UPTEC X 06 014
APR 2006

ISSN 1401-2138

JOHAN VIKLUND

ORFans within the alphaproteobacteria and their frequency in one environmental sample

Master's degree project

UPPSALA UNIVERSITET

Molecular Biotechnology Programme

Uppsala University School of Engineering

UPTEC X 06 014	**Date of issue 2006-04**

Author

Johan Viklund

Title (English)

ORFans within the alphaproteobacteria and their frequency in one environmental sample

Abstract

Microbial genomes contain a large proportion of ORFan genes. Our understanding of the mechanism that generate these and how common they are in nature is lacking. The aim of this project has been to identify these genes within the α-proteobacteria and to develop a method for studying their presence in nature. As a part of this project a database was built that contains all sequenced α-proteobacterial genomes. 25 different α-proteobacterial species were analyzed. 10000 genes uniquely present in the α-proteobacteria were identied, most were ORFans i.e. present in only a single species. Approximately half of these gave weak signals in BLAST searches against a small soil environmental dataset of 5,000 sequences, of which only 30 were mutually best hits.

Keywords

Supervisors

Siv Andersson

Department of evolution, genomics and systematics, Uppsala University

Scientific reviewer

David Ardell

The Linnaeus Centre for Bioinformatics, Uppsala university

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **16**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

ORFans within the alphaproteobacteria and their frequency in one environmental sample

Johan Viklund

Sammanfattning

Studier av genom från bakterier har visat att det finns en stor mängd gener som är specifika för enskilda bakteriearter. Dessa gener har ingen känd funktion och vi vet heller inte hur de dyker upp. I det här examensarbetet har utbredningen av sådana gener inom bakteriegruppen α-proteobakterier kartlagts.

α -proteobakterierna är en stor samling besläktade bakterier. De har mycket olika livsstil och har bland annat spelat en nyckelroll för utvecklingen av flercelliga organismer. En del av dem kan orsaka sjukdommar hos männsikor.

Lite drygt 10000 olika gener hittades som var unika för α-proteobakterier. De flesta av dessa förekom enbart hos ”enskilda arter”. Flertalet av dessa är förmodligen inte riktiga gener utan snarast felmärkta som gener. Det riktigt intressanta var de gener som delades mellan flera olika α-proteobakterier. Eftersom de förmodligen är nya, rättmärkta, gener.

För att försöka ta reda på hur vanliga de här generna är i naturen så användes ett metagenomiskt dataset. I en metagenomisk studie sekvenserar man direkt från naturen utan att veta vilken eller vilka arter som DNA-sekvenserna kommer ifrån. Ett ganska litet dataset användes i pilotförsöket. De gener som fick träffar var företrädelsevis från jordlevande bakterier, vilket var väntat då den metagenomiska studien var på jord.

Examensarbete 20 p i Civilingenjörsprogrammet för Bioinformatik

Uppsala universitet April 2006

Contents

List of Figures

List of Tables

1 Introduction

1.1 ORFans

ORFans are open reading frames (ORFs) that share no homology with other genomes. The word ORFan is a contraction of the two words orphan and ORF. For each new genome that is published, the total number of ORFans increases. Homologous ORFans are ORFans that share homology only with very closely related organisms [1]. In this study the term ORFan refers to both true ORFans and homologous ORFans. Most of these have no known function: some are known to code for proteins, but most are hypothetical. It has been showed by Amiri *et al.* [2] that some of these genes are parts of genes being deleted from the genome. Ochman [3] and Skovgaard *et al.* [4] have shown that there is a large proportion misannotated genes in this category, especially short ones get annotated as genes more often than they should.

Metagenomics, which is the study of the genomic content of complete eco-systems, offer a new approach to study these genes. The *α-proteobacteria* is a group of bacterial species where some genomes have been sequenced and published, they have varying lifestyles and are predominant in several different eco-systems.

1.2 Metagenomics

Most of our understanding of microbiology and genomics has been gained from cultured bacteria. Cultured bacteria are bacteria that can be grown in a lab. Estimates for the number of species that can be cultured today in the environment are less than one percent [5]. The fact that there are so few cultured species is a problem as it implies that our knowledge of microbiology is very biased.

In 1985, the first steps were taken towards the field that now is called metagenomics. Lane *et al.* sequenced 16S rRNA genes from the environment without culturing and used them to estimate the taxonomic diversity of their sample [6].

A modern definition of metagenomics is "the functional and sequence-based analysis of the collective microbial genomes that are contained in an environmental sample" [7]. Where the environmental sample can be taken from, for example, soil, water or the gastric tract.

As mentioned there are two basic types of metagenomic studies, functional and sequence-based. Functional studies have yielded knowledge about many proteins. For example new antibiotics and new antibiotic resistance genes have been discovered. The aim of functional studies is often to find new or better genes [8].

The sequence based approach is mainly used to determine the diversity of a sample [8], often by scanning for certain phylogenetic markers, such as 16S rRNA. Increasingly the method has been to sequence randomly from the environment sample. This can then be used to classify the sample and/or determine what to study in more detail.

One of the largest and most famous metagenomic studies has been done by Venter *et al.*, who sampled and sequenced surface waters from the Sargasso ocean [9]. They used shotgun sequencing over the complete dataset and sequenced over one billion base pairs and found 1.2 million new genes.

Metagenomics is a growing field and it will probably lead to many new and interesting discoveries.

1.3 Alpha-proteobacteria

The class *α-proteobacteria* contains a diverse group of species. Today the the genomes of 32 different species and strains of *α-proteobacteria* have been sequenced [10], of which three were sequenced at the department of Molecular Evolution in Uppsala. They live both in relationships with other cells and as free cells which can be found in most biota.

Recently metagenomic studies have suggested that *α-proteobacteria* is one of the most abundant classes in ocean surface waters [9, 11]. This is mainly due to the extreme commonality of the so called SAR11 clade which is *α-proteobacteria* [12]. A clade is a group of one or more species with a common ancestor. One species from the SAR11 clade has been cultured and sequenced, *Pelagibacter ubique*. It has the smallest genome size of a free-living bacteria sequenced to date [13, 14].

Many of the *α-proteobacteria* live in close interaction with eukaryotes as parasites or endosymbionts. Some of them can infect several different hosts, often mediated by a vector, e.g. louse or tick. In man they can cause several diseases, for example trench-fever (*Bartonella quintana*), cat-scratch disease (*Bartonella henselae*) and epidemic typhus (*Rickettsia prowazekii*) [15].

Their evolutionary history is also interesting. The intracellular lifestyle has evolved at least twice in the *α-proteobacteria*, once in the *Brucella/Bartonella* clade and once in the *Rickettsia/Wolbachia/Ehrlichia* clade. In both of these lineages, the transition to the intracellular environment has been associated with a genome reduction [15]. Andersson *et al.* [16] showed that the mitochondria and the *α-proteobacterial* species *R. prowazekii* are very closely related.

1.4 BLAST

The main tool used in this project was *Basic Local Alignment Search Tool* (BLAST) [17]. It is a program used to search for sequences similar to a query sequence. The algorithm scores each alignment (hit) it gets, the score is usually calculated from a BLOSUM matrix (if it is amino acids in the query-sequence). The matrix represents different costs for amino acid changes. In this project the BLOSUM62 matrix was used. The score is then converted into a E-value by this formula:

$$E = Kmne^{\lambda S}$$

where m and n are the length of the database and the query sequence, S is the score and K and λ are parameters. The E-value is the expected number of hits with score at least S.

For small values of E (less than 0.01), E is essentially the same as the probability of finding at least one hit with that score (sometimes referred to as the P-value).

2 Aims

The aims of this project were three-fold. The first part was to construct a database for the *α-proteobacteria*. The second to identify ORFans and homologous ORFans in the *α-proteobacteria*. The third part was to devise a method for quantifying the distribution of these genes in environmental datasets. Such a quantification might also give information about how common different *α-proteobacterial* species are in the environment.

3 Materials and Methods

3.1 Datasets

Two large gene sets were constructed, *prok* which was the set of all protein coding genes in all sequenced prokaryotes and *alpha* which was a subset of *prok* that only contained *α-proteobacteria*. The *alpha*-set contained 25 *α-proteobacterial* genomes (Table 1, p10). All sequences were downloaded from the National Center of Biotechnology Information (NCBI) [18].

An environmental dataset, *darm*, were retrieved from Treusch *et al.* [19]. They took three soil-samples outside Darmstadt, Germany, two grassland and one forest soil. The size of the fosmid library they constructed was estimated to three Gbp, but they have only sequenced four Mbp, or 5376 sequence reads.

3.2 Procedure

3.2.1 Extracting ORFans and homologous ORFans

To identify ORFans from the *α-proteobacteria* the genes from the *alpha*-set were blasted with *blastp* against the *prok*-set with a cutoff at $E = 10^{-3}$, *blast1* in Figure 2. If there were only hits against *α-proteobacteria* with $E \leq 10^{-10}$ and no hits at all above that limit, the hits from that query were added to a homologous cluster of ORFans. Duplicate clusters were removed.

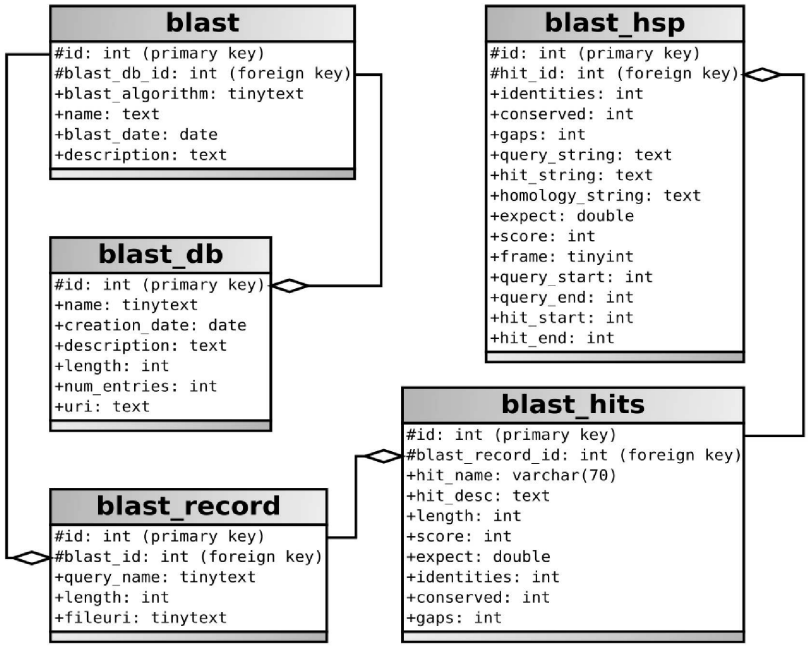

Figure 1: *Blast-schema:* The tables and their relations in the blast database.

3.2.2 Distribution of ORFan genes in *darm*

To estimate the frequency of the clustered ORFans in *darm*, the genes from the clusters were blasted with *tblastx* against *darm* (*blast2* in Figure 2). The cutoff used in the blast was $E = 10^{-3}$.

The hits from this blast were then blasted with *tblastx* against the full *prok* dataset (*blast3*). If the *darm* sequence did not hit against the query-sequence from the *alpha*-set, it was filtered out.

3.3 Database and Software

A database for storing genome information was developed as part of this project. A Perl front-end to the database was also developed. The database schema was based on the BioSQL project [20]. In this database all *α-proteobacterial* species in Table 1 are currently stored. It is possible to query the database for very different information, for example where a gene is positioned, which genome it is in, how long it is, what annotations it has and so forth. The perl-frontend can return bioperl [20] objects for some of the queries.

A database schema for storing blast-results was developed. The schema can be found in Figure 1. The database currently (2006-01-19) stores 187,803 blast queries. A Perl interface was written using `Class::DBI` from CPAN [21].

The database manager used in this project was MySQL 4.1, which is an open source

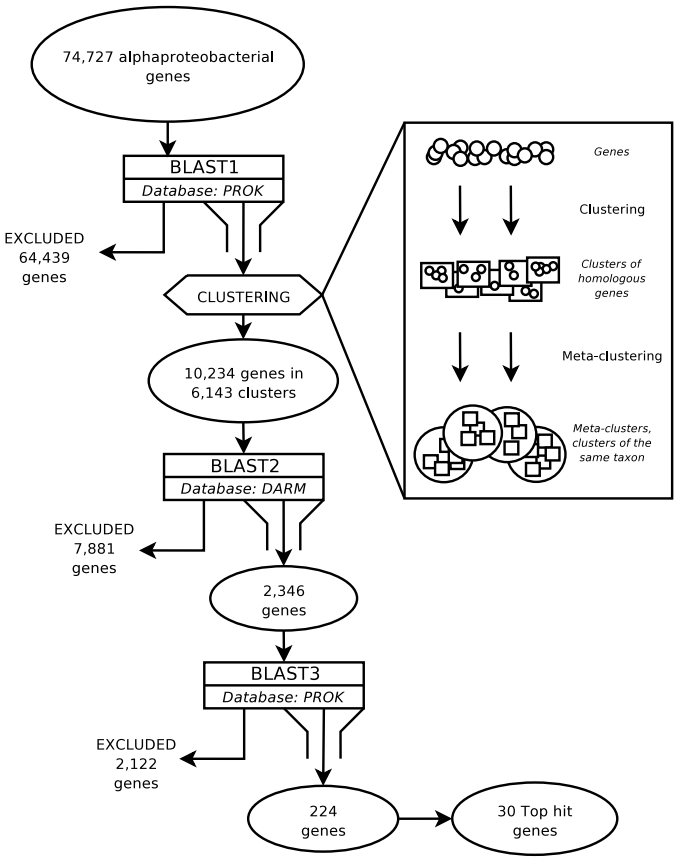

Figure 2: *Workflow:* The overall workflow and the clustering procedure. The input for the analysis is the *α-proteobacterial*-genes which then went through three blast steps with subsequent filtering. The database used in the respective blasts are shown in the figures, *prok* is all known protein coding genes in all sequenced prokaryotes while *darm* is an environmental dataset. The clustering was done according to homology, and the meta-clustering was based on species content of the clusters.

database manager [22].

4 Results

4.1 ORFan genes in the *α-proteobacterial* genomes

In this study 25 *α-proteobacterial* genomes were analyzed. They had sizes ranging from one to nine Mbp, and the number of protein-coding genes ranged between 805 and 8316. These were all placed in a database that was based on the BioSQL project.

To search for ORFan genes within the *α-proteobacteria*, the genes of the *α-proteobacterial* genomes in the database were blasted against all sequenced prokaryote genomes (*blast1*, Figure 2) with a cutoff at $E = 10^{-3}$. If there were only hits below $E \leq 10^{-10}$ against *α-proteobacteria* and no hits above that limit, the hits were considered to form a homologous

cluster of ORFans. This resulted in 6143 clusters of ORFans. These clusters were then collected into meta-clusters, where each meta-cluster contained all clusters with the same species content (Figure 2).

Table 1: Meta-clusters with one species. Mean and median lengths are in aminoacids (denoted as Mean and Median in the table), the genome size is in Mbp. The mean and median lengths are calculated both for all genes in each genome and for all the genes in each meta-cluster. ACS is the average cluster size. N is the number of clusters or the number of genes.

Species	Meta-clusters with one species				All genes			Genome
	Mean	Median	N	ACS	Mean	Median	N	size
Agrobacterium tumefaciens str. C58	146.47	118	577	1.44	313.52	286	10690	5.67
Anaplasma marginale str. St. Maries	401.25	242	86	1.20	359.58	283	949	1.20
Bartonella henselae str. Houston-1	127.25	120	58	1.80	314.12	252	1488	1.93
Bartonella quintana str. Toulouse	106.36	105	11	1.00	332.59	276	1142	1.58
Bradyrhizobium japonicum USDA 110	185.66	137	825	1.06	316.75	277	8316	9.11
Brucella abortus biovar 1 str. 9-941	121.40	126	5	1.00	288.09	255	3085	3.29
Brucella melitensis 16M	96.73	84	43	1.02	297.13	263	3198	3.30
Brucella suis 1330	78.55	59	33	1.00	284.41	254	3271	3.32
Caulobacter crescentus CB15	191.19	153	432	1.04	323.60	275	3727	4.02
Ehrlichia canis str. Jake	202.34	184	62	1.23	340.79	270	925	1.32
Ehrlichia ruminantium str. Gardel	80.38	70	13	1.00	335.39	270	950	1.50
Ehrlichia ruminantium str. Welgevonden	72.58	69	12	1.00	341.96	276	1846	1.51
Gluconobacter oxydans 621H	204.39	161	333	1.11	326.73	287	2664	2.92
Mesorhizobium loti MAFF303099	166.12	130	669	1.12	299.40	269	7272	7.60
Pelagibacter ubique HTCC1	174.45	141	112	1.01	307.61	267	1354	1.31
Rhodopseudomonas palustris CGA009	175.67	132	283	1.07	328.44	284	4819	5.47
Rickettsia conorii str. Malish 7	86.78	79	93	1.00	247.82	173	1374	1.27
Rickettsia felis URRWXCal2	180.34	139	94	1.15	291.74	239	1512	1.59
Rickettsia prowazekii str. Madrid E	203.80	90	5	1.00	334.42	282	835	1.11
Rickettsia typhi str. Wilmington	52.25	45	4	1.00	333.03	280	838	1.11
Silicibacter pomeroyi DSS-3	223.51	174	364	1.03	323.19	290	4252	4.60
Sinorhizobium meliloti 1021	145.99	118	356	1.09	309.01	281	6203	6.69
Wolbachia endosymbiont wBm	92.95	80	53	1.04	298.82	242	805	1.27
Wolbachia endosymbiont wMel	157.48	92	111	1.31	282.66	224	1195	1.08
Zymomonas mobilis subsp. mobilis ZM4	162.52	116	200	1.07	294.76	254	1998	2.06
Total	174.45	131	4834	1.13	311.34	272	74708	86.39

The size of a cluster is the number of genes in that cluster while the size of a meta-cluster is the number of clusters in that meta-cluster. There were 193 different meta-clusters. Most of these were quite small, 129 of them had a size of three clusters or less. 79% of the clusters contained only one species, these are all shown in Table 1. Of the large meta-clusters, containing 10 clusters or more, most contained only one species. For most of the intracellular organisms the meta-clusters containing several species were bigger than the ones containing one (Figure 3).

The average cluster size (ACS) of a meta-cluster is the average number of genes of all the clusters in a meta-cluster. The average cluster size for a meta-cluster with only one species is a measure of the amount of duplications of the ORFans in that species. The average cluster size of the clusters containing only one species ranged from 1.00 to 1.80. The average ACS was only 1.13, which indicates that duplications are not that common. The extremes with lots of duplications were *A. tumefaciens*, *B. henselae* and *W. wMel*, of these *A. tumefaciens* is free-living while the other two are intracellular and from completely different branches of the α-*proteobacterial* tree.

Table 2: Meta-clusters with more than one species. Mean and median lengths are in aminoacids (denoted as Mean and Median in the table). CS is the cluster size.

Species	Mean	Median	CS
B. japonicum, S. meliloti	190.30	180	10
A. tumefaciens, M. loti, S. meliloti	214.12	175	10
B. japonicum, M. loti, R. palustris	318.91	284	10
A. marginale, E. canis, E. ruminantium	338.34	277	10
A. tumefaciens, M. loti	209.36	166	13
A. tumefaciens, B. abortus, B. henselae, B. japonicum, B. melitensis, B. quintana, B. suis, M. loti, R. palustris, S. meliloti	243.64	196	14
A. marginale, E. canis, E. ruminantium, Wolbachia wBmu, Wolbachia wMel	265.14	209	16
B abortus, B. melitensis	136.71	107	17
A. tumefaciens, B. abortus, B. melitensis, B. suis, M. loti, S. meliloti	211.49	186	20
B. japonicum, M. loti	283.90	235	25
B. melitensis, B. suis	172.12	114	25
B. loti, B. meliloti	209.44	146	25
E. canis, E. ruminantium	499.13	342	32
B. henselae, B. quintana	137.46	111	43
A. tumefaciens, S. meliloti	163.45	132	43
Wolbachia wBmu, Wolbachia wMel	239.94	219	62
R. conorii, R. felis, R. prowazekii, R. typhi	323.74	241	90
B. abortus, B. suis	69.20	59	93
B. abortus, B. melitensis, B. suis	127.28	96	133
B. japonicum, R. palustris	214.89	175	137
R. conorii, R. felis	269.94	223	160

The average and median lengths for the genes in each meta-cluster were calculated and for all proteincoding genes in each genome. As can be seen the average and median lengths of the clusters are generally between half and a third of the lengths in the genome from which the ORFans came from (Table 1). There are some exceptions though. The average gene length in the meta-clusters containing more than one species is 252.42 which is longer than the average for the meta-clusters with only one species. The average and median gene length for the larger meta-clusters is presented in Table 2.

4.2 Presence of ORFan genes in Environment

To develop a method for quantifying the distribution of the ORFan genes of the *α-proteobacteria* in the environment, a small test-dataset acquired from Treusch *et al.* [19] was used (*darm*). This dataset was only four Mbp in size. Each gene from the *α-proteobacteria* clusters were blasted against the *darm*-dataset (*blast2*) with a cutoff at $E = 10^{-3}$. This resulted in 46163 hits to the *darm*-dataset, which were distributed over 2424 *darm*-sequences and 2346 *α-proteobacterial* genes.

To verify which of these were good hits, all of the *darm*-hits were blasted against the *prok*-dataset (*blast3*), again with a cutoff at $E = 10^{-3}$. If the *darm*-sequence did not hit against the query-sequence from the *α-proteobacterial*-sequence, it was excluded from further analysis. This limited the number of Darmstadt sequences further down to 540 sequences, or the number of *α-proteobacterial*-genes to 224. If all except best hits against the *α-proteobacterial*-sequences were discarded only 30 *darm*-sequences remained. These were

dominated by earth-living *α-proteobacteria*, especially by *M. Loti*, and where *B. henselae*, *C. crescentus* and *R. felis* were the exceptions.

The distribution between the *α-proteobacteria* of the hits in the two blast searches can be found in Table 3.

The average length of the hits in *blast3* was 293.12 and the average length of the top hits was 437.00. Almost all of the hits (91%) were against genes annotated as hypothetical.

Table 3: *Environmental hits*. The second column is the number of genes in the *alpha*-set that hit something in *darm* while the third is the number of hits. The fourth column is the number of *darm* sequences that hit the original *α-proteobacterial* sequence, the fifth the number of hits and the sixth the number of those that were top hits.

Species	Blast 2		Blast 3		
	Genes	Hits	*darm*seqs.	Hits	Best hits
Agrobacterium tumefaciens str. C58	228	1025	16	61	0
Anaplasma marginale str. St. Maries	5	10	1	2	0
Bartonella henselae str. Houston-1	6	154	3	100	7
Bartonella quintana str. Toulouse	3	9	1	1	0
Bradyrhizobium japonicum USDA 110	426	8778	27	56	3
Brucella abortus biovar 1 str. 9-941	25	43	3	4	0
Brucella melitensis 16M	17	33	2	2	0
Brucella suis 1330	26	45	3	3	0
Caulobacter crescentus CB15	356	18486	52	429	1
Ehrlichia canis str. Jake	3	3	1	1	0
Ehrlichia ruminantium str. Gardel	1	2	1	1	0
Ehrlichia ruminantium str. Welgevonden	2	4	2	2	0
Gluconobacter oxydans 621H	134	1140	8	109	0
Mesorhizobium loti MAFF303099	364	4880	30	177	13
Pelagibacter ubique HTCC1	17	67	7	29	0
Rhodopseudomonas palustris CGA009	208	5083	16	225	1
Rickettsia conorii str. Malish 7	5	7	1	1	0
Rickettsia felis URRWXCal2	25	46	9	11	1
Rickettsia prowazekii str. Madrid E	1	4	1	4	0
Rickettsia typhi str. Wilmington	4	7	2	5	0
Silicibacter pomeroyi DSS-3	292	4843	10	47	1
Sinorhizobium meliloti 1021	170	1180	16	33	2
Wolbachia endosymbiont wBm	3	8	1	1	0
Wolbachia endosymbiont wMel	7	13	2	3	0
Zymomonas mobilis subsp. mobilis ZM4	18	293	9	36	1
Sum	2346	46163	224	1343	30

5 Discussion

The meta-clusters containing only one taxon is bigger on average than the meta-clusters containing several taxa, almost 80% of all the clusters contained only one taxon. Most of these genes are short, the average length is 170, and most of these are probably not protein coding but rather misannotated [3, 4].

The meta-clusters containing several taxa have longer genes than the ones containing only one, which suggests that these genes probably are protein-coding. Even these genes are a bit shorter than the average for all the annotated protein-coding genes. But since they are shared between several genomes gives support to them being protein-coding. It

would be very interesting to try to find out what function these new genes have.

Some of the meta-clusters did not map onto the *α-proteobacterial* phylogeny. Especially the *Brucellas* where all possible combinations had some support among the clusters. These are all very closely related however. *M. loti*, *S. meliloti*, *A. tumefaciens* and *B. japonicum* also share a lot of genes in different combinations of the species, they are all soil-dwelling so this gene-sharing pattern might be due to horizontal transfer. The less likely, but in some cases plausible, explanation is that the genes have been lost in the other lineages in the phylogeny.

The intracellular parasites generally share a larger proportion of genes between them. Most of these are closer related than their free-living relatives so it is hard to determine if this is due to their lifestyle or their relatedness.

Some of the short ORFans are probably gene-fragments as shown in [2]. There are probably also bacteriophage genes among these short ORFans.

The test dataset gave little results but this is probably due to the small size of that dataset. It provided some insights though. Most of the top hits from *darm* to *prok* where against soil-dwelling *α-proteobacteria* which confirms the applicability of the method. The *darm*-sequences were random parts of genomes which means that they could contain the end of one gene and the beginning of another. So if a sequence hit one small part of the *darm*-sequence it might be more similar to other sequences. This fact might explain the big difference between the number of hits in *blast2* and *blast3*. Additionally my criteria were very strict to minimize the number of false positives.

The genes that did hit against the *darm*-set were longer than average. And, as mentioned previously, the longer genes are more likely to be expressed as proteins. So there might be

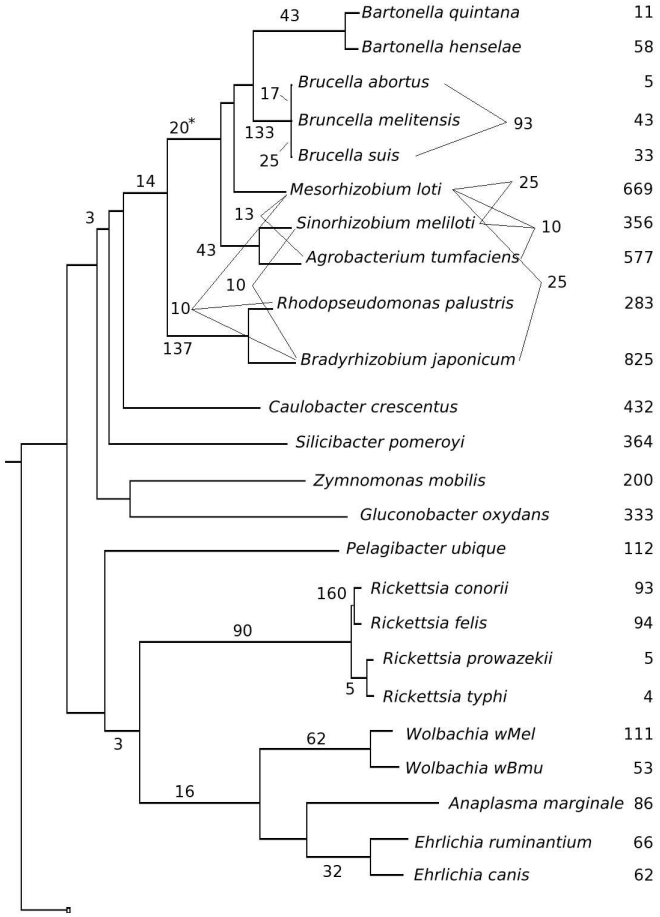

Figure 3: *Phylogenetic tree with meta-clusters.* Distribution of the number of clusters for each meta-cluster where the meta-cluster either support the phylogenetic tree or were the meta-cluster has 10 or more clusters supporting it. Numbers to the right indicate the number of clusters with only that species. On the branches the numbers indicate the number of clusters in that clade, except for * where the *Bartonella* is missing. The other numbers indicate the number of clusters having those species the lines point at. The tree was made by Björn Sällström with the method used in [15].

a bias towards expressed genes in the hits against environmental datasets, at least for *darm*.

The method for looking for ORFan genes in environmental datasets has been tested on a small dataset in this pilot study. The next step is to look in a bigger environmental dataset, for example the one collected by Venter et al. [9].

6 Acknowledgments

I would like to thank my supervisor Siv Andersson for giving me such a fun project to work on and for keeping me on the right track. I also want to thank Björn Sällström for answering questions and coming with insights and interesting views, and for reading this report. Alistair Darby for his comments on the report, without which it would have been really bad. David Ardell, my scientific reviewer, for reading this document so quickly and for the comments he had. I am also grateful to all the other people at the department of molecular evolution for giving me a great time during my degree project. My opponents for giving me good comments and reading the report.

And to my girlfriend Emmeli, for putting up with me and reading this project over and over and over again.

7 References

[1] N. Siew and D. Fischer. Twenty thousand orfan microbial protein families for the biologist? *Structure*, 11(1):7–9, January 2003.

[2] H. Amiri, W. Davids, and S. G. Andersson. Birth and death of orphan genes in rickettsia. *Mol Biol Evol*, 20(10):1575–1587, October 2003.

[3] H. Ochman. Distinguishing the orfs from the elfs: short bacterial genes and the annotation of genomes. *Trends Genet*, 18(7):335–337, July 2002.

[4] M. Skovgaard, L. J. Jensen, S. Brunak, D. Ussery, and A. Krogh. On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet*, 17(8):425–428, August 2001.

[5] R. I. Amann, W. Ludwig, and K. H. Schleifer. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev*, 59(1):143–169, March 1995.

[6] D. J. Lane, B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin, and N. R. Pace. Rapid determination of 16s ribosomal rna sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A*, 82(20):6955–6959, October 1985.

[7] E. E. Allen and J. F. Banfield. Community genomics in microbial ecology and evolution. *Nature Reviews Microbiology*, 3(6):489–498, June 2005.

[8] J. Handelsman. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*, 68(4):669–685, December 2004.

[9] C. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Yu-Hui Rogers, and H. O. Smith. Environmental genome shotgun sequencing of the sargasso sea. *Science*, 304(5667):66–74, April 2004.

[10] KEGG. The koyoto encyclopedia of genes and genomes. *http://www.genome.jp/kegg/catalog/org_list.html*, January 2006.

[11] R. M. Morris, M. S. Rappe, S. A. Connon, K. L. Vergin, W. A. Siebold, C. A. Carlson, and S. J. Giovannoni. Sar11 clade dominates ocean surface bacterioplankton communities. *Nature*, 420(6917):806–810, December 2002.

[12] S. J. Giovannoni, T. B. Britschgi, C. L. Moyer, and K. G. Field. Genetic diversity in sargasso sea bacterioplankton. *Nature*, 345(6270):60–63, May 1990.

[13] M. S. Rapp, S. A. Connon, K. L. Vergin, and S. J. Giovannoni. Cultivation of the ubiquitous sar11 marine bacterioplankton clade. *Nature*, 418(6898):630–633, August 2002.

[14] S. J. Giovannoni, H. J. Tripp, S. Givan, M. Podar, K. L. Vergin, D. Baptista, L. Bibbs, J. Eads, T. H. Richardson, M. Noordewier, M. S. Rapp, J. M. Short, J. C. Carrington, and E. J. Mathur. Genome streamlining in a cosmopolitan oceanic bacterium. *Science*, 309(5738):1242–1245, August 2005.

[15] B. Sällström and S. G. G. Andersson. Genome reduction in the alpha-proteobacteria. *Curr Opin Microbiol*, August 2005.

[16] S. G. Andersson, A. Zomorodipour, J. O. Andersson, T. Sicheritz-Pontn, U. C. Alsmark, R. M. Podowski, A. K. Näslund, A. S. Eriksson, H. H. Winkler, and C. G. Kurland. The genome sequence of rickettsia prowazekii and the origin of mitochondria. *Nature*, 396(6707):133–140, November 1998.

[17] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, September 1997.

[18] NCBI. National center for biotechnology information. *http://www.ncbi.nlm.nih.gov*, September 2005.

[19] A. H. Treusch, A. Kletzin, G. Raddatz, T. Ochsenreiter, A. Quaiser, G. Meurer, S. C. Schuster, and C. Schleper. Characterization of large-insert dna libraries from soil for environmental genomic studies of archaea. *Environ Microbiol*, 6(9):970–980, September 2004.

[20] OpenBIO. *http://obda.open-bio.org/*, August 2005.

[21] T. Bowden. Cpan, comprehensive perl archive network. *http://search.cpan.org/∼tmtm/Class-DBI-v3.0.14/lib/Class/DBI.pm*, September 2005.

[22] MySQL. *http://www.mysql.com*, October 2004.