Aspekte der bioinformatischen Analyse und Annotation des Genoms von *Rhodopirellula baltica*^{T}

•••••

Aspects of the bioinformatic analysis and annotation of the genome of *Rhodopirellula baltica*^{T}

•••••

Dissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
- Dr. rer. nat. -

•••••

dem Fachbereich Biologie/Chemie der
Universität Bremen
vorgelegt von

Hanno Teeling

im Februar 2004

Diese Arbeit entstand in der Zeit vom Oktober 1999 bis zum Februar 2004 im Rahmen des REGX-Projekts am Max-Planck-Institut für marine Mikrobiologie in Bremen. Das REGX-Projekt (REGX = Real Environmental Genomics; http://www.regx.de) ist eine vom BMBF und der Max-Planck-Gesellschaft geförderte, deutsche Initiative zur Totalsequenzierung, Annnotation und funktionellen Analyse mariner Bakterien.

1. Gutachter: Prof. Dr. Rudolf Amann
2. Gutachter: Dr. Wolfgang Ludwig

Tag des Promotionskolloquiums: 22.03.2004

Inhaltsverzeichnis

Abkürzungsverzeichnis

A	Adenin
Abb.	Abbildung
Anammox	*anaerobic ammonium oxidation*; anaerobe Ammonium-Oxidation
ARB	bekanntes Phylogenie-Programm (von lat. *arbor* = Baum)
ATP	Adenosintriphosphat
BAC	*bacterial artificial chromosome*
BLAST	*basic local alignment search tool*
BLASTp	BLAST-Suche einer Aminosäuresequenz gegen eine Datenbank von Aminosäuresequenzen (p = *protein*)
bp	*base pairs*; Basenpaare
BMBF	Bundesministerium für Bildung und Forschung
C	Cytosin
CAI	*codon adaptation index*
d.h.	das heißt
COG	*cluster of orthologous groups*
CRITICA	*coding region identification tool invoking comparative analysis*
DNA	Desoxyribonukleinsäure
DPS	*DNA-protein search*
EC	*Enyzme Commission*
EMBL	*European Molecular Biology Laboratory*
ER	Endoplasmatisches Retikulum
ERGO	Annotationssystem von Integrated Genomics Inc., Chicago
G	Guanin
GenDB	g*ene database* - *Open Source*-Annotationssystem der Universität Bielefeld
GENEWIZ	*gene wizard* - ein Programm zur Visualisierung von Genomstrukturen
Glimmer	*gene locator and interpolated Markov modeler*
GSP	*general secretion pathway*
HMMER	Programmpaket zur Erstellung von und Analyse mittels *hidden Markov*-Modellen
HSP	*heat shock protein*
HTML	*hypertext markup language*
ICM	*interpolated context model*
KI	künstliche Intelligenz
LGT	*lateral gene transfer*; lateraler Gentransfer
LPS	Lipopolysaccharid
MCP	*methyl accepting chemotaxis protein*
MORFind	*Max Planck Institute open reading frame finder*

MySQL	eine Open-Source-Variante der *structured query language* - einer Ende der 70er Jahre von IBM entwickelten Datenbankabfrage-Sprache
Nc	*number of codons*
NCBI	*National Center for Biotechnology Information*
nr	*non-redundant*
ORF	*open reading frame*
ORI	*origin of replication*; Replikationsursprung
ORPHEUS	ein Vorhersageprogramm für proteinkodierende Gene in Prokaryonten
PA	*predicted alien*
PDB	*protein data bank*
PEDANT	*protein extraction, description and analysis tool*
PERL	*practical extraction and report language*
pers.	persönlich
Pfam	*database of protein families*
PHX	*predicted highly expressed*
Pkt.	Punkt
PROSITE	*database of protein sites (families and domains)*
RDBMS	*relational database management system*
REGX	*Real Environmental Genomics*
resp.	respektive
RBS	*ribosomal binding site*
RNA	Ribonukleinsäure
rRNA	ribosomale Ribonukleinsäure
SCOP	*structural classification of protein*s
s.o.	siehe oben
sog.	sogenannt
SRS	*sequence retrieval system*
T	Thymin
Tab.	Tabelle
Tat	*twin arginine translocation*
TER	*terminus of replication*; Replikationsterminus
TIGR	*The Institute for Genomic Research*
TU	Technische Universität
UV	ultra-violett
vgl.	vergleiche
tRNA	transfer-Ribonukleinsäure
z.B.	zum Beispiel
z.T.	zum Teil

Zusammenfassung

In dieser Arbeit sind wesentliche Ergebnisse der bioinformatischen Analyse und Annotation des Genoms von *Rhodopirellula baltica*^{T} zusammengefaßt. *Rhodopirellula baltica*^{T} ist ein mariner Planktomycet, dessen Genom im Rahmen des REGX-Projekts mittels der *Shotgun*-Methode totalsequenziert wurde.

Zu Beginn des Projekts wurde eine bioinformatische Arbeitspipeline aus Tools zur Genvorhersage, Annotation und Visualisierung aufgebaut und etabliert. Anschließend wurde nahezu ein Jahr auf eine qualitativ-hochwertige, manuelle Annotation der Gene verwandt.

Wie die Annotation des Genoms zeigte, ist *Rhodopirellula baltica*^{T} auf die vollständige aerobe Oxidation komplexer Kohlenhydrate spezialisiert. In diesem Zusammenhang ist interessant, daß das Genom nicht weniger als 110 Gene enthält, die Sulfatasen kodieren. Diese Sulfatasen werden wahrscheinlich zu einem großen Teil sezerniert und ermöglichen dem Organismus, in marinem Milieu abundante, sulfatisierte Saccharide abzubauen (z.B. Chondroitin-sulfat oder Carragen). Zuckermonomere werden von *Rhodopirellula baltica*^{T} über die Glykolyse, den vollständigen Tricarbonsäurezyklus und die aerobe Respiration totaloxidiert. Letztere kann über zwei alternative terminale Oxidasen verlaufen, von denen eine vermutlich nur unter sauerstofflimitierenden Bedingungen exprimiert wird. Zudem hat *Rhodopirellula baltica*^{T} das genetische Potential, um Zuckermonomere vom Pentose-5-Phosphat-Weg abzweigend über die heterofermentative Milchsäuregärung vergären zu können. Dies ist überraschend, da *Rhodopirellula baltica*^{T} bislang nicht unter anaeroben Bedingungen kultiviert werden konnte. Ein weiteres, interessantes Ergebnis der Annotation ist, daß das *Rhodopirellula baltica*^{T}-Genom nahezu alle Gene des Tetrahydromethanopterin-abhängigen C1-Metabolismus enthält. Diese waren bislang nur von methylotrophen Proteobakterien und methanogenen Archaeen bekannt.

Phylogenetische Analysen einer Vielzahl von Markergenen aus dem *Rhodopirellula baltica*^{T}-Genom sowie aus dem noch ungeschlossenen Genom des limnischen Planktomyceten *Gemmata obscuriglobus* UQM2246^T legen eine Verwandtschaft der Planktomyceten mit den Chlamydien oder wenigstens mit dem Chlamydien-Spirochaeten-Supercluster nahe. Dieses Ergebnis bestätigt ältere, 16S rRNA-basierte phylogenetische Analysen und widerspricht der jüngst vorgeschlagenen Position der Planktomyceten als am tiefsten abzweigendem Phylum innerhalb der Domäne *Bacteria*. Außerdem finden sich im *Rhodopirellula baltica*^{T}-Genom viele Gene zur Peptidoglycan-Biosynthese, was darauf hindeutet, daß es sich bei den Proteinzellwänden der Planktomyceten um sekundäre Adaptionen und nicht um Relikte aus einer Zeit vor der Evolution des Peptidoglycans handelt. Zudem deuten Indizien, wie z.B. das Vorhandensein aller Gene zur Lipid A-Biosynthese, darauf hin, daß sich die Planktomyceten aus Gramnegativen Bakterien mit einer Cytoplasma- und einer äußeren Membran entwickelt haben.

Das ribosomenfreie, äußere Kompartiment der Planktomyceten ist daher vermutlich dem periplasmatischen Raum Gram-negativer Bakterien homolog.

Aus der Annotation des *Rhodopirellula baltica*^{T}-Genoms sind zwei eigenständige Softwareprojekte hervorgegangen. Das eine ist MORFind, ein flexibles und erweiterbares Meta-Programm zur *in silico* Vorhersage proteinkodierender Gene. Das zweite ist ein auf der Analyse von Tetranukleotid-Häufigkeiten in DNA-Sequenzen basierendes Verfahren, das dabei helfen kann, Genomfragmente aus Metagenomprojekten auch in Abwesenheit phylogenetischer Marker einander zuzuordnen. Dieses Verfahren wurde mit dem Programm TETRA in Form eines benutzerfreundlichen Programms implementiert.

Zusammengefaßt hat das Genom von *Rhodopirellula baltica*^{T} eine Fülle interessanter Einblicke in die Biologie eines marinen Planktomyceten gewährt und dazu beigetragen, wenigsten einige der lange unbeantworteten Fragen über Planktomyceten zu klären. Zudem wurden neue Fragen aufgeworfen und weitere Untersuchungen initiiert. Schließlich hat sich das *Rhodopirellula baltica*^{T}-Genom als geeignetes Modell zum Aufbau und zur Etablierung einer bioinformatischen Pipeline erwiesen, die bei der Analyse weiterer Genome und genomischer Fragmente im Rahmen des REGX-Projekts von großem Nutzen sein wird.

Summary

This thesis focuses on the bioinformatic analysis and annotation of the genome of the marine planctomycete *Rhodopirellula baltica*^{T}, whose genome sequence was determined via a shotgun approach within the framework of the REGX project.

In order to achieve a high quality annotation, a comprehensive bioinformatic pipeline was set up and established that, among other things, comprises gene prediction, annotation and visualization tools. In addition, considerable effort was put into the manual annotation process itself which took close to a year to finish.

The annotation of the genome of *Rhodopirellula baltica*^{T} revealed that this organism is specialized on the aerobic degradation of complex carbohydrates. Its genome harbors the exceptional number of 110 sulfatases, which are likely excreted to get access to the carbon skeletons of sulfated marine saccharides like chondroitin sulfate or carrageen. Sugar-monomers are taken up via dedicated transporters and subsequently degraded via glycolysis, the tricarboxic acid cycle and oxic respiration. The latter employs two alternate terminal oxidases, of which one is likely an adaption to cope with low oxygen concentrations. Furthermore, its genome harbors not only genes for the pentose-5-phosphate cycle but also for heterolactic acid fermentation, despite the fact that *Rhodopirellula baltica*^{T} has not yet been successfully cultured under anoxic conditions. Another interesting finding is the a near by complete set of genes for tetrahydro-methanopterin-dependent C1 carbon metabolism, which previously were only known from methylotrophic *Proteobacteria* and methanogenic *Archaea*.

Phylogenetic analysis of numerous marker genes from the genome of *Rhodopirellua baltica*^{T} and the unfinished genome of the freshwater planctomycete *Gemmata obscuriglobus* UQM2246^T support a relationship with the *Chlamydia* or at least an affiliation with the *Chlamydia-Spirochaetes* superclade. This is in agreement with earlier 16S rRNA-based analysis and contradicts the recently proposed deepest branching position of the *Planctomycetes* within the bacterial domain. The presence of many genes for peptidoglycan biosynthesis in the genome of *Rhodopirellula baltica*^{T} indicates that the proteinaceous cell envelopes of planctomycetes are secondary adaptions rather than a relict from a time preceding peptidoglycan cell walls. This is also consistent with findings from the *Rhodopirellula baltica*^{T} genome that indicate that the *Planctomycetes* originated from organisms with Gram negative-type cell envelopes (e.g. genes for lipid A biosynthesis) and suggest that the outer ribosome-free compartments of *Planctomycetes* descended from a former periplasmic space.

Two spin-off projects from the annotation of the *Rhodopirellula baltica*^{T} genome have lead to independent software developments. The first is MORFind, a flexible and extensible meta-tool for the *in silico* prediction of protein-coding genes. The second is a method for the probability-based assignment of genomic fragments originating from metagenome projects that is based on skewed tetranucleotide distributions. A dedicated program named TETRA has been developed for this purpose.

Altogether, the annotation of the genome of *Rhodopirellula baltica*^{T} has delivered many interesting answers to long time pending questions regarding the biology of the *Planctomycetes*. Just as well, many new questions arose that triggered further research. Finally, *Rhodopirellula baltica*^{T} proved to be a well-suited model organism for the establishment of a working bio-informatic pipeline aimed at future analysis of genomes and genomic fragments within the framework of the REGX project.

Teil I:
Darstellung der Ergebnisse im Gesamtzusammenhang

A Einleitung

1. Der wissenschaftliche Rahmen: Das REGX-Projekt

Das REGX-Projekt (Real Environmental Genomics) ist eine vom Max-Planck-Institut für marine Mikrobiologie in Bremen initiierte Gemeinschaftsinitiative deutscher Forschungsinstitute zur vollständigen Sequenzierung, bioinformatischen und funktionellen Analyse mariner Bakterien. Die Schlüsselmotivation zu diesem Projekt bildete die Tatsache, daß die Genomforschung zu Projektbeginn im wesentlichen medizinisch-pharmazeutisch, agrar-ökonomisch und biotechnologisch orientiert war. Infolgedessen handelte es sich bei den seinerzeit weniger als 40 totalsequenzierten Bakterien- und Archaeen-Genomen überwiegend um humanpathogene, pflanzenpathogene und extremophile Spezies. Die damit einhergehenden physiologischen und genomischen Eigenschaften dieser Spezies sowie die starke Über- und Unterrepräsentation einzelner Phyla ermöglichten so gut wie keine Rückschlüsse auf mikrobielle Prozesse an natürlichen Standorten - wenn man von einigen Habitaten mit physikochemisch extremen Bedingungen und geringer Diversität absieht.

Um einen besseren Einblick in die Biologie umweltrelevanter, mariner Bakterien zu gewinnen, wurden im Rahmen des REGX-Projekts die Genome dreier Isolate zur Totalsequenzierung ausgewählt: *Rhodopirellula baltica*^{T}, *Desulfotalea psychrophila* LSv54^T und *Desulfobacterium autotrophicum* HRM2^T.

Rhodopirellula baltica^T (Schlesner *et al.*, 2004) gehört zum Phylum der Planktomyceten (Garrity *et al.*, 2002), deren Vertreter trotz weiter Verbreitung, hoher interner Radiation sowie erstaunlicher morphologischer und mitunter auch physiologischer Eigenschaften bislang nur unzureichend untersucht wurden. Die zum Teil beachtlichen Abundanzen von Planktomyceten in den oberen Schichten mariner Sedimente [3-6% der Gesamtpopulation in Wattenmeersedimenten; (Llobet-Brossa *et al.*, 1998)] und in marinem Detritus [bis zu 22% der Gesamtpopulation; (DeLong, 1993)] legen nahe, daß Planktomyceten zumindest in manchen marinen Habitaten eine wichtige Rolle bei der aeroben Mineralisation komplexer organischer Substrate zufällt. Die Wahl von *Rhodopirellula baltica*^{T} zur Totalsequenzierung liegt in der Tatsache begründet, daß diese Spezies in den vergangenen zwanzig Jahren wiederholt aus der Kieler Bucht isoliert wurde und somit an diesem Standort abundant zu sein scheint (Griepenburg *et al.*, 1999). Außerdem waren zu Projektbeginn adäquate Kulturbedingungen bekannt

(Schlesner, 1994) und die Physiologie von *Rhodopirellula baltica*^{T} bereits partiell charakterisiert.

Desulfotalea psychrophila LSv54^T und *Desulfobacterium autotrophicum* HRM2^T gehören phylogenetisch zur Gruppe der sulfatreduzierenden Deltaproteobakterien, welche in anoxischen marinen Sedimenten eine Schlüsselrolle bei der anaeroben Mineralisation niedermolekularer organischer Substrate und somit eine wichtige Rolle als Destruenten im globalen Kohlenstoff-kreislauf spielen (Jørgensen, 1984). *Desulfotalea psychrophila* LSv54^T kann kein Acetat verwerten und gehört somit zur Gruppe der unvollständig oxidierenden Sulfatreduzierer, wohingegen *Desulfobacterium autotrophicum* HRM2^T den vollständigen Oxidierern zuzurechnen ist. Auch sonst unterscheiden sich beide Organismen im Hinblick auf ihre physiologische Flexibilität und optimalen Wachstumstemperaturen. Während *Desulfotalea psychrophila* LSv54^T ein obligat organoheterotrophes, psychrophiles Isolat aus arktischen Sedimenten bei Spitzbergen ist [optimale Wachstumstemp.: 10 °C; (Knoblauch *et al.*, 1999)], vermag das mesophile Bakterium *Desulfobacterium autotrophicum* HRM2^T auch mit H_2/CO_2 chemolithoautotroph zu wachsen [optimale Wachstumstemp.: 25-28 °C; (Brysch *et al.*, 1987)]. Zudem kann *Desulfobacterium autotrophicum* HRM2^T im Gegensatz zu *Desulfotalea psychrophila* LSv54^T neben Pyruvat auch Lactat vergären und organische Fettsäuren oxidieren. Obwohl *Desulfobacterium autotrophicum* HRM2^T aufgrund seines Temperaturoptimums als mesophil einzustufen ist, vermag es auch bei 4°C zu wachsen und ist daher psychrotolerant.

Allen drei ausgewählten Spezies war zu Projektbeginn gemein, daß aus ihren jeweiligen Taxa noch keine vollständig sequenzierten Genome vorlagen - eine Tatsache, die sich bei den Planktomyceten trotz andauernder Sequenzierungen der Genome von *Gemmata obscuriglobus* UQM2246^T (TIGR), *Gemmata* sp. Wa1-1 (Integrated Genomics) und *Kuenenia stuttgartiensis* (Genoscope/TU-München/Universität Nijmegen) bis heute nicht geändert hat. Aus der Gruppe der Deltaproteobakterien hingegen liegen mit *Geobacter sulfurreducens* PCA (Methe *et al.*, 2003) und *Bdellovibrio bacteriovorus* HD100 (Rendulic *et al.*, 2004) inzwischen zwei Genome vor, und nicht weniger als 13 weitere befinden sich zur Zeit in der Sequenzierung, darunter die der Sulfatreduzier *Desulfovibrio vulgaris* subsp. *vulgaris* (TIGR), *Desulfovibrio desulfuricans* G20 (Joint Genome Institute) und *Desulfuromonas acetoxidans* (Joint Genome Institute).

Dem institutsübergreifenden Charakter des REGX-Projekts entsprechend wurden die drei Bakterien durch unterschiedliche Institutionen sequenziert: Das Genom des Planktomyceten *Rhodopirellula baltica*^{T} (7.14 Mb) vom Max-Planck-Institut für molekulare Genetik in Berlin (vgl. Pkt. 1), das Genom von *Desulfotalea psychrophila* LSv54^T (3.52 Mb) von der Epidauros Biotechnologie AG (Bernried) und das Genom von *Desulfobacterium autotrophicum* HRM2^T

(~5.4 Mb) vom Göttingen Genomics Laboratory (G_2L). Letztere dauert an und wird voraussichtlich im Frühjahr 2004 zum Abschluß gebracht werden. An der Technischen Universität München wurde in Rahmen des REGX-Projekts das Programmpaket ARB um Funktionalitäten zur Verknüpfung von bioinformatischen Daten aus Genomsequenzierungsprojekten mit funktionellen Daten aus Proteom- und Transskriptomexperimenten erweitert (ARB-Genome) sowie die Hardware für das Annotationssystem bereitgestellt und gewartet. Dem Max-Planck-Institut für marine Mikrobiologie in Bremen oblag schließlich die Entwicklung einer bioinformatischen Arbeitspipeline (Genvorhersage, Bereitstellung und Pflege des Annotationssystems, Durchführung bioinformatischer Analysen, Softwareentwicklung), der Aufbau eines Proteomlabors mit der Durchführung entsprechender Experimente sowie die Projektkoordination. Die gesamte bioinformatische Analyse und die Annotation des Genoms von *Rhodopirellula baltica*^{T} wurden ebenfalls am Max-Planck-Institut für marine Mikrobiologie in Bremen durchgeführt und bilden die Schwerpunkte der vorliegenden Arbeit.

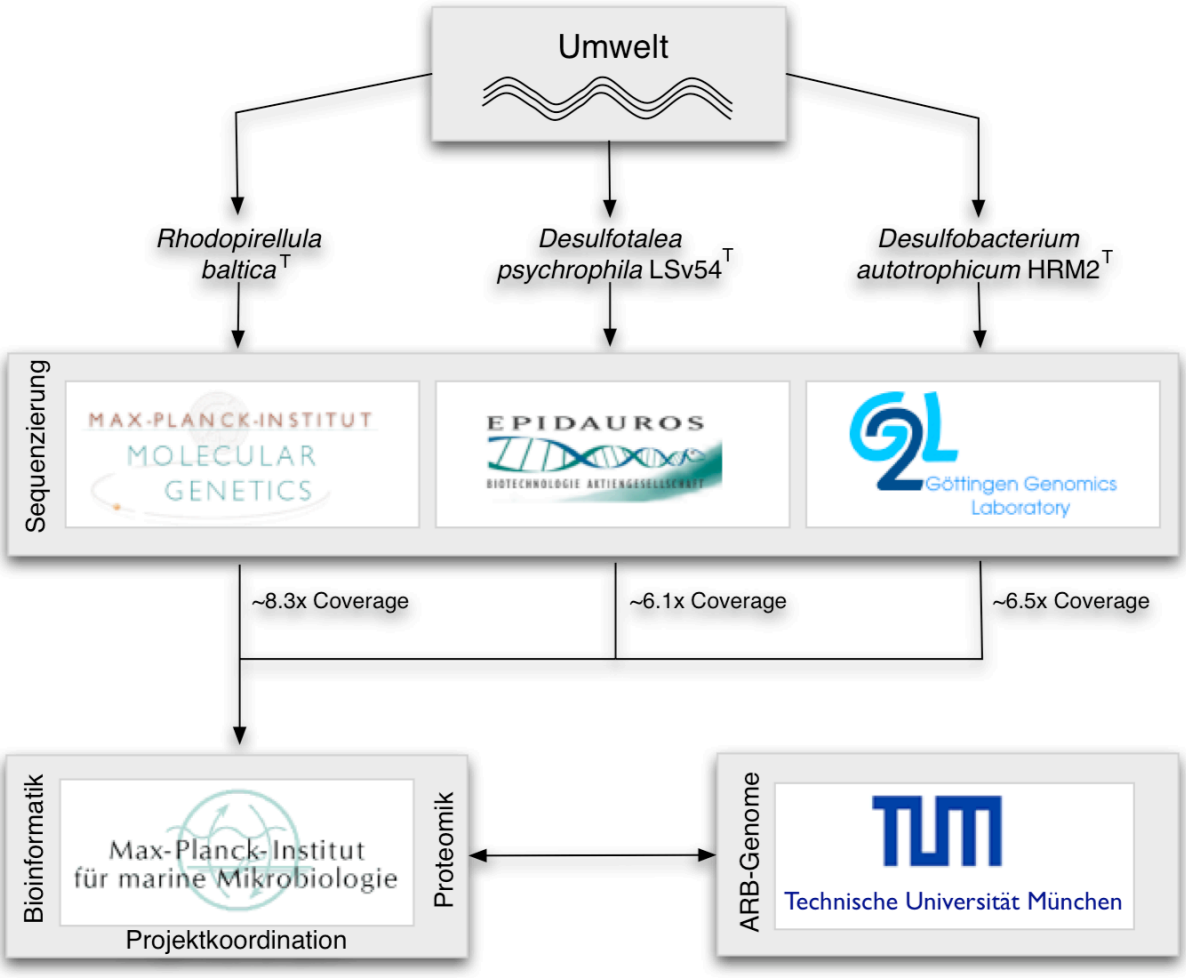


Abb. 1 Organisation des REGX-Projekts während der ersten Projektphase.

2. Eine kurze Einführung in die Gruppe der Planktomyceten

Innerhalb der Domäne *Bacteria* bilden die Planktomyceten das eigenständige Phylum *Planctomycetes* (Garrity *et al.*, 2002), welches derzeit lediglich aus der Ordnung *Planctomycetales* mit der alleinigen Familie *Planctomycetaceae* besteht. Letztere umfaßt die sechs Gattungen *Blastopirellula*, *Isosphaera, Gemmata, Pirellula, Planctomyces und Rhodopirellula* (Schlesner *et al.*, 2004). Allen bislang beschriebenen Planktomyceten sind eine Fülle morphologischer Besonderheiten gemein, die sie klar von den Vertretern anderer Phyla abgrenzen.

So sind die Zellen von Planktomyceten stark polar organisiert. Tochterzellen entstehen an einem dezidierten reproduktiven Zellpol in einer speziellen, inäqualen Zellteilung, welche der Knospung von Sproßhefen (*Saccharomycetaceae*) ähnelt. Bei den Gattungen *Pirellula* und *Rhodopirellula* sind die Tochterzellen monotrich begeißelt und reproduktionsinaktiv. Nach einiger Zeit des Umherschwimmens verlieren diese Schwärmerzellen ihre Geißeln und differenzieren erneut zu adulten, sessilen, reproduktionsaktiven Zellen (Tekniepe *et al.*, 1981). Planktomyceten der Gattung *Planctomyces* zeigen möglicherweise einen vergleichbaren Zellzyklus (Schlesner, 1994). Insgesamt gesehen ähnelt der Zellzyklus der Planktomyceten demjenigen von *Caulobacter crescentus* (Nierman *et al.*, 2001).

Dem reproduktiven Zellpol der Planktomyceten gegenüber liegt der generative Zellpol. Er dient oftmals der Anheftung mittels einer sezernierten Festhaltesubstanz (ähnlich den Byssusfäden bei *Mytilus edulis*) oder einer Prostheka (gr. *prostheka* = Anhängsel). Auf diese Weise können sich Planktomyceten nicht nur an die umgebende Matrix sondern auch aneinander festheften, wodurch charakteristische Rosetten entstehen (Fuerst, 1995).

Die Gattung *Isosphaera*, welche lediglich durch die Art *Isospaera pallida* (Giovannoni *et al.*, 1987) repräsentiert wird, nimmt unter den Planktomyceten insofern eine Sonderstellung ein, als daß *Isospaera pallida* nicht nur moderat thermophil ist (Wachstum bis 55 °C), sondern auch als einziger unter den bislang bekannten Planktomyceten Phototaxis sowie eine gleitende Fortbewegung zeigt.

Die Zellwände der Planktomyceten bestehen nicht aus Peptidoglycan, sondern aus via Disulfidbrücken vernetzen Proteinen (König *et al.*, 1984; Liesack *et al.*, 1986; Stackebrandt *et al.*, 1986). Die Oberflächen dieser Zellwände weisen kleine, kreisförmige Vertiefungen unbekannter Funktion auf, welche als *crateriform structures* bezeichnet werden. Diese befinden sich bei der Gattung *Pirellula* ausschließlich am generativen Pol, während sie bei der Gattung *Planctomyces* über die ganze Zelloberfläche verteilt sind (Liesack *et al.*, 1986).

Das auffälligste morphologische Charakteristikum der Planktomyceten ist jedoch ihre Kompartimentierung. Wie mit Hilfe von Ultradünnschnitten cryosubstituierter Zellen gezeigt werden konnte, besitzen alle bislang bekannten Planktomyceten vollständig von Lipid-Bilayern umschlossene, intracytoplasmatische Kompartimente, welche nicht mit der äußeren Cytoplasmamembran in Verbindung stehen (Lindsay *et al.*, 2001). Allen Gattungen ist ein großes Kompartiment eigen, dessen Plasma die Ribosomen enthält (Riboplasma), wohingegen das umgebende Plasma ribosomenfrei zu sein scheint (Paryphoplasma). Bei den Gattungen *Blastopirellula*, *Pirellula* und *Rhodopirellula* ist dies das einzige Kompartiment und wird als Pirellulosom bezeichnet (Lindsay *et al.*, 1997). Auch die Gattungen *Isosphera* und *Planctomyces* verfügen nur über dieses eine Kompartiment, welches sich jedoch durch eine charakteristische Invagination vom Pirellulosom unterscheidet. Im Bereich dieser Invagination scheinen Ribosomen - einem eukaryontischen, rauhen endoplasmatischem Retikulum (ER) gleich - mit der intracytoplasmatischen Membran assoziiert zu sein (Lindsay *et al.*, 2001). Bei der Gattung *Gemmata* (Franzmann & Skerman, 1984) ist das Chromosom von zwei zusätzlichen Membranen umgeben - eine Situation, die stark an den eukaryontischen Nukleus erinnert, weshalb man in Analogie von einem Nukleoid spricht (Fuerst & Webb, 1991).

Auch die Kandidatus-Gattungen *Brocadia* (Strous *et al.*, 1999) und *Kuenenia* (Schmid *et al.*, 2000) sind durch ein weiteres, membranumgebenes Kompartiment charakterisiert, das sog. Anammoxosom. Das Anammoxosom dient dem Anammox-Prozeß (http://www.anammox.com), einer Oxidation von Ammonium mit Nitrit zu elementarem Stickstoff. Die Membran des Anammoxosoms enthält ungewöhnliche Lipide aus konkatenierten Cyclobutanringen, sog. Ladderane. Diese machen die Membran nahezu impermeabel für passive Diffusion, wodurch toxisches Hydrazin, welches beim Anammox-Prozeß auf der Membraninnenseite entsteht, vom übrigen Stoffwechsel separiert wird. Aufgrund seiner potentiellen industriellen Applikation zur biologischen Abwasserklärung wird der Anammox-Prozeß derzeit intensiv beforscht und das Genom des Planktomyceten *Kuenenia stuttgartiensis* totalsequenziert.

Insgesamt betrachtet läuft die Kompartimentierung der Planktomyceten nicht nur der klassischen Definition einer prokaryontischen Zelle zuwider, sondern wirft auch Fragen nach einer damit verbundenen physiologischen Funktion und vor allem nach ihrer Entstehung auf. So deutet die Kompartimentierung der Planktomyceten darauf hin, daß der Vorgang der Endosymbiose, welcher zur Entstehung der Eukaryonten führte, für die Entstehung von kompartimentierten Zellen nicht zwingend notwendig ist.

Die ökologische Bedeutung der Planktomyceten wurde lange Zeit als gering eingeschätzt, da man glaubte, sie seien auf limnische Systeme beschränkte, durchweg obligat aerobe, organo-

heterotrophe Organismen. So stammte die 1924 erstmals nachweislich lichtmikroskopisch beobachtete Spezies aus einem Teich bei Budapest und wurde fälschlicherweise als planktischer Pilz klassifiziert (Gimesi, 1924), worauf im übrigen die Bezeichnung "Planktomyceten" (gr. *planktos* = Umherirrendes, Umhertreibendes; gr. *mykes* = Pilz) für die gesamte Organismengruppe zurückzuführen ist. Mit den Jahren zeigte sich jedoch, daß das Vorkommen von Planktomyceten nicht auf limnische Systeme beschränkt ist, sondern auch marine und terrestrische Habitate unterschiedlicher Salinität, Temperatur, Trophiestufe und Sauerstoff-Verfügbarkeit umfaßt. So wurden Planktomyceten in der Wassersäule und in anoxischen Sedimenten von Seen (Miskin *et al.*, 1999; Wang *et al.*, 2002), in der Wassersäule und Sedimenten mariner Habitate sowie in marinem Detritus (DeLong, 1993; Schlesner, 1994; Vergin *et al.*, 1998), Kläranlagen (Griepenburg *et al.*, 1999; Neef *et al.*, 1998), moderat heißen Quellen (Giovannoni *et al.*, 1987), oxischen und anoxischen Böden (Derakshani *et al.*, 2001; Liesack & Stackebrandt, 1992; Wang *et al.*, 2002) und sogar dem Hepatopancreas der Crustaceae *Penaeus monodon* (Fuerst *et al.*, 1991; Fuerst *et al.*, 1997) nachgewiesen. Jüngere Studien haben gezeigt, daß der Anammox-Prozeß im Bereich mariner Kontinentalschelfe mit bis zu 65% an der Entstehung von elementarem Stickstoff beteiligt ist und damit den Prozeß der Denitrifikation quantitativ überflügelt (Dalsgaard & Thamdrup, 2002; Thamdrup & Dalsgaard, 2002). Die inzwischen augenscheinliche, nahezu ubiquitäre Verbreitung von Planktomyceten hat dazu geführt, daß dieser Organismengruppe zumindest in einigen Habitaten eine wichtige ökologische Rolle zugesprochen wird.

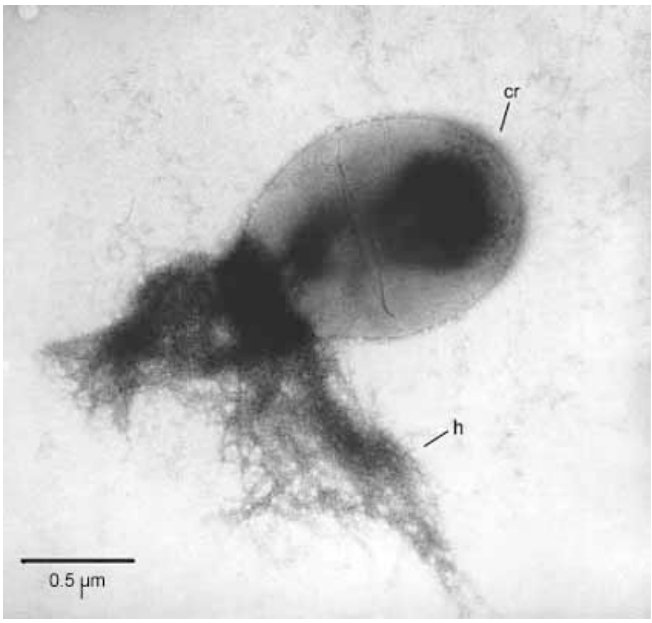


Abb. 2 *Rhodopirellula baltica*^{T}-Zelle mit Festhaltesubstanz (h) am vegetativen und *crateriform structures* (cr) am generativen Pol (Schlesner *et al.*, 2004). Die birnenförmige Morphologie ist typisch für die Gattungen *Pirellula, Blastopirellula* und *Rhodopirellula,* und haben ihnen ihren Namen eingetragen (lat. *pirum* = Birne).

3. Genomsequenzierungen und Bioinformatik

Die Technik der Genomsequenzierung hat seit 1995, als mit dem Genom von *Haemophilus influenzae* erstmals die Totalsequenzierung eines Genoms erfolgreich abgeschlossen wurde, kontinuierlich Fortschritte gemacht, und zwar sowohl im Hinblick auf die Geschwindigkeit und Qualität als auch auf die Kosten (Fraser *et al.*, 2002). Infolgedessen ist die Rate, mit der neue mikrobielle Genome sequenziert werden, seit Beginn der Genomsequenzierungen stetig gestiegen, und ein Ende dieses Trends ist auch für die kommenden Jahre nicht in Sicht. Mit Ende des Jahres 2003 waren beim NCBI 148 mikrobielle Genome als vollständig und weitere 123 als teilweise sequenziert aufgeführt. Die tatsächliche Anzahl der total- oder zumindest teilweise sequenzierten Genome dürfte jedoch ungleich höher sein, da viele der von der biotechnologischen Industrie sequenzierten Genome nicht oder erst nach Ablauf längerer Fristen öffentlich zugänglich gemacht werden. In der Datenbank der kommerziellen ERGO-Suite (Integrated Genomics Inc., Chicago) waren z.B. mit Ablauf des Jahres 2003 nicht weniger als 664 ganz oder teilweise sequenzierte Genome verzeichnet.

Neben der überwiegenden Zahl an Mikroorganismen mit medizinisch-pharmazeutischer, agrar-ökonomischer oder biotechnologischer Bedeutung finden sich darunter auch zunehmend solche von ökologischer Relevanz. Beispiele aus den vergangenen zwei Jahren sind neben dem Genom von *Rhodopirellula baltica*^{T} (Glöckner *et al.*, 2003) die Genome von *Oceanobacillus iheyensis* HTE831 (Takami *et al.*, 2002), *Geobacter sulfurreducens* PCA (Methe *et al.*, 2003), *Nitrosomonas europaea* (Chain *et al.*, 2003) und die der vier Cyanobakterien *Gloeobacter violaceus* PCC 7421 (Nakamura *et al.*, 2003), *Prochlorococcus marinus* subsp. *marinus* str. CCMP1375 (Dufresne *et al.*, 2003), *Prochlorococcus marinus* str. MIT 9313 (Rocap *et al.*, 2003) und *Prochlorococcus marinus* subsp. *pastoris* str. CCMP1986 (Rocap *et al.*, 2003).

Die rasante Entwicklung im Bereich der Genomsequenzierung ist jedoch nicht allein auf technische Fortschritte in der Sequenzierungstechnik zurückzuführen, sondern wurde auch maßgeblich durch Fortschritte in der Bioinformatik ermöglicht. Die 1965 vom Intel-Entwickler Gordon Moore geäußerte Prognose, daß sich die Zahl der Transistoren in Mikrochips jedes Jahr verdoppeln würde (*Moore's law*) hat bis heute nichts von ihrer Aktualität eingebüßt, und so erfreut sich die Wissenschaft seit Jahren an der exponentiellen Zunahme der zur Verfügung stehenden Rechenleistung. Brauchte es zur Assemblierung und bioinformatischen Analyse eines mikrobiellen Genoms noch vor wenigen Jahren teurer Workstations und Cluster aus dem *midrange*-Sektor, so lassen sich diese Aufgaben heute mit Hilfe weniger Desktop-Systeme bewältigen. Auf der anderen Seite hat die zunehmende Zahl totalsequenzierter Genome der

Bioinformatik mit der vergleichenden Genomik ein aufregendes, aber auch außerordentlich rechenintensives neues Forschungsgebiet beschert, welches ohne verteiltes Rechnen in leistungsfähigen Clustern derzeit nicht zu bewältigen ist. Als weiteres zukünftiges Arbeitsgebiet der Bioinformatik zeichnet sich die Integration und Auswertung biologischer Daten ab, wie z.B. die Verknüpfung von *in silico*-Analysen von Genomen oder Genom-Fragmenten aus der Metagenomik mit physikochemischen Standortparametern sowie Daten aus Proteom- und Transkriptomanalysen. Dies stellt nicht nur hohe Anforderungen an die Hardware, sondern auch hohe logistische Anforderungen an die verwendeten Datenbankarchitekturen und Auswertestrategien. Insgesamt gesehen hat sich die Bioinformatik in den letzten Jahren als eigenständiger Zweig der Biologie etabliert. Die Übergänge zu anderen Disziplinen sind dabei fließend und reichen von der Mathematik (Algorithmenentwicklung) und Informatik (Installation und Wartung von Hard- und Software, sowie Softwareentwicklung und -implementation) auf der einen Seite bis hin zu klassisch biologischen Feldern wie der Molekularbiologie oder Physiologie bei der Interpretation der bioinformatisch generierten Daten.

Mit der Ausweitung des Tätigkeitsfeldes von der Sequenzierung einzelner Genome hin zu Softwareentwicklungen und integrativen Ansätzen wie der vergleichenden Genomik sowie der Integration von Genom-, Metagenom- Proteom-, Transkriptom- und Umweltdaten folgt das REGX-Projekt aktuellen Entwicklungen in der Bioinformatik.

4. Zielsetzung dieser Arbeit

Im Vergleich zu vielen anderen bakteriellen Phyla sind die *Planctomycetes* bislang nur wenig untersucht worden. Zwar werden einige zum Anammox-Prozeß befähigte Planktomyceten gezielt in Richtung auf eine biotechnologische Anwendung hin erforscht, doch die Rolle von Planktomyceten an ihren natürlichen Standorten ist nach wie vor weitgehend ungeklärt. Auch die Herkunft und Bedeutung der ungewöhnlichen Zellmorphologie der Planktomyceten wurde bis dato kaum untersucht, und die Phylogenie der Planktomyceten war immer schon kontrovers und gerade in jüngster Zeit Gegenstand heftig ausgetragener Debatten (Brochier & Philippe, 2002; Di Giulio, 2003).

Ziel der Annotation des Genoms von *Rhodopirellula baltica*^{T} war es daher, mit Hilfe der Informationsfülle des Gesamtgenoms einen Beitrag zur Beantwortung der mannigfaltigen ungeklärten Fragen über Planktomyceten zu leisten. Neben Fragen zu Planktomyceten im allgemeinen, wie z.B. ihrer Phylogenie oder der Herkunft und Bedeutung ihrer ungewöhnlichen

Zellmorphologie, standen dabei insbesondere Fragen nach der Biologie und ökologischen Rolle von *Rhodopirellula baltica*^{T} als Stellvertreter eines marinen Planktomyceten im Vordergrund.

Da das Genom von *Rhodopirellula baltica*^{T} das erste totalsequenzierte Genom des gesamten Phylums ist, steht zu erwarten, daß sich die Annnotationen nachfolgender Planktomyceten-Genome an derjenigen von *Rhodopirellula* baltica^T orientieren werden (die derzeit in der Assemblierung befindlichen Planktomyceten-Genome sind unter Pkt. 1. angeführt). Aus diesem Grunde war eine qualitative hochwertige Annotation des Genoms unerläßlich, weshalb mit etwa drei Mannjahren ein ungewöhnlich hohes Maß an Zeit in die manuelle Annotation investiert wurde.

Aspekte dieser Annotation bilden zentrale Themen der vorliegenden Arbeit, darunter Untersuchungen zu sequenzimmanenten Eigenschaften und Genomorganisation, zur Physiologie, zu Motiliät und Taxis, zu Reaktionen auf Stressoren, zur Zellbiologie und -kompartimentierung, zur Replikation, zum Zellzyklus und zur Ökologie (Publikation 1). Einen weiteren Schwerpunkt stellten Untersuchungen zur Phylogenie der Planktomyceten dar (Publikation 2).

Um auf dem wichtigen Forschungsfeld der Genomik kompetitiv zu bleiben, war es zudem erklärtes Ziel, auf Basis der Erfahrungen der Analyse des *Rhodopirellula baltica*^{T}-Genoms eine Arbeitspipeline zur Prozessierung von Genomdaten aus weiteren Genomsequenzierungs- und Metagenomprojekten zu etablieren. Dazu wurden mit MORFind eine Software für eine qualitativ hochwertige Vorhersage proteinkodierender Gene entwickelt (Publikation 4) und ein nahtloser Übergang vom zuvor verwendeten Annotationssystem PEDANT-Pro (Frishman *et al.*, 2001) auf das leistungsfähigere Open-Source-Annotationssystem GenDB (Meyer *et al.*, 2003) sichergestellt. Außerdem wurde eine Identifikationshilfe für Metagenomfragmente entwickelt (Publikation 3) und mit TETRA in Form einer Anwendersoftware implementiert.

B Ergebnisse und Diskussion

Im folgenden Abschnitt werden Aspekte der im zweiten Teil dieser Arbeit aufgeführten Publikationen zusammengefaßt und übergreifend diskutiert. Er ergänzt die in den Publikationen dargelegten, detaillierten Diskussionen.

Zudem werden in diesem Abschnitt Ergebnisse dargestellt, die nicht oder nur teilweise Eingang in die Publikationen gefunden haben. Aufgrund der Fülle der Informationen aus der fast dreijährigen Analyse des *Rhodopirellula baltica*^{T}-Genoms konnten dabei nicht alle Ergebnisse berücksichtigt und viele Ergebnisse lediglich summarisch dargestellt werden.

1. Die Sequenzierung des Genoms von *Rhodopirellula baltica*^{T}

Das Genom von *Rhodopirellula baltica*^{T} wurde nach dem *Shotgun*-Verfahren sequenziert und ist mit 7.145.576 bp eines der größten bislang bekannten bakteriellen Genome. Mit einer durchschnittlichen *Coverage* von 8.3 und einem maximalen Sequenzierfehler von 1/10.000 bp ist die Qualität der Genomsequenz im Vergleich zu vielen anderen Genomsequenzierungen sehr hoch. Die Details der Sequenzierung sind in der Promotionsschrift von Dr. Michael Kube zusammengefaßt, die im Rahmen des REGX-Projekts am Max-Planck-Institut für molekulare Genetik in Berlin angefertigt, an der Universität Bremen eingereicht und im Juni 2003 am Max-Planck-Institut für marine Mikrobiologie verteidigt wurde.

2. Die Vorhersage proteinkodierender Gene

In den vergangenen Jahren wurden zahlreiche Progamme zur *in silico* Vorhersage proteinkodierender ORFs in Prokaryonten entwickelt, darunter Glimmer (Delcher *et al.*, 1999; Salzberg *et al.*, 1998), ORPHEUS (Frishman *et al.*, 1998), CRITICA (Badger & Olsen, 1999), GeneScan (Ramakrishna & Srinivasan, 1999), der 'frame-by-frame' Algorithmus (Shmatkov *et al.*, 1999), GeneMarkS (Besemer *et al.*, 2001), EasyGene (Larsen & Krogh, 2003) und jüngst ZCURVE (Guo *et al.*, 2003) und YACOP (Tech & Merkl, 2003). Unter diesen Programmen (ORF-Finder[1]) lassen sich solche unterscheiden, bei denen das Vorhersage-Modell *ab initio* berechnet wird, und solche, deren Vorhersage-Modell unter Einbeziehung von Sequenzähnlichkeiten zu bekannten

[1] In Anlehnung an den angloamerikanischen Sprachgebrauch („ORF-*finder*") wird nachfolgend der Begriff „ORF-Finder" verwendet, obwohl die entsprechenden Programme *de facto* nicht ORFs als solche, sondern *proteinkodierende* ORFs vorhersagen.

Proteinen berechnet wird. Letzteres geschieht in der Regel über Vergleiche zu Sequenzdatenbanken mittels heuristischer Algorithmen, wie z.B. BLAST (Altschul *et al.*, 1990) oder DPS (Huang, 1996). Die eigentliche Vorhersage basiert bei fast allen genannten ORF-Findern auf Markov-Modellen unterschiedlichster Couleur, neuronalen oder Bayes'schen Netzen, Fourier-Transformationen oder auf Kombinationen dieser Methoden.

Im Falle des Genoms von *Rhodopirellula baltica*^{T} wurden drei ORF-Finder eingesetzt: Glimmer, ORPHEUS und CRITICA. Vom Prinzip her ist Glimmer ein *ab initio* ORF-Finder, der auf einem sog. interpolierendem Kontextmodell (ICM) basiert. ORPHEUS ist ein vergleichender ORF-Finder, der auf genomspezifischen Charakteristika in der Basenzusammensetzung kodierender Sequenzen beruht, und CRITICA benützt ein kombiniertes Modell aus einer *ab initio* (*dicodon bias*) und einer vergleichenden Methode. Trotz ihrer unterschiedlichen Algorithmen war selbst nach Ausschöpfung möglicher Programmparameter-Optimierungen keiner der drei ORF-Finder zu einer schlüssigen Genvorhersage für das *Rhodopirellula baltica*^{T}-Genom imstande (eine ausführliche Diskussion findet sich im zweiten Teil dieser Arbeit in der Publikation Nr. 4). Während CRITICA mit lediglich 4531 deutlich zu wenig proteinkodierenden Gene für das 7.1 Mb große Genom von *Rhodopirellula baltica*^{T} vorhersagte, kam es bei ORPHEUS mit über 10.200 Genen zu einer drastischen Übervorhersage. Auch die Glimmer-Vorhersage erwies sich als problematisch und konnte nur unter Verwendung strikter Grenzen für mögliche Gen-Überlappungen von über 9.200 (Standardparameter) auf 7.579 gesenkt werden (Publikation 4, Tab. 1).

Die Tatsache, daß alle drei Programme und damit die ihnen zugrunde liegenden Algorithmen Schwierigkeiten mit der Gen-Vorhersage für das *Rhodopirellula baltica*^{T}-Genom hatten, ist wahrscheinlich multifaktoriell bedingt. So sind nahezu alle bis dato publizierten ORF-Finder für prokaryontische Genome zu wesentlichen Teilen an gut untersuchten Modellorganismen wie den totalsequenzierten *Escherichia coli*-Stämmen oder *Bacillus subtilis* subsp. *subtilis* 168 entwickelt, validiert und somit auf diese hin optimiert worden. Außerdem scheinen die Unterschiede zwischen kodierenden und nicht-kodierenden Sequenzen bei *Rhodopirellula baltica*^{T} im Vergleich zu vielen anderen bakteriellen Genomen nur schwach ausgebildet zu sein. Darauf deutet z.B. die Tatsache hin, daß der iterative Startvorhersage-Algorithmus von ORPHEUS an der Vorhersage der Consensus-Sequenz für die ribosomale Bindestelle (RBS) des *Rhodopirellula baltica*^{T}-Genom scheiterte (CTTCAC anstelle des dem 3'-Ende der ribosomalen RNA komplementären AAGGAG). Dies hatte ein falsches RBS-Modell und damit eine mehr oder weniger statistisch gestreute Startvorhersage zur Folge. Hinzu kommt, daß lediglich 2384 (32%) der schlußendlich annotierten 7325 Gene im *Rhodopirellula baltica*^{T}-Genom eine

Funktion zugeordnet werden konnte, was etwa um 20% unter der Annotationsquote für andere bakterielle Genome liegt (Fraser *et al.*, 2000). Dies spiegelt nicht nur die phylogenetische Alleinstellung der Planktomyceten und das derzeit noch geringe Wissen über diese Organismengruppe wider, sondern dürfte auch entscheidenden Einfluß auf die Güte des Trainingssets von CRITICA und ORPHEUS und damit auf die Güte ihrer Vorhersagen gehabt haben.

Eingedenk der beschriebenen Probleme wurden zur Vorhersage der proteinkodierenden Gene in *Rhodopirellula baltica*^{T} die Einzelvorhersagen von Glimmer, CRITICA und ORPHEUS kombiniert (Abb. 3). Dazu wurden redundant vorhergesagte ORFs vereinigt, wodurch sich die Zahl der zu annotierenden ORFs von ursprünglich 22.388 auf 13.331 reduzieren ließ. Diese wurden manuell annotiert, wodurch die Zahl der potentiellen Gene auf schlußendlich 7325 eingeengt werden konnte. Mit dieser Strategie wurde das Ziel verfolgt, so wenig wie möglich

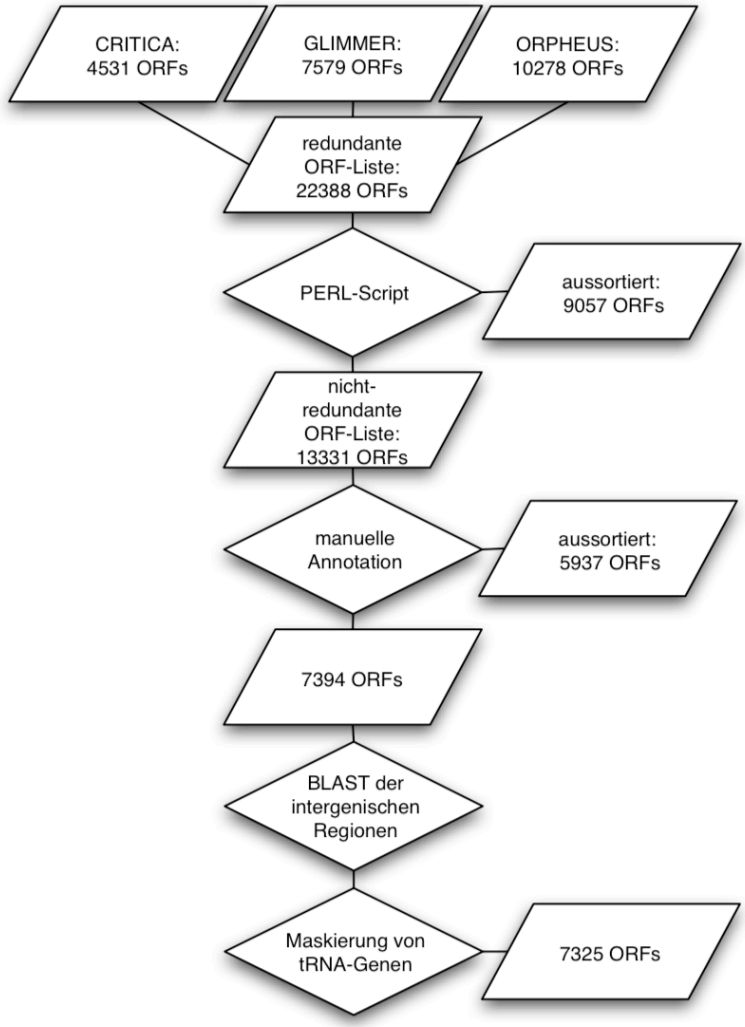


Abb. 3 Flußdiagramm der kombinierten Vorhersage proteinkodierender Gene für das Genom von *Rhodopirellula baltica*^{T}. Von zentraler Bedeutung ist das PERL-Skript, das - stark vereinfacht dargestellt - ORFs zusammenfaßt, die den gleichen Start haben und sich nur geringfügig in der Länge unterscheiden (maximal 10%, bezogen auf das jeweils längere ORF).

Gene zu übersehen. Die dadurch entstehende verstärkte Überannotation durch fälschlich vorhergesagte kurze Gene wurde dabei bewußt in Kauf genommen. Die meisten ORF-Finder weisen nämlich eine mehr oder weniger starke Übervorhersage von ORFs mit Längen kleiner als als 90 Codons auf, da der Informationsgehalt dieser ORFs zu gering ist, als daß bioinformatische Modelle kodierende und nicht-kodierende ORFs perfekt diskriminieren könnten. In dieser Hinsicht unterscheidet sich die Annotation des Genoms von *Rhodopirellula baltica*^{T} nicht von der überwiegenden Mehrzahl der publizierten prokaryontischen Genome (Skovgaard *et al.*, 2001). Auch die Tatsache, daß *Rhodopirellula baltica*^{T} mit 95% eine sehr hohe *gene coverage* aufweist und lediglich 5% (362.530 bp) der Genomsequenz auf intergenische Regionen entfallen, ist vor diesem Hintergrund zu sehen.

Da im Rahmen des REGX-Projekts auch in Zukunft Genome mariner Prokaryonten und Genomfragmente aus Metagenom-Bibliotheken mariner, mikrobieller Lebensgemeinschaften sequenziert und annotiert werden sollen (geplant sind z.B. die Annotation des Genoms von *Desulfobacterium autotrophicum* HRM2^T und die Totalsequenzierung des Genoms von *Cytophaga sp.* KT0803), war eine Verbesserung und weitgehende Automatisierung des ORF-Vorhersageprozesses unabdingbar. Zu diesem Zweck wurde "MORFind" programmiert (Publikation 4). MORFind stellt eine graphische Nutzeroberfläche für CRITICA, Glimmer und ORPHEUS zur Verfügung und erlaubt die Prozessierung der Einzelvorhersagen zu einer nicht-redundanten Liste. Dazu wurde ein aufwendiger, rekursiver Algorithmus entwickelt, der BLASTp-Treffer gegen eine Datenbank der proteinkodierenden Gene aller totalsequenzierten prokaryontischen Genome, Vorhersagen von Signalpeptiden (Nielson *et al.*, 1997) und Transmembranregionen (Sonnhammer *et al.*, 1998) sowie die Anzahl der vorhersagenden Programme verwendet, um zwischen einander widersprechenden ORFs zu entscheiden. Zusätzlich wird die Qualität der Startvorhersage mittels RBSfinder (Suzek *et al.*, 2001) verbessert (Publikation 4, Abb. 1). Gegenüber den einzelnen ORF-Findern führt die Postprozessierung durch MORFind zu einem im Großen und Ganzen verbesserten Verhältnis von Vorhersagesensitivität und -selektivität und minimiert das Problem der Übervorhersage (Publikation 4). MORFind wurde als Server/Client-basierte multiuser-Lösung implementiert, d.h. die Software läuft auf einem zentralen Linux-Server und kann von mehreren Benutzern zeitgleich über ein Netzwerk mit Hilfe dedizierter graphischer Clients angesprochen werden. Eine Anbindung an den graphischen Genom-Viewer Artemis (Rutherford *et al.*, 2000) und an das Annotationssystem GenDB (Meyer *et al.*, 2003) wurde über einen EMBL-Exportfilter realisiert, so daß mit der Kombination MORFind/Artemis/GenDB eine leistungsfähige Pipeline zur bioinformatischen Analyse von Genomsequenzdaten zur Verfügung steht.

3. Sequenzbasierte Eigenschaften des *Rhodopirellula baltica*^{T}-Genoms

Für Genome können eine ganze Reihen von sequenzbasierten Eigenschaften berechnet werden. Dazu gehört die Basenzusammensetzung (G+C-Gehalt, G+T-Gehalt etc.), Asymmetrien in der Basenzusammensetzung zwischen dem kontinuierlich und dem diskontinuierlich replizierten Strang (*GC skew* und *AT skew* resp. GC- und AT-Schiefe), Basenstapelkräfte (Ornstein & Rein, 1979), DNA-Torsion [*twist*; (Kabsch *et al.*, 1982)], sowie die lokale und globale DNA-Krümmung [*bending*, *curvature*; (Gabrielian & Bolshoy, 1999; Goodsell & Dickerson, 1994)].

Alle bislang bekannten prokaryontischen Genome weisen eine stark asymmetrische Basenzusammensetzung zwischen dem kontinuierlich und dem diskontinuierlich replizierten Strang auf, bei der die Anteile an Guanin und Thymin im kontinuierlich replizierten Strang gegenüber ihren komplementären, aminogruppentragenden Basen erhöht sind (G>C; T>A). Für die Ursache dieser Asymmetrie sind unterschiedliche Theorien vorgeschlagen worden (Frank & Lobry, 1999), wobei eine Assoziation mit der asymmetrisch gebauten, replikativen DNA-Polymerase III als am wahrscheinlichsten gilt. Bei einer kumulativen Auftragung der GC- und AT-Schiefe liegen der Replikationsursprung und -terminus an den Extrempunkten der entsprechenden Kurven (Frank & Lobry, 2000; Lobry, 1996). Ein Plot der GC- und AT-Schiefe für das *Rhodopirellula baltica*^{T}-Genom (Abb. 4) zeigt einen sehr ebenmäßigen Verlauf zwischen dem Replikationsterminus bei ~1.8 Mb und dem Replikationsursprung bei 5.45 Mb (5.2.1). Im zweiten Replichore hingegen kommt es zu einer lokalen Umkehr der GC- und AT-Schiefe im Bereich zwischen 87.500 bp und 431.000 bp. Dieser Bereich wird auf der einen Seite von fünf und auf der anderen Seite von vier Transposasen flankiert, von denen die äußeren ein identisches Paar bilden. Außerdem beherbergt dieser weniger als 5% des Gesamtgenoms umfassende Bereich 35% der im Genom vorkommenden Transposasen und IS-Elementen (27 von 81). Dies deutet darauf hin, daß die Unregelmäßigkeit auf eine Inversion des Bereichs zurückzuführen ist. Die Aufnahme von Fremd-DNA kann als Ursache nahezu ausgeschlossen werden, da sich der CAI der Gene in diesem Bereich nicht signifikant vom CAI der Gene im übrigen Genom unterscheidet (Abb. 4). Außerdem liegen eine Fülle von *housekeeping*-Genen in diesem Bereich, z.B. die Gene für die Histidyl-, Valyl-, und Seryl-tRNA Synthetasen, die Gene der ribosomalen Proteine *rps4* und *rps21*, *dnaE*, Glucose-6-phosphatisomerase, Gluconolactonase und die Flagellengene *fliP* und *flhB*. Von diesem größeren genomischen Rearrangement abgesehen, zeigt die kumulative Auftragung der GC- und AT-Schiefen jedoch eine fast perfekte Zickzack-Charakteristik. Dies deutet darauf hin, daß es im *Rhodopirellula baltica*^{T}-Genom in jüngerer Zeit (~1 Million Jahre) keine weiteren größeren Rearrangements gegeben hat. Das Genom von

Synechocystis sp. PCC6803 z.B. weist keine ausgeprägte GC-Schiefe auf, wofür die Fülle aktiver Transposasen im Genom verantwortlich gemacht wird (Kaneko *et al.*, 1996).

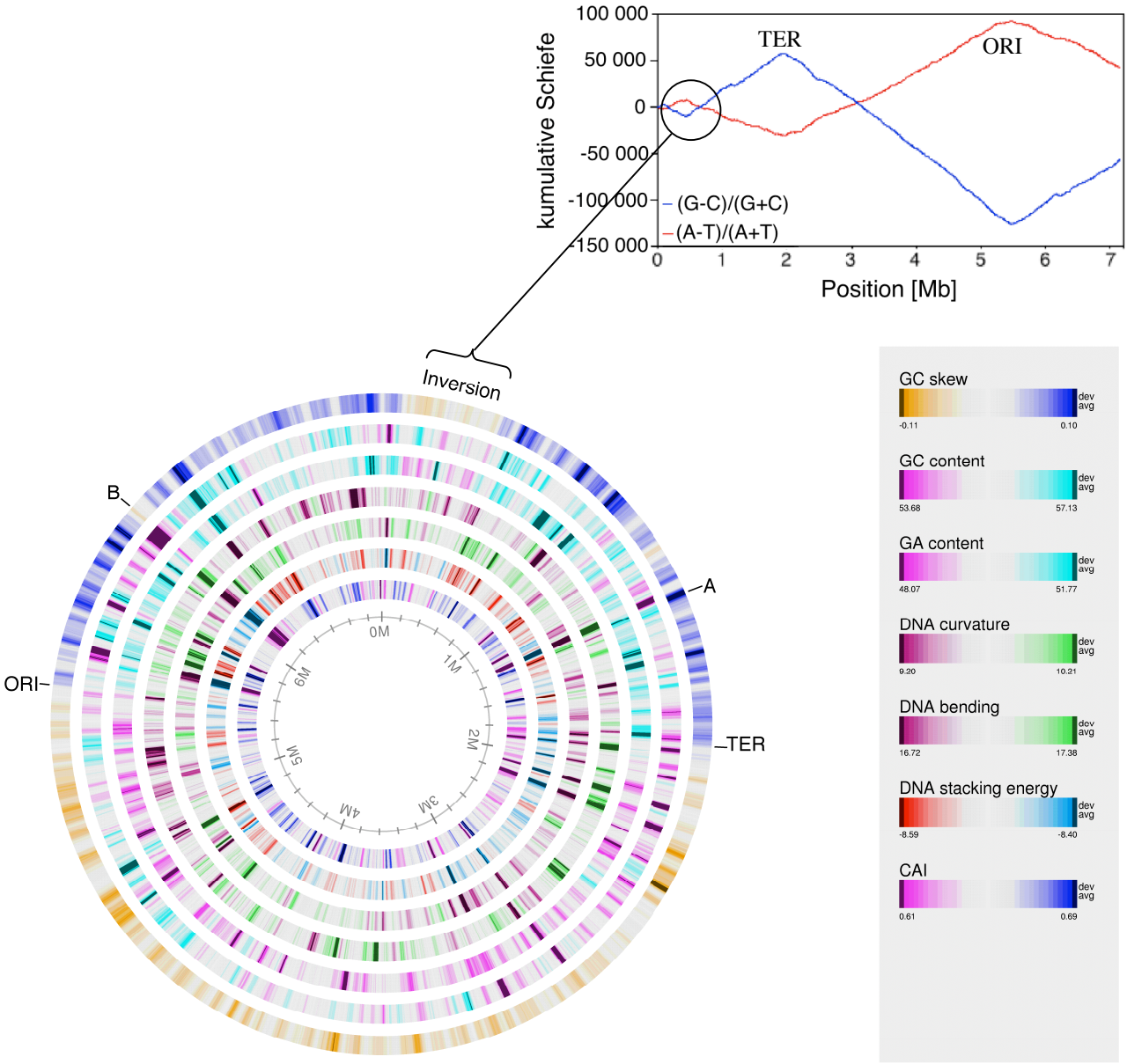


Abb. 4 Sequenzbasierte Eigenschaften des *Rhodopirellula baltica*^{T}-Genoms in hoher Auflösung. Von außen nach innen: GC-Schiefe, G+C-Gehalt [%], G+A-Gehalt [%], DNA *curvature* [°], DNA *bending* [°], Basenstapelkräfte [kcal mol^{-1} dincul.$^{-1}$]. Im innersten Kreis ist zusätzlich der CAI (*codon adaptation index*) der annotierten 7325 Gene dargestellt. GC-Schiefe, G+C-Gehalt und G+A-Gehalt wurden mit selbstgeschriebenen PERL-Skripten berechnet. DNA *curvature* und *bending* wurden mit dem Programm BANANA, Basenstapelkräfte mit dem Programm BTWISTED und der CAI (Sharp & Li, 1987) mit den Programmen CUSP und CAI aus dem EMBOSS-Paket berechnet (Olson, 2002; Rice *et al.*, 2000). Die Visualisierung erfolgte mit dem Programm GENEWIZ unter Verwendung einer Fensterbreite von 25.000 bp (Jensen *et al.*, 1999). Eingezeichnet sind eine vermutete Inversion, der Replikationsursprung (ORI) und -terminus (TER), sowie zwei Bereiche, in denen das Genom größere Unregelmäßigkeiten aufweist. Im ersten Bereich (A) liegt ein Cluster mit 41 sehr kurzen Genen, von denen 39 hypothetisch sind. Im zweiten Bereich (B) liegt ein Cluster aus 27 tRNA-Genen sowie ein sehr langes hypothetisches Gen (24.510 bp).

Der durchschnittliche G+C-Gehalt des *Rhodopirellula baltica*^{T}-Genoms liegt bei 55,4% und zeigt nur geringe Schwankungen (Abb 4, Abb 5). Dahingegen zeigen der GT-Gehalt, die DNA-Torsion, die Basenstapelkräfte sowie die DNA-Flexibilitätsmaße (*curvature* und *bending*) deutlich stärkere Schwankungen im Replichore mit der Inversion als im Replichore ohne Inversion. Die Inversion ist jedoch zu klein, um diese Schwankungen vollständig erklären zu können. Die etwaige biologische Bedeutung dieser Beobachtung ist zu diesem Zeitpunkt noch unklar.

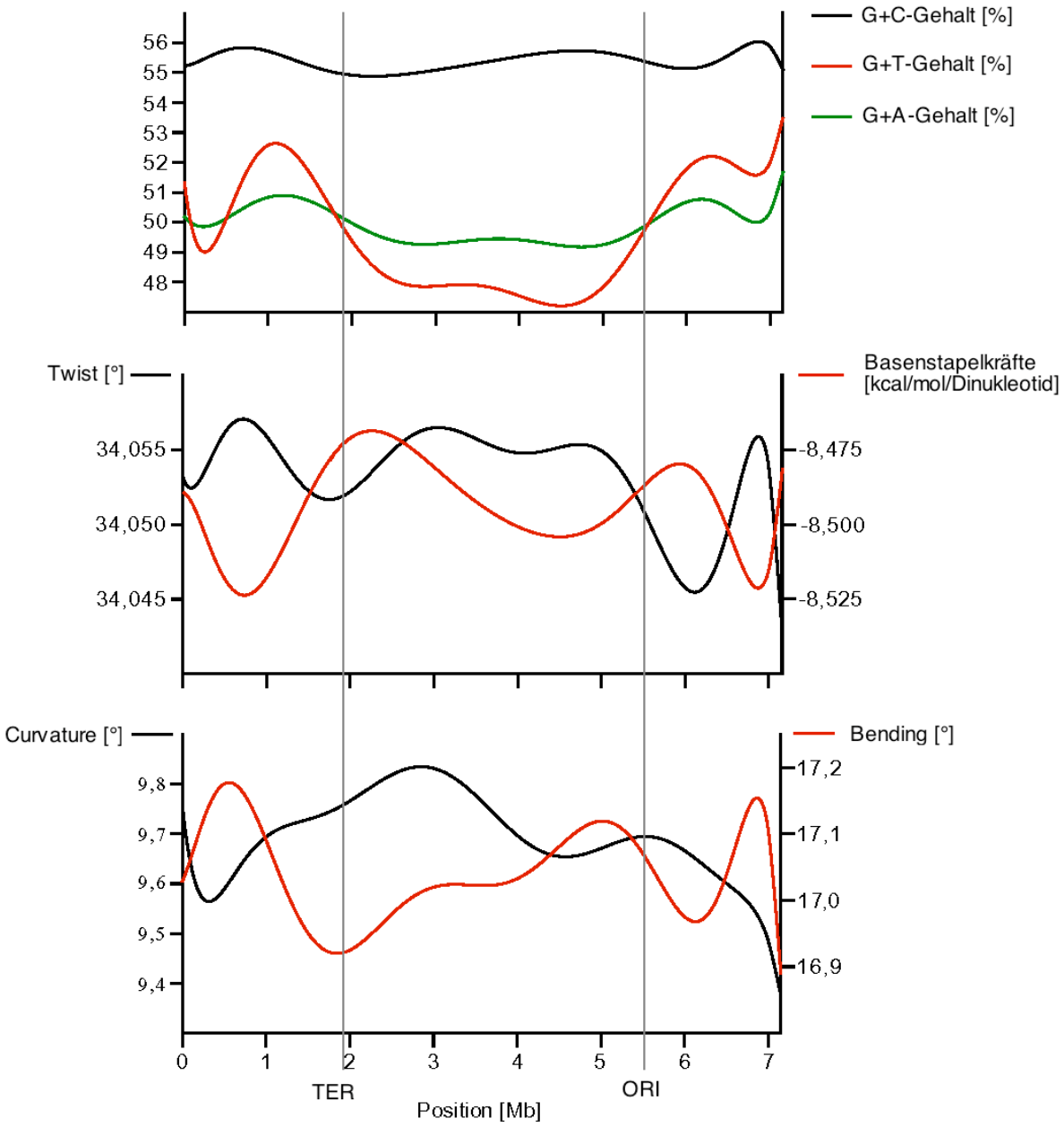


Abb. 5 Sequenzbasierte Eigenschaften des *Rhodopirellula baltica*^{1}-Genoms in niedriger Auflösung. Die Berechnung der Rohdaten erfolgte wie unter Abb. 3 beschrieben. Um die generellen Tendenzen im Genom aufzuzeigen, wurden die Verläufe der Werte anschließend durch Approximation mit Polynomen neunter Ordnung geglättet.

4. Das verwendete Annotationssystem: PEDANT-Pro

Zur Annotation der Genome von *Rhodopirellula baltica*^{T} und *Desulfotalea psychrophila* LSv54^T wurde das Annotationssystem PEDANT-Pro (Frishman *et al.*, 2001) eingesetzt (Biomax Informatics AG, Martinsried), mit dem bis dato 144 Bakterien-, 17 Archeaen- und 29 Eukaryonten-Genome automatisch annotiert wurden.

Das PEDANT-Pro-System erlaubt seinem Anwender, eine Fülle bioinformatischer Verfahren auf *in silico* Genvorhersagen anzuwenden und speichert die Ergebnisse in einem zentralen MySQL-RDBMS (MySQL *relational database management system*). Zu diesen Verfahren gehören BLAST-Suchen gegen verschiedene Datenbanken, wie der NCBI-nr-Proteindatenbank (ein nichtredundantes Kompilat unterschiedlicher Proteindatenbanken), der COG-Datenbank aus bekannten Familien orthologer Gene (Tatusov *et al.*, 1997), den SCOP- (Barton, 1994) und PDB- (Berman *et al.*, 2000) Datenbanken mit Proteinen bekannter Struktureigenschaften, der BLOCKS-Datenbank [einer Metadatenbank verschiedener kleinerer Datenbanken; (Henikoff & Henikoff, 1996)] sowie gegen eine Datenbank sog. genetischer Elemente (tRNAs, rRNAs, Transposons, Phagen). Hinzu kommen HMMER-Profil-Suchen gegen die PfamA und PfamB-Datenbanken (Sonnhammer *et al.*, 1997), sowie eine Suche nach Aminosäuremustern bekannter Funktion in der PROSITE-Datenbank (Bairoch, 1991). Die PROSITE-, BLOCKS- und Pfam-Datenbanken sind inzwischen in der InterPro-Metadatenbank aufgegangen (Apweiler *et al.*, 2000). Zusätzlich zu solchen extrinsischen Datenbankvergleichen berechnet das PEDANT-Pro-System eine Fülle intrinsischer Informationen für jeden ORF. Dazu gehören Vorhersagen von Transmembranregionen mit ALOM2 (Nakai & Kanehisa, 1992), sowie struktureller Proteineigenschaften [alpha-Helices, beta-Faltblätter, nicht globuläre Regionen, *coiled-coils* (Lupas *et al.*, 1991)].

Alle Ergebnisse lassen sich im PEDANT-Pro-System visuell inspizieren und sind über ein SRS (*sequence retrieval system*) mit den entsprechenden Online-Resourcen verbunden. Die Visualisierung erfolgt in Form von HTML-Ausgaben, welche mittels PERL-Skripten aus den in der MySQL-Datenbank gespeicherten Ergebnissen generiert werden (Abb. 6). Dies hat den Vorteil, daß sich die Ergebnisse plattformunabhängig mit nahezu jedem beliebigen Browser betrachten lassen, bringt aber zugleich den Nachteil mit sich, daß die Visualisierung der Information ausnahmslos statisch erfolgt. Das PEDANT-Pro-System faßt alle Ergebnisse für jeden ORF übersichtlich zusammen und stellt ein standardisiertes Formular für manuelle Annotationen bereit, welche in der MySQL-Datenbank gespeichert werden. Zudem versucht das PEDANT-Pro-System, proteinkodierenden ORFs anhand ihrer EC-Nummern sowie anhand von

Schlüsselwörtern automatisch in ein System funktioneller Kategorien einzuordnen und somit einen schnellen Überblick über die Eigenschaften eines Organismus zu schaffen.

Um eine qualitativ-hochwertige Annotation der Genome von *Rhodopirellula baltica*^{T} und *Desulfotalea psychrophila* LSv54^T zu gewährleisten, waren zahlreiche Modifikationen am PEDANT-Pro-System erforderlich. So mußten sämtliche zugrundeliegenden Datenbanken gegen aktuelle Versionen ersetzt werden. Da sich die Formate einiger Datenbank geändert hatten, waren hierzu auch Modifikationen einiger Systembestandteile erforderlich. Außerdem wurde die vom PEDANT-Pro-System verwendete BLAST-Version aktualisiert, das SRS (urspr. BioRS) geändert sowie eine externe Vorhersage von Signalpeptiden mit SignalP (Nielson *et al.*, 1997) implementiert. Letztere wurde mit Hilfe eines PERL-Skripts nachträglich in die PEDANT-MySQL-Datenbank eingebunden. Zudem wurden einige graphische Änderungen am HTML-Interface vorgenommen und ein PERL-Export-Filter geschrieben, der es erlaubt, die in diversen MySQL-Tabellen gespeicherten Annotationsergebnisse in eine EMBL-formatierte Textdatei zu exportierten. Auf diese Weise wurde nicht nur eine publikationsfähige EMBL-Datei erzeugt, sondern auch ein Reimport der Daten in das GenDB-Annotationssystem sichergestellt, das nach Auslaufen der Lizenz für das PEDANT-Pro-System im Rahmen des REGX-Projekts verwendet wird.

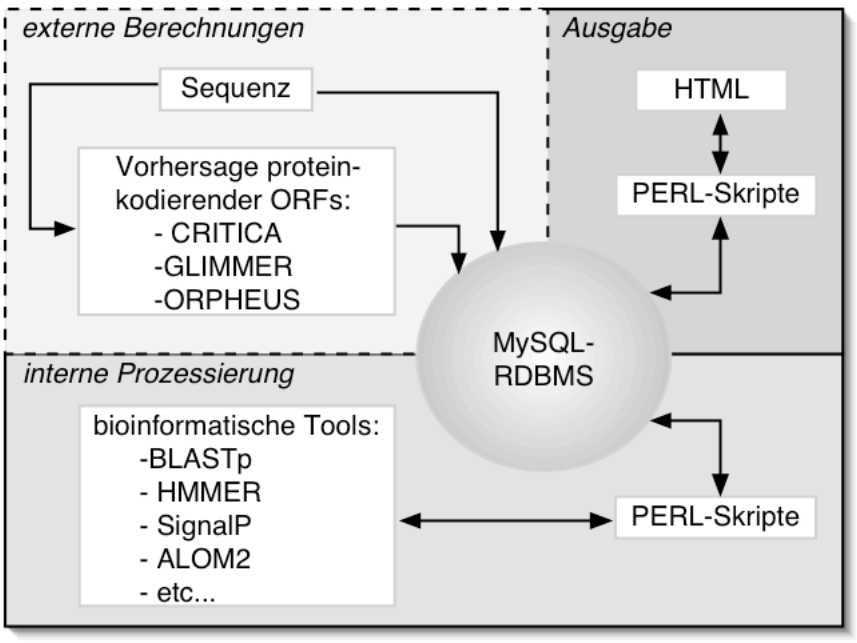


Abb. 6 Schema der Architektur des PEDANT-Pro-Annotationssystems (stark vereinfacht). Das Layout der zentralen MySQL-Datenbank ist zu komplex, um es an dieser Stelle en détail erörtern zu können. Ihm ist jedoch wegen seiner Redundanz deutlich anzumerken, daß das PEDANT-PRO-System ursprünglich auf Textdateien basierte.

5. Ausgewählte Aspekte der Annotation des *Rhodopirellula baltica*^{T}-Genoms

In diesem Kapitel sind einige interessante Aspekte der Annotation des *Rhodopirellula baltica*^{T}-Genoms zusammengefaßt. Vieles mußte dabei verkürzt oder vereinfacht werden, da eine erschöpfende Diskussion den Rahmen dieser Arbeit gesprengt hätte.

5.1 Physiologie

5.1.1 Das Substratspektrum

Das Substratspektrum eines Organismus läßt sich nicht ausschließlich anhand von bioinformatischen Analysen bestimmen. Zwar lassen sich Substrat-Transporter meistens eindeutig als solche annotieren, doch ihre Spezifität kann anhand von *in silico*-Analysen oftmals nicht eindeutig geklärt werden. Anhand von Wachstumsversuchen konnte jedoch gezeigt werden, daß *Rhodopirellula baltica*^{T} eine Fülle unterschiedlicher Mono- und Disaccharide zu verwerten vermag (Schlesner *et al.*, 2004). Aus der Annotation stammte der Hinweis auf die Verwertung von Glycerin, die durch entsprechende Laborexperimente bestätigt werden konnte. Außerdem kann N-Acetylglucosamin als einzige C- und N-Quelle genutzt werden (Rabus *et al.*, 2002). N-Acetylglucosamin ist in seiner Eigenheit als Monomer des Chitins und als Bestandteil des bakteriellen Peptidoglycans (Abb. 7) einer der häufigsten Aminozucker in marinen Habitaten (Riemann & Azam, 2002). Das Genom von *Rhodopirellula baltica*^{T} enthält jedoch keine Chitinase, so daß Chitin nur im Verbund mit Bakterien abgebaut werden kann, die Chitinasen zu sezernieren vermögen. Auf kurzkettigen, einwertigen Alkoholen und auf Aminosäuren vermag *Rhodopirellula baltica*^{T} nicht zu wachsen (Rabus *et al.*, 2002). Letzteres ist aus bioinformatischer Sicht erstaunlich. Zwar ist *Rhodopirellula baltica*^{T} Aminosäure-prototroph, d.h. es verfügt über die Gene zur Biosynthese sämtlicher proteinogenen Aminosäuren (mit

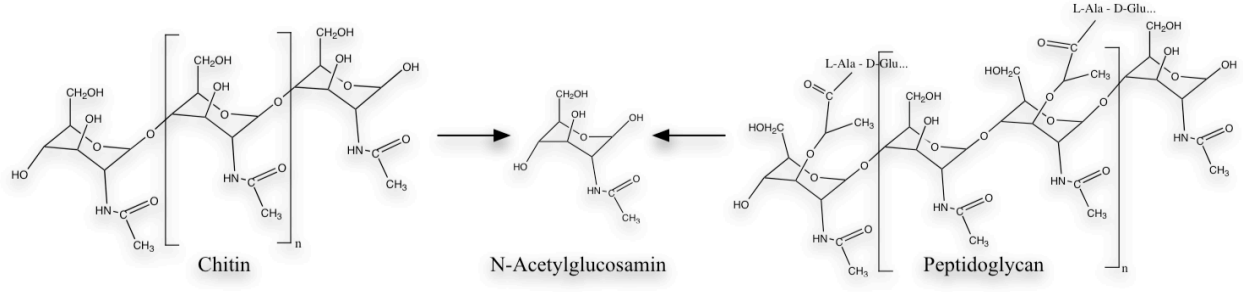


Abb. 7 N-Acetylglucosamin ist Monomer des Chitins und einer der Hauptbestandteile bakteriellen Peptidoglycans.

Ausnahme von Selenocystein), aber es konnten auch eindeutig Transporter für Prolin, Glutamat/Aspartat sowie Di- und Tripeptide annotiert werden.

Ein weiterer Widerspruch zwischen der funktionellen und der bioinformatischen Analyse findet sich im Bezug auf die Verwertung von Acetat. So findet sich im *Rhodopirellula baltica*^{T}-Genom ein Operon aus einem Acetatkinase- und einem Phosphat-acetyltransferase-Gen. Damit verfügt *Rhodopirellula baltica*^{T} über die genetische Ausstattung, um Acetat als C-Quelle nutzen zu können. Im Laborexperiment konnte jedoch kein Wachstum auf Acetat nachgewiesen werden (Schlesner *et al.*, 2004).

Insgesamt gesehen, stellt sich *Rhodopirellula baltica*^{T} anhand seines Substratspektrums als Kohlenhydratspezialist dar (Rabus *et al.*, 2002) und steht somit am Beginn der marinen Nahrungskette zur Mineralisation komplexer organischer Verbindungen.

5.1.2 Die Rolle der Sulfatasen

Eines der überraschendsten Ergebnisse der Annotation war, daß das *Rhodopirellula baltica*^{T}-Genom nicht weniger als 110 Gene enthält, die für Sulfatasen kodieren. Viele der Sulfatasen im *Rhodopirellula baltica*^{T}-Genom tragen Signalpeptide, weshalb die Vermutung naheliegt, daß es sich um Exoenzyme zum Aufschluß sulfatisierter Polysaccharide handelt. Diese werden z.B. von Algen produziert (so z.B. Carragene von Rhodophyceen) und sind Bestandteil von Fischknorpel (z.B. Chondroitin-sulfat). Sulfatisierte Polysaccharide sind daher im marinen Milieu abundant. Sulfatasen hydrolysieren das Sulfat vom Kohlenhydratgerüst sulfatisierter Polysaccharide. Da Sulfat im Meerwasser in hoher Konzentration vorliegt, dienen sie wohl kaum zur Akquisition zusätzlichen Sulfats, sondern um das Kohlenstoffgerüst sulfatisierter Polysaccharide für den weiteren enzymatischen Abbau zugänglich zu machen. Untermauert wird diese Hypothese dadurch, daß das *Rhodopirellula baltica*^{T}-Genom für zwei potentielle Carragenasen kodiert. Zwar wurde Carragen als Substrat bislang noch nicht in Wachstumsversuchen getestet (Schlesener, pers. Mitteilung), doch es konnte gezeigt werden, daß *Rhodopirellula baltica*^{T} tatsächlich mit Chondroitin-sulfat als alleiniger C-Quelle zu wachsen vermag (Schlesner *et al.*, 2004).

Im noch unvollständigen Genom des limnischen Planktomyceten *Gemmata obscuriglobus* UQM2246^T ließen sich bislang elf Sulfatasegene annotieren. Dies ist deutlich weniger als bei *Rhodopirellula baltica*^{T}, aber im Vergleich zu anderen Bakterien immer noch relativ viel. Dies deutet womöglich auf eine generelle Bedeutung von Sulfatasen für Planktomyceten hin.

5.1.3 Proteine mit Dockerin-Domäne - Hinweise auf ein Cellulosom?

Gram-positve, anaerobe Clostridien, wie z.B. *Clostridium thermocellum*, *Clostridium cellulolyticum* oder *Clostridium acetobutylicum* sezernieren ein Konglomerat aus etwa 30 unterschiedlichen Enzymen, das dem Abbau von Cellulose dient und in seiner Gesamtheit als Cellulosom bezeichnet wird. Diese Enzyme sind eng miteinander assoziiert, und zwar mittels sog. Dockerin-Domänen (Pfam PF01049, PF00028, PF00404). Neben Exo- und Endo-β-1,4-glucanasen enthält das auch eine Vielzahl von Xylanasen.

In anderen Bakterien finden sich Gene mit der Dockerin-Domäne meist gar nicht oder nur vereinzelt (Abb. 8). Im Genom von *Rhodopirellula baltica*^{T} finden sich jedoch 13 Gene mit zumeist multiplen Dockerin-Domänen. Außerdem sind im Genom nicht weniger als zehn Xylanasen annotiert, und eine Cellulase ist ebenfalls vorhanden. Ob *Rhodopirellula baltica*^{T} Cellulose abbauen kann, wurde bislang nicht untersucht. Aus Wachstumsversuchen ist jedoch bekannt, daß zumindestens das entsprechende Disaccharid Cellobiose abgebaut werden kann. Man könnte daher darüber spekulieren, daß *Rhodopirellula baltica*^{T} in seiner Eigenheit als Kohlenhydratspezialist Glykosylhydrolasen sezerniert, die denen eines Cellulosoms gleichen und über Dockerin-Domänen zu einem Multienzymkomplex aggregieren. Diese Glykosyl-

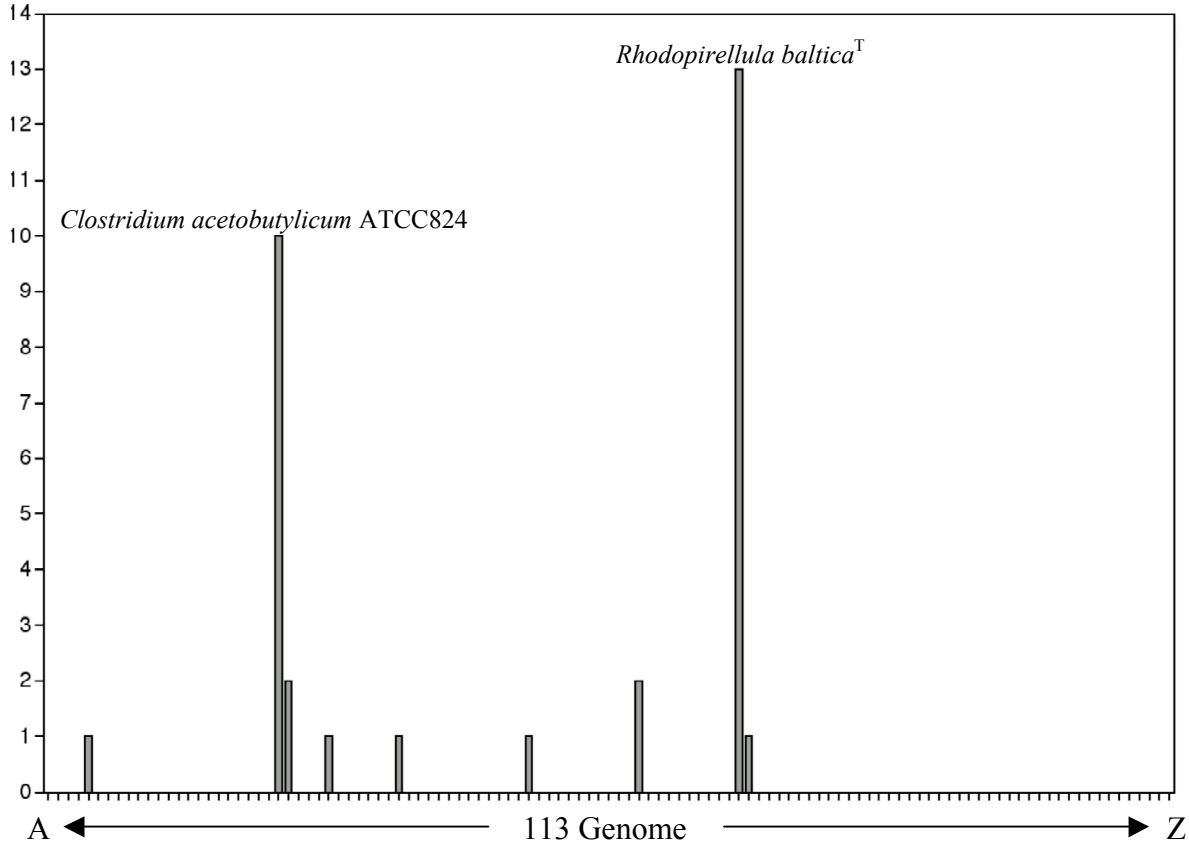


Abb. 8 Anzahl von Genen mit mindestens einer Dockerin-Domäne in 113 totalsequenzierten Genomen. Gesucht wurde mit den Pfam-Profilen PF01049, PF00028 und PF00404 bei einer Signifikanzschwelle von E=10E-5.

hydrolasen könnten z.B. über die intracytoplasmatische Membran in das Paryphoplasma sezerniert werden, womit das Pirellulosom die Funktion eines Verdauungsorganells hätte. Eine Untersuchung der Enzyme mit Dockerin-Domäne mit Pfam-Profilen aller bekannter Glykosylhydrolase-Familien erbrachte jedoch keine signifikanten Treffer, so daß es sich aller Wahrscheinlichkeit nach nicht um Glykosylhydrolasen handelt. Zudem weisen die als Xylanase-Gene annotierten Gene keine Pfam-Treffer zur zugehörigen Glykosylhydrolase-Familie zehn auf (Pfam-Motiv PF00331). Diese Gene sind somit höchstwahrscheinlich fehlannotiert und stellen eine Gruppe bislang unbekannter Glykosylhydrolasen dar. *Rhodopirllula baltica*^{T} verfügt daher wahrscheinlich über kein dem Cellulosom analoges System und die Funktion sowohl des Pirellulosoms als auch der Enyzme mit Dockerin-Domäne bleibt bislang unklar.

5.1.4 Der Energiestoffwechsel

Kohlenhydrate werden von *Rhodopirellula baltica*^{T} über die Glykolyse, den Tricarbonsäure-zyklus und die aerobe Endoxidation abgebaut. Daneben steht auch der oxidative Pentose-5-phosphat-Weg zur Verfügung, nicht jedoch der 2-Keto-3-phospho-6-phospho-gluconat-Weg. Die oxidative Decarboxylierung des aus der Glykolyse stammenden Pyruvats erfolgt über den Pyruvat-Dehydrogenase-Multienzymkomplex. Gene für die bekannten alternativen Wege zur Pyruvat-Oxidation konnten nicht gefunden werden. Der Tricarbonsäurezyklus in *Rhodopirellula baltica*^{T} ist vollständig. Ihm ist jedoch kein Glyoxylat-Zyklus beigestellt. Als anaplerotische Sequenzen stehen die Carboxylierung von Phosphoenolpyruvat und Pyruvat zur Verfügung. Bislang gibt es keine Untersuchungen darüber, an welcher der beiden Membranen in *Rhodopirellula baltica*^{T} die eigentliche Atmung stattfindet. Wie unter Pkt. 5.3.5. dargelegt wird, ist es jedoch vermutlich die intracytoplasmatische Membran des Pirellulosoms. Die Annotation des Genoms weist auf einige Besonderheiten der Atmungskette von *Rhodopirellula baltica*^{T} hin. Elektronen auf der Stufe des NADPHs werden zunächst durch eine NAD(P)-Transhydrogenase auf oxidiertes NAD übertragen und von dort mittels einer NADH-Dehydrogenase in die Atmungskette eingeschleust. Letztere transloziert keine Protonen, sondern Natrium-Kationen. Dies ist auch von anderen marinen Bakterien, wie z.B. von *Vibrio alginolyticus*, bekannt (Tan *et al.*, 1996). Von der NADH-Dehydrogenase gelangen die Elektronen in den Chinon-Pool, von wo aus sie entweder über c-Typ-Cytochrome via einer aa_3-Typ-Cytochromoxidase oder via einer d-Typ-Cytochromoxidase auf Sauerstoff übertragen werden können (ein o-Typ-Cytochrom ist nicht vorhanden). Von *Escherichia coli* ist bekannt, daß die d-Typ Cytochromoxidase eine höhere Affinität zum Sauerstoff hat und daher bei niedrigen Sauerstoffpartialdrücken exprimiert

wird (Green *et al.*, 1988). Es steht zu vermuten, daß die d-Typ-Cytochromoxidase bei *Rhodopirellula baltica*^{T} die gleiche Funktion hat. Die aa_3-Typ-Cytochromoxidase ähnelt in ihrer Aminosäuresequenz der von *Rhodothermus marinus* (die besten BLAST-Treffer für alle vier Untereinheiten entfallen auf dieses Bakterium). Dies ist insofern interessant, als daß es sich bei *Rhodothermus marinus* um ein thermophiles, marines Bakterium handelt, dessen aa_3-Typ Cytochromoxidase zahlreiche mechanistische Besonderheiten aufweist (Pereira *et al.*, 2000; Pereira *et al.*, 1999). Interessanterweise finden sich im Genom von *Rhodopirellula baltica*^{T} keine Gene für einen bc_1-Komplex, der bei den meisten aerob atmenden Bakterien und in den Mitochondrien von Eukaryonten zwischen den c-Typ-Cytochromen und der aa_3-Typ-Cytochromoxidase vermittelt und zudem Protonen transloziert. Möglicherweise verfügt *Rhodopirellula baltica*^{T} an dieser Stelle über einen alternativen Mechanismus, was jedoch nur mit Hilfe geeigneter Laborexperimente geklärt werden kann. Als weitere Besonderheit sind im *Rhodopirellula baltica*^{T}-Genom zwei Operons für ATPasen vom F1F0-Typ kodiert. So etwas ist bislang lediglich aus den Genomen von *Listeria innocua* CLIP 11262 und *Listeria monocytogenes* EGD-e bekannt. Eine hier nicht gezeigte phylogenetische Analyse hat *Pirellula marina* als nächsten Verwandten für die Gene des ersten und *Methanosarcina barkeri* als nächsten Verwandten für die Gene des zweiten Operons identifiziert. Auch die Anordnung der Gene im zweiten Operon entspricht exakt der in *Methanosarcina barkeri*. Archaeen haben normalerweise ATPasen vom V-Typ, und es steht daher zu vermuten, daß *Methanosarcina barkeri* seine F1F0-Typ ATPase via lateralem Gentransfer von einem Eubakterium erhalten hat. Dieses ATPase-Operon scheint anschließend durch einen zweiten lateralen Transfer von *Methanosarcina barkeri* (oder einem nahen Verwandten) zu *Rhodopirellula baltica*^{T} gelangt zu sein. Ob die entsprechenden Gene in *Rhodopirellula baltica*^{T} auch exprimiert werden, kann nur durch eine funktionelle Analyse, nicht jedoch mit bioinformatischen Methoden geklärt werden.

5.1.5 Das ungeklärte Verhältnis zum Sauerstoff

Mit der zuvor erwähnten d-Typ-Cytochromoxidase verfügt *Rhodopirellula baltica*^{T} bereits über einen Mechanismus, um Sauerstoff auch bei niedrigen Partialdrucken reduzieren zu können. Daneben finden sich im Genom alle erforderlichen Gene, um Kohlenhydrate vom Pentose-5-Phosphatweg abzweigend anaerob über die heterofermentative Milchsäuregärung abbauen zu können (Phosphoketolase, Actetatkinase, Lactat-Dehydrogenase). Diese Gene finden sich ebenfalls im noch unvollständig assemblierten Genom von *Gemmata obscuriglobus* UQM2246^T. Darüberhinaus konnten im *Rhodopirellula baltica*^{T}-Genom keine weiteren Hinweise auf

Gärungen oder anaerobe Atmungen gefunden werden. Wie zuvor erwähnt, sind adulte *Rhodopirellula baltica*^{T}-Zellen sessil und können daher nicht fliehen, wenn sie durch Gezeiten oder Strömungen in ein anoxisches Milieu geraten. Eine Möglichkeit zur fermentativen Energiegewinnung erscheint vor diesem Hintergrund überaus sinnvoll zu sein, doch ist es bislang nicht gelungen, *Rhodopirellula baltica*^{T} unter anoxischen Bedingungen zu kultivieren. Der Organismus ist daher auch als obligater Aerobier beschrieben worden (Schlesner *et al.*, 2004). Möglicherweise sind jedoch die geeigneten Kulturbedingungen noch nicht gefunden worden, oder aber die Energieausbeute ist unter anoxischen Bedingungen so gering, daß *Rhodopirellula baltica*^{T} zwar überlebt, sich aber nicht zu vermehren vermag.

5.1.6 C1-Metabolismus

Überraschenderweise finden sich im *Rhodopirellula baltica*^{T}-Genom nahezu alle Gene des Tetrahydromethanopterin-abhängigen C1-Stoffwechsels, der weitgehend reversibel ist und den man bis dato lediglich von methanogenen Archaeen und methylotrophen Proteobakterien kannte. Die Gene des Ribulosemonophosphat-Zyklus sind ebenfalls vorhanden (nicht die des Serin-Wegs). Die erforderlichen primären Oxidationssysteme, um auf C1-Verbindungen methylotroph wachsen zu können, fehlen dem *Rhodopirellula baltica*^{T}-Genom jedoch. Konsequenterweise konnte in Laborexperimenten kein Wachstum auf C1-Substraten wie Methanol, Formiat, Methylaminen oder Methylsulfonat nachgewiesen werden (Schlesner *et al.*, 2004). Im Genom von *Gemmata obscuriglobus* UQM2246^T finden sich ebenfalls Gene für den Tetrahydro-methanopterin-abhängigen Metabolismus von C1-Verbindungen, so daß es sich möglicherweise um eine universelle Eigenschaft von Planktomyceten handelt. Welche Rolle diese Gene spielen, ist derzeit unklar. Es könnte sich jedoch um einen Mechanismus zur Formalydehyd-Detoxifikation handeln. Eine Publikation über die möglichen physiologischen und phylo-genetischen Implikationen des Auftretens dieser C1-Gene in Planctomyceten befindet sich in der Revision (Bauer *et al.*, 2004).

5.1.7 Vitamine, Cofaktoren und Spurenelemente

Im *Rhodopirellula baltica*^{T}-Genom finden sich Hinweise auf die Biosynthese folgender Vitamine und Cofaktoren: Biotin, Coenzym A, F^{390}, FAD, Folat, Häm, Liponsäure, Molybdopterin, NAD, Pantothenat, Riboflavin, Tetrahydrobiopterin sowie Thiaminpyrophoshat. Aus Wachstumsversuchen ist bekannt, daß *Rhodopirellula baltica*^{T} auf die Zufuhr von Cobalamin (Vit. B_{12}) angewiesen ist (Schlesner *et al.*, 2004).

Die Makronährstoffe Schwefel und Stickstoff werden von *Rhodopirellula baltica*^{T} über die assimilatorische Reduktion von Sulfat und Nitrat gewonnen. Phophor wird über einen speziellen $H_2PO_4^-$/HPO_4^{2-}-Transporter aufgenommen. Außerdem stehen Transporter für diverse wichtige Ionen zur Verfügung (darunter Cu^{2+}, Fe^{2+}, Fe^{3+}, K^+, Mg^{2+}, Mn^{2+}). Fe^{3+} kann dabei scheinbar sowohl über ein dem Enterobactin homologes Siderophor, als auch über einen noch ungeklärten, Siderophor-unabhängigen Mechanismus aufgenommen werden (Angerer *et al.*, 1990).

5.1.8 Speicherstoffe und Zelleinschlüsse

Rhodopirellula baltica^T speichert vermutlich Eisen mittels eines Bakterioferritin-Eisen-Speicherproteins. Interessanterweise ist jüngst bekannt geworden, daß die adhäsiven Eigenschaften der aus Proteinen bestehenden Bysussfäden bei *Mytilus edulis* über Eisenionen reguliert werden (Sever *et al.*, 2004). Über die chemische Natur des Anheftungs-Polymers von *Rhodopirellula baltica*^{T} ist bislang nichts bekannt. Da sie eine den Bysussfäden analoge Funktion ausüben, ist jedoch ein ähnlicher chemischer Aufbau vorstellbar.

BLAST-Treffer zur Glykogen-Synthase und dem 1,4-alpha-Glucan-Verzweigungsenzym deuten außerdem auf Glykogen als Reservepolysaccharid hin. Ob die zahlreichen kleineren Einschlüsse, die in elektronenmikroskopischen Aufnahmen von *Rhodopirellula baltica*^{T}-Zellen beobachtet werden konnten (Schlesner *et al.*, 2004) mit diesen Reservestoffen in Verbindung stehen, ist bislang unklar.

5.1.9 Schleime

Im *Rhodopirellula baltica*^{T}-Genom finden sich einige sehr große Operons aus zahlreichen unterschiedlichen Glykosyltransferasen, die wegen ihrer Nachbarschaft mit Genen zu Colansäure-Biosynthese aller Wahrscheinlichkeit nach in der Produktion von Polysaccharid-Schleimen involviert sind (*Rhodopirellula baltica*^{T} ist ein starker Schleimproduzent).

5.1.10 Streßantwort

Rhodopirellula baltica^T gehört zu den wenigen bekannten Organismen, in denen neben der Superoxid-Dismutase alle vier bekannten Katalase-Typen vorkommen (Monohäm-Katalase, Dihäm-Katalase, Mangan-abhängige Katalase, Thiol-Peroxidase). Damit ist *Rhodopirellula baltica*^{T} gut gegen oxidativen Streß gewappnet. Wie eine genomübergreifende BLAST-Suche zeigte, kommen alle vier Katalasen nur in fünf weiteren unter den bislang totalsequenzierten

Bakterien vor (*Bacillus halodurans* C-125, *Escherichia coli* O157:H7 EDL933, *Escherichia coli* O157:H7 VT2-Sakai, *Salmonella typhi* CT18, *Salmonella typhimurium* LT2 SGSC 1412). Neben den Genen zur SOS-Antwort (*recA, lexA, uvrABC*) findet sich im *Rhodopirellula baltica*^{T}-Genom ein weiterer Mechanismus zur Verhinderung photoxidativer Schäden: Gene für die Biosynthese des Carotinoid-Vorläufers Phytoen liegen direkt neben einer Photolyase und bilden vermutlich ein Operon. Es hat somit den Anschein, als ob das Auftreten photochemisch entstandener Thymin-Dimere nicht nur durch die Expression einer Photolyase wieder voneinander getrennt werden, sondern zeitgleich ein Carotinoid zum UV-Schutz synthetisiert wird. Dies würde auch die rosa Pigmentierung von *Rhodopirellula baltica*^{T}-Zellen erklären, die jedoch laut Laborbefund eher konstitutiv zu sein scheint. Interessant ist auch das Auftreten eines Gens für bakterielles Hämoglobin. Bakteriellem Hämoglobin wird eine Rolle bei der NO-Detoxifikation zugeschrieben (Couture *et al.*, 1999; Yeh *et al.*, 2000).

Neben diesen Mechanismen gibt es im *Rhodopirellula baltica*^{T}-Genom eine Fülle von Genen für detoxifizierende Transportsysteme, so z.B. für Schwermetalle und Arsenat (weitere Reaktionen auf unterschiedliche Stressoren sind in Publikation 1 beschrieben).

5.1.11 Antibiotika und Xenobiotika

Die Annotation hat bislang keinen Hinweis auf einen vollständigen, bekannten Abbauweg für Xenobiotika, wie z.B. dem Aromatenabbau, erbracht. Es gibt jedoch Gene, die auf die Produktion von Polyketid-Antibiotika hinweisen. Diese Möglichkeit wird derzeit von Chris Würdemann im Rahmen einer Promotion mit Hilfe von Transkriptomanalysen über DNA-Chips erforscht (der Schwerpunkt dieser Arbeit liegt allerdings auf der Untersuchung der Sulfatasen).

5.2 Zellteilung und -zyklus

5.2.1. Der Replikationsursprung

Der Replikationsursprung (*oriC*) liegt bei *Rhodopirellula baltica*^{T}, wie bei vielen anderen Bakterien [z.B. bei *Thermus thermophilus* (Schaper *et al.*, 2000)], neben dem *dnaN*-Gen, welches die β-Untereinheit der replikativen DNA-Polymerase III kodiert. *Rhodopirellula baltica*^{T} hat, wie *Gemmata obscuriglobus* UQM2246^T, nicht nur ein sondern zwei unterschiedliche Gene für das Replikationsinitiationsprotein DnaA. Unter allen anderen totalsequenzierten Bakerien ist dies nur auch bei den Chlamydien der Fall (vgl. Pkt. 7.). Außerdem

liegen die *dnaA*-Gene nicht, wie bei den meisten anderen Bakterien, dicht beim oder am *oriC* (Fujita *et al.*, 1992; Ogasawara & Yoshikawa, 1992), sondern ungewöhnlich weit davon entfernt. Die *dnaA*-Gene selbst haben Promotoren mit DnaA-Bindestellen, was auf Autoregulation hindeutet. Der Aufbau des *oriC* von *Rhodopirellula baltica*^{T} ist dem anderer diesbezüglich untersuchter Bakterien recht ähnlich. Er besteht aus einer 70 bp langen AT-reichen Region, die auf der einen Seite von einer DnaA-Bindestelle und auf der anderen Seite von drei DnaA-Bindestellen sowie einigen Oligonukleotid-Palindromen flankiert wird. Außerdem weist der *oriC* zwei Methylierungsstellen auf.

5.2.2. DNA-Polymerase III

Die Replikation erfolgt bei den Bakterien durch die DNA-Polymerase III, während sie bei Archaeen und Eukaryonten durch die DNA-Polymerase II erfolgt. Letztere findet sich auch bei einigen Bakterien (z.B. *Bacillus subtilis*, *Escherichia coli, Vibrio cholerae, Pseudomonas* sp.), wo sie allerdings nicht der Replikation sondern der DNA-Reparatur dient.

Bei *Escherichia coli* ist die DNA Polymerase III gut untersucht und wie folgt aufgebaut: Ein Ring (*clamp*) aus β-Untereinheiten (*dnaN*) umfaßt die DNA und hält den Replikationsapparat in Position. Die Assoziation des Rings erfolgt mit Hilfe des γ-Komplexes (*clamp loader*), der aus den Untereinheiten γ (*dnaX*), δ (*holA*), δ' (*holB*), χ (*holC*), und ψ (*holD*) besteht. Über eine oder zwei τ-Untereinheiten (*dnaX*) sind zwei replikative Zentren aus den Untereinheiten α (*dnaE*), ε (*dnaQ*) und θ (*holE*) mit dem Rest des Replikationsapparats verbunden. Die α-Untereinheit ist die eigentlich katalytische Untereinheit, während die ε-Untereinheit 3'→5'-Exonucleaseaktivität und somit Korrekturfunktion hat. Von den zwei Replikationszentren repliziert eines den kontinuierlich- und das andere den diskontinuierlich-replizierten Strang. Der gesamte Komplex hat die Zusammensetzung $(\alpha\varepsilon\theta)_2X(\delta\delta'\chi\psi) + \beta_n$, wobei X entweder $\tau_2\gamma$ oder $\tau\gamma_2$ ist (Pritchard *et al.*, 2000). Wegen der Asymmetrie der Replikation ist die DNA-Polymerase III asymmetrisch gebaut (Glover & McHenry, 2001; Pritchard *et al.*, 2000).

Bei *Thermotoga maritima* MSB8^T, sowie den *Mollicutes* und *Firmicutes* kommt neben DnaE eine zweite Varinte der α-Untereinheit vor. Dieses PolC genannte Protein trägt sowohl die Replikationsfunktion als auch die 3'→5'-Exonuclease-Korrekturfunktion. Es wird vermutet, daß PolC für die kontinuierliche und DnaE für die diskontinuierliche Replikation verantwortlich ist (Dervyn *et al.*, 2001; Glover & McHenry, 2001; Rocha, 2002). Untersuchungen an

Streptococcus pyogenes weisen auf die Zusammensetzungen $(\alpha\varepsilon\theta)(\text{polC})(\tau_4)(\delta\delta') + \beta_n$ hin (Bruck & O'Donnell, 2000).

Die DNA-Polymerase III von *Rhodopirellula baltica*^{T} entspricht vermutlich keinem dieser Schemata, denn im *Rhodopirellula baltica*^{T}-Genom kommen zwei unterschiedliche Varianten des Gens *dnaE* vor. Das *dnaX*-Gen in *Rhodopirellula baltica*^{T} weist, wie in *Escherichia coli* (Flower & McHenry, 1990; Tsuchihashi, 1991; Tsuchihashi & Kornberg, 1990; Tsuchihashi & Brown, 1992) oder in *Thermus thermophilus* (Yurieva *et al.*, 1997), eine *frameshift*-Stelle auf und kodiert daher vermutlich neben der τ-Untereinheit (vollständiges Transkript) auch eine γ-Untereinheit (partielles Transkript). Neben den Genen für die α-, τ- und γ-Untereinheiten finden sich im *Rhodopirellula baltica*^{T}-Genom Gene für die ε-, und δ'-Untereinheit. Das *holA*-Gen für die δ-Untereinheit konnte nicht gefunden werden, ist jedoch generell sehr schlecht konserviert und wird daher bei vielen Genom-Annotationen übersehen (Bruck & O'Donnell, 2000). Gene für die θ-, χ- und ψ- Untereinheiten wurden dagegen nicht gefunden. Diese sind jedoch für eine funktionsfähige DNA-Polymerase III auch nicht erforderlich. In Analogie zur replikativen Polymerase der *Firmicutes* und *Mollicutes* mit zwei unterschiedlichen katalytischen Zentren (PolC, DnaE) erscheint es nicht unwahrscheinlich, daß in *Rhodopirellula baltica*^{T} die Produkte beider *dnaE*-Gene Bestandteil der replikativen DNA-Polymerase III sind. Damit ergäbe sich für die Polymerase III von *Rhodopirellula baltica*^{T} folgende Zusammensetzung: $(\alpha^1 1\varepsilon)(\alpha^2 1\varepsilon)X(\delta\delta') + \beta_n$ (s.o.). Ein solcher Aufbau wäre ein Novum, ist aber ohne experimentelle Beweise natürlich rein spekulativ.

Die phylogenetischen Implikationen des vermeintlichen Baus der replikativen DNA-Polymerase III von *Rhodopirellula baltica*^{T} sind nicht eindeutig. Es ist darüber spekuliert worden, daß das ursprüngliche *dnaE*-Gen noch vor der Aufspaltung zwischen den Gram-positiven und Gram-negativen Bakterien vor etwa 1.2 Milliarden Jahren dupliziert worden sein könnte. In den Gram-negativen Bakterien soll eine dieser Kopien verloren gegangen sein, während die Gene in den *Mollicutes* und *Firmicutes* zu den heutigen *polC*- und *dnaE*-Genen evolvierten, welche somit paralog wären (Koonin & Bork, 1996). Das Auftreten zweier *dnaE*-Gene in *Rhodopirellula baltica*^{T} könnte somit ein Indiz für ein Abzweigen dieses Phylums vor dem Verlust der zweiten Kopie in den Gram-negativen Bakterien sein, oder aber ein Indiz für eine weitere Genduplikation nach dem Verlust einer der beiden Kopien. Jüngere Untersuchungen zeigen jedoch, daß die Evolution der DNA-Polymerasen weitaus komplizierter ist und von vielen lateralen Transferereignissen bestimmt wird (Filee *et al.*, 2002).

5.2.3. Zellteilungsgene

Die Zellteilung der *Bacteria* ist gut untersucht und besteht aus einer fein abgestimmten Expression verschiedener Zellteilungsgene, darunter *ftsZ, ftsA, ftsI, ftsL, ftsQ, ftsN, zipA* und *ftsW*. Das *ftsZ*-Gen spielt eine Schlüsselrolle bei der Zellteilung, weil es das Rückgrat des sog. septalen Rings bildet. Dieser kontraktile Ring wird bei der Zellteilung direkt unterhalb der Cytoplasmamembran ausgebildet und ist über Proteine mit der Peptidoglycanzellwand verbunden. Das *ftsZ*-Gen kommt in allen Eubakterien mit Ausnahme von *Ureaplasma urealyticum parvum biovar* serovar 3 und den Chlamydien vor (Margolin, 2000). Auch die Archaeen haben, mit Ausnahme von *Aeropyrum pernix* $K1^T$, ein dem *ftsZ*-Gen homologes Gen. Überraschenderweise fehlen aber bei *Rhodopirellula baltica*^{T} mit Ausnahme von *ftsK* sämtliche *fts*-Gene, und auch im fast kompletten Genom von *Gemmata obscuriglobus* $UQM2246^T$ fehlt das *ftsZ*-Gen bislang. Die Zellteilung der Planktomyceten muß daher, wie bei den Chlamydien auch, nach einem noch völlig unbekannten Mechanismus erfolgen.

5.2.4. Der Zellzyklus

Wie zuvor erwähnt, weist *Rhodopirellula baltica*^{T} einen Zellzyklus auf, der in seinem Verlauf dem von *Caulobacter cresentus* CB15 ähnelt: Adulte Zellen produzieren begeißelte, aber reproduktionsinaktive Schwärmerzellen, die nach einiger Zeit ihre Geißeln verlieren und wieder zu adulten Zellen heranreifen. Der Zellzyklus von *Caulobacter cresentus* CB15 ist Dank jahrelanger Forschungen der Arbeitsgruppen um Lucy Shapiro und Urs Jenal gut untersucht, und hat zu Publikationen geführt, die zu zahlreich sind, um an dieser Stelle aufgeführt zu werden (Jenal *et al.*, 1995; Marczynski & Shapiro, 2002). Dreh- und Angelpunkt des Zellzyklus in *Caulobacter cresentus* CB15 ist das Master-Kontrollprotein CtrA, ein *reponse*-Regulator, der als Dimer an das Motif TTAA-N_7-TTAA im Replikationsursprung bindet und die Expression der am Zellzyklus beteiligten Gene über ein Kaskade von Zweikompenten-Systemen steuert. Im Genom des mit *Caulobacter cresentus* CB15 verwandten Alphaproteobakteriums *Rickettsia prowazekii* sind leicht abweichende Motive für ein CtrA-Analogon (CzcR) nachgewiesen worden (Brassinga *et al.*, 2002). Aus bioinformatischer Sicht sind Histidinkinasen und *response*-Regulatoren von Zweikomponentensystemen leicht als solche zu identifizieren, die Wirkungsweise der betreffenden Zweikomponentensysteme kann jedoch in aller Regel nicht durch Sequenzhomologien geklärt werden. So konnte im *Rhodopirellula baltica*^{T}-Genom auch kein dem *ctrA*-Gen homologer *reponse*-Regulator gefunden werden. Im Replikationsursprung von *Rhodopirellula baltica*^{T} finden sich jedoch Motive, die den CtrA und CzcR-Bindestellen ähneln

(z.B. TTAA-N_7-AAAC). Es erscheint daher wahrscheinlich, daß der Zellzyklus in *Rhodopirellula baltica*^{T} auf ähnliche Weise gesteuert wird wie in *Caulobacter crescentus* CB15.

5.3 Morphologie

5.3.1 Das Nukleoid

Die elektronenmikroskopische Sichtbarkeit der DNA von Planktomyceten deutet auf eine starke Aufwindung und einen damit verbundenen hohen Organisationsgrad hin. Eine vergleichende Untersuchung aller bislang sequenzierten Genome mit Helicase-Pfam-Profilen (PF00521, PF03989, PF00204, PF00986, PF00270, PF00772, PF03796, PF03457, PF04408, PF00271, PF00570, PF03880, PF00580, PF00176, PF01330) zeigte, daß das *Rhodopirellula baltica*^{T}-Genom mit vier Genen für Helicasen von SNF2-Typ die höchste Anzahl solcher Gene hat. SNF2-Typ-Helicasen sind für die Auf- und Entwindung von DNA verantwortlich. Allerdings zeigte sich auch, daß die Zahl der Helicasen in bakteriellen Genomen stark mit der Genomgröße korreliert, und daß die Zahl von vier SNF2-Typ Helicasen bei einer Genomgröße von über sieben Megabasen nicht ungewöhnlich ist.

Im *Rhodopirellula baltica*^{T}-Genom finden sich zwei Gene, die prokaryontische Histone kodieren. Diesem Hinweis folgend wurde versucht, über eine Periodizitätsanalyse Hinweise auf eine eventuelle Histonassoziaton zu erhalten. Dazu wurden die Basenperiodizitäten mit Hilfe eines eigens geschriebenen PERL-Skripts ermittelt und anschließend im Programm STATISTICA (http://www.statsoft.com) einer Fourier-Transformation mit Tappering, Mittelwertsubtraktion und Trendbereinigung unterzogen (Bailey *et al.*, 2000; Fukushima *et al.*, 2002; Widom, 1996). Die Analyse der Periodogramme zeigte zwar die für Eubakterien mit positiv gewundener DNA typische Periodizität von 11.5 bp (Gribaldo & Philippe, 2002), eine eindeutige Histonassoziation ließ sich jedoch nicht nachweisen.

5.3.2 Zellwandproteine

Obwohl die Aminosäurecharakteristika der Zellwandproteine verschiedener Planktomyceten bekannt sind (Giovannoni *et al.*, 1987; König *et al.*, 1984; Stackebrandt *et al.*, 1986), erweisen sich diese als zu unspezifisch, als daß im Genom von *Rhodopirellula baltica*^{T} potentielle Gene für die Zellwandproteine hätten identifiziert werden können. Tatsächlich erbrachte die gesamte

Annotation des *Rhodopirellula baltica*^{T}-Genoms keinen einzigen Hinweis auf die Gene für die Proteinzellwand.

5.3.3 Peptidoglycan-Biosynthese

Das *Rhodopirellula baltica*^{T}-Genom enthält einige der Gene zur Peptidoglycan-Biosynthese [*murB, murE, murG, ddlA* und *upk* (*bakA*)], wohingegen andere fehlen (*murA, murC, murD, murF, dda*) Dies deutet darauf hin, daß die Planktomyceten nicht, wie vormals vermutet, von Bakterien aus einer Entwicklungslinie von vor der Evolution des Peptidoglycans abstammen (Stackebrandt *et al.*, 1984). Stattdessen scheinen die Planktomyceten einst Peptidoglycan besessen zu haben, und es handelt sich bei den Proteinzellwänden um eine sekundäre Adaption. Diese Situation entspricht genau derjenigen im zweiten bekannten bakteriellen Phylum, in dem keine Peptidoglycanzellwände vorkommen, nämlich den Chlamydien. Die Infektionsstadien der Chlamydien haben, ebenso wie die Planktomyceten, Proteinzellwände (Hatch *et al.*, 1986). Ihre Genome weisen jedoch noch alle zur Peptidoglycan-Biosynthese erforderlichen Gene auf (Ghuysen & Goffin, 1999; Stephens *et al.*, 1998). Phylogenetische Analysen legen eine Verwandtschaft beider Gruppen nahe (Pkt. 7), und somit hat es den Anschein, als ob die Gene zur Peptidoglycan-Biosynthese entweder bei den Planktomyceten infolge schneller Evolution partiell verloren gegangen sind, oder aber bei den Chlamydien noch eine Rolle spielen.

5.3.4 Lipid A-Biosynthese

Anhand der Fettsäurezusammensetzung wurde bereits vermutet, daß Planktomyceten Lipid A produzieren (Kerger *et al.*, 1988). Tatsächlich enthält das *Rhodopirellula baltica*^{T}-Genom sämtliche zur Lipid A-Biosynthese erforderlichen Gene. Die Schlüsselgene zur Biosynthese einer O-spezifischen Seitenkette (O-Antigenligase, O-Antigenpolymerase) fehlen jedoch.

5.3.5 Das Problem des *protein targeting*

Die intracytoplasmatische Membran des Pirellulosoms in *Rhodopirellula baltica*^{T} stellt eine Barriere für den Transport von Proteinen zwischen dem Ribo- und dem Paryphoplasma dar. Dies wirft natürlich die Frage auf, wie Proteine gezielt in das Paryphoplasma gelangen (*protein trageting*). Wie unter Pkt. 2 erwähnt, weist die intracytoplasmatische Membran in *Isosphera pallida* und *Planctomyces* spp. eine Invagination auf, an der Ribosomen wie in einem eukaryontischen rauhen ER angelagert sind. Dies könnte in Analogie zu einem rauhen ER darauf

hindeuten, daß bei diesen Organismen die naszierenden Proteine von den Ribosomen durch die intracytoplasmatische Membran hindurch direkt in das Paryphoplasma sezerniert werden. Bei *Rhodopirellula baltica*^{T} ist eine dem ER analoge Struktur bislang nicht beobachtet worden, so daß das *protein targeting* vermutlich über die konventionellen Wege mit Hilfe von Signalpeptiden oder Protein-Translokationssystemen erfolgt.

Rhodopirellula baltica^T hat die Schlüsselgene für Proteintranslokationssysteme vom Sec-, Tat- und GSP-Typ. Wie eine vergleichende Analyse von 112 totalsequenzierten Bakteriengenomen mit Pfam-Profilen (PF00263, PF03958, PF02501, PF05134, PF00482, PF04612, PF01203, PF03840, PF00584, PF01043, PF02556, PF00344, PF00902) zeigte, hat *Rhodopirellula baltica*^{T} die höchste Zahl an Kopien von *secA*-Genen (3), Genen mit GSP Type-II F-Domäne (6) sowie Genen für das GSP Typ-II/III Secretionsprotein (9). Bei einem Screening nach Genen, die sowohl das Tat-Motif als auch eine Signalpeptid-Vorhersage aufwiesen, stellte sich *Rhodopirellula baltica*^{T} sowohl in absoluter (135) als auch in relativer Hinsicht (18.9 / Mb) als Spitzenreiter heraus. Gesucht wurde dabei nach dem Motif SRRxFLK, wobei höchstens zwei abweichende Aminosäuren erlaubt und die beiden Arginine konserviert sein mußten (Palmer & Berks, 2003).

Diese Befunde unterstreichen, daß die Translokation von Proteinen in *Rhodopirellula baltica*^{T} eine wichtige Rolle spielt.

5.3.6 Die Flagelle (Motilität und Taxis)

In *Rhodopirellula baltica*^{T} kommen sämtliche zur Biosynthese einer Flagelle erforderlichen Gene vor (*flgA, flgB, flgC, flgD, flgE, flgG, flgF, flgH, flgI, flgK, flgL, fliC, fliD, fliE, fliF, fliG, fliH, fliJ, fliK, fliL, fliM, fliN, fliP, fliQ, fliR, fliS, fliT, flhA, flhB, flhC, flhD, flhE, motA, motB*). Unter diesen befinden sich mit *flgH* und *flgI* auch die Gene für die L- und P-Ringe der Flagelle. Diese kommen ansonsten nur bei Gram-negativen Bakterien mit einer äußeren Membran vor und fehlen folglicherweise in den Genomen aller bislang totalsequenzierten flagellentragenden, Gram-positiven Bakterien (z.B. *Bacillus halodurans* C-125, *Bacillus subtilis* subsp. *subtilis* 168, *Clostridium actobutylicium* ATCC 824, *Clostridium tetani* E88, *Listeria innocua* CLIP 11262, *Listeria monocytogenes* EGD-e, *Oceanobacillus iheyensis* HTE831^T). Sechzehn der an der Flagellen-Biosynthese beteiligten Gene sind in Operons zu je vier Genen organisiert (*fli-fliQ-fliR-flhB* / *flgB-flgC-fliE-fliF* / *flddD-motA-ompA-fliN* / *flgI-flgH-flgG-flgF*). Diese werden vermutlich nacheinander in einer für die konzertierte Assemblierung der Flagelle erforderlichen Reihenfolge exprimiert. Erstaunlicherweise findet sich jedoch kein vollständiges System zur

Chemotaxis-Signaltransduktion im *Rhodopirellula baltica*^{T}-Genom. So findet sich kein Gen für das MCP (*methyl accepting chemotaxis protein*), und von den Chemotaxisgenen (*che*) ist lediglich *cheY* vorhanden. Eine genomübergreifende Analyse aller bislang totalsequenzierten bakteriellen Genome mit Hilfe von Pfam-Profilen zeigte, daß unter den begeißelten Bakterien neben *Rhodopirellula baltica*^{T} bislang nur drei weitere keine Chemotaxisproteine aufweisen. Diese sind *Aquifex aeolicus* VF5, *Nostoc* sp. PCC7120 und *Mesorhizobium loti* MAFF303099, wobei das Fehlen von Chemotaxisproteinen in der Erstpublikation des Genoms von *Aquifex aeolicus* VF5 explizit erwähnt wird (Deckert *et al.*, 1998). Für den Planktomyceten *Pirellula marina* hingegen sind bereits die Chemotaxisproteine *cheB* und *cheC* nachgewiesen worden (Jenkins *et al.*, 2002), und auch im noch unvollständigen Genom von *Gemmata obscuriglobus* UQM2246^T finden sich ein Fülle von Chemotaxisgenen (z.B. *cheA, cheB, cheR, cheY, mcpD*). Letztere liegen oftmals geclustert vor und bilden vermutlich Operons. Vor diesem Hintergrund erscheint es unwahrscheinlich, daß die Signaltransduktion zur Chemotaxis bei den Planktomyceten nach einem vollständig anderen Mechanismus als dem bislang bekannten verläuft. Als Alternative bliebe, daß sich die Schwärmerzellen von *Rhodopirellula baltica*^{T} ungerichtet bewegen, und die Besiedlung neuer Lebensräume nach dem Zufallsprinzip erfolgt. Die Verbreitungsstrategie von *Rhodopirellula baltica*^{T} entspräche damit derjenigen der meisten Pflanzen. Einfache Chemotaxisversuche könnten diese Frage klären helfen, wurden aber bis heute leider nicht durchgeführt.

5.3.7 Kompartimentierung

Aus dem *Rhodopirellula baltica*^{T}-Genom ergeben sich einige Hinweise, daß die Planktomyceten von Gram-negativen Bakterien mit einer inneren Cytoplasmamembran, einer dünnen Peptidoglycanschicht und einer äußeren Membran mit dem Lipopolysaccharid Lipid A abstammen. Zu diesen Hinweisen gehören das Vorhandensein sämtlicher Gene zur Lipid A-Biosynthese (Pkt. 5.3.4), die Struktur der DNA-Polymerase III (Pkt. 5.2.2), das Fehlen der für Gram-positive Bakterien typischen Signatursequenzen (Gupta & Griffiths, 2002) in den Genen *rpsL* (ribosomales Protein S12) und *secF* (Protein-Exportsystem), sowie das Vorhandensein der Gene für die L- und P-Ringe der Flagelle (Pkt. 5.3.6).

Die Proteinzellwand der Planktomyceten könnte sich daher durch eine Zunahme und Quervernetzung von peripheren Proteinen der äußeren Membran eines Gram-negativen Vorläufers entwickelt haben. Als diese Proteinschicht stabil genug war, um den osmotischen Druck des Protoplasten zu halten, wurde die dünne Peptidoglycanschicht entbehrlich und konnte

sukzessive reduziert werden (Abb. 9). Dies würde das Vorhandensein von Genen zur Peptidoglycan-Biosynthese im *Rhodopirellula baltica*^{T}-Genom erklären (Pkt. 5.3.3). Die rezente Cytoplasmamembran der Planktomyceten wäre somit der äußeren Membran des Gram-negativen Vorläufers homolog und die intracytoplamatische Membran der ehemaligen Cytoplasmamembran. Das Paryphoplasma wäre demzufolge aus dem ehemaligen periplasmatischen Raum hervorgegangen, was erklärte, warum es keine Ribosomen enthält.

Wenn diese Hypothese richtig ist, dann müßte die Cytoplasmamembran der Planktomyceten das Lipid A, und die intracytoplasmatische Membran die Komponenten der Atmungskette tragen. Leider liegen entsprechende Studien bislang nicht vor.

Vor dem Hintergrund dieser Befunde nimmt sich die Kompartimentierung der Planktomyceten weit weniger spektakulär aus, als ursprünglich angenommen. Insbesondere erscheinen Überlegungen fraglich, wonach die Planktomyceten das am tiefsten abzweigende bakterielle Phylum sein könnten sowie auch Spekulationen darüber, daß ihre Kompartimentierung derjenigen von Eukaryonten homolog sein könnte (Brochier & Philippe, 2002). Allerdings greift die hier vorgeschlagenen Hypothese zur Erklärung der Kompartimentierung in Planktomyceten zu kurz. Sie kann zwar das große Kompartiment in den Gattungen *Blastopirellula, Isophera, Pirellula, Planctomyces* und *Rhodopirellula* erklären, nicht jedoch die zusätzlichen Membranen in *Gemmata* spezies und den Anammox-Planktomyceten (Pkt. 2.), für die weitere Evolutionsschritte angenommen werden müssen (z.B. Invaginationen mit anschließender Abschnürung von Kompartimenten).

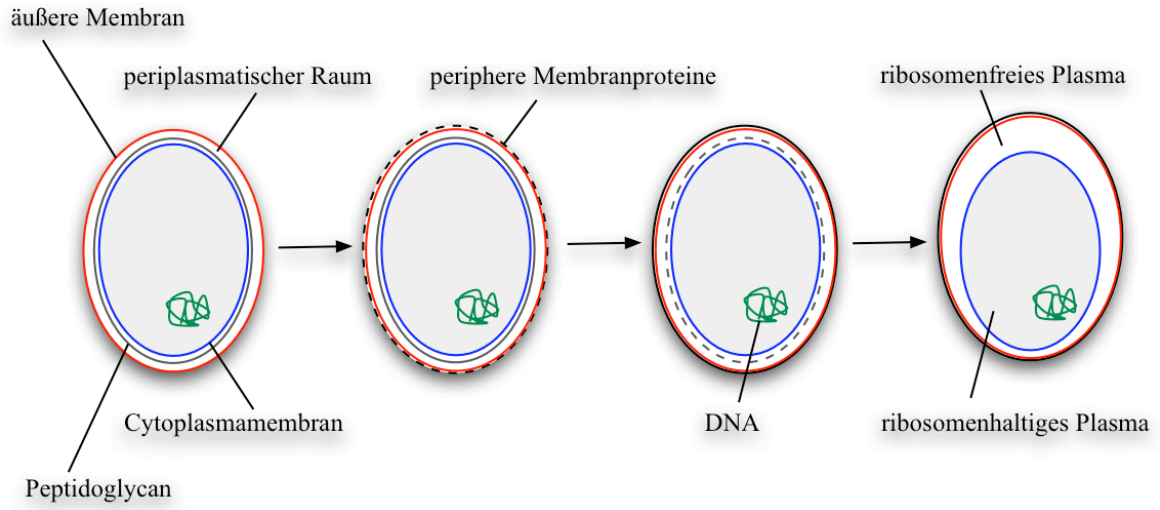


Abb. 9 Hypothese zur Entstehung der Kompartimentierung in Planctomyceten. Äußere Membran (LPS) = rot, Peptidoglycan-Zellwand = dunkelgrau, Protein-Zellwand = schwarz, Cytoplasmamembran = blau, Cytoplasma = hellgrau, DNA = grün.

5.4 Lateraler Gentransfer (LGT)

Eine systematische Untersuchung des Anteils potentiell lateral transferierter Gene im *Rhodopirellula baltica*^{T}-Genom steht noch aus. Es konnten jedoch einige Bereiche identifiziert werden, die mit großer Wahrscheinlichkeit aus lateral transferierten Genen bestehen. Diese Bereiche sind durch Gene mit z.T. atyptischem GC-Gehalt charakterisiert, die durchweg beste BLASTp-Treffer zu ein und demselben Organismus aufweisen und zudem ähnlich wie in diesem angeordnet sind (Abb. 10). Das markanteste Beispiel ist sicherlich das zweite ATPase-Operon vom F1F0-Typ in *Rhodopirellula baltica*^{T} (5.1.4), welches aller Wahrscheinlichkeit nach über LGT von einem Eubakterium zu *Methanosarcina barkeri* (oder einem engen Verwandten) und über ein weiteres LGT-Event von dort in *Rhodopirellula baltica*^{T} gelangte. Stellenweise lassen sich solche LGT-Bereiche durch einen deutlich abweichenden CAI oder anhand eines sog. Phyloatlas identifizieren (Appendix, Abb. 17). Neben ganzen LGT-Bereichen finden sich auch eine Reihe von Einzelgenen in *Rhodopirellula baltica*^{T}-Genom, die mit hoher Wahrscheinlichkeit via LGT in das Genom gelangt sind. Zu den eindeutigen Fällen gehören z.B. ein Gen aus der Streptomycin-Biosynthese von *Streptomyces griseus* (oder einem Verwandten) sowie ein archaeelles Membranprotein.

Untersuchungen an den Planktomyceten *Gemmata obscuriglobus* und *Pirellula marina* haben Gene zutage gefördert, die sonst typischerweise nur in Eukaryonten auftreten (Jenkins *et al.*, 2002). So finden sich in *Gemmata obscuriglobus* z.B. Gene zur Steroid-Biosynthese, was außer bei Eukaryonten bislang nur noch für das Archeaon *Methylococcus capsulatus* nachgewiesen werden konnte (Pearson *et al.*, 2003). Diese Gene sind höchstwahrscheinlich über LGT in diese Planktomyceten gelangt, aber auch über eine vorübergehende endosymbiontische Assoziation zwischen Planktomyceten und Archaeen im Laufe der Evolution ist spekuliert worden (Pearson *et al.*, 2003). Auch in *Rhodopirellula baltica*^{T} finden sich einige ungewöhnliche Gene, die auf LGT von Eukaryonten hindeuten. So tragen etwa 25% der in *Rhodopirellula baltica*^{T} aufgefundenen 110 Sulfatase-Gene typisch eukaryontische Merkmale (Publikation 1). Mindestens sechs Gene haben eine entfernte Ähnlichkeit mit eukaryontischem Myosin, und weiter sechs Gene haben zum Teil multiple Cadherin-Domänen, die normalerweise nur bei eukaryontischen Proteinen zur Zell-Zell-Anheftung vorkommen. Im Bereich um ~6.4 Mb findet sich zudem eine leichte Häufung von Genen mit besten BASTp-Treffern zu Eukaryonten. Dieser Bereich ist sowohl durch seine abweichende *codon usage* (Abb. 4) als auch im Phyloatlas (Appendix, Abb. 17) gut erkennbar.

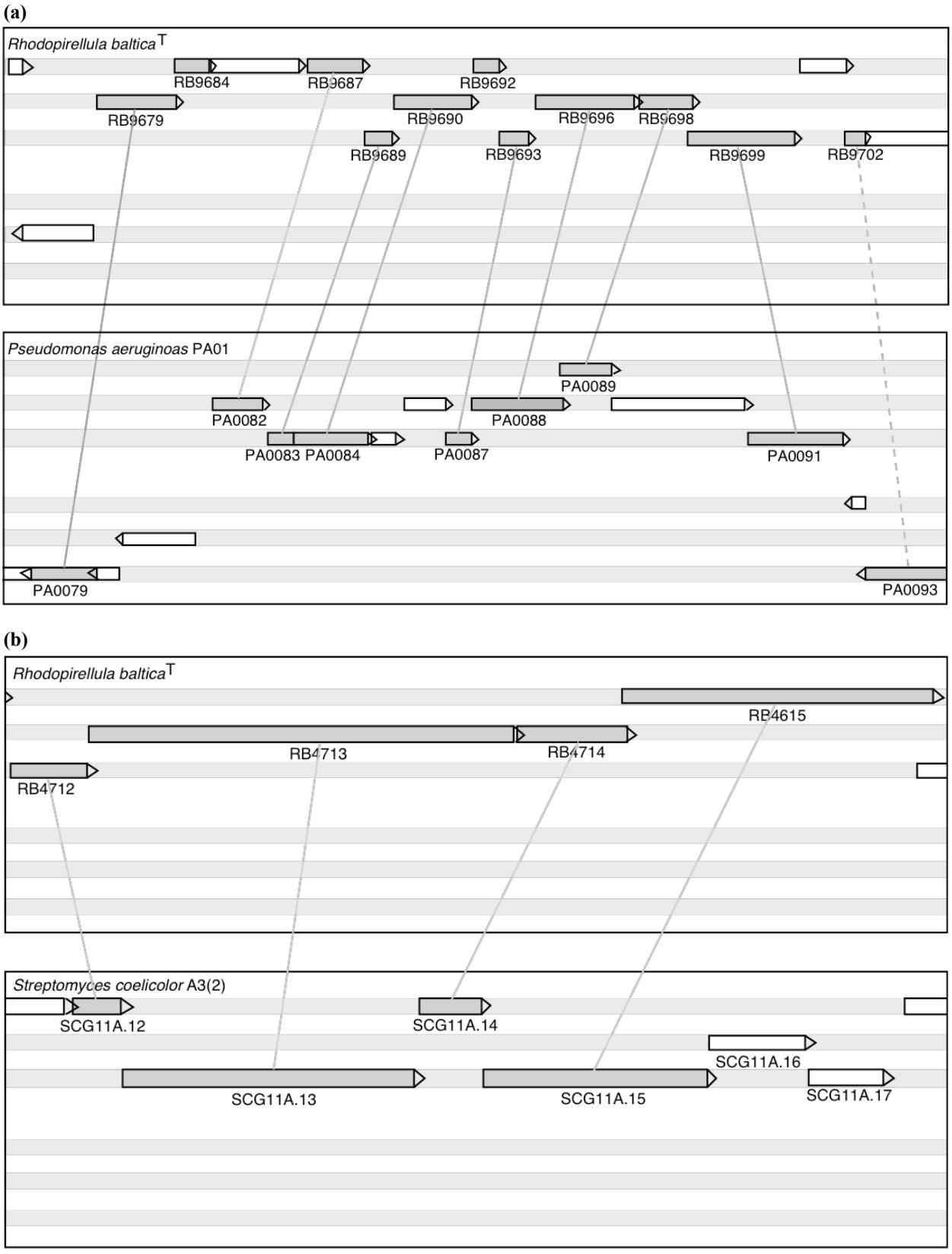


Abb. 10 Beispiele für Bereiche potentiellen lateralen Gentransfers im *Rhodopirellula baltica*^{T}-Genom. (a) Die grau eingezeichneten Gene haben beste BLASTp-Treffer zu Genen in *Pseudomonas aeruginosa* PA01, von denen alle bis auf zwei ähnlich wie in *Pseudomonas aeruginosa* PA01 angeordnet sind. Die Gene haben einen für *Rhodopirellula baltica*^{T} ungewöhnlich hohen GC-Gehalt von im Schnitt 57.1%. Ihre Funktion ist bislang unbekannt. (b) Die grau eingezeichneten Gene haben beste BLASTp-Treffer zu Genen in *Streptomyces coelicolor* A3(2) und sind in beiden Genomen ähnlich angeordnet. Auch in diesem Fall handelt es sich durchweg um Gene unbekannter Funktion.

5.5 Das entkoppelte rDNA-Operon in *Rhodopirellula baltica*^{T}

Bei den meisten Bakterien liegen die Gene für die ribsomale RNA (*rrn*) in Form eines Operons vor, in dem die 16S rDNA (*rrs*) vor der 5S rDNA (*rrf*) und der 23S rDNA (*rrl*) liegt: 5'-16S-23S-5S-3'. *Rhodopirellula baltica*^{T} hat einen einzigen Satz an rDNA-Genen, wobei jedoch lediglich die 5S- und 23S-rDNA ein Operon bilden, während die 16S-rDNA mehr als 460 kB davon entfernt liegt. Solche entkoppelten rDNA-Operons wurden bereits für andere Planktomyceten beschrieben, so z.B. für *Pirellula marina* (Liesack & Stackebrandt, 1989) und *Planctomyces limnophilus* (Menke *et al.*, 1991). Andere Planktomyceten, wie z.B. die Anammox-Bakterien *Broccardia anammoxidans* und *Kuenenia stuttgartiensis*, haben kein entkoppeltes rDNA-Operon (Schmid *et al.*, 2001), und auch in den Genomfragmenten des noch ungeschlossenen Genoms von *Gemmata obscuriglobus* UQM2246^T findet sich von vermuteten fünf rDNA-Operons (Menke *et al.*, 1991) bislang lediglich eines, welches nicht entkoppelt ist.

Außer in den Planktomyceten sind entkoppelte rDNA-Operons auch in einigen anderen Bakterien nachgewiesen worden, darunter *Buchnera*-Spezies (Munson *et al.*, 1993), den Rickettsien (Rurangirwa *et al.*, 2002), sowie den thermophilen Archaeen *Thermoplasma acidophilum* (Tu & Zillig, 1982), *Thermus thermophilus* (Hartmann *et al.*, 1987) und *Thermus aquaticus* (Menke *et al.*, 1991). In einigen Mycoplasmen ist es nicht die 16S-rDNA, sondern die 5S-rDNA, die von den übrigen rDNA-Genen getrennt ist (Chen & Finch, 1989; Taschke *et al.*, 1986), und im Proteobakterium *Beneckea harveyi* liegen die rDNA-Gene zwar beisammen, jedoch in einer unüblichen Reihenfolge (Lamfrom *et al.*, 1978).

Die Tatsache, daß entkoppelte rDNA-Operons auch bei Eukaryonten vorkommen, hat zur Hypothese geführt, daß auseinanderliegende rDNA-Gene den ursprünglichen Zustand in der Progenote dargestellt haben könnten (Clark, 1987). Da in Operons organisierte Gene einfacher zu regulieren sind, und die rDNA-Gene alle im selben Verhältnis benötigt und daher gleichermaßen exprimiert werden müssen, könnte der Zusammenschluß der rDNA-Gene zu einem Operon einen evolutionären Vorteil bedeutet haben. Dieser Zusammenschluß müßte dann jedoch in vielen Entwicklungslinien unabhängig voneinander stattgefunden haben.

5.6 Regulation

Die Regulation von Bakterien bioinformatisch zu untersuchen ist ein äußerst schwieriges Unterfangen. Regulation mittels Operatoren ist mit bioinformatischen Methoden kaum faßbar, da Operatorbindestellen zu kurz und inhomogen sind, um informatisch detektiert werden zu

können. Ähnlich verhält es sich mit der Attenuation. Mit dem Programm Transterm (Ermolaeva *et al.*, 2000) ließen sich 238 *Rho*-unabhängige Terminatoren im *Rhodopirellula baltica*^{T}-Genom finden, mit deren Hilfe einige konsekutive Gene als Operons identifiziert werden konnten. Im Großen und Ganzen scheint die Regulation über Operons in *Rhodopirellula baltica*^{T} jedoch eine wesentlich geringere Rolle zu spielen als z.B. bei *Escherichia coli*. So liegen in *Rhodopirellula baltica*^{T} eine Reihe von Genen vereinzelt vor, die bei *Escherichia coli* zu Operons zusammengefaßt sind. Ein Beispiel hierfür wären die Gene des *pur*-Operons zur Purin-Biosynthese. Die Regulation über Sigma-Faktoren ist von Thierry Lombardot im Rahmen seiner Promotion untersucht worden (Lombardot *et al.*, 2004) und wird daher hier nur der Vollständigkeit halber erwähnt. Eine Analyse der Häufigkeit unterschiedlicher Oligonukleotide im *Rhodopirellula baltica*^{T}-Genom zeigte, daß das 11-mer GTAGGCCAGGT mit 64 Kopien statistisch gesehen mehr als 37-fach überrepräsentiert ist und zudem fast immer zwischen annotierten Genen liegt. Eine regulatorische Funktion dieses 11-mers ist daher wahrscheinlich, muß aber noch bewiesen werden. Neben den erwähnten Mechanismen gibt es die Regulation über den Codegebrauch der Gene, resp. der mRNA. Obwohl aufgrund der Degeneriertheit des genetischen Codes 61 Tripletts zur Kodierung von 21 proteinogenen Aminosäuren zur Verfügung stehen, gibt es solche, die aufgrund der Zusammensetzung des tRNA-Pools effektiver translatiert werden können als andere (Merkl, 2003). Stark exprimierte bzw. mit hoher Effizienz translatierte Gene zeigen daher einen stark optimierten Codegebrauch. Für kürzere Proteine kann der effektive Codegebrauch unter Verwendung einer einfachen Statistik als Nc *codon usage* extrapoliert werden (Wright, 1990). In *Rhodopirellula baltica*^{T} hat z.B. das Gen für die α-Untereinheiten der DNA-abhängigen RNA-Polymerase einen Nc von 29.7, d.h. von 61 möglichen Codons werden lediglich etwa 30 benützt. Ein Histogramm (Abb. 11) der Nc *codon usage* aller im *Rhodopirellula baltica*^{T}-Genom annotierten Gene zeigt, daß *Rhodopirellula baltica*^{T} moderat über den Codegebrauch reguliert (relativ breite Basis der Kurve), ingesamt gesehen aber keinen allzu stark optimierten Codegebrauch hat (Kurvenmaximum bei etwa Nc = 47.5). Die Kurve für *Caulobacter crecentus* CB15 z.B. hat eine vergleichbar breite Basis, deren Maximum jedoch bei Nc = 32 liegt. Der Codegebrauch der Gene ist in *Caulobacter crecentus* CB15 daher sehr viel stärker optimiert als in *Rhodopirellula baltica*^{T}. Entsprechende Histogramme wurden für insgesamt 114 Bakteriengenome angefertigt und verglichen. Dabei zeichneten sich folgende Tendenzen ab: Umweltorganismen, insbesondere phototrophe Bakterien haben oftmals sehr breite, flach verlaufende Kurven. Die Notwendigkeit, flexibel auf wechselnde Umweltbedingungen reagieren zu müssen, spiegelt sich womöglich im variablen Codegebrauch der Gene dieser Organismen wider. Manche parasitische Bakterien, wie. z.B.

Mycoplasma penetrans HF-2, haben Kurven mit sehr enger Basis. Man könnte dies dahingehend interpretieren, daß diese Bakterien in der konstanten Umgebung ihres Wirts leben und ihren Stoffwechsel daher nicht aufwendig regulieren müssen. Allerdings gibt es auch parasitische Bakterien mit anderen Kurvencharakteristika, so daß sich diese Aussage keinesfalls generalisieren läßt. Überraschenderweise haben die Kurven vieler extremophiler Spezies weit rechts liegende Maxima. Dies deutet darauf hin, daß die Selektivität, mit der die tRNA-Moleküle an die mRNA binden, bei sehr hohen oder sehr niedrigen Temperaturen eingeschränkt ist, wodurch eine Regulation über den Codegebrauch nicht besonders effektiv ist. Besonders augenfällig wird dies beim Vergleich des psychrophilen sulfatreduzierenden Bakteriums *Desulfotalea psychrophila* LSv54T mit dem mesophilen Sulfatreduzierer *Desulfovibrio vulgaris* subsp. *vulgaris*. Möglicherweise nützen extremophile Spezies verstärkt andere Regulationsmechanismen.

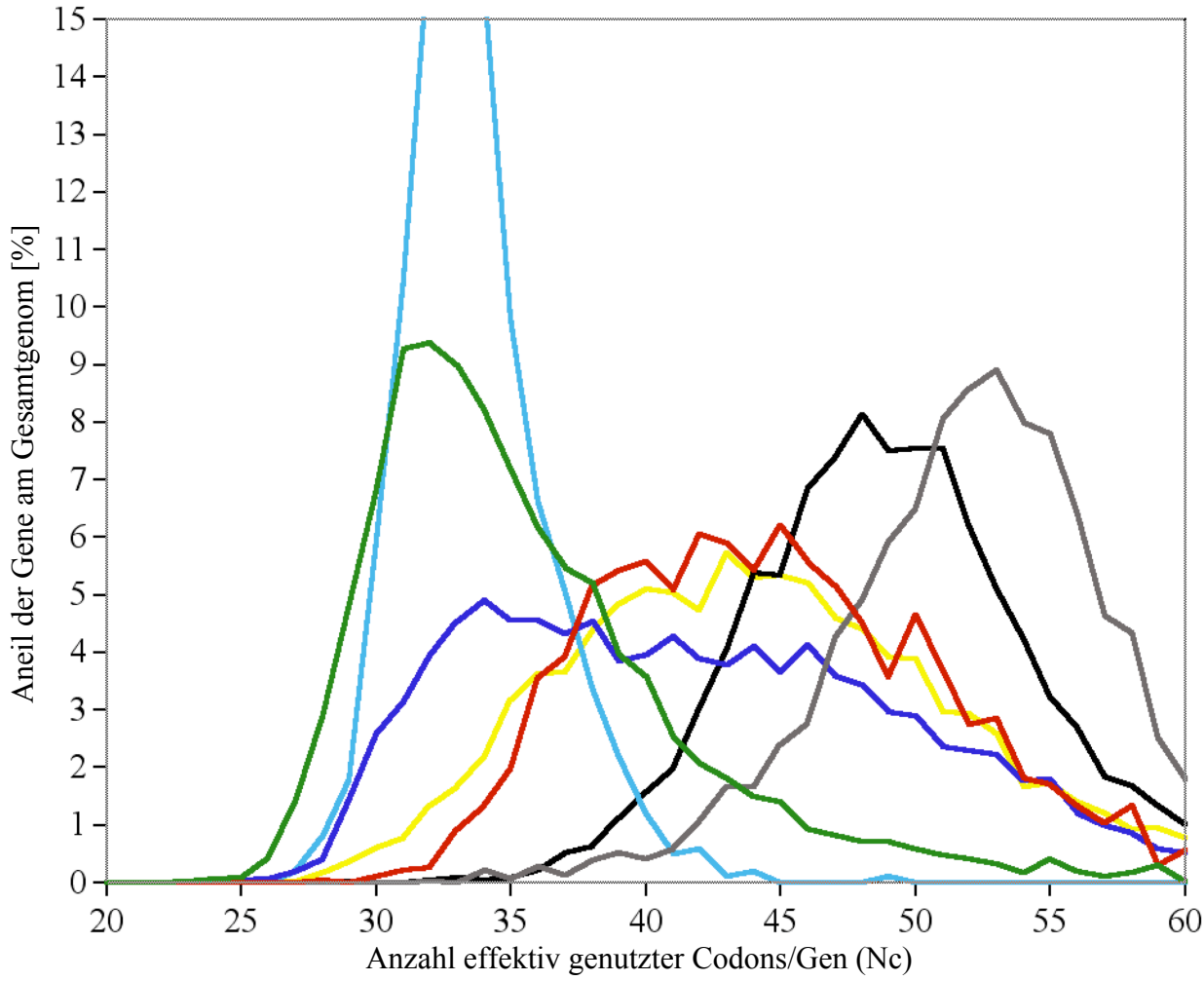


Abb 11 Codegebrauch in ausgewählten Genomen: *Caulobacter cresentus* CB15 (grün), *Chlorobium tepidum* TLS^T (rot), *Desulfotalea psychrophila* LSv54^T (grau), *Desulfovibrio vulgaris* subsp. *vulgaris* (gelb), *Gemmata obscuriglobus* UQM2246^T (blau), *Mycoplasma penetrans* HF-2 (cyan), *Rhodopirellula baltica*^{T} (schwarz). Für jedes Genom wurde die effektive Anzahl der in jedem Gen verwendeten Codons nach der Nc-Methode berechnet (Wright, 1990). Hierzu wurde das Programm CHIPS aus dem EMBOSS-Programmpaket verwendet (Olson, 2002; Rice *et al.*, 2000). Die Ergebnisse wurden für jedes Genom in Form von Histogrammen dargestellt.

6. Die Ökologie

Insgesamt gesehen läßt sich anhand der Genomannotation ein recht klares Bild der ökologischen Rolle von *Rhodopirellula baltica*^{T} rekonstruieren. *Rhodopirellula baltica*^{T} ist auf den Abbau häufiger mariner Mono- und Polysaccharide spezialisiert. Darunter fallen das Chitin-Monomere N-Acetylglucosamin ebenso wie sulfatisierte Polysaccharide aus Algen (z.B. Alginate, Carragen) und Fischknorpel (z.B. Chondroitin-sulfat). Adulte Zellen sind sessil und heften sich an Partikel in der oxischen Zone des Sediments oder aber an organischen Detritus an. Geraten sie unter sauerstofflimitierende Bedingungen, so überstehen sie diese durch die Expression einer hochsauerstoffaffinen d-Typ Cytochromoxidase. Anoxische Bedingungen können möglicherweise mittels der heterofermentativen Milchsäuregärung überdauert werden, und bei Substratmangel werden zur Vermeidung von Zellschäden Schutzproteine (z.B. *carbon starvation proteins* und *heat shock*-Proteine) exprimiert. Neue Lebensräume und Resourcen, werden mit Hilfe monotrich begeißelter Tochterzellen erschlossen. Diese verlieren ihre Geißel nach einiger Zeit des Umherschwimmens und kehren wieder zur sessilen Lebensweise zurück. Geraten die Zellen in

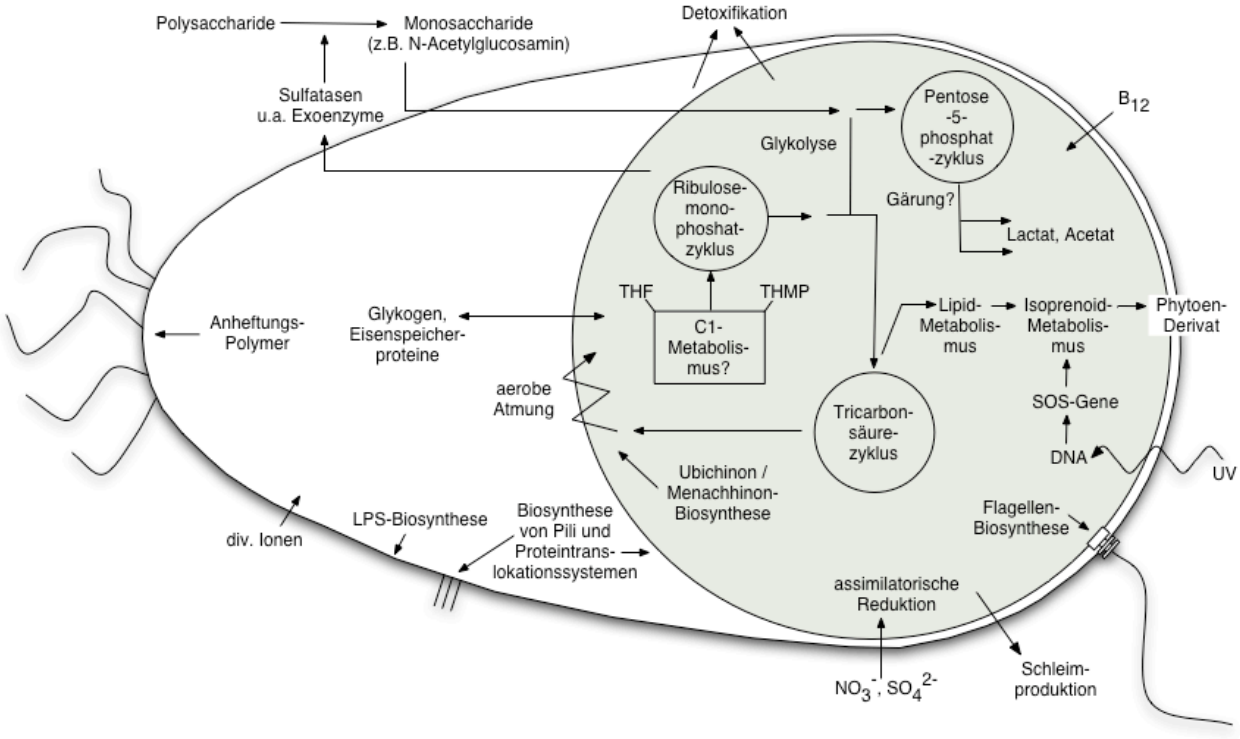


Abb 12 Ausgewählte Aspekte aus der Annotation des *Rhodopirellula baltica*^{T}-Genoms (schematisch). Die Lokalisation der einzelnen Prozesse sowie die Art der Flagelleninsertion sind rein hypothetisch. Da sich die Ribosomen mutmaßlich nur im Riboplasma befinden, können Proteine nur dort synthetisiert werden. Die Auslagerung von Speicherstoffen in das Paryphoplasma erscheint in Anbetracht der dort sichtbaren Einschlüsse (Schlesner *et. al*, 2004) wahrscheinlich. Nicht gezeigt sind die Biosynthesen der Aminosäuren und Cofaktoren, der Nukleotidstoffwechsel, die Transkription, Translation und Replikation und vieles mehr. LPS = Lipopolysaccharid; UV = ultra-violett; THF = Tetrahydrofolat; THMP = Tetrahydromethanopterin

Zonen starker Sonneneinstrahlung, so schützen sie sich durch ein Carotinoid vor der UV-Strahlung. Insgesamt gesehen scheint *Rhodopirellula baltica*^{T} über eine breite Zahl unterschiedlicher Regulationsmechanismen zu verfügen, die es dem Organismus erlauben, flexibel auf wechselnde Umweltbedingungen zu reagieren.

7. Die Phylogenie

Die Phylogenie der Planktomyceten ist, zahlreicher Studien zum Trotz, bis heute weitgehend ungeklärt. Es besteht zwar Einigkeit darüber, daß die Planktomyceten ein eigenes Phylum innerhalb der bakteriellen Domäne bilden, die Position dieses Phylums konnte jedoch bislang nicht eindeutig geklärt werden und war gerade in jüngster Zeit Gegenstand einer heftig ausgetragenen Debatte (Brochier & Philippe, 2002; Di Giulio, 2003).

Frühe phylogenetische Analysen auf der Basis von 16S rRNA-Sequenzen legten eine entfernte Verwandtschaft der Planktomyceten mit den Chlamydien nahe (Liesack & Stackebrandt, 1992; Weisburg *et al.*, 1986). Diese konnte jedoch in späteren phylogenetischen Analysen auf der Basis von Sequenzen der 16S- und 23S-rRNA (Ward *et al.*, 2000), des *dnaK*-Gens (Ward-Rainey *et al.*, 1997) und des *tuf*-Gens (Jenkins & Fuerst, 2001) nicht bestätigt werden. Die bislang von Planktomyceten bekannten 16S rRNA-Sequenzen weisen ein hohes Maß an Heterogenität auf. Dies wurde sowohl als Indiz für eine hohe Evolutionsgeschwindigkeit der Planktomyceten (tachytelische DNA) und somit für eine hohe interne Radiation gewertet (Bomar *et al.*, 1988; Liesack & Stackebrandt, 1992; Woese, 1987), als auch als Indiz für ein hohes Alter der gesamten Entwicklungslinie (Stackebrandt *et al.*, 1984). In zwei aktuelleren phylogenetischen Studien, in denen ausschließlich die langsam evolvierenden Positionen der 16S rRNA zur phylogenetischen Analyse benützt wurden, sind die Planktomyceten als das am tiefsten abzweigende Phylum (Brochier & Philippe, 2002) oder zumindest als tief abzweigendes Phylum (Di Giulio, 2003) innerhalb der bakteriellen Domäne beschrieben worden. Ersteres würde nicht nur bedeuten, daß die Planktomyceten die entwicklungsgeschichtlich älteste unter den bekannten bakteriellen Entwicklungslinien ist, sondern gleichzeitig auch der Hypothese widersprechen, daß die Bakterien auf einen thermophilen Vorläufer zurückgehen (Brochier & Philippe, 2002).

Die Tatsache, daß die phylogenetische Position der Planktomyceten bislang nicht eindeutig geklärt werden konnte, liegt zum Teil in den bisherigen Studien selbst begründet. So ist die 16S rRNA nach der 23S rRNA zwar das am besten geeignete Einzelgen für phylogenetische Studien, doch ihr Informationsgehalt ist zu gering, um die Reihenfolge der einzelnen Phyla

innerhalb der bakteriellen Domäne klar auflösen zu können (Ludwig *et al.*, 1998). Lediglich die beiden thermophilen Bakterien *Thermotoga maritima* MSB8^T und *Aquifex aeolicus* VF5 lassen sich als tief abzweigende Organismen deutlich gegenüber allen anderen Bakterien abgrenzen. Die übrigen Phyla liegen jedoch dicht beisammen und sind im Hinblick auf ihre Reihenfolge nur schlecht aufgelöst. Die Position von *Aquifex aeolicus* VF5 als tief abzweigendes Bakterium ist im Übrigen nicht unumstritten, da eine Reihe von Proteinphylogenien eher eine Verwandtschaft mit den Proteobakterien nahelegen. Ob eine Beschränkung der phylogenetischen Analyse auf die langsam evolvierenden Positionen der 16S rRNA tatsächlich, wie behauptet, eine Erhöhung der Auflösung für tief abzweigenden Phyla mit sich bringt, darf aufgrund des von vornherein begrenzten Informationsgehalts der 16S-rRNA bezweifelt werden. Die beiden oben erwähnten Studien haben jedenfalls gezeigt, daß die Ergebnisse bei dieser Methode in hohem Maße von den jeweils zur phylogenetischen Analyse ausgewählten Positionen abhängen. Auch das *tuf*-Gen, das den Elongationsfaktor TU kodiert, ist nur bedingt als phylogenetischer Marker geeignet. Der Informationsgehalt des *tuf*-Gens ist nicht nur aufgrund seiner geringen Länge begrenzt, sondern auch aufgrund seiner starken Konserviertheit. Dieser ist so stark, daß sich *tuf*-Gene (bzw. ihre Analoga) von Eukaryonten, Archaeen und Eubakterien problemlos alignieren lassen. Hinzu kommt, daß es zwei Paraloge des *tuf*-Genes gibt. In der oben genannten Studie wurde zudem nur eine *tuf*-Vollsequenz eines Planktomyceten (*Pirellula marina*) zur phylogenetischen Analyse verwendet. Noch weniger als das *tuf*-Gen ist das *dnaK*-Gen zur phylogenetische Analyse geeignet, welches für das 70 kD *heat shock*-Protein kodiert (Hsp70). Allein im *Rhodopirellula baltica*^{T}-Genom gibt es vier Paraloge dieses Gens, die sich so stark in ihren Sequenzen unterscheiden, daß eine sinnvolle phylogenetische Analyse praktisch unmöglich ist. Auch im noch unvollständigen Genom von *Gemmata obscuriglobus* UQM2246^T finden sich mindestens vier Paraloge des *tuf*-Gens. Die erwähnte Studie auf der Basis von 16S- und 23S-rRNA-Sequenzen schließlich krankt an der Tatsache, daß die publizierten Bäume nur mit der *neighbor-joining* Methode berechnet wurden (d.h., es wurde kein *likelihood*-basiertes Verfahren eingesetzt) und nur eine geringe Anzahl von Vergleichsspezies enthalten waren.

Während sich die bisherigen phylogenetischen Unterschungen von Planktomyceten durchweg auf Einzelgene stützten (typischerweise macht die 16S rDNA-Sequenz weniger als 0,2% der Sequenzinformation eines bakteriellen Genoms aus), steht mit dem Genom von *Rhodopirellula baltica*^{T} erstmals die Möglichkeit zur Verfügung, viele Gene oder gar die Informationsfülle des ganzen Genoms zu nutzen. Zwar ist die Auswahl an sinnvollen evolutionären Markern begrenzt (Harris *et al.*, 2003; Ludwig *et al.*, 1998), doch mit der Konkatenierung von ribosomalen Proteinen sowie von Untereinheiten der DNA-abhängigen

RNA-Polymerase stehen erprobte Verfahren zur Verfügung, bei denen ein größeres Maß an Informationen genutzt wird als bei den bislang angewandten Verfahren (Bocchetta *et al.*, 2000; Brochier *et al.*, 2002; Matte-Tailliez *et al.*, 2002). Neben der Verwendung von Markergenen gibt es die phylogenomischen Verfahren, bei denen nahezu die gesamte genomische Information in die Analyse einbezogen wird (Wolf *et al.*, 2001). Dazu gehören z.B. Verfahren, die auf der An- und Abwesenheit (Clarke *et al.*, 2002) oder der Anordnung von Genen beruhen (Wolf *et al.*, 2002) oder aber Verfahren, die auf der Basenzusammensetzung der Gene basieren (Pride *et al.*, 2003).

Eine auf normalisierten BLASTp-Treffern beruhende phylogenomische Analyse konnte die Phylogenie der Planktomyceten jedoch nicht klären (durchgeführt von Thierry Lombardot im Rahmen seiner Promotion), widerspricht aber der Hypothese, daß die Planktomyceten das am tiefsten abzweigende bakterielle Phylum seien (Publikation 2, Abb. 2). Die phylogenetische Analyse konkatenierter ribosomaler Proteine und konkatenierter Untereinheiten der DNA-abhängigen RNA-Polymerase von *Rhodopirellula baltica*^{T} und *Gemmata obscuriglobus* UQM2246^T hingegen zeigte eine deutliche Verwandtschaft der Planktomyceten mit den Chlamydien (Publikation 2, Abb. 1) oder wenigstens mit dem jüngst vorgeschlagenen Spirochaeten-Chlamydien-Supercluster (Clarke *et al.*, 2002; Wolf *et al.*, 2002; Wolf *et al.*, 2001). Die Details dieser Analysen sind in Publikation 2 dargelegt und ausführlich diskutiert. Sie sollen daher an dieser Stelle nicht noch einmal wiederholt werden.

Interessanterweise wird eine Verwandtschaft der Planktomyceten durch eine Reihe weiterer Indizien gestützt. So plaziert ein von Gupta (2002) vorgeschlagenes, auf Indels basierendes phylogenetisches System *Rhodopirellula baltica*^{T} ebenfalls in die Nähe der Spirochaeten und Chlamydien (Abb 13). Indels sind Insertionen und Deletionen in Schlüssel-proteinen, welche als spezifische Marker für bestimmte Phyla angesehen werden (Gupta, 2001; Gupta & Griffiths, 2002). Ob die lineare Natur des von Gupta vorgeschlagenen Indel-Systems sinnvoll ist, und ob dieses nur anhand von wenigen Genomen etablierte System universell anwendbar ist, darf allerdings bezweifelt werden.

Eine hier nicht gezeigte phylogenetische Analyse verschiedener Kombinationen konkatenierter ATPase-Untereinheiten zeigte, daß die zu *Pirellula marina* verwandten ATPase-Gene aus *Rhodopirellula baltica*^{T} (Pkt. 5.1.4) und die ATPase-Gene aus *Gemmata obscuriglobus* UQM2246^T mit dem einzigen Spirochaeten in der Analyse, *Leptospira interrogans* serovar *lai* str. 56601, clusterten. Chlamydien haben wegen eines vermuteten frühen LGT von Archaeen keine ATPasen von F1F0-Typ, sondern solche vom V-Typ. Das Clustern mit einem Spiro-chaeten in der phylogenetischen Analyse deutet jedoch eine Verwandtschaft der Plankto-

myceten mit dem Spirochaeten-Chlamydien-Supercluster an und widerspricht folglich einem besonders tiefen Abzweigen des Planctomyceten-Phylums.

Einen weiteren Hinweis stellt die phylogenetische Analyse des *recA*-Gens dar. In einer hier nicht gezeigten *neighbor-joining*-Analyse clusterten die beiden Planktomyceten mit hoher statistischer Unterstützung (*bootstrap:* 95) mit den Chlamydien. Abgesehen von den phylogenetischen Analysen gibt es eine Reihe von Parallelen zwischen den Planktomyceten und den Chlamydien. Die Organismen beider Phyla haben Zellwände aus Proteinen - die Planktomyceten permanent und die Chlamydien während ihres Infektionsstadiums (Hatch, 1996). In beiden Gruppen scheint es sich um eine sekundäre Adaption zu handeln, denn die Chlamydien verfügen über alle (Ghuysen & Goffin, 1999; Stephens *et al.*, 1998) und die Planktomyceten über einige der Gene zur Peptidoglyan-Biosynthese. Verbunden mit der Proteinzellwand ist vermutlich das Fehlen von *ftsZ* in beiden Gruppen, was auf einen noch unbekannten Modus der Zellteilung hindeutet. Außerdem kommen unter allen bislang totalsequenzierten Bakterien lediglich in den Chlamydien sowie in den Planktomyceten *Rhodopirellula baltica*^{T} und *Gemmata obscuriglobus* UQM2246^{T} zwei Kopien des Gens für das Replikationsinitiationsprotein DnaA vor. Im übrigen durchlaufen sowohl die Planktomyceten als auch die Chlamydien komplexe Zellzyklen. Hinzu

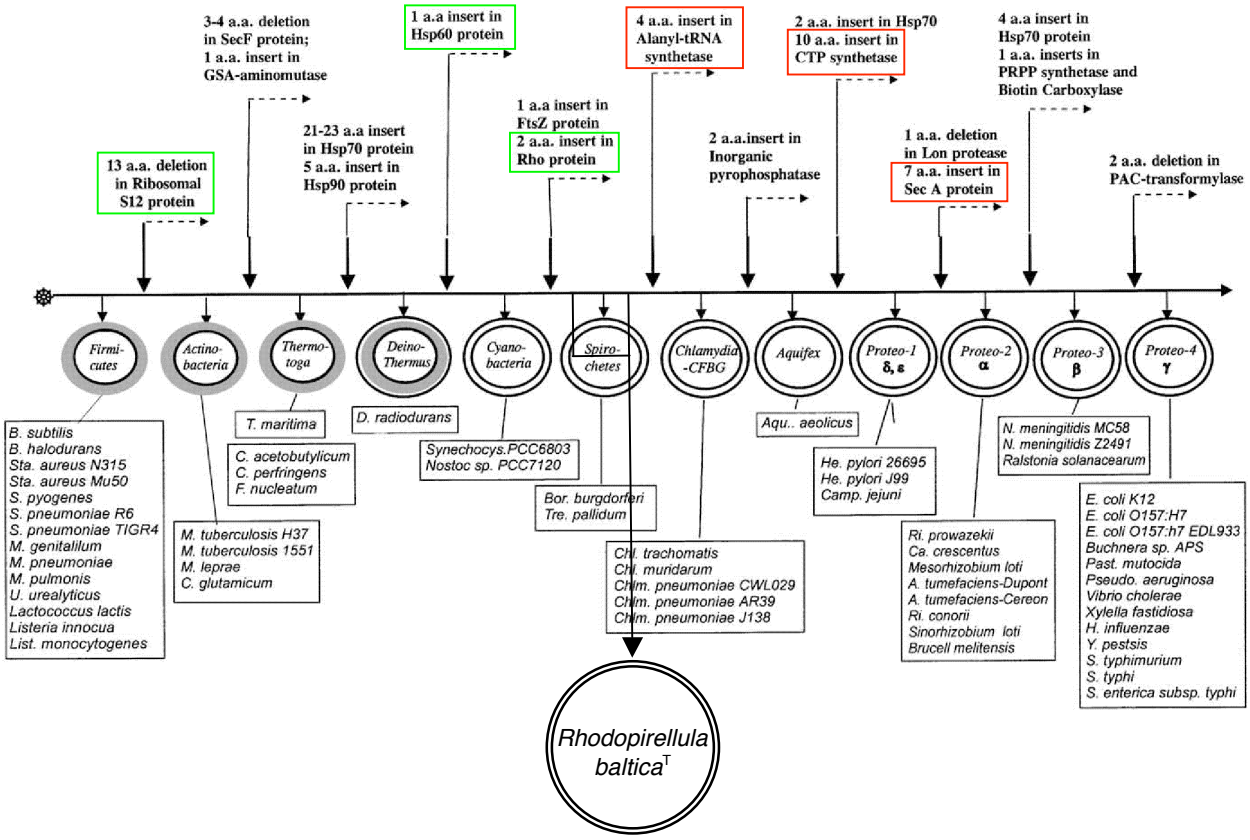


Abb. 13 Phylogenetische Position von *Rhodopirellula baltica*^{T} nach der Indel-Methode (Gupta, 2001; Gupta & Griffiths, 2002). Die im *Rhodopirellula baltica*^{T}-Genom vorhandenen Indels sind grün, die fehlenden rot gekennzeichnet. Die übrigen Marker ließen sich nicht verwenden: *SecF* ist im *Rhodopirellula baltica*^{T}-Genom mit *secD* fusioniert, Hsp70 und *hemL* (GSA-Aminomutase) haben Paraloge, Hsp90, *ftsZ*, die anorganische Pyrophosphatase und die *lon*-Protease fehlen. Abbildung nach Gupta (2002).

kommt, daß die DNA sowohl der Chlamydien (Hatch, 1996) als auch der Planktomyceten elektronenmikroskopisch als Nukleoid sichtbar ist. Außerdem kommen sowohl bei den Chlamydien (Karunakaran *et al.*, 2003) als auch in *Rhodopirellula baltica*^{T} drei Kopien von *groEL*-ähnlichen Genen vor, die in allen anderen Phyla fehlen. Schlußendlich weist die 16S rRNA in den Chlamydien und den Planktomyceten Signatursequenzen auf, die nur diesen beiden Gruppen zu eigen ist (Fuerst, 1995).

Zusammengenommen sind die Hinweise, die sich aus der Annotation des *Rhodopirellula baltica*^{T}-Genoms für eine Verwandtschaft der Planktomyceten mit den Chlamydien ergaben, konsistent, wohingegen keine Hinweise auf ein tiefes Abzweigen der Planktomyceten gefunden werden konnten.

8. Tetranukleotide als Identifikationsmerkmal in der Metagenomik

Wie molekularbiologische Studien gezeigt haben, lassen sich mit den heutigen Kultivierungstechniken nur weniger als etwa ein Prozent der Mikroorganismen aus der Umwelt in Reinkultur bringen (Amann *et al.*, 1995). Nur diese lassen sich jedoch gezielt physiologisch untersuchen, so daß das Wissen über Mikroorganismen an ihren natürlichen Standorten begrenzt ist. Einen zumindest teilweisen Ausweg aus diesem Dilemma bietet die Metagenomik. Bei diesem Verfahren werden Genomfragmente aus Umweltproben extrahiert und anschließend mit Hilfe von Vektoren, wie z.B. Fosmiden oder BACs (*bacterial artificial chromosomes*) kloniert. Die auf diese Weise erstellten Genbanken lassen sich nach spezifischen Genen oder aber nach bestimmten Stoffwechselleistungen durchsuchen.

Das Grundproblem des Metagenom-Verfahrens ist, daß sich mit Screening-Methoden zwar Genomfragmente mit den jeweils gesuchten Genen identifizieren lassen, diese aber nach erfolgter Sequenzierung oftmals nicht einem bestimmten Organismus oder einer Organismengruppe zugeordnet werden können. So hat ein typisches Fosmid-Genomfragmente nur etwa ein Länge von etwa 40 kb und trägt somit etwa 40 Gene. Von diesen lassen sich beim derzeitigen Stand des Wissens etwa die Hälfte funktionell annotieren, und nur auf etwa 5-10% der Genomfragmente finden sich geeignete phylogenetische Markergene. Wenn also geklärt werden soll, ob zwei nicht überlappende Fragmente vom gleichen Organismus stammen, dann ist es wahrscheinlich, daß mindestens eines der beiden Fragmente keinen geeigneten phylogenetischen Marker trägt. In solchen Fällen kann man die besten BLAST-Treffer der annotierbaren Gene, ihre *codon usage* und den G+C-Gehalt der Fragmente vergleichen. Diese Verfahren sind jedoch nur bedingt dazu geeigent, Genomfragmente einander zuzuordnen (Publikation 3).

Ein alternatives, rein sequenzbasiertes Verfahren konnte im Rahmen der phylogenomischen Untersuchungen des *Rhodopirellula baltica*^{T}-Genoms entwickelt werden. Wie in zahlreichen Studien gezeigt werden konnte, sind die Häufigkeiten kurzer Oligonukleotide in Genomen Spezies-spezifisch und tragen zudem ein phylogenetisches Signal. So wurden unterschiedliche Tetranukleotid-Häufigkeiten in den Genen ganzer Genome bereits als Basis für ein phylogenomisches Verfahren eingesetzt (Pride *et al.*, 2003). In eigenen Untersuchungen zeigte sich jedoch, daß dieses Verfahren trotz Verbesserungen unter anderem des zugrundeliegenden Markov-Modells eine unzureichende Auflösung besitzt, wenn man andere als die von den Autoren zur Validierung des Verfahrens verwendeten Genome einsetzt. In abgewandelter Form läßt sich das Verfahren jedoch als sequenzbasierte Zuordnungsmethode für Metagenomfragmente einsetzen. Die Diskriminierung zwischen zusammengehörigen und nicht-zusammengehörigen Genomfragmenten, also die Selektivität des Tetranukleotid-Verfahrens, ist dabei sehr viel höher als die des (ebenfalls sequenzbasierten) G+C-Gehalts.

Die Details des Verfahrens sowie eine vergleichende Evaluierung mit dem G+C-Gehalt sind in Publikation 3 ausführlich erörtert und diskutiert, weswegen sie an dieser Stelle nicht noch einmal wiederholt werden sollen. Stattdessen soll das in der Publikation nicht erwähnte Programm TETRA kurz vorgestellt werden (Abb. 14).

TETRA erlaubt es, statistische Analysen der Tetranukleotidverteilung einer beliebigen Zahl von DNA-Sequenzen durchzuführen. Die Ergebnisse der Einzelanalysen lassen sich in Tabellenform exportieren und können zur weiteren Auswertung z.B. in Microsoft-Excel reimportiert werden. Zusätzlich können die Tetranukleotidverteilungen gegeneinander geplottet und ihre Korrelationen berechnet werden. Außerdem lassen sich die Plots auf die Über- oder Unterrepräsentation spezifischer Tetranukleotide hin untersuchen, wodurch z.B. Hinweise auf Restriktionsschnittstellen erhalten werden können. Als zusätzliches Feature stehen eine statistische Auswertung der Basenzusammensetzungen von Sequenzen sowie einfache Sequenzeditiermöglichkeiten zur Verfügung. Eine Suchfunktion mit regulären Ausdrücken ist ebenfalls implementiert.

TETRA befindet sich derzeit in einem späten Betastadium und steht kurz vor der ersten Release. Das Programm wurde in Realbasic5 (REAL Software, Inc., Austin, Texas, USA; www.realsoftware.com) implementiert, einer objektorientierten Hochsprache, die eine plattformübergreifende Programmentwicklung ermöglicht. Die derzeitige TETRA-Version ist noch auf das Unix-basierte MacOS X beschränkt, läßt sich jedoch mit vergleichsweise geringem Aufwand für Windows-Systeme und Linux anpassen.

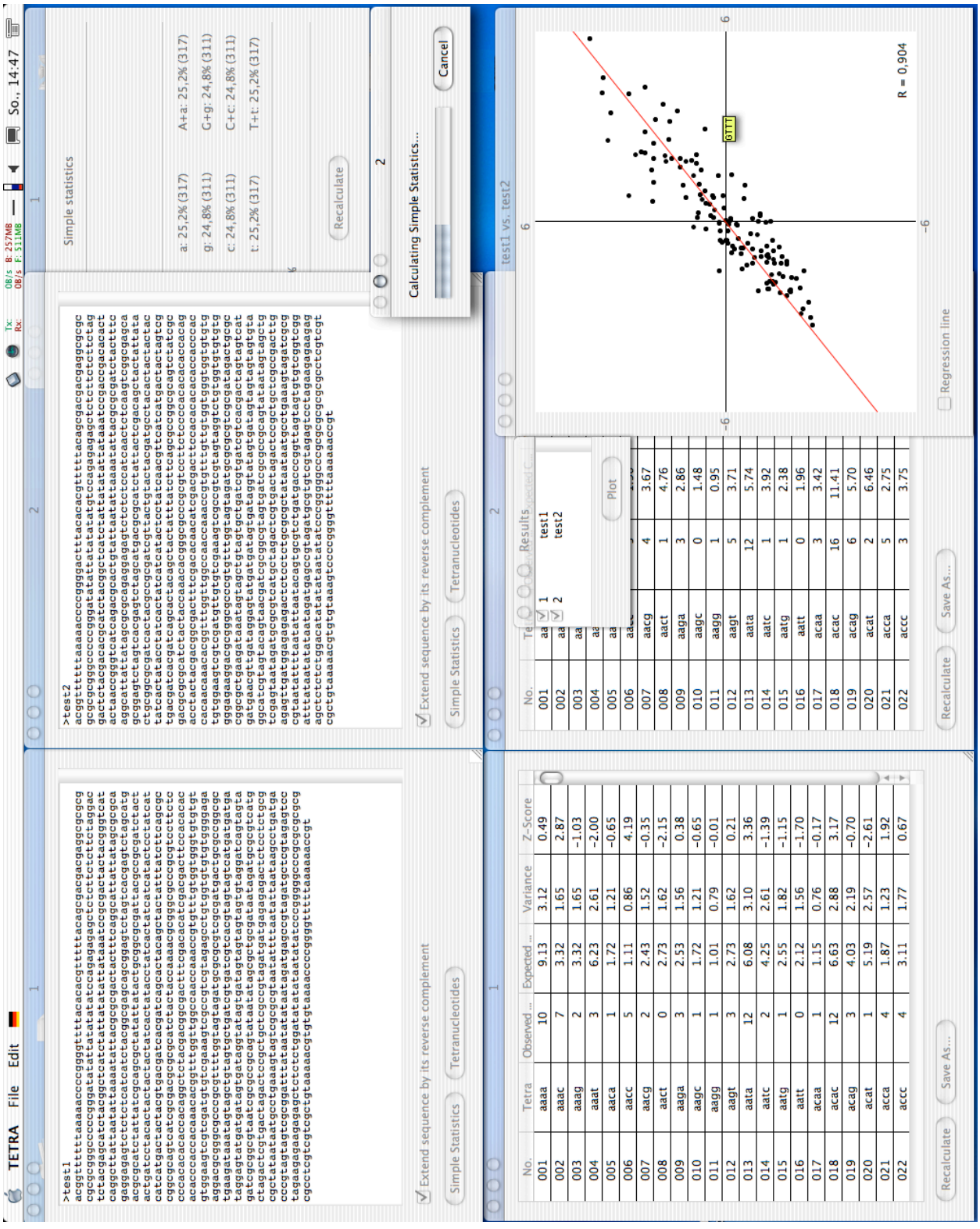


Abb. 14 *Screenshot* des Programms TETRA (v. 1.0b5). Mit TETRA können Tetranukleotid-Analysen beliebig vieler Sequenzen durchgeführt und zur Auswertung gegeneinander geplottet werden. TETRA wurde in REALbasic 5 (einer objektorientierten Hochsprache) implementiert und ermöglicht aufgrund seiner *multithreading*-Programmierung das zeitgleiche Durchführen mehrerer Berechnungen. Neben der Tetranukleotid-Analyse können auch Basenzusammensetzungen von Sequenzen und graphische Tetranukleotid-Musteranalysen durchgeführt werden.

9. Ausblick

Nach dem Aufbau einer funktionsfähigen bioinformatischen Arbeitspipeline und der erfolgreichen Annotation des *Rhodopirellula baltica*^{T}-Genoms wird sich der Arbeitsschwerpunkt der Arbeitsgruppe Genomik auf die Auswertung genomischer Daten mariner Bakterien konzentrieren. So sollen Metagenom-Daten aus marinen Habitaten ausgewertet, und mit *Cytophaga* sp. KT0803 mindestens ein weiteres marines Bakterium totalsequenziert und annotiert werden. Außerdem soll die Auswertung des *Rhodopirellula baltica*^{T}-Genoms fortgeführt werden. Die Tatsache, daß für nur etwa ein Drittel der Gene verläßliche Funktionsvorhersagen getroffen werden konnten, ist gleichbedeutend damit, daß das Gros der aus dem Genom zu gewinnenden Informationen noch brachliegt. Um die Annotation auf dem neuesten Stand des Wissens zu halten, wird es nötig sein, funktionell nicht charakterisierte Gene in regelmäßigen Abständen mit aktuellen Datenbanken abzugleichen und die Ergebnisse auszuwerten. Außerdem sollen Einzelaspekte des Genoms gezielt untersucht werden. Dazu gehört die Untersuchung des mRNA-Pools und seine Regulation mit Hilfe von DNA-Chips sämtlicher Gene (*whole genome chips*), welche im Rahmen des EU-Projekts "Marine Genomics" in einer Kooperation des Max-Planck-Instituts für marine Mikrobiologie mit der Universität Bielefeld erstellt werden sollen. Darüber hinaus sollen die Analysen des Proteoms von *Rhodopirellula baltica*^{T} fortgesetzt werden, die von Dr. Ralf Rabus und Dörte Gade am Max-Planck-Institut für marine Mikrobiologie durchgeführt wurden (Rabus *et al.*, 2002). An der Universität Greifswald schließlich sollen die von *Rhodopirellula baltica*^{T} sezernierten Proteine, das sog. Sekretom, mit den Mitteln der Proteomik untersucht werden. Diese Versuche haben zumindest das Potential, um erste Erkenntnisse über die Funktion des Pirellulosoms zu erhalten.

Zusätzliche Erkenntnisse über Gene bislang unbekannter Funktion versprechen auch Genomvergleiche von Planktomyceten. Diese werden möglich, sobald die Genome der derzeit noch in der Sequenzierungs- und Assemblierungsphase befindlichen Planktomyceten-Genome zur Verfügung stehen.

Da Genomvergleiche überaus rechenintensiv sind, wurde die Anschaffung eines kleineren Rechnerclusters beschlossen. Dieser soll Anfang 2004 in Betrieb genommen werden, was eine Reihe von softwareseitigen Anpassungen erforderlich machen wird.

Neben der Anpassung bereits existierender Software stellt die Entwicklung eigener Software ein weiteres zukünftiges Arbeitsfeld dar. Dazu gehört die Weiterentwicklung des ORF-Finders MORFind, bei dem ein Ersatz von ORPHEUS durch den rein intrinsischen ZCURVE-

Algorithmus (Guo *et al.*, 2003) angedacht ist. Außerdem soll die Software TETRA um Funktionalitäten erweitert und in Form einer *application note* veröffentlicht werden.

Bislang diente die Bioinformatik primär der Generierung, Auswertung und Speicherung relativ einfacher Daten. In Zukunft wird jedoch die Verknüpfung einer Vielzahl verschiedener Daten mit Hilfe von wissensbasierten Systemen und mit Methoden aus der KI-Forschung helfen, Wissen zu akquirieren und Hypothesen zu generieren, die sich einer rein manuellen Auswertungen entziehen. Diese Tendenz zeichnet sich auf dem Gebiet der vergleichenden Genomik und der semiautomatischen Annotation bereits heute ab. Bioinformatische Systeme werden infolgedessen zunehmend komplex und schwerer zu managen. Insofern stellt bioinformatisches know-how heute und in Zukunft eine der wichtigen Kernkompetenzen in der Genomik dar.

Ansonsten schreitet die Entwicklung der Bioinformatik mit einem derart rasanten Tempo voran, daß sich langfristige Entwicklungen nur schwer absehen lassen. Mit an Sicherheit grenzender Wahrscheinlichkeit wird es in zehn oder zwanzig Jahren interessante Anwendungsmöglichkeiten geben, die heute niemand auch nur erahnt.

C Literatur

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403-410.

Amann, R. I., Ludwig, W. & Schleifer, K. H. (1995). Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol Rev* **59**, 143-169.

Angerer, A., Gaisser, S. & Braun, V. (1990). Nucleotide sequences of the *sfuA*, *sfuB*, and *sfuC* genes of *Serratia marcescens* suggest a periplasmic-binding-protein-dependent iron transport mechanism. *J Bacteriol* **172**, 572-578.

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M., Servant, F., Sigrist, C. J. & Zdobnov, E. M. (2000). InterPro--an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16**, 1145-1150.

Badger, J. H. & Olsen, G. J. (1999). CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* **16**, 512-524.

Bailey, K. A., Pereira, S. L., Widom, J. & Reeve, J. N. (2000). Archaeal histone selection of nucleosome positioning sequences and the procaryotic origin of histone-dependent genome evolution. *J Mol Biol* **303**, 25-34.

Bairoch, A. (1991). PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res* **19** Suppl, 2241-2245.

Barton, G. J. (1994). Scop: structural classification of proteins. *Trends Biochem Sci* **19**, 554-555.

Bauer, M., Lombardot, T., Teeling, H., Amann, R., Ward, N. L. & Glöckner, F. O. (2004). Archaeal-like genes for C1-transfer enzymes in *Planctomycetes*: phylogenetic implications of their unexpected presence in this phylum. *J Mol Evol*, under revision.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242.

Besemer, J., Lomsadze, A. & Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**, 2607-2618.

Bocchetta, M., Gribaldo, S., Sanangelantoni, A. & Cammarano, P. (2000). Phylogenetic depth of the bacterial genera *Aquifex* and *Thermotoga* inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. *J Mol Evol* **50**, 366-380.

Bomar, D., Giovannoni, S. & Stackebrandt, E. (1988). A unique type of eubacterial 5S rRNA in members of the order *Planctomycetales*. *J Mol Evol* **27**, 121-125.

Brassinga, A. K., Siam, R., McSween, W., Winkler, H., Wood, D. & Marczynski, G. T. (2002). Conserved response regulator CtrA and IHF binding sites in the *alpha-proteobacteria Caulobacter crescentus* and *Rickettsia prowazekii* chromosomal replication origins. *J Bacteriol* **184**, 5789-5799.

Brochier, C. & Philippe, H. (2002). Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature* **417**, 244.

Brochier, C., Bapteste, E., Moreira, D. & Philippe, H. (2002). Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet* **18**, 1-5.

Bruck, I. & O'Donnell, M. (2000). The DNA replication machine of a gram-positive organism. *J Biol Chem* **275**, 28971-28983.

Brysch, K., Schneider, C., Fuchs, G. & Widdel, F. (1987). Lithoautotrophic growth of sulfate-reducing bacteria, and description of *Desulfobacterium autotrophicum* gen. nov., sp. nov. *Arch Microbiol* **148**, 264-274.

Chain, P., Lamerdin, J., Larimer, F., Regala, W., Lao, V., Land, M., Hauser, L., Hooper, A., Klotz, M., Norton, J., Sayavedra-Soto, L., Arciero, D., Hommes, N., Whittaker, M. & Arp, D. (2003). Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph Nitrosomonas europaea. *J Bacteriol* **185**, 2759-2773.

Chen, X. & Finch, L. R. (1989). Novel arrangement of rRNA genes in *Mycoplasma gallisepticum*: separation of the 16S gene of one set from the 23S and 5S genes. *J Bacteriol* **171**, 2876-2878.

Clark, C. G. (1987). On the evolution of ribosomal RNA. *J Mol Evol* **25**, 343-350.

Clarke, G. D., Beiko, R. G., Ragan, M. A. & Charlebois, R. L. (2002). Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J Bacteriol* **184**, 2072-2080.

Couture, M., Yeh, S. R., Wittenberg, B. A., Wittenberg, J. B., Ouellet, Y., Rousseau, D. L. & Guertin, M. (1999). A cooperative oxygen-binding hemoglobin from *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A* **96**, 11223-11228.

Dalsgaard, T. & Thamdrup, B. (2002). Factors controlling anaerobic ammonium oxidation with nitrite in marine sediments. *Appl Environ Microbiol* **68**, 3802-3808.

Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E., Overbeek, R., Snead, M. A., Keller, M., Aujay, M., Huber, R., Feldman, R. A., Short, J. M., Olsen, G. J. & Swanson, R. V. (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**, 353-358.

Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**, 4636-4641.

DeLong, E. F. (1993). Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. *Limnol Oceanogr* **38**, 924-934.

Derakshani, M., Lukow, T. & Liesack, W. (2001). Novel bacterial lineages at the (sub)division level as detected by signature nucleotide-targeted recovery of 16S rRNA genes from bulk soil and rice roots of flooded rice microcosms. *Appl Environ Microbiol* **67**, 623-631.

Dervyn, E., Suski, C., Daniel, R., Bruand, C., Chapuis, J., Errington, J., Janniere, L. & Ehrlich, S. D. (2001). Two essential DNA polymerases at the bacterial replication fork. *Science* **294**, 1716-1719.

Di Giulio, M. (2003). The ancestor of the *Bacteria* domain was a hyperthermophile. *J Theor Biol* **224**, 277-283.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Le Gall, F., Makarova, K. S., Ostrowski, M., Oztas, S., Robert, C., Rogozin, I. B., Scanlan, D. J., Tandeau de Marsac, N., Weissenbach, J., Wincker, P., Wolf, Y. I. & Hess, W. R. (2003). Genome sequence of the cyanobacterium

Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci U S A* **100**, 10020-10025.

Ermolaeva, M. D., Khalak, H. G., White, O., Smith, H. O. & Salzberg, S. L. (2000). Prediction of transcription terminators in bacterial genomes. *J Mol Biol* **301**, 27-33.

Filee, J., Forterre, P., Sen-Lin, T. & Laurent, J. (2002). Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *J Mol Evol* **54**, 763-773.

Flower, A. M. & McHenry, C. S. (1990). The gamma subunit of DNA polymerase III holoenzyme of *Escherichia coli* is produced by ribosomal frameshifting. *Proc Natl Acad Sci U S A* **87**, 3713-3717.

Frank, A. C. & Lobry, J. R. (1999). Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238**, 65-77.

Frank, A. C. & Lobry, J. R. (2000). Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* **16**, 560-561.

Franzmann, P. D. & Skerman, V. B. (1984). *Gemmata obscuriglobus*, a new genus and species of the budding bacteria. *Antonie Van Leeuwenhoek* **50**, 261-268.

Fraser, C. M., Eisen, J. A. & Salzberg, S. L. (2000). Microbial genome sequencing. *Nature* **406**, 799-803.

Fraser, C. M., Eisen, J. A., Nelson, K. E., Paulsen, I. T. & Salzberg, S. L. (2002). The value of complete microbial genome sequencing (you get what you pay for). *J Bacteriol* **184**, 6403-6405; discusion 6405.

Frishman, D., Mironov, A., Mewes, H. W. & Gelfand, M. (1998). Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res* **26**, 2941-2947.

Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A. & Mewes, H. W. (2001). Functional and structural genomics using PEDANT. *Bioinformatics* **17**, 44-57.

Fuerst, J. A. (1995). The *Planctomycetes*: emerging models for microbial ecology, evolution and cell biology. *Microbiology* **141 (Pt 7)**, 1493-1506.

Fuerst, J. A. & Webb, R. I. (1991). Membrane-bounded nucleoid in the eubacterium *Gemmatata obscuriglobus*. *Proc Natl Acad Sci U S A* **88**, 8184-8188.

Fuerst, J. A., Sambhi, S. K., Paynter, J. L., Hawkins, J. A. & Atherton, J. G. (1991). Isolation of a bacterium resembling *Pirellula* species from primary tissue culture of the giant tiger prawn (*Penaeus monodon*). *Appl Environ Microbiol* **57**, 3127-3134.

Fuerst, J. A., Gwilliam, H. G., Lindsay, M., Lichanska, A., Belcher, C., Vickers, J. E. & Hugenholtz, P. (1997). Isolation and molecular identification of planctomycete bacteria from postlarvae of the giant tiger prawn, *Penaeus monodon*. *Appl Environ Microbiol* **63**, 254-262.

Fujita, M. Q., Yoshikawa, H. & Ogasawara, N. (1992). Structure of the dnaA and DnaA-box region in the *Mycoplasma capricolum* chromosome: conservation and variations in the course of evolution. *Gene* **110**, 17-23.

Fukushima, A., Ikemura, T., Kinouchi, M., Oshima, T., Kudo, Y., Mori, H. & Kanaya, S. (2002). Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. *Gene* **300**, 203-211.

Gabrielian, A. & Bolshoy, A. (1999). Sequence complexity and DNA curvature. *Comput Chem* **23**, 263-274.

Garrity, G. M., Johnson, K. L., Bell, J. & Searles, D. B. (2002). *Taxonomic outline of the prokaryotes, Bergey's Manual of Systematic Bacteriology*, Second release, 3.0 edn. New York: Springer-Verlag.

Ghuysen, J. M. & Goffin, C. (1999). Lack of cell wall peptidoglycan versus penicillin sensitivity: new insights into the chlamydial anomaly. *Antimicrob Agents Chemother* **43**, 2339-2344.

Gimesi, N. (1924). Hydrobiologiai Tanulmányok (Hydrobiologische Studien). I. *Planctomyces Béfekii Gim*. nov. gen. et sp. *Budapest: Kiadja a Magyar Ciszterci Rend.*, 1-8.

Giovannoni, S. J., Godchaux, W., Schabtach, E. & Castenholz, R. W. (1987). Cell wall and lipid composition of *Isosphaera pallida*, a budding eubacterium from hot springs. *J Bacteriol* **169**, 2702-2707.

Glöckner, F. O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W., Gade, D., Beck, A., Borzym, K., Heitmann, K., Rabus, R., Schlesner, H., Amann, R. & Reinhardt, R. (2003). Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc Natl Acad Sci U S A* **100**, 8298-8303.

Glover, B. P. & McHenry, C. S. (2001). The DNA polymerase III holoenzyme: an asymmetric dimeric replicative complex with leading and lagging strand polymerases. *Cell* **105**, 925-934.

Goodsell, D. S. & Dickerson, R. E. (1994). Bending and curvature calculations in B-DNA. *Nucleic Acids Res* **22**, 5497-5503.

Green, G. N., Fang, H., Lin, R. J., Newton, G., Mather, M., Georgiou, C. D. & Gennis, R. B. (1988). The nucleotide sequence of the *cyd* locus encoding the two subunits of the cytochrome d terminal oxidase complex of *Escherichia coli*. *J Biol Chem* **263**, 13138-13143.

Gribaldo, S. & Philippe, H. (2002). Ancient phylogenetic relationships. *Theor Popul Biol* **61**, 391-408.

Griepenburg, U., Ward-Rainey, N., Mohamed, S., Schlesner, H., Marxsen, H., Rainey, F. A., Stackebrandt, E. & Auling, G. (1999). Phylogenetic diversity, polyamine pattern and DNA base composition of members of the order *Planctomycetales*. *Int J Syst Bacteriol* **49**, 689-696.

Guo, F. B., Ou, H. Y. & Zhang, C. T. (2003). ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res* **31**, 1780-1789.

Gupta, R. S. (2001). The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. *Int Microbiol* **4**, 187-202.

Gupta, R. S. & Griffiths, E. (2002). Critical issues in bacterial phylogeny. *Theor Popul Biol* **61**, 423-434.

Harris, J. K., Kelley, S. T., Spiegelman, G. B. & Pace, N. R. (2003). The genetic core of the universal ancestor. *Genome Res* **13**, 407-412.

Hartmann, R. K., Ulbrich, N. & Erdmann, V. A. (1987). An unusual rRNA operon constellation: in *Thermus thermophilus* HB8 the 23S/5S rRNA operon is a separate entity from the 16S rRNA operon. *Biochimie* **69**, 1097-1104.

Hatch, T. P. (1996). Disulfide cross-linked envelope proteins: the functional equivalent of peptidoglycan in *Chlamydiae*? *J Bacteriol* **178**, 1-5.

Hatch, T. P., Miceli, M. & Sublett, J. E. (1986). Synthesis of disulfide-bonded outer membrane proteins during the developmental cycle of *Chlamydia psittaci* and *Chlamydia trachomatis*. *J Bacteriol* **165**, 379-385.

Henikoff, J. G. & Henikoff, S. (1996). Blocks database and its applications. *Methods Enzymol* **266**, 88-105.

Huang, X. (1996). Fast comparison of a DNA sequence with a protein sequence database. *Microb Comp Genomics* **1**, 281-291.

Jenal, U., Stephens, C. & Shapiro, L. (1995). Regulation of asymmetry and polarity during the *Caulobacter* cell cycle. *Adv Enzymol Relat Areas Mol Biol* **71**, 1-39.

Jenkins, C. & Fuerst, J. A. (2001). Phylogenetic analysis of evolutionary relationships of the *Planctomycete* division of the domain *Bacteria* based on amino acid sequences of elongation factor Tu. *J Mol Evol* **52**, 405-418.

Jenkins, C., Kedar, V. & Fuerst, J. A. (2002). Gene discovery within the *Planctomycete* division of the domain *Bacteria* using sequence tags from genomic DNA libraries. *Genome Biol* **3**, research0031.0031-0031.0011.

Jensen, L. J., Friis, C. & Ussery, D. W. (1999). Three views of microbial genomes. *Res Microbiol* **150**, 773-777.

Jørgensen, B. B. (1984). Mineralization of organic matter in the sea bed - the role of sulphate reduction. *Nature* **296**, 643-645.

Kabsch, W., Sander, C. & Trifonov, E. N. (1982). The ten helical twist angles of B-DNA. *Nucleic Acids Res* **10**, 1097-1104.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M. & Tabata, S. (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions (supplement). *DNA Res* **3**, 185-209.

Karunakaran, K. P., Noguchi, Y., Read, T. D., Cherkasov, A., Kwee, J., Shen, C., Nelson, C. C. & Brunham, R. C. (2003). Molecular analysis of the multiple GroEL proteins of *Chlamydiae*. *J Bacteriol* **185**, 1958-1966.

Kerger, D., Mancuso, A., Nichols, P. D., White, D. C., Langworthy, T., Sittig, M., Schlesner, H. & Hirsch, P. (1988). The budding bacteria, *Pirellula* and *Planctomyces*, with atypical 16S rRNA and absence of peptidoglycan, show eubacterial phospholipids und uniquely high proportions of long chain beta-hydroxy fatty acids in the lipopolysaccharide lipid A. *Arch Microbiol* **149**, 255-260.

Knoblauch, C., Sahm, K. & Jorgensen, B. B. (1999). Psychrophilic sulfate-reducing bacteria isolated from permanently cold arctic marine sediments: description of *Desulfofrigus oceanense* gen. nov., sp. nov., *Desulfofrigus fragile* sp. nov., *Desulfofaba gelida* gen. nov., sp. nov., *Desulfotalea psychrophila* gen. nov., sp. nov. and *Desulfotalea arctica* sp. nov. *Int J Syst Bacteriol* **49**, 1631-1643.

König, E., Schlesner, H. & Hirsch, P. (1984). Cell wall studies on budding bacteria of the *Planctomyces/Pasteuria* group and on a *Prosthecomicrobium* sp. *Arch Microbiol* **138**, 200-205.

Koonin, E. V. & Bork, P. (1996). Ancient duplication of DNA polymerase inferred from analysis of complete bacterial genomes. *Trends Biochem Sci* **21**, 128-129.

Lamfrom, H., Sarabhai, A. & Abelson, J. (1978). Cloning of *Beneckea* genes in *Escherichia coli*. *J Bacteriol* **133**, 354-363.

Larsen, T. S. & Krogh, A. (2003). EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* **4**, 21.

Liesack, W. & Stackebrandt, E. (1989). Evidence for unlinked rrn operons in the planctomycete *Pirellula marina*. *J Bacteriol* **171**, 5025-5030.

Liesack, W. & Stackebrandt, E. (1992). Occurrence of novel groups of the domain *Bacteria* as revealed by analysis of genetic material isolated from an Australian terrestrial environment. *J Bacteriol* **174**, 5072-5078.

Liesack, W., König, H., Schlesner, H. & Hirsch, P. (1986). Chemical composition of the peptidoglycan-free cell envelopes of budding bacteria of the *Pirellula/Planctomyces* group. *Arch Microbiol* **145**, 361-366.

Lindsay, M. R., Webb, R. & Fuerst, J. A. (1997). Pirellulosomes: a new type of membrane-bounded cell compartment in planctomycete bacteria of the genus *Pirellula*. *Microbiology* **143**, 739-748.

Lindsay, M. R., Webb, R. I., Strous, M., Jetten, M. S., Butler, M. K., Forde, R. J. & Fuerst, J. A. (2001). Cell compartmentalisation in planctomycetes: novel types of structural organisation for the bacterial cell. *Arch Microbiol* **175**, 413-429.

Llobet-Brossa, E., Rossello-Mora, R. & Amann, R. (1998). Microbial community composition of Wadden Sea sediments as revealed by fluorescence in situ hybridization. *Appl Environ Microbiol* **64**, 2691-2696.

Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* **13**, 660-665.

Lombardot, T., Bauer, M., Teeling, H., Amann, R. & Glöckner, F. O. (2004). The transcriptional regulators pool of the marine bacterium *Pirellula* sp. strain 1 as revealed by whole genome comparisons. *Environ Microbiol*, in preparation.

Ludwig, W., Strunk, O., Klugbauer, S., Klugbauer, N., Weizenegger, M., Neumaier, J., Bachleitner, M. & Schleifer, K. H. (1998). Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis* **19**, 554-568.

Lupas, A., Van Dyke, M. & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science* **252**, 1162-1164.

Marczynski, G. T. & Shapiro, L. (2002). Control of chromosome replication in *Caulobacter crescentus*. *Annu Rev Microbiol* **56**, 625-656.

Margolin, W. (2000). Themes and variations in prokaryotic cell division. *FEMS Microbiol Rev* **24**, 531-548.

Matte-Tailliez, O., Brochier, C., Forterre, P. & Philippe, H. (2002). Archaeal phylogeny based on ribosomal proteins. *Mol Biol Evol* **19**, 631-639.

Menke, M. A. O. H., Liesack, W. & Stackebrandt, E. (1991). Ribotyping of 16S and 23S rRNA genes and organization of *rrn* operons in members of the bacterial genera *Gemmata, Planctomyces, Thermotoga, Thermus,* and *Verrucomicrobium*. *Arch Microbiol* **133**, 263-271.

Merkl, R. (2003). A survey of codon and amino acid frequency bias in microbial genomes focusing on translational efficiency. *J Mol Evol* **57**, 453-466.

Methe, B. A., Nelson, K. E., Eisen, J. A., Paulsen, I. T., Nelson, W., Heidelberg, J. F., Wu, D., Wu, M., Ward, N., Beanan, M. J., Dodson, R. J., Madupu, R., Brinkac, L. M., Daugherty, S. C., DeBoy, R. T., Durkin, A. S., Gwinn, M., Kolonay, J. F., Sullivan, S. A., Haft, D. H., Selengut, J., Davidsen, T. M., Zafar, N., White, O., Tran, B., Romero, C., Forberger, H. A., Weidman, J., Khouri, H., Feldblyum, T. V., Utterback, T. R., Van Aken, S. E., Lovley, D. R. & Fraser, C. M. (2003). Genome of *Geobacter sulfurreducens*: metal reduction in subsurface environments. *Science* **302**, 1967-1969.

Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R. & Puhler, A. (2003). GenDB--an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res* **31**, 2187-2195.

Miskin, I. P., Farrimond, P. & Head, I. M. (1999). Identification of novel bacterial lineages as active members of microbial populations in a freshwater sediment using a rapid RNA extraction procedure and RT-PCR. *Microbiology* **145**, 1977-1987.

Munson, M. A., Baumann, L. & Baumann, P. (1993). *Buchnera aphidicola* (a prokaryotic endosymbiont of aphids) contains a putative 16S rRNA operon unlinked to the 23S rRNA-encoding gene: sequence determination, and promoter and terminator analysis. *Gene* **137**, 171-178.

Nakai, K. & Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* **14**, 897-911.

Nakamura, Y., Kaneko, T., Sato, S., Mimuro, M., Miyashita, H., Tsuchiya, T., Sasamoto, S., Watanabe, A., Kawashima, K., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Nakazaki, N., Shimpo, S., Takeuchi, C., Yamada, M. & Tabata, S. (2003). Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Res* **10**, 137-145.

Neef, A., Amann, R., Schlesner, H. & Schleifer, K. H. (1998). Monitoring a widespread bacterial group: in situ detection of planctomycetes with 16S rRNA-targeted probes. *Microbiology* **144 (Pt 12)**, 3257-3266.

Nielson, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* **10**, 1-6.

Nierman, W. C., Feldblyum, T. V., Laub, M. T., Paulsen, I. T., Nelson, K. E., Eisen, J. A., Heidelberg, J. F., Alley, M. R., Ohta, N., Maddock, J. R., Potocka, I., Nelson, W. C., Newton, A., Stephens, C., Phadke, N. D., Ely, B., DeBoy, R. T., Dodson, R. J., Durkin, A. S., Gwinn, M. L., Haft, D. H., Kolonay, J. F., Smit, J., Craven, M. B., Khouri, H., Shetty, J., Berry, K., Utterback, T., Tran, K., Wolf, A., Vamathevan, J., Ermolaeva, M., White, O., Salzberg, S. L., Venter, J. C., Shapiro, L., Fraser, C. M. & Eisen, J. (2001). Complete genome sequence of *Caulobacter crescentus*. *Proc Natl Acad Sci U S A* **98**, 4136-4141.

Ogasawara, N. & Yoshikawa, H. (1992). Genes and their organization in the replication origin region of the bacterial chromosome. *Mol Microbiol* **6**, 629-634.

Olson, S. A. (2002). EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. *Brief Bioinform* **3**, 87-91.

Ornstein, R. L. & Rein, R. (1979). Energetics of intercalation specificity. I. Backbone unwinding. *Biopolymers* **18**, 1277-1291.

Palmer, T. & Berks, B. C. (2003). Moving folded proteins across the bacterial cell membrane. *Microbiology* **149**, 547-556.

Pearson, A., Budin, M. & Brocks, J. J. (2003). Phylogenetic and biochemical evidence for sterol synthesis in the bacterium *Gemmata obscuriglobus*. *Proc Natl Acad Sci U S A* **100**, 15352-15357.

Pereira, M. M., Verkhovskaya, M. L., Teixeira, M. & Verkhovsky, M. I. (2000). The caa(3) terminal oxidase of *Rhodothermus marinus* lacking the key glutamate of the D-channel is a proton pump. *Biochemistry* **39**, 6336-6340.

Pereira, M. M., Santana, M., Soares, C. M., Mendes, J., Carita, J. N., Fernandes, A. S., Saraste, M., Carrondo, M. A. & Teixeira, M. (1999). The caa3 terminal oxidase of the thermohalophilic bacterium *Rhodothermus marinus*: a HiPIP:oxygen oxidoreductase lacking the key glutamate of the D-channel. *Biochim Biophys Acta* **1413**, 1-13.

Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. (2003). Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**, 145-158.

Pritchard, A. E., Dallmann, H. G., Glover, B. P. & McHenry, C. S. (2000). A novel assembly mechanism for the DNA polymerase III holoenzyme DnaX complex: association of deltadelta' with DnaX(4) forms DnaX(3)deltadelta'. *Embo J* **19**, 6536-6545.

Rabus, R., Gade, D., Helbig, R., Bauer, M., Glockner, F. O., Kube, M., Schlesner, H., Reinhardt, R. & Amann, R. (2002). Analysis of N-acetylglucosamine metabolism in the marine bacterium *Pirellula* sp. strain 1 by a proteomic approach. *Proteomics* **2**, 649-655.

Ramakrishna, R. & Srinivasan, R. (1999). Gene identification in bacterial and organellar genomes using GeneScan. *Comput Chem* **23**, 165-174.

Rendulic, S., Jagtap, P., Rosinus, A., Eppinger, M., Baar, C., Lanz, C., Keller, H., Lambert, C., Evans, K. J., Goesmann, A., Meyer, F., Sockett, R. E. & Schuster, S. C. (2004). A predator unmasked: life cycle of *Bdellovibrio bacteriovorus* from a genomic perspective. *Science* **303**, 689-692.

Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-277.

Riemann, L. & Azam, F. (2002). Widespread N-acetyl-D-glucosamine uptake among pelagic marine bacteria and its ecological implications. *Appl Environ Microbiol* **68**, 5554-5562.

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., Johnson, Z. I., Land, M., Lindell, D., Post, A. F., Regala, W., Shah, M., Shaw, S. L., Steglich, C., Sullivan, M. B., Ting, C. S., Tolonen, A., Webb, E. A., Zinser, E. R. & Chisholm, S. W. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**, 1042-1047.

Rocha, E. (2002). Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol* **10**, 393-395.

Rurangirwa, F. R., Brayton, K. A., McGuire, T. C., Knowles, D. P. & Palmer, G. H. (2002). Conservation of the unique rickettsial rRNA gene arrangement in *Anaplasma*. *Int J Syst Evol Microbiol* **52**, 1405-1409.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944-945.

Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* **26**, 544-548.

Schaper, S., Nardmann, J., Luder, G., Lurz, R., Speck, C. & Messer, W. (2000). Identification of the chromosomal replication origin from *Thermus thermophilus* and its interaction with the replication initiator DnaA. *J Mol Biol* **299**, 655-665.

Schlesner, H. (1994). The development of media suitable for the microorganisms morphologically resembling *Planctomyces* spp., *Pirellula* spp., and other *Planctomycetales* from various aquatic habitats using dilute media. *Syst Appl Microbiol* **17**, 135-145.

Schlesner, H., Rathmann, M., Bartels, C., Tindall, B., Gade, D., Rabus, R., Pfeiffer, S. & Hirsch, P. (2004). Taxonomic heterogeneity within the *Planctomycetales* as derived by DNA/DNA-hybridization, description of *Rhodopirellula baltica* gen. nov., sp. nov. and transfer of *Pirellula marina* to the genus *Blastopirellula* gen. nov. as *Blastopirellula marina* comb. nov. *Int J Syst Evol Microbiol*, under revision.

Schmid, M., Schmitz-Esser, S., Jetten, M. & Wagner, M. (2001). 16S-23S rDNA intergenic spacer and 23S rDNA of anaerobic ammonium-oxidizing bacteria: implications for phylogeny and in situ detection. *Environ Microbiol* **3**, 450-459.

Schmid, M., Twachtmann, U., Klein, M., Strous, M., Juretschko, S., Jetten, M., Metzger, J. W., Schleifer, K. H. & Wagner, M. (2000). Molecular evidence for genus level diversity of bacteria capable of catalyzing anaerobic ammonium oxidation. *Syst Appl Microbiol* **23**, 93-106.

Sever, M. J., Weisser, J. T., Monahan, J., Srinivasan, S. & Wilker, J. J. (2004). Metal-mediated cross-linking in the generation of a marine mussel adhesive. *Angewandte Chemie International Edition*, 10.1002/anie.200352759.

Sharp, P. M. & Li, W. H. (1987). The codon adaptation index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 1281-1295.

Shmatkov, A. M., Melikyan, A. A., Chernousko, F. L. & Borodovsky, M. (1999). Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes. *Bioinformatics* **15**, 874-886.

Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D. & Krogh, A. (2001). On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet* **17**, 425-428.

Sonnhammer, E. L., Eddy, S. R. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405-420.

Sonnhammer, E. L., von Heijne, G. & Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* **6**, 175-182.

Stackebrandt, E., Wehmeyer, U. & Liesack, W. (1986). 16S ribosomal RNA- and cell wall analysis of *Gemmata obscuriglobus*, a new member of the order *Planctomycetales*. *FEMS Microbiol Lett* **37**, 289-292.

Stackebrandt, E., Ludwig, W., Schubert, W., Klink, F., Schlesner, H., Roggentin, T. & Hirsch, P. (1984). Molecular genetic evidence for early evolutionary origin of budding peptidoglycan-less eubacteria. *Nature* **307**, 735-737.

Stephens, R. S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R. L., Zhao, Q., Koonin, E. V. & Davis, R. W. (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**, 754-759.

Strous, M., Fuerst, J. A., Kramer, E. H., Logemann, S., Muyzer, G., van de Pas-Schoonen, K. T., Webb, R., Kuenen, J. G. & Jetten, M. S. (1999). Missing lithotroph identified as new planctomycete. *Nature* **400**, 446-449.

Suzek, B. E., Ermolaeva, M. D., Schreiber, M. & Salzberg, S. L. (2001). A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics* **17**, 1123-1130.

Takami, H., Takaki, Y. & Uchiyama, I. (2002). Genome sequence of *Oceanobacillus iheyensis* isolated from the Iheya Ridge and its unexpected adaptive capabilities to extreme environments. *Nucleic Acids Res* **30**, 3927-3935.

Tan, K., Beattie, P., Leach, D. R., Rich, P. R., Coulson, A. F. & Ward, F. B. (1996). Expression and analysis of the gene for the catalytic beta subunit of the sodium-translocating NADH-ubiquinone oxidoreductase of *Vibrio alginolyticus*. *Biochem Soc Trans* **24**, 12S.

Taschke, C., Klinkert, M. Q., Wolters, J. & Herrmann, R. (1986). Organization of the ribosomal RNA genes in *Mycoplasma hyopneumoniae*: the 5S rRNA gene is separated from the 16S and 23S rRNA genes. *Mol Gen Genet* **205**, 428-433.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science* **278**, 631-637.

Tech, M. & Merkl, R. (2003). YACOP: Enhanced gene prediction obtained by a combination of existing methods. *In Silico Biol* **3**, 441-451.

Tekniepe, B. L., Schmidt, J. M. & Starr, M. P. (1981). Life cycle of a budding and appendaged bacterium belonging to morphotype IV ofthe *Blastocaulis-Planctomyces* group. *Curr Microbiol* **5**, 1-6.

Thamdrup, B. & Dalsgaard, T. (2002). Production of N(2) through anaerobic ammonium oxidation coupled to nitrate reduction in marine sediments. *Appl Environ Microbiol* **68**, 1312-1318.

Tsuchihashi, Z. (1991). Translational frameshifting in the *Escherichia coli dnaX* gene in vitro. *Nucleic Acids Res* **19**, 2457-2462.

Tsuchihashi, Z. & Kornberg, A. (1990). Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. *Proc Natl Acad Sci U S A* **87**, 2516-2520.

Tsuchihashi, Z. & Brown, P. O. (1992). Sequence requirements for efficient translational frameshifting in the *Escherichia coli dnaX* gene and the role of an unstable interaction between tRNA(Lys) and an AAG lysine codon. *Genes Dev* **6**, 511-519.

Tu, J. & Zillig, W. (1982). Organization of rRNA structural genes in the archaebacterium *Thermoplasma acidophilum*. *Nucleic Acids Res* **10**, 7231-7245.

Vergin, K. L., Urbach, E., Stein, J. L., DeLong, E. F., Lanoil, B. D. & Giovannoni, S. J. (1998). Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order *Planctomycetales*. *Appl Environ Microbiol* **64**, 3075-3078.

Wang, J., Jenkins, C., Webb, R. I. & Fuerst, J. A. (2002). Isolation of *Gemmata*-like and *Isosphaera*-like planctomycete bacteria from soil and freshwater. *Appl Environ Microbiol* **68**, 417-422.

Ward, N. L., Rainey, F. A., Hedlund, B. P., Staley, J. T., Ludwig, W. & Stackebrandt, E. (2000). Comparative phylogenetic analyses of members of the order *Planctomycetales* and the division *Verrucomicrobia*: 23S rRNA gene sequence analysis supports the 16S rRNA gene sequence-derived phylogeny. *Int J Syst Evol Microbiol* **50 Pt 6**, 1965-1972.

Ward-Rainey, N., Rainey, F. A. & Stackebrandt, E. (1997). The presence of a *dnaK* (HSP70) multigene family in members of the orders *Planctomycetales* and *Verrucomicrobiales*. *J Bacteriol* **179**, 6360-6366.

Weisburg, W. G., Hatch, T. P. & Woese, C. R. (1986). Eubacterial origin of *Chlamydiae*. *J Bacteriol* **167**, 570-574.

Widom, J. (1996). Short-range order in two eukaryotic genomes: relation to chromosome structure. *J Mol Biol* **259**, 579-588.

Woese, C. R. (1987). Bacterial evolution. *Microbiol Rev* **51**, 221-271.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V. & Koonin, E. V. (2002). Genome trees and the tree of life. *Trends Genet* **18**, 472-479.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L. & Koonin, E. V. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol* **1**, 8.

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene* **87**, 23-29.

Yeh, S. R., Couture, M., Ouellet, Y., Guertin, M. & Rousseau, D. L. (2000). A cooperative oxygen binding hemoglobin from *Mycobacterium tuberculosis*. Stabilization of heme ligands by a distal tyrosine residue. *J Biol Chem* **275**, 1679-1684.

Yurieva, O., Skangalis, M., Kuriyan, J. & O'Donnell, M. (1997). *Thermus thermophilis dnaX* homolog encoding gamma- and tau-like proteins of the chromosomal replicase. *J Biol Chem* **272**, 27131-27139.

Teil II:
Publikationen

A Publikationsliste mit Erläuterungen

Diese Dissertation beruht in wesentlichen Teilen auf vier Publikationen. Die Beiträge der einzelnen Autoren zu diesen Publikationen sind im Folgenden aufgeführt:

1. **Glöckner, F. O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W., Gade, D., Beck, A., Borzym, K., Heitmann, K., Rabus, R., Schlesner, H., Amann, R. & Reinhardt, R. (2003)** Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc Natl Acad Scie U S A* **100**(14): 8298-8303

   - *Eigenbeiträge (entsprechend der Reihenfolge in der Publikation): Wesentliche Anteile an der ORF-Vorhersage und der Annotation, Erstellung der Genom-Atlanten (Abb. 3 und Abb. 4), Durchführung von Untersuchungen zur Genom-Organisation (u.a. Transposasen), zur Genomgröße (Anteil paraloger Gene), zum Energiestoffwechsel, zur Motilität, zur Streßantwort, zur Zellbiologie (Zellwand, Membran, Kompartmentierung, Zellteilung, Zellzyklus) und zur Verwandtschaft mit den Chlamydien*
   - *Konzept: F. O. Glöckner, R. Amann*
   - *Sequenzierung und Assemblierung: M. Kube, A. Beck, K. Borzym, K. Heitmann*
   - *Bioinformatische Untersuchungen: M. Bauer, H. Teeling, T. Lombardot, F. O. Glöckner*
   - *Zellanzucht, funktionelle Analysen und Proteomik: D. Gade, R. Rabus, H. Schlesner*
   - *Erstellung des Manuskripts: F. O. Glöckner unter redaktioneller Mitarbeit von M. Bauer, H. Teeling, T. Lombardot und R. Amann*

2. **Teeling, H., Lombardot, T., Bauer, M., Ludwig, L. & Glöckner, F. O. (2003)** Reevalutation of the phylogenetic position of the *Planctomycetes* by means of concatenated ribosomal protein sequences, DNA-directed RNA polymerase subunit sequences and whole genome trees. *Int J Sys Evol Microbiol*, published online December 5th - in press

   - *Konzept: H. Teeling, F. O. Göckner*
   - *Durchführung der Untersuchungen: H. Teeling (konkatenierte Sequenzen ribosomaler Proteine und Untereinheiten der DNA-abhängigen RNA-Polymerase), T. Lombardot (BLASTp-basierte Genom-Phylogenie)*
   - *Erstellung des Manuskripts: H. Teeling. unter redaktioneller Mitarbeit von F. O. Glöckner, T. Lombardot, M. Bauer und W. Ludwig*

3. **Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glöckner, F. O. (2004)** Application of Tetranucleotide Frequencies for the Assignment of Genomic Fragments. *Environ Microbiol* Special Issue on Metagenomics - in press

   - *Konzept: H. Teeling, F. O. Glöckner*
   - *Durchführung der Untersuchungen: H. Teeling (Bioinformatik), A. Meyerdierks (Metagenomik)*
   - *Erstellung des Manuskripts: H. Teeling unter redaktioneller Mitarbeit von F. O. Glöckner, A. Meyerdierks, M. Bauer und R. Amann*

4. **Teeling, H., Waldmann J., Bauer, M., & Glöckner, F. O.,** MORFind: improved gene-prediction by the combination of gene-finders. - in preparation for *Bioinformatics*

   - *Konzept: H. Teeling*
   - *Durchführung der Untersuchungen: H. Teeling (Bioinformatik), J. Waldmann (Implementation)*
   - *Erstellung des Manuskripts: H. Teeling unter redaktioneller Mitarbeit von F. O. Glöckner und M. Bauer*

1

Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1

F. O. Glöckner, M. Kube, M. Bauer, H. Teeling, T. Lombardot, W. Ludwig, D. Gade, A. Beck, K. Borzym, K. Heitmann, R. Rabus, H. Schlesner, R. Amann and R. Reinhardt

Proc Natl Acad Sci U S A 100(14): 8298-8303 (2003)

Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1

F. O. Glöckner*†, M. Kube‡, M. Bauer*†, H. Teeling*, T. Lombardot*, W. Ludwig§, D. Gade*, A. Beck‡, K. Borzym‡, K. Heitmann‡, R. Rabus*, H. Schlesner¶, R. Amann*, and R. Reinhardt†‡

*Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany; ‡Max Planck Institute for Molecular Genetics, Ihnestrasse 73, D-14195 Berlin, Germany; §Department of Microbiology, Technical University Munich, Am Hochanger 4, D-85350 Freising, Germany; and ¶Department for General Microbiology, University of Kiel, Am Botanischen Garten 1-9, D-24118 Kiel, Germany

Edited by Carl R. Woese, University of Illinois at Urbana–Champaign, Urbana, IL, and approved May 6, 2003 (received for review March 12, 2003)

***Pirellula* sp. strain 1 ("*Rhodopirellula baltica*") is a marine representative of the globally distributed and environmentally important bacterial order *Planctomycetales*. Here we report the complete genome sequence of a member of this independent phylum. With 7.145 megabases, *Pirellula* sp. strain 1 has the largest circular bacterial genome sequenced so far. The presence of all genes required for heterolactic acid fermentation, key genes for the interconversion of C1 compounds, and 110 sulfatases were unexpected for this aerobic heterotrophic isolate. Although *Pirellula* sp. strain 1 has a proteinaceous cell wall, remnants of genes for peptidoglycan synthesis were found. Genes for lipid A biosynthesis and homologues to the flagellar L- and P-ring protein indicate a former Gram-negative type of cell wall. Phylogenetic analysis of all relevant markers clearly affiliates the *Planctomycetales* to the domain *Bacteria* as a distinct phylum, but a deepest branching is not supported by our analyses.**

Pirellula sp. strain 1, which is in the process of being validly described as "*Rhodopirellula baltica*," is a marine, aerobic, heterotrophic representative of the globally distributed and environmentally important bacterial order *Planctomycetales*. Molecular microbial ecology studies repeatedly provided evidence that planctomycetes are abundant in terrestrial and marine habitats (1–5). For example, they inhabit phytodetrital macroaggregates in marine environments (6) and include one of the organisms known to derive energy from the anaerobic oxidation of ammonia (7). They catalyze important transformations in global carbon and nitrogen cycles. By their mineralization of marine snow particles planctomycetes have a profound impact on global biogeochemistry and climate by affecting exchange processes between the geosphere and atmosphere (8). From a phylogenetic perspective the order *Planctomycetales* forms an independent, monophyletic phylum of the domain *Bacteria* (9). It has recently been suggested to be the deepest branching bacterial phylum (10). Planctomycetes are unique in many other respects. Their cell walls do not contain peptidoglycan, the main structural polymer of most members of the domain *Bacteria*. They show a unique cell compartmentalization in which a single membrane separates a peripheral ribosome-free paryphoplasm from the inner riboplasm (pirellulosome). Within the riboplasm, all planctomycetes contain a condensed fibrillar nucleoid, which in *Gemmata* spp. is surrounded by a additional double membrane (11). These structures, together with an unusual fatty acid composition of the phospholipids, resemble eukaryotes rather than a representative of the bacterial domain (12).

Characteristic for *Planctomycetales* are the polar cell organization and a life cycle with a polar, yeast-like cell division. Cells attach to surfaces at their vegetative poles by means of an excreted holdfast substance or stalks (13). Further unusual features are the crateriform structures on the cell surface of all planctomycetes (14). They appear as electron-dense circular regions at the reproductive cell pole (*Pirellula* spp.) or on the whole cell surface (*Planctomyces* spp.) (15).

Currently, none of the members of this fascinating group have been investigated by a genomic approach. Here we report the complete, closed genome of *Pirellula* sp. strain 1, a Baltic Sea isolate from the Kiel Fjord (16).

Methods

Sequencing Strategy. Genome sequencing was performed by a combination of a clone-based and a whole-genome shotgun approach. Two plasmid libraries with 1.5- and 3.5-kb inserts and a cosmid library (Epicentre Technologies, Madison, WI) were built from *Pirellula* sp. strain 1 DNA. End sequences of inserts were determined by using Big Dye chemistry (ABI), M13 primers, and ABI 3,700 capillary sequencers (ABI) up to eightfold sequence coverage. All raw sequences were processed by PHRED (17) and controlled for vector or *Escherichia coli* contamination. Reads were assembled by PHRAP and manually finished by using GAP4 (18). The quality of the sequence data was finished to reach a maximum of 1 error within 10,000 bases. Gap closure and finishing of the sequence were done by resequencing clones, primer-walking, and long-range PCR. Locations and sequence of repetitive sequence elements were additionally controlled by PCR.

Open Reading Frame (ORF) Prediction. Three different programs were used for ORF prediction, GLIMMER (19), CRITICA (20), and ORPHEUS (21). A nonredundant list of ORFs was generated by parsing the results with a self-written Perl-script. The script applied performs in the following way: For all ORFs that are predicted identically by all three gene finders, only one is kept. If the script recognizes identical stop positions but different starts and the difference is below 10% of the sequence length, only the longer ORF is kept. If the difference is more than 10%, both ORFs are kept.

Annotation. The software package PEDANT PRO (22) was used for annotation. All automatically generated results were evaluated manually for final annotation. Obviously overpredicted ORFs, e.g., overlapping ORFs without functional assignment, were marked for deletion and deleted after cross-checking by at least two independent annotators.

Data Analysis. For origin and terminus determination a combination of compositional indexes and oligomer distribution skew was used. The following compositional indexes were determined with self-written Perl-scripts: (*i*) cumulative GC skew [sum of (G − C)/(G + C) over adjacent windows of 10 kb]; (*ii*) keto excess [sum(GT) − sum(AC)]; (*iii*) purine excess [sum(AG) − sum(TC)] and the external program OLIGOSKEW8 (www.tigr.org/~salzberg/oligoskew8). Repeats were detected by the software REPUTER (23). DNA flexibility such as curvature and bending

This paper was submitted directly (Track II) to the PNAS office.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. BX119912).

†To whom correspondence may be addressed. E-mail: fog@mpi-bremen.de or mbauer@mpi-bremen.de. Requests for sequencing details, sequences, and clones should go directly to R. Reinhardt. E-mail: rr@molgen.mpg.de.

www.pnas.org/cgi/doi/10.1073/pnas.1431443100

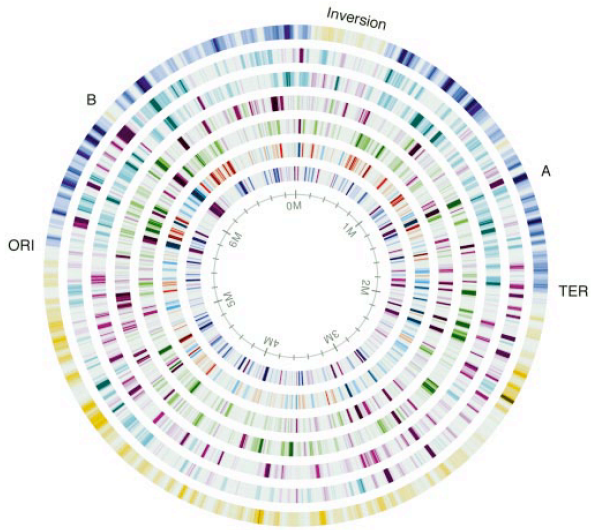


Fig. 1. Structural representation of the *Pirellula* sp. strain 1 genome. Circle 1 (from the outside in), GC skew; circle 2, G+C content; circle 3, G+A content; circle 4, DNA curvature; circle 5, DNA bending; circle 6, DNA stacking energy; circle 7, codon adaptation index (CAI). The origin of replication (ORI) and the terminus (TER) are indicated. A and B indicate minor irregularities. Ochre, pink, and red represent high values; blue, light blue, and green show low values.

was calculated with the BANANA program, and sequence twist was calculated with the program BTWISTED, both taken from the EMBOSS package (www.hgmp.mrc.ac.uk/Software/EMBOSS). Codon usage (codon adaptation index, CAI) was calculated with the CODONW-program (www.molbiol.ox.ac.uk/cu). Highly expressed and alien genes according to Karlin and Mrazek (24) were identified with self-written Perl-scripts. For the phylogenetic distribution of the best BLAST hits the SEALS package and the taxonomy of the National Center for Biotechnology Information was used. Tat signals were found by extracting all proteins containing twin arginines plus two additional amino acids of the conserved Tat pattern (SRRXFLK).

For whole-genome visualization the software tool GENEWIZ (25) was used. Total gene numbers were calculated by searches against all publicly available genomes with Pfam profiles (http://pfam.wustl.edu) by using GENDB 1.1 (46). For phylogenetic reconstructions the preliminary sequence of *Gemmata obscuriglobus* UQM 2246 was obtained from The Institute for Genomic Research (www.tigr.org). The program package ARB was used for phylogenetic analysis (www.arb-home.de).

Supporting Information. All supporting information (*Appendices 1–8*) is available on the PNAS web site, www.pnas.org. The complete annotation data and all supporting information are available on the home page of the REGX Project, www.regx.de. For fast searching a BLAST server is available for public use.

Results and Discussion

Genome Organization. With a size of 7,145,576 bases, *Pirellula* sp. strain 1 has the largest prokaryotic circular genome sequenced so far. Origin and terminus could be clearly identified by the change in cumulative GC and AT skews (Fig. 1). A single, unlinked rRNA operon was identified near the origin. Unlinked *rrn* operons have also been described for other planctomycetes (13) but the 460 kb separating the 16S from the 23S–5S rRNA genes in *Pirellula* sp. strain 1 are exceptional. A nonrandom distribution of the 81 transposases and the 13 integrases/recombinases was found: 68% (55) of all transposases and 85% (11) of all integrases/recombinases are located in the region between 0 and 3.6 megabases of the genome. General features of the genomic sequence are shown in Table 1.

Table 1. General features of the *Pirellula* sp. strain 1 genome

Component of chromosome	Property
Total size, bases	7,145,576
G+C content, %	55.4
Coding sequences	7,325
Coding density, %	95
Average gene length, bases	939
Genes with similarities in databases*	3,380 (46%)
Genes with functional assignments	2,582 (35%)
rRNAs	1× (16S) and (23S–5S)
tRNAs	70
Other stable RNAs	1 (ribozyme)

*Threshold for BLASTP *E* value $\leq 1 \times 10^{-3}$, includes hits to hypothetical proteins.

Irregularity. A large inversion at position 87,500 to 431,000 of 343.5 kb is indicated by cumulative GC-skew and other structural parameters (Fig. 1 and *Appendix 1*, which is published as supporting information on the PNAS web site). Analysis of this anomalous region did not give any indications for lateral gene transfer. On the contrary, a regular codon adaptation index and the localization of housekeeping genes (e.g., several tRNA synthetases, ribosomal proteins, and flagella proteins) in this region indicate that most probably an internal chromosomal inversion has occurred. This conclusion is supported by five and four flanking transposases. Two of them are identical and have reverse orientation. In total, 16% (13) of all transposase genes are located within this region, supporting a hot spot for large genomic rearrangements.

Gene and Functional Prediction. An initial nonredundant list of 13,331 potential ORFs was generated. By manual annotation, this ORF set could be reduced to 7,325 ORFs, which equals a gene coverage of 95%. A BLASTX of all intergenic regions confirmed that a comprehensive ORF set was achieved. More than half (56%, 4,148) of the predicted ORFs showed only weak or no similarities (*E* values higher than 0.9) compared with the currently available sequence databases. Only 32% (2,384) of all ORFs had reliable functional predictions, which is about 20% less than the numbers found in general (26). This low percentage reflects the distinct phylogenetic position and the lack of molecular studies performed on *Planctomycetales* so far. An overview of the localization of the functional genes according to our functional classification (27) is given in Fig. 2. The complete annotation results are available at www.regx.de.

Genome Size. We found that 1,301 genes (i.e., 17.6% of all genes), with an average length of 464 aa, have more than one copy within the genome. In total, multicopy genes make up for about 25.4% of the genome sequence. This is less than the 30% reported for *Bacillus subtilis* (28) and the 29% we calculated for *E. coli* K-12. Therefore extensive gene duplication is not the reason for the large genome size of *Pirellula* sp. strain 1. A large genome with an expanded genetic capability might be a prerequisite for environmental adaptability, as already discussed for the genome of *Pseudomonas aeruginosa* (29).

Potential Environmental Adaptations. ***Metabolism.*** The annotation process identified the standard pathways for heterotrophic bacteria such as glycolysis, the citrate cycle, and oxidative phosphorylation. *Pirellula* sp. strain 1 lacks the glyoxylate bypass and the Entner–Doudoroff pathway but exhibits the pentose-phosphate cycle. Fur-

ECOLOGY

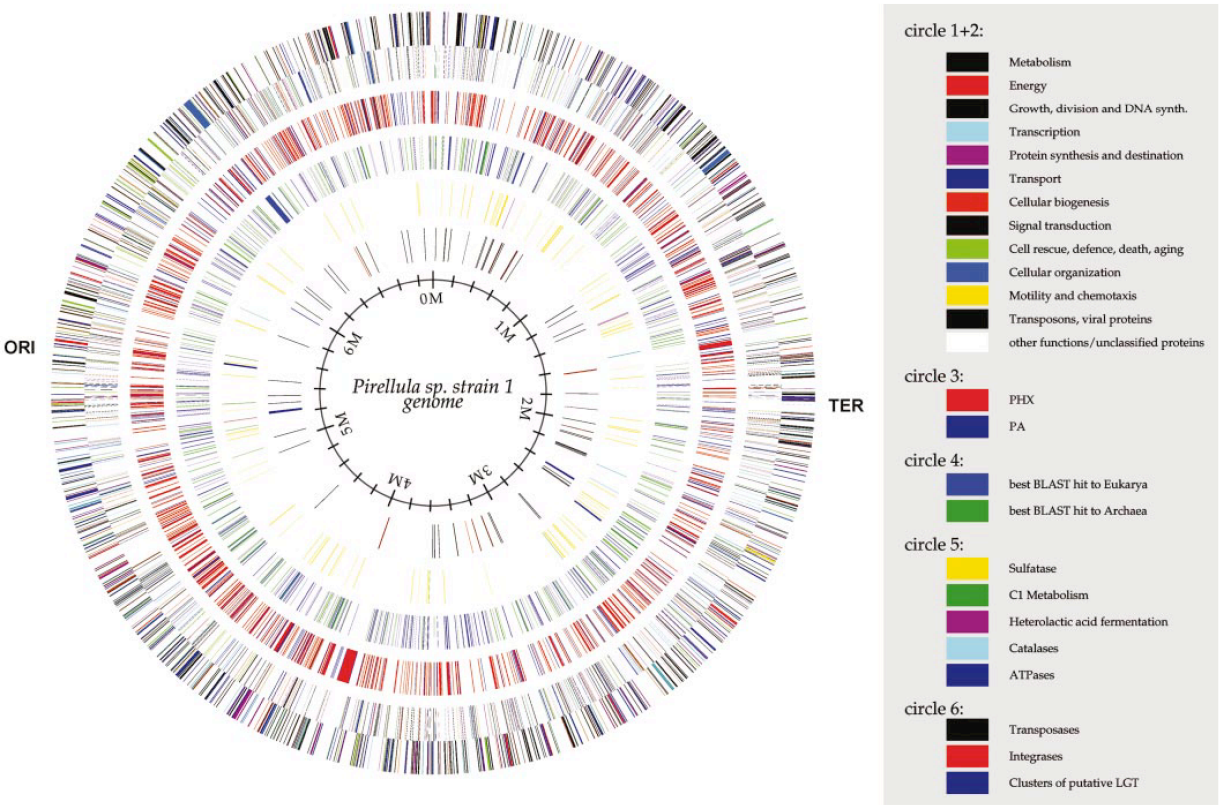


Fig. 2. Circular representation of the *Pirellula* sp. strain 1 genome. Circles 1 and 2 (from the outside in), all genes (reverse and forward strand, respectively) color-coded by function; circle 3, predicted highly expressed (PHX) and predicted alien (PA) genes; circle 4, potentially eukaryotic or archaeal genes; circle 5, sulfatase, C1-metabolism, heterolactic acid fermentation, catalase, and ATP-synthase genes; and circle 6, transposase, integrase, and clusters of putative laterally acquired genes. The origin of replication (ORI) and the terminus (TER) are indicated.

thermore, it seems to be capable of synthesizing all amino acids (*Appendix 2*, which is published as supporting information on the PNAS web site).

Recent growth studies did not provide evidence that *Pirellula* sp. strain 1 can grow under nitrate-reducing or fermentative conditions. Interestingly, however, all genes required for heterolactic acid fermentation are present (Fig. 2 and *Appendix 3*, which is published as supporting information on the PNAS web site). Expression of the genes is likely, because the key enzyme lactate dehydrogenase has been predicted to be highly expressed on the basis of codon usage. Furthermore, both *Pirellula marina* and *Planctomyces limnophilus* have been described to be capable of carbohydrate fermentation (30). This capability could explain why planctomycetes were found in anoxic marine and freshwater sediments and anoxic terrestrial habitats (1–3).

Motility. The life cycle of *Pirellula* sp. strain 1 consists of an aggregate-forming sessile form and a motile swarmer cell. In the genome all genes for a functional flagellum could be determined, whereas except for *cheY*, essential genes for chemotaxis such as *cheA*, *cheB*, *cheR*, *cheW*, and *cheZ* could not be identified.

Transporters. As a free-living organism, *Pirellula* sp. strain 1 was expected to have a wide range of transporters (29). A comparative study with Pfam profiles for ABC-transporters against all 70 publicly available prokaryotic genomes revealed that the 55 ABC-transporters found in *Pirellula* sp. strain 1 is close to the calculated mean of 49 transporters. In comparison with other free-living bacteria this is only about one-third of the 148 ABC-transporters found with the same method in *Streptomyces coelicolor* A3 (2), but similar to the 45 transporters of *Caulobacter crescentus* (*Appendix 4*, which is published as supporting information on the PNAS web site). Annotation revealed that ABC-transporters for ribose, oligopeptides, phosphate, manganese, nitrate, and sodium are present, but only one phosphotransferase system (PTS) specific for fructose could be identified. Exceptional is a set of ORFs for nitrate transport and nitrate/nitrite reduction that were predicted to be highly expressed (PHX) on the basis of codon usage (24). This set of ORFs could be essential in nitrogen-limited marine systems (31).

Stress response. The genome harbors homologues to superoxide dismutase and all four known types of catalases (Fig. 2). Methionine-sulfoxide reductases are present to repair oxidized methionine. By synthesis of a cytochrome *d* oxidase as an alternative to the regular cytochrome aa_3 *Pirellula* sp. strain 1 should be able to cope with low oxygen concentrations.

Many mechanisms are present to reduce the damaging effect of UV radiation. Besides the genes for SOS response (*recA*, *lexA*, *uvrA*, *uvrB*, and *uvrC*), *Pirellula* sp. strain 1 has a photolyase gene organized in an operon-like manner with genes encoding phytoene dehydrogenase and phytoene synthase. Probably, UV stress triggers the biosynthesis of a UV-protection carotenoid, which might be responsible for the pinkish color of *Pirellula* sp. strain 1. Regarding temperature stress, *Pirellula* sp. strain 1 has many homologues to heat and cold shock DNA-binding proteins. Detoxification seems to take place by means of unspecific export systems like cation efflux systems of the AcrB/AcrD/AcrF family or unspecific multidrug export systems for hydrophobic compounds. A cytochrome

P450 mono-oxygenase and an epoxide hydrolase are present for the detoxification of xenobiotics. Specific detoxification involves mercury reductase, arsenate reductase, and the ArsA arsenite-exporting ATPase. The harmful effect of D-tyrosine binding to tyrosyl-tRNA is minimized by D-tyrosyl-tRNA^{Tyr} deacylase. In addition, *Pirellula* sp. strain 1 has a gene encoding a bacterial hemoglobin, which is believed to detoxify NO by oxidation. Finally, the genome has some homologues to carbon-starvation proteins, including DNA-protection proteins.

Antibiotics. Several ORFs potentially coding for polyketide antibiotics and nonribosomal polypeptide antibiotics or a mixture of both have been determined in the genome of *Pirellula* sp. strain 1. In general the ORFs are unusually long, coding for proteins from 916 up to 3,665 aa.

Sulfatases. The *Pirellula* sp. strain 1 genome harbors 110 genes encoding proteins with significant similarity to prokaryotic (82 genes; 75%) and eukaryotic (28 genes; 25%) sulfatases. For instance, similarity was found to alkylsulfatase of *Pseudomonas aeruginosa*, to arylsulfatases of *Pseudomonas* sp., to mucin-desulfating sulfatase of *Prevotella* sp., and to archaeal arylsulfatase, as well as to mammalian iduronate-2-sulfatase and arylsulfatases A and B. In comparison, the analysis of 70 published prokaryotic genomes with a specific Pfam profile revealed a maximum of only 6 sulfatases found in the *Pseudomonas aeruginosa* PAO1 genome. In *Pirellula* sp. strain 1, the sulfatase genes are distributed across the genome in 22 clusters containing two to five genes (Fig. 2).

In *Pirellula* sp. strain 1, all detected sulfatase gene products, except for the three alkylsulfatases, are of the cysteine type; 85 (79%) of them show the canonical CXPXR motif and are hence considered as potentially functional (32). In contrast to the known bacterial cysteine-type sulfatases, which are cytosolic enzymes (e.g., arylsulfatase AtsA of *Pseudomonas aeruginosa*), for 26 (31%) of the 85 potentially functional sulfatases in *Pirellula* sp. strain 1 a signal peptide is predicted with high probability, suggesting an extracytosolic localization of the proteins.

The fact that the sulfatase genes in *Pirellula* sp. strain 1 outnumber those present in all other known prokaryotic genomes by two orders of magnitude raises the question about their physiological role. Bacterial sulfatases seem to be primarily used in sulfur scavenging, and their expression is known to be tightly regulated and dependent on the sensing of sulfur deprivation (32). As marine systems are characterized by high inorganic sulfate concentrations, sulfur limitations should not occur. Therefore, *Pirellula* sp. strain 1 might use its sulfatases to access more effectively the carbon skeleton of sulfated compounds as an energy source rather than to meet its sulfate requirements. Cleavage of sulfate esters in sulfated high molecular weight glycoproteins (mucins) to increase the efficiency of polymer degradation by other enzymes has been described for *Prevotella* sp. RS2. Seven of the 110 sulfatase genes in *Pirellula* sp. strain 1 encode proteins with high similarity to mucin-desulfating sulfatase of *Prevotela* sp. RS2, an enzyme that seems to be specific for the cleavage of sulfate from *N*-acetylglucosamine 6-sulfate in mucin side chains (33). Remarkably, one of the seven genes in *Pirellula* sp. strain 1 is located next to a gene encoding a protein with some similarity to *N*-acetylglucosamine-6-phosphate deacetylase, a protein involved in metabolism of *N*-acetylglucosamine. This compound is known to support growth of *Pirellula* sp. strain 1 as a sole source of nitrogen and organic carbon.

Furthermore, 17 (20%) of the potentially functional *Pirellula* sp. strain 1 sulfatase gene products exhibit similarity to proteins hydrolyzing sulfate ester bonds of sugar units in heteropolysaccharides. Interestingly, chondroitin sulfate is an excellent growth substrate for *Pirellula* sp. strain 1 (H.S., unpublished results).

In addition, 25 (29%) of the 85 genes coding for potentially functional sulfatases in *Pirellula* sp. strain 1 are found in a genomic context of genes involved in carbohydrate metabolism, e.g., carrageenase and other glycosyl hydrolases. For 3 of these 25 sulfatases as well as for an additional 4 potentially functional sulfatases a high expression level is predicted (PHX).

These data suggest that sulfatases are metabolically important in *Pirellula* sp. strain 1 and could play a role in the efficient degradation of sulfated glycopolymers. Such compounds (e.g., carrageen) are abundant in marine environments in the form of phytodetrital macroaggregates ("marine snow"), and planctomycetes have been shown to be components of the microbial communities on such aggregates (6).

C1 metabolism. Another intriguing feature derived form the genome sequence is the genetic potential for degrading C1 compounds (Fig. 2). Although *Pirellula* sp. strain 1 is not capable of utilizing methanol, methylamine, or methylsulfonate, and genes encoding enzymatic activities for the primary oxidation steps of such C1 compounds could not be identified in the genome, functional prediction revealed all enzymes necessary for the oxidation of formaldehyde to formate. The predicted pathway in *Pirellula* sp. strain 1 (M.B., unpublished results) resembles very closely a pathway of formaldehyde oxidation/detoxification in methylotrophic proteobacteria. These tetrahydromethanopterin (H_4MPT)-dependent enzymes were previously thought to be unique for anaerobic methanogenic and sulfate-reducing *Archaea*. Recently, however, they have been shown to play an essential physiological role in methylotrophic proteobacteria (34). *Pirellula* sp. strain 1 is, to our knowledge, the first bacterial organism outside the proteobacterial division found to contain genes encoding H_4MPT-dependent enzymes. In context with the fact that planctomycetes constitute an independent phylum, our finding revives the discussion on the evolutionary processes leading to the distribution of these archaeal genes.

Cell Biology. ***Cell wall.*** Planctomycetes are the only group of free-living members of the domain *Bacteria* known so far that have no peptidoglycan in their cell walls. Instead, they are stabilized by a protein sacculus with disulfide bonds (35). A systematic investigation for genes involved in peptidoglycan biosynthesis revealed that *murB*, *murE*, *murG*, *ddlA*, and *upk* (*bacA*) are present. Furthermore, *Pirellula* sp. strain 1 possesses the gene *glmS*, which is involved in the formation of *N*-acetyl-D-glucosamine, a precursor for peptidoglycan biosynthesis. Other key enzymes, such as MurA, MurC, MurD, MurF, and DdaA for the final cross-linking of peptidoglycan, are notably absent from the *Pirellula* sp. strain 1 genome. The preservation of at least some of the genes of the peptidoglycan synthesis pathway suggests that *Pirellula* sp. strain 1 is not a descendant from a bacterium evolving before the invention of peptidoglycan, as proposed earlier (36). It rather seems that after the development of a proteinaceous cell envelope in planctomycetes, genes for peptidoglycan biosynthesis were successively lost.

Membrane. It is noteworthy that the *Pirellula* sp. strain 1 genome harbors all genes required for biosynthesis of lipid A, the major constituent of the lipopolysaccharide (LPS) layer in Gram-negative bacteria. The presence of these genes is in line with earlier reports of presence of lipid A with unusual portions of long-chain 3-OH fatty acids in members of the *Pirellula*/*Planctomyces* group (12). Nevertheless, the key enzymes necessary for the biosynthesis of an O-specific side chain (O-antigen ligase; O-antigen polymerase) are absent from the *Pirellula* sp. strain 1 genome. The presence of lipid A and homologues to the flagellar L- and P-ring protein suggests that the cell envelope of planctomycetes was converted from a Gram-negative type of cell wall. Furthermore, *Pirellula* sp. strain 1 lacks the signature sequences in the ribosomal protein S12 and SecF typical for low and high G+C Gram-positive bacteria, respectively (37).

Compartmentalization. One of the most striking properties of the *Planctomycetales* is their complex internal structures (11). Ribosomes are located only within the riboplasm, therefore proteins targeted to the paryphoplasm have to overcome the intracytoplasmic membrane. This requires effective protein targeting. A com-

ECOLOGY

parative search with Pfam profiles against all publicly available prokaryotic genomes revealed that *Pirellula* sp. strain 1 has the highest number of hits to *secA* (3), the general secretory pathway (GSP) type II F-domain (6), and the GSP type II/III secretion system protein (9). Furthermore *Pirellula* sp. strain 1 has 1,271 genes with predicted signal peptides. In comparison with all other genomes investigated this number is matched only by *Pseudomonas aeruginosa* PA01 (1,277). When normalized on genome size, *Pirellula* sp. strain 1 comprises 178 signal peptides per megabase, which is again at the high end. An equally high proportion of proteins with signal peptides was also found in another "life cycle" bacterium, *C. crescentus*, which has 182 signal peptides per megabase (*Appendix 5*, which is published as supporting information on the PNAS web site). A comparison for Tat (twin arginine translocation) signal peptides showed that *Pirellula* sp. strain 1 has the highest number of all genomes investigated (135; 18.9 per megabase). Effective protein targeting might be the basis for the polar organization of *Pirellula* sp. strain 1 and for distinct features (e.g., stalks, holdfast substance, crateriform structures) present only in certain regions of the cells.

Cell division. Cell division involves a plethora of genes. The most important are *ftsZ*, *ftsA*, *ftsI*, *ftsL*, *ftsQ*, *ftsN*, *zipA*, and *ftsW* (38). Surprisingly, with the exception of *ftsK*, all genes are absent from the genome of *Pirellula* sp. strain 1. A lack of the key enzyme FtsZ, the major constituent of the septal replication ring, has so far been reported only for chlamydiae, the *Crenarchaeota*, and *Ureaplasma urealyticum* (38). Not much is known about replication in planctomycetes, especially how the cell compartments are distributed to the daughter cells. Altogether, cell division in *Pirellula* sp. strain 1 must follow a different pathway than reported for the model organisms *E. coli* and *B. subtilis*.

Life cycle. *Pirellula* sp. strain 1 exhibits a life cycle similar to *C. crescentus* (39). Surprisingly, no homologue to the master response regulator protein CtrA has been found within the genome of *Pirellula* sp. strain 1. However, the origin in *Pirellula* sp. strain 1 contains some patterns similar to the CtrA binding site pattern $TTAAN_7TTAA$ upstream of *dnaN* (e.g., $TTAAN_7AAAC$), which might indicate a similar control mechanism.

Regulation. Analysis of the *Pirellula* sp. strain 1 genome with 116 relevant Pfam 7.2 family models shows only 135 genes with motifs for predicted transcriptional regulators. No evidence for eukaryotic-like transcriptional regulators could be found. There are 68 response regulators, which allow microorganisms to respond to changes in their environment, but common bacterial regulators such as LysR are absent or underrepresented. A comparative analysis in all currently available bacterial genomes was performed. The results confirm earlier findings that the proportion of the genome encoding transcriptional regulators increases with genome size (29, 40) (*Appendix 6*, which is published as supporting information on the PNAS web site). Nevertheless with a genome size of more than 7 megabases and only 2% predicted regulatory genes, *Pirellula* sp. strain 1 clearly contradicts this trend. It remains to be determined whether these results reflect a lack of knowledge in the diversity of regulatory proteins or even unknown gene regulation mechanisms. For example, a unique family of predicted DNA-binding proteins has been reported in the genome of *S. coelicolor* A3(2) (40), which might constitute a family of *Streptomyces*-specific transcriptional regulators. Regulation of metabolic capacities of *Pirellula* sp. strain 1 were recently addressed by a proteomic approach, which revealed differential protein patterns in response to carbohydrates used for growth (41).

Sigma factors. *Pirellula* sp. strain 1 encodes for a total of 51 sigma factors, including 16 ECF (extracytoplasmic function) sigma factors. Currently, with 65 sigma factors, only *S. coelicolor* has a higher number. Therefore, it seems that in *Pirellula* sp. strain 1 gene regulation is based to a greater extent on altering the promoter specificity of the RNA polymerase. Further support for this hypothesis comes from the observation that genes for essential pathways such as purine or biotin and amino acid biosynthesis are not organized in operons, as is known for most of the prokaryotes. This scattering, together with the split rRNA operon, requires a different way of regulation.

Evolution. ***Phylogeny.*** The currently accepted bacterial systematics based on 16S rRNA assigns the *Planctomycetales* with the genera *Pirellula*, *Planctomyces*, *Isosphaera*, and *Gemmata* as an independent monophyletic phylum (9). For evaluation, phylogenetic trees for commonly used alternative markers such as the 23S rRNA, elongation factors Tu and G, ATP-synthase subunits, RecA, heat shock proteins HSP60 and HSP70, RNA polymerase, and DNA gyrase subunits as well as the aminoacyl-tRNA synthetases of *Pirellula* sp. strain 1 were constructed. A common *Pirellula–Gemmata* (*G. obscuriglobus* UQM_2246) cluster separated from all other phyla or major subgroups is seen in trees derived from 23S rRNA, elongation factors, ATP-synthase subunits α and β, DNA gyrase subunits A and B, the heat shock proteins HSP60 and HSP70, the recA protein, and most of the tRNA synthetases. Multiple copies are present in the case of ATP-synthase α and β as well as the DNA gyrase A and B subunits. The sequence divergence of the duplicates corresponds to the phylum level. Thus the individual monophyletic *Pirellula–Gemmata* pairs are separated from other phyla in the respective trees. However, the multiple copies of heat shock proteins HSP60 and HSP70 cluster in common groups with the *Pirellula* and *Gemmata* proteins phylogenetically intermixed. Keeping in mind the possible pitfalls and the differences in resolving power when functional genes are used for tree reconstruction (42), the status as an independent phylum affiliated to the domain *Bacteria* is clearly supported in the majority of analyses. A deepest branching of planctomycetes within the bacterial subtree as reported recently (10) is not convincingly supported by any of the markers, as evaluated by applying different tree-building methods, parameters, and significance tests.

ATP-synthases. *Pirellula* sp. strain 1 is the only bacterium described so far that contains two F_oF_1 ATP-synthases (Fig. 2). By gene organization (a, c, b, δ, α, γ, β, ε) and similarity one ATP-synthase resembles the "standard" ATP-synthase that is common in all *Bacteria* (43). Nevertheless, the J gene found in all *Bacteria* except for *Thermotoga maritima* is missing. The second set of F_oF_1-ATP-synthase genes has a different operon structure with β and ε genes followed by the J, X, a, c, and b genes and the α, γ genes (*Appendix 7*, which is published as supporting information on the PNAS web site). The strong conservation and the high similarity of this ATP-synthase operon to *Methanosarcina barkeri* indicates a lateral gene transfer event. It remains unclear whether the genes were transferred from the archaeal to the bacterial domain or vice versa.

Phylogenetic Distribution of Best BLAST Hits. All 7,325 potential proteins (ORFs) in the *Pirellula* sp. strain 1 genome were searched against the National Center for Biotechnology Information nonredundant database. By setting the cut-off for the BLASTP expectation value $\leq 1 \times 10^{-3}$, significant hits could be obtained for 3,380 genes. Of these genes, 83% were assigned to the domain *Bacteria*, 9% and 8% to the domains *Archaea* and *Eukarya*, respectively (Fig. 2). The large number of hits to eukaryotes is exceptional. *T. maritima*, for example, shows 24% hits to *Archaea* but only 2% to *Eukarya*, and in *E. coli* only 0.4% of the best hits assign to *Archaea* or *Eukarya*. Nevertheless, among the 270 genes found in *Pirellula* sp. strain 1 no trend for a certain organism of origin or a distinct functional category could be detected (*Appendix 8*, which is published as supporting information on the PNAS web site). Furthermore, the two genes for the integrin α-V and inter-α-trypsin inhibitor found in *Pirellula marina* and *G. obscuriglobus*, considered to be typical for *Eukarya* (44), could not be detected in the *Pirellula* sp. strain 1 genome.

Relationship to *Chlamydia*. The phylogenetic analysis of a comprehensive set of markers with different tree-building methods and confidence tests revealed that only the trees for DNA gyrase, RNA polymerase C, and lysyl- and valyl-tRNA synthetase supported a moderate relationship to *Chlamydia*. Using only distance matrix methods, we also found a remote relationship to *Chlamydia* in some ribosomal proteins, DnaA, Hsp60, Rho, the protein component of RNase P, and CTP synthase. According to the indel method of Gupta and Griffiths (37), *Pirellula* sp. strain 1 branches off between spirochetes and *Chlamydia*.

As already mentioned, *Pirellula* sp. strain 1 and *Chlamydia* share some noteworthy features: both lost their peptidoglycan cell walls in favor of a proteinaceous cell envelope. Furthermore, *Chlamydia* and *Pirellula* sp. strain 1 have two *dnaA* copies and lack *ftsZ*, indicating an unknown mode of cell division. In addition, the ribosomal protein L30 is missing from the *spc* operon in *Chlamydia* as well as in *Pirellula* sp. strain 1. Although the complete genome sequence does not confirm a close relationship of *Chlamydia* and planctomycetes, it is still possible that an ancient relatedness exists. In this case the extremely different habitats might have blurred the genetic records.

Conclusions. The genome sequence of *Pirellula* sp. strain 1 has revealed insights into adaptations of free-living marine bacteria. Environmental versatility demands an enhanced genetic complexity harbored by larger genomes (29, 40). In case of *Pirellula* sp. strain 1 no indication for a recent expansion of the genome could be detected. In contrast, except for some irregularities (Fig. 1), the plot of the cumulative GC-skew of *Pirellula* sp. strain 1 is very "smooth" with clear maxima and minima (*Appendix 1*). From the genome features it is now possible to propose a certain lifestyle of *Pirellula* sp. strain 1. In the water column *Pirellula* sp. strain 1 gains energy from the aerobic oxidation of mono- or disaccharides derived from the cleavage of sulfated polymers produced by algae. Protection systems for UV and the expression of carotenoids protect *Pirellula* sp. strain 1 from irradiation at the water surface. Nitrate transporters support growth even under limited-nitrogen conditions common in continental shelf areas. The holdfast substance enables *Pirellula* sp. strain 1 to attach to nutrient-rich marine snow particles slowly sedimenting to the sea floor. When the bacterium reaches the sediment the expression of cytochrome *d* oxidase allows survival under low oxygen conditions. Anoxic conditions force *Pirellula* sp. strain 1 to switch to heterolactic acid fermentation or pathways involving formaldehyde conversion if not to support growth, then at least to allow basic maintenance metabolism. With the expression of genes for carbon starvation *Pirellula* sp. strain 1 can even outlast periods of nutrient depletion. The formation of swarmer cells helps *Pirellula* sp. strain 1 to reach for new resources. The high number of sigma factors allows the tight control of gene expression under changing environmental conditions.

Exceptionally interesting will be the genome comparison of *Pirellula* sp. strain 1 with the freshwater isolate *G. obscuriglobus* UQM 2246 (45), currently being sequenced by The Institute for Genomic Research, and *Gemmata* sp. Wa1-1, being sequenced by Integrated Genomics. It will not only reveal common traits of the *Planctomycetales* but also give hints for the specific adaptations to the different habitats and the origin of the unique combination of morphological and ultrastructural properties.

We acknowledge Jörg Wulf for DNA extraction; Sven Klages, Ines Marquardt, Silvia Lehrack, Birol Köysüren, Özlem Ogras, and Roman Pawlik for technical assistance; Hauke Pfeffer, Michael Richter, Thomas Otto, Tim Frana, Stella Koufou, Andreas Schmitz, and Jost Waldmann for their expert help in annotation and programming; Folker Meyer and Alexander Goesmann for their continuous support with GenDB; Anke Meyerdierks and Erich Lanka for excellent critical discussions; and Hans Lehrach for his continuous support and critical discussions. Major funding of this project was provided by the Federal Ministry of Education and Research. Further support came from the Max Planck Society. *G. obscuriglobus* is being sequenced with support from the U.S. Department of Energy.

1. Wang, J., Jenkins, C., Webb, R. I. & Fuerst, J. A. (2002) *Appl. Environ. Microbiol.* **68,** 417–422.
2. Llobet-Brossa, E., Rosselló-Mora, R. & Amann, R. (1998) *Appl. Environ. Microbiol.* **64,** 2691–2696.
3. Miskin, I., Farrimond, P. & Head, I. (1999) *Microbiology* **145,** 1977–1987.
4. Neef, A., Amann, R., Schlesner, H. & Schleifer, K.-H. (1998) *Microbiology* **144,** 3257–3266.
5. Fuerst, J. A., Gwilliam, H. G., Lindsay, M., Lichanska, A., Belcher, C., Vickers, J. E. & Hugenholtz, P. (1997) *Appl. Environ. Microbiol.* **63,** 254–262.
6. DeLong, E. F., Franks, D. G. & Alldredge, A. L. (1993) *Limnol. Oceanogr.* **38,** 924–934.
7. Strous, M., Fuerst, J. A., Kramer, E. H. M., Logemann, S., Muyzer, G., van de Pas-Schoonen, K. T., Webb, R., Kuenen, J. G. & Jetten, M. S. M. (1999) *Nature* **400,** 446–449.
8. Alldredge, A. L. (2000) *Limnol. Oceanogr.* **45,** 1245–1253.
9. Garrity, G. M. & Holt, J. G. (2001) in *Bergey's Manual of Systematic Bacteriology*, eds. Boone, D. R. & Castenholz, R. W. (Springer, New York), 2nd Ed., Vol. 1, pp. 119–166.
10. Brochier, C. & Philippe, H. (2002) *Nature* **417,** 244–244.
11. Lindsay, M. R., Webb, R. I., Strous, M., Jetten, M. S., Butler, M. K., Forde, R. J. & Fuerst, J. A. (2001) *Arch. Microbiol.* **175,** 413–429.
12. Kerger, B. D., Mancuso, C. A., Nichols, P. D., White, D. C., Langworthy, T., Sittig, M., Schlesner, H. & Hirsch, P. (1988) *Arch. Microbiol.* **149,** 255–260.
13. Fuerst, J. A. (1995) *Microbiology* **141,** 1493–1506.
14. Schmidt, J. M. & Starr, M. P. (1978) *Curr. Microbiol.* **1,** 325–330.
15. Schlesner, H. & Hirsch, P. (1984) *Int. J. Syst. Bacteriol.* **34,** 492–495.
16. Schlesner, H. (1994) *Syst. Appl. Microbiol.* **17,** 135–145.
17. Ewing, B. & Green, P. (1998) *Genome Res.* **8,** 186–194.
18. Staden, R., Beal, K. F. & Bonfield, J. K. (2000) *Methods Mol. Biol.* **132,** 115–130.
19. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999) *Nucleic Acids Res.* **27,** 4636–4641.
20. Badger, J. H. & Olsen, G. J. (1999) *Mol. Biol. Evol.* **16,** 512–524.
21. Frishman, D., Mironov, A., Mewes, H. W. & Gelfand, M. (1998) *Nucleic Acids Res.* **26,** 2941–2947.
22. Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A. & Mewes, H. W. (2001) *Bioinformatics* **17,** 44–57.
23. Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J. & Giegerich, R. (2001) *Nucleic Acids Res.* **29,** 4633–4642.
24. Karlin, S. & Mrazek, J. (2000) *J. Bacteriol.* **182,** 5238–5250.
25. Jensen, L. J., Friis, C. & Ussery, D. W. (1999) *Res. Microbiol.* **150,** 773–777.
26. Fraser, C. M., Eisen, J. A. & Salzberg, S. L. (2000) *Nature* **406,** 799–803.
27. Mewes, H. W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S. G., *et al.* (1997) *Nature* **387,** 7–8.
28. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., *et al.* (1997) *Nature* **390,** 249–256.
29. Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., Brinkman, F. S. L., Hufnagle, W. O., Kowalik, D. J., Lagrou, M., *et al.* (2000) *Nature* **406,** 959–964.
30. Schlesner, H. & Stackebrandt, E. (1986) *Syst. Appl. Microbiol.* **8,** 174–176.
31. Agusti, S., Duarte, C. M., Vaque, D., Hein, M., Gasol, J. M. & Vidal, M. (2001) *Deep-Sea Res. Part II* **48,** 2295–2321.
32. Kertesz, M. A. (1999) *FEMS Microbiol. Rev.* **24,** 135–175.
33. Wright, D. P., Knight, C. G., Parker, S. G., Christie, D. L. & Roberton, A. M. (2000) *J. Bacteriol.* **182,** 3002–3007.
34. Chistoserdova, L., Vorholt, J. A., Thauer, R. K. & Lidstrom, M. E. (1998) *Science* **281,** 99–102.
35. Liesack, W., König, H., Schlesner, H. & Hirsch, P. (1986) *Arch. Microbiol.* **145,** 361–366.
36. Stackebrandt, E., Ludwig, W., Schubert, W., Klink, F., Schlesner, H., Roggentin, T. & Hirsch, P. (1984) *Nature* **307,** 735–737.
37. Gupta, R. S. & Griffiths, E. (2002) *Theor. Popul. Biol.* **61,** 423–434.
38. Margolin, W. (2000) *FEMS Microbiol. Rev.* **24,** 531–548.
39. Marczynski, G. T. & Shapiro, L. (2002) *Annu. Rev. Microbiol.* **56,** 625–656.
40. Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quail, M. A., Kieser, H., Harper, D., *et al.* (2002) *Nature* **417,** 141–147.
41. Rabus, R., Gade, D., Helbig, R., Bauer, M., Glöckner, F. O., Kube, M., Schlesner, H., Reinhardt, R. & Amann, R. (2002) *Proteomics* **2,** 649–655.
42. Ludwig, W. & Klenk, H. P. (2001) in *Bergey's Manual of Systematic Bacteriology*, eds. Boone, D. R. & Castenholz, R. W. (Springer, New York), 2nd Ed., Vol. 1, pp. 49–65.
43. Deckers-Hebestreit, G. & Altendorf, K. (1996) *Annu. Rev. Microbiol.* **50,** 791–824.
44. Jenkins, C., Kedar, V. & Fuerst, J. (2002) *Genome Biol.* **3,** 0031.1–0031.11.
45. Franzmann, P. D. & Skerman, V. B. D. (1984) *Antonie Leeuwenhoek* **50,** 261–268.
46. Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R., *et al.* (2003) *Nucleic Acids Res.* **31,** 2187–2195.

ECOLOGY

2

Reevaluation of the phylogenetic position of the *Planctomycetes* by means of concatenated ribosomal protein sequences, DNA-directed RNA polymerase subunit sequences and whole genome trees

Hanno Teeling, Thierry Lombardot, Margarete Bauer, Wolfgang Ludwig and Frank Oliver Glöckner

Int J Syst Evol Microbiol, published online December 5th, (2003) - in press

Reevaluation of the phylogenetic position of the *Planctomycetes* by means of concatenated ribosomal protein sequences, DNA-directed RNA polymerase subunit sequences and whole genome trees

Hanno Teeling*, Thierry Lombardot*, Margarete Bauer*, Wolfgang Ludwig[†] and Frank Oliver Glöckner*[‡]

*Max Planck Institute for Marine Microbiology, Department of Molecular Ecology, Genomics Group, D-28359 Bremen, Germany

[†] Department of Microbiology, Technical University Munich, D-85350 Freising, Germany

[‡] to whom correspondence should be addressed

SUMMARY

In recent years, *Planctomycetes* were recognized as a phylum of environmentally important bacteria with habitats ranging from soil and freshwater to marine ecosystems. The *Planctomycetes* form an independent phylum within the bacterial domain, whose exact phylogenetic position still remains controversial. With one finished planctomycete genome at hand and further genome sequences in process, it is now possible to reevaluate the phylogeny of the *Planctomycetes* based on multiple genes and genome trees in addition to single genes like the 16S ribosomal RNA or the elongation factor Tu. Here we present evidence based on the concatenated amino acid sequences of ribosomal proteins and DNA-directed RNA polymerase subunits from more than 90 publicly available genomes including *Pirellula* sp. strain 1 and *Gemmata obscuriglobus* UQM 2246^T that support a relationship of *Planctomycetes* and *Chlamydia*. Affiliation of both groups was reasonably stable regarding site selection since, during stepwise filtering of less conserved sites from the alignments, it was only broken when rigorous filtering was applied. In few cases, the *Planctomycetes* shifted to a deep branching position adjacent to the *Thermotoga/Aquifex* clade. These findings are in agreement with recent publications, but the deep branching position of the *Planctomycetes* was depending on site selection and treeing algorithm and thus not stable. A genome tree calculated from normalized BLASTP scores did not confirm a close relationship of *Planctomycetes* and *Chlamydia* but also indicated that the *Planctomycetes* do not emerge at the very root of the *Bacteria*. Therefore, these analyses rather contradict a deepest branching position of the *Planctomycetes* within the bacterial domain and reaffirm their earlier proposed relatedness to the *Chlamydia*.

INTRODUCTION

The phylum *Planctomycetes* (Garrity *et al.*, 2002) consists of bacteria, whose members share an astounding cell morphology making them unique within the bacterial domain. Characteristic for *Planctomycetales* are a polar cell organization and a life-cycle with a yeast-like budding mechanism (Schlesner, 1994). The cell walls of *Planctomycetes* are composed of proteins rather than peptidoglycan (König *et al.*, 1984; Liesack *et al.*, 1986; Giovannoni *et al.*, 1987). These cell walls exhibit crateriform structures, small pits which appear as electron-dense circular regions either on the reproductive pole (*Pirellula* spp.) or on the entire cellular surface (*Planctomyces* spp.) (Liesack *et al.*, 1986). The most striking morphological feature of *Planctomycetes*, however, is their compartmentalization. The cytoplasm of *Planctomycetes* is divided by an intracytoplasmic membrane into the peripheral

ribosome-free paryphoplasm and the inner ribosome-containing riboplasm. (Lindsay *et al.*, 1997; Lindsay *et al.*, 2001). The DNA of *Planctomycetes* is highly condensed and forms a nucleoid within the riboplasm, which in case of *Gemmata* sp. is surrounded by an additional double membrane (Lindsay *et al.*, 2001).

Planctomycetes are widespread and of environmental importance (Fuerst, 1995; Ward-Rainey *et al.*, 1996; Gade *et al.*, 2003). They have been found to be abundant in various habitats including terrestrial and aquatic habitats differing in salinity (from hypersaline to freshwater), oxygen availability (from the oxic watercolumn to anoxic sediments), trophic level (from oligotrophic lakes to eutrophic wastewater) and temperature (from cold-water marine snow to hot springs) (Giovannoni *et al.*, 1987; Kerger *et al.*, 1988; DeLong, 1993; Schlesner, 1994; Ward *et al.*, 1995; Vergin *et al.*, 1998; Miskin *et al.*, 1999; Wang *et al.*, 2002). *Planctomycetes* have even been isolated from the digestive tracts of crustaceans (Fuerst, 1995; Fuerst *et al.*, 1997).

In addition, *Planctomycetes* have interesting metabolic capabilities, e.g. the postulated anammox process, the anaerobic conproportionation of ammonia and nitrite to dinitrogen (Strous *et al.*, 1999; Schmid *et al.*, 2001).

Despite their outstanding morphology, ubiquitous occurrence and interesting physiology, the phylogeny of *Planctomycetes* still awaits resolution. All studies conducted so far agree on the phylogenetic distinctness of the *Planctomycetes* (Bomar *et al.*, 1988; Ward *et al.*, 2000) but they disagree on the position of the phylum within the tree of life. Early analyses based on 16S ribosomal RNA (rRNA) sequences suggested a distant relationship to *Chlamydia* (Weisburg *et al.*, 1986; Liesack *et al.*, 1992), whereas such a relationship could not be confirmed in later studies based on 16S/23S rRNA (Ward *et al.*, 2000), *dnaK* (Ward-Rainey, 1997) and EF-Tu (Jenkins & Fuerst, 2001). The broad level of sequence divergence within the 5S and 16S rRNA genes of *Planctomycetes* has been interpreted either as an indication that they are rapidly evolving (i.e. contain tachytelic DNA) (Woese, 1987; Bomar *et al.*, 1988; Liesack *et al.*, 1992) or that they represent a very deep branching phylum (Stackebrandt *et al.*, 1984). In two recent studies based on the slowly evolving positions of the 16S rRNA gene, the *Planctomycetes* have even been described as the deepest branching phylum within the bacterial domain (Brochier & Philippe, 2002) or as branching off deeply after the *Thermotoga/Aquifex* clade (Di Giulio, 2003).

With the recently finished genome of *Pirellula* sp. strain 1 (to be validly described as 'Rhodopirellula baltica') (Glöckner *et al.*, 2003), the nearly finished genome of *Gemmata obscuriglobus* UQM 2246^T, and the availability of more than 100 publicly available complete genome sequences, we are now for the first time in a position to exploit the wealth of information emerging from entire genomes to reassess the phylogeny of the *Planctomycetes*. In this study the results of two genomic approaches for phylogenetic tree reconstruction are compared: Concatenation of the amino acid (aa) sequences of subunits of large information processing proteins (ribosomal and DNA-directed RNA polymerase subunits) and genome trees based on normalized BLASTP scores (Clarke *et al.*, 2002).

METHODS

Sequences. The aa sequences of ribosomal proteins and DNA-directed RNA polymerase subunits were extracted from all bacterial genome sequences that were publicly available on the NCBI website in midyear 2003 (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/). Screening for ribosomal proteins was done with corresponding Pfam-profiles (http://www.sanger.ac.uk/Software/Pfam/) using hmmer 2.2g and with BLASTP (ftp://ftp.ncbi.nih.gov/blast), respectively. Remaining subunits were extracted directly from genome sequence annotations. Subunits of the DNA-directed RNA polymerase were searched with BLASTP or extracted directly from genome sequence annotations. Preliminary sequence data of *Gemmata obscuriglobus* UQM 2246^T was obtained from the website of 'The Institute for Genomic Research' (http://www.tigr.org). Sequencing of *Gemmata obscuri-*

globus UQM 2246^T is accomplished with support from the U.S. Department of Energy.

For the genome tree approach, EMBL-formatted annotated genome sequences were obtained from the EMBL website (ftp://ftp.ebi.ac.uk/pub/databases/embl/genomes) and imported into a local installation of the GenDB annotation system for further analysis (Meyer *et al.*, 2003). The final dataset in the GenDB-MySQL database comprised 85 species accounting for 231,509 open reading frames (ORFs).

Alignments. For the analysis of the ribosomal proteins, sequences that are known to be prone to lateral gene transfer (LGT), that have paralogues or that were absent in some of the species were excluded from further analysis. The resulting dataset comprised sequences of the following 39 ribosomal proteins from 91 bacterial species: *rpl1-rpl4, rpl6, rpl7/12, rpl9-rpl11, rpl13-rpl23, rpl27, rpl29, rpl34, rps2-rps9, rps11-rps13, rps15, rps17-rps20*. These were aligned independently using clustalW 1.83 (with settings gapopen 10 and gapext 0.2) and subsequently concatenated using a custom-made PERL script (9,377 aa positions).

For the analysis of the DNA-directed RNA polymerases, the aa sequences of the main subunits *rpoA*, *rpoB* and *rpoC* were extracted from whole genome sequences of 95 bacterial species. The other subunits were left out because *rpoC*1 is restricted to *Cyanobacteria*, *rpoE* is restricted to Gram-positives and *rpoZ* is rather small and seems to be absent from many genomes. The sequences of the main subunits were aligned independently (clustalW 1.83 with settings gapopen 10 and gapext 0.2) and then concatenated (5,277 aa positions).

Columns at which gaps were maximal were omitted from both initial alignments. Afterwards, nine filtered alignments were derived from each initial alignments by successively discarding columns with less than 10, 20, 30, 40, 50, 60, 70, 80 and 90% sequence conservation.

For the analysis with MrBayes (see below) the alignment with 30% positional conservation filtering was chosen. Species with ambiguously aligned stretches of sequence were removed from the datasets, resulting in final alignments of 83 species for the ribosomal proteins and 93 species for the DNA-directed RNA polymerase subunits.

Phylogenetic analysis. For each alignment, neighbor-joining, parsimony and maximum-likelihood trees were calculated using the programs PROTDIST/NEIGHBOR, PROTPARS and ProML. form the PHYLIP 3.6a4 package (http://evolution.genetics.washington.edu/phylip.html) and pfaat (http://pfaat.sourceforge.net/) (Johnson *et al.*, 2003). The PHYLIP programs were used with default settings and within pfaat, neighbor-joining trees were calculated using the BLOSUM62 substitution matrix and global column conservation weighting. Bootstrapping of neighbor-joining and parsimony trees was carried out with 100 replicates. Bootstrapping of the ProML trees was impossible due to the tremendous requirements in memory and computing power. In order to asses the branch support given by a likelihood-based method, trees were calculated for the alignments with 30% positional conservation filtering using MrBayes v3 (http://morphbank.ebc.uu.se/mrbayes/) (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003). MrBayes uses Bayesian inference estimations to assess phylogeny, is sufficiently fast to allow branch support evaluation by posterior probabilities and has been shown to be one of the most accurate likelihood programs available (Williams & Moret, 2003). Analysis was carried out using the Jones amino acid substitution model, four chains and an approximated gamma distribution of evolutionary rates with four categories. Visualization of trees was accomplished with ARB (http://www.arb-home.de/).

Genome trees. Genome trees were calculated from normalized BLASTP scores (Clarke *et al.*, 2002). In brief, all 231,509 ORFs were searched against each other using BLASTP. Different substitution matrices were tested (BLOSUM62, PAM70, PAM250). From the results, only the reciprocal BLASTP hits were extracted to avoid paraloguous hits. ORFs

involved in fewer than a given number of RBMs (reciprocal best matches between genome pairs with an E-value of 10^{-10} or better) were also omitted because they contain too little information (tresholds of 0, 4 and 10 were tested). No filtering of putatively laterally transferred ORFs was applied, because its effect has been proven to be small (Clarke *et al.*, 2002). The remaining data was transferred into a distance matrix in the following way: For each ORF in a given query genome, the bit-score of the RBM for this ORF in a given target genome was divided by the ORFs self-matching bit-score. The mean of these values for a given query-target pair was used as measure of the overall sequence similarity between the two. Distances were calculated as 1.0 minus the above mentioned similarity measure. The tree was calculated from the distance matrix with the program fitch from the PHYLIP 3.6a4 package (model of Fitch-Margoliash, global rearrangements, jumble 100).

RESULTS AND DISCUSSION

Throughout the last years it has become apparent that the extent of LGT is so high that it must be regarded as one of the major driving forces of evolution. The view that all genetic information in a given lineage traces back to one common ancestor simply does not apply in the world of prokaryotes where each protein has its own history. In this regard, the tree of life is a complex network of vertical and horizontal inheritance. While LGT is not specifically problematic for the phylogeny of closely related species, it becomes more severe when distantly related organisms are compared. The key question is, whether the extent of LGT is so high that it is impossible to trace the evolutionary relations between the major lineages or not. While some scientists believe that the degree of LGT is too high to trace organismal phylogeny on the basis of whole genomes or protein-coding genes (Nesbo *et al.*, 2001), others believe that there is a small set of proteins that forms a robust genetic core of an organism and carries signals of their evolutionary inheritance (Jain *et al.*, 1999; Wolf *et al.*, 2002, Daubin *et al., 2003*). Part of this core is built by so-called informational proteins. These are proteins of the transcription and translation apparatus like the subunits of the ribosome and those of the DNA-directed RNA polymerase (Harris *et al.*, 2003). According to the complexity hypothesis (Jain *et al.*, 1999), the physiological interactions and dependencies of these proteins are thought to be much more interwoven than those of operational genes. Therefore, once transferred to another organism via LGT, informational genes are unlikely to be capable to replace their counterparts. They simply would not fit into the fine-tuned regulation network of their new hosts. Because of their absence of function, they would rapidly accumulate mutations which would render them inactive first and finally cause them to vanish from their hosts' genomes. Despite the fact that the complexity hypothesis is debatable (Nesbo *et al.*, 2001; Daubin *et al.*, 2001), the informational ribosomal and DNA-directed RNA polymerase genes seem to be the best-suited genes encoding multi-subunit proteins whose sequences can be concatenated to infer phylogeny. In addition, it has been demonstrated recently, that the extent in which LGT affects typical phylogenetic protein markers might has been overestimated (Daubin *et al.*, 2003). Concatenated ribosomal protein sequences have been successfully applied in a number of phylogenetic studies before (Hansman & Martin, 2000; Wolf *et al.*, 2001; Brochier *et al.*, 2002; Matte-Tailliez *et al.*, 2002; Forterre *et al.*, 2002). They are supposed to have very strong resolution power in evaluating close and intermediate evolutionary distances, i.e. the relations between species and between major lineages (Wolf *et al.*, 2002).

Besides concatenation of protein sequences, three different methods to infer phylogeny from coding sequences of entire genomes have been developed in recent years. These methods are based on gene content (i.e. presence/absence of genes), gene order and normalized distances between orthologs (Wolf *et al.*, 2002). The gene content approach was not considered for this study as it is affected by artifacts caused by gene loss. For example,

parasitic bacteria with reduced genomes are artificially clustered in gene content trees (Wolf *et al.*, 2002). Gene order trees seemed inappropriate because gene order in general is only poorly conserved which is especially problematic with only one planctomycete genome and no close relative available. Therefore, trees based on normalized BLASTP scores were chosen.

Concatenated ribosomal proteins

Up to a positional conservation filtering of 30%, all trees calculated from concatenated ribosomal protein sequences successfully resolved the major phyla and in general confirmed the currently accepted 16S rRNA based phylogeny. *Spirochaetes* and *Chlamydia* formed a distinct superclade in the likelihood-based (Fig. 1a) and parsimony trees, whereas in some of the corresponding neighbor-joining trees these groups formed neighboring but independent clades (data not shown). The positions of *Chlorobium tepidum* TLS^T, *Deinococcus radiodurans* R1^T and *Thermoanaerobacter tengcongensis* MB4^T turned out not to be stable among the trees. *Chlorobium tepidum* TLS^T either affiliated to the *Spirochaetes* or branched off before the *Spirochaete/Chlamydia* clade in the neighbor-joining trees, whereas this species branched off before the *Epsilonproteobacteria* in the likelihood-based trees and clustered with the *Epsilonproteobacteria* in the parsimony trees. The position of *Thermoanaerobacter tengcongensis* MB4^T was dependent on the positional filtering. This species branched off next to the *Thermotoga/Aquifex* clade with the 10% positional conservation filter and affiliated to the *Firmicutes* with the respective 20% and 30% filters. The position of *Deinococcus radiodurans* R1^T varied considerably, but affiliated to the *Cyanobacteria/Actinobacteria* clade in the majority of trees.

These exceptions aside, the tree topologies exhibited only little dependence on the treeing algorithms used and all trees consistently placed the *Planctomycetes* at the base of the *Chlamydia*. Bootstrap support for the node grouping these phyla calculated from the alignment with 30% positional conservation filtering, was 97 (pfaat), 63 (PROTDIST/NEIGHBOR) and 56 (PROTPARS). The posterior probability in the corresponding MrBayes analysis was 100 (Fig. 1a). The *Planctomycete-Chlamydia* relationship was very stable regarding site selection. It was persistent up to a positional conservation filtering of 70% in the PROTDIST/NEIGHBOR neighbor-joining, 50% in the pfaat neighbor-joining, 40% in the ProML maximum-likelihood and 30% in the PROTPARS parsimony trees (Tab. 1). Stricter filtering lead to topologies, which were often not in agreement with the currently accepted 16S rRNA based phylogeny - especially regarding the position of the *Mollicutes* and *Epsilonproteobacteria*. Likewise, the position of the *Planctomycetes* became less stable and association with the *Chlamydia* was partly lost (Tab. 1). In three cases, the *Planctomycetes* shifted to the presumed root of the bacteria, adjacent to the *Thermotoga/Aquifex* clade. This position, however, was dependent on the treeing algorithm and site selection. For example, in the parsimony analysis, the *Planctomycetes* shifted to a deep branching position when a positional conservation filtering of 60% was applied. With the stricter 70% filter, however, the *Planctomycetes* swapped back into the *Spirochaete/Chlamydia* cluster (Tab. 1).

Concatenated DNA-directed RNA polymerase subunits

Trees based on concatenated aa sequences of DNA-directed RNA polymerase subunits resolved all major lineages known from 16S rRNA trees. *Spirochaetes* and the *Chlamydia* formed distinct clades in all trees. Up to a positional conservation filtering of 40%, the *Planctomycetes* branched off together with *Chlorobium tepidum* TLS^T from the chlamydial clade in all likelihood-based trees (Fig. 1b). However, in the majority of parsimony and neighbor-joining trees the *Planctomycetes* formed an independent clade that branched off either between the *Spirochaetes* and the *Chlamydia*, or between the *Chlamydia* and the *Espilonproteobacteria* (Tab. 2). Since the position of the *Planctomycetes* relative to the *Spirochaetes* and *Chlamydia* was dependent on site selection as

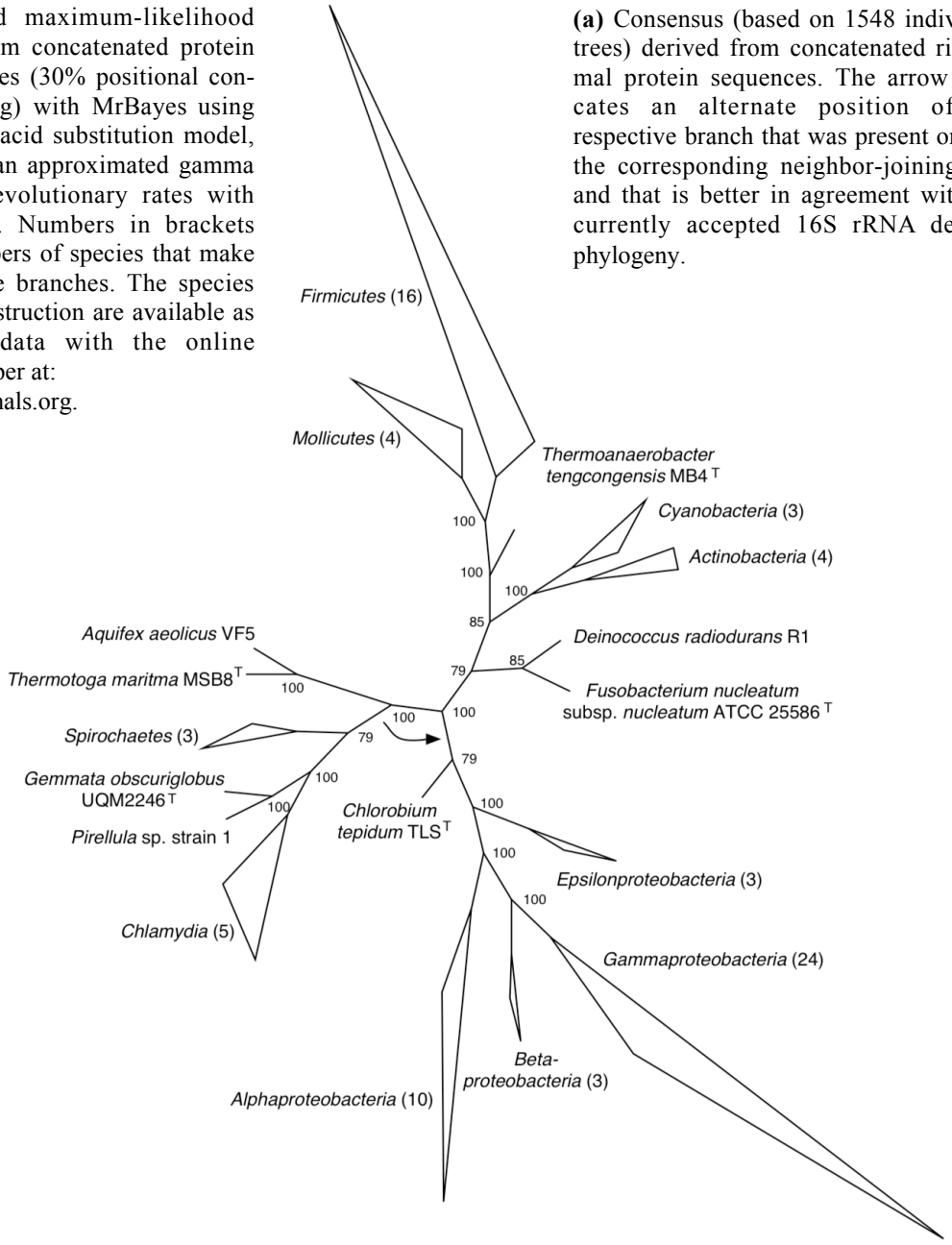
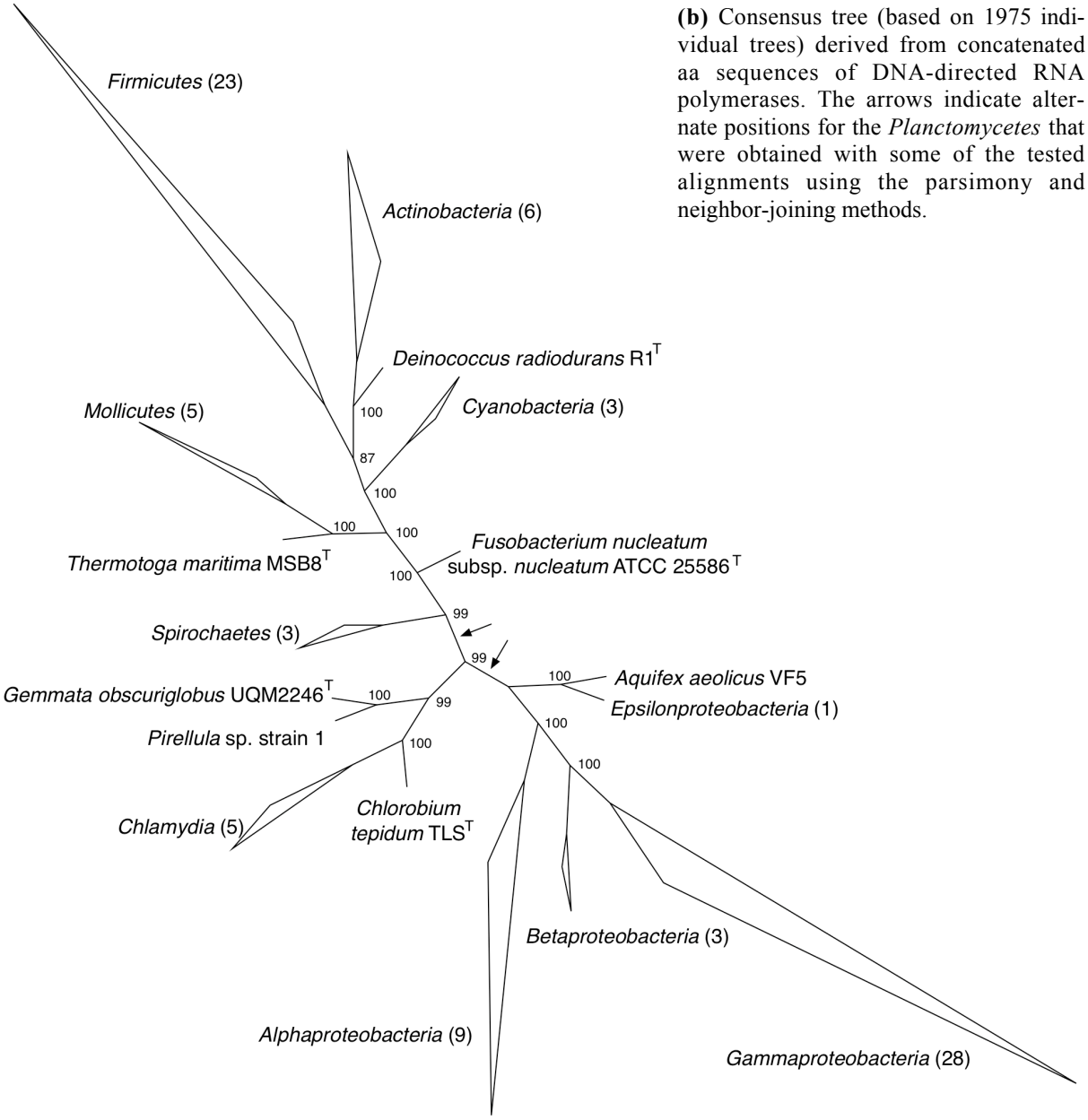


Fig. 1 Unrooted maximum-likelihood trees derived from concatenated protein subunits sequences (30% positional conservation filtering) with MrBayes using the Jones amino acid substitution model, four chains and an approximated gamma distribution of evolutionary rates with four categories). Numbers in brackets refer to the numbers of species that make up the respective branches. The species used for tree construction are available as supplementary data with the online version of this paper at: http://ijs.sgmjournals.org.

(a) Consensus (based on 1548 individual trees) derived from concatenated ribosomal protein sequences. The arrow indicates an alternate position of the respective branch that was present only in the corresponding neighbor-joining tree and that is better in agreement with the currently accepted 16S rRNA derived phylogeny.

filter	aa positions	pfaat	NEIGHBOR	PROTPARS	ProML	MrBayes
10%	5275	1	1	1	1	n.d.
20%	5238	1	1	1	1	n.d.
30%	4781	1	1	1	1	1
40%	4053	1	1	-	1	n.d.
50%	3320	1	1	-	-	n.d.
60%	2650	-	1	2	1	n.d.
70%	2144	2	1	1	1	n.d.
80%	1697	2	-	-	-	n.d.
90%	1201	-	-	-	-	n.d.

Tab. 1 Stability of the *Planctomycete/Chlamydia* relationship in phylogenetic analysis regarding the selection of sites from concatenated ribosomal protein sequences: (1) *Planctomycetes* branch off at the base of the *Chlamydia*; (2) *Planctomycetes* branch off near the presumed root of the *Bacteria* adjacent to the *Thermotoga/Aquifex* clade; (-) other topology; (n.d.) not determined



(b) Consensus tree (based on 1975 individual trees) derived from concatenated aa sequences of DNA-directed RNA polymerases. The arrows indicate alternate positions for the *Planctomycetes* that were obtained with some of the tested alignments using the parsimony and neighbor-joining methods.

filter	aa positions	pfaat	NEIGHBOR	PROTPARS	ProML	MrBayes
10%	2550	1	2	2	1	n.d.
20%	2529	1	2	2	1	n.d.
30%	2322	3	2	1	1	1
40%	2037	3	2	2	1	n.d.
50%	1697	3	-	1	4	n.d.
60%	1452	3	-	2	-	n.d.
70%	1266	-	-	1	4	n.d.
80%	1072	-	-	-	4	n.d.
90%	765	-	-	-	-	n.d.

Tab. 2 Stability of the *Planctomycete/Chlamydia* relationship in phylogenetic analysis regarding the selection of sites from concatenated sequences of the DNA-directed RNA polymerase: (1) *Planctomycetes* as well as *Chlorobium tepidum* TLS^T branch off at the base of the *Chlamydia*; (2) *Planctomycetes* branch off as independent clade between the *Chlamydia* and *Epsilonproteobacteria/Aquifex* (either with or without *Chlorobium tepidum* TLS^T); (3) *Planctomycetes* branch off as an independent clade between the *Spirochaetes* and the *Chlamydia* (either with or without *Chlorobium tepidum* TLS^T); (4) *Planctomycetes* branch off near the presumed base of the *Bacteria* adjacent to the *Thermotoga/Aquifex* clade; (-) other topology (partly not in accordance with the accepted 16S rRNA derived topology); (n.d.) not determined

well as on the treeing algorithm, a consensus tree would place these lineages within one multifurcating node. Likelihood-based treeing algorithms, however, must be regarded as superior to neighbor-joining and parsimony approaches. Therefore, using not too strict positional conservation filtering, concatenated aa sequences of the DNA-directed RNA polymerase support a relationship between the *Planctomycetes* and the *Chlamydia*, albeit with less strength as concatenated aa sequences of ribosomal proteins.

The positions of *Chlorobium tepidum* TLS^T, the *Cyanobacteria* and *Fusobacterium nucleatum* subsp. *nucleatum* ATCC 25586^T were not stable among the trees. *Chlorobium tepidum* TLS^T affiliated either to the *Planctomycetes*, the *Chlamydia* or the *Spirochaetes*. *Fusobacterium nucleatum* subsp. *nucleatum* ATCC 25586^T formed an independent lineage in most, but not in all trees and the *Cyanobacteria* had varying positions within the *Actinobacteria*/*Deinococcus*/*Firmicutes* clade. Thus, the overall topology was much more dependent on site selection as compared to trees derived from concatenated ribosomal protein sequences. While most trees in general exhibited an overall topology comparable to the 16S rRNA based phylogeny, there were some noteworthy exceptions. Most notably, *Deinococcus radiodurans* R1^T affiliated to the *Actinobacteria* and, in accordance with a previous study based on the DNA-directed RNA polymerase (Klenk *et al.*, 1999), *Aquifex aeolicus* VF5 did not cluster with *Thermotoga maritima* MSB8^T at the presumed root of the *Bacteria* but instead clustered with the *Epsilonproteobacteria*. Furthermore, the *Mollicutes* were clearly separated from the *Firmicutes* and branched of more deeply. Such a separation is consistent with trees based on fused 16S and 23S ribosomal RNA sequences (*Brochier* et al., 2002) and the results of different genome tree approaches (Tekaia *et al.*, 1999). A more deep branching position of the *Mollicutes* has been reported before for concatenated DNA-directed RNA polymerase subunits and has been attributed to an accelerated evolutionary rate and thus long branch attraction (Bocchetta, 2000). Since this could cause tree distortions, the dataset underlying Fig. 1(b) was reanalyzed without the *Mollicutes*. This did, however, neither change the position of the *Planctomycetes* nor the overall tree topology (data not shown).

With a positional conservation filtering of 50% and higher, most of the trees exhibited topologies, that were partly inconsistent with the currently accepted 16S derived topology. Interestingly, in the ProML maximum-likelihood analysis *Aquifex aeolicus* VF5 shifted adjacent to *Thermotoga maritima* MSB8^T and the *Planctomycetes* shifted to a position next to the newly formed *Thermotoga/Aquifex* clade (Tab. 2). This position was not found in the corresponding parsimony and neighbor-joining trees. In general, however, the shift of the *Planctomycetes* towards a deeper branching position with increased filtering of variable positions was more obvious with concatenated aa sequences of DNA-directed RNA polymerase subunits than with those of ribosomal proteins.

Genome trees

The genome tree derived from normalized BLASTP scores successfully resolved all major phyla (Fig. 2). Bootstrapping of the genome tree was not possible because of the enormous processing power and time required for its calculation. Thus, branch length and the overall topology were the only measures to assess the reliability of the tree. While separation of the major phyla was good (long branches), their branching pattern was only poorly resolved (very short branches). As in trees based on 16S rRNA analysis, the two thermophiles *Aquifex aeolicus* VF5 and *Thermotoga maritima* MSB8^T emerged at the very root of the *Bacteria*. *Spirochaetes* and *Chlamydia* formed a well-resolved superclade while *Pirellula* sp. strain 1, being the only *Planctomycete* in the tree, emerged as long independent branch between *Actino-* and *Cyanobacteria*. The exact position of the *Planctomycetes*, however, remained ambiguous because the branches of most phyla were too close together to reliably infer their branching pattern. Variation of the BLASTP scoring matrix (BLOSUM62, PAM70, PAM250) and the threshold for RBM filtering of species (0; 4; 10) retained the same overall

topology but did not improve the resolution of the tree (data not shown).

Comparison of trees

All trees derived from concatenated protein sequences consistently placed the *Planctomycetes* near or at the base of the *Chlamydia* as long as the underlying alignments were not restricted to the most conserved sites. When only highly conserved sites were used for tree construction, the *Planctomycetes* eventually shifted towards a deeper branching position. This position, however, was not consistent among different treeing algorithms and furthermore highly dependent on which sites were filtered. It is also noteworthy, that the overall branch length (and thus resolution) decreased with increased positional filtering and that the bootstrap support for a deep branching position of the *Planctomycetes* was low (around 40-50 in the neighbor-joining trees for ribosomal proteins).

The *Chlamydia* formed a distinct superclade with the *Spirochaetes* in most of the trees derived from concatenated ribosomal protein sequences and in the genome tree, while these two phyla formed neighboring but independent clades in most trees derived from concatenated DNA-directed RNA polymerase aa sequences. Neighborship of *Chlamydia* and *Spirochaetes* is consistent with the 16S rRNA based phylogeny and both groups forming a distinct superclade is supported by earlier studies on concatenated ribosomal proteins (Wolf *et al.*, 2001; Wolf *et al.*, 2002) and by previous genome trees (Clarke *et al.*, 2002). In

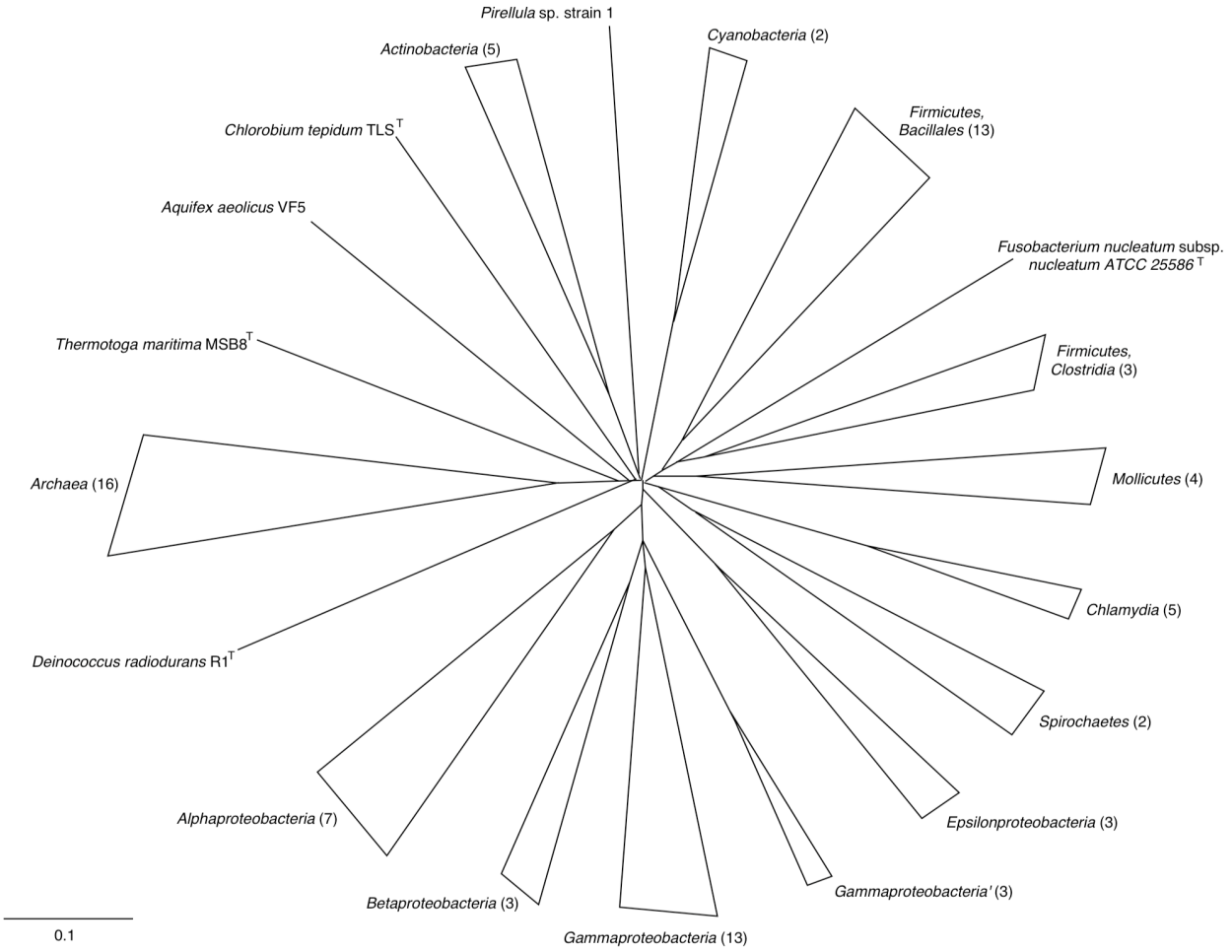


Fig. 2 Genome tree based on normalized BLASTP scores calculated with the Fitch-Margoliash distance matrix method (scoring matrix: BLOSUM62, RBM threshold: $E = 10^{-10}$, RBM filter: 4 species). The numbers in brackets refer to the number of species that make up the respective branches. The species used for tree construction are available as supplementary data with the online version of this paper at http://ijs.sgmjournals.org.

contrast to the trees inferred from concatenated protein sequences, the genome tree neither supports a close relationship of the planctomycete *Pirellula* sp. strain 1 and the *Chlamydia* nor a deep branching position. The exact position of the *Pirellula* sp. strain 1, however, is not well-resolved in the genome tree. Based on the current dataset, it is impossible to deduce whether the *Pirellula* sp. strain 1 did not occur within the *Chlamydia/Spirochaetes* superclade because of a lack in resolution or a contradiction to the trees inferred from concatenated protein sequences. A limitation of the genome tree is that so far only one *Planctomycete*, *Pirellula* sp. strain 1, has been completely sequenced and could be included in the tree. Clades consisting of few species often are less stable than those consisting of several species. Thus, inclusion of more than one *Planctomycete* would have been desirable.

The same rationale is valid for the position of *Chlorobium tepidum* TLS^T. This species affiliated with the *Spirochaetes*, *Planctomycetes* or *Chlamydia* in most of the trees inferred from concatenated protein sequences but not in the genome tree, where its exact position is unclear. Affiliation of *Chlorobium tepidum* TLS^T, the only representative of the *Chlorobi* sequenced so far, with *Chlamydia* and *Spirochaetes* is also consistent with the 16S rRNA based phylogeny (Nelson *et al.*, 2000).

The two thermophiles *Thermotoga maritima* MSB8^T and *Aquifex aeolicus* VF5 clustered consistently in trees based on concatenated ribosomal protein sequences and in the genome tree as they do in the currently accepted 16S rRNA tree. The *Bacteria* are rooted in the genome tree by inclusion of the *Archaea*. Because the two thermophiles emerged at the deepest branching positions within the *Bacteria*, the recently proposed deepest branching position of *Planctomycetes* (Brochier & Philippe, 2002) is not supported by the genome tree.

It is noteworthy, that in the trees inferred from concatenated aa sequences of DNA-directed RNA polymerase subunits, the two thermophiles *Thermotoga maritima* MSB8^T and *Aquifex aeolicus* VF5 did not cluster when lower positional conservation filtering was applied. Instead, *Aquifex aeolicus* VF5 affiliated to the *Epsilonproteobacteria*. There is in fact a debate going on whether in the currently accepted 16S rRNA based phylogenetic tree the thermophiles were placed at the root of the *Bacteria* erroneously due to long branch attraction (LBA) and whether the last common ancestor of the *Bacteria* was a thermophile at all (Daubin *et al.*, 2001; Gribaldo & Philippe, 2002). It has been proposed earlier that *Aquifex aeolicus* VF5 is closely related to *Proteobacteria* (Philippe & Laurent, 1998), which is for example indicated by gene content trees (Wolf *et al.*, 2002).

The recently proposed *Actinobacteria*/*Cyanobacteria*/*Deinococcus* superclade (Wolf *et al.*, 2001; Wolf *et al.*, 2002) was found in most of the trees based on concatenated protein sequences, but was not resolved in the genome tree. It is likely, that the resolution power of the genome tree method is blurred due to horizontally transferred genes. The inclusion of the *Archaea* also might have had a limiting effect on the resolution within the *Bacteria*, because genes that were laterally transferred between both domains minimize the distances within the *Bacteria*.

Additional support for a relationship of *Planctomycetes* and *Chlamydia*

A relationship of *Planctomycetes* and *Chlamydia* is further albeit weakly supported by indels. Indels are conserved insertions and deletions in key proteins that are assumed to be phylum-specific and thus suited for phylogeny. A system based on 18 indels was developed by Gupta (Gupta, 2001; Gupta & Griffith, 2002). According to this system, inserts in the termination factor *rho* and the alanyl-tRNA synthetase (*alaS*) are supposed to be diagnostic for the species that arose after the branching of the *Spirochaetes* and the *Chlamydia*, respectively. While *Pirellula* sp. strain 1 carries the first insert, it lacks the latter, which is consistent with a branching between *Spirochaetes* and *Chlamydia* (data for *Gemmata obscuriglobus* UQM2246^T are incomplete). However, like all phylogenetic methods, the indel method has its limitations (Philippe & Laurent, 1998; Gribaldo & Philippe, 2002). These are especially obvious

in the case of *Planctomycetes* where some of the markers are absent (*ftsZ*, *hsp*90, *lon* protease, inorganic pyrophosphatase), have paralogues (*hemL*, *dnaK*) or are fused with other genes (*secF*).

Further support for an affiliation of the *Planctomycetes* and the *Chlamydia/Spirochaetes* clade comes from trees derived from concatenated sequences of subunits of the well-conserved F1F0-ATPase operon (data not shown). In these trees, *Planctomycetes* consistently clustered with the only spirochaete in the dataset, *Leptospira interrogans* serovar lai 56601. *Chlamydia* do not have an F1F0-type ATPase and consequently were absent from the dataset. However, as mentioned above, *Spirochaetes* and *Chlamydia* are assumed to be relatives and thus clustering of the *Planctomycetes* with the *Spirochaetes* weakly supports an overall affiliation of the *Planctomycetes* with the *Spirochaetes/Chlamydia* superclade. Preliminary phylogenetic analysis of the *RecA* protein also indicated a close relationship of *Planctomycetes* and *Chlamydia* (data not shown).

Aside from the results of phylogenetic analysis, it seems at first rather surprising, that *Chlamydia* and *Planctomycetes* should have evolved from a common ancestor. *Chlamydia* are small, intracellular energy parasites with reduced genomes while *Planctomycetes* are free-living bacteria with genomes that are among the largest bacterial genomes known. There are however some noteworthy analogies between both groups, such as proteinaceous cell walls that are cross-linked via disulphide bonds. The *Planctomycetes* do have these in general (König *et al.*, 1984; Liesack *et al.*, 1986; Giovannoni *et al.*, 1987;) and the *Chlamydia* during their elementary body state (Hatch, 1996). The existence of all genes required for peptidoglycan biosynthesis in *Chlamydia* (Stephens *et al.*, 1998; Ghuysen & Goffin, 1999) and of some of these genes in the planctomycete *Pirellula* sp. strain 1 (Glöckner *et al.*, 2003) indicate that both groups once possessed peptidoglycan and that their proteinaceaous cell walls are secondary adaptations. *Chlamydia* and *Planctomycetes* do not only exhibit complex cell-cycles but also lack *ftsZ*, indicating an unknown mode of cell division (Brown & Rockey, 2000; Glöckner *et al.*, 2003). Related with their cell-division might also be the fact that from all genomes sequenced so far only *Chlamydia* and the planctomycetes *Pirellula* sp. strain 1and *Gemmata obscuriglobus* UQM2246^T harbor two copies of the gene *dnaA*. Moreover, these genes seem to be distantly related (Glöckner *et al.*, 2003). In addition, the genomes of all six sequenced *Chlamydia* (Karunakaran *et al.*, 2003) as well as that of *Pirellula* sp. strain 1 harbor three copies of *groeEL*-like genes. Furthermore, *Chlamydia* (Hatch, 1996) and *Planctomycetes* both have highly condensed DNA that is visible in electron micrographs as nucleoids. The 16S rRNA genes of both groups share signature positions that are not present in other bacteria (Fuerst, 1995) and finally, the ribosomal spc operon in all *Chlamydia* sequenced to date as well as in *Pirellula* sp. strain 1 and *Gemmata obscuriglobus* UQM2246^T is devoid of the ribosomal protein L30, like in some other bacteria, e.g. *Synechococcus* sp. PCC 6301 (Sugita, 1997).

Relationship to *Chlamydia* versus deep branching position

In two aforementioned studies, the slowly evolving positions of the 16S rRNA have been used to infer the phylogeny of the deepest branching species within the bacterial domain. The rational for this approach is that the phylogenetic signal of very ancient relationships is retained exclusively in the slowly evolving positions but might be obscured by the faster evolving ones if these are not filtered. Brochier & Philippe (2002) showed, that the *Planctomycetes* shift to the very root of the *Bacteria*, when only 751 slowly evolving positions are used for tree reconstruction and concluded, that the last universal common ancestor (LUCA) might not have been a thermophile. In a reevaluation of their results, Di Giulio (2003) demonstrated that a different selection of slowly evolving sites reestablishes the deepest branching position of the thermophiles (*Thermotoga maritima* MSB8^T, *Aquifex aeolicus* VF5) and places the *Planctomycetes* at a deep branching position after the thermophiles (this analysis is unfortunately devoid of *Chlamydia*). The

nature of the LUCA is beyond the scope of this study. Regarding the position of the *Planctomycetes*, however, our data as well exhibit an undeniable albeit inconsistent tendency to place the *Planctomycetes* at a deeper branching position within the *Bacteria*, when only highly conserved positions are used for tree reconstruction. Both of the mentioned studies used parsimony analysis to select the slowly evolving positions, while our analysis is based on a general filtering of variable positions. We therefore cannot exclude, that a different mode of site selection from our alignments would have lead to a more stable deep branching position of the *Planctomycetes*. However, one might ask, if filtering of the majority of unambiguously aligned positions from an alignment really reveals an otherwise obscured phylogenetic signal or if it rather introduces artifacts. In the end, the association of the *Planctomycetes* with the *Chlamydia* in our phylogenetic analysis was quite consistent, especially for the very large alignment of ribosomal protein sequences. Ignoring this signal in favor of a weakly supported deep branching position would be hard to justify. It is also noteworthy, that in numerous trees that were calculated from alignments with very strict positional conservation filtering, other species (e.g. *Campylobacter jejuni* subsp. *jejuni* NCTC 11168) often shifted to a deep branching position as well. Finally, a deep branching position of the *Planctomycetes* fails to explain the numerous analogies that exist between the *Planctomycetes* and the *Chlamydia*. Therefore, our analyses rather support a close affiliation of *Planctomycetes* with the *Chlamydia* than a deep branching position of the *Planctomycetes*.

Conclusions

The fact that different markers like concatenated protein sequences of ribosomal and DNA-directed RNA polymerase subunits reveal comparable overall topologies encourages the view that despite LGT prokaryote genomes retained a phylogenetic signal from which relationships can be reliably inferred. Phylogenetic analysis of concatenated aa sequences of ribosomal and DNA-directed RNA polymerase subunits concordantly indicate a close relationship of *Planctomycetes* and *Chlamydia*. It seems unlikely that both groups of proteins at the same time are affected by LGT from *Chlamydia*, that mimic an otherwise non-existing relationship. This case, however, can of course not be excluded with certainty. Also, a false affiliation of both group due to LBA is possible. These scenarios, however, seem unlikely because the relationship of *Planctomycetes* and *Chlamydia* is further supported - albeit with varying strength - by phylogenetic analysis of *RecA*, indels and some noteworthy analogies between both groups. Phylogenetic analysis of concatenated ATPase subunits weakly supported an affiliation of *Planctomycetes* and *Spirochaetes* and thus the *Chlamydia/Spirochaetes* superclade.

With respect to resolution power, concatenation of proteins sequences of ribosomal and DNA-directed RNA polymerase subunits did provide a better resolution for distant relationships than genome trees. Whether - as suggested by some authors (Wolf *et al.*, 2001) - concatenated sequences of ribosomal proteins are superior to 16S rRNA based phylogeny in assessing the overall topology of the bacterial tree of life, cannot be deduced from our data. Like with all protein-based phylogenies, concatenation of protein sequences has to face the problems of LGT and paralogy. In addition, site selection has a major impact on the weakly supported branches of the inferred trees, which especially affects the position of the *Chlamydia* (Hansmann & Martin, 2000). The *Planctomycete-Chlamydia* relationship in our trees based on concatenated protein sequences, however, was quite stable regarding site selection. In addition, trees based on concatenated sequences of ribosomal proteins from different workgroups show only slight differences in their branching patterns and are remarkably similar (Wolf *et al.*, 2002). 16S rRNA based phylogeny on the other hand also has to face the problem of paralogy, because most bacterial genomes harbor more than one set of rDNA genes. In addition, there are cases where the 16S rRNA sequences within one organism can vary considerably, and even LGT of the 16S rDNA gene has been described (Yap *et al.*, 1999). Moreover, information content of the 16S rRNA gene is

limited. The overall topology of the bacterial tree of life is beyond the scope of this study. However, bootstrap values and branch length (data not shown) of the trees derived from concatenated aa sequences of ribosomal proteins and DNA-directed RNA polymerase subunits indicate that they are well-suited to infer the phylogeny of the *Planctomycetes*.

It will be interesting to see if the phylogenetic relationship of the *Planctomycetes* with the *Chlamydia* that is indicated by most of our data will hold true as more planctomycete genomes or those of the recently discovered environmental *Parachlamydia* become available (http://www.microbial-ecology.net/edge.html). We hope, that an in depth analysis of future planctomycete genomes will help to gain further insights into their phylogenetic position.

ACKNOWLEDGEMENTS

We would like to acknowledge Tim Frana for screening all published bacterial genomes for ribosomal proteins. Preliminary sequence data of *Gemmata obscuriglobus* UQM 2246^T was obtained from The Institute for Genomic Research website at http://www.tigr.org. Sequencing of *Gemmata obscuriglobus* UQM 2246^T is accomplished with support from the U.S. Department of Energy.

REFERENCES

Bocchetta, M., Gribaldo, S., Sanangelaantoni, A. & Cammarano, P. (2000). Phylogenetic depth of the bacterial genera *Aquifex* and *Thermotoga* inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. *J Mol Evol* **50,** 366-380

Bomar, D., Giovannoni, S. & Stackebrandt, E. (1988). A unique type of eubacterial 5S rRNA in members of the order *Planctomycetales*. *J Mol Evol* **27**, 121-125

Brochier, C., Baptest, E., Moreira, D. & Philippe, H. (2002). Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet* **18**, 1-5

Brochier, C. & Philippe, H. (2002). Phylogeny: A non-hyperthermophilic ancestor for *Bacteria*. *Nature* **417,** 244-244

Brown, W. J. & Rockey, D. (2000). Identification of an antigen localized to an apparent septum within dividing *Chlamydiae*. *Infect Immun* **68**, 708-715

Clarke, G. D. P., Beiko, R. G., Ragan, M. A. & Charlebois R. L. (2002). Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J Bacteriol* **184**, 2072-2080

Daubin, V., Gouy, M. & Perrière, G. (2001). Bacterial molecular phylogeny using supertree approach. *Genome Informatics* **12,** 155-164

Daubin, V., Moran N. A. & Ochman H. (2003). Phylogenetics and the cohesion of bacterial genomes. *Science* **301,** 829-832

DeLong, E. F. (1993). Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. *Limnol Oceanogr* **38**, 924-934

Di Giulio, M. (2003). The ancestor of the *Bacteria* domain was a hyperthermophile. *J Theor Biol* **224**, 277-283

Forterre, P., Brochier, C. & Philippe, H. (2002). Evolution of the *Archaea*. *Theor Pop Biol* **61**, 409-422

Fuerst, J. A. (1995). The *Planctomycetes*: emerging models for microbial ecology, evolution and cell biology. *Microbiology* **141,** 1493-1506

Fuerst, J. A., Gwilliam, H. G., Lindsay, M., Lichanska, A., Belcher, C., Vickers, J. E. & Hugenholtz P. (1997). Isolation and molecular identification of planctomycete bacteria from postlarvae of the giant tiger prawn *Penaeus monodon*. *Appl Environ Microbiol* **63**, 254-262

Gade, D., Schlesner, H., Glöckner, F.O., Amann, R., Pfeiffer, S. & Thomm M. (2003). Identification of planctomycetes

with order-, genus- and strain-specific 16S rRNA-targeted probes. *FEMS Microbiol Ecol*, in press.

Garrity, G. M., Johnson, K. L., Bell, J. & Searles, D. B. (2002). Taxonomic outline of the prokaryotes, Bergey's manual of systematic bacteriology. 2nd edn. release 3.0. July, Springer Verlag, NY (http://dx.doi.org/10.1007/bergeysoutline 200210)

Ghuysen, J. M. & Goffin C. (1999). Lack of cell wall peptidoglycan versus penicillin sensitivity: new insights into the chlamydial anomaly. *Antimicrob Agents Chemother* **43**, 2339-2344

Giovannoni, S. J., Godchaux III, W., Schabtach, E. & Castenholz R. W. (1987). Cell wall and lipid composition of *Isosphaera pallida*, a budding eubacterium from hot springs. *J Bacteriol* **169**, 2702-2707

Glöckner, F. O., Kube, M., Bauer, M. & 11 other authors (2003). Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc Natl Acad Sci U S A* **100**, 8298-8303

Gribaldo S. & Philippe H. (2002). Ancient phylogenetic relationships. *Theor Pop Biol* **61**, 391-408

Gupta, R. S. (2001). The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. *Int Microbiol* **4**, 187-202

Gupta, R. S. & Griffith E. (2002). Critical issues in bacterial phylogeny. *Theor Pop Biol* **61**, 423-434

Hansmann, S. & Martin, W. (2000). Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from the analysis. *Int J Syst Evol Microbiol* **50**, 1655-1663

Harris, J. K., Kelley, S. T., Spiegelman, G. B. & Pace, N. R. (2003). The genetic core of the universal ancestor. *Genome Res* **13**, 407-412

Hatch, T. P. (1996). Minireview: disulphide cross-linked enveloppe proteins: the functional equivalent of peptidoglycan in *Chlamydiae*?. *J Bacteriol* **178**, 1-5

Huelsenbeck, J. P. & Ronquist, F. (2001). MRBAYES: bayesian inference of phylogenetic trees. *Bioinf* **17**, 754-755

Jain, R., Rivera, M. & Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* **96,** 3801-3806

Jenkins, C. & Fuerst, J. A. (2001). Phylogenetic analysis of evolutionary relationships of the *Planctomycete* division of the domain *Bacteria* based on amino acid sequences of elongation factor Tu. *J Mol Evol* **52,** 405-418

Johnson, J. M., Mason, K., Moallemi, C., Xi, H., Somaroo, S. & Huang, E. S. (2003). Protein family annotation in a multiple alignment viewer. *Bioinf* **19**, 544-545

Karunakaran, K. P., Noguchi, Y., Read, T. D., Cherkasov, A., Kwee, J., Shen, C., Nelson, C. C. & Brunham R. C. (2003). Molecular analysis of the multiple *GroEL* proteins of *Chlamydiae*. *J Bacteriol* **185**, 1958-1966.

Kerger, D., Mancuso, A., Nichols, P. D., White, D. C., Langworthy, T., Sittig, M., Schlesner, H. & Hirsch, P. (1988). The budding bacteria, *Pirellula* and *Planctomyces*, with atypical 16S rRNA and absence of peptidoglycan, show eubacterial phospholipids und uniquely high proportions of long chain beta-hydroxy fatty acids in the lipopolysaccharide lipid A. *Arch Microbiol* **149,** 255-260

Klenk, H. P., Meier, T. D., Durovic, P., Schwass, V., Lottspeich, F., Dennis, P. P. & Zillig, W. (1999). RNA polymerase of *Aquifex pyrophilus*: implications for the evolution of the bacterial *rpoBC* operon and extremely thermophilic bacteria. *J Mol Evol* **48,** 528-541

König, E., Schlesner, H. & Hirsch, P. (1984). Cell wall studies on budding bacteria of the *Planctomyces/Pasteuria* group and on a *Prosthecomicrobium* sp.. *Arch Microbiol* **138**, 200-205

Liesack, W., König, H., Schlesner, H. & Hirsch, P. (1986). Chemical composition of the peptidoglycan-free cell envelopes of budding bacteria of the

Pirellula/Planctomyces group. *Arch Microbiol* **145**, 361-366

Liesack, W., Söller, R., Steward, T., Haas, H., Giovannoni, S. & Stackebrandt, E. (1992). The influence of tachytelically (rapidly) evolving sequences on the topology of phylogenetic trees - intrafamily relationships and the phylogenetic position of *Planctomycetaceae* as revealed by comparative analysis of 16S ribosomal RNA sequences. *System Appl Microbiol* **15**, 357-362

Lindsay, M. R., Webb, R. I. & Fuerst J. A. (1997). Pirellulosomes: a new type of membrane-bounded cell compartment in planctomycete bacteria of the genus *Pirellula*. *Microbiology* **143**, 739-748

Lindsay, M. R., Webb, R. I., Strous, M., Jetten, M. S. M., Butler, M. K., Forde, R. J., & Fuerst, J. A. (2001). Cell compartmentalization in *Planctomycetes*: novel types of structural organization for the bacterial cell. *Arch Microbiol* **175,** 413-429

Matte-Tailliez, O., Brochier, C., Forttere, P. & Philippe, H. (2002). Archaeal phylogeny based on ribosomal proteins. *Mol Biol Evol* **19**, 631-639

Meyer, F., Goesmann, A., McHardy, A. C & 8 other authors (2003). GenDB--an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res* **31**, 2187-2195.

Miskin, I. P., Farrimond, P. & Head, I. M. (1999). Identification of novel bacterial lineages as active members of microbial populations in a freshwater sediment using a rapid RNA extraction procedure and RT-PCR. *Microbiology* **145**, 1977-1987

Nelson, K. E., Paulsen, I. T., Heidelberg, J. F. & Fraser, C. M. (2000). Status of genome projects for nonpathogenic *Bacteria* and *Archaea. Nat Biotechnol* 18, 1049-1054

Nesbo, C. L., Boucher, Y. & Doolittle, W. F. (2001). Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. *J Mol Evol* **53**, 340-350

Philippe, H. & Laurent, J. (1998). How good are deep phylogenetic trees?. *Curr Opin Genet Dev* **8**, 616-623

Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: bayesian phylogenetic inference under mixed models. *Bioinf* **19**, 1572-1574

Schlesner, H. (1994). The development of media suitable for the microorganisms morphologically resembling *Planctomyces* spp., *Pirellula* spp. and other *Planctomycetales* from various aquatic habitats using dilute media. *System Appl Microbiol* **17**, 135-145

Schmid, M., Schmitz-Esser, S., Jetten, M. & Wagner, W. (2001). 16S-23S rDNA intergenic spacer and 23S rDNA of anaerobic ammonium-oxidizing bacteria: implications for phylogeny and *in situ* detection. *Environ Micorbiol* **3**, 450-459

Stackebrandt, E., Ludwig, W., Schubert, W., Klink, F., Schlesner, H., Roggenton, T. & Hirsch, P. (1984). Molecular genetic evidence for early evolutionary origin of budding peptidoglycan-less *Eubacteria*, *Nature* **307**, 735-737

Stephens, R. S., Kalman, S., Lammel, C. & 9 other authors (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**, 754-759

Strous, M,. Fuerst, J. A., Kramer, E. H. M., Logemann, S., Muyzer, G., van de pas-Schoonen, K. T., Webb, R., Kuenen, J. G. & Jetten, M. S. M (1999). Missing lithotroph identified as new planctomycete. *Nature* **400**, 446-449

Sugita, M., Sugishita, H., Fujishiro, T., Tsuboi, M., Sugita, C., Endo, T. & Sugiura, M. (1997). Organization of a large gene cluster encoding ribosomal proteins in the cyanobacterium *Synechococcus* sp. strain PCC 6301-comparison of gene cluster among *Cyanobacteria*, *Eubacteria* and chloroplast genomes. *Gene* **195**, 73-79

Tekaia , F., Lazcano, A. & Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. *Genome Res.* **9**, 550–557

Vergin, K. L., Urbach, E., Stein, J. L., DeLong, E. F., Lanoil, B. D. & Giovannoni, S. J. (1998). Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order *Planctomycetales*. *Appl Environ Microbiol* **64,** 3075-3078

Wang, J., Jenkins, C., Webb, R. & Fuerst, J. A. (2002). Isolation of *Gemmata*-like and *Isosphaera*-like bacteria from soil and freshwater. *Appl Environ Microbiol* **68**, 417-422

Ward, N., Rainey, F. A., Stackebrandt, E. & Schlesner H. (1995). Unraveling the extent of diversity within the order *Planctomycetales*. *Appl Environ Microbiol* **61,** 2270-2275

Ward-Rainey, N., Rainey, F. A., Wellington, E. M. H. & Stackebrandt E. (1996). Physical map of the genome of *Planctomyces limnophilus*, a representative of the phylogenetically distinct planctomycete lineage. *J Bacteriol* **178**, 1908-1913

Ward-Rainey, N., Rainey, F. A. & Stackebrandt, E. (1997). The presence of a *dnaK* (HSP70) multigene family in members of the order *Planctomycetales* and *Verrucomicrobiales*. *J Bacteriol* **179**, 6360-6366

Ward, N. L., Rainey, F. A., Hedlund, B. P., Staley, J. T., Ludwig, W. & Stackebrandt, E. (2000). Comparative phylogenetic analyses of members of the order *Planctomycetales* and the division *Verrucomicrobia*: 23S rRNA gene sequence analysis supports the 16S rRNA gene sequence-derived phylogeny. *Int J Syst Evol Microbiol* **50**, 1965-1972

Weisburg, W. G., Hatch, T. P. & Woese, C. R. (1986). Eubacterial origin of *Chlamydiae*. *J Bacteriol* **167**, 570-574

Williams, T. L. & Moret B. M. E. (2003). An investigation of phylogenetic likelihood methods. *Proc 3rd IEEE Symp on Bioinformatics and Bioengeneering (BIBE '03)*, IEEEE Press, 79-86 (http://www.computer.org/proceedings/bibe/1907/1907toc.htm)

Woese, C. R. (1987). Bacterial evolution. *Microbiol Rev* **51,** 221-271

Wolf, Y. I., Rogozin, I.B., Grishin, N.V., Tatusov, R.L. & Koonin, E.V. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evolutionary Biology* **1**, 8

Wolf, Y. I., Rogozin, I. B., Grishin, N.V. & Koonin, E.V. (2002). Genome trees and the tree of life. *Trends Genet* **18**, 472-479

Yap, W. H., Zhang, Z. & Wang, Y. (1999). Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. *J Bacteriol* **151**, 5201-5209

SUPPLEMENTARY DATA

	ribosomal proteins tree	DNA-directed RNA polymerase tree	genome tree
Bacteria			
Agrobacterium tumefaciens strain C58 Cereon	+	+	+
Agrobacterium tumefaciens strain C58 UWash	+	+	-
Aquifex aeolicus VF5	+	+	+
Bacillus cereus ATCC 14579^T	-	+	-
Bacillus halodurans C-125	-	+	+
Bacillus subtilis subsp. *subtilis* 168	+	+	+
Bifidobacterium longum NCC 2705	+	+	-
Borrelia burgdorferi B31^T	+	+	+
Bradyrhizobium japonicum USDA 110	+	+	-
Brucella melitensis 16M	+	-	+
Brucella suis 1330	+	+	-
Buchnera aphidicola APS	+	+	+
Buchnera aphidicola Bp	+	-	-
Buchnera aphidicola Sg^T	+	+	+
Campylobacter jejuni subsp. *jejuni* NCTC 11168	+	+	+
Caulobacter crescentus CB15	+	+	+
Chlamydia muridarum strain Nigg	+	+	+
Chlamydia trachomatis serovar D	+	+	+
Chlamydophila caviae GPIC^T	-	+	-
Chlamydophila pneumoniae AR39	+	+	+
Chlamydophila pneumoniae CWL029	+	+	+
Chlamydophila pneumoniae J138	+	+	+
Chlorobium tepidum TLS^T	+	+	+
Clostridium acetobutylicum ATCC 824^T	-	+	+
Clostridium perfringens 13	-	+	-
Corynebacterium efficiens YS-314^T	+	+	+
Corynebacterium glutamicum ATCC 13032^T	+	-	-
Coxiella burnetii RSA 493	-	+	+
Deinococcus radiodurans R1^T	+	+	-
Enterococcus faecalis V583	+	+	-
Escherichia coli CFT073	+	+	+
Escherichia coli K-12-MG1655	+	+	+
Escherichia coli O157:H7 EDL933	+	+	+
Escherichia coli O157:H7 VT2-Sakai	+	+	+
Fusobacterium nucleatum subsp. *nucleatum* ATCC 25586^T	+	+	+
Gemmata obscuriglobus UQM 2246^T	+	+	-
Haemophilus influenzae Rd	+	+	+
Helicobacter pylori 26695	+	-	+
Helicobacter pylori J99	+	-	+
Lactobacillus plantarum WCFS1	+	-	-
Lactococcus lactis subsp. *lactis* IL1403	+	+	+
Leptospira interrogans serovar *lai* str. 56601	+	+	-
Listeria innocua CLIP 11262	+	+	+
Listeria monocytogenes EGD-e	+	+	+
Mesorhizobium loti MAFF303099	+	+	+
Mycobacterium leprae TN	+	+	+
Mycobacterium tuberculosis CDC1551	-	+	+
Mycobacterium tuberculosis H37Rv^T	-	+	+
Mycoplasma genitalium G37^T	+	+	+
Mycoplasma penetrans HF-2	+	+	-
Mycoplasma pneumoniae M129	+	+	+
Mycoplasma pulmonis UAB CTIP	+	+	+

Neisseria meningitidis serogroup A Z2491	+	+	+
Neisseria meningitidis MC58	+	+	+
Nostoc sp. PCC 7120	+	+	+
Oceanobacillus iheyensis HTE831^T	+	+	-
Pasteurella multocida PM70	+	+	+
Pirellula sp. strain 1	+	+	+
Pseudomonas aeruginosa PA01	+	+	+
Pseudomonas putida KT2440	+	+	-
Pseudomonas syringae pv. *tomato* DC3000	+	+	-
Ralstonia solanacearum GMI1000	+	+	+
Rickettsia conorii Malish 7^T	+	+	+
Rickettsia prowazekii Madrid E	+	+	+
Salmonella enterica subsp. *enterica* serovar *Typhi*	+	+	+
Salmonella typhi CT18	-	+	-
Salmonella typhimurium LT2 SGSC 1412	+	+	+
Shewanella oneidensis MR-1^T	+	+	-
Shewanella violacea DSS12^T	-	+	-
Shigella flexneri 2a strain 301	-	+	-
Shigella flexneri 2a 2457T	+	+	-
Sinorhizobium meliloti 1021	+	+	+
Staphylococcus aureus subsp. *aureus* Mu50	+	+	+
Staphylococcus aureus subsp. *aureus* MW2	+	+	+
Staphylococcus aureus subsp. *aureus* N315	+	+	+
Staphylococcus epidermidis ATCC 12228	-	+	-
Streptococcus agalactiae 2603 V/R	+	+	-
Streptococcus agalactiae NEM316	+	+	-
Streptococcus mutans UA159	+	+	-
Streptococcus pneumoniae R6	+	+	+
Streptococcus pneumoniae TIGR4	+	+	+
Streptococcus pyogenes SF370 serotype M1	+	+	+
Streptococcus pyogenes MGAS315	+	+	+
Streptococcus pyogenes MGAS8232	+	+	+
Streptococcus pyogenes SSI-1	+	-	-
Streptomyces coelicolor A3(2)	-	-	+
Synechocystis sp. PCC 6803	+	+	+
Thermoanaerobacter tengcongensis MB4^T	+	+	+
Thermosynechococcus elongatus BP-1	+	+	-
Thermotoga maritima MSB8^T	+	+	+
Treponema pallidum Nichols	+	+	+
Tropheryma whipplei TW08/27	+	+	-
Tropheryma whipplei strain Twist	+	+	-
Ureaplasma urealyticum parvum biovar serovar 3	+	+	+
Vibrio cholerae El Tor N16961	+	+	+
Vibrio parahaemolyticus RIMD 2210633	+	-	-
Vibrio vulnificus CMCP6	+	+	-
Wigglesworthia brevipalpis	+	+	-
Xanthomonas axonopodis pv. *citri* 306	+	+	+
Xanthomonas campestris pv. *campestris* ATCC 33913^T	+	+	+
Xylella fastidiosa 9a5c	+	+	+
Xylella fastidiosa Temecula1	+	+	-
Yersinia pestis KIM	+	+	+
Yersinia pestis CO92	+	+	+
Archaea			
Aeropyrum pernix K1^T	-	-	+
Archeoglobus fulgidus DSM 4304^T	-	-	+
Halobacterium sp. NRC-1	-	-	+
Methanocaldococcus janaschii DSM 2661^T	-	-	+
Methanopyrus kandleri AV19^T	-	-	+
Methanosarcina acetivorans C2A^T	-	-	+

Methanosarcina mazei Goe1	-	-	+
Methanothermobacter thermoautotrophicus delta-H^T	-	-	+
Pyrobaculum aerophilum IM2^T	-	-	+
Pyrococcus abyssi GE5	-	-	+
Pyrococcus furiosus DSM 3638^T	-	-	+
Pyrococcus horikoshii shinkaj OT3^T	-	-	+
Sulfolobus solfataricus P2	-	-	+
Sulfolobus tokodaii strain 7^T	-	-	+
Thermoplasma acidophilum DSM 1728^T	-	-	+
Thermoplasma volcanium GSS1^T	-	-	+

3

Application of Tetranulceotide Frequencies for the Assignment of Genomic Fragments

Hanno Teeling, Anke Meyerdierks, Margarete Bauer, Rudolf Amann and Frank Oliver Glöckner

Environ Microbiol Special Issue on Metagenomics
in press

Application of Tetranucleotide Frequencies for the Assignment of Genomic Fragments

Hanno Teeling, Anke Meyerdierks, Margarete Bauer, Rudolf Amann and Frank Oliver Glöckner

Max Planck Institute for Marine Microbiology, Department of Molecular Ecology,, Genomics Group, Celsiusstrasse 1, D-28359 Bremen, Germany

ABSTRACT

A basic problem of the metagenomic approach in microbial ecology is the assignment of genomic fragments to a certain species or taxonomic group, when suitable marker genes are absent. Currently, the (G+C)-content together with phylogenetic information and codon adaptation for functional genes is mostly used to assess the relationship of different fragments. These methods, however, can produce ambiguous results. In order to evaluate sequence-based methods for fragment identification, we extensively compared (G+C)-contents and tetranucleotide usage patterns of 9,054 fosmid-sized genomic fragments generated *in silico* from 118 completely sequenced bacterial genomes (40,982,931 fragment pairs were compared in total). The results of this systematic study show that the discriminatory power of correlations of tetranucleotide-derived z-scores is by far superior to that of differences in (G+C)-content and provide reasonable assignment probabilities when applied to metagenome libraries of small diversity. Using six fully sequenced fosmid inserts from a metagenomic analysis of microbial consortia mediating the anaerobic oxidation of methane (AOM), we demonstrate that discrimination based on tetranucleotide-derived z-score correlations was consistent with corresponding data from 16S ribosomal RNA sequence analysis and allowed us to discriminate between fosmid inserts that were indistinguishable with respect to their (G+C)-contents.

INTRODUCTION

Until the advent of molecular techniques, studies on the diversity and function of microorganisms were restricted by the need to obtain pure cultures. Although attempts to cultivate bacteria have been conducted over decades, and thousands of isolates are available in culture collections, the vast majority of microorganisms has not yet been characterized. From cultivation-independent techniques it is currently estimated that in many ecosystems less than 1% of the microbial diversity has been seen on agar plates (Amann *et al.*, 1995). This fact has dramatically shifted our understanding of microbial communities and their role and function within their natural habitats. To overcome these limitations, methods have been developed to explore the physiological potential of uncultivated microorganisms by extracting and analyzing not only genes but large genomic fragments directly from the environment (DeLong, 2002; Rondon *et al.*, 2000). In this so called metagenomic approach (Schloss and Handelsman, 2003), DNA is directly extracted from environmental samples and cloned into vectors such as cosmids, fosmids or bacterial artificial chromosomes (BACs). These metagenome clone libraries can be screened for inserts carrying specific functions or for sequences of known genes.

Employing metagenomics, new enzymes and large genomic fragments of as yet uncultured bacteria have been successfully retrieved from the environment. One of the

most prominent examples to date is the discovery of bacteriorhodopsin in marine *Gammaproteobacteria*, a protein that previously was believed to occur exclusively in archaeal *Halobacteria* (Béjà *et al*., 2000; Béjà *et al.*, 2001).

Despite the proven potential of the metagenomic approach to broaden our knowledge about the composition and function of natural microbial communities, several methodological problems remain to be solved. One of the major challenges is the identification of the fragments' organismal origin. Assuming an average genome size of 4 Mb, only about 5-10% of the clones within a fosmid library (40 kb insert size) harbor phylogenetic marker genes like 16S rDNA, *rpoA*, *recA* etc. and can therefore be assigned to a certain species or taxonomic group. The chance that new and interesting functional genes are located on a fragment that also carries a phylogenetic marker gene is even lower. Hence, there is a clear need for additional tools to provide evidence that e.g. two fragments belong to the same organism. Identification of unknown fragments is trivial if they overlap for at least one or two kilobases. In this case, fragments can be fused and genome walking techniques can be applied to reconstruct as much of the metagenome as possible. However, this is extremely time-consuming, and if a connecting clone cannot be found, the whole procedure is stalled. The sequence-based measure commonly used to assess whether two unlinked fragments belong to the same organism, is their similarity in (G+C)-content. In addition, best BLAST hits as well as codon usage of the genes residing on the fragments can provide valuable hints about their organismal origin. These measures, however, can be misleading. The (G+C)-content of prokaryotes can vary considerably within genomes, and does not carry a strong phylogenetic signal. Regarding gene content, a typical fosmid insert of ~40 kb harbors about 40 genes, from which on average only half yield significant hits when searched against the public databases, e.g. by BLAST. Frequently, these hits do not come from the same phylogenetic taxon and thus provide no hints on the insert's organismal origin. For example, many protein families are phylogenetically unspecific and the phylogenetic significance of others is affected by lateral gene transfer (LGT). Moreover, in many cases databases simply do not contain close relatives to the species under investigation, which affects the metagenome approach in particular, since the discovery of new lineages and functions is one of its key intentions. The codon usage of the genes can also be blurred by LGT, and varies with gene expression level rather than carrying a phylogenetic signal (Karlin and Mrázek, 2000).

In contrast, a rather pronounced phylogenetic signal can be found in the tetranucleotide usage patterns of the inserts' nucleotide sequences (Pride *et al.*, 2003). Frequencies of oligonucleotides in genomic sequences are known to carry species-specific signals (Karlin *et al.*, 1994; Karlin and Burge, 1995; Karlin, 1998; Karlin *et al.*, 1998; Nakashima *et al.*, 1998). Using the relative abundances of oligomers up to a length of four nucleotides, this has been shown for prokaryotes (Karlin *et al.*, 1994) as well as for eukaryotes (Karlin *et al.*, 1994; Gentles and Karlin, 2001). Species-specific signals for oligomers of different length have also been detected by means of neuronal networks (Abe *et al.*, 2003), using chaos game representations (Goldman 1993; Deschavanne *et al.*, 1999) and naïve Bayesian classifiers (Sandberg *et al.*, 2001). The cause for these signals has been attributed - at least for dinucleotides - to species-specific codon usage as well as a selective pressure on stacking energies and base-step conformational preferences, species-specific properties of DNA modification, replication and repair mechanisms, and selection by specific restriction endonucleases (Karlin *et al.*, 1998). The evolutionary significance of species-specific patterns that are observed with longer oligonucleotides is unclear so far.

Here we present the application of tetranucleotide-derived z-score correlations as a measure for the relatedness of genomic fragments as well as a comparative assessment of the method’s discriminatory power versus the discriminatory power of (G+C)-content differences for fosmid-sized genomic fragments. Furthermore, we demonstrate the successful application of the tetranucleotide

method for the assignment of fosmid inserts from metagenome libraries that were constructed from microbial consortia involved in the anaerobic oxidation of methane (AOM).

RESULTS AND DISCUSSION

In silico assessment of the discriminatory power of Δ(G+C) versus tetranucleotide-derived z-score correlations

Pairwise comparisons of 118 bacterial genomes revealed that, using differences in GC-content [Δ(G+C)], artificial fosmid-sized genomic fragments of 40 kb could not be matched correctly to their genomes in 36.0% of 6,903 possible comparisons. This can be explained by overlapping ranges of (G+C)-content, which in the worst case can be as extreme as illustrated in Fig. 1a for *Escherichia coli* K-12 and *Neisseria meningitidis* Z2491. For this genome pair, only a (G+C)-content of less than 48% or more than 54% would allow an unambiguous assignment of a given fragment. If a fosmid library was generated from a hypothetical habitat harboring solely these two bacteria, the combined probability for obtaining two fosmid inserts that could be assigned to their genomes beyond doubt would be only 1%. In contrast, z-scores derived from the respective fragments' tetranucleotide usage patterns exhibit a high correlation within both genomes (>0.69) and a low correlation between them (<0.49). This enables a perfect assignment in all cases (Fig. 1b).

It is obvious that overlapping histograms as presented in Fig. 1a are the more likely the more genomes are present. By this, the Δ(G+C)-method gets saturated quickly and discrimination between fosmids renders problematic, if the number of species in a library exceeds about ten to 20 species. The tetranucleotide method is less affected by such saturation, since the possible permutations of species-specific tetranucleotide usage patterns are extremely high. As long as the intra-genomic variation within the tetranucleotide usage is low, the addition of more genomes will only slightly decrease the method's

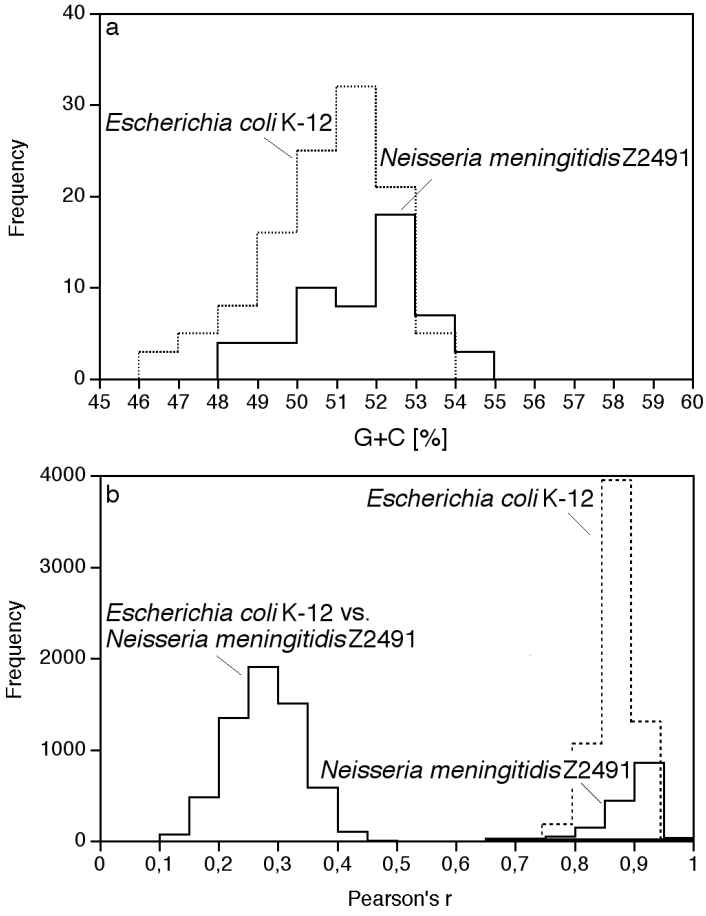


Fig. 1 (G+C)-content and correlation of tetranucleotide-derived z-scores of artificial 40 kb fragments from *Escherichia coli* K-12 and *Neisseria meningitidis* Z2491. In the upper part (a), histograms for the (G+C)-content of all 40 kb fragments from both genomes are shown. Discrimination between the two on the basis of (G+C)-content is impossible in most cases, because both histograms largely overlap. In the lower part (b), histograms of Pearson's correlation coefficients are shown for all possible pairwise comparisons of the fragment's tetranucleotide-derived z-scores. Correlation within both genomes is high (0.69 - 0.96) while being low between them (0.06 - 0.49). Discrimination of both genomes on the basis of tetranucleotide usage patterns is possible in all cases.

discriminatory power, since the correlations of tetranucleotide usage patterns will be high between fragments form the same genome and low between fragments from different genomes. Di- and trinucleotides provide less permutations and hence their discriminatory power is reduced (see Fig. 3 in Sandberg *et al.*, 2001).

Our systematic evaluation of the discriminatory power of the Δ(G+C) and tetranucleotide-derived z-score correlations revealed that, in the majority of cases, the discriminatory power of the latter exceeded that of Δ(G+C) (Fig. 2). If one compares the fraction of fragment pairs from two species that can be assigned to their original genomes, then discrimination by the Δ(G+C) was only superior regarding the number of genome pairings with a near to perfect discrimination (less than 4% non-assignable fragment pairs). If one includes, however, genome pairings with a higher percentage of non-assignable fragment pairs, tetranucleotide-derived z-score correlations outperformed Δ(G+C) considerably. For instance, the number of genome pairs with at most 35% non-assignable fragment pairs was 5,131 (74.3%) when the Δ(G+C) was used, but 6,397 (92.7%) when tetranucleotide-derived z-score correlations were used (Fig. 2, dotted lines). Discrimination was completely impossible for 463 (6.7%) genome pairings when Δ(G+C) was used, but only for 96 (1.4%) genome pairings when tetranucleotide-derived z-score correlations were applied.

In real metagenome projects, the situation is even more complicated. Neither Δ(G+C) values nor tetranucleotide-derived z-score correlations of all possible fragment pairs within and between all genomes are known for real metagenome libraries. Thus, in the absence of marker genes, there is at first no chance to assess whether a given Δ(G+C) or tetranucleotide-derived z-score correlation between two fragments supports or contradicts the assumption that they originate from the same group or species. In order to provide the background for a decision guideline, we investigated how well a given Δ(G+C) or tetranucleotide-derived z-score correlation discriminates on the phylogenetic levels of species, orders, classes, phyla and domains. The results are summarized in Tab. 1 and 2. For instance,

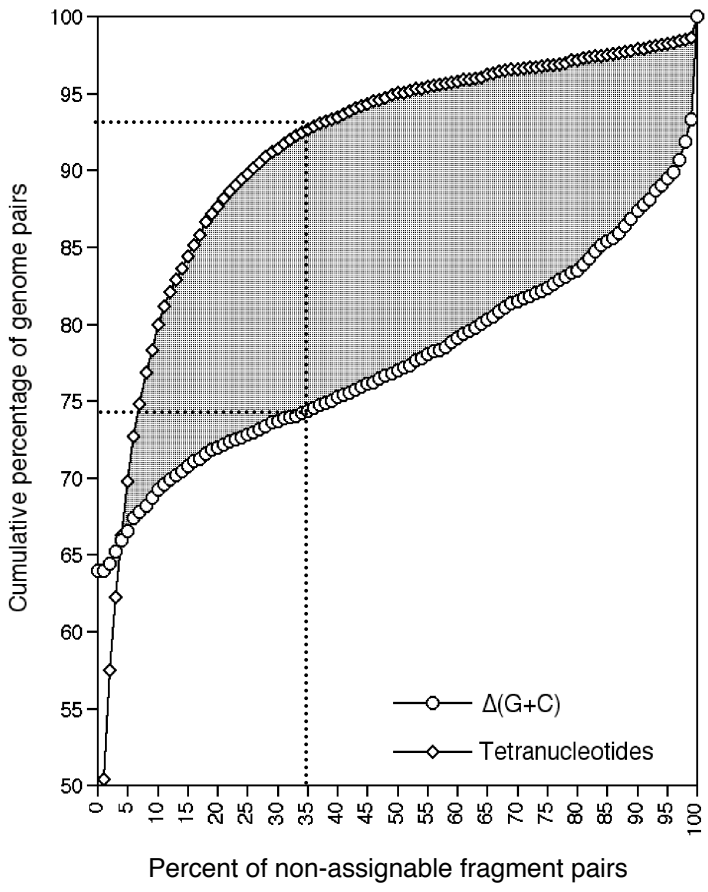


Fig.2 For all 6,903 pairwise comparisons of the 118 bacterial genomes investigated, the percentage of non-assignable fragment pairs was calculated for the Δ(G+C) and for tetranucleotide-derived z-score correlations (abscissa). Both, inter- and intra-genome fragment pairs were considered. The number of genome pairs having the indicated or a better (i.e. smaller) percentage of non-assignable fragment pairs was plotted on the ordinate (cumulative plot). The hatched area indicates the region, where tetranucleotide-derived z-score correlations provide a better resolution than the Δ(G+C) and the dotted lines refer to the maximum difference between both methods.

when two fragments are randomly chosen from the same species, the average probability of obtaining a Δ(G+C) of six percent or less is 98.0% (Tab. 1). When two fragments are randomly chosen from different species, the average probability for a Δ(G+C) of six percent or less is 21.4%. A tetranucleotide-derived z-score correlation of 0.6 corresponds to a Δ(G+C) of 6% or better, since both yield nearly the same results within species (98.0% vs. 98.2%, Tab. 2). However, the background frequency, i.e. the average probability of a correlation coefficient of 0.6 or better for fragments chosen from an arbitrary pair of genomes, is only 8.3%. This is almost three times less than the corresponding Δ(G+C) value of 21.4%. This elevated signal-to-noise ratio of tetranucleotide-derived z-score correlations becomes even more pronounced when smaller Δ(G+C) values and higher correlation coefficients are compared. In general, on each phylogenetic level, tetranucleotide-derived z-score correlations provide a better signal-to-noise ratio than corresponding Δ(G+C) values as long as the correlation coefficients are better than 0.5. This value can be considered the threshold at which a correlation coefficient begins to discriminate a signal (the relatedness of two fragments) from the noise (an arbitrary relationship).

	Species		Order		Class		Phylum		Domain	
Δ(G+C)	within	between	within	between	within	between	within	between	within	between
cases ≤ .5%	23.0 ± 6.1	2.2 ± 0.7	8.7 ± 6.4	1.9 ± 0.7	6.5 ± 6.0	1.7 ± 0.8	4.7 ± 5.0	1.7 ± 0.9	1.2 ± 0.1	1.2 ± 0.1
cases ≤ 1%	43.2 ± 10.3	4.3 ± 1.5	16.8 ± 11.9	3.8 ± 1.5	12.6 ± 11.0	3.6 ± 1.6	9.1 ± 9.0	3.3 ± 1.9	2.3 ± 0.6	2.3 ± 0.7
cases ≤ 2%	71.6 ± 12.4	8.7 ± 2.8	31.9 ± 21.2	7.6 ± 2.9	23.9 ± 19.8	7.2 ± 3.1	17.8 ± 16.7	6.8 ± 3.7	2.8 ± 0.2	2.8 ± 0.2
cases ≤ 3%	86.0 ± 10.0	13.0 ± 4.0	44.8 ± 27.1	11.5 ± 4.0	33.3 ± 25.2	10.9 ± 4.5	25.4 ± 21.3	10.3 ± 5.4	3.5 ± 0.2	3.5 ± 0.3
cases ≤ 4%	92.9 ± 7.2	17.2 ± 5.0	55.3 ± 30.7	15.4 ± 5.1	41.2 ± 28.5	14.7 ± 5.6	32.1 ± 23.8	13.9 ± 7.0	4.4 ± 0.3	4.5 ± 0.3
cases ≤ 5%	96.3 ± 4.9	21.4 ± 6.0	63.7 ± 32.0	19.3 ± 6.1	48.2 ± 30.1	18.4 ± 6.2	38.1 ± 25.2	17.5 ± 8.5	5.6 ± 0.3	5.7 ± 0.4
cases ≤ 6%	98.0 ± 3.3	25.4 ± 7.0	70.2 ± 31.4	23.3 ± 7.1	54.4 ± 30.7	22.2 ± 7.6	44.0 ± 25.9	21.2 ± 9.9	6.8 ± 0.4	7.1 ± 0.4
cases ≤ 7%	98.9 ± 2.2	29.5 ± 7.9	75.2 ± 30.3	27.3 ± 8.0	60.1 ± 30.9	26.1 ± 8.4	49.5 ± 26.6	24.9 ± 11.2	8.1 ± 0.4	8.6 ± 0.4
cases ≤ 8%	99.9 ± 1.4	33.4 ± 8.9	79.0 ± 29.0	31.2 ± 9.0	64.8 ± 30.9	30.0 ± 9.3	54.4 ± 27.3	28.6 ± 12.2	9.5 ± 0.4	10.1 ± 0.5
cases ≤ 9%	99.7 ± 0.9	37.2 ± 9.8	81.9 ± 27.7	35.1 ± 9.9	68.8 ± 30.9	33.8 ± 10.1	58.5 ± 27.4	32.3 ± 13.6	11.0 ± 0.4	11.7 ± 0.5
cases ≤ 10%	99.8 ± 0.6	40.9 ± 10.7	84.2 ± 26.7	38.9 ± 10.7	72.0 ± 30.9	37.6 ± 10.9	62.2 ± 27.9	36.0 ± 14.2	12.5 ± 0.4	13.4 ± 0.5
cases ≤ 11%	99.9 ± 0.4	44.6 ± 11.5	85.9 ± 25.8	42.7 ± 11.5	74.6 ± 30.6	41.4 ± 11.6	65.6 ± 27.9	39.6 ± 15.0	14.0 ± 0.5	15.1 ± 0.5
cases ≤ 12%	100.0 ± 0.3	48.2 ± 12.3	87.4 ± 24.8	46.5 ± 12.3	77.1 ± 29.8	45.2 ± 12.3	69.0 ± 27.6	43.2 ± 15.8	48.6 ± 2.2	52.0 ± 6.7

Tab. 1 Differences in (G+C)-content between 9,054 artificial 40 kb fragments from 118 entire genomes (126 chromosomes). The results of all 40,982,931 pairwise comparisons were summarized on the levels of species, orders, classes and phyla (taxonomy according to the list of bacterial names with standing in nomenclature (LBSN) - http://www.bacterio.cict.fr/). Tabulated values are expressed as average percentages of cases (mean ± standard deviation).

This can also be seen if one calculates the likelihood indicated by a given Δ(G+C) or tetranucleotide-derived z-score correlation that a pair of fragments belongs to the same species (Fig. 3). For the 118 genomes investigated, even a Δ(G+C) of zero translates to a probability of only 10.4%. In other words, the number of inter-genome fragment pairs with a Δ(G+C) of zero is about nine times as high as the intra-genome count. A high correlation of tetranucleotide-derived z-scores, however, discriminates far better. For instance, a correlation of 0.94 equals a probability of 79.5% that two fragments originate from the same species. Interestingly, higher correlation coefficients are very rare within genomes. Highly-correlated fragment-pairs can always be found by chance between genomes, however, because the number of inter-genome fragment pairs is two orders of magnitude higher than the number of intra-genome pairs. Therefore, the discriminatory power of tetranucleotide-derived z-score correlations drops dramatically for correlation coefficients above 0.94. It is also important to note that the discriminatory power of both methods decreases with the number of species that are present in the library (Fig. 3), because the number of intra-species fragment pairs (i.e. the signal) increases linearly with the number of species whereas the number of inter-species fragment pairs (i.e. the noise) increases qudratically. This implies that, in order to achieve a good signal-to-noise ratio for sequence-based fragment assignment, the overall diversity within metagenome libraries should be as low and the abundance of the species of interest should be as high as possible. The 118 genomes used in this study might not be representative for many natural

	Species		Order		Class		Phylum		Domain	
Pearson's r	within	between	within	between	within	between	within	between	within	between
cases ≥..95	0.2 ± 0.9	0.0 ± 0.2	0.2 ± 0.6	0.0 ± 0.0	0.1 ± 0.5	0.0 ± 0.0	0.0 ± 0.2	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
cases ≥ 0.9	12.6 ± 21.8	0.1 ± 0.2	2.5 ± 9.3	0.0 ± 0.0	1.5 ± 5.8	0.0 ± 0.0	0.3 ± 0.6	0.0 ± 0.0	0.1 ± 0.3	0.0 ± 0.0
cases ≥ 0.8	65.5 ± 30.4	1.1 ± 2.5	13.9 ± 22.3	0.1 ± 0.3	8.6 ± 14.9	0.1 ± 0.3	3.3 ± 5.9	0.0 ± 0.0	1.3 ± 2.8	0.0 ± 0.1
cases ≥ 0.7	91.1 ± 13.3	3.1 ± 3.9	33.7 ± 26.8	1.1 ± 2.4	21.7 ± 19.9	1.0 ± 2.6	10.4 ± 10.3	0.6 ± 1.5	3.7 ± 4.3	0.3 ± 0.9
cases ≥ 0.6	98.2 ± 3.2	8.3 ± 7.0	55.8 ± 29.6	5.3 ± 5.6	41.3 ± 26.3	4.7 ± 5.8	27.8 ± 19.8	4.0 ± 4.7	10.7 ± 7.3	1.5 ± 2.7
cases ≥ 0.5	99.5 ± 1.3	19.8 ± 12.0	65.5 ± 30.0	17.1 ± 10.4	56.9 ± 29.2	15.8 ± 9.8	41.0 ± 27.4	14.8 ± 9.8	22.3 ± 12.1	6.8 ± 6.4
cases ≥ 0.4	99.8 ± 0.9	40.4 ± 17.4	70.8 ± 28.0	38.5 ± 16.4	69.3 ± 27.8	37.1 ± 16.1	58.4 ± 28.1	35.7 ± 16.6	43.5 ± 17.3	23.5 ± 16.5
cases ≥ 0.3	99.9 ± 0.6	63.9 ± 18.4	75.3 ± 24.8	62.9 ± 18.1	78.1 ± 24.4	61.8 ± 18.1	72.4 ± 25.0	61.0 ± 18.1	66.5 ± 18.4	49.7 ± 22.8
cases ≥ 0.2	99.9 ± 0.3	82.3 ± 14.8	77.2 ± 23.6	81.9 ± 14.6	82.0 ± 21.7	81.4 ± 14.7	81.3 ± 20.5	81.3 ± 14.6	83.8 ± 14.5	74.4 ± 20.9
cases ≥ 0.1	100.0 ± 0.0	93.1 ± 9.1	78.4 ± 20.8	93.0 ± 8.9	85.1 ± 16.4	92.8 ± 9.0	87.2 ± 14.6	92.9 ± 8.8	93.7 ± 8.7	90.4 ± 13.2

Tab. 2 Correlation coefficients of tetranucleotide usage patterns between 9,054 artificial 40 kb fragments from 118 entire genomes (126 chromosomes). The results of all 40,982,931 pairwise comparisons were summarized on the levels of species, orders, classes and phyla (taxonomy according to the list of bacterial names with standing in nomenclature (LBSN) - http://www.bacterio.cict.fr/). Tabulated values are expressed as average percentages of cases (mean ± standard deviation).

sampling sites, but they indicate that discrimination of more than 100 species that contribute evenly to a library is difficult when using tetranucleotide frequencies and almost impossible when using the Δ(G+C). When, however, the species of interest clearly dominate a metagenome library, the noise might be low enough to allow for their discrimination even in the presence of a few hundred species of lower abundance. In this regard, optimal sampling sites are microbial consortia, enrichment cultures or extreme habitats where species dominate or the natural diversity is reduced. Based on log-normal distributions, it has been estimated, that the biodiversity of ocean water comprises only 160 different species per milliliter, whereas soil harbors several thousands of species per gram (Curtis *et al.*, 2002). While the tetranucleotide method is likely to perform well with ocean water samples, it will fail for the analysis of metagenome libraries from soil samples and other highly diverse habitats. In addition to these limitations, being a sequenced-based measure, the tetranucleotide method is affected by intragenomic fluctuations in base-composition. For example, fragments that exhibit an atypical tetranucleotide usage because they carry a high number of laterally transferred genes will not be assigned correctly. Also, the method is not likely to be able to interrelate fragments from genomes with a high degree of sequence polymorphism and thus inhomogeneous tetranucleotide usage.

Tetranucleotide usage patterns carry a phylogenetic signal. Therefore, discrimination based on tetranucleotide usage patterns is possible not only on the species-level but also on the level of higher-order taxa, albeit with decreased discriminatory power (Tab. 2). Pride *et al.* (2003) for example used distances of tetranucleotide frequencies calculated by a zero-order Markov model to construct a phylogenetic tree for 27 bacterial genomes. We found that their results can be improved when whole genome sequences and tetranucleotide-derived z-score correlations from a maximal-order Markov model are used as distance measure (unpublished data - phylogenetic trees

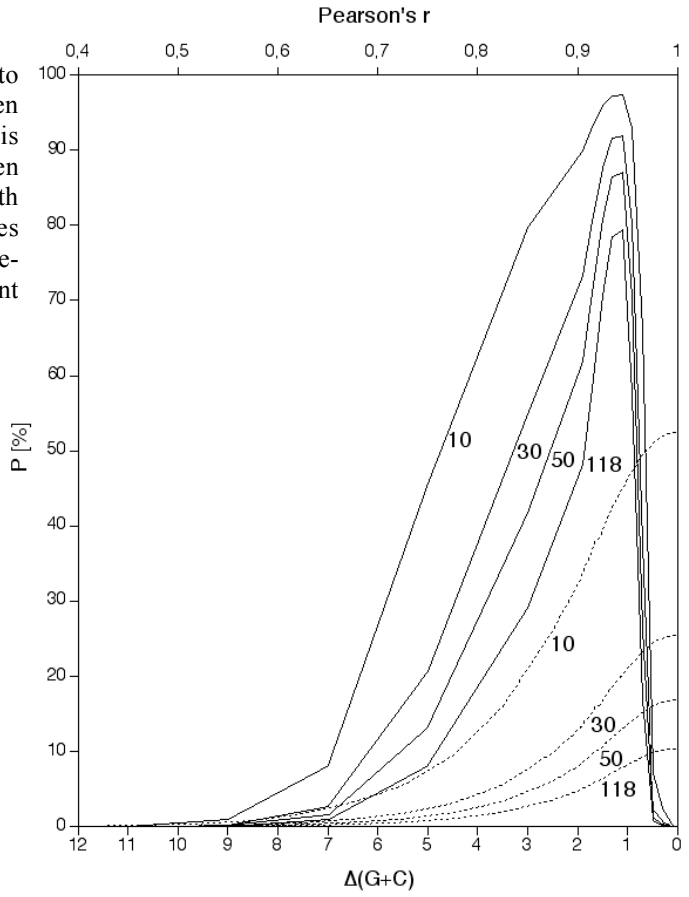


Fig. 3 Probability (P) for two 40 kb fragments to originate from the same species. Values have been calculated for the 118 species investigated in this study and for theoretical libraries containing even amounts of fragments from 10, 30 and 50 species with average-sized genomes of 3.1 Mb. Solid lines represent probabilities obtained for tetranucleotide-derived z-score correlations and dotted lines represent probabilities obtained for the Δ(G+C).

available on request). However, despite the evident phylogenetic signal in tetranucleotide frequencies, both methods fail to reconstruct phylogenetic trees for all publicly available genomes, that reflect the standard 16S rRNA-based topology. While closely related species are correctly grouped in most cases, more distant ones often are not. In other words, the phylogenetic signal in tetranucleotide usage patterns quickly fades in moving from the species level to the higher order taxa. Both, phylogenetic analysis and the data presented in table 2 show that distant relationships cannot be resolved on the basis of tetranucleotide usage patterns.

Application of the tetranucleotide method to real fosmid insert sequences

In addition to the *in silico* assessment of tetranucleotide-derived z-score correlations as a measure for the relatedness of genomic fragments, we applied the method to real fosmid insert sequences originating from two metagenome libraries. These libraries were constructed from samples of methane-rich habitats, that exhibit AOM activity and are characterized by high abundances of sulfate-reducing bacteria and *Archaea* of the ANME-2 (Boetius *et al.*, 2000) and the ANME-1 cluster (Michaelis *et al.*, 2002), respectively. Consequently, these libraries were dominated by few species. Phylogenetically, ANME-2 belongs to *Methanosarcinales* while ANME-1 is only distantly related to this order. In total, six insert sequences were analyzed - two non-overlapping inserts with 16S rRNA genes belonging to ANME-2 (ANME-2a; ANME-2c) and four overlapping inserts belonging to ANME-1 (Fig. 4). Two of the ANME-1 inserts carried identical 16S rRNA genes and

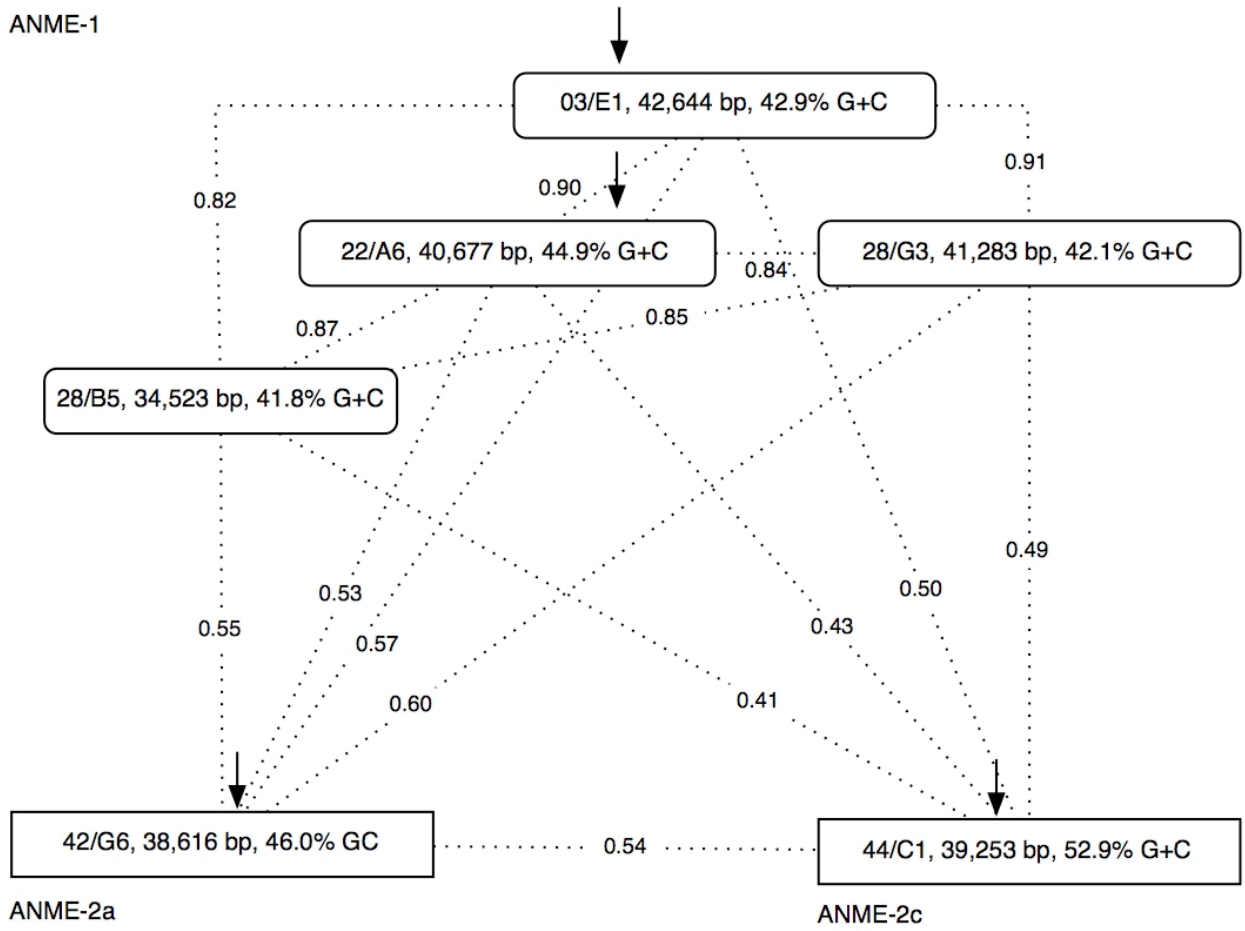


Fig. 4 Correlation coefficients of tetrancleotide-dervied z-scores between genomic fragments from bacterial communities mediating AOM. Fragments are represented as rounded boxes (ANME-1) or rectangles (ANME-2). The overlapping regions of the ANME-1 fragments correspond proportion-wise to their observed overlaps. Dotted lines and numbers correspond to the respective correlation coefficients and arrows indicate the positions of 16S rRNA genes.

overlapped by 21.0 kb, with 47 mismatches in the overlap. This indicates that they belong to closely related but different species. For each of these two inserts, a perfectly overlapping insert without the 16S rRNA gene is available (overlaps 17.5 and 10.4 kb, respectively). The ANME-1 sequences have an average (G+C)-content of 42.9% with a maximum Δ(G+C) of 3.1%. By this measure, they are clearly different from ANME-2c (52.9% G+C) but not from ANME-2a (46.0% G+C). Based on the values obtained for a hypothetical library of ten species (Fig. 3), a Δ(G+C) of 3.1% as observed within the ANME-1 fragments and between the ANME-1 fragments and ANME-2a corresponds to a probability of 21% that these fragments belong to the same species. The Δ(G+C) of 10% that has been found between ANME-1 and ANME-2c corresponds to a probability of less than 1%. Thus, from the Δ(G+C), it is highly unlikely that the ANME-1 and ANME-2c fragments belong to the same species. Discrimination between the ANME-1 and ANME-2a fragments, however, is not possible on the basis of their Δ(G+C), even though their 16S rDNA sequences indicate that they belong to different species (81% 16S rRNA identity). In contrast to this, discrimination results are congruent with the results of the 16S rRNA sequence analysis when tetranucleotide-derived z-score correlations are used. Correlation is high among the ANME-1 fragments (0.82-0.91; identical 16S rRNA) and low between fragments of ANME-1/ANME-2a (≤0.60; 81% 16S rRNA similarity), ANME-1/ANME-2c (≤0.50; 83% 16S rRNA similarity) and ANME-2a/ANME-2c (0.54; 90% 16S rRNA similarity). Referring to the ten-species curve (Fig. 3), a correlation of 0.91-0.82 corresponds to a probability of 75-93% that the ANME-1 fragments originate from the same species, while all other correlation coefficients correspond to a probability of less than 4%. Thus, in contrast to the Δ(G+C) method, tetranucleotide-derived z-score correlations are able to discriminate both, the ANME-2c fragment as well as the ANME-2a fragment from the others and in addition strongly suggest that all four ANME-1 fragments belong to the same or at least very closely related species.

We would like to emphasize that in the case of the ANME-1 insert sequences, tetranucleotide-derived z-score correlations were high not only between the overlapping inserts (which is expected, since they share considerable parts of their sequences), but also between the non-overlapping inserts. Therefore, we regard the tetranucleotide method as being well-suited to tackle the fragment identification problem in metagenomics.

Implications for metagenomics

Based on tetranucleotide usage patterns, genomic fragments derived for the same (or closely related) species could be assigned with reasonable probabilities even in the absence of suitable marker genes. Thus, together with such widely used identification approaches as marker genes, the Δ(G+C), or the gene's best BLAST hits and codon usage, the analysis of tetranucleotide-derived z-score correlations enhances our capability to classify genomic fragments. Further improvements of the method could include the combination of di-, tri- and tetranucleotide frequencies and the application of the self organizing map variant of neuronal networks. Using these methods, it has been shown that in most cases genomic fragments of 10 kb and sometimes even fragments of 1 kb retain species-specific information (Abe *et al.*, 2003). Even sequences as short as 400 bases can be correctly classified with 85% probability, when the sequences of the genomes they belong to are known and thus a model for signature oligonucleotides can be built. This has been demonstrated using a naïve Bayesian classifier for dimers up to nonamers (Sandberg *et al.*, 2001). These results indicate the large potential of genome linguistic approaches to solve the fragment identification problem in metagenomics. Hence, sequencing does not need to be restricted to genomic fragments carrying phylogenetic markers or functional genes of interest.

When the relatedness of genomic fragments is to be assessed on the basis of skewed oligonucleotide distributions, complete sequencing rather than the cost-effective end-sequencing is recommended since reliability improves with sequence length. We hope that

the tetranucleotide method will grow to be a valuable addition to the assignment tools available to scientists in the field of metagenomics.

EXPERIMENTAL PROCEDURES

Sequences for in silico evaluation of (G+C)-content and tetranucleotide usage patterns

Sequences of 116 publicly available prokaryote genomes were obtained from the NCBI website (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/). These were complemented by the recently published genome sequences of *Pirellula* sp. strain 1 (Glöckner *et al.*, 2003), the as yet unpublished genome sequence of *Desulfotalea psychrophila* from the REGX consortium (www.regx.de) and six insert sequences from metagenome libraries from methane-rich sites (Meyerdierks *et al.*, in preparation). In total, 126 chromosomes from the following 118 genomes were evaluated: *Aeropyrum pernix* K1, *Agrobacterium tumefaciens* strain C58 Cereon, *Agrobacterium tumefaciens* strain C58 UWash, *Aquifex aeolicus* VF5, *Archaeoglobus fulgidus* DSM 4304, *Bacillus cereus* ATCC 14579, *Bacillus halodurans* C-125, *Bacillus subtilis* subsp. *subtilis* 168, *Bacteroides thetaiotaomicron* VPI-5482, *Bifidobacterium longum* NCC2705, *Borrelia burgdorferi* B31, *Bradyrhizobium japonicum* USDA110, *Brucella melitensis* 16M, *Brucella suis* 1330, *Buchnera aphidicola* APS, *Buchnera aphidicola* Sg, *Buchnera aphidicola* Sg, *Campylobacter jejuni* subsp. *jejuni* NCTC 11168, *Caulobacter crescentus* CB15, *Chlamydia muridarum* strain Nigg, *Chlamydia trachomatis* serovar D, *Chlamydophila caviae* GPIC, *Chlamydophila pneumoniae* AR39, *Chlamydophila pneumoniae* CWL029, *Chlamydophila pneumoniae* J138, *Chlorobium tepidum* TLS, *Clostridium acetobutylicum* ATCC 824, *Clostridium perfringens* 13, *Clostridium tetani* E88, *Corynebacterium efficiens* YS-314, *Corynebacterium glutamicum* ATCC 13032, *Coxiella burnetii* RSA 493, *Deinococcus radiodurans* R1, *Desulfotalea psychrophila*, *Enterococcus faecalis* V583, *Escherichia coli* CFT073, *Escherichia coli* K12-MG1655, *Escherichia coli* O157:H7 VT2-Sakai, *Escherichia coli* O157:H7 EDL933, *Fusobacterium nucleatum* subsp. *nucleatum* ATCC 25586, *Haemophilus influenzae* Rd, *Halobacterium* sp. NRC-1, *Helicobacter pylori* 26695, *Helicobacter pylori* J99, *Lactococcus lactis* subsp. *lactis* IL1403, *Leptospira interrogans* serovar *lai* strain 56601, *Listeria innocua* CLIP 11262, *Listeria monocytogenes* EGD-e, *Mesorhizobium loti* MAFF303099, *Methanothermobacter thermoautotrophicus* delta H, *Methanocaldococcus jannaschii* DSM2671, *Methanopyrus kandleri* AV19, *Methanosarcina acetivorans* C2A, *Methanosarcina mazei* Goe1, *Mycobacterium leprae* TN, *Mycobacterium tuberculosis* CDC1551, *Mycobacterium tuberculosis* H37Rv, *Mycoplasma genitalium* G37, *Mycoplasma penetrans* HF-2, *Mycoplasma pneumoniae* M129, *Mycoplasma pulmonis* UAB CTIP, *Neisseria meningitidis* serogroup A Z2491, *Neisseria meningitidis* MC58, *Nostoc* sp. PCC 7120, *Oceanobacillus iheyensis* HTE831, *Pasteurella multocida* PM70, *Pirellula* sp. strain 1, *Pseudomonas aeruginosa* PA01, *Pseudomonas putida* KT2440, *Pyrobaculum aerophilum* IM2, *Pyrococcus abyssi* GE5, *Pyrococcus furiosus* DSM 3638, *Pyrococcus horikoshii shinkaj* OT3, *Ralstonia solanacearum* GMI1000, *Rickettsia conorii* Malish 7, *Rickettsia prowazekii* Madrid E, *Salmonella enterica* subsp. *enterica* serovar *Typhi*, *Salmonella typhi* CT18, *Salmonella typhimurium* LT2 SGSC1412, *Shewanella oneidensis* MR1, *Shigella flexneri* 2a strain 301, *Shigella flexneri* 2a 2457T, *Sinorhizobium meliloti* 1021, *Staphylococcus aureus* Mu50, *Staphylococcus aureus* subsp. *aureus* MW2, *Staphylococcus aureus* subsp. *aureus* N315, *Staphylococcus epidermidis* ATCC 12228, *Streptococcus agalactiae* A909, *Streptococcus agalactiae* 2603V/R, *Streptococcus agalactiae* NEM316, *Streptococcus mutans* UA159, *Streptococcus pneumoniae* R6, *Streptococcus pneumoniae* TIGR4, *Streptococcus pyogenes* MGAS315, *Streptococcus pyogenes* MGAS8232, *Streptococcus pyogenes* SF370 serotype M1, *Streptomyces coelicolor* A3(2), *Sulfolobus solfataricus* P2, *Sulfolobus tokodaii* strain 7, *Synechocystis* sp. PCC 6803, *Thermoanaerobacter tengcongensis* MB4(T), *Thermoplasma acidophilum* DSM 1728, *Thermoplasma volcanium* GSS1, *Thermo-*

synechococcus elongatus BP-1, *Thermotoga maritima* MSB8, *Treponema pallidum* Nichols, *Tropheryma whipplei* TW08 27, *Tropheryma whipplei* strain Twist, *Ureaplasma urealyticum parvum biovar* serovar 3, *Vibrio cholerae* El Tor N16961, *Vibrio vulnificus* CMCP6, *Wigglesworthia brevipalpis*, *Xanthomonas axonopodis* pv. *citri* 306, *Xanthomonas campestris* pv. *campestris* ATCC 33913, *Xylella fastidiosa* 9a5c, *Xylella fastidiosa* Temecula1, *Yersinia pestis* CO92, *Yersinia pestis* KIM.

Assessment of the resolution power of (G+C)-content and tetranucleotide usage patterns

All 118 genomes were split into artificial fosmid-sized 40 kb fragments (9,054 in total). The Δ(G+C) and tetranucleotide-derived z-score correlations were calculated for all 40,982,931 possible fragment pairs (460,084 intra- and 40,522,847 inter-genome pairs). Results were summarized on the levels of species, orders, classes and phyla (Tab. 1 and 2).

Calculation of tetranucleotide frequencies and correlation coefficients

In brief, all fragments were extended with their reverse complements. The observed frequencies of all 256 possible tetranucleotides and their corresponding expected frequencies were computed for these sequences. The differences between observed and expected values were transformed into z-scores for each tetranucleotide. The similarity be-tween two fosmids was assessed by calculating the Pearson correlation coefficient for their 256 tetranucleotide-derived z-scores.

In more detail, the expected frequencies and z-scores were computed according to the method published by Schbath *et al.* (Schbath *et al.*, 1995): If we denote the observed frequency of a tetranucleotide within a given sequence as N(n1n2n3n4), then the corresponding expected frequency E(n1n2n3n4) can be calculated by means of a maximal-order Markov model:

$$E(n1n2n3n4) = \frac{N(n1n2n3)N(n2n3n4)}{N(n2n3)}$$

The significance of the level of over- or under-representation, i.e. the divergence between observed and expected frequencies, can be evaluated using z-scores

$$Z(n1n2n3n4) = \frac{N(n1n2n3n4) - E(n1n2n3n4)}{\sqrt{\mathrm{var}(N(n1n2n3n4))}}$$

whereby the variance var(N(n1n2n3n4)) can be approximated as follows:

$$\mathrm{var}(N(n1n2n3n4)) = E(n1n2n3n4) * \frac{[N(n2n3) - N(n1n2n3)][N(n2n3) - N(n2n3n4)]}{N(n2n3)^2}$$

The question, if two genomic fragments exhibit similar patterns of over- and underrepresented tetranucleotides, can be addressed by calculating the Pearson correlation coefficient for their z-scores. Similar patterns correlate and thus have high correlation coefficients, whereas diverging patterns have low correlation coefficients.

Calculation of percentages of non-assignable fragment pairs

When comparing two genomes, fragment pairs can only be assigned with certainty to their genomes of origin when the respective Δ(G+C) or tetranucleotide-derived z-score correlations do not reside within regions where the values of both genomes overlap (Fig. 1). For each genome pairing, the number of fragment-pairs was determined that fulfill this condition (for both, intra- and inter-genome fragment pairs). The percentage of non-assignable fragment pairs was calculated by relating this number with the total number of possible fragment pairs (Fig. 2). For a pair of genomes (1, 2), with N1 and N2 40 kb fragments, this is the sum of all intra- and inter-genome fragment-pairs [N1(N1-1)/2 + N2(N2-1)/2 + N1N2].

Calculation of the probabilities to originate from the same species

The number of fragment pairs having a given Δ(G+C) or tetranucleotide-derived z-score

correlation was determined for all 460,084 intra- and 40,522,847 inter-genome pairs. For a hypothetical unbiased fosmid-library containing the 118 genomes investigated, the likelihood for two fragments to originate from the same species corresponds to the fraction that intra-genome fragment pairs make up from all fragment pairs having the respective Δ(G+C) or tetranucleotide-derived z-score correlation (Fig. 3). Assuming that the percentages of intra- and inter-genomic values obtained for the 118 genomes are representative for typical bacteria, numbers have also been calculated for hypothetical fosmid libraries of 10, 30 and 50 average-sized genomes, respectively (the average size of all publicly available bacterial genomes listed at NCBI to date is 3.1 Mb (http://www.ncbi.nlm.nih.gov/Genomes/)).

ACKNOWLEDGEMENTS

We would like to thank Tim Frana for cross-checking PERL scripts and results and the Max Planck society for funding this study. Retrieval of metagenome sequences was carried out within the framework of the Competence Network Göttingen 'Genome research on bacteria' (GenoMik) financed by the German Federal Ministry of Education and Research (BMBF).

REFERCENCES

Abe, T., Kanaya, S., Kinouchbi, M., Ichiba, Y., Kozuki, T. and Ikemura, T. (2003) Informatics for unveiling hidden genome signatures. *Genome Res* **13**: 693-705.

Amann, R.I., Ludwig, W., and Schleifer, K.H (1995) Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol Rev* **59**: 143-169.

Béjà, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P. *et al.* (2000a) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289** : 1902-1906.

Béjà, O., Suzuki, M.T., Koonin, E.V., Aravind, L., Hadd, A., Nguyen, L.P. *et al.* (2000b) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* **2**: 516-529.

Béjà, O., Spudich, E.N., Spudich, J.L., Leclerc, M. and DeLong, E.F. (2001) Proteorhodopsin phototrophy in the ocean. *Nature* **411**: 786-789.

Boetius, A., Ravenschlag, K., Schubert, C.J., Rickert, D., Widdel, F., Gieseke, A. *et al*. (2000) A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* **407**: 623-626.

Curtis, T.P., Sloan, W.T. and Scannell, J.W. (2002) Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci U S A* **99**: 10494-10499.

DeLong, E.F. (2002) Microbial population genomics and ecology. *Curr Opin Microbiol* **5**: 520-524.

Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., and Fertil, B. (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol* **16**: 1391-1399.

Gentles, A.J., and Karlin, S. (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res* **11**: 540-546.

Glöckner, F.O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W. *et al.* (2003). Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc Natl Acad Sci* **100**: 8298-82303.

Goldman, N. (1993) Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Res* **21**: 2487-2491.

Schloss, P.D., and Handelsman, J. (2003) Biotechnological prospects from metagenomics. *Curr Opin Biotechnol* **14**: 303-310.

Karlin, S. (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol* **1**: 598-610.

Karlin, S., and Burge, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11**: 283-290.

Karlin, S., Campbell, A.M., and Mrázek, J. (1998) Comparative DNA analysis across diverse genomes. *Annu Rev Genet* **32**: 185-225.

Karlin, S., and Ladunga, I. (1994) Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci* **91**: 12832-12836.

Karlin, S., Ladunga, I., and Blaisdell, B.E. (1994) Heterogeneity of genomes: measures and values. *Proc Natl Acad Sci* USA **91**: 12837-12841.

Karlin, S., and Mrázek, J. (2000) Predicted Highly Expressed Genes of Diverse Prokaryotic Genomes. *J Bacteriol* **182**: 5238-5250.

Michaelis, W., Seifert, R., Nauhaus, K., Treude, T., Thiel, V., Blumenberg, M. *et al.* (2002) Microbial reefs in the Black Sea fueled by anaerobic oxidation of methane. *Science* **297**: 1013-1015.

Nakashima, H., Ota, M., Nishikawa, K., and Ooi, T. (1998) Genes from nine genomes are separated into their organisms in the dinucleotide composition space. *DNA Res* **5**: 251-259.

Pride, D.T., Meinersmann, R.J., Wassenaar, T.M. and Blaser, M.J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**: 145-158.

Rondon, M.R., August, P.R., Bettermann, A.D., Bradly, S.F., Grossman, T.H., Liles, M.R. et al. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorgansims. *Appl Environ Microbiol* **66**: 2541-2547.

Sandberg, R., Winberg, G., Bränden, C.I., Kaske, A., Ernberg, I., and Cöster, J. (2001) Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genome Res* **11**: 1404-1409.

Schbath, S., Prum, B., and de Turckheim. E. (1995) Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J Comput Biol* **2**: 417-437.

4

MORFind: improved gene-prediction by the combination of gene-finders

Hanno Teeling, Jost Waldmann, Margarete Bauer and Frank Oliver Glöckner

in preparation for *Bioinformatics*

MORFind: improved gene-prediction by the combination of gene-finders

Hanno Teeling, Jost Waldmann, Margarete Bauer and Frank Oliver Glöckner

Max Planck Institute for Marine Microbiology, Department of Molecular Ecology, Genomics Group, Celsiusstrasse 1, D-28359 Bremen, Germany

ABSTRACT

Despite constant improvements in the *in silico* prediction of prokaryote protein-coding genes, this task can by no means be regarded as solved or trivial. Most respective gene-finders have been developed, optimized and evaluated for model organisms such as *Escherichia coli* or *Bacillus subtilis* and therefore their performance can vary considerably when applied to phylogenetically distant genomes. For instance, many widely-used gene-finders generate poor results regarding selectivity and sensitivity for the recently published genome of the planctomycete *Rhodopirellula baltica*^{T}. Prediction results can however be improved considerably when the results of different gene-finders are combined. Here we present MORFind, a new meta-tool for the prediction of protein-coding genes in prokaryote genomes that combines the CRITICA, Glimmer2 and ORPHEUS gene-prediction programs. MORFind allows to merge the results of these individual gene-finders into a non-redundant list via a sophisticated post-processing algorithm. The latter uses external BLASTp searches as well as signal peptide and transmembrane predictions in combination with a recursive algorithm considering the genetic neighborhood to decide between conflicting predictions. In a comparative evaluation performed on the genomes of *Escherichia coli* K-12-MG1655, *Bacillus subtilis* subsp. *subtilis* 168 and *Rhodopirellula baltica*^{T}, MORFind excelled the stand-alone gene-finders by providing a better balance between prediction selectivity and sensitivity. MORFind has been implemented as a multi-user system that runs on a UNIX server and can be operated via a dedicated, user-friendly graphical interface running either on Windows or Linux. Due to its modular architecture, the underlying gene-finders in MORFind can be changed without much effort. Hence, MORFind provides a flexible framework for the combination of different gene-finders.

INTRODUCTION

Within recent years, newly developed gene-prediction software for prokaryote genomes has been published in a more or less constant rate. Among these are Glimmer (Delcher *et al.*, 1999; Salzberg *et al.*, 1998), ORPHEUS (Frishman *et al.*, 1998), CRITICA (Badger & Olsen, 1999), GeneScan (Ramakrishna & Srinivasan, 1999), the 'frame-by-frame' algorithm (Shmatkov *et al.*, 1999), GeneMarkS (Besemer *et al.*, 2001), and more recently EasyGene (Larsen & Krogh, 2003) and ZCURVE (Guo *et al.*, 2003), to name just a few (a comprehensive overview on computational gene-recognition is maintained by W. Li on his website: http://www.nslij-genetics.org/gene). Different algorithms have been applied to the problem, like various variants of Markov models, Bayesian networks, Fourier transforms and combinations of these. Due to the increased prediction accuracy of recent prokaryotic gene-finders it is a common misconception that the task of gene-prediction has essentially been solved for prokaryotes. However, in fact many respective problems have not been ironed out. For instance, the problem of selectivity versus

sensitivity is still pending. Some gene-finders are too selective, i.e. their gene-finding algorithm is so strict that - although the accuracy of the predicted genes is high - many genes are overlooked. Others are too sensitive. They find genes nearly quantitatively, but at the price of an overprediction often exceeding 20%. Especially the prediction of short genes is problematic, because successful discrimination between coding and non-coding short open reading frames (ORFs) is handicapped by their limited information content. Consequently, many of the short ORFs predicted by over-sensitive gene-finders do not code for any known protein (Basrai *et al.*, 1997; Skovgaard *et al.*, 2001). This problem is especially prominent in AT-rich genomes, since stop codons are AT-rich and short non-coding ORFs are therefore more frequent in AT-rich than in GC-rich genomes. In extreme GC-rich genomes like those of *Halobacterium* species on the other hand, stop codons are less likely to occur by chance. This leads to unusually long non-coding ORFs that often overlap and that most gene-finders erroneously identify as genes. For larger genomes like those of *Sorangium cellulosum* SO cec56 (12.3 Mb), *Gemmata obscuriglobus* UQM2246^T (9.0 Mb) or *Rhodopirellula baltica*^{T} (7.1 Mb), 20% overprediction easily translates into 1,000 or more superfluous ORFs that have to be sorted out manually. Yet another unsolved problem besides overprediction is the accuracy of gene start prediction. Although appropriate algorithms have been implemented in many gene-finders (Besemer *et al.*, 2001; Frishman *et al.*, 1998; Hayes & Borodovsky, 1998; Yada *et al.*, 2003) and algorithms have been developed specifically with this task in mind (Buhler & Tompa, 2002; Hannenhalli *et al.*, 1999; Suzek *et al.*, 2001; Tompa, 1999; Walker *et al.*, 2002), proper start site prediction remains difficult. At least within some lineages this is caused by the low information content of the ribosomal binding site (Hayes & Borodovsky, 1998), which is mostly used for start site identification. Associated with the problem of proper start site prediction is the problem of overlapping ORFs. Although there seems to be a selective pressure towards genome compactness, genes with large overlaps are rare within the so far sequenced prokaryote genomes. They seem to originate mainly by chance from mutational stop codon loss (Fukuda *et al.*, 1999), but have also been identified as mechanism for functional coupling and co-regulation of genes (Krakauer, 2000). Most gene-finders have algorithms that assess alternate start sites in order to resolve conflicts between predicted overlapping genes. When start site relocation fails to resolve such a conflict, ORFs are eventually dismissed in favor of others. However, since start site prediction often is inaccurate, overlaps and thus ORF dismissal is possibly as well. In addition, all gene-finders published to date evaluate only pairs of genes but not the entire genetic neighborhood to resolve overlaps, which is another source of error.

In summary, a fully-automated gene-prediction pipeline that works perfectly for all prokaryotes does not exist. Sophisticated models have been applied to the problem of gene-prediction, but none seems to work in a universal fashion. It has been shown, however, that prediction accuracy can be improved by combining multiple gene-finders that rely on different prediction models (Bocs *et al.*, 2002; Makalowska *et al.*, 2001; Murakami & Takagi, 1998; Pavlovic *et al.*, 2002; Rogic *et al.*, 2002; Yada *et al.*, 2003).

Here we present such a meta-tool, MORFind, which is an acronym for MPI ORF finder. MORFind is a spin-off of the genome-annotation of the planctomycete *Rhodopirellula baltica*^{T} (Glöckner *et al.*, 2003), for which a useful gene-prediction could only be achieved by a combination of gene-finders. In comparisson with the individual gene-finders, it delivers enhanced prediction selectivities while maintaining high prediction sensitivities. MORFind also works well for the genome of *Rhodopirellula baltica*^{T} which indicates, that it works in a more universal fashion than the individual programs.

MATERIALS AND METHODS

Architecture and Implementation

In its current initial release, MORFind is based on three individual gene-finders: CRITICA, Glimmer2 and ORPHEUS (this composition is likely to change in future versions). These gene-finders have been selected because they are based on different algorithms and free of charge for academia.

Glimmer2 is an *ab initio* gene-finder, i.e. derives all information for its interpolated context prediction model from the provided sequence. Glimmer2 is very fast, since it does not need to extract a training set of coding ORFs via time-consuming database similarity searches. ORPHEUS on the other hand relies heavily on such database comparisons, that it not only uses to derive its codon weights-based gene-prediction model but also to derive a prediction model for the ribosomal binding site (RBS). Due to its complexity, ORPHEUS is the slowest of the three gene-finders used in MORFind. CRITICA finally uses a combination of comparative and non-comparative (*ab initio*) calculations for gene-prediction.

MORFind has a client/server based architecture with a usermanagement system that allows simultaneous access of multiple users to the system (Fig. 1). The client was implemented in Object Pascal with Kylix™ 3 Professional (Borland Software Corporation, CA, USA) and has been compiled for Windows and Linux. It provides the user with a graphical interface from which the settings of the individual gene-finders and the post-processing of their outputs can be controlled (Fig. 2). The upload of genomic sequences and the project-management are handled from the client as well. Multiple predictions can be set up, run in parallel and stored in order to compare the effects of different settings. The

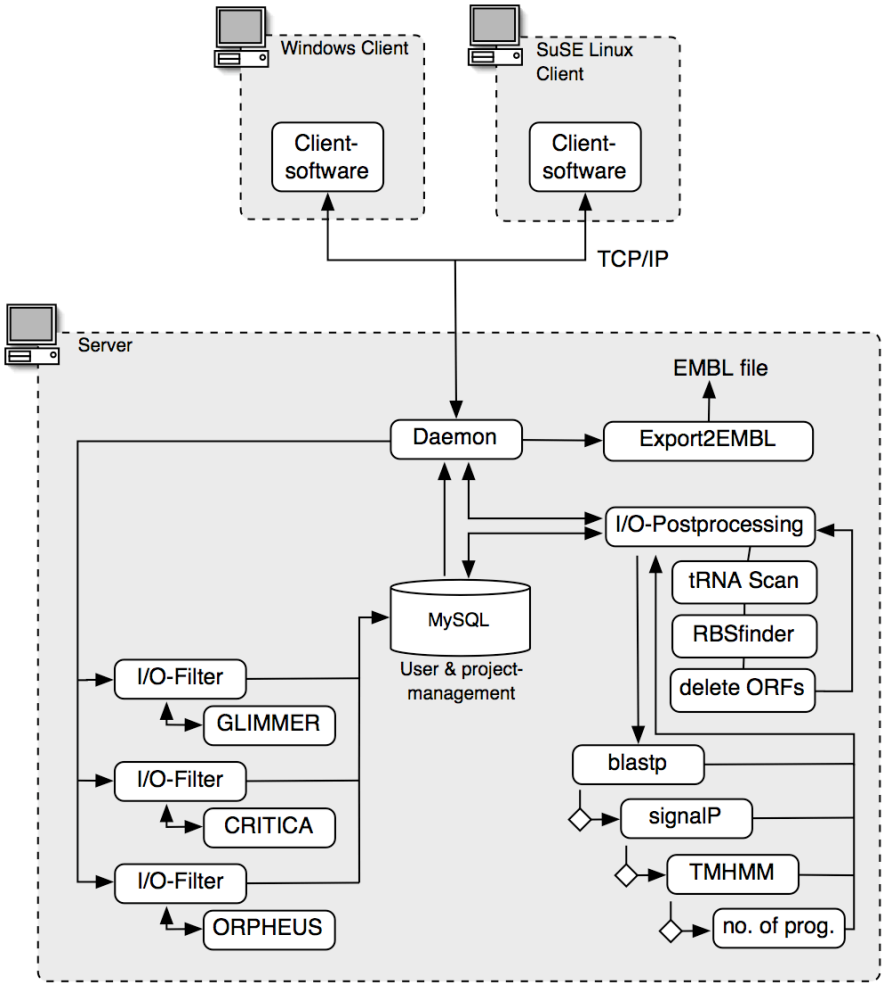


Fig. 1 Flowchart showing the basic architecture of MORFind (simplified)

client communicates through a simple, unencrypted TCP/IP-based protocol with a daemon residing on the server, which has been implemented in PERL. Input and output for the gene-finders on the server-side is handled by corresponding PERL I/O filters as is the communication with the accessory programs (Fig. 1). Thus, an extension or alternate composition of gene-finders can be implemented relatively easy. Prediction results are stored in a MySQL relational database on the server and can be exported as EMBL-formatted files for subsequent visual inspection by Artemis (Rutherford *et al.*, 2000) or imported directly into an annotation system as e.g. GenDB (Meyer *et al.*, 2003).

All programs that MORFind uses are free of charge for academia, but some require the signature of a dedicated license agreement. Likewise, MORFind can be obtained from the authors upon request without charges.

Post-processing

MORFind allows to merge the results of the individual gene-finders into a non-redundant ORF list by means of a post-processing algorithm that mimics aspects of a manual annotation:

First, tRNAs are detected with tRNAscan-SE (Lowe & Eddy, 1997). They are treated like ORFs with high-scoring BLASTp hits further down the post-processing pipeline. Then, identical ORFs and such ones that vary only in length (i.e. regarding their start codons) are merged, whereby alternate start positions are evaluated via a scoring scheme invoking RBSfinder (Suzek *et al.*, 2001), SignalP (Nielson *et al.*, 1997) and the number of gene finders that predicted a given start. In the next step, MORFind tries to resolve conflicts between overlapping ORFs. The threshold from which on MORFind regards overlaps as problematic can be adjusted in

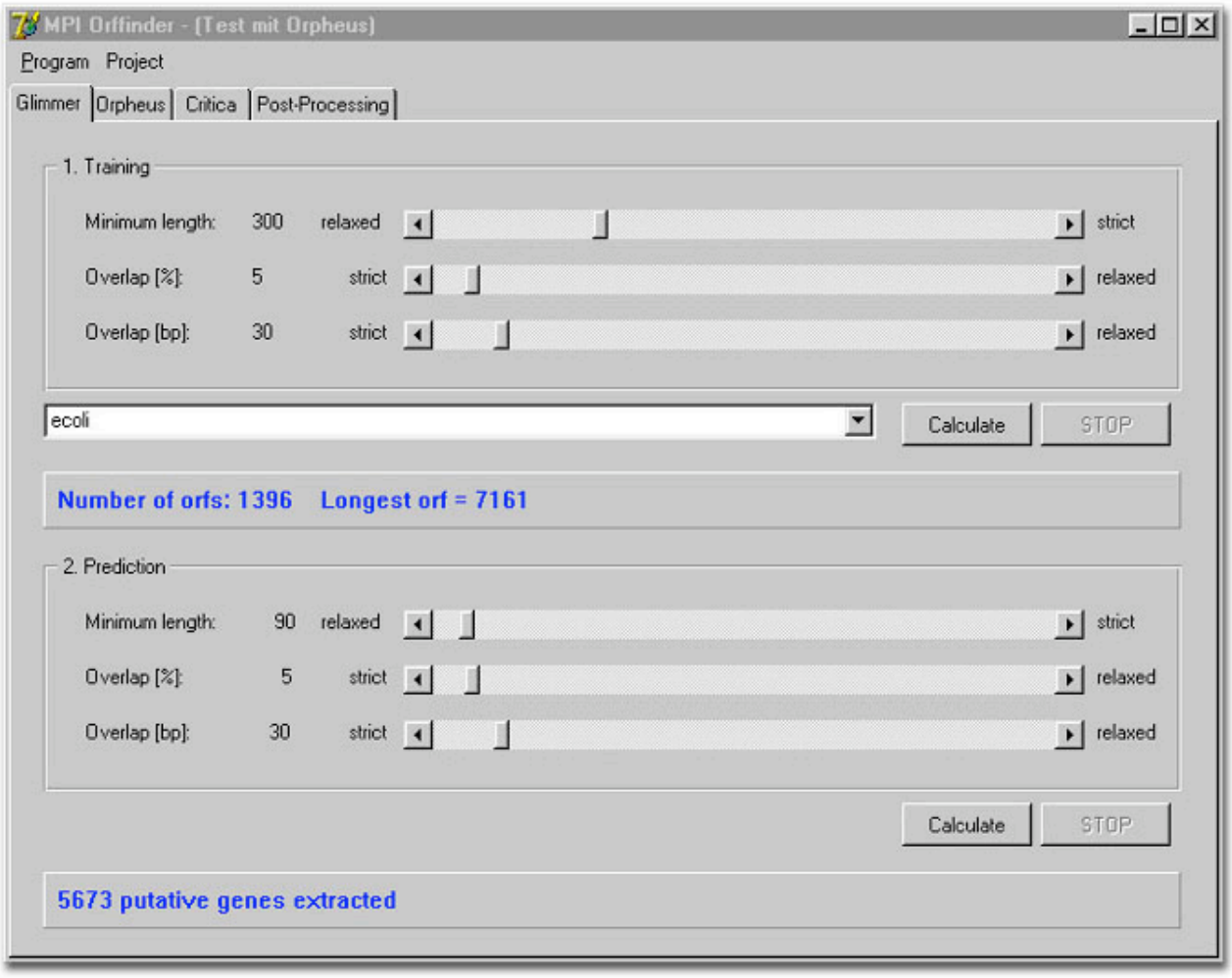


Fig. 2 Screenshot from the graphical user interface of the MORFind client for Windows.

absolute and in relative terms from the client's graphical user interface (i.e. ≥ 40 bp or ≥ 40% of the ORF's length). When two ORFs are oriented in a way that their overlap cannot be resolved by moving either start position, one of the ORFs is omitted via a scoring scheme that comprises (a) BLASTp (Altschul *et al.*, 1990) similarity searches against a small protein database compiled from all completely sequenced prokaryote genomes, (b) signal peptide predictions by SignalP, (c) prediction of transmembrane regions by TMHMM (Sonnhammer *et al.*, 1998), (e) the number of programs that predicted an ORF and (f) the ORFs' length. If the overlap cannot be resolved, both ORFs are kept. Afterwards, MORFind tries to resolve overlaps between pairs of ORFs where a relocation of at least one of the start positions is possible with RBSfinder. Remaining conflicting ORFs are then treated as chains which allows to take their genetic context into account. A chain is defined as any number of continuously arranged conflicting ORFs and is bounded by ORFs without problematic overlaps. Again, MORFind evaluates each ORF in a chain according to the above-mentioned scoring scheme invoking steps (a) - (f). Then, it recursively searches for the solution with the maximal overall score and a minimum of gap positions. Finally, MORFind allows to exclude all ORFs from the final list, that fall short of an adjustable length (i.e. ≤ 70 bp) and do not exhibit any significant BLASTp hits.

Server hardware and software requirements

The recommended minimum hardware for the server consists of a standard PC (Intel P4 2.8 GHz or speed-equivalent AMD or PPC, 1 GB RAM), running Linux with MySQL and PERL installations (SuSE Linux v. 8.1 and v. 8.2 have been tested). Since CRITICA and ORPHEUS rely on computation-intense similarity searches, faster hardware will speed up the prediction process considerably.

Evaluation

MORFind has been tested and evaluated against the genomes of *Escherichia coli* K-12-MG1655 and *Bacillus subtilis* subsp. *subtilis* 168, which must be regarded as among the best annotated and validated genomes to date. In addition, MORFind was evaluated against the genome of the planctomycete *Rhodopirellula baltica*^{T} (Glöckner *et al.*, 2003; Schlesner *et al.*, 2004), for whom MORFind was developed.

RESULTS AND DISCUSSION

In general, MORFind yielded improved results compared to the stand-alone gene-finders by providing higher selectivity than Glimmer2 and ORPHEUS and a higher sensitivity than CRITICA (Tab. 1).

With the genomes tested, ORFs predicted by CRITICA were indeed annotated as genes in ≥ 97.8% of all cases. However, this high selectivity came at the price of a lack in sensitivity. This was especially obvious for the genome of *Rhodopirellula baltica*^{T}, for which CRITICA predicted only 4,537 genes. This number is simply too low for a genome that exceeds 7.1 Mb and consequently relates only to 61.7% of the genes that were annotated. Hence, the combined comparative/non-comparative prediction model that CRITICA uses is very strict. Despite the fact that CRITICA did not quantitatively predict all annotated genes in the tested genomes, it is a valuable gene-finder since it can be used (a) to assess the quality of other gene-finders and (b) to construct training sets for other gene-finders.

The balance between sensitivity and selectivity turned out to be reversed when ORPHEUS was used. While ORPHEUS achieved high prediction sensitivities (≥ 95% for the genome of *Escherichia coli* K-12-MG1655 and even better for *Bacillus subtilis* subsp. *subtilis* 168), it ran into problems in predicting the annotated genes of *Rhodopirellula baltica*^{T}, for which it predicted around 10,000 genes This might be attributed to the fact, that the genome of *Rhodopirellula baltica*^{T} has an unusual low percentage of ORFs with database similarities to known proteins (Glöckner *et al.*, 2003). Thus, only a limited set of genes was available to ORPHEUS that it could use as training set to construct its prediction model. It is also

possible that the ORFs in the training set had a non-standard codon usage and therefore were not well-suited to build a selective model. ORPHEUS employs a sophisticated algorithm for the correct prediction of gene starts, which worked very well in the case of the *Bacillus subtilis* subsp. *subtilis* 168 and *Escherichia coli* K-12-MG1655 genomes. In *Rhodopirellula baltica*^{T} however, the model failed to predicted the correct RBS consensus sequence (CTTCAC instead of AAGGAG), which was also reflected by the fact that the final prediction contained nearly equal amounts of ORFs with ATG, GTG and TTG start codons (the annotations as well as the CRITICA and GLIMMER predictions showed a strong preference for ATG). This indicates that either the ORFs in the training set were not well-suited or that the iterative procedure that ORPHEUS uses to calculated the ribosomal binding site was trapped in a local maximum (possibly due to unfortunate start conditions). This, however, is probably not the fault of ORPHEUS since the information content of the ribosomal binding site in *Rhodopirellula baltica*^{T} is only 2.1, which is very low when compared to other genomes (Frishman *et al.*, 1999).

Like ORPHEUS, Glimmer2 is a gene finder with high sensitivity but low sensitivity - albeit with a lesser extent of overprediction. The prediction accuracy (i.e. true positives) was above 97% for the genomes of *Bacillus subtilis* subsp. *subtilis* 168 and *Escherichia coli* K-12-MG1655, but more than 10% lower for *Rhodopirellula baltica*^{T}. Overprediction was in the range of 10 - 30% and depended heavily on the treatment of overlaps in the last prediction step. The values of about 21% overprediction for the genomes of *Bacillus subtilis* subsp. *subtilis* 168 and *Escherichia coli* K-12-MG1655 correspond closely to those published by the Glimmer2 authors (Delcher *et al.*, 1999).

	gene prediction program	*in silico* prediction	hits to annotation	overlooked ORFs	'overpredicted' ORFs
B. subtilis (4103 genes)	Gimmer2	5131	4039 (98.4%)	64 (1.6%)	1092 *(21.3%)*
	Gimmer2*	4426	4005 (97.6%)	98 (2.4%)	421 *(9.5%)*
	CRITICA	3718	3667 (89.4%)	436 (10.6%)	51 *(1.4%)*
	ORPHEUS	5590	3982 (97.1%)	121 (2.9%)	1608 *(28.8%)*
	ORPHEUS*	5570	4015 (97.9%)	88 (2.1%)	1555 *(27.9%)*
	MORFind*	4444	4009 (97.7%)	94 (2.3%)	435 *(9.7%)*
E. coli (4255 genes)	Gimmer2	5270	4157 (97.7%)	98 (2,3%)	1113 *(21.1%)*
	Glimmer2*	5672	4182 (98.3%)	73 (1.7%)	1490 *(26.3%)*
	CRITICA	4023	3935 (92.5%)	310 (7.5%)	88 *(2.2%)*
	ORPHEUS	5903	4049 (95.2%)	206 (4.8%)	1854 *(31.4%)*
	ORPHEUS*	5878	4110 (96.6%)	145 (3.4%)	1768 *(30.1%)*
	MORFind*	4923	4171 (98.0%)	84 (2.0%)	752 *(15.3%)*
R. baltica (7325 genes)	Gimmer2	9204	6243 (85.2%)	1082 (14.8%)	2961 *(32.2%)*
	Gimmer2*	7578	6125 (83.6%)	1200 (16.4%)	1453 *(19.2%)*
	CRITICA	4537	4518 (61.7%)	2788 (38.3%)	19 *(0.4%)*
	ORPHEUS	10618	6434 (87.8%)	891 (12.2%)	4184 *(39.4%)*
	ORPHEUS*	9935	6506 (88.8%)	819 (11.2%)	3429 *(34.5%)*
	MORFind*	6716	5737 (78.3%)	1588 (21.7%)	979 *(14.6%)*

Tab. 1 Results of the MORFind evaluation against the genomes of *Escherichia coli* K-12-MG1655, *Bacillus subtilis* subsp. *subtilis* 168 and *Rhodopirellula baltica*^{T}. All gene finders were used with their default settings and – except but CRITICA - with modified settings that are used as defaults within MORFind (*). The total number of genes for the genomes were extracted from appropriate GenBank files from the NCBI website as they were in midyear 2001 (http://www.ncbi.nlm.nih.gov:80/genomes/static/eub_g.html). Percentages drawn in normal text refer to annotated genes, while percentages in italics refer to the gene predictions.

The higher selectivity and lower sensitivity of CRITICA versus Glimmer2 and ORPHEUS was also reflected in the distribution of the predicted ORF lengths (Fig. 3). The percentage of ORFs with less than 100 codons was much smaller for CRITICA than for Glimmer2 and ORPHEUS. The detection of short genes with their limited information content from the vast pool of short open reading frames in genomes is non-trivial. However, many important functions are encoded by short genes, e.g. some of the ribosomal proteins or stress response proteins. Our knowledge about small genes in prokaryotes is sparse, but studies indicate, that there might be more present than previously estimated (Basrai *et al.*, 1997). Therefore, strict cutoff values for short open reading frames bear the risk of overlooking important new genes (MORFind nevertheless provides such an option, which can be used to generate prediction sets with varying minimum ORF lengths).

On overall, MORFind yielded the best prediction results for all three genomes. For the genomes of *Bacillus subtilis* subsp. *subtilis* 168 and *Escherichia coli* K-12-MG1655 97.7 - 98.0% of the annotated genes were correctly predicted by MORFind, while overprediction was 9.7% and 15.3%, respectively. Thus, the post-processing algorithm of MORFind was able to maintain the high prediction sensitivities of Glimmer2 and ORPHEUS, while reducing the amount of overpredicted genes in all cases except but in one. The selectivity of MORFind was not as high as with CRITICA, which - as mentioned above - did overlook by far too many genes. As with the genomes of *Escherichia coli* K-12-MG1655 and *Bacillus subtilis* subsp. *subtilis* 168, MORFind exhibited less overprediction for the genome of *Rhodopirellula baltica*^{T} than Glimmer2 and ORPHEUS. However, MORFind found less of the annotated genes (77.8%) as Glimmer2 (84.8%) and ORPHEUS (88.7%). The reason for this is most likely not attributed to an inferior performance of MORFind, but lies within the nature of how the genome of *Rhodopirellula baltica*^{T} was annotated. During the annotation of the genome of *Rhodopirellula baltica*^{T}, the same gene-finders as within MORFind were used, however, their individual prediction results were simply merged to a non-redundant list, containing no less than 13.331 predicted ORFs. These were annotated manually, and all ORFs that were not conflicting with other ORFs were kept in order to achieve an as complete gene-prediction as possible. Conse-

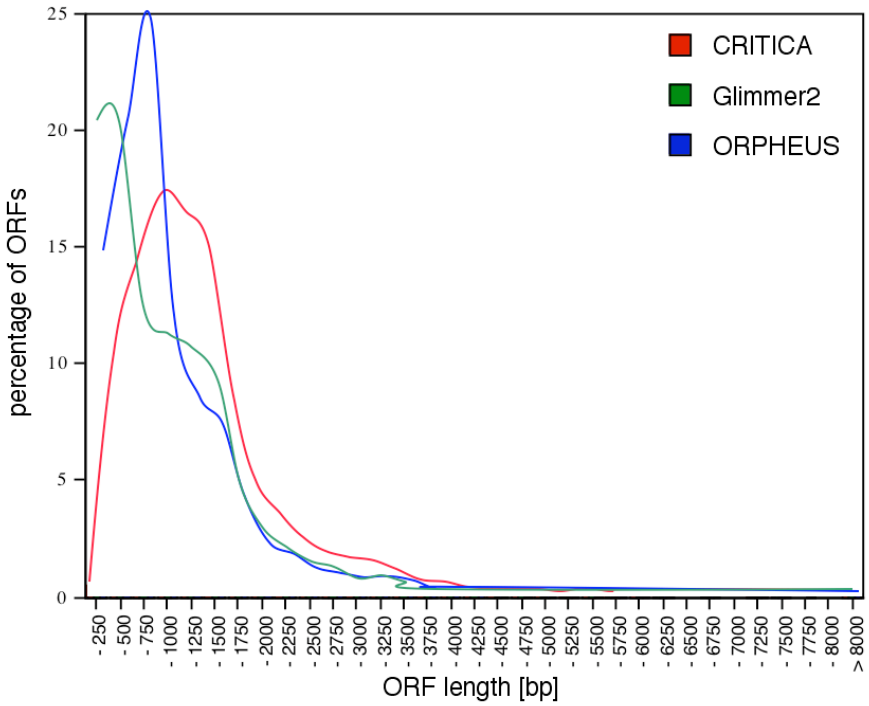


Fig. 3 Histograms for the lengths of ORFs that the three stand-alonegene-finders used in MORFind predicted for the *Rhodopirellula baltica*^{T} genome.

quently, many short ORFs were annotated as hypothetical genes, from which a large fraction (but not all) are likely overpredicitons. This also explains the high gene coverage of 95% that has been reported for the genome of *Rhodopirellula baltic*a^T (Glöckner *et al.*, 2003). MORFind did quantitatively find all genes with significant BLAST hits and all longer ORFs that have been annotated in the *Rhodopirellula baltic*a^T genome. In addition, the total number of predicted protein-coding genes that were reported by MORFind (6.716) is much more realistic than those predicted by CRITICA (4537), Glimmer2 (7578) or even ORPHEUS (9935). Therefore, the performance of MORFind is not only superior to the stand-alone gene-finders for the genomes of *Escherichia coli* K-12-MG1655 and *Bacillus subtilis* subsp. *subtilis*, 168, but most likely also for the overannoteted genome of *Rhodopirellula baltica*^{T}.

MORFind admittedly does not solve the problem of falsely predicted short genes, but it is a notable improvement over the stand-alone gene-finders it depends on. It also works in a more universal fashion than the individual gene-finders, since the short-comings and pitfalls that single gene-prediction algorithms exhibit for some genomes are compensated by the use of multiple gene-finders with different prediction models. This in general indicates the power of combinatory approaches, which consequently are more and more often applied to varying problems in bioinformatics.

CONCLUSIONS

The example of the planctomycete *Rhodopirellula baltica*^{T} illustrates that gene-finders that perform perfectly well for most known genomes are not guaranteed to work for a genome from a phylum, which is not well-studied in terms of genetics. Obviously, genomes differ in their coding sequence characteristics which affects different gene-finding algorithms to different extents. Thus, in order to minimize gene over- and under-predictions, the use of multiple gene-finders is strongly recommended when newly sequenced genomes are analyzed.

Far too many genomes are still being annotated on the basis of the prediction of a single gene-finder alone, despite the fact that the effort to use multiple gene-finders is small compared to the tedious manual detection of over- and underpredicted genes.

ACKNOWLEGEMENTS

This study was conducted within the framework of the REGX project (Real Environmental Genomics), a German initiative for sequencing and functional analysis of marine bacteria (www.regx.de). Major funding of the REGX project is provided by the German Federal Ministry of Education and Research and by the Max Planck Society.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* 215, 403-410.

Badger, J. H. & Olsen, G. J. (1999). CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* 16, 512-524.

Basrai, M. A., Hieter, P. & Boeke, J. D. (1997). Small open reading frames: beautiful needles in the haystack. *Genome Res* 7, 768-771.

Besemer, J., Lomsadze, A. & Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 29, 2607-2618.

Bocs, S., Danchin, A. & Medigue, C. (2002). Re-annotation of genome microbial coding-sequences: finding new genes and inaccurately annotated genes. *BMC Bioinformatics* 3, 5.

Buhler, J. & Tompa, M. (2002). Finding motifs using random projections. *J Comput Biol* 9, 225-242.

Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999). Improved

microbial gene identification with GLIMMER. *Nucleic Acids Res* 27, 4636-4641.

Frishman, D., Mironov, A. & Gelfand, M. (1999). Starts of bacterial genes: estimating the reliability of computer predictions. *Gene* 234, 257-265.

Frishman, D., Mironov, A., Mewes, H. W. & Gelfand, M. (1998). Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res* 26, 2941-2947.

Fukuda, Y., Washio, T. & Tomita, M. (1999). Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. *Nucleic Acids Res* 27, 1847-1853.

Glöckner, F. O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W., Gade, D., Beck, A., Borzym, K., Heitmann, K., Rabus, R., Schlesner, H., Amann, R. & Reinhardt, R. (2003). Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc Natl Acad Sci U S A* 100, 8298-8303.

Guo, F. B., Ou, H. Y. & Zhang, C. T. (2003). ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res* 31, 1780-1789.

Hannenhalli, S. S., Hayes, W. S., Hatzigeorgiou, A. G. & Fickett, J. W. (1999). Bacterial start site prediction. *Nucleic Acids Res* 27, 3577-3582.

Hayes, W. S. & Borodovsky, M. (1998). Deriving ribosomal binding site (RBS) statistical models from unannotated DNA sequences and the use of the RBS model for N-terminal prediction. *Pac Symp Biocomput*, 279-290.

Krakauer, D. C. (2000). Stability and evolution of overlapping genes. *Evolution Int J Org Evolution* 54, 731-739.

Larsen, T. S. & Krogh, A. (2003). EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* 4, 21.

Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25, 955-964.

Makalowska, I., Ryan, J. F. & Baxevanis, A. D. (2001). GeneMachine: gene prediction and sequence annotation. *Bioinformatics* 17, 843-844.

Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R. & Puhler, A. (2003). GenDB--an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res* 31, 2187-2195.

Murakami, K. & Takagi, T. (1998). Gene recognition by combination of several gene-finding programs. *Bioinformatics* 14, 665-675.

Nielson, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* 10, 1-6.

Pavlovic, V., Garg, A. & Kasif, S. (2002). A Bayesian framework for combining gene predictions. *Bioinformatics* 18, 19-27.

Ramakrishna, R. & Srinivasan, R. (1999). Gene identification in bacterial and organellar genomes using GeneScan. *Comput Chem* 23, 165-174.

Rogic, S., Ouellette, B. F. & Mackworth, A. K. (2002). Improving gene recognition accuracy by combining predictions from two gene-finding programs. *Bioinformatics* 18, 1034-1045.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* 16, 944-945.

Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* 26, 544-548.

Schlesner, H., Rathmann, M., Bartels, C., Tindall, B., Gade, D., Rabus, R., Pfeiffer, S. & Hirsch, P. (2004). Taxonomic heterogeneity within the *Planctomycetales* as derived by DNA/DNA-hybridization, description of

Rhodopirellula baltica gen. nov., sp. nov. and transfer of *Pirellula marina* to the genus *Blastopirellula* gen. nov. as *Blastopirellula marina* comb. nov. under revision.

Shmatkov, A. M., Melikyan, A. A., Chernousko, F. L. & Borodovsky, M. (1999). Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes. *Bioinformatics* 15, 874-886.

Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D. & Krogh, A. (2001). On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet* 17, 425-428.

Sonnhammer, E. L., von Heijne, G. & Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* 6, 175-182.

Suzek, B. E., Ermolaeva, M. D., Schreiber, M. & Salzberg, S. L. (2001). A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics* 17, 1123-1130.

Tompa, M. (1999). An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. *Proc Int Conf Intell Syst Mol Biol*, 262-271.

Walker, M., Pavlovic, V. & Kasif, S. (2002). A comparative genomic method for computational identification of prokaryotic translation initiation sites. *Nucleic Acids Res* 30, 3181-3191.

Yada, T., Takagi, T., Totoki, Y., Sakaki, Y. & Takaeda, Y. (2003). DIGIT: a novel gene finding program by combining gene-finders. *Pac Symp Biocomput*, 375-387.

Teil III:
Appendix

A Zusätzliche Genom-Atlanten

Alle nachfolgend gezeigten Genom-Atlanten wurden mit dem Programm GENEWIZ erstellt (Jensen *et al.*, 1999).

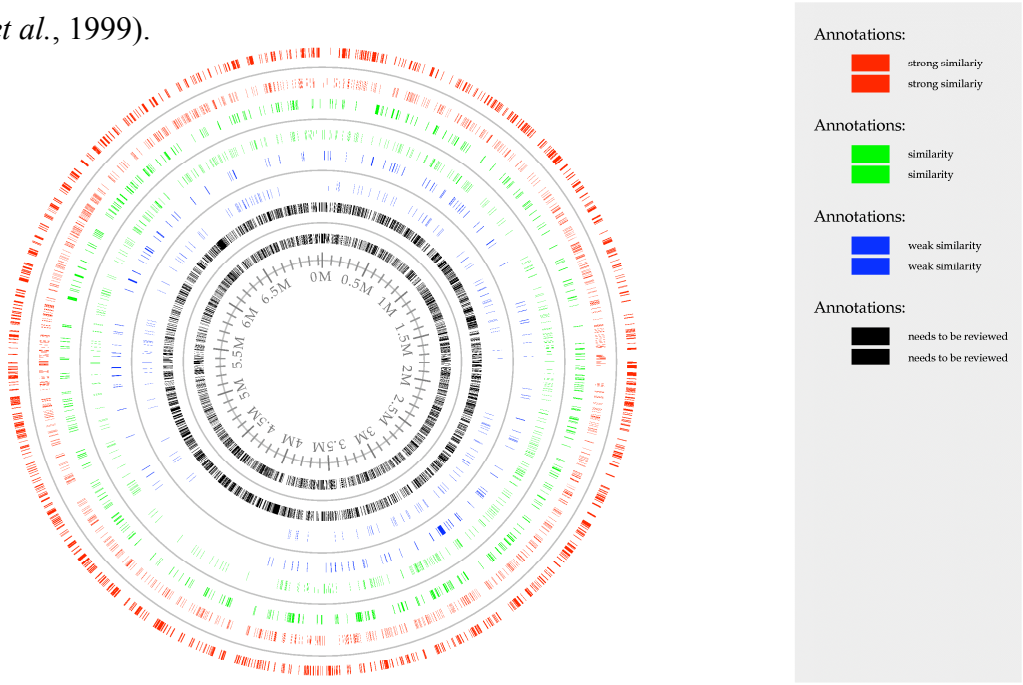


Abb. 15 Atlas aller im *Rhodopirellula baltica*^{T}-Genom annotierten 7325 Gene. Die Farben entsprechen den in der Annotation verwendeten Güteklassen für BLASTp-Treffer (rot: E ≤ E-15; grün: E-15 < E ≤ E-2; blau: E-2 < E ≤ 0,9; schwarz: 0,9 < E). Auf dem kontinuierlich replizierten Strang liegen 51% und auf dem diskontinuierlich replizierten Strang 49% der annotierten Gene. Eine starke ausgeprägte Bevorzugung des kontinuierlich replizierten Strangs für die Lokalisation der Gene, wie sie für viele andere Genome beschrieben wurde, gibt es bei *Rhodopirellula baltica*^{T} folglich nicht.

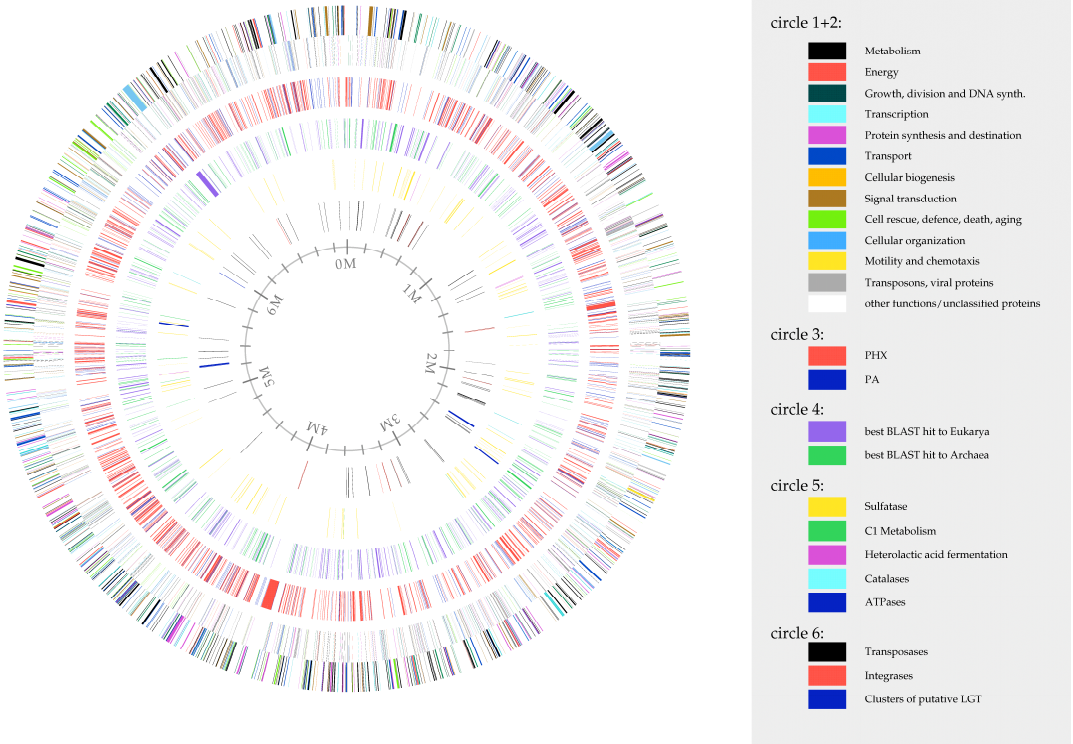


Abb. 16 Atlas einer Auswahl der im *Rhodopirellula baltica*^{T}-Genom annotierten Gene. Dargestellt sind einige der annotierten funktionellen Kategorien (1+2), theoretisch berechnete Expressionslevel (3), Gene mit besten BLASTp-Treffern zu Archaeen und Eukaryonten (4), Gene ausgewählter Stoffwechselwege (5) sowie mobile Elemente und vermeintliche Cluster lateralen Gentransfers (6).

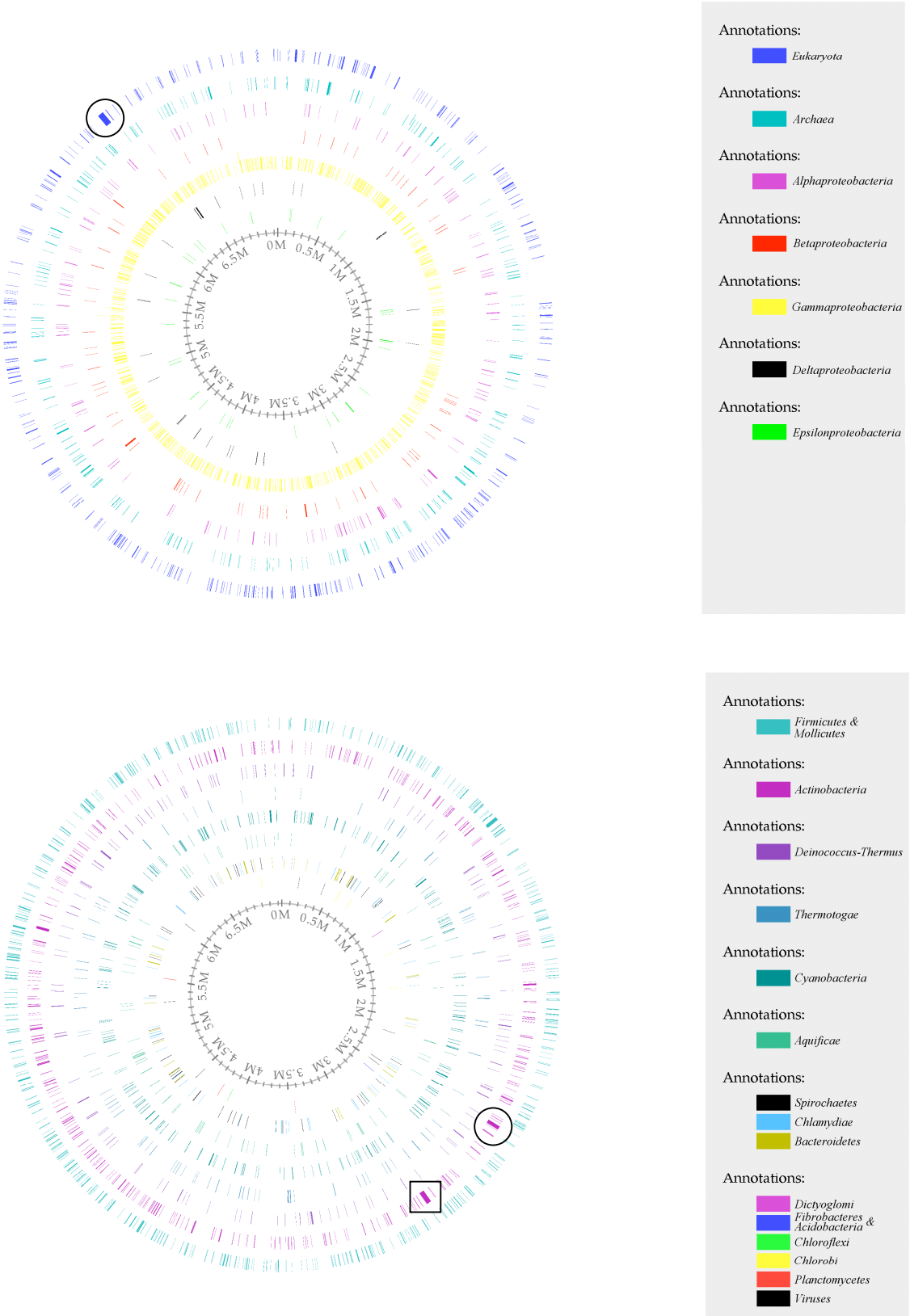


Abb. 17 Atlas der phylogenetischen Zuordnung der besten BLASTp-Treffer. Treffer zu *Eukarya* und *Archaea* sind auf der Stufe ihrer Domänen dargestellt, während die Treffer zur Domäne *Bacteria* nach Phyla aufgeschlüsselt sind. Da die ursprüngliche Annotation keine besten BLASTp-Treffer zu *Mollicutes* enthält, ist dieses Phylum nicht aufgeführt. Zwei Bereiche potentiellen lateralen Gentransfers sind durch Kreise gekennzeichnet. Der durch ein Quadrat gekennzeichnete Bereich geht auf ein sehr langes potentielles Gen von 26.672 Aminosäuren zurück.

B Danksagung

Bedanken möchte ich mich bei Prof. Dr. Rudolf Amann dafür, daß er mir die Gelegenheit bot, in seiner Arbeitsgruppe ein interessantes Thema an der Schnittstelle zwischen Biologie und Informatik zu bearbeiten. Vieles ist nicht immer leicht gewesen, zumal es kaum Erfahrungen gab, auf die man hätte aufbauen können, aber gerade darin lag auch der Reiz des Themas.

Mein besonderer Dank gilt Prof. Dr. Frank Oliver Glöckner. Er hat diese Arbeit fast durchweg in vorbildlicher Weise betreut und sich selbst dann Zeit genommen, wenn er selbst bis zum Hals in Arbeit steckte. Sein unverbesserlicher Optimismus ist mir zwar fremd, aber geholfen hat er ganz sicher. Hervorheben möchte ich seinen unkomplizierten Führungsstil. Ihn zeichnet die Fähigkeit aus, neuen Ideen unvoreingenommen gegenüber zu treten und sich gute Ideen zu eigen zu machen. Das bedeutet auch, daß er sich durch gute Argumente überzeugen läßt und seine Mitarbeiter bei Entscheidungen so gut wie nie übergeht.

Bei Dr. Margarete Bauer möchte ich mich ganz herzlich für die Mitbetreuung dieser Arbeit bedanken, für ihre vielen guten Tips, für die stellenweise sintflutartige Versorgung mit relevanten Publikationen und dafür, daß sie ist, wie sie ist. Aus unseren gelegentlichen Diskussionen weiß ich, daß sie aktuelle Entwicklungen auch außerhalb unseres Arbeitsgebietes oftmals gleich einschätzt wie ich. Sie zieht jedoch mitunter völlig unterschiedliche Schlüsse daraus, was mich dazu bringt, meine eigenen Positionen zu hinterfragen. Auch wenn mir die Verschiedenheit von Menschen manchmal Rätsel aufgibt, so macht diese uns erst zu Menschen und das Leben somit lebenswert.

Außerden möchte ich Prof. Dr. Frank Oliver Glöckner, Dr. Margarete Bauer und Thierry Lombardot von der Arbeitsgruppe Genomik für ihr großes Engagement danken, welches das REGX-Projekt zu einem Erfolg hat werden lassen. Wir sind vier völlig unterschiedliche Menschen und haben es dennoch fast immer verstanden, fruchtbar zusammenzuarbeiten. Dies ist nicht zuletzt der Tatsache zu verdanken, daß wir uns alle in den letzten Jahren positiv verändert haben. Für die Zukunft kann das nur Gutes bedeuten.

Bei den Mitgliedern der gesamten Arbeitsgruppe Molekulare Ökologie möchte ich mich für die gemeinhin gute Stimmung und das angenehme Arbeitsklima bedanken. Aus eigener Erfahrung weiß ich, daß dies keine Selbstverständlichkeit ist. Auch für die vielen Frühstücke sei allen meinen Kollegen an dieser Stelle ausdrücklich gedankt, sowie für die Versorgung mit Koffein in allen bekannten Variationen. Letzteres hat zwar zu Kreislaufproblemen, aber auch maßgeblich zum Gelingen dieser Arbeit beigetragen.

Da moderne Wissenschaft ohne entsprechende finanzielle Mittel kaum möglich ist, sei dem BMBF, der Max-Planck-Gesellschaft sowie dem Max-Planck-Institut für molekulare Genetik in Berlin für ihr finanzielles Engagement im Rahmen des REGX-Projekts gedankt.

Abschließend möchte ich mich noch bei meiner langjährigen Lebensgefährtin Silke Bolte für ihre unendliche Geduld und Liebe bedanken. In den letzten zwölf Jahren hat es das Leben nicht immer gut mit uns gemeint, und manchmal blicke ich verwundert zurück und frage mich, wie wir die ungewöhnlichen Widerigkeiten der Vergangenheit nur meistern konnten. Wenn ich eines ganz sicher weiß, dann daß ich mich auf sie immer verlassen kann. Ohne sie wäre ich nicht der Mensch, der ich heute bin. Ich kann mir keine bessere Partnerin vorstellen.

C Curriculum vitae

Name		Hanno Teeling
Geburtsdatum / -ort		03.12.1967 / Detmold
Nationalität		niederländisch
Familienstand		ledig

Schulbildung	08/74 - 07/78	Grundschule (Oldenburg)
	09/78 - 07/80	Orientierungsstufe (Oldenburg)
	08/80 - 05/87	Gymnasium (Oldenburg)
Schulabschluß	05/87	Abitur (Note: 1,2)

Studium	10/87	Beginn des Studiums der Chemie (Universität Oldenburg)
	04/90	Beginn des Studiums der Biologie (Universität Oldenburg)
	11/91	Vordiplom Chemie (Gesamtnote: „sehr gut“)
	04/93	Vordiplom Biologie (Gesamtnote: „sehr gut“)
	10/95 - 07/96	Diplomarbeit: *Einfluß von organischen Bleiverbindungen auf die Mikrobielle Aktivität* (Note: „sehr gut“)
Studienabschlüsse	07/96	Diplom-Chemiker (Gesamtnote: „sehr gut“)
	02/97	Diplom-Biologe (Gesamtnote: „sehr gut“)

wissenschaftlicher Werdegang	11/96 - 12/96	Wissenschaftlicher Mitarbeiter am *Institut für Chemie und Biologie des Meeres* (ICBM) in Oldenburg - Publikation von Ergebnissen -
	07/97 - 08/97	Wissenschaftlicher Mitarbeiter am *Zentrum für Rehabilitationsforschung* (ZRF) in Bremen - Datenmanagement und Statistik -
	10/97 - 02/98	Wissenschaftlicher Mitarbeiter am ICBM - Molekularbiologische Untersuchungen an grünen Schwefelbakterien -
	04/98 - 06/99	Wissenschaftlicher Mitarbeiter am ZRF - Statistik , Testtheorie, Arzneimittelstudie -
	07/99 - 05/00	unentgeltliche Projektarbeit am ICBM
	06/00 - 08/00	Wissenschaftlicher Mitarbeiter am ICBM - Molekularbiologische Untersuchungen an grünen Schwefelbakterien -
	10/00 - 02/04	Promotion am Max-Planck-Institut für marine Mikrobiologie in Bremen: *Aspekte der bioinformatischen Analyse und Annotation des Genoms von Rhodopirellula baltica*^{T}

Publikationen

Teeling, H., Cypionka, H. (1997) Microbial degradation of tetraethyl lead in soil monitored by microcalorimetry. *Appl Microbiol Biotechnol*, **48**(2): 275-279

Glöckner, F. O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W., Gade, D., Beck, A., Borzym, K., Heitmann, K., Rabus, R., Schlesner, H., Amann, R. & Reinhardt, R. (2003) Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc Natl Acad Scie U S A* **100**(14): 8298-8303

Teeling, H., Lombardot, T., Bauer, M., Ludwig, L. & Glöckner, F. O. (2003) Reevalutation of the phylogenetic position of the *Planctomycetes* by means of concatenated ribosomal protein sequences, DNA-directed RNA polymerase subunit sequences and whole genome trees. *Int J Sys Evol Microbiol*, published online December 5th - in press

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glöckner, F. O. (2004) Application of Tetranucleotide Frequencies for the Assignment of Genomic Fragments. *Environ Microbiol* Special Issue on Metagenomics - in press