GeneMark

A family of gene prediction programs developed at
Georgia Institute of Technology , Atlanta, Georgia, USA.

What's New: A new algorithm, **BRAKER1**, an RNA-seq based eukaryotic genome annotation pipeline -- using GeneMark-ET and AUGUSTUS

Supported by NIH

Gene Prediction in Bacteria, Archaea, Metagenomes and Metatranscriptomes

Novel genomic sequences can be analyzed either by the self-training program **GeneMarkS** (sequences longer than 50 kb) or by **GeneMark.hmm with Heuristic models**. For many species pre-trained model parameters are ready and available through the **GeneMark.hmm** page. Metagenomic sequences can be analyzed by **MetaGeneMark** , the program optimized for speed.

Gene Prediction in Eukaryotes

Novel genomes can be analyzed by the program **GeneMark-ES** utilizing unsupervised training. Note that GeneMark-ES has a special mode for analyzing fungal genomes. Recently, we have developed a semi-supervised version of GeneMark-ES, called GeneMark-ET that uses RNA-Seq reads to improve training. For several species pre-trained model parameters are ready and available through the **GeneMark.hmm** page.

Gene Prediction in Transcripts

Sets of assembled eukaryotic transcripts can be analyzed by the modified **GeneMarkS** algorithm (the set should be large enough to permit self-training). A single transcript can be analyzed by a special version of **GeneMark.hmm with Heuristic models**. A new advanced algorithm GeneMarkS-T was developed recently (manuscript sent to publisher); The GeneMarkS-T software (beta version) is available for download .

Gene Prediction in Viruses, Phages and Plasmids

Sequences of viruses, phages or plasmids can be analyzed either by the **GeneMark.hmm with Heuristic models** (if the sequence is shorter than 50 kb) or by the self-training program **GeneMarkS**.

All the software programs mentioned here are available for download and local installation.

The software of GeneMark line is a part of genome annotation pipelines at NCBI, JGI, Broad Institute as well as the following software packages:

- QUAST : quality assessment tool for genome assemblies
  -- using GeneMarkS
- MetAMOS : a modular and open source metagenomic assembly and analysis
  -- using MetaGeneMark
- MAKER2 : a eukaryotic genome annotation pipeline
  -- using GeneMark-ES (along with SNAP and AUGUSTUS)
- BRAKER1 : an RNA-seq based eukaryotic genome annotation pipeline
  -- using GeneMark-ET and AUGUSTUS

For more information see Background and Publications.

Borodovsky Group
Group news

Gene Prediction Programs

- GeneMark
- GeneMark.hmm
- GeneMarkS
- Heuristic models
- MetaGeneMark
- Mirror site at NCBI
- GeneMarkS+
- BRAKER1

Information

- Publications
- Selected Citations
- Background
- FAQ
- Contact

Downloads

- Programs
- Prebuild species models

Other Programs

- UnSplicer
- GeneTack
- Frame-by-Frame
- IPSSP

In silico Biology International Conferences

- 2015
- 2013
- 2011
- 2009
- 2007
- 2005
- 2003
- 2001
- 1999
- 1997

History of development of GeneMark family of gene finding tools

Early algorithms -- supervised training

(need large validated sequence sets for parameter estimation)

- **GeneMark** **prokaryotic and eukaryotic** 1993 (Computers & Chemistry)
- **GeneMark.hmm** **prokaryotic** 1998 (NAR)

Extrapolation of model parameters for short genomes and metagenomes

(no training sets for obvious reasons)

- **GeneMark.hmm with heuristic models** 1998 (NAR) **MetaGeneMark** 2010 (NAR)

Latest algorithms -- unsupervised training

(no need in validated sequence sets)

- **GeneMarkS**, **GeneMarkS-2** **prokaryotic genomes** 2001 (NAR) 2018 (Genome Research)
- **GeneMark-ES/ET/EP** **eukaryotic genomes** 2005 (NAR) 2005 (NAR) 2016 (Bioinformatics)

http://exon.gatech.edu/

GeneMark

The algorithm uses Markov chain models and Bayesian logic

Bayes Theorem

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

If A_i is mutually exclusive and exhaustive, then

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_i P(B \mid A_i)P(A_i)}$$

Using Bayes theorem to model sequences from different genomes

$$\mathrm{P(model_i|data)} = \frac{P(\mathrm{data} \mid \mathrm{model_i})P(\mathrm{model_i})}{\sum_i P(\mathrm{data} \mid \mathrm{model_i})P(\mathrm{model_i})}$$

Two genomes

Yeast (model 1) -- low G+C content 35%: $P_T^1 P_C^1 P_A^1 P_G^1$; $\{P_i^1\}$

Chlamidomonas reinhardtti (model 2) – highG+C content 60%: $P_T^2 P_C^2 P_A^2 P_G^2$; $\{P_i^2\}$

Two genomes

Yeast (model 1) -- low G+C content 35%: $P_T^1 P_C^1 P_A^1 P_G^1$; $\{P_i^1\}$

Chlamidomonas reinhardtti (model 2) – highG+C content 60%: $P_T^2 P_C^2 P_A^2 P_G^2$; $\{P_i^2\}$

$$P(\text{sequence } \mathrm{X_1X_2\cdots X_{200}} \mid \text{model}_1) = \prod_{i=1}^{200} P^1(X_i)$$

$$P(\text{sequence } \mathrm{X_1X_2\cdots X_{200}} \mid \text{model}_2) = \prod_{i=1}^{200} P^2(X_i)$$

$$P(\text{model}_1 \mid \text{sequence}) = \frac{\mathrm{P}(\text{sequence} \mid \text{model}_1)P(\text{model}_1)}{\sum_{i=1}^{2} \mathrm{P}(\text{sequence} \mid \text{model}_i)P(\text{model}_i)} = \frac{\prod_{i=1}^{200} P^1(X_i) \cdot \frac{1}{2}}{\frac{1}{2}(\prod_{i=1}^{200} P^1(X_i) + \prod_{i=1}^{200} P^2(X_i))}$$

Likelihood L is defined as:

$$L = \frac{P(\text{sequence} \mid \text{model}_1)}{P(\text{sequence} \mid \text{model}_2)} = \frac{\prod_{i=1}^{200} P_{X_i}^1}{\prod_{i=1}^{200} P_{X_i}^2} = \frac{(P_A^1)^{n_A}(P_C^1)^{n_C}(P_G^1)^{n_G}(P_T^1)^{n_T}}{(P_A^2)^{n_A}(P_C^2)^{n_C}(P_G^2)^{n_G}(P_T^2)^{n_T}}$$

Log Likelihood LL is defined as:

$$LL = \log\left[\left(\frac{P_A^1}{P_A^2}\right)^{n_A}\left(\frac{P_T^1}{P_T^2}\right)^{n_T}\left(\frac{P_C^1}{P_C^2}\right)^{n_C}\left(\frac{P_G^1}{P_G^2}\right)^{n_G}\right] = \sum_{\alpha=A,C,T,G} n_\alpha \log\left(\frac{P_\alpha^1}{P_\alpha^2}\right)$$

If L>1 (or LL>0), then the sequence is likely to be from model 1.
If L<1 (or LL<0), then the sequence is likely to be from model 2.
If L=1 (or LL=0), then cannot decide.

|LL| ~score because if |LL| is high, then we can be confident of the prediction, whereas if |LL| is low, then we cannot.

$$S_\alpha = \log\frac{P_\alpha^1}{P_\alpha^2}$$

$$LL = \sum n_\alpha S_\alpha$$

The degree of randomness can be measured by entropy.

$$H = -\sum_{i=1}^{n} P_i \log P_i = E(-\log P)$$

For n=4 (as in the case of DNA sequence), the entropy is highest for p=1/4.

$$LL = \sum_{\alpha=A,T,C,G} n_\alpha \log\left(\frac{P_\alpha^1}{P_\alpha^2}\right)$$

Relative entropy (Kullback-Liebler distance)

$$D(P^1 \parallel P^2) = \sum_{\alpha=A,T,C,G} P_\alpha^1 \log\left(\frac{P_\alpha^1}{P_\alpha^2}\right) \geq 0$$

$$D(P^2 \parallel P^1) = \sum_{\alpha=A,T,C,G} P_\alpha^2 \log\left(\frac{P_\alpha^2}{P_\alpha^1}\right) \geq 0$$

$$L = \frac{\prod P_{X_i}^1}{\prod P_{X_i}^2} = \prod\left(\frac{P_{X_i}^1}{P_{X_i}^2}\right)$$

$$LL = \sum_{i=1}^{n} \log\left(\frac{P_{X_i}^1}{P_{X_i}^2}\right) = \sum_{\alpha \in A} n_\alpha \log\left(\frac{P_\alpha^1}{P_\alpha^2}\right)$$

$$\frac{LL}{n} = \sum_{\alpha=A,T,C,G} \frac{n_\alpha}{n} \log\left(\frac{P_\alpha^1}{P_\alpha^2}\right)$$

$$E(LL) = \sum_{\alpha \in A} E(n_\alpha) \cdot \log\left(\frac{P_\alpha^1}{P_\alpha^2}\right)$$

If n is large and the sequence is from model 1, then $\frac{n_\alpha}{n} \to P_\alpha^1$

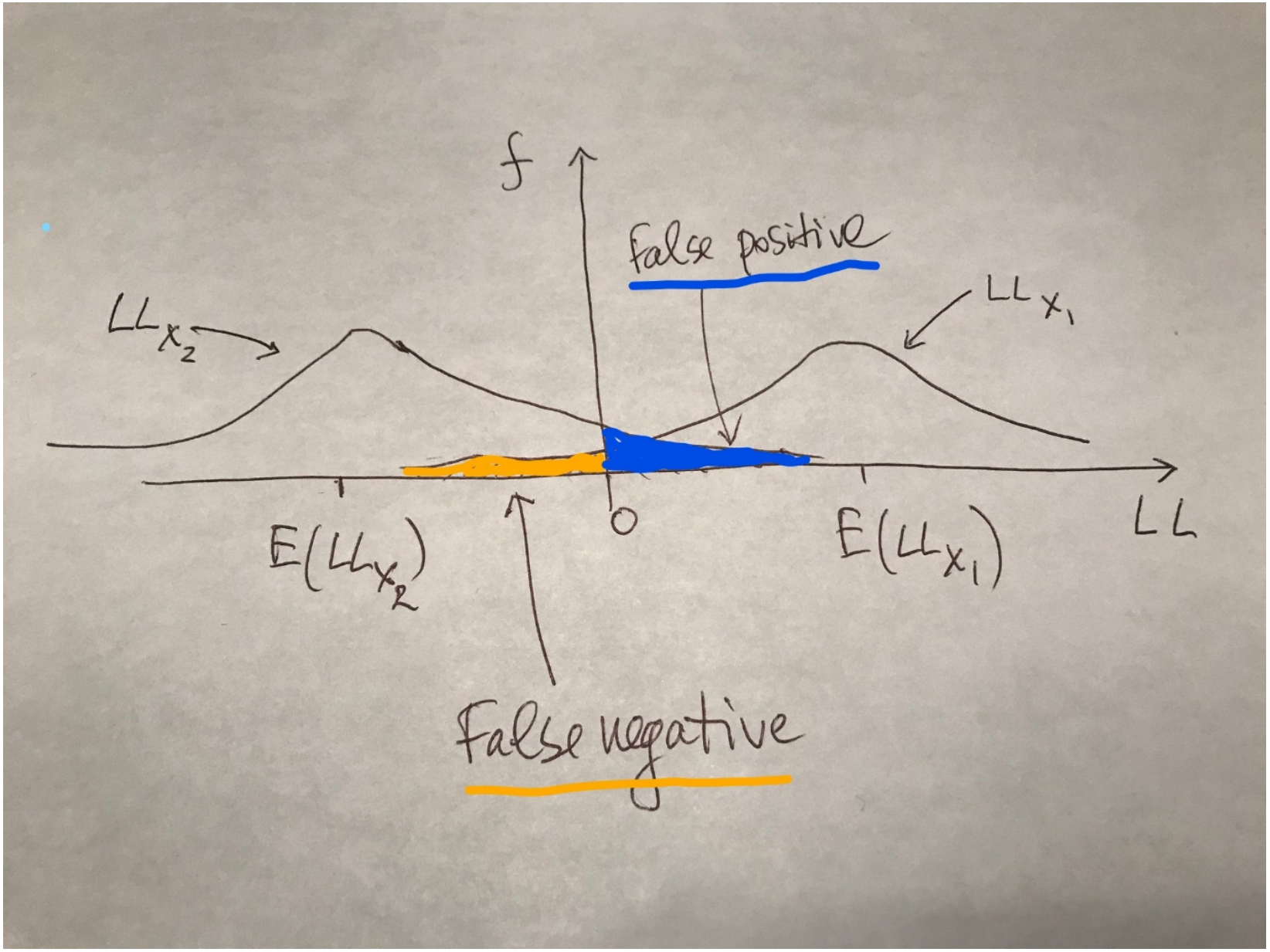
$$E(LL \mid x \in X_1) = \sum_{\alpha \in A} nP_\alpha^1 \cdot \log\left(\frac{P_\alpha^1}{P_\alpha^2}\right) = n\sum_{\alpha \in A} P_\alpha^1 \log\left(\frac{P_\alpha^1}{P_\alpha^2}\right) = nD(P^1 \parallel P^2) \geq 0$$

$$Var(LL \mid x \in X_1) = Var\left(\sum n_\alpha \log \frac{P_\alpha^1}{P_\alpha^2}\right) = \sum Var(n_\alpha) \cdot \log^2 \frac{P_\alpha^1}{P_\alpha^2} = \sum nP_\alpha^1(1 - P_\alpha^1) \cdot \log^2 \frac{P_\alpha^1}{P_\alpha^2}$$

If n is large and the sequence is from model 2, then $\frac{n_\alpha}{n} \to P_\alpha^2$

$$E(LL \mid x \in X_2) = \sum_{\alpha \in A} nP_\alpha^2 \cdot \log\left(\frac{P_\alpha^1}{P_\alpha^2}\right) = n\sum_{\alpha \in A} P_\alpha^2 \log\left(\frac{P_\alpha^1}{P_\alpha^2}\right) = -nD(P^2 \parallel P^1) \leq 0$$

$$Var(LL \mid x \in X_2) = Var\left(\sum n_\alpha \log \frac{P_\alpha^1}{P_\alpha^2}\right) = \sum Var(n_\alpha) \cdot \log^2 \frac{P_\alpha^1}{P_\alpha^2} = \sum nP_\alpha^2(1 - P_\alpha^2) \cdot \log^2 \frac{P_\alpha^1}{P_\alpha^2}$$


f
False positive
LL_{x_2}
LL_{x_1}
$E(LL_{x_2})$
$E(LL_{x_1})$
LL
False negative

Application to modeling coding and non-coding regions

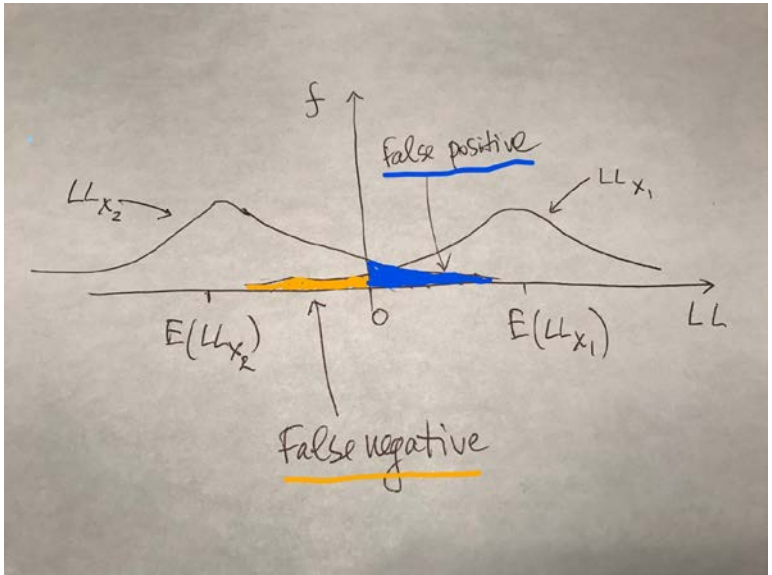
Experiment: a sample of 100nt-long segments of a bacterial genome is given
Need to identify the segments as coding or non-coding based on the LL values computed .

*False negative (type I) error: an error that identifies a true *coding* region as non-coding (coding is the null hypothesis – H_0, non-coding is the alternative one – H_1)

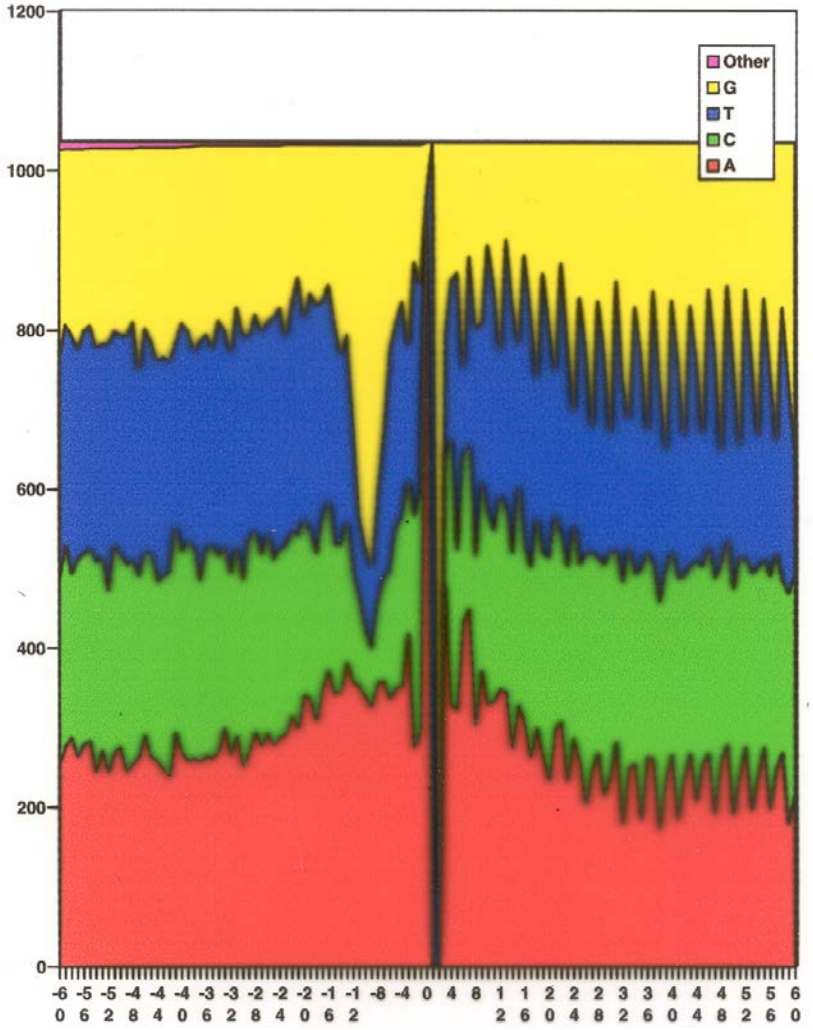
*False positive (type II) error: an error that identifies a true *non-coding* region as coding

*The smaller the errors, the higher the discrimination power of the test.
We want $nD(P^1||P^2)$ to be high and bring higher discrimination power.
We can increase n only within certain limits – characteristic length of the functional region
The higher the value of relative entropy $D(P^1||P^2)$, the smaller n can be to deliver the same level of discrimination power.

*The simplest multinomial model with frequencies of single nucleotides is not accurate enough to distinguish coding from non-coding regions when fragments of about 100 nt are used.



… compositional patterns observed in non-coding and protein-coding DNA
(visualized by alignment of 1000+ *E. coli* genes,
the sequences containing gene starts with flanking regions)



NUCLEOTIDE FREQUENCIES IN *M. JANNASCHII* GENOME

Coding regions

	Frame 1	Frame 2	Frame 3
T	20.6	33.3	35.3
C	8.6	15.8	9.4
A	37.8	37.0	39.9
G	32.9	13.9	15.4

Codon positions 1, 2, 3

Non-coding regions

T	**36.0**
C	**13.8**
A	**36.0**
G	**14.2**

NUCLEOTIDE FREQUENCIES IN *M. JANNASCHII* GENOME

Coding regions — positional frequencies

	Frame 1	Frame 2	Frame 3
T	20.6	33.3	35.3
C	8.6	15.8	9.4
A	37.8	37.0	39.9
G	32.9	13.9	15.4

Codon positions 1, 2, 3

Non-coding regions

T	**36.0**
C	**13.8**
A	**36.0**
G	**14.2**

POSITIONAL DINUCLEOTIDE FREQUENCIES IN *M. JANNASCHII* CODING REGIONS

Codon positions 1-2, 2-3, 3-1

	Frame 1	Frame 2	Frame 3
TT	11.39	13.58	7.44
TC	2.87	2.74	3.03
TA	4.37	12.22	11.53
TG	2.01	4.58	13.44
CT	2.31	5.87	3.30
CC	3.39	1.39	1.39
CA	2.87	8.00	3.38
CG	0.08	0.63	1.37
AT	12.54	12.68	7.31
AC	4.08	3.93	2.47
AA	15.66	13.32	17.11
AG	5.41	7.14	13.09
GT	6.88	3.32	2.58
GC	5.54	1.38	1.76
GA	14.18	6.45	5.66
GG	6.42	2.76	5.12

POSITIONAL DINUCLEOTIDE FREQUENCIES IN *M. JANNASCHII* CODING REGIONS

Codon positions 1-2, 2-3, 3-1

	Frame 1	Frame 2	Frame 3
TT	11.39	13.58	7.44
TC	2.87	2.74	3.03
TA	4.37	12.22	11.53
TG	2.01	4.58	13.44
CT	2.31	5.87	3.30
CC	3.39	1.39	1.39
CA	2.87	8.00	3.38
CG	0.08	0.63	1.37
AT	12.54	12.68	7.31
AC	4.08	3.93	2.47
AA	15.66	13.32	17.11
AG	5.41	7.14	13.09
GT	6.88	3.32	2.58
GC	5.54	1.38	1.76
GA	14.18	6.45	5.66
GG	6.42	2.76	5.12

INHOMOGENEOUS MARKOV MODEL

```
TTA | CGT | GCA | TGC | GTA | CGT | GCA
123 | 123 | ...
```

$$p^1 = \| p_{ij}^1 \|,\ p^2 = \| p_{ij}^2 \|,\ p^3 = \| p_{ij}^3 \|$$

$$p_{ij}^k = \frac{N^k(XY)}{N^k(X)}, i = X, j = Y$$

M. JANNASCHII

TRANSITION PROBABILITIES FOR CODON POSITIONS 1, 2

(PRIMARY CODING FUNCTION)

	T	C	A	G
T	0.552	0.138	0.212	0.098
C	0.262	0.397	0.333	0.008
A	0.336	0.108	0.412	0.142
G	0.209	0.169	0.428	0.194

TRANSITION PROBABILITIES FOR CODON POSITIONS 2, 3

(SYNONYMOUS CODON USAGE)

	T	C	A	G
T	0.404	0.083	0.366	0.148
C	0.370	0.087	0.505	0.039
A	0.340	0.107	0.360	0.193
G	0.237	0.099	0.469	0.196

TRANSITION PROBABILITIES FOR CODON POSITIONS 3, 1

	T	C	A	G
T	0.207	0.086	0.323	0.384
C	0.349	0.149	0.357	0.146
A	0.181	0.062	0.427	0.330
G	0.170	0.115	0.375	0.340

Computation of an a posteriori probability of protein-coding function in six frames for a given DNA fragment

$$P(S_n|COD_1) = P_0^1(x_1)P^1(x_2|x_1)P^2(x_3|x_2)\ldots P^2(x_n|x_{n-1}),$$

$$P(S_n|COD_2) = P_0^2(x_1)P^2(x_2|x_1)P^3(x_3|x_2)\ldots P^3(x_n|x_{n-1}),$$

$$P(S_n|COD_3) = P_0^3(x_1)P^3(x_2|x_1)P^1(x_3|x_2)\ldots P^1(x_n|x_{n-1}).$$

$$P(S_n|COD_4) = P_0^4(x_1)P^4(x_2|x_1)P^5(x_3|x_2)\ldots P^5(x_n|x_{n-1}),$$

$$P(S_n|COD_5) = P_0^5(x_1)P^5(x_2|x_1)P^6(x_3|x_2)\ldots P^6(x_n|x_{n-1}),$$

$$P(S_n|COD_6) = P_0^6(x_1)P^6(x_2|x_1)P^4(x_3|x_2)\ldots P^4(x_n|x_{n-1}).$$

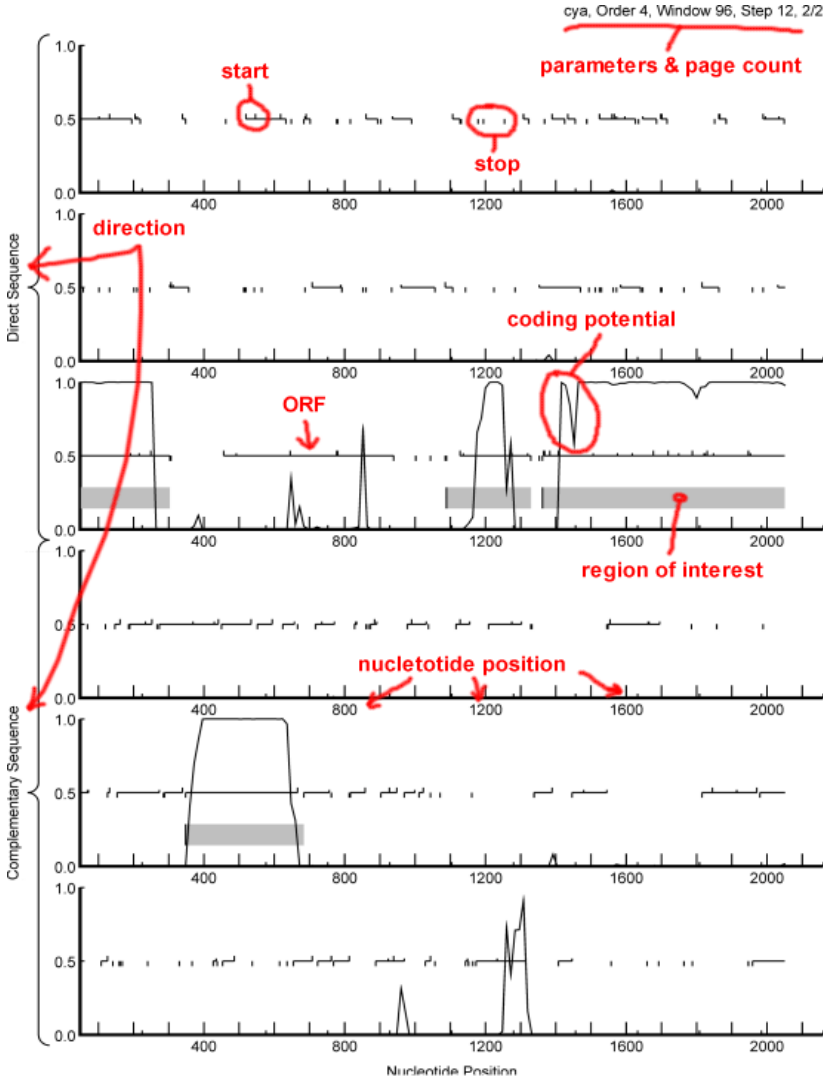
$$P(S_n|NON) = P_0^n(x_1)P^n(x_2|x_1)P^n(x_3|x_2)\ldots P^n(x_n|x_{n-1}).$$

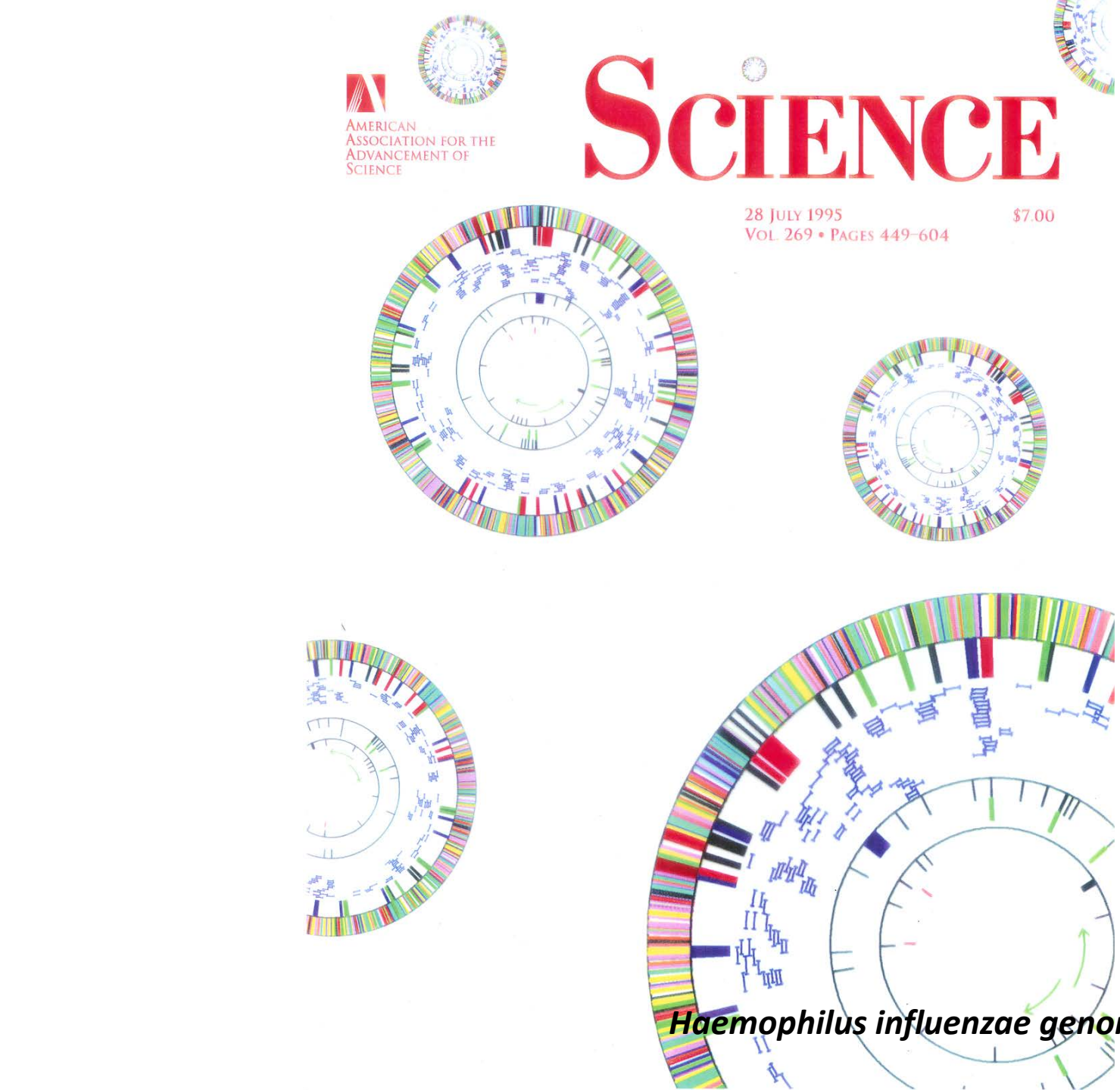
$$\mathrm{P}(\mathrm{COD}_i|S_n) = \frac{P(S_n|COD_i)P(COD_i)}{\sum_{i=1}^{6} \mathrm{P}(\mathrm{S_n}|COD_i)P(COD_i) + \mathrm{P}(\mathrm{S_n}|NON)P(NON)}$$

$$i = 1,2,3,4,5,6.$$

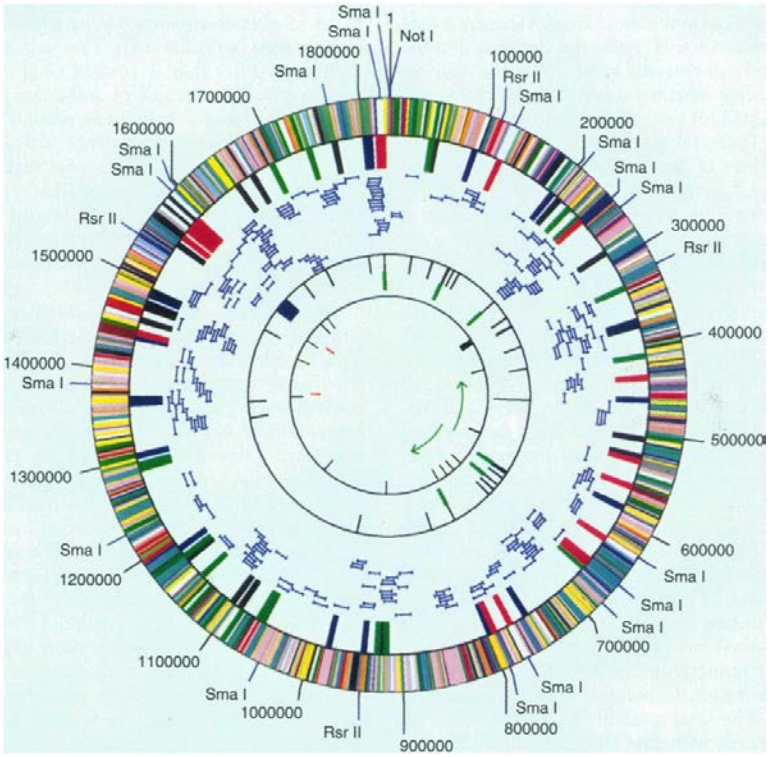
Borodovsky & McIninch, Computers & Chemistry 1993

A posteriori probabilities of protein coding function (computed in sliding windows by the GeneMark algorithm)





***Haemophilus influenzae* genome, 28 July 1995**

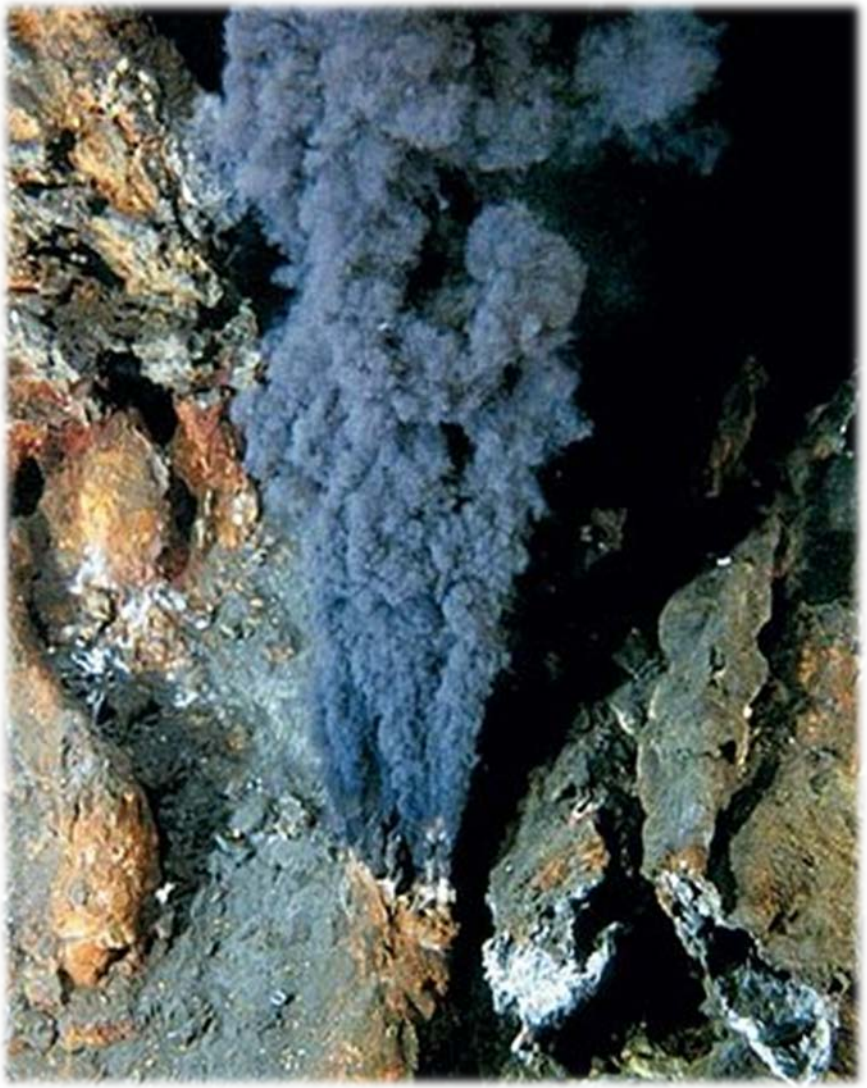


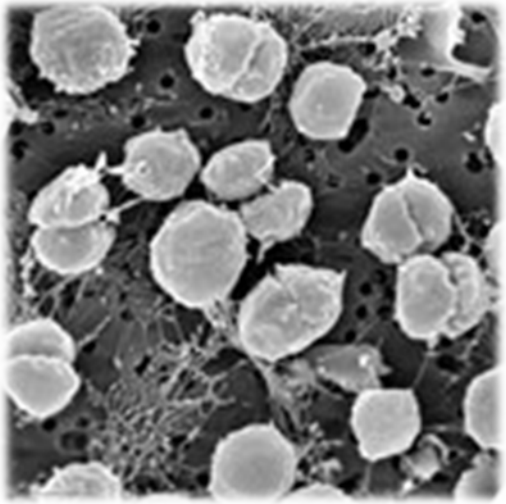
In 1995, *the genomic era started* with decoding 1,830,140bp genome of *Haemophilus influenzae,* an opportunistic pathogen, sequenced by the whole genome shotgun method.

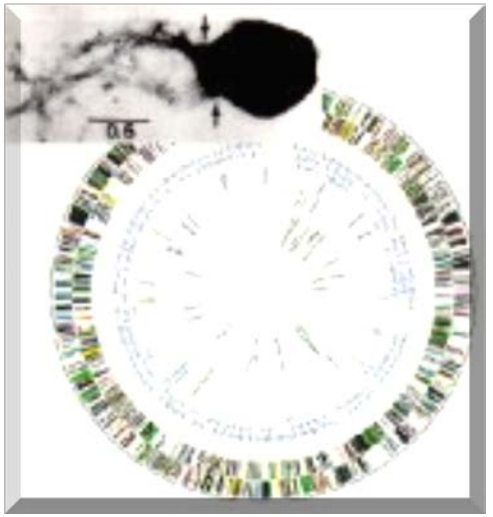
1740 protein-coding genes were identified by GeneMark

Fleischmann et al., Science 1995
Borodovsky & McInnch, 1993

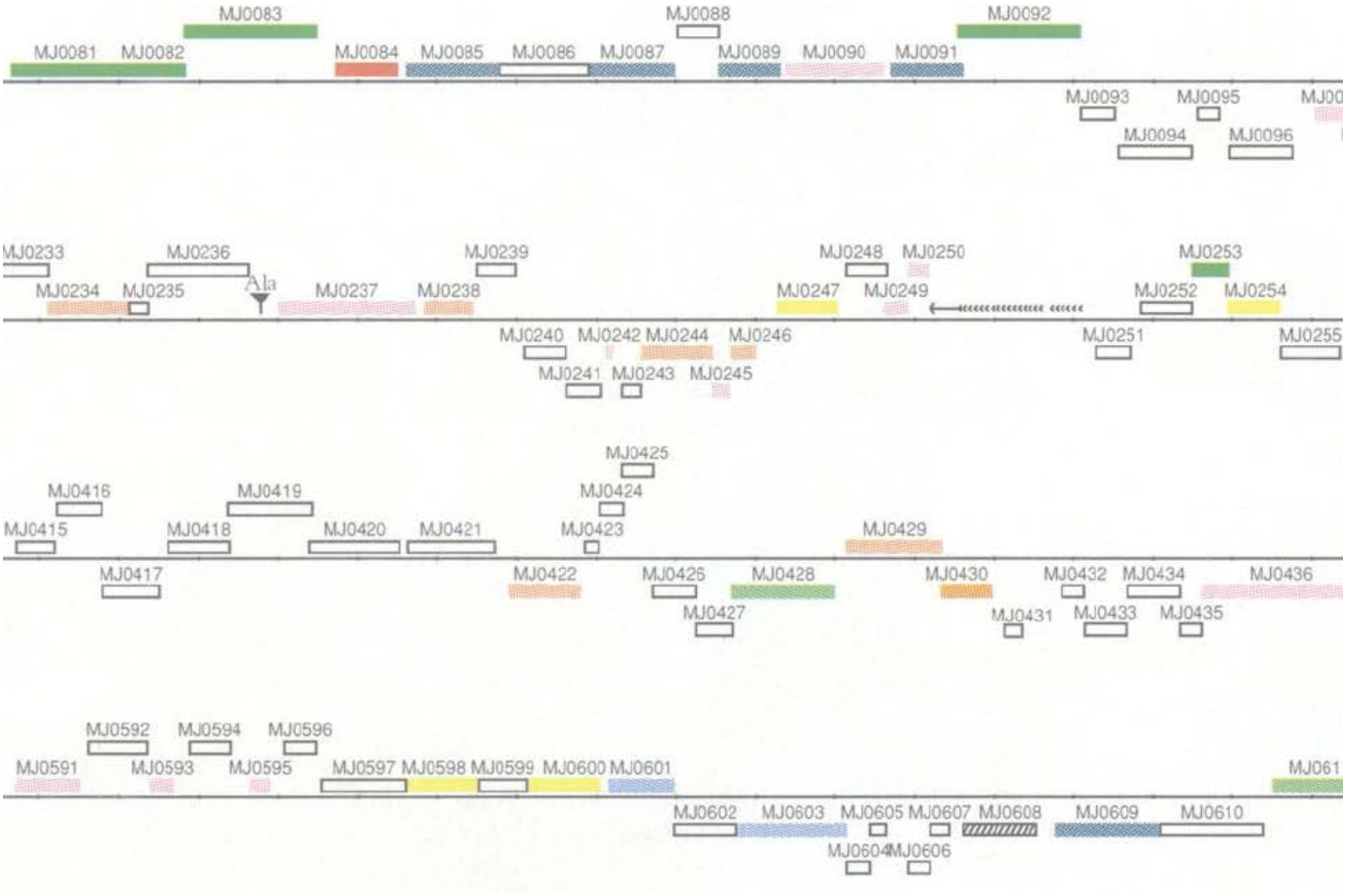
… followed with annotation of genomes of *E. coli, B. subtilis, M. jannaschii, H. pilory* in 1996-1997 .







Methanogenic archaeon *Methanococcus jannaschii, 1996*



Methanogenic archaeon, *Methanococcus jannaschii, Bult et al. 1996,* *Science* 273:5278

GeneMark.hmm

- The inhomogeneous (three matrix) Markov chains become incorporated in the Hidden Markov model
- They are needed to compute HMM emission probabilities for whole sequence fragments

Each nucleotide is ***labeled***:
n – non-coding, c – coding in direct strand, r – coding in reverse strand

```
DNA sequence:          AAGTTATGAACTATGGGACATGTTATCCGTAAGGTTATAAGC
Possible labels        nnnnnddddddddddddddddddddddddddddnnrrrrrrrr
                       nnnrrrrrrrrrrrrrrrrrnnnnnnnnnnnnnnnnrrrrr
                       nnnnnnnnnnnnnnnnnnnnnrrrrrrrrrrrrrrrrrr
                       nnnnnnnnnnnnddddddddddddddddddddddddddddd
```

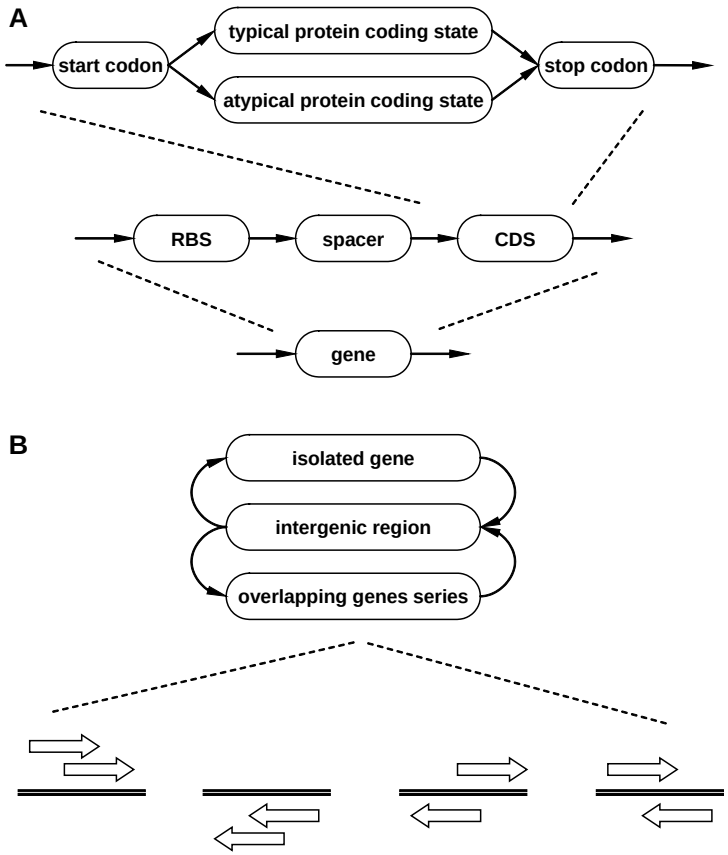
Effectively we do assign labels in process of genome annotation

```
DNA sequence:          AAGTTATGAACTATGGGACATGTTATCCGTAAGGTTATAAGC
labels:
        non-coding     *****                              **
      coding direct          ***************************
     coding reverse                                        ********
Annotation sequence    nnnnnddddddddddddddddddddddddddddnnrrrrrrrr
```

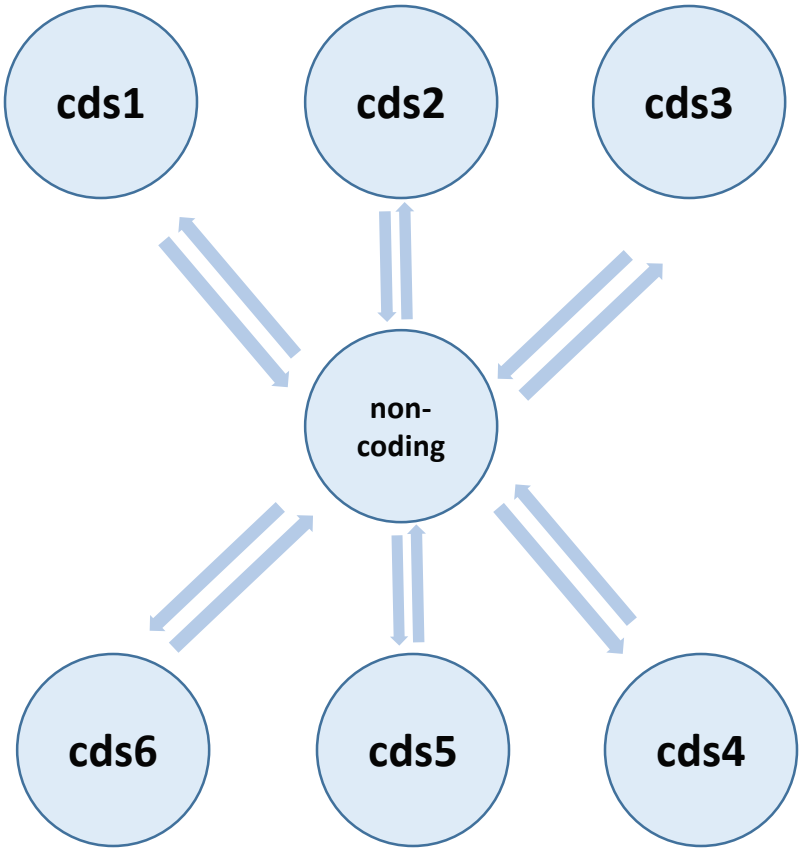
In fragments *carrying the same label patterns of* nucleotide ordering are 'label specific' and are described by the Markov chain models. Labels are interpreted as hidden states of an HMM.

The Viterbi algorithm computes most likely sequence of labels for a nucleotide sequence by finding max joint probability of a sequence of nucleotides and sequence of labels given the parameters of the HMM

Maximize Prob (seq of labels| DNA seq) ~ Prob (seq of labels & DNA seq)

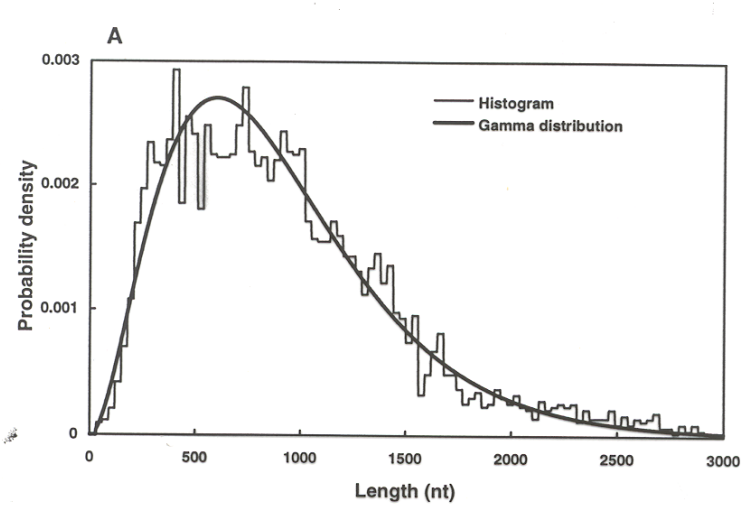


(**A**) Part of HMM representing in GeneMark.hmm a single gene with RBS. For simplicity, only the direct strand is shown. (**B**) In this simplified diagram of hidden state architecture, the state 'gene' represents a genomic segment including RBS, an RBS spacer and a protein-coding sequence (CDS).

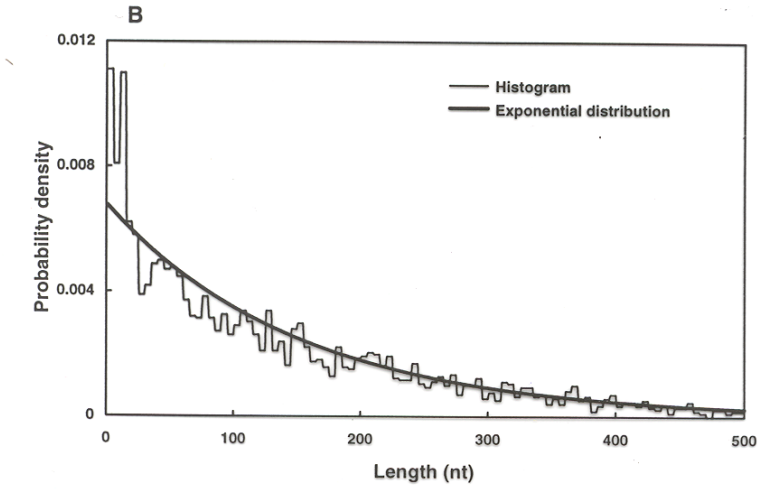


GeneMark could be interpreted as an approximation to *posterior decoding* (the HMM-like algorithm)

‘Classic’ HMM does not allow to correctly describe experimentally observed length distribution of protein-coding genes



An HMM with explicit distributions of durations of runs of hidden states is needed.



Lukashin & Borodovsky, Nucl Acids Res, 1998

CORE ALGORITHM

The algorithm determines the maximum likelihood path through hidden states $(q_1^* d_1^*)(q_2^* d_2^*)...(q_M^* d_M^*)$ provides the maximum probability value $P_{\max}$ of generating the given sequence $B_1 B_2 ... B_{L-1} B_L$ given the specified HMM:

$$P_{\max} = \max_{\substack{(q_1 d_1)(q_2 d_2)...(q_M d_M) \\ \sum_{s=1}^{M} d_s = L}} Prob\{(q_1 d_1)(q_2 d_2).....(q_M d_M), B_1 B_2 ... B_{L-1} B_L\}$$

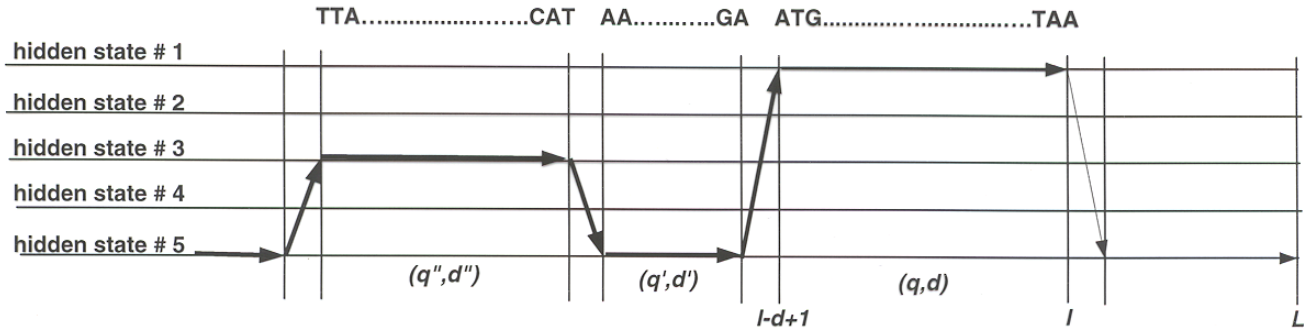
The probability that the maximum likelihood path reaching state q_m after the first l-d_m+1 nucleotides ends at l is

$$\delta_l(q_m, d_m) = \max_{\substack{(q_1 d_1)...(q_{m-1} d_{m-1}) \\ \sum_{s=1}^{m-1} d_s = l - d_m}} \left[Prob\{(q_1 d_1)...(q_{m-1} d_{m-1}), B_1 ... B_{l-d_m}\} a_{q_{m-1} q_m} \right] p_{q_m}(d_m) P_{q_m}(B_{l-d_m+1} ... B_l)$$

Direct coding

Reverse coding

Non-coding



By induction ($m \geq 2$)

$$\delta_l(q_m, d_m) = \max_{(q_{m-1} d_{m-1})} \left[\delta_{l-d_m}(q_{m-1}, d_{m-1}) a_{q_{m-1} q_m} \right] p_{q_m}(d_m) P_{q_m}(B_{l-d_m+1} ... B_l)$$

$$\left(q_l^*(q), d_l^*(q)\right) = \arg\max_{(q_m d_m)} \left[\delta_l(q_m, d_m) a_{q_m q} \right] \qquad 2 \leq l \leq L-1$$

$$P^* = \max_{(q_M d_M)} \delta_L(q_M, d_M)$$

$$\left(q_L^*, d_L^*\right) = \arg\max_{(q_M d_M)} \left[\delta_L(q_M, d_M) \right]$$

Lukashin & Borodovsky,
Nucl Acids Res, 1998

www.ncbi.nlm.nih.gov/genome/annotation_prok/process/

NCBI Resources How To

Genome

Genome

Limits Advanced

Prokaryotic Annotation Home | Documentation | Complete Genome Submission | WGS Genome Submission

NCBI Prokaryotic Genome Annotation Process

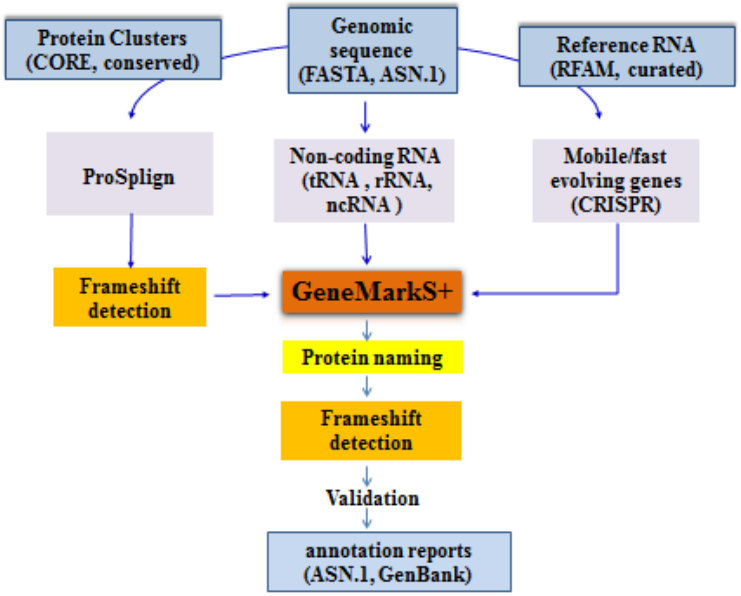
Go back to NCBI Prokaryotic Genome Annotation Pipeline

The NCBI Prokaryotic Annotation Pipeline combines a gene calling algorithm with similarity-based gene detection approach. The p
16S, 23S), tRNAs and small non-coding RNAs. Gene prediction algorithms rely only on statistical properties of DNA and a trainir
ab initio prediction are:

1. predicting coding genes in the region of non-coding RNA
2. predicting two or more coding regions in the region of a single frameshifted gene
3. giving a preference to a longer open reading frame and missing a conserved gene in the same region but different frame
4. incorrect prediction of atypical (phage-associated or horizontally transferred) genes, mobile elements and fast evolving syst

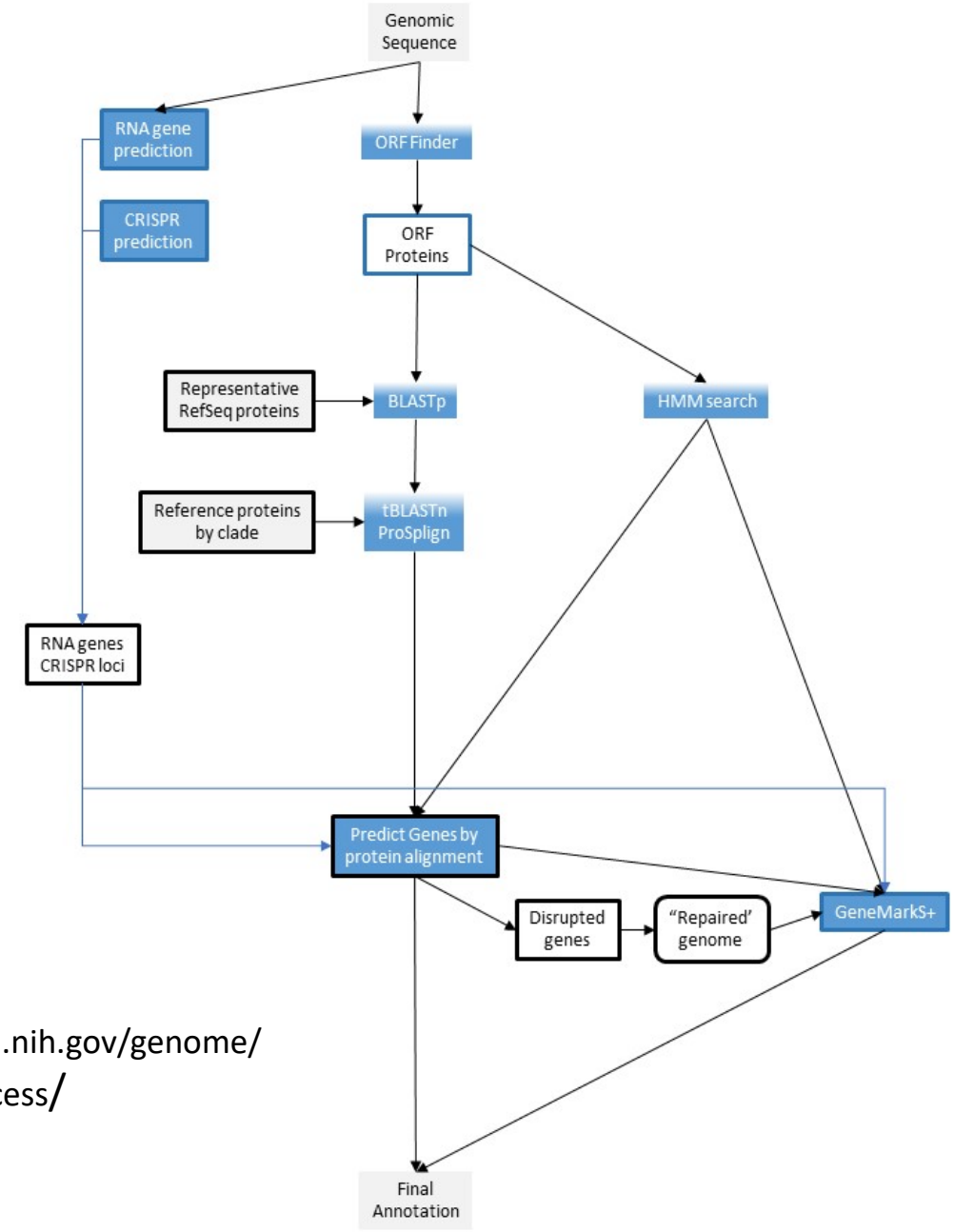
Corrections of these problems are typically implemented as a post-processing step. NCBI, in collaboration with the GeneMark tea
alignments and statistical predictions. A new application, GeneMarkS+, takes alignment data as an input and incorporates that inf

The flowchart below describes the major components of the pipeline:



Twenty years on

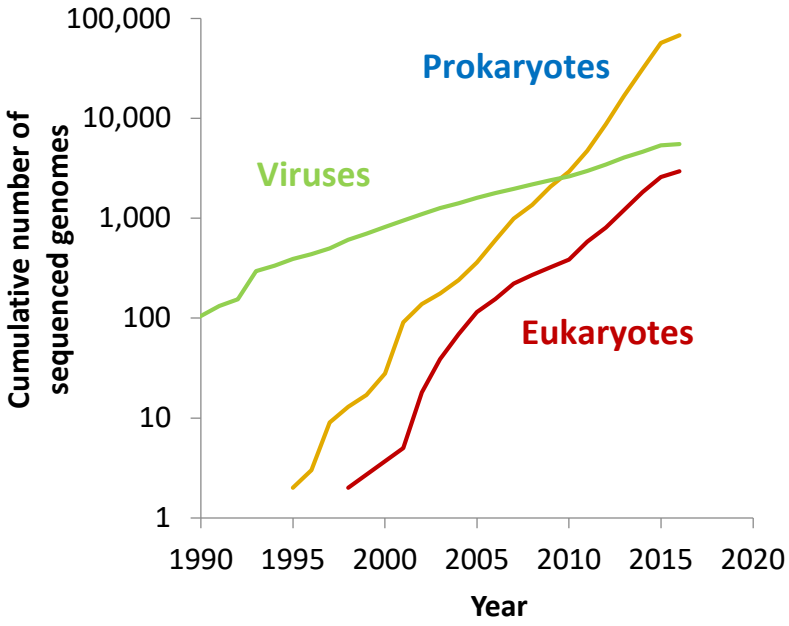
Tatusova et al., 2016, Nucl. Acids Res



https://www.ncbi.nlm.nih.gov/genome/annotation_prok/process/

The 2017 version of the pipeline

Growth of GenBank

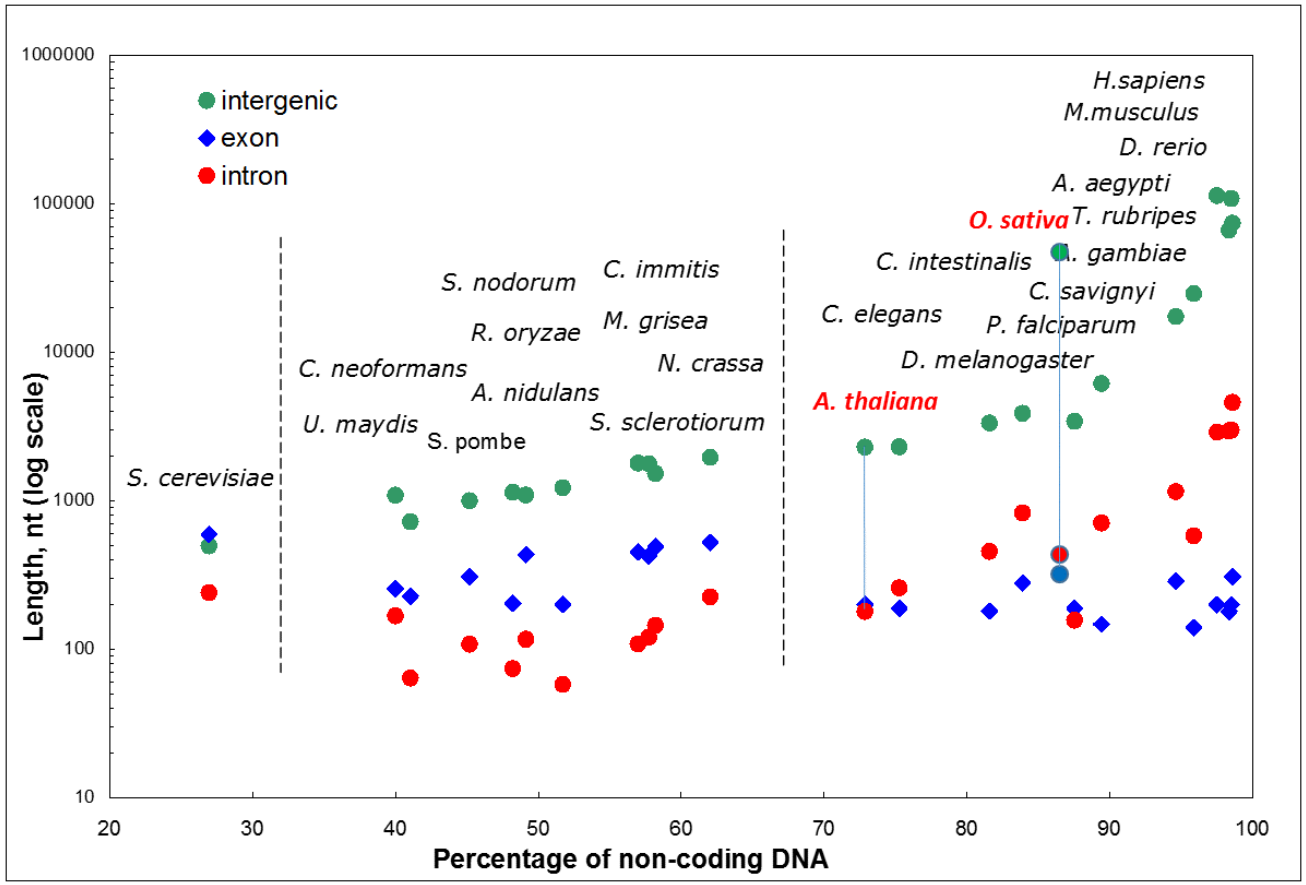


	Genomes
Eukaryotes	
Animals	**1,173**
Plants	**408**
Fungi	**2,637**
Protists	**509**
Prokaryotes	
Bacteria	**116,711**
Archaea	**1,883**

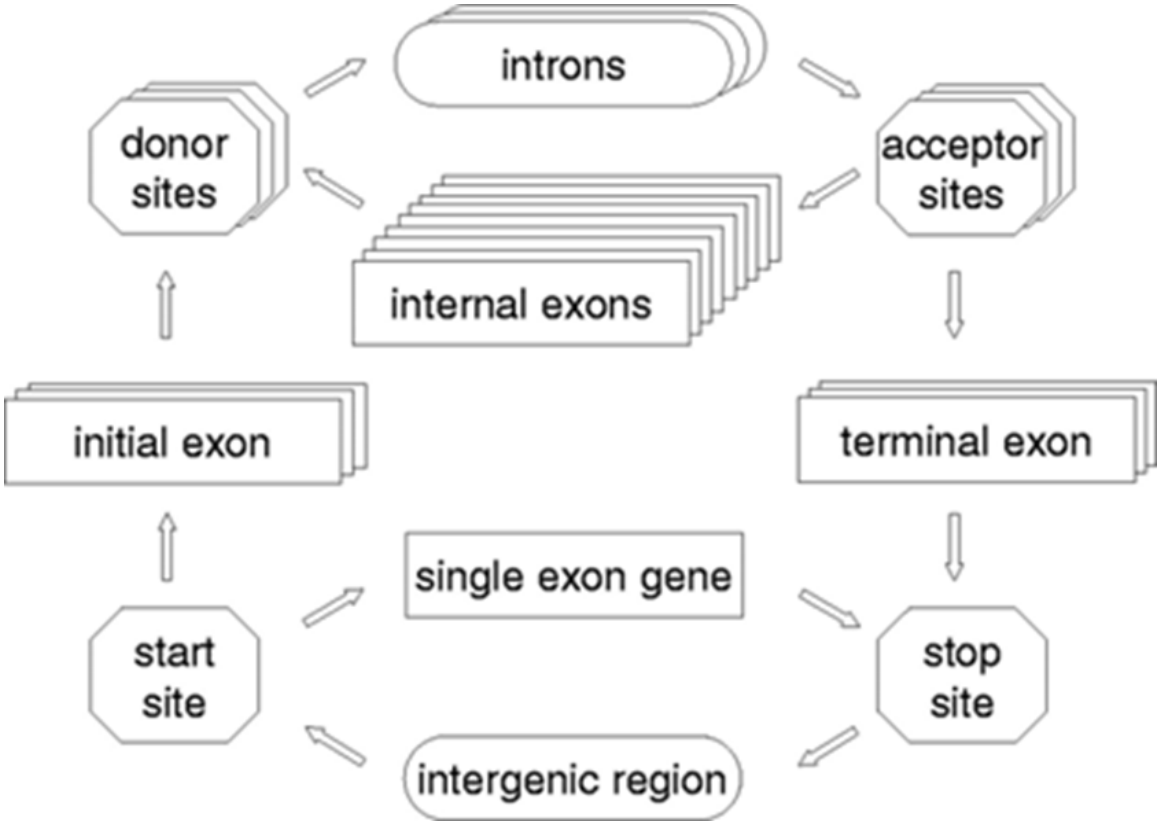
http://www.ncbi.nlm.nih.gov/genome/browse/#

Diversity of eukaryotic genome organization

average length of exon, intron and intergenic region
vs percentage of non-coding DNA in genome

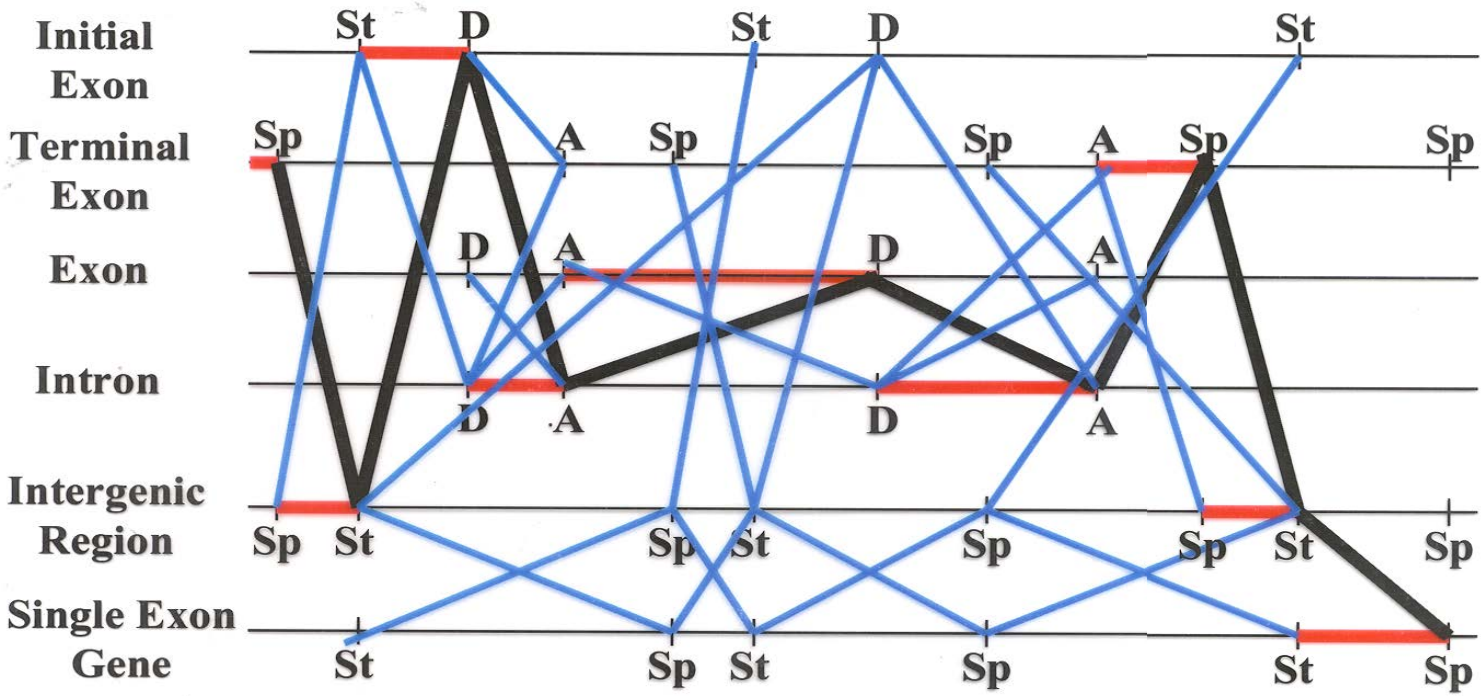


Hidden states of GMM used in eukaryotic GeneMark.hmm



28 states for direct strand; total – 57 states

Lomsadze et al. Nucl. Acids Res. 2005



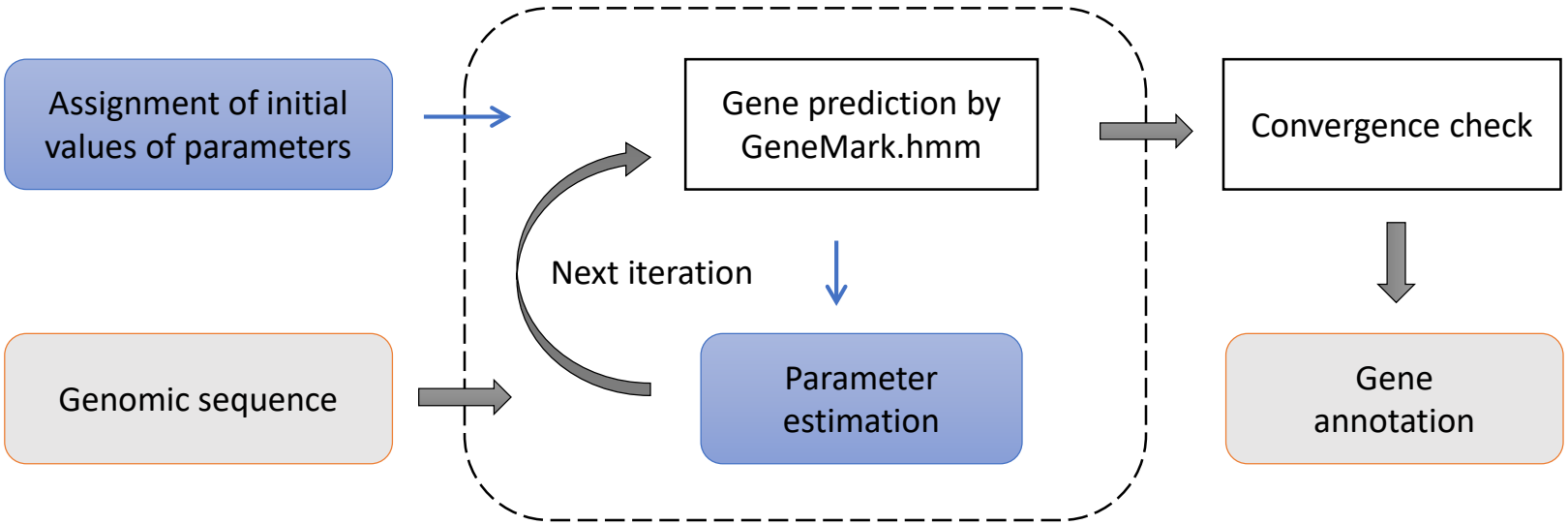
St – start codon, Sp – stop codon, D – donor site, A – acceptor site

possible elements of exon-intron structures (connections between sites)

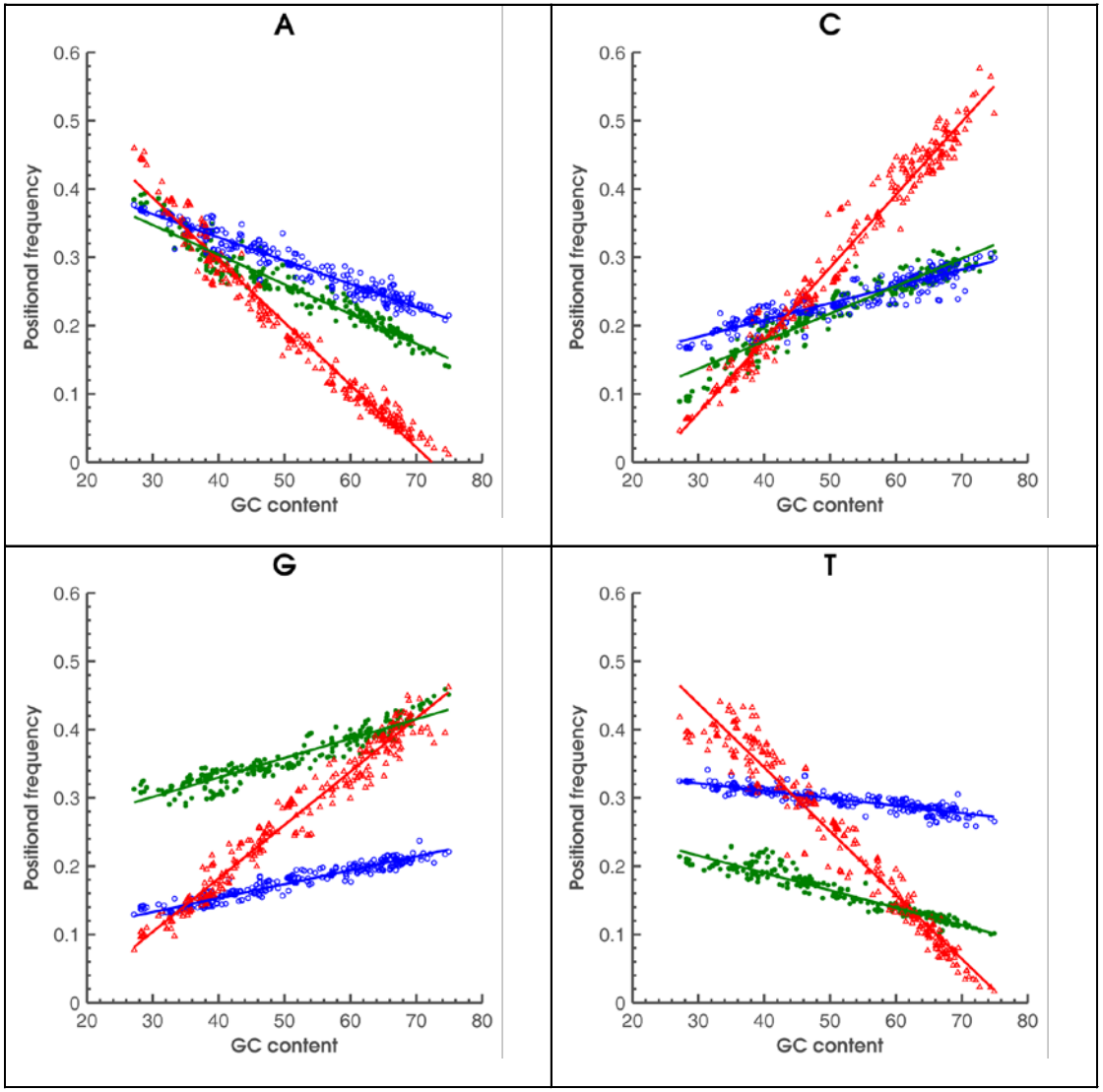
predicted chain of connections (by the Viterbi algorithm)

classified fragments of the sequence parse

Eukaryotic gene finder with unsupervised training - GeneMark-ES



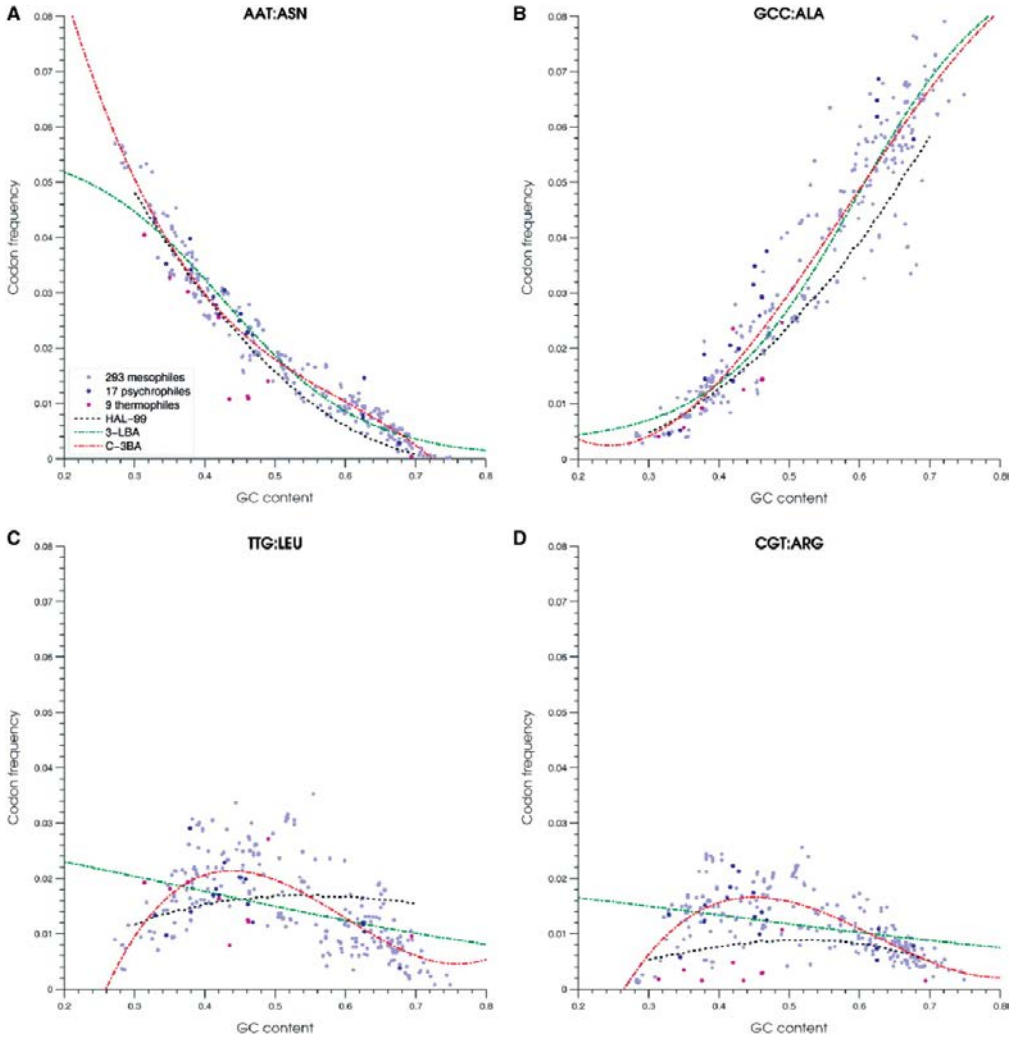
Lomsadze et al. Nucl. Acids Res. 2005



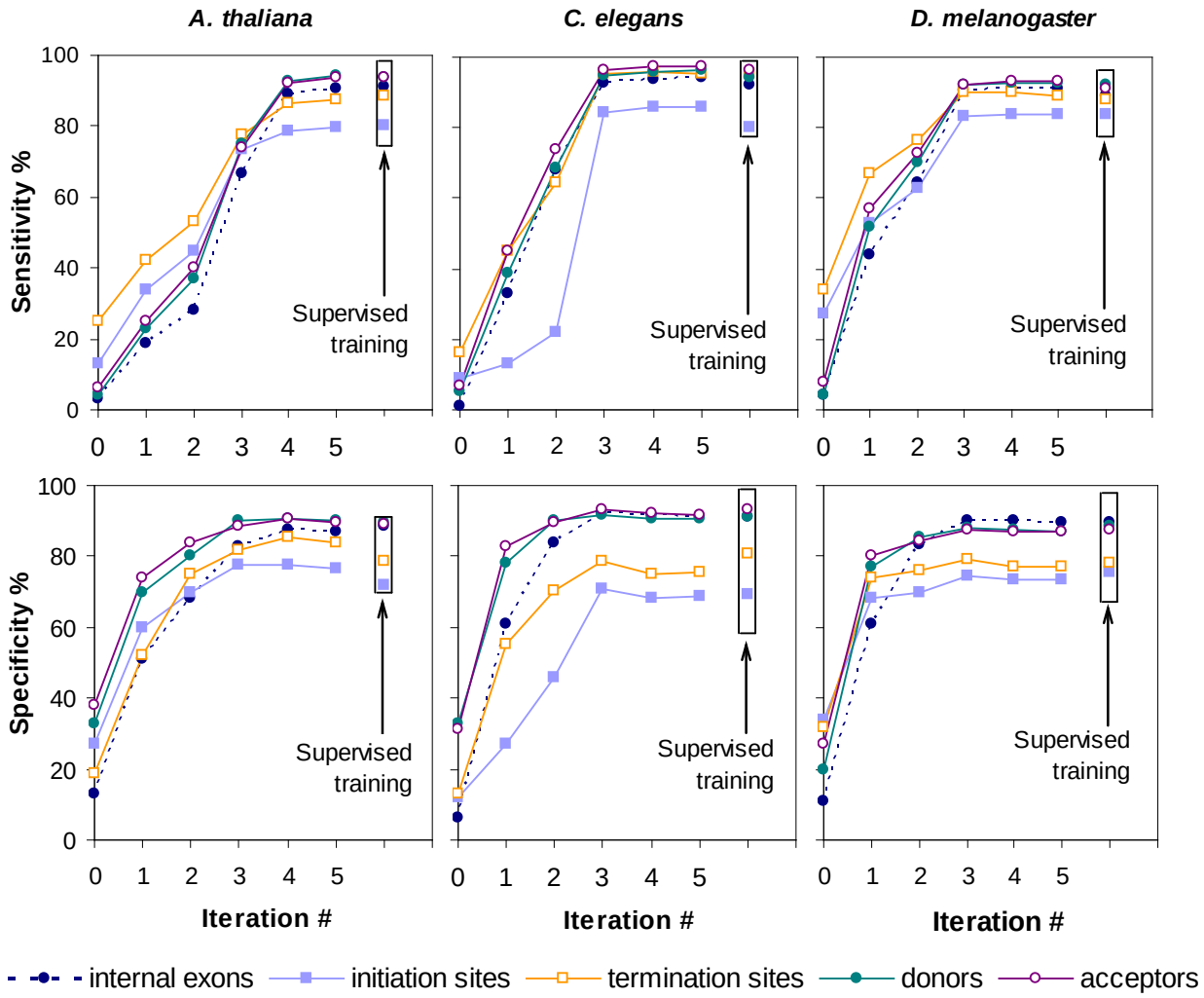
Observed frequencies of four nucleotides in the three codon positions (first- green, second- blue, third- red)
as functions of genome GC content for 319 bacterial genomes.

Besemer & Borodovsky, Nucl. Acids Res.1999
Zhu et al. Nucl. Acids Res. 2010

oligomer frequencies are predicted from sequence GC content



Zhu et al. Nucl. Acids Res. 2010



Gene prediction Sensitivity and Specificity as functions of iteration index.

Lomsadze et al. Nucl. Acids Res. 2005

Assessment of accuracy of gene finding tools on A. thaliana AraSet
a manually curated collection of 174 multi-gene sequences with 1028 exons

Program	Predicted exons	Correct exons	Overlap. exons	Wrong exons	Missing exons	Sensitivity	Specificity	Ratio WE
GENSCAN	938	652	204	82	175	0.63	0.70	0.09
GeneMark.hmm	1104	845	172	87	26	0.82	0.77	0.08
MZEF prior *p* = 0.01	641	401	153	87	480	0.39	0.63	0.14
MZEF prior *p* = 0.04	846	459	236	151	358	0.45	0.54	0.18
MZEF prior *p* = 0.10	998	490	298	210	283	0.48	0.49	0.21
FGENE	1061	569	300	192	213	0.55	0.54	0.18
GRAIL	1184	449	506	229	80	0.44	0.38	0.19
FEX	1745	562	484	699	155	0.55	0.32	0.40
FGENESP	737	433	195	109	403	0.42	0.59	0.15

Table reproduced from Pavy N.,Rombauts S., Dehais P., Mathe C., Ramana D.V.V., Leroy P. and Rouze P. (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics*, **15**, 887-899.

Sensitivity = number of correct exons / number of annotated exons
Specificity = number of correct exons / number of predicted exons
Ratio WE = wrong exons / predicted exons

Assessment of accuracy of gene finding tools on A. thaliana AraSet
a manually curated collection of 174 multi-gene sequences with 1028 exons

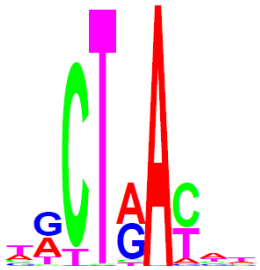
Program	Predicted exons	Correct exons	Overlap. exons	Wrong exons	Missing exons	Sensitivity	Specificity	Ratio WE
GENSCAN	938	652	204	82	175	0.63	0.70	0.09
GeneMark.hmm	1104	845	172	87	26	0.82	0.77	0.08
MZEF prior *p* = 0.01	641	401	153	87	480	0.39	0.63	0.14
MZEF prior *p* = 0.04	846	459	236	151	358	0.45	0.54	0.18
MZEF prior *p* = 0.10	998	490	298	210	283	0.48	0.49	0.21
FGENE	1061	569	300	192	213	0.55	0.54	0.18
GRAIL	1184	449	506	229	80	0.44	0.38	0.19
FEX	1745	562	484	699	155	0.55	0.32	0.40
FGENESP	737	433	195	109	403	0.42	0.59	0.15

Table reproduced from Pavy N.,Rombauts S., Dehais P., Mathe C., Ramana D.V.V., Leroy P. and Rouze P. (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics*, **15**, 887-899.

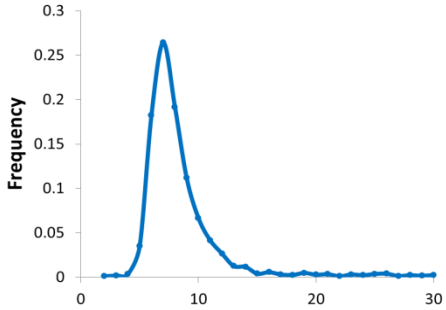
Sensitivity = number of correct exons / number of annotated exons
Specificity = number of correct exons / number of predicted exons
Ratio WE = wrong exons / predicted exons

Unusual structure of splice signals in fungal introns

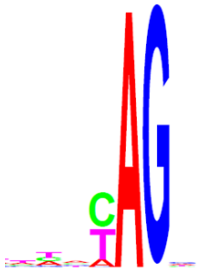
Fungi: strong and well localized branch point, weak acceptor signal



Branch point model logo



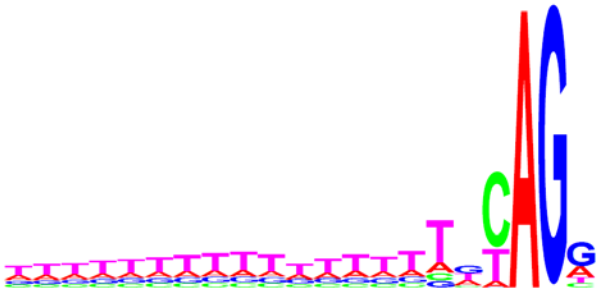
Distribution of distance between branch point and acceptor



Acceptor model logo

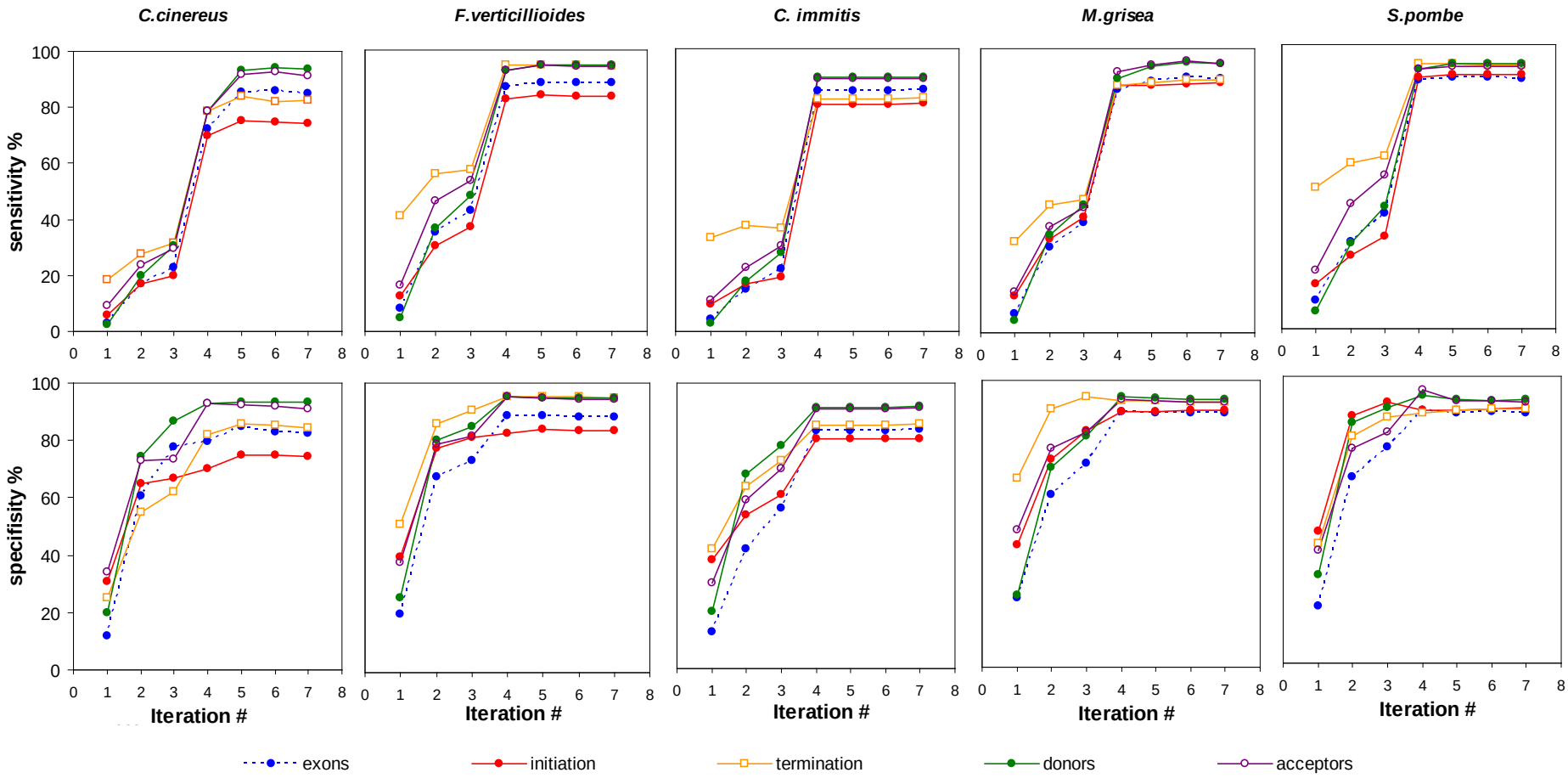
Rhynchosporium sp.

Plants and animals: weak and not localized branch point, strong acceptor signal



Polypyrimidine tract and acceptor model logo

A.thaliana



Change in accuracy of identification of elements of gene structure (Sn and Sp).

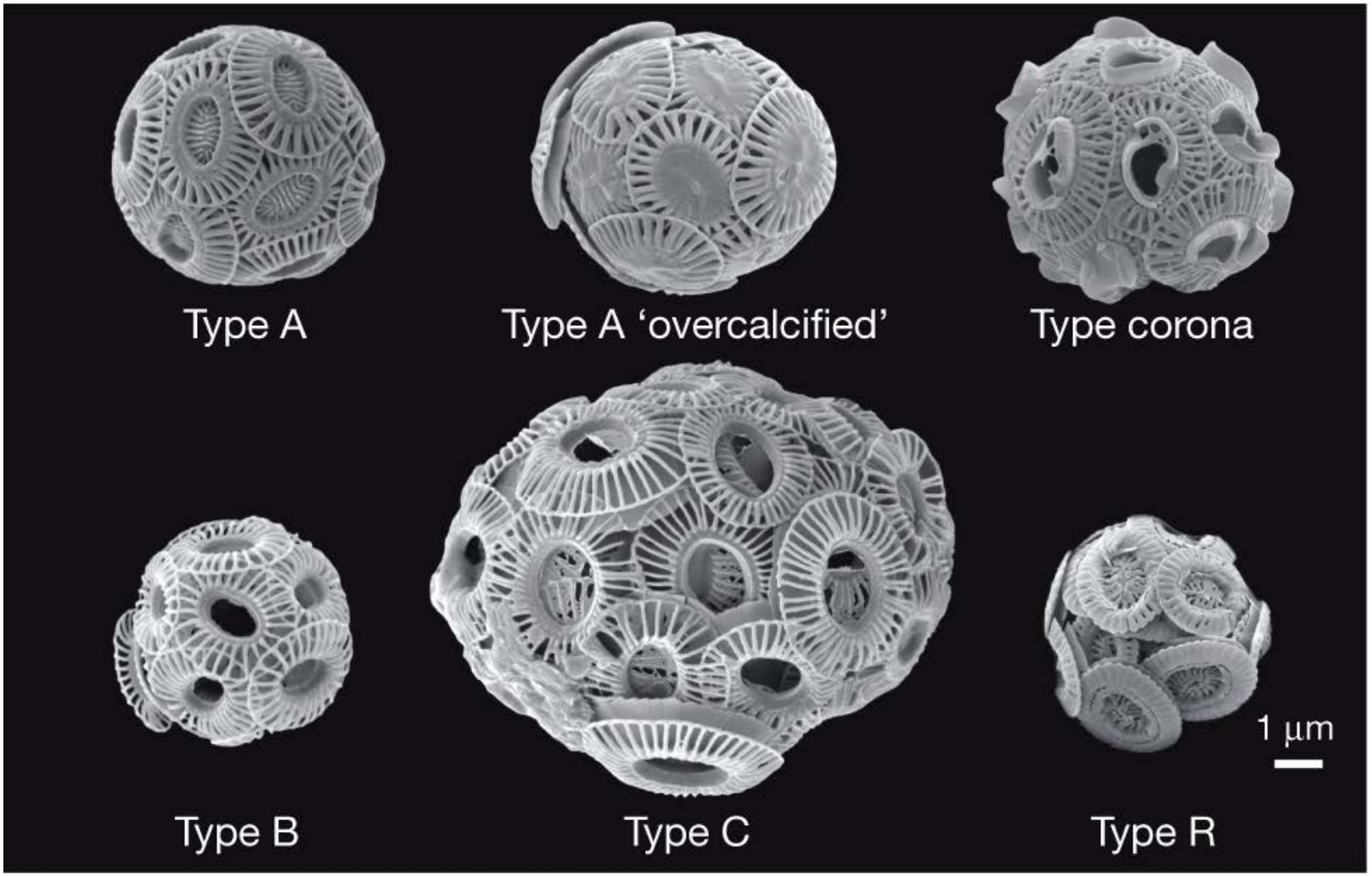
Ter-Hovhannisyan et al. Genome Research, 2008

Application of GeneMark-ES

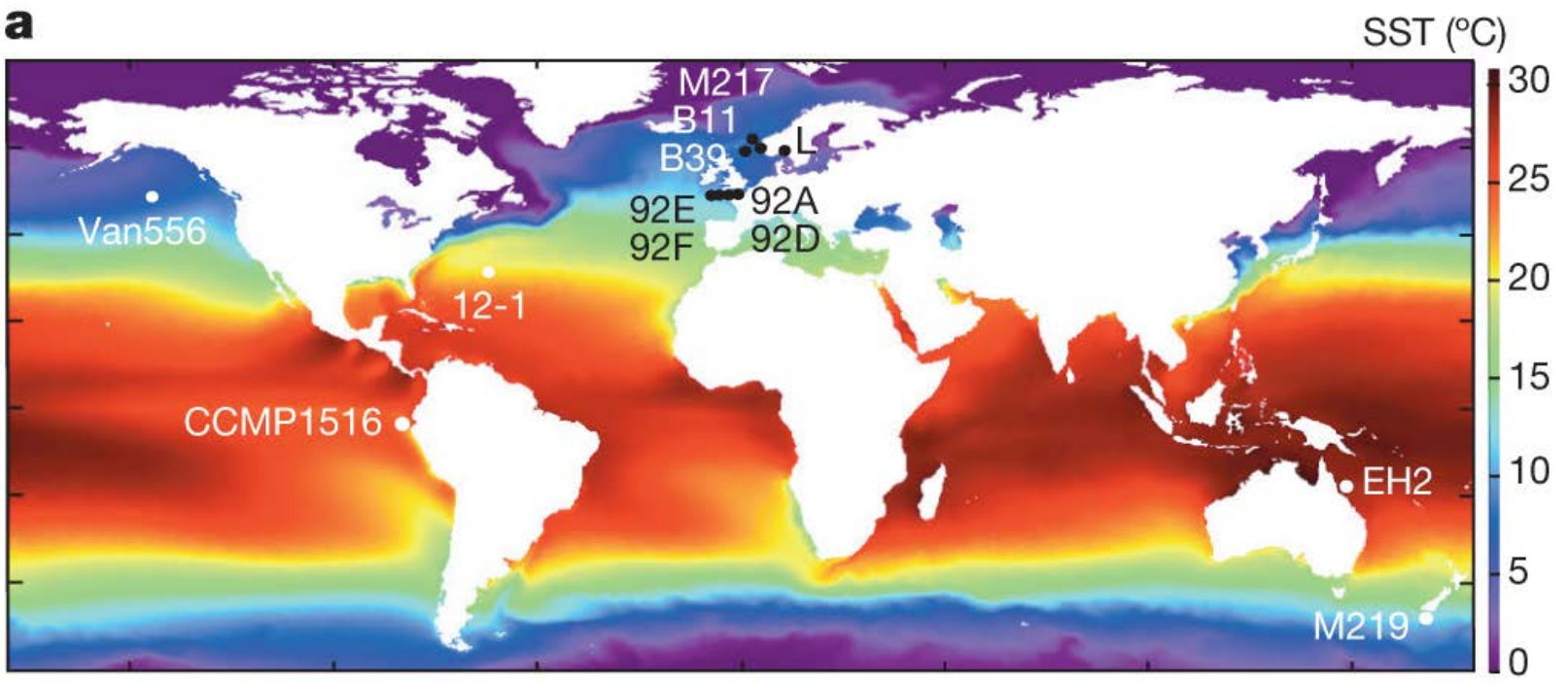
Cover image: Pictured are mature *Coprinopsis cinerea* mushrooms fruiting in a Petri dish. *Coprinopsis cinerea* is a model mushroom and important genetic system for studying development and meiosis.

Stajich et al. PNAS, 2010

Phytoplankton coccolithophore *Emiliania huxleyi*

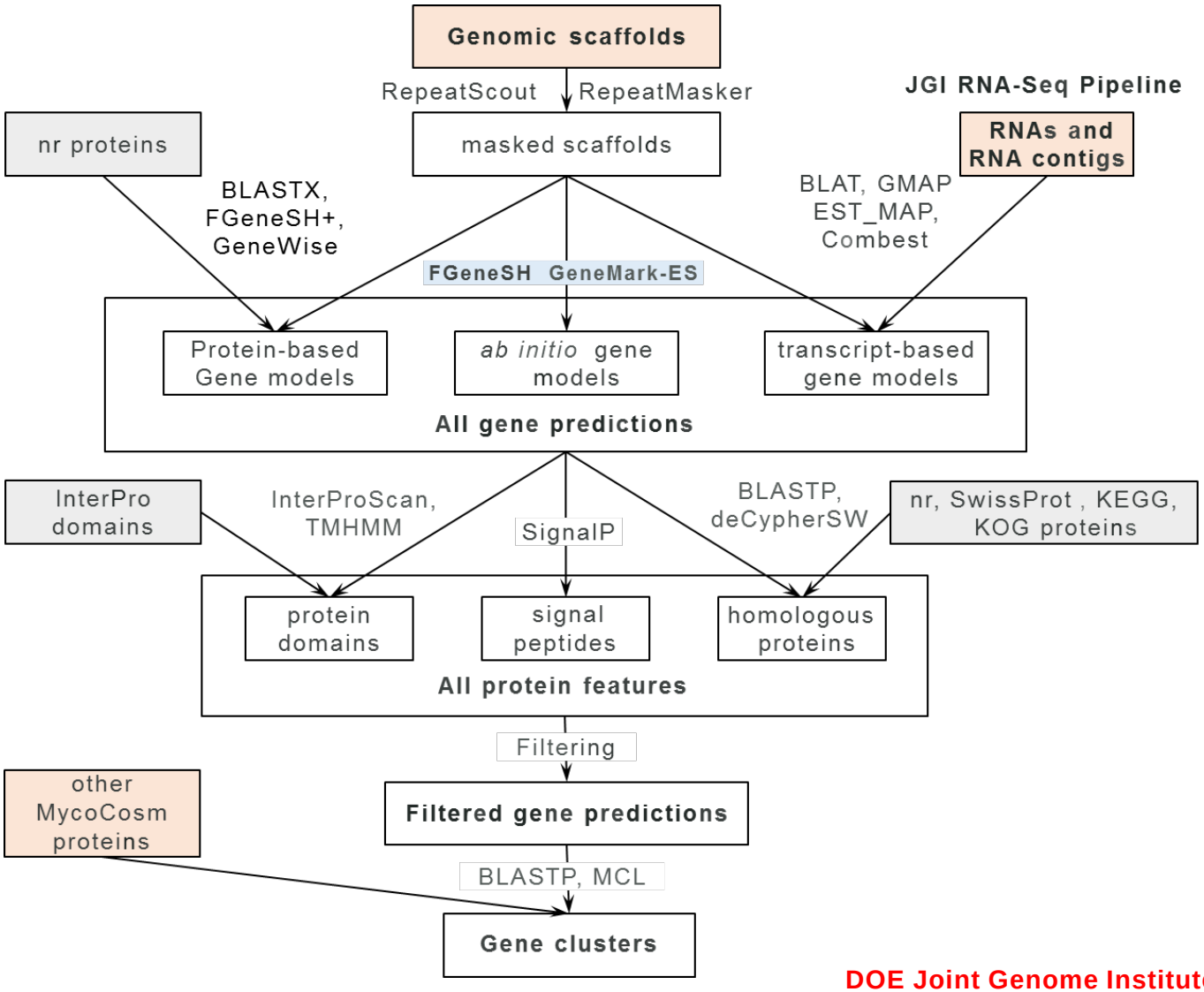


Phytoplankton coccolithophore *Emiliania huxleyi*

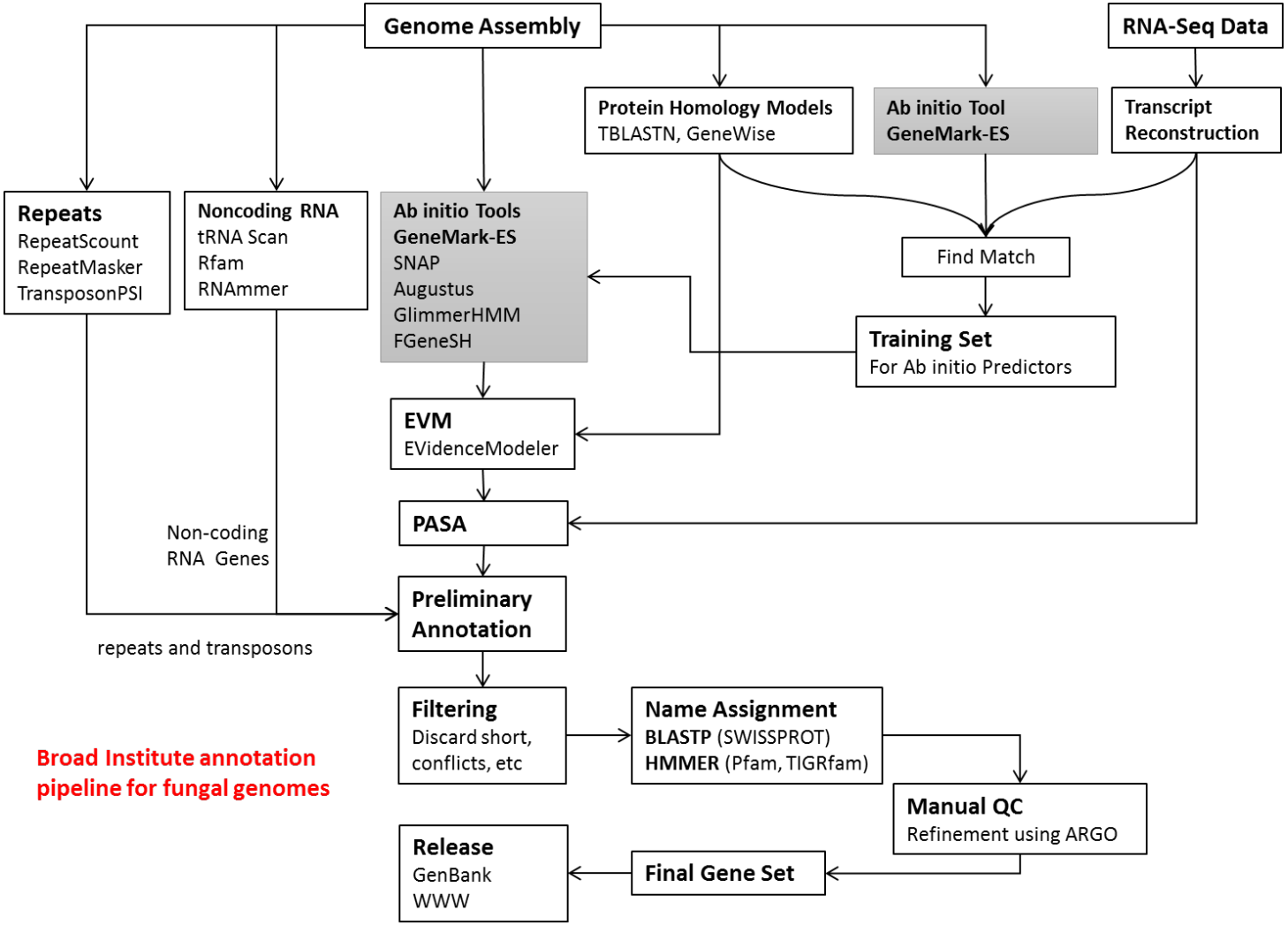


Isolation locations shown over the averaged Reynolds monthly sea-surface temperature (SST) climatology (1985–2007).

Read *et al. Nature*, (2013)



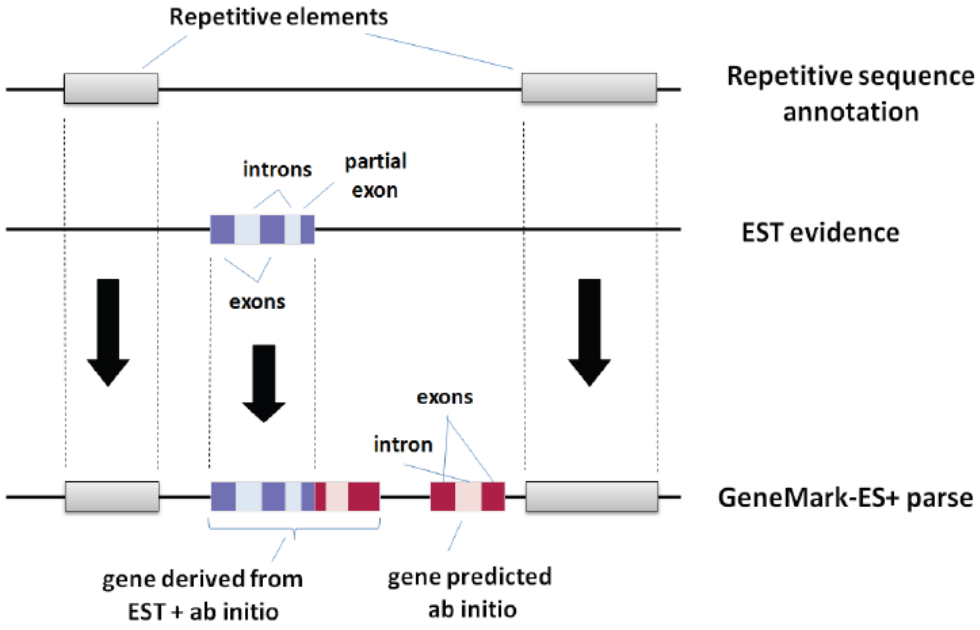
DOE Joint Genome Institute
Genome Annotation Pipeline



Broad Institute annotation pipeline for fungal genomes

Haas et al. Mycology, 2011

RNA-Seq/EST information can guide the segmentation of genomic sequence if integrated in the ab initio algorithm

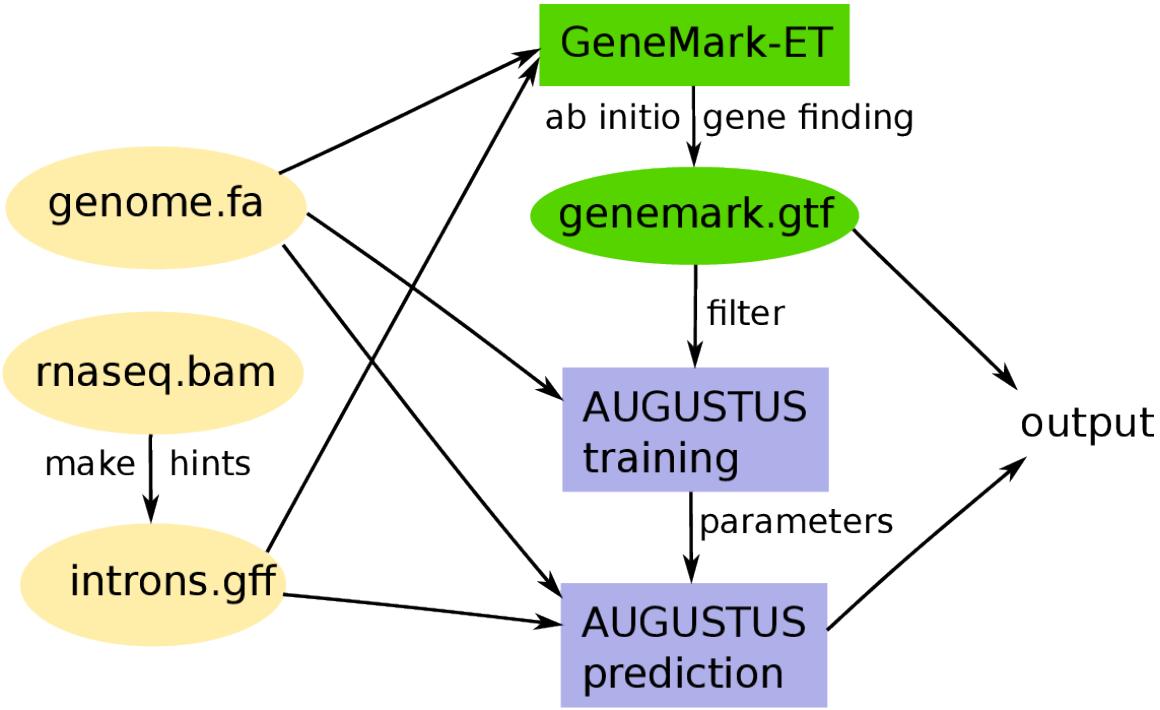


Conceptual scheme of integration of the ab initio and EST/TE in the GeneMark-ES+ algorithm

The genome of woodland strawberry (*Fragaria vesca)*
Shulaev et al. *Nature Genetics,* 2011

Potentilla migrantha

Flowchart of the BRAKER1 pipeline.



Benchmarking of BRAKER1 (continued)

	Arabidopsis thaliana		*Caenorhabditis elegans*		*Drosophila melanogaster*	
	MAKER2	**BRAKER1**	**MAKER2**	**BRAKER1**	**MAKER2**	**BRAKER1**
Exon sensitivity	76.1	**82.6**	69.4	**80.2**	64.9	**73.1**
Exon specificity	76.1	**79.0**	62.3	**85.5**	55.0	**67.0**

Full run of the pipeline includes run of AUGUSTUS with RNA-Seq derived hints

GeneMark

A family of gene prediction programs developed at
Georgia Institute of Technology , Atlanta, Georgia, USA.

What's New: A new algorithm, **BRAKER1**, an RNA-seq based eukaryotic genome annotation pipeline -- using GeneMark-ET and AUGUSTUS

Supported by NIH

Gene Prediction in Bacteria, Archaea, Metagenomes and Metatranscriptomes

Novel genomic sequences can be analyzed either by the self-training program **GeneMarkS** (sequences longer than 50 kb) or by **GeneMark.hmm with Heuristic models**. For many species pre-trained model parameters are ready and available through the **GeneMark.hmm** page. Metagenomic sequences can be analyzed by **MetaGeneMark** , the program optimized for speed.

Gene Prediction in Eukaryotes

Novel genomes can be analyzed by the program **GeneMark-ES** utilizing unsupervised training. Note that GeneMark-ES has a special mode for analyzing fungal genomes. Recently, we have developed a semi-supervised version of GeneMark-ES, called GeneMark-ET that uses RNA-Seq reads to improve training. For several species pre-trained model parameters are ready and available through the **GeneMark.hmm** page.

Gene Prediction in Transcripts

Sets of assembled eukaryotic transcripts can be analyzed by the modified **GeneMarkS** algorithm (the set should be large enough to permit self-training). A single transcript can be analyzed by a special version of **GeneMark.hmm with Heuristic models**. A new advanced algorithm GeneMarkS-T was developed recently (manuscript sent to publisher); The GeneMarkS-T software (beta version) is available for download .

Gene Prediction in Viruses, Phages and Plasmids

Sequences of viruses, phages or plasmids can be analyzed either by the **GeneMark.hmm with Heuristic models** (if the sequence is shorter than 50 kb) or by the self-training program **GeneMarkS**.

All the software programs mentioned here are available for download and local installation.

The software of GeneMark line is a part of genome annotation pipelines at NCBI, JGI, Broad Institute as well as the following software packages:

- QUAST : quality assessment tool for genome assemblies
  -- using GeneMarkS
- MetAMOS : a modular and open source metagenomic assembly and analysis
  -- using MetaGeneMark
- MAKER2 : a eukaryotic genome annotation pipeline
  -- using GeneMark-ES (along with SNAP and AUGUSTUS)
- BRAKER1 : an RNA-seq based eukaryotic genome annotation pipeline
  -- using GeneMark-ET and AUGUSTUS

For more information see Background and Publications.

Borodovsky Group
Group news

Gene Prediction Programs

- GeneMark
- GeneMark.hmm
- GeneMarkS
- Heuristic models
- MetaGeneMark
- Mirror site at NCBI
- GeneMarkS+
- BRAKER1

Information

- Publications
- Selected Citations
- Background
- FAQ
- Contact

Downloads

- Programs
- Prebuild species models

Other Programs

- UnSplicer
- GeneTack
- Frame-by-Frame
- IPSSP

In silico Biology International Conferences

- 2015
- 2013
- 2011
- 2009
- 2007
- 2005
- 2003
- 2001
- 1999
- 1997

Georgia Tech

Alexandre Lomsadze
Gena Tang
Karl Gemayel
Paul Burns
Wenhan Zhu
Vardges Ter-Hovhannesian
John Besemer

University of Greifswald, Germany

Katharina Hoff
Simona Lange
Mario Stanke

NCBI

Tatiana Tatusova
Mike DiCuccio
Azat Badretdin
Slava Chetvernin
James Ostell

DOE Joint Genome Institute

Asaf Salamov
Igor Grigoriev

Broad Institute

Christina Cuomo
Bruce Birren

National Institute of Health
US Centers for Disease Control
Georgia Tech
Moscow Institute for Physics and Technology