminal domain of Dyskerin-like proteins	59	DKC1_RAT	27	1	1	1	1	1	1	10	0
NUC020	RS11NT	N-terminal domain of ribosomal S11/S17 proteins	39	RS11_MAIZE	44	3	4	1	1	3	2	12	0
NUC021	RS13NT	N-terminal domain of ribosomal S13/S15 proteins	60	RS13_MAIZE	37	5	2	1	1	2	1	12	0
NUC023	RS4NT	N-terminal domain of Ribosomal S4 / S4e proteins; associated with KOW domains	41	RS4_DROME	45	8	3	1	2	3	2	9	0
NUC168	MRACN	central domain in nucleolar proteins of the multi-copy repressor of ras (Mra) family	79	MRA1_SCHPO	16	1	1	1	1	1	1	9	0
Domains detected in eubacteria and eukaryotes (3/1)													
NUC009	PADR2§	domain in poly(ADP-ribose) polymerases; associated with zf-PARP, BRCT, SAM, PARP and ankyrin repeats/domains	76	PPO2_HUMAN	35	5	3	3	1	3	0	0	(2)
NUC108	MLECT	C-terminal domain of maleless-like RNA helicase family	41	MLE_DROME	45	18	27	9	9	18	6	0	4
NUC185	DSHCT§	characteristic C-terminal domain of DOB1/SKI2/helY-like DEAD box helicases	202	DOB1	24	4	2	2	1	5	2	0	(2)
Ancient domains detected in eubacteria, archaebacteria and in eukaryotes (1/1)													
NUC060	SMChinge	SMC hinge region	153	XCPC_XENLA	86	9	7	5	4	6	4	7	6

Acc	Proposed Name	Comment	size (aa)	representative	Pfam-seq	hs	mm	ce	dm	at	sc	ar	eu
Domains detected only in eukaryotes (80)													
NUC003	TCOFD	Treacher Collins-Franceschetti syndrome 1 protein tandem repeat	65	TCOF_HUMAN	3	3	1	0	0	0	1	0	0
NUC004	P68HR	characteristic repeat of p68-like RNA helicases	35	DDX5_MOUSE	7	1	1	0	0	0	0	0	0
NUC006	P120R	characteristic repeat of proliferating cell nuclear antigen P120	23	Q922K7	3	2	2	0	0	0	0	0	0
NUC007	KI67R	KI67/Chmadrin-repeat	113	KI67_HUMAN	3	3	1	0	0	0	0	0	0
NUC008	PADR1	novel domain in poly(ADP-ribose)-synthetases; located between zf-PARP domains and BRCT repeats	57	PPOL_DROME	16	3	1	1	0	2	0	0	0
NUC010	UME	characteristic domain in UVSB PI-3 kinase, MEI-41 and ESR1; associated with FAT, FATC, PI3_PI4_kinase modules	110	ESR1_YEAST	11	1	0	0	1	1	1	0	0
NUC014	ROKNT	N-terminal domain in RNP K-like proteins with KH-domains	45	ROK_MOUSE	4	4	3	0	0	0	0	0	0
NUC016	PMC2NT	N-terminal domain in 3'-5'-exonucleases with HRDC domain; putative exosome components; Polymyositis autoantigen 2	98	PMC2_HUMAN	7	2	1	1	1	0	1	0	0
NUC017	RL6NT	N-terminal domain of ribosomal L6 proteins	57	Q9HBB3	8	11	3	0	1	0	0	0	0
NUC018	RL30NT	N-terminal domain of ribosomal L30 proteins	71	RL7_MOUSE	21	18	3	1	1	4	2	0	0
NUC029	DTHCT	C-terminal domain of DNA gyrases B / topoisomerase IV / HATPase proteins	110	TP2B_HUMAN	15	3	2	0	0	0	0	0	0
NUC031	BDHCT	C-terminal domain in Bloom's syndrome DEAD helicase subfamily	41	BLM_HUMAN	4	3	1	0	0	0	0	0	0
NUC034	CHDNT	N-terminal domain in PHD/RING finger and chromo domain-associated helicases	55	CHD4_HUMAN	7	4	2	2	1	0	0	0	0
NUC036	CHDCT1	C-terminal domain A in PHD/RING finger and chromo domain-associated CHD-like helicases	120	CHD4_HUMAN	14	6	3	2	0	1	0	0	0
NUC038	CHDCT2	C-terminal domain B in PHD/RING finger and chromo domain-associated CHD-like helicases	180	CHD4_HUMAN	11	6	3	2	0	0	0	0	0
NUC045	CAFNT	N-terminal domain in family of CCR4-associated factor-like proteins with zf-CCCH and R3H domains; part of the CCR4/NOT transcription complex	136	CNO7_MOUSE	34	9	6	3	1	14	1	0	0
NUC046	PARNUCT	C-terminal domain in Poly(A)-specific ribonucleases	46	O95453	9	2	3	2	0	2	0	0	0
NUC056	IPN	domain in ILF3/p122/NF45 transcription factors; associated with dsrm repeats	154	ILF3_HUMAN	43	12	7	2	2	0	0	0	0
NUC059	NOPS	C-terminal domain of NONA and PSP1 proteins	53	SFPQ_HUMAN	20	9	4	2	1	0	0	0	0

Acc	Proposed Name	Comment	size (aa)	representative	Pfam-seq	hs	mm	ce	dm	at	sc	ar	eu
NUC062	CBFMK21	characteristic domain of CCAAT-box binding transcription factors and MAK21-like proteins; implications in ribosome biogenesis and transcription regulation	40	CBF_HUMAN	19	3	4	3	4	2	3	0	0
NUC063	zf-RNPHF	novel putative zinc-binding domain (CHHC motif) in RNP H and F; rrm repeat-associated	36	ROH1_HUMAN	6	3	4	0	0	0	0	0	0
NUC064	RBM1CTR	C-terminal repeat in RBM1-like RNA binding hnRNPs; associated with rrm repeats in the N-terminus	46	O75526	15	21	3	0	0	0	0	0	0
NUC068	PrCBPCN	central domain in Poly(rC)-binding proteins; associated with KH domain	132	PCB2_HUMAN	14	12	5	0	1	0	0	0	0
NUC069	PRO8NT	N-terminal domain in pre-mRNA splicing factors of PRO8 family	155	YLJ6_CAEEL	13	6	1	1	1	2	1	0	0
NUC071	PROCN	central domain in pre-mRNA splicing factors of PRO8 family	426	YLJ6_CAEEL	13	5	1	1	1	2	1	0	0
NUC072	PRO8CT	C-terminal domain in pre-mRNA splicing factors of PRO8 family	129	YLJ6_CAEEL	13	5	1	1	1	2	1	0	0
NUC083	DIP2CT	novel domain C-terminal to WD40 repeats in Dom34p-interacting protein 2 from yeast; role in regulation of translation	103	DIP2_YEAST	8	2	1	1	1	1	1	0	0
NUC086	BysCR	conserved region in proteins of the Bystin family; interaction with trophinin, tastin and cytokeratin; unusual occurence in nucleolar protein	256	BYST_HUMAN	9	2	1	1	1	1	1	0	0
NUC087	NOGCT	C-terminal domain characteristic of NOG subfamily of nucleolar GTP-binding proteins	134	NOG1_TRYBB	15	3	1	1	1	2	1	0	0
NUC091	NGP1NT	N-terminal domain characteristic for subfamily of hypothetical nucleolar GTP-binding proteins similar to human NGP1	134	NGP1_HUMAN	14	2	1	2	1	1	1	0	0
NUC094	FerI	present in proteins of the Ferlin family; often central between two C2 domains	72	DYSF_HUMAN	10	9	7	1	1	0	0	0	0
NUC095	FerA	central domain A in proteins of the Ferlin family	67	DYSF_HUMAN	18	8	6	2	0	0	0	0	0
NUC096	FerB	central domain B in proteins of the Ferlin family	79	DYSF_HUMAN	18	11	7	2	0	0	0	0	0
NUC098	FerC	central domain C in proteins of the Ferlin family	120	DYSF_HUMAN	18	12	8	1	1	0	0	0	0
NUC102	TRAUB	C-terminal conserved domain of traube proteins	87	Q9JKX4	11	2	1	1	1	1	1	0	0
NUC103	NUC103/4	central domain hypothetical nucleolar proteins of novel family defined by alignments NUC103/104	156	YIJ1_YEAST	8	1	1	1	1	1	1	0	0
NUC104	NUC103/4	C-terminal domain hypothetical nucleolar proteins of novel family defined by alignments NUC103/104	149	YIJ1_YEAST	8	1	1	1	1	1	1	0	0
NUC105	MLENT	N-terminal domain of maleless-like RNA helicase family	128	MLE_DROME	6	2	0	1	1	1	1	0	0

Acc	Proposed Name	Comment	size (aa)	representative	Pfam-seq	hs	mm	ce	dm	at	sc	ar	eu
NUC109	DPOCT	central domain of proteins from DNA polymerase type V subfamily	71	DPO5_YEAST	7	2	1	0	0	0	1	0	0
NUC110	CDC5PAD	central domain between Ubox (RING-finger like) domain and WD40 repeats in spliceosome/cdc5p-associated proteins; possibly degenerated WD40 repeats	117	CWF8_SCHPO	10	2	1	1	1	2	0	0	0
NUC111	PESCNT	N-terminal domain in pescadillo-like proteins with BRCA1 C-terminus domain	139	YG2S_YEAST	10	3	3	1	1	1	1	0	0
NUC114	NUBF	N-terminal domain in UBF transcription factors; possibly degenerated HMG box	100	UBF1_MOUSE	9	4	2	0	0	0	0	0	0
NUC119	CPL	C-terminal domain in Penguin-like proteins associated with Pumilio repeats	159	PEN_DROME	3	1	1	1	1	0	1	0	0
NUC121	AARP2CN	AARP2 central domain; weakly similar to GTP-binding domain of elongation factor TU	91	Q94649	18	6	6	2	2	2	2	0	0
NUC123	AARP2CT	AARP2 family C-terminal domain	208	Q94649	19	11	6	2	2	2	2	0	0
NUC125	NUC125	central conserved domain in novel family of hypothetical proteins defined by NUC125	73	Q9Y3B9	9	1	1	1	1	1	1	0	0
NUC126	NUC126	novel family of hypothetical nucleolar proteins defined by NUC126	194	YQ52_CAEEL	12	1	1	1	1	1	1	0	0
NUC127	NOP5NT	N-terminal domain in RNA-binding proteins of the NOP5 family	68	NOP5_RAT	27	2	1	2	2	4	2	0	0
NUC129	NUC129	C-terminal domain in novel family of hypothetical nucleolar proteins defined by NUC129	63	Q9UMY1	4	1	2	0	0	0	0	0	0
NUC130	NUC130/3NT	N-terminal domain of novel nucleolar protein family defined by NUC130/133; weakly similar to TFIIF beta subunit	52	YBLE_SCHPO	8	3	1	1	1	2	1	0	0
NUC133	NUC130/3CT	C-terminal domain of novel family of nucleolar proteins defined by NUC130/133	114	YBLE_SCHPO	11	2	1	1	1	2	1	0	0
NUC135	NLE	redefined Nle domain of a family of proteins founded by fly notchless protein and yeast microtubule-associated protein YTM1; located N-terminal to WD40 repeats	71	YTM1_YEAST	13	4	2	1	2	1	1	0	0
NUC136	zf-LYAR	novel C2HC-type zinc finger in LYAR-like cell growth-regulating proteins; associated with rrm domains; present in one or two copies per protein	62	LYAR_MOUSE	8	3	12	2	7	2	1	0	0
NUC141	BING4CT	C-terminal domain in BING4 family of nucleolar WD40 repeat proteins	80	BIN4_HUMAN	12	2	1	1	1	1	1	0	0

Acc	Proposed Name	Comment	size (aa)	representative	Pfam-seq	hs	mm	ce	dm	at	sc	ar	eu
NUC142	KRSL	characteristic KR-rich domain for novel family of nucleolar proteins; SAS10 is a derepressor of silencing; LCP5 is a U3 snRNP component; domain is combined with a basic leucine zipper in one protein	69	LCP5_YEAST	12	2	2	1	2	1	1	0	0
NUC145	NNRR	central domain in NNP1/RRP1-like proteins	144	NNP1_HUMAN	10	2	3	1	1	0	1	0	0
NUC152	GUCT	C-terminal domain characteristic for RNA helicase II / Gu protein family	108	DD21_HUMAN	14	2	3	0	0	1	0	0	0
NUC153	NUC153	small domain in novel nucleolar protein family defined by NUC153	30	YG3J_YEAST	8	3	4	1	1	1	2	0	0
NUC156	NUC156	C-terminal domain in nucleolar proteins of family NUC156	151		6	1	1	1	1	1	0	0	0
NUC160	DBP10CT	characteristic C-terminal domain for Dbp10p subfamily of hypothetical RNA helicases	68	DBPA_YEAST	8	2	1	2	1	1	1	0	0
NUC161	CBFNT	N-terminal domain of CARG-binding factor A-like proteins; combined with rrm domains	76	Q98UD3	12	3	1	0	0	0	0	0	0
NUC162	RBB1NT	characteristic domain N-terminal to ARID/BRIGHT domain in DNA binding proteins of Retinoblastoma-binding protein 1 family	100	RBB1_HUMAN	4	6	2	0	1	0	0	0	0
NUC164	MAK16NT	N-terminal domain in MAK16-like proteins	139	MK16_YEAST	12	2	1	1	1	1	1	0	0
NUC167	Y112CN	central domain in nucleolar proteins of family NUC167	50	Y112_HUMAN	10	2	2	1	0	1	1	0	0
NUC169	BOP1NT	N-terminal domain in BOP1-like WD40 proteins	286	P97452	9	1	1	1	1	1	1	0	0
NUC173	NUC173	central domain of novel family of hypothetical nucleolar proteins defined by NUC173	203	Q9VYA7	8	2	2	1	1	2	1	0	0
NUC176	NOPP140CT	C-terminal domain in Nopp140-like proteins	72	Q91803	9	2	1	1	1	1	1	0	0
NUC177	TAHNT	N-terminal domain defining a novel family of nucleolar translational activator proteins with HEAT repeats	66	YAQ5_SCHPO	5	3	3	1	0	1	1	0	0
NUC188	POPLD	novel domain in family POP1-like nucleolar proteins	108	POP1_HUMAN	6	1	1	1	1	0	1	0	0
NUC189	NUC189	characteristic domain in NUC189 family of nucleolar proteins	90	Q9LFN2	9	1	2	1	1	2	2	0	0
NUC191	NUC191	domain A in the catalytic subunit of DNA-dependent protein kinases	515	PRKD_HUMAN	4	1	1	0	0	0	0	0	0
NUC194	NUC194	domain B in the catalytic subunit of DNA-dependent protein kinases	399	PRKD_HUMAN	4	1	1	0	0	0	0	0	0
NUC200	MPP10	characteristic domain in U3 snRNP mpp10-like proteins	88	MP10_YEAST	7	1	1	1	1	1	1	0	0
NUC201	NUC201	N-terminal domain in hypothetical nucleolar proteins with NUC202 tandem repeat	86	Q9DBD5	4	3	2	0	0	0	0	0	0
NUC202	NUC202	NUC202 repeat; characteristic for a novel family of nucleolar proteins	76	Q9DBD5	4	3	1	0	0	0	0	0	0

Acc	Proposed Name	Comment	size (aa)	representative	Pfam-seq	hs	mm	ce	dm	at	sc	ar	eu
NUC203	NUC203	C-terminal domain in novel family of hypothetical nucleolar WD40 repeat proteins	87	YC47_SCHPO	5	3	1	1	1	1	1	0	0
NUC205	NUC205	characteristic domain for novel family NUC205 of nucleolar proteins	44	Q9VW10	3	1	1	0	1	0	0	0	0
NUC209	BP28NT	N-terminal domain of BAP28-like nucleolar proteins	286	BP28_DROME	6	1	2	1	1	1	1	0	0
NUC211	BP28CT	C-terminal domain of BAP28-like nucleolar proteins	171	BP28_DROME	6	1	1	1	1	1	1	0	0
NUC213	NUC213	N-terminal domain in hypothetical nucleolar proteins of novel NUC213 family	36	YEV6_YEAST	14	2	4	1	1	1	1	0	0

References

1. Melese T, Xue Z. The nucleolus: an organelle formed by the act of building a ribosome. Curr Opin Cell Biol 1995;7(3):319-324.
2. Olson MO, Dundr M, Szebeni A. The nucleolus: an old factory with unexpected capabilities. Trends Cell Biol 2000;10(5):189-196.
3. Schneiter R, Kadowaki T, Tartakoff AM. mRNA transport in yeast: time to reinvestigate the functions of the nucleolus. Mol Biol Cell 1995;6(4):357-370.
4. Politz JC, Yarovoi S, Kilroy SM, Gowda K, Zwieb C, Pederson T. Signal recognition particle components in the nucleolus. Proc Natl Acad Sci U S A 2000;97(1):55-60.
5. Gerbi SA, Lange TS. All small nuclear RNAs (snRNAs) of the [U4/U6.U5] Tri-snRNP localize to nucleoli; Identification of the nucleolar localization element of U6 snRNA. Mol Biol Cell 2002;13(9):3123-3137.
6. Mitchell JR, Wood E, Collins K. A telomerase component is defective in the human disease dyskeratosis congenita. Nature 1999;402(6761):551-555.
7. Bertrand E, Houser-Scott F, Kendall A, Singer RH, Engelke DR. Nucleolar localization of early tRNA processing. Genes Dev 1998;12(16):2463-2468.
8. Andersen JS, Lyon CE, Fox AH, Leung AK, Lam YW, Steen H, Mann M, Lamond AI. Directed proteomic analysis of the human nucleolus. Curr Biol 2002;12(1):1-11.
9. Doerks T, Copley RR, Schultz J, Ponting CP, Bork P. Systematic identification of novel protein domain families associated with nuclear functions. Genome Res 2002;12(1):47-56.
10. de la Cruz J, Kressler D, Linder P. Unwinding RNA in Saccharomyces cerevisiae: DEAD-box proteins and related families. Trends Biochem Sci 1999;24(5):192-198.
11. Bycroft M, Hubbard TJ, Proctor M, Freund SM, Murzin AG. The solution structure of the S1 RNA binding domain: a member of an ancient nucleic acid-binding fold. Cell 1997;88(2):235-242.
12. Aravind L, Koonin EV. Novel predicted RNA-binding domains associated with the translation machinery. J Mol Evol 1999;48(3):291-302.
13. Kyrpides NC, Woese CR, Ouzounis CA. KOW: a novel motif linking a bacterial transcription factor with ribosomal proteins. Trends Biochem Sci 1996;21(11):425-426.
14. Burd CG, Dreyfuss G. Conserved structures and diversity of functions of RNA-binding proteins. Science 1994;265(5172):615-621.
15. Palm GJ, Billy E, Filipowicz W, Wlodawer A. Crystal structure of RNA 3'-terminal phosphate cyclase, a ubiquitous enzyme with unusual topology. Structure Fold Des 2000;8(1):13-23.
16. Mitchell P, Petfalski E, Shevchenko A, Mann M, Tollervey D. The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'-->5' exoribonucleases. Cell 1997;91(4):457-466.
17. Allmang C, Mitchell P, Petfalski E, Tollervey D. Degradation of ribosomal RNA precursors by the exosome. Nucleic Acids Res 2000;28(8):1684-1691.

18. Lafontaine DL, Bousquet-Antonelli C, Henry Y, Caizergues-Ferrer M, Tollervey D. The box H + ACA snoRNAs carry Cbf5p, the putative rRNA pseudouridine synthase. Genes Dev 1998;12(4):527-537.

19. Kelley WL. The J-domain family and the recruitment of chaperone power. Trends Biochem Sci 1998;23(6):222-227.

20. Prasad TK, Stewart CR. cDNA clones encoding Arabidopsis thaliana and Zea mays mitochondrial chaperonin HSP60 and gene expression during seed germination and heat shock. Plant Mol Biol 1992;18(5):873-885.

21. Hemmingsen SM, Woolford C, van der Vies SM, Tilly K, Dennis DT, Georgopoulos CP, Hendrix RW, Ellis RJ. Homologous plant and bacterial proteins chaperone oligomeric protein assembly. Nature 1988;333(6171):330-334.

22. Qi Y, Pei J, Grishin NV. C-terminal domain of gyrase A is predicted to have a beta-propeller structure. Proteins 2002;47(3):258-264.

23. Wigley DB, Davies GJ, Dodson EJ, Maxwell A, Dodson G. Crystal structure of an N-terminal fragment of the DNA gyrase B protein. Nature 1991;351(6328):624-629.

24. Durocher D, Henckel J, Fersht AR, Jackson SP. The FHA domain is a modular phosphopeptide recognition motif. Mol Cell 1999;4(3):387-394.

25. Stark H, Rodnina MV, Rinke-Appel J, Brimacombe R, Wintermeyer W, van Heel M. Visualization of elongation factor Tu on the Escherichia coli ribosome. Nature 1997;389(6649):403-406.

26. Vernet C, Ribouchon MT, Chimini G, Pontarotti P. Structure and evolution of a member of a new subfamily of GTP-binding proteins mapping to the human MHC class I region. Mamm Genome 1994;5(2):100-105.

27. Aravind L, Koonin EV. Phosphoesterase domains associated with DNA polymerases of diverse origins. Nucleic Acids Res 1998;26(16):3746-3752.

28. Harvey SH, Krien MJ, O'Connell MJ. Structural maintenance of chromosomes (SMC) proteins, a family of conserved ATPases. Genome Biol 2002;3(2).

29. Goodwin GH. Isolation of cDNAs encoding chicken homologues of the yeast SNF2 and Drosophila Brahma proteins. Gene 1997;184(1):27-32.

30. Martin JL. Thioredoxin--a fold for all reasons. Structure 1995;3(3):245-250.

31. Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, Grayhack EJ, Phizicky EM. A biochemical genomics approach for identifying genes by the activity of their products. Science 1999;286(5442):1153-1155.

32. Hung LW, Wang IX, Nikaido K, Liu PQ, Ames GF, Kim SH. Crystal structure of the ATP-binding subunit of an ABC transporter. Nature 1998;396(6712):703-707.

33. Tavernarakis N, Driscoll M, Kyrpides NC. The SPFH domain: implicated in regulating targeted protein turnover in stomatins and other membrane-associated proteins. Trends Biochem Sci 1999;24(11):425-427.

34. Anantharaman V, Koonin EV, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res 2002;30(7):1427-1464.

35. Vellai T, Takacs K, Vida G. A new aspect to the origin and evolution of eukaryotes. J Mol Evol 1998;46(5):499-507.

36. Omer AD, Lowe TM, Russell AG, Ebhardt H, Eddy SR, Dennis PP. Homologs of small nucleolar RNAs in Archaea. Science 2000;288(5465):517-522.

37. Mayer C, Suck D, Poch O. The archaeal homolog of the Imp4 protein, a eukaryotic U3 snoRNP component. Trends Biochem Sci 2001;26(3):143-144.

38. Davies C, Gerstner RB, Draper DE, Ramakrishnan V, White SW. The crystal structure of ribosomal protein S4 reveals a two-domain molecule with an extensive RNA-binding surface: one domain shows structural homology to the ETS DNA-binding motif. Embo J 1998;17(16):4545-4558.

39. Zwickl P, Lupas A, Baumeister W. The Thermoplasma acidophilum rpl15 gene encodes a homologue of eukaryotic ribosomal proteins L15/YL10. Biochem Biophys Res Commun 1995;209(2):684-688.

40. Koonin EV. Multidomain organization of eukaryotic guanine nucleotide exchange translation initiation factor eIF-2B subunits revealed by analysis of conserved sequence motifs. Protein Sci 1995;4(8):1608-1617.

41. Peat TS, Newman J, Waldo GS, Berendzen J, Terwilliger TC. Structure of translation initiation factor 5A from Pyrobaculum aerophilum at 1.75 A resolution. Structure 1998;6(9):1207-1214.

42. Si K, Maitra U. The Saccharomyces cerevisiae homologue of mammalian translation initiation factor 6 does not function as a translation initiation factor. Mol Cell Biol 1999;19(2):1416-1426.

43. Song H, Mugnier P, Das AK, Webb HM, Evans DR, Tuite MF, Hemmings BA, Barford D. The crystal structure of human eukaryotic release factor eRF1-- mechanism of stop codon recognition and peptidyl-tRNA hydrolysis. Cell 2000;100(3):311-321.

44. Burley SK, Xie X, Clark KL, Shu F. Histone-like transcription factors in eukaryotes. Curr Opin Struct Biol 1997;7(1):94-102.

45. Hermann H, Fabrizio P, Raker VA, Foulaki K, Hornig H, Brahms H, Luhrmann R. snRNP Sm proteins share two evolutionarily conserved sequence motifs which are involved in Sm protein-protein interactions. Embo J 1995;14(9):2076-2088.

46. Cramer P, Bushnell DA, Fu J, Gnatt AL, Maier-Davis B, Thompson NE, Burgess RR, Edwards AM, David PR, Kornberg RD. Architecture of RNA polymerase II and implications for the transcription mechanism. Science 2000;288(5466):640-649.

47. Birney E, Kumar S, Krainer AR. Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. Nucleic Acids Res 1993;21(25):5803-5816.

48. Morozov V, Mushegian AR, Koonin EV, Bork P. A putative nucleic acid-binding domain in Bloom's and Werner's syndrome helicases. Trends Biochem Sci 1997;22(11):417-418.

49. Grishin NV. The R3H motif: a domain that binds single-stranded nucleic acids. Trends Biochem Sci 1998;23(9):329-330.

50. Moser MJ, Holley WR, Chatterjee A, Mian IS. The proofreading domain of Escherichia coli DNA polymerase I and other DNA and/or RNA exonuclease domains. Nucleic Acids Res 1997;25(24):5110-5118.

51. Gray MD, Shen JC, Kamath-Loeb AS, Blank A, Sopher BL, Martin GM, Oshima J, Loeb LA. The Werner syndrome protein is a DNA helicase. Nat Genet 1997;17(1):100-103.

52. Smith TF, Gaitatzes C, Saxena K, Neer EJ. The WD repeat: a common architecture for diverse functions. Trends Biochem Sci 1999;24(5):181-185.

53. Bork P, Hofmann K, Bucher P, Neuwald AF, Altschul SF, Koonin EV. A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. Faseb J 1997;11(1):68-76.

54. Bustin M, Lehn DA, Landsman D. Structural features of the HMG chromosomal proteins and their genes. Biochim Biophys Acta 1990;1049(3):231-243.

55. Kuwano Y, Olvera J, Wool IG. The primary structure of rat ribosomal protein L38. Biochem Biophys Res Commun 1991;175(2):551-555.

56. Gallagher RA, McClean PM, Malik AN. Cloning and nucleotide sequence of a full length cDNA encoding ribosomal protein L27 from human fetal kidney. Biochim Biophys Acta 1994;1217(3):329-332.

57. Birse DE, Kapp U, Strub K, Cusack S, Aberg A. The crystal structure of the signal recognition particle Alu RNA binding heterodimer, SRP9/14. Embo J 1997;16(13):3757-3766.

58. Aravind L, Koonin EV. G-patch: a new conserved domain in eukaryotic RNA-processing proteins and type D retroviral polyproteins. Trends Biochem Sci 1999;24(9):342-344.

59. Thomas JO, Travers AA. HMG1 and 2, and related 'architectural' DNA-binding proteins. Trends Biochem Sci 2001;26(3):167-174.

60. Smith S. The world according to PARP. Trends Biochem Sci 2001;26(3):174-179.

61. Koonin EV, Zhou S, Lucchesi JC. The chromo superfamily: new members, duplication of the chromo domain and possible role in delivering transcription regulators to chromatin. Nucleic Acids Res 1995;23(21):4229-4233.

62. Aasland R, Stewart AF. The chromo shadow domain, a second chromo domain in heterochromatin- binding protein 1, HP1. Nucleic Acids Res 1995;23(16):3168-3174.

63. Gregory SL, Kortschak RD, Kalionis B, Saint R. Characterization of the dead ringer gene identifies a novel, highly conserved family of sequence-specific DNA-binding proteins. Mol Cell Biol 1996;16(3):792-799.

64. Aravind L, Koonin EV. SAP - a putative DNA-binding motif involved in chromosomal organization. Trends Biochem Sci 2000;25(3):112-114.

65. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 1997;389(6648):251-260.

66. Ito T, Tyler JK, Bulger M, Kobayashi R, Kadonaga JT. ATP-facilitated chromatin assembly with a nucleoplasmin-like protein from Drosophila melanogaster. J Biol Chem 1996;271(40):25041-25048.

67. Zhang M, Yu L, Xin Y, Hu P, Fu Q, Yu C, Zhao S. Cloning and mapping of the XRN2 gene to human chromosome 20p11.1-p11.2. Genomics 1999;59(2):252-254.

68. Till DD, Linz B, Seago JE, Elgar SJ, Marujo PE, Elias ML, Arraiano CM, McClellan JA, McCarthy JE, Newbury SF. Identification and developmental expression of a 5'-3' exoribonuclease from Drosophila melanogaster. Mech Dev 1998;79(1-2):51-55.

69. Dykstra CC, Kitada K, Clark AB, Hamatake RK, Sugino A. Cloning and characterization of DST2, the gene for DNA strand transfer protein beta from Saccharomyces cerevisiae. Mol Cell Biol 1991;11(5):2583-2592.

70. Amberg DC, Goldstein AL, Cole CN. Isolation and characterization of RAT1: an essential gene of Saccharomyces cerevisiae required for the efficient nucleocytoplasmic trafficking of mRNA. Genes Dev 1992;6(7):1173-1189.

71. Redinbo MR, Stewart L, Kuhn P, Champoux JJ, Hol WG. Crystal structures of human topoisomerase I in covalent and noncovalent complexes with DNA. Science 1998;279(5356):1504-1513.

72. Cavallo R, Rubenstein D, Peifer M. Armadillo and dTCF: a marriage made in the nucleus. Curr Opin Genet Dev 1997;7(4):459-466.

73. Yano R, Oakes ML, Tabb MM, Nomura M. Yeast Srp1p has homology to armadillo/plakoglobin/beta-catenin and participates in apparently multiple nuclear functions including the maintenance of the nucleolar structure. Proc Natl Acad Sci U S A 1994;91(15):6880-6884.

74. Moroianu J, Blobel G, Radu A. The binding site of karyopherin alpha for karyopherin beta overlaps with a nuclear localization sequence. Proc Natl Acad Sci U S A 1996;93(13):6572-6576.

75. Warner JR. The economics of ribosome biosynthesis in yeast. Trends Biochem Sci 1999;24(11):437-440.

76. Ponting CP, Parker PJ. Extending the C2 domain family: C2s in PKCs delta, epsilon, eta, theta, phospholipases, GAPs, and perforin. Protein Sci 1996;5(1):162-166.

77. Bosotti R, Isacchi A, Sonnhammer EL. FAT: a novel domain in PIK-related kinases. Trends Biochem Sci 2000;25(5):225-227.

78. Barton GJ, Newman RH, Freemont PS, Crumpton MJ. Amino acid sequence analysis of the annexin super-gene family of proteins. Eur J Biochem 1991;198(3):749-760.

79. Bohm H, Benndorf R, Gaestel M, Gross B, Nurnberg P, Kraft R, Otto A, Bielka H. The growth-related protein P23 of the Ehrlich ascites tumor: translational control, cloning and primary structure. Biochem Int 1989;19(2):277-286.

80. Chitpatima ST, Makrides S, Bandyopadhyay R, Brawerman G. Nucleotide sequence of a major messenger RNA for a 21 kilodalton polypeptide that is under translational control in mouse tumor cells. Nucleic Acids Res 1988;16(5):2350.

81. Thaw P, Baxter NJ, Hounslow AM, Price C, Waltho JP, Craven CJ. Structure of TCTP reveals unexpected relationship with guanine nucleotide-free chaperones. Nat Struct Biol 2001;8(8):701-704.

82. Martin W. A briefly argued case that mitochondria and plastids are descendants of endosymbionts, but that the nuclear compartment is not. Proc R Soc Lond 1999;266:1387-1395.

83. Margulis L. Archaeal-eubacterial mergers in the origin of Eukarya: phylogenetic classification of life. Proc Natl Acad Sci U S A 1996;93(3):1071-1076.

84. Margulis L, Dolan MF, Guerrero R. The chimeric eukaryote: origin of the nucleus from the karyomastigont in amitochondriate protists. Proc Natl Acad Sci U S A 2000;97(13):6954-6959.

85. Horiike T, Hamada K, Kanaya S, Shinozawa T. Origin of eukaryotic cell nuclei by symbiosis of Archaea in Bacteria is revealed by homology-hit analysis. Nat Cell Biol 2001;3(2):210-214.

86. Poole A, Penny D. Does endo-symbiosis explain the origin of the nucleus? Nat Cell Biol 2001;3(8):E173-174.

87. Rivera MC, Jain R, Moore JE, Lake JA. Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci U S A 1998;95(11):6239-6244.

88. Lake JA, Rivera MC. Was the nucleus the first endosymbiont? Proc Natl Acad Sci U S A 1994;91(8):2880-2881.

89. Rotte C, Martin W. Does endo-symbiosis explain the origin of the nucleus? Nat Cell Biol 2001;3(8):E173-174.

90. Martin W, Muller M. The hydrogen hypothesis for the first eukaryote. Nature 1998;392(6671):37-41.

91. Martin W, Russell MJ. On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. Philos Trans R Soc Lond B Biol Sci 2003;358(1429):59-83; discussion 83-55.

92. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2002;30(1):276-280.

93. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305(3):567-580.

94. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int J Neural Syst 1997;8(5-6):581-599.

95. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science 1991;252(5010):1162-1164.

96. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 1994;18(3):269-285.

97. Sonnhammer EL, Durbin R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 1995;167(1-2):GC1-10.

98. Mott R. Accurate formula for P-values of gapped local sequence and profile alignments. J Mol Biol 2000;300(3):649-659.

99. Pearson WR. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 2000;132:185-219.

100. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389-3402.

101. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. Multiple sequence alignment with Clustal X. Trends Biochem Sci 1998;23(10):403-405.

102. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14(9):755-763.

9 Diskussion

9.1 Überblick

Die vorliegende Arbeit setzt sich aus sieben Studien zusammen, die sich mit der Identifizierung von Proteindomänen oder -motiven in Proteinsequenzen befassen. Ziel der ersten fünf Studien ist es, durch detaillierte Proteinsequenzanalysen neue Proteindomänen zu entdecken, die Rückschlüsse auf Struktur, Funktion oder Evolution ihrer Proteine zulassen (9.2-9.6). Zwei weitere Studien haben Anwendungen der genomweiten Identifizierung von Proteindomänen zum Thema: Ziel der einen Anwendung ist die Verbesserung der genomweiten Identifizierung von kurzen Proteinmotiven, den Immunorezeptor Tyrosin-basierten inhibitorischen Motiven, durch die Einbeziehung des Domänenkontexts (9.7). Die andere Anwendung besteht in einer umfassenden Charakterisierung des Proteindomänen-repertoires des Nukleolus, mit dem Ziel durch eine komparative Analyse dieser Proteindomänen in multiplen Genomen Aufschlüsse über die Evolution des Nukleolus zu erhalten (9.8). Die Diskussion schließt mit einer Einschätzung des aktuellen Stands der Forschung an Proteindomänen und einem Ausblick in ihre Zukunft.

9.2 Die DAPIN-Domäne als vierter Subtyp der Death-Domain-Superfamilie

Die DAPIN wurde als eine gemeinsame Domäne von Proteinen identifiziert, die in unterschiedlichen Krankheitsprozessen von Vertebraten auffällig geworden sind: dem Pyrin Protein (hereditäres Familiäres Mediterranes Fieber, FMF), dem ASC Protein (programmierter Zelltod, Brustkrebs), einigen Interferon-induzierbaren Proteinen (Entzündung und Virusantwort), dem AIM2 Protein (Melanome) und den viralen M013L/G013L Proteinen (Myxoma- und Fibromavirus). Die DAPIN liegt in diesen Proteinen in Kombinationen mit verschiedenen anderen Domänen vor, kann also als evolutionär mobiles Modul angesehen werden. Besonders bemerkenswert ist, dass die DAPIN in manchen Proteinen mit bekannten Domänen aus Apoptoseproteinen kombiniert wird, wie etwa der „Caspase Recruitment Domain“ (CARD) [1] oder der katalytischen Domäne von Caspasen, die zu den ausführenden Komponenten des programmierten Zelltods gehören [2]. Dies steht im Einklang mit der Funktion der DAPIN Proteine in inflammatorischen Prozessen. Die Vorhersage der Sekundärstruktur ließ auf eine ausschließlich α-helikale dreidimensionale Faltung schließen. Die Länge der DAPIN-Sequenz beträgt etwa 95 Aminosäuren. Das postulierte Faltungsmotiv der DAPIN, ihre Länge, die Kombination von DAPIN

und CARD im ASC Protein und die Präsenz in apoptose- und entzündungsrelevanten Proteinen ließen den Schluss zu, dass die DAPIN eine vierte Subfamilie von apoptotischen Adapterdomänen der Death-Domain-Superfamilie darstellt. Diese Superfamilie beinhaltete bisher drei Domänenfamilien: die „Death Domain", die „Death Effector Domain" und die „Caspase Recruitment Domain" [3]. Alle drei Subfamilien haben ein charakteristisches dreidimensionalen Faltungsmotiv aus sechs α-Helices gemeinsam, obwohl sie eine sehr geringe Ähnlichkeit auf Sequenzebene aufweisen (<20% Identität). Eine große Anzahl derjenigen Proteine, die eine dieser drei Domänen besitzen, fungieren in der apoptotischen Signaltransduktion. Dabei spielen Domänen der Death-Domain-Superfamilie die Rolle von Adaptermodulen, welche die Interaktionen der apoptotischen Signaltransduktionsproteine steuern. Die Eigenschaften der DAPIN ließen eine ähnliche Faltung und eine verwandte Funktion erwarten.

Nach Veröffentlichung unserer Resultate lieferten diverse experimentelle Studien Belege für eine regulatorische Rolle von DAPIN-Proteinen in der Apoptose und im NF-κB Signaltransuktionsweg, dem eine besondere Rolle in der Immunantwort und in inflammatorischen Prozessen zukommt (siehe Überblicksartikel [4,5]). Weitere humane Erbkrankheiten, das Kälte-induzierte autoinflammatorische Syndrom und das Muckle-Wells-Syndrom, konnte auf Mutationen des DAPIN-kodierenden Gens CIAS1 zurückgeführt werden [6]. Die erste Struktur einer DAPIN wurde vor kurzem durch eine NMR Studie im Labor von Prof. Dr. Gottfried Otting am Karolinska Institut Stockholm aufgeklärt, an deren Interpretation ich beteiligt war (siehe Manuskript im Anhang). Demnach weist die DAPIN des ASC Proteins eine typische „Death Domain"-artige Struktur aus sechs α-Helices auf. Diese experimentellen Resultate sind ein Beleg für die Validität der von mir in dieser Arbeit vielfach benutzten Methoden der Domänenidentifizierung und Strukturvorhersage auf Proteinsequenzbasis.

9.3 Der Ssty/Spin-Repeat: Einblicke in die Evolution einer Proteinfamilie mit Funktion in der Gametogenese von Vertebraten

Nur zwei orthologe Genprodukte der Spin/Ssty-Genfamilie, die SPIN-Proteine der Maus und des Huhns, wurden bisher molekularbiologisch untersucht. Sie spielen eine Rolle in der Ausbildung des Spindelapparats während der Oogenese und werden im Zuge der intrazellulären Signaltransduktion während der Meiose phosphoryliert. Das Transkript des homologen Gens Ssty gehört zu den häufigsten Transkripten in Samenzellen der Maus.

Diese Studie ist die erste detaillierte Analyse der Sequenzen von Spin/Ssty-Genprodukten. Sie stellt zudem die erste umfassende Suche nach paralogen

Spin/Ssty-Sequenzen in Genomen von Vertebraten dar.
Die Analyse der intramolekularen Struktur von Spin/Ssty-Sequenzen ergab, dass jedes Protein dieser Familie aus einer dreifach wiederholten Einheit besteht. Die Sequenzen dieser Einheiten weisen nur noch geringe Ähnlichkeit miteinander auf. Jede Einheit bildet mit hoher Wahrscheinlichkeit eine Struktur aus vier β-Strängen. Methoden der 3D-Strukturvorhersage lieferten keine signifikanten Vorhersagen über die Verwandtschaft mit einem bekannten Faltungsmotiv. Eine phylogenetische Analyse der Proteinsequenzen dieser Einheiten deutet darauf hin, dass diese Proteinarchitektur bereits im gemeinsamen Vorfahren von Vertebraten präsent gewesen ist. Die repetitive Architektur von Spin/Ssty-Proteinen muss demnach durch zwei aufeinanderfolgende Duplikationen des Strukturmoduls entstanden sein, bevor eine Reihe von Genduplikationen zur Entstehung einer großen Genfamilie führte.
Untersuchungen der Spin/Ssty-Genstrukturen stützen diese Hypothese. In der Genstruktur des humanen Gens *SPIN* liegt jede strukturelle Einheit auf einem separaten Exon. Die paralogen Gene dieser Familie besitzen keine Introns. Da es unwahrscheinlich ist, dass die Introns im *SPIN* Gen exakt an den Grenzen der strukturellen Proteinmodule inseriert wurden, kann man annehmen, dass die Genstrukur von *SPIN* wohl die ursprüngliche Genstruktur dieser Genfamilie ist. Zwei aufeinanderfolgende Exonduplikationen in einem ursprünglichen Spin/Ssty-Gen sollten somit zur Ausbildung der repetitiven Proteinarchitektur der heutigen Spin/Ssty-Proteine geführt haben. Im Zuge der nachfolgenden Duplikationen des *SPIN*-Gens müssen die Introns früh verloren gegangen sein, da alle Paralogen keine Introns besitzen. Ein möglicher Mechanismus ist die Retrotransposition eines bereits gespleißten *SPIN*-Trankripts durch reverse Trankription und Reintegration ins Genom [7]. In einem solchen Prozess verliert das duplizierte Gen alle Introns. Den Schlüssel zur weiteren Aufklärung der evolutionären Geschichte der Spin/Ssty- Genfamilie könnte die Genomsequenzen von Chordaten bereithalten. Leider konnte im Genom der Seescheide *Ciona intestinalis* bisher kein Gen der Spin/Ssty-Familie gefunden werden.

9.4 Die strukturelle Rolle des CSPG-Repeats in NG2/MCSP-Proteinen und seine Ähnlichkeit zu Cadherin-Repeats

Das humane Protein MCSP und das orthologe Protein NG2 der Ratte spielen in Angiogenese-abhängigen Prozessen wie der Wundheilung und der Entwicklung von Tumoren eine Rolle. MCSP dient daher seit langem als Zielmolekül für die Therapeutikaentwicklung. Die Ektodomäne von MCSP/NG2 wird durch die kovalente Bindung von zahlreichen Chondroitinsulfatketten posttranslational

modifiziert. Auf Grundlage einer früheren elektronenmikroskopischen Charakterisierung des NG2 Proteins wurden bisher drei Bereiche der NG2-Ektodomäne unterschieden: ein globulärer N-Terminus, ein flexibler stäbchenförmiger zentraler Bereich und ein globulärer C-Terminus [8].

Die vorliegende Sequenzanalyse lieferte neue Hinweise auf die Domänenstruktur der NG2/MCSP-Proteinfamilie. Die Entdeckung einer neuen Familie von evolutionär mobilen, repetitiven Proteindomänen, hier genannt CSPG-Repeat, ermöglichte eine Feineinteilung der Domänenstruktur von NG2 und eine neue Interpretation der elektronenmikroskopischen Resultate. Demnach besteht der zentrale flexible Teil der NG2-Ektodomäne aus 15 Kopien des CSPG-Repeats. Tillet et al. hatten die Länge des zentrale flexiblen Teils der Ektodomäne auf 30-110 nm geschätzt. Die Länge der 15 CSPG-Repeats beträgt etwa 1700 Aminosäuren. Wäre dieser Bereich unstrukturiert, so hätte die maximal ausgestreckte Polypeptidkette eine Länge von etwa 612nm. Im Vergleich zu einer maximal gestreckten Polypeptidkette ist der zentrale Bereich der NG2-Ektodomäne also etwa um den Faktor 10 kürzer. Die Entdeckung des CSPG-Repeats als strukturelle Einheit des zentralen flexiblen Bereichs kann zur Erklärung dieser Diskrepanz herangezogen werden: die Faltung der CSPG-Repeats muss demnach die elektronenmikroskopisch zu beobachtende Länge des flexiblen Teils der Ektodomäne etwa um den Faktor 10 kürzen. Welche dreidimensionale Struktur nimmt der CSPG-Repeat ein? Verschiedene Methoden der Sekundärstrukturvorhersage deuten auf eine β-Faltblattstruktur des CSPG-Repeats hin. Zudem besitzt eine Kopie des Repeats eine niedrige, aber signifikante Sequenzähnlichkeit zu den repetitiven Einheiten in Proteinen der Cadherin Familie. In Kristallstrukturen verschiedener Cadherine liegen diese repetitiven Einheit in β-Faltblattstruktur vor [9,10]. Die sogenannten Cadherin-Repeats komprimieren die Polypeptidkette etwa um den Faktor 10 zu einer Kette aus sogenannten „β-Sandwich“ Einheiten. Die Ähnlichkeit zwischen Cadherin-Repeats und CSPG-Repeats auf Sequenzebene, die vorhergesagte Sekundärstruktur des CSPG-Repeats und seine Länge legen nahe, dass er entfernt mit dem Cadherin-Repeat verwandt ist und dass er im NG2 Protein eine ähnliche Struktur einnimmt.

Der CSPG-Repeat wurde im Zuge der Evolution in unterschiedlichen Proteinarchitekturen wiederverwendet. Er wurde kombiniert mit EGF-ähnlichen Domänen, Laminin-G Domänen und Calx-β Repeats. Er wird in diversen Vielzellern gefunden, nicht jedoch in Hefen oder anderen einzelligen Eukaryonten. Nur ein einzelner Prokaryont, das Cyanobakterium *Nostoc PCC9229* besitzt ein Protein mit einem einzelnen CSPG-Repeat. Was bedeutet das für den evolutionären Ursprung von CSPG-Repeats? Es ist unwahrscheinlich, dass der CSPG-Repeat prokaryontischen Ursprungs ist: Dies würde bedeuten, dass Gene mit CSPG-

Repeats in niederen Eukaryonten, wie etwa Hefen oder Protozoen, und in den bakteriellen Vorläufern von Eukaryontenzellen, den Archaebakterien und α-Proteobakterien, mehrfach unabhängig voneinander komplett deletiert worden sind. Eher wahrscheinlich ist, dass der CSPG-Repeat in einem Vorfahren heutiger Vielzeller entstand. Die Präsenz des CSPG-Repeats in einem prokaryotischen Protein eines Cyanobakteriums ist demnach ein Hinweis auf horizontalen Gentransfer aus einem marinen vielzelligen Lebewesen in dieses Cyanobakterium.

9.5 Die Rolle des EPTP-Repeats in verschiedenen hereditären Epilepsie-Syndromen

Vor kurzem konnte gezeigt werden, dass in zwei verschiedenen Familien mit autosomal dominanter lateraler temporaler Epilepsie (ADLTE) zwei verschiedene Mutationen des humanen Gens Leucine-rich Glioma Inactivated 1 vorliegen [11]. Die eine Mutation ist eine Deletion und bewirkt eine Leserasterverschiebung, durch die ein Protein mit verändertem und verkürztem C-Terminus gebildet wird. Die andere Mutation ist eine C→T Transition, die in einem verfrühtem Stop-Codon und somit ebenfalls in einer Trunkierung des LGI1 C-Terminus resultiert. Diese Entdeckungen warfen die Frage nach der Funktion des LGI1 C-Terminus auf. Sie waren der Anlass, eine detaillierte Sequenzanalyse des LGI1 C-Terminus durchzuführen.
In dieser Arbeit beschreibe ich die Entdeckung einer repetitiven Sequenzeinheit, dem EPTP-Repeat, im C-Terminus des LGI1 Proteins. Im Zuge der Analyse wurden die zu LGI1 paralogen Gene LGI2, LGI3 und LGI4 der Maus und des Menschen entdeckt. Alle Genprodukte der LGI Genfamilie weisen sieben EPTP-Repeats im C-Terminus auf. Die N-Termini aller Proteine der LGI-Familie bestehen aus Leucin-reichen Repeats und deren charakteristischen flankierende N- und C-terminalen Domänen. Die Entdeckung des EPTP-Repeats in LGI-Proteinen ermöglichte die Konstruktion von Sequenzmodellen zur sensitiven Suche nach entfernter verwandten Sequenzen. So konnten in zwei weiteren Proteinen EPTP-Repeats entdeckt werden. In diesen Proteinen sind die EPTP-Repeats allerdings mit anderen Domänen kombiniert als in LGI-Proteinen. Das Protein „Very Large G-Protein Coupled Receptor 1“ (VLGR1) ist ein Membranprotein mit sieben Transmembranhelices, das neben den sieben EPTP-Repeats eine Vielzahl von Calx-β Repeats und eine für G-Protein gekoppelte Rezeptoren typische GPS Domäne besitzt. Das zweite Protein wurde aus humanen ESTs und genomischen Sequenzen hergeleitet. Es besitzt neben dem EPTP-Repeats eine zum N-Terminus von Thrombospondin homologe Domäne und wurde demnach TNEP1 genannt. EPTP-Repeats wurden somit als evolutionär mobiles Modul für die Evolution von Proteinen unterschiedlicher Domänenarchitektur genutzt.

Es ist ein besonderes Charakteristikum von EPTP-Repeats, dass sie in sieben Kopien auftreten. Laut Sekundärstrukturvorhersage besteht ein EPTP-Repeats aus vier β-Strängen und ist etwa 50 Aminosäuren lang. WD40 Domänen, die häufig als Adapterdomänen in intrazellulären Signaltransduktionsproteinen wie zum Beispiel in den β-Untereinheiten von G-Proteinen vorkommen, bilden ein charakteristische Struktur aus sieben bis acht radial angeordneten β-Faltblattstrukturen mit je vier β-Strängen [12]. Obwohl EPTP-Repeats keine signifikante Sequenzähnlichkeit mit WD40 Domänen aufweisen, legen Strukturvorhersage, Periodizität und Länge der EPTP-Repeats die Vermutung nahe, dass sie ebenfalls eine β-Propeller-Struktur ausbilden.

Die besondere Bedeutung des EPTP-Repeats als charakteristisches Motiv von Epilepsie-assoziierten Proteinen wird durch die Analyse der chromosomalen Lokalisationen der humanen und murinen Gene mit EPTP-Repeats belegt. Auf die Bedeutung des humanen LGI1 Gens für die Entstehung von ADLTE wurde bereits hingewiesen. In einem Mausmodell für Epilepsie des „non-channel" Typs, der sogenannten Frings Maus, ist das murine *MASS1* Gen mutiert [13]. *MASS1* ist ortholog zum humanen Gen VLGR1. Das humane VLGR1 Gen liegt in der chromosomalen Region 5q14.1, die mit dem humanen Epilepsiesyndrom „familial febrile convulsions type 4" (FEB4) assoziiert wird [14]. Durch Studium der OMIM Datenbank fand ich heraus, dass das humane LGI4 Gen in der chromosomalen Region 19q13.12 liegt, die mit einem dritten Epilepsie-Syndrom assoziiert ist („benign familial infantile convulsions", BFIC) [15]. Daher ist LGI4 ein attraktives Kandidatengen für die Erforschung der Ursache von BFIC. Das humane TNEP1 Gen liegt in der chromosomalen Region 21q22.3 nahe der sogenannten Down-Syndrom kritischen Region [16]. Dies macht TNEP1 zu einem Kandidatengen für die Erforschung des mental-retardierten Phänotyps des Down-Syndroms. Weil der Zusammenhang mit neurologischen Krankheiten für zwei Gene bereits gezeigt ist (LGI1, VLGR1) und die chromosomalen Loci von zwei weiteren Genen mit neurologisch-auffälligen Phänotypen assoziiert sind (LGI4, TNEP1), lässt sich eine essentielle Funktion des EPTP-Repeats in der Aufrechterhaltung der Gehirnfunktion postulieren.

9.6 Die Bedeutung der Sequenzähnlichkeit zwischen NtrY- und HIG-Proteinen

Die Phosphorylierung von Serin-, Threonin- oder Tyrosin-Seitenketten von Proteinen dient häufig als Mechanismus der intrazellulären Signaltransduktion in Eukaryonten. Dagegen ist die bei Prokaryonten verbreitete Signaltransduktion durch Phosphorylierung von Histidin-Seitenketten in Eukaryonten wenig erforscht.

Im Zuge der sogenannten Zwei-Komponenten-Signalübertragung in Prokaryonten werden extrazelluläre Signale durch sensorische Histidinkinase-Rezeptoren detektiert. Diese Rezeptoren besitzen eine sensorische extrazelluläre Region, die von zwei Transmembranhelices flankiert wird. Die Rezeptoren dimerisieren als Antwort auf ein Signal, was zur Autophosphorylierung von Histidinen in ihren cytoplasmatischen Regionen führt. Die Phosphatgruppen werden anschließend auf sogenannten Receiver-Domänen von intrazellulären Regulatoren übertragen, die häufig direkt als Transkriptionsfaktoren dienen und die Transkription von Zielgenen steuern. Man weiß, dass phosphorylierte Histidine auch in der eukaryotischen Bäckerhefe etwa 6% aller phosphorylierten Aminosäuren in Kernproteinen ausmachen. Über die Phosphorylierung von Histidin in Säugetieren gibt es nur ungenaue Schätzungen. Es ist möglich, dass die Bedeutung der Histidin-Phosphorylierung in Eukaryonten bisher nicht voll erfasst worden ist.

Im Rahmen dieser Arbeit führte ich eine Analyse der Proteinsequenz des humanen Hypoxie-induzierbaren Gens (HIG) durch. Dabei zeigte sich, dass die Familie der HIG-ähnlichen eukaryotischen Proteine eine schwache Ähnlichkeit zu Proteinen der NtrY-Subfamilie bakterieller Histidinkinasen aufweist. Die Ähnlichkeit erstreckt sich ausschließlich über den Bereich der sensorischen Domäne der Histidinkinasen, die den extrazellulären Bereich und Teile der flankierenden transmembranen α-Helices umfasst. Mit Standardmethoden der paarweisen Sequenzsuche in Datenbanken wurde nur eine marginale Signifikanz der Ähnlichkeit gezeigt. Daher wurden zwei weitere Methoden eingesetzt, die auf dem Vergleich von zwei kompletten Alignments beruhen und daher sensitiver sind: COMPASS und LAMA. Beide zeigen, dass die Ähnlichkeit der NtrY- und HIG-Sequenzen deutlich höher ist, als man aufgrund von Zufallseffekten erwarten könnte.

Wegen der Einbeziehung von Transmembranhelices in den Sequenzvergleich könnte man eventuell argumentieren, dass eine ähnliche Aminosäurenzusammensetzung der Hauptgrund für die festgestellte interfamiliäre Sequenzähnlichkeit ist. Der paarweise Vergleich von Sequenzen aus beiden Familien mit dem Programm PRSS zeigte allerdings, dass die Reihenfolge der Aminosäuren in den HIG- und NtrY-Sequenzen für das Alignment sehr wichtig ist. Auch ein mit der LAMA Methode assoziiertes Programm stellte keine auffällige Unausgewogenheit in der Aminosäurenzusammensetzung unserer NtrY/HIG-Alignments fest. Dies bedeutet, dass die paarweise Ähnlichkeit der Sequenzen nicht vorrangig durch eine simple Ähnlichkeit der Aminosäurenzusammensetzung zustande gekommen ist. Daher ist die Sequenzähnlichkeit zwischen sensorischen Regionen von HIG- und NtrY-Proteinen ein starker Hinweis auf die Homologie dieser Familien.

Die bisher verfügbaren Daten über die zellulären Funktionen beider Proteinfamilien stehen in Einklang mit der Hypothese ihrer Homologie. Die NtrY-Proteine fungieren in Bakterien als Regulatoren des Stickstoffmetabolismus und sind vermutlich Sensoren für die Wahrnehmung der Sauerstoff- oder Stickstoffkonzentration [17]. Die Expression der HIG mRNA des Hypoxie-toleranten Fisches *Gillichthys mirabilis* wird während der zellulären Reaktion auf Hypoxie hochreguliert [18]. Die Funktion beider Proteine ist also von der extrazellulären Konzentration von Sauerstoff oder Stickstoff abhängig.
Weil die Sequenzähnlichkeit zwischen sensorischen Domänen der NtrY- und HIG-Proteine von manchen Methoden der Sequenzanalyse nur als marginal signifikant beurteilt wurde, sollte die von uns aufgestellt Hypothese, dass die sensorischen Domänen von NtrY- und HIG-ähnlichen Proteinen homolog zueinander sind, durch zukünftige experimentellen Studien zur biochemischen Funktion der Proteine überprüft werden. Wenn diese Experimente ebenfalls eine funktionelle Homologie der NtrY- und HIG-Proteine zeigen können, dann ist unsere Entdeckung die erste, die eine Homologie zwischen einem tierischen Protein und einem Protein aus der bakteriellen Zwei-Komponenten-Signaltransduktion beschreibt. HIG-ähnliche Proteine könnten dann einen vielversprechenden Ansatzpunkt für die Suche nach den Mechanismen der Histidin-Phosphorylierung in Eukaryonten darstellen.

9.7 Anzeichen für ITIM-abhängige Signaltransduktion in bisher unbeachteten Proteinen und humanen Geweben

Immunorezeptor Tyrosin-basierte inhibitorische Motive (ITIMs) sind kurze Proteinsequenzmotive mit der Consensussequenz {ILV}-x-x-Y-x-{LV} in den cytoplasmatischen Regionen von Immunrezeptoren. Die Phosphorylierung des Tyrosins in ITIMs ist ein wichtiger regulatorischer Mechanismus zur Kontrolle der Aktivierung verschiedener Zellen des Immunsystems. Die Verfügbarkeit der humanen Genomsequenz machte es möglich, eine breit angelegte Suche nach neuen ITIM Rezeptoren in allen humanen Proteinsequenzen durchzuführen. Allerdings haben herkömmliche Suchverfahren nach kurzen Motiven mit einer inakzeptabel hohen Rate an falsch positiven Vorhersagen zu kämpfen. Verwendet man etwa reguläre Ausdrücke zur Suche nach ITIMs, so wird für 30% der Proteine der humanen RefSeq Proteinsequenzdatenbank mindestens ein ITIM vorhergesagt.
In dieser Arbeit stelle ich eine neue Strategie zur Suche nach kurzen Proteinmotiven in großen Sequenzdatenbanken am Beispiel der Suche nach ITIMs vor. Um die Zahl der vorhergesagten ITIMs sinnvoll einzuengen, benutzte ich den Sequenzkontext eines ITIMs, das heißt Information über vorhergesagte Signalpeptide, Transmembranhelices oder bekannte Proteindomänen aus

Signaltransduktionsproteinen oder extrazellulären Proteinen. Mit Hilfe des neuen Suchalgorithmus konnte die Anzahl der vorhergesagten ITIM Rezeptoren gegenüber der Suche mit regulären Ausdrücken um etwa das 45-fache auf letztlich 109 von 16177 untersuchten Proteinen reduziert werden. Von diesen 109 Proteinen wurden 36 bereits in der Literatur als ITIM Rezeptoren beschrieben. Nur zwei uns bekannte Typ-I-Transmembranproteine mit ITIMs konnten nicht identifiziert werden: das SHP-2-interagierende transmembrane Adapterprotein (SIT), das keine extrazellulären Domänen besitzt, und der Interleukin-Rezeptor 4a (IL4R), dessen ITIM nicht der Consensussequenz entspricht.
Es konnten 29 orthologe Proteine der Maus identifiziert werden, in denen viele der bekannten sowie der neuen ITIMs konserviert sind. Dies ist ein zusätzlicher Hinweis auf die Validität der Vorhersage der humanen ITIMs. Um die Gewebespezifität der ITIM Rezeptoren zu untersuchen, wurde ein öffentlich verfügbarer Datensatz über die mRNA-Expression von etwa 12.000 Genen in humanen Geweben ausgewertet. Wie eine Analyse der mRNA-Expression der vorhergesagten ITIM Rezeptoren zeigt, ist ihre Expression nicht auf Blutzellen beschränkt. ITIM Rezeptoren scheinen in den unterschiedlichsten soliden Organen exprimiert zu werden. Bewertet man dieses Resultat mit Blick auf die ubiquitären Expressionsmuster der SHP-Phosphatasen [19,20], die wichtige Vermittler des ITIM Signals darstellen, so erscheint es vernünftig zu postulieren, dass ITIM-vermittelte Signaltransduktion nicht auf Blutzellen beschränkt ist, sondern in vielen humanen Geweben eine Rolle spielt.

9.8 Evidenz für einen chimären Ursprung und eine graduelle Evolution des Nukleolus durch Analyse seines Proteindomänen-repertoires

Vor kurzem wurde die erste massenspektrometrische Studie des humanen Nukleolus vorgestellt. Diese führte zur Identifizierung von 271 Proteinsequenzen, die mit dem Nukleolus assoziiert sind, unter ihnen viele bisher unbekannt Proteine. Dieser Datensatz ist eine wertvolle Quelle für eine Analyse der Evolution des Proteindomänenreservoires des Nukleolus. Ziel dieser Arbeit war es, die evolutionären Wurzeln der nukleolaren Proteindomänen im Reich der Bakterien auszumachen, die bekanntlich keine Nukleoli besitzen. Ausgehend von den 271 Proteinen und der PFAM Datenbank bekannter Proteindomänen, entwickelte ich ein semiautomatisches Sequenzanalyseprotokoll zur Identifizierung aller bereits bekannten und bisher unbekannten konservierten Proteindomänen des Nukleolus. Nach einzelner Begutachtung aller Sequenzalignments ergab dessen Anwendung einen Satz von 115 bekannten Proteindomänen, sowie die Entdeckung von 91

neuen Domänen.
Durch die systematische Suche nach diesen Domänen in Proteindatenbanken verschiedener kompletter Genome wurde die Präsenz jeder Domäne in verschiedenen Archaebakterien, Eubakterien und Eukaryonten ermittelt. Seit langem ist bekannt, dass die Translationsmaschinerie von Eukaryonten enger mit derjenigen von Archaebakterien als mit der von Eubakterien verwandt ist. Da Nukleoli die Orte der Entstehung von Ribosomen sind, liegt die Vermutung nahe, dass die evolutionäre Quelle der nukleolaren Proteindomänen ebenfalls eher bei den Archaebakterien zu suchen ist.
Insgesamt sind 59 Proteindomänen des Nukleolus sowohl in Archaebakterien als auch in Eubakterien zu finden. Diese Domänen waren demnach bereits im sogenannten *Last Universal Common Ancestor* (LUCA) vorhanden. Sie spiegeln die universelle Bedeutung der Translationsmaschinerie für alle Organismen wieder. Die 59 Domänen bilden den Kern der Maschinerie, die für die Reifung von Ribosomen benötigt wird. Da jedoch ein beträchtlicher Teil aller Proteindomänen des Nukleolus nicht in allen Reichen der Bakterien vorhanden waren, kann LUCA noch keinen Nukleolus im heutigen Sinn besessen haben. Dies steht in Einklang mit der Auffassung, dass alle heute lebenden Bakterien keinen Nukleolus besitzen und demnach ihr gemeinsamer Vorfahr auch keinen Nukleolus besessen hat.
Desweiteren habe ich 25 Domänen identifiziert, die zwar in Archaebakterien, nicht aber in Eukaryonten vorkommen. Dagegen stehen 13 Domänen, die anscheinend aus Eubakterien stammen. Dies beweist einen chimären Ursprung des Nukleolus. Die Entstehung des Nukleolus muss demnach vor dem Ereignis in der frühen eukaryotischen Evolution liegen, bei dem das Genom eines Archaebakteriums mit dem eines Eubakteriums in einer Zelle vereint wurde. Unter den archaebakteriellen Nukleolusdomänen haben viele eine Funktion in der Reifung der Ribosomen oder im Translationsapparat selbst. Dies ist ein Beleg dafür, dass der „urtümliche“ Teil des Nukleolus aus Archaebakterien stammt und bestätigt die bereits in anderen Arbeiten mehrfach postulierte Verwandtschaft des sogenannten „informations-verarbeitenden“ Apparats von Archaebakterien und Eukaryonten, also den Proteinen aus DNA-Replikation, Transkription und Translation. Die eubakteriellen Proteindomänen des Nukleolus besitzen keine klassischen Ribosomen-assoziierten Funktionen. Diese Proteindomänen sind wahrscheinlich erst im Laufe der frühen Eukaryontenevolution zu einer prä-nukleolaren Struktur rekrutiert worden. Legt man zum Beispiel die Hydrogenosomentheorie der Mitochondrienevolution zugrunde [21], geschah dies nachdem der eubakterielle Vorläufer von Mitochondrien, höchstwahrscheinlich ein α-Proteobakterium, durch Endosymbiose in ein Archaebakterium aufgenommen worden ist. Das besagt gleichzeitig, dass der

Nukleolus nicht älter sein kann als die Mitochondrien oder ihre Vorläufer. Weiterhin ist eine große Anzahl von nukleolaren Proteindomänen ausschließlich in Eukaryonten zu finden. Das kann teilweise ein Resultat der mangelnden Sensitivität von Methoden der Sequenzanalyse sein, die nicht immer in der Lage sind Homologie zwischen Proteinsequenzen zu entdecken, die seit über einer Milliarde Jahren divergieren. Allerdings deutet die Vielzahl der Eukaryonten-spezifischen Domänen in jedem Fall auf substantielle Veränderungen der heutigen nukleolaren Proteine und auf die Entstehung neuer Funktionen während der Evolution von Eukaryonten hin.
Die kontinuierliche, graduelle Evolution des Proteindomänenrepertoires des Nukleolus, dokumentiert durch die eubakteriellen Kontributionen von Proteindomänen und durch die zahlreichen eukaryotischen Neuentwicklungen von Domänen, spricht für eine langsame, schrittweise Entwicklung des Nukleolus in frühen Eukaryonten. Auch sein chimärer Charakter zeigt, dass der Nukleolus als subnukleare Struktur nicht durch ein einzelnes endosymbiotisches Ereignis in den ersten Eukaryonten entstanden ist. Somit sprechen die hier dargestellten Resultate auch gegen die umstrittene Hypothese eines endosymbiotischen Ursprungs des Nukleus [22].

9.9 Ausblick

Die Suche nach bekannten Proteindomänen in einer neuen Proteinsequenz hat sich über Jahrzehnte als eine erfolgreiche Strategie erwiesen, eine erste Prognose über die Funktion eines neuen Proteins zu erhalten. Die Zahl der bekannten Protein-sequenzen ist in den letzten Jahren exponentiell gewachsen. Die experimentelle Charakterisierung von Proteinen ist dagegen ein vergleichsweise langsamer Prozess. Die *in silico* Funktionsvorhersage mittels bekannter Domänen wird also in Zukunft eine noch wichtigere Rolle einnehmen.
Neue experimentelle Befunde über Proteine und deren Domänen verlangen eine ständige Aktualisierung der Annotationen in Domänendatenbanken. Ein erheblicher Teil des Wissens über individuelle Domänen ist schon heute in zahlreichen Publikationen verborgen und nur schwer zugänglich. Es ist eine große Herausforderung für die Bioinfomatik, dieses Wissen mit intelligenten Systemen zu erschließen und für Sequenz- oder Domänendatenbanken nutzbar zu machen.
Die Aktualisierung von Domänendatenbanken bedeutet auch eine Erweiterung der existierenden Domänenalignments durch neue Sequenzen. Nur dadurch kann die Qualität der Domänenmodelle auf ausreichend hohem Niveau gehalten werden, so dass die Sensitivität der Domänensuche mit dem Wachstum der Sequenzdaten-banken Schritt hält. Gerade die Konstruktion der Alignments von stark divergenten

Proteindomänen ist eine Aufgabe, die stark von Expertenwissen profitiert und bisher nur unzureichend automatisiert werden kann. Die hohe Qualität der Alignments war über Jahre der Schlüssel zum Erfolg der manuell gepflegten Domänendatenbanken wie SMART (http://smart.embl-heidelberg.de/) oder PFAM (http://www.sanger.ac.uk/Pfam/). Es ist zu hoffen, dass Fortschritte in der automatischen Erstellung von multiplen Alignments in Zukunft die Aktualisierung von Domänenkollektionen erleichtern können.

Der Aufwand der Aktualisierung von manuell gepflegten Domänendatenbanken hängt letztlich auch von der Zahl der zu pflegenden Domänen ab. Dies führt zu der Frage, wie viele Proteindomänen noch zu entdecken sind. In der Einleitung wurde bereits darauf hingewiesen, dass Schätzungen über die Zahl der verschiedenen Faltungsmotive im Proteinuniversum zwischen 400 und 8.000 variieren [23]. Ein strukturelles Faltungsmotiv wird auf Sequenzebene häufig durch mehrere Domänenfamilien repräsentiert, wie u.a. die in dieser Arbeit beschriebene DAPIN-Domänensubfamilie der Death-Domain-Superfamilie zeigt. Die Schätzungen für die Gesamtzahl von Proteinsequenzfamilien liegen daher höher, bei etwa 1000 bis 30.000 [23].

Tatsächlich lassen sich bereits heute sehr viele der heute aufgeklärten 3D-Strukturen von Proteinen in bekannte Faltungsklassifikationen einfügen. Dagegen steigt bislang die Zahl der durch Sequenzähnlichkeit definierten Proteinfamilien jedoch unaufhaltsam an. Während die Zahl der Einträge in der SMART Domänendatenbank in den letzten Jahren annähernd linear auf heute etwa 650 Datensätze gewachsen ist, hat die PFAM Datenbank von Domänen und Proteinfamilien in den letzten Jahren ein weit stärkeres Wachstum auf heute über 7200 Datensätze gezeigt. Es ist wahrscheinlich, dass die Entdeckung neuer Proteindomänen erst dann gebremst wird, wenn die Genomsequenzierung eine deutlich bessere Spezies-Abdeckung aller Zweige des Lebens erreicht hat.

Das unterschiedlich starke Wachstum von SMART und PFAM erklärt sich aus den verschiedenen Zielen der Datenbanken. SMART ist ausschließlich auf evolutionär mobile Proteindomänen fokussiert, die strukturelle Einheiten bilden. Das Ziel von PFAM ist eine möglichst gute Abdeckung des Proteinsequenzraums. Daher werden auch weniger divergente Proteinfamilien modelliert, die durchaus bisher unentdeckte gemeinsame Domänen mit anderen Proteinfamilien besitzen können. Es kommt häufig vor, dass die PFAM-Einträge zweier solcher Familien durch die spätere Entdeckung einer neuen gemeinsamen Domäne neu definiert werden müssen. Der Kreuzvergleich von Proteinfamilien und die Redefinition von Domänen und Familien wird einen signifikanten Teil der zukünftigen Arbeit an Protein-familiendatenbanken ausmachen.

Als positive Konsequenz aus ihrer Strategie hat die PFAM Datenbank eine sehr gute Abdeckung des bekannten Sequenzraums. Heute haben etwa 75% der bekannten Proteine Ähnlichkeit mit mindestens einem PFAM-Eintrag. Dies entspricht einer Abdeckung von etwa 50% der bekannten Sequenz. Für Proteine aus neu sequenzierten Genomen ist die Abdeckung allerdings deutlich geringer. Das ist ein weiteres Anzeichen dafür, dass noch immer Raum für weitere Neuentdeckungen von Proteindomänen und Proteinfamilien vorhanden ist.

Auch das lineare Wachstum der SMART Datenbank und die Monat für Monat erscheinenden Publikationen neuer evolutionär mobiler Domänen in Fachzeitschriften lassen erwarten, dass solche Domänen auch weiterhin entdeckt werden können. Ein Grund dafür ist die zunehmende Dichte des Sequenzraums, die u.a. bewirkt, dass bisher nicht detektierbare Ähnlichkeiten zwischen entfernt verwandten Sequenzen durch die Vermittlung neuer intermediärer Sequenzen entdeckt werden können. Es ist allerdings wahrscheinlich, dass zukünftige Entdeckungen von neuen Proteindomänen weniger die bereits gut erforschten Modellorganismen betreffen. Die Proteinsequenzen bisher weniger beachteter Modellorganismen bieten einen größeren Raum für neue Entdeckungen. Spezies- oder Phylum-spezifische Proteindomänen könnten sich in Zukunft als besonders wertvoll erweisen, um Aufschlüsse über die molekularen Ursachen Organismen-spezifischer Entwicklungsprozesse zu erhalten.

Die zwei Anwendungen von Domänenkollektionen im Rahmen dieser Arbeit sind exemplarisch für zahlreiche weitere Möglichkeiten, Nutzen aus dem Wissen über Proteindomänen zu ziehen. Ähnlich wie für ITIMs in dieser Arbeit gezeigt wurde, kann der Sequenzkontext für die Vorhersage vieler anderer kurzer Proteinmotive genutzt werden. Die Erstellung von phylogenetischen Profilen von Domänen-kollektionen kann zur Analyse der Evolution von Organellen oder subzellulären Einheiten dienen, wie es in dieser Arbeit am Beispiel des Nukleolus demonstriert wurde. Die Nutzung phylogenetischer Profile von Domänen könnte man durch die Erstellung phylogenetischer Profile von Domänenarchitekturen komplementieren. Solche Methoden ließen sich, ähnlich wie für den Nukleolus gezeigt, in der Analyse von Substrukturen von Protein-Protein-Interaktionsnetzwerken oder von Signal-transduktionswegen einsetzen. Dies könnte Hinweise auf den evolutionären Ursprung einzelner Netzwerk-Substrukturen geben oder allgemeine Erkenntnisse über die Co-Evolution von Proteindomänen und Proteinnetzwerken liefern.

9.10 Referenzen der Diskussion

1. Bouchier-Hayes L, Martin SJ. CARD games in apoptosis and immunity. EMBO Rep 2002;3(7):616-621.
2. Cohen GM. Caspases: the executioners of apoptosis. Biochem J 1997;326 (Pt 1):1-16.
3. Weber CH, Vincenz C. The death domain superfamily: a tale of two interfaces? Trends Biochem Sci 2001;26(8):475-481.
4. Tschopp J, Martinon F, Burns K. NALPs: a novel protein family involved in inflammation. Nat Rev Mol Cell Biol 2003;4(2):95-104.
5. Mariathasan S, Vucic D. POPing the fire into the pyrin? Biochem J 2003;373(Pt 1):1-2.
6. Hoffman HM, Mueller JL, Broide DH, Wanderer AA, Kolodner RD. Mutation of a new gene encoding a putative pyrin-like protein causes familial cold autoinflammatory syndrome and Muckle-Wells syndrome. Nat Genet 2001;29(3):301-305.
7. Cooper DN. Chapter 3: Introns, exons, and evolution. Human Gene Evolution. Oxford: BIOS Scientific Publishes Ltd.; 1999.
8. Tillet E, Ruggiero F, Nishiyama A, Stallcup WB. The membrane-spanning proteoglycan NG2 binds to collagens V and VI through the central nonglobular domain of its core protein. J Biol Chem 1997;272(16):10769-10776.
9. Tamura K, Shan WS, Hendrickson WA, Colman DR, Shapiro L. Structure-function analysis of cell adhesion by neural (N-) cadherin. Neuron 1998;20(6):1153-1163.
10. Boggon TJ, Murray J, Chappuis-Flament S, Wong E, Gumbiner BM, Shapiro L. C-cadherin ectodomain structure and implications for cell adhesion mechanisms. Science 2002;296(5571):1308-1313.
11. Morante-Redolat JM, Gorostidi-Pagola A, Piquer-Sirerol S, Saenz A, Poza JJ, Galan J, Gesk S, Sarafidou T, Mautner VF, Binelli S and others. Mutations in the LGI1/Epitempin gene on 10q24 cause autosomal dominant lateral temporal epilepsy. Hum Mol Genet 2002;11(9):1119-1128.
12. Smith TF, Gaitatzes C, Saxena K, Neer EJ. The WD repeat: a common architecture for diverse functions. Trends Biochem Sci 1999;24(5):181-185.
13. Skradski SL, Clark AM, Jiang H, White HS, Fu YH, Ptacek LJ. A novel gene causing a mendelian audiogenic mouse epilepsy. Neuron 2001;31(4):537-544.
14. Nakayama J, Hamano K, Iwasaki N, Nakahara S, Horigome Y, Saitoh H, Aoki T, Maki T, Kikuchi M, Migita T and others. Significant evidence for linkage of febrile seizures to chromosome 5q14-q15. Hum Mol Genet 2000;9(1):87-91.
15. Guipponi M, Rivier F, Vigevano F, Beck C, Crespel A, Echenne B, Lucchini P, Sebastianelli R, Baldy-Moulinier M, Malafosse A. Linkage mapping of benign familial infantile convulsions (BFIC) to chromosome 19q. Hum Mol Genet 1997;6(3):473-477.
16. Delabar JM, Theophile D, Rahmani Z, Chettouh Z, Blouin JL, Prieur M, Noel B, Sinet PM. Molecular mapping of twenty-four features of Down syndrome on chromosome 21. Eur J Hum Genet 1993;1(2):114-124.

17. Pawlowski K, Klosse U, de Bruijn FJ. Characterization of a novel Azorhizobium caulinodans ORS571 two-component regulatory system, NtrY/NtrX, involved in nitrogen fixation and metabolism. Mol Gen Genet 1991;231(1):124-138.

18. Gracey AY, Troll JV, Somero GN. Hypoxia-induced gene expression profiling in the euryoxic fish Gillichthys mirabilis. Proc Natl Acad Sci U S A 2001;98(4):1993-1998.

19. Li L, Dixon JE. Form, function, and regulation of protein tyrosine phosphatases and their involvement in human diseases. Semin Immunol 2000;12(1):75-84.

20. Plutzky J, Neel BG, Rosenberg RD. Isolation of a src homology 2-containing tyrosine phosphatase. Proc Natl Acad Sci U S A 1992;89(3):1123-1127.

21. Martin W, Muller M. The hydrogen hypothesis for the first eukaryote. Nature 1998;392(6671):37-41.

22. Horiike T, Hamada K, Kanaya S, Shinozawa T. Origin of eukaryotic cell nuclei by symbiosis of Archaea in Bacteria is revealed by homology-hit analysis. Nat Cell Biol 2001;3(2):210-214.

23. Wolf YI, Grishin NV, Koonin EV. Estimating the number of protein folds and families from complete genome data. J Mol Biol 2000;299(4):897-905.

10 Zusammenfassung

Diese Arbeit setzt sich aus sieben Studien zusammen, die sich mit der Identifizierung von Proteindomänen oder -motiven in Proteinsequenzen befassen. Ziel der ersten fünf Studien ist es, durch detaillierte Proteinsequenzanalysen neue Proteindomänen zu entdecken (a-e). Zwei weitere Manuskripte haben Anwendungen der genomweiten Identifizierung von Proteindomänen zum Thema (f-g).

(a) Die „Domain in Apoptosis and Interferon Response“ (DAPIN) wurde als gemeinsames Proteinmodul von Proteinen identifiziert, die in Krankheitsprozessen von Vertebraten auffällig geworden sind, wie dem Pyrin Protein (hereditäres familiäres mediterranes Fieber), dem ASC Protein (Apoptose, Brustkrebs), einigen Interferon-induzierbaren Proteinen (Entzündung und Virusantwort), oder dem AIM2 Protein (Melanome). Aufgrund der Kombination der DAPIN mit bekannten Domänen aus Apoptoseproteinen in einigen Proteinen und den Ergebnissen der Strukturvorhersage folgerte ich, dass die DAPIN-Familie eine vierte Subfamilie von Adapterdomänen der Death-Domain-Superfamilie darstellt.

(b) Die Spin/Ssty-Proteinfamilie spielt eine Rolle in der Ausbildung des Spindelapparats während der Gametogenese von Vertebraten. Im Zuge dieser Arbeit wurden vier menschliche Proteine der Spin/Ssty-Familie beschrieben, sowie drei der Maus, zwei des Huhns, eins des Rinds und eins des Medaka-Fischs. Alle Spin/Ssty-Proteine bestehen aus einer sich dreifach wiederholenden Einheit, dem Spin/Ssty-Repeat, der wahrscheinlich eine β-Faltblattstruktur bildet. Die phylogenetische Analyse der Repeats und die Analyse der Genstrukturen ergaben, dass die repetitive Architektur von Spin/Ssty-Proteinen durch zwei aufeinanderfolgende Duplikationen von Exons in einem gemeinsamen Vorfahren der heutigen Vertebraten entstanden sein muss.

(c) Das humane MCSP Protein und das orthologe NG2 Protein der Ratte spielen sowohl in der Wundheilung als auch in der Entwicklung von Tumoren eine Rolle. Die Entdeckung einer neuen repetitiven Proteindomäne, dem CSPG-Repeat, ermöglichte eine Feineinteilung der Domänenstruktur der NG2/MCSP-Proteine. Der zentrale flexible Teil der NG2-Ektodomäne besteht aus 15 Kopien des CSPG-Repeats. Jede bildet höchstwahrscheinlich eine β-Faltblattstruktur aus. Der CSPG-Repeat ist entfernt verwandt mit dem Cadherin-Repeat. Eine Kopie des CSPG-Repeats in einem cyanobakteriellen Protein entstammt wahrscheinlich einem horizontalem Gentransfer von einem marinen Vielzeller zu einem Cyanobakterium.

(d) Mutationen im humanen Gen LGI1 führen zu einer veränderten Expression des C-Terminus des LGI1 Proteins. Ich entdeckte eine repetitive Sequenzeinheit, den

EPTP-Repeat, im C-Terminus von LGI1. Der EPTP-Repeat ist das gemeinsame Sequenzmodul der Proteine LGI1, LGI2, LGI3 und LGI4, des G-Protein gekoppelten Rezeptors VLGR1 und des TNEP1 Proteins. In einem Mausmodell für Epilepsie ist das VLGR1 Gen mutiert. Auch das humane Gen LGI4 liegt einer chromosomalen Region, die mit einem Epilepsie-Syndrom assoziiert wird. Es ist somit wahrscheinlich, dass der EPTP-Repeat eine essentielle Funktion in der Aufrechterhaltung der Gehirnfunktion hat.

(e) Die Suche nach den molekularen Ursachen der Histidin-Phosphorylierung in Eukaryonten war die Triebfeder zur Sequenzanalyse der Proteine HIG und NtrY. Durch verschiedene Methoden der Sequenzanalyse konnte gezeigt werden, dass die sensorische Region von bakteriellen Histidinkinase-Rezeptoren der NtrY-Familie und der membran-proximale Bereich von eukaryotischen HIG-ähnlichen Proteinen eine signifikante Sequenzähnlichkeit besitzen. Da die Sequenzanalyse auch Regionen von Transmembranhelices umfasst, sollte die Homologie der beiden Proteinfamilien zusätzlich auf experimenteller Ebene belegt werden. HIG-Proteine wären die ersten Proteine von Metazoen, die homolog zu Histidinkinase-Rezeptoren von Bakterien sind.

(f) Immunorezeptor Tyrosin-basierte inhibitorische Motive (ITIMs) haben eine wichtige Funktion in der Kontrolle der Aktivierung von Immunzellen. Die von mir verwendete Strategie zur Suche nach ITIMs in großen Sequenzdatenbanken nutzt den Sequenzkontext, also Informationen über eventuelle Signalpeptide, Transmembranhelices oder Domänen in einem Protein, um die hohe Rate von falsch-positiven ITIM-Vorhersagen herkömmlicher Verfahren der Proteinmotivsuche zu reduzieren. So konnten 109 humane ITIM-Rezeptoren identifiziert werden. Von diesen wurden 36 bereits in der Literatur als ITIM-Rezeptoren beschrieben. Nur zwei bekannte Typ-I ITIM-Rezeptoren wurden nicht gefunden. Eine Analyse öffentlicher Datenquellen über die Gewebeexpression humaner Gene ergab, dass die Expression von ITIM-Rezeptoren nicht auf Blutzellen beschränkt ist.

(g) In 271 Proteinen, die kürzlich in einer massenspektrometrischen Studie des Nukleolus entdeckt wurden, konnten im Zuge dieser Arbeit 115 bekannte und 91 neue Proteindomänen identifiziert werden. Die phylogenetischen Profile dieser Domänen in Archaebakterien, Eubakterien und Eukaryonten deuten auf einen archaebakteriellen Ursprung derjenigen Proteine des Nukleolus hin, die urtümliche Funktionen in der Ribosomenreifung besitzen, zeigen aber deutlich seinen insgesamt chimären Ursprung aus eubakteriellen, archaebakteriellen und eukaryotischen Proteindomänen. Die Ergebnisse sprechen für eine langsame, kontinuierliche Entwicklung des Nukleolus in den ersten Eukaryonten und damit gegen die umstrittene Hypothese eines endosymbiotischen Ursprungs des Nukleus.

11 Anhang

11.1 Manuskript „The death-domain fold of the ASC PYRIN domain, presenting a basis for PYRIN/PYRIN recognition.“

Diese Arbeit entstand unter der Leitung von Prof. Dr. Gottfried Otting am Karolinska Institut in Stockholm und an der Australian National University in Canberra. Mein Anteil an der Arbeit beschränkt sich allein auf das Verfassen der Einleitung, die Erstellung eines Alignments und Beiträgen zur Diskussion. Die Resultate dieser Publikation sind mir also nicht anzurechnen.

Hier aufgeführt ist dieses Manuskript allein deswegen, weil es ein experimenteller Beleg der in Kapitel 2 aufgestellten Hypothese ist, dass die DAPIN-Domäne eine ähnliche Struktur wie die übrigen drei Domänenfamilien der Death-Domain-Superfamilie einnimmt (siehe Kapitel 2 und 9). Dies bestätigt die Validität der von mir eingesetzten Methoden der Sequenzanalyse. Die DAPIN-Domäne wird hier PYRIN-Domäne genannt, da sich dieser Name in der Literatur durchgesetzt hat.

doi:10.1016/j.jmb.2003.07.007 J. Mol. Biol. (2003) 332, 1155–1163

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

The Death-domain Fold of the ASC PYRIN Domain, Presenting a Basis for PYRIN/PYRIN Recognition

Edvards Liepinsh[1], Raitis Barbals[2], Edgar Dahl[3], Anatoly Sharipo[2] Eike Staub[3] and Gottfried Otting[1,4]*

[1]*Department of Medical Biochemistry and Biophysics Karolinska Institute, S-17177 Stockholm, Sweden*

[2]*Biomedical Research and Study Centre, University of Latvia, LV-1067 Riga, Latvia*

[3]*metaGen Pharmaceuticals Oudenarder Str. 16, D-13347 Berlin, Germany*

[4]*Research School of Chemistry Australian National University Canberra, ACT 0200, Australia*

**Corresponding author*

The PYRIN domain is a conserved sequence motif identified in more than 20 human proteins with putative functions in apoptotic and inflammatory signalling pathways. The three-dimensional structure of the PYRIN domain from human ASC was determined by NMR spectroscopy. The structure determination reveals close structural similarity to death domains, death effector domains, and caspase activation and recruitment domains, although the structural alignment with these other members of the death-domain superfamily differs from previously predicted amino acid sequence alignments. Two highly positively and negatively charged surfaces in the PYRIN domain of ASC result in a strong electrostatic dipole moment that is predicted to be present also in related PYRIN domains. These results suggest that electrostatic interactions play an important role for the binding between PYRIN domains. Consequently, the previously reported binding between the PYRIN domains of ASC and ASC2/POP1 or between the zebrafish PYRIN domains of zAsc and Caspy is proposed to involve interactions between helices 2 and 3 of one PYRIN domain with helices 1 and 4 of the other PYRIN domain, in analogy to previously reported homophilic interactions between caspase activation and recruitment domains.

Keywords: PYRIN domain; human ASC; NMR spectroscopy; three-dimensional protein structure; death-domain superfamily

Introduction

The death domain fold is the unifying structural motif of a superfamily of protein domains comprising the death domain (DD) itself,[1] the death effector domain (DED)[2] and the caspase recruitment domain (CARD).[3] Their names express the prominent roles of these domains in programmed cell death. Domains from all three subfamilies occur as modules in diverse human apoptosis proteins in a variety of domain contexts. They all form α-helical bundles acting as adapters in signalling pathways and recruiting other proteins into signalling complexes.[4] Domains from the different death domain subfamilies tend to interact with each other, suggesting that their common fold was frequently reused as a module during the evolution of apoptotic adapter proteins, providing the structural backbone of the signalling pathways that control programmed cell death. Commensurate with their biological importance and despite often poor solubility due to self-association, several

Abbreviations used: AIM2, absent in melanoma 2; ANGIN, angiogenin inhibitor-like protein; ASC, apoptosis-associated speck-like protein containing a caspase recruitment domain; CARD, caspase activation and recruitment domain; CIAS, cold autoinflammatory syndrome; DD, death domain; DED, death effector domain; DEFCAP, death effector filament-forming Ced-4-like apoptosis protein; DQF-COSY, double-quantum filtered two-dimensional correlation spectroscopy; FADD, Fas-associated death domain protein; FMF, familial Mediterranean fever; IFI16, interferon gamma-inducible protein 16; LPS, lipopolysaccharide; MATER, maternal-antigen-that-embryos-require; MCMI1, mast cell maturation inducible protein-like protein; MNDA, myeloid cell nuclear differentiation antigen; NALP1, NACHT-, LRR-, and PYD-containing protein 1; NOE, nuclear Overhauser effect; NOESY, two-dimensional NOE spectroscopy; POP1, pyrin-only protein 1; PYPAF1, PYRIN-containing APAF1-like protein 1; TMS1, target of methylation-induced silencing 1; TNFR, tumour necrosis factor receptor; TOCSY, total correlation spectroscopy.

E-mail address of the corresponding author: gottfried.otting@anu.edu.au

0022-2836/$ - see front matter

structures have been determined for DDs,[1,5–9] DEDs[2,10] and CARDs.[3,11–14]

The PYRIN domain, also called DAPIN, PAAD or PYD, is a recently identified domain that has been suggested to present a new member of the DD superfamily.[15–20] No experimentally determined structure of a PYRIN domain has been reported to date. An attempt to solve the structure of the PYRIN domain of CARD7 failed due to limited solubility.[17]

PYRIN domains are located at the N terminus of proteins that are linked intimately to a variety of human diseases, ranging from cancer to inflammatory syndromes.[15,16,18,21,22] The PYRIN domain was originally found in pyrin, the product of the familial Mediterranean fever (FMF)-associated gene, which is involved in a hereditary hyper-inflammatory response syndrome,[23] and in ASC/TMS1/PYCARD, which functions as a positive mediator of apoptosis.[19,24] Inflammation and apoptosis upregulate ASC in neutrophils and, depending on the cellular context, it can either inhibit or activate NF-κB.[25,26] ASC contains both a PYRIN and a CARD domain. Homophilic and heterophilic interactions of both domains have been reported to be involved in self-association and filament-like aggregation of ASC *in vivo*.[27] ASC and PYRIN seem to interact *via* their PYRIN domains,[19] while the CARD domain of ASC was shown to bind to the CARD domain of caspase-1.[24,28,29] The PYRIN domain of ASC was further shown to bind to POP1/ASC2, a small protein consisting of a single PYRIN domain with a high level of amino acid sequence similarity to the PYRIN domain of ASC.[30] The interaction between ASC and POP1/ASC2 results in a modulation of NF-κB and pro-caspase-1 regulation.[30] Finally, there is evidence that ASC and caspase-1, together with NALP1 (another PYRIN-domain protein) and caspase-5, form a pro-apoptotic complex, named inflammasome, which is essential for innate immunity involving LPS-induced apoptosis.[31]

Two further human hereditary diseases were recently attributed to the PYRIN-domain protein NALP3/CIAS1/PYPAF1, Muckle–Wells syndrome and familial cold autoinflammatory syndrome.[22,32,33] CIAS1 assembles with ASC and regulates the activation of NF-κB.[34] A homologous protein with identical domain architecture, PYPAF7, also binds ASC, activates caspase-1 and regulates NF-κB dependent transcription.[35]

Although PYRIN domains occur in more than 20 human proteins, only a few additional PYRIN-domain proteins have been characterized functionally. Almost all of them appear to be involved in apoptosis and inflammation.[36–40] To the best of our knowledge, no mutation analysis is available for any PYRIN domain. Considering that PYRIN-domain proteins interact frequently with other PYRIN-domain proteins, PYRIN/PYRIN interactions are likely an important feature of PYRIN-domain function. Besides the binding between the PYRIN domains of ASC and POP1/ASC2,[28] conclusive data on PYRIN/PYRIN interactions come from the zebrafish orthologue of ASC (zAsc) and the caspase Caspy, where the PYRIN domains in both proteins were shown to be required for mutual binding *in vitro*.[41]

Here, we present the three-dimensional structure of the PYRIN domain from human ASC. The structure establishes the PYRIN domain fold and corrects previous sequence alignments with other members of the DD superfamily. It suggests a PYRIN/PYRIN interaction mode related to that observed between the CARD domains of Apaf-1 and procaspase-9.[12]

Results and Discussion

PYRIN domains belong to the DD superfamily

The structure of the PYRIN domain is composed of six helices that are arranged in the classical DD fold (Figure 1). A search of the Protein Data Bank (PDB) with the program DALI[42] yielded the death effector domain (DED) from human FADD[2] as the structurally most closely related protein, followed by the procaspase-9 prodomain[12] which belongs to the CARD group of proteins, and the tumor necrosis factor receptor-1 (TNFR) death domain.[9] All three proteins could be aligned to the PYRIN domain with backbone rmsd values of less than 2.5 Å for at least 75 aligned residues. Figure 1 shows that, although the proteins share the six-helix fold of the DD superfamily, the interhelical angles, lengths of helices and lengths of loop segments between helices are variable (Figure 2). It is remarkable that the three proteins that are structurally most similar to the ASC PYRIN domain are representatives of the DED, CARD and DD proteins, i.e. from the three other subfamilies of the DD superfamily. This illustrates the significant structural diversity within the DED, CARD, and DD subfamilies that can exceed the diversity between the subfamilies. Apart from differences in function, the boundaries between the subfamilies of the DD superfamily are thus primarily defined by sequence similarities. The sequence similarities of PYRIN domains with DED, CARD, and DD domains are rather limited. We note that all previously reported sequence alignments of PYRIN domains with members of the DD superfamily differ, at least in some part, from the structure-based sequence alignment of Figure 2, causing inaccuracies in subsequent model building.[16,17,19]

Sequence comparison between PYRIN and DED domains

We identified 24 PYRIN domains in the human genome and a smaller number of murine PYRIN domains (because the annotation of the mouse genome is less complete) (Figure 3). The sequence comparison between PYRIN domains and the

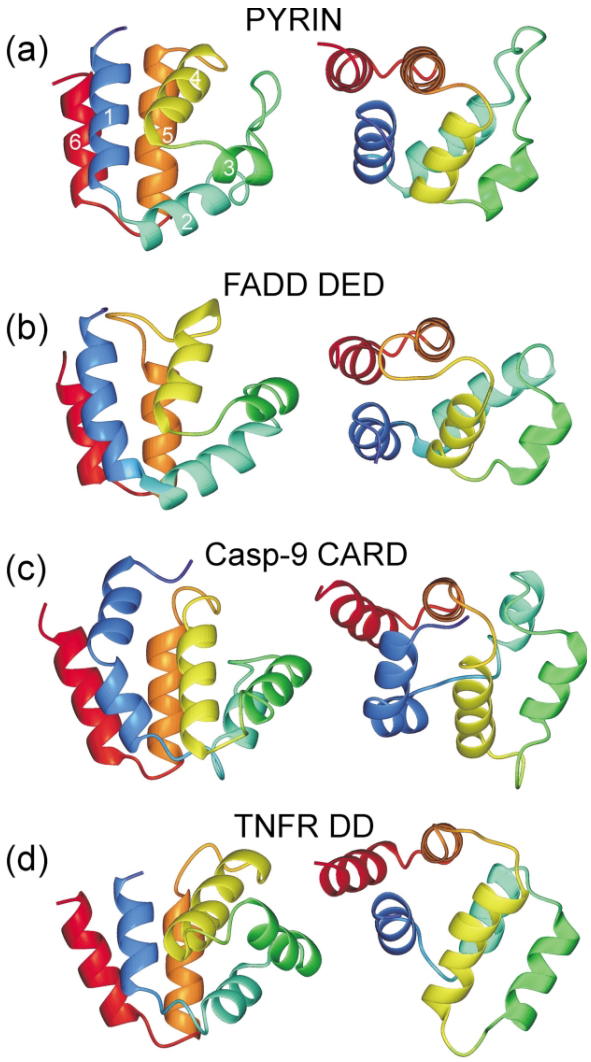

Figure 1. Structure comparison of PYRIN, DED, CARD and DD domains. All four proteins are of human origin. (a) PYRIN domain from ASC. The six helices are numbered. (b) Death effector domain from FADD.[2] (c) Procaspase-9 prodomain (Casp-9 CARD).[12] (d) Tumor necrosis factor receptor−1 death domain.[9] The views in the left and right panel differ by a 90° rotation of the molecules about a horizontal axis. The structures were oriented for best match of helix 5.

structurally most closely related domain of the DD superfamily, the FADD DED, shows significant conservation of buried residues at positions with little solvent accessibility (Figure 3), as expected for a conserved domain fold. Using the DED consensus sequence as reported by the Pfam data base[43] and the alignment of 38 DEDs reported by Kaufmann *et al.*,[44] four positions with characteristic differences from DED sequences could be identified in the PYRIN domains (arrows in Figure 3). Thus, DEDs have a charged or polar residue at the position of Leu12, single-residue insertions between helices 3 and 4, and between helices 4 and 5, and a polar or charged residue at the position of Gly65. Leu12 and Gly65 are buried in the ASC PYRIN domain, while the corresponding residues in the known DED structures (FADD DED and PEA-15 DED)[2,10] are significantly more solvent-exposed. In contrast, the insertion between helices 2 and 3, which is a prominent feature of the PYRIN domain in the structural comparison of Figure 1, is not present in all PYRIN domains (Figure 3).

Binding between PYRIN domains

The PYRIN domain from human ASC is a highly bipolar molecule, with most of the positively charged side-chains located in helices 2 and 3 and the connecting loop, while most of the negatively charged side-chains reside in helices 1 and 4, and immediately adjacent regions (Figure 4(b)). The large electrostatic dipole moment observed in the ASC PYRIN domain suggests that charge–charge interactions may play an important role in the association between PYRIN domains, in analogy to CARDs[3,12–14] and death domains.[4,6,9,45] In the crystal structure of the complex between the CARD domains of the procaspase-9 prodomain and Apaf-1, helices 2 and 3 of the procaspase-9 prodomain pack against helices 1 and 4 of Apaf-1 CARD in a complex that is determined largely by charged residues in the interaction surface.[12] A similar interface, termed a type I interaction, has been postulated for a hexameric complex between Fas and FADD death domains.[45] As helices 2 and 3 have been shown to be crucial for self-association

Figure 2. Structure-based amino acid sequence alignment of representatives of the four different subfamilies of the death-domain superfamily. The folds of the domains are shown in Figure 1. Boxes delineate the helix boundaries. Yellow bands identify structurally conserved residues contributing to the core of the domains. The alignment of the FADD DED and CASP-9 CARD domains and the structurally conserved residues are taken from Weber & Vincenz.[4]

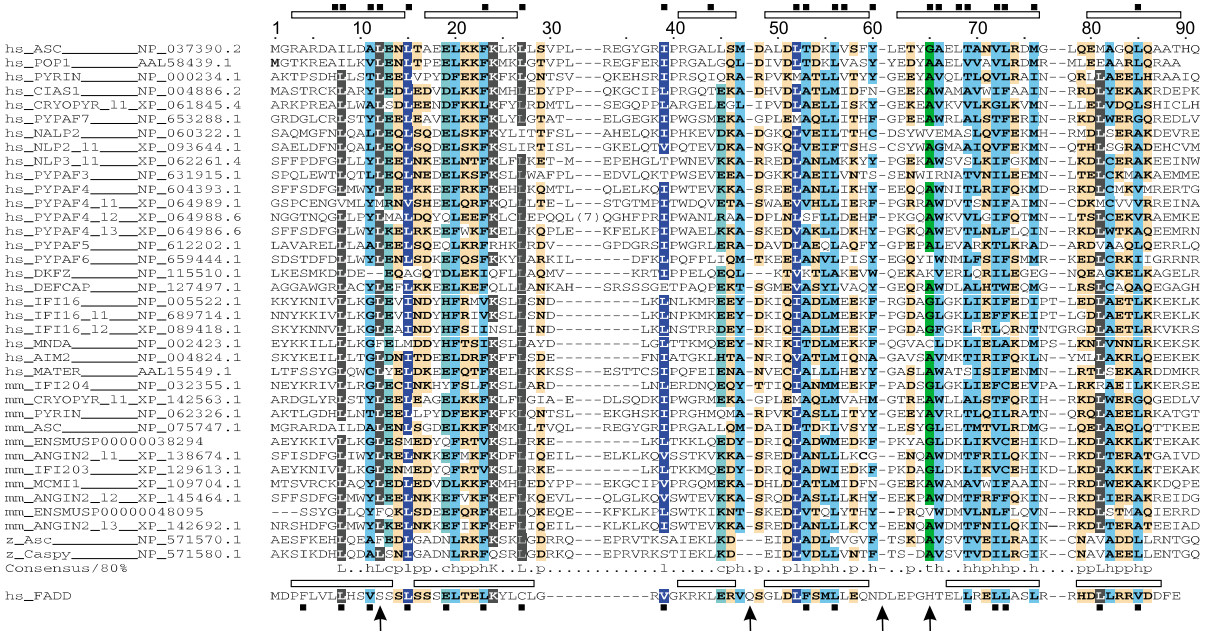

Figure 3. Amino acid sequence alignment of PYRIN domains, PYRIN domain structure and distribution of charged residues. A, Amino acid sequence alignment of 35 PYRIN domains from human and mouse and two from zebrafish. The PYRIN domain of ASC is shown at the top, together with its sequence numbering. The amino acid sequence of the FADD DED domain is shown at the bottom for comparison. The location of the helices and buried side-chains with less than 5% solvent exposure is indicated for the ASC PYRIN domain and the FADD DED domain by bars and filled squares, respectively. Arrows identify positions of significant differences between the PYRIN domains and the DED consensus. The consensus sequence of PYRIN domains is indicated below the PYRIN sequences, where upper case letters indicate conservation of distinct amino acids and lower case letters indicate conservation of (h)ydrophobic, (c)harged, (p)olar, a(l)iphatic, and (t)iny amino acid side-chains. Consensus characters were assigned when >80% of the residues of a column belong to the same amino acid class. hs, *Homo sapiens*; mm, *Mus musculus*; PYRIN, pyrin protein; z, zebrafish; ASC, apoptosis-associated speck-like protein containing a CARD; POP1, pyrin-only protein 1; CIAS1, cold autoinflammatory syndrome 1 (also called PYPAF1 or cryopyrin); CRYOPYR_l1, cryopyrin-like protein 1; NALP2, NACHT-, LRR-, and PYD-containing protein 2 (also called PYPAF2); NLP2_l1, NALP2-like protein 1; NLP3_l1, NALP3-like protein 1; PYPAF1-7: PYRIN-containing APAF1-like protein 1-7; PYPAF4_l1-3, PYPAF4 like proteins 1, 2 and 3; DEFCAP, death effector filament-forming Ced-4-like apoptosis protein; IFI16, interferon gamma-inducible protein 16; IFI16_l1-2, IFI16-like proteins 1 and 2; IFI203 and IFI204, interferon-activatable proteins 203 and 204; MNDA, myeloid cell nuclear differentiation antigen; AIM2, absent in melanoma 2; MATER, maternal-antigen-that-embryos-require; ANGIN2_l1-3, angiogenin inhibitor 2-like proteins 1–3; MCMI1, mast cell maturation-inducible protein-like protein; DKFZ, predicted protein from DKFZ institute transcript; ENSMUSP00000038294 and ENSMUSP00000048095: hypothetical proteins predicted by the ENSEMBL genome annotation project. hs_POP1 is also called ASC2.[30,46]

and intermolecular binding of Fas DD,[1] TNFR-1 DD,[9] FADD DED,[2] FADD DD[7] and CARD/CARD interactions,[3,12] type I interactions appear to be common among members of the DD superfamily. A different DD/DD binding mode was observed in the co-crystal structure of Pelle and Tube death domains, involving the loops between helices 1 and 2, and 5 and 6, and the opposite side of the domain.[6] Such a type II interaction[45] would be almost perpendicular to the electrostatic dipole moment of the PYRIN domain of ASC, making it a less likely interaction mode with other PYRIN domains.

Narrow linewidths in the NMR spectra of the PYRIN domain of ASC indicate that the domain is monomeric in solution and not prone to self-association (data not shown). A similar situation seems to prevail for POP1/ASC2 (hs_POP1 in Figure 3) for which NMR assignments were reported earlier.[46]

Specific binding was reported between the PYRIN domains of ASC and POP1/ASC2.[30] The two PYRIN domains share 63% sequence identity (Figure 3). Most importantly, the charged residues are conserved or conservatively substituted, except for six positions (10, 37, 63, 81, 84, and 87) and a shift of a positively charged residue in ASC (Arg3) by one position in POP1/ASC2 (Lys4). Consequently, the overall charge distribution is very similar in both proteins. A type I interaction between both PYRIN domains, where helices 1 and 4 of one PYRIN domain pack against helices 2 and 3 of the other PYRIN domain, would agree with the negative charges on helices 1 and 4 and the positive charges on helices 2 and 3. Except for residue 10, the charge conservation between these

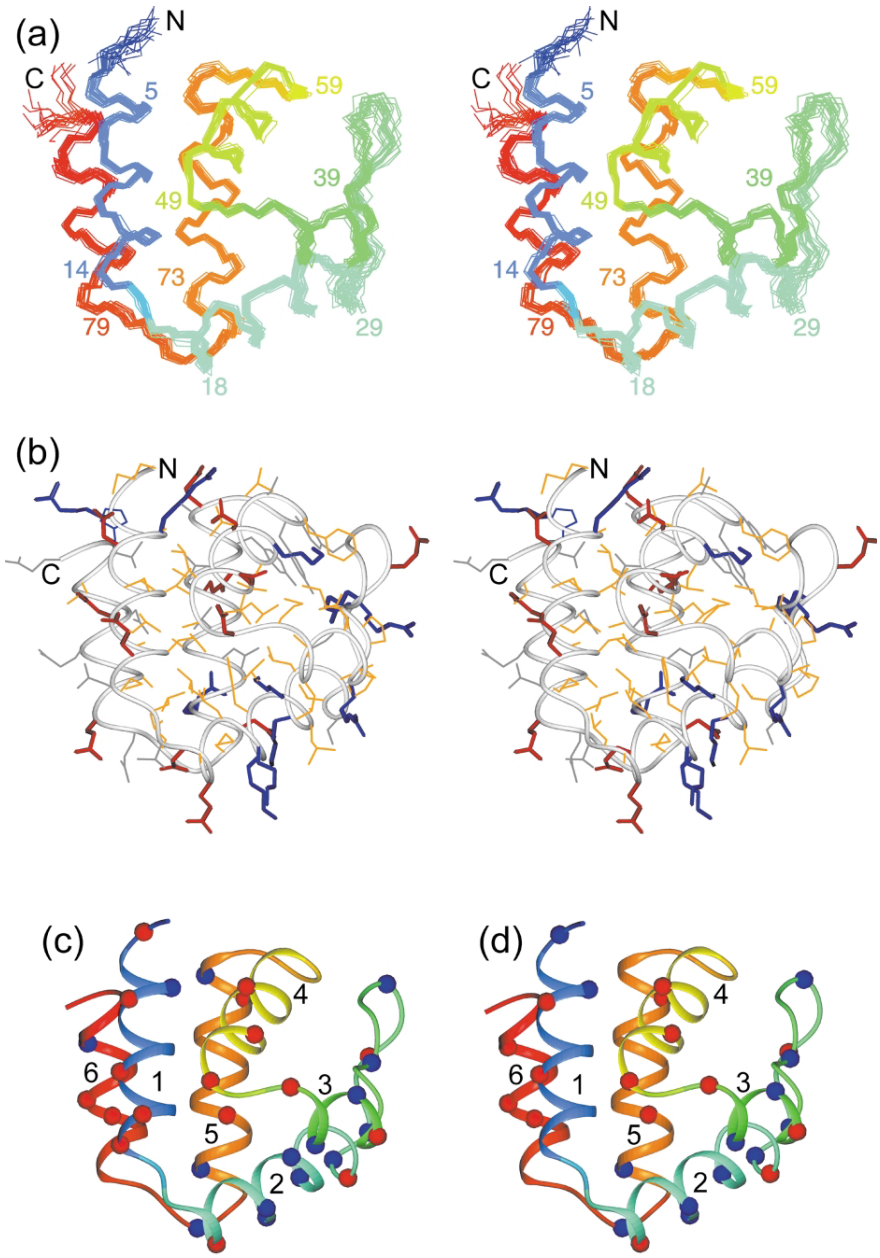

Figure 4. PYRIN domain structure and distribution of charged residues. (a) Stereo view of a superposition of the backbone atoms in the 20 conformers representing the NMR structure of the ASC PYRIN domain (Table 1). Numbers identify sequence positions. (b) Stereo view of the conformer closest to the mean structure of the 20 conformers shown in (a). The following colors were used for the side-chains: blue, Arg, Lys, His; red, Glu, Asp; yellow, Ala, Cys, Ile, Leu, Met, Phe, Pro, Trp, Val; grey, Asn, Gln, Ser, Thr, Tyr. Bold lines identify charged side-chains of Arg, Lys, Asp and Glu. The molecule is oriented so that most of the negatively and positively charged side chains are located, respectively, in the left and right half of the molecule. (c) and (d) Ribbon drawing of the PYRIN domain. Spheres identify the positions of C^{α} atoms, where positively (blue) and negatively (red) charged side-chains are located in the zebrafisch PYRIN domains of zAsc (c) and Caspy (d). The six helices are numbered.

helices is complete between both proteins (Figure 3).

Similarly, specific interactions have been reported between the PYRIN domain of the zebrafish orthologue of ASC, zASC, and the PYRIN domain of the zebrafish caspase Caspy.[41] These two PYRIN domains share 72% sequence identity, including a very similar charge distribution. Mapping the locations of the charged residues of these PYRIN domains on the NMR structure of the PYRIN domain from human ASC (Figure 4(c) and (d)) reveals a charge distribution similar to that of human ASC (Figure 4(b)), suggesting that type I interactions between PYRIN domains may be conserved. More detailed models of the respective complexes are difficult to establish, however,

because the CARDs of Casp-9 and Apaf-1 (Figure 1(c)) are the only proteins for which the structure of a type I interaction complex has been solved,[12] and the structural homology between these proteins and the ASC PYRIN domain is limited (Figure 1).

Concluding remarks

The present structure determination establishes PYRIN domains unambiguously as a fourth branch in the superfamily of DD-type proteins. None of these domains has been found in yeast or bacteria. CARD and DD domains have been identified in nematodes and higher organisms, while DED and PYRIN domains seem to be limited to vertebrates.[20] Despite their recent evolutionary origin, however, PYRIN domains are widespread and diverse. The structural similarity of their fold to that of DED, CARD and DD domains suggests that PYRIN domains may interact with PYRIN domains as well as with other members of the DD superfamily. The NMR structure of the PYRIN domain of human ASC presents a basis for future interaction studies.

Materials and Methods

Cloning of the ASC PYRIN domain

The ASC PYRIN-domain encoding 272 bp DNA fragment was PCR amplified from Marathon-Ready cDNAs prepared from human lymphocytes (Clontech) using *Bam* HI (ATTCGGATCCATGGGGCGCGCGCGACG CCA) and *Hin*dIII (GAATAAGCTTCTACTGGTGCG TGGCCGCCT) oligonucleotide primers designed according to the sequence of the human ASC gene (NCBI protein number BAA87339). PCR cycle conditions were: ten seconds denaturation at 95 °C, 30 seconds annealing at 58 °C, and 1.5 minutes of polymerization at 72 °C. The first one minute pre-denaturation step at 95 °C was followed by 35 PCR cycles. The reaction mixture contained 10 pmol of each primer, 10 ng of the template, 1 mM dNTP mixture and 2.5 units of *Taq* polymerase (Fermentas) in 50 μl of PCR buffer (Fermentas) with 5 mM $MgCl_2$. The 272 bp *Bam* HI/*Hin*dIII fragment bearing the PYRIN domain was cloned into the corresponding sites of pQE30 (QIAGEN) with N-terminal His_6-tag. Recombinant plasmid was transformed into *Escherichia coli* DH5a competent cells (Invitrogen) and the resulting recombinant gene sequenced in both strands by the dideoxy method.

Expression of the ASC PYRIN domain

Recombinant protein was expressed in *E. coli* M15 (QIAGEN), harvested five hours after IPTG-induction and purified by affinity chromatography on Ni-NTA Sepharose (QIAGEN) under denaturing condition according to the manufacturer's basic protocol. The protein was renatured by diluting rapidly into a buffer containing standard phosphate-buffered saline (PBS) at pH 3.7, 1% (v/v) Triton X-100, 1 mM DTT, 10% (v/v) glycerol and 50 mM glycine. The final yield of soluble protein was about 10 mg per liter of bacterial culture.

NMR measurements

NMR spectra were recorded at pH 3.7, 28 °C, using ca 1 mM solutions of the ASC PYRIN domain construct including the His_6-tag. Samples were prepared in 90% H_2O/10% ^2H_2O or 100% ^2H_2O and measured at a ^1H NMR frequency of 800 MHz on a Varian Unity INOVA 800 NMR spectrometer. Sequence-specific resonance assignments were obtained from double-quantum filtered two-dimensional correlation spectroscopy (DQF-COSY), clean- total correlation spectroscopy (TOCSY) (70 ms mixing time), two-dimensional NOE spectroscopy (NOESY) (40, 70 and 150 ms mixing time) and ω_1-decoupled NOESY (150 ms mixing time)[47] spectra, recorded with unlabelled protein.

NMR spectral evaluation

The cross-peaks in the NOESY spectra were assigned and integrated using the program XEASY.[48] Most of the NOE restraints were collected from the NOESY spectrum recorded with 40 ms mixing time, $t_{1max} = 100$ ms, $t_{2max} = 225$ ms, and three days total recording time. $^3J(H^N,H^\alpha)$ couplings were measured using the program INFIT to fit the lineshapes observed in the NOESY spectrum.[49] $^3J(H^\alpha,H^\beta)$ couplings were estimated as 11.0(±3.00) Hz and 4.0(±3.0) Hz, respectively, when COSY, TOCSY and NOESY cross-peaks indicated the presence of large and small couplings, respectively, together with staggered conformations around the C^α–C^β bond.

Structure calculations and evaluation

The NMR structure was calculated using the program DYANA[50] starting from 50 random conformers. As no long-range NOE could be observed for the N-terminal His_6-tag residues, only the residues of the PYRIN domain were included in the structure calculations. The 20 conformers with the lowest residual restraint violations were energy minimized in water using the program OPAL[51] with standard parameters. The Ramachandran plot was analyzed using PROCHECK-NMR.[52] Table 1 shows an overview of the restraints used and structural statistics. Secondary structure elements and rmsd values

Table 1. Structural statistics for the NMR conformers of the ASC PYRIN domain

Number of assigned NOE cross-peaks	1744
Number of non-redundant NOE upper-distance limits	1118
Number of scalar coupling constants[a]	206
Number of dihedral-angle restraints	233
Intra-protein AMBER energy (kcal/mol)	−3788 ± 53
Maximum NOE-restraint violations (Å)	0.1 ± 0.0
Maximum dihedral-angle restraint violations (deg.)	2.1 ± 0.23
rmsd to the mean for N, C^α and C′ (Å)[b]	0.38 ± 0.06
rmsd to the mean for all heavy atoms (Å)[b]	0.81 ± 0.06
Ramachandran plot appearance[c]	
Most favoured regions (%)	92.6
Additionally allowed regions (%)	7.4
Generously allowed regions (%)	0.0
Disallowed regions (%)	0.0

[a] 78 $^3J(H^N,H^\alpha)$, 138 $^3J(H^\alpha,H^\beta)$.
[b] For residues 3–90.
[c] From PROCHECK-NMR.[52]

were calculated using the program MOLMOL,[53] which was used also to create Figures of the structures. Side-chain solvent accessibilities were measured with a spherical probe of 1.4 Å radius and calculated as the percentage of the accessibilities measured for a fully extended side-chain of residue X in a helical Gly-X-Gly peptide.[54] The values obtained were averaged over the 20 NMR conformers.

Data Bank accession codes

The atomic coordinates have been deposited in the Protein Data Bank with accession code 1UCP. The NMR chemical shifts have been deposited at the BioMagResBank (BMRB) under accession code BMRB 5780.

Acknowledgements

This work was supported by the Swedish and Australian Research Councils, and by the Karolinska Institute. G.O. thanks the Australian Research Council for a Federation Fellowship.

References

1. Huang, B. H., Eberstadt, M., Olejniczak, E. T., Meadows, R. P. & Fesik, S. W. (1996). NMR structure and mutagenesis of the Fas (Apo-1/CD95) death domain. *Nature*, **384**, 638–641.
2. Eberstadt, M., Huang, B., Chen, Z., Meadows, R. P., Ng, S., Zheng, L. *et al.* (1998). NMR structure and mutagenesis of the FADD (Mort1) death-effector domain. Solution structure of the RAIDD CARD and model for CARD/CARD interaction in caspase-2 and caspase-9 recruitment. *Nature*, **392**, 941–945.
3. Chou, J. J., Matsuo, H., Duan, H. & Wagner, G. (1998). Solution structure of the RAIDD CARD and model for CARD/CARD interaction in caspase-2 and caspase-9 recruitment. *Cell*, **94**, 171–180.
4. Weber, C. H. & Vincenz, C. (2001). The death domain superfamily: a tale of two interfaces? *Trends Biochem. Sci.* **26**, 475–481.
5. Liepinsh, E., Ilag, L. L., Otting, G. & Ibáñez, C. F. (1997). NMR structure of the death domain of the p75 neurotrophin receptor. *EMBO J.* **16**, 4999–5005.
6. Xiao, T., Towb, P., Wasserman, S. A. & Sprang, S. R. (1999). Three-dimensional structure of a complex between the death domains of Pelle and Tube. *Cell*, **99**, 545–555.
7. Jeong, E. J., Bang, S., Lee, T. H., Park, Y. I., Sim, W. S. & Kim, K. S. (1999). The solution structure of FADD death domain—structural basis of death domain interactions of Fas and FADD. *J. Biol. Chem.* **274**, 16337–16342.
8. Berglund, H., Olerenshaw, D., Sankar, A., Federwisch, M., McDonald, N. Q. & Driscoll, P. C. (2000). The three-dimensional solution structure and dynamic properties of the human FADD death domain. *J. Mol. Biol.* **302**, 171–188.
9. Sukits, S. F., Lin, L., Hsu, S., Malakian, K., Powers, R. & Xu, G. (2001). Solution structure of the tumor necrosis factor receptor-1 death domain. *J. Mol. Biol.* **310**, 895–906.
10. Hill, J. M., Vaidyanathan, H., Ramos, J. W., Ginsberg, M. H. & Werner, M. H. (2002). Recognition of ERK MAP kinase by PEA-15 reveals a common docking site within the death domain and death effector domain. *EMBO J.* **21**, 6494–6504.
11. Day, C. L., Dupont, C., Lackmann, M., Vaux, D. L. & Hinds, M. G. (1999). Solution structure and mutagenesis of the caspase recruitment domain (CARD) from Apaf-1. *Cell Death Differ.* **6**, 1125–1132.
12. Qin, H., Srinivasula, S. M., Wu, G., Fernandes-Alnemri, T., Alnemri, E. S. & Shi, Y. (1999). Structural basis of procaspase-9 recruitment by the apoptotic protease-activating factor 1. *Nature*, **399**, 549–557.
13. Zhou, P., Chou, J., Olea, R. S., Yuan, J. & Wagner, G. (1999). Solution structure of Apaf-1 CARD and its interactino with caspase-9 CARD; a structural basis for specific adaptor/caspase interaction. *Proc. Natl Acad. Sci. USA*, **96**, 11265–11270.
14. Humke, E. W., Shriver, S. K., Starovasnik, M. A., Fairbrother, W. J. & Dixit, V. M. (2000). ICEBERG: a novel inhibitor of interleukin-1β generation. *Cell*, **103**, 99–111.
15. Staub, E., Dahl, E. & Rosenthal, A. (2001). The DAPIN family: a novel domain links apoptotic and interferon response proteins. *Trends Biochem. Sci.* **26**, 83–85.
16. Martinon, F., Hofmann, K. & Tschopp, J. (2001). The PYRIN domain: a possible member of the death domain-fold family implicated in apoptosis and inflammation. *Curr. Biol.* **11**, R118–R120.
17. Fairbrother, W. J., Gordon, N. C., Humke, E. W., O'Rourke, K. M., Starovasnik, M. A., Yin, J. P. & Dixit, V. M. (2001). The PYRIN domain: a member of the death domain-fold superfamily. *Protein Sci.* **10**, 1911–1918.
18. Pawlowski, K., Pio, F., Chu, Z., Reed, J. C. & Godzik, A. (2001). PAAD—a new protein domain associated with apoptosis, cancer and autoimmune diseases. *Trends Biochem. Sci.* **26**, 85–87.
19. Richards, N., Schaner, P., Diaz, A., Stuckey, J., Shelden, E., Wadhwa, A. & Gumucio, D. L. (2001). Interaction between PYRIN and the apoptotic speck protein (ASC) modulates ASC-induced apoptosis. *J. Biol. Chem.* **276**, 39320–39329.
20. Aravind, L., Dixit, V. M. & Koonin, E. V. (2001). Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. *Science*, **291**, 1279–1284.
21. Bertin, J. & DiStefano, P. S. (2000). The PYRIN domain: a novel motif found in apoptosis and inflammation proteins. *Cell Death Differ.* **7**, 1273–1274.
22. Gumucio, D. L., Diaz, A., Schaner, P., Richards, N., Babcock, C., Schaller, M. & Cesena, T. (2002). Fire and ICE: the role of PYRIN domain-containing proteins in inflammation and apoptosis. *Clin. Exptl. Rheum.* **20**, S45–S53.
23. International FMF Consortium (1997). Ancient missense mutations in a new member of the *RoRet* gene family are likely to cause familial mediterranean fever. *Cell*, **90**, 797–807.
24. Masumoto, J., Taniguchi, S., Ayukawa, K., Sarvotham, H., Kishino, T., Niikawa, N. *et al.* (1999). ASC, a novel 22-kDa protein, aggregates during apoptosis of human promyelocytic leukemia HL-60 cells. *J. Biol. Chem.* **274**, 33835–33838.

25. Shiohara, M., Taniguchi, S., Masumoto, J., Yasui, K., Koike, K., Komiyama, A. & Sagara, J. (2002). ASC, which is composed of a PYD and a CARD, is up-regulated by inflammation and apoptosis in human neutrophils. *Biochem. Biophys. Res. Commun.* **293**, 1314–1318.
26. Stehlik, C., Fiorentino, L., Dorfleutner, A., Bruey, J. M., Ariza, E. M., Sagara, J. & Reed, J. C. (2002). The PAAD/PYRIN-family protein ASC is a dual regulator of a conserved step in nuclear factor κB activation pathways. *J. Exptl. Med.* **196**, 1605–1615.
27. Masumoto, J., Taniguchi, S. & Sagara, J. (2001). PYRIN N-terminal homology domain- and caspase recruitment domain-dependent oligomerization of ASC. *Biochem. Biophys. Res. Commun.* **280**, 652–655.
28. Masumoto, J., Taniguchi, S., Nakayama, J., Shiohara, M., Hidaka, E., Katsuyama, T. *et al.* (2001). Expression of apoptosis-associated speck-like protein containing a caspase recruitment domain, a PYRIN N-terminal homology domain-containing protein, in normal human tissues. *J. Histochem. Cytochem.* **49**, 1269–1275.
29. Srinivasula, S. M., Poyet, J. L., Razmara, M., Datta, P., Zhang, Z. & Alnemri, E. S. (2002). The PYRIN-CARD protein ASC is an activating adaptor for caspase-1. *J. Biol. Chem.* **277**, 21119–21122.
30. Stehlik, C., Krajewska, M., Welsh, K., Krajewski, S., Godzik, A. & Reed, J. C. (2003). The PAAD/PYRIN-only protein POP1/ASC2 is a modulator of ASC-mediated NF-κB and pro-Caspase-1 regulation. *Biochem. J.* **373**, 101–113.
31. Martinon, F., Burns, K. & Tschopp, J. (2002). The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proIL-β. *Mol. Cell*, **10**, 417–426.
32. Hoffman, H. M., Mueller, J. L., Broide, D. H., Wanderer, A. A. & Kolodner, R. D. (2001). Mutation of a new gene encoding a putative PYRIN-like protein causes familial cold autoinflammatory syndrome and Muckle-Wells syndrome. *Nature Genet.* **29**, 301–305.
33. Aganna, E., Martinon, F., Hawkins, P. N., Ross, J. B., Swan, D. C., Booth, D. R. *et al.* (2002). Association of mutations in the NALP3/CIAS1/PYPAF1 gene with a broad phenotype including recurrent fever, cold sensitivity, sensorineural deafness, and AA amyloidosis. *Arthritis Rheum.* **46**, 2445–2452.
34. Manji, G. A., Wang, L., Geddes, B. J., Brown, M., Merriam, S., Al-Garawi, A. *et al.* (2002). PYPAF1, a PYRIN-containing Apaf1-like protein that assembles with ASC and regulates activation of NF-κB. *J. Biol. Chem.* **277**, 11570–11575.
35. Wang, L., Manji, G. A., Grenier, J. M., Al-Garawi, A., Merriam, S., Lora, J. M. *et al.* (2002). PYPAF7, a novel PYRIN-containing Apaf1-like protein that regulates activation of NF-κB and caspase-1-dependent cytokine processing. *J. Biol. Chem.* **277**, 29874–29880.
36. Fiorentino, L., Stehlik, C., Oliveira, V., Ariza, M. E., Godzik, A. & Reed, J. C. (2002). A novel PAAD-containing protein that modulates NF-κB induction by cytokines tumor necrosis factor-α and interleukin-1β. *J. Biol. Chem.* **277**, 35333–35340.
37. Harton, J. A., Linhoff, M. W., Zhang, J. & Ting, J. P. (2002). Cutting edge: CATERPILLER: a large family of mammalian genes containing CARD, PYRIN, nucleotide-binding, and leucine-rich repeat domains. *J. Immunol.* **169**, 4088–4093.
38. Grenier, J. M., Wang, L., Manji, G. A., Huang, W. J., Al-Garawi, A., Kelly, R. *et al.* (2002). Functional screening of five PYPAF family members identifies PYPAF5 as a novel regulator of NF-κB and Caspase-1. *FEBS Letters*, **530**, 73–78.
39. Xie, J. P., Briggs, J. A. & Briggs, R. C. (1997). MNDA dimerizes through a complex motif involving an N-terminal basic region. *FEBS Letters*, **408**, 151–155.
40. Hlaing, T., Guo, R. F., Dilley, K. A., Loussia, J. M., Morrish, T. A., Shi, M. M. *et al.* (2001). Molecular cloning and characterization of DEFCAP-L and -S, two isoforms of a novel member of the mammalian Ced-4 family of apoptosis proteins. *J. Biol. Chem.* **276**, 9230–9238.
41. Masumoto, J., Taniguchi, S., Ayukawa, K., Sarvotham, H., Kishino, T., Niikawa, N. *et al.* (2003). Caspy, a zebrafish caspase, activated by ASC oligomerization is required for pharyngeal arch development. *J. Biol. Chem.* **278**, 4268–4276.
42. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–136.
43. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R. *et al.* (2002). The Pfam protein families database. *Nucl. Acids Res.* **30**, 276–280.
44. Kaufmann, M., Bozic, D., Briand, C., Bodmer, J., Zerbe, O., Kohl, A. *et al.* (2002). Identification of a basic surface area of the FADD death effector domain critical for apoptotic signaling. *FEBS Letters*, **527**, 250–254.
45. Weber, C. H. & Vincenz, C. (2001). A docking model of key components of the DISC complex: death domain superfamily interactins redefined. *FEBS Letters*, **492**, 171–176.
46. Espejo, F., Green, M., Preece, N. E. & Assa-Munt, N. (2002). NMR assignment of human ASC_2, a self contained protein interaction domain involved in apoptosis and inflammation. *J. Biomol. NMR*, **23**, 151–152.
47. Brüschweiler, R., Griesinger, C., Sørensen, O. W. & Ernst, R. R. (1984). Combined use of hard and soft pulses for ω_1 decoupling in two-diemsnional NMR spectroscopy. *J. Magn. Reson.* **59**, 178–185.
48. Bartels, C., Xia, T., Güntert, P., Billeter, M. & Wüthrich, K. (1995). The program XEASY for computer-supported NMR spectral analysis. *J. Biomol. NMR*, **5**, 1–10.
49. Szyperski, T., Güntert, P., Otting, G. & Wüthrich, K. (1992). Determination of scalar coupling constants by inverse Fourier transformation of in-phase multiplets. *J. Magn. Reson.* **99**, 552–560.
50. Güntert, P., Mumenthaler, C. & Wüthrich, K. (1997). Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283–298.
51. Luginbühl, P., Güntert, P., Billeter, M. & Wüthrich, K. (1996). The new program OPAL for molecular dynamics simulations and energy refinements of biological macromolecules. *J. Biomol. NMR*, **8**, 136–146.
52. Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R. & Thornton, J. M. (1996). AQUA and ROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR*, **8**, 477–486.
53. Koradi, R., Billeter, M. & Wüthrich, K. (1996).

MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.
54. Sevilla-Sierra, P., Otting, G. & Wüthrich, K. (1994). Determination of the nuclear magnetic resonance structure of the DNA-binding domain of the P22 c2 repressor(1-76) in solution and comparison with the DNA-binding domain of the 434 repressor. *J. Mol. Biol.* **235**, 1003–1020.

Edited by M. F. Summers

(Received 6 June 2003; received in revised form 24 July 2003; accepted 28 July 2003)

11.2 *Curriculum vitae*

Persönliche Angaben

	Dipl. Biotechnol. Eike Staub
Wohnort	Binzstr. 12, D-13189 Berlin
Telefon privat	030 - 49910136
Telefon dienstlich	030 - 8413 1157
Fax dienstlich	030 - 8413 1152
Email	eike.staub@molgen.mpg.de

Geburtstag und -ort	12.07.1972 in Osnabrück
Familienstand	verheiratet mit Claudia Staub, geborene Rußwurm

Schulbildung

07/1979 - 06/1983	Grundschule Bissendorf
08/1983 - 03/1984	Orientierungsstufe Bissendorf
04/1984 - 07/1985	Orientierungsstufe Dom, Osnabrück
08/1985 - 06/1992	Ratsgymnasium Osnabrück Abitur, allgemeine Hochschulreife (Note 1,4)

Zivildienst

09/1992 - 09/1993	Beschützende Werkstätten Schledehausen

Studium

10/1993 - 08/1999	Studium der Biotechnologie an der Technischen Universität Braunschweig
09/95	Vordiplom (Note „sehr gut“)
02/96 – 06/96	Studienarbeit an der University of Glasgow im Institute of Biomedical and Life Sciences (IBLS) bei Dr. James E. Milner-White, Thema: „The β-sheet properties of amino acids“, Note „sehr gut”

07/98 – 04/99	Diplomarbeit in der Firma metaGen Gesellschaft für Genomforschung mbH bei Prof. Dr. A. Rosenthal, Thema: „Expressionsanalyse von Prostatakarzinomzellen mittels DNA Microarrays“, Note „sehr gut“
07/99	Diplom der Biotechnologie, Note „sehr gut“

Promotion

09/1999 – 02/2002	Anfertigung der Arbeiten dieser Dissertation als Promotionsstipendiat der Firma metaGen GmbH, Betreuung durch Prof. Dr. A. Rosenthal und Prof. Dr. Dr. T. Braun

Berufserfahrung

03/2002 – 06/2003	Wissenschaftler bei der Firma metaGen im Bereich Bioinformatik/DNA- und Proteinannotation und Leitung des Labors „Expressionsanalyse des Kolonkarzinoms“
seit 09/2003	Wissenschaftler am Max Planck Institut für Molekulare Genetik in Berlin, Abteilung „Computational Molecular Biology“ bei Prof. Dr. M. Vingron, AG Protein Families and Evolution

Halle (Saale), den 15.November 2003

11.3 Publikationsliste

(1) Staub E, Dahl E, Rosenthal A.
The DAPIN family: a novel domain links apoptotic and interferon response proteins.
Trends in Biochemical Sciences (2001) Band 26, S.83-85.

(2) Staub E, Mennerich D, Rosenthal A.
The Spin/Ssty repeat: a new motif identified in proteins involved in vertebrate development from gamete to embryo.
Genome Biology (2002) Band 3, Ausgabe 1 (RESEARCH0003).

(3) Morante-Redolat JM, Gorostidi-Pagola A, Piquer-Sirerol S, Saenz A, Poza JJ, Galan J, Gesk S, Sarafidou T, Mautner VF, Binelli S, Staub E, Hinzmann B, French L, Prud'homme JF, Passarelli D, Scannapieco P, Tassinari CA, Avanzini G, Marti-Masso JF, Kluwe L, Deloukas P, Moschonas NK, Michelucci R, Siebert R, Nobile C, Perez-Tur J, Lopez de Munain A.
Mutations in the LGI1/Epitempin gene on 10q24 cause autosomal dominant lateral temporal epilepsy.
Human Molecular Genetics (2002) Band 11, S.1119-1128.

(4) Martin Reczko, Petko Fiziev, Eike Staub, and Artemis Hatzigeorgiou.
Finding Signal Peptides in Human Protein Sequences Using Recurrent Neural Networks.
Lecture Notes in Computer Science (2002) Band 2452, S.60-67.

(5) Kasper G, Taudien S, Staub E, Mennerich D, Rieder M, Hinzmann B, Dahl E, Schwidetzky U, Rosenthal A, Rump A.
Different structural organization of the encephalopsin gene in man and mouse.
Gene (2002) Band 295, S.27-32.

(6) Staub E, Hinzmann B, Rosenthal A.
A novel repeat in the melanoma-associated chondroitin sulfate proteoglycan defines a new protein family.
FEBS Letters (2002) Band 527, S.114-118.

(7) Staub E, Perez-Tur J, Siebert R, Nobile C, Moschonas NK, Deloukas P, Hinzmann B.
The novel EPTP repeat defines a superfamily of proteins implicated in epileptic disorders.
Trends in Biochemical Sciences (2002) Band 27, S.441-444.

(8) Grutzmann R, Pilarsky C, Staub E, Schmitt AO, Foerder M, Specht T, Hinzmann B, Dahl E, Alldinger I, Rosenthal A, Ockert D, Saeger HD.
Systematic isolation of genes differentially expressed in normal and cancerous tissue of the pancreas.
Pancreatology (2003) Band 3, S.169-178.

(9) Liepinsh E, Barbals R, Dahl E, Sharipo A, Staub E, Otting G.
The Death-domain Fold of the ASC PYRIN Domain, Presenting a Basis for PYRIN/PYRIN Recognition.
Journal of Molecular Biology (2003) Band 332, S.1155-1163.

(10) Gruetzmann R, Foerder M, Alldinger I, Staub E, Brummendorf T, Ropcke S, Li X, Kristiansen G, Jesnowski R, Sipos B, Lohr M, Luttges J, Ockert D, Kloppel G, Saeger HD, Pilarsky C.
Gene expression profiles of microdissected pancreatic ductal adenocarcinoma.
Virchows Archiv (2003). Band 443, S. 508-517.

(11) Himmelfarb M, Klopocki E, Grube S, Staub E, Klamann I, Hinzmann B, Rosenthal A, Duerst M, Dahl E.
ITIH5, a novel member of the inter-alpha trypsin inhibitor heavy chain family is downregulated in breast cancer
Cancer Letters (2003). Im Druck. Online-Publikation seit 19.November 2003.

(12) Staub E, Fiziev P, Rosenthal A, Hinzmann B.
Systematic identification of immunoreceptor tyrosine-based inhibitory motifs (ITIMs) in the human proteome.
Cellular Signalling (2003). Im Druck. Online-Publikation seit 30.Oktober 2003.

(13) Staub E, Fiziev P, Rosenthal A, Hinzmann B.
Insights into the evolution of the nucleolus by an analysis of its protein domains.
BioEssays. Zur Publikation angenommen.

(14) Staub E, Braun T.
Significant sequence similarity between prokaryotic nitrogen-sensing histidine kinases and vertebrate hypoxia-inducible proteins.
Cellular Signalling. Zur Publikation eingereicht.

11.4 Erklärung gemäß § 5 (2) b) der Promotionsordnung

Hiermit versichere ich, dass die vorliegende Dissertation selbständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet wurden. Die Passagen der Arbeit, die anderen Schriften entnommen sind, wurden unter Angabe der Quellen kenntlich gemacht.

Mehrere Autoren treten als Co-Autoren der Einzelmanuskripte auf, aus denen der Kern dieser Arbeit besteht. Die Beiträge der Co-Autoren betreffen in keinem Fall die Planung der Projekte, die Durchführung der praktischen Arbeiten oder das Verfassen von Texten. Die Fremdbeiträge beschränken sich auf Denkanstößen zu Projekten, auf die kollegiale Diskussion über die Bedeutung von Ergebnissen und auf Anstöße zur Verbesserung des sprachlichen Stils von Manuskripten.

Um die Beiträge jedes Co-Autors transparent zu machen, wurde eine Sammlung von Erklärungen verfasst, in denen die Beiträge jedes einzelnen Co-Autors zu jedem einzelnen Manuskript kenntlich gemacht worden sind. Die dort gemachten Angaben sind von allen Co-Autoren für korrekt befunden worden. Für jedes einzelne Manuskript bezeugte dies jeder der Co-Autoren mit seiner Unterschrift. Die Sammlung der originalen unterschriebenen Erklärungen ist dem Promotionsantrag beigelegt.

Ich erkläre zudem, diese wissenschaftliche Arbeit an keiner anderen wissenschaftlichen Einrichtung zur Erlangung eines akademischen Grades eingereicht zu haben.

Halle (Saale), den 15.November 2003