CMSC423: Bioinformatic Algorithms, Databases and Tools
Lecture 17

Gene finding

Signals in DNA

- we have the genome sequence... now what?
- ...see chapter 9 ...
- Motifs are a kind of “signal” - pattern of DNA that is “unexpected” in the genome of an organism
- Uncovering new motifs – already did this – Gibbs sampling (local multiple alignment).
- Given a motif – how do we find where it occurs in a genome?
- Remember? Motif=
  - k consecutive positions
  - frequency of occurrence of each base at these positions

Finding/scoring motifs

- Given motif M of length k – can be represented as a Position Weight Matrix (PWM) – same thing as a multiple alignment profile

$$pwm_M = \{ p_{c,i} | \forall (1 \leq i \leq k , c \in \sigma) \}$$

- Scoring a region of the genome according to motif? Given consecutive characters s_1,...,s_k

$$p(M | s_{1,} .., s_k) = \prod_{1 \leq i \leq k} p_{s_i , i}$$

- How surprising is this? Need to compare to background probabilities

$$p(M | s_{1,} .., s_k) = \prod_{1 \leq i \leq k} p_{s_i , i} / q_{s_i}$$

where q_{s_i} is background probability of character s_i in genome

Scoring motifs

- Note: Score usually presented as a log-likelihood $(\log(p(M|s_1...s_k))$
- The p/q ratios in the motif are often called Position Specific Scoring Matrix (PSSM)
- The program psi-blast can search a sequence against a database of PSSMs

- Motifs are just one piece of the puzzle
- How do we handle more complex “signals”

Gene finding/prediction

- Given a string of DNA, identify regions that might be genes
- Question: What does a gene look like?

- Start codon: ATG
- Stop codon: TGA, TAG, TAA
- Splicing: GT...intron...AG

- Also, DNA composition is different in genes – mutations are more likely in the third position of codons.

Simple gene finder (in bacteria)

- Find all stop-codons in the genome
- For each stop-codon, identify an in-frame start-codon upstream of it.
- Each section between a start and a stop is called an ORF – open reading frame.
- The long ORFs are likely genes – evolution prevented stop codons from occurring

- 3 stop codons, 64 possible codons => in random DNA every 22nd codon is a stop.

GGC **TAG** **ATG** AGG GCT CTA ACT **ATG** GGC GCG **TAA**

Gene finding as machine learning

- Main question: does the ORF look like a gene?

- Given a set of examples – genes we already know
- and a string of DNA (e.g. ORF)
- compute the likelihood that the ORF is a gene.
- Note: more complex than motif finding

- Codon usage bias – not all codons for a same amino-acid are equally likely
- K-mer (e.g. 6-mer) frequencies (instead of single-base frequencies in motif finding)

Bacillus anthracis codon usage

UUU	F	0.76	UCU	S	0.27	UAU	Y	0.77	UGU	C	0.73
UUC	F	0.24	UCC	S	0.08	UAC	Y	0.23	UGC	C	0.27
UUA	L	0.49	UCA	S	0.23	UAA	*	0.66	UGA	*	0.14
UUG	L	0.13	UCG	S	0.06	UAG	*	0.20	UGG	W	1.00
CUU	L	0.16	CCU	P	0.28	CAU	H	0.79	CGU	R	0.26
CUC	L	0.04	CCC	P	0.07	CAC	H	0.21	CGC	R	0.06
CUA	L	0.14	CCA	P	0.49	CAA	Q	0.78	CGA	R	0.16
CUG	L	0.05	CCG	P	0.16	CAG	Q	0.22	CGG	R	0.05
AUU	I	0.57	ACU	T	0.36	AAU	N	0.76	AGU	S	0.28
AUC	I	0.15	ACC	T	0.08	AAC	N	0.24	AGC	S	0.08
AUA	I	0.28	ACA	T	0.42	AAA	K	0.74	AGA	R	0.36
AUG	M	1.00	ACG	T	0.15	AAG	K	0.26	AGG	R	0.11
GUU	V	0.32	GCU	A	0.34	GAU	D	0.81	GGU	G	0.30
GUC	V	0.07	GCC	A	0.07	GAC	D	0.19	GGC	G	0.09
GUA	V	0.43	GCA	A	0.44	GAA	E	0.75	GGA	G	0.41
GUG	V	0.18	GCG	A	0.15	GAG	E	0.25	GGG	G	0.20

Questions

- Given the G/C content for a genome (fraction of letters in the genome that are G or C), what is the expected distance between two stop codons? - requires Poisson statistics