Aus dem Institut für Physiologie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Repräsentation von Biomolekülen und Bioereignissen durch Profile und Matrizen, deren Vergleichbarkeit durch Metriken und ihre Bedeutung in der Biomedizin

zur Erlangung des akademischen Grades
Doctor rerum medicarum (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Diplom-Chemiker Thomas Meinel

aus Berlin

Gutachter:

1. Priv.-Doz. Dr. rer. medic. Robert Preißner
2. Prof. Dr. rer. nat. Ina Koch
3. Prof. Dr. rer. nat. Dr. h. c. Edda Klipp

Datum der Promotion:

22. März 2013

Inhaltsverzeichnis

DISSERTATION

Repräsentation von Biomolekülen und Bioereignissen durch Profile und Matrizen, deren Vergleichbarkeit durch Metriken und ihre Bedeutung in der Biomedizin

Thomas Meinel

2013

Zusammenfassung

Übersicht

Profile und Matrizen sind in den Biowissenschaften häufig angewendete Konstrukte, um Vergleiche von Hochdurchsatzdaten durchzuführen oder deren Visualisierungen zu ermöglichen. Ein mathematischer Vergleich solcher Profile setzt bestimmte Algorithmen und in bestimmten Fällen Vergleichsmetriken, z.B. Distanzmetriken, voraus. Im Zusammenhang dieser Arbeit versteht sich ein Profil als eine Aneinanderreihung von Eigenschaften einer physikalischen Entität über eine Reihe von Objekten, und die Eigenschaften dieser Objekte charakterisieren dann diese Entität. Wenn Profile mehrerer Entitäten aufeinander gestapelt werden, ergibt sich eine Matrix von Entitäten über Objekte. In einer solchen Matrix gilt die orthogonale Sichtweise genauso: ein Objekt lässt sich über ein Profil von biophysikalischen Entitäten beschreiben.

In den vier mit dieser Arbeit zusammengefassten Publikationen wird eine Reihe von Profilanwendungen präsentiert, die direkt oder indirekt Bedeutung in der Biomedizin erlangt haben: Topologien in Phylogenomischen Bäumen für eine Reihe genereller Bakterienstämme, methodisch neu eingeführt als Tree Topology Profiling; die Charakterisierung von Spezien über die Gesamtheit aller Genfamilien; Chemosensitivitätsprofile sowie Genexpressionsprofile als Charakterisierungen von chemischen Substanzen oder Genen über eine Reihung von Krebszelllinien; Substanz-Zielgen-Matrizen als Charakteristikum des Interaktionspotentials z. B. in Krebs-relevanten Signalwegen. Eine weitere Anwendung besteht in der Darstellung von Genexpressionsprofilen innerhalb einer Genfamilie für zwei biologische Subjekte, die vergleichbaren experimentellen oder medizinischen Behandlungen unterworfen waren.

In der ersten Publikation werden Genomphylogenien, die in der Literatur in sehr heterogener Form als Resultate der jeweiligen Generierung zu finden sind, in Bezug zu ihrem algorithmischen Hintergrund gestellt. Aus den Ergebnissen wird die allgemeine Bedeutung verwendeter Distanzmetriken für Profilanwendungen herausgearbeitet. Die Datenbank CancerResource, http://bioinf-data.charite.de/cancerresource/, für Substanz-Zielgen-Interaktionen untermauert mit experimentellen Daten sowie die Web-basierte Software Ortho2ExpressMatrix für Genexpressionsereignisse innerhalb von Genfamilien, http://bioinf-data.charite.de/o2em/cgi-bin/o2em.pl, werden in drei weiteren Publikationen beschrieben. Sie haben zentrale Bedeutung auf den Gebieten der Medikament-relevanten Krebsforschung und der Vergleichenden Genomik.

1. Einleitung mit Zielstellungen

Im Zeitalter der Genom-, Transkriptom-, Proteom-, oder Metabolom-weiten Erfassung von biophysikalischen oder biochemischen Zuständen in Zellen von Lebewesen sind Profile und Matrizen mathematische oder programmatische Konstrukte, die eine Bewältigung entsprechender Datenmengen oft erst erlauben. Nach Transformation der Daten ermöglicht die Bildung solcher Konstrukte einerseits die weitere Verarbeitung im Computer und andererseits Repräsentationen oder Visualisierungen, die einem Betrachter die Rezeption dieser komplexen Zusammenhänge vereinfachen. Dabei ist mit Transformation gemeint, dass im Vorfeld dem jeweiligen Wissenschaftsfeld angepasste Abstraktionen oder Modellierungen notwendig sind. Die Wissenschaftsfelder, bei denen Profile oder Matrizen zum Zwecke der Darstellung eingesetzt werden, haben stetig zugenommen. Sie sind äußerst vielfältig, und Beispiele von ihnen werden in den Publikationen präsentiert, die mit dieser Arbeit vorgestellt werden. Profile sind zentrales Thema in zwei Veröffentlichungen, Matrizen in einer weiteren.

In den Biowissenschaften werden Profilvergleiche für verschiedenste biophysikalische Größen und biologische Objekte durchgeführt. Damit sind Profilvergleiche nicht zuletzt für die Biomedizin relevant. Die untersuchten Objekte sind Spezien, Gewebetypen, Zelllinien, Individuen in einer Kohorte, aber auch schwer beschreibbare Subjekte wie Topologien in Phylogenomischen Bäumen. Die mathematische Grundlagen zum Vergleich zweier Profile mit binären Charakteren hinsichtlich ihrer Ähnlichkeit wurden mit der Numerischen Taxonomie zu Beginn des 20. Jahrhunderts eingeführt.[1-3]

‚Phylogenetic Profiling' ist im Jahr 1999 in der Bioinformatik etabliert worden.[4, 5] Mit Phylogenetischen Profilen wird das Vorhandensein eines Gens in einer definierten Reihe biologischer Arten, Spezien, aufgezeigt; die so charakterisierten Genfamilien werden über diese Profile vergleichbar gemacht. Mit dem ‚Aufeinanderstapeln' aller dieser Profile wird eine Phylogenetische Ereignismatrix gebildet. - Die zum ‚Phylogenetic Profiling' orthogonale Sichtweise auf die Ereignismatrix charakterisiert die untersuchten Spezien: Anhand der Gegenwart (oder Abwesenheit) einer jeden in der Matrix existierenden Genfamilie kann die Ähnlichkeit über mehrere tausend (Zahlenbeispiel PhyloMatrix:[6] 19374) Familien zwischen zwei vollständig sequenzierten Spezien (PhyloMatrix: 106 Spezien) untereinander berechnet werden. Aufgrund solcher Ereignismatrizen und einer Vielzahl von Abwandlungen davon, aber auch aufgrund

gänzlich anderer Methodiken, die auf Bäumen von Genfamilien oder auf Super-Alignments von Sequenzen basieren, wurden aufgrund rein molekularbiologischer Daten eine Vielzahl von Phylo*genomischen* Bäumen publiziert.[7-9] Solche Genombäume, aber auch Phylo*genetische* Bäume oder der Taxonomische Baum werden erstellt, um dem Baum des Lebens, dem ‚Tree of Life' (ToL), nahe zu kommen, wie ihn Charles Darwin in seinen Aufzeichnungen bereits 1837/1838 skizziert hat.[10] Für publizierte Genom-Bäume fällt auf, dass deren Generierung im Ergebnis sehr unterschiedlich ausfällt, und es bestand Klärungsbedarf, inwieweit sich die Umstände einer Generierung, insbesondere die Art der Berechnung der Ähnlichkeiten zwischen den Spezien, auf Genomphylogenien auswirken. Mit bestimmten Heuristiken und Distanzmaßen wird ein Teil dieser Genombäume als ‚gene content' Phylogenien erzeugt, indem die Profile über sämtliche Genfamilien ausgewertet wurden, die die Spezien charakterisieren. Die Literatur bietet eine Vielzahl von Publikationen zu diesem Thema, und in einer Meta-Analyse sollten die publizierten Bäume vergleichend beschrieben und in Beziehung zu dem Hintergrund ihrer Generierung gestellt werden. Gleichzeitig sollte die Möglichkeit genutzt werden, Molekulardaten von SYSTERS-PhyloMatrix[6] Genfamilien zu verwenden, indem ‚gene content' Genom-Bäume mit verschiedenen Techniken erzeugt und in diese Meta-Analyse einbezogen werden. Neben diesen Gesichtspunkten werden in der *Publikation 1* die bei der Erzeugung dieser Phylogenien verwendeten Distanzmetriken studiert, um Auswirkungen bei einer Anwendung in der Biomedizin abschätzen zu können.

Interaktionen von chemischen Substanzen mit deren Zielgenen werden in vielen Bereichen der Biowissenschaften beschrieben. Oftmals sind jedoch die entsprechenden Web-Ressourcen einseitig auf einen bestimmten Bereich beschränkt. Die projektierte Datenbank *CancerResource* sollte die Resultate eines Datamining für Interaktionen von chemischen Substanzen mit Zielgenen auf Krebserkrankungen fokussieren und mit experimentellen Daten, die dem biologisch-medizinischen Forschungsfeld entspringen, unterfüttert werden. Deren Visualisierung durch Profile und Matrizen war zu prüfen und in das Konzept der zu erstellenden Ressource zu integrieren. Aktivitätsprofile auf Basis der Chemosensitivität sind bereits seit mehreren Jahrzehnten ein anerkanntes methodisches Instrument zur Identifizierung der biologischen Wirksamkeit chemischer Substanzen. Der Grundansatz hierfür ist die Ermittlung der Inhibition des Zellwachstums nach Applikation einer chemischen Substanz auf ein Spektrum von Zelllinien, die in eine definierte Reihung gebracht werden. Ein etabliertes Werk-

zeug hierfür sind die bekannten 60 Zelllinien humaner Herkunft (NCI-60 Zelllinien),[11] die aus Gewebeproben verschiedener Tumore abgeleitet wurden. Diese Profile, die biologisch wirksame Substanzen über deren Chemosensitivität charakterisieren, werden auch als ‚Daumenabdrücke über Zelllinien' (‚cellular fingerprints') bezeichnet. Es war naheliegend, in ihrer Chemosensitivität ähnliche Substanzen auch in der Ähnlichkeit von Struktureigenschaften zu vergleichen.[12] Darüber hinaus war die Aufdeckung der Substrat-Spezifität in Signalweg-Beziehungen Gegenstand intensiver Forschung.[13, 14] Die Quantifizierung von Genen (Expression) ist Grundlage für eine zweite Quelle experimenteller Daten, die in der klinischen Diagnostik am häufigsten mit der Mikroarray-Technologie erhoben werden. Für eine genomweite Repräsentation der Resultate werden üblicherweise zwei-dimensionale Heatmaps durch Clustering erzeugt. Dabei werden in den zwei Dimensionen sowohl die Gene (durch die entsprechenden Sonden) als auch die experimentellen Ansätze als Profil über den jeweilig anderen Parameter beschrieben. Die Ähnlichkeit der experimentellen Ansätze (wie auch der Gene) zueinander wird in Dendrogrammen ausgedrückt, die neben die Heatmap gezeichnet werden. In der projektierten Datenbank sollte ein Benutzer auf die Profile beider Datensammlungen und in beiden Dimensionen in geeigneter Weise zugreifen können. Ein weiterer Gesichtspunkt sollte die Darstellung von Genen in Krebs-relevanten Signalwegen sein, um dem Betrachter den bestehenden funktionalen Kontext integral erfassen lassen zu können. Daher sollte im Sinne des Profil- oder Matrixgedankens ermöglicht werden, neben einer üblichen bildlichen Darstellung in einem Graphen für Signalwege das multiple und redundante Auftreten dieser Interaktionen aus Eventmatrizen heraus zu erfassen (*Publikationen 2* und *3*).

Wie oben bereits beschrieben, visualisieren Phylogenetische Profile das Vorhandensein von Genen und sind dafür konstruiert, einen möglichen funktionalen Kontext sichtbar zu machen. Offen blieb an dieser Stelle die Frage, in welcher Quantität diese Gene in den Spezien vorhanden (exprimiert) sind. Weiterhin würde es ein Phylogenetisches Profil präzisieren, wenn man funktionelle Orthologie aufzeigen könnte. Funktionelle Orthologie führt einzelne Proteine z. B. zweier Spezien zusammen, die gleichzeitig eine identische, also beibehaltene, Funktion bei gemeinsamem evolutionären Ursprung besitzen.[15] Eine Berücksichtigung dieser zwei Aspekte entspräche also einer Aufschlüsselung oder Präzisierung eines „Charakters" in einem Phylogenetischen Profil. Die Frage, ob mehrere funktionell zusammenhängende Gene sich über gleichsinnige Expression, also Koexpression, identifizieren lassen, wurde in

verschiedenen Ansätzen beantwortet,[16, 17] ohne jedoch explizit die Frage nach den Paralogen zu berücksichtigen. Ein in diese Richtung weisender systematischer Forschungsansatz würde etwa mit der Frage berührt, welche Paraloge in zwei verschieden Spezien eine identische funktionelle Aufgabe besitzen. In der Literatur bot die Frage nach der Krankheits-Relevanz[18] bzw. die Funktionalität von Paralogen in verschiedenen Spezien und hier besonders in vergleichbaren Geweben Anlass zu verschiedenen Untersuchungen, und es wurde gefunden, dass die Expressionsprofile von ‚one-to-one' Orthologen höher konserviert sind als von ‚many-to-many' Orthologen.[19] Zwischen vergleichbaren Geweben von Spezien scheint es dennoch eine hohe Konservierung der Genexpression zu geben.[20] Zusammengefasst besteht für viele Bereiche in den Biowissenschaften in einer zentralen Frage, nämlich der nach der Koexpression innerhalb von Genfamilien, ein permanenter und vielfältig aufkommender Forschungsbedarf. Die in *Publikation 4* beschriebene Software *Ortho2ExpressMatrix* sollte daher breit und unter dem Gesichtspunkt universeller Verwendbarkeit ausgelegt sein, Expressionsdaten im Kontext von Genfamilien genomweit visualisieren zu können.

2. Ergebnisse mit Diskussion: Publikationen

2.1 Publikation 1: „Meta-Analysis of General Bacterial Subclades in Whole-Genome Phylogenies Using Tree Topology Profiling"

2.1.1 Bibliographische Angabe

Thomas Meinel, Antje Krause (2012). *Evolutionary Bioinformatics Online* 8: 489-525

Doi: 10.4137/EBO.S9642

2.1.2 Zusammenfassung Publikation 1

Diese Forschungsarbeit umfasst die Meta-Analyse von 47 publizierten Genomphylogenien, die aus 30 Publikationen zusammengetragen wurden. Es werden die darin gezeichneten und zumeist bifurkierenden Bäume in der Weise analysiert, dass Topo-

logien von in ihrem Genom vollständig untersuchten Bakteriengruppen in einen eigens erstellten Topologie-Katalog eingeordnet werden. Die in den Bäumen gefundenen reinen Topologiebeschreibungen wurden einer Generalisierung unterworfen und in Scorewerte übersetzt. Dieses Vorgehen eröffnet nun die Möglichkeit einer algorithmischen Behandlung, womit das mit dieser Veröffentlichung eingeführte ‚Tree Topology Profiling' umschrieben ist. Die sehr heterogenen Genombäume - sie unterscheiden sich hauptsächlich in der Anzahl der implizierten Spezien, im zugrunde liegenden Datenmodell und in der Methodik der Erstellung der jeweiligen Phylogenie - wurden damit durch ein sehr einfaches Verfahren quantitativ vergleichbar gemacht. Das erste Ziel, rekonstruierte Phylogenien mit deren Datenmodellen und Generierungsverfahren für eine Diskussion gegenüberzustellen, konnte so erreicht werden.

Als zweites Ziel konnte für die Gruppe der ‚Gene Content' Phylogenien durch Simulation mit neuen Daten gezeigt werden, welches die innerhalb dieser Gruppe besser geeigneten Verfahren für die Generierung dieser Phylogenien sind. Zentraler Ansatz war die Charakterisierung der Spezien durch binäre Profile über SYSTERS-PhyloMatrix[6] Proteinfamilien mit mindestens drei Spezien. Es zeigte sich, dass für die Generierung der Bäume Distanz-basierte Heuristiken (Neighbor Joining) mit Distanzmetriken nach Simpson[3] oder nach Korbel[21] am realistischsten sind.

Mit dieser Veröffentlichung lässt sich durch den Vergleich der unterschiedlichen Rekonstruktionen des Baums des Lebens die Nähe bestimmter Bakteriengruppen zueinander bestätigen, wie etwa die Nähe der Spirochäten zu den Chlamydien und die der Actinobakterien zu den Cyanobakterien. Dies ist bemerkenswert, weil in neueren Phylogenien[22, 23] eine Bifurkation für diesen Teil des Baums des Lebens aus systematischen Gründen vernachlässigt wird oder aufgegeben wurde.

Das zugrunde liegende Datenmodell SYSTERS-Phylomatrix[6] erweist sich mit dieser Untersuchung als biologisch sinnvoll, weil erwähnte Distanzmetriken sinnvolle Phylogenien erzeugen, und zwar im gleichen Maße wie für publizierte Phylogenien. Das Ergebnis dieser Meta-Analyse trägt somit auch übergeordnet zur Bestätigung des gesamten SYSTERS Projektes[24] bei und reiht sich in frühere Analysen ein, wie die Analyse zur Spezifität von Proteinfamilien für Spezien und Gruppen von Spezien[25] oder eine Betrachtung der 20 Aminoacyl-tRNA-Synthetase-Familien, die die biologisch sinnvolle Partitionierung des gesamten Proteinsequenzraumes durch die in SYSTERS angewendete Methodik anhand ausgewählter Protein-Familien zeigte.[26]

In der Publikation sind die Arbeitsschritte für die durchgeführte Meta-Analyse sowie das Clustering-Resultat als ‚Heatmap' abgebildet. Im Anhang befinden sich sieben SYSTERS-Phylomatrix ‚Gene Content' Genomphylogenien, die mit Parsimony-Methoden und mit Neighbor Joining erzeugt wurden.

2.1.3 Bedeutung im Kontext dieser Dissertation

Erst das durch diese Publikation eingeführte ‚Tree Topology Profiling' ermöglichte methodisch die Quantifizierbarkeit der zunächst nur deskriptiven Topologiemerkmale. Durch deren Übertragung auf Profile ließ sich eine Vergleichbarkeit der extrem unterschiedlichen Genomphylogenien erzielen. Die zweite Bedeutung dieser Arbeit liegt in der Profil-gestützten Generierung von ‚Gene Content'-Phylogenien, bei deren Berechnung die Distanzmetriken nach Simpson oder Korbel am erfolgreichsten waren.

2.1.4 Perspektiven und Ansätze für weitere Entwicklungen

Nachdem gezeigt werden konnte, welche Distanzmetriken besonders wertvoll für einen speziellen Bereich der Biologie sind, erhebt sich die Frage, inwiefern sich diese zusammen mit den entsprechenden Distanz-basierten Algorithmen auch für den Vergleich von Chemosensitivitätsprofilen biologisch aktiver Substanzen eignen. Dies wäre dann eine Alternative zur Berechnung über die Pearson-Korrelation (siehe Kap. 2.2.3). Mit dem gleichen Ansatz könnte auch die automatisierte Kategorisierung einer großen Anzahl dieser Substanzen geführt werden - was seinerseits Einordnungen nach struktur-chemischen Gesichtspunkten kontrastieren würde.

2.2 Publikation 2: „CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge"

2.2.1 Bibliographische Angabe

Jessica Ahmed, Thomas Meinel, Manuela S. Murgueitio, Robert Adams, Corinna Blasse, Andreas Eckert, Saskia Preissner, Robert Preissner (2011).

Nucleic Acids Research 39:D960-D967 (Database issue)

Doi: 10.1093/nar/gkq910

2.2.2 Zusammenfassung Publikation 2

Die Datenbank CancerResource komplementiert umfassende Literatur-Recherchen über den Zusammenhang von Wirksubstanzen und Zielgenen durch die Integration einer Vielzahl von Experimentaldaten. Ihre Ausrichtung besitzt damit zwei Stufen: Die Recherche als erste Stufe beinhaltet ein Datamining, bei dem über 19 Mio. Zusammenfassungen (Abstracts) von in PubMed verfügbaren Publikationen automatisiert durchsucht wurden; das Ergebnis, 8000 Hits, wurde anschließend manuell validiert. Zu den mehr als 900 Publikationen mit relevanter Wirksubstanz-Zielgen-Annotierung in hoher Qualität konnten Informationen anderer Repositorien hinzugefügt werden,[27-30] die das gefundene Interaktionsspektrum erheblich erweiterten. Die Fokussierung der Datenbank auf Krebs-Relevanz ist im Wesentlichen durch die Zugehörigkeit der entsprechenden Gene zu Krebs-relevanten KEGG-Signalwegen[31] definiert. Auf Grundlage der in der zweiten Stufe hinzugefügten experimentellen Daten zur Chemosensitivität und Genexpression sind diese Interaktionen über die entsprechenden Parameter - auch in Form vollständiger Profile - in der Datenbank auffindbar. Diese experimentellen Daten untermauern somit die Wirksubstanz-Zielgen-Interaktionen.

CancerResource findet in der Literatur inzwischen großes Interesse. Hier war immer die Frage nach Zielgenen und entsprechenden biologisch wirksamen Substanzen gestellt, einerseits bei der Analyse von Interaktionsbeziehungen (‚druggability') neu gefundener Gene und deren Mutationen im Zusammenhang mit Lungenkrebs[32] oder bei der Priorisierung von Genen in regulatorischen Netzwerken im Zusammenhang von Ovarialkarzinomen.[33] Wegen ihrer Integrativität wurde CancerResource für eine Signalweg-Assoziierungs-Vorhersage ausgeschöpft[34], um einzelne Krebstypen mit spezifischen Wirkstoffen behandeln zu können.

In der Publikation werden Schaudiagramme dargestellt, die die Datenintegration beschreiben und dem Benutzer zwei mehrstufige Arbeitsschrittkaskaden (‚Pipelines') vorschlagen.

2.2.3 Bedeutung im Kontext dieser Dissertation

An vier Stellen wird durch diese Publikation das Kernthema der vorliegenden Dissertation berührt: 1. durch die Integration von in beiden Dimensionen durchsuch- und analysierbaren Genexpressionsprofilen, 2. durch die Integration von Profilen von *differentieller* Genexpression, also vor und nach Einfluss von Wirksubstanzen (sowie mit einem Vergleich zu Resultaten aus den Literaturrecherchen), 3. durch die Integration von Chemosensitivitätsprofilen, und 4. durch die dynamische Erzeugung einer Ereignismatrix, in der aufgezeigt wird, für welche Substanzen die entsprechenden Zielgene gefunden werden konnten.

Zu 1.: Profilierte Genexpressionsdaten erlauben den Vergleich von Genen aber auch von Gewebeproben eines bestimmten Krebstypes. Gene werden durch ihre Expression in den 60 NCI-Krebszelllinien[11] in einem im Web-Browser horizontalen Profil charakterisiert. Die Normalisierung der Daten erfolgt über die ‚relative abundance' analog zu einer Berechnung über zu vergleichende Gewebe in Mensch und Maus,[35] der Vergleich dynamisch über die Berechnung mit der Pearson Korrelation. Diese so standardisierten Daten zur Genexpression sind im Vergleich mit Balkendiagrammen anderer Publikationen[36] in Farbbalken übersetzt, der Farbcode unterscheidet sich dabei von dem für die differentielle Genexpression, siehe Punkt 2. Ähnliche Profile für Gene sind suchbar oder werden in den Kontext von Protein-Protein-Interaktionen[37] gestellt. Eine beliebige Gewebeprobe kann, auch mit Mikroarray-Daten, genom-weit über die (für NCI-60 Zelllinien im Web-Browser vertikal orientierten) Expressionsprofile mit der Datenbank abgeglichen werden.

Zu 2.: Aus Daten zur differentiellen Genexpression können indirekte Wirksubstanz-Zielgen-Interaktionen abgeleitet werden. Auf entsprechende Ergebnisse einer externen Datenquelle (BROAD Institute, Cambridge, MA, USA),[38] die den Einfluss von Wirksubstanzen auf Gene in ausgewählten Zelllinien aufzeigt, wird über einen Web Service zugegriffen, vgl. *Publikation 3*. Neben den dynamisch erzeugten Profilen dieser experimentellen Resultate ist eine Gegenüberstellung mit Resultaten der Literaturrecherche auf einer nachgeschalteten Internetseite einsehbar.

Zu 3.: Chemosensitivitätsprofile von Wirksubstanzen charakterisieren in einer festgelegten Reihung von Krebs-Zelllinien - analog zu Genexpressionsdaten - diese Substanzen über die halb-maximale Wachstumsinhibition (GI-50) in jeder dieser Zellli-

nien. Dabei sind die Profile bereits in Bit-Muster übersetzt, als solche in der Datenbank abgelegt und mit einer entsprechenden schnellen Routine bereits auf Datenbankebene vergleichbar.[14]

Zu 4.: Ein weiterer Visualisierungstyp, der mit dieser Arbeit beschrieben wird, besteht in Ereignismatrizen, die Interaktionen von Zielgenen und deren Wirksubstanzen aufzeigen. Eine solche Ereignismatrix wird mit einer Karte in Verbindung gebracht, die einen Krebs-spezifischen Signalweg der KEGG Datenbank[31] visualisiert. Über vierzig solcher (interaktiver) Karten sind wesentlicher Bestandteil der CancerResource. Diese Matrix gibt Hinweise auf ‚Repositioning' oder auf Nebenwirkungseffekte, also auf das multiple Einwirken einer chemischen Substanz auf die in diesem Signalweg vorhandenen Gene, oder auf Substanzen, die alternativ auf ein Gen einwirken.

2.2.4 Perspektiven und Ansätze für weitere Entwicklungen

Eine Überarbeitung der Chemosensitivitätsprofile könnte mit dem Ziel erfolgen, eine aus den Profilen selbst heraus erzeugte Gruppierung sowie eine Ähnlichkeitssuche zu ermöglichen bzw. dass eine bereits publizierte Gruppierung[39] überprüft werden kann. Dies sollte sich auch nach den Erkenntnissen richten, die sich aus den Ergebnissen des Kapitels 2.1 bezüglich der Distanzmetriken ergeben. Der bezeichnende Unterschied zu Genexpressionsdaten liegt in der jeweiligen Beteiligung der über die Krebszelllinien charakterisierten biophysikalischen Entitäten Gen bzw. Substanz. Während Gene bei der Aufnahme von Expressionsprofilen nur charakterisiert werden, also die äußere Ursache Krebs und ein Gen als inhärenter Teil der (unbehandelten) Krebszelle eine passive Größe ist, ist in Chemosensitivitätsprofilen die chemische Substanz die Ursache, die aktiv auf die Zelllinien einwirkt und gleichzeitig den zu beschreibenden Parameter GI-50 erzeugt. Die Kombination der beiden Einheiten Gen und Substanz über eine Linearkombination beider Profile könnte allenfalls die Sensitivität einer möglichen Therapie herausstellen.[40] Deren Zusammenführung wäre eine weitere Perspektive hinsichtlich der Weiterentwicklung der Weboberfläche.

2.3 Publikation 3: „SOAP/WSDL-based Web Services for Biomedicine: Demonstrating the Technique with the CancerResource“

2.3.1 Bibliographische Angabe

Thomas Meinel, Manuela S. Müller, Jessica Ahmed, Reha Yildirimman, Mathias Dunkel, Ralf Herwig, Robert Preissner (2010). In: Panagiotis D. Bamidis and Nicolas Pallikarakis (Eds.): *XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010 (IFMBE Proceedings)*, Chalkidiki, Greece. Berlin Heidelberg: Springer. 835-838

Doi: 10.1007/978-3-642-13039-7_211

Beitrag zur Konferenz: Web-based Applications in Health Care & Biomedicine, MEDICON 2010. 27-30 May 2010, Porto Carras, Chalkidiki, Greece.

2.3.2 Zusammenfassung Publikation 3

In der vorgestellten Publikation werden drei inhaltlich unterschiedliche Implementierungen von Web-Service-Klienten beschrieben, die Beispiele für deren Auslegung sind. Erst wird ein einfacher Datenzugriff auf die SYSTERS Datenbank für Protein-Familien beschrieben (die extra hierfür Server-seitig erweitert wurde), dann die Anbindung einer Software der KEGG Datenbank[41] über einen Web Service zur dynamischen Erzeugung von kolorierten Signalwechselwegen. Weiterhin werden vom Server der Datenbank GenomeMatrix vorberechnete Daten zur differentiellen Genexpression bezogen, die in Zelllinien nach Behandlung mit chemischen Wirksubstanzen gemessen werden konnten.[38] Dieses Web Server Modul ist als Arbeitsfluss (workflow) mehrerer Web Service Komponenten gestaltet und bietet dem Benutzer interaktiv Auswahlmöglichkeiten an.

Dieser durch ‚peer-reviewing' begutachtete, nicht indizierte Konferenzbeitrag zeigt die neuen technischen Möglichkeiten auf, die Web-Services bieten können. In einer früheren Veröffentlichung wurde die Bedeutung von Web-Services in der Biomedizin herausgestellt, diese die Interoperabilität von Computer untereinander ermöglichen und so Ergebnisse dynamisch entstehen lassen können.[42] Diese Technologie, die

programmatisch ganz unterschiedlich ausgelegt sein kann, ist in der Lage, Datenquellen, Software aber auch medizinische Geräte miteinander zu verbinden - sie ist daher eine Schlüsseltechnologie für die Verbindung dezentraler, delokalisierter und daher prinzipiell unabhängiger Datenquellen. Dies konnte anhand der CancerResource Anwender-seitig mit mehreren Web-Service-Klienten gezeigt werden.

Es werden ein einfaches Programmierbeispiel in der Programmiersprache PHP und ein Flussdiagramm für eine Funktionsabfolge gezeigt, die die oben erwähnten Möglichkeiten zusammenfasst.

2.3.3 Bedeutung im Kontext dieser Dissertation

Einer der beschriebenen Web-Services dient der Beschaffung der Daten, die für die Erzeugung der Profile der differentiellen Genexpression notwendig sind.

2.3.4 Perspektiven und Ansätze für weitere Entwicklungen

Web Services auf Basis der ‚Web Service Description Language' WSDL wären geeignet, (z.B. im Arbeitskreis) existierende Datenbanken miteinander zu vernetzen, ohne Eingriffe in existierende Datenstrukturen zu machen. Für eine solche Aufgabe sind mit dieser Publikation technische und inhaltliche Möglichkeiten für die Etablierung von Web-Service-Klienten aufgezeigt.

2.4 Publikation 4: **„Ortho2ExpressMatrix—a web server that interprets cross-species gene expression data by gene family information"**

2.4.1 Bibliographische Angabe

Thomas Meinel, Michal-Ruth Schweiger, Andreas Hanno Ludewig, Ramu Chenna, Sylvia Krobitsch, Ralf Herwig (2011). *BMC Genomics* 12:483

Doi: 10.1186/1471-2164-12-483

2.4.2 Zusammenfassung Publikation 4

Ortho2ExpressMatrix ist eine Internet-basierte Software, die zwei Kernbereiche der Bioinformatik zusammenführt, die der Gruppierung ähnlicher Gene (vornehmlich in Proteinfamilien oder allgemein Genfamilien) mit deren Quantifizierung (in Form der Genexpression). Dabei wird jede dieser Genfamilien in der Ortho2ExpressMatrix als Entität aufgefasst und in einer Matrix visualisiert, an deren beiden Achsen nach Maßgabe des Benutzers zwei zu vergleichende biologische Objekte notiert werden. Diese sind im Verhältnis zueinander gleichrangig und können bestimmten Formen des Lebens wie Spezien, Patienten(-kohorten) oder Gewebe eines Individuums kennzeichnen. Die Zellen der Matrix symbolisieren die Ähnlichkeitsbeziehung einer jeden Gen-Einheit der einen Achse zu der der anderen. Diese erste experimentelle Größe, die im Vorfeld auch zur Generierung solcher Proteinfamilien beiträgt, ist über die Sequenzähnlichkeit in Form eines BLAST-e-values,[43] und dieser wiederum stufenweise über einen Farbkode, quantifiziert. Die Achsen führen mehrfach Annotierungen wie die Bezeichnungen für Gen, Transkript, oder Mikroarray-Sonden sowie, als zweiten experimentellen Parameter, ausschließlich vom Benutzer definierte Werte der differentiellen Expression.

Die Organisation in Genfamilien kann wichtige Hinweise mit besonderer biomedizinischer Bedeutung enthalten. Bei der Inspektion mancher Proteinfamilien fällt auf, dass mehrere Gene einer Familie zu einem gemeinsamen Proteinkomplex gehören können. Neben der notwendigerweise bestehenden Sequenzähnlichkeit innerhalb einer Genfamilie sind entsprechende Familienmitglieder dann auch über Funktionalität miteinander verbunden. Über Genexpressionsdaten ist mit Ortho2ExpressMatrix leicht bestimmbar, welche Komponente eines solchen Komplexes das Nadelöhr in der Regulation sein kann. Allerdings versteht sich Ortho2ExpressMatrix nicht explizit als Werkzeug zur Identifizierung von Koexpression. Mit Ortho2ExpressMatrix kann der Benutzer, bei identischer Ursächlichkeit (zum Beispiel derselben Krankheit oder Behandlung mit demselben Medikament), mit eigenen Genexpressionsdaten explorieren, ob Paraloge eine entsprechende Funktionalität übernommen haben können, ob unterscheidbare Isoformen eines Gens ähnlich exprimiert sind oder wie sich, im Vergleich über vorgegebene Annotierungen, Entsprechungen von Genen in ihren Expressionsmustern ergeben. Das Aufzeigen von Ambiguität, die an verschiedenen Stellen der Assoziierung auftritt, stellt bei einem solchen Projekt jedoch ein großes

Problem aber auch eine Chance dar. So sind oftmals Orthologiebeziehungen zwischen zwei Spezien nicht eineindeutig. Dies ist in der Biomedizin von Relevanz, wenn ein Vergleich von Genmaterial humanen Ursprungs mit dem von verschiedenen Modellorganismen ein wichtiges Hilfsmittel zur Erkundung genetischer Ursachen von Krankheiten sein soll. Als sehr unbeliebt gilt auch die ambige Annotierung von Mikroarray-Sonden zu Genen; dies kann mit Ortho2ExpressMatrix für eine Vielzahl von Mikroarray-Plattformen in jeder einzelnen Proteinfamilie aufgeklärt werden. Ortho2ExpressMatrix ermöglicht nicht nur den Blick auf die Resultate grundsätzlich verschiedener (in der aktuellen Software-Version sind dies vier) Generierungsverfahren für Proteinfamilien, sondern - als unabhängige Alternativen zu Protein-kodierenden Genen - auch den Blick auf Familien von microRNAs. Hier sind zwei Ansätze implementiert, erstens microRNA-Familien für microRNAs mit ähnlichen Sequenzen für reife microRNAs (gemäß der Annotierung in der Datenbank miRBase[44]) und zweitens für funktions-orientierte microRNA-Familien der Datenbank TargetScan,[45] in der microRNAs mit gleicher Zielgen-Orientierung als Familie zusammengefasst sind.

In Ortho2ExpressMatrix wird für jede Genfamilie automatisch die Über-Repräsentierung signifikant differentiell exprimierter Gene (oder deren Repräsentanten wie Sonden, Isoformen, Proteine) angegeben, und zwar separat für beide biologische Objekte sowie für beide Expressionsrichtungen: für jede dieser vier Gengruppen wird mittels des Hypergeometrischen Tests ein p-value berechnet.

In der Publikation wird bildhaft neben der prinzipiellen Anordnung einer Genfamilien-Matrix schematisch der Datenfluss während der genomweiten Generierung des Ergebnisses beschrieben. Ein Suchergebnis wird in zwei Beispielen mit Daten aus dem Expressionsdatenrepositorium Gene Express Omnibus GEO[46] erklärt.

2.4.3 Bedeutung im Kontext dieser Dissertation

Ortho2ExpressMatrix bringt zwei biologische Objekte, die optisch orthogonal zueinander gestellt sind, in die vergleichende Beziehung einer Matrix. Die in den Profilen längs der Achsen visualisierten Datenreihen unterscheiden sich jedoch von Profilen der vorher erwähnten Publikationen dadurch, dass die Anzahl der Charaktere in den Wertereihen über ein Profil, die Sequenzähnlichkeit in den Zellen der Matrix oder die Reihung der Genexpression an den Achsen der Matrix in den einzelnen Genfamilien variieren. In Profilen der vorangehenden Beispiele ist die Reihenfolge dieser Charak-

tere einmalig festgelegt und für jeden Charakter variiert nur der Status (also ein binärer oder ein numerischer Wert). Damit wird in Ortho2ExpressMatrix gegenüber den klassischen Profilen ein grundsätzlich anderer Typ einer Profil-basierten Visualisierung verwendet.

2.4.4 Perspektiven und Ansätze für weitere Entwicklungen

Die aktuelle Version der Internet-basierten Software Ortho2ExpressMatrix vergleicht in den zwei Achsen der zweidimensionalen Matrix zwei gleichrangige Parameter, die Matrix ist in diesem Sinne symmetrisch: ein Austausch der Achsen verändert die Aussage der Visualisierung nicht, selbst wenn Gene zweier verschiedener Spezien in einer Genfamilie miteinander verglichen werden. Die programmatische Struktur der Software ist hingegen so ausgelegt, dass nur wenige Veränderungen nötig sind, um auch Parameter anzeigen zu können, die zueinander in asymmetrischer Beziehung stehen. Hier sei an solche Fälle gedacht wie an microRNAs und ihre Zielgene oder, um ein Beispiel mit dem Hintergrund der Evolution zu geben, an Pseudogene und deren ursprüngliche Verwandte. Damit ist Ortho2ExpressMatrix nicht nur ein in seiner Entwicklung abgeschlossenes Internetwerkzeug sondern auch eine technische wie intellektuelle Blaupause für neue Visualisierungen, wenn es darauf ankommt, zwei voneinander abhängige Parameter, die hierarchisch organisiert sind, gegenüber zu stellen.

3. Zusammenfassende Betrachtung

Mit dieser Arbeit werden anhand der vorgestellten Publikationen Profile und Matrizen als Kernthema hervorgehoben. Gleichwohl könnte aus denselben Publikationen auch ein anderer gemeinsamer Schwerpunkt herausgearbeitet werden wie der der evolutionären Beziehungen zwischen Genen oder der der Bildung von (Gen-)Familien. Eine weitere Abstraktion dieses Aspektes liefe auf eine Hierarchisierung von (genetischen) Parametern hinaus, deren Bildung und Verwendung in den Biowissenschaften eine zunehmende Bedeutung erlangen wird. Integrative Ebenen von biologischen Einheiten, also beispielsweise Genfamilien für Gene wie microRNA-Familien anstelle einzelner microRNAs, werden punktuell eine wachsende Rolle in der Analyse von genomischen Daten spielen. So finden beispielsweise kollektive Effekte immer mehr

Beachtung - die Gemeinsamkeit eines Zielgenes für alle Mitglieder einer microRNA-Familie wäre ein Beispiel dafür. Die Aufklärung solcher Zusammenhänge könnte Werkzeugen wie der *Ortho2ExpressMatrix* zufallen. Mit den besprochenen Publikationen konnten einige Beispiele der Profilbildung und Beispiele der Anwendbarkeit von Profilen und Matrizen in der Biomedizin aufgezeigt werden. Genexpressionsprofile oder Chemosensitivitätsprofile sind bereits seit geraumer Zeit etabliert - das Potential solcher Datenreihen wird im Fall der *CancerResource* vielfältig aufgezeigt. Datenbanken oder, allgemein, visualisierende Software sind, wie auch gezeigt werden konnte, eine bevorzugte Spielwiese für den Einsatz von Profilen. Nicht zuletzt können schwer quantifizierbare, zunächst nur deskriptive Beschreibungen der Natur mit dem Trick einer Profilbildung darstellbar gemacht werden, wie es beim ‚tree topology profiling' für Genomphylogenien erfolgreich versucht wurde.

4. Literaturverzeichnis

1. Jaccard, P. Nouvelles recherches sur la distribution florale. Bull. Soc. Vaud. Sci. Nat. 44 (1908).
2. Dice, L. R. Measures of the Amount of Ecologic Association Between Species. Ecology, 297-302 (1945).
3. Simpson, G. G. Notes on the measurement of faunal resemblance. American Journal of Science 258-A, 300-311 (1960).
4. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. Nature 402, 83-6 (1999).
5. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96, 4285-8 (1999).
6. Meinel, T., Krause, A., Luz, H., Vingron, M. & Staub, E. The SYSTERS Protein Family Database in 2005. Nucleic Acids Res 33, D226-9 (2005).
7. Philippe, H., Delsuc, F., Brinkmann, H. & Lartillot, N. Phylogenomics. Annu Rev Ecol Evol Syst 36, 541-62 (2005).
8. Snel, B., Huynen, M. A. & Dutilh, B. E. Genome trees and the nature of genome evolution. Annu Rev Microbiol 59, 191-209 (2005).
9. Gogarten, J. P., Fournier, G. & Zhaxybayeva, O. Gene Transfer and the Reconstruction of Life's Early History from Genomic Data. Space Sci Rev 135, 115-131 (2008).
10. Darwin, C. R. (ed. Murray, J.) (Darwin Online, http://darwin-online.org.uk/ 1837-1838).
11. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer 6, 813-23 (2006).
12. Fullbeck, M., Dunkel, M., Hossbach, J., Daniel, P. T. & Preissner, R. Cellular fingerprints: a novel approach using large-scale cancer cell line data for the identification of potential anticancer agents. Chem Biol Drug Des 74, 439-48 (2009).
13. Huang, R., Wallqvist, A., Thanki, N. & Covell, D. G. Linking pathway gene expressions to the growth inhibition response from the National Cancer Institute's anticancer screen and drug mechanism of action. Pharmacogenomics J 5, 381-99 (2005).
14. Gunther, S., Neumann, S., Ahmed, J. & Preissner, R. in 2nd Brazilian Conference on Advances in Bioinformatics and Computational Biology (BSB'07) (eds. Sagot, M.-F. & Walter, M. E. M. T.) 167-170 (Springer-Verlag, Berlin, Heidelberg, Angra dos Reis, Brazil, 2007).
15. Dolinski, K. & Botstein, D. Orthology and functional conservation in eukaryotes. Annu Rev Genet 41, 465-507 (2007).
16. Kuhn, M. et al. STITCH 2: an interaction network database for small molecules and proteins. Nucleic Acids Res 38, D552-6 (2010).
17. Obayashi, T. & Kinoshita, K. COXPRESdb: a database to compare gene coexpression in seven model animals. Nucleic Acids Res 39, D1016-22 (2011).
18. Forslund, K., Schreiber, F., Thanintorn, N. & Sonnhammer, E. L. OrthoDisease: tracking disease gene orthologs across 100 species. Brief Bioinform 12, 463-73 (2011).
19. Huminiecki, L. & Wolfe, K. H. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res 14, 1870-9 (2004).
20. Chan, E. T. et al. Conservation of core gene expression in vertebrate tissues. J Biol 8, 33 (2009).
21. Korbel, J. O., Snel, B., Huynen, M. A. & Bork, P. SHOT: a web server for the construction of genome phylogenies. Trends Genet 18, 158-62 (2002).
22. Moran, N. A., McCutcheon, J. P. & Nakabachi, A. Genomics and evolution of heritable bacterial symbionts. Annu Rev Genet 42, 165-90 (2008).
23. Muller, J. et al. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res 38, D190-5 (2010).

24. Krause, A., Stoye, J. & Vingron, M. Large scale hierarchical clustering of protein sequences. BMC Bioinformatics 6, 15 (2005).
25. Meinel, T., Vingron, M. & Krause, A. in Proceedings of the German Conference on Bioinformatics (GCB03) 103-108 (Belleville, München, 2003).
26. Meinel, T. in Handbook of Research on Systems Biology Applications in Medicine (ed. Daskalaki, A.) 143-166 (IGI Global, Hershey, Pennsylvania, USA, 2009).
27. Hernandez-Boussard, T. et al. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. Nucleic Acids Res 36, D913-8 (2008).
28. Wishart, D. S. et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36, D901-6 (2008).
29. Davis, A. P. et al. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic Acids Res 37, D786-92 (2009).
30. Zhu, F. et al. Update of TTD: Therapeutic Target Database. Nucleic Acids Res 38, D787-91 (2010).
31. Kanehisa, M. et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34, D354-7 (2006).
32. Xiong, D. et al. Exome sequencing identifies MXRA5 as a novel cancer gene frequently mutated in non-small cell lung carcinoma from Chinese patients. Carcinogenesis (2012).
33. Madhamshettiwar, P. B., Maetschke, S. R., Davis, M. J., Reverter, A. & Ragan, M. A. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. Genome Med 4, 41 (2012).
34. Ma, H. & Zhao, H. iFad: an integrative factor analysis model for drug-pathway association inference. Bioinformatics (2012).
35. Liao, B. Y. & Zhang, J. Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol 23, 530-540 (2006).
36. Gmeiner, W. H., Reinhold, W. C. & Pommier, Y. Genome-wide mRNA and microRNA profiling of the NCI 60 cell-line screen and comparison of FdUMP[10] with fluorouracil, floxuridine, and topoisomerase 1 poisons. Mol Cancer Ther 9, 3105-14 (2010).
37. Kamburov, A. et al. ConsensusPathDB: toward a more complete picture of cell biology. Nucleic Acids Res 39, D712-7 (2011).
38. Lamb, J. et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313, 1929-35 (2006).
39. Holbeck, S. L., Collins, J. M. & Doroshow, J. H. Analysis of Food and Drug Administration-approved anticancer agents in the NCI60 panel of human tumor cell lines. Mol Cancer Ther 9, 1451-60 (2010).
40. Scherf, U. et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet 24, 236-44 (2000).
41. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38, D355-60 (2010).
42. Meinel, T. & Herwig, R. in Web-Based Applications in Healthcare and Biomedicine (ed. Lazakidou, A.) 101-116 (Springer US, 2010).
43. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-402 (1997).
44. Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. miRBase: tools for microRNA genomics. Nucleic Acids Res 36, 154-158 (2008).
45. Friedman, R. C., Farh, K. K., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 19, 92-105 (2009).
46. Barrett, T. et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 37, D885-90 (2009).

Anteilserklärung

Impact Factor (IF) nach Thompson Reuters, Journal Citations Report, ISI Web of Knowledge, Stand 6.6.2012

Publikation 1:
„Meta-Analysis of General Bacterial Subclades in Whole-Genome Phylogenies Using Tree Topology Profiling“

Thomas Meinel, Antje Krause (2012)

Evolutionary Bioinformatics Online 8: 489-525

Doi: 10.4137/EBO.S9642

PMID: 22915837

Idee / Konzept	90 %	
Programmierung / Datenauswertung	100 %	
Manuskript	90 %	
Gesamt	95 %	
Erstautorenschaft	ja	
Peer-reviewed	ja	
Impact Factor (IF)	2.684	(5-jähriger IF: 12.620)

Publikation 2:
„CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge“

Jessica Ahmed*, Thomas Meinel*, Manuela S. Murgueitio, Robert Adams, Corinna Blasse, Andreas Eckert, Saskia Preissner, Robert Preissner (2011)

Nucleic Acids Research 39:D960-D967 (Database issue)

Doi: 10.1093/nar/gkq910

(* gleichberechtigte Erstautorenschaft)

PMID: 20952398

Idee / Konzept	20 %	
Programmierung / Datenauswertung	40 %	
Manuskript	90 %	
Gesamt	60 %	
Erstautorenschaft	ja, geteilt	
Peer-reviewed	ja	
Impact Factor (IF)	7.836	(5-jähriger IF: 7.314)

Publikation 3:
„SOAP/WSDL-based Web Services for Biomedicine: Demonstrating the Technique with the CancerResource“

Thomas Meinel, Manuela S. Müller, Jessica Ahmed, Reha Yildirimman, Mathias Dunkel, Ralf Herwig, Robert Preissner (2010)

In: P. D. Bamidis and N. Pallikarakis (Eds.): *XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010 (IFMBE Proceedings)*, Chalkidiki, Greece. Berlin Heidelberg: Springer. 835-838

Doi: 10.1007/978-3-642-13039-7_211

Beitrag zur Konferenz: Web-based Applications in Health Care & Biomedicine, MEDICON 2010. 27-30 May 2010, Porto Carras, Chalkidiki, Greece.

Idee / Konzept	100 %	
Programmierung / Datenauswertung	80 %	
Manuskript	95 %	
Gesamt	90 %	
Erstautorenschaft	ja	
Peer-reviewed	ja	(Konferenzbeitrag)
Impact Factor (IF)	(kein)	

Publikation 4:
„Ortho2ExpressMatrix—a web server that interprets cross-species gene expression data by gene family information“

Thomas Meinel, Michal-Ruth Schweiger, Andreas Hanno Ludewig, Ramu Chenna, Sylvia Krobitsch, Ralf Herwig (2011)

BMC Genomics 12:483

Doi: 10.1186/1471-2164-12-483

PMID: 21970648

Idee / Konzept	100 %	
Programmierung / Datenauswertung	100 %	
Manuskript	90 %	
Gesamt	95 %	
Erstautorenschaft	ja	
Peer-reviewed	ja	
Impact Factor (IF)	4.073	(5-jähriger IF: 4.328)

Unterschriften

Berlin, 11. Juli 2012 Thomas Meinel

Berlin, 11. Juli 2012 PD Dr. Robert Preißner

Druckexemplare der ausgewählten Publikationen oder elektronische Verweise

Publikation 1:

„Meta-Analysis of General Bacterial Subclades in Whole-Genome Phylogenies Using Tree Topology Profiling“

http://dx.doi.org/10.4137/EBO.S9642

(Die Seiten 23-46 sind im Druckexemplar enthalten oder online erhältlich.)

Publikation 2:

„CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge“

http://dx.doi.org/10.1093/nar/gkq910

(Die Seiten 48-55 sind im Druckexemplar enthalten oder online erhältlich.)

Publikation 3:

„SOAP/WSDL-based Web Services for Biomedicine: Demonstrating the Technique with the CancerResource“

http://dx.doi.org/10.1007/978-3-642-13039-7_211

(Die Seiten 57-62 sind im Druckexemplar enthalten oder online erhältlich.)

Publikation 4:

„Ortho2ExpressMatrix—a web server that interprets cross-species gene expression data by gene family information“

http://dx.doi.org/10.1186/1471-2164-12-483

(Die Seiten 64-74 sind im Druckexemplar enthalten oder online erhältlich.)

Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Berlin, 22. März 2013 Thomas Meinel

Vollständige Publikationsliste

Poster:

T. Meinel, A. Krause (2012)

Introducing Tree Topology Profiling for Meta-Analysis of Whole-Genome Phylogenies

GCB 2012 - German Conference on Bioinformatics 2012, Jena, Germany.

Fachpublikation:

T. Meinel, A. Krause (2012)

Meta-Analysis of General Bacterial Subclades in Whole-Genome Phylogenies Using Tree Topology Profiling

Evolutionary Bioinformatics Online 8: 489-525.

Doi: 10.4137/EBO.S9642

Fachpublikation:

T. Meinel, M.-R. Schweiger, A. H. Ludewig, R. Chenna, S. Krobitsch, R. Herwig (2011)

Ortho2ExpressMatrix—a Web Server that Interprets Cross-Species Gene Expression Data by Gene Family Information

BMC Genomics 12:483.

Doi: 10.1186/1471-2164-12-483

Poster:

A. H. Ludewig, T. Meinel, F. Döring (2011)

Identification of Evolutionary Conserved Regulators of Dietary Restriction Using the "Ortho2ExpressMatrix"

18th International C. elegans Meeting 2011. Los Angeles, California, USA.

Fachpublikation:

J. Ahmed*, T. Meinel*, M. S. Murgueitio, R. Adams, C. Blasse, A. Eckert, S. Preissner, R. Preissner (2011) (* gleichberechtigte Erstautorenschaft)

CancerResource: a Comprehensive Database of Cancer-Relevant Proteins and Compound Interactions Supported by Experimental Knowledge

Nucleic Acids Research 39 (Database issue): D960-D967.

Doi: 10.1093/nar/gkq910

Konferenzbeitrag:

T. Meinel, M.S. Müller, J. Ahmed, R. Yildirimman, M. Dunkel, R. Herwig, R. Preissner (2010)

SOAP/WSDL-based Web Services for Biomedicine: Demonstrating the Technique with the CancerResource

In: P. D. Bamidis and N. Pallikarakis (Eds.): *XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010 (IFMBE Proceedings)*, Chalkidiki, Greece. Berlin Heidelberg: Springer. 835-838.

Doi: 10.1007/978-3-642-13039-7_211

Buchartikel:

T. Meinel, R. Herwig (2010)

SOAP/WSDL-based Web Services for Biomedicine

In: A. Lazakidou (Ed.): *Web-Based Applications in Healthcare and Biomedicine*. New York, USA: Springer. Chapter 7: 101-116.

Doi: 10.1007/978-1-4419-1274-9_7

Poster:

T. Meinel (2009)

Influence of Methodological Factors on the Inference of Gene Content Trees

Celebrating Darwin: From The Origin of Species to Deep Metazoan Phylogeny. DMP 2009. Berlin, Germany.

Buchartikel:

T. Meinel (2009)

Sequence Similarity and Function of Homologous Proteins - Phylogenetic Profiling

In: A. Daskalaki (Ed.): *Handbook of Research on Systems Biology Applications in Medicine*. Hershey, Pennsylvania, USA: IGI Global. Section III: Genomics and Bioinformatics for Systems Biology, Chapter VIII: 143-166.

Doi: 10.4018/978-1-60566-076-9

Fachpublikation:

T. Meinel, A. Krause, H. Luz, M. Vingron, E. Staub (2005)

The SYSTERS Protein Family Database in 2005

Nucleic Acids Research 33 (Database issue): D226-D229.

Doi: 10.1093/nar/gki030

Poster:

T. Meinel, A. Krause, E. Staub, M. Vingron (2004)

PhyloMatrix: a Tool for Phylogenetic Profiling within the SYSTERS Protein Family Web Server

ISMB/ECCB 2004. Glasgow, UK.

Poster:

T. Meinel, A. Krause, E. Staub, H. Luz, S. Hartmann, U. Krämer, J. Selbig, M. Vingron (2004)

The SYSTERS Protein Family Web Server: Shortcut from Large-Scale Sequence Information to Phylogenetic Information

ARABIDOPSIS meeting 2004. Berlin, Germany.

Konferenzbeitrag:

T. Meinel, M. Vingron, A. Krause (2003)

The SYSTERS Protein Family Database: Taxon-Related Protein Family Size Distributions and Singleton Frequencies

In: *Proceedings of the German Conference on Bioinformatics*. Munich, Germany: Belleville. 103-108.

Poster:

T. Meinel, A. Krause, M. Vingron (2003)

SYSTERS Protein Family Database: Taxonomy Web Interface and Taxon-related Cluster Frequencies

ECCB 2003 Paris, France.

Poster:

T. Meinel, A. Krause, M. Vingron (2003)

A Taxonomical View on the SYSTERS Protein Family Database

RECOMB 2003 Berlin, Germany.

Poster:

A. Krause, T. Meinel, J. Stoye, H.A. Schmidt, H. Luz, M. Vingron (2002)

The SYSTERS Protein Family Webserver

ECCB 2002 Saarbrücken, Germany.

Selbständigkeitserklärung

Ich, Thomas Meinel, erkläre, dass ich die vorgelegte Dissertation mit dem Thema:

> ‚Repräsentation von Biomolekülen und Bioereignissen durch Profile und Matrizen, deren Vergleichbarkeit durch Metriken und ihre Bedeutung in der Biomedizin'

selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, ohne die (unzulässige) Hilfe Dritter verfasst und auch in Teilen keine Kopien anderer Arbeiten dargestellt habe.

Berlin, 22. März 2013 Thomas Meinel

Danksagung

Der Erfolg dieser Arbeit wurde erst durch den Einsatz und das Zusammenspiel vieler Seiten ermöglicht. Zuallererst möchte ich Robert Preissner für die weitgehend freie Hand bei der Ausgestaltung dieser Arbeit, die freundliche Aufnahme in seinen Arbeitskreis und die Möglichkeit danken, an bestimmten Projekten Anteil haben zu dürfen. Damit konnten die Grundlagen zu dieser Arbeit gelegt werden, die während der Zeit am Institut für Physiologie der Charité in seiner Gruppe „Strukturelle Bioinformatik“ entstanden. So konnte die das Projekt CancerResource nur mit dem Beitrag vieler Kollegen wie Jessica Ahmed, Mathias Dunkel, Robert Adams, Corinna Blasse, Manuela Sabrina Murgueitio, Andreas Eckert entstehen. Vielen Dank Euch allen.

Ortho2ExpressMatrix entstand aus dem Bestreben, Biologen ein geeignetes Werkzeug in die Hand geben zu wollen, das zwei Kernbereiche der Bioinformatik, Berechnung der Orthologie von Genen und Ermittlung von Genexpression, miteinander verbindet. Dieses Werkzeug wäre nicht ohne eine Anfrage von Hanno Andreas Ludewig entstanden. Dass Andreas das daraufhin entstandene Webtool nun wissenschaftlich nutzt und auf Konferenzen bewirbt, ist höchste Ehre für den Programmierer. Vielen Dank auch Sylvia Krobitsch und Michal-Ruth Schweiger, die durch ihre Anmerkungen und ihre Diskussionsbereitschaft wesentlich zur Entwicklung des Tools beitrugen. Ralf Herwig habe ich sehr schätzen gelernt, auch bei dieser Entwicklung, als master mind behind. - Diese Publikation dokumentiert meine tiefste Überzeugung, dass ein Projekt nur dadurch entstehen kann, wenn interessierte Menschen gemeinsam zu einer Sache beitragen. Wir haben das hier so gemacht - Dank Euch allen.

Die Publikationen mit Antje Krause sind Dokumente gemeinsam gelebter Wissenschaft. Vielen Dank, liebe Antje, für die fachliche Einführung in die Welt und die Organisation der Eiweißsequenzen und die vielen angeschlossenen Diskussionen. Mit Protein-Familien fing für mich vor mehr als zehn Jahren die Bioinformatik an, und es schließt sich mit der kürzlich erschienenen Publikation zu dieser Thematik einer der vielen gemeinsamen Kreise, ohne den diese Arbeit auch nicht möglich war.

Finanzielle Ausstattungen der beschriebenen Publikationen erfolgten durch: Technische Fachhochschule Bingen (Tree Topology Profiling); Bundesministerium für Bildung und Forschung BMBF, MedSys Projekt ‘Therapeutic Systems Immunology’ (0315450A) (CancerResource); Max-Planck Gesellschaft (Ortho2ExpressMatrix).