Review of available software for gene prediction

Maciej Bielecki

Genome annotation: standard pipeline

Sequencing

↓

Quality control

↓

Assembly

↓

Gene prediction

↓

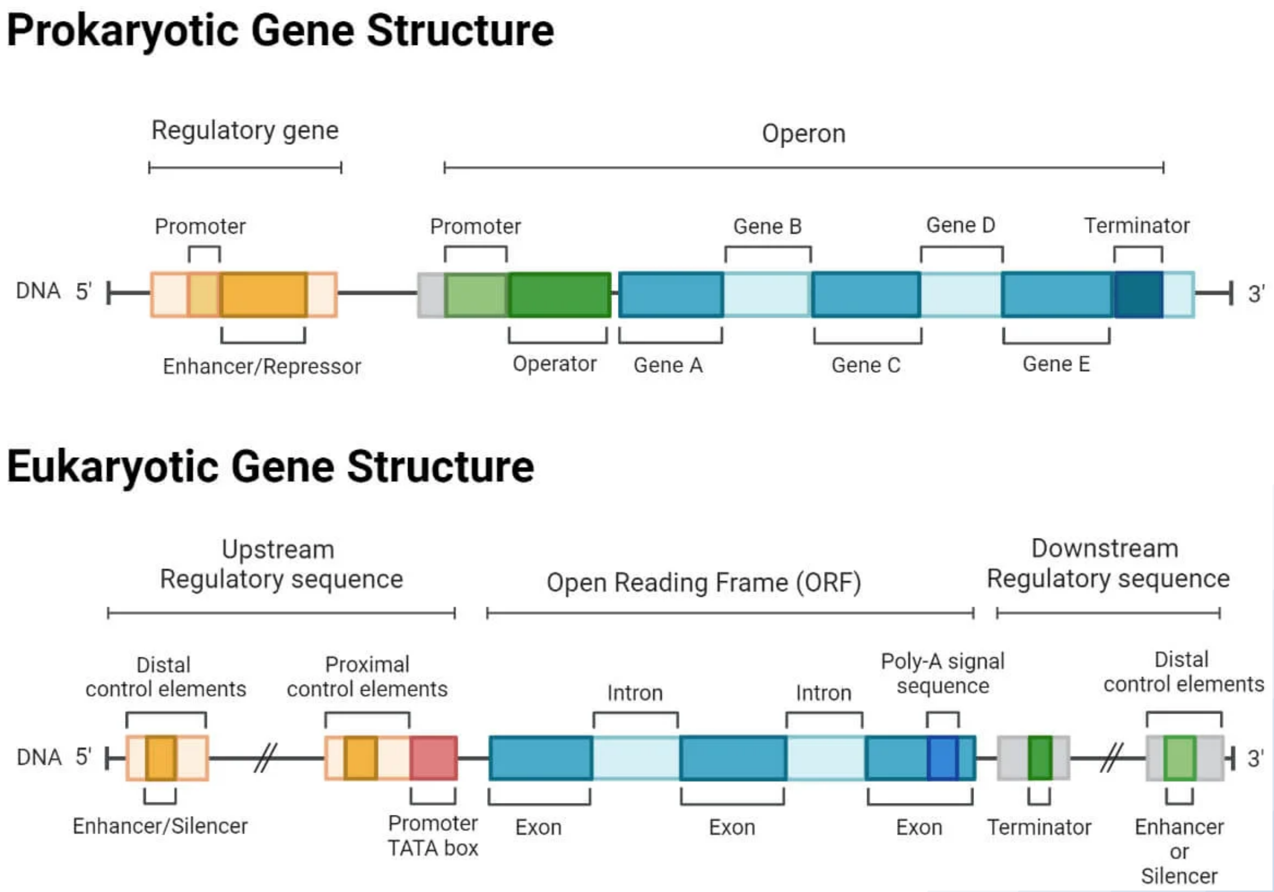
Functional annotation

Methods of gene prediction

- Similar sequence-based
  - Based on known sequences of homologue genes/mRNA/proteins
  - Less likely to produce false positives
  - Pre-existing data is required

- *Ab initio* methods
  - Based on processing sequence data intrinsically
  - Potentially more thorough detection
  - Higher likelihood of false positives

Prokaryotic and eukaryotic genes



Structure of prokaryotic and eukaryotic genes (Kulkarni, 2023)

GeneMark

- Suite of predictors for prokaryotes and eukaryotes as well as metagenomic and metatransciprtomic data
- Part of the NCBI’s pipeline for prokaryotic genome assembly, among other pipelines
- Developed at the Georgia Institute of Technology, proprietary

AUGUSTUS

- *Ab initio* predictor for eukaryotes, based on a generalized hidden Markov model
- Written in C++, open source (Artistic License)
- Available locally and remotely
- Part of various larger frameworks for genomic data processing

Prodigal

- *Ab initio* predictor designed for prokaryotic and metagenomic data
- Designed to be simple in use
- Written in C, open-source (GPL3)

GlimmerHMM

- *Ab initio* predictor for eukaryotes, based on a generalized hidden Markov model
- Written in C, open source (Artistic License)
- Highly efficient in regards to computation time and memory usage

GeMoMa

- Flexible homology-based predictor
- Written in Java, open-source (GPL3)
- Available as a Conda package and through a web server

GenomeScan

- Homology-based predictor, designed specifically for vertebrates, maize and *Arabidopsis*
- Developed in the Biology Department at MIT
- Only available through web server

Conclusions

- Seemingly low demand for this type of software
  - A lot of predictors are no longer being maintained
  - Ostensibly no new predictors being developed
- AUGUSTUS, Prodigal and GeneMark make up most of ”market share”
- Not much literature comparing existing software
- Benchmarking has proved *ab initio* methods highly sensitive

Based on this:
gene prediction is considered to be a solved problem?