**Comparative Genomics Practical 02**

**Gene Prediction**

**Group 6: Tianlin He Xueqing Wang**

**Summary**

In this practical, we make use of the bioinformatics tools GLIMMER and GENSCAN to predict and locate genes from the five genomes that we were given. With GLIMMER, we first extracted long ORF from the genome, and trained the interpolated Markov model with these ORFs. Together with determining the location of ribosome binding site and frequency of start codon, the prediction can be obtained.[[1]](#endnote-1) Using the prediction result, we plotted a histogram which illustrates the frequency versus gene size. In addition to GLIMMER, GENSCAN is a probabilistic-based model which enables us to identify multiple eukaryotic genes according to general compositional features of known eukaryotic genes. The amino acid sequence fragments of genes are given out separately after running prediction. We further converted the predicted protein sequences into original DNA sequences, thus the genes on the genome are predicted.

**Glimmer:**

**1. Find long ORF**

tigr-glimmer long-orfs -n -t 1.15 09.fa.txt 09.long-orf-coords

**2. Extract long ORF**

tigr-glimmer extract -t 09.fa.txt 09.long-orf-coords > 09.longorf

**3. Prepare training set**

tigr-glimmer build-icm -r 09.icm < 09.longorf

**4. Run Glimmer**

tigr-glimmer glimmer3 -o50 -g110 -t30 09.fa.txt 09.icm 09.glimmertigr

**5. Long ORFs are provided to construct training set, what other two sources of**

**sequences can be used instead of or in addition to long ORFs?**

Annotated “known” genes from NCBI complete genomes can be used as an alternative; start codon frequencies and ribosome binding site motif can be used as supplement to long ORFs.

**6. Is Glimmer suitable for all genomes? Why?**

Glimmer is only suitable for bacterial genes, but not suitable for eukaryotic genes. This is because of the difference in gene sequence and transcription mechanism between eukaryotes and prokaryotes.

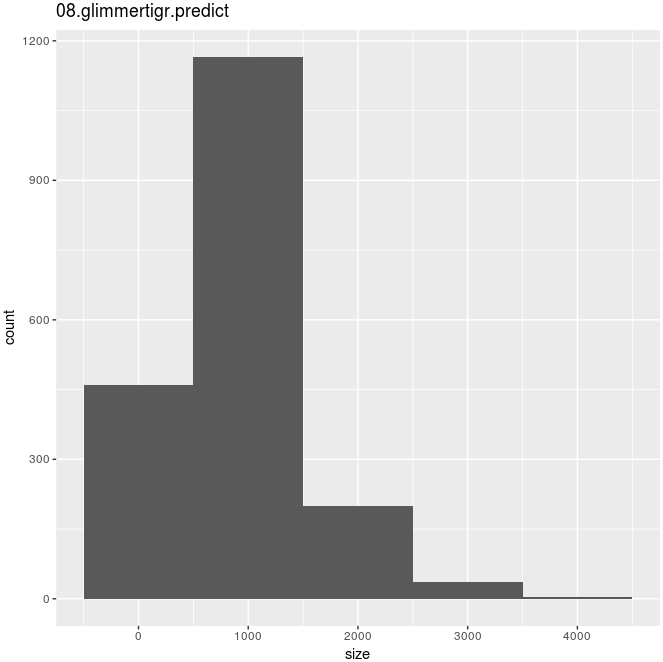
- Eukaryotic genes contain non-coding introns and coding exons, but prokaryotic genes are intronless

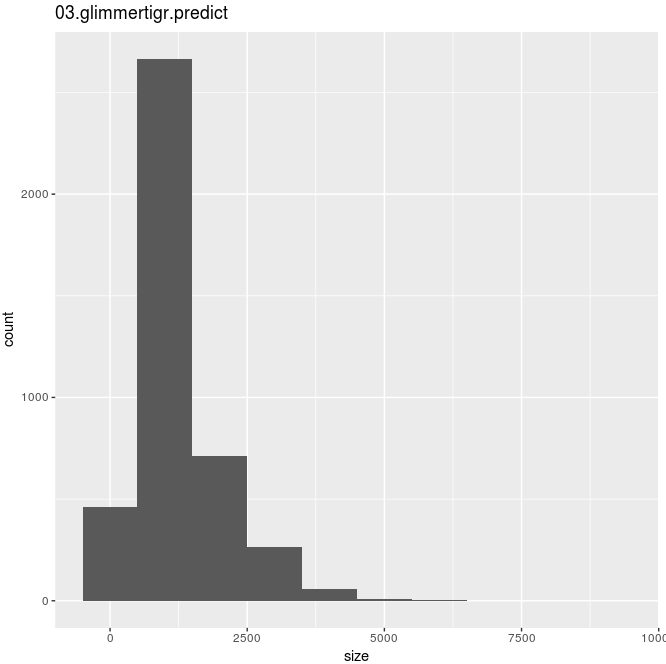
- Eukaryotic genes carry out alternative splicing, so that a same segment of gene can give rise to different protein products; on the other hand, prokaryotic genes are mostly not overlapped.

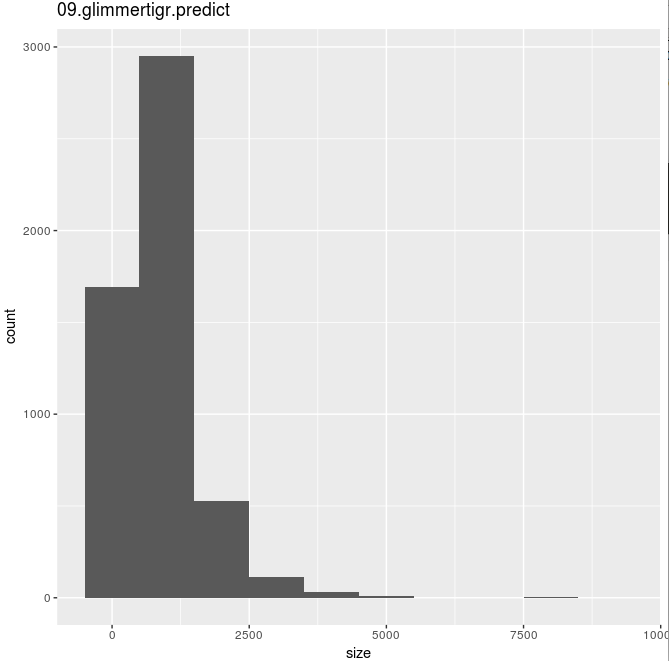
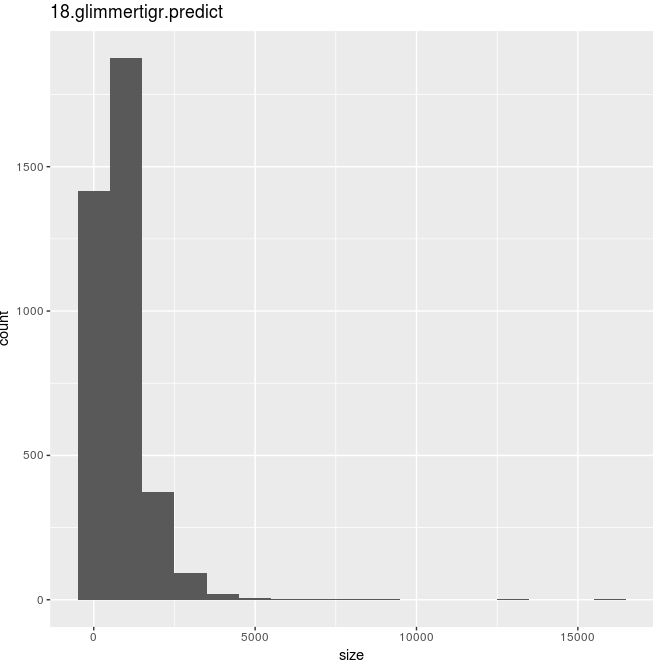
- Prokaryotic genome are gene-dense, which means gene prediction can be easily achieved with high accuracy but also challenged by high false-positives; while less than 10% of eukaryotic genomes are coding, it is more difficult to identify genes.

- Start codon in eukaryotes is Methionine instead of f-Met.

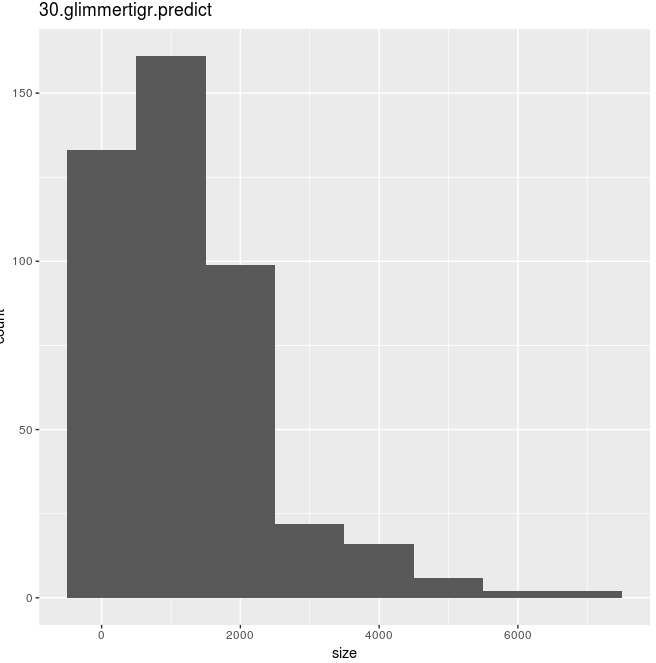
**7. Make a histogram of predicted gene lengths for each genome in R**

03: B. thetaiotaomicron 08: D. turgidum



 09: E.coli 18: Synechocystis sp.

30: S. cerevisiae



**8. Do all gene sizes follow the same distribution in all genomes?**

Yes. In all five genomes, all of them have the highest frequency at size around 1000 bases. On the right side of this peak, frequency drops dramatically when gene size increase. The largest gene size that a genome can attain ranges from 4000 to 6000, except for there are few outliers with abnormally large size for 09 (E. coli) and 18 (Synechocystis sp.)

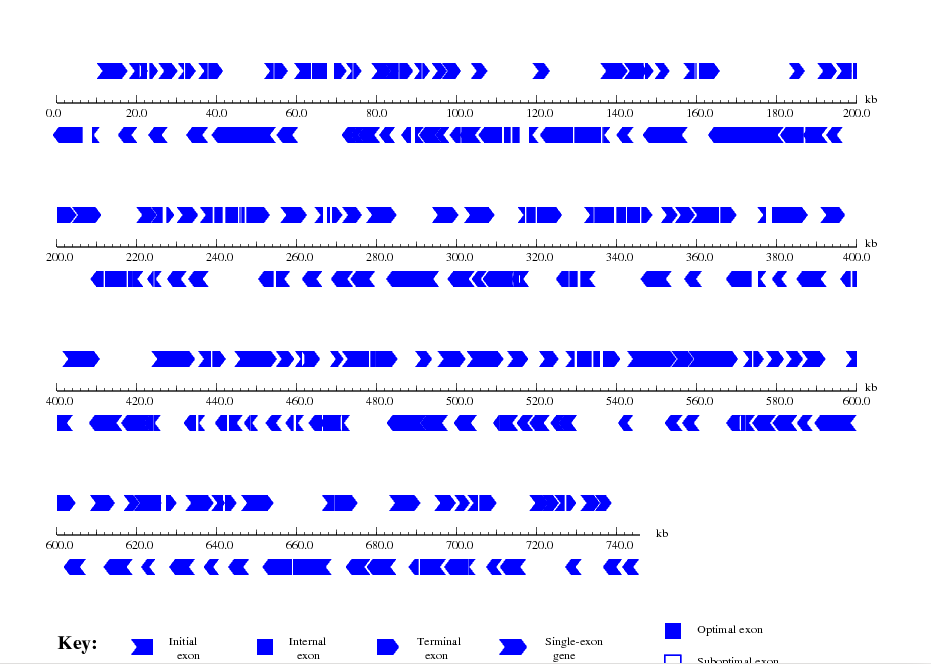
**GENSCAN**

1. **Run GENSCAN for the eukaryote genomes using HumanIso.smat.**

In the terminal, firstly go to the directory where Genscan is installed. Then run the eukaryota genome file using HumanIso.smat by typing:

./genscan [path for HumanIso.smat] [path for 30.fa.txt] -v –ps [path\_for\_output\_plot\_file/name\_of\_plot\_file.ps]

We got the gene prediction as follows, as well as the specific amino acid codes for different genes.



1. **GENSCAN requires a hefty amount of memory, what implications might this have?**

The computer chosen must have enough memory in order to run sequences efficiently. As a rule, to run sequences with efficient length of N kb, the computer should have at least N/2 Megabytes of RAM. When we want to try longer sequences the outcomes may be inaccurate, in which case we can separate the whole work into several parts and test them one by one.[[2]](#endnote-2)

The reason that GENSACN needs huge memory might be interpreted as that it is aimed for eukaryotic gene prediction in the way of complete exon/intron identification. Thus it needs more calculation power than prokaryote gene prediction.

1. It might be good to try with short test file.

**Discussion**

GLIMMER and GENSCAN are both capable of identifying genes from genome and both of them are based on the probabilistic Markov model. However, they are different in a way that GLIMMER is specialized in identifying genes of prokaryotic origins, while GENSCAN is able to find out eukaryotic gene components. In addition, GLIMMER runs locally as it can be trained with the long ORFs generated from the test genome; while in GENSCAN, model parameters are obtained from database constructed from collection of known genes and cDNAs.[[3]](#endnote-3)

Analyzing the result from GLIMMER, it is found that length of genes predicted mainly fall in the range from a few hundreds to one or two thousands bases, which is compatible with our understanding of prokaryotic genes.[[4]](#endnote-4) Therefore, the outliers from no. 9 and no. 18 should be re-assessed due to their uncommon large size. On the other hand, results from no. 30 should be excluded as it is eukaryotic, not prokaryotic.

To obtain results from GENSCAN, we first go to the directory where GENSCAN is installed, and run the eukaryotic genome (no. 30) file using HumanIso.smat. Here the flag “-v” is used to show all the gene-coding sequences out on the terminal, while flag “-ps” is used to plot out the outcome of the whole genome. After running it we get a postscript plotting output as well as the specification of all peptide sequences. The amino acid sequences were saved into a fasta file. GENSCAN gives out the specific predicted distribution of initial exon, internal exon, terminal exon and so on, for the given eukaryotic genome.

1. Arthur L. Delcher, Kristen A. Bratke, Edwin C. Powers, Steven L. Salzburg. Identifying bacterial genes and endosymbiont DNA with Glimmer. 2007. *Bioinformatics*: 6, 673-679. [↑](#endnote-ref-1)
2. Christopher Burge. 1997. *Documentation for Command-line Version of GENSCAN*. <http://genes.mit.edu/README> [↑](#endnote-ref-2)
3. Chris Burge and Samuel Karolin. 1997. Prediction of Complete Gene Structures in Human Genomic DNA. *Journal of Molecular Biology*: 268, 78-94. [↑](#endnote-ref-3)
4. L Xu, H Chen, X Hu, R Zhang, Z Zhang, ZW Luo. 2006. Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Mol. Biol. Evol.*: 23, 1107-1108. [↑](#endnote-ref-4)