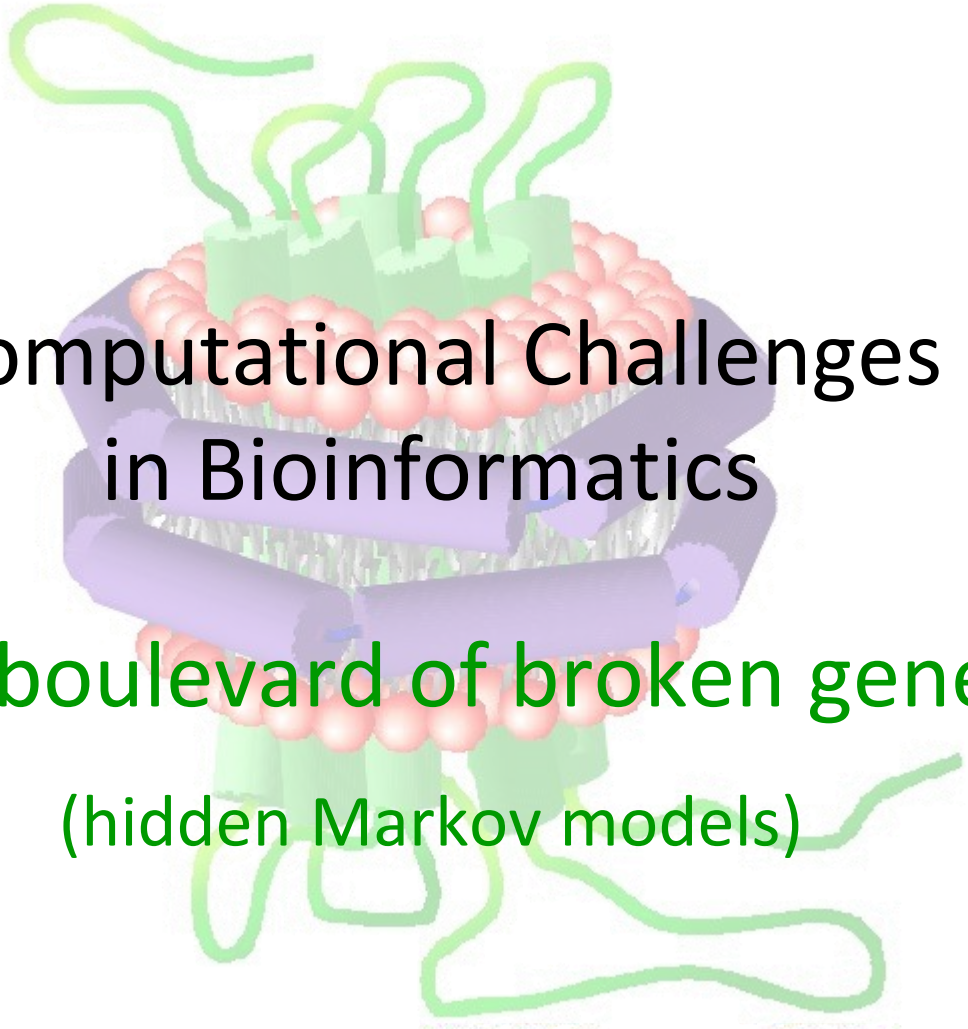
Master of Mathematical Informatics
Master of Statistical Data Analysis



Computational Challenges in Bioinformatics

The boulevard of broken genes

(hidden Markov models)

Prof. Dr. Peter Dawyndt

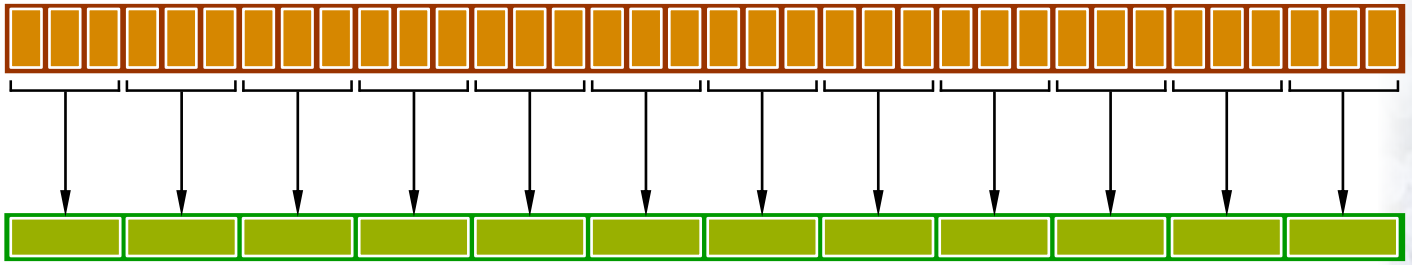
Peter.Dawyndt@UGent.be

UNIVERSITEIT GENT

Problem statement

- gene prediction in metagenomics DNA reads
  - partial gene fragments

  - read errors (indels → frameshifts)

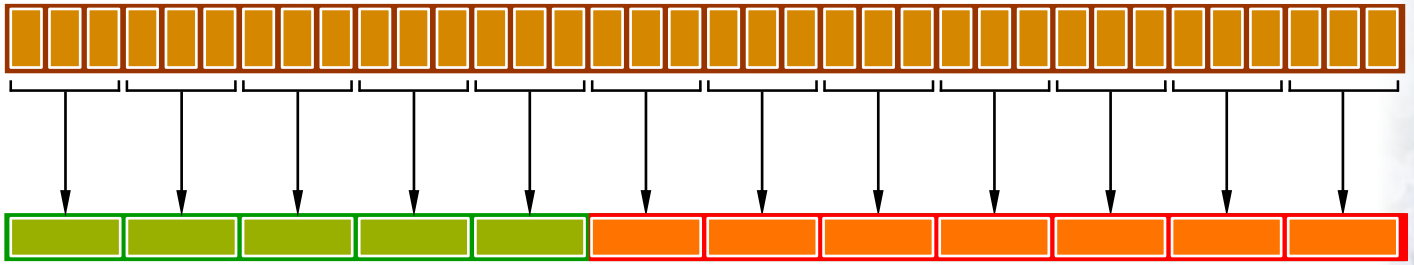


UNIVERSITEIT GENT

Problem statement

- gene prediction in metagenomics DNA reads
  - partial gene fragments

  - read errors (indels → frameshifts)

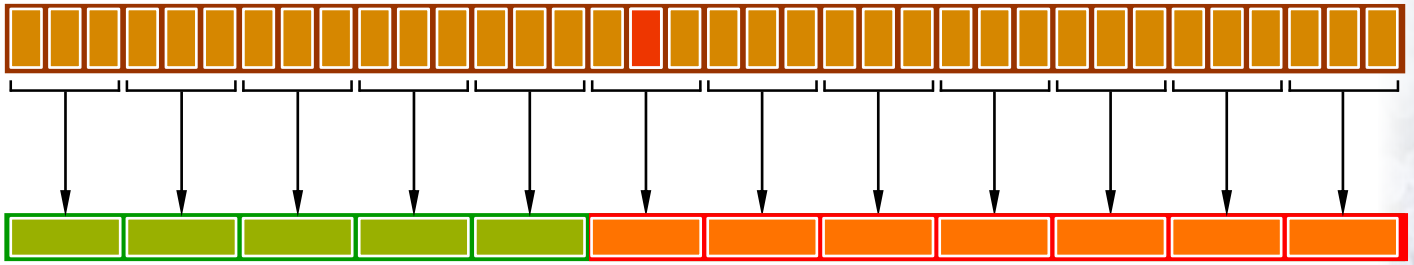


UNIVERSITEIT GENT

Problem statement

- gene prediction in metagenomics DNA reads
  - partial gene fragments

  - read errors (indels → frameshifts)



  - mixed population (bacteria, archaea, viruses, eukaryotes)

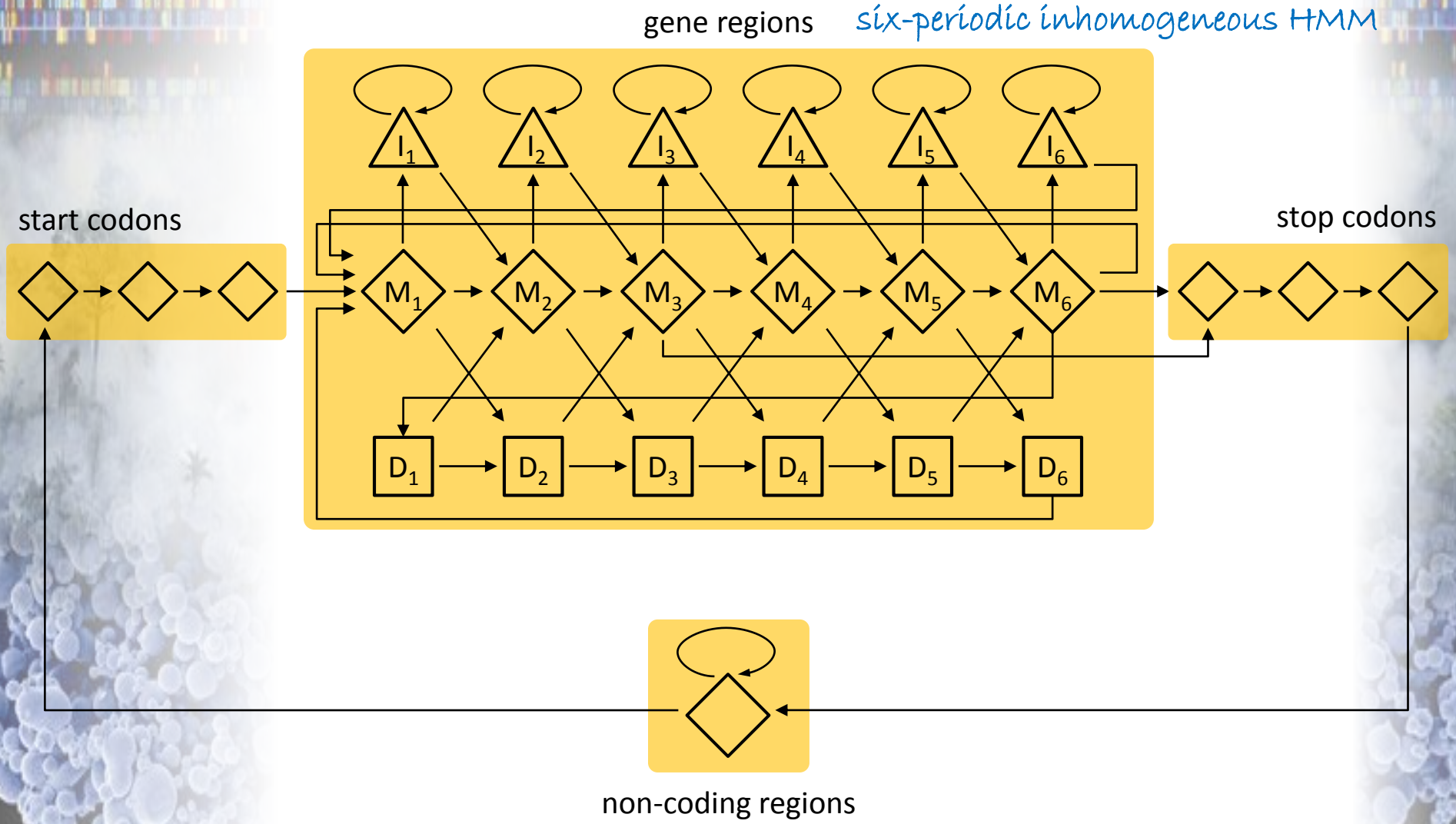
UNIVERSITEIT GENT

FragGeneScan

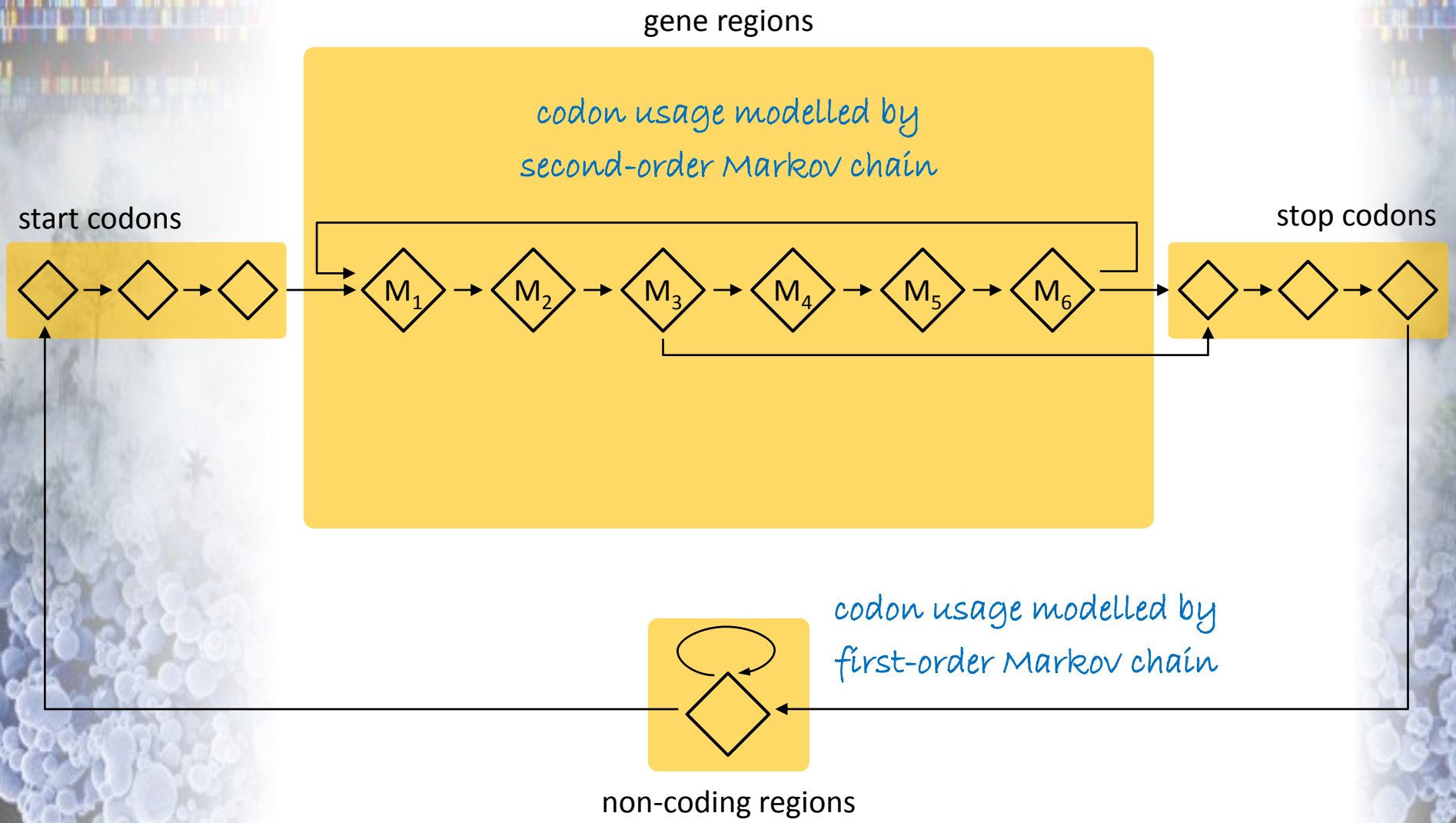
- algorithm
  - hidden Markov model
    - codon usage bias
    - sequencing error models
    - start/stop codon patterns
  - best path of hidden states (Viterbi algorithm)
  - gene reported if
    - length of gene ≥ 60 bp
    - gene starts in
      - start state (start codon)
      - match state (internal region of genes)
    - gene stops in
      - stop state (stop codon)
      - match state (internal region of genes)

UNIVERSITEIT GENT

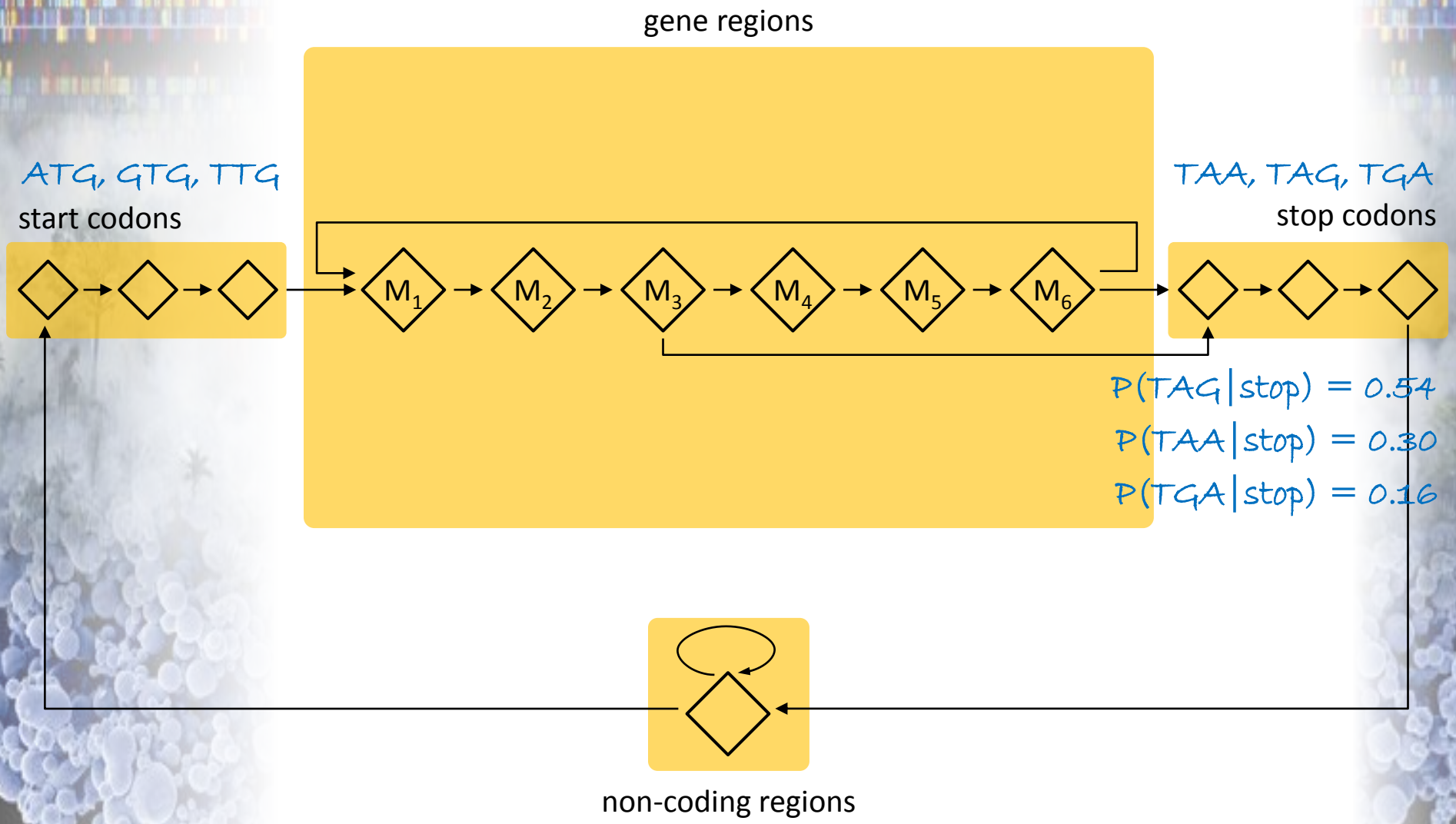
FragGeneScan HMM



FragGeneScan HMM

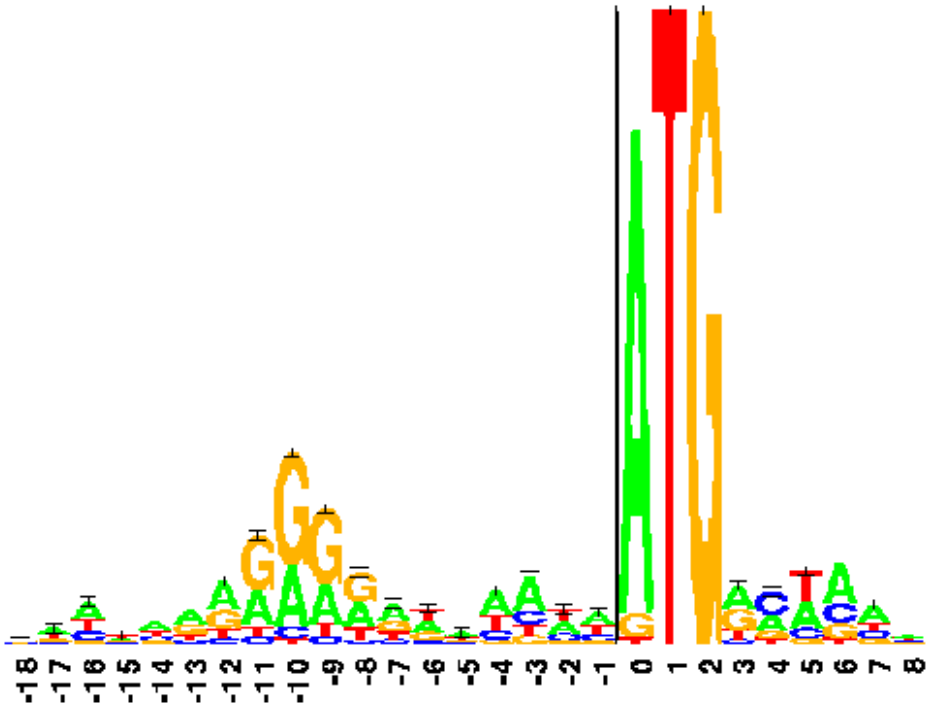


FragGeneScan HMM



Start state modelling

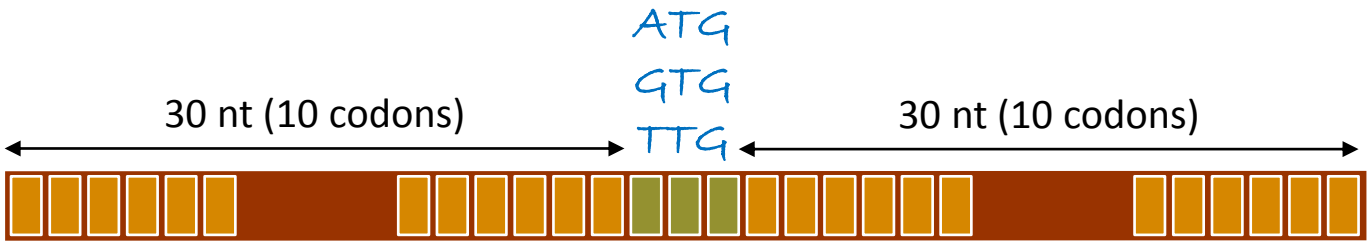
- sequence pattern around real start codons
  - AT-rich region
  - Shine-Dalgarno sequence (**AGGAG**)



UNIVERSITEIT GENT

Start state modelling

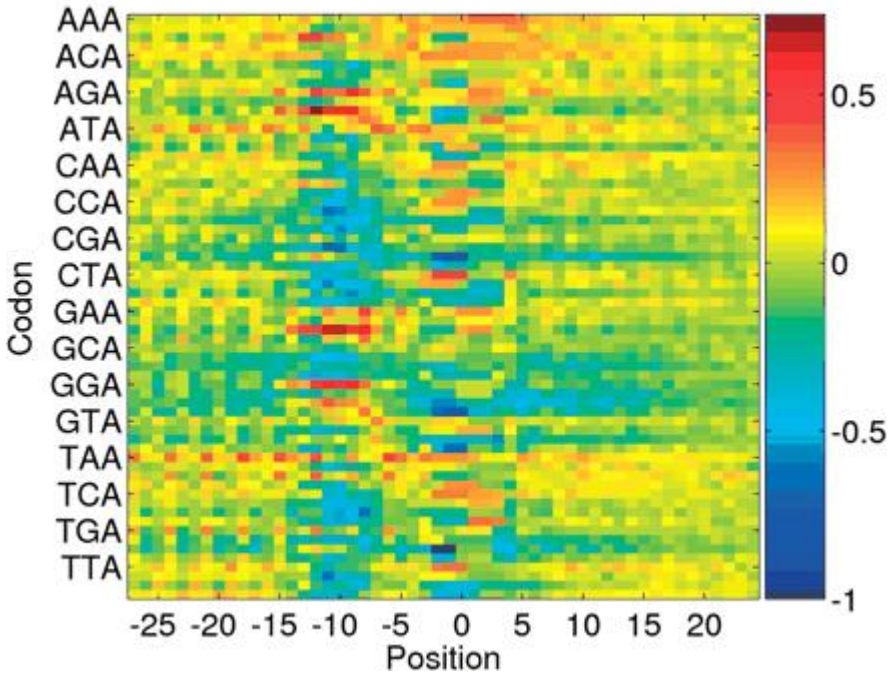
- sequence pattern around real start codons
  - AT-rich region
  - Shine-Dalgarno sequence (**AGGAG**)
  - triple-A downstream box



UNIVERSITEIT GENT

Start state modelling

- sequence pattern around real start codons
  - AT-rich region
  - Shine-Dalgarno sequence (**AGGAG**)
  - triple-A downstream box
- compute positional weight matrix (PWM)



UNIVERSITEIT GENT

Start state modelling

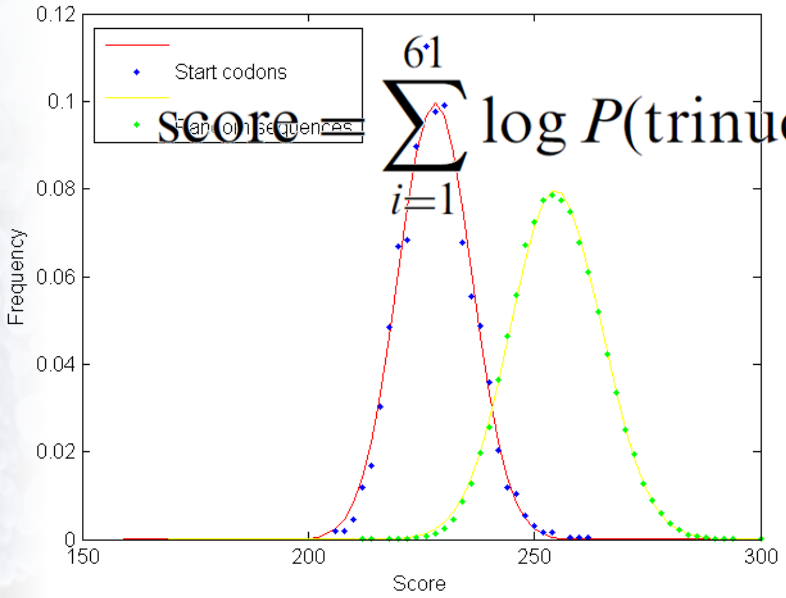
- sequence pattern around real start codons
  - AT-rich region
  - Shine-Dalgarno sequence (**AGGAG**)
  - triple-A downstream box
- compute positional weight matrix (PWM)
- compute score for each putative start codon

$$\text{score} = \sum_{i=1}^{61} \log P(\text{trinucleotide}_i | \text{PWM})$$

UNIVERSITEIT GENT

Start state modelling

- sequence pattern around real start codons
  - AT-rich region
  - Shine-Dalgarno sequence (**AGGAG**)
  - triple-A downstream box
- compute positional weight matrix (PWM)
- compute score for each putative start codon

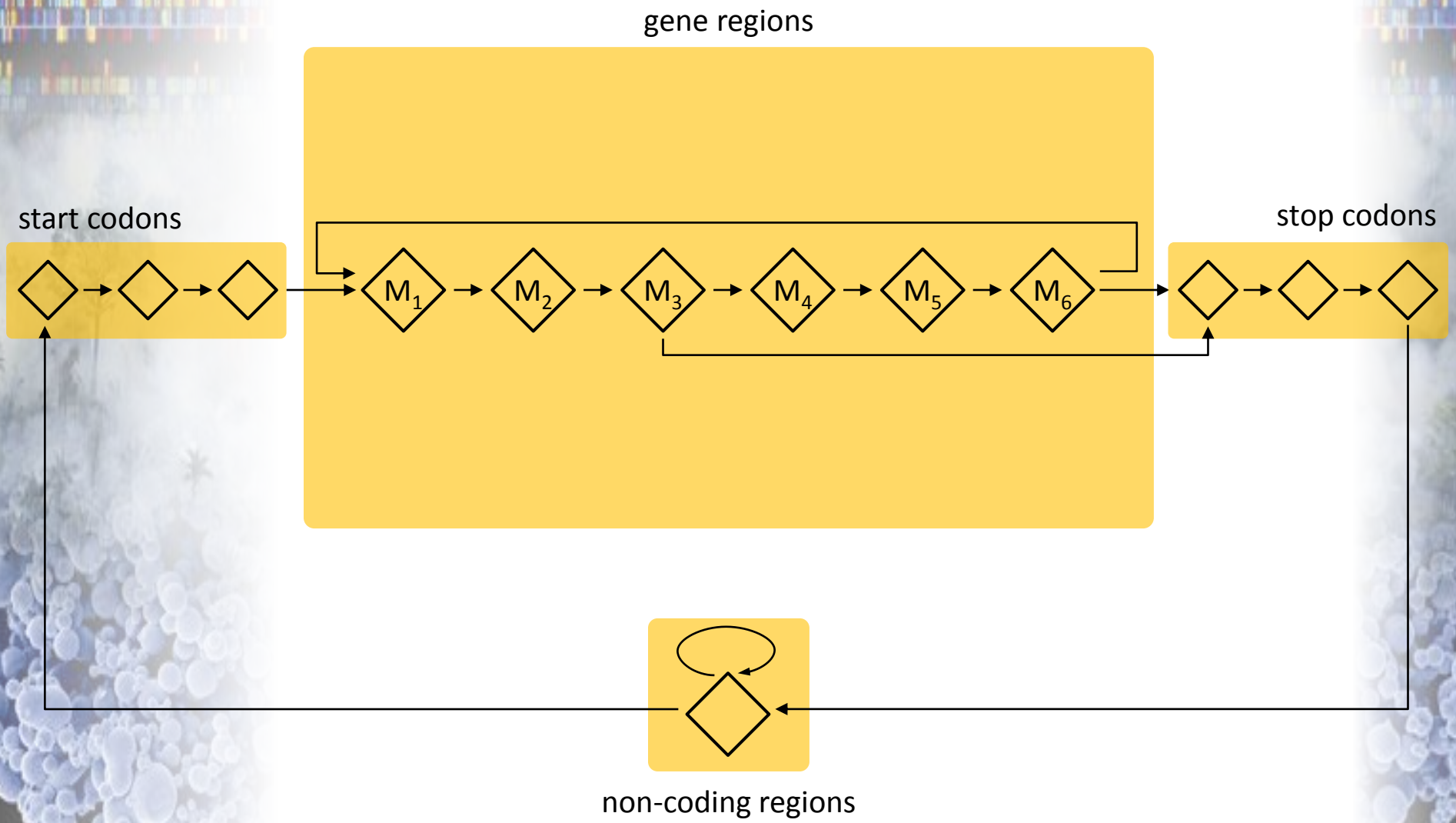


$$\text{score} = \sum_{i=1}^{61} \log P(\text{trinucleotide}_i | \text{PWM})$$

naïve Bayesian classifier

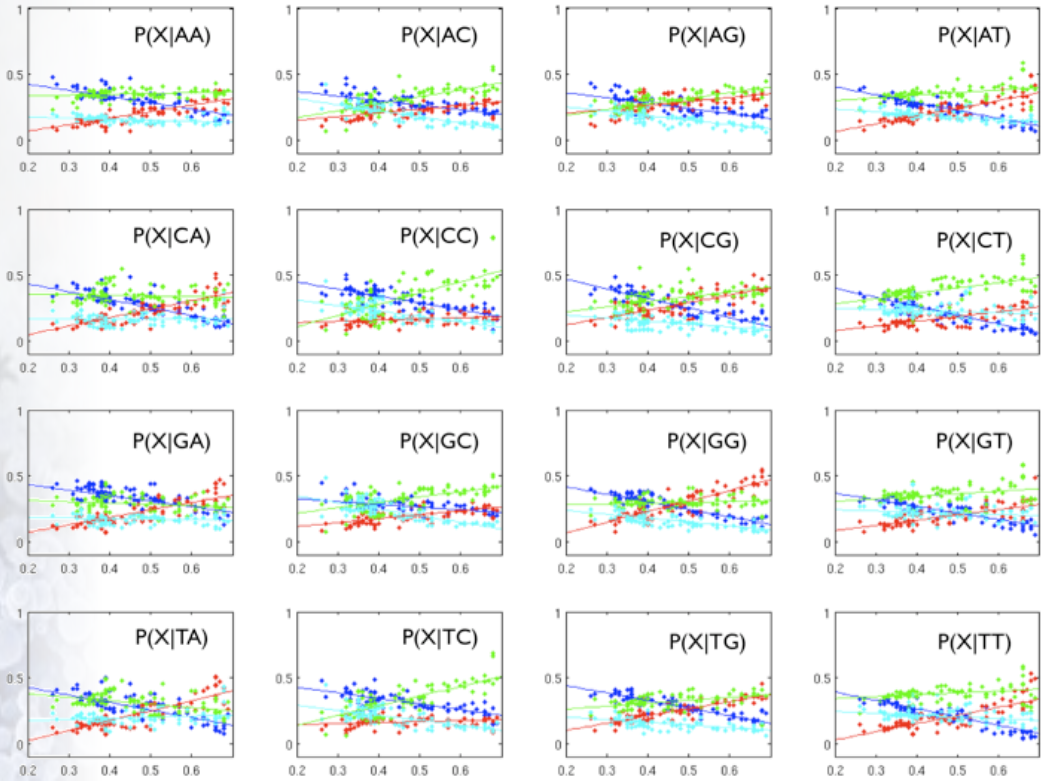
UNIVERSITEIT GENT

HMM parameter estimation



HMM parameter estimation

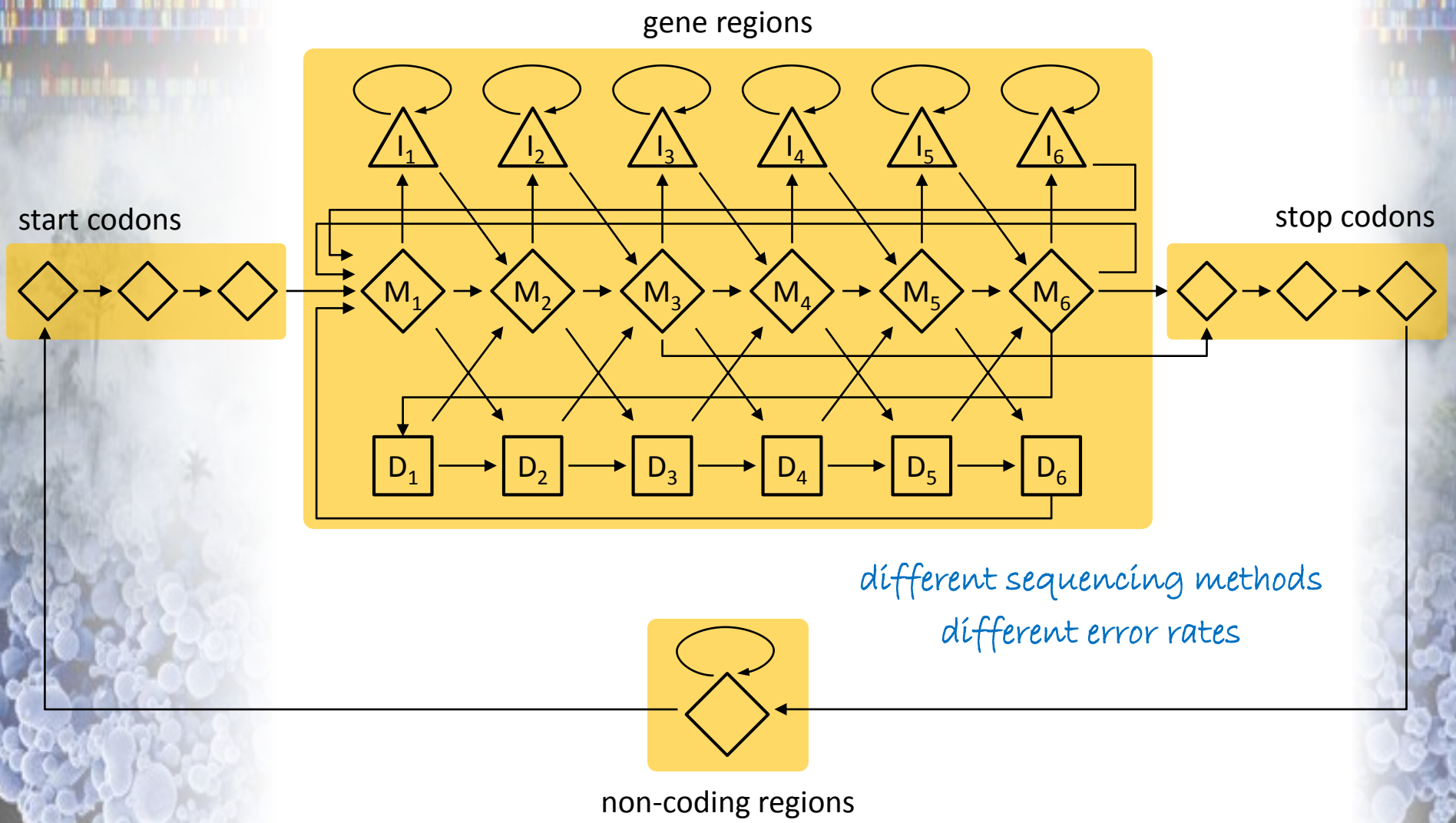
- 139 complete bacterial genomes
- gene annotations taken from NCBI



Complete genomes of 139 bacteria were used to estimate parameters of second-order Markov chains for all match states. The x-axis denotes GC contents and y-axis denotes the conditional probability P. The parameters show linear correlation with GC contents, and therefore a linear regression was applied to give estimations of parameters for various GC contents. Note that *FragGeneScan* does not need training for gene prediction in individual genomes or datasets of short reads. Given a dataset of short reads, *FragGeneScan* estimates GC contents independently for each read and uses the corresponding set of pre-computed parameters based on the GC content for gene prediction in that read.

UNIVERSITEIT GENT

FragGeneScan HMM



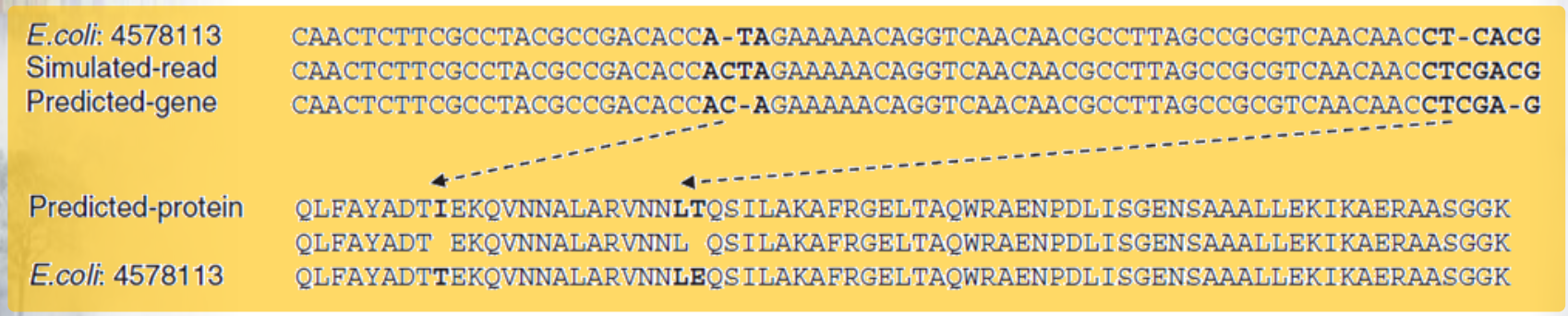
Benchmark

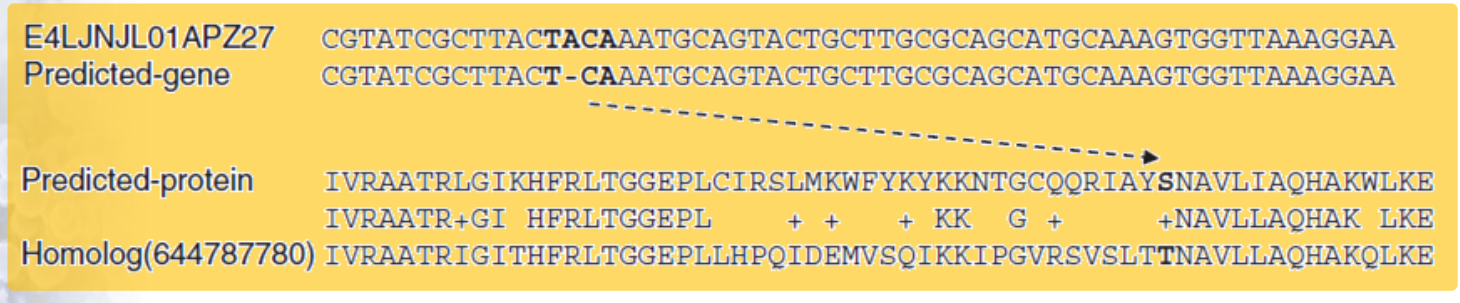
Organisms	Read length (bp)[a]	FragGeneScan			MetaGene		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
B. aphidicola	100	79.16	80.12	79.64	49.59	55.24	52.41
	200	83.56	84.20	83.88	31.32	28.92	30.12
	400	84.75	81.58	83.16	17.63	13.73	15.68
	700	89.92	74.64	82.28	45.89	32.42	39.16
B. pseudomallei	100	75.79	64.78	70.28	18.64	49.63	34.14
	200	86.56	78.01	82.29	46.97	43.86	45.41
	400	90.40	82.57	86.48	31.03	25.91	28.47
	700	91.57	82.50	87.04	54.42	42.10	48.26
B. subtilis	100	72.36	65.96	69.16	31.21	55.81	43.51
	200	83.39	79.06	81.22	34.03	36.18	35.10
	400	88.24	83.51	85.88	19.83	19.25	19.54
	700	92.17	84.37	88.27	47.93	39.67	43.80
C. jeikeium	100	75.46	71.04	73.25	33.30	60.11	46.71
	200	83.75	80.93	82.34	39.65	39.27	39.46
	400	86.94	84.44	85.69	24.65	22.06	23.35
	700	90.21	85.72	87.97	49.81	39.14	44.47
C. tepidum	100	73.45	65.20	69.33	28.90	58.64	43.77
	200	81.54	77.22	79.38	40.41	40.71	40.56
	400	84.37	83.02	83.70	24.42	22.73	23.58
	700	86.51	85.86	86.19	49.33	42.55	45.94
E. coli	100	75.24	65.99	70.62	31.33	57.64	44.48
	200	85.78	78.52	82.15	39.78	37.85	38.81
	400	89.19	82.76	85.98	23.54	19.57	21.56
	700	92.86	84.19	88.53	50.97	38.26	44.62
H. pylori	100	72.69	71.69	72.19	41.94	54.58	48.26
	200	82.81	81.39	82.10	30.28	29.83	30.05
	400	84.34	78.25	81.29	17.68	15.64	16.66
	700	88.63	81.79	85.21	45.79	34.87	40.33
P. marinus	100	73.30	75.05	74.16	45.45	57.01	51.23
	200	80.00	81.39	80.69	32.04	31.01	31.52
	400	80.02	77.85	78.94	18.89	16.63	17.76
	700	86.63	82.35	84.49	47.27	36.51	41.89
W. endosymbiont	100	70.71	55.90	63.30	38.83	45.39	42.11
	200	77.56	60.10	68.83	33.23	26.81	30.02
	400	80.43	61.78	71.10	18.05	13.57	15.81
	700	86.66	61.16	73.91	47.90	31.11	39.51

[a]Reads were simulated with 1% sequencing error rate for lengths of 100, 200 and 400 bp, and 0.5% sequencing error rate for length of 700 bp, respectively. The nine genomes are the same as those in Table 2, and were used for testing gene prediction in short reads (16,18).

UNIVERSITEIT GENT

Frameshift error prediction

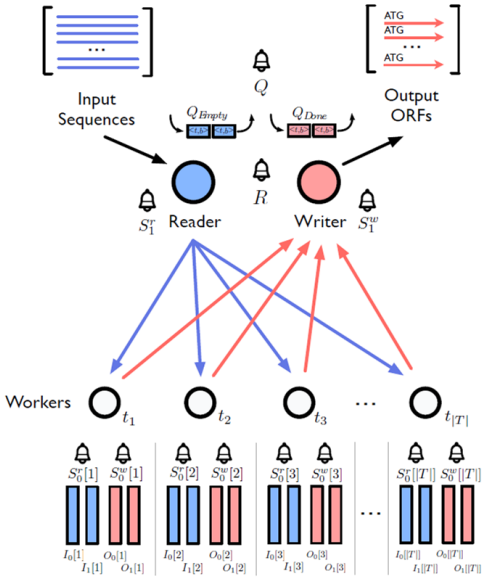




UNIVERSITEIT GENT

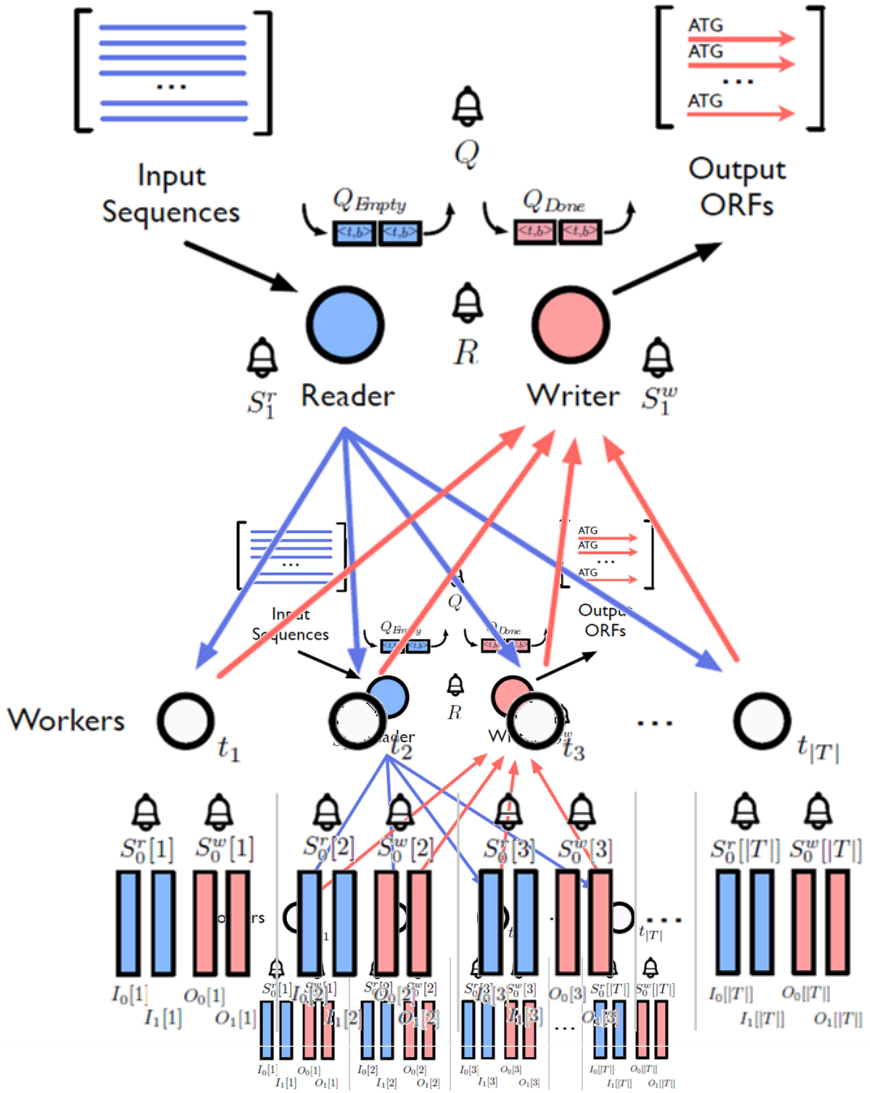
FragGeneScan-Plus

- implementation
  - algorithmic thread synchronization
  - efficient in-memory data management
  - non-blocking I/O operations
  - upfront global parameter computation
  - reduced function calls



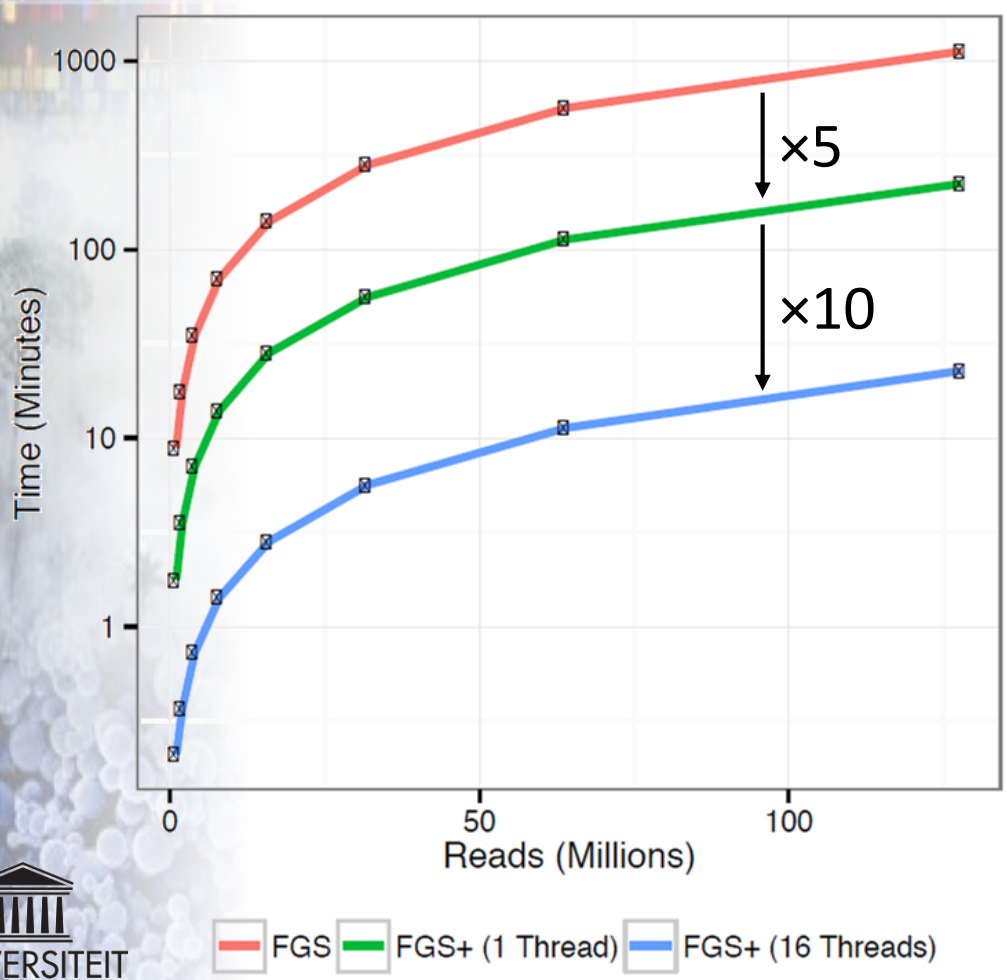
UNIVERSITEIT GENT

FragGeneScan-Plus



UNIVERSITEIT GENT

FragGeneScan-Plus



Empirical performance of *FragGeneScan* and *FragGeneScan-Plus* on progressively larger subsamples from an Illumina-sequenced soil metagenome. *FragGeneScan-Plus* when run with 16-threads (blue), on a hyper-threaded 8-core machine, is approximately 50-times faster than the original implementation of *FragGeneScan* (red). This improvement is attributable to both serial and parallel improvements. *FragGeneScan-Plus* single-threaded (green) is approximately 5-times faster, owing to serial improvements in system calls, memory management, upfront global parameter computation, and code logic. The remaining approximate 10-times speedup can be attributed to the parallel implementation and algorithmic thread synchronization.

UNIVERSITEIT GENT

References

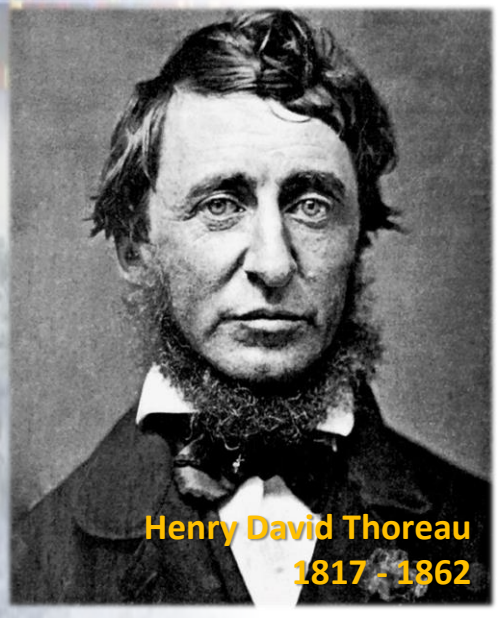
- **Rho M, Tang H, Ye Y (2010)**. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic acids research* **38(20)**, e191-e191.
- **Kim D, Hahn AS, Wu SJ, Hanson NW, Konwar KM, Hallam SJ (2015)**. FragGeneScan-Plus for scalable high-throughput short-read open reading frame prediction. *In* Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference, 1-8.

UNIVERSITEIT GENT

Questions

UNIVERSITEIT GENT

The sky is the limit…



"It is not enough to be busy; so are the ants.
The question is: What are we busy about?"

— Henry David Thoreau

UNIVERSITEIT GENT