DNA Sequence Informatics - II

BINP29

Student: Chandrashekar CR, ch1131ch-s@student.lu.se

Lecturer: Dag Ahren

Report

Please send in the following through Canvas:

1. A file with a README documenting your work in a reproducible way. Make sure that you describe your choice of filters AND describe your reasoning for those choices. Each command should be present and links to downloaded data (including versions of data and software).
2. A short report answering the questions from the exercise.

Data Collection

Background

Malaria is caused by parasites belonging to the phylum Apicomplexa. One of the more well known genera causing malaria is Plasmodium. The species of this genus infect not only humans, but also other mammals, birds and "reptiles''. The single one species that causes most malaria infections in humans is Plasmodium falciparum. This is also the most serious of malaria infections. Many genomes of Plasmodium have been sequenced. Recently the first genome of an avian (bird) malaria parasite has been sequenced. The species' name is Haemoproteus tartakovskyi. The genomes we will work with are:

Species	Host	Genome Size (bp)	Genes	Genomic GC (%)
Plasmodium berghei	Rodents	17,954,629	7,235	23.71
Plasmodium cynomolgi	Macaques	26,181,343	5,787	39.08
Plasmodium falciparum	Humans	23,270,305	5,207	19.36
Plasmodium knowlesi	Lemurs	23,462,346	4,953	37.54
Plasmodium vivax	Humans	27,007,701	5,682	42.20
Plasmodium yoelii	Rodents	22,222,369	4,919	20.78
Toxoplasma gondii	Humans	128,105,889	15,892	52.20

Table 1: Genomic characteristics of various apicomplexan parasites, including genome size, number of genes, and GC content.

There is a controversy among researches whether Plasmodium falciparum is related to the other mammalian parasites, or if it originates from a bird malaria parasite which has changed its host. We may solve this by making an extensive phylogenetic analysis.

Problem 1: Do you think that in a phylogenetic tree the parasites that use similar hosts will group together?

Not necessarily. While host similarity can be a factor in shaping parasite evolution, phylogenetic trees are based on genetic relatedness rather than host preference alone. If host shifts are rare and parasites co-evolve with their hosts over long periods, we might expect parasites infecting similar hosts (e.g., primates) to cluster together. However, if host shifts occur frequently (as seen in some parasites), a species like Plasmodium falciparum might be more closely related to an avian parasite like Haemoproteus tartakovskyi rather than other mammalian Plasmodium species. Horizontal gene transfer, recombination, or convergent evolution could further blur the host-based grouping. Thus, while we might see some clustering by host type, the true relationships will depend on the genetic divergence patterns revealed by our phylogenetic analysis.

Plasmodium data

The first task will be performed on the server. Remember to organize your files! All the genomes of the Plasmodium species together with the outgroup taxon Taxoplasma gondii, another member of Apicomplexa, are found at the course web site in one single file. The file name is plasmodiumGenomes.tgz. With one of the genome files you will later encounter a problem that you have to solve. First a gene prediction using GeneMark should be carried out (the program name is gmes_petap.pl, see the gene prediction and annotation exercise from the previous course). Since prediction takes time, divide the work among the students so that one particular student makes the gene prediction for one particular genome. The genome of Toxoplasma gondii takes too long to run and can be skipped. The produced gff-files should be copied to the /tmp/Prediction directory on the course server. From here, all students can copy the gff-files. The Toxoplasma gff-file can be downloaded from the course web site with the name Tg.gff.

```
# Copy all the gtf files and also the raw .genome files from the courser server
cp /temp/Prediction/* .
cp /resources/binp29/Data/malaria/plasmodiumGenomes.tgz .
```

Processing of Haemoproteus tartakovskyi data

Clean geneome sequence

The genome of Haemoproteus tartakovskyi was sequenced using 454 technology, with both shotgun and paired-end sequencing methods employed. The scaffold file for this genome is available on the course server under the name Haemoproteus_tartakovskyi.genome. Since the input reads come from both the bird and the parasite, bird scaffolds need to be removed. For the filtering process, I have set the GC content threshold at **28%**, and any scaffolds with GC content above this threshold should be excluded. Additionally, scaffolds shorter than **3000** nucleotides will also be removed. These steps will help in cleaning the data before performing gene prediction.

```
# Step2: Cleaning the genome sequence of Haemoproteus tartakovskyi
# Removing all scaffolds that have a GC content greater than 28% and a length shorter than
    3000 nucleotides.
echo "Removing all the scaffolds based on GC content and nucleotide content"
python3 "$scripts/filtering_scafolds.py" --fasta_file
    "$data/03_haemoproteus/Haemoproteus_tartakovskyi.raw.genome" --output_file
```

```
    "$data/03_haemoproteus/Haemoproteus_tartakovskyi_filtered.genome" --gc_content 28
    --scaffold_len 30000
echo "Sequence before filtering is $(cat
    $data/03_haemoproteus/Haemoproteus_tartakovskyi.raw.genome | grep "^>" | wc -l)"
echo "Sequence after filtering is $(cat
    $data/03_haemoproteus/Haemoproteus_tartakovskyi_filtered.genome | grep "^>" | wc -l)"
```

Make a gene prediction

With the new genome file, make a gene prediction. You will probably still have some scaffolds that derive from the bird. These should be short.

```
# Step3: Gene Prediction for the new haemoproteus tartakovskyi filtered file.
mkdir $data/03_haemoproteus/haemoproteus_gene_prediction
echo "Running gene prediction on harmoproteus_tartakovskyi..."
## Min contig is 100000, is because we have the genome of the host and the parasite as well.
    We need to keep only the shorter contigs, but not too short ones because that would lead
    to no proper results.
gmes_petap.pl --ES --sequence
    $data/03_haemoproteus/Haemoproteus_tartakovskyi_filtered.genome --min_contig 10000
    --work_dir $data/03_haemoproteus/haemoproteus_gene_prediction --cores 60
```

> **Problem 2: Why are the remaining bird-derived scaffolds short?**
>
> We are filtering out bird-derived contigs and retaining only parasite contigs with a GC content between 19-45%. Additionally, we keep only those contigs that are longer than 3000 base pairs.

To create fasta sequences from the gff file and the genome file use gffParse.pl found at the course server. It's recommended that you use the -c option.

```
# Formatting the gtf file for the perl script to run
cat
    $data/03_haemoproteus/haemoproteus_gene_prediction/haemoproteus_tartakovskyi_gene_predict.gtf
    | sed "s/ length=.*\tGeneMark.hmm/\tGeneMark.hmm/" >
    $data/03_haemoproteus/haemoproteus_gene_prediction/haemoproteus_tartakovskyi_gene_predict_formatted.gt

# Create FASTA sequences from the gff file.
perl $bin_dir/gffParse.pl -c -p -F -i
    $data/03_haemoproteus/Haemoproteus_tartakovskyi_filtered.genome -g
    $data/03_haemoproteus/haemoproteus_gene_prediction/haemoproteus_tartakovskyi_gene_predict_formatted.gt
mv $scripts/gffParse* $data/03_haemoproteus/
```

Remove scaffolds that have genes that are from avian origin. For this purpose use blastx or blastp on the course server. Use SwissProt as database. It is already installed with that name (SwissProt). If you want to look for the path to the BLAST databases, do:

echo $BLASTDB

If this environmental variable is set you don't have to specify the path, it's enough giving the database name like: -db SwissProt. As you experience the BLAST search takes a lot of time. SwissProt's size is

only 1% to that of UniProt so even on a more powerful computer the search will take considerable time. If you don't want to wait download the Ht.blastp file from the course server. Since the students have chosen different thresholds for the GC-content, the BLAST output file maybe not reflects the fasta file entirely.

```
## Step4: Blast Search
#mv $data/03_haemoproteus/gffParse.fna
    $data/03_haemoproteus/haemoproteus_tartakovskyi_gene_predict_formatted.fna
#mv $data/03_haemoproteus/gffParse.faa
    $data/03_haemoproteus/haemoproteus_tartakovskyi_gene_predict_formatted.faa

#echo "Running Blast.."
#blastx -query $data/03_haemoproteus/haemoproteus_tartakovskyi_gene_predict_formatted.fna
    -db SwissProt -out $data/04_blast/haemoproteus_tartakovskyi_blast.blastx -outfmt 6
    -evalue 1e-5 -num_threads 35
# Create a symlink for the taxonomy file
#ln -s /resources/binp29/Data/malaria/taxonomy.dat $data/04_blast/taxonomy.dat
#ln -s /resources/binp29/Data/malaria/uniprot_sprot.dat $data/04_blast/uniprot_sprot.dat

# A script is provided to retrieve the host scaffolds
#echo "Script to retrieve the host scaffolds"
#python $scripts/datParser.py $data/04_blast/haemoproteus_tartakovskyi_blast.blastx
    $data/03_haemoproteus/haemoproteus_tartakovskyi_gene_predict_formatted.fna
    $data/04_blast/taxonomy.dat $data/04_blast/uniprot_sprot.dat
```

The scaffolds.txt file did not contain any sequneces.

Problem 3: Insert the missing data in the above table. Use bash, not inter- net!

Species	Host	Genome Size (bp)	Genes	Genomic GC (%)
Plasmodium berghei	Rodents	17,954,629	7,235	23.71
Plasmodium cynomolgi	Macaques	26,181,343	5,787	39.08
Plasmodium falciparum	Humans	23,270,305	5,207	19.36
Plasmodium knowlesi	Lemurs	23,462,346	4,953	37.54
Plasmodium vivax	Humans	27,007,701	5,682	42.20
Plasmodium yoelii	Rodents	22,222,369	4,919	20.78
Toxoplasma gondii	Humans	128,105,889	15,892	52.20

Table 2: Genomic characteristics of various apicomplexan parasites, including genome size, number of genes, and GC content.

Problem 4: Compare the genome sizes with other eukaryotes and bacteria. Discuss with your partner (that is student partner) the reason for the observed genome sizes.

The genome sizes of Plasmodium species and Toxoplasma gondii range from approximately 18 Mb (megabases) to 128 Mb. In contrast, bacterial genomes typically range between 1–10 Mb, making these apicomplexan genomes significantly larger. However, compared to other eukaryotes, such as mammals (which often have genomes in the range of 2–3 Gb (gigabases)) and plants (which can

exceed 10 Gb), their genomes are relatively compact. This variation in genome size is influenced by factors such as life cycle complexity, gene content, and evolutionary pressures. Many parasitic organisms experience genome reduction due to their reliance on host organisms, leading to the loss of genes that are no longer necessary for survival. However, some parasites retain relatively large genomes due to expanded gene families involved in host interaction, immune evasion, and life cycle regulation.

Problem 5: What may cause the biased GC-contents in some of the species?

The variation in GC content among species can be influenced by several evolutionary and biological factors:

1. Mutation and Selection Pressure: Different species experience varying rates of mutation, with some favoring GC-to-AT mutations and others maintaining higher GC content due to selective advantages.
2. Genome Stability: GC-rich regions contribute to higher thermal stability of DNA, which can be beneficial in certain environmental conditions.
3. Codon Usage and Gene Expression: Some organisms exhibit GC-biased codon usage, which can affect transcription efficiency and translation accuracy, impacting protein expression levels.
4. Recombination Frequency: GC-rich regions are often associated with higher recombination rates, which can influence genome evolution and adaptation.
5. Parasitic Lifestyle Adaptation: Some parasites undergo genome reduction, which can lead to skewed nucleotide compositions depending on the genes retained or lost during host adaptation.

Phylogenetic Trees

Identify othrologs

Species trees are in the majority of cases inferred from gene or protein trees. However, it is today well known that a single gene may show a different topology than the "true' ' species tree. One important criteria when using genes for reconstructing a species tree is that orthologous genes, or in short orthologs, are used. Orthologs are descendants from the same gene in an ancestor of the two species. Paralogous genes, or paralogs, derive from a gene duplication within the species. It is complicated to identify orthologs. There are four main tools used for this purpose:

1. OrthoMCL (probably the most well known)
2. InParanoid
3. proteinortho
4. BUSCO

We will use the last two. proteinortho uses a reciprocal BLAST or Diamond search to find orthologs, while BUSCO uses HMM to identify a core set of proteins that normally should be found in the lineage that the species under investigation belong to. In the case of BUSCO input sequences could be a genome

sequence, a trancriptome or a proteome. We will compare the output of the two programs. Print the protein sequences in fasta format for each genome with the aid of the gffParse.pl program downloaded earlier. Use the -b option to give the protein file a shorter name as well as the -c option. Use Pb for Plasmodium berghei, Pc for Plasmodium chabaudi and so on. Then install proteinortho. This can be done with conda. More details about the program and installation can be found at: https://www.bioinf.uni-leipzig.de/Software/proteinortho/

```
nohup proteinortho6.pl {Ht,Pb,Pc,Pf,Pk,Pv,Py,Tg}.faa
```

Problem 6: What does the curly braces notation stand for?

The curly braces notation Ht,Pb,Pc,Pf,Pk,Pv,Py,Tg in the context of a bash command line stands for brace expansion or alternation.
It's a feature of the shell that generates multiple command-line arguments from a single pattern. In this specific case, it expands to a list of filenames, effectively running proteinortho6.pl on each of the specified FASTA files.

While running take a look at the output files, especially myproject.blast-graph. If you run top and then type c you will see the entire blast command that proteinortho uses. (note: On the new server, the proteinortho run will be fast if many cores are free, so you might not 'get time' to do this inspection) When the first blast results have been produced inform the teacher and continue with the BUSCO analysis. Busco can be installed using conda. More information about the software and installation can be found at https://busco.ezlab.org/.

```
# Example on how to run BUSCO:
busco -i ../Fasta/Pb.faa -o Pb -m prot -l apicomplexa
```

what does the flag -l stands for? and why do we choose apicomplexa? This time also Toxoplasma should be run. Take a look at the full_table_??.tsv files. We will only use the proteins that have been flagged as Complete or Duplicated. When it comes to the duplicated genes chose only one of these so that you in the end get one to one orthologs.

Problem 7: Compare how many BUSCOs (orthologues proteins) that are found in each proteome. Do the investigated parasites have close to complete numbers of BUSCOs?

Species	Complete BUSCOs	Percentage
Plasmodium berghei	367	76.3
Plasmodium cynomolgi	468	97.3
Plasmodium falciparum	476	99.0
Plasmodium knowlesi	349	72.6
Plasmodium vivax	480	99.8
Plasmodium yoelii	473	98.3
Toxoplasma gondii	382	79.4

Table 3: Complete BUSCOs

Problem 8: Is the assembly of the Haemoproteus tartakovskyi genome a reasonable approximation of the true genome?

The BUSCO results indicate 62.4% genome completeness, with 7.3% fragmented and 30.4% missing genes:

```
C:62.4%[S:62.0%,D:0.4%],F:7.3%,M:30.4%,n:481
300   Complete BUSCOs (C)
298   Single-copy BUSCOs (S)
2  Duplicated BUSCOs (D)
35 Fragmented BUSCOs (F)
146   Missing BUSCOs (M)
```

The assembly is moderately complete but lacks 30.4% of expected genes. Some sequences may be misassembled or missing, requiring further refinement. Additional long-read sequencing and hybrid assembly approaches could improve completeness.

Problem 9: How many of the BUSCOs are found in all eight organisms?

146 BUSCOs were common amongst the 8 different species.

Problem 10: If Toxoplasma is removed, how many BUSCOs are shared among the remaining seven species. Interpret!

176 BUSCOs were common amongst the 7 different species.

Appearently, we can't use the Toxoplasma data. We included this species to serve as an outroup for the phylogenetic analysis. By using bash or Python, create one protein fasta file per BUSCO based on the eight full_table_??.tsv files. In each file all eight species should be present with one sequence. We are only using one to one orthologs (why?). For information on what gene corresponds to a particular BUSCO, look in the directory hmmer_output. Use the same sequence id for the same organism in all created files, like Pf. Output file names should probably be the names of the BUSCOs.

Alignment and Trees

We will now make alignments for all individual BUSCO fasta files. Use clustalo for this. Then run raxml for all the alignments. Make a shell script for this. The trees can be visualized at: http://itol.embl.de/ You can install clustalo and BUSCO using conda: conda install -c bioconda clustalo raxml Run clustalo on the BUSCO protein fasta files you created. The -o is the option for specifying an outgroup. Here we choose Toxoplasma. Why? If you excluded Toxoplasma from the alignment files, choose another suitable outgroup. Use consense from the phylip package to merge all individual trees conda install -c bioconda phylip

Problem 11: Does all protein trees reflect the "true' ' species tree?

No, it is highly unlikely that all protein trees will perfectly reflect the "true" species tree.

1. Gene Tree vs. Species Tree: Gene trees represent the evolutionary history of individual genes, while the species tree represents the evolutionary history of the species. Gene trees can dif-

fer from the species tree due to gene duplication, gene loss, horizontal gene transfer, and incomplete lineage sorting.

2. Phylogenetic Signal: Some genes may have a stronger phylogenetic signal than others. Genes with high rates of evolution or subject to strong selection may produce trees that are inconsistent with the species tree.

3. Methodological issues: The methods used to create the gene trees (alignment, tree building) can introduce errors.

4. Stochasticity: Evolution is a stochastic process.

A consensus of many gene trees is often used to approximate the species tree, as it can help to minimize the effects of individual gene tree errors. Single gene trees are useful for understanding the evolution of a particular gene.

Problem 12: What is the phylogenetic position of Plasmodium falciparum?

Plasmodium falciparum (Pf) is closely related to Plasmodium yoelii (Py), Plasmodium berghei (Pb), Plasmodium knowlesi (Pk), Plasmodium vivax (Pv), and Plasmodium chabaudi (Pc). It forms a clade with these species, suggesting a relatively recent common ancestor compared to the other species in the tree, Haemoproteus tartakovskyi (Ht) and Toxoplasma gondii (Tg).
Specifically, Pf is basal to the clade containing Py, Pb, Pk, Pv, and Pc, indicating it diverged slightly earlier from the common ancestor of those five species.

Problem 13: Do you think that the GC contents have an impact on the tree topology?

Yes, GC content can potentially impact tree topology.

1. Codon Bias: GC content can influence codon usage bias, which can affect amino acid composition and therefore the inferred protein sequences.

2. Evolutionary Rate Regions of the genome with high GC content may have different evolutionary rates than regions with low GC content.

3. Alignment Artifacts: Differences in GC content can lead to alignment artifacts, which can in turn affect tree topology.

4. Long-branch attraction If GC bias is extreme, it can cause long branch attraction.

If there is a strong correlation between GC content and tree topology, it may be necessary to use methods that account for GC content bias. It is important to consider the potential impact of GC content when interpreting phylogenetic trees.

Problem 14: Do you think that the host range has an impact on the tree topology?

Yes, host range can potentially impact tree topology.

1. Adaptation: Adaptation to different hosts can lead to convergent evolution, which can obscure true phylogenetic relationships.

2. Gene Transfer: Host-parasite interactions can facilitate horizontal gene transfer, which can complicate phylogenetic analysis.

3. Selection: Host range can exert selective pressure on parasite genomes, leading to changes in gene content and sequence.

If there is a strong correlation between host range and tree topology, it may be necessary to use methods that account for host-associated adaptation. It is important to consider the potential impact of host range when interpreting phylogenetic trees.

Problem 15: Are the BUSCO proteins also found as orthologs in the proteinortho output?

Problem 16: Make a script that concatenates the alignments for each organism and BUSCO into one fasta file that in the end should contain seven sequences. Alternatively, use bash.