Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science

Draft genome assembly and bioinformatic analysis of a denovo sequenced genome of Treponema phagedenis like strain V1

Mamoona Mushtaq

Department of Animal Breeding and Genetics
Examensarbete 334
Uppsala 2010

Master's Thesis, 30 HEC

One-Year Master's Programme in Biology – Bioinformatics

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Draft genome assembly and bioinformatic analysis of a denovo sequenced genome of Treponema phagedenis like strain V1

Mamoona Mushtaq

Supervisors:
Erik Bongcam-Rudloff, SLU, Department of Animal Breeding and Genetics
Märit Pringle, SLU, Department of Biomedical Sciences and Veterinary Public Health
Hans Henrik Fuxelius, SLU, Department of Animal Breeding and Genetics
Examiner:
Göran Andersson, SLU, Department of Animal Breeding and Genetics

Credits: 30 HEC
Course title: Degree project in Animal Science
Course code: BI1021
Programme: One-Year Master's Programme in Biology
- Bioinformatics
Level: Advanced, A2E

Place of publication: Uppsala
Year of publication: 2010
Name of series: Examensarbete 334
Department of Animal Breeding and Genetics, SLU
On-line publication: http://epsilon.slu.se

Key words: *Treponema phagedenis* like strain V1, Next Generation Sequencing Technologies, Denovo Assembly, Digital dermatitis, lipoproteins

Table of Contents

ABSTRACT

In this study, a draft genome assembly of *Treponema phagedenis*-like strain V1 was obtained using Mira3 assembler. *T. phagedenis*-like strain V1 was isolated and characterized from a digital dermatitis lesions, a skin disease causing lameness in cattle. The draft assembly of Mira consisted of 812 contigs with 105 contigs of length >10,000 bp and 707 contigs of length <10,000 bp. The total number of bases in all the contigs was 3560210 with an average GC content of 40.53%. The assembly obtained with Mira3 was of good quality, average consensus quality was 79 with average contig coverage 29.39. Annotation of the draft genome sequence using an in-house annotation pipeline predicted 3507 open reading frames (ORFs) of which, 1018 encode for proteins that have homologues in *T. denticola* and 382 ORFs encode for proteins having homologues in *T. pallidum*. Genes encoding for lipoproteins, hemolysins, proteases, lipases and different types of antigens were also predicted, these genes could be involved in the pathogenesis and virulence in *T. phagedenis*-like. Genome comparison of *T. phagedenis*-like with *T. pallidum* and *T. denticola* showed more similarity of the *T. phagedenis*-like genome with that of *T. denticola* than with *T. pallidum*. Scaffolding and finishing steps of the assembly will be performed further and complete annotation including the structural and functional annotation of all the genes and proteins will be performed, and interesting findings will be validated using molecular biology techniques.

INTRODUCTION

Next generation sequencing techniques are generating a huge amount of sequence data at high speed and low cost making them suitable for population studies of various important pathogens, crops and livestock. A large number of genomes of pathogens have been sequenced to date, accelerating research in the field of biomedicine. Among sequenced genomes there are a large number of viruses and bacteria including different spirochetes.

LITERATURE STUDY

Digital Dermatitis

Digital dermatitis (DD) is an infectious and painful skin disease causing lameness in cattle (Murray *et al.,* 1996). Digital dermatitis also known as papillomatous digital dermatitis, interdigital papillomatosis, hairy heel wart and hairy foot wart (Read *et al.,* 1992) was first reported in Italy in 1974 (Cheli and Mortellaro, 1974) and since then it has been reported in several European countries, North America and Japan (Rebhun *et al.,* 1992; Blowey *et al.,* 1988; Gourreau *et al.,* 1992; Kimura *et al.,* 1993). Digital dermatitis is characterized by a mild, superficial dermatitis in the beginning that later becomes an erosive lesion that is typically red,

flat and ulcerative. The tissue becomes granulated and form hair-like projections on the plantar surface of the heel or within the interdigital cleft (Blowey *et al.,* 1988; Read *et al.,* 1992; Rebhun *et al.,* 1980). The disease results in decreased body weight and milk production that eventually leads to economic losses and welfare problems (Losinger *et al.,* 2006; Read *et al.,* 1998). DD has a connection to constant exposure to fecal slurry. Unsuccessful isolation of viruses and the resolution of lesions after topical or parental antibiotic therapy, suggest bacteria to be the causative agents of the disease. A number of bacteria of different genera including spirochetes, other anaerobic bacteria and microaerophilic organisms have been isolated from DD lesions. Among all the genera, spirochetes are most prevalent in DD lesions and also in deeper tissues where other bacteria are rarely observed.

Spirochetes

Spirochetes belong to the phylum *Spirochaetes*, having long, helically coiled (spiral-shaped) cells. Spirochetes are inhabitants of a wide range of environments. They are either linked to certain organisms such as animals and humans or present as free cells in the environment (Iida *et al.,* 2000). The location of the flagella within the periplasmic space, makes spirochetes different from other bacteria. The flagella runs lengthwise between the cell wall and outer membrane. They cause twisting motion which allows spirochetes to move.

Spirochetes are divided into three main families:

- *Brachyspiraceae*
- *Leptospiraceae*
- *Spirochaetaceae*

Some of the disease-causing members of the spirochetes are *Leptospira* spp. causing Leptospirosis, *Borrelia burgdorferi* causing Lyme disease (**Burgdorfer** *et al.,* 1982), *Borrelia recurrentis* causing Relapsing fever (Meri *et al.,* 2006), *Treponema pallidium* subsp. *pallidum* causing Syphilis and *Treponema pertenue* causing Yaws disease. The first *in vitro* cultivation of DD-associated spirochetes from Californian dairy cattle was reported by Walker *et al.,* 1995. Spirochetes isolated from DD lesions belong to the genus *Treponema*. Different strains of treponemes isolated from DD lesions are related to *T. phagedenis*, *T. denticola* and *T. vincentii* (Walker *et al.,* 1995; Choi *et al.,* 1997, Collighan and Woodward 1997). *Treponema phagedenis*-like strains have also been isolated and characterized from DD lesions in dairy cattle by Trott *et al.,* 2003 and Pringle *et al.,* 2008. Yano *et al.,* 2009 successfully islolated and characterized 40 strains of *T. phagedenis*-like spirochetes from dairy cattle with DD. Among all spirochetal genomes published to date, the following *Treponema* spp. genomes have been published:

- *Treponema denticola* ATCC 35405, Accession number: AE017226 (Seshadri *et al.,* 2004)

- *Treponema pallidum* subsp. *pallidum* Nichols, Accession number*:* AE000520 (Fraser *et al.*, 1998)
- *Treponema pallidum* subsp. *pallidum* SS14, Accession number: CP000805 (Matejková *et al.*, 2008)
- *Treponema pallidum* subsp. *pallidum* Chicago, Accession number: CP001752 (Giacani *et al.*, 2010)

Next generation Sequencing and Assembly

Sequencing

In order to determine the nucleotide sequence of a DNA molecule, next generation high throughput sequencing techniques are widely used. Next generation sequencing techniques were selected by Nature Methods as the "method of the year" in 2007 (Schuster *et al.*, 2008). Before the advent of next generation sequencing techniques Sanger enzymatic dideoxy technique (Sanger *et al.*, 1977) and Maxam and Gilbert chemical degradation method (Maxam and Gilbert, 1977) were used for sequencing purposes. Initially these techniques were used for the sequencing of small fragments ranging from 1000 to 2000 bp until whole genome sequencing was first introduced by Sanger in 1980 using shotgun sequencing. In shotgun sequencing longer sequences of DNA are subdivided into smaller fragments, these fragments are then sequenced and reassembled. Shotgun sequencing started with the sequencing of small genomes such as genomes of the bacteriophage λ (Sanger *et al.*, 1982) and viruses (Fiers *et al.*, 1978; Chee *et al.*, 1990). Whole genome shotgun sequencing had been quite expensive and required a great deal of manual labour for the assembly of the sequenced fragments until the advent of automated assembly programs that decreased both the time and manual labour required for assembly. Another important step towards decreasing the time required for sequencing was use of paired end, in which both the ends of the DNA are sequenced simultaneously, allowing large and complex genomes to be sequenced in lesser time.

A typical sequencing process involves the following steps:

1. DNA fragmentation
2. PCR amplification
3. Photodetection
4. Size selection
5. Sequencing by synthesis (polymerase) or ligation (ligase)
6. Basecalling/colorspace
7. Analysis

Below are some of the next generation sequencing techniques that are in use these days:

Technology	Approach	Read length	Bp per run	Company name
Automated Sanger sequencer	Synthesis in the presence of dye terminators	Up to 900 bp	96 kb	Applied Biosystems
ABI3730xl 454/Roche FLX system	Pyrosequencing on solid support	200–300 bp	80–120 Mb	Roche Applied Science
Illumina/ Solexa	Sequencing by synthesis with reversible terminators	30–40 bp	1 Gb	Illumina, Inc.
ABI/SOLiD Massively parallel	sequencing by ligation	35 bp	1–3 Gb	Applied Biosystems

Table 1: Advances in DNA sequencing (Morozova et al., 2009)

454 Sequencing

454 Sequencing is a sequencing technology applying the principle of pyrosequencing (Ronaghi, 2001). Pyrosequencing determines the DNA sequence using "Sequencing by Synthesis" approach. It involves taking a single strand of DNA to be sequenced and then synthesizing its complementary strand enzymatically, one base at a time. A chemoluminescent enzyme is added in the reaction mixture. Upon the incorporation of each nucleotide, this enzyme emits light that is detected to determine the nucleotide incorporated.

The complete 454 sequencing workflow is comprised of the following four steps:

Generation of a single-stranded template DNA library

In this step, the sample to be sequenced is collected. Samples from a variety of materials including genomic DNA, PCR products, BACs (bacterial artificial chromosome), and cDNA can be sequenced using this technique. Large samples such as genomic DNA and BACs are fractionated into small, 300- to 800-bp fragments. Smaller samples do not require fragmentation. A and B adapters are then added to these fragments using different molecular biology techniques, these adapters are short DNA fragments comprised of sequences complementary to a PCR primer,

sequencing primer and a key sequence. Each fragment is then immobilized to a DNA capture bead and loaded in microreactors, each microreactor containing one bead.

Emulsion-based clonal amplification of the library

PCR amplification of each unique sample library fragment is performed, all fragments are amplified in parallel, resulting in a copy number of several million per bead.

Data generation via sequencing-by-synthesis

Beads with PCR-amplified fragments are then loaded on the sequencing plate, each well of the plate having one bead. Nucleotides are then added to the fragment attached with the beads, resulting in a reaction that generates light signal. A CCD camera records the light signal emitted. Signal strength is proportional to the number of nucleotides added.

Data analysis using different bioinformatics tools

Sequencing software is then used to determine the nucleotide(s) incorporated using the colour and the strength of the signal emitted.

Assembly

After the completion of the sequencing, the sequenced fragments have to be joined to create a representation of the original chromosome from which the DNA originated. The process of joining the fragments (reads) is known as assembly. A genome assembly algorithm takes all the reads, align them to one another and detect all the places where these reads overlap. Overlapping reads are then joined to make contig(s). Ideally one contig should be produced by an assembly program for each chromosome but this is difficult because some portions of the genomes remain unsequenced because of sequencing errors and those unsequenced portions cause gaps, producing more than one contig. Another assembly problem is the presence of large numbers of identical sequences, known as repeats. The reads originating from different copies of a repeat appear identical to the assembler and an assembler incorrectly combines those reads causing assembly errors.

Assembly algorithms

The assembly algorithms fall into two main categories:

- ***De novo* Assembly**

  It is the process of assembling all the reads to form a new, unknown sequence. *De novo* assembly makes use of Greedy, Overlap-Layout-Consensus and Eulerian path algorithms.

- **Mapping/Reference Assembly**

  In mapping or reference assembly reads are assembled by aligning them to a previously known backbone sequence. It uses the Align-Layout-Consensus algorithm.

Available assemblers

There is a large number of assemblers, some commercial and some freely available are used for the assembly purpose depending upon the read types and the technology used. Some of the most commonly used assemblers are given below:

- MAQ (Mapping and Assembly with Quality) (http://maq.sourceforge.net/)
- AMOS (http://sourceforge.net/apps/mediawiki/amos/)
- MOSAIK (http://bioinformatics.bc.edu/marthlab/Mosaik)
- MIRA(http://sourceforge.net/apps/mediawiki/mira-assembler/index.php?title=Main_Page)

Scaffolding

The contigs are then ordered and oriented along a chromosome, using additional linking information like size of fragments generated through the shotgun and orientation of the read corresponding to the DNA.

Finishing

The final step of assembly is to close the gaps known as finishing or gap closure. It is a time and cost intensive process (**Nagarajan *et al*., 2010)** that requires 90 percent of the total time consumed in assembly. Gap closure is performed through different laboratory experiments and extensive manual curation is performed to validate the correctness of assembly.

Annotation

Once the sequencing and assembly is done, the next step is to perform the annotation of the sequence. The main objective of the sequence annotation is to mine useful information from a large dataset that has been generated as a result of improved sequencing techniques. Generation of large amounts of sequence data has resulted in development of databases and repositories for genomic sequences. Most of the databases have now incorporated many analytical tools that facilitate genome annotation (Bhattacharyya, 2009). Annotation is of two types, structural annotation that involves localization of gene elements *e.g*. ORF prediction, gene structure, coding regions and prediction of regulatory motifs and functional annotation that involves assignment of function to the gene or DNA sequence in question. It assigns biochemical property, gene ontology and biological function. Regulatory and interaction networks are also predicted (Aubourg, 2001) and assigned to the obtained sequence. Different tools have been used to annotate the newly discovered sequence. BLAST provides some basic level of annotation. GENETOOLS is a group of web-based tools that integrates information from a broad range of resources and is useful for genome wide analysis. NMC Annotation Tool and eGOn V2.0 are connected to this database. NMC Annotation Tool provides information from UniGene (NCBI), EntrezGene, SwissProt and

Gene Ontology (GO). *eGOn V2.0* facilitates interpretation of GO annotation. GOanna is another tool that permits users to quickly add more GO annotations by transferring GO annotations from annotated gene products in other species based on a standard Blast searches against databases that contain GO annotated sequences. GOAnnotator is the tool that links uncurated annotations to the literature (Couto *et al.*, 2006)

Aim of the Study

For this study reads from a 454 sequence run of a *Treponema phagedenis*-like genome were available. The size of the genome ranges from 3.2-3.5 Mbp and consists of one chromosome. Only a few genes has been sequenced for the human variant of *T. phagedenis*. These sequences show high similarity to the corresponding genes in our *T. phagedenis*-like strain and possibly they belong to the same species. The genome sequence of a human strain is underway (http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genomeprj&cmd=ShowDetailView&TermToSearch=48199) and comparison of this genome with the genome of bovine strains will sort out the taxonomyThe main aim of this study was to assemble and annotate the sequence using different bioinformatics tools and pipelines and to identify genes that could be important in the pathogenicity of *Treponema phagedenis*-like treponemes.

MATERIALS AND METHODS

To achieve the aim of the study the following methodology was adopted

1. Input data was gathered.
2. Sequences were assembled.
3. Assembled sequences were then visualized.
4. Sequence annotation was performed.
5. Analysis on the annotation was performed.
6. Lipoproteins were predicted.

1. Data Gathering

The data was received from Märit Pringle after sequencing of a *Treponema phagedenis*-like spirochete, strain V1, that was isolated and characterized from digital dermatitis lesions in Swedish dairy cattle (Pringle *et al.*, 2008).

Input data

Input data was organized into three files.

454Reads.fna

It contained the nucleotide sequence of all the reads in fasta format.

454Reads.qual

It consists of the base quality score in fasta format corresponding to each read in the 454Reads.fna. The reads in this file are in the same order as in the corresponding 454Reads.fna file. For each base of the read, the quality value is shown. It contained the numerical quality for each nucleotide in the read. Quality file was also in fasta format.

Sff file

Sff stands for Standard Flowgram Format, it contains information on the signal strength for each flow and contains the traces. This file is used as the input file for different assembly softwares like Newbler that Roche/454 provides with the instrument, and for some other assemblers like Mira.

2. Assembly

Mira3 (Chevreux *et al.,* 1999) was used to perform Assembly. Mira is a whole genome shotgun and EST sequence assembler for Sanger, 454 and Solexa/ Illumina data. The algorithm in the Mira assembler works in the following steps:

a. Data preprocessing

In data preprocessing stage, High confidence regions (HCR) within each read is selected, based on the base quality values present there and also Low Confidence Regions are removed which usually contains the sequencing vector that is present at the start of each read.

b. Read Scanning

Read Scanning is considered to be the start of the assembly process. In this step all the reads are compared with all the other reads and their reverse complement to find the possible overlaps between the reads.

c. Systematic match inspection

Overlaps found during the Read scanning phase are inspected using the Smith Waterman Algorithm for the local alignment of all the overlaps. Score of the expected length of the overall and the overall computed score of the overlap are calculated. Then the overlaps with computed score within the threshold of expected score and with a reasonable length is accepted as true overlap, overlaps that does not fulfill the expected criteria are identified and rejected from the further assembly.

d. Building contigs

Overlaps found and verified in the systematic match inspection phase are then assembled into contigs using iterative pairwise sequence alignment.

e. Path finder and contig interaction

In the path finder and Contig interaction phase, reads are added to the contigs based on their alignment with the consensus. If the new read integrate nice into the contig, it will be accepted if not it will be added to the same contig at some other position or skipped. The new read can then be added to another contig where possible or left as a single read contig if it does not fit in any of the contigs.

f. Consensus approval methods

The consensus formed in the above phases is then approve using different methods.

g. Read Extension

Two types of read extension is performed on the data, intracontig read extension and extra contig read extension. Intracontig read extension increases the coverage of contigs while extra contig read extension allows existing contigs to be joined after the assembly process. After extracontig extension, the contigs could be joined.

h. Contig linking and editing

This is the last step of the assembly process where the contigs are linked wherever possible and errors in the assembly are corrected using the automatic finisher.

Assembly Parameters

In order to run *de novo* assembly of 454 reads Mira uses fasta file containg reads, Fasta quality file containing quality values for reads and XML trace info file containing additional information like adaptor sequences. An adaptor sequence contains PCR primer, sequencing primer and a key and could be present in the read. Mira uses this information from XML file to trim these extra bases from the original sequence. All the three files were extracted from the sff file using SFF extract, a python script that inputs sff file and outputs fasta, fasta quality and XML trace info file. Mira was run using the following parameters:

Assembly type: whole genome, accurate.

Job : de novo

Technology: 454

Merge XML info: yes

All the other settings were used as default for 454 technology.

3. Assembly Visualization:

Assembly obtained from Mira was visualized and examined using Hawkeye (Schatz *et al.,* 2007). Hawkeye is a visual analytics tool for genome assemblies. Hawkeye is used for viewing and correcting errors from the assemblies. All levels of an assembly along with the summary statistics and assembly metrics can be analyzed using Hawkeye and mis-assemblies can be identified. It can be used for the visualization of assemblies of all sizes and it also perform various statistics besides visualization, performs different statistics that could detect contigs whose coverage is too deep and other possible mis-assemblies. Hawkeye is freely available and released as part of the open source AMOS project. In order to view assembly in Hawkeye, it must be first converted into the AMOS file type, ace file of Mira result was converted into Amos using toAmos script present in the AMOS project. The file was then used for Hawkeye. Assembly statistics including statistics on contigs length, reads used, GC contents, detailed graphical and statistical view of each contig including number of reads in it, GC content in indiviual contig and its length was performed.

4. Annotation

Annotation was performed in following steps.

- Contigs of length greater then 100bp were selected and annotated using an in-house annotation pipeline GeneComp.
- Open reading frames (ORFs) were predicted using Glimmer. Glimmer is a tool that is especially designed to predict genes in bacterial and viral genomes. It uses interpolated Markov models to predict coding sequences. (Delcher *et al.*, 1999)
- ORFs with length greater than 100 bases were selected and translated into its respective peptide sequence.
- All the peptide sequences are than checked for their homologous proteins using Basic Local Alignment Search Tool (Blast).
- Multiple sequence alignment of the Blast results were than performed using Muscle (multiple sequence comparison by log-expectation).

5. Analysis

For the analysis and visualization of the results of GeneComp, Artemis was used. Artemis is a DNA sequence visualization and annotation tools (**Rutherford *et al.,* 2000).** Artemis can be used to analyse the results of compact genomes of bacteria, archea and lower eukaryotes. It can handle sequences of small genes to large genomes. Annotations and analysis in EMBL (Baker *et al.,* 2000) and GenBank (Benson *et al.,* 2000) can be visualized in Artemis. Besides visualization, Artemis can also be used as an annotation tool, it can be used to run blast and pfam searches on

the selected sequence. It was used to calculate the percentage of GC content in all the Contigs, genes that have homologues in *T. denticola, T. pallidum* and other bacteria.

6. Lipoproteins Prediction

For the prediction of lipoproteins, LipoP 1.0 Server (Juncker *et al.,* 2003) and Splip were used and their results were cross-validated.

LipoP is used to predict lipoprotein signal peptides in Gram-negative eubacteria. It has the ability to distinguish between lipoproteins (SPaseII-cleaved proteins), SPaseI-cleaved proteins, cytoplasmic proteins, and transmembrane proteins. It also predicts cleavge sites, CleavI: Cleavage site for (signal peptidase I) and CleavII: Cleavage sites for (signal peptidase II) along with amino acid at position+2 after the Cleav2, where aspartic acid "D" shows that it is attached to the inner membrane, others are attached to the outer membrane. The Plot of scores is also shown for all the predicted classes, with their cleavage sites. Amino acid sequences of all the translated CDS in fasta format were given to LipoP server to find out the lipoproteins among all those sequences.

SpLip (Setubal *et al.,* 2006) is used to predict lipoproteins in spirochaetal genomes. It makes use of already predicted lipoproteins of spirochaetes like *T. denticola, T. pallidum* etc, making a training set for the query sequence. SpLip takes amino acid sequence as an input and predict that sequence to be a probable lipoprotein, possible lipoprotein or not a lipoprotein. The annotation "probable lipoprotein" is more likely being a *bona fide* lipoprotein as compared to the annotation "possible lipoprotein". In the result, SpLip shows the best hit and secondary hit with the cleavage site and score and size of the H-region, N-region and the lipobox.

Lipoproteins having their homologous protein sequences in *T. pallidum* were selected and analysed further.

7. Phylogenetic Analysis

Phylogenetic analysis was performed using 16S rRNA sequence of *T. phagedenis* like strain V1 (Genbank accession number: DQ470655). The phylogenetic tree was constructed using neighbour joining method (Saitou and Nei, 1987) using a distance matrix comprised of 1239 nucleotides.

RESULTS

Assembly:

After performing assembly with Mira3, 812 contigs were formed, with average coverage of 29.39 and average consensus quality 79. Assembly statistics of Mira3 assembly has been shown in Table 1 and its graphical view also showing the length distribution of contigs is shown in Figure 1. GC Contents of the Contigs range from 25.29% to 63% with the average GC contents of all the

contigs 40.53%. GC content distribution can be seen in Figure 2.

Number of Contigs	812
Number of large Contigs (size >10,000bp)	105
Largest Contig size	96846
Number of small Contigs	707
Total reads used in all Contigs	295597
Total number of bases in all Contigs	3560210
Average Consensus quality	79
Average GC contents	40.53%
Average Contig coverage	29.39
Lowest mismatches from consensus	0
Highest mismatch from consensus	3.8

Table 2. Assembly statistics of the Mira3 assembly.

Graphical view of the assembly using Hawkeye

Assembly Statistics

Length distributions of all contigs and their GC contents were retrieved. Following window shows the statistics of the assembly and the length distribution of all the contigs.

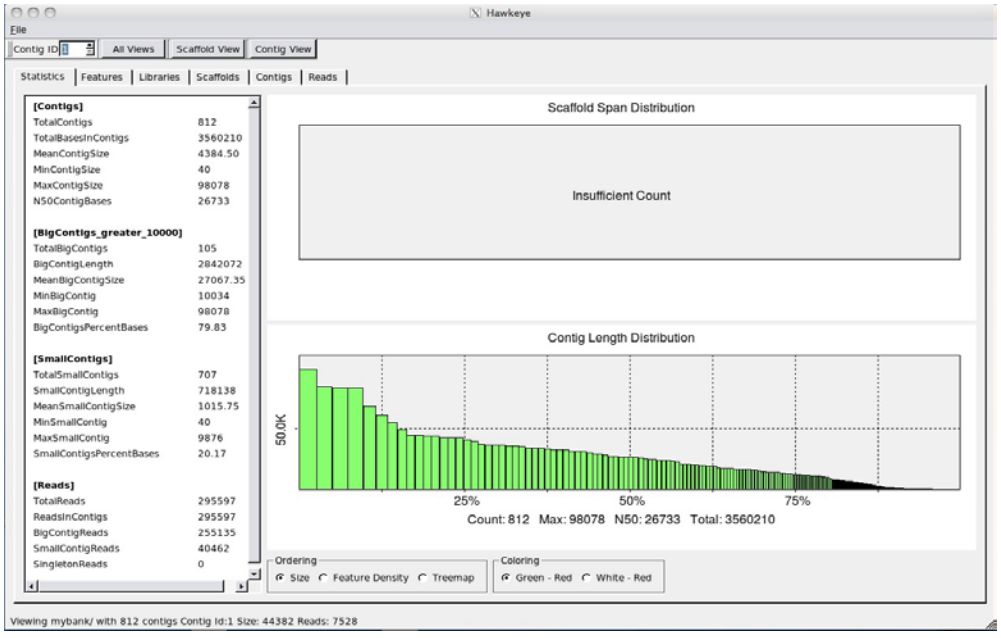

Figure 1. Graphical view of the Mira3 assembly

GC Content

GC contents of the contigs range from 25.29% to 63% with the average GC contents of all the contigs 40.53%.

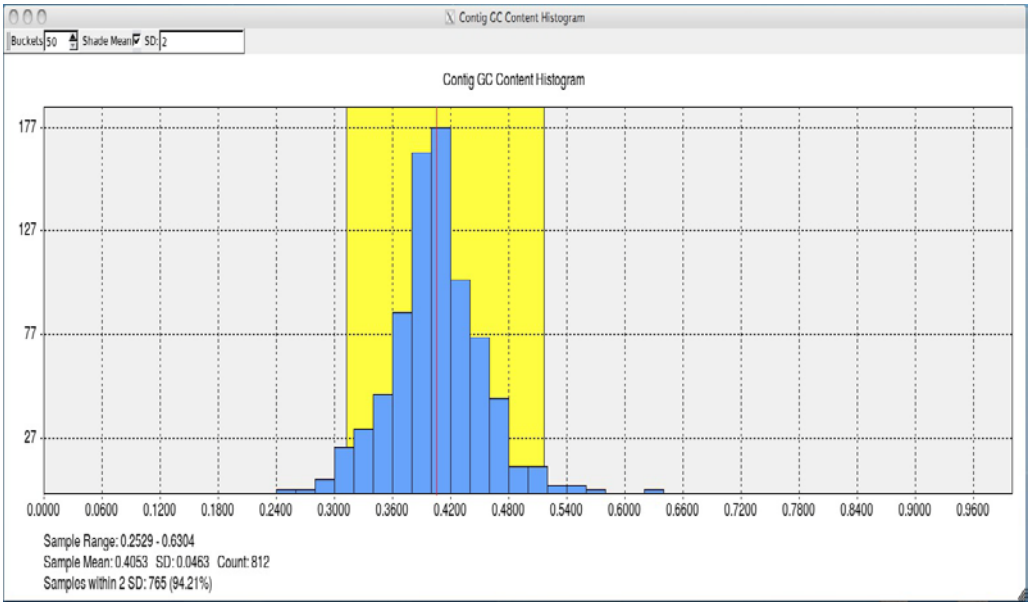

Figure 2. GC Content distribution graph of Mira3 contigs

Contig View

Figure 3. Contig view showing part of contig 18 (largest contig)

Newbler Assembly:

Number of Contigs: 309
GC Content: 39.53%
Mismatch percentage: 0-1

Velvet Assembly:

Number of contigs: 18976
Average Mismatch: 25%

Annotation

With GeneComp 3507 ORFs were predicted containing 2581590 base pairs. Gene density, GC percentage and the calculated number of homologous genes of the predicted CDS are shown in Table 3.

Predicted ORFs (>100bp)	3507
Number of bases in all ORFs	2581590
Density	1.159 genes per kb (862 bases per gene)
GC percentage	39.92
Predicted ORFs encoding for homologous proteins in *T. denticola*	1018
Predicted ORFs encoding for homologous proteins in *T. pallidum*	382
Predicted ORFs encoding for homologous proteins in different other bacteria	721
Predicted ORFs encoding for hypothetical proteins	1394

Table 3. General information of the annotation performed with GeneComp

Artemis View

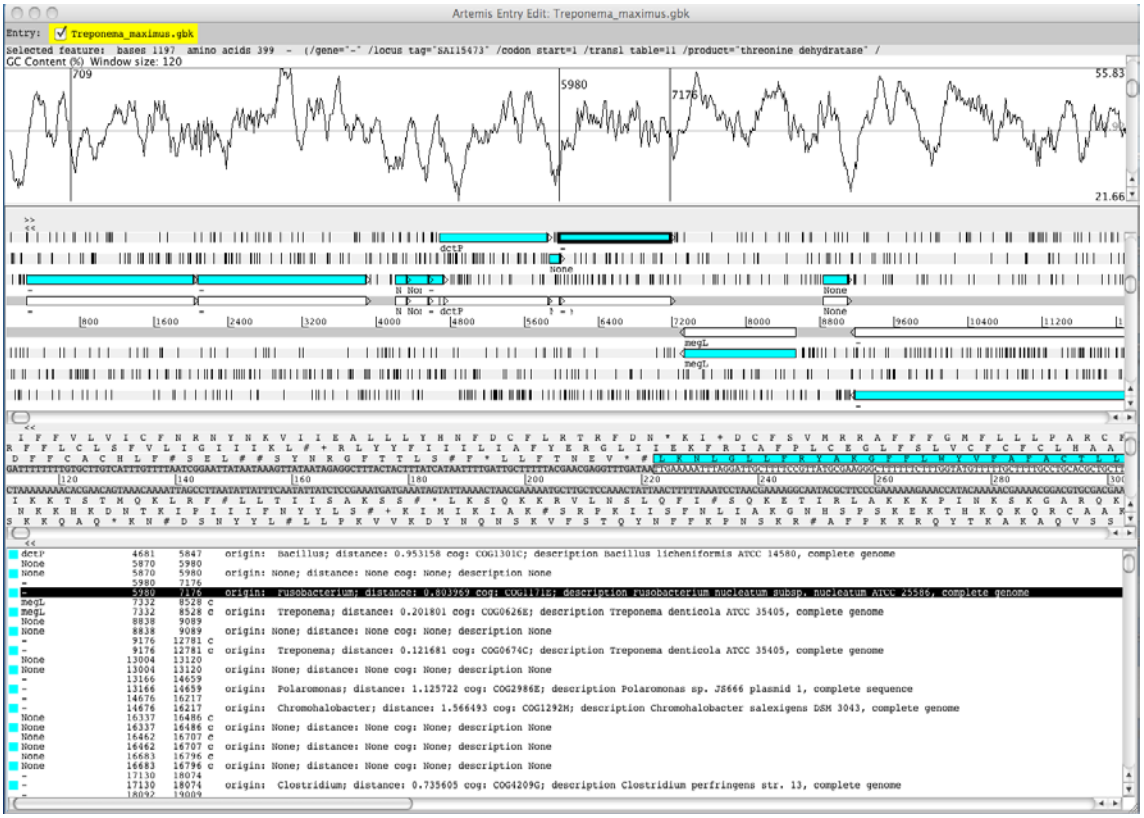

Figure 4. Graphical view of the GeneComp annotation in Artemis

Genes with Potential Role in Pathogenesis and Virulence

Following table shows the number of predicted ORFs that could be involved in the pathogenesis and virulence.

Putative gene	No of genes
ORFs encoding Lipoproteins	123
ORFs encoding Hemolysins	3
ORFs encoding Proteases	17
ORFs encoding Lipases	1
ORFs encoding Antigens	11

Table 4. The number of predicted genes with possible role in pathogenesis and virulence.

Lipoproteins

LipoP predicted 123 lipoproteins among all the translated predicted ORFs. Among these 123 lipoproteins, 49 were homologous to *T. denticola* proteins and 16 to *T. pallidum* proteins. Splip

predicted total 140 lipoproteins including 10 possible and 130 probable lipoproteins. Among all 140, 58 were homologous to *T. denticola* proteins and 15 to *T. pallidum* proteins. Sample output from LipoP is shown in Table 5 and Figure 5 and from that of SpLip is shown in Table 6.

 # SAI15544 SpII score=26.4858 margin=15.4067 cleavage=19-20 Pos+2=T

Cut-off=-3

SAI15544	LipoP1.0:Best	SpII	1	1	26.4858		
SAI15544	LipoP1.0:Margin	SpII	1	1	15.4067		
SAI15544	LipoP1.0:Class	SpI	1	1	11.0791		
SAI15544	LipoP1.0:Class	CYT	1	1	-0.200913		
SAI15544	LipoP1.0:Signal	CleavII	19	20	26.4858	# VLLVS|CTKSK	Pos+2=T
SAI15544	LipoP1.0:Signal	CleavI	25	26	10.2489	# TKSKT|ESEQG	
SAI15544	LipoP1.0:Signal	CleavI	23	24	8.82631	# SCTKS|KTESE	
SAI15544	LipoP1.0:Signal	CleavI	24	25	7.9721	# CTKSK|TESEQ	
SAI15544	LipoP1.0:Signal	CleavI	21	22	6.54629	# LVSCT|KSKTE	
SAI15544	LipoP1.0:Signal	CleavI	27	28	6.23338	# SKTES|EQGAV	
SAI15544	LipoP1.0:Signal	CleavI	19	20	4.42421	# VLLVS|CTKSK	
SAI15544	LipoP1.0:Signal	CleavI	26	27	4.35515	# KSKTE|SEQGA	
SAI15544	LipoP1.0:Signal	CleavI	22	23	4.19607	# VSCTK|SKTES	
SAI15544	LipoP1.0:Signal	CleavI	17	18	3.06006	# SAVLL|VSCTK	
SAI15544	LipoP1.0:Signal	CleavI	20	21	1.03796	# LLVSC|TKSKT	
SAI15544	LipoP1.0:Signal	CleavI	28	29	0.208823	# KTESE|QGAVS	
SAI15544	LipoP1.0:Signal	CleavI	18	19	-0.859783	# AVLLV|SCTKS	
SAI15544	LipoP1.0:Signal	CleavI	30	31	-1.40653	# ESEQG|AVSKT	
SAI15544	LipoP1.0:Signal	CleavI	16	17	-2.83057	# VSAVL|LVSCT	

Table 5. Output of LipoP for protein sequence having id SAI15544

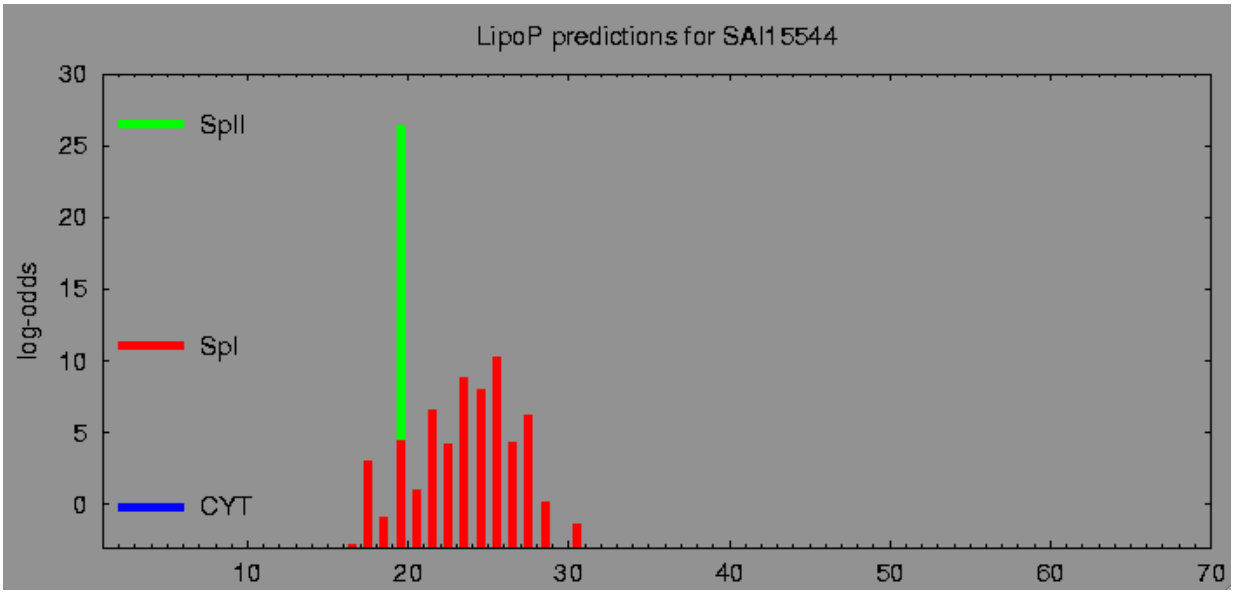

Figure 5. Output for protein sequence having id SAI15544

Results for ORF SAI15544 – gamma-glutamyl transpeptidase 71588:73366 forward MW:63735: PROBABLE LIPOPROTEIN

	BEST HIT	SECONDARY HIT
cleavage site	19	-1
score H-Region	23.500000	-100.000000
score Lipobox	8.231000	100
size of H-Region	10	-5
score N-Region	3	0

Table 6. Output of SpLip for protein sequence having id SAI15544

Analysis of predicted lipoproteins having homologues in T. pallidum

Below is the table providing the analysis of the lipoprotein sequences that have homologous proteins in *T. pallidum*:

Protein ID	Protein Length	SpII Length	Lipobox sequence	T. pallidum homologue's accession number	Function
SAI15793	307	17	ILTG	NP_218816.1	Hypothetical protein

SAI15862	309	22	LLVG	NP_218602.1	ABC transporter, periplasmic binding protein (troA)
SAI15971	344	21	CFTG	NP_219205	Membrane protein (tmpA)
SAI16225	236	18	IFSS	NP_219209	Hypothetical protein TP0772
SAI16565	346	20	LLGA	NP_219092	Spermidine/putrescine ABC transporter, periplasmic binding protein
SAI16582	342	18	LLIC	ADD72890	Thiamine biosynthesis lipoprotein ApbE
SAI16975	273	20	LFFS	NP_218749	Amino acid ABC transporter, periplasmic binding protein (hisJ)
SAI17190	81	20	AAGF	ADD73081	P26 [*Treponema pallidum* subsp. *pallidum* str. Chicago]
SAI17208	353	21	GFGS	ADD73092	Conserved hypothetical protein
SAI17236	194	19	LFVS	ADD72615	Conserved hypothetical protein
SAI17243	284	21	FFSS	NP_218894	Hypothetical protein TP0453
SAI17436	311	18	CFAS	ADD72776	Conserved hypothetical protein
SAI18198	501	28	VSAR	ADD72934	Periplasmic serine protease DO
SAI18227	177	20	AAVS	ADD72398	Putative LysM domain protein
SAI18264	568	20	LIIS	NP_218985	Periplasmic serine protease, putative

SAI18386	470	22	GLIA	NP_219391	Hypothetical protein TP0954

Table 7. Analysis of the lipoprotein sequences that have homologous proteins in T. pallidum. Table contains lipoproteins predicted by both LipoP and Splip.

Phylogenetic Analysis

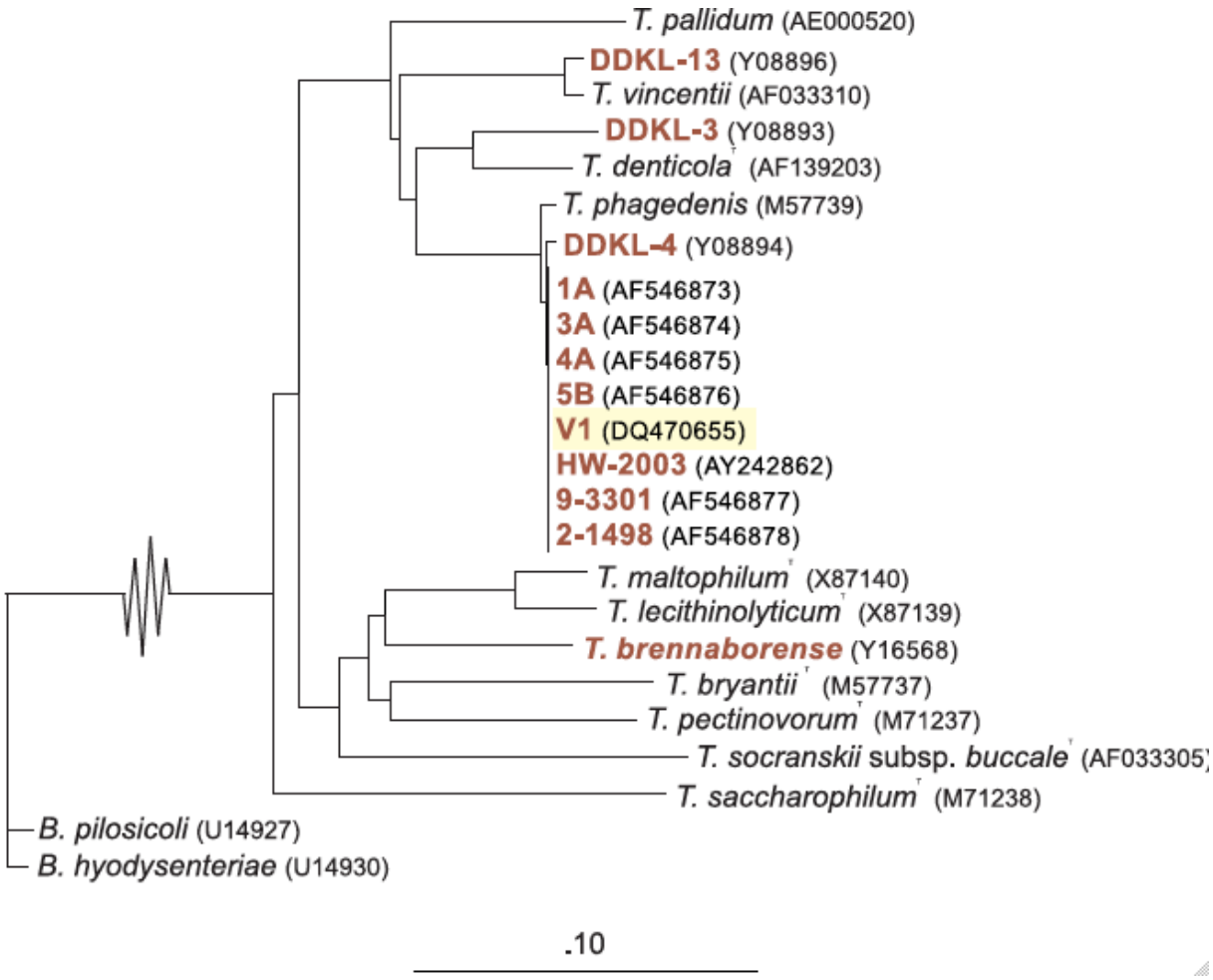

Figure 6. Phylogenetic tree based on a distance matrix analysis of 1239 positions from 16S rRNA gene sequences. The Brachyspira hyodysenteriae and Brachyspira pilosicoli sequences served as outgroup. The scale bar indicates 10 substitutions per 100 nucleotide positions. Isolates from cattle are highlighted in red.

DISCUSSION

Assembly

Recent development in the high throughput sequencing techniques has enormously increased the amount of data generated from the sequencing of genomes. The usefulness of this data depends upon the quality of the draft genome sequence obtained from this huge data. In this study we performed assembly with Mira3 Assembler. Results obtained from Mira when analysed and compared with the assembly obtained from other assembly programs came out to be of good quality. One run of assembly was done by Newbler that comes along with the Roche 454 Sequencing system and another assembly was performed using Velvet assembler. The assembly obtained with Mira was inspected with Hawkeye assembly visualization tool to find out the miss assemblies and area where those miss assemblies exist but no large miss assemblies were found, only rare mismatches in the reads were observed that could be because of sequencing errors and those mismatches were not affecting the consensus sequence. Mira assembly generated 812 contigs containing 3560210 base pairs with largest contig comprised of 96846 base pairs. From the total of 812 contigs, 105 contigs were large contigs of length greater then 10000 bp and remaining 707 contigs were of length less then 10000 bp, considered to be small contigs. Contigs obtained from Mira were compared with the assembly obtained from Newbler. The Newbler assembly consisted of 309 contigs containing 3025138 bp showing much difference in the number of contigs but less difference in the number of bp. The difference in the size and numbers of contigs can be mainly because of the “Uniform read distribution” feature of Mira. According to this feature reads should be evenly distributed in the project. In our assembly average coverage was 29.39, during the assembly process Mira would have enforced the coverage near this value and would have rejected the reads from a contig, keeping them reserve for other copies of that repeat. If those reads were not actually the repetitive sequences it would have joined them with some other false reads having an overlapping region, building a new and a small contig. Overall assembly quality seems good making it an acceptable draft assembly. Other assembly information from Mira assembly including GC contents, mismatches percentage was also inspected and compared with Newbler's assembly. Average GC contents in the Mira contigs was 40.53% and in Newbler's contigs, it was 39.53% that shows consistency in the two assemblies. More bases in the contigs formed by Mira could also be contributable to the slight increase in the GC contents of Mira. Reads mismatches from the consensus sequences were also low, most of the contigs have mismatch value ranging from 0% to 1%. Percentage mismatches between Newbler's and Mira assembly were also consistent showing almost same mismatch percentage from the consensus. Another assembly run was performed using Velvet, but the results obtained from velvet showed high deviation from the results of Newbler and Mira. The number of contigs formed was 18976

with the 25% average mismatch percentage. Reason for bad assembly could be the use of only fasta file without having quality values for the reads and trace data such as adapters information. Assembly obtained from Velvet was thus discarded and was not used for further analysis. As the data contained only single ended reads and no reference sequence for *T. phagedenis* like genome was available, the assembly had to be stopped without joining the contigs into scaffolds and finishing the assembly. Thus, the assembly performed here gives just a draft genome sequence, present in number of contigs instead of one as it should be in the finished assembly.

Annotation

The *T. phagedenis*-like strain V1 genome is predicted to encode 3507 CDS comprised of 2581590 bp. For annotation, only those ORFs having at least 100 nucleotides were selected to be potential functional genes. Among all the 3507 predicted genes, 1018 are found to have homologous genes in *T. denticola*, 382 genes have homologues in *T. pallidium*, 721 genes have their homologues in a number of other bacteria and 1394 genes are found to encode hypothetical proteins. Gene density is predicted to be 1.159 genes per kb with 862 bases per gene. In the draft genome sequence of *T. phagedenis*-like strain V1 a number of genes were predicted to be potential candidates for causing pathogenesis and virulence. These genes include three genes (CDS) encoding hemolysins, 17 encoding proteases, one encoding a lipase and 11 genes (CDS) were predicted to encode antigens. Besides these 123 genes (CDS) encoding for Lipoproteins are also predicted using LipoP 1.0 server.

Lipoprotein Prediction

Two different sets of results were obtained using LipoP and SpLip. 123 lipoproteins were predicted with LipoP having 49 homologous proteins in *T. denticola* and in *T. pallidum.* Splip predicted total 140 lipoproteins including 10 possible and 130 probable lipoproteins. Among all 140, 58 were homologous with *T. denticola* proteins and 15 with *T. pallidum* proteins. Lipoproteins that LipoP was unable to predict were predicted by SpLip. The reason for high sensitivity of SpLip has been designed specifically for the prediction of spirochetal lipoproteins. Because lipobox sequence in spirochetes exhibit more plasticity compared with those of the other bacteria, it is difficult for the prediction algorithm to identify those lipoboxes that are different from other bacteria. SpLip algorithm makes use of lipobox weight constructed using the already known lipoproteins of different spirochetes and assigns the scores to the query lipobox based on that matrix. Thus, it can predict all those lipobox sequences that are unique to spirochetes and can not be predicted by other prediction algorithms like the one used in lipoP. In order to compare the results of LipoP and SpLip, lipoproteins having homologoues in *T. pallidum* have been analysed. 15 lipoproteins were common in both results, LipoP predicted an additional protein SAI17190

having homology with P26 (*T. pallidum* subsp. *pallidum* str. Chicago) The analysis of these proteins shows that four out of 16 proteins have role as the periplasmic binding protein that include their role as amino acid transporters and serine protease. While seven out of 16 are hypothetical proteins and one of the 16 predicted lipoproteins contain LysM domain which was originally identified in the bacterial cell wall degrading enzymes (Bateman *et al.,* 2000). Large numbers of proteins that have this domain were found to be involved in the bacterial pathogenesis *e.g*. Staphylococcal IgG binding proteins. Another protein shows similarity with lipoprotein ApbE that is involved in thiamine biosynthesis.

Genome Comparison

Comparison of the draft genome of *T. phagedenis*-like strain V1 with that of *T. denticola* and *T. pallidum* shows that the *T. phagedenis*-like genome is more similar to *T. denticola*. Genome length of *T. denticola* is closer to *T. phagedenis*-like than that of *T. pallidum*. Also *T. denticola* shares more nucleotide identity with *T. phagedenis*, 1018 genes in *T. phagedenis* (CDS) have their matches to CDS in *T. denticola* whereas only 382 genes have matches with that of *T. pallidum.* GC content in the draft genome of *T. phagedenis* like is 39.92% that is also closer to that of the *T. denticola* that is 37.9% than *T. pallidum* that is 52.8%. Average predicted CDS size of *T. phagedenis* (862 bases per gene) is also closer to *T. denticola* (939 bases per gene) than *T. pallidum* (1,017 bases per gene) (Seshadri *et al.,* 2004). The phylogenetic analysis based on 16S rRNA gene sequences also shows that *T. phagedenis*-like strain V1 is more closely related to *T. denticola* than *T. pallidum.*

CONCLUSION

Assembly job performed with Mira3 produces good quality *de-novo* assembly of 454 reads and it can be used for obtaining a working draft of the genome sequence. Different genes predicted from the draft genome sequence of the *T. phagedenis*-like strain V1 can have a role in the pathogenesis and virulence. Among them are genes encoding for hemolysins, proteases, lipases, antigens and lipoproteins. These genes could be validated with different molecular biology techniques and can for example be used for vaccine production. Genome comparison of *T. phagedenis*-like strain V1 with that of *T. denticola* and *T. pallidum* shows that the *T. phagedenis*-like genome is more similar to *T. denticola*.

FUTURE PROSPECTS

Genome assembly and annotation of *T. phagedenis* like strain V1 is yet incomplete. Scaffolding and finishing steps of the assembly will be performed further and complete annotation including

the structural and functional annotation of all the genes and proteins will be performed, and interesting findings will be validated using molecular biology techniques.

ACKNOWLEDGEMENT

I would like to thank the following persons for helping me in completion of my MS thesis:

- Märit Pringle, my co-supervisor for providing the data and for discussions and suggestions on the result and written work.
- Anna Rosander for suggesting me the type of analysis that should be performed on the data.
- HEC (Higher Education Commission) Pakistan for financing me to do my studies in Sweden.

REFERENCES

- Aubourg, S., Rouzé, P. (2001) Genome annotation Plant Physiology and Biochemistry, **39:** 181-193
- Baker, W., Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G., Tuli, M. A. (2000) The EMBL Nucleotide Sequence Database. *Nucleic Acids Research*, **28:** 19-23
- Bateman, A., Bycroft, M. (2000) The structure of a LysM domain from E. coli membrane-bound lytic murein transglycosylase D (MltD). J Mol Biol, **299(4)**:1113-9.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E. W. (2000) GenBank. *Nucleic Acids Research*, **37:** 26-31
- Bhattacharyya, A. Genome Sequence Databases: Annotation. (2009) Encyclopedia of Microbiology, (Third edition) 174-184
- Blowey, R. W., Sharp, M. W. (1988) Digital dermatitis in dairy cattle. *Vet. Rec*., **122**: 505-508 *BMC Microbiol*. **8:** 76.
- Burgdorfer, W., Barbour, A. G., Hayes, S. F., Benach, J. L., Grunwaldt, E., Davis, J. P. (1982) Lyme disease-a tick-borne spirochetosis? *Science*, **216:** 1317-1319
- Chee, M. S., Bankier, A. T., Beck, S., Bohni, R., Brown, C. M. (1990) Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. *Curr Top Microbiol Immunol*, **154**: 125–169
- Cheli, R., Mortellaro, C. M. (1974) La dermatite digitale del bovino, p. 208-213. *In* P. Gallarati (ed.), Proceedings of the 8th International Conference on Diseases of Cattle. Piacenza, Milan, Italy
- Chevreux, B., Wetter, T., Suhai, S. (1999) Genome sequence assembly using trace signals and additional sequence information. *Proc German Conf Bioinformatics*, **99:** 45-56.
- Choi, B.K., Nattermann, H., Grund, S., Haider, W., Gobel, U. B. (1997) Spirochetes from digital dermatitis lesions in cattle are closely related to treponemes associated with human periodontitis. *Int. J. Syst. Bacteriol*. **47:** 175-181 .
- Collighan, R. J., Woodward**,** M. J., 1997. Spirochaetes and other bacterial species associated with bovine digital dermatitis. *FEMS Microbiol. Lett*. **156:** 37-41.
- Couto1, F.M., Silva1, M.J., Lee,V., Dimmer, E., Camon, E., Apweiler, R., Kirsch, H., Schuhmann, D.R. (2006) GOAnnotator: linking protein GO annotations to evidence text. *JBDC*, **1:**19
- Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, A. (1978) Complete nucleotide sequence of SV40 DNA. *Nature*, **273**: 113–20
- Fraser, C. M., Norris, S. J., Weinstock, G. M., White, O., Sutton, G. G., Dodson, R., Gwinn, M., Hickey, E. K., Clayton, R., Ketchum, K. A., Sodergren. E., Hardham, J. M., McLeod, M. P., Salzberg, S., Peterson, J., Khalak, H., Richardson, D., Howell, J. K., Chidambaram, M., Utterback, T., McDonald, L., Artiach, P., Bowman, C., Cotton, M. D., Fujii, C., Garland, S., Hatch, B., Horst, K., Roberts, K., Sandusky, M., Weidman, J., Smith, H. O., Venter, J. C. (1998) Complete genome sequence of Treponema pallidum, the syphilis spirochete. *Science **281:** 375 - 388*
- Giacani, L., Jeffrey, B. M., Molini, B. J., Le, H. T., Lukehart, S. A., Centurion-Lara, Arturo., Rockey, D. D. (2010) Complete Genome Sequence and Annotation of the *Treponema pallidum* subsp. *pallidum* Chicago Strain. *Journal of Bacteriology*, **192:** 2645-2646
- Gourreau, J. M., Scott, D. W., Rousseau, J. F. (1992) La dermatite digitee des bovins. *Le Point Veterinaire*, **24**: 49-57.
- Iida, T., Ohkuma, M., Ohtoko, K., Kudo, T. (2000) Symbiotic spirochetes in the termite hindgut: Phylogenetic identification of ectosymbiotic spirochetes of oxymonad protists. *FEMS Microbiol. Ecol***, 34:**17-26

- Juncker, A. S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H., Krogh, A.Prediction of lipoprotein signal peptides in Gram-negative bacteria. Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby 2800, Denmark.
- Kimura, Y., Takahashi, M., Matsumoto, N., Tsukuda, H., Satoh, M., Okhawara, K., Kaoe, M., Gotoh, N., Kubo, M., Aoki, O., Hataya, M. (1993) Verrucose dermatitis and digital papillomatous in dairy cows. *J.Vet. Med., Jpn.* **46:**680-685.
- Losinger, W.C. (2006) Economic impacts of reduced milk production associated with papillomatous digital dermatitis in dairy cows in USA. *J Dairy Res.,***73**: 244-256.
- *Matějková,* P., Strouhal, M., Šmajs, D., Norris, S. J., Palzkill, T., Petrosino, J. F., Sodergren, E., Norton, J. E., Singh, J., Richmond, T. A., Molla,M. N., Albert, T. J., Weinstock, G. M. (2008)
- Maxam, A. M., Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A* **74**: 560–564.
- Meri, T., Cutler, S. J., Blom, A. M., Meri, S., Jokiranta, T. S. (2006) Relapsing Fever Spirochetes Borrelia recurrentis and B. duttonii Acquire Complement Regulators C4b-Binding Protein and Factor H. *Infection and Immunity***, 74:** 4157-4163
- Morozova, O., Hirst, M., Marra, M. A., (2009) Applications of New Sequencing Technologies for Transcriptome Analysis. *Annu. Rev. Genomics Hum. Genet,* **10:** 135–51
- Murray, R. D., Downham, D. Y., Clarkson, M. J., Faull, W. B., Hughes, J.W., Manson, F. J., Merritt, J. B., Russell, W.B., Sutherst, J. E., Ward, W.R. (1996) Epidemiology of lameness in dairy cattle: description and analysis of foot lesions. *Vet. Rec*. **138**:586-591.
- Pringle**,** M., Bergsten, C., Fernström, L.L., Höök, H., Johansson, K.E. (2008) Isolation and characterization of Treponema phagedenis-like spirochetes from digital dermatitis lesions in Swedish dairy cattle. *Acta Veterinaria Scandinavica,* **50:** 40.
- Read, D. H., Walker, R. L., Castro, A. E., Sundberg, J. P., Thurmond, M. C. **(**1992) An invasive spirochaete associated with interdigital papillomatosis of dairy cattle. *Vet. Rec.* **130:**59-60.
- Rebhun, W. C., Payne, R. M., King, J. M., Wolfe, M., Begg, S. N. (1980) Interdigital dermatitis in dairy cattle. *J.Am. Vet. Med. Assoc.* **177:** 437-440.
- Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics,* **16:** 944-5
- Saitou, N., Nei, M. (1987) The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol*. **4(4):** 406-425.
- Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F., Petersen, G. B. (1982) Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol,* **162:** 729–773
- Sanger, F., Nicklen, S., Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *PNAS***, 74:** 5463–5467
- Schatz, M.C., Phillippy, A.M., Shneiderman, B., Salzberg, S.L. (2007) Hawkeye: a visual analytics tool for genome assemblies. *Genome Biology,* **8:** R34
- Schuster, S.C. (2008) Next-generation sequencing transforms today's biology. *Nat. Methods,* **5:** 16–18
- Seshadri, R., Myers, G. S., Tettelin, H., Eisen, J. A., Heidelberg, J. F., Dodson, R. J., Davidsen, T. M., DeBoy, R. T., Fouts, D. E., Haft, D. H., Selengut, J., Ren, Q,. Brinkac, L. M., Madupu, R., Kolonay, J., Durkin, S. A., Daugherty, S. C., Shetty, J., Shvartsbeyn, A., Gebregeorgis, E., Geer, K., Tsegaye, G., Malek, J., Ayodeji, B., Shatsman, S., McLeod, M. P., Smajs, D., Howell, J. K., Pal, S. , Amin, A., Vashisth, P., McNeill, T. Z., Xiang, Q., Sodergren, E., Baca, E., Weinstock, G. M., Norris, S. J., Fraser, C. M., Paulsen, I. T. (2004) Comparison of the genome of the oral pathogen Treponema denticola with other spirochete genomes. *PNAS ,**101:** 5646-5651*
- Setubal, J. C., Reis, M., Matsunaga, J., Haake, D. A. (2006) Lipoprotein computational prediction in spirochaetal genomes. *Microbiology,* **152:** 113–121

- Trott, D.J., Moeller, M.R., Zuerner, R.L., Goff, J.P., Waters, W.R., Alt, D.P., Walker, R.L., Wannemuehler, M.J. (2003) Characterization of Treponema phagedenis-like spirochetes isolated from pappilomatous digital dermatitis lesions in dairy cattle.*J Clin Microbiol* **41:** 2522-2529.
- Walker, R. L., Read, D.H., Loretz, K. J., Nordhausen**,** R. W. (1995) Spirochetes isolated from dairy cattle with papillomatous digital dermatitis and interdigital dermatitis. *Vet. Microbiol.* **47:** 343-355.
- Yano, T., Yamagami, R., Misumi, K., Kubota, C., Moe, K. K., Hayashi, T., Yoshitani, K., Ohtake, O., Misawa, N. (2009). Genetic Heterogeneity among Strains of Treponema phagedenis-Like Spirochetes Isolated from Dairy Cattle with Papillomatous Digital Dermatitis in Japan. *J. Clin. Microbiol.* **47:** 727-733.
- Ronaghi, M. (2001) Pyrosequencing Sheds Light on DNA Sequencing. *Genome Res.* **11**: 3-11
- Nagarajan, N., Cook, C., Bonaventura, M.D., Ge,H., Richards, A., Bishop-Lilly, K.A., DeSalle, R., Read, T.D., Pop, M. (2010) Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. *BMC Genomics.* **11**:242