Antonie van Leeuwenhoek (2006) 90:69–91
DOI 10.1007/s10482-006-9061-2

Signature proteins that are distinctive characteristics of *Actinobacteria* and their subgroups

Beile Gao, Ragi Paramanathan and Radhey S. Gupta*
*Department of Biochemistry and Biomedical Science, McMaster University, Hamilton, Canada, L8N3Z5; *Author for correspondence (e-mail: gupta@mcmaster.ca; phone: +1-905-525-9140, extn. 22639; fax: +1-905-522-9033)*

Received 28 November 2005; accepted in revised form 20 January 2006

Key words: *Actinobacteria*, Actinobacterial taxonomy, Bacterial phylogeny, Branching order, CMN bacteria, Group-specific proteins, *Magnetospirillum*, Mycobacteria

Abstract

The *Actinobacteria* constitute one of the main phyla of *Bacteria*. Presently, no morphological and very few molecular characteristics are known which can distinguish species of this highly diverse group. In this work, we have analyzed the genomes of four actinobacteria (viz. *Mycobacterium leprae* TN, *Leifsonia xyli* subsp. xyli str. CTCB07, *Bifidobacterium longum* NCC2705 and *Thermobifida fusca* YX) to search for proteins that are unique to *Actinobacteria*. Our analyses have identified 233 actinobacteria-specific proteins, homologues of which are generally not present in any other bacteria. These proteins can be grouped as follows: (i) 29 proteins uniquely present in most sequenced actinobacterial genomes; (ii) 6 proteins present in almost all actinobacteria except *Bifidobacterium longum* and another 37 proteins absent in *B. longum* and few other species; (iii) 11 proteins which are mainly present in *Corynebacterium, Mycobacterium* and *Nocardia* (CMN) subgroup as well as *Streptomyces, T. fusca* and *Frankia* sp., but they are not found in *Bifidobacterium* and *Micrococcineae*; (iv) 8 proteins that are specific for *T. fusca* and *Streptomyces* species, plus 2 proteins also present in the *Frankia* species; (v) 13 proteins that are specific for the *Corynebacterineae* or the CMN group; (vi) 14 proteins only found in *Mycobacterium* and *Nocardia*; (vii) 24 proteins unique to different *Mycobacterium* species; (viii) 8 proteins specific to the *Micrococcineae*; (ix) 85 proteins which are distributed sporadically in actinobacterial species. Additionally, many examples of lateral gene transfer from *Actinobacteria* to *Magnetospirillum magnetotacticum* have also been identified. The identified proteins provide novel molecular means for defining and circumscribing the *Actinobacteria* phylum and a number of subgroups within it. The distribution of these proteins also provides useful information regarding interrelationships among the actinobacterial subgroups. Most of these proteins are of unknown function and studies aimed at understanding their cellular functions should reveal common biochemical and physiological characteristics unique to either all actinobacteria or particular subgroups of them. The identified proteins also provide potential targets for development of drugs that are specific for actinobacteria.

Introduction

Gram-positive bacteria with high G+C DNA content are currently recognized as a distinct phylum, *Actinobacteria*, on the basis of their branching in 16S rRNA trees (Balows et al. 1992; Boone, 2001; Collier et al. 1998; Ludwig and Klenk 2001; Stackebrandt et al. 1997; Stackebrandt and Schumann 2000). This phylum constitutes one of the largest groups among *Bacteria*, comprising of five subclasses and fourteen suborders (Stackebrandt and Schumann 2000; Boone 2001). Actinobacterial species exhibit high level of diversity in terms of their morphology and physiology and play important roles in medicine, industry and environment; some species are major antibiotic producers while many others can cause serious human, animal and plant diseases (Lechevalier and Lechevalier 1967; Goodfellow and Williams 1983; Embley and Stackebrandt 1994; Collier et al. 1998; Stackebrandt and Schumann 2000). However, except for their branching pattern in the 16S rRNA tree, until recently no other biochemical or molecular characteristics were known that could distinguish species of this group from all other bacteria (Embley and Stackebrandt 1994; Stackebrandt and Schumann 2000; Ludwig and Klenk 2001; Gao and Gupta 2005). In our recent work (Gao and Gupta, 2005) we have identified three conserved indels (i.e. inserts and deletions) in widely distributed proteins (viz. a 2 aa deletion in cytochrome c oxidase I, a 4 aa insert in CTP synthetase, and a 5 aa insert in glutamyl-tRNA synthetase), and also confirmed the actinobacterial specificity of a large insert in the 23S rRNA (Roller C et al. 1992), which are distinctive characteristics of the *Actinobacteria* and can be used to circumscribe this phylum. Additionally, a few inserts in variable regions of the RNA polymerase β subunit that might be specific for actinobacteria have also been described (Morse et al. 2002). In phylogenetic trees based on the 16S rRNA gene sequence, actinobacterial species form a compact cluster, beyond which it has proven difficult to resolve the branching order or interrelationships among its different constituent subgroups (Garrity and Holt 2001; Ludwig and Klenk 2001; Stackebrandt et al. 1997; Stackebrandt and Schumann 2000).

The availability of whole genome sequence has opened new windows for discovering novel molecular characteristics that are unique for different groups of bacteria and can be used for their identification as well as for biochemical and functional studies (Karlin et al. 1998; Lerat et al. 2003; Bentley and Parkhill 2004; Fraser et al. 2004; Kainth and Gupta 2005; Mazumder et al. 2005). To date, the genomes of 19 different actinobacterial strains have been completely sequenced and an additional 25 genomes are in progress (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). The complete genomes are from 17 species belonging to 10 genera (some are multiple strains of the same species) and they are as follows: *Bifidobacterium longum*, *Corynebacterium diphtheriae, Corynebacterium efficiens, Corynebacterium glutamicum, Corynebacterium jeikeium, Leifsonia xyli, Mycobacterium avium, Mycobacterium bovis, Mycobacterium leprae*, *Mycobacterium tuberculosis, Nocardia farcinica, Propionibacterium acnes, Streptomyces avermitilis, Streptomyces coelicolor, Symbiobacterium thermophilum,Thermobifida fusca and Tropheryma whipplei* (Domenech et al. 2001; Cole et al. 2001; Schell et al. 2002; Bentley et al. 2002; Fleischmann et al. 2002; Bentley et al. 2003; Cerdeno-Tarraga et al. 2003; Garnier et al. 2003; Kalinowski et al. 2003; Ikeda et al. 2003; Nishio et al. 2003; Raoult et al. 2003; Ishikawa et al. 2004; Bruggemann et al. 2004; Monteiro-Vitorello et al. 2004; Ueda et al. 2004; Tauch et al. 2005). The sequenced genomes differ considerably from each other in various regards (such as genome sizes, numbers of identified proteins or open reading frames (ORF) and GC content (see Table 1; Bentley and Parkhill 2004)) and they provide valuable resources for identifying novel molecular characteristics that are useful for biochemical, taxonomic and evolutionary studies on actinobacteria.

Comparative genomic studies have previously been carried out only on some closely related actinobacterial species. Extensive work has been done on *Mycobacterium* genomes to identify possible virulence factors or new drug targets (Domenech et al. 2001; Cole et al. 2001; Cole 2002). Sutcliffe and Harrington (2004) have analyzed the *M. tuberculosis* genome to identify various genes/proteins that are involved in the synthesis and regulation of cell envelope lipoproteins. Studies have also been done on the *Streptomyces* genomes to identify proteins/enzymes that are possibly involved in production of useful secondary metabolites (Zazopoulos et al. 2003; Ikeda et al. 2003; McAlpine et al. 2005). However, thus far no study has been carried out aimed at identifying different gene/proteins that are uniquely present either in all *Actinobacteria* or in various subgroups that make up this large phylum. Such studies are of much interest in order to understand what unifying molecular characteristics are shared by various actinobacterial species beneath their highly diverse phenotypes.

In our earlier work, we have identified a large number of conserved indels in broadly distributed

Table 1. Actinobacterial species with sequenced genomes.

Strain name	Genome project	Genome size (Mb)	GC content (%)	Protein number
Streptomyces avermitilis MA-4680	Complete	9.12	72.0	7577
Streptomyces coelicolor A3(2)	Complete	9.05	72.1	7769
Nocardia farcinica IFM 10152	Complete	6.29	70.7	5683
Mycobacterium avium subsp. paratuberculosis str. k10	Complete	4.83	69.3	4350
Mycobacterium tuberculosis H37Rv	Complete	4.41	65.6	3991
Mycobacterium tuberculosis CDC1551	Complete	4.4	65.6	4189
Mycobacterium bovis AF2122/97	Complete	4.35	65.6	3920
Thermobifida fusca YX	Complete	3.64	67.5	3110
Symbiobacterium thermophilum IAM 14863	Complete	3.57	68.7	3337
Corynebacterium glutamicum ATCC 13032	Complete	3.31	53.8	2993
Mycobacterium leprae TN	Complete	3.27	57.8	1605
Corynebacterium efficiens YS-314	Complete	3.15	63.1	2950
Leifsonia xyli subsp. xyli str. CTCB07	Complete	2.58	67.7	2030
Propionibacterium acnes KPA171202	Complete	2.56	60.0	2297
Corynebacterium diphtheriae NCTC 13129	Complete	2.49	53.5	2272
Corynebacterium jeikeium K411	Complete	2.46	61.4	2137
Bifidobacterium longum NCC2705	Complete	2.26	60.0	1727
Tropheryma whipplei str. Twist	Complete	0.93	46.0	808
Tropheryma whipplei TW08/27	Complete	0.93	46.3	783
Arthrobacter sp. FB24	Incomplete	–	65.4	–
Brevibacterium linens BL2	Incomplete	4.37*	62.8	–
Frankia sp. CcI3	Incomplete	5.4*	70.1	–
Frankia sp. EAN1pec	Incomplete	–	70.9	–
Kineococcus radiotolerans SRS30216	Incomplete	4.89*	74.2	–
Rubrobacter xylanophilus DSM 9941	Incomplete	3.17*	70.4	–

Note: *indicates that the genome size is estimated. (–) denotes that the information is not available at present due to incomplete sequencing.

proteins that are distinctive characteristics of different groups of bacteria including *Actinobacteria* and which can be used for their identification and characterization (Gupta 1998, 2000, 2004; Gao and Gupta 2005; Griffiths et al. 2005). The objectives of our recent comparative genomic studies are to identify whole proteins or ORFs that are uniquely present in either all species from particular groups (phyla) of bacteria, or in various higher taxonomic groups (e.g. Order, Family, Genus, etc.) among them. By this approach, a large number of proteins that are specific for alpha proteobacteria and *Chlamydiae* have been identified (Kainth and Gupta 2005; Griffiths et al. 2006). In the work presented, we have applied this approach to protein sequences from actinobacterial genomes to identify signature proteins that are unique to *Actinobacteria* or its various subgroups. In addition to their values as molecular and taxonomic markers for the phylum *Actinobacteria*, the study of these unique proteins should also prove instrumental in identifying important physiological characteristics that are distinctive of *Actinobacteria.*

Methods

Identification of Actinobacteria-specific proteins

To identify proteins which are specific for *Actinobacteria* or its various subgroups, all proteins in the genomes of *M. leprae* TN (ML), *L. xyli* subsp. xyli str. CTCB07 (Lxx), *B. longum* NCC2705 (BL) and *T. fusca* YX (Tfu) were analyzed (Cole et al. 2001; Schell et al. 2002; Raoult et al. 2003). BLAST searches were carried out on each individual protein in these genomes to identify all other organisms containing proteins with similar sequences (Karlin and Altschul 1990; Altschul et al. 1997). Protein–protein BLAST was performed with default parameters as set by the BLAST program against sequences from all organisms in the GenBank and the results were visually inspected for homologues showing specificity to *Actinobacteria*. Expected values (E-values) were analyzed as described in our earlier work (Kainth and Gupta 2005; Griffiths et al. 2006) to identify putative *Actinobacteria*-specific proteins. The results of BLAST searches were inspected for sudden increase in E-values from the last actinobacterial species in the search to the first non-actinobacterial organism. This increase in E-values was important when the first non-actinobacterial BLAST hit was in a higher range, such as more than 10^{-5}. Scores above this value suggest that the BLAST matches represent a weak level of similarity that could occur by chance. However, higher E-values are sometimes acceptable for smaller proteins as the magnitude of the E-value depends upon the length of the query sequence (Altschul et al. 1997). A protein was considered to be *Actinobacteria*-specific if all BLAST hits with acceptable E-values corresponded to actinobacterial species. We have retained a few proteins where, besides *Actinobacteria*, 1 or 2 isolated species from other groups of bacteria also had acceptable E-values. We consider these proteins to be also *Actinobacteria*-specific and the presence of a related homologue in isolated other species is very likely due to lateral gene transfer (LGT).

For all *Actinobacteria*-specific signature proteins described here, E-values were recorded for each actinobacterial hit as well as the first non-actinobacterial organism in a given search. The length of each hit protein is also shown in brackets beside the E-values. All proteins indicated in the Tables 2–9 are specific for the *Actinobacteria* based on these criteria unless otherwise mentioned. In the description of these proteins in various Tables, the ‘ML’, ‘Lxx’, ‘BL’ and ‘Tfu’ part of the descriptors indicate the original source of the query protein sequence from *M. leprae* TN, *L. xyli* subsp. xyli str. CTCB07, *B. longum* NCC2705 and *T. fusca* YX genomes, respectively.

Results and discussion

The goal of this study was to identify signature proteins (or ORFs), which are specific for *Actinobacteria* or some of the subgroups from this phylum at different phylogenetic depths. To search for these molecules, comprehensive analysis of four actinobacterial genomes was carried out. Of these, *M. leprae* was chosen because of its small protein numbers (see Table 1). One expects that most proteins that are distinctive characteristics of all actinobacteria should be present in it. Additionally, the analysis of proteins in this genome should also prove useful in identifying proteins that are specific for the suborder *Corynebacterineae* (comprised of *Corynebacterium, Mycobacterium* and *Nocardia*; CMN subgroup). The *L. xyli* genome was chosen because it offered the possibility of identifying proteins that are specific for the suborder *Micrococcineae*, which is an important subgroup within *Actinobacteria*. The *T. fusca* and *B. longum* genomes were chosen because they belong to different suborders branching deeply within the *Actinobacteria* and analyses of proteins that are uniquely shared by these bacteria and other groups could provide useful information regarding interrelationships among various subgroups of *Actinobacteria*. The BLAST searches on each ORF from these four genomes have led to identification of 233 proteins that are unique to *Actinobacteria* and generally do not have homologues in any other bacterial group. We have grouped these 233 signature proteins in nine arbitrary groups based on their distribution pattern. Most of these proteins are of unknown functions. In the few cases where some information regarding their functions is available, it is mentioned in the discussion that follows.

Signature proteins specific for all Actinobacteria

We have identified 29 proteins that are present in nearly all actinobacterial species and are not found in any other *Bacteria* with a few exceptions (see Table 2). In Table 2(a), the first five proteins ML0257, ML0642, ML1009, ML1029, and ML1306 are present in all sequenced actinobacterial genomes including *Rubrobacter xylanophilus* DSM 9941. The observed E-values for these proteins from actinobacterial species are very low, close to 0 (i.e. $< e^{-200}$), indicating that the proteins in various actinobacteria are homologous to the query sequence. In the 16S rRNA tree, *Rubrobacter* species are distantly related to other actinobacterial species and form an outgroup of the other actinobacteria (Stackebrandt et al. 1997; Stackebrandt and Schumann 2000; Ludwig and Klenk 2001; Gao and Gupta 2005). Presently, *Table 2*. Signature proteins specific for *Actinobacteria*.

(a)

Protein	ML0257	ML0642	ML1009	ML1029	ML1306	ML0760[4]	ML0804[4]	ML0857
	[NP_301312]	[NP_301530]	[NP_301746]	[NP_301762]	[NP_301939]	[NP_301589]	[NP301614]	[NP_301645]
Length	167 aa	479 aa	326 aa	273 aa	274 aa	89 aa	84 aa	250 aa
Possible function	Unknown	Unknown	Unknown	Unknown	Unknown	WhiB	WhiB	Unknown
Mycobacterium leprae	1e-91 (167)	0 (479)	5e-177 (326)	8e-153 (273)	1e-147 (274)	6e-47 (89)	8e-45 (84)	3e-97 (250)
Mycobacterium tuberculosis	1e-75 (163)	0 (472)	9e-155 (324)	3e-97 (259)	3e-122 (292)	9e-41 (89)	2e-43 (84)	8e-80 (250)
Mycobacterium avium	3e-73 (162)	0 (459)	9e-152 (323)	2e-98 (257)	5e-134 (300)	5e-39 (121)	1e-43 (84)	2e-80 (250)
Mycobacterium bovis	5e-71 (155)	0 (472)	2e-156 (324)	1e-96 (259)	3e-122 (292)	9e-41 (89)	2e-43 (84)	8e-80 (250)
Nocardia farcinica	2e-57 (175)	4e-141 (470)	4e-109 (310)	1e-51 (299)	6e-109 (294)	8e-34 (92)	5e-38 (84)	4e-56 (247)
Corynebacterium glutamicum	4e-48 (184)	3e-93 (482)	3e-66 (319)	4e-38 (287)	7e-18 (319)	1e-29 (104)	2e-36 (86)	3e-36 (260)
Corynebacterium efficiens	2e-48 (184)	3e-92 (483)	6e-66 (319)	2e-35 (274)	2e-17 (319)	2e-29 (110)	4e-37 (86)	2e-36 (282)
Corynebacterium diphtheriae	2e-46 (188)	2e-89 (463)	9e-62 (313)	9e-35 (239)	4e-15 (313)	2e-29 (99)	2e-36 (86)	2e-36 (257)
Corynebacterium jeikeium	2e-44 (188)	1e-78 (494)	9e-65 (359)	7e-30 (233)	3e-16 (359)	1e-28 (121)	9e-35 (88)	6e-45 (259)
Streptomyces avermitilis	1e-47 (174)	3e-98 (483)	1e-60 (312)	6e-20 (254)	4e-78 (334)	1e-30 (87)	4e-32 (85)	6e-29 (233)
Streptomyces coelicolor	7e-48 (186)	6e-100 (487)	2e-58 (312)	8e-18 (250)	4e-77 (333)	1e-30 (87)	3e-32 (85)	3e-29 (234)
Thermobifida fusca	2e-38 (180)	1e-50 (453)	1e-41 (338)	5e-10 (244)	4e-35 (261)	1e-28 (85)	8e-31 (141)	4e-21 (256)
Propionibacterium acnes	3e-40 (188)	4e-92 (427)	1e-67 (313)	4e-15 (240)	2e-69 (281)	5e-30 (95)	1e-30 (90)	4e-24 (242)
Nocardioides sp.	4e-40 (200)	5e-84 (447)	1e-15 (323)	2e-17 (213)	2e-78 (323)	3e-29 (84)	5e-31 (83)	7e-25 (238)
Frankia sp. CcI3	6e-45 (210)	3e-90 (455)	5e-74 (312)	7e-21 (237)	2e-79 (280)	1e-28 (210)	9e-30 (82)	4e-26 (288)
Frankia sp. EAN1pec	3e-46 (206)	4e-88 (645)	3e-73 (307)	6e-21 (218)	2e-79 (280)	1e-28 (213)	1e-28 (82)	1e-26 (322)
Kineococcus radiotolerance	2e-40 (182)	2e-88 (544)	3e-59 (311)	8e-18 (275)	8e-73 (294)	4e-28 (155)	1e-30 (82)	6e-14 (243)
Brevibacterium linens	4e-37 (176)	8e-63 (481)	2e-11 (265)	8e-15 (287)	9e-30 (265)	3e-29 (103)	1e-28 (82)	5e-08 (243)
Arthrobacter sp.	6e-39 (308)	6e-96 (483)	4e-59 (312)	1e-16 (234)	1e-60 (301)	2e-27 (181)	2e-29 (82)	2e-14 (254)
Leifsonia xyli	8e-36 (170)	4e-72 (447)	9e-54 (309)	2e-13 (167)	1e-50 (298)	5e-27 (88)	4e-26 (82)	2e-16 (232)
Tropheryma whipplei Twist	7e-30 (175)	7e-48 (424)	3e-08 (307)	3e-13 (207)	8e-41 (307)	2e-24 (115)	1e-22 (129)	2e-08 (232)
Tropheryma whipplei TW08/27	7e-30 (162)	2e-48 (424)	3e-08 (296)	3e-13 (191)	8e-41 (296)	2e-24 (102)	1e-22 (77)	2e-08 (232)
Bifidobacterium longum	1e-28 (188)	1e-33 (581)	9e-15 (285)	5e-17 (284)	1e-31 (285)	3e-27 (99)	4e-24 (92)	0.39 (260)
Rubrobacter xylanophilus	1e-08 (145)	4e-07 (340)	1e-24 (299)	0.071 (233)	7e-51 (299)	–*	–*	–*
Non-*Actinobacteria*	See note 1	0.007 (391) *Chloroflexus aurantiacus*	5.0 (863) Human enterovirus	See note 2	See note 3	2.6 (377) *Rattus norvegicus*	7.4 (520) *Rhodopirellula baltica*	2.8 (399) *Novosphingobium aromaticivorans*

Protein	ML0869	ML1016	ML1026	ML2073	ML2137	ML2204	ML0013
	[NP_301656]	[NP_301752]	[NP_301759]	[NP_302382]	[NP_302410]	[NP_302445]	[NP_301140]
Length	124 aa	107 aa	100 aa	231 aa	251 aa	62 aa	93 aa
Possible function	Unknown	Unknown	Unknown	MerR	Unknown	Unknown	Unknown
Mycobacterium leprae	1e-64 (124)	1e-58 (107)	5e-51 (100)	2e-128 (231)	1e-141 (251)	3e-29 (62)	2e-48 (93)
Mycobacterium tuberculosis	3e-50 (236)	3e-36 (82)	8e-49 (100)	7e-109 (225)	5e-109 (253)	3e-18 (60)	6e-45 (93)
Mycobacterium avium	2e-53 (230)	7e-35 (79)	1e-48 (100)	3e-102 (225)	6e-106 (254)	7e-17 (61)	3e-45 (93)
Mycobacterium bovis	3e-50 (236)	3e-36 (82)	8e-49 (100)	7e-109 (225)	5e-109 (253)	3e-18 (60)	6e-45 (93)

Table 2. Continued.

Protein	ML0869 [NP_301656]	ML1016 [NP_301752]	ML1026 [NP_301759]	ML2073 [NP_302382]	ML2137 [NP_302410]	ML2204 [NP_302445]	ML0013 [NP_301140]
Length	124 aa	107 aa	100 aa	231 aa	251 aa	62 aa	93 aa
Possible function	Unknown	Unknown	Unknown	MerR	Unknown	Unknown	Unknown
Nocardia farcinica	2e-29 (231)	1e-24 (81)	2e-41 (100)	2e-74 (185)	8e-59 (324)	5e-16 (68)	5e-24 (87)
Corynebacterium glutamicum	1e-15 (266)	8e-21 (79)	9e-22 (97)	1e-59 (191)	1e-29 (311)	1e-08 (71)	7e-13 (90)
Corynebacterium efficiens	5e-19 (273)	1e-20 (300)	2e-18 (106)	1e-61 (207)	3e-31 (350)	4e-09 (70)	3e-11 (90)
Corynebacterium diphtheriae	9e-22 (224)	6e-21 (79)	5e-22 (97)	2e-58 (186)	3e-30 (349)	7e-07 (68)	3e-12 (89)
Corynebacterium jeikeium	5e-22 (242)	1e-19 (80)	4e-23 (99)	4e-55 (170)	4e-17 (387)	3e-07 (85)	3e-12 (90)
Streptomyces avermitilis	1e-14 (208)	1e-06 (98)	7e-33 (98)	5e-54 (211)	1e-25 (348)	1e-07 (83)	2e-05 (84)
Streptomyces coelicolor	3e-15 (251)	1e-06 (97)	7e-33 (98)	3e-55 (228)	1e-24 (352)	5e-07 (84)	3e-05 (84)
Thermobifida fusca	8e-13 (214)	9e-06 (98)	1e-28 (98)	3e-50 (264)	3e-28 (330)	6e-07 (84)	–
Propionibacterium acnes	0.060 (251)	1e-08 (80)	1e-30 (99)	2e-46 (195)	1e-13 (359)	8e-07 (79)	0.002 (93)
Nocardioides sp.	2e-13 (194)	5e-09 (107)	4e-28 (96)	2e-54 (198)	1e-20 (318)	2e-06 (62)	–*
Frankia sp. CcI3	6e-12 (209)	–*	3e-23 (97)	8e-55 (204)	4e-22 (593)	1e-08 (68)	3e-06 (87)
Frankia sp. EAN1pec	9e-14 (209)	0.003 (74)	1e-22 (97)	1e-54 (208)	4e-22 (755)	2e-05 (73)	2e-05 (88)
Kineococcus radiotolerance	2e-15 (225)	2e-07 (107)	4e-31 (99)	4e-55 (199)	4e-28 (?)	3e-07 (85)	8e-04 (110)
Brevibacterium linens	2e-15 (148)	7e-08 (95)	6e-33 (?)	2e-54 (172)	2e-18 (357)	6e-08 (83)	3e-07 (274)
Arthrobacter sp.	3e-10 (200)	3e-09 (118)	7e-32 (99)	3e-55 (198)	7e-17 (492)	8e-07 (106)	4e-07 (84)
Leifsonia xyli	4e-13 (213)	2e-05 (81)	3e-21 (99)	2e-44 (200)	2e-11 (365)	3e-04 (73)	4e-08 (86)
Tropheryma whipplei Twist	5e-06 (173)	1e-02 (81)	3e-16 (92)	2e-40 (158)	6e-05 (320)	2e-03 (41)	6e-04 (69)
Tropheryma whipplei TW08/27	5e-06 (173)	1e-02 (81)	3e-16 (92)	2e-40 (158)	6e-05 (307)	2e-03 (41)	6e-04 (69)
Bifidobacterium longum	3e-06 (171)	7e-09 (97)	1e-19 (120)	6e-33 (210)	2e-03 (352)	3e-05 (129)	4e-07 (156)
Rubrobacter xylanophilus	–*	–*	–*	–*	–*	–*	–*
Non-*Actinobacteria*	1.9 (221) *Bacillus cereus*	7.4 (230) *Cytophaga hutchinsonii*	0.009 (1306) *Arabidopsis thaliana*	2e-04 (168) *Nostoc* sp. 42/76 (55%)	0.73 (637) *Drosophila melanogaster*	7.4 (664) *Prochloro-coccusmarinus*	1.5 (951) *Dechloromonas aromatica*

(b)

Protein	ML0007 [NP_301135]	ML0580 [NP_301492]	ML0921 [NP_301704]	ML1439 [NP_302017]	ML1610 [NP_302109]	ML2207 [NP_302448]	ML1439 [NP_302017]	ML0256 [NP_301311]	ML0775 [NP_301599]	Proteins showing similar specificity
Length	303 aa	265 aa	96 aa	111 aa	101 aa	131 aa	111 aa	227 aa	589 aa	
Possible function	Unknown	OpcA	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	LpqB	
Mycobacterium leprae	7e-161 (303)	2e-147 (265)	2e-35 (96)	1e-45 (111)	2e-52 (101)	3e-69 (131)	1e-45 (111)	1e-94 (227)	0 (589)	ML0761 [NP_301590]
Mycobacterium tuberculosis	3e-68 (304)	1e-87 (303)	2e-31 (96)	2e-44 (111)	5e-52 (101)	4e-50 (129)	2e-44 (111)	1e-60 (228)	0 (587)	ML0814 [NP_301620]
Mycobacterium avium	8e-72 (283)	2e-80 (303)	8e-31 (96)	6e-44 (111)	7e-51 (101)	9e-43 (131)	6e-44 (111)	2e-56 (225)	0 (585)	ML1649 [NP_302131]
Mycobacterium bovis	3e-68 (304)	1e-87 (303)	2e-31 (96)	2e-44 (111)	5e-52 (101)	4e-50 (129)	2e-44 (111)	1e-60 (228)	0 (583)	ML1666 [NP_302145]
Nocardia farcinica	6e-25 (389)	1e-53 (302)	3e-21 (101)	3e-37 (111)	4e-48 (101)	1e-32 (128)	3e-37 (111)	3e-14 (223)	7e-101 (604)	ML2142 [NP_302413]
Corynebacterium glutamicum	3e-05 (114)	7e-29 (319)	7e-11 (95)	1e-30 (132)	2e-38 (101)	3e-32 (125)	1e-30 (132)	4e-07 (180)	3e-50 (568)	For details, see Supplemental Table 1(a)

Corynebacterium efficiens	3e-05 (114)	3e-29 (321)	2e-10 (130)	3e-27 (120)	1e-38 (101)	2e-33 (142)	3e-27 (120)	3e-06 (150)	7e-49 (563)
Corynebacterium diphtheriae	2e-08 (114)	7e-27 (319)	2e-08 (95)	4e-29 (129)	9e-39 (101)	2e-31 (131)	4e-29 (129)	4e-06 (177)	8e-45 (581)
Corynebacterium jeikeium	2e-12 (217)	4e-31 (358)	2e-11 (96)	1e-31 (125)	3e-41 (101)	5e-18 (126)	1e-31 (125)	1e-08 (236)	1e-60 (583)
Streptomyces avermitilis	4e-10 (204)	8e-28 (311)	5e-12 (94)	2e-19 (124)	4e-39 (102)	6e-21 (265)	2e-19 (124)	5e-08 (164)	2e-11 (610)
Streptomyces coelicolor	2e-10 (185)	3e-27 (351)	5e-12 (94)	2e-19 (124)	4e-39 (102)	6e-21 (202)	2e-19 (124)	3e-09 (174)	2e-07 (615)
Thermobifida fusca	2e-08 (230)	2e-24 (308)	4e-12 (101)	3e-18 (129)	3e-38 (107)	1e-20 (169)	3e-18 (129)	1e-07 (179)	9e-17 (626)
Propionibacterium acnes	1e-11 (210)	1e-19 (310)	3e-05 (96)	7e-13 (110)	5e-32 (103)	5e-10 (280)	7e-13 (110)	5e-06 (217)	5e-10 (591)
Nocardioides sp.	4e-14 (172)	6e-23 (303)	2e-12 (96)	1e-15 (128)	4e-35 (102)	2e-17 (119)	1e-15 (128)	9e-05 (184)	1e-19 (582)
Frankia sp. CcI3	8e-09 (249)	3e-26 (370)	9e-09 (88)	0.001 (198)	7e-38 (98)	9e-19 (270)	0.001 (198)	–*	–*
Frankia sp. EAN1pec	5e-13 (645)	6e-23 (340)	1e-07 (84)	7e-24 (214)	4e-38 (98)	2e-19 (336)	7e-24 (214)	3e-09 (134)	–*
Kineococcus radiotolerance	2e-11 (171)	1e-37 (406)	5e-12 (101)	4e-14 (115)	1e-38 (108)	1e-21 (130)	4e-14 (115)	1e-06 (212)	–*
Brevibacterium linens	9e-10 (153)	–*	4e-11 (99)	1e-10 (113)	9e-33 (105)	1e-22 (133)	1e-10 (113)	–*	1e-10 (562)
Arthrobacter sp.	5e-10 (215)	1e-30 (313)	5e-12 (95)	5e-15 (115)	3e-37 (113)	3e-21 (137)	5e-15 (115)	6e-06 (229)	1e-15 (573)
Leifsonia xyli	1e-05 (137)	5e-30 (320)	2e-10 (98)	6e-09 (123)	2e-30 (107)	9e-22 (118)	6e-09 (123)	3e-04 (177)	8e-18 (557)
Tropheryma whipplei Twist	–	–	–	–	–	–	–	–	–
Tropheryma whipplei TW08/27	–	–	–	–	–	–	–	–	–
Bifidobacterium longum	7e-10 (188)	2e-26 (341)	2e-06 (100)	3e-10 (115)	1e-33 (104)	7e-08 (177)	3e-10 (115)	1e-07 (203)	1e-10 (576)
Non-*Actinobacteria*	0.004 (2528)	0.002 (384)	2.5 (88)	2.2 (265)	0.059 (344)	1.5 (425)	2.2 (265)	0.60 (407)	1.1 (969)
	Magnaporthe grisea	*Chloroflexus aurantiacus*	*Wolbachia endosymbiont*	*Leptospira interrogans*	*Gallus gallus*	*Thermoanaerobacter tengcongensis*	*Leptospira interrogans*	*Arabidopsis thaliana*	*Bacillus anthracis*

These proteins were identified by BLASTP searches as detailed in the Methods section. The top line is the protein ID number in genome of *M. leprae* TN (ML), which was used as probe to perform the blast search. Accession numbers for these proteins are shown in square brackets. The second line and the third line describe the sequence length and possible function of each query protein. The left column lists the actinobacterial strains that have been completely sequenced or draft assembled. The expected (E) values for various actinobacterial species as well as the first non-actinobacterial species in the BLAST results are shown here. The values in brackets after the E-values represent the length of the hit protein. Proteins not found in a genome are indicated with dash (–). * indicates that the genome is incompletely sequenced so it is possible that the protein is present in the genome but not identified at the moment. *Note 1*. The first 3 non-actinobacterial hits to ML0257 correspond to *Thermotoga maritima* MSB8 with E-value of 4e-14 (170 aa) [NP_228884]; *T. neapolitana* with E-value of 1e-12 (150 aa) [CAA07517]; and *Aquifex aeolicus* VF5 with E-value of 6e-04 (147 aa) [NP_214081]. The next non-actinobacterial hit is *Trypanosoma cruzi* with E-value of 0.035 (271 aa). *Note 2*. The first non-actinobacterial hit for ML1029 is found in *M. magnetotacticum* with E-value of 2e-15 (235 aa) [ZP_00049023]; the next non-actinobacterial hit is *Microbulbifer degradans* with E-value of 0.30 (1245 aa). *Note 3*. A low scoring homologue to ML1306 is also found in *Dehalococcoides ethenogenes* with E-value of 8e-09 (276 aa) [YP_181269]; the next non-actinobacterial hit is *Archaeoglobus fulgidus* with E-value of 0.17 (239 aa). *Note 4*. These 2 proteins are paralogous gene products recognized as WhiB. All sequenced actinobacterial species contain several copies of *whiB* gene. Some phages also have homologous gene as observed by their low E-values. These phage proteins include: protein [AAD17616] from Mycobacteriophage TM4 (76 aa); protein [NP_958255] from Bacteriophage VWB (81 aa); and protein [AAN01709] from Mycobacteriophage CJW1 (86 aa).

there are no biochemical or molecular characteristics (other than the 16S rRNA gene sequence analyses) known that support a specific relationship of *Rubrobacter* species to the *Actinobacteria*. In our recent work, a number of conserved indels in 23S rRNA and several proteins (viz. CTP synthetase, CoxI and GluRS) that were uniquely shared by various other actinobacteria, were described (Gao and Gupta 2005). However, these indels were either not present or information for them was lacking for *Rubrobacter* species, thus failing to reveal a specific relationship of this group to *Actinobacteria* (Gao and Gupta 2005). In this context, the shared presence of these five signature proteins in *R. xylanophilus* and various other actinobacteria is of much interest. The simplest and most logical explanation for the shared presence of these five proteins is that the genes for these proteins evolved only once in a common ancestor of *R. xylanophilus* and various other actinobacteria and then passed on to various members of the *Actinobacteria* phylum through vertical descent. This observation, in conjunction with the phylogenetic relationship of *R. xylanophilus* to other *Actinobacteria* in 16S rRNA gene sequence analyses, provides evidence that this species is a part of the phylum *Actinobacteria*.

For three of the proteins described above, 1–2 hits with acceptable E-values are also present in other unrelated bacteria. For example, a single hit with low E-value for ML1029 and ML1306 was also found, respectively, in *Magnetospirillum magnetotacticum* MS-1 (an α-proteobacterium) and *Dehalococcoides ethenogenes* (a green non-sulfur (GNS) bacterium). Because, homologues of these proteins were not present in any other α-proteobacteria or GNS bacteria and both these groups are phylogenetically unrelated to actinobacteria, these exceptions are very likely due to a non-specific event such as lateral gene transfer (LGT). Similarly, for the protein ML0257, homologues with low E-values are also present in two *Thermotoga* species. The phylogenetic position of *Thermotoga* is not reliably known (Gupta 1998; Ludwig and Klenk 2001; Griffiths and Gupta, 2004), thus the possible significance of the shared presence of this protein in these two groups of species is not clear.

The remaining 10 proteins in Table 2(a) are found in almost all sequenced actinobacterial species except *R. xylanophilus*. The genome of *R. xylanophilus* is still not completely sequenced and it is possible that some of these proteins may be found in the *Rubrobacter* genome upon its completion. However, if these proteins are truly absent in *R. xylanophilus*, then based upon its deep branching in the rRNA trees (Stackebrandt et al. 1997; Stackebrandt and Schumann 2000; Gao and Gupta 2005), the most likely explanation for this observation will be that the genes for these proteins have evolved in a common ancestor of *Actinobacteria* after the divergence of *Rubrobacter*. In the tables shown here, we have also included information for *Frankia* sp., *Kineococcus radiotolerans, Nocardioides* sp., *Atrhrobacter* sp. and *Brevibacterium linens*, whose genomes are only draft assemblies. It is possible that for some of the proteins from these species for which sequence information is presently lacking (denoted by asterisks in the tables) this information will become available at a later time.

In Table 2(b), we list 14 additional proteins that show similar distribution as the proteins listed in Table 2(a), but which are missing in the two *T. whipplei* strains. *T. whipplei* is an intracellular pathogen and the genomes of these strains have undergone massive gene decay (to only 0.93 Mb), as many proteins are not required in the intracellular environment (Moran and Wernegreen 2000; Raoult et al. 2003; Bentley et al. 2004). Thus, the absence of these genes in the two *T. whipplei* strains represents a special situation, which is not characteristic of other *Actinobacteria*. Therefore, despite their absence in *T. whipplei*, we still regard these proteins as distinctive characteristics of various other *Actinobacteria*. Note that for the protein ML2204, the E-values for several actinobacterial homologues are higher than our indicated BLAST cut off value (10^{-5}), but these higher E-values are acceptable in this case because of the very short length of this protein (62 amino acids) and the fact that besides *Actinobacteria* no hits for other bacterial species were observed.

Among the *Actinobacteria*-specific proteins listed in Table 2, ML0760 and ML0804 are very similar to each other and they are homologous to the developmental regulator gene *whiB* in *S. coelicolor*. WhiB is a short DNA-binding protein that is essential for sporulation of aerial hyphae in *S. coelicolor* (Soliveri et al. 2000). Our observation that *whiB*-like genes are present in all sequenced actinobacterial genomes including the non-spore-forming intracellular pathogens *T. whipplei* and *L. xyli*, suggests that this protein, in addition to its role in sporulation, also performs a more generalized function common to all *Actinobacteria*. Most actinobacterial species contain multiple copies of the *whiB*-like gene. There are five copies of *whiB* in *M. leprae*, which include ML0639, ML2307 and ML0382, in addition to the two discussed above. The protein lengths of the BLAST hits for the WhiB (ML0760 and ML0804) in some species (viz. *Frankia* sp., *K. radiotolerans* and *T. fusca*) were found to be almost twice the length of the query sequence but their matching regions are highly conserved. It is possible that these species have acquired additional protein domains during the course of evolution. The genes related to *whiB* are also present in some actinophages, which have likely acquired them from actinobacteria (Pedulla et al. 2003).

Of the other *Actinobacteria*-specific proteins with predicted functions, ML2073 (Table 2a) and also ML2075 (in Table 3b) are MerR proteins. MerR is a transcriptional regulator of the mercury resistance genes (Rother et al. 1999). All gram-positive and some gram-negative bacteria are resistant to a broad range of mercuric compounds, and other bacteria besides actinobacteria possess proteins that are annotated as MerR family proteins (Ravel et al. 2000). However, they share very little similarity with actinobacterial MerR sequences as seen by the results of BLAST analyses. Therefore, it is possible that the *merR* gene in *Actinobacteria* has evolved differently from other bacteria and may possess different functional characteristics. The absence of homologues of ML2075 in some actinobacterial species (viz. *B. longum* and *T. fusca*) is likely due to gene loss events.

For all of the 29 *Actinobacteria*-specific proteins identified in the present work (Table 2), no homologues were detected in the *S. thermophilum* genome. *S. thermophilum* is presently placed in the *Actinobacteria* phylum based on its high GC content (Ueda et al. 2001). However, recent genomic analyses indicate that this species is much more closely related to *Bacilli* and *Clostridia* than to *Actinobacteria* (Ueda et al. 2004). In our recent work, this species was also found to be lacking conserved indels in various proteins (viz. Cox1, CTP synthetase and GluRS) as well as the 23S rRNA gene that are distinctive characteristics of most other actinobacteria (Gao and Gupta 2005). These observations strongly indicate that *S. thermophilum* is distinct from all other actinobacteria and it should not be placed in the phylum *Actinobacteria* (Gao and Gupta 2005).

Signature proteins specific for actinobacterial subgroups or providing information regarding their branching order

In phylogenetic trees based on 16S rRNA gene sequences, bifidobacteria generally form a deep branch within the phylum *Actinobacteria* (Stackebrandt et al. 1997; Stackebrandt and Schumann 2000; Ludwig and Klenk 2001; Gao and Gupta 2005). In this study, we have identified six proteins that are found in almost all sequenced actinobacterial species with the exception of *B. longum* (Table 3a). Among these proteins, three are present in all other completely sequenced actinobacterial strains but missing in *Bifidobacterium*, whereas the remaining three are also missing in one isolated species or genus. For example, homologues of the protein ML1781 were also found to be missing in all four *Corynebacterium* species as well as *Bifidobacterium*. The most parsimonious explanation for this observation is that this protein was introduced in an actinobacterial antecedent after the divergence of *Bifidobacterium* and subsequently lost in a common ancestor of *Corynebacterium*. Therefore, these proteins provide us with useful evolutionary information that bifidobacteria very likely constitute one of the earliest branching lineages within *Actinobacteria*, which is consistent with its branching in the 16S rRNA trees (Stackebrandt and Schumann 2000; Ludwig and Klenk 2001; Gao and Gupta 2005).

Besides these six proteins, we have also found 31 additional proteins, which are present in most actinobacterial species but missing in *B. longum* and a few other species (see Table 3b). In a large number of cases, these proteins were absent from the *T. whipplei* and *L. xyli* genomes, which are intracellular pathogens with greatly reduced genomes (Moran and Wernegreen 2000; Raoult et al. 2003). As discussed earlier, the gene loss in these cases represents a special situation and for actinobacterial species that are free-living, these proteins show similar specificity as those listed in Table 3a. Six additional proteins are mainly absent in *B. longum*, *T. whipplei*, *L. xyli* and

Table 3. Signature proteins specific for *Actinobacteria*, except *Bifidobacterium longum*.

(a)

Protein	ML0762 [NP_301591]	ML0876 [NP_301662]	ML1027 [NP_301760]	ML1041 [NP_301768]	ML1176 [NP_301858]	ML1781 [NP_302210]
Length	165 aa	139 aa	157 aa	196 aa	119 aa	170 aa
Possible function	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
Mycobacterium leprae	1e-87 (165)	3e-58 (139)	7e-89 (157)	3e-102 (196)	8e-66 (119)	1e-95 (170)
Mycobacterium tuberculosis	7e-59 (163)	2e-54 (139)	9e-71 (161)	5e-87 (210)	1e-40 (120)	6e-90 (177)
Mycobacterium avium	5e-65 (165)	9e-51 (139)	4e-68 (161)	5e-85 (211)	2e-52 (116)	2e-90 (?)[2]
Mycobacterium bovis	7e-59 (163)	2e-54 (139)	9e-71 (161)	5e-87 (210)	1e-40 (120)	2e-89 (177)
Nocardia farcinica	8e-31 (125)	2e-36 (138)	1e-45 (176)	2e-6 (183)	8e-26 (127)	6e-56 (181)
Corynebacterium glutamicum	2e-20 (130)	2e-14 (143)	5e-26 (192)	5e-48 (208)	8e-15 (118)	–
Corynebacterium efficiens	1e-21 (151)	1e-13 (143)	5e-26 (183)	6e-49 (205)	1e-13 (173)	–
Corynebacterium diphtheriae	4e-19 (137)	1e-12 (143)	9e-26 (170)	1e-46 (193)	5e-12 (118)	–
Corynebacterium jeikeium	4e-13 (172)	2e-16 (143)	7e-25 (161)	2e-42 (207)	4e-11 (121)	–
Streptomyces avermitilis	1e-20 (157)	1e-12 (132)	5e-12 (153)	5e-38 (223)	2e-13 (112)	2e-38 (164)
Streptomyces coelicolor	6e-21 (140)	2e-10 (132)	4e-12 (154)	3e-32 (238)	1e-12 (109)	2e-36 (164)
Thermobifida fusca	4e-16 (148)	3e-16 (141)	–	8e-32 (193)	8e-09 (141)	4e-39 (197)
Propionibacterium acnes	1e-14 (153)	2e-06 (131)	2e-04 (176)	1e-21 (189)	2e-10 (190)	9e-20 (165)
Nocardioides sp.	1e-16 (119)	1e-18 (132)	2e-14 (153)	5e-30 (209)	2e-07 (115)	3e-42 (208)
Frankia sp. CcI3	1e-17 (147)	2e-09 (133)	–*	2e-22 (312)	4e-05 (140)	4e-34 (180)
Frankia sp. EAN1pec	1e-17 (170)	3e-08 (133)	–*	5e-23 (299)	1e-08 (142)	7e-33 (177)
Kineococcus radiotolerance	4e-19 (133)	7e-19 (133)	5e-07 (?)	5e-23 (208)	1e-10 (127)	1e-33 (210)
Brevibacterium linens	2e-12 (124)	3e-10 (133)	1e-08 (151)	3e-24 (207)	3e-08 (145)	2e-23 (119)
Arthrobacter sp.	1e-20 (129)	5e-15 (133)	2e-11 (174)	2e-26 (202)	7e-08 (170)	8e-40 (167)
Leifsonia xyli	1e-15 (70)	3e-07 (110)	2e-13 (148)	4e-27 (198)	9e-09 (147)	1e-33 (167)
Tropheryma whipplei Twist	7e-05 (110)	2e-08 (139)	2e-12 (155)	4e-15 (207)	0.005 (145)	4e-25 (164)
Tropheryma whipplei TW08/27	–	2e-08 (139)	2e-12 (155)	4e-15 (183)	0.005 (145)	4e-25 (160)
Non-*Actinobacteria*	See note 1	–	0.73 (203)	2.2 (623)	0.046 (447)	See note 2
		–	*Cupriavidus necator*	*Shewanella baltica*	*Ralstonia metallidurans*	

(b)

Actinobacterial signature proteins which are not present in *B. longum* and *T. whipplei*	Actinobacterial signature proteins which are not present in *B. longum*, *T. whipplei* and *L. xyli*	Actinobacterial signature proteins mainly lost in *B. longum*, *T. whipplei, L. xyli* and *Corynebacterium* species
ML1485 [NP_302044] oxidoreductase	ML0169 [NP_301248]	ML0115 [NP_301211]
ML2075 [NP_302384] MerR	ML0284 [NP_301324]	ML1299 [NP_301933]
ML0234 [NP_301294] Lsr2	ML0540 [NP_301459] mIHF	Tfu_0030 [YP_288091]
ML0898 [NP_301682]	ML0561 [NP_301475]	Tfu_0751 [YP_288812] acetyltransferase
ML0904 [NP_301687]	ML0589 [NP_301498] ABC-2	Tfu_1240 [YP_289301]
ML0986 [NP_301731]	ML0816 [NP_301622]	Tfu_1340 [YP_289401]
ML1067 [NP_301785]	ML0899 [NP_301683]	
ML1165 [NP_301800] LpqZ	ML1300 [NP_301934]	
ML1165 [NP_301850] Clp	ML1312 [NP_301943]	
ML1166 [NP_301851]	ML1706 [NP_302175]	
ML1927 [NP_302300]	ML2030 [NP_302360] Rpf2	
ML2064 [NP_302376]	ML2428A [NP_302573]	
ML2156 [NP_302419]	ML2442 [NP_302583]	
ML2200 [NP_302442]	ML2446 [NP_302585] lipoprotein	
Lxx03620 [YP_061480]	ML2687 [NP_302714]	
Lxx08190 [YP_061831]	Tfu_0365 [YP_288426]	
Lxx10090 [YP_061981] RDD		
Lxx16410 [YP_062531] Abi		
Tfu_0515 [YP_288576]		
Tfu_2498 [YP_290554]		

Table 3. Continued.

(b): The protein ID number starting with Lxx or Tfu represents query protein from the genome of *L. xyli* subsp. xyli str. CTCB07 (Lxx) or *T. fusca* YX (Tfu). The possible cellular functions of some of these proteins are noted. For other proteins the cellular functions are not known, the E values are provided in the Supplemental Table 1 (b), (c), and (d). *Note 1*. A homologue to ML0762 is also found in *M. magnetotacticum* with E-value of 1e-18 (117 aa) [ZP_00049347]; the next non-actinobacterial hit is *Oryza sativa* with E-value of 0.50 (130 aa). *Note 2*. A homologue to ML1781 is also found in *M. magnetotacticum* MS-1 with E-value of 8e-29 (138 aa) [ZP_00051058]; the next non-actinobacterial hit is *Pan troglodytes* with E-value of 0.92 (1491 aa).

Corynebacterium species (see Table 3b). In these cases, in addition to the gene loss in the intracellular pathogens, an additional gene loss event has occurred in the ancestor of *Corynebacterium* species. We believe that these genes have also most likely evolved in a common ancestor of other actinobacteria after the divergence of bifidobacteria and been subsequently lost in a few groups, due to different reasons. However, the possibility that some of these genes were also lost in *B. longum* cannot be excluded.

Another group of 11 *Actinobacteria*-specific proteins are mainly present in the CMN subgroup, *Streptomyces, Thermobifida*, and *Frankia* but were not found in *B. longum* and species of*Micrococcineae* (*L. xyli, T. whipplei*, *Arthrobacter* sp. FB24 and *B. linens*; Table 4). The shared presence of these proteins in the CMN subgroup and *Streptomyces, Thermobifida*, and *Frankia* species indicates a closer relationship among these groups. The branching order of different subgroups within the phylum *Actinobacteria* is presently not clear. The absence of these proteins in *B. longum* and *Micrococcineae* suggests that these groups have likely evolved prior to the branching of CMN subgroup and *Streptomycineae*.

The BLAST searches on proteins from the *L. xyli* genome have led to identification of 8 proteins (viz. Lxx12820, Lxx05060, Lxx12850, Lxx05560, Lxx08840, Lxx10900, Lxx13550, and Lxx24950; see Supplemental Table 2) that are only present in members of the suborder *Micrococcineae*. Presently, only two *Micrococcineae* species, *L. xyli* and *T. whipplei*, have been completely sequenced, while the genomes of two additional members, *Arthrobacter* sp. FB24 and *B. linens* BL2, are in progress. Five of these proteins (Lxx05560, Lxx24950, Lxx08840, Lxx10900 and Lxx13550) are absent from the genomes of the two *T. whipplei* strains, which is again probably caused by the massive genome shrinkage in these bacteria. However, the presence of these genes in *L. xyli*, which colonizes the xylem vessels of sugarcane (Monteiro-Vitorello et al. 2004), indicates that the cellular environment of this bacterium is quite different from that of *T. whipplei*, with the result that the gene losses in its genome are quite different. This may also explain why, in our analysis, so few proteins that are specific for *Micrococcineae* were identified.

Our BLAST searches with proteins from *T. fusca* genome have revealed eight proteins that are specific to *T. fusca* and two *Streptomyces* species (see first eight proteins in Table 5). *T. fusca* belongs to the suborder *Streptosporangineae*. In some 16S rRNA trees, species from this suborder form a cluster with *Streptomycineae* species, which suggests these two suborders are close relatives (Gao and Gupta 2005). The eight signature proteins that are uniquely present in these two groups of actinobacteria now strongly indicate that species from these two subgroups are closely related and they likely shared a common ancestor exclusive of other actinobacteria. The remaining two proteins in Table 5 (viz. Tfu_2750 and Tfu_2037) are also present in the two *Frankia* strains, in addition to the *Streptomyces* and *Thermobifida*. *Frankiae* are developmentally complex species, which grow by hyphal branching and tip extension and thus resemble the *Streptomyces* spp. (Balows et al. 1992; Benson and Silvester 1993; Collier et al. 1998). Currently, *Frankineae* is recognized as a distinct suborder within the phylum *Actinobacteria* but its phylogenetic relationship to other actinobacterial groups is unclear (Stackebrandt et al. 1997; Boone 2001; Ludwig and Klenk 2001). The two commonly shared proteins are consistent with a closer relationship of *Frankia* to *Streptomyces* and *Thermobifida*.

Signature proteins specific for the CMN subgroup

We have identified 13 proteins which are only found in *Corynebacterium* (C), *Mycobacterium*

Table 4. Signature proteins which are mainly present in CMN subgroup, *Streptomyces, Thermobifida*, and *Frankia* but not found in *Bifidobacterium* and *Micrococcineae*.

Protein	ML0591 [NP_301500]	ML1544 [NP_302075]	ML2435 [NP_302579]	Tfu_2483 [YP_290539]	ML2473 [NP_302601]	ML2570 [NP_302647]	ML2705 [NP_302726]	Proteins showing similar specificity
Length	593 aa	506 aa	277 aa	150 aa	159 aa	1405 aa	259 aa	
Possible function	Unknown	Unknown	Unknown	Unknown	Unknown	Coagulation factor	Unknown	
Mycobacterium leprae	0 (593)	0 (506)	e-158 (277)	–	1e-83 (159)	0 (1405)	2e-136 (259)	ML2199 [NP_302441]
Mycobacterium tuberculosis	0 (591)	0 (506)	e-127 (296)	2e-18 (169)	2e-74 (166)	0 (1400)	3e-95 (244)	ML2289 [NP_302489]
Mycobacterium avium	0 (567)	0 (505)	e-128 (267)	5e-18 (170)	1e-74 (173)	0 (1393)	8e-99 (250)	ML2581 [NP_302650]
Mycobacterium bovis	0 (591)	0 (506)	e-127 (296)	2e-18 (169)	2e-74 (173)	0 (1400)	3e-95 (244)	Tfu_0540 [YP_288601]
Nocardia farcinica	e-118 (538)	3e-78 (495)	2e-94 (286)	6e-13 (169)	5e-56 (179)	0 (1377)	5e-50 (255)	For details, see Supplemental Table 3
Corynebacterium glutamicum	3e-61 (602)	6e-17 (419)	4e-57 (301)	7e-12 (168)	8e-27 (164)	1e-93 (1007)	6e-19 (206)	
Corynebacterium efficiens	1e-73 (590)	3e-14 (451)	6e-52 (324)	2e-14 (168)	2e-29 (161)	3e-96 (1003)	2e-19 (222)	
Corynebacterium diphtheriae	4e-60 (520)	5e-16 (432)	3e-50 (289)	1e-11 (168)	2e-27 (150)	6e-90 (1025)	–	
Corynebacterium jeikeium	2e-69 (579)	1e-14 (431)	1e-62 (318)	1e-13 (168)	3e-31 (175)	4e-112 (1199)	7e-16 (311)	
Streptomyces avermitilis	3e-38 (466)	2e-09 (507)	4e-60 (289)	3e-24 (171)	–	1e-102 (1514)	3e-33 (205)	
Streptomyces coelicolor	–	4e-07 (508)	4e-59 (271)	1e-25 (167)	–	–	1e-33 (205)	
Frankia sp. CcI3	3e-22 (654)	–*	–*	5e-79 (150)	3e-06(163)	1e-108 (1403)	8e-35 (212)	
Frankia sp. EAN1pec	3e-20 (681)	–*	–*	–*	9e-05 (161)	–*	5e-35 (212)	
Thermobifida fusca	–	–	–	2e-12 (118)	4e-17 (166)	2e-137 (1381)	–	
Kineococcus radiotolerans	2e-10 (473)	–*	5e-54 (305)	2e-17 (274)	–*	8e-87 (1322)	2e-35 (244)	
Nocardioides sp.	1e-27 (564)	–*	5e-46 (256)	7e-19 (163)	1e-20 (161)	8e-87 (1383)	8e-25 (224)	
Propionibacterium acnes	1e-26 (526)	–	–	5e-15 (169)	3e-10 (243)	–	–	
Other organisims	0.006 (457) *Ralstonia eutropha*	0.012 (3376) *Strongylocentrotus purpuratus*	1.1 (306) *Homo sapiens*	See note 1	0.04 (389) *Bacteroides thetaiotaomicron*	0.13 (773) *Magnetospirillum magnetotacticum*	0.019 (339) *Escherichia coli*	

Note 1: A homologue to Tfu_2483 is found in *M. magnetotacticum* with E-value of 2e-16 (176 aa) [ZP_00051391]; the next non-actinobacterial hit is *Coxiella burnetii* with E-value of 0.063 (503 aa).

Table 5. Signature proteins specific to *Streptomyces* and *Thermobifida*.

Protein	Tfu_0721 [YP_288782]	Tfu_0828 [YP_288889]	Tfu_0884 [YP_288945]	Tfu_1708 [YP_289766]	Tfu_1938 [YP_289994]
Length	85 aa	355 aa	506 aa	285 aa	282 aa
Possible function	Unknown	Unknown	Unknown	Unknown	Unknown
Thermobifida fusca	8e-42 (85)	1e-160 (355)	0 (506)	4e-170 (285)	4e-109 (282)
Streptomyces avermitilis	3e-12 (109)	4e-30 (412)	1e-68 (519)	5e-72 (288)	4e-14 (311)
Streptomyces coelicolor	3e-11 (78)	7e-31 (355)	1e-70 (497)	3e-66 (281)	3e-13 (317)
Other hit	1.2 (770) *Mycobacterium bovis*	3.0 (330) *Chlorobium limicola*	0.44 (580) *Pseudomonas syringae*	6e-04 (323) *Solibacter usitatus* 47/112 (41%)	3.6 (396) *Burkholderia ambifaria*
Protein	Tfu_2046 [YP_290102]	Tfu_2377 [YP_290433]	Tfu_2886 [YP_290942]	Tfu_2037 [YP_290093]	Tfu_2750 [YP_290806]
Length	308 aa	329 aa	79 aa	119 aa	347 aa
Possible function	Unknown	Unknown	Unknown	Unknown	Unknown
Thermobifida fusca	5e-150 (308)	1e-177 (329)	2e-37 (79)	7e-59 (119)	0 (347)
Streptomyces avermitilis	3e-31 (307)	3e-18 (368)	5e-07 (77)	6e-16 (115)	3e-25 (281)
Streptomyces coelicolor	3e-32 (308)	–	1e-07 (77)	8e-15 (115)	5e-29 (283)
Other hit	0.002 (264) *Nocardia farcinica*	0.71 (594) *Thermus thermophilus*	1.2 (1213) *Mesorhizobium loti*	See note 1	See note 2

Note 1: A homologue to Tfu_2037 is found in *Frankia* sp. EAN1pec with E-value of 6e-08 (112 aa) [ZP_00574167] and *Frankia* sp. CcI3 with E-value of 2e-09 (112 aa) [ZP_00548765]; the next hit is *Caulobacter crescentus* CB15 with E-value of 0.33 (102 aa) [AAK24225]. *Note 2*: A homologue to Tfu_2750 is found in *Frankia* sp. EAN1pec with E-value of 4e-24 (337 aa) [ZP_00570691] and *Frankia* sp. CcI3 with E-value of 9e-26 (276 aa) [ZP_00547325].

(M) and *Nocardia* (N) species, but not found in any other bacteria (Table 6). These bacteria, commonly referred as the CMN group (Balows et al. 1992; Embley and Stackebrandt 1994; Collier et al. 1998), share similar ultrastructure and chemical composition of their cell envelopes, which is composed of a tripartite structure consisting of the ubiquitous cytoplasmic membrane, the cell wall and an outer layer (Daffe and Draper 1998; Brennan 2003). The outer layer is formed by mycolic acids which are covalently linked to the arabinogalactan (Brennan and Nikaido 1995; Sutcliffe 1998; Puech et al. 2001; Sutcliffe and Harrington 2004). Mycolic acids are found only in bacteria belonging to the CMN subgroup and it is a defining feature of this subgroup. The 13 signature proteins listed in Table 6 show high specificity for the three representative genera of this suborder for which sequence information is available and it is likely that they will also be found in other members of this suborder.

Among these proteins, ML0104, ML0105 and ML0106 (encoding for EmbA, EmbB, and EmbC, respectively) are clustered together in the genome, and they play important role in the biosynthesis of the cell envelope (Belanger et al. 1996; Berg et al. 2005). These three ORFs are paralogous genes which are found in all sequenced genomes of *Mycobacterium* spp. and *Nocardia farcinica*, whilst the complete genomes of related *Corynebacterium* species contain only one homologue of the *emb* genes, suggesting that gene duplication has occurred in *Mycobacterium* and *Nocardia* genomes after their divergence from *Corynebacterium*. In *Mycobacterium*, their products are the sites of resistance to the anti-tuberculosis drug ethambutol (EMB). EmbA and EmbB contribute to the synthesis of arabinogalactan, whereas EmbC is involved in the synthesis of lipoarabinomannan (Belanger et al. 1996; Berg et al., 2005). Overall, these 13 signature proteins provide important molecular markers for distinguishing the CMN subgroup from other actinobacteria and functional studies on them should be helpful in the identification of new biochemical properties that are characteristics of this subgroup.

Our analyses have also identified 14 proteins that are shared by *Mycobacterium* and *Nocardia* species but are not found in any other organisms including *Corynebacterium* (Table 7). The existence of this group of proteins suggests that these two genera are more closely related to each other

Table 6. CMN subgroup-specific proteins.

Protein	ML0054 [NP_301167]	ML0096 [NP_301194]	ML0099 [NP_301197]	ML0104[2] [NP_301201]	ML0105[2] [NP_301202]	ML0106[2] [NP_301203]	ML0107 [NP_301204]
Length	481 aa	649 aa	336 aa	1083 aa	1111 aa	1070 aa	632 aa
Possible function	Unknown	Membrane protein	Unknown	EmbB	EmbA	EmbC	Unknown
Mycobacterium leprae	0 (481)	0 (649)	6e-180 (336)	0 (1083)	0 (1111)	0 (1070)	0 (632)
Mycobacterium tuberculosis	0 (480)	0 (641)	3e-135 (336)	0 (1098)	0 (1094)	0 (1094)	0 (643)
Mycobacterium avium	4e-68 (495)	0 (639)	2e-136 (336)	0 (1065)	0 (1108)	0 (1091)	0 (697)
Mycobacterium bovis	0 (480)	0 (627)	3e-135 (336)	0 (1098)	0 (1094)	0 (1094)	0 (643)
Nocardia farcinica	3e-68 (495)	e-136 (647)	4e-75 (325)	0 (1080)	0 (1080)	0 (1081)	e-161 (700)
Corynebacterium glutamicum	5e-18 (419)	5e-75 (686)	3e-53 (309)	0 (1157)	5e-141 (1157)	0 (1157)	1e-82 (677)
Corynebacterium efficiens	4e-17 (451)	2e-80 (676)	2e-50 (306)	0 (1157)	2e-130 (1157)	0 (1157)	1e-86 (703)
Corynebacterium diphtheriae	6e-20 (432)	5e-68 (562)	1e-50 (303)	7e-173 (1141)	3e-118 (1141)	e-169 (1141)	4e-79 (694)
Corynebacterium jeikeium	4e-17 (431)	–	3e-52 (310)	2e-176 (1154)	1e-130 (1154)	0 (1154)	4e-74 (681)
Non-CMN	0.027 (508) *Streptomyces coelicolor*	See note 1	0.003 (227) *Neurospora crassa*	0.27 (670) *Ictalurid herpesvirus*	0.21 (316) *Caenorhabditis briggsae*	1.7 (264) *Anaeromyxobacter dehalogenans*	0.25 (608) *Polaromona* sp.

Protein	ML0281 [NP_301322]	ML0703 [NP_301560]	ML0810 [NP_301617]	ML0990 [NP_301735]	ML1077 [NP_301790]	ML1270 [NP_301915]
Length	229 aa	423 aa	407 aa	209 aa	139 aa	265 aa
Possible function	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
Mycobacterium leprae	3e-97 (229)	0 (423)	3e-180 (407)	3e-95 (209)	4e-75 (139)	2e-146 (265)
Mycobacterium tuberculosis	2e-48 (215)	8e-176 (420)	4e-145 (407)	2e-58 (204)	8e-50 (154)	3e-73 (235)
Mycobacterium avium	1e-44 (231)	2e-167 (437)	7e-131 (408)	7e-51 (202)	2e-47 (133)	5e-48 (143)
Mycobacterium bovis	2e-48 (215)	8e-176 (420)	1e-144 (407)	2e-58 (204)	8e-50 (154)	3e-73 (235)
Nocardia farcinica	1e-24 (224)	1e-108 (414)	7e-47 (409)	4e-31 (248)	4e-20 (159)	9e-20 (240)
Corynebacterium glutamicum	5e-05 (256)	9e-60 (409)	1e-25 (400)	6e-05 (243)	6e-05 (164)	1e-11 (219)
Corynebacterium efficiens	1e-06 (308)	1e-57 (409)	1e-23 (401)	2e-04 (251)	4e-04 (125)	1e-15 (221)
Corynebacterium diphtheriae	6e-04 (243)	7e-60 (410)	4e-15 (420)	1e-06 (212)	6e-12 (137)	8e-10 (191)
Corynebacterium jeikeium	6e-04 (297)	2e-59 (420)	3e-16 (434)	1e-06 (243)	8e-05 (122)	3e-12 (244)
Non-CMN	8.9 (654) *Pan troglodytes*	6.3 (1807) *Trypanosoma brucei*	0.41 (967) *Homo sapiens*	4.3 (271) *Desulfovibrio vulgaris*	1e-04 (574) *Mus musculus*	7e-04 (222) *Pseudomonas putida* 68/138 (49%)

Note 1: Low scoring homologues with E-value of 5e-10 (605 aa) [NP_712164] and 1e-09 (605 aa) [YP_001868] are also present in *Leptospira interrogans* serovar Lai and *L. interrogans* serovar Copenhageni, respectively; the next non-actinobacterial hit is *Oryza sativa* with E-value of 0.15 (880 aa). *Note 2*: ML0104, ML0105 and ML0106 are paralogous proteins and the genomes of *Corynebacterium* species possess only one copy of this gene.

than to *Corynebacterium*. A close relationship between these two CMN genera is also supported by phylogenetic trees based on 16S rRNA gene sequences (Stackebrandt and Schumann 2000; Gao and Gupta 2005). Another group of 24 proteins that we have identified are unique to the *Mycobacterium* species (Table 8). Of the proteins that are specific for either *Mycobacterium*, or *Mycobacterium* and *Nocardia*, five proteins (viz. ML0319 (LpqE), ML0557 (LprG), ML1116 (LprC), ML0676 (LprJ) and ML1966 (LpqH)) are putative lipoproteins (Sutcliffe and Harrington 2004; Sutcliffe and Russell 1995). Two additional putative lipoproteins (ML2592 (Mce1D) and ML2589 (Mce1A); listed in Table 9) are also found, in addition to these bacteria, in the two *Streptomyces* species as well as *Nocardioides* sp. (Table 9; Supplemental Table 3)

Of the proteins listed in Tables 7 and 8, four are clustered together, namely ML1180, ML1181, ML1182 and ML1183. The functions of the former two are unknown, whilst ML1182 belongs to the PPE family and ML1183 belongs to the PE family. Because these four ORFs are tightly clustered, only spaced by 50–60 bp, they probably form an operon and have related functions. We have found a total of six PE-family proteins and five PPE-family proteins, which were either specific to *Mycobacterium* and *Nocardia* or unique to *Mycobacterium* species. PE and PPE protein families are very large and the genomes of *M. tuberculosis* and *M. bovis* contain 99 PE proteins and 67 PPE proteins (Bentley et al. 2004; Cole et al. 1998; Gordon et al. 2001). It is likely that we would have found more mycobacterial specific PE and PPE proteins if the BLAST searches were carried out using another *Mycobacterium* genome rather than that of *M. leprae*, which has a greatly reduced genome (Cole et al. 2001). Both PE and PPE family proteins have a conserved N-terminal domain, but their C-terminal domains vary in size, sequence and repeat copy number. The extensive diversity in the sequence of PE and PPE proteins likely contributes to differences among tubercle strains and to play a role in their virulence by varying their antigenic repertoire (Gordon et al. 2001). Another virulence-associated protein, ML2055, is encoded by the *modD* gene. This protein contains a fibronectin binding motif, which helps mycobacteria in attachment to fibronectin of host cells (Schorey et al. 1995). The other mycobacteria-specific proteins, whose functions are not known at present, are also likely to play important physiological roles that contribute to the characters that distinguish mycobacteria from other bacteria.

Actinobacteria-specific proteins with a sporadic distribution pattern

We have also identified 85 other *Actinobacteria*-specific proteins that show a somewhat sporadic distribution in actinobacterial species (see Table 9). Some of these proteins are present in many actinobacterial genomes, but they are not found in several species. Also, the species that do not contain these proteins are not closely related according to our current understanding of actinobacterial phylogeny. Thus, it is likely that gene losses for these proteins have occurred independently in a number of actinobacterial species. For many other proteins in Table 9, their distribution can be accounted for by groupings such as those shown in Tables 3–7, followed by 1 or 2 gene loss events in particular groups or species of bacteria. To avoid extensive gaps in the main tables, many of these proteins have been included in the Table 9. A large number of proteins in this table are more randomly distributed among a limited number (between 3 and 6) of sequenced actinobacterial species. There are two possible explanations that can account for their sporadic distribution: first, it is possible that some of these genes are the remnants of ancestral sequences that were introduced in the common ancestor of *Actinobacteria* but have been selectively lost in many species because they are not required for growth. It is now known that gene loss provides a selective advantage for pathogenic organisms and this process may contribute to their virulence, particularly as gene loss can play an important role in the adaptation of intracellular organisms to the physiologically stable environments of their host cells (Moran and Wernegreen 2000; Coenye et al. 2005). Alternatively, the sporadic presence of these genes in a number of actinobacterial species can also be, in principle, explained if some of these genes were originally introduced in a particular group or species of *Actinobacteria* and then transferred to other actinobacteria by LGT. Given the specificity of these genes/proteins for *Actinobacteria*, one would have

Table 7. Signature proteins specific for *Mycobacterium* and *Nocardia*.

Protein	ML0071 [NP_301179]	ML0319 [NP_301346]	ML0520 [NP_301445]	ML0557 [NP_301471]	ML0614 [NP_301514]	ML0984 [NP_301729]	ML1115 [NP_301812]
Length	177 aa	183 aa	202 aa	238 aa	95 aa	164 aa	188 aa
Possible function	Unknown	LpqE	Unknown	LprG	Unknown	Unknown	Unknown
Mycobacterium leprae	2e-98 (177)	5e-97 (183)	9e-98 (202)	2e-115 (238)	2e-49 (95)	5e-90 (164)	e-106 (188)
Mycobacterium tuberculosis	8e-95 (177)	1e-56 (182)	3e-56 (230)	6e-76 (236)	8e-18 (67)	4e-33 (149)	5e-83 (185)
Mycobacterium avium	2e-94 (177)	8e-65 (188)	2e-57 (223)	6e-82 (235)	2e-17 (67)	1e-37 (143)	4e-82 (185)
Mycobacterium bovis	8e-95 (177)	1e-56 (182)	3e-56 (230)	6e-76 (236)	8e-18 (67)	4e-33 (149)	4e-83 (185)
Nocardia farcinica	4e-79 (176)	2e-14 (232)	1e-14 (159)	2e-23 (268)	0.007 (85)	2e-11 (139)	2e-15 (177)
Non-*Myco* & *Nocarida*	0.99 (442) *Aspergillus nidulans*	0.020 (189) *Corynebacterium glutamicum*	0.36 (580) *Leishmania major*	See note 1	2.6 (2659) *Pyrobaculum aerophilum*	0.019 (127) *Erwinia carotovora*	1.2 (973) *Anaplasmamarginale*
Protein	ML1116 [NP_301813]	ML1182 [NP_301862]	ML1380 [NP_301981]	ML1991 [NP_302342]	ML2141 [NP_302412]	ML2349 [NP_302528]	ML2463 [NP_302596]
Length	187aa	421 aa	187 aa	468 aa	91 aa	423 aa	264 aa
Possible function	LprC	PPE	Unknown	PPE	Unknown	Unknown	Unknown
Mycobacterium leprae	2e-93 (187)	0 (421)	8e-92 (187)	0 (468)	1e-35 (91)	0 (423)	2e-153 264)
Mycobacterium tuberculosis	3e-77 (180)	5e-70 (396)	3e-74 (187)	e-118 (463)	2e-24 (91)	0 (422)	2e-138 (264)
Mycobacterium avium	1e-80 (189)	2e-42 (395)	2e-71 (186)	9e-55 (493)	1e-14 (91)	–	3e-138 (264)
Mycobacterium bovis	3e-77 (180)	8e-68 (396)	3e-74 (187)	e-118 (463)	2e-24 (91)	0 (422)	2e-138 (264)
Nocardia farcinica	4e-21 (190)	2e-09 (385)	2e-12 (376)	2e-06 (385)	2e-04 (79)	4e-41 (417)	5e-53 (243)
Non-*Myco* &*Nocarida*	7.4 (1144) *Trypanosoma cruzi*	2e-05 (675) *Dissostichus mawsoni*	0.18 (857) *Thiobacillus denitrificans*	1.7 (814) *Mus musculus*	2.5 (447) *Brevibacterium linens*	2e-5 (15281) *Tolypocladium inflatum*	3e-04 (250) *Streptococcus pyogenes* 78/181 (43%)

Note 1: A low scoring homologue to ML0557 is also found in *Nocardioides* sp. JS614 with E-value of 1e-09 (283 aa) [ZP_00659609], and *Chloroflexus aurantiacus* J-10-fl with E-value of 1e-07 (240 aa) [EAO59897]; the next non-actinobacterial hit is *Cytophaga hutchinsonii* with E-value of 0.003 (3828 aa).

Table 8. *Mycobacterium*-specific proteins.

Protein	ML0030 [NP_301154]	ML0051 [NP_301164]	ML0410 [NP_301390]	ML0431 [NP_301401]	ML0538 [NP_301457]	ML0539 [NP_301458]	ML0676 [NP_301547]	ML0748 [NP_301579]
Length	113 aa	302 aa	100 aa	259 aa	102 aa	538 aa	158 aa	92 aa
Possible function	Unknown	PPE	PE	Unknown	PE	PPE	LprJ	Unknown
Mycobacterium leprae	1e-19 (113)	1e-160 (302)	3e-50 (100)	e-103 (259)	6e-52 (102)	0 (538)	6e-67 (158)	4e-34 (92)
Mycobacterium avium	9e-08 (113)	1e-05 (301)	2e-06 (102)	3e-52 (253)	2e-20 (102)	5e-66 (538)	1e-26 (203)	3e-22 (93)
Mycobacterium bovis	2e-05 (115)	2e-43 (368)	1e-10 (98)	2e-57 (273)	2e-32 (102)	e-133 (539)	5e-19 (129)	5e-22 (93)
Mycobacterium tuberculosis	3e-05 (115)	2e-43 (368)	1e-10 (98)	2e-57 (273)	2e-32 (102)	e-133 (539)	2e-19 (129)	5e-22 (93)
Non-*Mycobacterium*	–	0.60 (1194)	1.5 (107)	–	0.39 (488)	–	8.0 (1220)	–
	–	*Yarrowia lipolytica*	*Nocardia farcinica*	–	*Homo sapiens*	–	*Ustilago maydis*	–
Protein	ML0813 [NP_301619]	ML0878 [NP_301664]	ML1180 [NP_301860]	ML1181 [NP_301861]	ML1183 [NP_301863]	ML1232 [NP_301893]	ML1357 [NP_301967]	ML1607 [NP_302108]
Length	195 aa	212 aa	95 aa	100 aa	99 aa	358 aa	61 aa	96 aa
Possible function	Unknown	Unknown	Unknown	Unknown	PE	PE	Unknown	Unknown
Mycobacterium leprae	2e-75 (195)	8e-116 (212)	4e-41 (95)	6e-52 (100)	6e-43 (99)	0 (358)	2e-25 (61)	4e-51 (96)
Mycobacterium avium	5e-34 (191)	2e-64 (210)	1e-25 (?)[1]	7e-27 (98)	3e-12 (99)	9e-75 (376)	5e-14 (61)	3e-13 (98)
Mycobacterium bovis	3e-39 (186)	2e-67 (214)	6e-23 (94)	7e-25 (98)	1e-13 (99)	e-115 (359)	2e-14 (58)	8e-20 (128)
Mycobacterium tuberculosis	2e-39 (187)	1e-67 (214)	6e-23 (94)	1e-25 (99)	1e-13 (98)	e-115 (359)	2e-14 (58)	8e-20 (128)
Non-*Mycobacterium*	–	0.047 (504)	1.7 (95)	0.027 (105)	–	0.77 (424)	0.67 (88)	0.001 (122)
	–	*Streptomyces coelicolor*	*Corynebacterium diphtheriae*	*Corynebacterium diphtheriae*	–	*Bacillus clausii*	*Nocardia farcinica*	*Kineococcus radiotolerans*
Protein	ML1828 [NP_302239]	ML1835 [NP_302244]	ML1966 [NP_302330]	ML2055 [NP_302372]	ML2532 [NP_302627]	ML2534 [NP_302628]	ML2596 [NP_302663]	ML2616 [NP_302675]
Length	572 aa	227 aa	161 aa	287 aa	98 aa	102 aa	325 aa	170 aa
Possible function	PPE	Unknown	LpqH	modD	PE	PE	Unknown	Unknown
Mycobacterium leprae	0 (572)	1e-104 (227)	2e-86 (161)	e-128 (287)	1e-81 (98)	1e-83 (102)	2e-176 (325)	3e-85 (170)
Mycobacterium avium	e-101 (585)	2e-78 (221)	4e-26 (161)	7e-74 (368)	5e-51 (97)	8e-48 (102)	6e-106 (323)	–
Mycobacterium bovis	e-106 (556)	5e-81 (242)	5e-24 (159)	2e-84 (325)	6e-53 (97)	5e-47 (102)	1e-88 (322)	2e-46 (167)
Mycobacterium tuberculosis	e-106 (556)	5e-81 (242)	5e-24 (159)	2e-84 (325)	6e-53 (97)	5e-47 (102)	7e-88 (322)	2e-46 (167)
Non-*Mycobacterium*	2e-05 (385)	2e-04 (90)	3e-04 (1367)	0.92 (416)	0.72 (429)	0.23 (371)	0.23 (283)	0.10 (244)
	Nocardia farcinica 76/172 (44%)	*Arthrobacter* sp.	*Saccharomyces cerevisiae*	*Plasmodium yoelii*	*Chromohalobacter salexigens*	*Geobacillus kaustophilus*	*Magnetospirillum magnetotacticum*	*Syntrophobacter fumaroxidans*

Table 9. Actinobacteria-specific proteins with sporadic distribution.

Gene ID, accession number and possible function			
ML0271 [NP_301317]	Lxx04780 [YP_061569]	Lxx21720 [YP_062966]	Tfu_1028 [YP_289089]
ML0889 [NP_301674]	Lxx05200 [YP_061603]	Lxx22880 [YP_063058]	Tfu_1067 [YP_289128]
ML1526 [NP_302067]	Lxx05320 [YP_061610]	Lxx23490 [YP_063102] oxidoreductase	Tfu_1088 [YP_289149]
ML1593 [NP_302072]	Lxx06110 [YP_061675]	Lxx24290 [YP_063167]	Tfu_1137 [YP_289198]
ML1704 [NP_302173]	Lxx06130 [YP_061677]	Lxx24410 [YP_063172]	Tfu_1203 [YP_289264]
ML2070 [NP_302380]	Lxx06210 [YP_061684]	Tfu_0012 [YP_288075]	Tfu_1264 [YP_289325]
ML2143 [NP_302414]	Lxx06980 [YP_061735]	Tfu_0015 [YP_288078]	Tfu_1426 [YP_289487]
ML2199 [NP_302441]	Lxx07270 [YP_061760]	Tfu_0332 [YP_288393]	Tfu_1606 [YP_289664]
ML2289 [NP_302489]	Lxx07570 [YP_061782]	Tfu_0342 [YP_288403]	Tfu_1754 [YP_289812]
ML2581 [NP_302650]	Lxx08430 [YP_061848]	Tfu_0355 [YP_288416]	Tfu_1957 [YP_290013]
ML2589 [NP_302656] mce1A	Lxx08745 [YP_061874] ABC transporter	Tfu_0458 [YP_288519]	Tfu_2111 [YP_290167]
ML2592 [NP_302659] mce1D	Lxx09730 [YP_061949] secreted protein	Tfu_0510 [YP_288571]	Tfu_2127 [YP_290183]
ML2600 [NP_302666]	Lxx10335 [YP_062003]	Tfu_0540 [YP_288601]	Tfu_2164 [YP_290220]
ML2689 [NP_302716]	Lxx10420 [YP_062012]	Tfu_0565 [YP_288626]	Tfu_2237 [YP_290293]
BL0571 [NP_695759]	Lxx11560 [YP_062109] electron transport	Tfu_0596 [YP_288657]	Tfu_2238 [YP_290294]
BL0679 [NP_695864]	Lxx11715 [YP_062125]	Tfu_0741 [YP_288802]	Tfu_2265 [YP_290321]
BL0895 [NP_696072]	Lxx12500 [YP_062199]	Tfu_0860 [YP_288921]	Tfu_2382 [YP_290438]
BL1007 [NP_696179]	Lxx18330 [YP_062679]	Tfu_0870 [YP_288931]	Tfu_2579 [YP_290635]
BL1224 [NP_696395]	Lxx18480 [YP_062690]	Tfu_0889 [YP_288950]	Tfu_2706 [YP_290762]
BL1333 [NP_696497]	Lxx19690 [YP_062790]	Tfu_0967 [YP_289028]	Tfu_2899 [YP_290955]
BL1479 [NP_696638]	Lxx20480 [YP_062866]	Tfu_0998 [YP_289059]	Tfu_3004 [YP_291060]
BL1484 [NP_696643]			

Note: The possible cellular functions of some proteins are noted. The other proteins are of unknown function. The E values for these proteins from BLAST searches are provided in Supplemental Table 3.

to postulate that the LGT in these cases is highly selective and limited to only within *Actinobacteria*.

Gene transfer from Actinobacteria to Magnetospirillum magnetotacticum

One interesting and surprising observation from the present work is that for a number of proteins that are *Actinobacteria*-specific, homologous proteins (as indicated by their low E-values and similar protein lengths) are also present in the genome of *M. magnetotacticum* MS-1. *M. magnetotacticum* is a magenetotactic bacteria belonging to the α-proteobacteria subdivision (Bazylinski and Frankel 2004; Gupta 2005; Kainth and Gupta 2005). It forms internal crystals of magnetite in membrane enclosed bodies which it uses to swim along geomagnetic field lines (Bazylinski and Frankel 2004). In the present work, we have identified a total of 14 proteins (viz. ML1029, ML1666, ML0761, ML0762, ML1781, Lxx08190, Tfu_1340, Tfu_2483, BL0895, Lxx08745, ML1526, Tfu_2164, BL1224 and Lxx11715) for which a related homologue is found in *M. magnetotacticum*. Most of these genes/proteins from *M. magnetotacticum* exhibit highest similarity to the corresponding genes/proteins from *Streptomyces* species. When BLAST searches were carried out on these proteins from *M. magnetotacticum*, all of the hits with highest similarity were from actinobacterial species and no proteobacterial hits with low E-values were observed (results not shown). In view of the fact that besides *M. magnetotacticum*, no other α-proteobacterial species was found to contain any of these proteins, it is very likely that these genes in *M. magnetotacticum* have been acquired from actinobacterial species by means of LGT. The genome project of *M. magnetotacticum* is still in progress (DOE Joint Genome Institute; http://www.genome.jgi-psf.org/draft_microbes/magma/magma.home.html), but it is known to have a very large genome (ca. 9.2 Mb) with very high GC content (66.4%), similar to those of *Actinobacteria*. The lateral transfer of these genes to *M. magnetotacticum* seems to have occurred in a highly specific manner as, other than *M. magnetotacticum*, very few and only isolated examples of the presence of these gene/proteins in other groups of bacteria were observed. These results

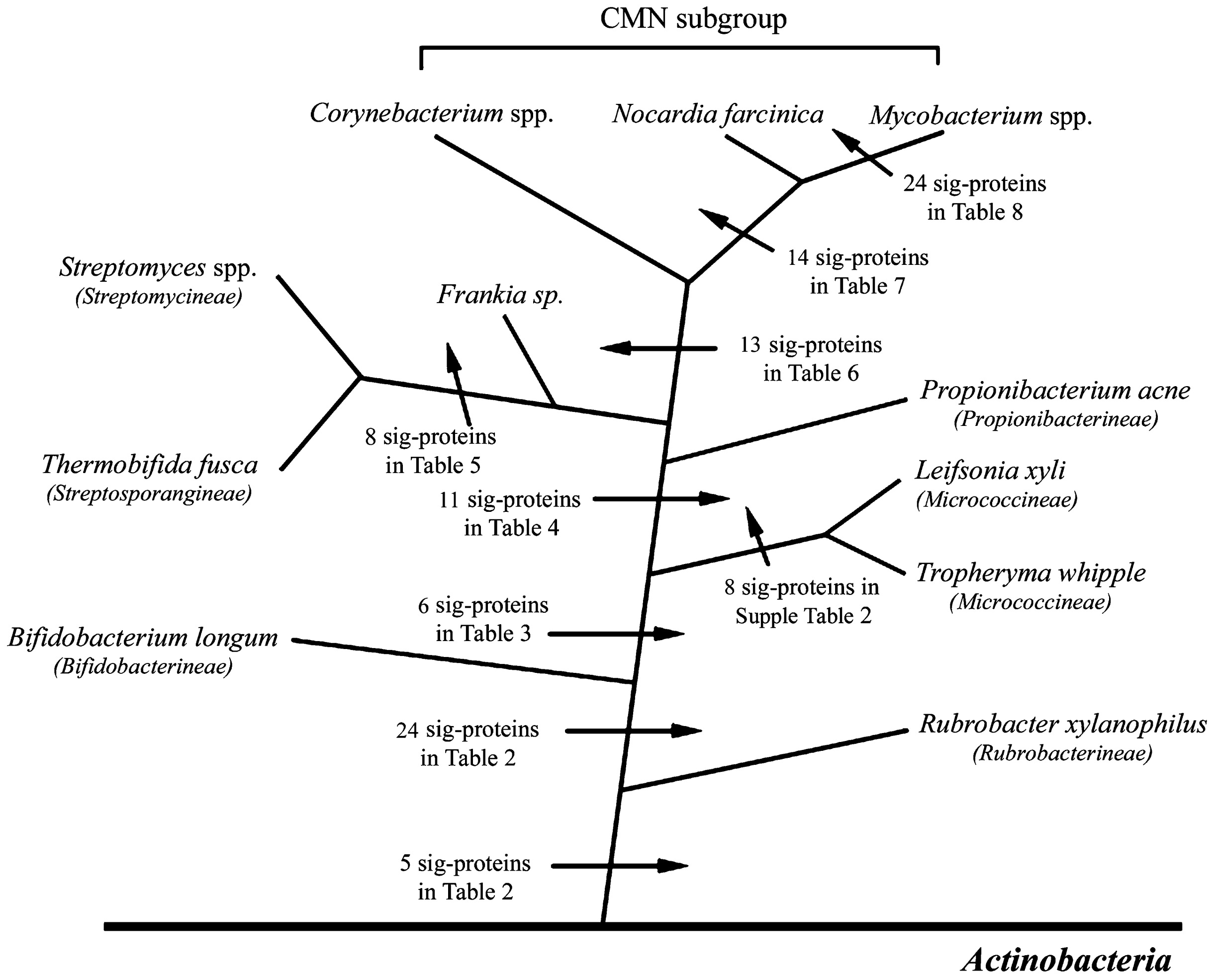
CMN subgroup
Corynebacterium spp.
Nocardia farcinica
Mycobacterium spp.
24 sig-proteins in Table 8
14 sig-proteins in Table 7
Streptomyces spp. (Streptomycineae)
Frankia sp.
13 sig-proteins in Table 6
Propionibacterium acne (Propionibacterineae)
8 sig-proteins in Table 5
Thermobifida fusca (Streptosporangineae)
11 sig-proteins in Table 4
Leifsonia xyli (Micrococcineae)
Tropheryma whipple (Micrococcineae)
8 sig-proteins in Supple Table 2
6 sig-proteins in Table 3
Bifidobacterium longum (Bifidobacterineae)
24 sig-proteins in Table 2
Rubrobacter xylanophilus (Rubrobacterineae)
5 sig-proteins in Table 2
Actinobacteria

Figure 1. Summary diagram showing the distribution patterns of various *Actinobacteria*-specific proteins. The arrows indicate the evolutionary stages where these signature proteins were likely introduced.

provide evidence against the widespread lateral transfer (Gogarten and Townsend 2005) of the genes for *Actinobacteria*-specific proteins to other bacteria. The possible functional significance of the genes, which have been apparently laterally transferred from *Actinobacteria* to *M. magnetotacticum* remains to be determined.

Conclusions

Our comparative analyses of actinobacterial genomes have identified 233 signature proteins that are uniquely found in *Actinobacteria*. Some of these proteins are present in all sequenced actinobacterial genomes, whereas others are limited to various subgroups of *Actinobacteria* at different phylogenetic depths. In addition to providing novel molecular markers that are distinctive characteristics of the entire phylum *Actinobacteria*, based on these proteins, a number of major subgroups within this phylum (viz. *Micrococcineae*, CMN subgroup, *Streptosporangineae* and *Streptomycineae*) can now be delineated. Within the CMN subgroup, a large number of proteins that are unique to either *Mycobacterium* and *Nocardia* species, or only the *Mycobacterium* species have been identified. The absence of all of these proteins in *S. thermophilum* indicates that this species should not be grouped with *Actinobacteria*, an inference which is also supported by other lines of evidences (Ueda et al. 2004; Gao and Gupta 2005).

In addition to these signature proteins, we have also identified a large number of conserved indels that are distinctive of the above groups or subgroups of *Actinobacteria* (Gao and Gupta 2005, and unpublished results).

The distribution pattern of these *Actinobacteria*-specific proteins provides valuable information regarding the relative branching order and interrelationships among various subgroups that comprise the *Actinobacteria* phylum. Based upon their distribution pattern, a tentative model concerning the branching order among a number of subgroups within this phylum, and the evolutionary stages where many of these proteins have been introduced, can be proposed (Figure 1). Our analyses suggest that *Rubrobacterales* constitute one of the deepest branches within the phylum *Actinobacteria* and this is followed by the emergence of *Bifidobacteriales* and *Micrococcineae*. Species belonging to the suborders *Streptomycineae*, *Streptosporangineae, Frankineae* and *Corynebacterineae* (CMN subgroup) are indicated as late branching groups within *Actinobacteria* and within them a closer relationship is generally observed among the *Streptomycineae*, *Streptosporangineae* and *Frankineae* suborders. The deduced relationships are generally in accordance with the 16S rRNA trees (Stackebrandt and Schumann 2000; Ludwig and Klenk 2001; Gao and Gupta 2005).

Most of the actinobacteria-specific proteins identified in the present work are of unknown function. The GC contents of these proteins are very similar to the rest of their genomes and their Ka/Ks ratios (i.e., substitution rates at non-synonymous versus synonymous sites) are less than 0.1 (results not shown). These results strongly indicate that the identified ORFs very likely correspond to functional proteins and they are not due to errors in gene annotation (Daubin and Ochman 2004; Yang 2005). Because of the specificity of these proteins for either all *Actinobacteria* or certain subgroups within this phylum, it is highly likely that these proteins carry out certain unique functions that are limited to these groups of bacteria. Therefore, studies aimed at understanding the functions of these *Actinobacteria*-specific proteins should be of great interest, as they will likely provide important insights into unique biochemical and physiological characteristics that distinguish these bacteria (or specific subgroups among them) from all other bacteria. Because of their specificity for *Actinobacteria* or certain groups within this phylum, many of which are important human pathogens (e.g. *M. leprae*, *M. tuberculosis* and *N. farcinica*), these proteins potentially also provide novel targets for development of drugs that are specifically directed against these bacteria.

Acknowledgments

This work was supported by research grants from the National Science and Engineering Research Council of Canada and the Canadian Institute of Health Research. We thank the editor and two anonymous reviewers for various helpful suggestions toward improvement of the manuscript.

Electronic supplementary material

Supplementary material is available for this article at http://www.dx.doi.org/10.1007/s10482-006-9061-2 and is accessible for authorized users.

References

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J.H., Zhang Z., Miller W. and Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Balows A., Trüper H.G., Dworkin M., Harder W. and Schleifer K.H. 1992. The Prokaryotes. Springer-Verlag, New York.

Bazylinski D.A. and Frankel R.B. 2004. Magnetosome formation in prokaryotes. Nat. Rev. Microbiol. 2: 217–230.

Belanger A.E., Besra G.S., Ford M.E., Mikusova K., Belisle J.T., Brennan P.J. and Inamine J.M. 1996. The embAB genes of Mycobacterium avium encode an arabinosyl transferase involved in cell wall arabinan biosynthesis that is the target for the antimycobacterial drug ethambutol. Proc Natl Acad Sci USA 93: 11919–11924.

Benson D.R. and Silvester W.B. 1993. Biology of Frankia Strains, Actinomycete Symbionts of Actinorhizal Plants. Microbiol. Rev. 57: 293–319.

Bentley S.D., Brosch R., Gordon S.V., Hopwood D.A. and Cole S.T. 2004. Genomics of Actinobacteria, the high G + C Gram-positive bacteria. In: Fraser C.M., Read T.D. and Nelson K.E. (eds.), Microbial Genomes, Humana Press, Totowa, NJ, pp. 333–359.

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., Bateman A., Brown S., Chandra G., Chen C.W., Collins M., Cronin A., Fraser A., Goble A., Hidalgo J., Hornsby T., Howarth S., Huang C.H., Kieser

T.Larke L., Murphy L., Oliver K., O'Neil S., Rabbinowitsch E., Rajandream M.A., Rutherford K., Rutter S., Seeger K., Saunders D., Sharp S., Squares R., Squares S., Taylor K., Warren T., Wietzorrek A., Woodward J., Barrell B.G., Parkhill J. and Hopwood D.A. 2002. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). Nature 417: 141–147.

Bentley S.D., Maiwald M., Murphy L.D., Pallen M.J., Yeats C.A., Dover L.G., Norbertczak H.T., Besra G.S., Quail M.A., Harris D.E., von Herbay A., Goble A., Rutter S., Squares R., Squares S., Barrell B.G., Parkhill J. and Relman D.A. 2003. Sequencing and analysis of the genome of the Whipple's disease bacterium Tropheryma whipplei. Lancet 361: 637–644.

Bentley S.D. and Parkhill J. 2004. Comparative genomic structure of prokaryotes. Annu. Rev. Genet. 38: 771–792.

Berg S., Starbuck J., Torrelles J.B., Vissa V.D., Crick D.C., Chatterjee D. and Brennan P.J. 2005. Roles of conserved proline and glycosyltransferase motifs of embC in biosynthesis of lipoarabinomannan. J. Biol. Chem. 280: 5651–5663.

Boone D.R. 2001. Bergey's Manual of systematic bacteriology, Springer.

Brennan P.J. 2003. Structure, function, and biogenesis of the cell wall of Mycobacterium tuberculosis. Tuberculosis 83: 91–97.

Brennan P.J. and Nikaido H. 1995. The envelope of mycobacteria. Annu. Rev. Biochem. 64: 29–63.

Bruggemann H., Henne A., Hoster F., Liesegang H., Wiezer A., Strittmatter A., Hujer S., Durre P. and Gottschalk G. 2004. The complete genome sequence of Propionibacterium acnes, a commensal of human skin. Science 305: 671–673.

Cerdeno-Tarraga A.M., Efstratiou A., Dover L.G., Holden M.T.G., Pallen M., Bentley S.D., Besra G.S., Churcher C., James K.D., De Zoysa A., Chillingworth T., Cronin A., Dowd L., Feltwell T., Hamlin N., Holroyd S., Jagels K., Moule S., Quail M.A., Rabbinowitsch E., Rutherford K.M., Thomson N.R., Unwin L., Whitehead S., Barrell B.G. and Parkhill J. 2003. The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129. Nucleic Acids Res. 31: 6516–6523.

Coenye T., Gevers D., de Peer Y.V., Vandamme P. and Swings J. 2005. Towards a prokaryotic genomic taxonomy. FEMS Microbiol. Rev. 29: 147–167.

Cole S.T. 2002. Comparative and functional genomics of the Mycobacterium tuberculosis complex. Microbiology 148: 2919–2928.

Cole S.T., Brosch R., Parkhill J., Garnier T., Churcher C., Harris D., Gordon S.V., Eiglmeier K., Gas S., Barry C.E., Tekaia F., Badcock K., Basham D., Brown D., Chillingworth T., Conner R., Davies R., Devlin K., Feltwell T., Gentles S., Hamlin N., Holroyd S., Hornsby T., Jagels K., Krogh A., McLean J., Moule S., Murphy L., Oliver K., Osborne J., Quail M.A., Rajandream M.A., Rogers J., Rutter S., Seeger K., Skelton J., Squares R., Squares S., Sulston J.E., Taylor K., Whitehead S. and Barrell B.G. 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence (vol 393, pg 537, 1998). Nature 396: 190–198.

Cole S.T., Eiglmeier K., Parkhill J., James K.D., Thomson N.R., Wheeler P.R., Honore N., Garnier T., Churcher C., Harris D., Mungall K., Basham D., Brown D., Chillingworth T., Connor R., Davies R.M., Devlin K., Duthoy S.Feltwell T., Fraser A., Hamlin N., Holroyd S., Hornsby T., Jagels K., Lacroix C., Maclean J., Moule S., Murphy L., Oliver K., Quail M.A., Rajandream M.A., Rutherford K.M., Rutter S., Seeger K., Simon S., Simmonds M., Skelton J., Squares R., Squares S., Stevens K., Taylor K., Whitehead S., Woodward J.R. and Barrell B.G. 2001. Massive gene decay in the leprosy bacillus. Nature 409: 1007–1011.

Collier L., Balows A. and Sussman M. 1998. Topley & Wilson's Microbiology and Microbial Infections, Vol. 2, Systematic Bacteriology. Arnold, London.

Daffe M. and Draper P. 1998. The envelope layers of mycobacteria with reference to their pathogenicity. Adv. Microb. Physiol. 39: 131–203.

Daubin V. and Ochman H. 2004. Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. Genome Res. 14: 1036–1042.

Domenech P., Barry C.E. and Cole S.T. 2001. Mycobacterium tuberculosis in the post-genomic age. Curr. Opin. Microbiol. 4: 28–34.

Embley T.M. and Stackebrandt E. 1994. The molecular phylogeny and systematics of the actinomycetes. Annu. Rev. Microbiol. 48: 257–289.

Fleischmann R.D., Alland D., Eisen J.A., Carpenter L., White O., Peterson J., Deboy R., Dodson R., Gwinn M., Haft D., Hickey E., Kolonay J.F., Nelson W.C., Umayam L.A., Ermolaeva M., Salzberg S.L., Delcher A., Utterback T., Weidman J., Khouri H., Gill J., Mikula A., Bishai W., Jacobs W.R., Venter J.C. and Fraser C.M. 2002. Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. J. Bacteriol. 184: 5479–5490.

Fraser C.M., Read T.D. and Nelson K.E. (eds), 2004. Microbial Genomes. Humana Press, Totowa, NJ.

Gao B. and Gupta R.S. 2005. Conserved indels in protein sequences that are characteristic of the phylum *Actinobacteria*. Int. J. Syst. Evol. Microbiol. 151: 2647–2657.

Garnier T., Eiglmeier K., Camus J.C., Medina N., Mansoor H., Pryor M., Duthoy S., Grondin S., Lacroix C., Monsempe C., Simon S., Harris B., Atkin R., Doggett J., Mayes R., Keating L., Wheeler P.R., Parkhill J., Barrell B.G., Cole S.T., Gordon S.V. and Hewinson R.G. 2003. The complete genome sequence of Mycobacterium bovis. Proc. Natl. Acad. Sci. USA 100: 7877–7882.

Garrity G.M. and Holt J.G. 2001. The road map to the manual. In: Boone D.R. and Castenholz R.W. (eds.), Bergey's Manual of Systematic Bacteriology, Springer-Verlag, Berlin, pp. 119–166.

Gogarten J.P. and Townsend J.P. 2005. Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. 3: 679–687.

Goodfellow M. and Williams S.T. 1983. Ecology of Actinomycetes. Annu. Rev. Microbiol. 37: 189–216.

Gordon S.V., Eiglmeier K., Garnier T., Brosch R., Parkhill J., Barrell B., Cole S.T. and Hewinson R.G. 2001. Genomics of Mycobacterium bovis. Tuberculosis 81: 157–163.

Griffiths E. and Gupta R.S. 2004. Signature sequences in diverse proteins provide evidence for the late divergence of the order *Aquificales*. Intl. Microbiol. 7: 41–52.

Griffiths E., Petrich A. and Gupta R.S. 2005. Conserved indels in essential proteins that are distinctive characteristics of *Chlamydiales* and provide novel means for their identification. Microbiology 151: 2647–2657.

Griffiths E., Ventresca M.S., Gupta R.S. 2006. BLAST screening of chlamydial genomes to identify signature proteins that are unique for the *Chlamydiales, Chlamydiaceae, Chlamydophila* and *Chlamydia* groups of species. BMC Genomics 7:14.

Gupta R.S. 1998. Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol. Mol. Biol. Rev. 62: 1435–1491.

Gupta R.S. 2000. The phylogeny of Proteobacteria: relationships to other eubacterial phyla and eukaryotes. FEMS Microbiol. Rev. 24: 367–402.

Gupta R.S. 2004. The Phylogeny and Signature Sequences characteristics of *Fibrobacters, Chlorobi* and *Bacteroidetes*. Crit. Rev. Microbiol. 30: 123–143.

Gupta R.S. 2005. Protein signatures distinctive of Alpha proteobacteria and its subgroups and a model for Alpha proteobacterial evolution. Crit. Rev. Microbiol. 31: 135.

Ikeda H., Ishikawa J., Hanamoto A., Shinose M., Kikuchi H., Shiba T., Sakaki Y., Hattori M. and Omura S. 2003. Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. Nat. Biotechnol. 21: 526–531.

Ishikawa J., Yamashita A., Mikami Y., Hoshino Y., Kurita H., Hotta K., Shiba T. and Hattori M. 2004. The complete genomic sequence of Nocardia farcinica IFM 10152. Proc. Natl. Acad. Sci. USA 101: 14925–14930.

Kainth P. and Gupta R.S. 2005. Signature proteins that are distinctive of alpha proteobacteria. BMC Genomics 6: 94.

Kalinowski J., Bathe B., Bartels D., Bischoff N., Bott M., Burkovski A., Dusch N., Eggeling L., Eikmanns B.J., Gaigalat L., Goesmann A., Hartmann M., Huthmacher K., Kramer R., Linke B., McHardy A.C., Meyer F., Mockel B., Pfefferle W., Puhler A., Rey D.A., Ruckert C., Rupp O., Sahm H., Wendisch V.F., Wiegrabe I. and Tauch A. 2003. The complete Corynebacterium glutamicum ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. J. Biotechnol. 104: 5–25.

Karlin S. and Altschul S.F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA 87: 2264–2268.

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

Lechevalier H.A. and Lechevalier M.P. 1967. Biology of actinomycetes. Annu. Rev. Microbiol. 21: 71–100.

Lerat E., Daubin V. and Moran N.A. 2003. From gene trees to organismal phylogeny in prokaryotes:the case of the gamma-proteobacteria. PLoS. Biol. 1: E19.

Ludwig W. and Klenk H.-P. 2001. Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systamatics. In: Boone D.R. and Castenholz R.W. (eds.), Bergey's Manual of Systematic Bacteriology, Springer-Verlag, Berlin, pp. 49–65.

Mazumder R., Natale D.A., Murthy S., Thiagarajan R. and Wu C.H. 2005. Computational identification of strain-, species- and genus-specific proteins. BMC Bioinform. 6: 279.

McAlpine J.B., Bachmann B.O., Piraee M., Tremblay S., Alarco A.M., Zazopoulos E. and Farnet C.M. 2005. Microbial Genomics as a guide to drug discovery and structural elucidation: ECO-02301, a novel antifungal agent, as an example. J. Nat. Prod. 68: 493–496.

Monteiro-Vitorello C.B., Camargo L.E.A., Van Sluys M.A., Kitajima J.P., Truffi D., do Amaral A.M., Harakava R., de Oliveira J.C.F., Wood D., de Oliveira M.C., Miyaki C., Takita M.A., da Silva A.C.R., Furlan L.R., Carraro D.M., Camarotte G., Almeida N.F., Carrer H., Coutinho L.L., El Dorry H.A., Ferro M.I.T., Gagliardi P.R., Giglioti E., Goldman M.H.S., Goldman G.H., Kimura E.T., Ferro E.S., Kuramae E.E., Lemos E.G.M., Lemos M.V.F., Mauro S.M.Z., Machado M.A., Marino C.L., Menck C.F., Nunes L.R., Oliveira R.C., Pereira G.G., Siqueira W., de Souza A.A., Tsai S.M., Zanca A.S., Simpson A.J.G., Brumbley S.M. and Setubal J.C. 2004. The genome sequence of the gram-positive sugarcane pathogen Leifsonia xyli subsp xyli. Mol. Plant Microb. Interact. 17: 827–836.

Moran N.A. and Wernegreen J.J. 2000. Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends Ecol. Evol. 15: 321–326.

Morse R., O'Hanlon K. and Collins M.D. 2002. Phylogenetic, amino acid content and indel analyses of the beta subunit of DNA-dependent RNA polymerase of gram-positive and gram-negative bacteria. Int. J. Syst. Evol. Microbiol. 52: 1477–1484.

Nishio Y., Nakamura Y., Kawarabayasi Y., Usuda Y., Kimura E., Sugimoto S., Matsui K., Yamagishi A., Kikuchi H., Ikeo K. and Gojobori T. 2003. Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of Corynebacterium efficiens. Genome Res. 13: 1572–1579.

Pedulla M.L., Ford M.E., Houtz J.M., Karthikeyan T., Wadsworth C., Lewis J.A., Jacobs-Sera D., Falbo J., Gross J., Pannunzio N.R., Brucker W., Kumar V., Kandasamy J., Keenan L., Bardarov S., Kriakov J., Lawrence J.G., Jacobs W.R., Hendrix R.W. and Hatfull G.F. 2003. Origins of highly mosaic mycobacteriophage genomes. Cell 113: 171–182.

Puech V., Chami M., Lemassu A., Laneelle M.A., Schiffler B., Gounon P., Bayan N., Benz R. and Daffe M. 2001. Structure of the cell envelope of corynebacteria: importance of the non-covalently bound lipids in the formation of the cell wall permeability barrier and fracture plane. Microbiology 147: 1365–1382.

Raoult D., Ogata H., Audic S., Robert C., Suhre K., Drancourt M. and Claverie J.M. 2003. Tropheryma whipplei twist: a human pathogenic Actinobacteria with a reduced genome. Genome Res. 13: 1800–1809.

Ravel J., DiRuggiero J., Robb F.T. and Hill R.T. 2000. Cloning and sequence analysis of the mercury resistance operon of *Streptomyces* sp. strain CHR28 reveals a novel putative second regulatory gene. J. Bacteriol. 182: 2345–2349.

Roller C., Ludwig W. and Schleifer K.H. 1992. Gram-positive bacteria with a high DNA G+C content are characterized by a common insertion within their 23S rRNA genes. J. Gen. Microbiol. 138: 167–175.

Rother D., Mattes R. and Altenbuchner J. 1999. Purification and characterization of MerR, the regulator of the broad-spectrum mercury resistance genes in Streptomyces lividans 1326. Mol. Gen. Genet. 262: 154–162.

Schell M.A., Karmirantzou M., Snel B., Vilanova D., Berger B., Pessi G., Zwahlen M.C., Desiere F., Bork P., Delley M., Pridmore R.D. and Arigoni F. 2002. The genome sequence of Bifidobacterium longum reflects its adaptation to the human gastrointestinal tract. Proc. Natl. Acad. Sci. USA 99: 14422–14427.

Schorey J.S., Li Q.L., Mccourt D.W., Bongmastek M., Clarkcurtiss J.E., Ratliff T.L. and Brown E.J. 1995. A mycobacterium-leprae gene encoding a fibronectin-binding protein is used for efficient invasion of epithelial-cells and schwann-cells. Infect. Immun. 63: 2652–2657.

Soliveri J.A., Gomez J., Bishai W.R. and Chater K.F. 2000. Multiple paralogous genes related to the Streptomyces coelicolor developmental regulatory gene whiB are present in Streptomyces and other actinomycetes. Microbiol.-UK 146: 333–343.

Stackebrandt E., Rainey F.A. and WardRainey N.L. 1997. Proposal for a new hierarchic classification system, Actinobacteria classis nov. Int. J. Syst. Bacteriol. 47: 479–491.

Stackebrandt E., Schumann P., (2000). Introduction to the taxonomy of actinobacteria. In: Dworkin M., et al. (eds) The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community. Springer-Verlag, New York, http://www.141.150.157.117:8080/prokPUB/chaprender/jsp/showchap.jsp?chapnum = 291.

Sutcliffe I.C. 1998. Cell envelope composition and organisation in the genus *Rhodococcus*. Antonie van Leeuwenhoek 74: 49–58.

Sutcliffe I.C. and Harrington D.J. 2004. Lipoproteins of Mycobacterium tuberculosis: an abundant and functionally diverse class of cell envelope components. FEMS Microbiol. Rev. 28: 645–659.

Sutcliffe I.C. and Russell R.R. 1995. Lipoproteins of gram-positive bacteria. J. Bacteriol. 177: 1123–1128.

Tauch A., Kaiser O., Hain T., Goesmann A., Weisshaar B., Albersmeier A., Bekel T., Bischoff N., Brune I., Chakraborty T., Kalinowski J., Meyer F., Rupp O., Schneiker S., Viehoever P. and Puhler A. 2005. Complete genome sequence and analysis of the multiresistant nosocomial pathogen Corynebacterium jeikeium K411, a lipid-requiring bacterium of the human skin flora. J. Bacteriol. 187: 4671–4682.

Ueda K., Ohno M., Yamamoto K., Nara H., Mori Y., Shimada M., Hayashi M., Oida H., Terashima Y., Nagata M. and Beppu T. 2001. Distribution and diversity of symbiotic thermophiles, Symbiobacterium thermophilum and related bacteria, in natural environments. Appl. Environ. Microbiol. 67: 3779–3784.

Ueda K., Yamashita A., Ishikawa J., Shimada M., Watsuji T., Morimura K., Ikeda H., Hattori M. and Beppu T. 2004. Genome sequence of Symbiobacterium thermophilum, an uncultivable bacterium that depends on microbial commensalism. Nucleic Acids Res. 32: 4937–4944.

Yang Z. 2005. The power of phylogenetic comparison in revealing protein function. Proc. Natl. Acad. Sci. USA 102: 3179–3180.

Zazopoulos E., Huang K.X., Staffa A., Liu W., Bachmann B.O., Nonaka K., Ahlert J., Thorson J.S., Shen B. and Farnet C.M. 2003. A genomics-guided approach for discovering and expressing cryptic metabolic pathways. Nat. Biotechnol. 21: 187–190.