Chapter 8

Gene Prediction

Overview

1. Introduction
2. Introduction to Biological Databases
3. Pairwise Sequence Alignment
4. Database Similarity Searching
5. Multiple Sequence Alignment
6. Profiles and Hidden Markov Models
7. Protein Motifs and Domain Prediction
8. Gene Prediction
9. Promoter and Regulatory Element Prediction
10. Phylogenetics Basics
11. Phylogenetic Tree Construction Methods and Programs
12. Protein Structure Basics
13. Protein Structure Visualization, Comparison and Classification
14. Protein Secondary Structure Prediction
15. Protein Tertiary Structure Prediction
16. RNA Structure Prediction
17. Genome Mapping, Assembly and Comparison
18. Functional Genomics
19. Proteomics

Why do we need gene prediction?

Whole-genome sequencing is increasingly common and leads to rapid growth of sequence databases, but mere genome sequences are not enough!

- Without further analysis, it is not clear where the actual genes are
- Error-free gene identification is a prerequisite for accurately deducing protein sequences and functions
- Computational gene prediction is amongst the most difficult problems in bioinformatics, particularly for eukaryotic genomes (splicing!)

Categories of gene prediction methods

- *Ab initio* (*i.e.* from first principles)
- Homology-based
- Consensus-based, *i.e.* a combination of both

Ab-initio gene prediction methods

Gene prediction based on the given sequence alone, making use of two phenomena:

1. **"Gene signals"**, *i.e.* sequence elements such as start and stop codons, intron splice signals, transcription factor binding sites, ribosomal binding sites and polyadenylation sites
2. **Nucleotide content**: composition and statistical patterns in coding regions differ significantly from those in non-coding regions

Homology-based gene prediction methods

Gene prediction based on on significant resemblance of the query sequence to known genes (*e.g.* from a closely related species whose genome has already been analysed and annotated)

- Underlying idea: if a translated DNA sequence is found to be highly similar to a known protein or protein family from a database search, this is a strong indication that the region encodes a protein
- Potential problem: *pseudogenes* (ancient genes that have become inactivated during evolution)

Gene prediction in prokaryotes

Prokaryotic genomes are easiest to deal with:

- Limited size: 0.5 – 10 Mbp
- High gene density: 90% of a genome consists of protein-encoding sequences
- Very few non-coding/repetitive sequences
- No introns! Each gene is composed of a single contiguous stretch of open reading frame (ORF), encoding a protein or RNA
- Prokaryotic genes usually comprise several readily recognisable sequence elements, both upstream and downstream of the ORF

Prokaryotes: upstream sequence elements

- Most genes have **ATG** for a start codon (**AUG** in mRNA, codes for methionine), sometimes **GTG** or **TTG**
- Genes have a *Shine-Dalgarno (SD)* sequence:
  - SD acts as a ribosomal binding site: purine-rich sequence (often **AGGAGGT**) complementary to 16 S rRNA in the ribosome
  - Often located immediately downstream of the transcription initiation site (although prokaryotic genes may be transcribed together as one operon), and always just upstream of the translation start codon

Prokaryotes: downstream sequence elements

- Three possible stop codons at the end of the protein coding region (**TAA**, **TGA** or **TAG**)
- Often there are several consecutive stops at the end of an ORF
- Genes are followed by a *transcription terminator* sequence: a palindromic sequence that forms a distinct stem-loop structure, which is followed by a string of T residues

Prokaryotic gene structure

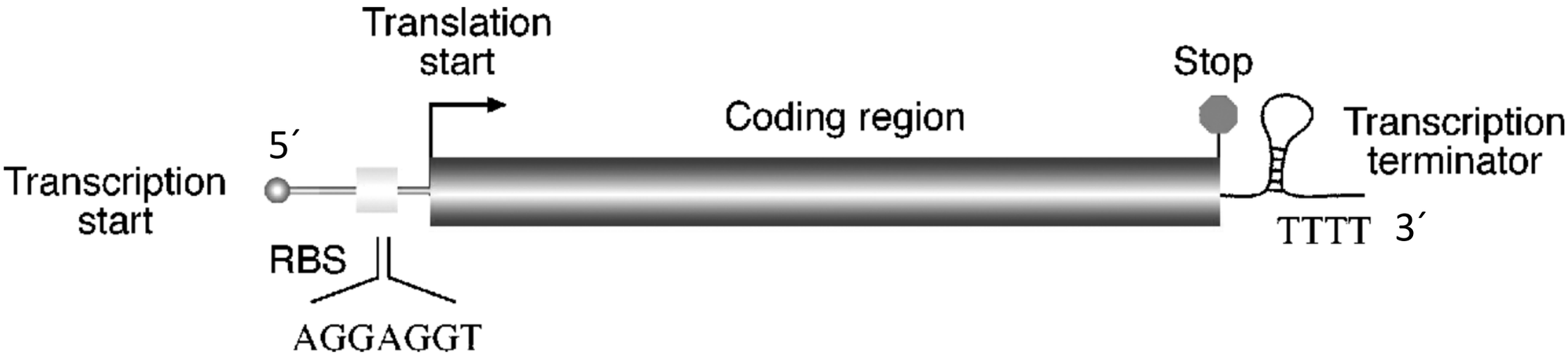


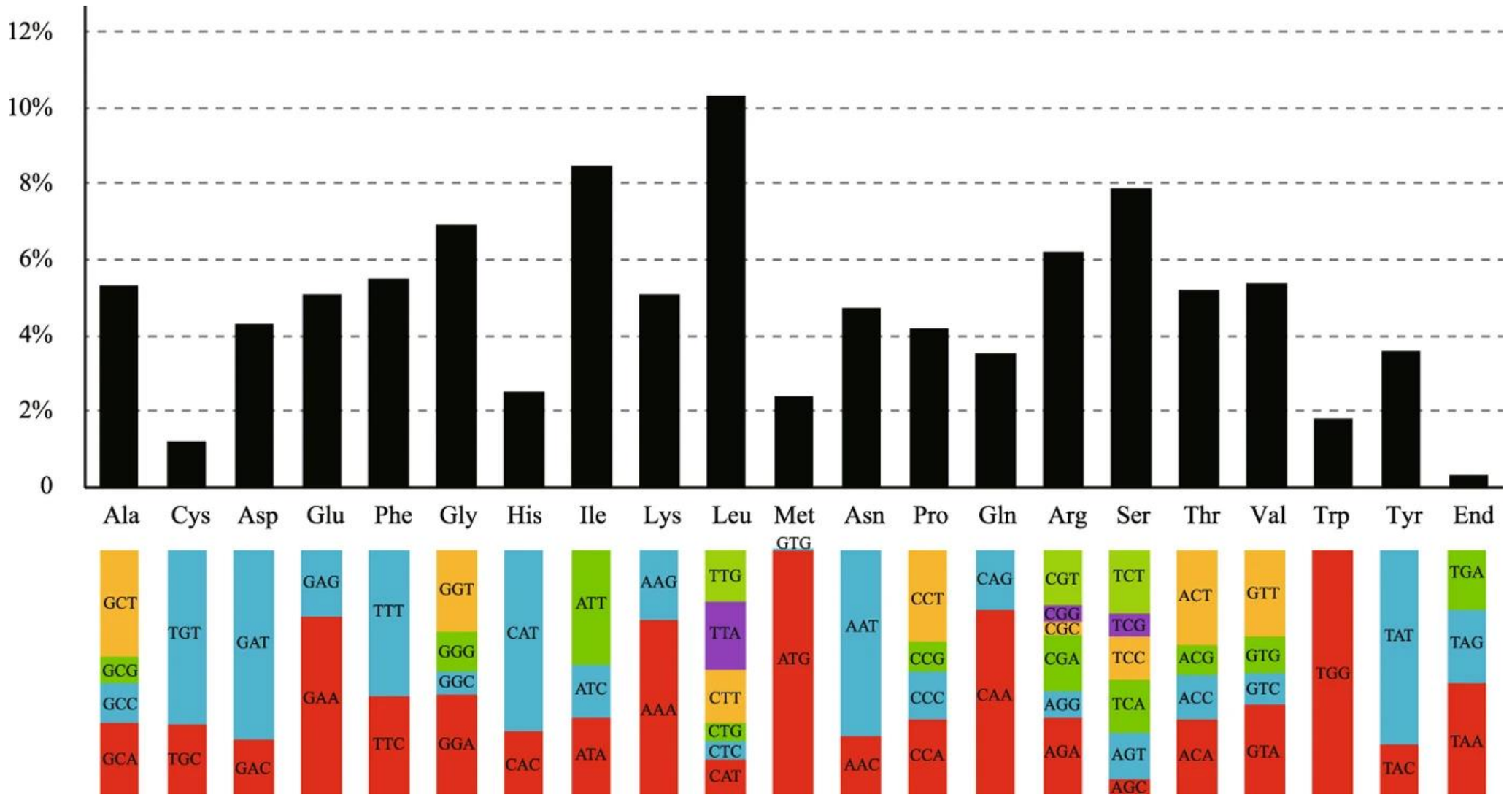
Figure 8.1: Structure of a typical prokaryotic gene structure. *Abbreviation:* RBS, ribosome binding site.

Conventional determination of ORFs

- First, translate a genomic region in all 6 possible frames
- In a non-coding region, stop codons occur once in about every twenty codons (by chance)
  - ➢ Identify ORFs with a certain <u>minimum length</u>
  - ➢ Threshold usually set at <u>50 or 60 codons</u>
- Putative gene then validated by the presence of other signals such as a start codon and Shine-Dalgarno sequence
- Subsequent search for homologs in a protein database to further confirm the protein-coding ORF

Coding regions differ from non-coding (“random”) ones

- Not all amino acids occur equally frequently in proteins
- Not all of the triplets that encode a particular amino acid are used equally often ("codon preference")



Third-position GC bias

Nucleotide at the third position of a codon:

- Coding sequences have a preference of G or C over A or T at this position
- Regions with GC values significantly above the random level are indicative of the presence of a gene
- Statistical patterns have to be computed for the six possible reading frames

Third-position repeats – TESTCODE

TESTCODE (part of the commercial GCG package):

- Uses the fact that the third-position nucleotides of codons in a coding region tend to repeat themselves
- Often used to confirm GC composition analysis

Problem: both methods (GC bias and TESTCODE) tend to miss "atypical" genes!

Comparison of GC bias and TESTCODE output

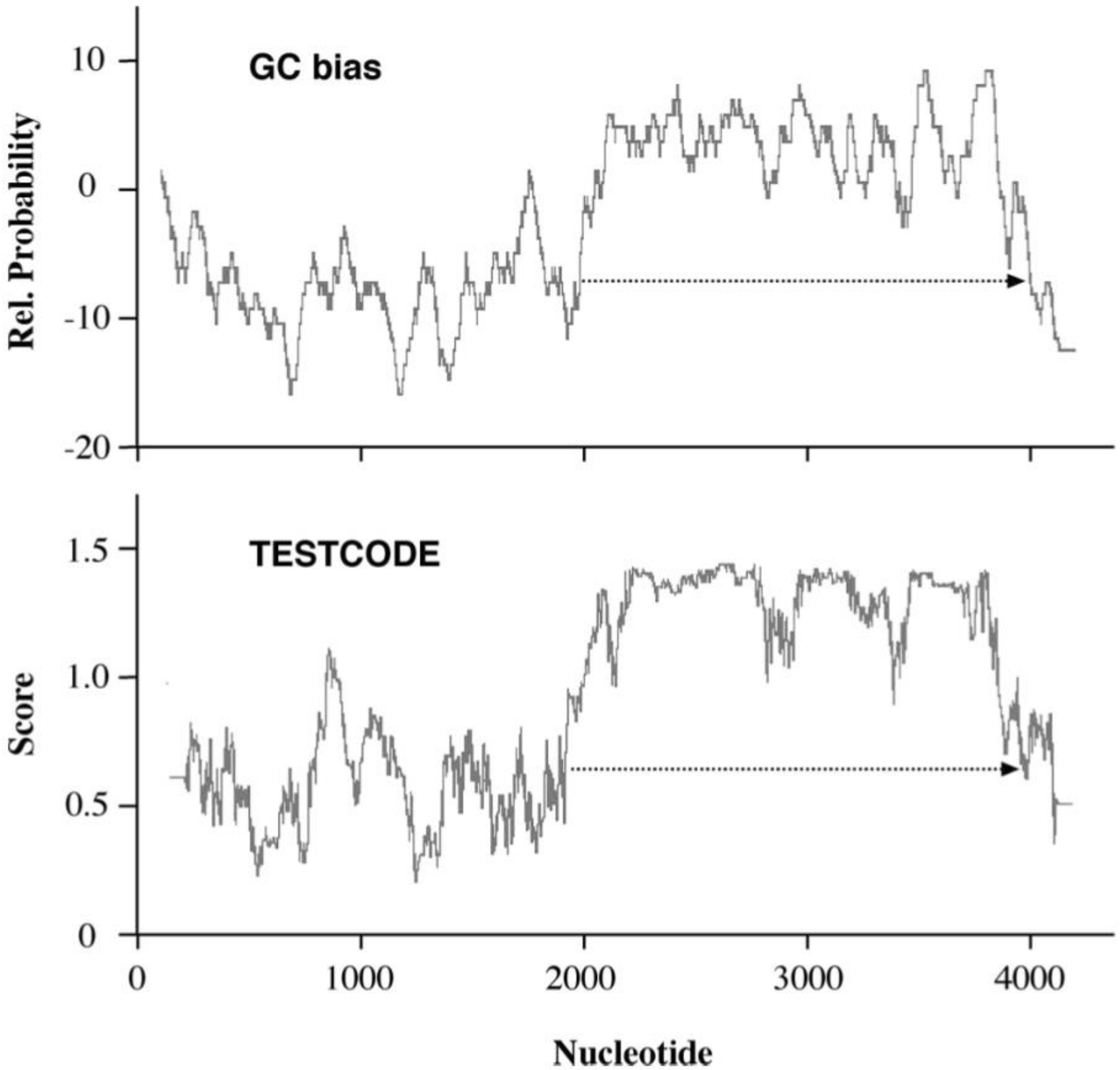


Figure 8.2: Coding frame detection of a bacterial gene using either the GC bias or the TESTCODE method. Both result in similar identification of a reading frame (*dashed arrows*).

Gene prediction using (Hidden) Markov Models

A Markov model describes the residue probabilities at a particular sequence position depending on the *k* previous positions (*k* is the order of the Markov model)

- Zero-order Markov model assumes that each base occurs independently with a given probability
  - ➢ This might be expected for non-coding sequences
- First-order Markov model assumes that the occurrence of a base depends on the base preceding it

Gene prediction using (Hidden) Markov Models

- Second-order model looks at the preceding two bases to determine residue probabilities
  - ➢ More characteristic of codons in a coding sequence
- The higher the order of a Markov model, the more accurately it can predict a gene (using more information)
- As a protein-encoding gene is composed of nucleotides in triplets as codons, more effective Markov models are built in sets of three nucleotides, describing non-random distributions of trimers (second order MM) or hexamers (fifth order MM), and so on
- Fifth order most common in gene prediction programs

Gene prediction using (Hidden) Markov Models

- Parameters of a Markov model have to be "trained" using a set of sequences with known (*e.g.* experimentally verified) gene locations
- Resulting Markov model is used to find non-random distributions of trimers or hexamers in a new sequence compatible with the statistical profile of the genes in the training set

Gene prediction using (Hidden) Markov Models

Fifth-order Markov model calculating probabilities of hexamer bases is most often used

- In short gene sequences the method's efficacy may be limited if there are not enough hexamers within the ORF

"Interpolated" Markov model (IMM):

- Variable-length Markov model
- Samples sequence patterns with *k* ranging from 1 to 8 (dimers to 9-mers), depending on ORF length

Gene prediction using (Hidden) Markov Models

Typical genes are in the range of 100 to 500 codons

Atypical genes are shorter or longer with different nucleotide statistics:

- Atypical genes tend to escape detection using the typical gene model
- More than one Markov model is needed

Solution: combination of different Markov models representing typical and atypical nucleotide distributions

Example of gene prediction using a Markov Model

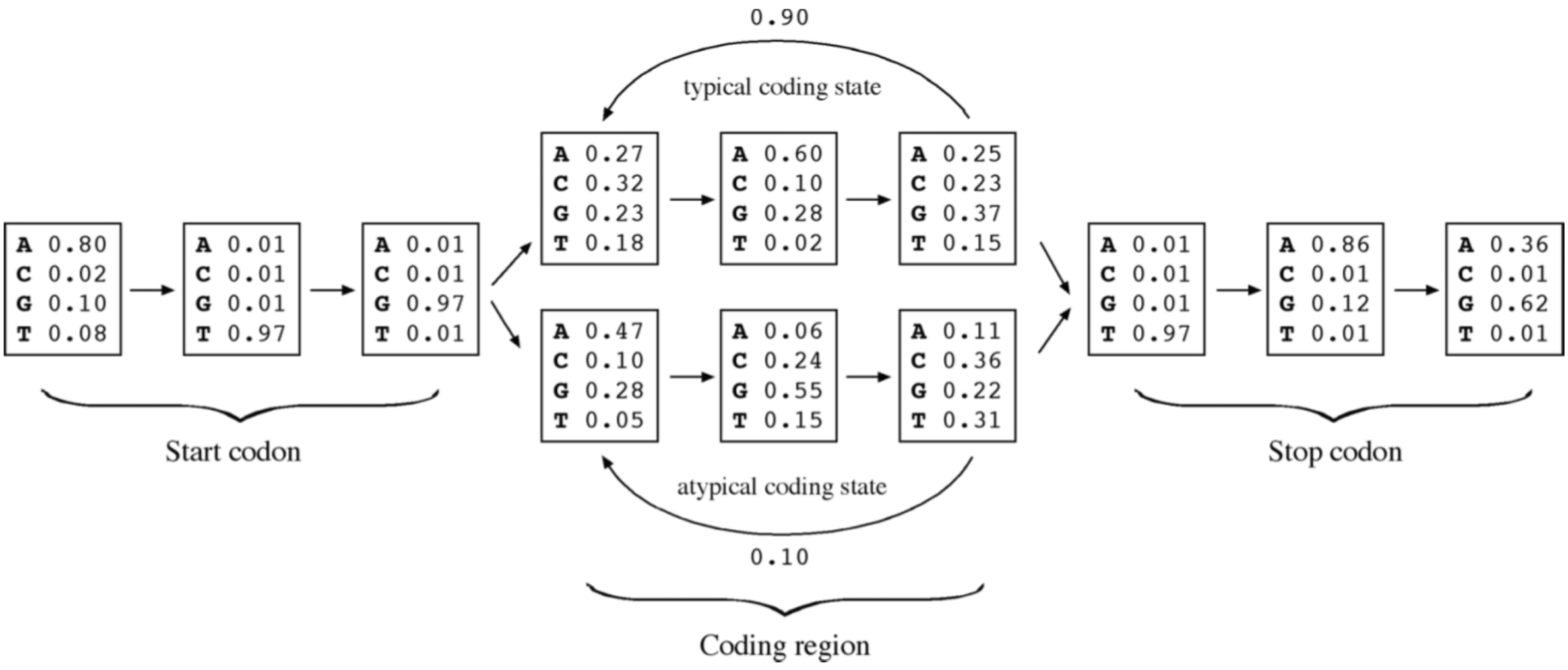


Figure 8.3: A simplified second-order HMM for prokaryotic gene prediction that includes a statistical model for start codons, stop codons, and the rest of the codons in a gene sequence represented by a typical model and an atypical model.

Example: GeneMark

- http://opal.biology.gatech.edu/GeneMark
- Suite of gene prediction programs based on fifth-order HMMs
- GeneMark.hmm:
  - Program trained on a number of complete microbial genomes
  - For a non-listed organism the most closely related organism can be chosen as the basis for computation
- **GeneMarkS**: self-trained program that can be used for new organisms, if user can provide at lead 100 kbp of sequence
- Variant for eukaryotic gene prediction is also available

Other programs

Glimmer (Gene Locator and Interpolated Markov Modeler), uses uses the IMM algorithm:
https://ccb.jhu.edu/software/glimmer/index.shtml

FGENESB
http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb

Performance evaluation

Sensitivity (Sn) = TP / (TP + FN)

Specificity (Sp) = TP / (TP + FP)

TP: true positive, TN: true negative, FN: false negative, FP: false positive

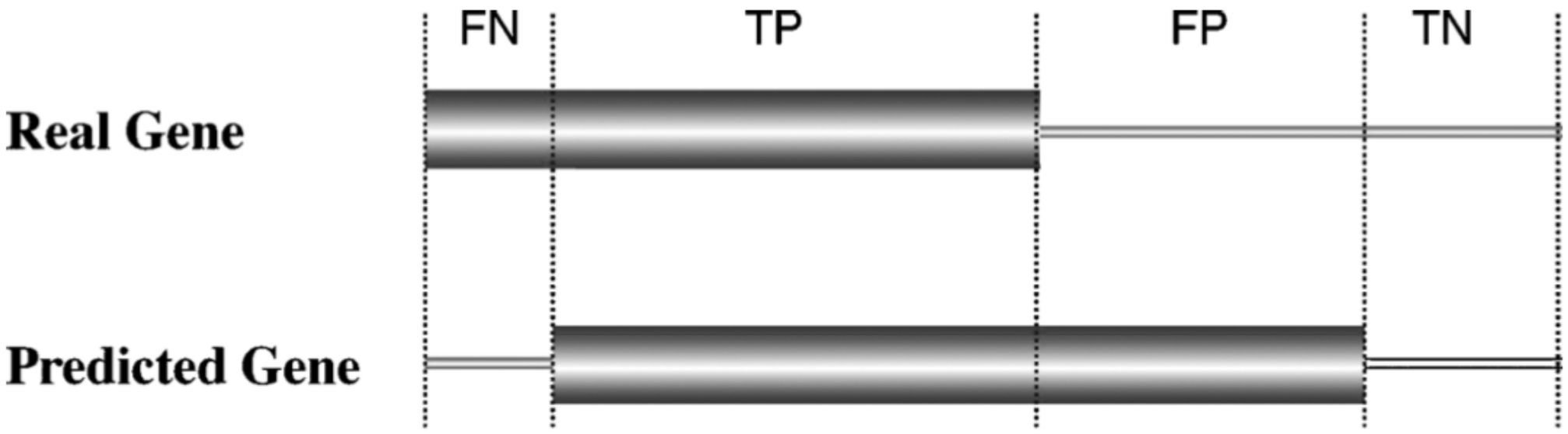


Figure 8.4: Definition of four basic measures of gene prediction accuracy at the nucleotide level. *Abbreviations:* FN, false negative; TP, true positive; FP, false positive; TN, true negative.

Performance evaluation

- **Sensitivity** is the ability to include true coding regions; if the sensitivity is low, the program lacks predictive power
- **Specificity** is the ability to exclude non-coding regions; if the specificity is low, the program has the tendency to over-predict

Performance evaluation

Neither sensitivity nor specificity alone can fully describe accuracy (both are important!)

Correlation coefficient (CC):

$$CC = \frac{TP \bullet TN - FP \bullet FN}{\sqrt{(TP + FP)(TN + FN)(FP + TN)}}$$

- **+1** means predictions are always correct
- **-1** means predictions are always incorrect

Performance evaluation

TABLE 8.1. Performance Analysis of the Glimmer Program for Gene Prediction of Three Genomes

Species	GC (%)	FN	FP	Sensitivity	Specificity
Campylobacter jejuni	30.5	10	19	99.3	98.7
Haemophilus influenzae	38.2	3	54	99.8	96.1
Helicobacter pylori	38.9	6	39	99.5	97.2

Note: The data sets were from three bacterial genomes (Aggarwal and Ramaswamy, 2002).
Abbreviations: FN, false negative; FP, false positive.

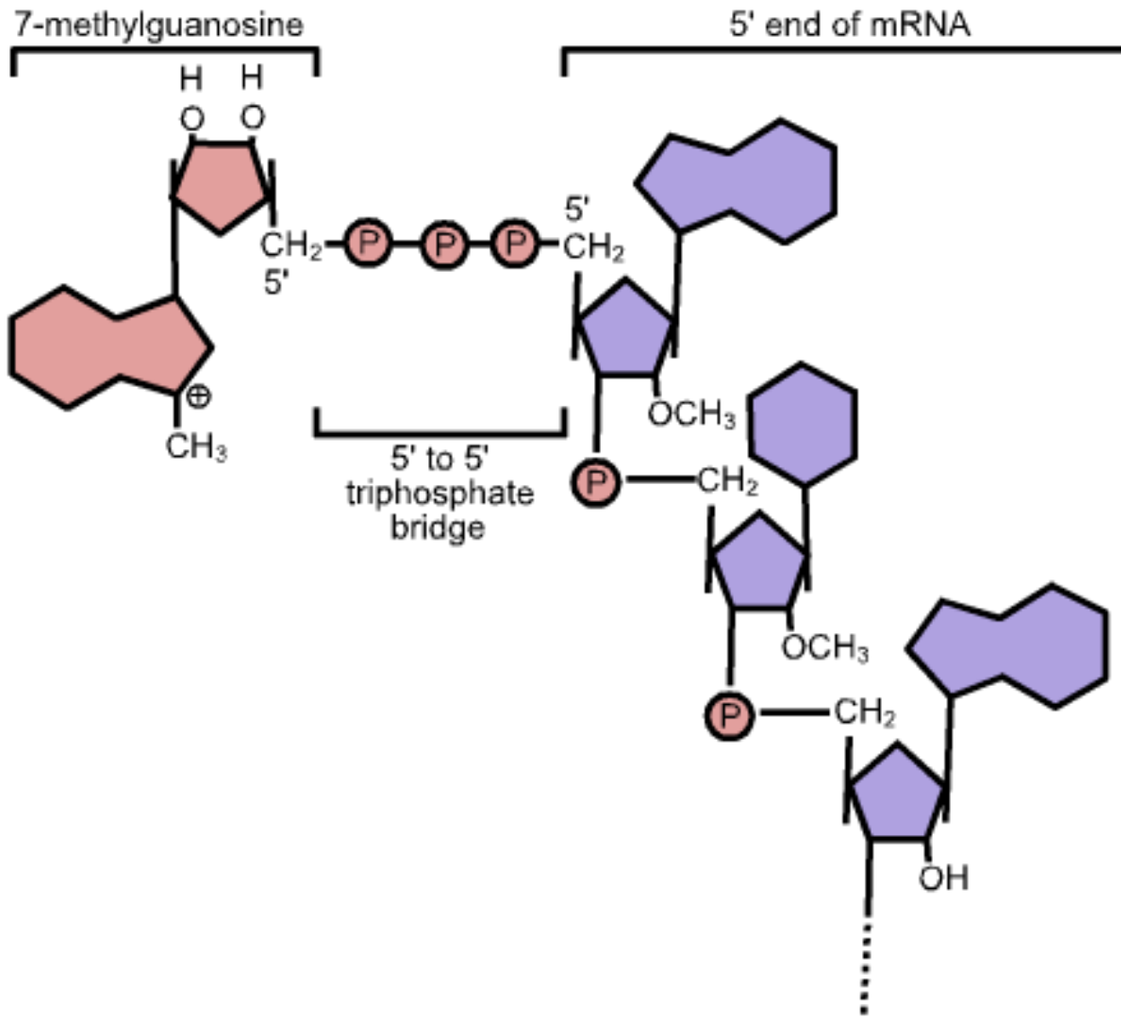
Gene prediction in eukaryotes

Eukaryotic genomes present a number of problems:

- <u>Much</u> larger: 10 Mbp to 670 Gbp
- Tend to have very low gene density, *e.g.* in humans only 3% of the genome contains the actual genes
- Space between genes is often very large and rich in repetitive sequences, transposable elements, *etc*.
- Many more pseudogenes
- "Mosaic" organisation: gene is split into pieces (*exons*) by intervening noncoding sequences (*introns*)

Three transcript modifications in eukaryotes

1. *Capping* at the 5'-end of the transcript, *i.e.* the attachment of a 7-methylguanosine *via* a 5'-to-5' triphosphate bridge and methylation of the initial residue



Three transcript modifications in eukaryotes

2. *Splicing*: removal of introns and joining of exons (moreover, *alternative splicing* can generate functional diversity in eukaryotic cells)

3. *Polyadenylation*: addition of a stretch of ~250 A-residues at the 3'-end of the RNA, controlled by a *poly-A signal* downstream of the ORF with consensus **CAATAA(T/C)**

Transcript modifications in eukaryotes

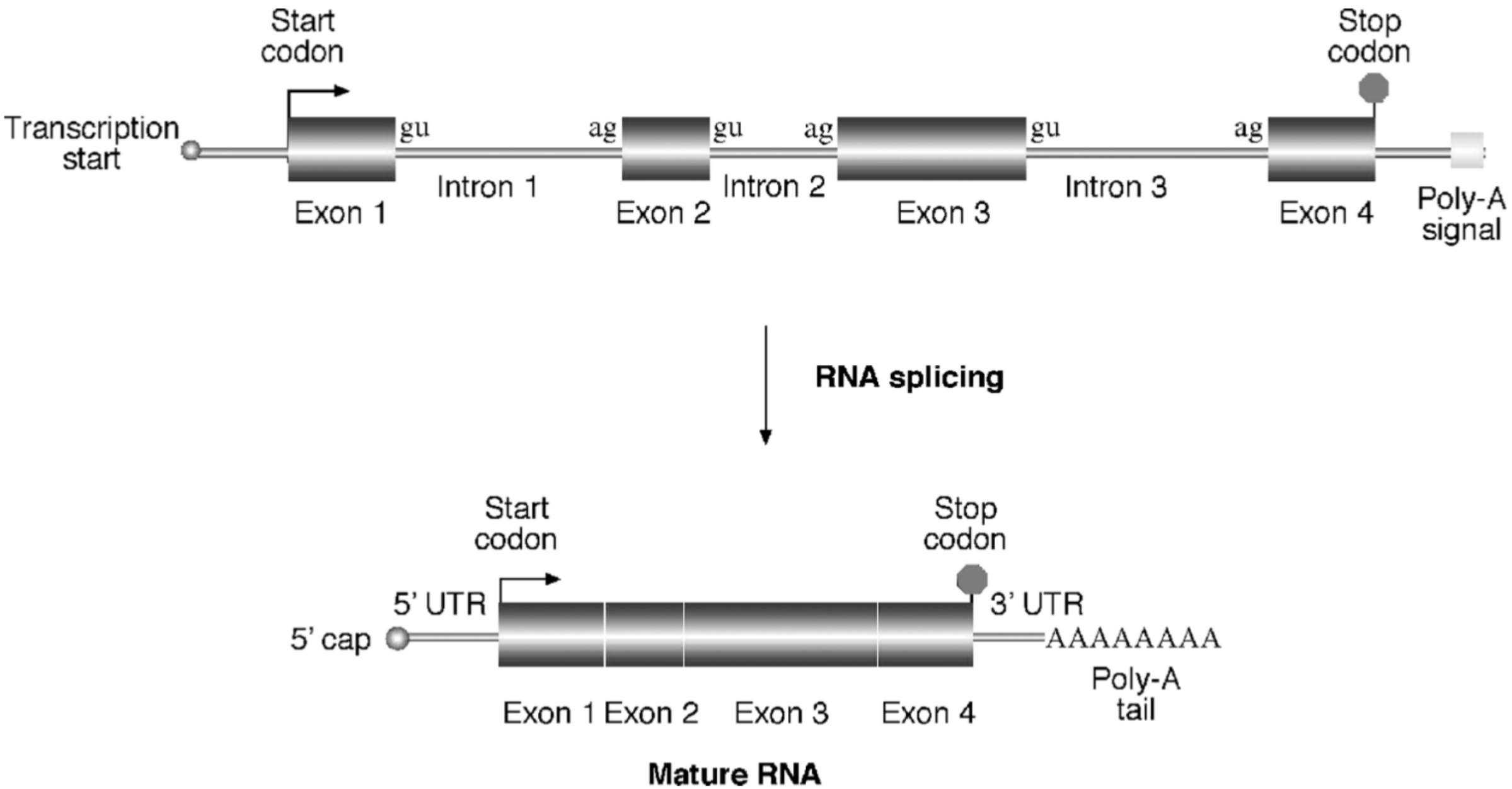


Figure 8.5: Structure of a typical eukaryotic RNA as primary transcript from genomic DNA and as mature RNA after posttranscriptional processing. *Abbreviations:* UTR, untranslated region; poly-A, polyadenylation.

Consequences of splicing for gene prediction

Main issue: correctly identifying the exons, introns and splice sites ("finding a broken needle in a haystack")

Conserved sequence features of splice sites (the so-called "GT-AG rule"):

- At the 5' splice junction (the beginning) there is a consensus motif of **GTAAGT**
- At the 3' splice junction (the end) there is a consensus motif of **$(Py)_{12}$NCAG**
- However, these sequence signatures are short and not always exactly the same, therefore sometimes difficult to identify; the possibility of alternative splicing complicates matters even further!

Transcript modifications in eukaryotes

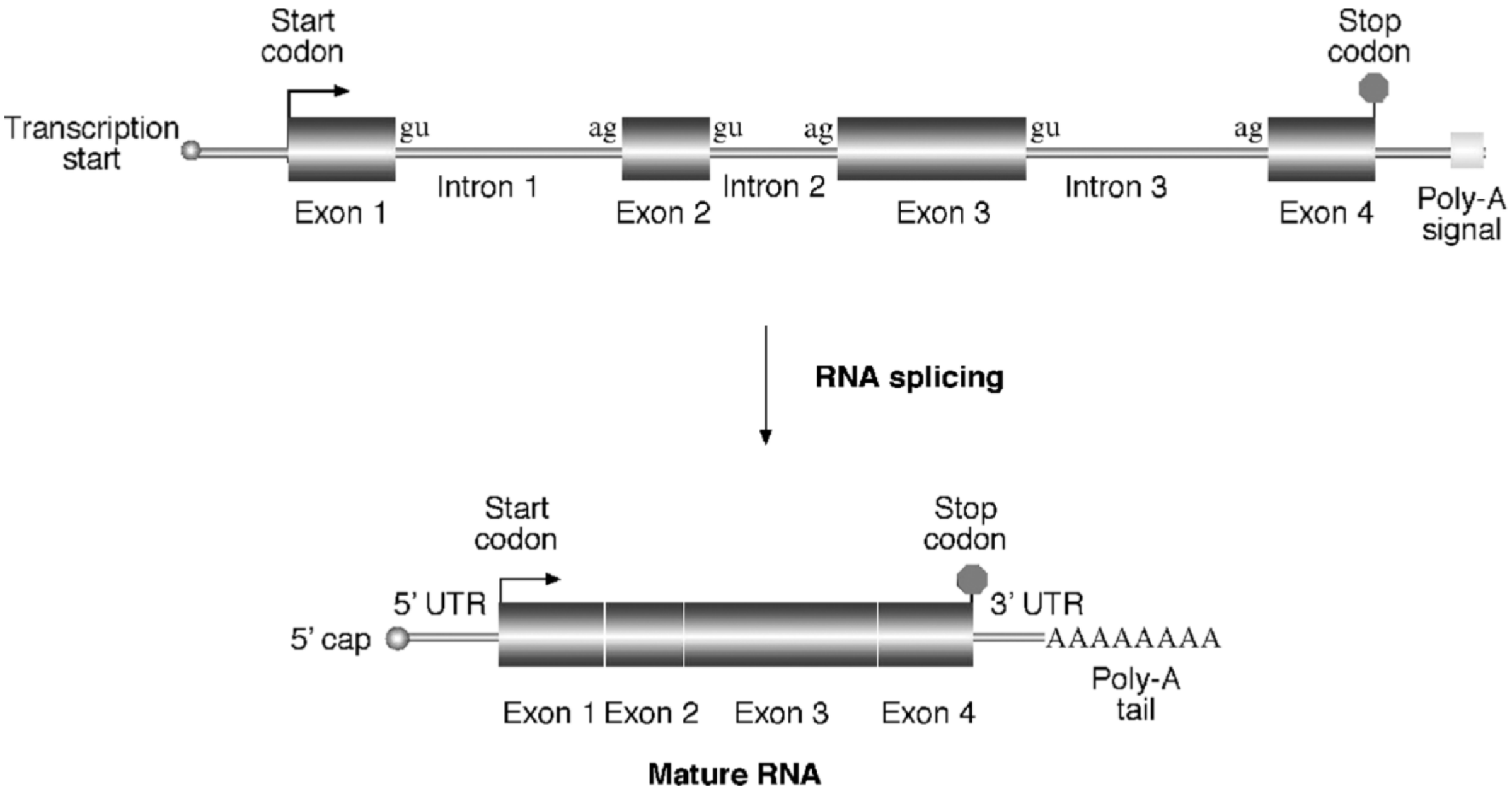


Figure 8.5: Structure of a typical eukaryotic RNA as primary transcript from genomic DNA and as mature RNA after posttranscriptional processing. *Abbreviations:* UTR, untranslated region; poly-A, polyadenylation.

Gene prediction in eukaryotes: methods

As for prokaryotes, three categories of algorithms:

- *Ab initio*
- Homology-based
- Consensus-based

Most programs are organism-specific as training data sets have to be derived from similar species: particularly in eukaryotes, *much of the information that gene prediction relies on strongly depends on the species!*

Ab initio methods

Approaches are generally comparable to those used for prokaryotic gene identification and, just like these, use:

Gene <u>signals</u>:

- ✓ Gene start and stop sites: most vertebrate genes use **ATG** as translation start codon and have a conserved flanking sequence (**CCGCC<u>ATG</u>G** or *Kozak* sequence)
- ✓ Most genes have a high density of **CG** dinucleotides near the transcription start site (*CpG island*)
- ✓ Splice sites within the gene
- ✓ Poly-A sites following the ORF

Gene <u>content</u>:

Nucleotide composition and pattern bias (*e.g.* non-random hexamer frequencies in coding regions)

Gene prediction using neural networks (“AI”)

Neural network:

- Complex statistical model with neuron-inspired architecture for pattern recognition and classification
- Network of mathematical variables that resembles the biological nervous systems
- Variables or nodes connected by weighted functions that are analogous to synapses
- Ability to "learn" and to make predictions after being trained
- Training by optimisation of the weights

Gene prediction using neural networks

Neural networks are constructed with multiple layers:

- Input: gene sequence with intron and exon signals (hexamer frequencies, splice sites, GC composition)
- One or more hidden layers
- Output: probability of an exon structure

Gene prediction using neural networks

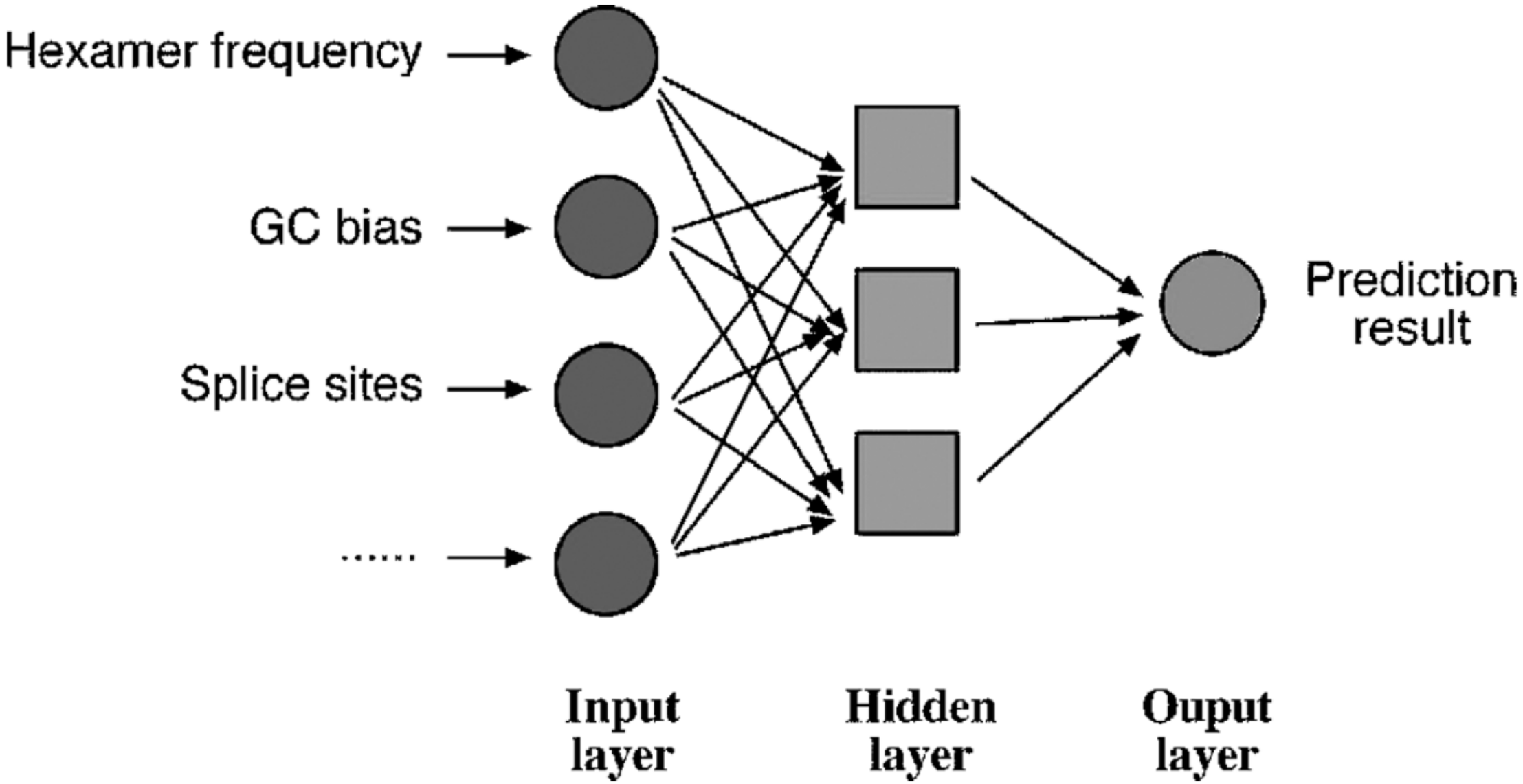


Figure 8.6: Architecture of a neural network for eukaryotic gene prediction.

Gene prediction using discriminant analysis

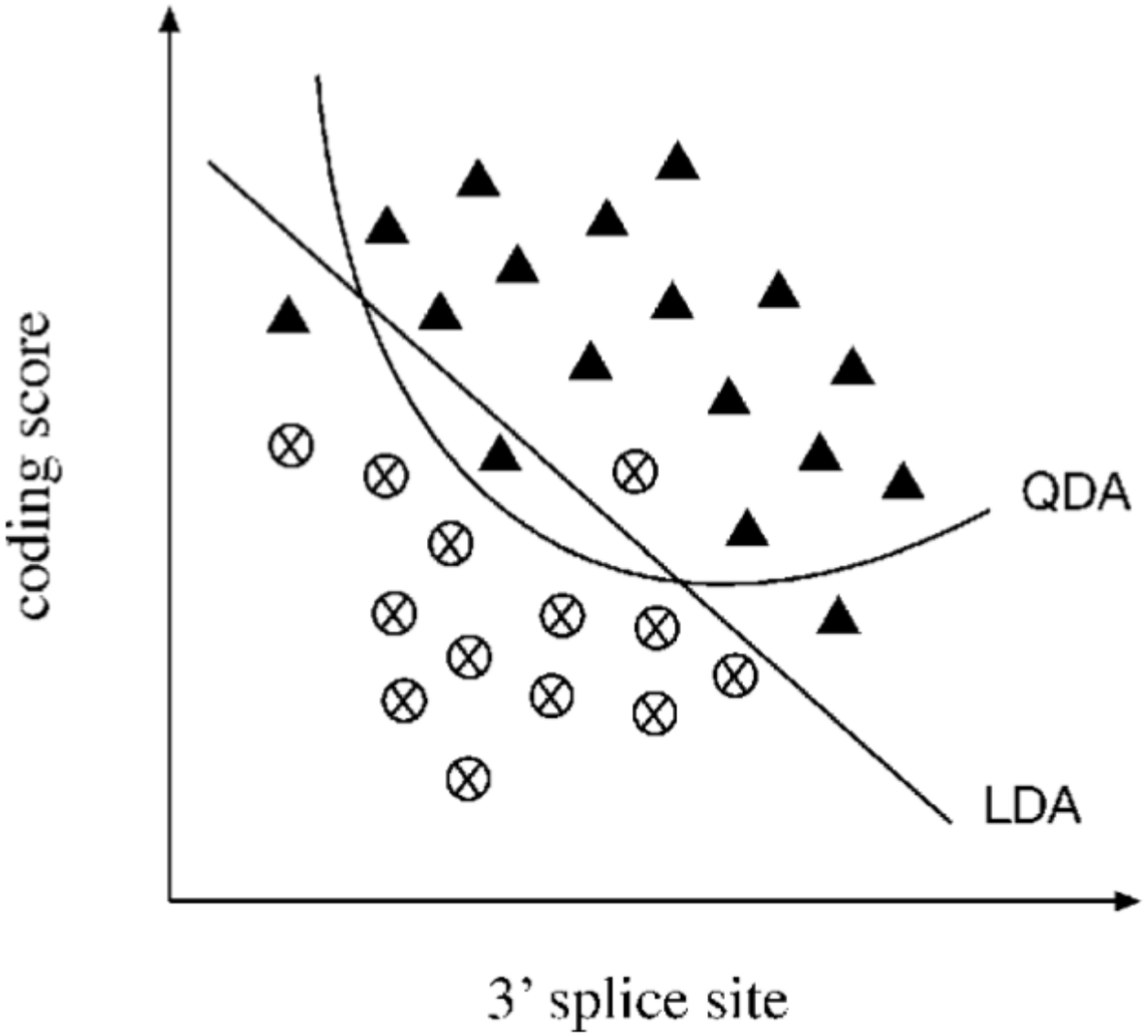


Figure 8.7: Comparison of two discriminant analysis, LDA and QDA. ▲ coding features; ⊗ noncoding features.

Examples of *ab initio* programs

HMMgene (HMM-based)

https://services.healthtech.dtu.dk/services/HMMgene-1.1

GENSCAN (HMM-based)

http://hollywood.mit.edu/GENSCAN.html

Homology-based programs

- Based on the fact that, in spite of splicing, the final products (processed mRNA and protein) are highly conserved between related species
- Potential coding frames are translated and align with proteins in databases, matches reveal exons
- Very powerful if mRNA/protein sequences are aligned for which experimental evidence exists (*e.g.* cDNA)
- Cannot be used for the discovery of novel genes in new species

Examples of homology-based programs

GenomeScan (combines GENSCAN with with BLASTX similarity searches)

http://hollywood.mit.edu/genomescan.html

EST2Genome (aligns known cDNAs/ESTs using dynamic programming algoritm)

https://www.bioinformatics.nl/cgi-bin/emboss/est2genome

Consensus-based approach

- Use combined results of multiple programs based on consensus
- Retain common predictions agreed by most programs and remove inconsistent predictions
- Results in higher specificity (removes false positives), though possibly lowering sensitivity (increases false negatives)
- For eukaryotic genes, a consensus-based approach is the only sensible way of doing things

Performance evaluation

- For a correctly predicted gene, all nucleotides and all exons have to be predicted correctly
- One single error at the nucleotide level can negate the entire gene prediction
- Consequently the accuracy values on exon and gene level are much lower than at nucleotide level

Performance evaluation

TABLE 8.2. Accuracy Comparisons for a Number of Ab Initio Gene Prediction Programs at Nucleotide and Exon Levels

	Nucleotide level			Exon level				
	Sn	Sp	CC	Sn	Sp	(Sn + Sp)/2	ME	WE
FGENES	0.86	0.88	0.83	0.67	0.67	0.67	0.12	0.09
GeneMark	0.87	0.89	0.83	0.53	0.54	0.54	0.13	0.11
Genie	0.91	0.90	0.88	0.71	0.70	0.71	0.19	0.11
GenScan	0.95	0.90	0.91	0.70	0.70	0.70	0.08	0.09
HMMgene	0.93	0.93	0.91	0.76	0.77	0.76	0.12	0.07
Morgan	0.75	0.74	0.74	0.46	0.41	0.43	0.20	0.28
MZEF	0.70	0.73	0.66	0.58	0.59	0.59	0.32	0.23

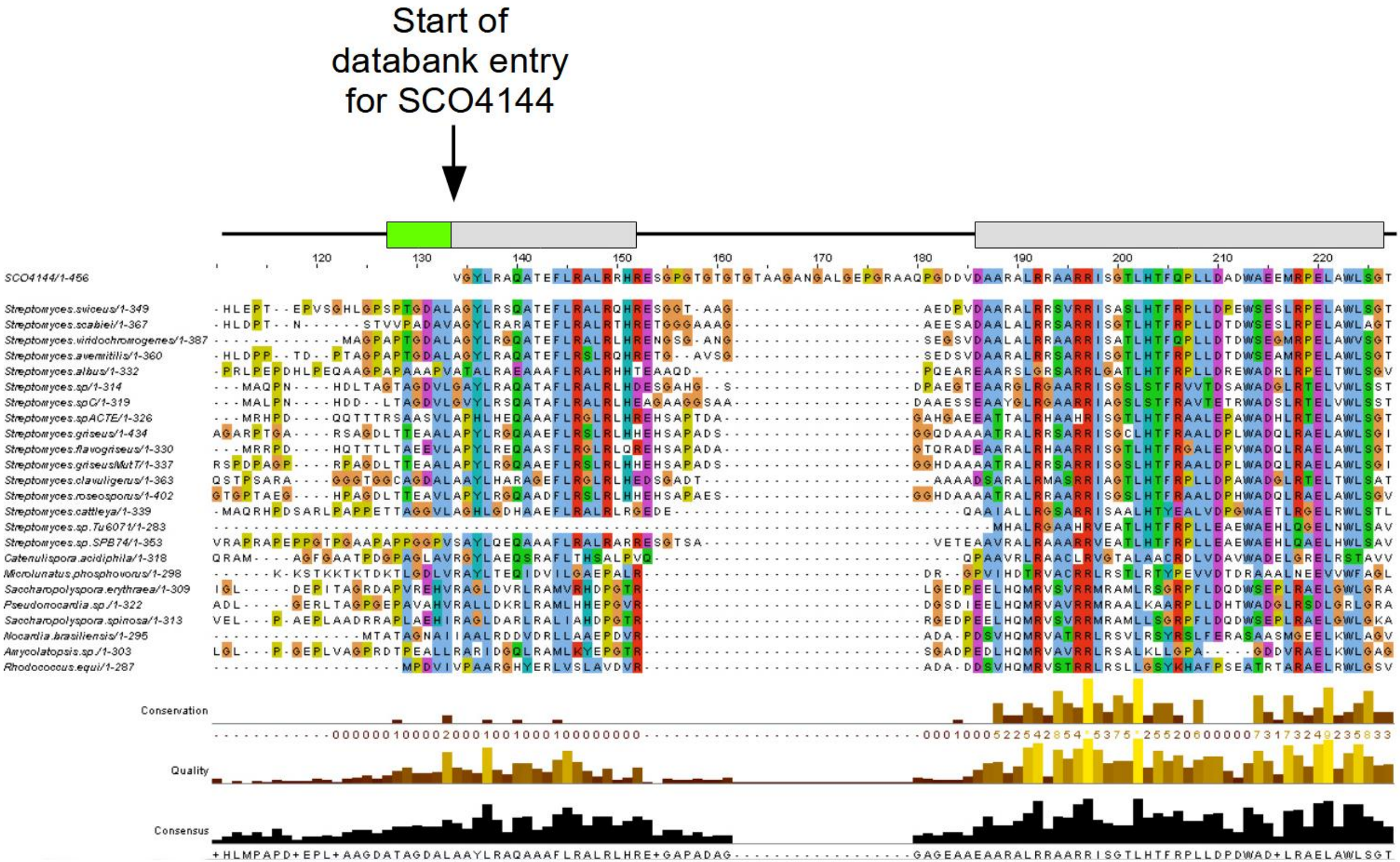
Note: The data sets used were single mammalian gene sequences (performed by Sanja Rogic, from www.cs.ubc.ca/~rogic/evaluation/tablesgen.html.

Abbreviations: Sn, sensitivity; Sp, specificity; CC, correlation coefficient; ME, missed exons; WE, wrongly predicted exons.

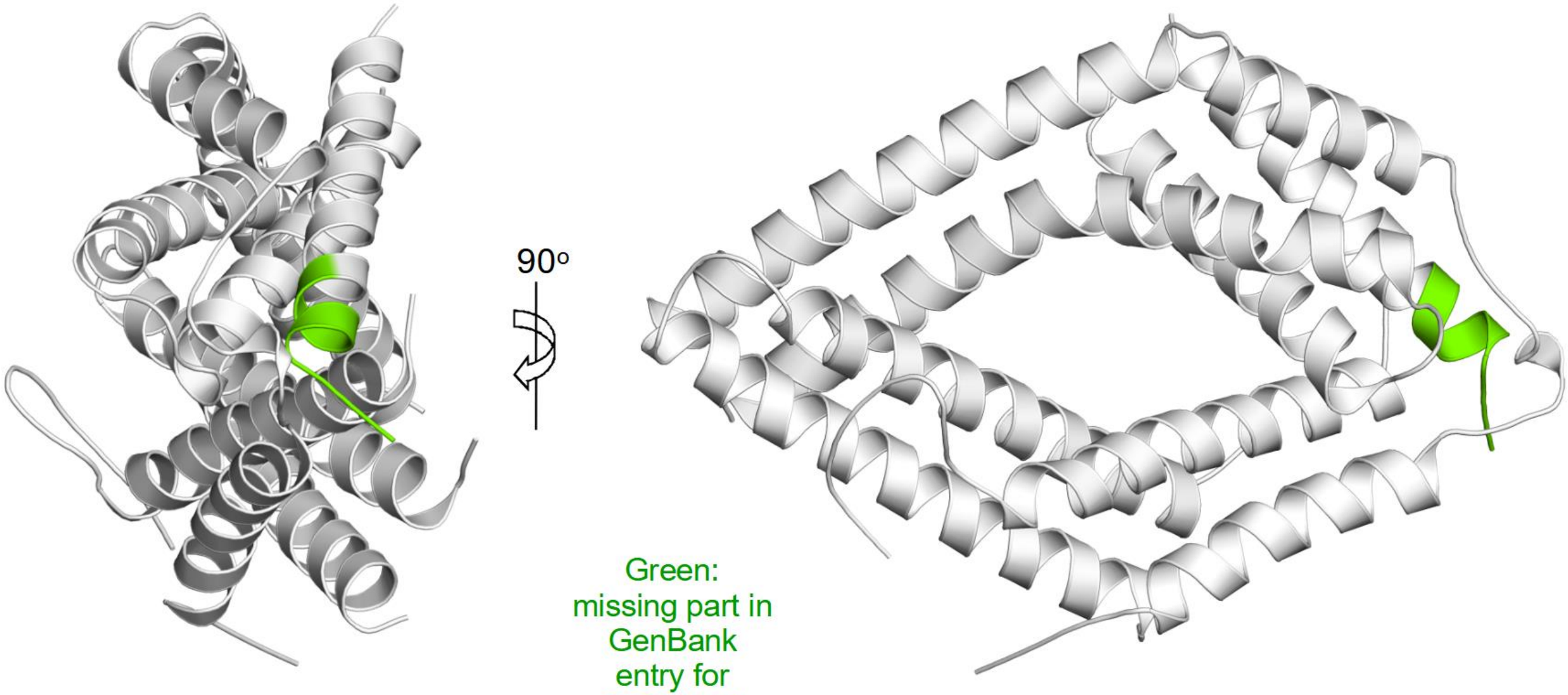
Performance evaluation

- Accuracy level of gene prediction program is usually published, but based on particular datasets optimised for the program
- When the programs are used for truly unknown eukaryotic genomic sequences, the accuracy can become much lower
- Therefore, it is difficult to estimate the true accuracy of the current prediction tools
- No single program is able to produce consistent superior results
- Most popular programs can not predict more than 40% of the genes exactly right in complex genomes

Databases can be wrong: an example from GenBank



Databases can be wrong: an example from GenBank



“side view”

“front view”