

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 unbekanntem 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 unbekanntes 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 unbekanntes 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 Brusttumorgewebe weit häufiger repräsentiert als in EST-Datenbanken von normalen Brustgewebe. 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 Strukturmodell 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 Histidinkinase Rezeptoren 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 Archaeobakterien, 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. Moulton 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, Kyripides 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. POPping 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.