GENE DETECTION

Domen Lušina

Bioinformatics algorithms

2018/19

HEURISTIC APPROACH TO DERIVING MODELS FOR GENE FINDING

ABOUT THE ALGORITHM

- John Besemer and Mark Borodovsky, 1999
- Markov models
- Heauristics
- finding genes in prokaryotes, organelles, viruses, phages and plasmids
- sequence longer than 400 nt

MATERIALS

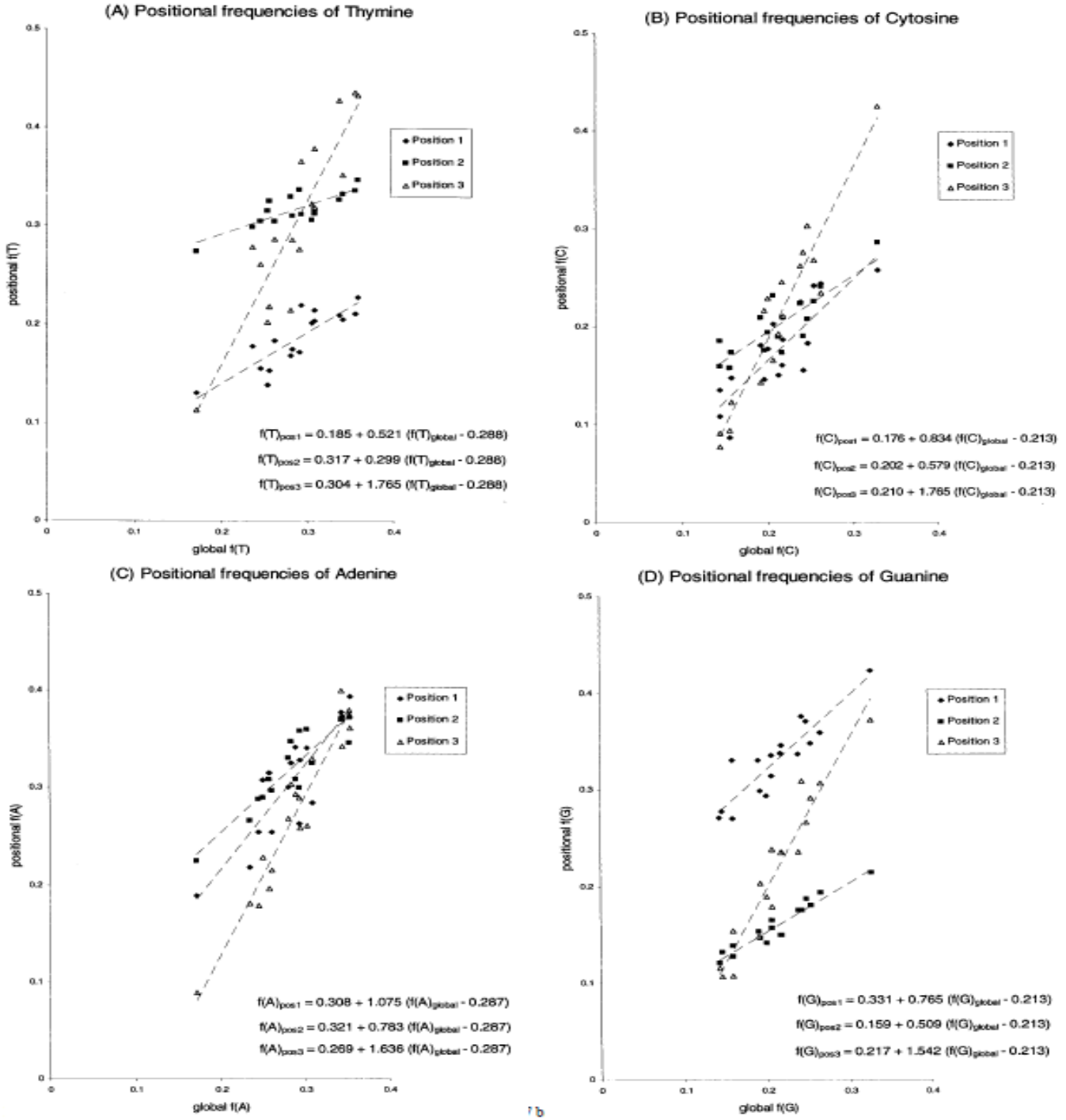
- 17 complete bacterial genomes
- 10 of these used for testing
- Observed in annotated sequences:
  - amino acid frequencies
  - positional nucleotide frequencies

HEURISTIC METHOD (1)

- relathionship between:
  - positional nt frequencies and global nucleotide frequncies
  - the amino acid frequencies and the global GC%

- linear regression

- „Z-pattern“
- difference in frequencies for 1st and 2nd codon positions

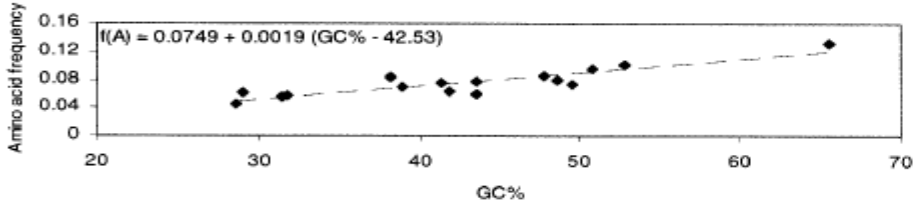


HEURISTIC METHOD (2)

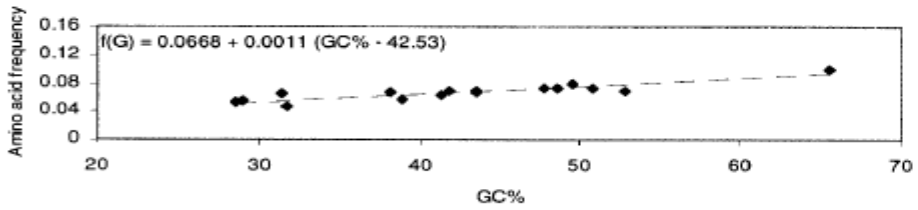
- frequency of 10 amino acids changes significantly (%GC)
- 4 SSN type codon (C or G) + valine
  - alanine
  - glycine
  - proline
  - arginine
- 5 WWN type codon (A or T)
  - phenylalanine
  - isoleucine
  - lysine
  - asparagine
  - tyrosine

1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA	Stop (*Ochre*) [B]	TGA	Stop (*Opal*) [B]	A
	TTG		TCG		TAG	Stop (*Amber*) [B]	TGG	(Trp/W) Tryptophan	G
C	CTT		CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G

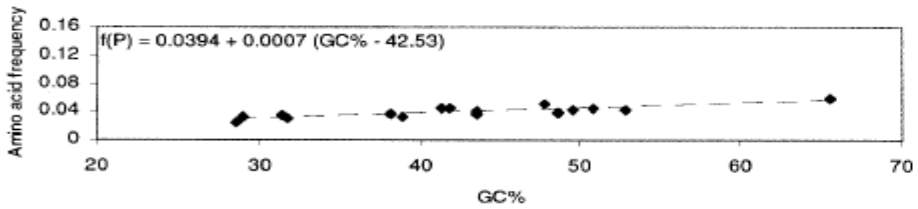
(A) Frequency of Alanine



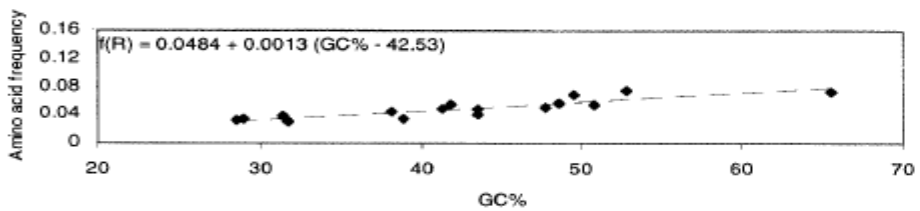
(B) Frequency of Glycine



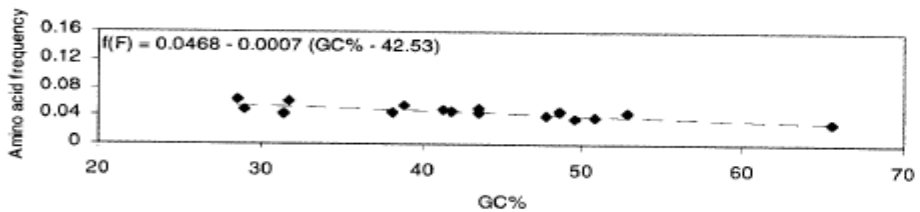
(C) Frequency of Proline



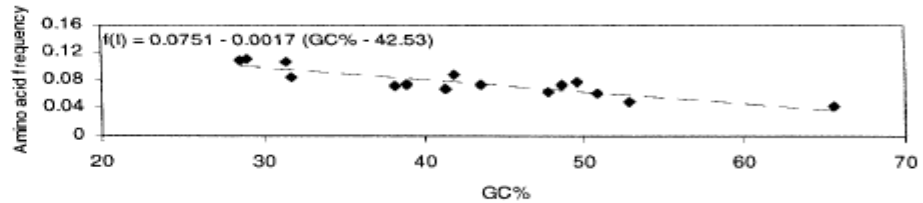
(D) Frequency of Arginine



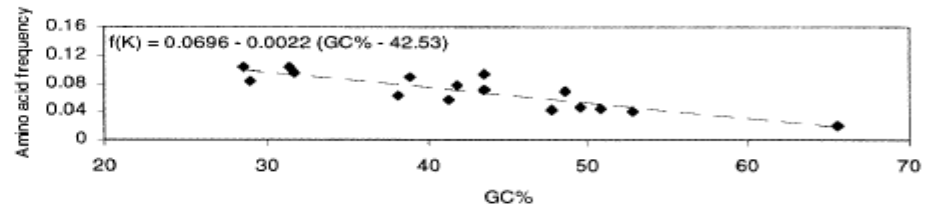
(E) Frequency of Phenylalanine



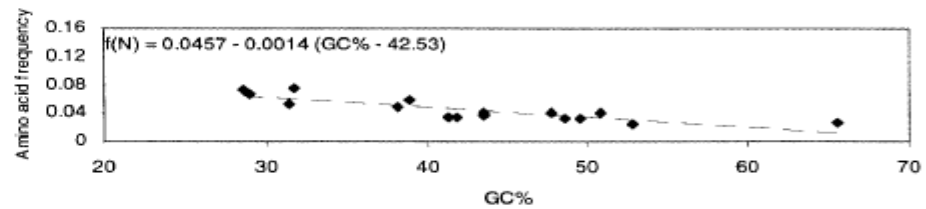
(F) Frequency of Isoleucine



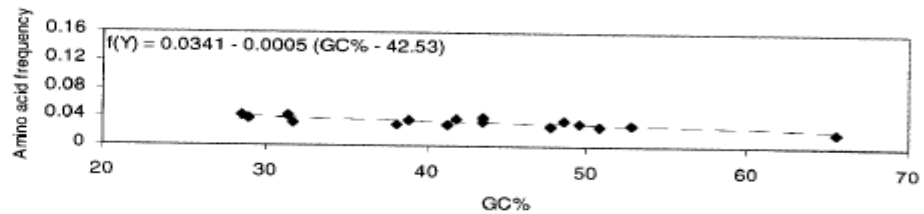
(G) Frequency of Lysine



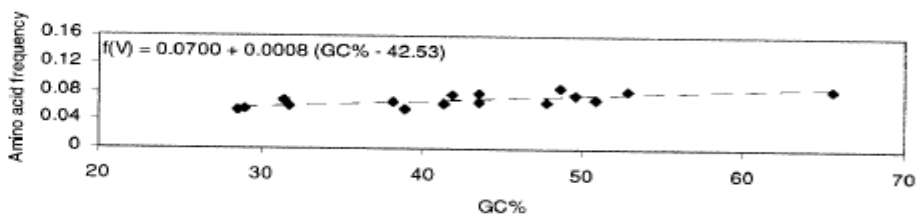
(H) Frequency of Asparagine



(I) Frequency of Tyrosine



(J) Frequency of Valine



HEURISTIC METHOD (3)

- set of Markov models
  - 3 periodic models for coding sequences (order zero, one and two)
  - one zero coding model for non-coding sequence

- from global nucleotide frequencies we determine nucleotide frequencies for each of three codon positions
- we compute initial frequency values for 61 codons $f_l(XYZ)$
- refine frequency
  - E. g. alanine codon CGT

$$f_R(GCT) = f_{alanine}(GC\%) \times [f_l(GCT) / (f_l(GCC) + f_l(GCA) + f_l(GCG) + f_l(GCT))]$$

- heuristically built codon usage table for the input genomic sequence

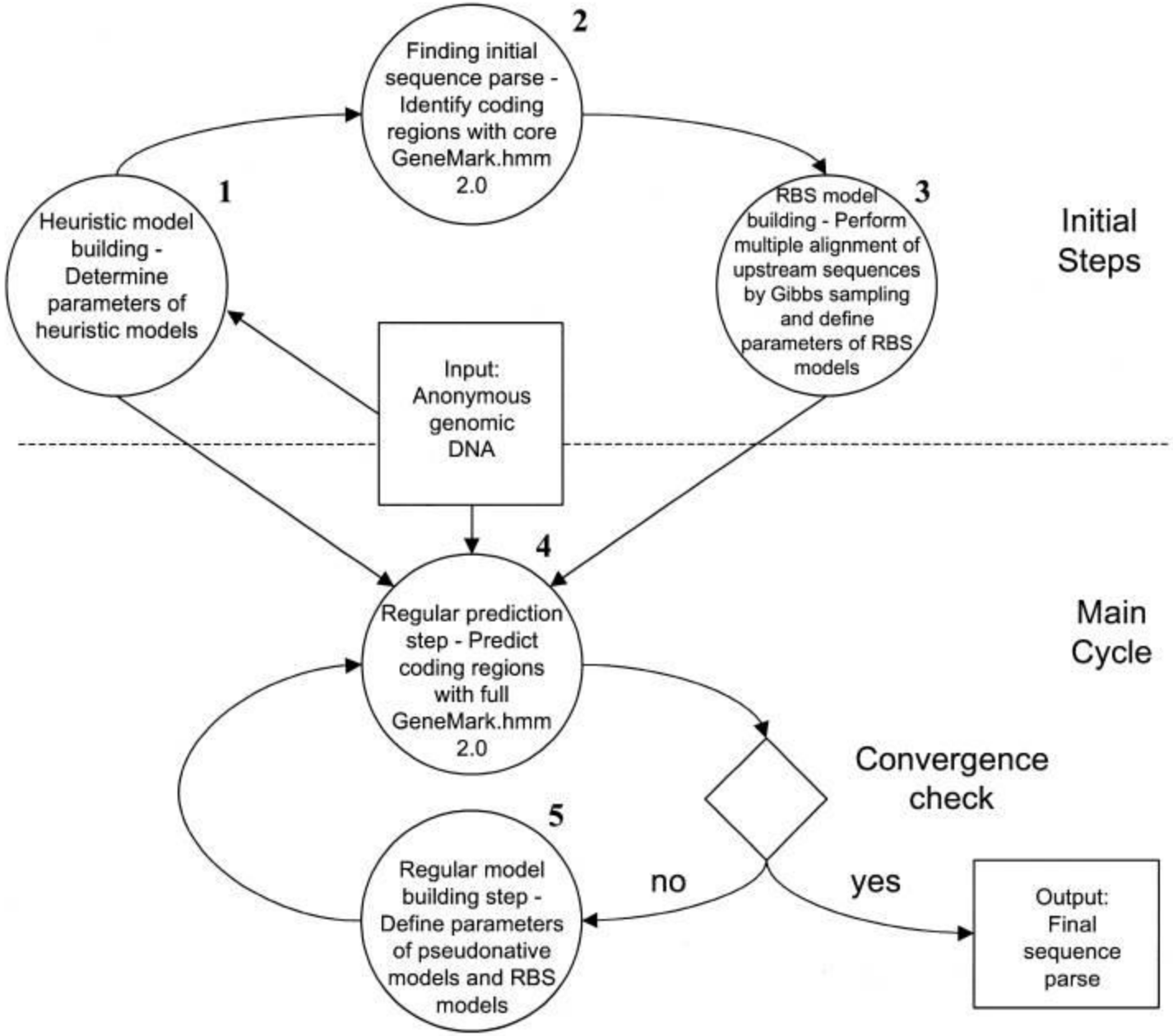
HEURISTIC METHOD (4)

- zero order Markov model
  - coding: use codon table
  - non-coding: global frequencies of nucleotides
- first order Markov model
  - ASSUMPTION: occurrences of adjacent codons are independent events
  - P(X→Y) for (..X|Y..) equal P(Y) in 1st position of codon defined by zero order MM
- second order Markov model
  - P(XY→Z) for (.XY|Z..) equal P(Z) in the 1st position of zero order MM
  - P(XY→Z) for (..X|YZ.) equal P(Y →Z), Z in 2nd position and Y in 1st position in first order MM

GENEMARKS

ABOUT THE ALGORITHM

- John Besemer, Alexandre Lomsadze and Mark Borodovsky, 2001
- Hidden Markov model based algorithm
- non-supervised training procedure
- uses gene finding program GeneMark.hmm
- Gibbs sampling multiple alignment program


2
Finding initial sequence parse - Identify coding regions with core GeneMark.hmm 2.0
1
Heuristic model building - Determine parameters of heuristic models
3
RBS model building - Perform multiple alignment of upstream sequences by Gibbs sampling and define parameters of RBS models
Initial Steps
Input: Anonymous genomic DNA
4
Regular prediction step - Predict coding regions with full GeneMark.hmm 2.0
Main Cycle
Convergence check
5
Regular model building step - Define parameters of pseudonative models and RBS models
no
yes
Output: Final sequence parse

RESULTS

	Heuristic HMM	Heuristic HMM (strict)	GenmarkS	GenmarkS (strict)
Azoarcus sp.	0.978	0.825	0.978	0.851
Bartonella tribocorum CIP	0.970	0.791	0.985	0.861
Bacillus subtilis	0.949	0.761	0.960	0.852
Campylobacter jejuni	0.984	0.886	0.991	0.908
Clostridium perfringens	0.980	0.782	0.961	0.782
E. Coli	0.937	0.708	0.947	0.723
Listeria	0.975	0.632	0.985	0.697
Salmonella enterica	0.921	0.697	0.935	0.721
Vibrio cholerae	0.974	0.769	0.992	0.883
Vibrio anguillarum	0.965	0.740	0.977	0.855
	0.963	0.758	0.971	0.813