Technische Universität München
Lehrstuhl für Genomorientierte Bioinformatik
GSF - Forschungszentrum für Umwelt und Gesundheit
Institut für Bioinformatik

Analyse konservierter Nachbarschaftsbeziehungen in vollständig sequenzierten Genomen

Dirk Haase

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Dimitri Frischmann
Prüfer der Dissertation:	1. Univ.-Prof. Dr. Hans-Werner Mewes 2. Univ.-Prof. Dr. John Parsch (Ludwig-Maximilians-Universität München)

Die Dissertation wurde am 10.05.2007 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 22.08.2007 angenommen.

Danksagung

Das Titelblatt dieser Arbeit führt – da es sich um eine Dissertation handelt – nur einen Autor auf. Dennoch gibt es viele, die wichtiges zum Gelingen beigetragen haben. Einigen von ihnen möchte ich hier besonderen Dank aussprechen.

An erster Stelle danke ich Prof. Dr. Hans-Werner Mewes, der mir durch die Anstellung an seinem Institut für Bioinformatik (MIPS) die Promotion ermöglichte und mich davon überzeugte, diesen Weg zu gehen. Er hat mich mit wissenschaftlich relevanten Projekten betraut, deren Ergebnisse ich in renommierten Zeitschriften veröffentlichen und schließlich in Form dieser Doktorarbeit zusammen tragen konnte.

Ein wesentlicher Teil der Arbeiten zur Attribut-Cluster Analyse wurde während meiner Beschäftigung beim Kompetenzzentrum Pathogenomik an der Julius-Maximilians-Universität Würzburg durchgeführt. Mein Dank geht an das von Prof. Dr. Werner Göbel geleitete Team. Der Firma Biomax Informatics AG (insbesondere Dr. Kaj Albermann) danke ich für die Überlassung ihrer Annotationsdaten für dieses Projekt. Dr. Andreas Ruepp gab mir wertvolle Hinweise zur Interpretation der Ergebnisse.

Biologische Expertise war in den in diese Arbeit eingegangenen Projekten immer essentiell. Bei den pflanzenbezogenen Analysen waren es Dr. Klaus Mayer und Dr. Heiko Schoof, bei den Pilzgenomen Dr. Gertrud Mannhaupt, Dr. Ulrich Güldener und Dr. Martin Münsterkötter, die mir ihr großes Fachwissen stets gern zur Verfügung gestellt haben.

Über die Zeit einer Promotion ist moralische Unterstützung oft ebenso wichtig wie fachliche Hilfe. Hierfür danke ich meinen Eltern und Freunden, vor allem aber den Kollegen am MIPS. Besonders Gabi Kastenmüller als ausdauernde Weggefährtin auf der Spur zum Doktorhut möchte ich hier hervorheben.

Susanne, Dir einfach Dank für alles!

Inhaltsverzeichnis

1 Einleitung **4**

2 Nachbarschaftsbeziehungen auf DNA-Ebene **7**

2.1 Einleitung 7

2.2 Hintergrund: Segmentale Duplikationen 8

2.2.1 Genom-Verändernde Prozesse 8

2.2.2 Segmentale Duplikationen in Eukaryonten-Genomen . . . 11

2.2.3 In-Silico-Methoden zum Nachweis segmentaler Duplikationen 12

2.3 Segmentale Duplikationen in *Arabidopsis* 14

2.3.1 Ansatz 14

2.3.2 Ergebnisse 20

2.4 Diskussion 26

2.4.1 Polyploidie oder Multiple Duplikationen? 26

2.4.2 Vergleich mit Ergebnissen anderer Autoren 27

2.4.3 Weitere Anwendungen 29

3 Nachbarschaftsbeziehungen auf Gen-Ebene **32**

3.1 Einleitung 32

3.2 Theoretische Ansätze 33

3.3 Eigene Arbeiten 38

3.3.1 *A. thaliana* vs. Reis 38

3.3.2 conSynteny 42

3.4 Diskussion 51

4 Beziehungen auf Funktionaler Ebene **53**

4.1 Einleitung 53

4.1.1 Begriff 53

4.1.2 Biologischer Hintergrund 54
4.1.3 Ziele 59
4.2 Literatur 60
4.2.1 Überblick 60
4.2.2 Konservierte Nachbarschaften 61
4.2.3 Phylogenetische Profile 64
4.2.4 Gen-Fusionen 65
4.2.5 Weitere Methoden 67
4.3 Attribut Cluster 68
4.3.1 Konzept 68
4.3.2 Ergebnisse 77
4.3.3 FunCat in Bakteriellen Genomen 77
4.4 Diskussion 92

5 Zusammenfassung und Ausblick **97**

Kapitel 1

Einleitung

Die Gesamtheit der genetischen Information, also das Genom eines Lebewesens, wird bisweilen auch als „Buch des Lebens“ bezeichnet. Vor allem für das menschliche Genom ist diese Metapher verbreitet. Das Bild eines Buches ist allerdings irreführend. Es impliziert, dass sich bereits aus der Nukleinsäuresequenz alle Informationen herauslesen lassen, die zum detaillierten Verständnis eines Organismus wichtig sind. Tatsächlich ist ein vollständig sequenziertes Genom jedoch zunächst nur Datenbasis für eine aufwendige Prozessierung, an deren Ende ein Erkenntnisgewinn bezüglich des betroffenen Organismus stehen kann. Stein beschreibt dies als aufeinander aufbauende Schichten von Interpretation, die um die Sequenz herum geschichtet werden (Stein 2001).

Für viele der hierzu erforderlichen Schritte liefern bioinformatische Methoden einen entscheidenden Beitrag. Allein die Menge an Sequenzdaten, die seit Aufklärung der ersten Genome Mitte der 90er Jahre zugänglich sind, zwingt zum Einsatz leistungsfähiger Algorithmen und Datenstrukturen.

Mit vollständig sequenzierten Genomen stehen erstmalig umfassende Informationen über die relative Anordnung einzelner genetischer Objekte zur Verfügung. Zwar sind auch mit klassischen Methoden Kartierungen möglich. Diese genetischen Karten können jedoch trotz hohen experimentellen Aufwands nur eine begrenzte Anzahl von Markern abbilden (Doerge 2002). Im Gegensatz dazu repräsentiert die komplette Sequenz eines Chromosoms alle enthaltenen Gene in ihrem genomischen Kontext.

Die vorliegende Arbeit führt systematische Analysen dieser Kontext-Information durch. Im Mittelpunkt steht dabei der Begriff der konservierten Nachbarschaft. Dieser Terminus beinhaltet drei Aspekte:

1. Eine Gruppe von Objekten liegt benachbart innerhalb eines bestimmten Genom-Abschnitts
2. Diese Objekte sind über eine geeignete Ähnlichkeitsfunktion auf andere, nicht in der Ausgangsgruppe enthaltene Objekte abbildbar
3. Die getroffenen Objekte liegen ihrerseits benachbart zueinander

Die folgende Definition gibt eine Formalisierung des Begriffs:

Definition 1.1 (Konservierte Nachbarschaft)
Betrachtet werden auf einen Chromosom lokalisierbare Objekte. Gegeben seien zwei disjunkte geordnete Objekt-Mengen A und B. Gesucht sind dann Teilmengen $A' \subset A$ und $B' \subset B$, so dass gilt

$$\begin{aligned} n(A') &= 1 && \text{(Nachbarschaft auf } A') \\ \forall a \in A' \exists b \in B' : s(a,b) > c \wedge \forall b \in B' \exists a \in A' : s(a,b) > c && \text{(Abbildbarkeit)} \\ n(B') &= 1 && \text{(Nachbarschaft auf } B') \end{aligned}$$

wobei eine Nachbarschaft indizierende Funktion $n : \{A, B\} \mapsto \{0, 1\}$, eine Abbildung $s : \{A, B\} \mapsto \mathcal{R}^+$ (Ähnlichkeitsfunktion) sowie eine Mindestähnlichkeit $c \in \mathcal{R}^+$ geeignet zu definieren sind. Die Gruppen benachbarter Objekte werden auch als Cluster bezeichnet.

Je nachdem, welche Objekte man betrachtet und wie man die Nachbarschaften identifiziert, kann diese sehr allgemein gehaltene Beschreibung in konkrete Fragestellungen umgesetzt werden. In den auf diese Einleitung folgenden Kapiteln untersuche ich drei unterschiedliche Konkretisierungen, die verschiedenen Abstraktionsniveaus entsprechen: Kapitel 2 analysiert konservierte Nachbarschaften auf DNA-Ebene. Am Beispiel von *A. thaliana* wird gezeigt, wie segmentale Duplikationen durch Identifizierung von Gruppen relativ kurzer homologer Sequenz-Abschnitte nachgewiesen werden können.

Kapitel 3 beschäftigt sich mit konservierten Gen-Nachbarschaften. Dabei wird von der Nukleinsäuresequenz abstrahiert. Ein Chromosom wird als Abfolge von Genen aufgefasst. Mit einer Modifikation des Verfahrens aus dem vorangegangenen Kapitel werden unter anderem syntenische Bereiche zwischen fungalen Genomen ermittelt.

Kapitel 4 behandelt Nachbarschaften auf funktionaler Ebene. In den Kapiteln 2 und 3 wurden zuerst alle (also auch nicht benachbarte) Objekte ermittelt, die sich über die gewählte Relation (Sequenzähnlichkeit) aufeinander abbilden lassen. Die Cluster gehen dann aus einem Filterprozess hervor. Hier erfolgt nun die Bestimmung der lokalen Cluster als primärer Schritt. Als Kriterium werden dazu funktionale Attribute herangezogen. Die einzelnen Cluster treten als statistisch auffällige Häufung bestimmter Genfunktionen hervor. Meine Analyse zeigt, dass in einigen bakteriellen Genomen bis zu 50% der Gene solchen funktionalen Clustern zugeordnet werden können.

Die Reihenfolge der Kapitel entspricht einer Steigerung im Abstraktionsniveau, auf dem ein Genom analysiert wird. Kapitel 2 betrachtet die genomische Sequenz an sich, ein Chromosom ist als String über dem Nukleinsäure-Alphabet {A, G, C, T} repräsentiert. In Kapitel 3 wird ein Chromosom als geordnete Liste einzelner Gene aufgefasst. Deren genaue Gestalt, ihre Sequenz in Amino- oder Nukleinsäure, selbst Länge und Abstände auf der DNA spielen dabei keine Rolle. Allenfalls die Orientierung wird in einigen Analysen berücksichtigt. Kapitel 4 schließlich transformiert das Chromosom von einer Liste von Genen in

einen binären Vektor, in dem jede Position nur noch das Vorhandensein eines bestimmten funktionalen Attributs indiziert.

Ziel der Arbeit ist es, durch Übertragung des Prinzips konservierter Nachbarschaftsbeziehungen auf jedes dieser Abstraktionslevel einen Beitrag zum Verständnis der Biologie der betrachteten Organismen zu leisten.

Kapitel 2

Nachbarschaftsbeziehungen auf DNA-Ebene

2.1 Einleitung

Als erste der drei in dieser Arbeit untersuchten Abstraktionsebenen wird in diesem Kapitel die DNA-Ebene betrachtet. Die hier beschriebene Analyse war eingebettet in ein internationales Projekt zur Sequenzierung des Genoms von *Arabidopsis thaliana*.

Die zur Familie der Brassicaceen gehörende Blütenpflanze *Arabidopsis thaliana* ist seit Mitte der 80er Jahren der wichtigste Modell-Organismus im Bereich der Botanik (Somerville und Koornneef 2002). Gründe dafür sind u.a. eine kurze Generationszeit, einfache Kultivierbarkeit sowie das Vorhandensein detaillierter genetischer Karten und einer Vielzahl an charakterisierten Mutanten (Meinke et al. 1998). Vorteile gegenüber anderen Pflanzen sind darüber hinaus ihr mit ca. 130 Mb relativ kompaktes Genom und ein geringer Anteil an repetitiven Sequenzen. Damit bot sie optimale Voraussetzungen für die erste vollständige Sequenzierung einer Pflanze. Mit diesem Ziel gründete sich 1996 die AGI (Arabidopsis Genome Initiative) als internationale Kooperation amerikanischer, japanischer und europäischer Forscher (Wambutt et al. 2000).

Eine der Fragen, die erst durch die vollständige Aufklärung des Erbguts eines Organismus geklärt werden kann, ist die nach dem Umfang an Redundanz des genetischen Materials. Darunter wird hier hauptsächlich das mehrfache Auftreten gleicher oder sehr ähnlicher Gene verstanden. Weitere Redundanz entsteht durch Transposons (ca. 10% der Gesamt-DNA bzw. 20% der inter-genischen DNA), sowie kurze repetitive Muster, die insbesondere in den Centromeren und Telomeren konzentriert sind. Außerdem gibt es eine hohe Zahl von Kopien ribosomaler DNA, geclustert in zwei sog. Nucleolus Organisierenden Regionen (NOR), jeweils eine auf den Chromosomen 2 und 4 (The Arabidopsis Genome Initiative 2000).

Redundante Gene können in drei unterschiedlichen Formationen auftreten:

- Verstreut: Die einzelnen Kopien eines Gens liegen ohne erkennbaren Bezug an unterschiedlichen Orten des Genoms
- Tandem: Kopien liegen direkt nebeneinander auf dem Chromosom
- Segmental: Längere chromosomale Abschnitte finden sich mehrfach innerhalb des Genoms

Unter den letzten Punkt lassen sich auch die Sonderfälle der Verdopplung ganzer Chromosomen oder gar des gesamten Genoms subsumieren.

Tandem-Duplikationen sind in *Arabidopsis* sehr häufig: 1528 solcher Formationen mit bis zu 23 Elementen wurden entdeckt, insgesamt enthalten sie 4140 Gene (17% der Gesamtzahl). Daneben gibt es eine hohe Anzahl an verstreuten Duplikaten, die zu sog. Gen-Familien gehören. Große Familien mit mehr als fünf Mitgliedern machen mehr als 37% aller Gene aus (The Arabidopsis Genome Initiative 2000).

Fokus meiner Arbeit innerhalb dieses Projektes ist die Aufklärung segmentaler Duplikationen in der Geschichte der Evolution von *A. thaliana*. Um eine Ausprägung des Problems konservierter Nachbarschaften handelt es sich hierbei vor allem deshalb, weil die Kontinuität der duplizierten Segmente durch vielfache lokale Mutationen und Umordnungen gestört ist. Es geht also nicht darum, perfekte Wiederholungen von DNA-Sequenzen über mehrere hunderttausend Basen zu entdecken, sondern kürzere Folgen konservierter Abschnitte, die an ihren jeweiligen Fundorten benachbart aber nicht zwingend zusammenhängend sind.

Objekte im Sinne der Definition 1.1 sind hier also DNA-Abschnitte unbekannter Länge. Welche Objekte als benachbart gelten, wird über einen maximalen Abstand, gemessen in Basen definiert. Dies ist nur möglich für Abschnitte auf dem gleichen Chromosom, daher sind die fünf Chromosomen des *Arabidopsis*-Genoms einzeln zu betrachten. Die Abbildung zwischen den Objekten erfolgt über Sequenzähnlichkeit.

Vor der Darstellung der eigenen Arbeiten folgt zunächst ein allgemeiner Überblick über segmentale Duplikationen und deren Nachweis.

2.2 Hintergrund: Segmentale Duplikationen

2.2.1 Genom-Verändernde Prozesse

Das Erbgut eines Organismus ist ständigen Veränderungen ausgesetzt. In wie weit sich diese in einer Population manifestieren können, hängt vom betroffenen Locus ab. Die Populationsgenetik analysiert unter anderem, welche Gene polymorph vorkommen und welche unter dem Einfluss positiver Selektion stehen. Einen aktuellen Überblick über die Populationsgenetik in Pflanzen gibt (Wright und Gaut 2005).

Je nach Umfang des betroffenen Bereichs haben solche Mutationen mehr oder weniger gravierende Auswirkungen auf die Architektur des Genoms. Am unteren Ende dieser Skala rangiert der Austausch einzelner Basen, der phänotypisch nur dann erkennbar wird, falls er in einer codierenden Region stattfindet und eine maßgebliche Änderung in der Aminosäure-Sequenz zur Folge hat. Dennoch führen bereits Punktmutationen zu – wenn auch zunächst minimalen – Störungen in der Identität ehemals gleicher DNA-Abschnitte. Bei hinreichender Häufung kann dies zur vollständigen Verdeckung der Gemeinsamkeiten der betroffenen Sequenzen führen.

Zu den wichtigen Mechanismen bei der Genom-Evolution zählt die Duplikation von Teilsequenzen. Diese ist, wie auch die anderen Typen chromosomaler Mutationen (Deletionen, Translokationen und Inversionen) in der Regel auf 'Unfälle' bei Rekombinationsereignissen zurückzuführen. Von einer solchen Duplikation können selbstverständlich auch Gene betroffen sein. Dies ist somit einer der Mechanismen, bei der die Anzahl der Gene innerhalb des Genoms erhöht wird. Deletionen bewirken das Gegenteil, hierbei geht genetisches Material verloren.

Unter einer Translokation versteht man die Verschiebung eines chromosomalen Segments an einen anderen Ort des Genoms. Wird dieses Segment in umgekehrter Lage eingefügt, spricht man von einer Inversion. Erkennbar wird eine solche Umkehrung u.a. an der Codierungsrichtung der enthaltenen Gene. Angenommen, innerhalb einer Region mit fünf Genen $ABCDE$ werden die mittleren drei invertiert, ergibt sich dann eine Konfiguration $A(-D)(-C)(-B)E$, wobei das negative Vorzeichen entgegengesetzte Orientierung notiert. Ein Überblick über Formalisierungen des Rearrangement-Problems folgt in Abschnitt 3.2.

Neben den bereits angesprochenen Duplikationen in kleinerem Maßstab kann es auch zur Verdopplung großer Bereiche der DNA kommen, hierfür wird dann meist der Ausdruck 'segmentale Duplikation' verwendet. Eine allgemein gültige Festlegung, ab welchem Umfang ein Duplikationsereignis segmental genannt wird gibt es allerdings nicht.

Ein weiterer Typ von Genom-Veränderungen betrifft die Anzahl an Chromosomen. Durch Unregelmäßigkeiten bei der Gameten-Bildung kann es zu überzähligen oder fehlenden Exemplaren einzelner Chromosomen kommen (Aneuploidie). Bei diploiden Organismen, also doppelter Satzzahl, spricht man entsprechend von Trisomien (dreifaches Auftreten) oder Monosomien (nur ein Chromosom).

Werden gar ganze Chromosomensätze vermehrt, wird das Ergebnis als Polyploidie bezeichnet. Zwei Sub-Typen werden unterschieden: bei der Autopolyploidie existieren mehrere Sätze des exakt gleichen Genoms, während es bei Allopolyploidie zur Verschmelzung des Erbmaterials nah verwandter Arten kommt. Polyploidie ist im Pflanzenreich häufig anzutreffen. Wendel schätzt, dass 40–70% aller Land-Pflanzen einen polyploiden Vorfahren haben (Wendel 2000).

Bei wichtigen Kulturpflanzen kann Polyploidisierung offenbar einen positiven Effekt auf den Ertrag haben (Paterson et al. 2000). Im Vergleich der beiden Subtypen scheint besonders die Allopolyploidie eine evolutionär erfolgreiche Strategie zu sein (Liu und Wendel 2003). Duplikationen (lokale und segmenta-

le), zusätzliche Chromosomen und erst Recht Polyploidie führen zum Auftreten von Paralogen, zum Teil in erheblichem Ausmaß.

Zur Rolle duplizierter Gene bei der Evolution neuer biologischer Merkmale gibt es konkurrierende Hypothesen. Das klassische Modell nach Ohno (Ohno 1970) ging davon aus, dass nach Verdopplung eines Gens die entsprechende Funktion redundant vorhanden ist. Folglich sei eine der beiden Kopien von jeglichem Selektionsdruck befreit und werde schnell Punktmutationen ansammeln, die sich bei nur einfachem Vorhandensein des Gens schnell schädlich ausgewirkt hätten. In den weitaus meisten Fällen würde dies dazu führen, dass sich das Duplikat zum Pseudogen entwickelt und in der Folge eventuell sogar komplett verloren geht, selten jedoch könne auf diesem Weg eine neue Funktion hervor gehen (Prince und Pickett 2002).

Die große Anzahl an Gen-Familien, die man in vielen Genomen gefunden hat, widerspricht jedoch Ohnos Redundanzmodell, denn die Anzahl an funktionell erhaltenen Gen-Duplikaten ist sehr viel höher als vorhergesagt (Lynch und Conery 2000). Ein alternatives Szenario, das sog. DDC-Modell (für Duplication-Degeneration-Complementation) geht hingegen davon aus, dass nach einer Duplikation beide Exemplare degenerative Mutationen ansammeln. Diese verhalten sich jedoch komplementär, in jeder Kopie wird also nur eine abgegrenzte Teilfunktion gestört. Die ursprüngliche Gesamtfunktion bleibt erhalten, wird aber nun von zwei Genen ausgefüllt, die somit beide essentiell sind und im Genom verbleiben (Force et al. 1999), (Lynch und Force 2000). DDC kommt demnach ausschließlich für Gene in Betracht, deren Architektur hinreichend modular gestaltet ist.

Die beschriebene Sub-Funktionalisierung kann sich auf Regulations-Ebene abspielen. Voraussetzung hierfür ist, dass die Duplikation auch cis-regulatorische Elemente umfasst. Anschließende Mutationen in unterschiedlichen Elementen führen zu veränderter Transkriptionsregulation der beiden Kopien. Die Aufgabenteilung der beiden modifizierten Gene kann sich also etwa auf unterschiedliche Gewebe oder Entwicklungsstadien beziehen (Force et al. 1999). In Hefe konnte gezeigt werden, dass sich bei einem Großteil duplizierter Gene das Expressionsmuster sehr bald nach dem Duplikationsereignis divergent entwickelt (Gu et al. 2002).

Lynch und Conery weisen darauf hin, dass Duplikationen aufgrund von Polyploidie möglicherweise per se größere Überlebens-Chancen besitzen, weil hier die Balance der Gen-Produkte erhalten bleibt, wenn auch auf einem erhöhten Niveau. Lokale Duplikationen hingegen stören dieses Gleichgewicht (Lynch und Conery 2000). Kondrashov et al. argumentieren, dass das Schicksal einer Duplikation sehr früh entschieden wird, nur solche Ereignisse werden fixiert, die sofort einen selektiven Vorteil für die Spezies bringen (Kondrashov et al. 2002). Die Auswahl konservierter Gen-Verdopplungen sei daher stark von der Funktion abhängig. Favorisiert seien solche Gene, bei denen eine Dosis-Erhöhung des entsprechenden Produktes positive Effekte haben.

2.2.2 Segmentale Duplikationen in Eukaryonten-Genomen

Das erste vollständig sequenzierte Genom in dem umfangreiche segmentale Duplikationen gefunden wurden, war das der Bäcker-Hefe *Saccharomyces cerevisiae* (Goffeau et al. 1996). Die Anzahl der identifizierten Blöcke variiert mit der angewendeten Methode. Berichtet wurden 53 (Mewes et al. 1997), 46 (Coissac et al. 1997), 55 (Wolfe und Shields 1997) und 39 (Friedman und Hughes 2001).

Die Frage, ob diese Segmente auf eine Verdopplung des gesamten Genoms durch Polyploidisierung oder auf einer Folge unabhängiger Duplikations-Ereignisse beruhen, wurde lange kontrovers diskutiert. Wolfes Argumente für eine Komplett-Verdopplung sind im wesentlichen die Konservierung der Orientierung der Blöcke bezüglich ihrer Lage zum Centromer sowie die Abwesenheit von Triplets, also Genen, die dreifach vorhanden sind (Wolfe und Shields 1997). Das vorgeschlagene Szenario konnte man außerdem mit Simulationen stützen (Seoighe und Wolfe 1998). Friedman und Hughes bestätigten die Hypothese in soweit, als dass sie zumindest 28 der von ihnen gefundenen 39 duplizierten Segmente auf ein gemeinsames (sehr hohes) Alter datierten. Allerdings schlossen sie auch die Möglichkeit von Einzel-Ereignissen nicht aus (Friedman und Hughes 2001). Llorente et. al. verglichen das Redundanz-Niveau in 13 weiteren Hefe-Arten mit *S. cerevisiae*, ihre Ergebnisse lassen eher auf eine Vielzahl von Duplikationen schließen (Llorente et al. 2000).

Eine Veröffentlichung aus dem Jahre 2004 liefert jedoch starke Indizien für die Polyploidie-These. Nach Sequenzierung einer relativ nah verwandten Art, *Kluyveromyces waltii* konnten 75% des Genoms auf das *S. cerevisiae*-Genom abgebildet werden. Das besondere daran: jeder Abschnitt S_K aus *K. waltii* hat zwei Entsprechungen im Referenz-Genom: S_S^1 und S_S^2, deren Gen-Gehalt jedoch zumeist komplementär ist. Das heißt, ein Gen aus S_K kommt entweder nur in S_S^1 oder in S_S^2, selten aber in beiden Segmenten des Hefe-Genoms vor. Dies lässt sich am einfachsten damit erklären, dass nach der Aufspaltung der beiden Arten das Genom in *S. cerevisiae* komplett verdoppelt wurde. Anschließend gingen die meisten Duplikate verloren, und zwar ohne Präferenz für bestimmte Bereiche des Chromosoms. Aus einem duplizierten Segment $S_S^1 = ABCDEF, S_S^2 = A'B'C'D'E'F'$ könnte auf diese Weise nach und nach zwei kaum noch verwandte Segmente $ACDF$ und $B'D'E'$ hervorgegangen sein. Erst durch Bezug zum entsprechenden Block in *K. waltii*, in dem noch alle Gene erhalten sind, wird der Ursprung wieder erkennbar (Kellis et al. 2004).

In den ebenfalls sequenzierten Eukaryonten-Genomen von *Drosophila melanogaster* (Rubin et al. 2000) und *Caenorhabditis elegans* wurden Duplikationen in sehr viel geringerem Ausmaß gefunden. Für den Fadenwurm werden einige kleinere Blöcke berichtet, die allerdings nicht als Hinweis auf eine poloyploide Phase in der Geschichte des Genoms gedeutet werden (Friedman und Hughes 2001). Im humanen Genom existieren ebenfalls segmentale Duplikationen mit Längen von 1–200 Kb, die etwa 5% des Gesamtgenoms ausmachen. Auch hier wird Polyploidie als Ursache ausgeschlossen. Davon unberührt gibt es eine kontrovers geführte Diskussion, ob in der Historie der Wirbeltier-Evolution eine oder sogar

zwei Genom-Verdopplungen (1R- bzw. 2R-Hypothese) stattgefunden hat (siehe etwa pro: (Larhammar et al. 2002), (Panopoulou et al. 2003), kontra: (Hughes et al. 2001), (Makalowski 2001)).

2.2.3 In-Silico-Methoden zum Nachweis segmentaler Duplikationen

Prinzipiell lassen sich große Abschnitte identischer oder sehr ähnlicher Sequenzen durch das Alignment vollständiger Chromosomen detektieren. Ein Problem beim Nachweis segmentaler Duplikationen liegt jedoch darin, dass die aus der Verdoppelung hervorgegangenen Segmente durch nachfolgende Mutationen mehr und mehr fragmentiert werden. Besonders einschneidende Veränderungen bringt eine Rückkehr zur ursprünglichen Satzzahl mit sich, wenn sich also ein ehemals diploides Genom nach Verdopplung (tetraploide Phase) wieder zum zweifachen Chromosomensatz entwickelt. Die genauen Mechanismen dieser Diploidisierung sind weitgehend ungeklärt (Wolfe 2001).

Hinzu kommt eine Reihe kleinerer Veränderungen. Die Feinstruktur duplizierter Bereiche unterliegt einer hohen Dynamik (Bennetzen 2000; Bancroft 2001). Eine Kombination aus Gen-Verlusten, Deletionen kleineren Maßstabs, Einfügen neuer DNA oder Translokationen zu anderen Stellen im Genom zerstört die Kontinuität des verdoppelten Materials. Zu berücksichtigen ist auch, dass bei segmentalen Duplikationen sowohl funktionelle (Exons, regulatorische Elemente) wie auch nicht funktionelle Abschnitte verdoppelt werden. Letztere unterliegen keinem Selektionsdruck, die Veränderungsrate ist somit nicht einheitlich für ein Segment. Zur Abschätzung relativer Häufigkeiten von Insertionen, Deletionen und Nukleotid-Substitutionen in nicht codierenden Abschnitten gibt es neuere Untersuchungen (Zhang und Gerstein 2003; Denver et al. 2004), die jedoch zu teilweise widersprüchlichen Ergebnissen kommen.

Alles in allem muss man also damit rechnen, nur noch mosaikartige Spuren selbst größter Duplikationsereignisse vorzufinden. In einem Alignment zweier Chromosomen müssten diese Spuren als hoch konservierte Abschnitte sichtbar werden. Genaue Lage und Ausdehnung dieser Bereiche sind aber nicht bekannt. Der Nachweis segmentaler Duplikationen erfordert daher zunächst die Identifizierung homologer Abschnitte auf den einzelnen Chromosomen.

Der Vergleich von Sequenzen mit dem Ziel, ähnliche Teilabschnitte zu finden, ist eines der klassischen Probleme in der Bioinformatik. In diesem Bereich haben sich zwei Heuristiken etabliert: BLAST (Altschul et al. 1990) und FASTA (Pearson und Lipman 1988). Sie unterscheiden sich in der Art, wie sie die Anfrage-Sequenz vorverarbeiten und wie Treffer-Sequenzen in der Datenbank lokalisiert werden. Beiden gemein ist jedoch, dass sie in Hinblick auf das Filtern von Proteinen entwickelt wurde und somit auf entsprechende Sequenzlängen hin ausgelegt sind. Um segmentale Duplikationen zu finden, ist es jedoch erforderlich, Sequenzen in Größenordnungen ganzer Chromosomen, im Falle von *A. thaliana* also im Bereich von 10^7 Basen zu vergleichen. Eine einfache BLAST-Suche kommt hierzu also nicht in Frage.

Zum Alignment im Genom-Maßstab sind daher sind andere Herangehensweisen erforderlich. Eine Möglichkeit ist die Verwendung von sog. Suffix-Trees. Dabei handelt es sich um eine baumartige Datenstruktur zur Repräsentation von Sequenzen. Jeder Teilabschnitt einer Sequenz, der sich bis zum letzten Zeichen erstreckt, ist ein Suffix dieser Sequenz (incl. der Gesamt-Sequenz, eine Sequenz der Länge n hat also auch n Suffixe mit einer Länge > 0). Der Suffix-Tree ist so konstruiert, dass alle Suffixe darin enthalten sind. Interne Knoten des Baumes repräsentieren mehrfach vorkommende Teilsequenzen, Blattknoten dagegen einmalig vorkommende Vervollständigungen zu Suffixen. Somit kann jedes Suffix durch Verfolgen eines Pfades von der Wurzel des zu einem Blatt rekonstruiert werden.

MUMmer (Delcher et al. 1999) benutzt Suffix-Trees, um sog. MUMs (für maximal unique matching strings) in zwei Sequenzen (Genomen) zu finden. Das sind Abschnitte, die in jedem Genom genau einmal auftreten, deren benachbarte Zeichen in den beiden Genomen aber nicht mehr gleich sind. Im nächsten Schritt werden die gefundenen MUMs gefiltert, so dass alle verbleibenden eine kollineare (nicht überkreuzende) Anordnung aufweisen. Dieses geschieht mit einem Standard-Algorithmus zur Bestimmung der längsten aufsteigenden Sequenz von Integer-Werten (LIS, siehe Kapitel 2.3.1). Damit ergibt sich gewissermaßen ein Gerüst für das Alignment, in dem jeder MUM eine Stützstelle darstellt. Die verbleibenden Lücken werden in einem weiteren Schritt behandelt, das Ergebnis schließlich mit einer Implementierung des Smith-Waterman-Algorithmus (Smith und Waterman 1981) aligniert.

Ein weiterer auf Suffix-Trees aufbauender Ansatz ist der REPuter (Kurtz und Schleiermacher 1999). Auch diese Software ist explizit zur Verarbeitung ganzer Chromosomen konzipiert. Ermöglicht wird dies dadurch, dass Suffix-Trees sehr kompakt gestaltet werden können (die Implementierung des REPuters benötigt beispielsweise 12,5 Bytes pro Zeichen der Eingabe-Sequenz[1]). Außerdem kann er in linearer Zeit konstruiert werden.

Beide Systeme kommen für die Suche nach segmentalen Duplikationen jedoch nicht in Betracht. MUMmer ist eher für Sequenzen mit relativ hoher Ähnlichkeit geeignet. In (Delcher et al. 1999) wird der Algorithmus etwa auf die Genome zweier Stämme von *Mycoplasma tuberculosis* sowie auf Sequenzen aus dem Maus- bzw. Human-Genom, deren relativ hohe Ähnlichkeit bereits bekannt war. Eine entsprechende Ähnlichkeit ist für *Arabidopsis* aber nicht zu erwarten, zumindest nicht über den Bereich ganzer Chromosomen (die Segmente sind ja eben noch nicht bekannt). Der REPuter ist für die Suche nach exakten Wiederholungen und Palindromen gedacht. Von den zu suchenden Segmenten kann aber nicht vorhergesagt werden, in wie weit perfekte Sequenz-Gleichheit noch besteht.

Eine dem Suffix-Tree verwandte Datenstruktur, der sog. HPT hatte seine prinzipielle Eignung zwar anhand des Hefe-Genoms bewiesen (Heumann et al. 1996), stand jedoch für das *Arabidopsis*-Projekt und damit für meine Arbeit aus tech-

[1] Zum Auffinden von Palindromen muss jedoch das Palindrom der Gesamt-Sequenz mit eingegeben werden, es ergeben sich also 25 Bytes pro Zeichen der Ausgangs-Sequenz.

nischen Gründen nicht zur Verfügung. Seit Beendigung des Projektes sind einige neue Ansätze, insbesondere mit Blick auf das Maus- und Human-Genom, veröffentlicht worden, einen guten Überblick bietet (Ureta-Vidal et al. 2003).

2.3 Eigene Arbeit: Segmentale Duplikationen in *A. thaliana*

2.3.1 Ansatz

Wie in der Einleitung dieses Kapitels erwähnt, war die Zielsetzung meiner Analyse die Bestimmung genomischer Redundanz als Folge segmentaler Duplikationen in *Arabidopsis thaliana*. Dazu wurde eine dreistufige Strategie entwickelt. Im ersten Schritt gilt es, die Homologie-Beziehungen innerhalb des Genoms möglichst vollständig zu erfassen und in geeigneter Form abzulegen. Ausgehend von dieser Vielzahl an Einzel-Signalen werden dann große zusammengehörige Blöcke bestimmt und schließlich in intuitiver Form dargestellt.

Homologie-Analyse

Für den ersten Schritt, also die grundlegende Homologie-Analyse bietet es sich an, auf den BLAST-Algorithmus zurückzugreifen. Die entsprechenden Programme sind frei verfügbar und werden bereits seit mehr als zwanzig Jahren gepflegt und weiterentwickelt. Sie weisen daher ein hohes Maß an Fehler-Freiheit und Geschwindigkeit auf. Allerdings wurde BLAST zur Ermittlung lokaler Ähnlichkeiten innerhalb von Proteinen entwickelt (s.o.), die im Maximal-Fall wenige tausend Aminosäuren, also in aller Regel weniger als 10.000 Basenpaare umfassen. Vorliegend geht es jedoch um vollständige Chromosomen, von denen das kürzeste bereits mehr als 18.000.000 Basenpaare lang ist. Konzeptionsbedingt eignet sich BLAST also nicht, um Homologien in DNA-Sequenzen dieser Dimension zu erfassen.

Um dennoch die Vorteile von BLAST nutzen zu können, habe ich daher ein Verfahren entwickelt, das ausgehend von zwei komplett sequenzierten Chromosomen A und B folgende Schritte umfasst:

1. Zerschneide jedes Chromosom in Fragmente einheitlicher Länge
2. Indiziere die Gesamtheit aller Fragmente eines Chromosoms als BLAST-Datenbank
3. Benutze jedes Fragment aus A als Anfrage-Sequenz gegen die Datenbank für B
4. Speichere die Koordinaten aller Treffer in einer relationalen Datenbank
5. Visualisiere die Ergebnisse in einem geeigneten Plot

Eine Schematische Übersicht zeigt Abbildung 2.1.

Ziel dieser Homologie-Analyse ist es, Signale mit hoher Sensitivität zu sammeln. Die Länge der Fragmente kann daher relativ kurz angesetzt werden, längere homologe Teilsequenzen werden im weiteren Verlauf wieder zusammengesetzt. Allerdings besteht die Gefahr, Signale zu verlieren, wenn sequenzähnliche Abschnitte um Schnittpunkte herum angesiedelt sind. Darum ist es wichtig, dass sich die einzelnen Fragmente überlappen. Die in dieser Arbeit zusammengefassten Ergebnisse wurden mit einer Fragmentlänge von 2500 Basen und einem Überlapp von 500 erzielt.

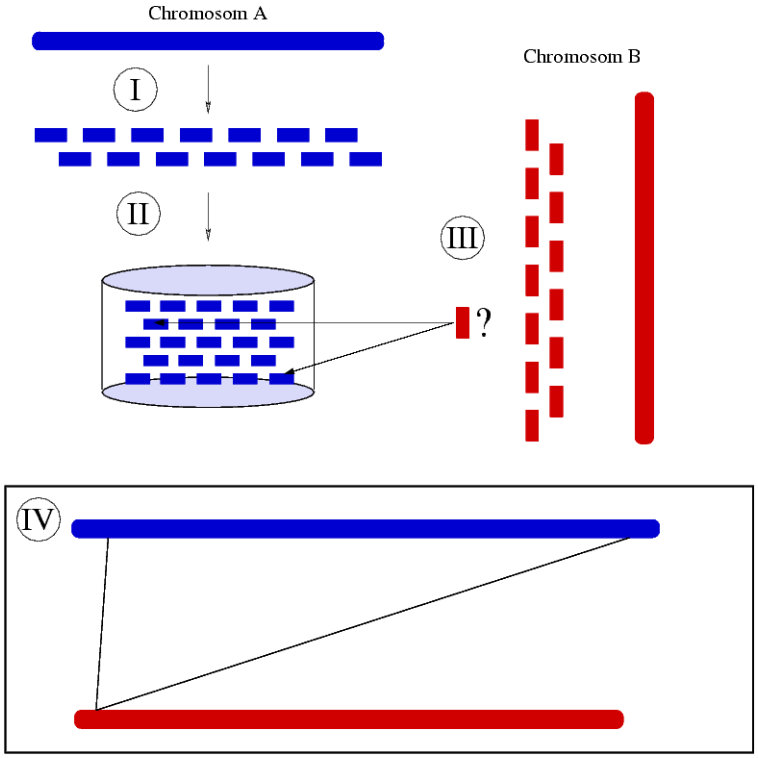


Abbildung 2.1: Schema der Homologie-Analyse: (I) Fragmentierung des Chromosoms in überlappende Abschnitte, (II) Indizierung als BLAST-Datenbank, (III) BLAST-Suche eines Fragments von Chromosom B in der Datenbank zu Chromosom A, (IV) Auftragen der gefundenen Treffer im Plot

Die einzelnen Fragmente erhalten fortlaufende Nummern, so dass sich die Start-Koordinate eines Fragments innerhalb des Chromosoms berechnen lässt. Die exakten Grenzen eines bestimmten Treffers ergeben sich dann als Summe des jeweiligen Fragment-Starts (Offset) und der von BLAST ausgegebenen Hit-Koordinate. Jeder Treffer wird in der Ergebnis-Datenbank gespeichert, wobei ein Datensatz die Koordinaten beider beteiligten Teilsequenzen und die Kennzahlen der Treffer-Qualität wie Score, E-Value, Prozent Identitäten etc. enthält.

Die Einführung von Überlappungen zwischen den Fragmenten führt zu künstlichen Redundanzen in den BLAST-Ergebnissen. Ein Treffer, der im Überlapp-Bereich liegt, kann grundsätzlich für beide angrenzenden Fragmente angezeigt

werden. Die Umrechnung in globale Koordinaten ermöglicht es jedoch, solche Dubletten zu filtern. Ein Treffer, der vollständig in einem anderen bereits gespeicherten liegt, wird nicht nochmals eingetragen.

Das Verfahren eignet sich auch zur Berechnung intra-chromosomaler Homologie. In diesem Fall sind die Fragment-Mengen, die als Datenbank und als Anfrage-Ressource dienen, identisch. Zur Vereinfachung werden hier jedoch nur Treffer gespeichert, die in Fragmenten mit höherer Ordnungsnummer gefunden werden. Der Treffer im identischen Fragment wird ignoriert, für die übrigen wird eine Symmetrie-Vermutung eingeführt (d.h. wenn ein Fragment a_i ein Fragment a_{i-n} trifft, wird vermutet, dass dieser Treffer bereits in der umgekehrten Richtung bei der Anfrage von a_{i-n} gefunden wurde). Diese idealisierende Vermutung ist beim BLAST-Algorithmus zwar nicht immer zutreffend, hat bei der hier vorliegenden Anwendung aber keine Auswirkungen: die entstehenden Ungenauigkeiten sind so gering, dass sie das weitere Verfahren nicht beeinflussen.

Das Ergebnis der Suche nach den Homologie-Signalen lässt sich durch einen einfachen Plot visualisieren. Dabei dienen zwei waagerechte Balken zur symbolischen Repräsentation der untersuchten Chromosomen. Jeder Treffer wird als Linie zwischen den Start-Koordinaten der beteiligten Teil-Sequenzen dargestellt (die Länge des alignierten Bereichs spielt dabei keine Rolle, da selbst ein Treffer über ein komplettes Fragment kürzer als ein Pixel wäre). Abbildung 2.2 zeigt einen solchen Plot für die intra-chromosmale Analyse von Chromosom 4. Aufgrund der Symmetrie-Vermutung (s.o.) sind treten nur Diagonalen in einer Richtung auf[2]. Die Interpretation des Plots folgt im Ergebnis-Teil im Abschnitt 2.3.2.

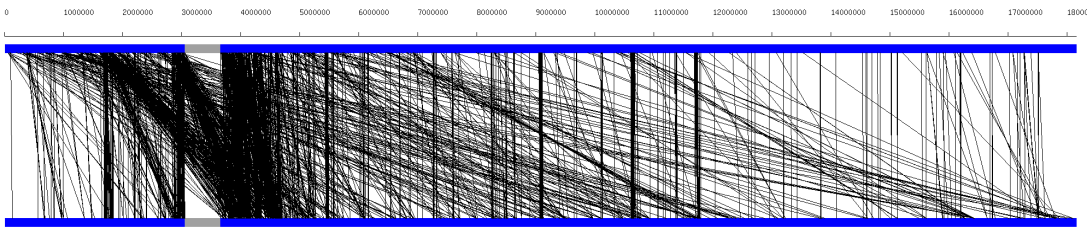


Abbildung 2.2: Plot der Homologie-Signale innerhalb von Chromosom 4

Filtern der Homologie-Signale

Bei der Ermittlung der Homologie-Signale werden nur sehr geringe Anforderungen an die Treffer-Qualität gestellt. Die Kenndaten aller BLAST-Treffer (Koordinaten und Informationen zur Treffer-Güte) werden gesammelt und in

[2]Senkrechte Linien dürfte es demnach auch nicht geben, die Auflösung reicht aber in einigen Fällen nicht aus, um die existierende Differenz in den X-Koordinaten des Anfangs- und Endpunktes darzustellen.

einer Datenbank gespeichert. Das eigentliche Filtern erfolgt erst bei der Cluster-Ermittlung (siehe nächster Abschnitt).

Außer Qualitäts-Kriterien ist auch die Berücksichtigung der Annotation möglich. Dabei handelt es sich um Zuordnung biologisch relevanter Information zu einzelnen Abschnitten eines Chromosoms (formale Definition folgt in Abschnitt 4.3.1, Definition 4.2). Bei der Ermittlung segmentaler Duplikationen spielt insbesondere die Lage von ORFs (resp. von Exons) eine wichtige Rolle, da vor allem funktionale DNA-Segmente konserviert sein sollten.

Ermittlung von Clustern

Plots wie der in Abbildung 2.2 enthalten eine enorme Zahl an Signalen. Es ist möglich, die Linien-Dichte durch Anwendung sehr restriktiver Filter-Kriterien zu verringern. Dabei besteht die Gefahr, wichtigen Informationsgehalt zu verlieren. Gefragt ist daher ein Verfahren, das geeignet ist, die hinsichtlich der Fragestellung relevanten Strukturen hervorzuheben.

Ziel der Analyse ist das Aufdecken großer Duplikationen während der Entwicklung des Genoms. Ein solches Ereignis führt dazu, dass eine große Anzahl aufeinander folgender Fragmente eines Chromosoms hohe Homologie zeigt zu Fragmenten, die auf einem anderen Chromosom ebenfalls benachbart liegen (konservierte Nachbarschaft). Im Plot äußert sich dies als Menge dicht beieinander liegender paralleler, bei invertierter Lage anti-paralleler Linien. Gesucht sind also Teilmengen (Cluster) von Hits, die (1) einen gewissen Maximalabstand nicht überschreiten (Abstandskriterium) und die (2) parallele oder anti-parallele Formation aufweisen (Kollinearitätskriterium).

Zur Beschreibung des Verfahrens ist zunächst die Definition einiger grundlegender Begriffe und Notationen erforderlich.

Definition 2.1 (Teil-String)
Ein Chromosomen sei definiert als String über dem Nukleinsäure-Alphabet $\mathcal{N}$, das Symbole für das Auftreten von Adenin, Guanin, Cytosin und Thymin enthält. Ein Chromosom A mit der Gesamtlänge n kann notiert werden als $A = a_1 a_2 \cdots a_n \in \mathcal{N}^n$. Ein beliebiger Teil $a_{i+1} \cdots a_j$, $j \geq i$ dieses Chromosoms wird als $i{:}a{:}j$ geschrieben. Seine Länge berechnet sich dann als $|i{:}a{:}j| = (j - i)$. Darin heißt $(i+1)$ Startkoordinate und j Stoppkoordinate. Der Abstand zweier Teil-Strings auf einem Chromosom $i{:}a{:}j$, $r{:}a{:}s$ mit $0 \leq i \leq j \leq r \leq s$ ist definiert als Anzahl der Symbole zwischen den Teil-Strings und beträgt $(r - s)$.

Definition 2.2 (Hilfsfunktionen *init* und *stop*)
Gegeben sei ein Teil-String $i{:}a{:}j$ gemäß Definition 2.1. Zwei Funktionen *init* und *stop* seien so definiert, dass sie die Teil-Strings auf ihre Grenzen abbilden:

$$init(i{:}a{:}j) = i$$

$$stop(i{:}a{:}j) = j$$

Definition 2.3 (Treffer (Hit), Instanz)
Gegeben sei eine Ähnlichkeitsfunktion $f : \mathcal{N}^* \times \mathcal{N}^* \mapsto \mathbb{R}^+$. Ein Treffer oder auch Hit ist definiert als Paar zweier beliebiger Strings $(i{:}a{:}j, r{:}b{:}s)$, so dass $f(i{:}a{:}j, r{:}b{:}s) > t$, wobei ein Schwellwert t festgelegt sei. Die beiden Strings des Treffers werden in den folgenden Definitionen auch als Instanzen bezeichnet.

Definition 2.4 (Hilfsfunktionen *first* und *second*)
Gegeben sei ein Treffer gemäß Definition 2.3 $h = (i{:}a{:}j, r{:}b{:}s)$. Seien zwei Funktionen *first* und *second* so definiert, dass sie den Treffer auf ihre erste bzw. zweite Instanz abbilden:

$$first(i{:}a{:}j, r{:}b{:}s) = i{:}a{:}j$$

$$second(i{:}a{:}j, r{:}b{:}s) = r{:}b{:}s$$

Definition 2.5 (Halbordnung bezüglich einer Instanz)
Eine Menge von Treffern kann bezüglich einer Instanz geordnet werden, sofern sie Teil-Strings des gleichen Strings sind. Seien h_1, h_2 zwei Treffer gem. Def. 2.3 mit $first(h_1) = i{:}a{:}j$ und $first(h_2) = k{:}a{:}l$. Dann gilt:

$$h_1 \leq^1 h_2 \Leftrightarrow \begin{cases} i < k & \text{falls } i \neq k \\ j \leq l & \text{sonst} \end{cases}$$

Analog die Halbordnung nach der zweiten Instanz.

Definition 2.6 (Partition)
Gegeben sei eine Treffermenge $H = \{h_1, h_2, \ldots, h_n\}$ sowie ein Schrankenwert $maxhole \in \mathbb{N}$. Eine Teilmenge $P = \{h_k, \ldots, h_l\} \subseteq H$ heißt Partition falls gilt:

1. Ist P geordnet nach der ersten Instanz, gilt für alle Paare (h_i, h_{i+1}):

$$init(first(h_{i+1})) - stop(first(h_i)) \leq maxhole$$

2. Ist P geordnet nach der zweiten Instanz, gilt für alle Paare (h_i, h_{i+1}):

$$init(second(h_{i+1})) - stop(second(h_i)) \leq maxhole$$

Eine Partition ist also eine Teilmenge von Treffern, in der der Abstand zweier benachbarter Hits durch einen bestimmten Wert beschränkt ist, unabhängig davon, nach welcher Instanz die Menge geordnet ist.

Die Menge aller Partitionen für ein gegebenes Chromosomenpaar kann durch einen rekursiven Prozess ermittelt werden. Die Ausgangsmenge wird zunächst nach der ersten Instanz geordnet. Treffer werden so lange zu einer Teilmenge zusammengefasst, bis durch Zufügen des nächsten Hits der maximale Abstand überschritten wird. Jede dieser Teilmengen wird dann nach der zweiten Instanz geordnet und in gleicher Weise weiter aufgesplittet. Erst wenn eine Menge bezüglich beider Instanzen stabil bleibt, stoppt die Rekursion und eine gültige Partition ist gefunden.

Eine Partition erfüllt somit den ersten Teil der Forderungen an einen Cluster (s.o.). Zum Herausfiltern der kollinearen Hits dient ein Standard-Algorithmus (Gusfield 1997) zur Ermittlung der *longest increasing subsequence* (längsten aufsteigenden Subsequenz), kurz LIS, in einer Liste von Integer-Werten. Im Unterschied zum Begriff der Teil-Sequenz nach Definition 2.1 müssen die einzelnen Elemente einer solchen LIS in der Ausgangsliste nicht aufeinanderfolgend sein. Beispiel: in einer Liste $(1,8,3,5,7,6,10)$ sind $(1,8)$, $(3,7,10)$ oder $(1,6,10)$ aufsteigende Subsequenzen. Eine Längste Aufsteigende Subsequenz (LIS) ist $(1,3,5,7,10)$[3].

Innerhalb einer Partition werden alle Treffer nach Startkoordinate auf dem Ausgangs-Chromosom geordnet und ihnen entsprechende Ordnungsnummern zugeteilt. Aus der Reihenfolge des Auftretens der Treffer auf dem Ziel-Chromosom ergibt sich dann eine geordnete Liste dieser Ordinalzahlen. Die LIS dieser Liste repräsentiert somit den maximalen Anteil nicht überschneidender (paralleler) Treffer. Eine vergleichbare Verwendung des LIS-Algorithmus findet sich in MUMmer (Delcher et al. 1999), einem Werkzeug zum Alignment langer DNA-Sequenzen bis hin zu ganzen Genomen, dessen Einsatz jedoch hohe Ähnlichkeit der zu vergleichenden Sequenzen voraussetzt.

Visualisierung

Die Darstellung der kollinearen Cluster erfolgt in sehr ähnlicher Weise wie die der Homologie-Signale (vgl. Abbildung 2.2). Allerdings werden die quasi dimensionslosen Linien durch Trapeze ersetzt, die die Ausdehnung der Cluster wiedergeben. Im Falle anti-paralleler Formation der Treffer kommen Sechsecke zum Einsatz, eine Form die entsteht, wenn die Eckpunkte des Trapezes nicht mit dem Linien-Zug ABCDA, sondern ABDCA verbunden werden[4].

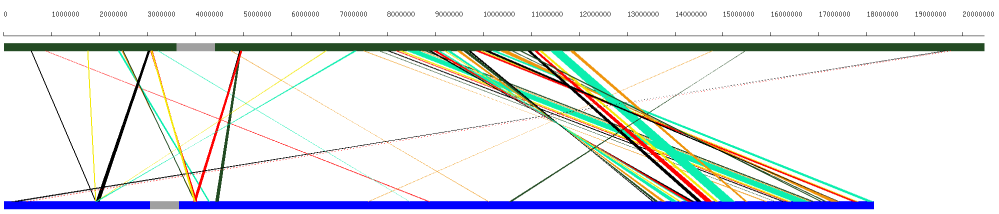


Abbildung 2.3: Ergebnis der Cluster-Analyse zwischen den Chromosom 2 (grün) und 4 (blau). Die grauen Abschnitte innerhalb der Chromosomen stehen für die (nicht sequenzierten) Centromer-Regionen.

Abbildung 2.3 zeigt die kollinearen Cluster zwischen den Chromosomen 2 und 4. Zur besseren Unterscheidbarkeit einzelner Cluster werden sie unterschiedlich

[3] Auch $(1,3,5,6,10)$ ist ein LIS, es muss also keine eindeutige Lösung geben

[4] Damit wird eine Verdrillung des Trapezes angedeutet, um die invertierte Lage des Clusters zu verdeutlichen

eingefärbt. Die Färbung erfolgt zufällig, ein bestimmter Farb-Code wird nicht verwendet. Auf diese Weise fällt es leichter, größere Strukturen in der Cluster-Menge zu erkennen.

2.3.2 Ergebnisse

Segmentale Duplikationen

Die Diagramme, die sich mit dem beschriebenen Verfahren ergeben, zeigen im Vergleich zu den ungeclusterten Homologie-Plots wie dem in Abb. 2.2 ein erheblich ausgedünntes Gesamtbild. Dennoch beherrschen nach wie vor Signale mit relativ begrenzter Ausdehnung das Bild. Cluster, die mehrere hundert Kilobasen umfassende segmentale Duplikationen repräsentieren könnten, sind kaum enthalten. Allerdings sind etwa in Abb. 2.3 außer einigen singulären Signalen zwei auffällige Formationen mehrerer Cluster zu erkennen.

- Cluster, deren Ausgangspunkte auf einem Chromosom verteilt sind, im Zielchromosom jedoch in einem Punkt zusammenlaufen.
- Größere Gruppen von benachbarten Clustern, die einen gleichartigen Verlauf zeigen.

Ein Beispiel für strahlenförmig zusammenlaufende Cluster ist in Abb. 2.3 die Region um 4 Megabasen auf Chromosom 4. Eine genauere Analyse dieses Abschnittes (The European Union Arabidopsis Genome Sequencing Consortium et al. 1999) zeigte, dass hier eine große Anzahl von Transposon-Elementen konzentriert ist. Solche Transposon-Hot-Spots treten in allen fünf Chromosomen auf, bevorzugt im peri-centromerischen Bereich. Folglich kommt es zwischen diesen Bereichen häufig zu einer großen Anzahl an BLAST-Signalen, die sich dann in Clustern der genannten Art widerspiegeln.

Weitaus interessanter im Hinblick auf die Suche nach segmentalen Duplikationen sind aber die Formationen, die zwischen ca. 7,5 Mb und 12 Mb auf Chromosom 2 bzw. zwischen etwa 13,5 Mb und 18 Mb auf Chromosom 4 zu beobachten sind. Hier lassen sich größere Strukturen ausmachen, die aus mehreren einzelnen Clustern bestehen, und als Meta-Cluster (also Cluster von Clustern) bezeichnet werden können. Besonders auffällig ist, dass parallele Cluster (viereckige Bänder) auch parallele Meta-Formationen bilden, anti-parallel ausgerichtete Cluster (verdrillte Bänder) sich jedoch kreuzen, der Meta-Cluster somit gleichfalls eine anti-parallele Form bildet.

Abb. 2.4 zeigt den betroffenen Ausschnitt und verdeutlicht mit farbigen halbtransparenten Polygonen die Zusammenfassung zu Meta-Clustern zwischen den Chromosomen 2 und 4. Benennt man die markierten Blöcke nach der Reihenfolge ihres Auftretens in Chromosom 2 $ABCDE$, ergibt sich auf Chromosom 4 die Formation $(-B)EA(-D)C$, wobei die negierte Form invertierte Lage anzeigt. Auffällig, dass sich trotz Umordnung und teilweiser Invertierung der Segmente auf beiden beteiligten Chromosomen ein fast lückenlos zusammenhängender Bereich von etwa 4,5 Mb ergibt.

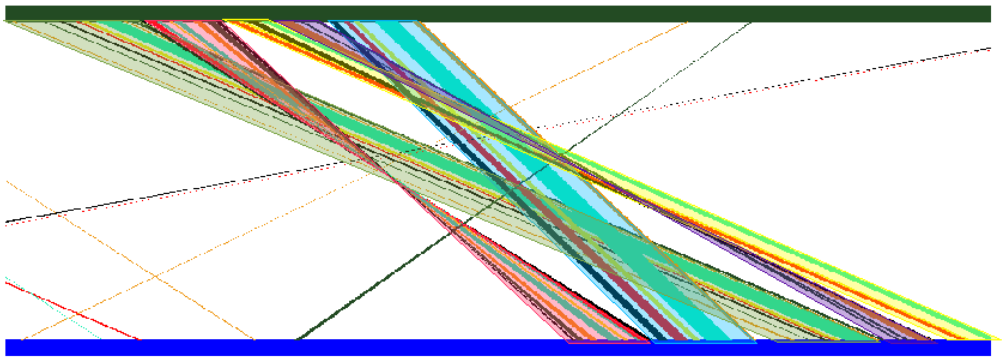


Abbildung 2.4: Ausschnitt aus dem Ergebnis der Cluster-Analyse zwischen Chromosom 2 (oben) und 4 (unten) wie in Abb. 2.3. Halbtransparente Flächen (grün, rot, gelb, violett, blau in der Reihenfolge auf Chromosom 2) verdeutlichen die Zusammenfassung zu Meta-Clustern.

Meta-Cluster zeigen die Charakteristika, die von Spuren historischer segmentaler Duplikationen zu erwarten sind: über lange Abschnitte des Genoms verteilen sich einzelne Homologie-Signale, die aufgrund ihrer Gleich-Gerichtetheit eindeutig auf ein einzelnes Ereignis zurückzuführen sind. Die Tatsache, dass sie nicht als Cluster, sondern erst als übergeordnete Struktur mehrerer Cluster sichtbar werden, liegt an der Vielzahl an Mutationen kleineren Ausmaßes, die jeweils eine Störung in der Kollinearität der duplizierten Segmente verursachen.

Entsprechende Analysen wurden für alle fünf Chromosomen von *Arabidopsis thaliana* durchgeführt, die Ergebnisse im Rahmen der Publikation zum Abschluss der Sequenzierung des Genoms in (The Arabidopsis Genome Initiative 2000) sowie in (European Union Chromosome 3 Arabidopsis Sequencing Consortium et al. 2000) veröffentlicht. Abb. 2.5 enthält eine Übersicht über alle gefundenen segmentalen Duplikationen. Es zeigt sich, dass alle Chromosomen daran beteiligt sind, wenn auch in unterschiedlichem Maße. Intrachromosomale Duplikationen gibt es in großem Umfang auf Chromosom 1, weniger stark ausgeprägt auch auf den Chromosomen 4 und 5. Zwischen allen möglichen Paaren von Chromosomen mit Ausnahme von (2,5) werden Duplikationen gefunden.

Zu betonen ist, dass Abb. 2.5 nicht das direkte Ergebnis eines Berechnungsprozesses ist, sondern die nachträgliche Interpretation der Ergebnisse der oben beschriebenen Cluster-Analysen. Die Auflösungs-Schärfe bei diesem Vorgehen ist beschränkt (siehe dazu auch Abschnitt 2.4). Daher wurden nur Duplikationen aufgenommen, die zumindest auf einer Seite mindestens 0,5 Mb umfassen. Kleinere Segmente sind nicht berücksichtigt. Dennoch wird bereits hier der hohe Grad an Redundanz deutlich. Alle Duplikationen zusammen umfassen fast 68 Mb und damit knapp 60% des gesamten Genoms. Die Tabelle 2.1 auf Seite 24 enthält die jeweils auf volle 500 Basen gerundeten Koordinaten aller Segmente[5]

[5]Die Daten weichen aufgrund von nachträglichen Sequenz-Revisionen leicht von den in (The Arabidopsis Genome Initiative 2000) veröffentlichten ab.

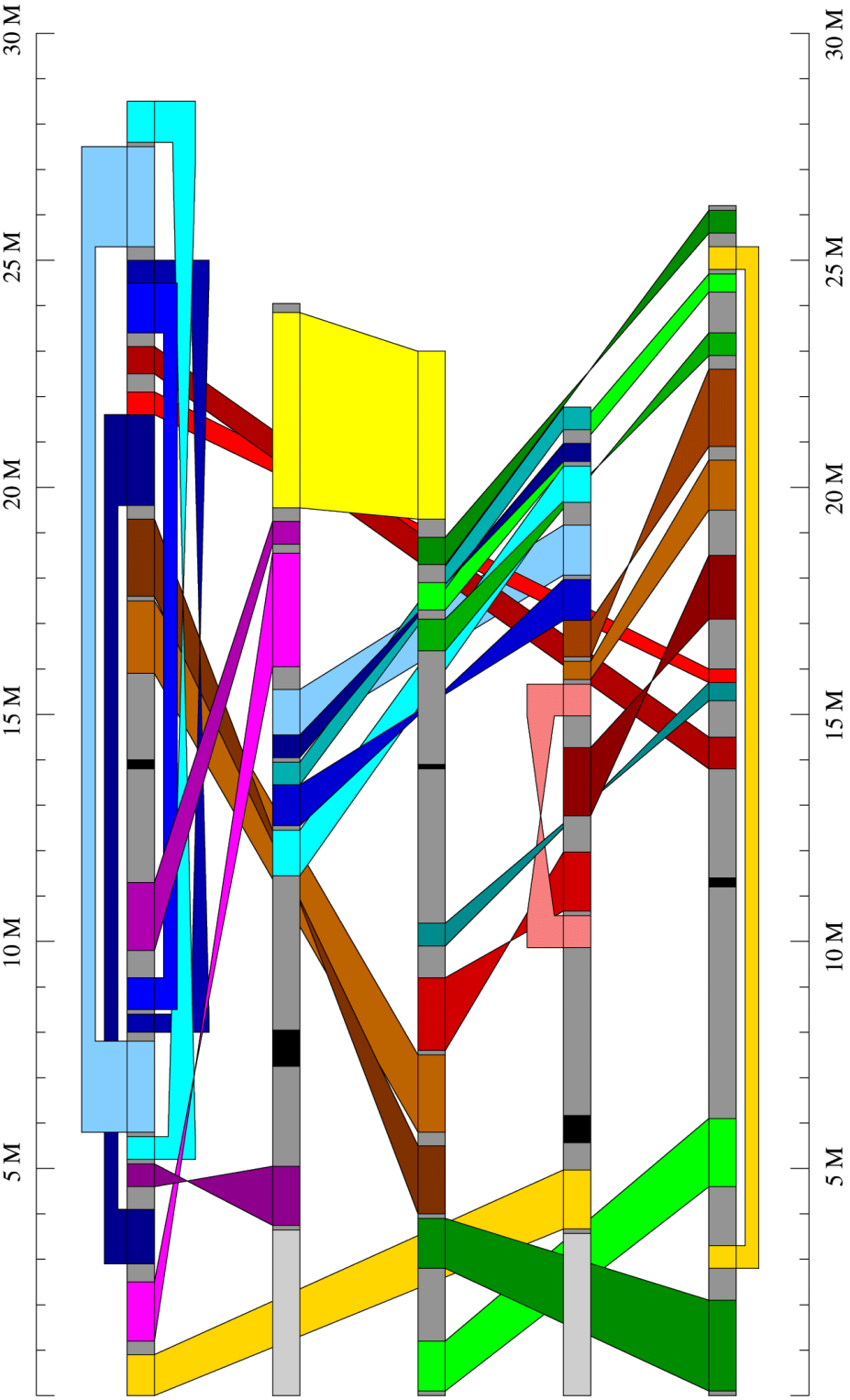


Abbildung 2.5: Segmentale Duplikationen im Genom von *Arabidopsis thaliana*. Horizontale graue Balken repräsentieren die Chromosomen 1 (oben) bis 5 (unten), darin schwarz gekennzeichnet die Centromer-Regionen. Die hellgrauen Segmente am Anfang der Chromosomen 2 und 4 entsprechen rDNA-Clustern (Duplikationen nicht dargestellt). Farbige Bänder verbinden duplizierte Regionen, entgegengesetzte Orientierung wird durch eine Verdrillung des entsprechenden Bandes angedeutet. Maßeinheit der Skala ist Megabasen (Mb=1.000.000 Basen). Aus: (The Arabidopsis Genome Initiative 2000)

Analyse einzelner Segmente

Wie bereits beschrieben, liegen die duplizierten Segmente nicht in Form perfekter Duplikate vor. Sowohl Konservierung als auch Kollinearität schwanken zwischen den einzelnen Segmenten beträchtlich. Die wenigsten Störungen enthält noch die Verdopplung zwischen den Chromosomen 2 und 3 (BC1 in Tabelle 2.1). Wie Abb. 2.6 zeigt, sind die einzelnen Cluster hier nur durch relativ kleine Lücken getrennt, ihre relative Abfolge ist auf beiden Chromosomen gleich. Bis auf wenige Ausnahmen sind alle Cluster parallel angeordnet. 47% der in den Clustern enthaltenen Gene haben einen paralogen Partner im korrespondierenden Segment (BLASTP, E-Value $< 10^{-30}$).

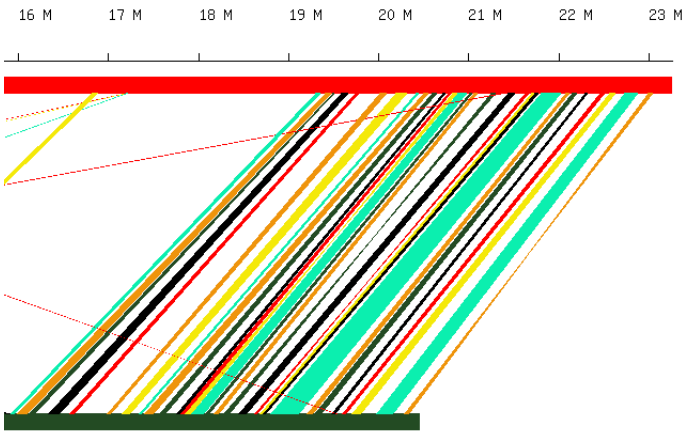


Abbildung 2.6: Die best-erhaltene segmentale Duplikation zwischen den Chromosomen 2 und 3.

In anderen Segmenten kommen auch größere Lücken zwischen den Clustern vor, in den meisten Fällen gibt es jedoch eine eindeutig vorherrschende Ausrichtung. Die Ausnahmen sind Tabelle 2.1 zu entnehmen ('x/p' bzw. 'p/x' in der Spalte 'Form'). Der Anteil an Genen mit hoch konserviertem Gegenstück sinkt auf bis zu 20%.

Ein Beispiel für die Verteilung paraloger Gene in korrespondierenden Clustern gibt Abb. 2.7. Dargestellt ist ein Cluster aus dem Segment BE1. Gezeigt sind die Gene innerhalb eines invertierten Blocks, was sich in jeweils entgegengesetzten Codierungsrichtungen der paralogen Partner niederschlägt.

Eine noch genauere Analyse der Verteilung der homologen Abschnitte ermöglicht das Programm DIALIGN (Morgenstern 1999). Dabei handelt es sich um eine Software, die besonders gut geeignet ist, lokale Homologie-Inseln in Regionen mit nur geringer globaler Ähnlichkeit aufzulösen. Beispielhaft zeigt Abb. 2.8, wie sich die homologen Sequenz-Abschnitte fast ausschließlich auf die codierenden Bereiche beschränken.

Name	ChrA	ChrB	Start A	Stop A	Start B	Stop B	Form
AA1	1	1	3045500	4325500	21019000	23111500	p
AA2	1	1	5533500	5876500	29211500	29703500	x
AA3	1	1	5892000	7907000	26889500	29177500	p
AA4	1	1	7938500	8439000	26191500	26907500	x
AA5	1	1	9204500	9458000	25797500	26088000	p
AB1	1	2	1211000	2406000	12454000	14311500	x
AB2	1	2	4745500	5189500	106000	1051500	x
AB3	1	2	10119000	11009000	14507500	14891500	p
AC1	1	3	17452500	19028000	5870808	7516000	p
AC2	1	3	19100500	20785500	4061553	5598000	x
AD1	1	4	52000	821500	159500	1293500	x/p
AE1	1	5	23207500	23758000	16485000	16786500	p
AE2	1	5	24132000	24744500	14613500	15323500	p
BC1	2	3	15209000	19391500	19733500	23257000	p
BD1	2	4	7201000	8220500	15701000	16543500	p
BD2	2	4	8235500	9091500	13206500	14074500	x
BD3	2	4	9284000	9676500	17393500	17779000	p
BD4	2	4	9753500	10281500	16549500	17055000	x
BD5	2	4	10285500	11299000	14085500	15153000	p
CD1	3	4	7595500	9210500	6795500	8060500	x
CE1	3	5	2000	1074500	4605000	6054500	x/p
CE2	3	5	1380000	1778500	9334500	9916500	x/p
CE3	3	5	2063000	2506500	19535000	20084000	x
CE4	3	5	2700500	3909000	42000	2157500	p
CE5	3	5	10126000	10327500	16149000	16310000	x
CE6	3	5	16948000	17652500	23786500	24279000	x
CE7	3	5	17808500	18292500	25078500	25427000	p
CE8	3	5	18733500	19340000	26298000	26961500	p/x
DD1	4	4	5922000	6446500	11197500	11639500	x
DE1	4	5	8854500	10454000	17831000	19364000	x
DE2	4	5	11956500	12301500	20452500	21353000	p
DE3	4	5	12327000	13123500	21500500	23355000	x
EE1	4	5	2265500	2564500	24552000	24958000	x
EE2	4	5	2780000	3343000	25472386	26255500	p

Tabelle 2.1: Koordinaten und Form ('p' = parallel, 'x' = anti-parallel) der segmentalen Duplikationen in *Arabidopsis thaliana*

Abbildung 2.7: BLASTP Treffer für einen invertierten Cluster zwischen Chromosom 4 (links) und 5 (rechts). Die Gene sind nicht maßstabsgetreu dargestellt, Pfeilrichtungen geben Codierungsrichtung wieder, Annotation gemäß (Schoof et al. 2004). Farb-Code für Treffer-Qualität: gelb: E $< 10^{-10}$, rot: E $< 10^{30}$, schwarz: E $< 10^{60}$

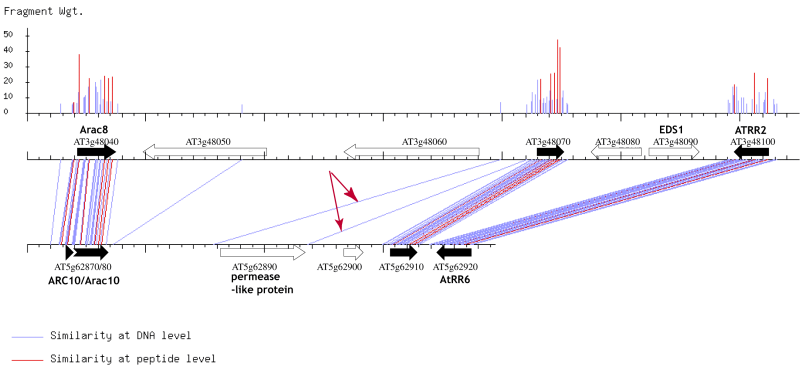


Abbildung 2.8: Verteilung der Homologie-Signale innerhalb eines Clusters mittels DIALIGN. DIALIGN-Ausgabe erweitert um die skizzierte Lage der enthaltenen Gene (schwarz: dupliziert, weiß: nicht dupliziert). Die mit zwei roten Pfeilen gekennzeichneten Signale könnten Spuren des relativ jungen Verlustes eines Paralogs zu At5g62890 aus der dargestellten Region auf Chromosom 3 darstellen.

2.4 Diskussion

2.4.1 Polyploidie oder Multiple Duplikationen?

Meine Analyse konnte zeigen, dass etwa 60% des *Arabidopsis*-Genoms zu duplizierten Segmenten gehören. Bei unabhängigen Ereignissen müsste man davon ausgehen, dass zumindest einige Segmente über größere Bereiche dreifach auftreten. Tatsächlich wurden keine Überlappungen gefunden. Dies ist ein deutliches Indiz dafür, dass es in der Geschichte des Genoms von *A. thaliana* eine tetraploide Phase gegeben hat. Ein weiteres Argument für eine Gesamt-Verdoppelung ergibt sich aus der Anzahl der Chromosomen. Die meisten nah verwandten Arten weisen acht Chromosomen pro Satz (n=8) auf (Koch et al. 2000). Das lässt auf n=4 für den letzten nicht polyploiden Vorläufer schließen. Im Zuge der Diploidisierung wären demnach in *A. thaliana* drei Chromosomen verloren gegangen.

Die Tatsache, dass nicht mal die Hälfte der Gene in den identifizierten Segmenten dupliziert vorliegt, deutet auf eine hohe Anzahl an Deletionen in. Dies ist konsistent mit den gängigen Theorien zum Schicksal duplizierter Gene (Lynch und Conery 2000). Das Ausmaß an Genverlusten ist nur zu erklären, wenn Deletionen ganze Blöcke entfernt haben. Solche Ereignisse konnten experimentell in synthetischen Allopolyploiden nachgewiesen werden. Die Eliminierung großer Sequenzabschnitte beginnt sehr schnell nach der Polyploidisierung (Song et al. 1995; Ozkan et al. 2001).

Der Prozess der Diploidisierung ist gekennzeichnet von einer umfangreichen Reorganisation des Genoms (Soltis und Soltis 1999). Die Topologie etwa der 5 duplizierten Segmente zwischen den Chromosomen 2 und 4 (BD1-BD5, siehe Tabelle 2.1) ist ein klarer Hinweis auf solche Umordnungs-Ereignisse. Dieses Phänomen ist auch auf Ebene der Cluster zu erkennen, wenn in an sich eindeutig parallel ausgerichteten Segmenten einzelne Cluster anti-parallele Formation zeigen.

Eine Abschätzung des Zeitpunktes der Duplikation für die einzelnen Blöcke wurde nicht durchgeführt. Eine gleichmäßige Verteilung wäre hier ein weiterer Anhaltspunkt für eine tetraploide Phase (vgl. nächsten Abschnitt). Einziger Hinweis auf ein unterschiedliches Alter ist die erwähnte Divergenz im Anteil hoch konservierter Gene. Dies kann jedoch auch ein Indiz für eine Allopolyploidie sein, denn bei der Vereinigung zweier, wenn auch nah verwandter Genome weisen einander entsprechende Regionen eine gewisse Divergenz auf, die dem phylogenetischen Abstand der Ursprungs-Genome entspricht (Wolfe 2001).

Über weitere Runden vollständiger Genom-Duplikation kann aufgrund meiner Analyse keine Aussage getroffen werden. Die Anwendung recht strikter Filter-Kriterien, etwa bei der Cluster-Analyse ein Loch-Kriterium von 30.000 Basen sowie strikte Kollinearität auf Cluster- und Segment-Ebene sind vermutlich nicht geeignet, weiter entfernt liegende Ereignisse zu detektieren. Die Tatsache, dass die Segmente gemäß der hier verwendeten Definition (mindestens 0,5 Mb zumindest auf einer Seite und homogene Kollinearität) nicht überlappen, schließt nicht aus, das solche Überschneidungen in kleinerem Maßstab existieren.

2.4.2 Vergleich mit Ergebnissen anderer Autoren

Außer der hier vorgestellten Analyse, deren Ergebnisse in (The European Union Arabidopsis Genome Sequencing Consortium et al. 1999), (The Arabidopsis Genome Initiative 2000) und (European Union Chromosome 3 Arabidopsis Sequencing Consortium et al. 2000) eingeflossen sind, gibt es eine Reihe weiterer Veröffentlichungen anderer Autoren, die sich mit den Duplikationen innerhalb des *Arabidopsis*-Genoms befassen. Die Studie von Paterson et. al. vergleicht die Chromosomen verschiedener Pflanzen-Arten hauptsächlich aufgrund genetischer Karten, enthält aber auch einen auf etwa 90% des Genoms gestützten Überblick der duplizierten Bereiche in *A. thaliana*. Die Identifikation beruht auf DNA-Vergleichen mittels BLASTN (Paterson et al. 2000). Obwohl die vorgestellte Skizze kaum eine konkrete Zuordnung zwischen den einzelnen Segmenten erlaubt, sind – wie zu erwarten – weitgehende Übereinstimmungen festzustellen. Auch die wesentlichen Ergebnisse und Interpretationen decken sich mit denen in (The Arabidopsis Genome Initiative 2000): Ausmaß an Duplikation und fast vollständige Abwesenheit von Triplets deuten auf einen tetraploiden Vorgänger. Ebenfalls auf BLASTN-Analyse innerhalb einer noch unvollständigen Version des Genoms fußt eine Publikation von Blanc et. al. Auch hier zeigt sich große Parallelität zu Ergebnissen und Schlussfolgerungen (eine Duplikation des vollständigen Genoms) meiner Arbeit (Blanc et al. 2000).

Eine auf den ersten Blick dramatisch abweichende Kartierung[6] duplizierter Segmente enthält (Vision et al. 2000). Diese, auf Proteinvergleichen basierende Untersuchung identifiziert 103 duplizierte Blöcke mit mind. sieben paralogen Genen, die sich zum Teil massiv gegenseitig überlagern (bis zu fünf-fach). Die Autoren nehmen auch eine Datierung der einzelnen Segmente vor und leiten daraus vier Duplikationsereignisse großen Umfangs (wahrscheinlich Polyploidisierungen) vor 100, 140, 170 und 200 Millionen Jahren ab. Damit ergänzen sie die frühere Schätzung aufgrund eines Vergleichs von *Arabidopsis* und Tomate, die zwei solcher Ereignisse annahm (Ku et al. 2000).

Eine weitere auf Protein-Vergleich basierte Untersuchung aus dem Jahr 2003 (Blanc et al. 2003) kommt zu weitreichenden Übereinstimmungen mit der Duplikations-Karte von Vision et. al., wendet aber eine andere Methode zur Datierung an. Deren Altersbestimmung war kritisiert worden, weil sie von einer gleichmäßigen Mutationsrate auf Aminosäure-Ebene ausgeht (Wolfe 2001; Sankoff 2001). Blanc et. al. beziehen ihre Altersberechnung hingegen auf den synonymen Austausch von Nukleinsäuren, also solchen, die nicht zur Änderung der Aminosäure-Sequenz führen. Basierend auf diesen Daten errechnen sie zwei eindeutig abgrenzbare Altersklassen unter den Duplikationen.

Zwei weitere Veröffentlichung beschäftigen sich mit den segmentalen Duplikationen in *A. thaliana*. Simillion et al. finden mit ihrer BLASTP basierten Methode 137 Blöcke mit mind. fünf duplizierten Genen. Diese Zahl wird weiter erhöht durch einen Transitions-Schritt: Wenn ein Segment *A* Ähnlichkeiten zu zwei

[6] eine genauere Analyse zeigt, dass sich die Unterschiede auf wenige leicht erklärbare Differenzen beschränken, siehe den Vergleich zur Karte von (Blanc et al. 2003) weiter unten

anderen Segmenten B und C aufweist, diese beiden aber keine offensichtlichen Kollinearität aufweisen, wird auf eine versteckte Block-Duplikation geschlossen. D.h. aus den offensichtlichen Duplikationen ($A \mapsto B$) und ($A \mapsto C$) wird eine weitere Duplikation ($B \mapsto C$) hergeleitet, selbst wenn diese nur wenige oder gar keine duplizierten Gene aufweisen. Insgesamt werden so zusätzliche 53 versteckte Blöcke identifiziert, einige Segmente sind dann bis zu achtfach[7] im Genom vertreten. Aus dieser Zahl schließen die Autoren auf drei vollständige Genom-Duplikationen (ein bestimmtest Gen ist nach der ersten Runde zweifach, nach der zweiten vierfach, nach der dritten achtfach vorhanden, wenn alle Exemplare erhalten bleiben). Diese Hypothese stützen sie durch eine Altersbestimmung der einzelnen Blöcke, die drei differenzierbare Klassen ergibt (Simillion et al. 2002). Allerdings fließen nur 44% der Blöcke und 23% der duplizierten Gene in diese Berechnung ein. Drei Polyploidisierungen ist auch das Ergebnis in (Ziolkowski et al. 2003), die sich im wesentlichen auf eine Fein-Analyse der Segmente AB3 und DE1 stützt.

Große Uneinigkeit besteht bezüglich der Datierung, vor allem der jüngsten Tetraploidie-Phase. Die Schätzungen reichen hier von 24-40 (Blanc et al. 2003) über 65 (Lynch und Conery 2000), 75 (Simillion et al. 2002), 100 (Vision et al. 2000) bis zu 112 (Ku et al. 2000) Millionen Jahren. Bis auf die beiden letztgenannten benutzen alle Autoren prinzipiell die gleiche Herangehensweise, nämlich die Bestimmung der Sättigung an Veränderungen in den synonymen Positionen paraloger Gene. Dieses Verhältnis wird als K_S-Wert angegeben. Offensichtlich besteht hier eine Ungleichheit in der Kalibrierung der Methode, zumal alle Arbeiten von fast gleichen Werten ($K_S = 0,8$) ausgehen.

Eine genauere Analyse der Unterschiede zwischen der in Abb. 2.5 dargestellten Karte und der in (Blanc et al. 2003) zeigt, dass sie keine signifikanten Widersprüche enthalten. Die Topologie der Blöcke meiner Analyse ist in der Blanc-Karte praktisch identisch erhalten. Dies wird besonders deutlich, wenn man in der interaktiven Ansicht auf (Blanc et al.) die älteren Duplikationen ausblendet. Die verbleibenden Unterschiede beschränken sich hauptsächlich auf die genauen Grenzen der einzelnen Blöcke, in einigen Fällen sind Segmente vereinigt (die Blocke AA2 und AA3 gem. Tabelle 2.1 sind z.B. nicht getrennt). Die wenigen zusätzlichen Blöcke weisen entweder eine sehr inhomogene Formation oder große Lücken auf. Gleiches gilt bezüglich der Differenzen meiner Kartierung zur Darstellung in (Vision et al. 2000).

Zusammenfassend lässt sich festhalten, dass meine Untersuchung die Spuren der jüngeren und von allen Autoren bestätigten Polyploidisierung im Genom von *Arabidopsis thaliana* identifizieren konnte. Die Parameter meiner Analyse wurden sehr konservativ gewählt, um die Gefahr falsch positiv vorhergesagter Segmente zu minimieren. Dies scheint insofern gelungen, als dass sich das in Abb. 2.5 dargestellte Segment-Set als eine Art Konsensus-Menge aller diesbezüglicher Studien herausgestellt hat.

Zumindest eine zweite, ältere Duplikationswelle, wahrscheinlich ebenfalls auf Polyploidisierung zurückzuführen, wird von einer Mehrzahl von Autoren an-

[7] Eines sogar neunfach

genommen. Auffällig ist, dass deren Spuren nur mit BLASTP, also Protein basierten Ansätzen gefunden werden. Das Auflösungspotenzial der DNA fokussierten Methoden scheint hierzu nicht auszureichen. Zwar werden in meiner ersten Analysephase Homologie-Signale mit hoher Sensitivität gefunden. Bei der anschließenden, aufgrund der Signal-Dichte notwendigen Filterung wird dann jedoch wertvolle Information ausgeblendet. Protein basierte Methoden kommen aufgrund geringerer Signal-Dichte mit weniger restriktiven Filtern aus und sind somit weniger anfällig dafür, tatsächliche Duplikationen zu übersehen.

Die hier präsentierten Analyse fand breites Interesse und Anerkennung in Form zustimmender Kommentare (Sankoff 2001), Abdruck der Grafik 2.5 (z.B. in (Stein 2001; Waterman et al. in Vorbereitung)) und Verwendung der Segment-Grenzen in weitergehenden Untersuchungen (etwa in (Baumgarten et al. 2003)). Auch die interaktive Version der Karte, programmiert von Martin Ruopp und enthalten auf der CD-ROM-Beilage zu (The Arabidopsis Genome Initiative 2000) sowie bereitgestellt als WWW-Dienst (Ruopp und Haase 2000), wurde als Werkzeug zur *Arabidopsis*-Forschung angenommen (z.B. (Baumbusch et al. 2001; Anderson et al. 2004)).

2.4.3 Weitere Anwendungen

Das zur Suche nach historischen segmentalen Duplikationen in *Arabidopsis* angewendete Verfahren und die hierfür entwickelten Programm-Module konnten auch bei der Analyse weiterer Genome eingesetzt werden. In einer Zusammenarbeit mit Burkhard Morgenstern und Oliver Rinner diente die Software als Präprozessierung für eine Anwendung zur Gen-Vorhersage. Außerdem wurde die Suche nach segmentalen Duplikationen in zwei Chromosomen des Genoms von *Neurospora crassa* durchgeführt.

Neurospora crassa

Die Anwendung des in Abschnitt 2.3.1 vorgestellten Ansatzes ist selbstverständlich nicht auf das Genom von *Arabidopsis thaliana* beschränkt. Ein weiteres Sequenzierungsprojekt, bei dem die entwickelte Software zum Einsatz kam, betrifft *Neurospora crassa*, ein zu den Fadenpilzen gehörender Organismus.

Analysiert wurden die Kopplungsgruppen II und V[8] auf großflächige Homologie-Signale. Im Gegensatz zu *Arabidopsis* konnten jedoch keine Anzeichen für segmentale Duplikationen entdeckt werden. Lediglich vier kleinere, sehr AT-reiche Abschnitte zwischen 4 und 10 Kilobasen Länge liegen dupliziert vor. Keiner der Abschnitte enthält codierende Bereiche; die Sequenz-Identität reicht von 57% bis immerhin 74%. Bedeutung und Duplikations-Mechanismen dieser Abschnitte sind unklar, offenbar handelt es sich nicht um Transposon induzierte Duplikationen. Die Abwesenheit segmentaler Duplikationen ist möglicherweise

[8] In der *N. crassa*-Gemeinde ist es üblich, von Kopplungsgruppen zu sprechen, tatsächlich weicht die Nummerierung der Chromosomen von der der Kopplungsgruppen ab.

auf einen in *N. crassa* aktiven spezifischen Abwehrmechanismus gegen Transposon induzierte Duplikationen (*RIP*) zurückzuführen (Mannhaupt et al. 2003).

Gen-Vorhersage

Die verlässliche Identifizierung von Genen auf Grundlage der chromosomalen DNA-Sequenz (Gen-Vorhersage) ist insbesondere in höheren Eukaryonten ein nicht triviales Problem. Ein exaktes Gen-Modell beinhaltet die richtigen Start- und Stopp-Koordinaten aller beteiligten Exons. Programme zur Gen-Vorhersage neigen zum Teil dazu, zu viele oder zu wenig Exons einzubeziehen oder bestimmen die Exongrenzen ungenau.

Ein Ansatz, der diese Schwierigkeiten zumindest teilweise umgehen kann, liegt im Vergleich genomischer Sequenzen relativ nah verwandter Arten. Wie beim phylogenetischen Footprinting (siehe nächsten Abschnitt) nutzt man hier aus, dass Mutationen in der DNA nicht gleichmäßig angesammelt werden. Veränderungen funktioneller Abschnitte mit negativem Einfluss auf die Funktion werden ausselektiert. Mutationen in Segmenten ohne Funktion haben keine Auswirkungen auf den Phänotyp und unterliegen deshalb keiner Selektion. Im Vergleich homologer DNA-Abschnitte sind die enthaltenen Exons folglich besser konserviert als Introns oder intergenische Sequenzen und auf diese Weise detektierbar.

Erforderlich ist also ein Algorithmus, der in der Lage ist, die Grenzen zwischen gut und schlecht konservierten Abschnitten zweier längerer DNA-Segmente mit hoher Genauigkeit aufzulösen. DIALIGN (Morgenstern 1999) ist hierfür geeignet (siehe auch Abschnitt 2.3.2, Analyse einzelner Segmente). Obwohl es gut mit Sequenzlängen von bis zu einigen hundert Kb umgehen kann, ist es nicht möglich, ganze Chromosomen zu vergleichen.

Um also Kandidaten-Regionen auszuwählen, die dann mit DIALIGN weiterverarbeitet werden können, wird ein Vorverarbeitungsschritt benötigt. Hier bietet sich das in Abschnitt 2.3.1 vorgestellte Verfahren an. Es ist auf vollständige Chromosomen ausgelegt und ermittelt bereits Cluster kollinearer Hits (DIALIGN ist ebenfalls auf Kollinearität der Signale angewiesen).

In (Morgenstern et al. 2002) konnte gezeigt werden, dass der Ansatz, Exongrenzen mit Hilfe von DIALIGN zu bestimmen, grundsätzlich viel versprechend ist. Insbesondere bei der Spezifität der Ergebnisse bezüglich Übereinstimmung mit codierenden Sequenzen hat DIALIGN Vorteile gegenüber anderen Programmen zum Genom-Alignment.

In einer von Burkhard Morgenstern betreuten Diplomarbeit entwickelte Oliver Rinner AGenDA (Rinner und Morgenstern 2001), ein System zur Gen-Vorhersage aufbauend auf dem Alignment von DIALIGN. Dazu wird innerhalb der alignierten Bereiche nach Splice-Signalen gesucht. Mit meinem Verfahren konnte ich die Entwicklung dieses Systems durch Auswahl diverser aussichtsreicher Regionen unterstützen. Dabei wurden nicht nur *Arabidopsis*-interne, sondern auch inter-genomische (*A. thaliana*-Tomate) Cluster zur weiteren Prozessierung mit AGenDA identifiziert.

Verwendung der Daten für Phylogenetisches Footprinting

Der Begriff phylogenetisches Footprinting (Tagle et al. 1988) bezeichnet ein Verfahren zum Auffinden regulatorischer Sequenzen. Dabei macht man sich die Tatsache zu Nutzen, dass sich Teile der DNA die nicht zu genetischen Elementen gehören schneller auseinander entwickeln als solche, die für den Prozess der Protein-Biosynthese von Bedeutung sind. Demnach lassen sich regulatorische Sequenzen als konservierte Teilstücke in den im Allgemeinen kaum konservierten Bereich strangaufwärts orthologer Genpaare detektieren. Üblicherweise werden für dieses Verfahren folglich die Genome zweier unterschiedlicher Organismen herangezogen, wobei deren Verwandtschaftsgrad eine wichtige Rolle spielt. Bei zu nah verwandten Genomen sind die nicht translatierten Bereiche unter Umständen noch nicht ausreichend divergent, bei zu großem phylogenetischen Abstand ist das Konservierungs-Signal nur noch schwer auszumachen (Ureta-Vidal et al. 2003).

Bei segmentalen Duplikationen werden nicht nur die codierenden Bereiche, sondern auch die umgebenden nicht exprimierten Regionen verdoppelt. Unmittelbar nach einem solchen Ereignis muss also das komplette Ensemble aus Gen und zugehörigen regulatorischen Sequenzen dupliziert vorgelegen haben. In der Folge wird es zu einem ähnlichen auseinander Driften der beiden Instanzen kommen wie es bei orthologen Genen in getrennten Spezies der Fall ist. Damit ergibt sich die Möglichkeit, phylogenetisches Footprinting auf die paralogen Genpaare anzuwenden, die aus segmentalen Duplikationen hervorgegangen sind (nach einem Vorschlag von Wolfe auch Ohnologe bezeichnet (Wolfe 2001)).

Bislang sind nur sehr wenige regulatorische Bereiche experimentell charakterisiert (wie etwa RANTES, (Werner et al. 2003)). Die Idee, Ohnologe zur Identifizierung regulatorischer Module zu verwenden, wurde daher in Form einer Diplomarbeit (Hindemitt 2003) weiter verfolgt. Die ermittelten Grenzen der duplizierten Segmente (Tabelle 2.1) bildeten dabei einen Teil der Ausgangsdaten. Die Arbeit konnte zeigen, dass phylogenetisches Footprinting bei Vorliegen segmentaler Duplikationen durchaus auch intra-genomisch angewendet werden kann.

Kapitel 3

Nachbarschaftsbeziehungen auf Gen-Ebene

3.1 Einleitung

In diesem Kapitel verschiebt sich der Fokus der Analysen konservierter Nachbarschaften von der DNA hin zu den Genen. Prinzipiell müssten sich alle Ergebnisse, die man mit Berechnungen auf Gen-Ebene erhalten kann, auch durch DNA gestützte Analysen erzielen lassen; denn letztlich sind Gene 1:1 auf DNA repräsentiert. Beim Vergleich verschiedener Genome ist aber die Erhöhung des Abstraktionsniveaus äußerst sinnvoll. Unterschied in der Codon-Präferenz verschiedener Organismen (Jansen et al. 2003) führen beispielsweise dazu, dass bei orthologen Genen die Ähnlichkeit der Nukleinsäuresequenzen deutlich geringer ist als auf Aminosäure-Ebene.

Zur Bezeichnung des Phänomens, dass Gene in mehreren Organismen gleiche Nachbarschaften bilden, wird meist die Vokabel 'Syntenie' herangezogen. Dabei enthält die eigentliche Wortbedeutung – eine Zusammensetzung der Vorsilbe *syn* (griechisch für zusammen) und dem griechischen Wort für Band, *tainia* – keinen Hinweis auf einen Genom vergleichenden Inhalt. Tatsächlich wurde der Begriff in den frühen 70er Jahren in Folge neuer experimenteller Methoden zur Gen-Kartierung geprägt und beinhaltete zunächst lediglich, dass sich zwei Marker auf dem gleichen Chromosom *eines* (insbesondere des menschlichen) Genoms befanden (Passarge et al. 1999). Korrekterweise sollte man also von konservierter Syntenie sprechen, wenn man sich auf mehrere Genome bezieht. Diese Arbeit behandelt ausschließlich konservierte Syntenie, weshalb auf die Präzisierung zum Teil verzichtet wird (vor allem bei adjektivischem Gebrauch 'syntenisch'). Eine formale Definition folgt im Abschnitt 3.3.2.

Interessant ist das Studium von Gen-Nachbarschaften vor allem deshalb, weil sie Auskunft geben über die Vorgänge, die die Architektur eines Genoms über evolutionäre Zeiträume hinweg formen. Die zugrunde liegende Annahme dabei ist die, dass Nachbarschaften in zwei Genomen deshalb beobachtet werden können, weil sie auch im letzten gemeinsamen Vorgänger bereits bestanden ha-

ben. Denkbar wäre auch ein konvergentes Szenario, dass also ursprünglich nicht zusammenhängende Gene unabhängig in beiden betrachteten Spezies Nachbarschaften gebildet haben. Im Allgemeinen schätzt man die Wahrscheinlichkeit hierfür aber wesentlich geringer ein. Erst recht dann, wenn mehr als zwei Genome ähnliche Konstellationen aufweisen (Nadeau und Taylor 1984). Eine Sonderstellung nehmen in diesem Zusammenhang funktionelle Nachbarschaften, insbesondere prokaryotische Operons ein. Modelle zur Entstehung und Erhaltung werden in der Einleitung zu Kapitel 4 im Abschnitt 4.1.2 diskutiert.

Eine sehr einfache Möglichkeit, syntenische Bereiche zu erkennen, ist die Verwendung sog. Dot-Plots. Darin spannen die betrachteten Genome eine Matrix auf, jedes Gen entspricht einer Skalen-Einheit. Somit ist jede mögliche Paarung von Genen der beiden Organismen durch einen Punkt in der Fläche repräsentiert. Positionen, die orthologen Paaren entsprechen, werden farblich markiert. Für konservierte Nachbarschaften ergeben sich Punkt-Cluster in Form diagonaler Linien (in Richtung der Hauptdiagonalen bei konservierter Orientierung, sonst entgegengesetzt). Abb. 3.1 zeigt ein Beispiel mit dem Vergleich zweier Listerien-Arten.

Solche Plots eignen sich gut, um einen Überblick über syntenische Bereiche in nah verwandten Organismen zu gewinnen. Wann sie zuerst für diesen Zweck eingesetzt wurden, lässt sich kaum zurückverfolgen, zumal sie mit geringem Aufwand zu erstellen sind. Eine spezialisierte Software wurde kürzlich vorgestellt (Celamkoti et al. 2004), die allerdings auf Genome bis zu 2 Mb beschränkt ist (die beiden Listerien-Genome in Abb. 3.1 wären bereits zu groß).

Bei weniger nah verwandten Organismen stößt der Einsatz von Dot-Plots jedoch an seine Grenzen. Gerade in größeren Genomen sind kleinere Diagonalen kaum noch sichtbar. Auch die Darstellung von Umordnungen in der Gen-Reihenfolge ist eher unübersichtlich. Im Rahmen der hier vorgestellten Arbeit werden daher Visualisierungen präsentiert, die nicht den genannten Limitierungen unterliegen.

Die Subsumtion des Problems Syntenie unter die allgemeine Formulierung konservierter Nachbarschaften gemäß Definition 1.1 ist offensichtlich. Als Objekte fungieren (vorhergesagte) Gene. Je nach Datenlage werden vollständige Chromosomen oder Contigs (aufgefasst als geordnete Listen von Genen) betrachtet, die Abbildung erfolgt über Ähnlichkeit in der Aminosäuresequenz.

3.2 Theoretische Ansätze

In aller Regel liegen orthologe Gene in syntenischen Bereichen zweier Genome nicht in exakt gleicher Reihenfolge vor. Die Differenzen in der Gen-Abfolge sind einer mathematischen Formulierung zugänglich. Die Modellierung von Genom-Umordnungen und Entwicklung entsprechender Algorithmen ist daher eine der traditionellen Domänen im Bereich der theoretischen vergleichenden Genom-Forschung.

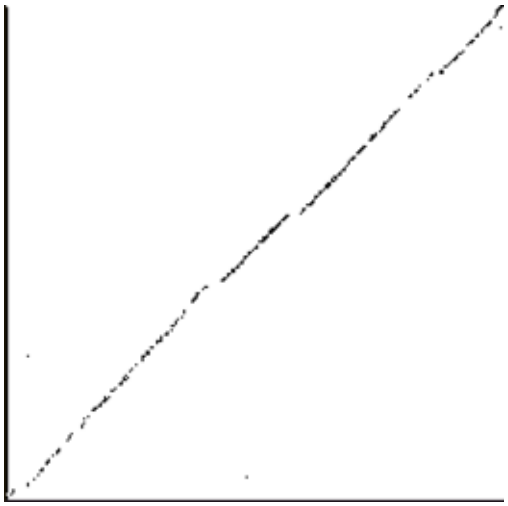


Abbildung 3.1: Dot-Plot: *Listeria innocua* (horizontale Achse) vs. *Listeria monocytogenes* (vertikale Achse). Gut zu erkennen sind Unterbrechungen, die auf Phagen-Einschlüsse in *L. innocua* zurückzuführen sind.

Von der DNA als Träger genetischer Information wird zumeist vollkommen abstrahiert; ein Genom wird als geordnete Liste von Genen aufgefasst. Auch Merkmale wie Gen-Länge, Intron-Exon-Struktur oder Sequenz-assoziierte Charakteristika wie Codon-Usage oder GC-Gehalt spielen keine Rolle. Beim Vergleich zweier Genome werden nur orthologe Paare berücksichtigt. Nummeriert man nun die Gene in einem Genom, ergibt sich aus der Position der orthologen Partner eine Permutation der Zahlen $1 \cdots n$, wobei n der Anzahl der Orthologen der betrachteten Genome entspricht.

Ausgehend von der beobachteten Permutation ergeben sich zwei Probleme: (1) Rekonstruktion der Prozesse, die die festgestellte Reihenfolge herstellen können, (2) Bestimmung eines auf der Permutation basierten Abstands der beiden Genome. Diese beiden Fragestellungen sind nicht unabhängig voneinander: Eine mögliche Abstands-Definition besteht darin, die Anzahl der notwendigen Schritte abzuschätzen, um Genom A in die Konfiguration B zu überführen. Eine verbreitete Herangehensweise definiert daher zunächst einen Satz an Operationen, um dann Pfade zu berechnen, die mit einer minimalen Anzahl dieser Operationen die Permutation erzeugen (Edit-Distanz).

Konkrete Problem-Formulierungen unterscheiden sich in der Definition der zulässigen Operationen und darin, ob die Codierungsrichtung der Gene einbezogen wird (Genome mit oder ohne Vorzeichen). Vorzeichenlose Modellierungen entsprechen eher den klassischen genetischen Karten, die die Position von bestimmten Marker-Genen in relativen Abständen (gemessen an der Häufigkeit, mit die beiden Gene durch Crossing-Over Ereignisse getrennt werden) enthalten. Kartierungen, die aus Sequenzierung hervorgehen, enthalten immer eine Richtungsinformation, obwohl auch diese nur relativ ist, nämlich bezogen auf das analysierte Teilstück. In welcher Richtung der Abschnitt dann z.B. in das Gesamt-Chromosom eingefügt wird, ist zunächst offen.

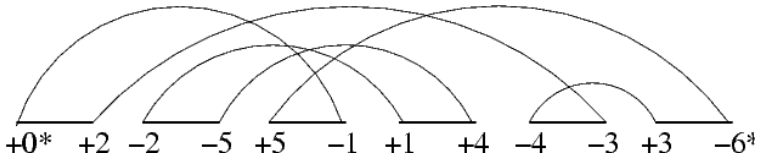


Abbildung 3.2: Breakpoint-Graph für eine Permutation $(1,2,3,4,5) \mapsto (-2,5,1,-4,3)$, dies entspricht der beobachteten Umordnung der Segmente zwischen den Chromosomen 2 und 4 in *A. thaliana* (BD1 – BD5). Die Knoten +0 und -6 werden aus technischen Gründen eingefügt. Der Graph enthält zwei geschlossene Zyklen. Die Permutation ist mit minimal vier Inversionen generierbar (Tesler 2002b) und entspricht damit der Abschätzung $D_i = n + 1 - c$.

Eine häufig verwendete Operation ist die Inversion. Dabei wird ein zusammenhängender Abschnitt des Genoms in seiner Reihenfolge gedreht. Sie transferiert ein Genom $A = a_1 \cdots a_i \cdots a_j \cdots a_n$, $1 \leq i \leq j \leq n$ zu $A' = a_1 \cdots a_j \cdots a_i \cdots a_n$. Bei Verwendung von Vorzeichen würden alle von der Operation betroffenen Gene negiert: $A' = a_1 \cdots (-a_j) \cdots (-a_i) \cdots a_n$.

Eine Transposition bewegt ein zusammenhängendes Teilstück an eine andere Position: $A = a_1 \cdots a_i a_{i+1} \cdots a_{j-1} a_j \cdots a_{k-1} a_k \cdots a_n$, $1 \leq i < j < k \leq n$ wird zu $A = a_1 \cdots a_i a_j \cdots a_{k-1} a_{i+1} \cdots a_{j-1} a_k \cdots a_n$. Auch die Kombination mit einer Inversion, genannt Transversion, bei der der transponierte Abschnitt zusätzlich gedreht wird, findet teilweise Verwendung.

Die Inversions-Distanz auf Genomen mit Vorzeichen wurde zuerst in (Kececioglu und Sankoff 1994) behandelt, Hannenhalli und Pevzner legten einen Algorithmus polynomialer Komplexität vor (Hannenhalli und Pevzner 1995). Dieser verwendet sog. Breakpoint-Graphen, eine spezielle Graph-Repräsentation der beobachteten Permutation (siehe Beispiel in Abb. 3.2). Zwei Gene, die in Genom A benachbart sind, nicht aber in Genom B, stellen einen Bruch (engl. *breakpoint*) dar. Die Mindest-Anzahl an erforderlichen Inversionen lässt sich anhand der Anzahl an geschlossenen Zyklen innerhalb des Breakpoint-Graphen gut abschätzen. Für viele biologisch relevante Probleme gilt sogar: $D_i = n + 1 - c$ (n Anzahl der betrachteten Gene, c die Anzahl der Zyklen). Verbesserungen dieses Verfahrens wurden beispielsweise in (Kaplan et al. 1997) und (Moret et al. 2001) vorgelegt. Einen Algorithmus mit linearer Laufzeit publizierten Bader et al. (Bader et al. 2001).

Für den Fall vorzeichenloser Genome lieferten wiederum Kececioglu und Sankoff einen ersten Algorithmus (Kececioglu und Sankoff 1995), (Capra-ra 1997) konnte hier jedoch den Beweis führen, dass es sich um ein NP-hartes Problem handelt. Einen Algorithmus zur approximativen Berechnung der Transpositions-Distanz stellten Bafna und Pevzner vor (Bafna und Pevzner 1998).

Die Komplexität des Problems vereinfacht sich erheblich, wenn man auf die Berechnung des Übergangs-Pfades verzichtet. Diesen Ansatz verfolgt die sog. Breakpoint-Distanz. Sie ist definiert als die Anzahl an Brüchen für ein gegebenes Paar von Genomen. Dieser Abstand ist in linearer Zeit berechenbar und zeigt

immerhin eine gewisse Korrelation mit den oben beschriebenen Edit-Distanzen (Sankoff und Blanchette 1999).

Die Schwierigkeiten mit der Komplexität von Edit-Distanzen verschärfen sich, sobald mehr als zwei Genome betrachtet werden. Selbst bei Verwendung der einfachsten Variante, der Inversions-Distanz auf Genomen mit Vorzeichen, ist die Rekonstruktion eines sog. Median für drei gegebenen Genome NP-hart (Caprara 1999). Dabei handelt es sich um eine Konfiguration von Genen, die die Summe der Abstände zu den Ausgangsgenomen minimiert und deshalb als gemeinsamer Vorfahre interpretiert werden kann.

Dennoch gibt es Ansätze, die Gen-Reihenfolge zur Rekonstruktion phylogenetischer Bäume zu nutzen. Eine Möglichkeit ist die Verwendung der Breakpoint-Distanz, obwohl auch hier die Verallgemeinerung auf mehr als zwei Genome NP-hart ist (Pe'er und Shamir 1998). Allerdings gibt es eine Reduktion auf das Traveling-Salesman-Problem, das ausführlich studiert ist und für das effiziente Heuristiken existieren (Blanchette et al. 1997). Kritisiert wird, dass die Breakpoint-Distanz nicht immer der minimal notwendigen Anzahl an Umordnungs-Ereignissen entspricht. Außerdem ist unklar, wie sie auf Genome mit mehreren Chromosomen ausgedehnt werden kann (Bourque und Pevzner 2002).

Die Autoren des letztgenannten Artikels bevorzugen daher die Inversions-Distanz und entwickelten dazu eine greedy Heuristik. Das bedeutet, dass in jedem einzelnen Schritt eine Bewertung aller möglichen Operationen vorgenommen und die günstigste ausgewählt wird. Eine über den Einzelschritt hinausgehende Strategie gibt es nicht. Für den allgemeinen Fall ist damit keine optimale Lösung garantiert. Gemäß Bourque et al. findet der Algorithmus aber für typische Anwendungsfälle Lösungen, die nur gering von der (vermeintlichen) tatsächlichen Entwicklung abweichen. Durch Einbeziehung weiterer Operationen wie Translokationen sowie Fusion und Spaltung von Chromosomen kann der Algorithmus auch auf typische Eukaryonten-Genome mit mehreren Chromosomen angewendet werden (Bourque und Pevzner 2002).

Dennoch gibt es einige grundsätzliche Kritikpunkte bezüglich der Verwendung von Edit-Distanzen in Heuristiken zur Rekonstruktion phylogenetischer Bäume. Eines der zentralen Probleme ist die weitgehende Unkenntnis relativer Häufigkeiten verschiedener Ereignisse und der daraus folgenden Kosten, die den entsprechenden Operationen zuzuweisen sind (Blanchette et al. 1997; Andersson und Eriksson 2000). Außerdem steigt mit wachsenden Unterschieden in der Gen-Abfolge die Anzahl optimaler Pfade, was die Ableitung der Bäume kompliziert. In diesen Fällen unterschätzt die minimierte Edit-Distanz die Zahl tatsächlicher Ereignisse mehr oder minder deutlich (Andersson und Eriksson 2000).

Die Einschränkungen der Methoden führten dazu, dass sie lange Zeit ausschließlich auf kleine und übersichtliche Genome wie Mitochondrien, Plastiden oder allenfalls Prokaryonten anwendbar waren. In Eukaryonten können Orthologie-Beziehungen nur schwer korrekt bestimmen werden, die korrekte Abbildung zwischen Orthologen ist jedoch essentiell. Eine neuere Analyse (Pevzner und

Tesler 2003a) der Genome von Mensch und Maus umgeht dieses Problem, indem sie syntenische Bereiche nicht aufgrund von Genen, sondern mittels besonders signifikant alignierter DNA-Abschnitte (auch als 'Anker' bezeichnet) identifiziert.

Ein Satz von insgesamt 558.678 solcher Anker wurde im Rahmen des Maus-Sequenzierungs-Projektes ermittelt (Mouse Genome Sequencing Consortium 2002). Dabei handelt es sich um relativ kurze Abschnitte mit durchschnittlich 340 nt, die teils innerhalb, teils außerhalb codierender Regionen lokalisiert sind. Ein mehrstufiger Filterprozess namens GRIMM (Tesler 2002a) fasst die Anker zu insgesamt 319 Clustern mit einer Mindestlänge von 1 Mb zusammen. Die Anwendung eines Kollinearitäts-Filters reduziert die Anzahl nochmals auf 281 Syntenie-Regionen. Das minimale Szenario zur Rekonstruktion dieser Beobachtungen enthält insgesamt 245 Operationen (Inversionen, Translokationen, Chromosomen-Spaltungen und -Fusionen). Viele der Syntenie-Blöcke weisen lokalen Störungen auf, das Ausmaß variiert jedoch stark (zwischen 0 und 40 sog. Mikro-Rearrangements wurden festgestellt).

Außer den 281 großen Regionen ($>$ 1 Mb) identifizieren Pevzner und Tesler viele weitere, wesentlich kleinere syntenische Blöcke. Außerdem beobachten sie, dass bestimmte, zum Teil sehr kleine chromosomale Abschnitte mehrfach als Bruchstelle gedient haben. Aus diesen beiden Funden schließen sie, dass das lange Zeit unumstrittene Modell chromosomaler Umordnungen nicht uneingeschränkt gültig ist (Pevzner und Tesler 2003b). Nadeau und Taylor hatten in den 80er Jahren die Hypothese aufgestellt, Bruchstellen seien zufällig über die Chromosomen verteilt (Nadeau und Taylor 1984), viele nachfolgende Beobachtungen schienen dies auch zu bestätigen. Pevzner und Tesler teilen Chromosomen nun auf in sehr stabile Segmente einerseits und besonders fragile Regionen andererseits. Solche 'hot spots' für chromosomale Umstrukturierungen wurden inzwischen bestätigt (Bailey et al. 2004).

Eine weitere Syntenie-Analyse zwischen Mensch und Maus (Kent et al. 2003) kommt trotz sehr unterschiedlicher Vorgehensweise (hier stützt man sich auf überlappende Genom-Alignments) zu ähnlichen Ergebnissen. Auch hier betonen die Autoren die Wichtigkeit der Differenzierung in 'große' und 'kleine' syntenische Blöcke. Der gewählte Maßstab ist hier jedoch sehr viel kleiner, als Grenze gilt hier die Marke von 100 kb. Dieser Wert ist, ebenso wie die 1 Mb von Pevzner und Tesler, letztlich willkürlich gewählt. Zunehmende Erfahrung in der Analyse komplexer Genome sollte zukünftig eine mehr biologisch motivierte Klassifizierung ermöglichen (Sankoff und Nadeau 2003).

3.3 Eigene Arbeiten

3.3.1 *A. thaliana* vs. Reis

Syntenie zwischen Mono- und Dikotyledonen?

A. thaliana gilt für die Dikotyledonen als Modell-Organismus; eine vergleichbare Stellung besetzt *Oryza sativa* (Reis) für die zweite große Gruppe der Blütenpflanzen, die Einkeimblättrigen (Monokotyledonen). Die Zeit seit der Aufspaltung dieser Linien wird auf etwa 200 Millionen Jahre geschätzt (Wolfe et al. 1989). Obwohl zwischen phylogenetisch derart weit auseinander liegenden Spezies auf chromosomaler Ebene große Re-Organisationen zu erwarten sind, wurde das Ausmaß an konservierter Syntenie für kleinere Bereiche vor der Sequenzierung von *Arabidopsis* recht optimistisch eingeschätzt. Über Abschnitte von bis zu 3 Centimorgan wurde vorhergesagt, dass etwa 50% der enthaltenen Gene eine kollineare Struktur bewahrt hätten (Paterson et al. 1996). Das würde bedeuten, dass etwa Gen- oder Funktions-Vorhersagen zwischen beiden Gruppen in hohem Grade übertragbar sind (Bevan und Murphy 1999). Spätere Ergebnisse deuteten jedoch darauf hin, dass diese Schätzungen zu hoch angesetzt waren (Devos et al. 1999; van Dodeweerd et al. 1999).

Ansatz

Im Rahmen der Sequenzierung und Annotation eines 340 kb langen Abschnitts des Reis-Chromosoms 2 galt es, syntenische Bereiche in *A. thaliana* zu identifizieren. Die manuelle Gen-Vorhersage stützte sich auf vier verschiedene Programme, bezog außerdem Übereinstimmungen mit ESTs und Homologie-Suchen in Datenbanken ein. Auf diese Weise wurden insgesamt 56 Gene auf diesem Segment ermittelt (Mayer et al. 2001).

Aufgrund des großen phylogenetischen Abstands ist eine hohe Konservierung über große Segmente auf DNA-Ebene zwischen Reis und *Arabidopsis* nicht zu erwarten, die Syntenie-Analyse konzentrierte sich darum auf Gen-Vergleiche mittels BLASTP. Dabei wurde mit allen vorhergesagten Proteinsequenzen des Reis-Segments eine Suche gegen die Datenbank aller *Arabidopsis*-Gene durchgeführt. Berücksichtigung fanden nur Treffer mit einem P-Value $\leq 10^{-5}$. Dieser relativ wenig stringente Schwellwert (bei der gegebenen Datenbankgröße von ca. 25.000 Genen sind $10^{-5} \cdot 25.000 = 2,5$ Zufallstreffer zu erwarten) berücksichtigt grundsätzliche Unterschiede der beiden Genome hinsichtlich ihrer Sequenz-Komposition. Differenzen bestehen auf DNA-Ebene bezüglich GC-Gehalt und Codon-Usage, aber auch in der Aminosäure-Nutzung (Wong et al. 2002). Außerdem soll die Homologiesuche – wie bereits bei der intragenomischen Analyse in *Arabidopsis* – mit hoher Sensitivität erfolgen.

Der angewendete Algorithmus zum Auffinden kollinearer Cluster stellt eine Modifikation des in Abschnitt 2.3.1 geschilderten Verfahrens dar. Treffer sind in diesem Falle keine Paare von Teil-Sequenzen, sondern Paare von Genen. Dabei wird jedes Gen durch die Position seines Auftretens auf dem entsprechenden Chro-

mosom repräsentiert, es ergeben sich also Paare von Ordinalzahlen. Eine Partitionierung im engeren Sinne wird nicht vorgenommen, der LIS-Algorithmus sucht also nach kollinearen Hits unter allen Treffern auf einem Chromosom.

Darüber hinaus wird das Kollinearitäts-Kriterium gelockert. Nicht nur die Paare, die Teil des LIS sind, werden berücksichtigt, sondern auch Treffer, die in beiden Instanzen innerhalb einer festgelegten Spanne liegen. Auf diese Weise werden kleinere lokale Umordnungen nicht verworfen (siehe Beispiel in Abb. 3.3).

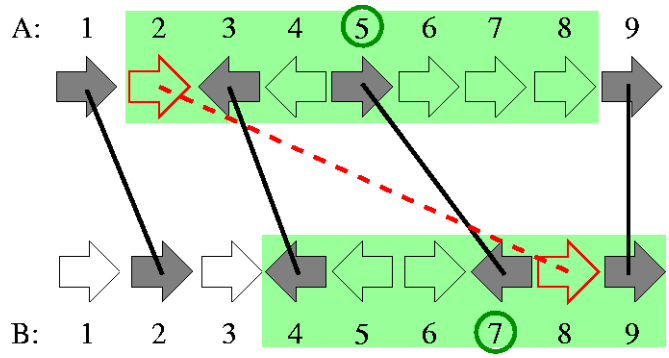


Abbildung 3.3: Gelockertes Kollinearitäts-Kriterium: Dargestellt sind zwei Regionen auf den Genomen A und B mit jeweils neun Genen. Orthologe Gene sind durch Linien verbunden. Die Paare $[(1,2),(3,4),(5,7),(9,9)]$ bilden eine streng kollineare Konfiguration (LIS). Hellgrün dargestellt eine Toleranz-Spanne von 3 um den Treffer $(5,7)$. Diese führt dazu, dass auch der rot markierte Treffer $(2,8)$ einbezogen wird.

Ergebnisse

Das beschriebene Verfahren identifiziert fünf Cluster innerhalb des *Arabidopsis*-Genoms mit mehr als drei kollinearen Treffern. Am weitesten ausgeprägt sind die Übereinstimmungen mit einem Segment auf Chromosom 4, hier finden sich acht Gene mit sehr hoher Homologie (P-Values zwischen 10^{-22} und 10^{-143}). Sechs der acht stimmen in ihrer Codierungsrichtung überein. Die zwei Ausnahmen sind auch bezüglich ihrer Reihenfolge vertauscht, so dass man hier von einer lokalen Inversion ausgehen kann (siehe auch Abb. 3.4). Weitere Regionen liegen auf den Chromosomen 2,3 und 5, auf Chromosom 4 gibt es einen zweiten Abschnitt, der getroffen wird. Die Anzahl der Hits ist hier jedoch geringer (7, 5, 7 bzw. 5).

Insgesamt zeigen 22 der 56 Reis-Gene Homologie zu mindestens einem Gen in den genannten Regionen, neun davon haben Partner in nur einem *Arabidopsis*-Segment, drei treffen drei verschiedenen Regionen, die verbleibenden zehn weisen Treffer in zwei Segmenten auf. Es ergibt sich ein Syntenie-Netzwerk mit insgesamt 33 Stützstellen (Hits), dargestellt in Abb. 3.4. Die strikte Übereinstimmung bezüglich der Codierungsrichtung wie für das Segment 4(a) festgestellt trifft jedoch nur noch auf die Regionen auf Chromosom 5 zu, bei allen anderen gibt es eine oder zwei Abweichungen.

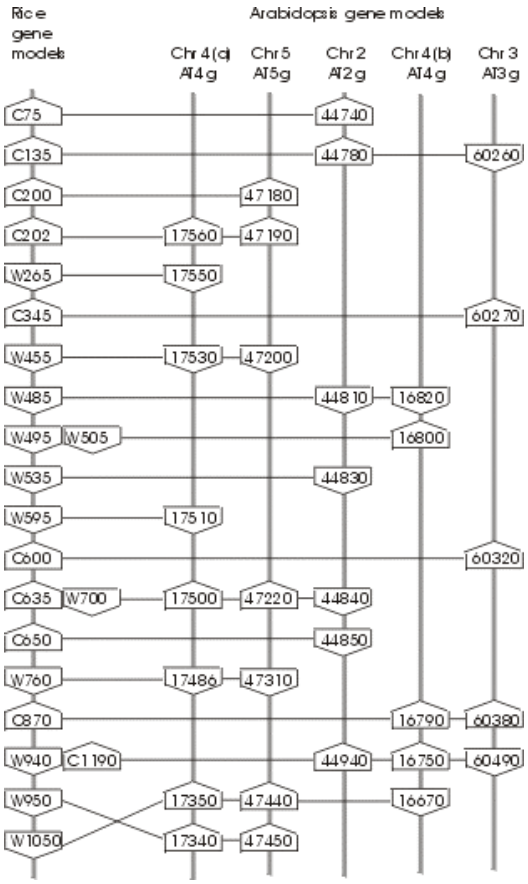


Abbildung 3.4: Organisation des analysierten Reis-Segments und der fünf syntenischen Bereiche innerhalb von *A. thaliana*. Die Spitzen der Polygone zeigen die Codierungsrichtung, mehrfach getroffenen Reis-Gene sind nebeneinander angeordnet an der Position des Gens mit der kleineren Ordnungsnummer. Gene ohne Treffer sind nicht dargestellt (Mayer et al. 2001).

Zwei weitere, 145 kb und 156 kb große Reis-BACs (P0406H10 und P0436E04) wurden in gleicher Weise analysiert, 26 bzw. 25 Gene konnten hier identifiziert werden. Die Syntenie-Analyse resultierte bei P0406H10 in drei Abschnitten des *Arabidopsis*-Genoms auf den Chromosomen 1, 3 und 4, im Fall von P0436E04 in zwei unterschiedlichen Regionen auf Chromosom 5. Die einzelnen Regionen überlappen in ähnlicher Weise wie in Abb. 3.4, allerdings mit weniger Stützstellen (P0406H10: zwölf Hits zu insgesamt neun Reis-Genen, P0436E04: acht Treffer zu fünf Reis-Genen).

Diskussion

Die Ergebnisse zeigen, dass durchaus Spuren von konservierter Syntenie zwischen Mono- und Dikotyledonen existieren. Sie verdeutlichen aber auch das Ausmaß der Veränderungen, denen die beiden Gruppen seit ihrer Aufspaltung unterworfen waren. Den höchsten Grad an Kollinearität zeigt das analysierte 340 kb Reis-Segment (bzw. etwa 2/3 davon) und ein etwa 83 kb umfassender Abschnitt auf Chromosom 4 von *Arabidopsis*. Aber selbst hier sind nur neun von 37 Genen konserviert, wenn auch mit ausgezeichneten Homologie-Werten und konsistenter Orientierung. D.h. Deletionen und Re-Organisationen haben dazu geführt, dass fast 75% der Gene keinen Partner mehr in der syntenischen Region aufweisen. Offenbar wurde der Gen-Gehalt der Regionen also durch eine Vielzahl an Ereignissen, die nur einzelne Gene oder kleinere Gruppen betreffen, verändert. Dies widerspricht der in (Paterson et al. 1996) geäußerten Annahme, dass Pflanzen-Genome hauptsächlich durch große Re-Organisationen umgestaltet werden und widerlegt letztlich auch die optimistischen Vorhersagen bezüglich des Grades an Syntenie zwischen ein- und zweikeimblättrigen Pflanzen.

Später durchgeführte ähnliche Analysen kommen trotz Unterschieden in Details zur übereinstimmenden Aussagen: zwischen den Genomen von Reis und *Arabidopsis* existiert konservierte Syntenie, aber nur auf sehr klein skalierten Abschnitten (Liu et al. 2001; Salse et al. 2002). Alle Studien entdecken Segmente in Reis, die zu mehr als einer Region in *A. thaliana* Ähnlichkeiten aufweisen.

Schwierig ist eine Interpretation der Ergebnisse in Hinblick auf Genom-Duplikationen. Das in Abb. 3.4 dargestellte Netz ist konsistent mit mehrfachen Polyploidisierungen in *thaliana*. Es bestätigt die Ergebnisse des von Ku et al. durchführten Vergleichs zwischen *Arabidopsis* und Tomate. Das dort ermittelte Geflecht syntenischer Beziehungen wurde als Indiz für zwei Komplett-Verdoppelungen interpretiert (Ku et al. 2000). Die Anzahl von fünf unterschiedlichen Regionen könnte sogar auf drei Tetraploidisierungen hindeuten, die Anzahl der Stützstellen ist allerdings viel zu gering, um nicht auch Duplikationen einzelner Segmente ausschließen zu können.

Die Interpretation wird dadurch erschwert, dass eindeutige Orthologie-Beziehungen zwischen Reis und *Arabidopsis* aufgrund der hohen Zahl an Duplikationen in beiden Linien kaum zu identifizieren sind (Yu et al. 2002). Dieser Umstand kann zu einer Überschätzung der Syntenie führen (Salse et al. 2002).

Dennoch ist die konservierte Nachbarschaft von mehreren Genen in Kombination mit hoher Sequenzähnlichkeit ein wichtiger Anhaltspunkt für einen gemeinsamen Ursprung.

Die systematische Einbeziehung von Syntenie-Daten zur Ermittlung orthologer Gen-Sets wurde im Vergleich *Caenorhabditis briggsae* und *Caenorhabditis elegans* (Stein et al. 2003) sowie zwischen Mensch und Maus (Zheng et al. 2005) angewendet. Auch das Ensembl Projekt integriert in seiner Orthologie-Berechnung Information über syntenische Blöcke (Clamp et al. 2003).

3.3.2 conSynteny

Die bislang vorgestellten Verfahren (2.3.1, 3.3.1) implizieren einen sehr eng gefassten Begriff konservierter Nachbarschaften, der Umordnungen innerhalb der Cluster gar nicht bzw. nur in sehr begrenztem Umfang erlaubt. In diesem Abschnitt wird die Methode so modifiziert, dass beliebige Konfigurationen der aufeinander abgebildeten Nachbarschaften zulässig sind. Ergebnisse aus Vergleichen zwischen einigen Arten aus der Klasse der *Chlamydiae* sowie zwischen Pilz-Genomen zeigen beispielhaft die Anwendung der Software.

Ansatz

Das Grundproblem zum Auffinden syntenischer Bereiche kann wie folgt formuliert werden: Finde Gruppen von Genen in zwei Genomen A, B, die in A nah beieinander liegen und potentielle Orthologe in B haben, die dort ebenfalls benachbart auftreten. Eine bestimmte Konfiguration (etwa parallele oder anti-parallele Ausrichtung) der einzelnen Treffer ist nicht erforderlich.

Definition 3.1 (Verallgemeinerte Syntenie)
Gegeben seien zwei Genome $A = (a_1, a_2, \ldots, a_n)$ und $B = (b_1, b_2, \ldots, b_m)$ als geordnete Listen von Genen sowie eine Ähnlichkeitsfunktion $s : A \times B \mapsto \mathbb{R}^+$. Zwei Bereiche $A' = (a_{i_1}, a_{i_2}, \ldots, a_{i_k}) \subset A$ mit $1 \leq i_1 < i_2 \cdots < i_k \leq n$ sowie $B' = (b_{i_1}, b_{i_2}, \ldots, b_{i_l}) \subset B$ mit $1 \leq i_1 < i_2 < \cdots < i_l \leq m$ heißen syntenisch, wenn gilt: $\forall a_i \in A' \; \exists \; b_i \in B' : \; s(a_i, b_i) > c$, wobei eine Mindest-Ähnlichkeit c festzulegen ist. Außerdem gilt für beide Teillisten A', B': $i_{\alpha+1} - i_\alpha \leq h$ mit einem definierten Maximal-Abstand (Lochgröße) h.

Es bietet sich folgendes Vorgehen an: stelle eine Liste aller Gen-Paare (Treffer) zusammen, deren Ähnlichkeit den Mindestwert übersteigt. Wende auf die sich ergebende Liste von Positions-Paaren den Partitionierungs-Algorithmus aus 2.3.1 an. Eine legale Partition der Treffer-Liste repräsentiert dann ein Paar syntenischer Bereiche.

Für die folgenden Anwendungen werden noch zwei weitere Parameter eingeführt. (1) Mindestgröße des syntenischen Bereichs, wobei die Größe als Anzahl an Treffern definiert ist, die den Bereich aufspannen. (2) Mindest-Vielfalt, d.h. die Anzahl an unterschiedlichen Genen, die in jedem Genom mindestens getroffen werden müssen. Dieser Parameter dient zum Unterdrücken von Tandem-

Duplikationen, die ggf. zu unerwünschten Ergebnissen führen können. So generiert ein in beiden Genomen vorhandenes Tandem vier signifikante Treffer, eine typische Mindestgröße von drei würde somit erreicht. Eine solche Konfiguration ist aber nicht unbedingt als konservierte Nachbarschaft zu interpretieren. Tandem-Duplikationen sind ein relativ häufiges Ereignis und können unabhängig in den beiden Genomen aufgetreten sein. Ein alternativer Ansatz besteht darin, in einem Vorverarbeitungs-Schritt redundante Gene in Tandem-Konfiguration auszuschließen (z.B. (Simillion et al. 2002; Vision et al. 2000)).

Anwendung: *Chlamydiae*

Chlamydien vs. Parachlamydien Zu den medizinisch besonders interessanten Bakterien gehört die Klasse der *Parachlamydien*, vor allem wegen einiger human-pathogener Arten in der Familie der *Chlamydiaceae* wie etwa *Chlamydia trachomatis*, *Chlamydophila pneumoniae* und *Chlamydophila caviae*. Erst seit relativ kurzer Zeit bekannt sind verwandte Arten, die als Endosymbionten in Amoeben leben. Eine dieser Parachlamydien-Spezies wurde kürzlich sequenziert und annotiert (Horn et al. 2004).

Dabei handelt es sich um *Parachlamydia* sp. UWE25, dessen Genom mit ca. 2,4 Mb und 2031 Genen etwa doppelt so groß ist wie das seiner pathogenen Verwandten (siehe Tabelle 3.2). Nach der Aufspaltung der beiden Linien vor geschätzt etwa 700 Millionen Jahren konnten die obligatorisch intrazellulär lebenden Chlamydien durch extreme Anpassung an den jeweiligen Wirt ihr Genom massiv verkleinern (metabolische Pfade werden überflüssig, wenn die Endprodukte aus dem umgebenden Medium aufgenommen werden können). Amoeben bieten eine weit weniger stabile Umgebung als etwa Säugetier-Zellen, so dass den Parachlamydien eine komplettere genetische Ausstattung erhalten blieb (Horn et al. 2004).

Syntenie-Analyse Angesichts des Unterschiedes in den Genom-Größen und des Zeitraumes seit dem letzten gemeinsamen Vorfahren ist nicht zu erwarten, dass ausgeprägte Parallelität zwischen *Parachlamydia* sp. UWE25 und den Chlamydiaceen existiert. So ist denn auch insgesamt kaum konservierte Gen-Reihenfolge zu beobachten (Horn et al. 2004), entsprechende Dotplots zeigen einige verstreute Punkte-Cluster, jedoch kaum ausgeprägte Diagonalen (Thomas Rattei, unveröffentlicht). Anhand des beschriebenen Verfahrens zum Auffinden syntenischer Bereiche wurden paarweise Analysen von UWE25 gegen einige Chlamydien-Arten durchgeführt.

Zunächst wurden potentielle Orthologe mit Hilfe von FASTA identifiziert (score $\geq$ 0,2 vom selfscore). Auf die resultierenden Treffer-Listen wurde der Partitionierungs-Algorithmus mit folgenden Parametern angewendet: zulässige Lochgröße = 10, Mindestgröße = 5, Mindest-Variabilität = 3.

Ergebnisse Auf diese Weise wurden 33 – 39 syntenische Blöcke identifiziert (vgl. Tabelle 3.2). Dabei fällt auf, dass jeweils etwa 50% der Orthologen in die-

Organismus	Kürzel	Referenz
Parachlamydia sp. UWE25	UWE25	(Horn et al. 2004)
Chlamydia muridarum MoPn	Cmuridarum	(Read et al. 2000)
Chlamydophila caviae GPIC	Ccaviae	(Read et al. 2003)
Chlamydophila pneumoniae J138	J138	(Shirai et al. 2000)
Chlamydophila pneumoniae CWL029	CWL029	(Kalman et al. 1999)
Chlamydia trachomatis sv D	Ctrachomatis	(Stephens et al. 1998)

Tabelle 3.1: Syntenie-Analyse: einbezogene Genome

sen Regionen lokalisiert sind. Die ermittelten Blöcke enthalten durchschnittlich neun Gen-Paare, der größte umfasst 40 Treffer (Cluster ribosomaler Gene). Dies widerlegt die Annahme, dass keine Konservierung bezüglich der Ordnung der Gene vorliegt.

Die Verteilung der syntenischen Regionen über die Genome deutet allerdings auf eine Vielzahl an Re-Organisationen hin (siehe Abb. 3.5). Die Feinstruktur der Blöcke unterstützt diesen Eindruck. Immer wieder gibt es kleinere Inversionen, außerdem enthalten die Regionen viele nicht orthologe Gene.

Genom	Größe [Mb]	Anzahl Gene	Anzahl Orthologe	Anzahl Blöcke	Gene in Blöcken
Cmuridarum	1,07	921	658	35	319
Ccaviae	1,17	1009	710	39	358
J138	1,23	1072	686	35	339
CWL029	1,23	1073	689	35	339
Ctrachomatis	1,04	894	652	33	310

Tabelle 3.2: Genom-Größe, Anzahl Orthologer zu UWE25, Anzahl der syntenischen Blöcke und der darin enthaltenen Gene für die beteiligten Organismen.

Abbildung 3.5 zeigt die Lage der syntenischen Regionen zwischen UWE25 (Mitte) und *Chlamydophila caviae* (oben) bzw. *Chlamydia trachomatis* (unten). Es fällt auf, dass in beiden Vergleichen in UWE25 überwiegend gleiche Abschnitte des Chromosoms betroffen sind. Die Lage in den Vergleichs-Genomen unterscheidet sich in vielen Fällen jedoch deutlich. Diese Beobachtung ist für alle durchgeführten Analysen gültig. Für die Funktion der Gene, die diese Regionen definieren, ist es also offenbar essentiell, dass sie sich in gegenseitiger Nähe auf dem Chromosom befinden.

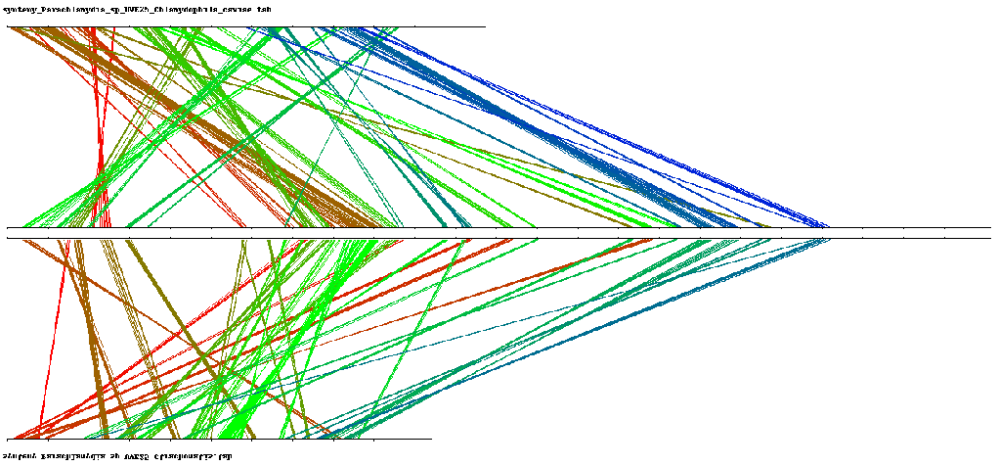


Abbildung 3.5: Vergleich *Parachlamydium* (Mitte) vs. *C. caviae* (oben) bzw. *C. trachomatis* (unten). Die Genome sind linearisiert, der Anfang entspricht dem Replikations-Ursprung. Farbige Linien verbinden orthologe Gene, die innerhalb syntenischer Regionen lokalisiert sind. Die Einfärbung dient zur Unterscheidung und spiegelt keine Qualitätsstufen wieder. (Plots: Thomas Rattei, Zusatz-Material zu (Horn et al. 2004))

Im folgenden werden beispielhaft einige der syntenischen Blöcke in einer profilartigen Darstellung gezeigt. Die erste Spalte enthält die Bezeichnung der UWE25 Gene, die weiteren Spalten geben die Positionen in den einzelnen Genomen wieder, sofern dort vorhanden.

Gen	**UWE25**	**Cm**	**Cc**	**J138**	**CWL029**	**Ct**
pc0815	1293	-	518	247	242	-
pc0817	1295	-	516	249	244	-
pc0818	1296	-	531	-	-	199
pc0819	1297	450	528	238	233	188
pc0821	1299	449	529	237	232	187
pc0823	1301	448	-	236	231	186
pc0824	1302	447	-	235	230	185
pc0825	1303	446	-	234	229	184
pc0827	1305	-	-	-	-	210

Tabelle 3.3: Beispiel-Region I: besonders große Divergenz

Das erste Beispiel (siehe Tabelle 3.3) zeigt eine besonders divergente Region. Es gibt eine Kern-Region von pc0819 – pc0825, die in allen Genomen außer *Chlamydophila caviae* vollständig und perfekt kollinear auftritt. Dabei fällt auf, dass die Kontinuität dieses Blockes nur in UWE25 leicht gestört ist, weil pc0820 keine Treffer in den anderen Genomen aufweist. Neben diesem Kern gibt es in fast allen anderen Organismen (Ausnahme: *Chlamydia muridarum*) benachbarte Treffer, die aber nicht mehr kollinear liegen und zum Teil den Maxi-

malabstand innerhalb einer Region (10 unbeteiligte Gene) voll ausnutzen. Die überwiegende Mehrzahl der identifizierten syntenischen Segmente verhalten sich einheitlicher als es hier der Fall ist.

Beispiel II in Tabelle 3.4 stellt zwei ineinander verschränkte Regionen dar. D.h. ein zusammenhängender Abschnitt in UWE25 wird auf zwei getrennte Segmente in den Ziel-Genomen abgebildet. Ein Block wird durch die Gene pc0244, pc0246, pc0247, pc0249 und pc0251 definiert, der zweite durch die übrigen im Profil aufgeführten.

Gen	**UWE25**	**Cm**	**Cc**	**J138**	**CWL029**	**Ct**
pc0240	718	329	427	348	336	66
pc0241	719	329	427	348	336	66
pc0244	722	719	179	555	541	455
pc0246	724	714	184	550	536	450
pc0247	725	720	178	556	542	456
pc0249	727	721	176	558	544	457
pc0250	728	329	427	348	336	66
pc0251	729	722	175	559	545	458
pc0253	731	330	428	347	335	67
pc0256	734	332	429	346	334	68
pc0257	735	333	430	345	333	69
pc0258	736	334	431	344	332	70
pc0259	737	335	432	343	331	71
pc0260	738	336	433	342	330	72
pc0261	739	337	434	341	329	73

Tabelle 3.4: Beispiel-Region II: verschränkte Regionen

Besonders interessant ist hier, dass mit pc0240, pc0241 und pc0250 drei der insgesamt fünf Nukleotid-Transporter beteiligt sind. In den pathogenen Chlamydien gibt es nur zwei (Horn et al. 2004). Einer davon ist in diesem Beispiel enthalten inklusive der Nachbarschaft zu einer ABC-Transporter-Kassette (pc0256–pc0260 in UWE25) sowie zwei weiteren uncharakterisierten Genen (pc0253, pc0261). Die Tatsache, dass der zweite Block (im Profil grau unterlegt) in den Chlamydien räumlich getrennt auftritt, weist auf einen modifizierten Funktions-Kontext hin.

Das dritte und letzte Beispiel (Tabelle 3.5) zeigt, dass durch eine vollständige Unterdrückung von Tandem-Duplikationen Information verloren gehen kann. Dargestellt ist die Region pc1501–pc1506. Dabei handelt es sich um das Oligopeptid-Permease-Operon OppA-OppB-OppC-OppD-OppF. Dieses Operon existiert in gleicher Konfiguration in Listerien, dort konnte für OppA eine wichtige Rolle beim Wachstum bei tiefen Temperaturen (5°C) nachgewiesen werden. Seine Funktion steht im Zusammenhang mit dem Überleben in intrazellulären Umgebungen wie etwa in Makrophagen (Borezee et al. 2000). Oligopeptid-Aufnahmesysteme sind auch an der Virulenz von *Streptococcus agalactiae* beteiligt (Samen et al. 2004).

OppA wurde in den verschiedenen Chlamydien-Arten offenbar unterschiedlich oft dupliziert und liegt dort in Tandem-Formation vor. In *Chlamydophila caviae* ist es zwei-, in den beiden *C. pneumoniae*-Stämmen dreifach vorhanden. Die doppelten Treffer für pc1506 sind dagegen auf die große Ähnlichkeit zwischen OppD und OppF zurückzuführen.

Gen	UWE25	Cm	Cc	J138	CWL029	Ct
pc1502	1980	463	588,589	194,195,197	189,190,192	200
pc1503	1981	464	590	198	193	201
pc1504	1982	465	591	199	194	202
pc1505	1983	466	592	200	195	203
pc1506	1984	466	592,593	200,201	195,196	203,204

Tabelle 3.5: Beispiel-Region III: Tandems in unterschiedlicher Anzahl

Einige weitere interessante Resultate ergeben sich aus dem Vergleich der syntenischen Regionen mit den Attribut-Clustern (siehe Kapitel 4). Als Beispiel sei hier nur der ribosomale Cluster genannt, der laut Annotation in UWE25 die Gene pc0412–pc0434 umfasst. Die Syntenie zu den Chlamydien erstreckt sich jedoch wesentlich weiter, nämlich bereits ab pc0379, allerdings mit einigen Lücken und lokalen Umordnungen.

Anwendung: Fungi

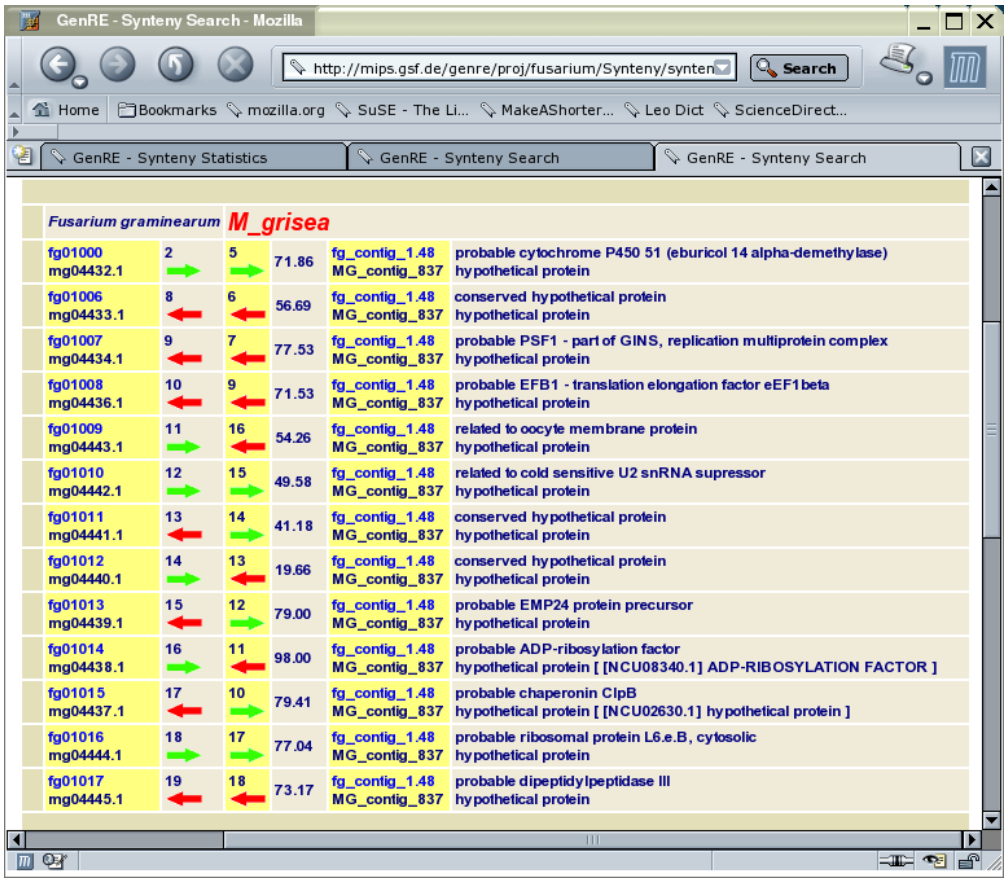
Fusarium graminearum Eine weitere Anwendung der Suche nach konservierter Syntenie ergibt sich im Zusammenhang mit *Fusarium graminearum*. Dabei handelt es sich um einen pflanzenpathogenen Fadenpilz, der wichtige Nutzpflanzen wie Gerste, Weizen und Mais befällt. Er verursacht hohe wirtschaftliche Schäden durch massive Qualitäts- und Ertrags-Einbußen. Darüber hinaus produziert er pathogene Toxine und stellt somit auch eine ernste Gesundheitsgefährdung für Mensch und Nutztiere dar (O'Donnell et al. 2000).

Die Genom-Sequenzierung von *F. graminearum* wurde am Whitehead Institut für Biomedizinische Forschung durchgeführt. Im Mai 2003 wurde eine erste Version veröffentlicht (Center for Genome Research). Aufbauend auf dieser Analyse versuchen Gruppen des Zentrums für Angewandte Genetik an der Universität für Bodenkultur in Wien und des Instituts für Bioinformatik an der GSF, München (IBI) die genetischen Grundlagen der Pathogenität des Schädlings aufzuklären. Insbesondere der Vergleich mit dem relativ nah verwandten, aber nicht pathogenen *Neurospora crassa* verspricht wertvolle Erkenntnisse.

Syntenie-Analyse Dr. Ulrich Güldener (IBI) hat die Syntenie-Analyse mittels des beschriebenen Verfahrens durchgeführt, die Ergebnisse in die 'Fusarium Graminearum Genome Database' (FGDB) eingepflegt und für die Präsentation im WWW aufbereitet. Sie stehen dort in Form eines 'Synteny Viewers' (Fusarium Synteny Viewer) zur Verfügung. Gegenüber der Chlamydien-Analyse wurden geänderte Parameter verwendet: Mindestgröße=2, Maximalabstand=5,

Variabilität=2. Orthologe Gene wurden mit BLASTP ermittelt. Die Qualität eines Treffers wurde berechnet als Produkt aus Anzahl identischer Aminosäuren und der Gesamtlänge, dividiert durch die Trefferlänge (Wilson et al. 2000). Berücksichtigt wurden alle Treffer, bei denen dieser Wert 15% oder mehr beträgt.

Ergebnisse Abb. 3.6 zeigt ein Beispiel. Dargestellt ist eine Region konservierter Syntenie zu *M. grisea*. Die linke Spalte enthält die Gen-Bezeichner, daneben die Position in *F. graminearum* (2. Spalte) und im Ziel-Genom (3. Spalte). Grüne oder rote Pfeile stehen für die Codierungsrichtung. In den letzten beiden Spalten sind die betroffenen Contigs sowie die Funktions-Beschreibungen der Gene aufgeführt. Abgebildet ist ein Bereich mit 13 Treffern. Ein Teilabschnitt mit sechs Genen (Positionen 10 – 16) ist in *M. grisea* invertiert. Die Codierungsrichtungen sind in diesem Bereich entgegengesetzt, ansonsten gleichen sie den Orthologen in *F. graminearum*.



Fusarium graminearum		*M_grisea*			
fg01000 mg04432.1	2 →	5 →	71.86	fg_contig_1.48 MG_contig_837	probable cytochrome P450 51 (eburicol 14 alpha-demethylase) hypothetical protein
fg01006 mg04433.1	8 ←	6 ←	56.69	fg_contig_1.48 MG_contig_837	conserved hypothetical protein hypothetical protein
fg01007 mg04434.1	9 ←	7 ←	77.53	fg_contig_1.48 MG_contig_837	probable PSF1 - part of GINS, replication multiprotein complex hypothetical protein
fg01008 mg04436.1	10 ←	9 ←	71.53	fg_contig_1.48 MG_contig_837	probable EFB1 - translation elongation factor eEF1beta hypothetical protein
fg01009 mg04443.1	11 →	16 ←	54.26	fg_contig_1.48 MG_contig_837	related to oocyte membrane protein hypothetical protein
fg01010 mg04442.1	12 →	15 →	49.58	fg_contig_1.48 MG_contig_837	related to cold sensitive U2 snRNA supressor hypothetical protein
fg01011 mg04441.1	13 ←	14 →	41.18	fg_contig_1.48 MG_contig_837	conserved hypothetical protein hypothetical protein
fg01012 mg04440.1	14 →	13 ←	19.66	fg_contig_1.48 MG_contig_837	conserved hypothetical protein hypothetical protein
fg01013 mg04439.1	15 ←	12 →	79.00	fg_contig_1.48 MG_contig_837	probable EMP24 protein precursor hypothetical protein
fg01014 mg04438.1	16 →	11 ←	98.00	fg_contig_1.48 MG_contig_837	probable ADP-ribosylation factor hypothetical protein [[NCU08340.1] ADP-RIBOSYLATION FACTOR]
fg01015 mg04437.1	17 ←	10 →	79.41	fg_contig_1.48 MG_contig_837	probable chaperonin ClpB hypothetical protein [[NCU02630.1] hypothetical protein]
fg01016 mg04444.1	18 →	17 →	77.04	fg_contig_1.48 MG_contig_837	probable ribosomal protein L6.e.B, cytosolic hypothetical protein
fg01017 mg04445.1	19 ←	18 ←	73.17	fg_contig_1.48 MG_contig_837	probable dipeptidylpeptidase III hypothetical protein

Abbildung 3.6: Snapshot der Ergebnisanzeige des Synteny Viewers (Fusarium Synteny Viewer).

Analysiert wurden diverse Bakterien-Genome sowie sieben Pilz-Arten. Die im Rahmen dieser Arbeit vorgestellten Ergebnisse beschränken sich allerdings auf

letztere (die ermittelten Regionen in Bakterien sind in aller Regel auf ausgedehnte Tandem-Duplikationen zurückzuführen). Objekte der Analyse waren in diesem Fall oft keine Chromosomen, sondern einzelne Contigs, da die Zuordnung zu Chromosomen in vielen Fällen nicht abgesichert ist.

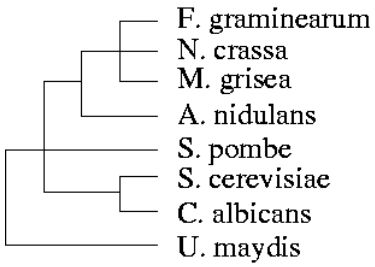


Abbildung 3.7: Phylogenie der beteiligten Fungi-Spezies. Die Kantenlängen sind nicht proportional zum tatsächlichen phylogenetischen Abstand.

Die Verwandtschafts-Verhältnisse der beteiligten Spezies zeigt Abb. 3.7. Darin wird jedoch ausschließlich auf Zugehörigkeit zu gleichen taxonomischen Gruppen abgestellt (Quelle: (Entrez Taxonomy Browser)), die Darstellung ist nicht proportional zum Alter der Aufspaltung der einzelnen Linien. Beispielsweise werden *N. crassa* und *Magnaporthe grisea* äquidistant zu *F. graminearum* aufgeführt, weil alle drei zur Gruppe der Sordariomyceten gehören. Dies schließt aber nicht aus, dass eine der beiden Arten nähere Verwandtschaft zu Fusarium aufweist.

	Blockgröße			
Genom	**> 13**	**$11 - 13$**	**$8 - 10$**	**$5 - 7$**
N. crassa	2	7	51	184
M. grisea	1	5	24	149
A. nidulans	0	1	8	57
S. pombe	0	0	6	14
C. albicans	0	0	0	17
S. cerevisiae	0	0	0	13
U. maydis	0	0	0	2

Tabelle 3.6: Größenverteilung der syntenischen Blöcke zwischen *F. graminearum* und diversen Fungi

Tabelle 3.6 fasst die Ergebnisse der Syntenie-Analyse zwischen den verschiedenen Pilz-Genomen zusammen und zeigt große Konsistenz mit dem Stammbaum. Aufgeführt ist die Anzahl an ermittelten Regionen, differenziert nach der Größe der Blöcke. Wie zu erwarten nimmt also die Größe der syntenischen Regionen mit phylogenetischem Abstand ab. Nach diesen Daten wäre *Neurospora* näher verwandt zu *F. graminearum* als *Magnaporthe*. Allerdings wirkt sich hier aus, dass die Contigs in *M. grisea* wesentlich kürzer sind. Die 11109 vorhergesagten ORFs (Quelle: PEDANT (Frishman et al. 2003)) verteilen sich hier auf 1917 Contigs; das *Fusarium*-Genom enthält 11640 Gene in nur 463 Contigs. Da ei-

ne syntenische Region niemals Contig-Grenzen überschreiten kann, ist also der Erwartungswert in *Magnaporthe* wesentlich kleiner.

Als Beispiel sei der Anfang des Contigs fg_contig_1.322 in *F. graminearum* genannt. Die Gene mit den Positionen 3, 4, 5, 7, 8, 9 und 10 haben Orthologe in drei verschiedenen *M. grisea*-Contigs, die Bezeichner (mg00863.1-mg00865.1, mg00867.1, mg00871.1, mg00872.1 und mg00874.1) deuten jedoch darauf hin, dass diese Gene unmittelbar benachbart liegen. Bei größerem Contig wäre hier also eine Region der Größe 7 identifiziert worden. Ein besseres Assembly von *M. grisea* würde also die Aussagekraft der Syntenie-Analyse bedeutend steigern. Umgekehrt zeigt das genannte Beispiel, dass Syntenie unter Umständen zur Assemblierung genutzt werden kann.

Gen	***F. graminearum***	***M. grisea***	***A. nidulans***	***N. crassa***
fg07949	8	-	165	-
fg07950	9	-	166	-
fg07951	10	-	167	-
fg07952	11	7	-	15
fg07953	12	8	-	17
fg07954	13	9	172	21
fg07955	14	10	172	19
fg07956	15	11	173	24
fg07957	16	12,13	175	22,23
fg07958	17	14	168	25
fg07959	18	15	-	26
fg07960	19	16	171	28
fg07963	22	-	-	3
fg07964	23	-	-	5
fg07965	24	1*	-	4
fg07966	25	3*	-	11
fg07967	26	4*	-	10
fg07968	27	-	-	14
fg07969	28	-	-	13

Tabelle 3.7: Syntenie-Profil für die Contigs fg_contig1.322, MG_contig_689 bzw. *= MG_contig_690, an_contig_1.61, LGVII:7nc520

Die jeweils längsten Syntenie-Regionen zeigen keine Überschneidungen, dennoch gibt es einige größere Abschnitte mit konservierten Teilen in den vier Arten, die den höchsten Verwandtschaftsgrad aufweisen. Profile für diese Blöcke zeigen die Tabellen 3.7 und 3.8. Einzelne Contigs weisen eine sehr hohe Anzahl an Genen in syntenischen Bereichen zu den andern Pilz-Genomen auf. Herausragend ist hier fg_contig_1.185: hier werden 85 Gene in *N. crassa*, 75 in *M. grisea*, 73 in *A. nidulans*, 39 in *S. pombe*, 22 in *C. albicans*, 16 in *S. cerevisiae* und 15 in *U. maydis* zugeordnet, die meisten allerdings in sehr kleinen Blöcken.

Gen	***F. graminearum***	***M. grisea***	***A. nidulans***	***N. crassa***
fg08857	16	11	-	17,24
fg08859	18	12	92	18
fg08860	19	13	93	19
fg08862	21	20	-	15
fg08863	22	14	94	10
fg08864	23	15	95	11
fg08866	25	16	-	11
fg08867	26	18	98	13
fg08868	27	19	97	14
fg08870	29	21	-	9
fg08871	30	22	-	-
fg08872	31	23	-	7
fg08873	32	-	-	8
fg08874	33	-	-	12
fg08875	34	-	-	11
fg08876	35	-	-	18
fg08879	38	7	-	16
fg08880	39	8	82	15
fg08881	40	-	87	-
fg08883	42	12	-	20
fg08884	43	11	-	19
fg08885	44	10	86	13
fg08886	45	13	80	21
fg08887	46	17	75	33
fg08888	47	19	-	34
fg08889	48	16	76	-

Tabelle 3.8: Syntenie-Profil für die Contigs fg_contig_1.358, MG_contig_723, an_contig_1.80 und LGI:9a75. Angegeben sind die Ordnungsnummern der orthologen Gene im jeweils zugeordneten Contig.

3.4 Diskussion

Die in diesem Kapitel vorgestellten Methoden verschieben den Fokus der Analyse von der unstrukturierten linearen DNA zum Gen. Das im vorigen Kapitel entwickelte Instrumentarium zur Suche nach konservierten Nachbarschaftsbeziehungen bedarf dazu nur leichter Modifikationen. Auf der Software-Ebene werden zum Teil identische Module verwendet. Die Implementierung des LIS-Algorithmus arbeitet beispielsweise auf einer Liste von generischen 'Redundanz-Objekten'. Diese können sowohl DNA (also etwa TBLASTX-Treffer) als auch Gene repräsentieren, die mit einer Datenbanksuche wie FASTA oder BLASTP ermittelt wurden.

Eine unmittelbare Folge des veränderten Abstraktionsniveaus ist die drastische Abnahme an zu verarbeitenden Signalen. Dies ermöglicht eine weniger restriktive Filter-Strategie: Das Kollinearitätskriterium wurde gelockert (Syntenie-Analyse Arabidopsis-Reis) bzw. ganz fallen gelassen (Untersuchung der Chlamydien und Fungi). Die Ergebnisse zeigen, dass die identifizierten Regionen oftmals keine perfekte Kollinearität aufweisen. Es gibt eine Vielzahl an lokalen Störungen, die zum teil ihrerseits wiederum kollinear verlaufen (etwa die Inversion in der in Abb. 3.6 dargestellten Region). Die den gesamten syntenischen Bereich betreffende Kollinearität wird dadurch bis zur Unkenntlichkeit verschleiert.

Kollinearität ist also in vielen Fällen kaum noch erkennbar. Dennoch gibt es in praktisch allen identifizierten Blöcken zumindest ein 'Skelett' an parallel oder anti-parallel konfigurierten Treffern. Insofern ist der Einsatz von Kollinearitätsfiltern ein legitimes Mittel, wenn man auf Filter zurückgreifen muss. Die Regionen, in denen wir konservierte Syntenie entdecken, wurden von Ereignissen unterschiedlicher Größenordnung geformt. Große Ereignisse schaffen Kollinearität, lokale Ereignisse verwischen sie. Die Frage, wo die Grenze zwischen 'großen' und 'kleinen' Ereignissen zu setzen ist, hängt sicher auch von den Fragestellungen ab, die untersucht werden sollen.

In diesem Zusammenhang stellt sich auch die Frage nach den 'richtigen' Parametern für die Suche nach syntenischen Regionen. Bei der Durchsicht der Ergebnisse stößt man immer wieder auf Fälle, in denen etwa die Veränderung der maximal zulässigen Lücke eine andere Block-Konfiguration ergeben hätte. Eine befriedigende Antwort lässt sich jedoch nicht geben. Allenfalls die genaue Kenntnis über die Häufigkeit und Ausmaß chromosomaler Veränderungen könnte hier konkrete Anhaltspunkte geben. Solche Daten liegen aber in der Regel nicht vor.

Die Suche nach Syntenie zwischen *Parachlamydia* sp. UWE25 und den Chlamydien ist ein gutes Beispiel für die Grenzen von Dot-Plots. Allein aufgrund der Größenunterschiede der Genome fehlt einem solchen Plot die Aussagekraft. Die vorhandenen Signale verlieren sich und fallen kaum ins Auge. Würde man die Darstellung andererseits auf den gemeinsamen Gen-Gehalt beschränken, käme es zu großen Verzerrungen, weil zwei drittel des UWE25-Genoms ausgeblendet wären. Die durchgeführte Analyse zeigt jedoch, dass etwa die Hälfte aller orthologen Gen-Paare in Syntenie-Blöcken lokalisiert ist.

Beim Studium der tabellarischen Ausgabe der Syntenie-Software, die auch die Funktionsbeschreibungen der Gene enthält, fällt immer wieder auf, dass die Gene innerhalb eines Blocks verwandte Funktionen erfüllen (siehe etwa das Profil in Tabelle 3.5 mit dem Oligopeptid-Transporter-Operon). Hier drängt sich die Frage auf, ob diese Ähnlichkeit in der Gen-Funktion für die Gen-Cluster eine entscheidende Rolle spielt. Diese Frage wurde im Ansatz „Nachbarschaft auf funktionaler Ebene" untersucht.

Kapitel 4

Nachbarschaftsbeziehungen auf Funktionaler Ebene

4.1 Einleitung

Die im letzten Kapitel exemplarisch beobachtete Parallelität zwischen syntenischen Bereichen und funktionaler Assoziation der beteiligten Gene soll in diesem Kapitel systematisch untersucht werden. Die daraus resultierende Frage lautet: Gibt es eine Korrelation der genomischen Topologie mit der Funktion der codierten Proteine? Beispiele für funktionale Cluster sind in Prokaryonten seit langer Zeit bekannt. Studienobjekte sind daher in diesem Teil der Arbeit bakterielle Genome.

Im Gegensatz zu den vorangegangenen Kapiteln treten in diesem Fall relevante Nachbarschaften nicht unmittelbar als Mengen konservierter Objekte hervor. Ein Teil der Aufgabe besteht darin, zunächst einmal in Frage kommende Gruppen von Genen innerhalb eines Genoms zu identifizieren. Zur Cluster-Definition werden nun funktionale Attribute der codierten Proteine herangezogen. Nach Betrachtung von DNA- und Gen-Ebene wird jetzt das Abstraktionsniveau nochmals erhöht.

Die Subsumption auf die allgemeine Formulierung konservierter Nachbarschaften ist ähnlich wie im vorigen Kapitel: betrachtet werden Chromosomen, repräsentiert als geordnete Listen vorhergesagter Gene; die Abbildung zwischen Clustern verschiedener Genome über Sequenzähnlichkeit. Die Nachbarschaftsrelation wird über ein Wahrscheinlichkeitsmaß implementiert.

4.1.1 Begriff

Die Existenz von Gruppen von Genen, die auf dem Chromosom nebeneinander liegen und deren Funktion in einem gemeinsamen Kontext steht, ist seit langem bekannt. Bereits in den 50er Jahren wurde entdeckt, dass in den Modell-Organismen *E. coli* und *Salmonella typhimurium* die Gene, die die einzelnen Schritte zur Synthese von Tryptophan codieren, in direkter Abfolge auf dem

Chromosom lokalisiert sind (Yanofski und Lennox 1959),(Demerec und Hartman 1956). Diese und viele weitere ähnliche Formationen werden als lineares Transkript sämtlicher enthaltener Gene in eine einzige mRNA überschrieben. Auf diese Weise erfolgt die koordinierte Expression mehrerer Schritte eines metabolischen Pfades. Jacob und Monod beschrieben 1961 diese Art von Gen-Clustern und führten für sie den Begriff 'Operon' ein (Jacob und Monod 1961).

Operons werden in Prokaryonten häufig, in Eukaryonten nur selten beobachtet. Ausnahmen bilden lediglich Nematoden und Trypanosomen, die molekularen Mechanismen unterscheiden sich hier allerdings deutlich vom Jacob-Monod-Modell (Blumenthal 1998). Immerhin 25% der Gene in *Caenorhabditis elegans* befinden sich laut (Blumenthal und Spieth 1996) in solchen Operons. Eine Studie aus dem Jahre 2003 belegt darüber hinaus, dass Gene eines Pathways in einigen weiteren Eukaryonten signifikant gehäuft vorliegen. Das Ausmaß dieser Ungleichverteilung variiert allerdings stark zwischen den einzelnen Genomen, auch die betroffenen metabolischen Pfade unterscheiden sich (Lee und Sonnhammer 2003).

Eine weitere bekannte Klasse von funktional assoziierten Gen-Clustern sind die sog. Pathogenizitäts-Inseln (kurz PAI). Dabei handelt es sich um benachbarte Gene, die einem Organismus die Fähigkeit vermitteln, Wirtsorganismen zu befallen und sich dort zu reproduzieren. Darunter fallen Funktionen wie das Anheften an oder Eindringen in Wirtszellen (Adhesine, Invasine) sowie das Ausschleusen von Substanzen aus der Zelle (Sekretions-Faktoren). In *E. coli*, *S. typhimurium*, *Vibrio cholerae*, *Helicobacter pylori* und einigen anderen wurden Pathogenizitäts-Inseln identifiziert (Karlin 2001). Mehrere solcher Cluster sind auch im Genus *Listeriae* bekannt. Definitionsgemäß beschränkt sich ihr Auftreten jedoch auf die pathogenen Arten *L. monocytogenes* und *L. ivanovii* (detaillierte Darstellung in (Kreft et al. 2002)).

4.1.2 Biologischer Hintergrund

Zur Erklärung der im vorigen Abschnitt beschriebenen Phänomene gibt es im wesentlichen vier Modelle. Die folgende Darstellung orientiert sich an der ausführlichen Diskussion in (Lawrence und Roth 1996):

Natales Modell Der älteste Ansatz geht davon aus, dass insbesondere Pfade zur Synthese von essentiellen Substanzen durch einen Prozess wiederholter Gen-Duplikation und anschließender Modifikation des Duplikats entstanden sind. Demnach wurde ein bereits existierender Pfad sukzessive als Reaktion auf die Limitierung von Vorläufer-Substanzen in der Umgebung ausgebaut. Operons in modernen Organismen wären demnach unveränderte Relikte der Genom-Historie und spiegelten in ihrer Gen-Abfolge die Entstehung der Pathways wider (daher die Bezeichnung 'Natales' Modell).

Tatsächlich gibt es Beispiele, bei denen sich die Reihenfolge von Syntheseschritten eins zu eins in der Gen-Folge wiederfindet (`trp` und `his` Operons in *S.*

typhimurium). Allerdings kann das Natale Modell allenfalls die Entstehung, nicht aber die Konservierung von Gen-Clustern erklären. In den meisten Fällen ist keinerlei Sequenz-Ähnlichkeit zwischen einzelnen Genen eines Operons feststellbar, so dass die These, dass sich diese aus Tandem-Duplikationen gebildet hätten, nicht haltbar ist. Ein weiteres Gegenargument lieferte die Entdeckung von Gen-Familien: einzelne Vertreter größerer Familien treten in verschiedenen Operons auf. Gene, die auf einen gemeinsamen Vorläufer zurückgehen, werden also unterschiedlich mit anderen kombiniert. Das aber weist darauf hin, dass Operons aus ehemals unabhängigen Teilen quasi zusammengesetzt wurden.

Fisher-Modell Das Fisher Modell führt die Entstehung von Gen-Clustern auf eine Co-Adaption der beteiligten Elemente zurück. Ein optimales Zusammenwirken spezifischer Allele führt zu einer zunehmenden Bindung der entsprechenden Gene. Je näher diese auf dem Chromosom beieinander liegen, desto geringer die Wahrscheinlichkeit, dass sie bei einem Rekombinations-Ereignis voneinander getrennt werden.

Lawrence und Roth (Lawrence und Roth 1996) halten dem vor allem entgegen, dass es einer großen Anzahl von Rekombinationen bedarf, um die Bildung von Clustern im beobachteten Ausmaß zu erklären. Insbesondere in Prokaryonten ist diese Voraussetzung nicht erfüllt.

Coregulations-Modell Diese Erklärung leitet sich direkt aus dem Jacob-Monod'schen Operonmodell ab. Die beiden Forscher hatten gezeigt, dass diese Form der Organisation dem Organismus ein Effizienzsteigerung bei der Transkription vermittelt. Gleichzeitig ist so sichergestellt, dass immer gleich viele mRNA-Moleküle aller für eine Synthese-Kette benötigten Enzyme synthetisiert werden. Daraus ergibt sich ein Selektionsvorteil für Zellen, die über diese Art der Regulation verfügen.

Coregulation kann nach Ansicht der Autoren von (Lawrence und Roth 1996) jedoch allenfalls die Manifestierung von Operons erklären, nicht aber ihre Entstehung. Der Selektionsvorteil ergebe sich erst mit dem finalen Operon, dass Co-Transkription ermöglicht. Für Intermediäre Stadien gelte er nicht. Die Existenz gemeinsam regulierter aber nicht geclusterter Gene zeige außerdem, dass der Selektionsdruck nicht zwingend in Richtung Zusammenlegung der betroffenen Segmente wirke.

In Eukaryonten ist eine effiziente Translations-Initiation abhängig von der Nähe des Start-Codons zur 5'-Cap-Struktur. Das Ribosom löst sich nach dem Ende eines Translationsvorgangs sehr schnell von der mRNA. Diese beiden Faktoren führen dazu, dass die Re-Initiation zur Übersetzung eines zweiten Polypeptids in Eukaryonten sehr ineffizient ist. In der weit überwiegenden Mehrzahl der Fälle codiert daher eukaryontische mRNA nur für ein Protein (Alberts et al. 1995). Der vom Coregulations-Modell postulierte Selektionsvorteil einer koordinierten Expression aller im Operon enthaltenen Gene ist somit in Eukaryonten nicht relevant. Insofern ist das Modell konsistent mit der weitgehenden Abwesenheit von Operons in Eukaryonten.

Eigennützige Operons Schließlich entwickeln Lawrence und Roth in (Lawrence und Roth 1996) eine eigene Hypothese. Zentrale Bedeutung kommt hierin dem horizontalen Gen-Transfer zu (siehe folgenden Abschnitt). Die Anzahl an vollständigen Genen, die dabei übertragen werden kann, ist abhängig vom jeweiligen Mechanismus, in jedem Fall aber begrenzt. Dies ist von großer Bedeutung bei der Weitergabe von Funktionen, die durch mehr als ein Gen codiert sind. Je näher alle erforderlichen Gene beieinander liegen, desto größer die Wahrscheinlichkeit, dass alle gemeinsam transferiert werden; nur dann wird tatsächlich die kombinierte Funktion aller Elemente auf die Rezeptorzelle übertragen.

Die Nähe der Gene auf dem Chromosom nutzt insofern nicht dem Donor, denn die Funktion ist auch bei gestreuten Genen gegeben. Sie begünstigt aber die Chance, als funktionelle Einheit auf den Rezeptor übertragen zu werden und diesem einen evolutionären Vorteil zu vermitteln. Eine weite Verbreitung wiederum verringert die Gefahr, durch Deletionen aus dem genetischen Pool zu verschwinden. Darum sprechen Lawrence und Roth vom eigennützigen Operon („selfish operon“,(Lawrence und Roth 1996).

Einen großen Vorteil ihres Modells sehen die Autoren darin, dass es auch die Bildung von Clustern erklären kann. Demnach führt ein hoher Selektionsdruck zu beschleunigter Deletion der Regionen, die in der transferierten DNA zwischen den gemeinsam agierenden Genen liegen. Die positive Selektion kann nur die neu erworbene Funktion propagieren; führen die zusätzlich erworbenen redundanten Gene zu einer Störung des metabolischen Gleichgewichts im Rezeptor-Organismus, kommt eine negative Selektion verstärkend hinzu. Mit jeder Deletion von Genen, die nicht zu der vom Cluster vermittelten Funktion beitragen, verringert sich der Abstand der kooperierenden Gene. So kommt es zu einer schrittweisen Annäherung dieser Gene, also im Ergebnis zu einem Cluster.

Keines der beschriebenen Modelle kann als universelle Erklärung für alle in bakteriellen Genomen beobachteten Gen-Cluster angesehen werden. Zur Beurteilung, welcher Mechanismus bei der Bildung eines bestimmten Clusters vorherrschend war, ist vor allem die Frage entscheidend, in wie weit er auf horizontalen Gen-Transfer zurückgeführte werden kann (Andersson und Eriksson 2000).

Das Lawrence-Roth-Modell der eigennützigen Operons kommt nur für Funktionen in Betracht, die zumindest zeitweise für die beteiligten Organismen nicht essentiell waren, die also nur unter schwachem Selektionsdruck standen. Demnach sollten lebensnotwendige Funktionen nicht geclustert auftreten. Tatsächlich aber gibt es solche Cluster essentieller Gene: das prominenteste Beispiel sind ca. 40 ribosomale Gene, die in fast allen Genomen in unmittelbarer Nachbarschaft lokalisiert sind (Keeling et al. 1994), (Watanabe et al. 1997). Der ribosomale Cluster ist jedoch in vielfacher Hinsicht eine Besonderheit und vieles spricht dafür, dass sein Ursprung bereits im letzten gemeinsamen Vorfahren aller modernen Organismen liegt (Lawrence und Roth 1996).

Horizontaler Gen-Transfer scheint jedoch an vielen Clustern zumindest beteiligt zu sein. Dies gilt insbesondere für Synthese- und Degradations-Pfade so-

wie für Transport-Systeme, in denen mehrere Gene zusammen arbeiten (ABC-Transporter) (Andersson und Eriksson 2000). Auch für Pathogenizitäts-Inseln bietet lateraler Transfer eine plausible Erklärung, ermöglicht er doch den Erwerb kompletter physiologischer Fähigkeiten auf relativ einfache und vor allem sehr schnelle Weise, i.e. mit nur einer Übertragung kann ein komplettes Funktionsmodul aufgenommen werden (Lawrence 1999).

Horizontaler Gen-Transfer

Die übliche Art der Weitergabe von genetischem Material ist der von der parentalen auf die Filial-Generation, also Vererbung im ursprünglichen Sinne. Dieser Weg wird als vertikaler Transfer bezeichnet. Insbesondere in Prokaryonten ist jedoch auch die Übertragung von DNA zwischen Zellen bekannt, die völlig unabhängig ist von der Reproduktion. Da die beteiligten Individuen hier nicht Mitglieder einer Verwandtschaftsline sind, wird diese Richtung als horizontal oder auch lateral charakterisiert. Dabei wird die Herkunftszelle des genetischen Materials Donor, die aufnehmende Zelle Rezeptor genannt.Im wesentlichen gibt es drei Mechanismen, die dem Austausch von genetischem Material zwischen adulten Zellen dienen (Ochman et al. 2000):

- Konjugation: Zwischen Donor und Rezeptor kommt es zu physischem Kontakt, während dessen autonome DNA-Elemente (zumeist Plasmide) übertragen werden.
- Transformation: Aufnahme nackter DNA aus der Umgebung, erfordert ein entsprechendes System beim Rezeptor („Kompetenz“).
- Transduktion: Als Transportvehikel fungieren Phagen, die bei der Replikation DNA des Donors integrieren und dieses bei Infektion des Rezeptors mit einschleusen.

Ein beobachtbarer Fall von horizontalem Gen-Transfer (HGT) liegt erst dann vor, wenn sich das neu erworbene Material in der Zelle manifestieren kann und (über den vertikalen Weg) an die Nachkommen weitergegeben wird. Zu einer Verbreitung der neu erworbenen Gene in der Population wird es hauptsächlich dann kommen, wenn sie einen Selektionsvorteil vermitteln. Dazu müssen sie jedoch funktionieren, d.h. erfolgreich exprimiert werden (Eisen 2000). Daraus folgt unmittelbar, dass entweder gleichzeitig ganze funktionale Einheiten aus mehreren Genen übertragen werden müssen oder aber solche, die für ihre Funktion kaum auf spezifische Interaktionen angewiesen sind. Für beides gibt es Beispiele.

Metabolische Fähigkeiten (Ochman et al. 2000) oder Pathogenizitäts-Inseln (Hacker et al. 2003) gelten als wichtigste Fälle von gemeinsam übertragenen Clustern, diese stützen das Modell der eigennützigen Operons (s.o.). Auch für sog. Restriktions-Modifikations-Systeme, die zum Teil aus zwei oder drei Genen bestehen, wird Verbreitung über HGT angenommen (Kobayashi et al. 1999), (Kobayashi 2001). Als Beispiele für Gene, die erfolgreich einzeln transferiert

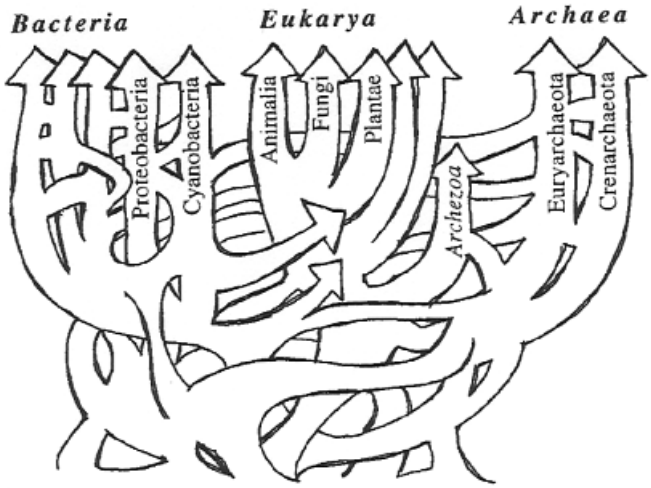


Abbildung 4.1: Der 'Stammbaum' des Lebens unter Berücksichtigung häufiger HGT Ereignisse (entnommen aus (Doolittle 1999)

werden können, wurden früh Amino-Acyl tRNA Synthetasen ausgemacht, die häufig auch zwischen Bakterien und Archaea ausgetauscht werden (Doolittle 1999).

HGT funktioniert grundsätzlich zwischen Individuen der gleichen oder eng verwandter Spezies, aber auch bei Organismen mit großem verwandtschaftlichem Abstand. Lange hielt man diesen Weg der Übertragung genetischen Materials für ein sehr seltenes Ereignis. Das änderte sich mit der Veröffentlichung von Schätzungen, wonach 18% der ORFs in *E. coli* seit Abspaltung von der *Salmonella*-Linie durch horizontalen Transfer erworben seien (Lawrence und Ochman 1998). Eine derartige Häufigkeit von HGT hätte enorme Auswirkungen auf die gesamte Evolutionsforschung, die gängige Sicht auf die Abstammungsbeziehungen als baumartige Struktur wäre in Frage gestellt. Demnach wäre eine Netz-Topologie (vgl. Abb. 4.1) die angemessenere Repräsentation zumindest der Beziehungen zwischen den Prokaryonten, also Bakterien und Archaea (Doolittle 1999).

Die Arbeit von Doolittle eröffnete eine angeregte Diskussion über die Häufigkeit von HGT und die sich daraus ergebenden Konsequenzen (siehe etwa die 'Technischen Kommentare' (Huynen et al. 1999). Die Extrempositionen werden in (Gogarten et al. 2002) bzw. in (Kurland et al. 2003) dargestellt. Erstgenannter Artikel hält HGT für die wichtigste Quelle prokaryontischer Evolution überhaupt, während Kurland et. al. die Frequenz von HGT-Ereignissen vor allem aufgrund methodischer Probleme extrem überschätzt sehen und einen nennenswerten Einfluss auf die Evolution von Genomen bestreiten.

Gewissermaßen eine Synthese-Position formuliert Carl Woese (Woese 1998). Er sieht HGT als treibende Kraft der Evolution in simplen Vorläufern moder-

ner Organismen, den sog. Progenoten. Diese, noch durch sehr unzulänglichen Translationsapparat gekennzeichneten Zellen, hätten praktisch ungehinderten Austausch genetischen Materials betrieben und auf diese Weise kleinste Verbesserungen schnell über die gesamte Population verteilt. Mit zunehmender Verbesserung und steigender Komplexität der zellulären Subsysteme (besonders der Translation) sei es im Verlaufe dieser Entwicklung nach und nach zu Inkompatibilitäten gekommen, die den Gen-Transfer dann immer mehr beschränkten. Woese benutzt eine Analogie zu physikalischen Abkühlungsprozessen und spricht von der beginnenden Kristallisation von Systemen. In der Folge hätten sich zunächst Sub-Populationen gebildet, die intern weiterhin häufig DNA transferierten, von den Fortschritten der anderen Gruppen aber nicht mehr profitieren konnten.

Der letzte gemeinsame Vorfahre allen modernen Lebens ist in dieser Hypothese der gesamte Pool der vollständig kompatiblen Progenoten. Mit den ersten Hindernissen im horizontalen Transfer erscheint der tiefste Verzweigungs-Punkt im Stammbaum des Lebens. Die sich abspaltenden Sub-Populationen sind demnach die Vorläufer der drei großen Domänen Bakterien, Archaea und Eukaryonten (Woese 1998).

In modernen Organismen sieht Woese HGT weiterhin als wichtigen Prozess, nämlich als einzigen, mit dem ein Organismus vollständig neue Fähigkeiten erwerben kann. Die Erhaltung der Universalität des genetischen Codes – ungeachtet etlicher Ausnahmen (siehe Überblick in (Knight et al. 2001)) verwendet die weit überwiegende Anzahl an Organismen das Standard-Schema – könnte sogar auf beständigen lateralen Austausch zurückzuführen sein. Haupteinfluss auf die Veränderungen in Genomen sind seiner Ansicht nach jedoch vertikale Prozesse (Woese 2000). Gestützt wird diese Meinung durch eine Abschätzung der relativen Häufigkeiten der wichtigsten evolutionären Mechanismen (Kunin und Ouzounis 2003). Demnach ist der Verlust von Genen dreimal so häufig wie HGT. Neuerwerb von Genen durch Duplikation mit anschließender Modifikation tritt immerhin noch doppelt so oft wie HGT auf, so dass insgesamt die vertikale Vererbung zwar deutlich deutlich überwiegt, Duplikation und Gentransfer andererseits relevante Mechanismen der funktionellen Expansion von Organismen sind.

4.1.3 Ziele

Ausgehend von den beschriebenen Mechanismen, die zur Bildung und Konservierung von Clustern funktional assoziierter Gene führen können, ergeben sich für eine entsprechende Analyse in prokaryontischen Genomen mehrere denkbare Ansätze:

Quantifizierung funktionaler Assoziation Einige Beispiele für funktionale Cluster sind seit langem bekannt, allen voran die für ribosomale Gene sowie eindeutig identifizierte Operons. Andere, wie etwa PAI oder Restriktions-Modifikationssysteme werden erst seit einigen Jahren intensiv studiert. Sys-

tematische Studien zur Quantifizierung funktionaler Beziehungen genomischer Nachbarschaften gibt es jedoch kaum. Die dazu erforderlichen Daten (vollständig bekannte Genom-Topologien und hinreichende funktionale Annotation (siehe unten Def. 4.2 sind für immer mehr Organismen gegeben (vgl. auch Abschnitt 4.3.1). Diese Art der Analyse ist zunächst rein intragenomisch orientiert.

Funktionsvorhersage Im Umkehrschluss kann das Signal funktionaler Nachbarschaften zur Vorhersage der Funktion unzureichend annotierter Gene genutzt werden. Sobald gezeigt ist, dass ein Cluster bei der Erfüllung einer Aufgabe zusammen wirkt, ist die Hypothese gerechtfertigt, dass auch bislang noch nicht charakterisierte Gene in dieser Gruppe daran beteiligt sind. Für diese Strategie der Funktionsvorhersage wurde der Begriff 'Kontext basiert' geprägt, denn entsprechende Methoden ziehen Inferenzen über die Funktion aus dem genomischen Zusammenhang. Zur Abgrenzung gegenüber klassischen Ansätzen, die auf Sequenz-Ähnlichkeit beruhen, ist auch die Bezeichnung 'non-homology-methods' verbreitet. Eine Darstellung der wichtigsten Methoden aus diesem Bereich folgt im Abschnitt 4.2.

Genomvergleiche Die Ermittlung der funktional assoziierten Cluster liefert eine Beschreibung eines Genoms und damit die Grundlage für den funktionellen Vergleich mit anderen Genomen. Dabei gilt es zunächst, durch Homologiesuche korrespondierende Cluster einander zuzuordnen. Daraus kann man Gemeinsamkeiten und Unterschiede in Umfang, Zusammensetzung und Abfolge der beteiligten Gene bestimmen. Solche Vergleiche liefern gegebenenfalls wichtige Unterstützung für die Funktionsvorhersagen: ein konservierter Cluster hat höhere Aussagekraft bezüglich der gemeinsamen Funktion als einer, der nur in einem Organismus auftritt (vgl. auch Abschnitt 4.3.1).

Phylogenie Cluster, die in sehr vielen Organismen auftauchen, können auch Grundlage phylogenetischer Untersuchungen sein. Gerade angesichts der Diskussion über den Einfluss von HGT auf die Berechnung von Stammbäumen könnte eine auf funktionalen Clustern basierende Metrik von Interesse sein. Bei Gruppen, die durch lateralen Transfer weitergegeben wurden, ist evtl. eine Datierung möglich: je kürzer das Ereignis zurück liegt, desto weniger Zeit stand für Umschichtungen in den Genomen zur Verfügung, desto größer sollten also die Übereinstimmungen in den Clustern sein.

4.2 Literatur

4.2.1 Überblick

Durch die vollständige Sequenzierung von Genomen steht erstmals die gesamte Erbinformation der entsprechenden Organismen zur Verfügung. Zuvor war man auf Kartierungen angewiesen, die bedingt durch die aufwändige Methodik

nur eine sehr beschränkte Menge von Marker-Genen enthalten. Mit dem ersten vollständig aufgeklärten Genom eines Bakteriums (*Haemophilus influenzae* (Fleischmann et al. 1995)) und weiteren Meilensteinen wie dem Hefe-Genom (Goffeau et al. 1996) oder *E. coli* (Blattner et al. 1997) waren seit der zweiten Hälfte der neunziger Jahre Daten verfügbar, die Studien auf Basis der Gesamtheit aller Gene eines Organismus erlaubten.

(Tamames et al. 1997) enthält eine erste systematische Suche nach Clustern, die auf einer umfassenden Analyse benachbarter Gene beruhte. Die zunehmende Anzahl bekannter Genome führte in den folgenden Jahren zur Entwicklung einer ganzen Reihe von Methoden, die auf der Auswertung von genomischer Kontext-Information basieren. Sie unterscheiden sich von früheren Ansätzen dadurch, dass sie nicht auf ein einzelnes Gen, sondern auf ganze Genome fokussieren. Einen Überblick geben einige Artikel aus dem Jahr 2000: (Galperin und Koonin 2000), (Marcotte 2000) oder (Huynen et al. 2000a). Die für die vorliegende Arbeit relevanten Arbeiten lassen sich in drei Klassen einteilen und werden in den folgenden Abschnitten kurz dargestellt.

4.2.2 Konservierte Nachbarschaften

Ausgehend vom Operon-Modell lässt sich folgender Ansatz formulieren: alle Gene innerhalb eines Operons sind in ihrer Funktion gekoppelt, also ist es sinnvoll, Operons zu identifizieren, um Aussagen über die Funktion der betroffenen Gene treffen zu können. Die rein Computer gestützte Identifikation von Operons wird jedoch dadurch erschwert, dass die Abfolge der Gene in der Evolution der Bakterien/Archaea kaum konserviert ist (Mushegian und Koonin 1996). Die Reihenfolge der Gene ändert sich sehr viel schneller als die Primär-Struktur der Proteine: bereits Spezies, deren orthologe Gene noch 50% Identität in der Aminosäure-Sequenz aufweisen, zeigen kaum noch Parallelität in der relativen Abfolge der Gene (Huynen und Bork 1998). Lediglich Cluster von physikalisch interagierenden Proteinen bilden die Ausnahme zu dieser Regel (Mushegian und Koonin 1996).

Eine der ersten Studien, die die systematische Nutzung von konservierter Gen-Reihenfolge beinhaltete, zielte entsprechend in erster Linie auf die Identifizierung von konservierten Interaktionen (Dandekar et al. 1998). Untersucht wurden jeweils drei Dreier-Gruppen von Genomen[1]. Berücksichtigt wurden hier nur Paare oder Cluster von Genen, die in allen drei Genomen einer Gruppe in gleicher Reihenfolge auftreten.

Einen weiteren viel zitierten Ansatz stellten Overbeek et. al. zunächst in (Overbeek et al. 1998), später weiter ausgearbeitet in (Overbeek et al. 1999) vor. Als benachbart gelten hier alle Gene innerhalb eines sog. 'Runs', einer Menge von Genen, die auf dem gleichen Strang codiert sind und in der Start- und Stopp-Codon aufeinander folgender Gene durch maximal 300 bp getrennt sind. Die Methode sucht nun Paare innerhalb eines Runs, für die in anderen Genomen

[1] eine Gruppe von Proteo-Bakterien, eine von Gram-Positiven Bakterien sowie eine Gruppe Archaea

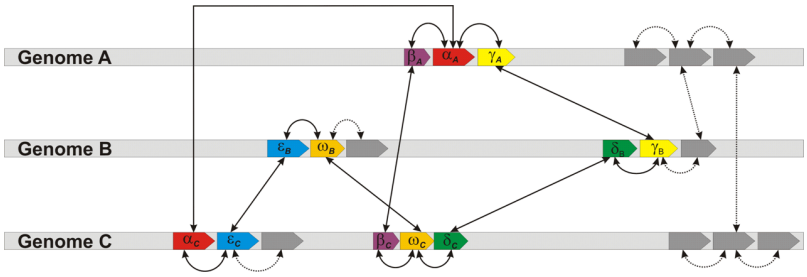


Abbildung 4.2: Beispiel für einen SN-Zyklus: orthologe Gene sind in gleicher Farbe dargestellt und werden durch S-Kanten verbunden, die bogenförmigen Pfeile repräsentieren N-Kanten (Kolesov et al. 2001).

orthologe Gene existieren, die dort ebenfalls innerhalb eines Runs liegen. Die Autoren zeigen, dass solche 'PCBBH' (für 'pair of close bidirectional best hits', also etwa: 'Paar eng benachbarter gegenseitig bester Treffer') in vielen Fällen funktional gekoppelt sind.

Dieser Ansatz basiert darauf, dass Module, die aus mehreren Genen bestehen, die in ihrer Funktion gegenseitig voneinander abhängig sind, über große phylogenetische Distanzen hinweg konserviert sein sollten. Entsprechend werden solche PCBBH von einer nachgeschalteten Gewichtungsfunktion besser bewertet, je geringer die Verwandtschaft der beteiligten Genome.

Das 2001 publizierte SNAP (Kolesov et al. 2001), (Kolesov et al. 2002) (Akronym aus 'Similarity Neighbourhood APproach') nimmt diese Idee auf und verallgemeinert einige Aspekte im Berechnungsprozess. Die Methode operiert auf einer Menge von Genomen, für die die Abfolge der Gene sowie ihre Codierungsrichtungen bekannt sein müssen. Darüber hinaus ist die Kenntnis der Orthologie-Beziehungen zwischen allen involvierten Genomen Voraussetzung.

Die Software betrachtet die Gesamtheit aller Gene als Knotenmenge eines zu ermittelnden Graphen. Als Kanten kommen einerseits Sequenz-Ähnlichkeit zwischen zwei Genen aus unterschiedlichen Genomen (sog. S-Beziehungen), andererseits die Nachbarschaft zweier Gene im gleichen Genom (sog. N-Beziehungen) in Betracht. Ein SN-Graph besteht aus alternierenden S- und N-Kanten, d.h. in einem Pfad wechseln sich intragenomische Schritte immer ab mit solchen, die zwei unterschiedliche Genome verbinden (siehe auch Abb. 4.2).

Ziel ist es, in diesem Graphen zyklische Subgraphen zu finden. Der kleinste denkbare Zyklus beinhaltet vier Gene aus zwei Genomen und entspricht im wesentlichen einem PCBBH wie in (Overbeek et al. 1998) definiert. Innerhalb dieser Zyklen, so die Theorie hinter diesem Ansatz, sind alle involvierten Gene funktional assoziiert. Die Autoren wiesen nach, dass das Auftreten insbesondere längerer SN-Zyklen in hohem Maße unzufällig ist. Vorrangig werden Gene innerhalb metabolischer oder regulatorischer Pfade zu einem zyklischen Subgraphen gruppiert.

Die Ergebnisse von SNAP sind durch die Tatsache limitiert, dass in den N-Beziehungen nur sehr eng benachbarte Gene berücksichtigt werden können. Grund ist, dass es bei der Pfadsuche mit Vergrößern der zulässigen Nachbarschaft zu einem exponentiellen Wachstum des Suchraums kommt. In den veröffentlichten Anwendungen der Methode wurden nur Gene mit einem Abstand von maximal zwei (also der direkte Nachbar und das darauf folgende Gen jeweils in beiden Richtungen) einbezogen.

Auch die Gruppe um Eugene Koonin beschäftigt sich mit dem Phänomen konservierter Gen-Abfolgen. Ein auf Gen-Strings operierender Algorithmus zum lokalen Alignment von Genomen bestätigte frühere Ergebnisse, dass die Reihenfolge von Genen über weitere phylogenetische Distanzen nur sehr schwach konserviert sind. So konnten denn auch nur wenige Ausnahmen von dieser Regel identifiziert werden (Wolf et al. 2001).

Ein weiterer Ansatz aus dieser Gruppe arbeitet mit einem dreistufigen Verfahren (Rogozin et al. 2002): zunächst werden ähnlich wie in (Overbeek et al. 1998) konservierte Gen-Paare ermittelt, die in mindestens drei nicht nah verwandten Genomen in gleicher Reihenfolge auftreten und durch maximal zwei andere Gene voneinander getrennt sind. Im zweiten Schritt werden sog. 'gene-arrays' quasi als transitive Hülle über diese Paare gebildet (d.h. wenn Paare AB und BC existieren, formen ABC ein gene-array). Schließlich folgt ein Clustering über die gene-arrays, das sukzessive jeweils zwei arrays, die eine hinreichend große Schnittmenge an Genen aufweisen, zu einem Cluster vereinigt.

Die sich ergebenden Cluster (im Artikel Gen-Nachbarschaften genannt) sind abstrakte Gebilde, die in der Regel in ihrer Gesamtheit in keinem realen Genom wiederzufinden sind. Allerdings können sie auf die Ausgangs-Genome zurück projiziert werden. Bei der Projektion treten dann Gruppen gleich orientierter Gene (in (Salgado et al. 2000) 'Directons' getauft) hervor, die auch nicht im Cluster enthaltene ORFs beinhalten können.

Gerade diese ursprünglich nicht geclusterten Gene sind von besonderem Interesse. Sie zeigen das Potential der Methode zur Funktionsvorhersage für bislang uncharakterisierte Gene. Bei den identifizierten Nachbarschaften ist meist ein funktionaler Zusammenhang erkennbar (nicht immer sehr strikt, die Autoren sprechen vom 'Thema mit Variationen' (Rogozin et al. 2002). Die so identifizierte Funktion wird als Vorhersage auf die bislang nicht annotierten Elemente im Direkton übertragen.

Allerdings tauchen auch bereits charakterisierte Gene in Nachbarschaften mit völlig unverwandter Funktion auf. Zwei mögliche Erklärungen geben Rogozin et. al. für diese Fälle. (1) Es gibt einen funktionalen Zusammenhang, der jedoch noch nicht bekannt ist. (2) Die Gene haben tatsächlich keine direkte funktionale Verwandtschaft zu ihrer Nachbarschaft, sie nutzen jedoch die von der Umgebung vorgegebene Regulation der Gen-Expression. Demnach handelte es sich hier um eine Art genomisches Schmarotzertum, bei dem sich Gene in existierende Module einfügen, um von der Regulation dieses Moduls zu profitieren

(die Autoren prägen den Begriff 'genomic hitchhiking'[2](Rogozin et al. 2002). Insbesondere bei sehr hohem Expressions-Niveau, wie es etwa bei den ribosomalen Genen der Fall ist, erscheint es plausibel, dass die „Mitbenutzung" der Regulation durch weitere Gene, einen Selektionsvorteil vermittelt. Tatsächlich gibt es einige Funde von potentiellem 'genomic hitchhiking' im Zusammenhang mit dem Apparat für Translation und RNA-Modifikation. Eine Unterscheidung zwischen unbekannter Funktion und Hitchhiking ist nur durch experimentelle Prüfung möglich, sofern keine Daten aus anderen Organismen herangezogen werden können.

Alle Methoden, die auf konservierten Nachbarschaften beruhen, suchen nach Gruppen von Genen, die über große phylogenetische Abstände hinweg in genomischen Clustern auftreten. Die grundlegende Annahme, dass solche Gruppen einen Hinweis auf die funktionale Assoziation der enthaltenen Elemente darstellen, konnten Yanai et. al. mit einer auf 41 Genomen basierenden Analyse stützen (Yanai et al. 2002a). In 80% der Genpaare, die in mindestens drei phylogenetischen Gruppen benachbart auftreten und bei denen beide Gene in KEGG (Kanehisa und Goto 2000) enthalten sind, tauchen beide in einem gemeinsamen metabolischen Pfad auf.

Eine weitere Methode zur Funktionsvorhersage von Genen mit Hilfe von Kontext-Information ist STRING. In seiner ersten Version aus dem Jahre 2000 stützte sich das Programm noch ausschließlich auf konservierte Nachbarschaften. Entsprechend stand das Kürzel STRING damals noch für „Search Tool for Recurring Instances of Neighbouring Genes"(Snel et al. 2000a). Ausgehend von einem vom Benutzer eingegebenen Protein, sucht die Software Gruppen von Genen, die wiederholt in der Nachbarschaft des Anfrage-Gens[3] auftreten.

Der verwendete Algorithmus arbeitet in mehreren Iterations-Stufen. Der erste Durchgang sucht zunächst nur nach Orthologen des Ausgangs-Gen selbst sowie dessen unmittelbarer Nachbarschaft. Diese ist definitionsgleich mit den 'Runs' in (Overbeek et al. 1998) (s.o.). Jedes so ermittelte Gen bildet dann den Ausgangspunkt für die nachfolgenden Schritte. Die Suche endet nach einer vorgegebenen Anzahl von Iterationen oder bei Konvergenz, wenn also kein zusätzliches Gen mehr gefunden wird.

STRING ist in Form eines WWW-Dienstes (STRING) frei verfügbar. Die aktuelle Version integriert zwei weitere kontextbasierte Methoden, nämlich Gen-Fusionen und phylogenetische Profile (von Mering et al. 2003a), die in den folgenden Abschnitten ausführlich dargestellt werden.

4.2.3 Phylogenetische Profile

Eine weitere Gruppe kontextbasierter Methoden lässt sich unter dem Begriff der phylogenetischen Profile zusammenfassen. Ähnlich wie die konservierten Nachbarschaften ist auch dieser Ansatz erst mit einer zunehmenden Anzahl vollständig sequenzierter Genome realisierbar geworden. Als Grundlage dient

[2] sinngemäß etwa: Regulation per Anhalter
[3] genauer: des Anfrage-Protein codierenden Gens

hier die Verteilung orthologer Gene in Organismen. Der Begriff der Orthologie spielt also auch hier wieder eine zentrale Rolle. Ein einfaches System zur Notation der Verteilung eines Gens über die verschiedenen Genome war bereits in der ersten Version der COG-Datenbank (Tatusov et al. 1997) enthalten (hier als 'phylogenetische Muster' bezeichnet), es wurde allerdings noch nicht systematisch zur Funktions-Vorhersage genutzt.

Den grundlegenden Gedanken formulierten u.a. Huynen und Bork in (Huynen und Bork 1998): Proteine, die in ihrer Funktion aneinander gekoppelt sind, sollten entweder in allen Genomen gemeinsam oder gar nicht auftreten, denn bei Verlust des einen kann das andere seine Aufgabe nicht mehr erfüllen. Dies ist entweder letal oder führt mit großer Wahrscheinlichkeit in der Folge zur Deletion des verbliebenen Partners.

(Pellegrini et al. 1999) ist ein erster systematischer Versuch, phylogenetische Profile zur Funktionsvorhersage zu Nutzen. Ausgehend von der Menge aller Gene in *E. coli* wurden Orthologe in 16 anderen Organismen identifiziert. Das Auftreten/Nicht-Auftreten eines Gens wird mit einer 1/0 markiert. Für jedes Gen ergibt sich ein String aus 16 Bit, in dem jede Position fix einem Organismus zugeordnet ist. Schließlich werden die Profile unter Verwendung der Hamming-Distanz[4] nach Ähnlichkeit sortiert, so dass gleiche oder sehr ähnliche Strings nebeneinander angeordnet werden. Die Autoren zeigten, dass vielfach Nachbarschaft im Profil und Funktion korrelieren.

Enault et. al. rücken in (Enault et al. 2003) von der 0/1-Notation ab und verwenden statt dessen einen normalisierten BLASTP Bit-Score. Die Profile können nun als numerische Vektoren aufgefasst werden, die die Verwendung komplexerer Distanz-Funktionen zulassen. Damit erreichen die Autoren eine Effizienzsteigerung der Methode gegenüber (Pellegrini et al. 1999).

Daneben gibt es einige weitere Arbeiten, die phylogenetische Profile nutzen: (Martin et al. 2003) legt den Schwerpunkt allerdings auf den Genom-Vergleich und nicht auf Funktionsvorhersage; (Zheng et al. 2002a) beschreibt eine Kombination aus konservierten Nachbarschaften mit phylogenetischen Profilen, bei der eine Position im Profil nicht mehr für ein einzelnes Gen, sondern für konservierte Gen-Paare steht. Eine weitere Variante stellt Profile für Proteindomänen auf und erreicht so eine höhere Auflösung. Außerdem wird kein rechenaufwändiger all-against-all Vergleich der Gene der betrachteten Organismen benötigt (Pagel et al. 2004).

4.2.4 Gen-Fusionen

Viele Proteine bestehen aus mehreren Untereinheiten, meist Domänen genannt, die klar voneinander unterscheidbare Beiträge zur Gesamtfunktion liefern. Die Zusammenfassung mehrerer Domänen zu einzelnen Proteinen variiert jedoch teilweise zwischen unterschiedlichen Organismen. D.h. eine Funktion, die in einer Spezies A zwei separate Proteinen erfüllen, wird in Art B von einem

[4] Hamming-Abstand = Anzahl unterschiedlicher Positionen

einzelnen Protein umgesetzt. Dieses Phänomen spiegelt sich im Genom wieder: ein Gen in *B* entspricht zweien in *A*. In diesem Fall spricht man von einer Gen-Fusion.

Die Identifikation von Genen, die in anderen Organismen fusioniert sind, stellt zwischen diesen also unmittelbar einen funktionalen Zusammenhang her. Der bereits mehrfach zitierte Artikel (Huynen und Bork 1998) enthält auch einen Hinweis, dass beim Vergleich der Genome von *E. coli* und *H. influenzae* eine Reihe solcher Fälle entdeckt wurden.

In Anlehnung an den berühmten Stein, der zur Entschlüsselung der alt-ägyptischen Hieroglyphen führte, prägte die Gruppe um David Eisenberg den Begriff 'Rosetta Stone' für Fusions-Gene[5] (Marcotte et al. 1999). Die Methode kann einerseits Hinweise auf die Funktion der beteiligten Gene geben (wenn entweder das fusionierte Gen oder eines der beiden separaten Gene bereits charakterisiert sind), andererseits aber auch unbekannte physikalische Interaktionen bekannter Gene aufdecken.

Eine ebenfalls 1999 publizierte Arbeit von Enright et. al. (Enright et al. 1999) verwendete eine etwas andere Methodik, zielte aber ebenfalls auf die Identifizierung von Gen-Fusionen. Hier fand man in den Genomen von *E. coli*, *H. influenzae* und *Methanococcus janaschii* insgesamt 215 an Fusionen beteiligte Gene (2,5%). Auch drei Fälle mit mehr als zwei fusionierten Einheiten wurden identifiziert (Enright et al. 1999). Durch Erweiterung auf insgesamt 24 Genome konnte die Anzahl der beobachteten Genfusionen erwartungsgemäß deutlich gesteigert werden. Der Anteil an Genen, die über Fusionen miteinander in Beziehung gebracht werden können, wird mit durchschnittlich 9% angegeben (Enright und Ouzounis 2001). Durch Clustering aller ermittelter Fusions-Gene auf Basis ihrer Sequenz-Ähnlichkeit (Enright und Ouzounis 2000) konnten die gleichen Autoren insgesamt 621 Gen-Familien ermitteln. Ein hierauf aufbauender phylogenetischer Baum zeigte keine wesentlichen Unterschiede zu gängigen Verwandtschaftsanalysen, was darauf hindeutet, dass der zu Grunde liegende biologische Prozess in allen bekannten Lebewesen verbreitet ist (Enright und Ouzounis 2001).

Durch Einbeziehen der phylogenetischen Beziehungen ist es möglich, zwischen Gen-Fusionen und dem entgegengesetzten Vorgang, der Aufspaltung in zwei Gene, zu unterscheiden (Snel et al. 2000b). Eine Analyse von 17 Genomen zeigte dabei vor allem zwei Ergebnisse: (1) Fusionen sind häufiger als Spaltungen; was zu erwarten ist, da eine Fusion einen Effizienzgewinn mit sich bringt (Marcotte et al. 1999). (2) In thermophilen Organismen gibt es einen Trend zu Spaltungen. Dies ist entweder auf eine erhöhte Mutationsrate zurückzuführen oder unter den extremen Bedingungen, in denen thermophile Spezies leben, gibt es einen Selektionsvorteil für kürzere Gene, möglicherweise aufgrund höherer Fehlerraten während der Protein-Biosynthese (Snel et al. 2000b). Der Anteil an fusionierten Genen korreliert außerdem mit der Genom-Größe (Snel et al. 2000b), (Enright und Ouzounis 2001).

[5] sachlich nicht ganz korrekt: der Stein enthielt identischen Inhalt in unterschiedlichen Codierungen; Gen-Fusionen enthalten komplementären Inhalt in gleicher Codierung

Einen weiteren Beitrag zum Thema Gen-Fusionen liefert (Yanai et al. 2002b). Motivation ist hier nicht die Vorhersage von Gen-Funktionen oder -Interaktionen, sondern die Frage, wie die Verbreitung von Fusions-Genen erklärt werden kann. Die Autoren fanden Hinweise darauf, dass dies vielfach auf horizontalen Gen-Transfer zurückzuführen ist. Das alternative Szenario, nämlich mehrfache unabhängige Fusions-Ereignisse wurde weniger häufig angetroffen (in etwa einem Viertel bis zu einem Drittel der untersuchten Fälle, (Yanai et al. 2002b).

4.2.5 Weitere Methoden

Neben den bereits beschriebenen Strategien gibt es ein weites Spektrum weiterer Methoden, Operons in Genomen zu identifizieren. Ein Ansatz, der dem der konservierten Nachbarschaften nahe steht, ist die Analyse der intergenischen Abstände. Für *E. coli* konnte gezeigt werden, dass durch Kombination von Codierungsrichtung und Abstand die Grenzen der Transkriptionseinheiten mit guter Trefferquote vorauszusagen sind. Durch zusätzliche Einbeziehung funktionaler Klassifikation konnten bis zu 75% der bekannten Einheiten korrekt bestimmt werden (Salgado et al. 2000). Die in *E. coli* gemachten Beobachtungen bezüglich der Gen-Abstände innerhalb von Operons und derer an den Grenzen von Transkriptionseinheiten scheinen in allen Bakterien und auch in den Archaea gültig zu sein (Moreno-Hagelsieb und Collado-Vides 2002).

Die Abstands-Analyse nach (Salgado et al. 2000) funktioniert im Unterschied zu den bisher beschriebenen Verfahren rein intra-genomisch. Eine von Ermolaeva et. al. präsentierte Methode zielt ebenfalls darauf ab, Transkriptionseinheiten durch Identifikation von Gen-Paaren innerhalb gleicher Operons zu berechnen. Allerdings werden hier wieder Genom-Vergleiche herangezogen: je häufiger ein bestimmtes Paar benachbart auftritt und je weniger Verwandtschaft die entsprechenden Genome aufweisen, desto größer die Wahrscheinlichkeit, dass das Paar zum gleichen Operon gehört (Ermolaeva et al. 2001).

Einen rein graphbasierten Ansatz verfolgt die in (Zheng et al. 2002b) vorgestellte Methode. Grundidee ist hier, linearisierte Teil-Graphen aus der KEGG-Repräsentation metabolischer Pfade (Kanehisa und Goto 2000) zu extrahieren und mit Gen-Clustern (hier im Sinne von im Genom aufeinander folgenden Genen) abzugleichen. Algorithmisch anders, jedoch auf dem gleichen Prinzip beruhend, arbeitet das in der Kanehisa-Gruppe entwickelte Verfahren der 'korrelierten Cluster'. Dieses lässt sich nicht nur auf KEGG als Referenz-Graph anwenden, sondern auch auf Graph-Repräsentationen anderer Genome (Ogata et al. 2000).

Yamanishi et. al. entwickeln eine Korrelationsanalyse, die speziell auf Daten angepasst ist, die nicht in Form von Vektoren vorliegen. Ziel ist es, Korrelationen in heterogenen genomischen Datensätzen zu identifizieren, da diese Hinweise auf – möglicherweise versteckte – biologische Phänomene (als Beispiel werden Operons genannt) geben können. Nachfolgend werden die Gene extrahiert, die für die Auffälligkeiten in den Daten verantwortlich sind. In der vorgestellten Anwendung nutzen die Autoren metabolische Pfade, Gen-Expressions-Experimente

und Genom-Architektur als Datenquellen, um Operons zu finden (Yamanishi et al. 2003).

Andere Methoden zur Operon-Vorhersage stützen sich auf Sequenz immanente Signale und fallen daher nicht unter die Kontext basierten Verfahren. Dazu zählen der Einsatz von Hidden Markov Modellen (Yada et al. 1999) oder probabilistischen Cluster-Verfahren zur Identifizierung von regulatorischen Sequenzen (van Nimwegen et al. 2002). Bockhorst et. al. verwenden sowohl Kontext- als auch Sequenz-Daten und verarbeiten diese mit stochastischen kontextfreien Grammatiken (Bockhorst et al. 2003b) oder Bayes-Netzen (Bockhorst et al. 2003a).

4.3 Attribut Cluster

4.3.1 Konzept

Das Problem, das mit der in diesem Abschnitt vorgestellten Methode angegangen werden soll, lässt sich wie folgt skizzieren: Gegeben sei ein Genom, dessen Gentopologie (also die relative Abfolge der einzelnen Gene) bekannt ist. Darüber hinaus muss eine hinreichend vollständige Annotation vorliegen. Gesucht sind signifikante Häufungen von Genen ähnlicher Funktion, also Gruppen von Genen, die in enger Nachbarschaft im Genom auftreten und deren Genprodukte verwandte Aufgaben erfüllen. Für solche Gruppen wird im folgenden der Begriff Attribut Cluster verwendet.

Abgesehen von der (durchaus beabsichtigten) Unschärfe dieser Problem-Charakterisierung ergeben sich unmittelbar drei Fragen:

- Wann ist eine Annotation hinreichend vollständig?
- Wie bemisst sich die Ähnlichkeit in der Funktion von Genen (bzw. Genprodukten)?
- Wie ist eine “enge Nachbarschaft” definiert?

Vollständigkeit der Annotation Unter der Annotation eines Genoms versteht man im Allgemeinen die Zuordnung biologischer Information zu den in einem vorangegangenen Schritt identifizierten potentiellen Genen (meist als ORFs für Open Reading Frames bezeichnet). Im Rahmen von Genom-Sequenzierungsprojekten geschieht diese Zuordnung mit Hilfe von Homologie zu bereits charakterisierten Proteinen.

Im Idealfall eines vollständig annotierten Genoms könnte man also alle Attribut Cluster für alle in der Annotation berücksichtigten Klassen von Information bestimmen. Zur Funktionsvorhersage von Proteinen des Referenzgenoms müsste die Methode in diesem Fall nichts mehr beitragen, dennoch könnten wertvolle Erkenntnisse mit Hilfe inter-genomischer Vergleiche erzielt werden.

Im anderen Extremfall, dem vollständig unannotierten Genom fehlt die Basis zur Anwendung der Methode. Gerade im Bereich der Prokaryonten ist es aber heute möglich, große Teile des Genoms mit hinreichender Verlässlichkeit mit sequenzbasierten Methoden zu annotieren (vgl. auch Abschnitt 4.3.2). Die Attribut Cluster Analyse kann also dann zur Anwendung kommen, wenn die Basis-Annotation abgeschlossen ist. Tendenziell gilt: je vollständiger die Basis-Annotation, desto größer die Wahrscheinlichkeit, tatsächlich vorhandene Cluster zu identifizieren.

Funktionsähnlichkeit Die Architektur des Genoms ist aufgrund ihrer Linearität leicht mit einer recht simplen Metrik zu fassen (selbst wenn Codierungsstrang und Überschneidungen berücksichtigt werden sollen). So kann man den Abstand zweier Gene beispielsweise als die Anzahl von Basen zwischen Startcodon des einen und Stoppcodon des zweiten Gens definieren, oder auch einfach die Differenz in der relativen Position innerhalb des Chromosoms.

Ganz anders verhält es sich mit der biologischen Funktion von Proteinen. Hier handelt es sich nicht um einen topologischen Raum. Für bestimmte Teilaspekte der Proteinfunktion gibt es Metriken, wie beispielsweise den zeitlichen Verlauf des Expressionslevels. Das Problem beim „messen“ der Funktion liegt aber in der Heterogenität funktionaler Klassifikation. Die Menge an Genprodukten unter bestimmten Bedingungen ist nur eine Dimension zur Erfassung der Funktion des betrachteten Gens. Sie sagt nichts aus über andere funktionale Aspekte wie etwa metabolische oder regulatorische Eigenschaften, Interaktionen mit anderen Proteinen, biochemische Charakteristika, Beteiligung an Signal-Transduktions-Pfaden; viele weitere sind denkbar. Angesichts dieser Vielfalt funktioneller Attribute ist es nicht möglich, Proteinfunktion mit einer eindimensionalen Metrik vollständig zu beschreiben.

Dennoch gibt es verschiedene Möglichkeiten, Verwandtschaft in der biologischen Funktion greifbar zu machen. Weit verbreitet ist die Strategie, die Menge aller Proteine in Klassen aufzuteilen, so dass alle Proteine innerhalb einer Klasse eine gewisse Eigenschaft aufweisen. Die Möglichkeiten zur Klassifizierung sind so vielfältig wie die Biologie selbst. Einige Beispiele für solche Schemata:

- Beteiligung an einem bestimmten metabolischen Pfad (Bsp.: Zitratzyklus)
- Lokalisierung innerhalb der Zelle (Bsp.: Zellkern)
- Bindung an bestimmte andere Komponenten (Bsp.: DNA-bindend)
- Beteiligung an einem zellulären Prozess (Bsp.: Transkription)
- Zeitlicher Verlauf der Expression (Bsp.: erhöhtes Expressionsniveau während der Teilungsphase)

Diese Aufzählung deutet bereits an, dass verschiedene Klassifizierungen orthogonal zueinander sind, so dass ein bestimmtes Protein in mehreren Klassen auftauchen kann. Ein Transkriptionsfaktor etwa ist im Zellkern lokalisiert

und bindet an die DNA. Selbst innerhalb eines Klassifzierungsschemas kann es Mehrfach-Zuordnungen geben, etwa wenn ein Protein an mehreren metabolischen Pfaden beteiligt ist. Für die Zwecke der Attribut-Cluster-Identifizierung ist jedoch nur wichtig, dass es möglich ist, einzelne Proteine bestimmten Klassen zuzuordnen. Deshalb kommt prinzipiell jedes Klassifizierungsschema in Frage, für das experimentelle Methoden oder Datenbestände existieren.

Definition 4.1 (Attribut, Attributgruppe, Attributwert)
Im Sprachgebrauch dieser Arbeit wird ein bestimmtes Klassifizierungsschema als Attributgruppe bezeichnet, eine bestimmte Methode der Zuweisung funktionaler Eigenschaften zu Proteinen als Attribut. Auch bestimmte Informationen aus Datenbanken können als Attribut fungieren, wenn sie Wissen bezüglich eines Klassifizierungsschemas repräsentieren und von einzelnen Methoden abstrahieren (siehe Beispiel unten). Der Wertebereich eines Attributs ergibt sich als Menge aller möglichen Bezeichner, die die entsprechende Methode einzelnen Proteinen zuordnen kann.

Beispiel: *Metabolische Pfade* könnte eine Attributgruppe sein, KEGG (Kanehisa und Goto 2000) (genauer: Zuordnung zu metabolischen Pfaden gem. der Datenbank KEGG) ein Attribut, *Zitratzyklus* ein Attributwert. Dem Gen lmo1566 aus *L. monocytogenes* wird dieser Attributwert zugeordnet, ebenso vier weiteren Genen desselben Genoms, allen anderen nicht. Finden sich alle oder eine Teilmenge dieser fünf Gene in enger Nachbarschaft (s.u.) handelt es sich um einen Cluster für den Wert Zitratzyklus[6].

Darüber hinaus gibt es verschiedene Ansätze, dem Raum der Proteinfunktionen unabhängig von bestimmten experimentalen Methoden eine generelle Einteilung zu geben. Bei COG (Tatusov et al. 1997) etwa werden die Proteine in drei Hauptklassen (Informations-Speicherung und -Verarbeitung, Zelluläre Prozesse und Metabolismus) mit insgesamt 16 Unterteilungen eingruppiert. Wesentlich feinere Einteilungen bieten GO (Gene Ontologies) (The Gene Ontology Consortium 2001) und der MIPS-Funktions-Katalog FunCat (Ruepp et al. 2004). Beide Systeme sind hierarchisch strukturiert, ein Protein, dass einer bestimmten Unterkategorie zugeordnet ist, erfüllt also immer auch die Voraussetzungen für alle übergeordneten Klassifizierungslevel.

Sowohl GO wie auch FunCat beinhalten ein sehr breites Spektrum von Kategorien, um die Vielfalt von Aspekten zur Klassifizierung der Proteinfunktion möglichst optimal abzudecken. Nicht alle Kategorien sind im Hinblick auf die Analyse von Nachbarschaftsbeziehungen gleich gut geeignet. Ausgehend von den Operon-Modellen (besonders Coregulations-Modell und Eigennützige Operons) erscheinen Kategorien besonders gut anwendbar, die in der Lage sind, ganze funktionale Module abzubilden. Als Beispiel seien Aminosäure-Synthese-Pfade genannt.

Ungeeignet sind beispielsweise solche Kategorien, die die molekulare Funktionweise beschreiben, also etwa die enzymatische Funktion oder Bindung bestimmter Cofaktoren. Ein Verfahren zur Identifizierung funktionaler Nachbarschaften

[6] Streng genommen sollte man also eher von einem Attribut*wert* Cluster sprechen

sollte aber – unabhängig von der (vermuteten) Eignung – für alle Kategorien in gleicher Weise anwendbar sein.

Klassifizierungssysteme wie FunCat oder GO haben zwei große Vorteile gegenüber Ergebnissen aus Einzel-Experimenten. Zum einen finden sie zunehmend Verwendung in Genom-Sequenzierungsprojekten, zum anderen bieten sie unmittelbar eine Vielzahl von Ähnlichkeitsklassen: jede Ober- wie Unterkategorie mit einer Mindestanzahl an zugeordneten Genen im analysierten Genom kann als Attributwert zur Cluster Analyse verwendet werden. So bilden beispielsweise die 17 Gene in *L. monocytogenes*, denen die FunCat-Kategorie 01.01.09.06.01 (Tryptophan-Biosynthese) zugeordnet ist eine Klasse, für die man eine Cluster Analyse durchführen kann.

Zur Identifizierung von Attribut Clustern wird also keine Ähnlichkeitsfunktion im engeren Sinne für Proteinfunktionen definiert, sondern auf externe Klassifizierungsschemata zurückgegriffen. Innerhalb einer Cluster Analyse gelten alle Mitglieder einer Klasse als (funktions-)ähnlich zueinander und unähnlich zu allen anderen Proteinen. Jede durch eine Kategorie definierte Klasse bildet die Basis für eine Nachbarschaftssuche.

Enge Nachbarschaft Die gegenseitige Lage zweier Gene auf einem Genom lässt sich leicht durch ihren Abstand, also die Anzahl der zwischen beiden gelegenen Genen beschreiben. Daher ist es nahe liegend, die Nachbarschaft mit Hilfe maximaler Abstände zu definieren. Die Festlegung dieses Maximums wirft jedoch Probleme auf. Ein zu kleiner Wert birgt das Risiko, viele tatsächlich vorhandene Cluster zu übersehen. Je unvollständiger ein Genom annotiert ist, desto größer die Wahrscheinlichkeit für falsch negative Klassenzuordnungen. Damit steigt auch die Gefahr, dass ein nicht annotiertes Gen die Identifizierung eines echten Clusters verhindert. Dies trifft insbesondere dann zu, wenn nur unmittelbar benachbarte Gene Berücksichtigung finden (wie bei SNAP und etlichen anderen Verfahren, vgl. 4.2).

Andererseits kann ein zu hoch gewählter Abstandswert zu kombinatorischer Explosion der Möglichkeiten führen (vgl. Erläuterungen zu SNAP in Abschnitt 4.2). Darüber hinaus besteht gerade bei Klassen mit vielen Proteinen ein Risiko weit ausgedehnte Cluster zu finden, die aus relativ wenigen positiven Genen bestehen, deren Abstände jeweils knapp unter dem zugelassenen Maximum liegen.

Mit einem über ein Wahrscheinlichkeitsmaß definierten Nachbarschaftsbegriff lassen sich die genannten Probleme umgehen. D.h. enge Nachbarschaft wird als statistisch signifikante Häufung unter Berücksichtigung von Genom- und Klassengröße interpretiert. Ein weiterer Vorteil dieser Herangehensweise liegt in der Zuordnung eines Erwartungswertes für jeden Cluster. Dieser sog. E-Value bietet einen Schätzwert für die Relevanz eines Clusters, so dass man nicht auf eine 'ganz-oder-gar-nicht-Strategie' angewiesen ist. Der Cutoffwert zum Verwerfen von potentiellen Clustern kann während der Suche relativ großzügig gewählt werden (späteres Filtern der Ergebnisse ist jederzeit möglich), die Gefahr von falsch negativen reduziert sich.

Aus algorithmischen Gesichtspunkten wird nicht auf einen Maximalabstand verzichtet. Dieser dient aber nur zur Festlegung eines look-ahead Wertes während der Initialisierung der Cluster (siehe Algorithmus 4.1). Der Einfluss dieses Parameters auf das Ergebnis ist unbedeutend, weil nicht er zur Definition der Nachbarschaft verwendet wird, sondern der E-Value. Der Maximalabstand kann deshalb auf einen sehr hohen Wert gesetzt werden, ohne damit große Lücken innerhalb der Cluster zu riskieren.

Ermittlung der Attribut Cluster innerhalb eines Genoms

Definitionen

Definition 4.2 (Annotation für Attributwert)
Gegeben sei ein vollständig sequenziertes Genom $G = (g_0, g_1, \ldots, g_{n-1})$ der Größe $|G| = n$. Die Reihenfolge der Gene g_i sei bekannt, die relative Position mit dem Index i gekennzeichnet. Gegeben sei außerdem ein Attributwert V im beschriebenen Sinne (siehe Abschnitt Annotation). Dann sei eine Zuordnung $\mathcal{A}_V : G \mapsto \{0,1\}$ wie folgt definiert:

$$\mathcal{A}_V(g_i) = \begin{cases} 1 & \text{falls } g_i \text{ erfüllt Attributwert} \\ 0 & \text{sonst} \end{cases}$$

Diese Zuordnung heißt Annotation für V in G.

Definition 4.3 (Attribut-Cluster, match, insert, Lücke, Clusterung)
Eine Gruppe von $m > 1$ Genen $C^G_{\mathcal{A}_V} = (c_1, c_2, \ldots, c_m) \subset G$, die auf unten beschriebene Weise ermittelt wurde, heißt Attribut-Cluster in G für $\mathcal{A}_V$. Jedes c_i, dem von $\mathcal{A}_V$ der Wert 1 zugeordnet ist, wird *match* oder *positiv* bezeichnet, alle anderen heißen *insert* oder *negativ*. Die Größe $|C|$ eines Clusters ist als Anzahl aller enthaltener Gene, also sowohl matches wie auch inserts, definiert. Das Auftreten eines oder mehrerer aufeinander folgender inserts innerhalb eines Clusters wird auch *Lücke* genannt.

Die Menge $\mathfrak{C}$ aller in G gefundener Cluster heißt *Clusterung* von G für $\mathcal{A}_V$.

Statistisches Modell Das statistische Modell soll eine Abschätzung der Relevanz eines (potentiellen) Clusters geben. Es soll schnell berechenbar sein, denn während der Suche wird die Funktion vielfach aufgerufen (vgl. Algorithmus 4.1). Aus dem gleichen Grund muss der Wert für jedes Einzelcluster berechenbar sein; Modelle, die nur Gesamtclusterungen berücksichtigen, kommen also nicht in Frage.

Die Modellierung der Funktionsähnlichkeit mit Hilfe binärer Werte (ein Gen bzw. Protein trägt die gesuchte Eigenschaft oder nicht) ermöglicht die Übertragung des klassischen Urnenmodells: gegeben eine endliche Menge von Objekten (Losen), die bezüglich einer gewissen Eigenschaft in zwei disjunkte Klassen geteilt ist (Treffer/Nieten). Die Gesamtzahl der Treffer ist bekannt. Die Analogie

liegt auf der Hand: das Genom entspricht der Objektmenge (alle Lose in der Urne), die merkmalstragenden Gene den Treffern.

Die Frage nach der Wahrscheinlichkeit, bei n Zügen k Treffer zu erhalten, errechnet sich nach der Hypergeometrischen Verteilung:

$$P(k) = \binom{pN}{k}\binom{N(1-p)}{n-k} / \binom{N}{n} \tag{4.1}$$

Dabei bezeichnet N die Anzahl aller Objekte, p die relative Häufigkeit der Treffer (also die Anzahl aller Treffer dividiert durch N), n die Größe der Stichprobe und k die Anzahl der Treffer in der Stichprobe. In der Anwendung kann ein identifizierter Cluster als Stichprobe aufgefasst werden, n entspricht also der Clustergröße, k ist dann die Anzahl der matches innerhalb des Clusters.

Die Hypergeometrische Verteilung stellt einen Grenzfall der Binomialverteilung dar:

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4.2}$$

Im Urnenmodell ist die Hypergeometrische Verteilung auf den Fall „ohne Zurücklegen der Lose nach jedem Zug“ anwendbar, die Binomialverteilung gilt im Prinzip für die Situation mit Zurücklegen. Bei größer werdender Gesamtmenge und relativ kleiner Stichprobengröße fällt das Zurücklegen jedoch nicht mehr ins Gewicht, so dass unter diesen Umständen die Binomialverteilung die angemessenere Modellierung bietet. Gemäß (Stöcker 1993) liegen diese Grenzen bei $N \geq 2000$ und $n/N \leq 0,1$. Für die Anwendung auf die Attribut Cluster sind diese Grenzwerte in den in dieser Arbeit betrachteten Organismen meist überschritten: Größen von 2000 und mehr Genen weisen mit einer Ausnahme alle verwendeten Genome auf. Lediglich *H. pylori* liegt mit 1576 Genen darunter (vgl. Tabelle 4.2). Die Cluster sind in aller Regel wesentlich kleiner als ein Zehntel des Genoms.

Die Werte (genannt P-Values), die auf diese Weise den Clustern zugeordnet werden, sind zu interpretieren als die Wahrscheinlichkeit, eine identische Konfiguration in einem vollständig ungeordneten Genom (Zufallsgenom) der gleichen Größe mit der gleichen Anzahl an positiven Genen anzutreffen. Je geringer dieser Wert, desto höher ist die Relevanz des Clusters einzuschätzen.

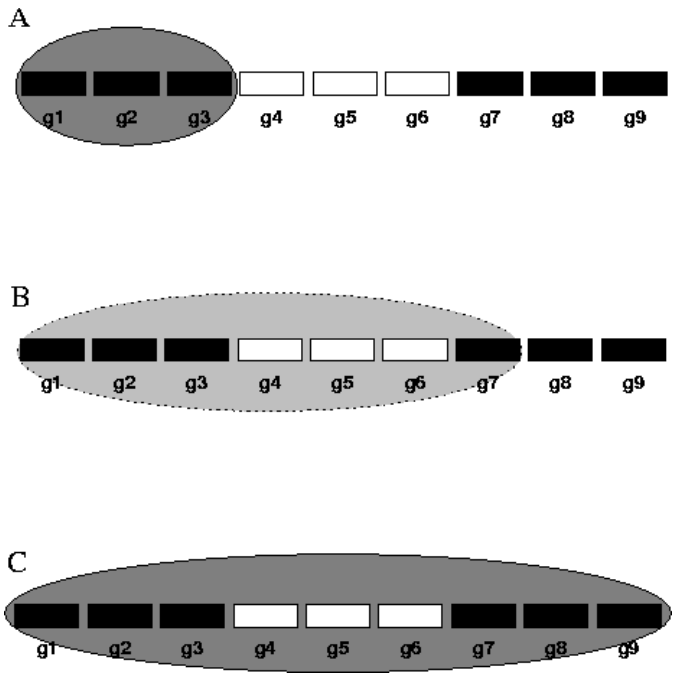


Abbildung 4.3: Initialisierung der Cluster. A: Die Gene g1-g3 wurden zu einem vorläufigen Cluster zusammengefasst. B: Bei Hinzufügen von g7 würde sich das Verhältnis Positive/Clustergröße ungünstig entwickeln, der Cluster wird daher zunächst abgeschlossen. Stattdessen wird in Phase 1 ein weiterer Cluster g7-g9 gebildet (nicht gezeigt). C: Erst in Phase 2 werden die beiden vorläufigen Cluster dann zu einem vereinigt.

Algorithmus Seien Genom G und Annotation $\mathcal{A}_V$ eines Attributwertes V wie oben definiert gegeben. Daraus lässt sich unmittelbar die Positionsliste der matches $P = \{i \mid 0 <= i < |G|, \mathcal{A}_V(g_i) = 1\}$ ableiten.

Nach aufsteigender Sortierung der Positionsliste gliedert sich der angewendete Algorithmus in zwei Phasen. In der ersten werden initiale Cluster gebildet, indem benachbarte Elemente der Liste zu Clustern vereinigt werden, sofern ihr Abstand unterhalb des festgelegten Maximums liegt und der P-Value sich durch das Hinzufügen des nächsten Elementes verbessert. Der P-Value verbessert sich mit der Anzahl der matches, gleichzeitig wirkt sich aber eine zunehmende Clustergröße ungünstig auf den Wert aus. Daher kann es zu Situationen kommen, in denen ein vorläufiger Cluster zunächst abgebrochen wird, obwohl die Einbeziehung weiterer Matches einen verbesserten P-Value bringen würde (vgl. Abb. 4.3).

Darum wird eine zweite Phase nachgeschaltet, in der benachbarte Cluster vereinigt werden – wiederum unter der Bedingung, dass der P-Value des neuen Clusters besser ist als der der beiden Ausgangscluster. Dieses Verschmelzen benachbarter Cluster wird wiederholt bis die Clusterung unverändert bleibt.

Formulierung des Algorithmus in einer halb-formalen Notation[7]:

Algorithmus 4.1

```
INPUT: P, D_max
Externe Funktion eval
Sortiere P aufsteigend
Phase 1: Initiale Clusterermittlung
p_0 := shift P
𝔈 := {}
Initialisiere C_temp := {p_0}
WHILE P ≠ ∅
      p_i := shift P
      p_last := tail C_temp
      IF ( p_i − p_last > D_max AND
           eval(C_temp ∪ {p_i} ) < eval(C_temp) )
           C_temp := C_temp ∪ {p_i}
      ELSE
           𝔈 := 𝔈 ∪ {C_temp}
           C_temp := {p_i}
Phase 2: Cluster Merge
FOREACH Cluster C_i der Clusterung 𝔈
      C_merge := C_i ∪ C_i+1
      e_merge := eval(C_merge)
      IF ( e_merge < eval(C_i) AND
           e_merge < eval(C_i+1) )
          𝔈 := 𝔈 \ {C_i, C_i+1}
          𝔈 := 𝔈 ∪ {C_merge}
REPEAT Phase 2 bis 𝔈 konvergiert
```

Phase 2 führt immer zur Konvergenz von $\mathfrak{C}$, da ausschließlich Merge-Operationen angewendet werden. Einmal vereinigte Cluster werden nicht mehr getrennt. Somit verringert sich die Anzahl der Cluster in $\mathfrak{C}$ in jedem Iterationsschritt um mindestens 1 oder das Verfahren stoppt. Im äußersten Fall werden alle Cluster zu einem vereint, spätestens dann konvergiert die Clusterung.

Zur Abschätzung der Komplexität des Algorithmus muss man die beiden Phasen einzeln betrachten. In Phase eins wird jede Position aus der Eingabeliste einmal betrachtet, dieser Teil verhält sich also linear bezogen auf die Länge der Positionsliste. Für Phase zwei ist die Anzahl der initialen Cluster $k = |\mathfrak{C}_0|$ Ausschlag gebend. Im worst case wird Phase zwei $(k-1)$ Mal durchlaufen, in jedem Schritt verringert sich die Anzahl der Cluster um 1. D.h. es sind zwischen $(k-1)$ im ersten Durchlauf und 1 (im letzten Iterationsschritt) mögliche Clustervereinigungen zu prüfen. Daraus ergibt sich ein Laufzeitverhalten quadratisch bezüglich k. Insgesamt ergibt sich eine Abschätzung von $O(|P| + k^2)$.

[7] Der Operator `shift` verhalte sich wie die gleichnamige Funktion von PERL: angewendet auf eine nichtleere Liste gibt sie deren erstes Element zurück und löscht dieses gleichzeitig aus der Liste. Der Operator `tail` liefert das letzte Element einer Liste – ohne es aus aus der Liste zu entfernen. *eval* berechnet den P-Value eines Clusters.

Der beschriebene Algorithmus ist den agglomerativ hierarchischen Clusterverfahren zuzurechnen[8], da einmal gebildete Cluster nicht mehr geteilt werden. Allerdings garantiert er keine optimale Clusterung. Beurteilt man die Güte einer Clusterung anhand der P-Values (also etwa die Summe der negativen dekadischen Logarithmen der P-Values aller Cluster, geteilt durch die Anzahl der Cluster), kann die strikte Vorgehensweise von links nach rechts möglicherweise suboptimale Clustervereinigungen zur Folge haben. Ein einzelner Treffer, der zwischen zwei Clustern liegt, wird immer dem linken zugeordnet (sofern sich dessen P-Value verbessert). Unter Umständen könnte aber die Vereinigung mit dem rechten Cluster global gesehen günstiger sein. Auf das Endergebnis hat diese suboptimale Zuordnung aber nur dann Einfluss, wenn die beiden Cluster nicht ohnehin in einer nachfolgenden Iteration vereinigt werden. Die möglichen Einbußen sind jedoch zu vernachlässigen, da in der praktischen Anwendung nur sehr selten unterschiedliche Mergealternativen möglich sind. Außerdem führt ein suboptimaler Merge nicht dazu, dass ein existierender Cluster verschluckt wird. Ziel des Verfahrens ist letztlich auch nicht die globale Optimierung der P-Values, sondern die Identifizierung biologisch relevanter Cluster.

Abbildung 4.4 fasst die Suche nach Attribut-Clustern innerhalb eines Genoms nochmal schematisch zusammen.

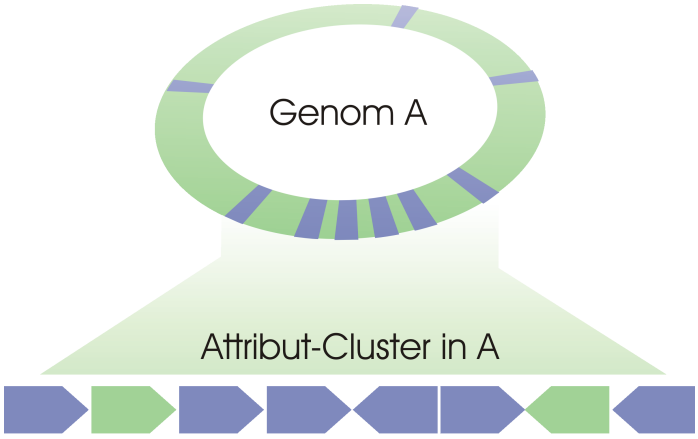


Abbildung 4.4: Schema Attribut-Cluster: der grüne Ring stellt ein zirkuläres Genom A dar. Gesucht wird nach einer bestimmten Klasse von Genen (blau markiert); es gibt einige verstreute Vorkommen sowie einen Cluster. Die Ausschnitt-Vergrößerung zeigt, dass dieser aus fünf positiven sowie zwei inserts besteht. Pfeilrichtungen deuten die Codierungsrichtung an.

Intergenomischer Vergleich

Das bislang beschriebene Verfahren arbeitet rein intra-genomisch. Von großem Interesse ist jedoch auch der Vergleich zwischen verschiedenen Genomen auf

[8] Im Unterschied zu den meisten Vertretern dieser Klasse wird das Clustering jedoch nicht weitergeführt bis nur noch ein Cluster besteht, sondern bricht in der Regel vorher ab

Basis der identifizierten Attribut-Cluster (vgl. auch Abschnitt 4.1.3). Angenommen, für ein Genom A wurden alle Cluster für ein Attribut $\mathcal{A}$ ermittelt. Zum Vergleich mit einem weiteren Genom B bietet sich dann folgendes Verfahren an: (1) ermittle die Clusterung für $\mathcal{A}$ in B, (2) bestimme Paare (C_i^A, C_j^B) einander entsprechender Attribut-Cluster.

Die Zuordnung von Clustern aus den unterschiedlichen Genomen kann mit Hilfe von Homologie basierten Methoden wie BLAST oder FASTA, angewendet auf die einzelnen Gene innerhalb der Cluster, erfolgen.

Definition 4.4 (Cluster-Zuordnung)
Gegeben seien zwei Genome A, B sowie die Clusterungen $\mathfrak{C}^A$ und $\mathfrak{C}^B$ gemäß Definition 4.3 in diesen Genomen für ein beliebiges Attribut. Sei außerdem $f : A \times B \mapsto \mathbb{R}^+$ eine Funktion, die zwei Genen einen Ähnlichkeitswert zuordnet.

Zwei Attribut-Cluster $C_i^A = \{a_1^i, \ldots, a_m^i\} \in \mathfrak{C}^A$ und $C_j^B = \{b_1^j, \ldots, b_n^j\} \in \mathfrak{C}^B$ können einander zugeordnet werden, wenn es mindestens ein Paar (a_k^i, b_l^j) gibt mit $a_k^i \in \mathcal{C}_i^A$ und $b_l^j \in \mathcal{C}_j^B$ und $f(a_k^i, b_l^j) \geq t$, wobei t ein festzusetzender Mindestwert für die Ähnlichkeit ist.

Je nach konkreter Anwendung sind Varianten dieser Definition denkbar. Beispielsweise wäre bei der Analyse möglicher HGT-Ereignisse eine Zuordnung aufgrund eines einzelnen homologen Gen-Paares zu wenig restriktiv. Definition 4.4 kann außerdem zu Mehrfach-Zuordnungen führen, unter Umständen ist jedoch eine klare 1:1 Abbildung wünschenswert.

Abbildung 4.5 zeigt schematisch den Idealfall einer solchen Zuordnung. Fast alle enthaltenen Treffer-Gene haben ein Pendant im Vergleichs-Cluster. Außerdem erstreckt sich die Homologie sogar auf ein Paar von Insert-Genen. Das erhöht die Verlässlichkeit der Attribut-Cluster basierten Funktionsvorhersage beträchtlich: ein Zusammenhang zwischen Funktion und Nachbarschaft wird umso wahrscheinlicher, je häufiger eine entsprechende Clusterung angetroffen wird. Die Aussagekraft steigt dabei außerdem mit phylogenetischem Abstand der verglichenen Organismen.

4.3.2 Ergebnisse

4.3.3 FunCat in Bakteriellen Genomen

Versuchsbeschreibung

Der MIPS Funktions-Katalog (Mewes et al. 2000) („FunCat“) stellt ein Klassifizierungsschema im Sinne der Definition 4.1 dar. Er erfüllt somit die Voraussetzung, um als Basis für die Berechnung von Attribut-Clustern in bakteriellen Genomen zu fungieren. Jede einzelne Kategorie definiert eine Klasse von Proteinen und kann somit als Attributwert Grundlage einer Suche nach Attribut-Clustern genutzt werden. Dank des streng hierarchischen Aufbaus des FunCat gilt das gleiche für alle Oberkategorien.

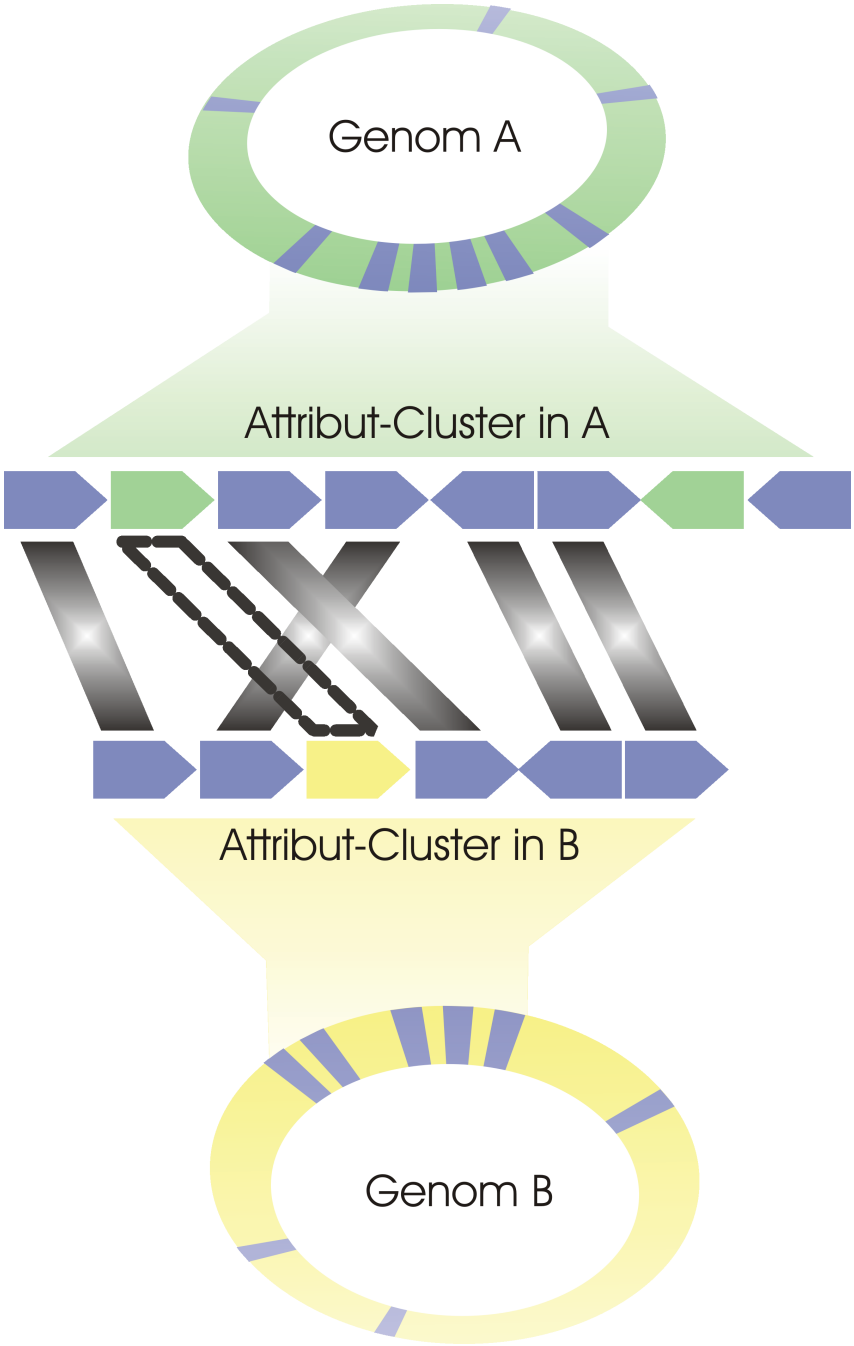


Abbildung 4.5: Schema Intergenom-Vergleich: Der obere Teil des Schemas entspricht Abbildung 4.4. Im gelb dargestellten Vergleichsgenom wurde ebenfalls ein Attribut-Cluster gefunden. Die schwarzen Balken im Zentrum der Graphik symbolisieren Homologie-Beziehungen zwischen den geclusterten Genen. Für die Funktionsvorhersage besonders interessant sind konservierte insert Gene (gestrichelter Balken).

Der Funktions-Katalog wird als Schema für die Annotation von kompletten Genomen benutzt. Für die Methode der Attribut-Cluster bietet dies den Vorteil, das mit einer Versuchsserie eine umfassende Beschreibung des Genoms bezüglich signifikanter funktionaler Nachbarschaftsbeziehungen erstellt werden kann. Diese Serie von Anwendungen muss dazu lediglich alle in einem Genom verwendeten Kategorien und Oberkategorien einbeziehen.

Es liegt nahe, für solch eine Versuchsserie zunächst auf eine qualitativ möglichst hochwertige Datenbasis zurückzugreifen. Ausgewählt wurden daher manuell annotierte bakterielle Genome. Bis zum Ende des Bearbeitungszeitraums der Arbeit (Februar 2004) standen dafür sechs verschiedene Genome zur Verfügung (Details siehe Tabelle 4.1).

Spezies	Stamm	Annotation
L. monocytogenes	EGD	MIPS
L. innocua	Clip11262	MIPS
H. pylori	26695	MIPS
B. subtilis	168	Biomax
E. coli	K12	Biomax
Parachlamyidium	sp UWE25	TU München[9]

Tabelle 4.1: Verwendete Genome

Die Daten für die MIPS-annotierten Genome wurden dem Pedant System (Pedant) entnommen. Die übrigen Daten waren noch nicht öffentlich verfügbar und wurden mir dankenswerte Weise von der Firma Biomax Informatics AG, Martinsried bzw. von der TU München, Lehrstuhl für Genomorientierte Bioinformatik zur Verfügung gestellt.

Parameter

Statistisches Modell Grundsätzlich wurde aus den oben beschriebenen Gründen die Binomialverteilung gegenüber der Hypergeometrischen Verteilung vorgezogen. Zur Bewertung der Relevanz eines ermittelten Clusters wurde das Verhältnis des Wahrscheinlichkeitswertes nach der Binomialverteilung P_{binom} zur Hintergrundwahrscheinlichkeit (berechnet als Häufigkeit des Auftretens eines Attributs dividiert durch die Genomgröße: $P_{hintergrund} = N/G$) verwendet.

$$PValue = P_{binom}/P_{hintergrund} \tag{4.3}$$

Damit wird auch ausgeschlossen, dass ein sehr seltenes Attribut mit einem einzelnen Auftreten bereits einen gültigen Cluster bilden kann.

Maximale Lücke Wie in Abschnitt 4.3.1 dargestellt, ist eine Beschränkung der Lücke aus algorithmischen Gründen erforderlich, muss allerdings nicht sehr

[9] In Zusammenarbeit mit MIPS

restriktiv definiert werden. In den im folgenden beschriebenen Versuchen wurde dieser Wert auf zehn Gene festgelegt, da es unwahrscheinlich erscheint, dass konservierte Gruppen noch größere Abstände untereinander aufweisen.

P-Value Dieser Wert setzt eine Qualitätsschranke. Cluster mit schlechterer relativer Wahrscheinlichkeit werden verworfen. Um die Anzahl falsch vorhergesagter Cluster gering zu halten, wurde hier ein Maximumwert von 1e-03 verwendet.

Minimale Clustergröße Cluster, die weniger als drei Treffer in einer funktionellen Klasse aufweisen, werden in den Ergebnissen nicht berücksichtigt.

Klassengröße Durch die minimale Clustergröße ergibt sich bereits, dass Attribute, die weniger als drei Mal im Genom auftreten, keine Ergebnisse liefern können. Andererseits legt die Konzeption der Attribut-Cluster auch nahe, keine zu großen Klassen zu verwenden; je häufiger ein Merkmal, desto geringer seine Aussagekraft. Die Verwendung von Oberkategorien kann jedoch zu sehr großen Klassen führen. Bei der Festlegung der oberen Schranke sollte berücksichtigt werden, dass bei der Entscheidung gegen die hypergeometrische Verteilung die Stichprobengröße (hier also: Clustergröße) eine entscheidende Rolle spielt. Demnach sollte ein Cluster ein Zehntel des Genoms nicht überschreiten. Daher liegt es nahe, nur Klassen mit maximal $G/10$ Genen zu verwenden.

Statistik

Die Tabelle 4.2 fasst die Ergebnisse der Experimente in einigen Kennzahlen zusammen. Bei der Anzahl der Cluster ist zu berücksichtigen, dass verschiedene Versuche identische Cluster liefern können. Häufig tritt dies zwischen Ober- und Unterkategorien auf: jeder Cluster der Kategorie $x_1.x_2.x_3$ ist zwangsläufig ebenso ein Cluster für die Oberkategorie $x_1.x_2$. Liegen in der Nachbarschaft weitere Gene, die zwar der Ober- nicht aber der Unterkategorie zugeordnet sind, ergeben sich nicht-identische aber überlappende Cluster. Die Angaben in der Tabelle beinhalten solche Überlappungen, identische Cluster sind jedoch nur einmal gezählt.

Organismus	Genom Größe	Gene Annotiert	Attribut Cluster	Summe Positive	Positive ÷ Annotierte Gene [%]	Summe Inserts
L. monocytogenes	2846	1546	176	587	37.97	246
L. innocua	2968	1549	184	644	41.58	353
H. pylori	1576	870	57	248	28.51	165
B. subtilis	4106	2790	335	1296	46.45	457
E. coli	4289	2824	478	1406	49.79	718
Parachlamyidium	2031	755	95	281	37.22	191

Tabelle 4.2: Cluster Statistik

Neben der Anzahl der gefundenen Cluster ist vor allem die Summe der Positiven über alle Cluster von Interesse. In einem vollständig bekannten Genom wäre dieser Wert ein Maß für die Häufigkeit der Kopplung zwischen Proteinfunktion und genomischer Nachbarschaft. In nur teilweise annotierten Genomen, wie sie in den vorliegenden Versuchen analysiert wurden, ist die Summe der Positiven durch die Anzahl der Gene mit vorliegender Annotation begrenzt. Die sechste Spalte stellt darum den prozentualen Anteil der annotierten Gene dar, die als Treffer in mindestens einem Cluster auftreten. Für die meisten Genome liegt dieser Wert bei 40% (± 3 Punkte), *H. pylori* liegt als einziges Genom mit gut 28% deutlich darunter, *B. subtilis* liegt mit gut 46% etwas darüber. *E. coli* liefert mit fast 50% den Spitzenwert.

Die letzte Spalte listet die Summe der Inserts über alle gefundenen Cluster auf. Gezählt werden hier nur Gene, die *ausschließlich* als Insert auftreten. Nicht enthalten sind also solche Gene, die mindestens einmal auch als Treffer in einem Cluster enthalten sind. Damit ist dieser Wert eine obere Schranke für die Anzahl an Funktionsvorhersagen, die mit Hilfe der Attribut-Cluster generiert werden. Erwartungsgemäß korreliert dieser Wert recht gut mit der Summe der Positiven im Sinne der Definition 4.3.

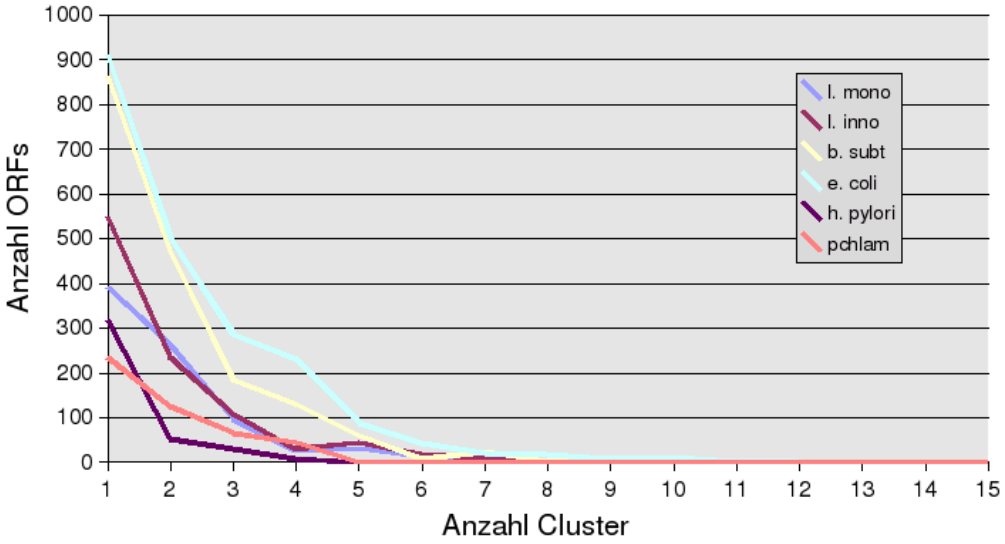


Abbildung 4.6: Überlappungen der Cluster in den einzelnen Genomen gemessen als Anzahl der Cluster, denen ein bestimmter ORF zugeordnet ist.

Den Grad der Überlappungen zwischen Clustern zeigt Abb. 4.6. Als Überlappung wird ein ORF bezeichnet, der mehreren nicht-identischen Clustern zugeordnet ist. Aufgezeichnet ist auf der X-Achse die Anzahl an Clustern pro ORF. Ein X-Wert von 1 kennzeichnet ORFs, die nur einem Cluster angehören (also keine Überlappung darstellen). Die Y-Koordinate zeigt die Anzahl von ORFs in jedem untersuchten Genom, die den einzelnen Überlappungsklassen zuzuordnen sind.

Mit Ausnahme von *E. coli* liegen in allen Genomen mind. die Hälfte (Listerien: knapp darunter) aller in Attribut-Clustern auftretenden ORFs in nur einem Cluster, drei Viertel in maximal zwei verschiedenen Clustern. Bis zu vierfache Zuordnung kommt in allen untersuchten Genomen vor. In den meisten Fällen sind diese hohen Werte auf den beschriebenen Effekt zwischen Ober- und Unterkategorien zurückzuführen.

Extreme Überlappungen resultieren oft aus Redundanzen innerhalb des FunCats: manche Funktionalitäten betreffen viele unterschiedliche Teilhierarchien innerhalb des Kategorienschemas. Als Beispiel sei aus *B. subtilis* der Bereich zwischen den ORFs mit den Positionen 1617 (bg10237) und 1645 (bg10750) genannt. Hier häufen sich Gene, die bei zur Bildung des Flagellums benötigt werden. Gleichzeitig haben sie damit eine Funktion für die Chemotaxis. Diese Bereiche sind im FunCat mehrfach abgedeckt:

- 13.05.07 ciliary/flagellar motility
- 13.11.03.03 chemotaxis
- 40.32 flagellum
- 30.32 flagellum

Hinzu kommt, dass einige dieser Gene auch noch den Kategorien '10.01.01 unspecified signal transduction' sowie '63.01 protein binding' zugeordnet sind. Im Ergebnis sind sieben ORFs insgesamt neun verschiedenen Clustern zugeordnet (dies ist der Maximalwert für *B. subtilis*).

Die Kurven der Abb. 4.6 sind konsistent mit den Absolutzahlen der annotierten Gene. Notwendige Voraussetzung für das Finden eines Clusters ist eine ausreichend dichte Annotation. Je mehr Gene also annotiert sind, desto mehr Cluster können gefunden werden. Damit steigt auch der Erwartungswert für Überlappungen zwischen den Clustern. Auffallend sind jedoch die sehr hohen Werte für drei- und vier-faches Auftreten in der Kurve für *E. coli*. Diese haben auch zur Folge, dass in *E. coli* nur ca. 43% (andere Genome: um 50%) der ORFs in nur einem bzw. zwei Drittel (andere Genome: >= 75%) in maximal zwei Clustern enthalten sind. In diesem Genom sind einzelne ORFs bis zu zwölf Clustern zugeordnet.

Der Grund dafür dürfte darin liegen, dass die Zahl an Mehrfach-Zuordnungen wesentlich größer ist als etwa in *B. subtilis*, das in Genomgröße und Anzahl annotierter Gene ja durchaus vergleichbar ist (siehe Tabelle 4.2). Insgesamt gibt es jedoch in *E. coli* 23.150 verschiedene FunCat-Zuordnungen, das sind über 30% mehr als in *B. subtilis*. Besonders augenfällig ist etwa die Handhabung der Kategorie '63 PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)' mit seinen Unterkategorien. In *E. coli* gibt es hier insgesamt 1229 Zuordnungen, in *B. subtilis* lediglich 135. In sehr vielen Fällen wird diese Kategorie als zusätzliche Beschreibung gewählt (die Tatsache dass, ein Protein eine Bindung eingeht, sagt noch nichts über seine eigentliche Funktion aus). Die ribosomalen Gene zum Beispiel sind in *E.*

coli nicht nur '05.01.01 ribosomal proteins' zugeordnet, sondern auch '63.03.03 RNA binding' — in *B. subtilis* ist dies nicht der Fall.

Zu solchen systematischen Unterschieden in der manuellen Annotation kommt es fast zwangsläufig, wenn sie von unterschiedlichen Gruppen ausgeführt werden. Jede Zuordnung einer Kategorie zu einem ORF durch einen Annotator stellt eine fachliche Bewertung der vorliegenden Evidenzen dar. Um zumindest innerhalb eines Annotationsprojektes einheitliche Standards zu gewährleisten, werden gewisse Richtlinien aufgestellt. Unterschiedliche Richtlinien verschiedener Teams können dann zu den beobachteten Differenzen führen.

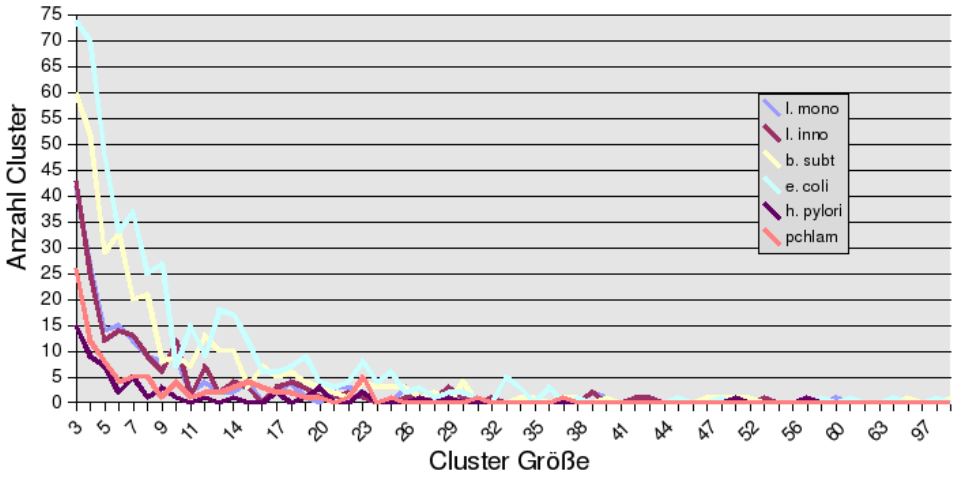


Abbildung 4.7: Größenverteilung der Cluster in den einzelnen Genomen

Die Grafik in Abb. 4.7 zeigt die Verteilung der Clustergrößen in den Genomen. Der überwiegende Teil der Cluster besteht aus drei bis sieben ORFs, in allen Genomen gibt es aber auch wesentlich größere Module. Der größte identifizierte Attribut-Cluster befindet sich in *B. subtilis* und umfasst 185 Gene. Es handelt sich hier um einen ins Genom integrierten Phagen und damit einen Cluster für die Kategorie '29.06 phage proteins'. Von den insgesamt zwölf Clustern mit mehr als 50 ORFs sind vier auf solche Phagengenome zurückzuführen. In *E. coli* treten auch sehr große Cluster für die Kategorie 63 (vgl. Diskussion Überlappungen) auf. Diese enthalten zwar ausgedehnten Lücken, dennoch liegen die entsprechenden P-Values bei bis zu 1e-12.

Die Häufigkeit bestimmter Qualitätsklassen (gemessen am P-Value) in den einzelnen Genomen stellt Abb. 4.8 dar. Auch hier liegt die Kurve für *E. coli* wieder oberhalb derer der übrigen Organismen – was aber aufgrund der absoluten Anzahl an Clustern auch zu erwarten ist. Die auf der X-Achse aufgetragenen Stufen sind Mindestwerte, d.h. alle Kurven beginnen bei 1e-03 mit 100% aller im Genom gefundenen Attribut-Cluster. Der qualitative Verlauf ist in allen Kurven sehr ähnlich. In etwa jeder zweite Cluster hat einen P-Value von 1e-05 oder bes-

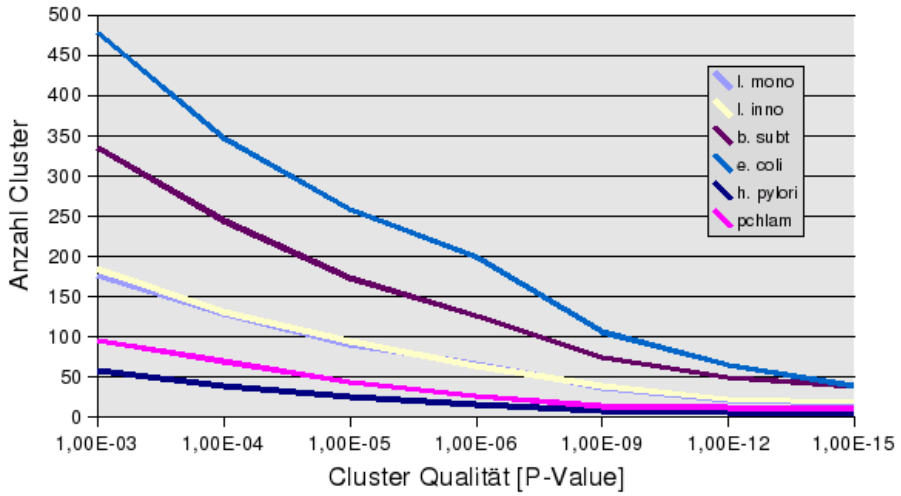


Abbildung 4.8: Cluster Qualität in den einzelnen Genomen

ser. Die beste aufgeführte Stufe von 1e-15 erreichen zwischen fünf (*H. pylori*) und 11 Prozent (*B. subtilis*).

Offen bleibt zunächst die Aussagekraft der P-Values. Weder das Signal-Rausch Verhältnis noch das Niveau falsch positiver Ergebnisse lässt sich aus dieser Graphik abschätzen. Immerhin deutet der flache Kurvenverlauf (die Häufigkeits-Werte fallen nicht exponentiell!) an, dass die ermittelten Cluster tatsächlich signifikant sind. Zur besseren Bewertung wurden Simulationen mit permutierten Genomen durchgeführt.

Randomisierte Genome

Das in dem beschriebenen Verfahren verwendete statistische Modell berücksichtigt keine Abhängigkeiten zwischen verschiedenen Experimenten. Bereits die Konzeption des FunCat führt jedoch dazu, dass solche Abhängigkeiten in den zugrunde liegenden Daten existieren. Eine angemessene statistische Modellierung wäre aber sehr aufwändig bzw. ohne vorhergehende Explikation der Abhängigkeiten gar nicht möglich.

Um trotzdem eine Fehler-Abschätzung vornehmen zu können, wurden Simulationen auf randomisierten Genomen durchgeführt. Ausgehend von echten Genomen (einschließlich ihrer manuellen Annotation) wurden zufällige Permutationen der Gen-Abfolge gebildet, die sich daraus ergebenden Zufallsgenome dann in gleicher Weise analysiert wie die realen. Diese artifiziellen Genome sind damit bezüglich der Häufigkeit aller Attributwerte identisch zu ihrem realen Pendant, deren Verteilung auf dem Genom allerdings unterliegt dem Zufall.

Wünschenswert wäre es, für jeden analysierten Organismus eine Versuchsreihe dieser Art durchzuführen. Dies ist jedoch aufgrund der hohen Anzahl an er-

forderlichen Berechnungen kaum durchführbar. Daher wurden exemplarisch *B. subtilis* und *L. monocytogenes* ausgewählt. Hinsichtlich der Annotationsdichte unterscheiden sich die Genome (37 gegenüber 46 Prozent), auch ist das *B. subtilis*-Genom ist deutlich größer. Die unten präsentierten Ergebnisse beruhen auf 5000 Permutationen des Listerien-Genoms, und 2500 auf Basis von *B. subtilis*.

	1E-03		1E-04		1E-05		1E-06	
	real	random	real	random	real	random	real	random
L. mono	176	4,35	129	0,54	90	0,06	66	< 0,01
B. subt	335	8,70	244	1,04	173	0,12	126	0,01

Tabelle 4.3: Vergleich Cluster Qualität reale vs. Zufallsgenome. Basis für die randomisierten Werte sind Versuche mit 5000 auf *L. monocytogenes* bzw. mit 2500 auf *B. subtilis* beruhenden permutierten Genomen. Angegeben ist für vier Qualitätsstufen jeweils die absolute Anzahl der ermittelten Cluster, für die Zufallsgenome zusätzlich der Durchschnitt pro Versuch.

Tabelle 4.3 gibt die Anzahl identifizierter Cluster getrennt nach den Qualitätsstufen wieder. Aufgeführt sind die Mittelwerte über alle Zufallsgenome. Im Gegensatz zu den realen Genomen zeigt sich hier ein exponentieller Abfall der Werte.

Die Zahlen deuten darauf hin, dass (in grober Näherung) in einem zufällig angeordneten Genom etwa zehn Cluster mit P-Value $\leq$ 1e-03 anzutreffen ist, somit ein Cluster der Stufe 1e-04. In jedem zehnten bzw. hundertsten Zufallsgenom sind demnach Cluster mit P-Value besser als 1e-05 respektive 1e-06 zu erwarten. Verglichen mit den realen Werten bedeutet dies eine Quote von unter 5% falsch positiven Clustern.

Exkurs: Manuelle vs. Automatische Annotation

Die in Tabelle 4.1 aufgeführten Genome wurden deshalb ausgewählt, weil sie manuell mit Hilfe des FunCat-Schemas annotiert wurden. Die Menge der Organismen mit vorhandenen FunCat-Zuordnungen ist wesentlich größer, die weitaus meisten dieser Genome wurden jedoch nur automatisch prozessiert. Eine automatische Annotation beruht im wesentlichen auf Sequenzhomologie zu bereits bekannten und bezüglich des Funktionskatalogs klassifizierten Genen. Für jedes Gen im zu annotierenden Genom wird also eine BLAST-Suche durchgeführt. Bei Überschreiten eines gewissen Ähnlichkeits-Schwellwerts werden die Zuordnungen des getroffenen Proteins auf das neue Gen übertragen.

Sequenz-Ähnlichkeit allein ist jedoch nicht unbedingt ausreichend, um die Funktion eines Gens vorherzusagen. Zwar ist sie ein Indiz für den gemeinsamen Ursprung zweier Proteine. Ein Protein kann aber im Laufe seiner Evolution neue Funktionen annehmen. Bei der Beurteilung der Aussagekraft eines Sequenzvergleichs ist darum die Unterscheidung zwischen Orthologen und Paralogen entscheidend. Alte Proteinfamilien haben viele nicht orthologe Homologe, für die eine Übertragung der Funktion oft nicht zulässig ist.

Besonders komplex ist die Situation bei Proteinen, die aus mehreren Domänen zusammengesetzt sind. Hier sind vor allem die Homologiebeziehungen der einzelnen Domänen relevant. Dies gilt besonders beim sog. Domain-Shuffling, wenn also die Abfolge der Domänen verändert wird (Abascal und Valencia 2003). Hinzu kommt, dass in Multi-Domänen-Proteinen der Zusammenhang von Sequenz- und Funktionsähnlichkeit generell weniger strikt ist (Hegyi und Gerstein 2001).

Unterschiede in der Evolutionsgeschwindigkeit innerhalb von Proteinfamilien sind ein weiteres Problem bei der Anwendung strikter Ähnlichkeits-Thresholds. Kellis et al. fanden in einer Studie über vier Hefen sowohl Beispiele für extrem schnelle (13% Aminosäure-Identität) als auch für extrem langsame (100% Identität) Veränderung orthologer Proteine (Kellis et al. 2003).

Manuelle Annotation erlaubt im Vergleich zur automatischen eine größere Flexibilität (Galperin und Koonin 1998). Die besondere Schwierigkeit eines rein automatisierten Vorgehens besteht darin, feste Einstellungen für wichtige Parameter wie Sequenzähnlichkeit zu wählen. Setzt man die Schwelle zu hoch, entgehen viele (auch richtige) Funktionsvorhersagen. Zu restriktive Filter führen andererseits zu einer hohen Anzahl falscher Zuordnungen. Existierende automatische FunCat-Annotationen gehen hier einen Mittelweg. Das heißt aber eben auch, dass man mit einer gewissen Quote von Fehlern in diesen Daten rechnen muss.

Fraglich ist, welche Auswirkungen diese Fehler bei der Berechnung von Attribut-Clustern haben. Zumindest die rein intragenomische Anwendung ist unabhängig von Sequenz-Ähnlichkeit, insofern könnte die Methode theoretisch auch als Filter fungieren.

Zur Abschätzung des Effektes wurden daher für vier Genome auf automatischer FunCat-Annotation basierende Attribut-Cluster berechnet: *Bacillus subtilis*, *Escherichia coli* sowie die beiden Listerien Arten. Um die Vergleichbarkeit zu erhöhen, wurden dabei Tabellen aus den gleichen Datenbanken[10] herangezogen, die auch die manuelle Annotation enthalten. Zeitlich weit auseinander liegende Prozessierungen (Bsp.: *B. subtilis* im öffentlich zugänglichen Pedant vs. Biomax-Annotation) basieren auf unterschiedlichem Wissensstand und sind daher für diese Studie schlecht geeignet. Tabelle 4.4 fasst die wichtigsten Zahlen hinsichtlich des Vergleichs automatischer zu manueller Annotation zusammen.

Am einfachsten ist die Bewertung der Ergebnisse bei *E. coli*. Die automatisch generierten Zuordnungen resultieren in sehr viel weniger Clustern als die manuellen. D.h. die Sensitivität sinkt beträchtlich. Ein Vergleich der einzelnen Attribut-Cluster zeigt, dass nur 51 (13% bezogen auf die Zahl für den automatischen FunCat) Cluster identisch sind mit denen, die bei manueller Annotation gefunden werden. Die weit überwiegende Zahl von Clustern wird demnach also unkorrekt[11] hinsichtlich des Match/Insert Musters und/oder der Ausdehnung berechnet.

[10] Im Pedant sind alle Informationen zu einem Genom innerhalb einer Datenbank organisiert. Automatisch berechnete FunCat-Zuordnungen sind in der Regel in der Tabelle 'funcat' gespeichert, die Ergebnisse manueller Annotation in 'sel_funcat'.

[11] Nimmt man die Ergebnisse für die manuelle Annotation als Standard of Truth

Organismus		Gene Annotiert	Zuordnungen	Attribut Cluster	Summe Positive	Positive ÷ Annotierte Gene [%]	Summe Inserts
Listeria monocyt.	man	1546	3865	176	587	37.97	246
	auto	1915	10179	219	679	35.46	315
Listeria innocua	man	1549	3685	184	644	41.58	353
	auto	1833	9768	239	681	37.15	363
Bacillus subtilis	man	2790	17557	335	1296	46.45	457
	auto	2998	19224	451	1355	45.20	662
Escherichia coli	man	2824	23150	478	1406	49.79	718
	auto	2352	18424	374	826	35.12	469

Tabelle 4.4: Kenndaten zum Vergleich manuelle vs. automatische Annotation

Tabelle 4.4 weist *E. coli* allerdings in vieler Hinsicht als Sonderfall aus; vor allem ist es das einzige Genom, in dem die Anzahl der annotierten Gene und auch Zuordnungen im manuellen Fall größer ist als bei Auto-Annotation. Bei den anderen Genomen ist es umgekehrt. Dies äußert sich auch in den Ergebnissen: die Auto-FunCat finden hier mehr Attribut-Cluster als die manuellen Analysen. Abbildung 4.9 zeigt jedoch, dass dies vorrangig für die schwächeren Qualitätsstufen gilt.

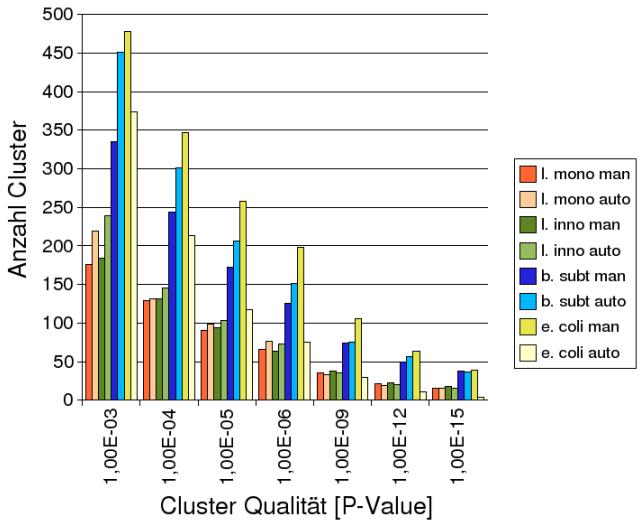


Abbildung 4.9: Cluster Qualitäten: automatische und manuelle Annotation im Vergleich. Gleiche Organismen sind jeweils in ähnlichen Farben dargestellt, wobei der manuellen Annotation der jeweils dunklere Ton zugewiesen ist.

Bei den Listerien beschränken sich die nennenswerten Unterschiede auf Cluster der schwächsten Güteklasse. Mit automatischer Annotation lassen sich hier also immerhin (mindestens) gleich viele Cluster finden wie mit manueller. Auch der Wert der vollständig richtigen Cluster liegt mit gut 41% sehr viel höher als bei den anderen beiden Genomen. Andererseits gibt es eine Reihe (23 = 10,5% in

monocytogenes, 28 = 11,7% in innocua) an falsch Positiven, also Clustern, die ausschließlich für die auto FunCats gefunden werden. Auch die Quote der falsch Negativen – berechnet als Anzahl der Cluster für die manuelle Zuordnung, die kein Pendant (Überschneidung in keinem ORF) in der automatischen Annotation haben – liegt um die 10%-Marke. Folglich sind auch hier deutliche Verluste in Sensitivität und Spezifität zu beobachten.

Die Situation in *B. subtilis* stellt sich nochmals anders dar. Mit 29% erreicht die Quote an exakten Übereinstimmungen mit der manuellen Annotation einen mittleren Wert. Nur 12 Cluster der Auto-Funcats haben keine Überschneidung in der Hand-Annotation, umgekehrt fehlen 24, die Zahlen für falsch Negative und falsch Positive scheinen also moderat. Diese Werte werden jedoch begünstigt durch die Tatsache, dass viele Cluster für die Auto-Annotation sehr stark verlängert sind: ein (mit falschen Treffern) verlängerter Cluster kann mit mehreren echten Clustern überlappen und begünstigt damit vor allem die Anzahl nicht gefundener Cluster. Die automatischen Zuordnungen liefern dreimal so viele Cluster mit mehr als 30 ORFs wie der Vergleichsfall (36 gegenüber 12).

Das führt zu einem unverhältnismäßig großen Anstieg in der Summe der Inserts (siehe Tabelle 4.4, letzte Spalte) um fast 45%, während die Summe der Positiven kaum wächst. Auch die Überlappungen innerhalb der Clusterung nehmen stark zu: während es in der Hand-Annotation zu maximal 9-facher Überlappung kommt (7 ORFs), treten bis zu 14-fache bei den Auto-FunCat auf (insgesamt 92 ORFs in Regionen mit mindestens 9-fachem Overlap).

Die detaillierte Analyse in *B. subtilis* zeigt, wie sich fehlerhafte Zuordnungen auf das Ergebnis der Clusterung auswirken können. Einzelne Treffer fungieren als „Brücken“zwischen den Inserts, so dass im Vergleich zur manuellen Annotation überlange Cluster mit relativ schwachen P-Values entstehen.

Weitere Beobachtungen

Jeder Attribut-Cluster beinhaltet die Zuordnung eines bestimmten funktionalen Attributs zu allen enthaltenen Genen. Für Inserts ist diese Zuordnung nicht in den Ausgangsdaten enthalten und impliziert somit eine neue Funktionsvorhersage. Deren experimentelle Überprüfung war nicht Gegenstand der vorliegenden Arbeit. Systematische Funktionsanalysen zur Beurteilung der Vorhersagequalität bieten sich jedoch als wünschenswerte Ergänzung an.

Ebenso würden die Ergebnisse dieser Arbeit nahe legen, die Annotation der untersuchten bakteriellen Genome zu redigieren. Ein solches Projekt wäre eine interessante Herausforderung für ein Annotationsteam. Aus Tabelle 4.2 ergibt sich, dass mehrere hundert ORFs pro Organismus einer intensiven manuellen Analyse unterzogen werden müssten.

Offen ist jedoch, für wie viele Inserts dabei tatsächlich ein Ergebnis gefunden werden könnte, der aktuelle Kenntnisstand ist bereits in die manuellen Annotationen eingeflossen. Für viele der betroffenen ORFs dürfte mit herkömmlichen Annotationsmethoden daher keine Bestätigung oder Zurückweisung der Vorhersage möglich sein.

Allerdings gibt es eine interessante Ausnahme, nämlich einen Cluster für die Kategorie '01.03.01.03 purine nucleotide anabolism' in *L. innocua* und umfasst die Gene mit den Bezeichnern gi_16414377 – gi_16414388 (Koordinaten: 1889719 – 1902449). Von den zwölf Genen sind elf der genannten Kategorie zugeordnet, lediglich der ORF gi_16414384 (lin1883) taucht als Insert auf. Der Theorie der Attribut-Cluster zu Folge sollte sich die Klassifizierung auch auf diesen ORF erstrecken. Tatsächlich liefert eine BLAST-Suche bei NCBI (NCBI BLAST) einige sehr gute Treffer (E-Values von 2e-18 bzw. 3e-17), die das Gen als `purS` (Phosphoribosylformylglycinamidin Synthase, PurS Untereinheit) ausweisen. Dabei handelt es sich um Einträge für *B. subtilis* und zwei Stämme von *Bacillus cereus*, die jedoch alle nach Abschluss der Annotation von *L. innocua* in die öffentlichen Datenbanken eingestellt wurden. Ein analoger Cluster existiert auch in *L. monocytogenes*, hier ist es der ORF lmo1771, der als Insert auftaucht.

In diesem Fall wird also die Vorhersage durch aktualisierte Daten bestätigt, was aber wie gesagt die Ausnahme ist. Dennoch ist es möglich, qualitative Aussagen zu treffen.

Erwartete Cluster Zunächst kann man feststellen, dass viele der erwarteten Cluster tatsächlich gefunden werden. Dazu gehören Phagen, bekannte Operons, ABC Transporter Kassetten, ribosomale Cluster oder auch Pathogenizitäts-Inseln. Leider gibt es offenbar keine umfassenden Listen mit solchen Modulen, anhand derer man eine echte Positiv-Kontrolle vornehmen könnte.

Ungenauigkeiten in der Annotation In einigen identifizierten Clustern fällt auf, dass Insert-Gene nach Name und Beschreibung eigentlich Treffer für die Cluster definierende Kategorie sein sollten, es gibt aber keine diesbezügliche Zuordnung. Oft drängt sich hier der Verdacht auf, dass es sich um irrtümlich nicht vorgenommenen Annotationen handelt. Ein gutes Beispiel ist das oben bereits erwähnte Gen `purS`. In der manuellen Annotation für *B. subtilis* ist das Ortholog sogar entsprechend benannt; trotzdem wurde keine Kategorie aus dem Unterbaum '01.03 nucleotide metabolism' zugewiesen. Dieses Beispiel zeigt, dass die hier vorgestellte Methode eine ideale Vorbereitung einer systematischen Redigierung darstellt. Insbesondere unterstützt sie die Überprüfung der Vollständigkeit einer Annotation.

Die Gefahr von Auslassungen bei der Annotation ist zum Teil auf das Klassifikationschema des FunCat zurückzuführen. Die Abbildung des Raumes der Proteinfunktionen auf Kategorien ist sehr ungleichmäßig. D.h. für einige Aspekte gibt es gleich mehrere Kategorien, andere sind unter-repräsentiert und sicher fehlen auch gewisse Kategorien noch ganz. Für den Annotator stellt dies ein Problem dar, vor allem weil die semantischen Beziehungen und Abhängigkeiten im Schema kaum explizit gemacht sind. Auch gibt es (bislang) keine Software-Unterstützung, die beispielsweise auf Basis einer definierten Regelmenge Vorschläge generieren könnte („Wenn Kategorie X zutrifft, prüfe auch Kategorie Y"). Die Ungleichmäßigkeit bezüglich der Dichte unterschiedlicher Kategorien

gilt für GO in mindestens gleichem Maße. Hinzu kommt hier die sehr große Anzahl an Attributen, die eine einheitliche Annotation erschwert.

Übervorsichtige Annotation Eine weitere Beobachtung betrifft Gene, bei denen die Konfidenz der Standardmethoden im Grau-Bereich liegen. In diesen Fällen gibt es zwar Hinweise auf die Funktion, die Signalstärke reicht aber nicht aus, um eine bestimmte Zuordnung zu rechtfertigen.

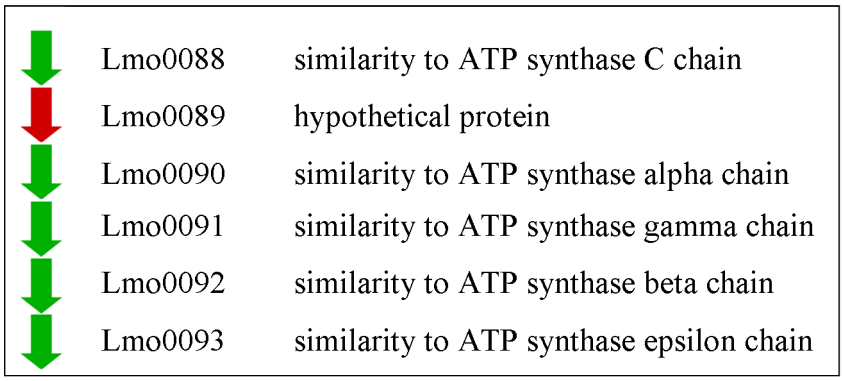


Abbildung 4.10: Cluster für die Kategorie '02.45.15 energy generation' in *L. monocytogenes* mit lmo0089 als Insert. Der berechnete P-Value beträgt 4,6e-09.

Als konkretes Beispiel kann hier ein Cluster in *L. monocytogenes* genannt werden, siehe Abbildung 4.10. Fünf Treffer für die Kategorie '02.45.15 energy generation' umrahmen das Gen lmo0089. Eine diesbezügliche Abfrage bei NCBI (NCBI BLAST) ergibt einen schwachen BLASTP Hit (E-Value 0.46) sowie eine konservierte Domäne mit einem E-Value von 1e-05. Beide Methoden weisen auf eine ATP Synthase Delta Kette hin, die Konfidenzwerte sind aber keineswegs überzeugend. Das Ergebnis der Attribut-Cluster Analyse geht jedoch in die gleiche Richtung und gibt somit zusätzliche Unterstützung für die Funktionsvorhersage der Sequenz basierten Methoden.

Wäre die Nachbarschaft bei der Annotation mit einbezogen worden, hätte dies den Ausschlag geben können, zumindest eine vermutliche Funktionszuordnung vorzunehmen. In der Regel wird jedoch Topologie-Information bei manueller Annotation nicht genutzt.

Auch hier zeigt sich, wie Attribut-Cluster in einem Genom-Annotationsprojekt gewinnbringend eingesetzt werden können: Nach Abschluss einer ersten Runde manueller Annotation werden die Cluster berechnet, um in einem zweiten Durchgang auffällige Inserts wie lmo0089 nochmals zu prüfen. Attribut-Cluster stellen eine Möglichkeit dar, Information, die in Nachbarschaftsbeziehungen enthalten ist, explizit zu machen und auf systematische Weise in den Annotationsprozess zu integrieren.

Metacluster Mit dem Begriff Metacluster sind Regionen gemeint, in denen sich Cluster für mehrere Funktionen überlagern („Cluster von Clustern“). Im einfachsten Fall ist dies darauf zurückzuführen, dass die Mehrzahl der enthaltenen Gene beiden Funktionen zugeordnet ist. Interessanter ist aber die Alternative, bei der den ORFs überwiegend nur entweder Funktion A oder Funktion B zugeordnet ist. Beispiel ist eine Region in *B. subtilis* (Koordinaten 3458839 – 3470032), in der ein Cluster für ‘14.10.90 other programmed cell death’ sich in eine Lücke für einen umspannenden Cluster der Kategorie ‘11.99 other cell rescue activities’ einfügt.

Diese Daten weisen darauf hin, dass es sich um ein Modul von Genen handelt, das der Zelle Verhaltensweisen in bestimmten Krisensituationen bereit stellt. Die real vorhandene Nachbarschaftsbeziehung kann jedoch von den verwendeten Suchkriterien nicht vollständig abgebildet werden. Eine Kategorie, die die beiden genannten subsumiert, würde diese Region aber zu einem einzigen Cluster zusammenfassen. Dieses Beispiel zeigt, dass Attribut-Cluster geeignet sind, Ansatzpunkte für experimentell nachprüfbare Hpypothesen bezüglich der Funktion bestimmter Gene zu liefern.

Intergenomischer Vergleich In vielen Fällen zeigen die ermittelten Cluster hohe Signifikanz. Unsicher ist allerdings, in wie weit es gerechtfertigt ist, daraus auch Rückschlüsse auf die Genfunktion zu ziehen. Ein Weg, die Konfidenz solcher Vorhersagen zu erhöhen, besteht darin, Übereinstimmungen in Clustern für andere Genome zu finden.

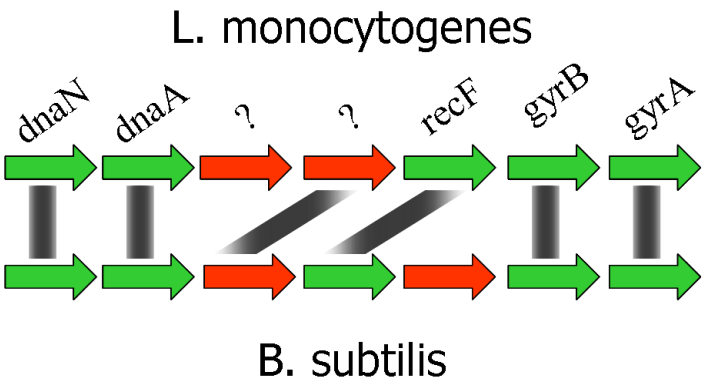


Abbildung 4.11: Die Homologie (angedeutet durch die schwarzen Balken) zwischen diesen Clustern erstreckt sich auch auf ein Insert. Die Codierungsrichtungen stimmen ebenfalls überein.

Eine systematische Auswertung in dieser Richtung gibt es noch nicht, aber ein sehr schönes Beispiel ist in Abb. 4.11 gezeigt. In den Listerien erstreckt sich über die ersten sieben Gene ein Cluster für die Kategorie ‘03.01.03 DNA synthesis and replication’. Alle fünf Treffer innerhalb dieses Clusters haben ein orthologes Gen in einem entsprechenden Cluster in *B. subtilis*. Besonders interessant aber:

die Homologie erstreckt sich auch auf eines der Insert-Gene. Die funktionale Nachbarschaftsbeziehung ist hier also immerhin über Genus-Grenzen hinweg konserviert.

Abschätzung: Häufigkeit Funktionaler Assoziation Die Frage, in wie weit Nachbarschaft und Funktion in bakteriellen Genomen gekoppelt sind, ist eine nach wie vor offene Frage. Die Methode der Attribut-Cluster Berechnung erlaubt hier jedoch eine Abschätzung. Tabelle 4.2 zeigt, das bis zu knapp 50% aller annotierten Gene als Match in mind. einem Cluster auftauchen (im Falle von *E. coli*). Die Quote falsch positiver Cluster ist gemäß der Abschätzung mit Hilfe der randomisierten Genome nur sehr gering. Hinzu kommt, dass der Anteil mit zunehmender Aufklärung des Genoms weiter steigen dürfte. Insgesamt erscheint die Abschätzung, dass jedes zweite Gen funktional mit seiner Nachbarschaft assoziiert ist, damit durchaus plausibel. Diese Einschätzung deckt sich mit einer Aussage in (Huynen et al. 2000b), wonach bei der Hälfte der Gene in *M. genitalium* genomische Kontext-Information von Bedeutung ist. Allerdings ist die Aussagekraft der genomischen Nachbarschaft nicht für alle Gene gleich groß: sie ist beispielsweise größer für Gene in Biosynthese-Pfaden als für solche in Degradations-Pfaden (von Mering et al. 2003b).

4.4 Diskussion

Abgrenzung gegen andere Begriffe Der vorgestellte Begriff der Attribut-Cluster ist ein Konzept, mit dem die Ungleichverteilung von Genfunktionen innerhalb von Genomen explizit gemacht werden kann. Ein Attribut-Cluster ist das Ergebnis eines evolutionären Prozesses, in dem die Abfolge der Gene auf einem Chromosom verändert wurde. Damit ergeben sich Überschneidungen zu den etablierten Begriffen Operon (Gruppe gemeinsam transkribierter Gene) und Regulon (Gruppe von Genen mit gleicher Expressions-Regulation).

Jedes Operon stellt einen Attribut-Cluster dar, denn unabhängig vom Erklärungsmodell (vgl. Abschnitt 4.1.2) können nur gemeinsam agierende Gene ein derart eng gekoppeltes Modul bilden. Ein bestimmtes Operon kann jedoch nur dann als Attribut-Cluster beschrieben werden, wenn die gemeinsam erfüllte Funktion als Suchkriterium gewählt wird (Bsp.: das Tryptophan-Operon wird nur sichtbar bei der Suche nach der Funktion 'Tryptophan-Biosynthese'). Umgekehrt *kann* sich ein identifizierter Attribut-Cluster als Operon herausstellen, muss es aber nicht. Die Methode ist ausdrücklich nicht als dedizierte Operon-Suche zu verstehen, denn über gemeinsame Transkription können die herangezogenen Daten keine Auskunft geben.

Ähnlich verhält es sich mit dem Begriff des Regulons. Komplexe Funktionen können die gekoppelte Regulation mehrerer Gene erforderlich machen. Kommt (aus welchen Gründen auch immer) eine lokale Massierung dieser Gruppe hinzu, kann dies als Attribut-Cluster erkennbar werden. Die molekularen Mechanismen der gemeinsamen Regulation (beispielsweise bestimmte Transkriptions-

Faktoren) gehen aber nicht in die Ermittlung der Attribut-Cluster ein, insofern handelt es sich auch nicht um eine Regulon-Suche.

Zusammenfassend kann man sagen, dass Attribut-Cluster eine formale Generalisierung verschiedener biologischer Phänomene darstellen, darunter auch Operons, Regulons (sofern nicht verstreut) oder Pathogenizitäts-Inseln.

Vergleich mit anderen Methoden Die Suche nach Attribut-Clustern ist eine Kontext basierte Methode im engsten Sinne, denn sie besteht im wesentlichen darin, relevante Kontexte zu identifizieren. Sie unterscheidet sich in einigen Punkten deutlich von den im Abschnitt 4.2 skizzierten Ansätzen.

Ein wichtiges Differenzierungsmerkmal zu den Methoden, die konservierte Nachbarschaften analysieren (siehe Abschnitt 4.2.1, u.a. SNAP oder STRING), besteht darin, dass die Identifizierung der Ziel-Entitäten (hier also der Attribut-Cluster) rein intra-genomisch betrieben wird. Damit besteht weitgehende Unabhängigkeit von einer exakten Bestimmung der Orthologie-Beziehungen zwischen zwei Genomen, insbesondere die Unterscheidung zwischen echten Orthologen und Paralogen fällt nicht ins Gewicht. Außerdem werden Aussagen über ORFs möglich, die keine Entsprechung in anderen Organismen haben. Die Anzahl dieser Gene, die sich rein inter-genomischen Methoden prinzipiell entziehen, schwankt mit dem betrachteten Taxon, kann aber einen bedeutenden Teil des Genoms ausmachen.

Die Reihenfolge der Gene ist selbst in konservierten Clustern nur schlecht erhalten (Huynen und Bork 1998). Auch dieses Problem umgeht der Ansatz der Attribut-Cluster, weil er an keiner Stelle auf eine bestimmte Reihenfolge abstellt, auch nicht beim Vergleich mit anderen Genomen. Dies ist ein weiterer Unterschied zu einigen anderen Verfahren, u.a. solchen, die auf konservierten Gen-Paaren aufsetzen.

Andere auf konservierter Nachbarschaft beruhende Methoden setzen zwar nicht zwingend konservierte aufeinander folgende Paare voraus, beschränken aber den Radius der Suche mehr oder weniger strikt (vgl. Definition der 'Runs' in (Overbeek et al. 1998), auch genutzt für STRING (Snel et al. 2000a); SNAP (Kolesov et al. 2001)). Erforderlich sind solche Beschränkungen bezüglich Orientierung und inter-genischen Abständen vor allem, um die Berechenbarkeit zu gewährleisten, obwohl sie zum Teil auch biologisch gerechtfertigt sind. Letzteres gilt allerdings nur, soweit es um die Suche nach Operons geht, Attribut-Cluster sind aber gerade ein Versuch, vom Operon-Konzept zu abstrahieren (siehe vorangegangenen Abschnitt).

Relevanz der Ergebnisse Für die Ergebnisse, die mit einer Attribut-Cluster-Analyse erzielt werden, gilt eine generelle Aussage aus (Wolf et al. 2001): Kontext basierte Methoden helfen bei der Einordnung uncharakterisierter Gene in das Gesamt-System, treffen Aussagen über die Rolle eines bestimmten Proteins, können aber keine Prognosen zu den biochemischen Funktionen eines Proteins treffen. Bei Sequenz basierten Methoden verhält es sich umgekehrt.

Die tatsächliche biologische Relevanz eines Clusters kann nur durch sorgfältige manuelle Überprüfung im Einzelfall bewertet werden. Die Ergebnisse mit randomisierten Genomen zeigen, dass das Auftreten von Attribut-Clustern kein zufälliges Ereignis ist, die Quote an falsch positiven Ergebnissen ist niedrig. Eine weitere Verbesserung wäre möglich durch nachträgliches Filtern auf kleine inter-genische Abstände innerhalb der Cluster oder gleiche Orientierung. Übergeordnete Strukturen wie die beschriebenen Metacluster würden dadurch jedoch ausgeblendet. Gerade diese Metastrukturen können aber wichtige Hinweise auf das Zusammenwirken einzelner zellulärer Prozesse liefern. Dieser Gedanke wurde im Zusammenhang mit sog. Uber-Operons bereits in (Lathe III et al. 2000) formuliert.

Ein Uber-Operon ist eine Gruppe von Genen, die in einer Vielzahl an Genomen in Clustern anzutreffen sind. Die Reihenfolge der Gene in den Clustern und selbst die Zusammensetzung der einzelnen Cluster variiert dabei, in allen Genomen jedoch ist der Pool dieser Gene in Form eines oder mehrerer Cluster organisiert. Uber-Operons verbinden oft klassische Operons mit anderen funktional gekoppelten Genen, beispielsweise den Translations-Faktor *tufA* mit einem ribosomalen Operon (Lathe III et al. 2000). Einzelne Cluster eines Uber-Operons werden als Attribut-Cluster sichtbar.

Ein Vorteil der hier präsentierten Methode liegt in der unmittelbaren Aussagekraft der Ergebnisse. Die Eigenschaft, die den Cluster definiert, ist genau das Attribut, mit dessen Hilfe er gefunden wurde. Bei anderen Methoden wie STRING oder SNAP kann dieses Merkmal nur indirekt über die Beschreibung der einzelnen enthaltenen Gene abgeleitet werden. Dies ist zwar für den Benutzer in vielen Fällen relativ leicht, setzt aber entsprechendes Expertenwissen voraus. Attribut-Cluster hingegen sind einer maschinellen Prozessierung eher zugänglich.

Diese Eigenschaft verdankt die Methode dem Umstand, dass sie auf Ausgangsdaten mit einem sehr hohen Abstraktionsniveau operiert. Darin unterscheidet sie sich nicht nur von den Verfahren, die in den ersten Kapiteln dieser Arbeit vorgestellt wurden, sondern auch von anderen kontextbasierten Ansätzen wie SNAP. Input für die Attribut-Cluster Suche sind keine Sequenzdaten sondern Annotation, die ihrerseits Endprodukt einer komplexen Informationsverarbeitung ist. STRING nimmt in dieser Hinsicht eine Zwischenstellung ein, weil es auf Zuordnungen von Genen zu COGs aufbaut. Diese abstrahieren zwar von den Aminosäuresequenzen, basieren aber auf Sequenzdaten. Die Such nach Attribut-Clustern stellt dagegen vollständig auf extrinsische Information ab.

Daraus ergibt sich allerdings auch, dass Quantität und vor allem Qualität der Ergebnisse in hohem Maße von der Basis-Annotation abhängen. Eine lückenhafte Zuordnung führt zu einer geringen Anzahl von gefundenen Clustern. Fehlerhafte Ausgangsdaten haben entweder ebenfalls das Verpassen existierender Cluster oder sogar falsche Cluster zur Folge. Gefahr besteht besonders bei systematischen Fehlern, wie sie bei automatischer Annotation nicht selten vorkommen.

Empfohlenes Einsatzgebiet sind demzufolge vor allem Projekte zur manuellen Annotation kompletter Genome. Hier jedoch kann sich eine Attribut-Cluster-Analyse als sehr sinnvoll erweisen. Sie bietet einen Weg zur systematischen Einbeziehung genomischer Kontext-Information. Signifikante Häufungen bestimmter Funktionen innerhalb des Genoms sind bei der Annotation zumindest zum Teil auch ohne diese Methode aufgefallen. Nun aber werden sie explizit gemacht und systematisiert, können in einer Datenbank abgespeichert und jederzeit abgefragt oder visualisiert werden.

Attribut-Cluster bieten dem Annotator ein Instrument, mit dem der evolutionäre Kontext eines bestimmten ORFs sichtbar wird. Er kann abschätzen, ob Information über benachbarte Gene bei der Charakterisierung eines Gens zu berücksichtigen ist. In Kombination mit sequenzbasierten Methoden kann der Annotator diesen Kontext als zusätzliches Kriterium in Zweifelsfällen hinzu ziehen. Auslassungen in der Basis-Annotation werden aufgedeckt.

Je mehr Gene funktional charakterisiert sind, desto größer werden die Erfolgsaussichten für die Attribut-Cluster. Auch die sequenzbasierten Methoden konnten anfangs nur auf einen kleinen Pool an Genen zurückgreifen. Erst mit zunehmender experimenteller Charakterisierung, Verfeinerung der Algorithmen und indirekter Übertragung von Annotation[12] wuchs die Daten-Basis und damit auch die Erfolgsquote. Ein ähnlicher Prozess könnte auch den Kontext basierten Ansätzen bevor stehen.

Ausblick Der hier vorgestellte Ansatz bietet noch vielerlei Möglichkeiten zur Erweiterung. Eine Kombination mit anderen Kontext Methoden ist denkbar, etwa das bereits erwähnte nachgelagerte Suche nach 'Runs' (Overbeek et al. 1998) innerhalb von Attribut-Clustern. Wünschenswert wäre auch ein interaktives Visualisierungs-Tool, das solche und ähnliche Filterfunktionen anbietet.

Der inter-genomische Vergleich von Clustern ist bislang nur recht rudimentär implementiert. Eine systematische Prozessierung dieser Daten könnte wichtige Daten für die Evolution funktionaler Cluster und zur Aufklärung horizontaler Gen-Transfers liefern. In diesem Zusammenhang wäre eine Erweiterung des statistischen Modells auf multiple Genome sehr hilfreich.

Andere Datensätze als die Annotation nach dem MIPS-FunCat können mit wenig Aufwand integriert werden, hierin liegt eine der Stärken des Verfahrens. Prinzipiell kann jede beliebige Menge an Genen Ausgangspunkt einer Suche nach Attribut-Clustern sein, sinnvoll sind allerdings nur solche, die eine Interpretation in Hinblick auf die Gen-Funktion erlauben. Versuchsweise wurden Adapter für EC-Nummern sowie für Daten metabolischer Pfade implementiert. Nennenswerte Ergebnisse wurden dabei nicht erzielt. EC-Nummern bilden in aller Regel keine Attribut-Cluster. Daten über metabolische Pfade versprechen mehr Erfolg, der verfügbare Datensatz war allerdings zu klein. Ziel dieser Anwendungen war auch vorrangig die technische Erprobung.

[12] ein nicht zu unterschätzendes Potential für Fehler-Propagierung

Denkbar wäre eine Schnittstelle, die die Verarbeitung Benutzer spezifischer Datensätze erlaubt. Eine solche Funktion ist auch als Teil eines WWW-Dienstes vorstellbar. Darin sollte dann auch eine Suche nach in der Datenbank enthaltenen Genen (über Name oder Sequenz-Ähnlichkeit) sowie die Visualisierung bekannter Cluster integriert sein.

Wie bereits mehrfach erwähnt, ist die Qualität der Ergebnisse der Attribut-Cluster Analyse stark abhängig von der Vollständigkeit und Richtigkeit der Ausgangsdaten. Daraus ergibt sich, dass die Methode von zukünftigen Verbesserungen in der automatischen Annotation sehr stark profitieren wird.

Kapitel 5

Zusammenfassung und Ausblick

Im Mittelpunkt dieser Arbeit steht die Identifizierung konservierter Nachbarschaften in Genomen. Dabei wurden drei verschiedene Abstraktionsniveaus betrachtet. Organismen unterschiedlicher Domänen wurden untersucht, die Analysen teils intra- teils inter-genomisch ausgerichtet. Trotz dieser Unterschiede im Detail liegt allen Untersuchungen eine gemeinsame Strategie zu Grunde: die Identifizierung von Konfigurationen, die aufgrund ihrer konstanten Zusammensetzung vor dem variablen Hintergrund hervortreten.

Die Evolution als andauernder Prozess führt zu ständigen Veränderungen in der Architektur von Genomen, sowohl im kleinen (einzelne Basen) wie auch im großen Maßstab (Chromosomen). Die in dieser Arbeit analysierten Nachbarschaftsbeziehungen werden deshalb erkennbar, weil sie ab einem gewissen Zeitpunkt konserviert blieben, also eben nicht von den verändernden Kräften getrennt wurden.

Kapitel 2 betrachtet intra-genomisch die Nucleotidsequenz-Ebene. Konservierte Nachbarschaften treten hier hervor als segmentale Duplikationen. Wichtigstes Ergebnis dieser Analyse ist die Identifizierung eines Satzes solcher Duplikationen in *Arabidopsis thaliana*, der auf eine vollständige Verdoppelung des Genoms schließen lässt.

Das Genom hat sich inzwischen weit von einem polyploiden Zustand entfernt. Die beobachteten Segmente sind die Abschnitte, die bislang nicht durch Deletions- oder Umordnungsprozesse auseinander gerissen wurden. Offen ist die Frage, warum gerade diese Segmente noch nicht rearrangiert wurden oder vollständig verloren gegangen sind. Denkbar ist, dass seit dem Duplikationsereignis noch nicht genug Zeit vergangen ist, um alle Duplikate soweit zu verändern, dass sie nicht mehr als segmentale Duplikation erkennbar sind. Möglich ist aber auch, dass die Erhaltung dieser Nachbarschaften einen evolutionären Vorteil bietet.

Eine genaue Analyse der in den Segmenten vertretenen Gen-Funktionen könnte hier weitere Erkenntnisse liefern. Von erheblichem Interesse wäre auch der

Vergleich mit Arten, die sich nach der Genom-Duplikation von *Arabidopsis thaliana* getrennt haben. Entsprechende Daten werden in Kürze verfügbar sein, da die Genome zweier naher verwandter Pflanzen (*Arabidopsis lyrata* und *Capsella rubella*) zur Zeit sequenziert werden. Auch die Sequenzierung von *Brassica rapa* ist in Planung. Dieses Genom weist einen etwas größeren phylogenetischen Abstand zu *A. thaliana* auf, hat sich gemäß Datierung in (Blanc et al. 2003) aber ebenfalls erst nach der Genom-Verdopplung abgespalten.

Die Ermittlung der segmentalen Duplikationen in *A. thaliana* lieferte wichtige Daten für Studien zur evolutionären Dynamik duplizierter Gene. Die Differenzierung zwischen Paralogen, die aus segmentalen Duplikationen hervorgegangen sind, und anderen Arten von Duplikationen spielt etwa bei Analysen von Gen-Familien eine Rolle (Parenicova et al. 2003). Der Vergleich zwischen Tandem- und segmentalen Duplikationen steht in (Haberer et al. 2004) im Vordergrund. Hier wird insbesondere die Konservierung in den Promotoren untersucht, außerdem werden Unterschiede und Gemeinsamkeiten in der Transkription analysiert. Genexpression ist auch der Fokus einer Studie, die zwischen Genom-Duplikationen und kleinen Duplikationsereignissen differenziert (Casneuf et al. 2006). Weitere Publikationen zeigen eine Überrepresentation bestimmter Genfunktionen innerhalb der segmental duplizierten Gene (Seoighea und Gehring 2004; Blanc und Wolfe 2004)

In Kapitel 3 stehen Gene im Mittelpunkt der Analyse. Konservierte Nachbarschaften werden hier sichtbar als syntenische Bereiche zwischen verschiedenen Genomen. Die Nachbarschaft der betroffenen Gene auf dem Chromosom wurde also jeweils seit dem letzten gemeinsamen Vorfahren der betrachteten Organismen beibehalten. Erwartungsgemäß konnten bei der Analyse für *Fusarium graminearum* mit wachsendem phylogenetischem Abstand der verglichenen Genome weniger syntenische Bereiche gefunden werden.

Das vielleicht interessanteste Ergebnis dieses Abschnitts wurde in der Analyse von *Parachlamydia* sp. UWE25 erzielt. Aufgrund des Größenunterschieds gegenüber den verglichenen Chlamydien-Arten ist der hier vorgestellte Algorithmus einer rein grafischen Methode wie etwa Dot-Plots überlegen. So wird deutlich, dass etwa die Hälfte aller orthologen Gen-Paare in syntenischen Blöcken auftreten.

Eine Kombination der Syntenie-Analyse mit einer Suche nach Attribut-Clustern könnte genauere Erkenntnisse liefern, in wie weit Gene in syntenischen Blöcken auch einen funktionalen Kontext bilden.

Kapitel 4 schließlich rückt die Genfunktion in den Mittelpunkt. Nachbarschaften werden hier evident als genomische Cluster von Genen, die ein bestimmtes funktionales Attribut gemeinsam haben. Das Alter solcher Attribut-Cluster ist ohne weiter gehende inter-genomische Analysen nicht abzuschätzen. Wichtigste Ergebnisse dieses Teils der Arbeit sind die statistische Untermauerung des Konzepts sowie die Abschätzung, dass bis zu 50% eines mikrobiellen Genoms solchen funktionalen Nachbarschaften zugeordnet werden können.

Die hier gezeigte Konservierung von Attribut-Clustern impliziert, dass es Prozesse geben muss, die zur Neubildung solcher Nachbarschaften führen. Wie diese Prozesse genau ablaufen ist unklar. Offensichtlich aber bietet es evolutionäre Vorteile, wenn die Gene eines Clusters eng benachbart auf dem Chromosom liegen.

Ein solcher Vorteil kann sich aus einer vereinfachten Koordination der Expression der beteiligten Gene ergeben. Für Untereinheiten von Proteinkomplexen kann durch enge Nachbarschaft eine gemeinsame Regulation besser gewährleistet werden. Außerdem bleibt im Falle segmentaler Duplikationen das Gleichgewicht der einzelnen Komponenten bestehen (Teichmann und Veitia 2004). Im Extremfall kann der evolutionäre Druck bei physikalisch interagierenden Proteinen sogar zu Gen-Fusionen führen (vgl. etwa (Enright et al. 1999; Marcotte et al. 1999)).

Auch die Modelle, die die Entstehung und Konservierung von Operons beschreiben (vgl. Abschnitt 4.1), können auf Attribut-Cluster übertragen werden. Zwar ist die Kopplung bei (nicht-Operon-) Attribut-Clustern weniger ausgeprägt, da keine gemeinsame Transkription stattfindet. Dennoch gelten die gleichen Rahmenbedingungen wie in Operons: Nachbarschaft begünstigt die Wartung des Moduls.

Weitgehend unbearbeitet ist die Frage, in wie weit Attribut-Cluster in Eukaryonten-Genomen existieren. Obwohl Operons im klassischen Sinne in höheren Eukaryonten nicht existieren, heißt das nicht, das gar keine Beziehungen zwischen Gen-Funktion und -Nachbarschaft bestehen. Vielfach wurde zumindest signifikante Koexpression benachbarter Gene gefunden (einen Überblick gibt (D.Hurst et al. 2004)). Nicht immer geht korrelierte Expression allerdings auch mit funktionaler Ähnlichkeit einher (siehe etwa (Spellman und Rubin 2002)).

Die vorliegende Arbeit entwickelte Verfahren zur Analyse konservierter Nachbarschaftsbeziehungen. In unterschiedlichen Anwendungsfällen konnten durch Einsatz dieser neuen Methoden wesentliche Beiträge zur Genom-Analyse geleistet werden. Auf den verschiedenen Abstraktionsebenen reichten die getroffenen Aussagen von der Beurteilung chromosomaler Evolutionsprozesse bis zur Unterstützung von Annotationsprojekten hinsichtlich der Funktionsvorhersage von Genen.

Die vorgestellten Ansätze zur Untersuchung konservierter Nachbarschaft eignen sich auf den jeweiligen Abstraktionsebenen für vielseitige weitere Anwendungen. Besonders attraktiv erscheint jedoch die Perspektive, Analysen auf verschiedenen Abstraktionsebenen zu verknüpfen. Großes Potential verspricht beispielsweise die Kombination der Suche nach segmentalen Duplikationen mit einer Attribut-Cluster-Analyse. Im humanen Genom machen segmentale Duplikationen etwa 5% aus. Es ist durchaus vorstellbar, dass Gene in diesen Bereichen Auffälligkeiten bezüglich gemeinsamer Funktions-Attribute aufweisen, die durch das hier vorgestellte Methoden-Spektrum aufgezeigt würden.

Literaturverzeichnis

Abascal, Federico und A. Valencia (2003). *Automatic annotation of protein function based on family identification*. Proteins, 53(3):683–692.

Alberts, Bruce, D. Bray, J. Lewis, M. Raff und J. D. Watson (1995). *Molekularbiologie der Zelle*. VCH Verlagsgesellschaft, Weinheim, Bundesrepublik Deutschland, 3 Aufl.

Altschul, S.F., W. Gish, W. Miller, E. Myers und D. Lipman (1990). *Basic local alignment search tool*. Journal of Molecular Biology, 215:403 –410.

Anderson, Garrett H., N. D. Alvarez, C. Gilman, D. C. Jeffares, V. C. Trainor, M. R. Hanson und B. Veit (2004). *Diversification of Genes Encoding Mei2-Like RNA Binding Proteins in Plants*. Plant Molecular Biology, 54(5):653–670.

Andersson, Siv G. und K. Eriksson (2000). *Dynamics of gene order structures and genomic architectures*. In: Sankoff, David und J. H. Nadeau, Hrsg.: *Comparative Genomics*, Bd. 1 d. Reihe *Computational Biology Series*, S. 267–280. Kluwer Academic Publishers.

Bader, David A., B. M. Moret und M. Yan (2001). *A Linear-Time Algorithm for Computing Inversion Distance between Signed Permutations with an Experimental Study*. Journal Of Computational Biology, 8(5):483–491.

Bafna, Vineet und P. A. Pevzner (1998). *Sorting by Transpositions*. SIAM Journal on Discrete Mathematics, 11(2):224–240.

Bailey, Jeffrey A., R. Baertsch, W. J. Kent, D. Haussler und E. E. Eichler (2004). *Hotspots of mammalian chromosomal evolution*. Genome Biology, 5(4):R23.

Bancroft, Ian (2001). *Duplicate and diverge: the evolution of plant genome microstructure*. Trends in Genetics, 17(2):89 –93.

Baumbusch, Lars O., T. Thorstensen, V. Kraus, A. Fischer, K. Naumann, R. Assalkhou, I. Schulz, G. Reuter und R. B. Aalen (2001). *The Arabidopsis thaliana genome contains at least 29 active genes encoding SET domain proteins that can be assigned to four evolutionary conserved classes*. Nuclear Acids Research, 29:4319–4333.

Baumgarten, Andrew, S. Cannon, R. Spangler und G. May (2003). *Genome-Level Evolution of Resistance Genes in Arabidopsis thaliana*. Genetics, 165(1):309–319.

Bennetzen, Jeffrey L. (2000). *Comparative Sequence Analysis of Plant Nuclear Genomes: Microlinearity and Its Many Exceptions*. The Plant Cell, 12:1021–1029.

Bevan, Michael und G. Murphy (1999). *The small, the large and the wild: the value of comparison in plant genomics*. TRENDS in Genetics.

Blanc, Guillaume, A. Barakat, R. Guyot, R. Cooke und M. Delseny (2000). *Extensive Reshuffling in the Arabidopsis Genome*. The Plant Cell, 12:1093 –1101.

Blanc, Guillaume, K. Hokamp und K. H. Wolfe. *Interactive maps of duplicated blocks*. http://wolfe.gen.tcd.ie/athal/index.html.

Blanc, Guillaume, K. Hokamp und K. H. Wolfe (2003). *A Recent Polyploidy Superimposed on Older Large-Scale Duplications in the Arabidopsis Genome*. Genome Res., 13(2):137 –144.

Blanc, Guillaume und K. H. Wolfe (2004). *Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution*. Plant Cell, 16(7):1679–1691.

Blanchette, Mathieu, G. Bourque und D. Sankoff (1997). *Breakpoint Phylogenies*. Genome Informatics, 8:25 –34.

Blattner, F.R., G. I. Plunkett, C. Bloch, N. Perna, V. Burland, M. Riley, J. Collado-Vides, J. Glasner, C. Rode, G. Mayhew, J. Gregor, N. Davis, H. Kirkpatrick, M. Goeden, D. Rose, B. Mau und Y. Shao (1997). *The complete genome sequence of Escherichia coli K-12*. Science, 277(5331):1453–1474.

Blumenthal, Thomas (1998). *Gene clusters and polycistronic transcription in eukaryotes*. BioEssays, 20(6):480–487.

Blumenthal, Thomas und J. Spieth (1996). *Gene structure and organization in Caenorhabditis elegans*. Current Opinion in Genetics And Development, 6(6):692–698.

Bockhorst, Joseph, M. Craven, D. Page, J. Shavlik und J. Glasner (2003a). *A Bayesian network approach to operon prediction*. Bioinformatics, 19(10):1227–1235.

Bockhorst, Joseph, Y. Qiu, J. Glasner, M. Liu, F. Blattner und M. Craven (2003b). *Predicting bacterial transcription units using sequence and expression data*. Bioinformatics, 19(90001):i34–i43.

Borezee, Elise, E. Pellegrini und P. Berche (2000). *OppA of Listeria monocytogenes, an Oligopeptide-Binding Protein Required for Bacterial Growth at Low Temperature and Involved in Intracellular Survival*. Infect. Immun., 68(12):7069–7077.

Bourque, Guillaume und P. A. Pevzner (2002). *Genome-Scale Evolution: Reconstructing Gene Orders in the Ancestral Species*. Genome Res., 12(1):26–36.

Caprara, Alberto (1997). *Sorting by Reversals is Difficult*. In: *Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97)*, S. 75–83, New York. ACM.

Caprara, Alberto (1999). *Formulations and Complexity of Multiple Sorting by Reversals*. In: Istrail, Sorin, P. Pevzner und M. Waterman, Hrsg.: *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB 99)*, S. 84–93, New York. ACM Press.

Casneuf, Tineke, S. De Bodt, J. Raes, S. Maere und Y. Van de Peer (2006). *Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana*. Genome Biology, 7(2):R13.

Celamkoti, Srikanth, S. Kundeti, A. Purkayastha, R. Mazumder, C. Buck und D. Seto (2004). *GeneOrder3.0: Software for comparing the order of genes in pairs of small bacterial genomes*. BMC Bioinformatics, 5(1):52.

Center for Genome Research. *Fusarium graminearum Sequencing Project*. http://www.broad.mit.edu.

Clamp, M., D. Andrews, D. Barker, P. Bevan, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, T. Hubbard, A. Kasprzyk, D. Keefe, H. Lehvaslaiho, V. Iyer, C. Melsopp, E. Mongin, R. Pettett, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik und E. Birney (2003). *Ensembl 2002: accommodating comparative genomics*. Nucl. Acids Res., 31(1):38–42.

Coissac, Eric, E. Maillier und P. Netter (1997). *A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres*. Mol. Biol. Evol., 14(10):1062–1074.

Dandekar, Thomas, B. Snel, M. Huynen und P. Bork (1998). *Conservation of Gene Order: a Fingerprint of Proteins that Physically Interact*. Trends in Biochemical Sciences, 23:324–328.

Delcher, Arthur L., S. Kasif, R. D. Fleischmann, J. Peterson, O. White und S. L. Salzberg (1999). *Alignment of whole genomes*. Nucleic Acids Research, 27:2369–2376.

Demerec, M. und P. Hartman (1956). *Tryptophan mutants in Salmonella typhimurium*. Carnegie Inst. Washington Publ., 612:5–33.

Denver, Dee R., K. Morris, M. Lynch und W. K. Thomas (2004). *High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome*. Nature, 430(7000):679–682.

Devos, Katrien M., J. Beales, Y. Nagamura und T. Sasaki (1999). *Arabidopsis-Rice: Will Colinearity Allow Gene Prediction Across the Eudicot-Monocot Divide?*. Genome Research, 9:825–829.

D.Hurst, Laurence, C. Pal und M. J. Lercher (2004). *The evolutionary dynamics of eukaryotic gene order*. Nature Reviews Genetics, 5:299–310.

Dodeweerd, A.M. van, C. Hall, E. Bent, S. Johnson, M. Bevan und I. Bancroft (1999). *Identification and analysis of homoeologous segments of the genomes of rice and Arabidopsis thaliana*. Genome, 42(5):887–892.

Doerge, Rebecca W. (2002). *MAPPING AND ANALYSIS OF QUANTITATIVE TRAIT LOCI IN EXPERIMENTAL POPULATIONS*. Nature Reviews Genetics, 3(1):43–52.

Doolittle, W. Ford (1999). *Phylogenetic Classification and the Universal Tree*. Science, 284:2124–2128.

Eisen, Jonathan A. (2000). *Horizontal gene transfer among microbial genomes: new insights from complete genome analysis*. Current Opinion in Genetics And Devlopment, 10:606–611.

Enault, F., K. Suhre, C. Abergel, O. Poirot und J.-M. Claverie (2003). *Annotation of bacterial genomes using improved phylogenomic profiles*. Bioinformatics, 19(90001):i105–i107.

Enright, Anton J., I. Iliopoulos, N. C. Kyrpides und C. A. Ouzounis (1999). *Protein interaction maps for complete genomes based on gene fusion events*. Nature, 402:86–90.

Enright, Anton J. und C. Ouzounis (2000). *GeneRAGE: a robust algorithm for sequence clustering and domain detection*. Bioinformatics, 16(5):451–457.

Enright, Anton J. und C. Ouzounis (2001). *Functional association of proteins in entire genomes by means of exhaustive detection of gen fusions*. Genome Biology, 2(9):research0034.1–research0034.7.

Entrez Taxonomy Browser. *Entrez Taxonomy Browser*. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy.

Ermolaeva, Maria D., O. White und S. L. Salzberg (2001). *Prediction of Operons in Microbial Genomes*. Nucleic Acids Research, 29(5):1216–1221.

European Union Chromosome 3 Arabidopsis Sequencing Consortium, The Institute for Genomic Research und Kazusa DNA Research Institute (2000). *Sequence and Analysis of Chromosome 3 of the plant Arabidopsis thaliana*. Nature, 408:820–822.

FGDB. *Fusarium Graminearum Genome Database*. http://mips.gsf.de/genre/proj/fusarium/.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J.-F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. G. Sutton, W. FitzHugh, C. A. Fields, J. D. Gocayne, J. D. Scott, R. Shirley, L. I. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith und J. C. Venter (1995). *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd*. Science, 269(5223):496–512.

Force, Allan, M. Lynch, F. B. Pickett, A. Amores, Y.-l. Yan und J. Postlethwait (1999). *Preservation of Duplicate Genes by Complementary, Degenerative Mutations*. Genetics, 151(4):1531–1545. formulation of the DDC model for duplicated genes.

Friedman, Robert und A. L. Hughes (2001). *Gene Duplication and the Structure of Eukaryotic Genomes*. Genome Res., 11:373–381.

Frishman, D., M. Mokrejs, D. Kosykh, G. Kastenmüller, G. Kolesov, I. Zubrzycki, C. Gruber, B. Geier, A. Kaps, K. Albermann, A. Volz, C. Wagner, M. Fellenberg, K. Heumann und H.-W. Mewes (2003). *The PEDANT genome database*. Nucl. Acids. Res., 31(1):207–211.

Fusarium Synteny Viewer. *Fusarium Graminearum Synteny Viewer*. http://mips.gsf.de/genre/proj/fusarium/Synteny.

Galperin, Michael Y. und E. V. Koonin (1998). *Sources of Systematic Error in Functional Annotation of Genomes: Domain Rearrangement, Non-Orthologous Gene Displacement and Operon Disruption*. In Silico Biology, 1(1):55–67.

Galperin, Michael Y. und E. V. Koonin (2000). *Who's your neighbor? New computational approaches for functional genomics*. Nature Biotechnology, 18(6):609–613.

Goffeau, A., B. Barrell, H. Bussey, R. Davis, B. Dujon, H. Feldmann, F. Galibert, J. Hoheisel, C. Jacq, M. Johnston, E. Louis, H. Mewes, Y. Murakami, P. Philippsen, H. Tettelin und S. Oliver (1996). *Life with 6000 genes*. Science, 274(5287):546, 563–567.

Gogarten, J. Peter, W. F. Doolittle und J. G. Lawrence (2002). *Prokaryotic Evolution in Light of Gene Transfer*. Mol Biol Evol, 19(12):2226–2238.

Gu, Zhenglong, D. Nicolae, H. H.-S. Lu und W.-H. Li (2002). *Rapid divergence in expression between duplicate genes inferred from microarray data*. Trends in Genetics, 18(12):609–613.

Gusfield, Dan (1997). *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press.

Haberer, Georg, T. Hindemitt, B. C. Meyers und K. F. Mayer (2004). *Transcriptional Similarities, Dissimilarities, and Conservation of cis-Elements in Duplicated Genes of Arabidopsis*. Plant Physiol., 136(2):3009–3022.

Hacker, Jörg, U. Hentschel und U. Dobrindt (2003). *Prokaryotic chromosomes and disease*. Science, 301(5634):790–793.

Hannenhalli, S. und P. A. Pevzner (1995). *Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals)*. In: *Proceedings of the 27th Annual ACM-SIAM Symposium on the Theory of Computing*, S. 178–189, New York. ACM.

Hegyi, Hedi und M. Gerstein (2001). *Annotation Transfer for Genomics: Measuring Functional Divergence in Multi-Domain Proteins*. Genome Res., 11(10):1632–1640.

Heumann, Klaus, C. Harris und H. W. Mewes (1996). *A Top-Down Approach to Whole Genome Visualization*. In: *Proceedings of the ISMB*, Bd. 4, S. 98–108.

Hindemitt, Tobias (2003). *Entwicklung und Applikation eines bioinformatischen Systems zur Analyse regulatorischer Bereiche homologer Genpaare in Arabidopsis thaliana*. Diplomarbeit, Ludwig-Maximilians-Universität München.

Horn, Matthias, A. Collingro, S. Schmitz-Esser, C. L. Beier, U. Purkhold, B. Fartmann, P. Brandt, G. J. Nyakatura, M. Droege, D. Frishman, T. Rattei, H.-W. Mewes, und M. Wagner (2004). *Illuminating the Evolutionary History of Chlamydiae*. Science, 304:728–730.

Hughes, Austin L., J. da Silva und R. Friedman (2001). *Ancient Genome Duplications Did Not Structure the Human Hox-Bearing Chromosomes*. Genome Res., 11(5):771–780.

Huynen, Martijn, B. Snel, W. Lathe III und P. Bork (2000a). *Exploitation of Gene Context*. Current Opinion in Structural Biology, 10:366–370.

Huynen, Martijn, B. Snel, W. Lathe III und P. Bork (2000b). *Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences*. Genome Res., 10(8):1204–1210.

Huynen, Martijn, B. Snel, P. B. J. W. Stiller, B. D. H. R. S. Gupta und B. J. S. W. F. Doolittle (1999). *Lateral Gene Transfer, Genome Surveys, and the Phylogeny of Prokaryotes*. Science, 286:1443a.

Huynen, Martijn A. und P. Bork (1998). *Measuring genome evolution*. PNAS, 95:5849–5856.

Jacob, F. und J. Monod (1961). *Genetic regulatory mechansims in the synthesis of proteins*. J. Mol. Biol., 3:318–356.

Jansen, Ronald, H. J. Bussemaker und M. Gerstein (2003). *Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models*. Nucl. Acids Res., 31(8):2242–2251.

Kalman, Sue, W. Mitchell, R. Marathe, C. Lammel, J. Fan, R. W. Hyman, L. Olinger, J. Grimwood, R. W. Davis und R. Stephens (1999). *Comparative genomes of Chlamydia pneumoniae and C. trachomatis*. Nat Genet., 21(4):385–389.

Kanehisa, M. und S. Goto (2000). *KEGG: kyoto encyclopedia of genes and genomes*. Nucleic Acids Res., 28(1):27–30.

Kaplan, H., R. Shamir und R. Tarjan (1997). *Faster and Simpler Algorithm for Sorting Signed Permutations by Reversals*. In: *Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms*, S. 344–351, New York. ACM.

Karlin, Samuel (2001). *Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes*. TRENDS in Microbiology, 9(7):335–343.

Kececioglu, J. und D. Sankoff (1994). *Efficient Bounds for Oriented Chromosome Reversal Distance*. In: Crochemore, M. und D. Gusfield, Hrsg.: *Proceedings of the Fifth Symposium on Combinatorial Pattern Matching*, Bd. 807 d. Reihe *Lecture Notes in Computer Science*, S. 307–325, New York. Springer-Verlag.

Kececioglu, J. und D. Sankoff (1995). *Exact and Approximation Algorithms for Sorting by Reversals, with Application to Genome Rearrangement*. Algorithmica, 13:180–210.

Keeling, P.J., R. Charlebois und W. Doolittle (1994). *Archaabacterial genomes. Eubacterial form and eukaryotic content*. Current Opinion in Genetics and Development, 4:816–822.

Kellis, Manolis, B. W. Birren und E. S. Lander (2004). *Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae*. Nature.

Kellis, Manolis, N. Patterson, M. Endrizzi, B. Birren und E. S. Lander (2003). *Sequencing and comparison of yeast species to identify genes and regulatory elements*. Nature, 423:241–254.

Kent, W. James, R. Baertsch, A. Hinrichs, W. Miller und D. Haussler (2003). *Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes*. PNAS, 100(20):11484–11489.

Knight, Robin D., S. J. Freeland und L. F. Landweber (2001). *Rewiring the keyboard: evolvability of the genetic code*. Nat Rev Genet., 2(1).

Kobayashi, Ichizo (2001). *Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution*. Nucl. Acids. Res., 29(18):3742–3756.

Kobayashi, Ichizo, A. Nobusato, N. Kobayashi-Takahashi und I. Uchiyama (1999). *Shaping the genome – restriction-modification systems as mobilile genetic elements*. Current Opinion in Genetics and Development, 9:649–656.

Koch, Marcus A., B. Haubold und T. Mitchell-Olds (2000). *Comparative Evolutionary Analysis of Chalcone Synthase and Alcohol Dehydrogenase Loci in Arabidopsis, Arabis, and Related Genera (Brassicaceae)*. Mol Biol Evol, 17(10):1483–1498.

Kolesov, Grigory, H.-W. Mewes und D. Frishman (2001). *SNAPing up Functionally Related Genes Based on Context Information: A Colinearity-free Approach*. J. Mol. Biol., 311:639–656.

Kolesov, Grigory, H.-W. Mewes und D. Frishman (2002). *SNAPper: Gene Order Predicts Gene Function*. Bioinformatics, 18:1017–1019.

Kondrashov, Fyodor A., I. B. Rogozin, Y. I. Wolf und E. V. Koonin (2002). *Selection in the evolution of gene duplications*. Genome Biology, 3(2):research0008.1–0008.9.

Kreft, J., J.-A. Vazquez-Boland, S. Altrock, G. Dominguez-Bernal und W. Goebel (2002). *Pathogenicity Islands and Other Virulence Elements in Listeria*. Current Topics in Microbiology and Immunology, 264(2):109–125.

Ku, Hsin-Mei, T. Vision und S. D. Tanksley (2000). *Comparing sequenced segments of the tomato and Arabidopsis genomes: Large-scale duplication followed by selective gene loss creates a network of synteny*. PNAS, 97(16):9121–9126.

Kunin, Victor und C. A. Ouzounis (2003). *The Balance of Driving Forces During Genome Evolution in Prokaryotes*. Genome Res., 13(7):1589–1594.

Kurland, C. G., B. Canback und O. B. Berg (2003). *Horizontal gene transfer: A critical view*. PNAS, 100(17):9658–9662.

Kurtz, S und C. Schleiermacher (1999). *REPuter: fast computation of maximal repeats in complete genomes*. Bioinformatics, 15(5):426–427.

Larhammar, Dan, L.-G. Lundin und F. Hallbook (2002). *The Human Hox-bearing Chromosome Regions Did Arise by Block or Chromosome (or Even Genome) Duplications*. Genome Res., 12(12):1910–1920.

Lathe III, Warren C., B. Snel und P. Bork (2000). *Gene context conservation of a higher order than operons*. Trends in Biochemical Sciences, 25(10):474–479.

Lawrence, Jeffrey (1999). *Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes*. Current Opinion in Genetics and Development, 9:642–648.

Lawrence, Jeffrey G. und H. Ochman (1998). *Molecular archeology of the Escherichia coli genome*. PNAS, 95:9413–9417.

Lawrence, Jeffrey G. und J. R. Roth (1996). *Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters*. Genetics, 143:1843–1860.

Lee, Jennifer M. und E. L. L. Sonnhammer (2003). *Genomic Gene Clustering Analysis of Pathways in Eukaryotes*. Genome Res., 13(5):875–882.

Liu, Bao und J. F. Wendel (2003). *Epigenetic phenomena and the evolution of plant allopolyploids*. Mol Phylogenet Evol., 29(3):365–379.

Liu, Hong, R. Sachidanandam und L. Stein (2001). *Comparative Genomics Between Rice and Arabidopsis Shows Scant Collinearity in Gene Order*. Genome Res., 11:2020–2026.

Llorente, B., P. Durrens, A. Malpertuy, M. Aigle, F. Artiguenave, G. Blandin, M. Bolotin-Fukuhara, E. Bon, P. Brottier, S. Casaregola, B. Dujon, J. de Montigny, A. Lepingle, C. Neuveglise, O. Ozier-Kalogeropoulos, S. Potier, W. Saurin, F. Tekaia, C. Toffano-Nioche, M. Wesolowski-Louvel, P. Wincker, J. Weissenbac, J. Souciet und C. Gaillardin (2000). *Genomic exploration of the hemiascomycetous yeasts: 20. Evolution of gene redundancy compared to Saccharomyces cerevisiae*. FEBS Lett. 2000 Dec 22;487(1):122-33, 487(1):122–133.

Lynch, Michael und J. S. Conery (2000). *The Evolutionary Fate and Consequences of Duplicate Genes*. Science, 290:1151–1155.

Lynch, Michael und A. Force (2000). *The Probability of Duplicate Gene Preservation by Subfunctionalization*. Genetics, 154(1):459–473.

Makalowski, Wojciech (2001). *Are We Polyploids? A Brief History of One Hypothesis*. Genome Res., 11(5):667–670.

Mannhaupt, Gertrud, C. Montrone, D. Haase, H. W. Mewes, V. Aign, J. D. Hoheisel, B. Fartmann, G. Nyakatura, F. Kempken, J. Maier und U. Schulte (2003). *What's in the genome of a filamentous fungus? Analysis of the Neurospora genome sequence*. Nucleic Acids Research, 31(7):1944–1954.

Marcotte, Edward M. (2000). *Computational Genetics: Finding Protein Function by Nonhomology Methods*. Current Opinion in Structural Biology, 10:359–365.

Marcotte, Edward M., M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, und D. Eisenberg (1999). *Detecting Protein Function and Protein-Protein Interactions from Genome Sequences*. Science, 285(5428):751–753.

Martin, Maria J., J. Herrero, A. Mateos und J. Dopazo (2003). *Comparing Bacterial Genomes Through Conservation Profiles*. Genome Res., 13(5):991–998.

Mayer, Klaus, G. Murphy, R. Tarchini, R. Wambutt, G. Volckaert, T. Pohl, A. Düsterhöft, W. Stiekema, K.-D. Entian, N. Terryn, K. Lemcke, D. Haase, C. R. Hall, A.-M. van Dodeweerd, S. V. Tingey, H.-W. Mewes, M. W. Bevan und I. Bancroft (2001). *Conservation of Microstructure between a Sequenced Region of the Genome of Rice and Multiple Segments of the Genome of Arabidopsis thaliana*. Genome Research, 11:1167–1174.

Meinke, David W., J. M. Cherry, C. Dean, S. D. Rounsley und M. Koornneef (1998). *Arabidopsis thaliana: A Model Plant for Genome Analysis*. Science, 282:662,679–682.

Mering, Christian von, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork und B. Snel (2003a). *STRING: a Database of Predicted Functional Associations between Proteins*. Nucleic Acids Research, 31(1):258–261.

Mering, Christian von, E. M. Zdobnov, S. Tsoka, F. D. Ciccarelli, J. B. Pereira-Leal, C. A. Ouzounis und P. Bork (2003b). *Genome evolution reveals biochemical networks and functional modules*. PNAS, 100(26):15428–15433.

Mewes, H. W., D. Frishman, C. Gruber, B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt, F. Pfeiffer, C. Schuller, S. Stocker und B. Weil (2000). *MIPS: a database for genomes and protein sequences*. Nucl. Acids. Res., 28(1):37–40.

Mewes, H.W., K. Albermann, M. Baehr, D. Frishman, A. Gleissner, J. Hani, K. Heumann, K. Kleine, A. Maierl, S. Oliver, F. Pfeiffer

und A. Z. A. (1997). *Overview of the yeast genome*. Nature, 387(6632 Suppl.):7–66.

Moreno-Hagelsieb, Gabriel und J. Collado-Vides (2002). *A Powerful Non-Homology Method for the Prediction of Operons in Prokaryotes*. Bioinformatics, 18(Suppl. 1):S329–S336.

Moret, Bernard M.E., L.-S. Wang, T. Warnow und S. K. Wyman (2001). *New approaches for reconstructing phylogenies from gene order data*. Bioinformatics, 17(Suppl. 1):S165–S173.

Morgenstern, Burkhard (1999). *DIALIGN2: Improvement of the segment-to-segment approach to mulitple sequence alignment*. Bioinformatics, 15:211–218.

Morgenstern, Burkhard, O. Rinner, S. Abdeddaïm, D. Haase, K. F. Mayer, A. Dress und H.-W. Mewes (2002). *Exon Prediction by Genomic Sequence Alignment*. Bioinformatics, 18:777–787.

Mouse Genome Sequencing Consortium (2002). *Initial sequencing and comparative analysis of the mouse genome*. Nature, 420:520–562.

Mushegian, Arcady R. und E. V. Koonin (1996). *Gene order is not conserved in bacterial evolution*. Trends in Genetics, 12(8):289–290.

Nadeau, Joseph H. und B. A. Taylor (1984). *Lengths of chromosomal segments conserved since divergence of man and mouse*. PNAS, 81(3):814–818.

NCBI BLAST. *NCBI BLAST Search*. http://www.ncbi.nlm.nih.gov/BLAST/.

Nimwegen, Erik van, M. Zavolan, N. Rajewsky und E. D. Siggia (2002). *Probabilistic Clustering of Sequences: Inferring New Bacterial Regulons by Comparative Genomics*. PNAS, 99(11):7323–7328.

Ochman, Howard, J. G. Lawrence und E. A. Groisman (2000). *Lateral gene transfer and the nature of bacterial innovation*. Nature, 405:299–304.

O'Donnell, Kerry, H. C. Kistler, B. K. Tacke und H. H. Casper (2000). *Gene genealogies reveal global phylogeographic structure and reproductive isolation among lineages of Fusarium graminearum, the fungus causing wheat scab*. PNAS, 97(14):7905–7910.

Ogata, Hiroyuki, W. Fujibuchi, S. Goto und M. Kanehisa (2000). *A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters*. Nucl. Acids. Res., 28(20):4021–4028.

Ohno, Susumu (1970). *Evolution by Gene Duplication*. Springer-Verlag, Heidelberg, Germany.

Overbeek, Ross, M. Fonstein, M. D'Souza und N. Maltsev (1999). *The Use of Gene Clusters to Infer Functional Coupling*. PNAS, 96:2896–2901.

Overbeek, Ross, M. Fonstein, M. D'Souza, G. Pusch und N. Maltsev (1998). *Use of Contguity on the Chromosome to Predict Functional Coupling*. In Silico Biology, 1:0009.

Ozkan, Hakan, A. A. Levy und M. Feldman (2001). *Allopolyploidy-Induced Rapid Genome Evolution in the Wheat (Aegilops-Triticum) Group*. Plant Cell, 13(8):1735–1747.

Pagel, Philipp, P. Wong und D. Frishman (2004). *A Domain Interaction Map Based on Phylogenetic Profiling*. J. Mol. Biol., 344(5):1331–1346.

Panopoulou, Georgia, S. Hennig, D. Groth, A. Krause, A. J. Poustka, R. Herwig, M. Vingron und H. Lehrach (2003). *New Evidence for Genome-Wide Duplications at the Origin of Vertebrates Using an Amphioxus Gene Set and Completed Animal Genomes*. Genome Res., 13(6a):1056–1066.

Parenicova, Lucie, S. de Folter, M. Kieffer, D. S. Horner, C. Favalli, J. Busscher, H. E. Cook, R. M. Ingram, M. M. Kater, B. Davies, G. C. Angenent und L. Colombo (2003). *Molecular and Phylogenetic Analyses of the Complete MADS-Box Transcription Factor Family in Arabidopsis: New Openings to the MADS World*. Plant Cell, 15(7):1538–1551.

Passarge, Eberhard, B. Horsthenke und R. A. Farber (1999). *Incorrect use of the term synteny*. Natuer Genetics, 23:387.

Paterson, Adndrew H., J. E. Bowers, M. D. Burow, X. Draye, C. G. Elsik, C.-X. Jiang, C. S. Katsar, T.-H. Lan, Y.-R. Lin, R. Ming und R. J. Wright (2000). *Comparative Genomics of Plant Chromosomes*. The Plant Cell, 12:1523–1539.

Paterson, A.H., T. Lan, K. Reischmann, C. Chang, Y. Lin, S. Liu, M. Burow, S. Kowalski, C. Katsar, T. DelMonte, K. Feldmann, K. Schertz und J. Wendel (1996). *Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence*. Nature Genetics, 14(4):380–382.

Pearson, W.R. und D. Lipman (1988). *Improved tools for biological sequence comparison*. PNAS, 85(8):2444–2448.

Pedant. *Protein Extraction, Description and Analysis Tool*. http://pedant.gsf.de.

Pe'er, Itsik und R. Shamir (1998). *The median problems for breakpoints are NP-complete*. ECCC Report TR98-071, Electronic Colloquium on Computational Complexity.

Pellegrini, Matteo, E. M. Marcotte, M. J. Thompson, D. Eisenberg und T. O. Yeates (1999). *Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles*. PNAS, 96(8):4285–4288.

Pevzner, Pavel und G. Tesler (2003a). *Genome Rearrangements in Mammalian Evolution: Lessons From Human and Mouse Genomes*. Genome Res., 13(1):37–45.

Pevzner, Pavel und G. Tesler (2003b). *Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution*. PNAS, 100(13):7672–7677.

Prince, Victoria E. und F. B. Pickett (2002). *Splitting pairs: the diverging fates of duplicated genes*. Nat Rev Genet., 3(11):827–837.

Read, T. D., R. C. Brunham, C. Shen, S. R. Gill, J. F. Heidelberg, O. White, E. K. Hickey, J. Peterson, T. Utterback, K. Berry, S. Bass, K. Linher, J. Weidman, H. Khouri, B. Craven, C. Bowman, R. Dodson, M. Gwinn, W. Nelson, R. DeBoy, J. Kolonay, G. McClarty, S. L. Salzberg, J. Eisen und C. M. Fraser (2000). *Genome sequences of Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39*. Nucl. Acids. Res., 28(6):1397–1406.

Read, T. D., G. S. A. Myers, R. C. Brunham, W. C. Nelson, I. T. Paulsen, J. Heidelberg, E. Holtzapple, H. Khouri, N. B. Federova, H. A. Carty, L. A. Umayam, D. H. Haft, J. Peterson, M. J. Beanan, O. White, S. L. Salzberg, R. c. Hsia, G. McClarty, R. G. Rank, P. M. Bavoil und C. M. Fraser (2003). *Genome sequence of Chlamydophila caviae (Chlamydia psittaci GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae*. Nucl. Acids. Res., 31(8):2134–2147.

Rinner, Oliver und B. Morgenstern (2001). *AGenDA: Gene prediction by comparative sequence analysis*. In Silico Biology, S. 0018.

Rogozin, Igor B., K. S. Makarova, J. Murvai, E. Czabarka, Y. I. Wolf, R. L. Tatusov, L. A. Szekely und E. V. Koonin (2002). *Connected Gene Neighborhoods in Prokaryotic Genomes*. Nucleic Acids Research, 30(10):2212–2223.

Rubin, G.M., M. Yandell, J. Wortman, G. G. Miklos, C. Nelson, I. Hariharan, M. Fortini, P. Li, R. Apweiler, W. Fleischmann, J. Cherry, S. Henikoff, M. Skupski, S. Misra, M. Ashburner, E. Birney, M. Boguski, T. Brody, P. Brokstein, S. Celniker, S. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R. Galle, W. Gelbart, R. George, L. Goldstein, F. Gong, P. Guan, N. Harris, B. Hay, R. Hoskins, J. Li, Z. Li, R. Hynes, S. Jones, P. Kuehl, B. Lemaitre, J. Littleton, D. Morrison, C. Mungall, P. O'Farrell, O. Pickeral, C. Shue, L. Vosshall, J. Zhang, Q. Zhao, X. Zheng

und S. Lewis (2000). *Comparative genomics of the eukaryotes*. Science, 287:2204–2215.

Ruepp, Andreas, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Guldener, G. Mannhaupt, M. Munsterkotter und H. W. Mewes (2004). *The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes*. Nucl. Acids Res., 32(18):5539–5545.

Ruopp, Martin und D. Haase (2000). *Arabidopsis Redundancy Viewer*. http://mips.gsf.de/proj/thal/db/gv/rv/view.html.

Salgado, Heladia, G. Moreno-Hagelsieb, T. F. Smith und J. Collado-Vides (2000). *Operons in Escherichia coli: Genomic analyses and predictions*. PNAS, 97(12):6652–6657.

Salse, Jerome, B. Piegu, R. Cooke und M. Delseny (2002). *Synteny between Arabidopsis thaliana and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project*. Nucl. Acids. Res., 30(11):2316–2328.

Samen, Ulrike, B. Gottschalk, B. J. Eikmanns und D. J. Reinscheid (2004). *Relevance of Peptide Uptake Systems to the Physiology and Virulence of Streptococcus agalactiae*. J. Bacteriol., 186(5):1398–1408.

Sankoff, David (2001). *Gene and Genome Duplication*. Current Opinion in Genetics and Development, 11:681–684.

Sankoff, David und M. Blanchette (1999). *Phylogenetic Invariants for Genome Rearrangements*. Journal of Computational Biology, 6:431–445.

Sankoff, David und J. H. Nadeau (2003). *Chromosome rearrangements in evolution: From gene order to genome sequence and back*. PNAS, 100(20):11188–11189.

Schoof, Heiko, R. Ernst, V. Nazarov, L. Pfeifer, H.-W. Mewes und K. F. X. Mayer (2004). *MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource for plant genomics*. Nucl. Acids. Res., 32(90001):D373–376.

Seoighe, Cathal und K. H. Wolfe (1998). *Extent of genomic rearrangement after genome duplication*. PNAS, 95:4447–4452.

Seoighea, Cathal und C. Gehring (2004). *Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome*. Trends in Genetics, 20(10):461–464.

Shirai, Mutsunori, H. Hirakawa, M. Kimoto, M. Tabuchi, F. Kishi, K. Ouchi, T. Shiba, K. Ishii, M. Hattori, S. Kuhara und T. Nakazawa (2000). *Comparison of whole genome sequences of Chlamydia pneumoniae J138 from Japan and CWL029 from USA*. Nucl. Acids. Res., 28(12):2311–2314.

Simillion, Cedric, K. Vandepoele, M. C. E. Van Montagu, M. Zabeau und Y. Van de Peer (2002). *The hidden duplication past of Arabidopsisthaliana*. PNAS, 99(21):13627–13632.

Smith, T.F. und M. Waterman (1981). *Identification of common molecular subsequences*. J. Mol. Biol., 147:195–197.

Snel, B., G. Lehmann, P. Bork und M. Huynen (2000a). *STRING: A Web-Server to Retrieve and Display the Repeatedly Occurring Neighbourhood of a Gene*. Nucleic Acids Research, 28(18):3442–3444.

Snel, Berend, P. Bork und M. Huynen (2000b). *Genome evolution: gene fusion versus gene fission*. Trends in Genetics, 16(1):9–1.

Soltis, Douglas E. und P. S. Soltis (1999). *Polyploidy: recurrent formation and genome evolution*. Trends in Ecology and Evolution, 14(9):348–352.

Somerville, Chris und M. Koornneef (2002). *A fortunate choice: the history fo Arabidopsis thaliana as a model plant*. Nature Reviews Genetics, 3(11):883–889.

Song, K, P. Lu, K. Tang und T. Osborn (1995). *Rapid Genome Change in Synthetic Polyploids of Brassica and Its Implications for Polyploid Evolution*. PNAS, 92(17):7719–7723.

Spellman, Paul T. und G. M. Rubin (2002). *Evidence for large domains of similarly expressed genes in the Drosophila genome*. Journal of Biology, 1(1):5.1–5.8.

Stein, Lincoln (2001). *Genome Annotation: From Sequence to Biology*. Nature Reviews Genetics, 2:493–505.

Stein, Lincoln D., Z. Bao, D. Blasiar, T. Blumenthal, M. R. Brent, N. Chen, A. Chinwalla, L. Clarke, C. Clee, A. Coghlan, A. Coulson, P. D'Eustachio, D. H. A. Fitch, L. A. Fulton, R. E. Fulton, S. Griffiths-Jones, T. W. Harris, L. W. Hillier, R. Kamath, P. E. Kuwabara, E. R. Mardis, M. A. Marra, T. L. Miner, P. Minx, J. C. Mullikin, R. W. Plumb, J. Rogers, J. E. Schein, M. Sohrmann, J. Spieth, J. E. Stajich, C. Wei, D. Willey, R. K. Wilson, R. Durbin und R. H. Waterston (2003). *The Genome Sequence of Caenorhabditis briggsae: A Platform for Comparative Genomics*. PLoS Biology, 1(2):166–192.

Stephens, Richard S., S. Kalman, C. Lammel, J. Fan, R. Marathe, L. Aravind, W. Mitchell, L. Olinger, R. L. Tatusov, Q. Zhao, E. V. Koonin und R. W. Davis (1998). *Genome Sequence of an Obligate Intracellular Pathogen of Humans: Chlamydia trachomatis*. Science, 23(5389):754 – 759.

STRING. *STRING - Search Tool for the Retrieval of Interacting Genes/Proteins*. http://string.embl.de/.

Stöcker, Prof. Dr. Horst, Hrsg. (1993). *Taschenbuch mathematischer Formeln und moderner Verfahren*. Verlag Harri Deutsch, Thun und Frankfurt am Main, 2 Aufl.

Tagle, D.A., B. Koop, M. Goodman, J. Slightom, D. Hess und R. Jones (1988). *Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints*. J. Mol. Biol., 203:439–455.

Tamames, Javier, G. Casari, C. Ouzounis und A. Valencia (1997). *Conserved Clusters of Functionally Related Genes in Two Bacterial Genomes*. Journal of Molecular Evolution, 44:66–73.

Tatusov, Roman L., E. V. Koonin und D. J. Lipman (1997). *A Genomic Perspective on Protein Families*. Science, 278(5338):631–637.

Teichmann, Sarah Amalia und R. A. Veitia (2004). *Genes Encoding Subunits of Stable Complexes Are Clustered on the Yeast Chromosomes: An Interpretation From a Dosage Balance Perspective*. Genetics, 167(4):2121–2125.

Tesler, Glenn (2002a). *Efficient algorithms for multichromosomal genome rearrangements*. Journal of Computer and System Sciences, 65:587–609.

Tesler, Glenn (2002b). *GRIMM: genome rearrangements web server*. Bioinformatics, 18(3):492–493.

The Arabidopsis Genome Initiative (2000). *Analysis of the Genome Sequence of the Flowering Plant Arabidopsis thaliana*. Nature, 408:796–815.

The European Union Arabidopsis Genome Sequencing Consortium, The Cold Spring Harbor, Washington University in St Louis und PE Biosystems Arabidopsis Sequencing Consortium (1999). *Sequence analysis of chromosome 4 of the plant Arabidopsis thaliana*. Nature, 402:769–777.

The Gene Ontology Consortium (2001). *Creating the Gene Ontology Resource: Design and Implementation*. Genome Res., 11(8):1425–1433.

Ureta-Vidal, Abel, L. Ettwiller und E. Birney (2003). *Comparative genomics: genome-wide analysis in metazoan eukaryotes*. Nature Reviews Genetics, 4:251–262.

Vision, Todd J., D. G. Brown und S. D. Tanksley (2000). *The Origins of Genomic Duplications in Arabidopsis*. Science, 290:2114–2117.

Wambutt, R., C. Bielke, D. Frishman, D. Haase, K. Lemcke, H. Mewes, S. Stocker, P. Zaccaria, K. Mayer, C. Schueller und M. Bevan (2000). *Progress in Arabidopsis genome sequencing and functional genomics*. Journal of Biotechnology, 78(3):281–292.

Watanabe, H., H. Mori, I. Takeshi und T. Gojobori (1997). *Genome plasticity as a paradigm of eubacterial evolotion*. Journal of Molecular Evolution, 44:S57–S64.

Waterman, Michal S., S. Tavare und R. C. Deonier (in Vorbereitung). *Computational Genome Analysis: An Introduction*. Springer-Verlag, Heidelberg, Berlin, New York.

Wendel, Jonathan F. (2000). *Genome evolution in polyploids*. Plant Mol. Biol., 42:225–249.

Werner, Thomas, S. Fessele, H. Maier und P. J. Nelson (2003). *Computer modeling of promoter organization as a tool to study transcriptional coregulation*. FASEB J., 17(10):1228–1237.

Wilson, Cyrus A., J. Kreychman und M. Gerstein (2000). *Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores*. J Mol Biol., 297(1):233–249.

Woese, Carl (1998). *The universal ancestor*. PNAS, 95(12):6854–6859.

Woese, Carl R. (2000). *Interpreting the universal phylogenetic tree*. PNAS, 97(15):8392–8396.

Wolf, Yuri I., I. B. Rogozin, A. S. Kondrashov und E. V. Koonin (2001). *Genome Alignment, Evolution of Prokaryotic Genome Organization, and Prediction of Gne Function Using Genomic Context*. Genome Res., 11:356–372.

Wolfe, Kenneth H. (2001). *Yesterday's polyploids and the mystery of diploidization*. Nat Rev Genet., 2(5):333–341.

Wolfe, Kenneth H. und D. C. Shields (1997). *Molecular evidence for an ancient duplication of the entire yeast genome*. Nature, 387:708–713.

Wolfe, K.H., M. Gouy, Y.-W. Yang, P. Sharp und W.-H. Li (1989). *Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data*. Proc. Natl. Acad. Sci., 86:6201–6205.

Wong, Gane Ka-Shu, J. Wang, L. Tao, J. Tan, J. Zhang, D. A. Passey und J. Yu (2002). *Compositional Gradients in Gramineae Genes*. Genome Res., 12(6):851–856.

Wright, Stephen I. und B. S. Gaut (2005). *Molecular Population Genetics and the Search for Adaptive Evolution in Plants*. Mol Biol Evol, 22(3):506–519.

Yada, Tetsushi, M. Nakao, Y. Totoki und K. Nakai (1999). *Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models*. Bioinformatics, 15(12):987–993.

Yamanishi, Y., J.-P. Vert, A. Nakaya und M. Kanehisa (2003). *Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis*. Bioinformatics, 19(90001):i323–i330.

Yanai, Itai, J. C. Mellor und C. DeLisi (2002a). *Identifying functional links between genes using conserved chromosomal proximity*. TRENDS in Genetics, 18(4):176–179.

Yanai, Itai, Y. I. Wolf und E. V. Koonin (2002b). *Evolution of gene fusions: horizontal transfer versus independent events*. Genome Biology, 3(5):research0024.1–0024.13.

Yanofski, C. und E. S. Lennox (1959). *Transduction and recombination study of linkage relationships among the genes controlling tryptophan synthesis in Escherichia coli*. Virology, 8:425–447.

Yu, Jun, S. Hu, J. Wang, G. K.-S. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, X. Huang, W. Li, J. Li, Z. Liu, L. Li, J. Liu, Q. Qi, J. Liu, L. Li, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Zhang, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Ren, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, J. Wang, W. Zhao, P. Li, W. Chen, X. Wang, Y. Zhang, J. Hu, J. Wang, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, G. Li, S. Liu, M. Tao, J. Wang, L. Zhu, L. Yuan und H. Yang (2002). *A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica) subspecies*. Science, 296:79–92.

Zhang, Zhaolei und M. Gerstein (2003). *Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes*. Nucl. Acids Res., 31(18):5338–5348.

Zheng, Xiangqun H., F. Lu, Z.-Y. Wang, F. Zhong, J. Hoover und R. Mural (2005). *Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs*. Bioinformatics, 21(6):703–710.

Zheng, Yu, R. J. Roberts und S. Kasif (2002a). *Genomic functional annotation using co-evolution patterns of gene clusters*. Genome Biology, 3(11):research0060.1–research0060.9.

Zheng, Yu, J. D. Szustakowski, L. Fortnow, R. J. Roberts und S. Kasif (2002b). *Computational Identification of Operons in Microbial Genomes*. Genome Res., 12:1221–1230.

Ziolkowski, Piotr A., G. Blanc und J. Sadowski (2003). *Structural divergence of chromosomal segments that arose from successive duplication events in the Arabidopsis genome*. Nucl. Acids. Res., 31(4):1339–1350.