

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 Proteindomänenrepertoires 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 beinhaltet bisher drei Domänenfamilien: die „Death Domain“, die „Death Effector Domain“ und die „Caspase Recruitment Domain“ (3). Alle drei Subfamilien haben ein charakteristisches dreidimensionales 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- κ B Signaltransduktionsweg, 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 paraloge 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 Genstruktur 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 Paraloge keine Introns besitzen. Ein möglicher Mechanismus ist die Retrotransposition eines bereits gespleißten *SPIN*-Transkripts durch reverse Transkription 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 Archaeobakterien 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 ⁽¹⁴⁾. 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 Leucinreichen 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 Aminosäurezusammensetzung 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 Aminosäurezusammensetzung unserer NtrY/HIG-Alignments fest. Dies bedeutet, dass die paarweise Ähnlichkeit der Sequenzen nicht vorrangig durch eine simple Ähnlichkeit der Aminosäurezusammensetzung 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 Proteindomänenreservoirs 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 unbekannt 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 Archaeobakterien, Eubakterien und Eukaryonten ermittelt. Seit langem ist bekannt, dass die Translationsmaschinerie von Eukaryonten enger mit derjenigen von Archaeobakterien 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 Archaeobakterien zu suchen ist.

Insgesamt sind 59 Proteindomänen des Nukleolus sowohl in Archaeobakterien 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 Archaeobakterien, 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 Archaeobakteriums mit dem eines Eubakteriums in einer Zelle vereint wurde. Unter den archaeobakteriellen 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 Archaeobakterien stammt und bestätigt die bereits in anderen Arbeiten mehrfach postulierte Verwandtschaft des sogenannten „informationsverarbeitenden“ Apparats von Archaeobakterien 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 Archaeobakterium 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 Eukaryontenspezifischen 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 Proteinsequenzen 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 Bioinformatik, 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 Sequenzdatenbanken 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 Domänenbanken 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 Domänenbanken 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 Domänenbank 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 Proteinfamiliendatenbanken 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 Organismenspezifischer 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 Domänenkollektionen 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 Signaltransduktionswegen 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. POPping 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, Vigeveno 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.