UPTEC X 06 042
OKT 2006

ISSN 1401-2138

GREGER HÄLLTORP

A phylogenomic study of *Francisella tularensis*

Master's degree project

UPPSALA
UNIVERSITET

Molecular Biotechnology Programme

Uppsala University School of Engineering

UPTEC X 06 042

Date of issue 2006-10

Author

Greger Hälltorp

Title (English)

A phylogenomic study of *Francisella tularensis*

Title (Swedish)

Abstract

Francisella tularensis is one of the most infectious organisms known, requiring as few as 10 bacteria to cause infection. In this study a phylogenomic analysis of *F. tularensis* was performed, constructing phylogenetic trees for 210 *F. tularensis* genes that had homologues in each of 15 other γ-proteobacteria and *Bacillus anthracis* Ames. The results indicate that *F. tularensis* is not closely related to any of the other included γ-proteobacteria.

Keywords

Francisella tularensis, tularaemia, phylogeny, phylogenomic, phylome

Supervisors

Siv Andersson
Institutionen för evolution, genomik och systematik, Uppsala universitet

Scientific reviewer

Mikael Thollesson
Institutionen för evolution, genomik och systematik , Uppsala universitet

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **36**

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

A phylogenomic study of *Francisella tularensis*

Greger Hälltorp

Sammanfattning

Francisella tularensis orsakar sjukdomen tularemi, i Sverige mest känd som harpest. Det är en av de mest infektiösa organismer som finns. Eftersom organismen så lätt kan infektera en människa finns det farhågor för att den skulle kunna användas av terrorister i en bioterrorattack. Det här arbetet var en del av ett sekvenseringsprojekt för *F. tularensis*.

I analysen av en organism är det av intresse att se hur den är besläktad med andra organismer i samma familj. För att undersöka sådana släktband gör man så kallade fylogenier, som oftast representeras av fylogenetiska träd där släktskapet mellan de olika ingående organismerna framgår. Fylogenetiska träd kan skapas utifrån olika typer av data och utvecklingen av sekvensering och bioinformatik har lett till att en vanlig typ av data är sekvensen hos organismens gener eller proteiner. Ett fylogenetiskt träd kan skapas där ett specifikt protein i den organism man är intresserad av, jämförs med motsvarande proteiner i andra besläktade organismer. För att sedan få en helhetsbild av hur organismerna är besläktade kan samma analys göras för alla, eller en så stor andel som möjligt, av organismens proteiner. Sammantaget kan alla dessa fylogenier ge en bild av hur organismen är besläktad med de andra organismerna i undersökningen. I det här arbetet gjordes en sådan analys; 210 proteiner hos *F. tularensis* jämfördes med motsvarande proteiner hos 15 andra bakterier i samma familj.

Examensarbete 20 p i Molekylär bioteknikprogrammet

Uppsala universitet Oktober 2006

Contents

Introduction

1.1 Biological Warfare and Bioterrorism

In the post 9/11-era the threat of terrorism is felt all over the world. For instance, the felt and real terrorist threat to the Olympic Summer Games held in Athens in 2004 lead to a total bill for security and anti-terrorist measures of $1.5 billion. This is four times more than that of the Olympic Games in Sydney in 2000.

The deliberation and earnestness of the attacks on the WTC and the Pentagon show that terrorists today are quite ruthless. Moreover they do not seem overly concerned with their own survival, sometimes even coveting their "martyr" deaths. The attacks also showed that modern terrorists are capable of launching massive attacks with unconventional methods. Taken together, all these facts point toward a very scary scenario: it seems unlikely that modern terrorists would be hesitant about using Weapons of Mass Destruction (WMD) if they but had the means and the materials at hand.

It has been suggested[9] that out of the three branches of WMD: Atomic (Nuclear), Biological and Chemical, the one most likely to be used in terrorism is that of Biological weapons. The reason for this is that although biological weapons are capable of causing almost as great damage as nuclear weapons, the lack of control makes them much easier to obtain. The threat of a terrorist attack using some sort of biological weapon has gone from being a part of the plot of a thriller set in the future to being a real threat today but the concept of biological weapons is not a new idea.

1.1.1 A historical perspective

We often think of biological warfare and biological weaponry as being rather modern ideas and in the implementations we usually imagine today, they are. The general idea of biological warfare, however, can be traced back to the 6th century B.C. and the methods used have been quite ingenious and diverse:[14,15]

6^{th} century B.C. The earliest known uses of biological weapons occurred in the 6^{th} century B.C. when the Assyrians poisoned enemy wells with a parasitic fungus called rye ergot. In the same era the Athenians under king Solon contaminated the water in an aqueduct leading water to the besieged city of Cirrha using black hellebore roots.

1346 In 1346 the Tartars were laying siege to the Genoese city of Kaffa. During the siege there was an outbreak of bubonic plague in the Tartar army. As the plague weakened the Tartars they resorted to biological warfare, catapulting the bodies of soldiers that had succumbed to the illness over the walls of the city. This tactic worked, weakening the city so much that it had to surrender. Before the surrender many of the Genoese fled to Italy, using ships that were believed to be clean of the disease. The precautions were inadequate and the fleeing Genoese brought the plague with them to mainland Italy. This is believed, by some medical historians, to be the starting point of the Black Death epidemic that swept over Europe during the 14^{th} and 15^{th} century, wiping out more than 25 million people.

1763 During the French and Indian war in the 18^{th} century the British were suffering from smallpox when Sir Jeffrey Amherst, commander of the British forces had the idea to use this to the advantage of the British. As a token of friendship, the tribes that were friendly to the French were sent gifts of blankets and handkerchiefs that were infected with the smallpox virus. The captain who handed over the gifts, Simon Ecuyer, later wrote: “I hope it will have the desired effect”. His wish was granted, a few months later the disease was killing the Native Americans in large numbers.

1863 During the American Civil War, biological weapons were used on at least two different occasions. The confederates were responsible for both of these. While retreating from Vicksburg, the troops under Gen. Johnston contaminated wells and ponds with animal carcasses. Furthermore, a Dr. Luke Blackburn, who later became governor of Kentucky, sold clothing infected with smallpox and yellow fever to Union troops.

1915 The discovery in 1870 that microorganisms were the causative agents of disease enabled more sophisticated biological warfare. The German used this new knowledge during WWI, when a German-American physician living in Washington, Dr. Anton Dilger, grew cultures of *Bacillus anthracis* and *Pseudomonas mallei.* The cultures were given to German agents working as dockworkers in Baltimore and they used them to infect livestock that was to be shipped to the Allied troops in Europe. These actions caused several hundred troops to be infected with anthrax and glanders.

1937-1945 The Japanese constructed a biological weapons laboratory, named unit 731, in 1937 and used the achievements from this facility to attack China during World War II, when they, among other things, dropped plague-infested fleas mixed with rice on China. The rice attracted rats on which the fleas could live and when the rats came in contact with people the fleas spread the disease to them, resulting in the deaths of up to 10,000 people. When the laboratory was destroyed in 1945 it was estimated that as many as 200,000 people - Chinese soldiers, private citizens and POW's - had died from the experiments in unit 731.

1979 In April 1979 Soviet citizens in the vicinity of Sverdlovsk began experiencing symptoms of poisoning. The Soviet Ministry of Health claimed that the outbreak was a case of contaminated meat but autopsies showed that the 66 people who died where infected with four different strains of anthrax in the lungs and lymph nodes. The most credible explanation is that a cloud of aerosolised *Bacillus anthracis* spores was released from a nearby research facility, indicating that the Soviet Union was developing biological weapons in violation of the Biological Weapons Convention.

1984 In 1984 the first incident of bioterrorism in the USA occurred in Wasco County, Oregon. The Rajneeshee cult were planning to take political control over the county and they were planning to do so by incapacitating a large proportion of the electorate using salmonella. In a test run cult members sprinkled salmonella on salad bars in 10 restaurants in The Dalles, causing more than 750 people to become seriously ill. The outbreak attracted the attention of the authorities and large groups of health officials and investigators swarmed the area, effectively inhibiting the group from carrying out their sinister plans on election day.

2001 In 2001, in the wake of the terrorist attacks on the World Trade Center and the Pentagon, several news media companies and government officials in the US received letters containing high-grade, finely textured anthrax. 22 people contracted the disease from inhaling the anthrax bacteria and 5 people died subsequently.

1.1.2 Modern bioweapons and bioterrorism

While the history of biological warfare is ancient, dating back more than 2,500 years, the efficiency of biological weapons took a major leap during the previous century. The identification of bacteria and viruses as the causative agents of diseases gave new insights and the further elucidation of the mechanisms behind diseases and their spread, leading to the emergence of modern microbiology, enabled vast improvements to biological weapons. During the cold war most major countries funded projects researching biological weapons, projects that lead to several important discoveries and developments to increase the effectiveness and ease of use of biological weapons; the agents themselves were improved as well as the technological means for spreading them.

The development of genetics opened yet another new door since it allows for conscious, directed modifications to the agents so as to give them desired features. While the first attempts to modify the agents were perhaps directed at transferring resistance to antibiotics and vaccines, future modifications may be much more sinister. As knowledge of molecular biology grows, so does the amount of possible modifications to the agents of biological weapons.

In the Biological Weapons Convention, which was signed in April 1972 and entered into force in March 1975[1], all parties agreed to discontinue their research into biological weapons and to dismantle their existing arsenals. Even though these tasks were said to have been carried out, there is a fear, fueled by such incidents as the one in Sverdlovsk, that e.g., the USSR did not entirely dismantle their program and that there might exist "rogue" labs and disgruntled scientists in Russia that could spread knowledge of and equipment for biological warfare to terrorists and rogue states.

The Centers for Disease Control and Prevention (CDC) maintains a three-tiered list of organisms (included in table 1.1) that are monitored closely as they are believed to be likely candidates for bioterrorism. The first category, 'category A', contains the agents that are considered to be the most potent bioterror weapons. The criteria are that they[5]

- can be easily disseminated or transmitted from person to person;
- result in high mortality rates and have the potential for major public health impact;
- might cause public panic and social disruption; and
- require special action for public health preparedness.

Category A	Category B
Anthrax (*Bacillus anthracis*)	Brucellosis (*Brucella* species)
Botulism(*Clostridium botulinum* toxin)	Epsilon toxin of *Clostidrium perfringens*
Plague (*Yersinia pestis*)	Food safety threats, such as: *Salmonella species*, *Escherichia coli O157:H7*, *Shigella*
Smallpox (variola major)	
Tularemia (*Francisella tularensis*)	
Viral hemorrhagic fevers (filoviruses [e.g., Ebola, Marburg], arenaviruses [e.g., Lassa, Machupo])	
	Glanders (*Burkholderia mallei*)
	Melioidosis (*Burkholderia pseudomallei*)
	Psittacosis (*Chlamydia psittaci*)
	Q fever(*Coxiella burnetii*)
	Ricin toxin from *Ricinus communis*
	Staphylococcal enterotoxin B
	Typhus fever (*Rickettsia prowazekii*)
	Viral encephalitis (alphaviruses [e.g., Venezuelan equine encephalitis Eastern equine encephalitis Western equine encephalitis])
	Water safety threats (e.g., *Vibrio cholerae*, *Cryptosporidium parvum*)

Table 1.1: CDC Categories for bioterrorism agents

To be included in the list of agents considered second highest priority ('category B'), agents must fulfill criteria that are similar to those for category A but less severe.

There is also a third group (category 'C') that includes “emerging pathogens that could be engineered for mass dissemination in the future” and which includes i.e., Nipah virus and hantavirus.

1.2 *Francisella tularensis*

Until recently, the complete genome sequences of all the CDC category A agents were known except for that of *Francisella tularensis.* This in spite of the fact that this organism has been known since 1912[13] and even though it is one of the most virulent bacteria known to man.

1.2.1 Biology and epidemiology

The genus *Francisella*, placed in the γ subdivision of the proteobacteria, is the only genus of the family *Francisellaceae* and it contains two species: *F. tularensis* and *F. philomiragia*. *F. philomiragia* is an opportunistic pathogen that rarely causes disease in humans.

F. tularensis is a small (0.2 μm × 0.2-0.7μm)[4], rod-shaped or coccoid, faintly staining, strictly aerobic, Gram-negative bacterium[16]. It is a nutritionally fastidious, facultative intracellular parasite and is the causative agent of the disease tularemia. *Francisella tularensis* is spread throughout most of the northern hemisphere within the range 30 ° to 71 ° latitude[16]. There are four recognized subspecies to the species, two of which are most commonly isolated[17]: *F. tularensis* subsp. *holarctica*, which is the predominant subspecies in Europe and Asia, and *F. tularensis* subsp. *tularensis*, which is the predominant subspecies on the American continent but has also been found in Central Europe[10]. *F. tularensis* spp. *holarctica* is the less virulent of the two and it is rarely lethal in humans even though it can cause severe disease. *F. tularensis* spp. *tularensis*, on the other hand, is one of the most infectious pathogens known.

The natural reservoir of *F. tularensis* is as of yet not known. Outbreaks of tularemia occur at irregular intervals, the period of which have been shown to have a relationship with the peaks in vole and hare populations[18]. This has lead to the suspicion that *F. tularensis* is maintained in the environment mainly in rodents[7], a theory that is challenged by the fact that rodents and lagomorphs generally do not survive the infection. It has been shown that *F. tularensis* can survive for extended periods (months) in water and mud and that the geographic distribution of the outbreaks of tularemia are clearly related to the incidence of natural water in the vicinity. Taken together with the fact that as many as 53 % of beavers have antibodies to the organism, the evidence suggest that water is somehow an important part of the lifecycle of *F. tularensis*. Possibly some water-borne non-mammalian host cell, similar to the protozoan reservoir for *Legionella pneumophila*, may harbor the organism between outbreaks, allowing it to replicate intracellularly[16].

The most common route of infection with *F. tularensis* is a bite of a blood-feeding arthropod vector (mainly ticks or mosquitoes) which has previously fed on an infected animal. This route of infection leads to the disease known as ulceroglandular tularemia, characterized by a skin ulcer and swollen, tender lymph nodes. Hunters and trappers handling infected meat while having cuts or abrasions on the hands can also contract the disease in this form. The infective dose of *F. tularensis tularensis* in humans when the bacteria are injected under the skin is extremely low, requiring as few as 10 bacteria[16].

Another route of infection is through the inhalation of an aerosol containing bacteria. This is most common among agricultural workers, especially during the handling of hay where the disruption of carcasses or sites of residence of rodents may result in the release of such an aerosol. The pneumonia-like disease resulting from this route of infection is known as respiratory tularemia and this route also requires a very low dose to be infective in humans, as few as 25 bacteria[16] are needed. Both these most common routes of infection put rural populations at a higher risk for infection. Members of the rural population tend to spend more time outdoors, increasing the exposure to ticks and mosquitoes, and they also come into contact with the carriers of the disease, such as rodents, more often than do members of the urban population.

Oropharyngeal or gastrointestinal tularemia occur when infected food or water is ingested. Oropharyngeal tularemia often presents enlarged tonsils and a painful sore throat while the effects of gastrointestinal tularemia can range from a mild diarrhea to a fatal disease with extensive ulceration of the bowel. A special case of ulceroglandular tularemia is oculoglandular tularemia, with the conjunctiva as the initial site of infection. This form of tularemia is probably often contracted by scratching the eyes while having bacteria on the fingertips and it presents ulcers and nodules on the conjunctiva.

Tularemia is treatable using antibiotics. The aminoglycosides streptomycin and gentamicin are the most effective and give a low risk of relapse. Tetracycline and chloramphenicol are also often used but there is a much higher risk of relapse when using these drugs[16, 7] .

Several attempts have been made to produce an effective vaccine against tularemia. In the 1930s-1940s the Soviet Union performed a large number of vaccinations using an LVS that had been attenuated by repeated growth of fully virulent strains in media supplemented with antiserum and by drying the strains. This LVS, called Strain Moscow, was demonstrated to have weakened virulence and high immunogenicity and was used to vaccinate several thousand people before it was lost.[7] Live vaccines have been produced later as well, for instance a live attenuated vaccine was routinely used in the USA by laboratory workers working with the organism[4]. This vaccine was not, however, fully effective. Work on developing new vaccines is ongoing but at present there is no publicly available vaccine.

1.2.2 Genetics

The strain of *F. tularensis* examined in this study is the strain SCHU S4. It is a fully virulent strain of the *Francisella tularensis* subspecies *tularensis* and has been proposed as the type strain for this subspecies. SCHU S4 was isolated from a human ulcer found on a tularemia patient in Ohio in 1941[6].

The genome of SCHU S4 consists of a single, circular chromosome with a length of 1,892,819 bp. The genome contains 1,804 predicted coding sequences, including 201 sequences predicted as pseudogenes or gene fragments.

1,281 genes in the *Francisella tularensis* SCHU S4 genome have homologues ($E < 1 \cdot 10^{-10}$) in at least one other γ-proteobacterial genome. These genes are randomly distributed around the genome with the exception of two regions outlined below.

The overall G+C content of the *F. tularensis* SCHU S4 genome is 32.9% which falls within the range of typical G+C content found in small (0.9 – 2.0 Mb) bacterial genomes (25 – 40 %). Two copies of a completely duplicated region of 33,910 bp, found in the ranges 1,374,371 – 1,408,281 and 1,767,715 – 1,801,625, stand out from the general genome with a G+C content for the predicted coding sequences of 27.5 %. These regions also differ from the rest of the genome in that the predicted coding sequences have no homologs in any of 16 other γ-proteobacteria. In fact, the genes in these regions do not show significant homology with any other genes in GenBank, making it difficult to elucidate their origin[20].

1.2.3 Potential as a biological weapon

Francisella tularensis has been viewed as a potential biological weapon for a long time. It was studied by the Japanese and the Soviets and it is known that it was one of the biological weapons kept by the US military[4].

The extreme virulence of *F. tularensis* and the fact that it can infect humans via such divergent routes makes it very potent and dangerous as a biological weapon even though it is nutritionally fastidious and not easily transmitted between humans.

A study by the World Health Organisation in 1969 estimated that the release of 50 kg of dried *Francisella tularensis* over a metropolitan area with a population of 5 million would result in 250,000 incapacitating casualties including 19,000 deaths[21]. Another study, performed by the CDC, placed the potential death toll even higher. It predicts that an attack on a metropolitan suburb with a population of 100,000 would result in 82,500 cases of tularemia, resulting in 6,188 deaths. This report estimates that such an aerosol attack using *F. tularensis* would cause total base costs to society of \$3.9 bn – \$5.4 bn, using low estimates for costs[12].

In short, *F. tularensis* is a possible biological weapon with an enormous destructive potential. A terrorist attack using *F. tularensis* would cause very much suffering and a large number of deaths, producing massive economic losses to society. Clearly a greater understanding of *F. tularensis* would be beneficial, especially if it could shed some light on the mechanisms behind the extreme virulence of the organism.

1.3 Bioinformatics

1.3.1 Phylogenetics

Phylogenies and phylogenetic trees

Phylogeny is the evolutionary history of a species or group of related species. More specifically phylogenies are patterns of shared history between biological replicators, such as species or genes. *Phylogenetic inference* aims at analyzing the phylogeny and presenting a well-corroborated hypothesis of this shared history.

The most common way to represent the results of a phylogenetic analysis (phylogenetic inference) is in a *phylogenetic tree*. A tree in itself is a mathematical representation consisting of nodes and branches (figure 1.1). A node with a degree (the number of adjacent branches attached to the node) greater than three is called a polytomy and a tree containing such a node is called a polytomous tree. A tree containing no polytomies is called a fully resolved or dichotomous tree.

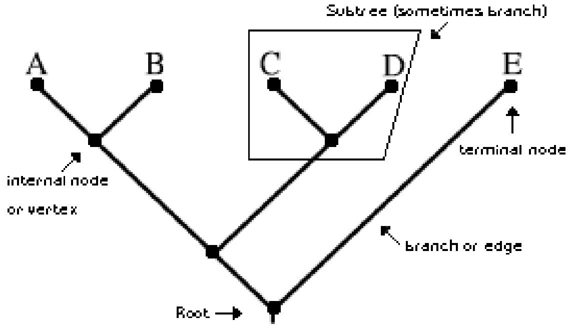

Figure 1.1: Example of a tree

Since a node represents a divergence, there are two possibilities for how a polytomy might occur. A polytomy might occur due to (i) simultaneous divergence, all the descendants evolved at the same time in history (this is called a “hard polytomy”) or (ii) the true history of divergence cannot be resolved (a “soft polytomy”). A common case is a soft polytomy that is introduced because the divergence happened so rapidly that it is not possible to reconstruct the history exactly. Such a soft polytomy is effectively a hard polytomy.

Most of the tools used to construct trees produce dichotomous trees.

The "simplest" form of phylogenetic tree is the *cladogram* (figure 1.2).

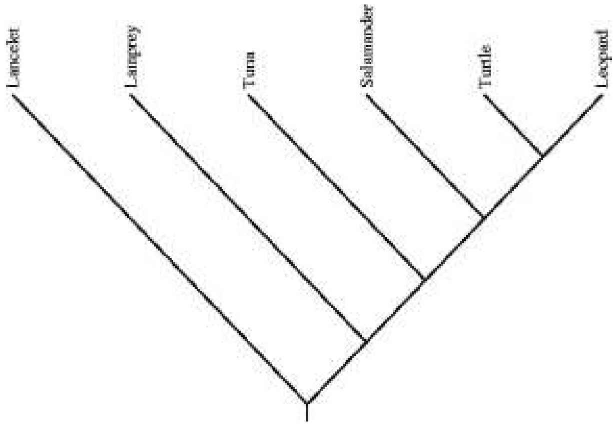

Figure 1.2: Cladogram example

The cladogram depicts the relationships between the terminal nodes but does not contain any other information.

The tree has a direction, meaning that the order in which the nodes occur in the phylogenetic tree represents the temporal order in which the different variations of the biological replicators occurred. In other words, the sequence of branching symbolizes historical chronology. If we're talking about species, this means that if the branch point at which two species diverged comes earlier in the tree, the speciation event that separated these two species occurred earlier in time. This also means that these two species are more distantly related. In Figure 1.2, the fact that the branch leading to Tuna diverges earlier than the branch leading to Salamander means that the last common ancestor to Leopard and Tuna lived longer ago than the last common ancestor to Leopard and Salamander. The cladogram does not, however, contain any information about the exact quantities of time or change that separate the nodes.

For a tree to have direction, it must be rooted. Many of the common tree-building tools used produce unrooted trees, which must then be rooted. The most common way to root an unrooted tree is to include an *outgroup*. An outgroup is a species which is known to have a more distant relation to any of the "ingroups" than any of the "ingroups" have to each other. This means that the root can be placed on the branch leading to the outgroup, as illustrated by figure 1.3, where 'O' represents the outgroup.

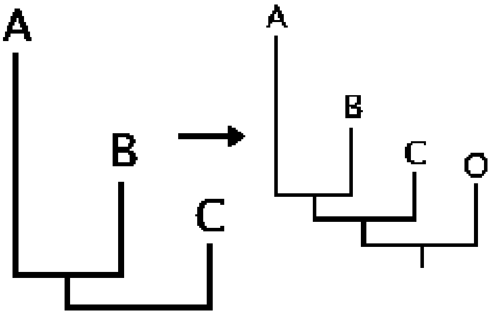

Figure 1.3: Including an outgroup to root an unrooted tree

For a tree containing additional information we turn to the additive tree

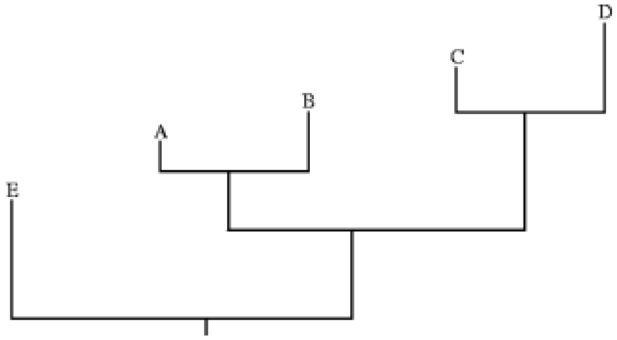

Figure 1.4: Phylogram example

or *phylogram* (figure 1.4). A phylogram is drawn so that the lengths of the branches are proportional to some cumulative variable that is of interest (such as change).

The final category of phylogenetic trees is the ultrametric tree or dendrogram (figure 1.5) in which all the terminal nodes are equidistant from the root of the tree. This means that the tree can depict i.e., evolutionary time from a certain point in time forward until today. (as in the example) The length of each branch then represents time.

For further analysis of a tree it is interesting to study the character states of the nodes, as illustrated in figure 1.6. A node that shares the same base as the common ancestor of all the sequences being studied is said to be a *plesiomorphic* or ancestral (primitive) state, otherwise it is said to be a *apomorphic* or derived state. If a derived character state is unique in the set it is an *autapomorphy*, shared derived states are *synapomorphies*. A shared de-

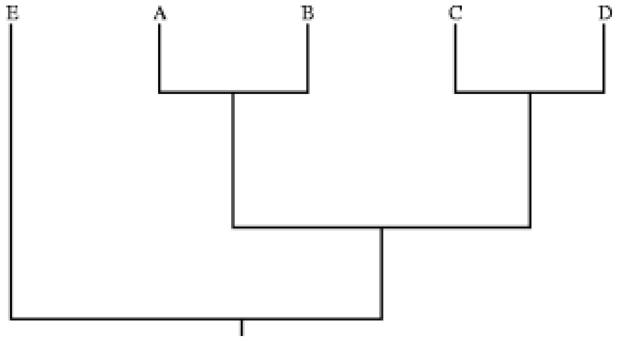

Figure 1.5: Dendrogram example

rived state may be inherited directly from a common ancestor, exemplifying *homology*, or it may have occurred due to independent evolution, in which case it is called *homoplasy.* In the figure the synapomorphies also illustrate homology.

While performing phylogenetic analyzes on molecular sequences, the sequences studied must be homologous to convey information about the phylogeny. Homologous sequences can be further subdivided into *orthologus* and *paralogous* sequences, as illustrated in figure 1.7. Two homologous genes are orthologous if their most recent common ancestor did not undergo a gene duplication, paralogous genes include the duplication.

Analyzes of genes must include only orthologous sets, since including paralogues in the analyzes may would give incorrect results. Considering the figure, if only sequences 1,3 and 5 were part of the analysis, the resulting tree would be ((A,C),B), a result that is incorrect.

Methods for constructing phylogenetic trees

A number of different methods are used for constructing phylogenetic trees from molecular data, methods which take different approaches to evaluating the available data. They can be divided into four different categories, according to two criteria.

The first criterion is whether the method is a distance method or a discrete method. Distance methods use pairwise distance matrices, which are constructed from the aligned sequences, whereas discrete methods use the sequences directly. Distance methods have the disadvantage that converting the aligned sequences to a distance matrix is “destructive”, i.e., some information is lost. Constructing the distance matrix can be a time-consuming task and there are several different methods for doing it. The simplest

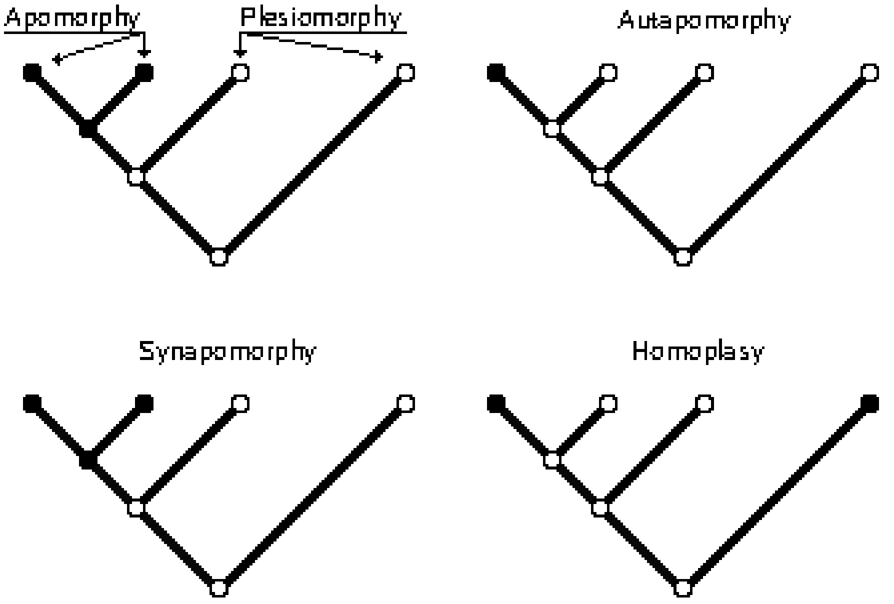

Figure 1.6: Patterns of ancestral and derived character states

method for constructing the distance matrices is to simply count the number of amino acid differences between the sequences. This approach can yield relatively good results, especially if the proteins under study diverged recently or have been highly conserved. If these conditions do not apply, the distance measured by direct comparison will be smaller than the actual distance, due to multiple substitutions at the same site. This underestimation of distance will increase with increasing evolutionary distance between the species compared and to compensate for it statistical methods can be used. Another aspect that is missed by both the direct method and the statistical methods is the empirical observation[2] that amino acid substitutions are not entirely random. Substitutions are more likely to occur between similar amino acids (e.g., in polarity or size) and certain amino acids are only substituted very rarely. To incorporate this aspect into the construction of distance matrices Dayhoff et. al. invented a new method wherein an amino acid substitution matrix is used[3]. This matrix contains data from empirical observations and is used to predict the probability of each amino acid substitution. The Jones-Taylor-Thornton[11] matrix, which was used in this work is very similar to the Dayhoff matrix but is based on a much larger set of empirical data and so is likely to yield better results.

The second criterion is whether the method is a clustering method or a

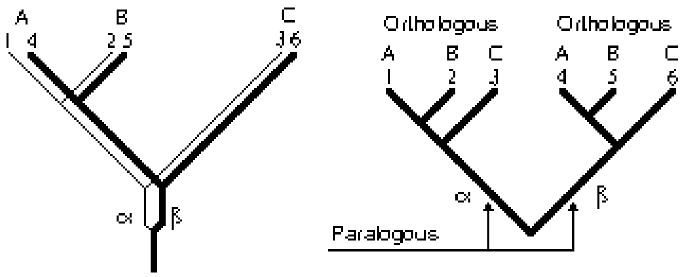

Figure 1.7: Orthologous and Paralogous replicators

search method. A clustering method uses an algorithm to construct the tree by starting at some point and then adding branches at the appropriate places according to the algorithm. Search methods use optimality criteria to choose the best tree among the set of all possible trees. The biggest advantages of clustering methods is that they are fast and that they always produce a single tree. The disadvantages are that the result may vary with the order of input and that they do nothing more than produce a tree, they do not allow for evaluation of competing hypotheses. For search algorithms the major disadvantage is speed. There are two major problems for search algorithms: what is the value of the optimality criterion for each tree, and which of all the possible trees has the maximum value of this criterion? For small numbers of sequences it is often possible to find the optimal tree but for larger sets it will not be possible. The solution is called *heuristic*, "quick and dirty" methods. One such method is "hill-climbing", which starts with some tree and goes on to rearrange it, maintaining any rearrangement that produces a better tree. A major problem with hill-climbing is that it can easily get stuck on local maxima. It is simply not certain that the resulting tree is the best of all possible.

The most common methods for constructing evolutionary trees are UPGMA, Neighbor joining (NJ), Minimum evolution (ME), Maximum parsimony (MP) and Maximum likelihood (ML). UPGMA and NJ are distance methods using clustering algorithms; ME is a distance method using an optimality criterion; MP and ML are discrete methods using an optimality criterion.

While the impact of the method on the accuracy of the phylogeny is naturally large, another possible source of error is, of course, the data itself. If the data is incomplete or contains sampling errors, the quality of the method does not matter. In fact, the method could be perfect and give the exact correct result for the available data and the results would still be erroneous.

For sampling error, a good way to get a grip on its size would be to take multiple samples. However, this approach is often too expensive to be feasible. Calculating confidence levels for a tree is a very difficult and often impossible task. Instead confidence in a phylogenetic hypothesis is often calculated using bootstrapping. Bootstrapping performs re-sampling from the existing sample, producing pseudoreplicates. A pseudoreplicate is produced by sampling at random and with replacement from the original data set until the new data set contains as many sites as the original one. Sampling with replacement means that each site is "returned" to the data set before the next sample is taken. This means that certain sites in the original data set may occur more than once in the new data set, while some other sites might not occur at all. Still, the new data set will only contain sites found in the original data set. Normally 100 to 1,000 pseudoreplicates are produced in the bootstrapping and the pseudoreplicates are then used to produce trees. Normally these trees are then combined into a consensus tree, showing the support for each node. As mentioned above, bootstrapping is a method for calculating precision, not accuracy.

Phylogenomics

While phylogenies and phylogenetic trees can be constructed from observed phenotypical differences, this work focuses on phylogenies constructed from molecular data. Constructing phylogenetic trees using this kind of data was, for a long time a very arduous task, which took immense amounts of time. Moores law and optimized algorithms have changed this very much. Simpler tasks, such as bootstrapping with 1,000 replicates or constructing a single NJ-tree can be completed in mere seconds and more calculation-intense tasks, such as calculating distance matrices or constructing trees with optimality criteria do not take more than hours.

All this computational power can be put to good use by processing large amounts of data, e.g., for doing a phylogenetic analysis of the entire genome of a species: a phylogenomic analysis. The goal of this work was to attempt a phylogenomic analysis of the genome of *Francisella tularensis*.

Materials and Methods

2.1 Genetic materials

The research group at the department participated in an international consortium on the sequencing of the *F. tularensis* genome.

As mentioned, the genome of *Francisella tularensis* contains 1805 putative genes, including pseudogenes and gene fragments, 1281 of which have homologues (BLAST E-value $< 1 \cdot 10^{-10}$) in one or more γ-proteobacterial genomes.

To perform a study of the phylome, i.e., the total set of trees generated from the complete set of genes, of *F. tularensis*, a database was constructed, consisting of the protein sequences encoded by all the genes in the genomes of a group of 16 γ-proteobacteria (shown in table 2.1).

No.	Name
1	*Escherichia coli* K12
2	*Haemophilus influenzae*
3	*Pasteurella multocida*
4	*Pseudomonas aeruginosa*
5	*Salmonella enterica* serovar Typhi
6	*Salmonella enterica* serovar Typhimurium LT2
7	*Vibrio cholerae*
8	*Xanthomonas axonopodis*
9	*Xanthomonas campestris*
10	*Xylella fastidiosa*
11	*Yersinia pestis*
12	*Shewanella oneidensis*
13	*Shigella flexneri* 2a
14	*Coxiella burnetii* RSA 493
15	*Bacillus anthracis* Ames
16	*Legionella pneumophila*

Table 2.1: The organisms included in the survey.

The 1805 putative proteins found in the genome of *Francisella tularensis* were run against all the proteins in the database using BLAST* without using any E-value threshold. The results of these runs, including the scores and E-values, were inserted into a new database, indexed by the corresponding *F. tularensis*-protein. After the completion of this search, all of the entries in the new database were searched for the proteins of *F. tularensis* that had homologues in all 16 of the γ-proteobacteria in table 2.1 and where the BLAST E-value threshold was $1 \cdot 10^{-10}$. These positives were further evaluated. To establish that the results represented homologues, the hits were further analyzed; e.g., the annotations of the genomes where searched and those proteins that had the same annotations were retained. Furthermore, certain complex proteins, such as transporters, were excluded for clarity.

2.2 Methods

The phylogenetic analyzes were performed using two slightly different methods. Overall, the goal was to produce a collection of trees representing the phylogenies of the different homologous genes found through the searches described above.

For the construction of the phylogenetic trees, two main methods were used: NJ and MP. As described in the phylogenetics section above, a key step in the construction of NJ trees is the construction of distance matrices later used to construct the trees. One of the analyzes performed did not use a specific method for computing these distances but relied on the built-in distance calculation of one of the programs used. The two methods are described in detail below.

Method 1

The sequences in each group were aligned using ClustalW, version 1.83[19] to produce multiple alignments. The alignments were performed using the GONNET weight matrix for both pairwise and multiple alignments. For the pairwise alignments the penalty for opening a gap was set to 35 and for extending a gap the penalty was set to 0.75; for the multiple alignments the corresponding penalties were 15 for opening a gap and 0.3 for extending it.

Each alignment was bootstrapped using *seqboot* from the PHYLIP package[8], producing 1,000 pseudoreplicates.

The NJ trees were constructed using the tree constructing capabilities of ClustalW, meaning that the trees were constructed using uncorrected,

[*] Version 2.2.8 was used throughout.

observed distances.

As a further analysis, the above analysis was repeated on a concatenated alignment of 10 proteins. The proteins selected were *dnaA*, *ftsA*, *mfd*, *mraY*, *murB*, *murC*, *parC*, *recA*, *recG* and *rpoC*. In addition to the Neighbor Joining tree a Maximum Parsimony tree was also constructed for the concatenated alignment. In the construction of the MP tree, the same pseudoreplicates were used as in the construction of the NJ tree and the tree construction was performed using the *protpars* program from the PHYLIP package.

Method 2

Alignments and bootstrapping were performed using the same methods as described above.

Subsequent phylogenetic analyzes were performed using the PHYLIP package. Distance calculations were performed using *protdist*, using the Jones-Taylor-Thornton matrix as the model for the distance calculations. After the distance calculations, Neighbor Joining trees were constructed for all the datasets using *neighbor* and finally the resulting set of trees were processed using *consense* to produce the consense neighbor joining trees for each of the datasets.

The above analysis of the concatenated alignment of 10 proteins was performed again using the same methods as in the construction of the individual trees.

Results

The primary database search yielded 405 sets of homologous proteins, each set consisting of one protein from *F. tularensis* and at least one for each of

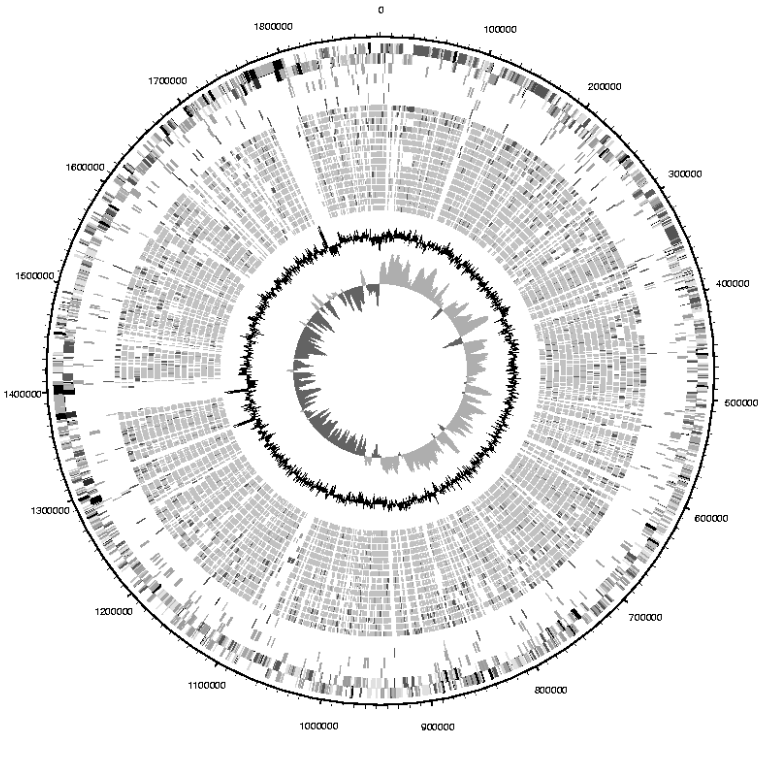

Figure 3.1: Blast searches (The illustration was used with permission from Hans-Henrik Fuxelius at the Department of Molecular Evolution, Uppsala universitet)

the other organisms. In figure 3.1, circles 7 through 22 show the blast search hits for, in order, *L. pneumophila*, *P. aeruginosa*, *V. cholerae*, *C. burnetii* RSA 493, *B. anthracis* Ames, *S. oneidensis*, *E. coli* K12, *H. influenzae*, *P. multocida*, *S. enterica* serovar Typhi, *S. enterica* serovar Typhimurium LT2, *X. axonopodis*, *X. campestris*, *Y. pestis*, *S. flexneri* 2a and *X. fastidiosa.*

A number of these sets contained paralogous rather than orthologous proteins and were excluded. Also, certain proteins were e.g., parts of protein complexes, such as transporters. These proteins were excluded because of the difficulty of getting clear results when analyzing them.

After these exclusions, the end result consisted of 210 sets of homologues, which were then subjected to phylogenetic analyzes as described in section 2.2, resulting in 210 separate trees, exhibiting several different topologies.

Though two different methods were used in constructing the trees, these two methods yielded virtually the same results.

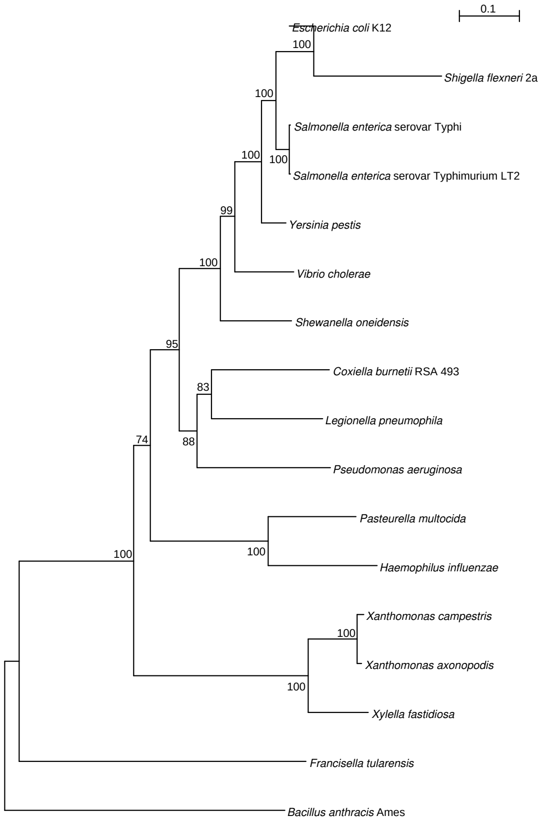

Figure 3.2: *dnaA*

The most common topology, represented in 40% of the trees, displayed *Francisella tularensis* as a separate outgroup to all the other species included in the analysis (as exemplified in figure 3.2, depicting the tree constructed for the protein *dnaA*). 56 % of the cases with *F. tularensis* as an outgroup to all the other species were well supported individually, having a bootstrap support exceeding 75 %.

The fact that the largest distinct group of phylogenies shows *F. tularensis* as the most deeply diverging lineage lends support to the theory that despite the similarities in lifestyle and genomic size, *Francisella tularensis*, *Coxiella burnetii* and *Legionella pneumophila* are not sister clades. These results make it more probable that the similarities shared between these organisms have developed through independent, convergent evolution. The phylogenetic analysis of the ten concatenated alignments (as described in section 2.2 yielded the tree included in figure 3.3. This alignment provides further support for the theory that *Francisella tularensis* is not closely related to any of the other γ-proteobacteria.

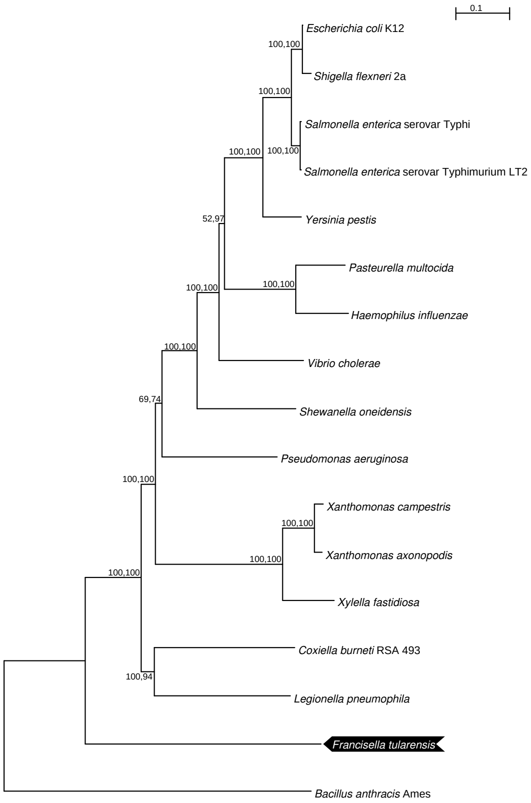

Figure 3.3: The tree of the concatenation of the genes

Most of the trees with high confidence levels displayed the configuration with *F. tularensis* as outgroup to all other organisms but there were a few exceptions. For instance the tree of the protein *130*, as displayed in figure 3.4, places a group containing *Xanthomonas axonopodis*, *Xanthomonas*

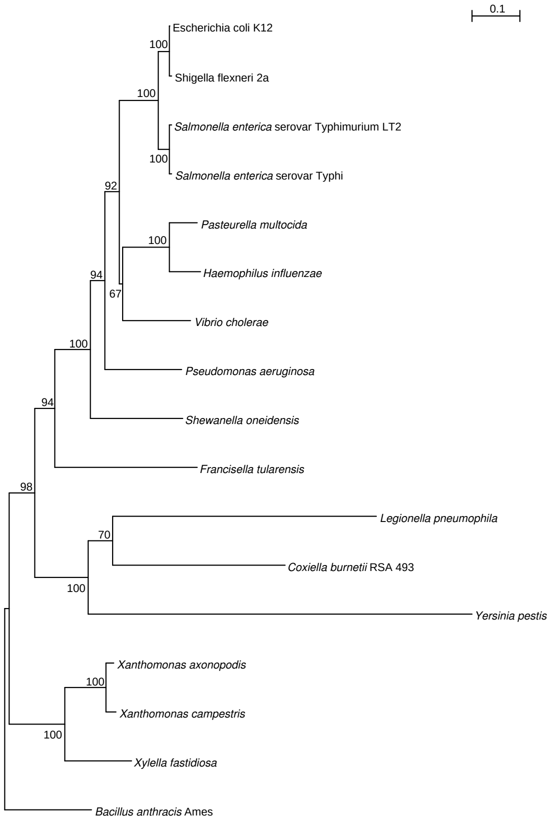

Figure 3.4: *130*

campestris and *Xylella fastidiosa* as the ”‘outmost”’ specimen. This group is maintained with complete confidence throughout the entire set of trees and can reasonably be treated as a specimien in itself.

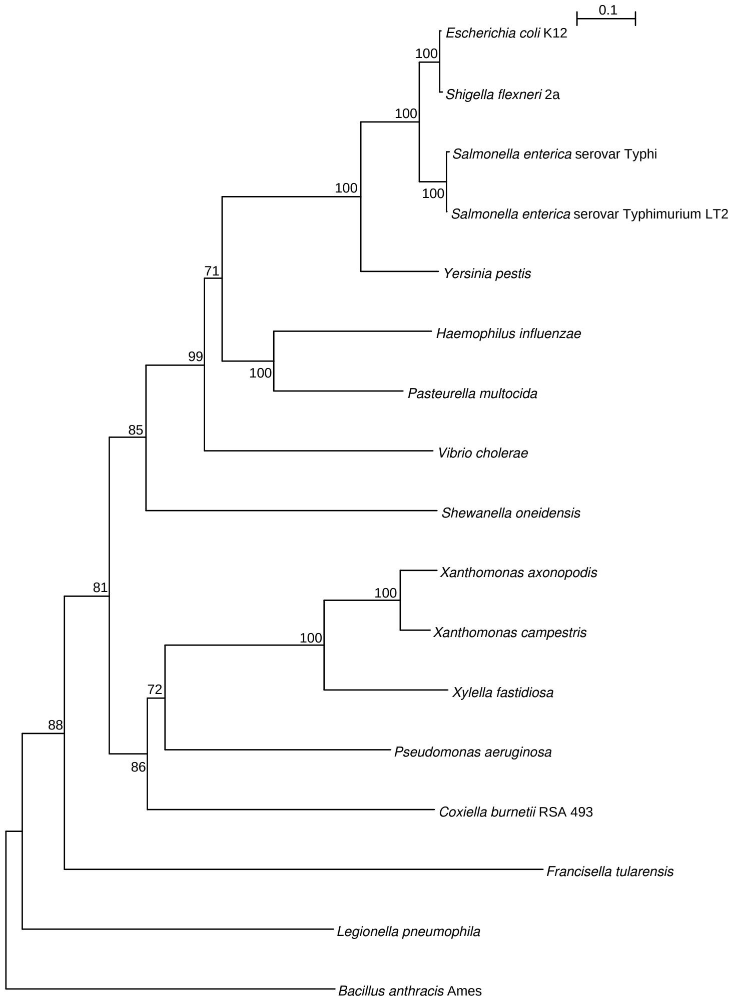

Figure 3.5: *murE*

Figure 3.5 (*murE*) shows a high confidence-tree where *Legionella pneumophila* is placed outside of *Francisella tularensis*, while figure 3.6 (*mutL*) shows *Coxiella burnetii* outside *F. tularensis*.

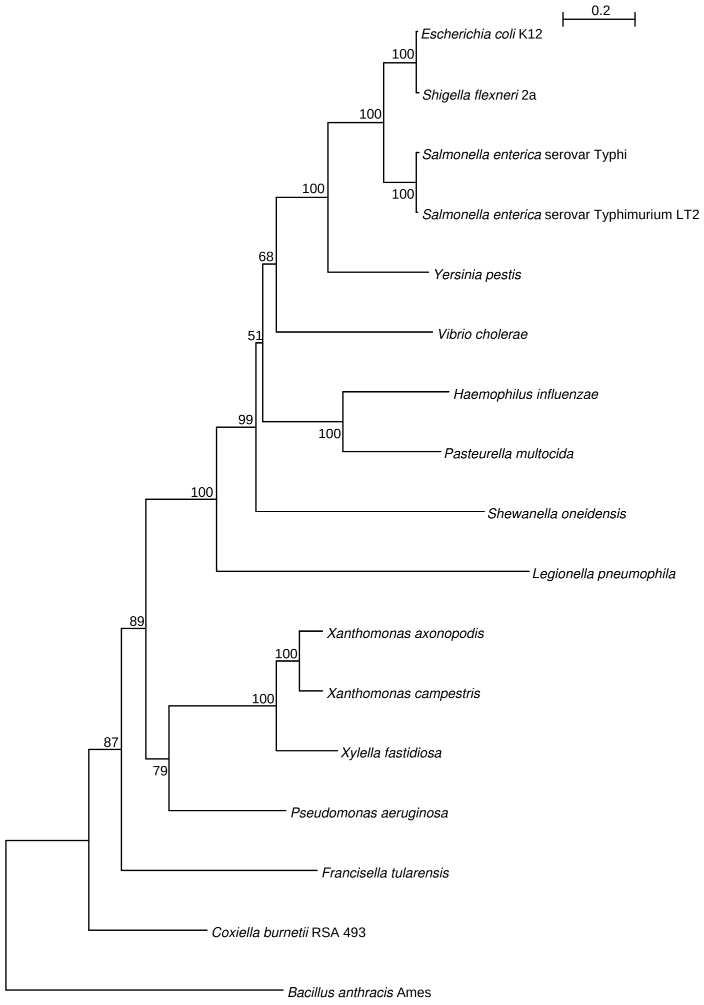

Figure 3.6: *mutL*

In figure 3.7 (*lipA*), *F. tularensis* and *L. pneumophila* are grouped together and their group is placed as an outgroup to all other organisms. This pattern is repeated in figure 3.8 (*murB*) but with *F. tularensis* grouped with *C. burnetii* and also in figure 3.9 (*leuS*), where all three organisms are grouped together and positioned as an outgroup to all other organisms.

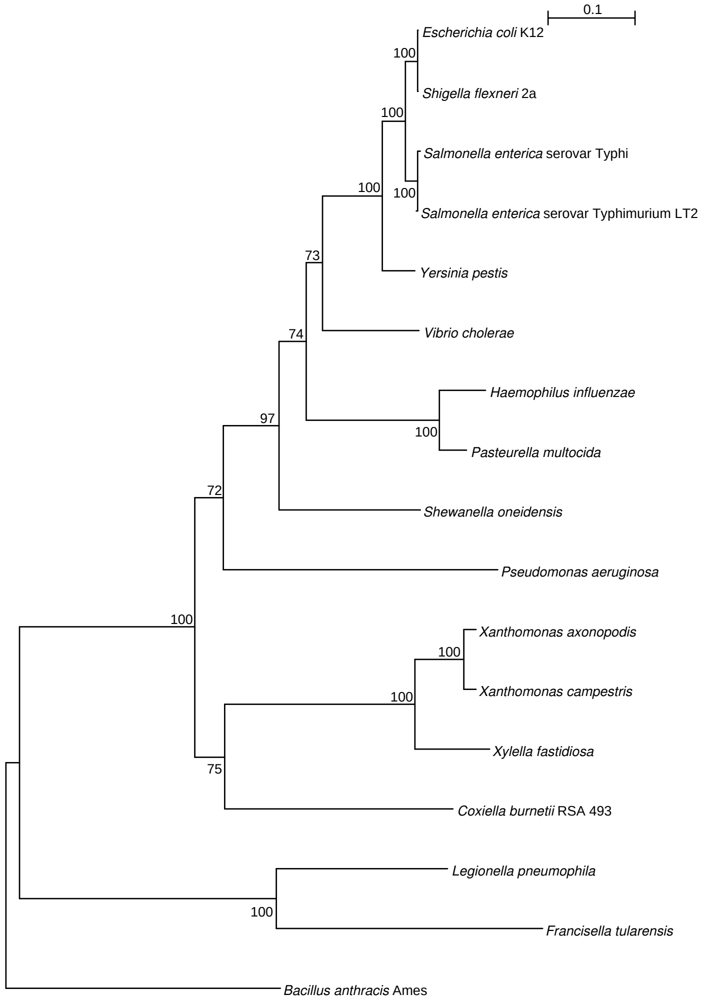

Figure 3.7: *lipA*

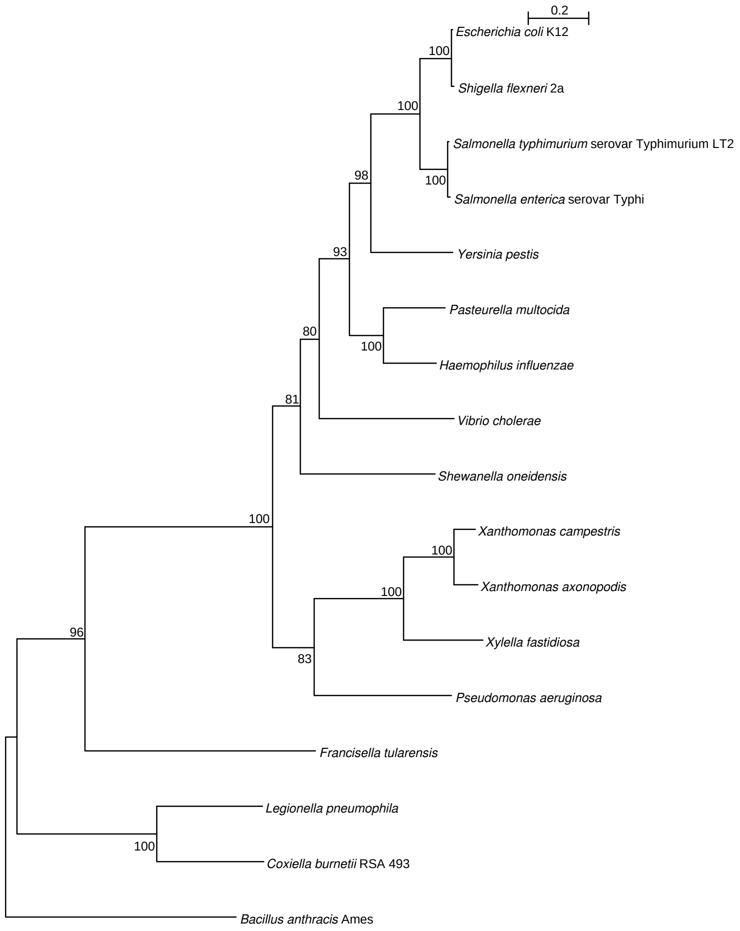

Figure 3.8: *murB*

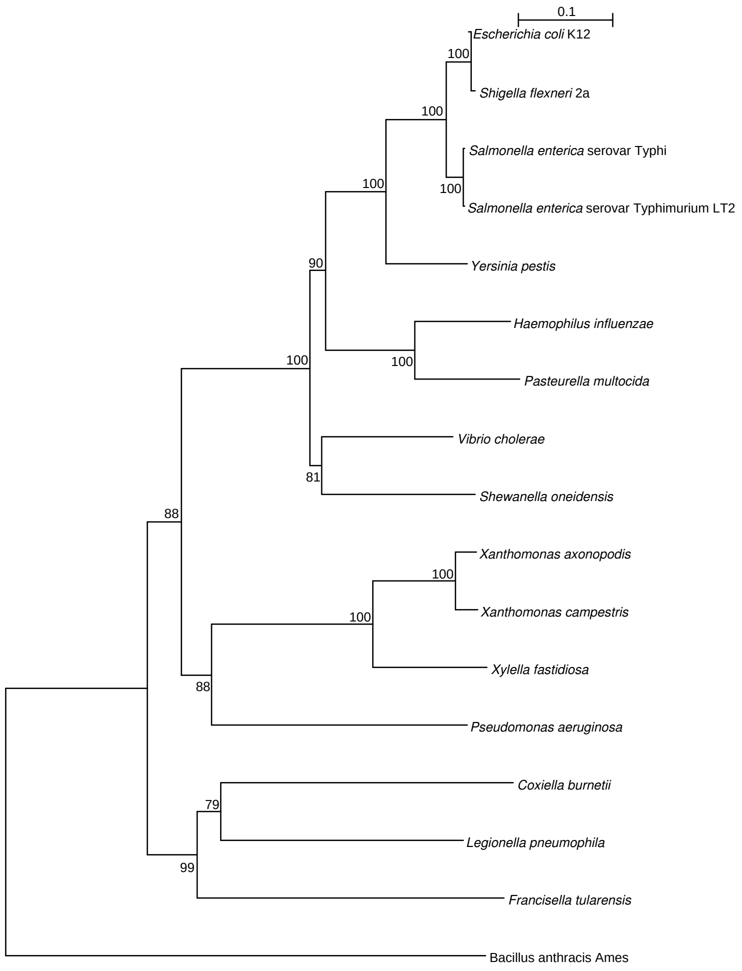

Figure 3.9: *leuS*

Discussion

This work was part of a larger effort to sequence and analyze the genome of *Francisella tularensis*, the results of which were published in Nature Genetics[20]. The results of the analysis indicate that "The genome sequence of *F. tularensis* SCHU S4 shows extensive inactivation of genes and a duplicated region that is strongly implicated in virulence and may be a pathogenicity island. The origins of the pathogenicity islands are not known, and the function of the genes in this region cannot be inferred on the basis of sequence homology with gene products of known functions. This finding raises the possibility that new mechanisms of virulence operate in *F. tularensis*"

The high level of gene activation, including a high proportion of disrupted metabolical or biosynthetic pathways, indicate that a lot of the evolution of the *F. tularensis* genome has occured through the loss of genetic information, although the pathogenicity islands may very well be acquired.

The results of the phylogenomic analysis indicates, as mentioned earlier, that while *Francisella tularensis*, *Legionella pneumophila* and *Coxiella burnetii* have similar lifestyles and genome sizes, they are not closely related genetically.

The results from the sequencing of the *F. tularensis tularensis* genome may in the future allow for better methods of diagnosing the infection; one major complication with tularemia is its intermittent occurrences, often with many years of dormancy in between. This causes hospital staff to be unfamiliar with the disease, which in turn can lead to problems in diagnosing the disease or even to misdiagnoses. The genome sequence may very well yield useful information that can lead to the development of better diagnositical tools. Also, the genome sequence could be of great help in the development of a functional vaccine. Both these developments would not only lead to less risk and better care for those who are infected with *F. tularensis* naturally, but would also lead to great reductions in the costs caused by a terrorist attack using the organism.

An interesting fact in this work is that both methods described in section 2.2 gave such similar results even though they differ quite markedly in their distance calculation procedures. The results from the direct distance calculations would, as mentioned in section 1.3.1, be expected to differ from those of the corrected distance calculations as the uncorrected distance calculations would tend to underestimate the distance between two sequences. On closer inspection it was discovered that the results did differ somewhat on a lower level. The results in topology or bootstrap supports for some of the individual proteins differed between the methods. Statistically, however, the results were inseparable. It seems likely these observed differences averaged out because of the large number of trees analyzed in the study. Thus my conclusion is that the results from the analysis are valid.

References

[1] Department of Peace Studies of the University of Bradford (-). Biological and Toxin Weapons Convention website. Published online (24 okt 2006). URL `http://www.opbw.org/`.

[2] M.O. Dayhoff (1972). *Atlas of protein sequence and structure.* National Biomedical Research Foundation, Silver Springs, Md.

[3] M.O. Dayhoff, R.M. Schwartz & B.C. Orcutt (1979). A model of evolutionary change in proteins. In *Atlas of protein sequence and structure, volume 5, supplement 3, 1978*, MO Dayhoff, editor, 345–352. National Biomedical Research Foundation, Silver Springs, Md.

[4] D.T. Dennis, T.V. Inglesby, D.A. Henderson, J.G. Bartlett, M.S. Ascher, E. Eitzen, A.D. Fine, A.M. Friedlander, J. Hauer, M. Layton, S.R. Lillibridge, J.E. McDade, M.T. Osterholm, T. O'Toole, G. Parker, T.M. Perl, P.K. Russell & K. Tonat (2001). Tularemia as a biological weapon: medical and public health management. *JAMA* **285**(21), 2763–2773.

[5] Centers for Disease Control & Prevention (2006). Bioterrorism Overview. Published online (24 okt 2006). URL `http://www.bt.cdc.gov/bioterrorism/overview.asp`.

[6] H.T. Eigelsbach, W. Braun & R.D. Herring (1951). Studies on the variation of Bacterium tularense. *J Bacteriol* **61**(5), 557–569.

[7] J. Ellis, P.C. Oyston, M. Green & R.W. Titball (2002). Tularemia. *Clin Microbiol Rev* **15**(4), 631–646.

[8] J Felsenstein (2004). PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, Universi ty of Washington, Seattle.

[9] C.M. Fraser & M.R. Dando (2001). Genomics and future biological weapons: the need for preventive action by the biomedical community. *Nat Genet* **29**(3), 253–256.

[10] D. Gurycov (1998). First isolation of Francisella tularensis subsp. tularensis in Europe. *Eur J Epidemiol* **14**(8), 797–802.

[11] David T. Jones, William R. Taylor & Janet M. Thornton (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**(3), 275–282.

[12] A.F. Kaufmann, M.I. Meltzer & G.P. Schmid (1997). The economic impact of a bioterrorist attack: are prevention and postattack intervention programs justifiable? *Emerg Infect Dis* **3**(2), 83–94.

[13] G.W. Mc Coy (1912). *Bacterium tularense*, the cause of a plaguelike disease of rodents. *Public Health Bull* **53**, 17–23.

[14] Center for Nonproliferation Studies (2001). Chronology of State Use and Biological and Chemical Weapons Control. Published online (24 okt 2006). URL `http://cns.miis.edu/research/cbw/pastuse.htm`.

[15] NOVA Online (2002). History of Biowarfare. Published online (24 okt 2006). URL `http://www.pbs.org/wgbh/nova/bioterror/hist_nf.html`.

[16] A. Tärnvik & L. Berglund (2003). Tularaemia. *Eur Respir J* **21**(2), 361–373.

[17] A. Tärnvik, H.S. Priebe & R. Grunow (2004). Tularaemia in Europe: an epidemiological overview. *Scand J Infect Dis* **36**(5), 350–355.

[18] A. Tärnvik, G. Sandström & A. Sjöstedt (1996). Epidemiological analysis of tularemia in Sweden 1931-1993. *FEMS Immunol Med Microbiol* **13**(3), 201–204.

[19] J.D. Thompson, D.G. Higgins & T.J. Gibson (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**(22), 4673–4680.

[20] R.W. Titball, P.C. Oyston, P. Chain, M.C. Chu, M. Duffield, H.H. Fuxelius, E. Garcia, G. Hälltorp, D. Johansson, K.E. Isherwood, P.D. Karp, E. Larsson, Y. Liu, S. Michell, J. Prior,

R. Prior, S. Malfatti, A. Sjöstedt, K. Svensson, N. Thompson, L. Vergez, J.K. Wagg, B.W. Wren, L.E. Lindler, S.G. Andersson, M. Forsman & P. Larsson (2005). The complete genome sequence of Francisella tularensis, the causative agent of tularemia. *Nat Genet* **37**(2), 153–159.

[21] WHO (1970). *Health aspects of chemical and biological weapons*, 105–107. World Health Organization, Geneva, 1st edition.