CISC 889 Bioinformatics (Spring 2004)

Gene Finding

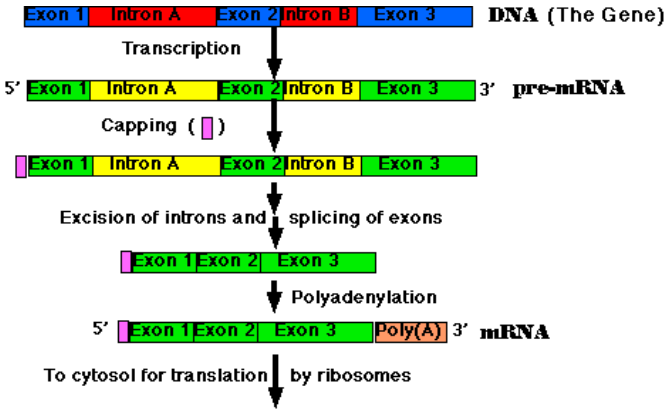
Gene prediction strategies

- Content-based
  - Codon usage
  - Periodicity of repeats
  - Compositional complexity
- Site-based
  - Binding sites for transcription factors
  - polyA tracts,
  - Donor and acceptor splice sites
  - Start and stop codons
- Comparative
  - BLAST

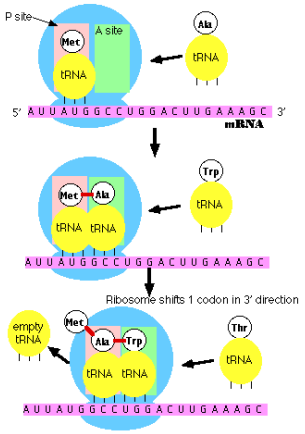
Gene expression

Transcription: DNA → mRNA

Translation: mRNA → Protein



Kimball's Biology page



Kimball's Biology page

Universal Genetic Code

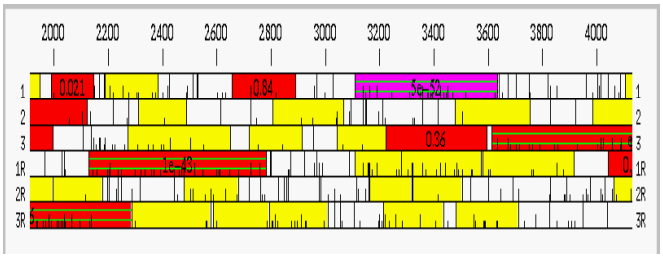
5'	2nd Position				3'
	T	C	A	G	
T	Phe	Ser	Tyr	Cys	T
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	T
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	T
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

ACCUUAGCGUA
Thr Leu Ala — Reading frame 1

ACCUUAGCGUA
Pro Stop Arg — Reading frame 2

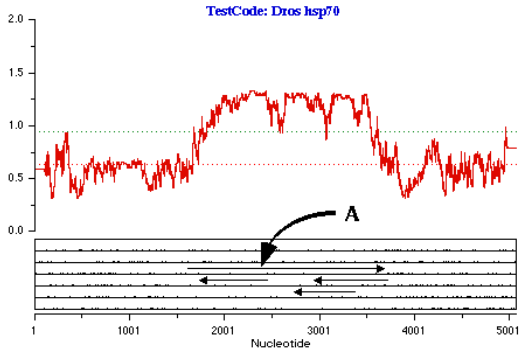
ACCUUAGCGUA
Leu Ser Val — Reading frame 3

Open Reading Frame (ORF)



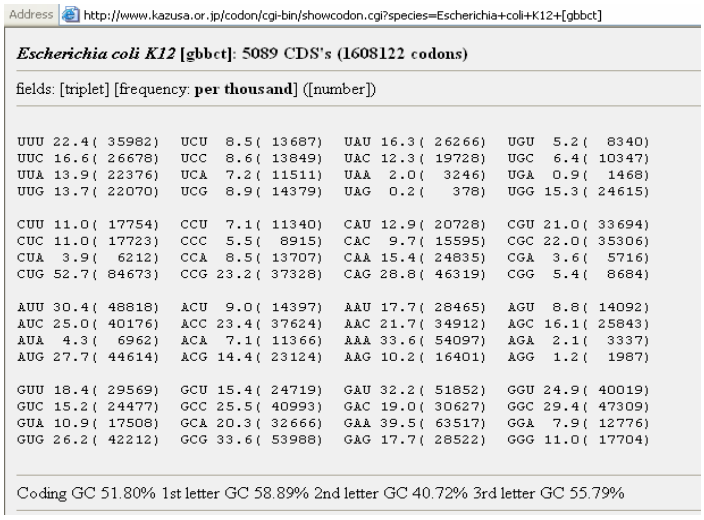
- Prokaryotic
  - Most regions of DNA are coding regions
  - No introns
- Eukaryotic
  - Introns and Exons

- Fickett's rule (1982)
  - In ORFs, every third base tends to be the same one more often than by chance alone.
    - Regardless species,
    - No knowledge of codon preference is required.

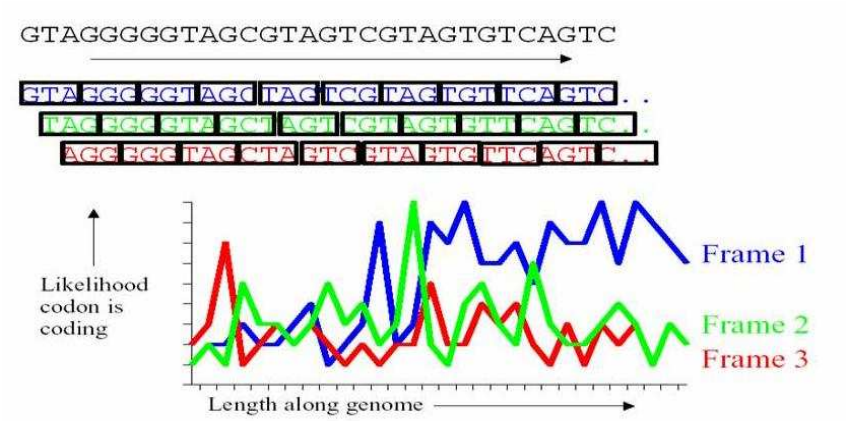


- Codon Usage Index
  - There are 64 codons but 20 amino acids to code, therefore some AAs are coded by multiple codons.
    - For example, 6 codons for Leu, and 4 for Ala, but only one for Try.
    - For random DNA sequences, the frequency of having these three AAs would be 6/4/1 for Lue:Ala:Trp. In real protein sequences, ratio was found to be 6.9/6.5/1, which implies coding DNA sequence is not random;
    - some codons are preferred (depending on species.)

Codon usage database: www.kazusa.or.jp/codon



Using codon usage for prediction ORFs



Can identify regions where there is a high frequency of 'preferred' codons

ORFs as Markov chains

- Glimmer: Interpolated Markov models (IMM)
  - 1st order model: p(a|a), p(a|c), …, p(t|t). Probability of having a amino acid given its previous neighbor.
  - 2nd order model: p(a|xx)
  - Up to 8th order model (0th for random, 1st to 6th for 6 reading frames, why higher order? Why stop at 8th ?
    - E.g, 5th order model, need 4^6 conditional probabilities p(a|xxxxx). In a genome of 1.8Mb, for each 6mer, we can observe about 1.8Mb/4096 samples. But the higher *k*, the less number of samples for *k*mers .
  - Interpolation (linear combination of models of different orders)

    $P(S|M) = \sum_{x=1}^{n} IMM_8(S_x)$

    where S_x is the oligomer ending at position x and n is the sequence length. The interpolated Markov model score is

    $IMM_k(S_x) = \lambda_k (S_{x-1}) P_k(S_x) + [1- \lambda_k (S_{x-1})] IMM_{k-1}(S_x)$

    where $\lambda_k (S_{x-1})$ is the numerical weight associated with the *k*mer ending at position x-1, and $P_k(S_x)$ is the probability of having S_x, predicted by k-th order model.

- Glimmer (cont'd)

  Results for H. *Influenzae*:

model	Found	Missed	New
Glimmer	1680	37	209
5^{th} order	1574	143	104

Self-identification (Audic and Claverie '98)

- Probability of sequence W of length L is generated by a k-th order Markov chain

  $P(W|M) = P(S_0) \prod_{i=k}^{L-1} P(n_i| S_{i-k})$ eq(1)

  where S_i is a *k*mer starting at position i in W. The model contains all the probabilities for any possible kmers to be followed by one of nucleotides A, C, G, or T. k = 5 is used.
- Which model is better?

  $P(M_j | W) = P(W|M_j)P(M_j) / \sum_{r= 1 \text{ to } N} P(W|M_r)P(M_r)$ eq(2)

  where *a priori* probability $P(M_j)$ is assumed to be equally probable for N models, i.e., is 1/N.

  If we have three models corresponding to coding, reverse coding, and non-coding, then the posterior probability tells what sequence W is more likely to be.

- Model building (no training data is required)
  - How to build the three models?

    If we have regions labeled as coding, reverse coding, and non-coding, then we can count the frequencies to train the transition matrices.
  - Self consistent
    - Randomly cut into non-overlapping pieces of w bases long, and assign them randomly into three distinct subsets, and build three Markov models M1, M2, and M3 respectively.
    - Scan genomic sequence using size w window. For each window segment, determine its class using eq(2). Slide the window by 5 bases, and repeat the process.
    - If a region is covered by n (for $5n \geq w$) successive windows of M_j type, it is qualified to be assigned into the j data set.
    - After finishing one scan of the genomic sequence, the 3 subsets are updated, and new Markov models are built for each of the 3 subsets.
    - Repeat the whole process until convergence is reached.

- Results (k = 5, w = 100)
  - Convergence is reached after 50 iterations
  - H. pylori
    - Correct rate: 95%, 94%, 93.8% for coding, reverse coding and non-coding respectively.

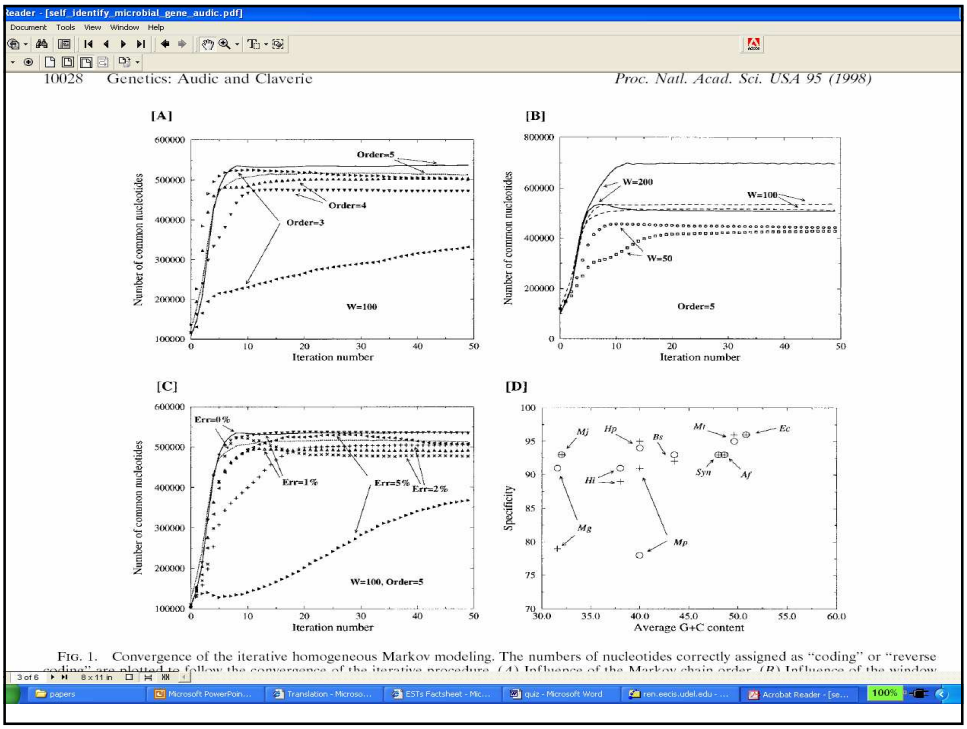


FIG. 1. Convergence of the iterative homogeneous Markov modeling. The numbers of nucleotides correctly assigned as "coding" or "reverse coding" are plotted to follow the convergence of the iterative procedure. (A) Influence of the Markov chain order. (B) Influence of the window

More markov based tools

- **HMMgene**
  http://www.cbs.dtu.dk/services/HMMgene/
  - Hidden Markov model
    - whole genes
    - partial genes
    - Cosmids or even longer sequences.
- GeneScan
- Genemark
  - http://opal.biology.gatech.edu/GeneMark/

- **blastx** compares a nucleotide query sequence translated in all reading frames against a protein sequence database
- **tblastn** compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
- **tblastx** compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Please note that tblastx program cannot be used with the nr database on the BLAST Web page.

Prediction Assessment

- (Baxevanis & Ouellette, 2nd ed, page 246)
  - TP (true positive), FP (false positive), TN (true negative), FN (false negative)
  - Actual positive, negative; Predicted positive, negative
  - TP + TN + FP + FN = N
  - Sensitivity = TP /(TP + FN)
  - Specificity = TP /(TP + FP)
  - Correlation Coefficient = TP• TN - FP• FN / √ (PP• PN • AP• AN)
  - CC = [-1, 1]

Strategies for Gene finding

- Challenges remain:
  - partial genes, non-coding RNA genes, etc.
- MZEF best for single exons
- GENSCAN best for whole genes
- shall try one more method for cross reference
- Shall resort to comparative method, e.g., run BLAST against dbEST and/or protein databases.