KB8019 Comparative Genomics
Group 6 Xueqing Wang Tianlin He

Constructing an ORF prediction and phylogenetic analysis of five unknown genomes

1. Introduction

With the recent advancement in sequencing technologies, the whole genome database is experiencing an explosion: 14 years after the completion of human genome project, nowadays, we have the complete genomic information of more than 20,000 species. However, the biggest challenge now becomes how we make sense of these astronomical amount of data: for example, the phylogenetic relationship between these species remains to be a question.[i] In this project, we took the approach of constructing multiple distance matrices, making use only the composition of five unknown genomes that we were given. In addition, we also explored their proteome compositions, such as the amino acid frequencies and differential usage of stop codons in different species, which can be achieved through building an opening reading frame (ORF) predictor that reads a genome and outputs its encoding proteome based on the biological features of protein-coding regions. [ii]

To begin with, we determined the identities of those 5 genomes by blasting it against the NCBI database. The results are shown below in Table 1:

	Species	Kingdom	Proteome Size
03	B. *thetaiotaomicron*	Prokaryote	4816
08	D. *turgidum*	Prokaryote	1742
09	E. *coli*	Prokaryote	4140
18	Synechocystis *sp.*	Prokaryote	3513
30	S. *cerevisiae* (chr. X)	Eukaryote	405

(For simplification, these 5 species were named 03, 08, 09, 18, 30 in this article.)

The core of our phylogenetic analysis is constructing an ORF predictor. ORF is a fraction of sequence on the genome of a species that have the potential to produce a certain peptide sequence. It begins from the start codon and stops right before the stop codon. Such a fraction has the capacity to undergo all the way from transcription to translation, encoding the final protein with physiological functions.

ORFs play a crucial role in exploring the genomes of different species: they build a bridge between nucleotide sequences and peptides, filtering out the “trash” section on the genome and reserve the actual encoding parts that have significant biological meaning. Thus, the detection of ORFs has long been a study field of interest. [ii] Reliably predicted ORFs can reveal the true coding parts in a species’ genome, and by comparing the ORFs of different species we can further infer the evolutionary relationship between them.

Both prokaryotes and eukaryotes require promoters (a fraction of DNA strand which initiates transcription) to begin transcription. In prokaryotes, most of the promoters contain two short elements at roughly -10 and -35 nucleotide positions. [iii] The -10 site (the Pribnow box) has consensus sequence TATAAT, while the -35 site bears a consensus sequence TTGACA. Although “consensus” as their names suggest, there are actually few sequences in real promoters that fulfill these sequences: most organisms only have 3 to 4 bases in accordance with these sequences in both sites; in other words, these consensus sequences are only profile of the most probable nucleotides. The majority of the putative TSSs were located between 20 to 40 nucleotides from the translational start site. [iv] Moreover, before the start codon (AUG) on mRNA, the ribosome binding sites (RBSs) are needed to initiate translation. In prokaryotes this is called Shine-Dalgarno sequence, which is the complementary sequence of 16s rRNA 3’ end that binds upstream to the start codon. It is usually around 8-base upstream of the start codon.

When it comes to eukaryotes, the identifying of signals becomes more challenging. For the transcription signals there are TATA box, CAAT box, regulatory sequences, general transcription factors and so on which can form complex with surrounding molecules to initiate transcription. [v] Before the start codon, Kozak consensus sequence is always observed, with high conservation falling on -3 and +4 position. [vi] The pattern of Kozak sequence is illustrated below in Fig. 1.

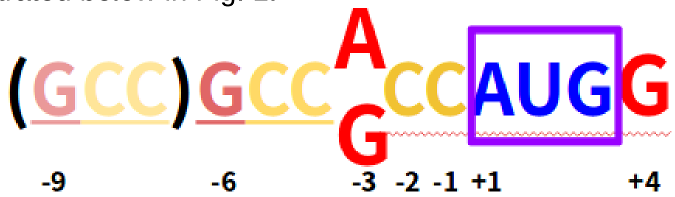


Figure 1: Kozak Consensus Sequence

With whole genome sequences and predicted ORFs available, we obtained various distance matrices by calculating the difference in genome and proteome composition in these 5 organisms (the principle and formula are detailed in *methods*). During the time course of evolution, sequences that descend from a common ancestor experience mutation, therefore, their distance can be reflected by the deviation in nucleotide and amino acid frequencies, which is in terms a classical way to construct a phylogenetic tree. Here, single and di-nucleotide frequencies, GC content from genome as well as single amino acid frequencies from predicted ORFs were used; in addition, we also investigated into the differential usage of three stop codons in these 5 organisms, and construct a distance matrix from it.[vii]

Finally, with the use of the program belvu, we compared the phylogenetic trees generated from the distance matrices to that from multiple sequences alignment of 16s rRNA, which is widely considered as standard in phylogenetic study.[viii]

2. Methods

2.1 Composition of Genome and Proteome

The frequencies of single, di-nucleotide, GC content, single amino acid, di-amino acid, stop codons, were calculated using python scripts Nucleotide.py, Dinucleotide.py, Aminoacid.py, DiAminoAcid.py and Distance_Stopcodon.py respectively.

2.2 ORF

All six reading frames are taken into consideration in our ORF predictor, with both direct and reverse strands encoding proteins. Apart from the start and stop codon, for prokaryotes we tested the Shine-Dalgarno sequence with the examination for the whole consensus sequence (AGGAGGT) and the partial sequence (GAGG), on the upstream regions of ATG ([-50, -2]) and changed the parameters to get better outcomes. In the final version of the predictor, (GAGG) is searched in the upstream internal [-50,-2] of genomes 03 and 08, while [-30,-2] for genomes genomes 09 and 18. For the eukaryote, we searched for A and G on position -3 as well as G on position +4 of the start codon.

For the evaluation of the result, we compared our predicted ORFs to the annotated proteomes of the identical species (downloaded from NCBI or UniProt), counting for the difference in total numbers. We run BLAST locally with our predicted proteomes against the annotated ones, counted for the e-value of the alignment of each sequence, and calculated the percentage of sequences with e-value less than 10^{-5} in all the predicted ORFs (accuracy) as well as in each annotated proteome (efficiency). The python scripts for this part is included in ORFforProkaryotes.py and ORFforEukaryote.py.

2.3 Distance matrix

Without loss of generality, distance between two subjects x and y simply refers to the operation such that a) d(x, y) > 0 if and only if x ≠ y; b) d(x,y) = d(y,x); c) d(x, x) = d(y, y).[ix] In this project, distance between the genomes or proteomes of 03, 08, 09, 18 and 30 were calculated according to the difference in frequency of 1) 4 nucleotide 2) 16 di-nucleotide 3) GC content 4) amino acid in predicted ORFs 5) 3 stop codons in predicted ORFs. Take nucleotide frequency as an example, the distance between organism 03 and 08 is defined with the following formula:

$$d(03,08)=d(08,03)=\sqrt{(A_{03}-A_{08})^2+(T_{03}-T_{08})^2+(C_{03}-C_{08})^2+(G_{03}-G_{08})^2}$$

The python script for calculating the 5 distance matrices were described in Distance_genome.py, Distance_gene.py and Distance_Stopcodon.py respectively.

2.4 Tree construction

Phylogenetic trees were built with the program belvu, which is a powerful distance estimator used by us for tree construction from multiple sequences alignment and bootstrapping during the labs. [x] Here, we only explored the simple function which takes up a distance matrix. The command line for running belvu is:

```
/common/courses/comparative_genomics/belvu_Ubuntu_12.04.3_64bit / -T R belvuInput.txt
```

3. Results

3.1 ORF Predictor

The translation signals are searched around the start codon, with variation in size of searching interval and the consensus percentage, so that more accurate prediction can be generated. The results of our ORF predictor together with the those from Glimmer/Genscan are shown Figure 2.

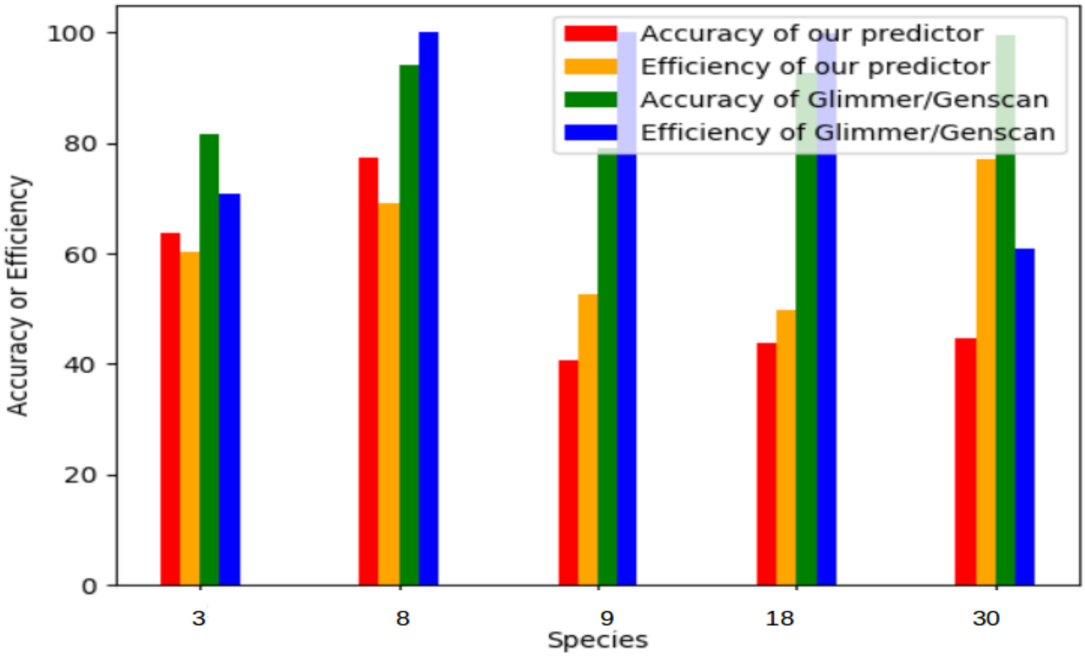
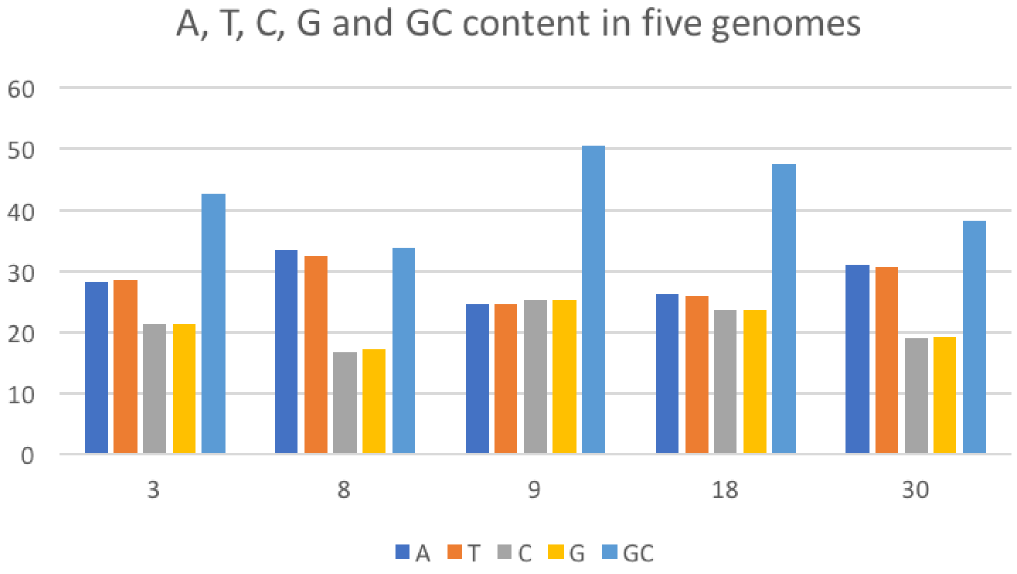


Figure 2: Comparison between our ORF Predictor, Glimmer and Genscan

3.2 Frequencies of nucleotides, amino acids and stop codons

The frequencies of 4 nucleotides A, T, C, G in all 5 genomes were illustrated in the below bar chart Fig. 3



Due to the limitation in space, only one value per 16 dinucleotides, 20 amino acids and 400 di-amino acids were displayed below in Table 2.

	3	8	9	18	30
Di-nucleotides					
AA	8.94%	12.72%	7.25%	9.13%	10.87%
Amino acids					
Alanine	6.7016%	5.2171%	9.4490%	8.3809%	5.5301%
Di-amino acids					
Ala-Ala	0.517076%	0.240027%	0.954983%	0.671384%	0.412511%
Stop codons					
TAA	53.8%	57.1%	48.8%	38.6%	42.2%

3.3 Distance matrices and Phylogenic Trees

Five 5 X 5 distance matrices were obtained using the frequencies calculated above. However, frequencies of di-amino acids were not included because of the enormous computational power it requires. Here, the distance matrix generated from differential usage of stop codons between five genomes were listed as an example:

0.0	0.065889375	0.124419828	0.225999921	0.144377705
0.065889375	0.0	0.189749114	0.239420054	0.198723869
0.124419828	0.189749115	0.0	0.268366617	0.116053547
0.225999921	0.239420054	0.268366617	0.0	0.160762898
0.144377704	0.198723869	0.116053547	0.160762898	0.0

In the matrix, there are five "0"s across the diagonal, and all numbers in the blocks range from 0 to 1, indicating that it fulfills the criteria listed in *methods.* These matrices were used to construct 5 phylogenetic trees using belvu. They are illustrated below from Fig.4a to 4e.

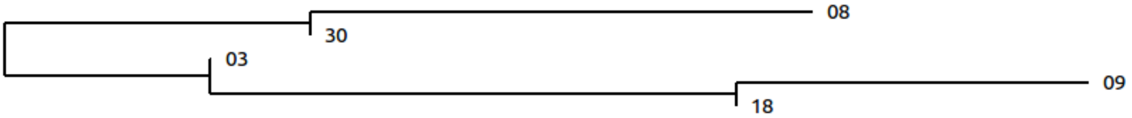


Fig. 4a Tree from nucleotide content

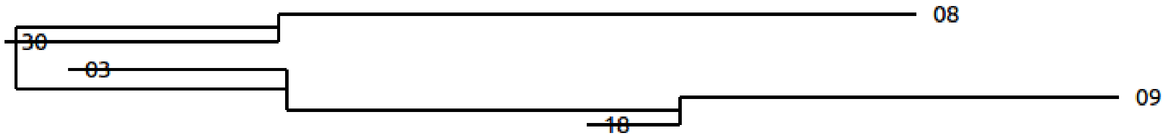


Fig. 4b Tree from di-nucleotide content

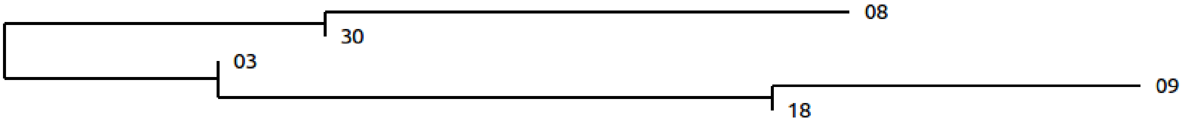


Fig. 4c Tree from GC content

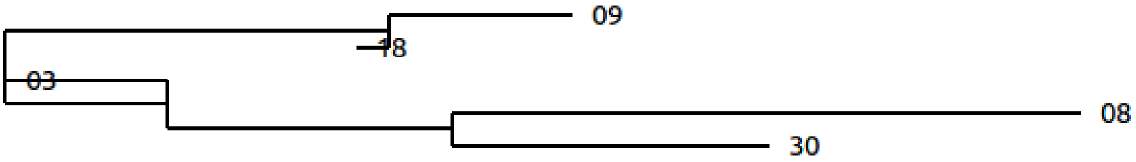


Fig. 4d Tree from amino acids frequencies

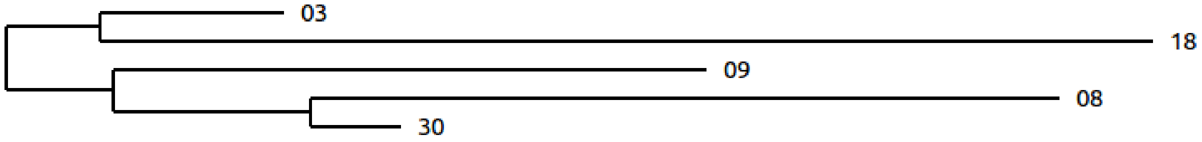


Fig. 4e Tree from 3 stop codons frequencies

4. Discussion

4.1 Open Reading Frames

For prokaryotes, we set the interval of Shine-Dalgarno sequence in a much larger range than nature does: when this sequence is supposed to be found at around position -8, we actually searched it all the way from -50 to -2. This is because that when we stuck to the actual positions we could only find a tiny fraction of all the actual ORFs, which can be interpreted by the characteristic of consensus sequence itself: it is actually not perfectly reserved, and thus should not be strictly restricted in our project. The searching criteria we used, although not having significant biological meaning, does have remarkable similarity to annotated proteomes in terms of the results.

However, when predicting prokaryote ORFs we found a large fraction of predicted ORFs which are similar to the same protein sequence in the annotated proteome. This might because that the ORFs we identified overlap severely (Overlapping is also a feature of prokaryotic genome), which has also been inspected in the XML file got out of local BLAST. We then removed the overlapping ORFs from our outcomes; the total number of predicted ORFs dropped but the accuracy is enhanced.

As for the eukaryotes, as stated before, we only counted for A and G on position -3 as well as G on position +4 of the start codon. This is also because of the biological fact of Kozak consensus sequence: the conservation on -3 and +4 have the most significant effect in the enhancement of translation efficiency. The other positions are more variable and therefore less effective than these two sites. Still, we predicted more sequences than the proteome from NCBI, which might because that we didn’t count for the other regulatory factors, as well as introns.

We also ran BLAST for the predicted proteomes of Glimmer and Genscan, and compared our results with them. We didn’t count for the transcription signal sites, as they have variable distances to the start codon and thus not easy to be put into the script.

4.2 Phylogenetic Tree Building Based on distance Matrices

If the tree obtained from 3 stop codons was not taken into consideration, a consensus tree can be built from the other 4 phylogenetic trees we generated using frequencies of nucleotides and amino acids in the genome or predicted ORFs, because they share a common topology:

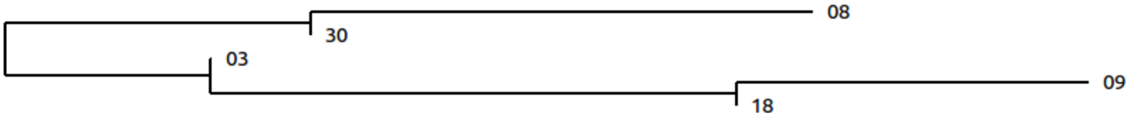


Fig. 5 Consensus tree from 5 “small” trees

Therefore, we are able to conclude that various compositional components in genomes and proteomes are highly consistent with each other, which means we could probably only take one of them to construct a distance matrix, saving plenty of time.

In order to assess the reliability of our consensus tree, we compare it with the phylogenetic tree generated from aligning multiple sequences of 16S rRNA in organisms 03, 08, 09 and 18, which was performed in practical 3. In this tree, organism 30 (baker’s yeast) is absent as eukaryotes use 18s rRNA in translational instead of 16s rRNA, therefore the alignment with other prokaryotes would be irrelevant.[xi]

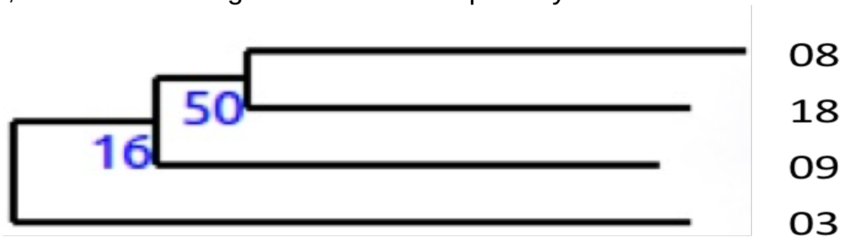


Fig. 6 Tree from 16s rRNA in organisms 03, 08, 09 and 18.

Unfortunately, it is found that our consensus tree resembles poorly the standard 16s rRNA tree. This result actually well illustrates the fact that only with partial information about the genome or proteome composition, it is not sufficient to draw a precise conclusion about the evolutionary relationship between species. To achieve maximum accuracy in hierarchy construction, it is believed that identification of conserved gene -> multiple sequence alignment -> calculating substitution rate would be the most reliable way.[xii] However, this method would undoubtedly require much more effort compared with distance matrix.

5. Conclusion

We successfully built an ORF predictor for four prokaryotes and one eukaryote. We took into account the translation starting signal (Shine-Dalgarno and Kozak sequence), without counting for the effect of transcription starting signals due to the limitation of time. In terms of accuracy, our predictions have great distinction to Glimmer/Genscan; as for efficiency, our predictor has produced more true ORFs than Genscan, with the cost of low accuracy. The distinctions can be interpreted from the difference in algorithm: Both Glimmer and Genscan are based on machine learning methods such as IMM score computation and training with long ORFs. [xiii] With the limitation of purely examining the regulatory signals before start codons on the genome, it's actually hard to improve the quality of predicted ORFs in our predictor; to acquire better result, machine learning and heuristic algorithms need to be considered.

To improve the accuracy of our distance matrix, information about coordination of bases in the genome or amino acids in the proteomes should be incorporated into the calculation, such as transformation distance, cosine similarity, compression distance, and even the corrected Kolmogorov complexity. [iv] The principle is that the more information we use to establish a distance matrix, the better it represents the whole genome.

References:

i Pavlopoulos GA, Soldatos TG, Barbosa-Silva A, Schneider R. (2010) A reference guide for tree analysis and visualization. B*ioData Mining* 3: 1.

ii Guo FB, et al. (2003) ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Research* 31:1780–1789.

iii Estrem ST, Ross W, Gaal T, Chen ZW, Niu W, Ebright RH, Gourse RL. (1999). Bacterial promoter architecture: Subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase alpha subunit. *Genes & Development* 13 (16): 2134–2147.

iv Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B. (2009) Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in *E. coli*. PLoS ONE 4(10): e7526.

v Jennifer B, Kadonaga JT. (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & Dev.* 16 (20): 2583–2592.

vi Kozak, M. (1987) At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J. Mol. Biol.* 196: 947–950.

vii Xu L, Kuo J, Liu JK, Wong TY. (2012) Bacterial phylogenetic tree construction based on genomic translation stop signals. *Microbial Informatics and Experimentation* 2: 6.

viii Ibrahim A, Goebel BM, Liesack W, Griffiths M, Stackebrandt E. (1993) The phylogeny of the genus Yersinia based on 16 S rDNA sequences. *FEMS Microbiol Lett* 114:173–177.

ix Li M, Badger JH, Chen X, Kwong Sam, Kearney P, Zhang H. (2001) An information-based sequence distance and its application to whole mitochondrial genome. *Bioinformatics* 17(2): 149-154.

x Sonnhammer EL, Hollich V. (2005) Scoredist: A simple and robust protein sequence distance estimator. *BMC Bioinformatics* 6:108.

xi Woese CR, Fox GE. (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *PNAS* 75(11): 5088-5090.

xii Zvelebil M, Baum JO. (2008) *Understanding Bioinformatics.* Garland Science.

xiii Delcher AL, Bratke KA, Powers EC, Salzberg SL. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer, Bioinformatics 23(6): 673-679.