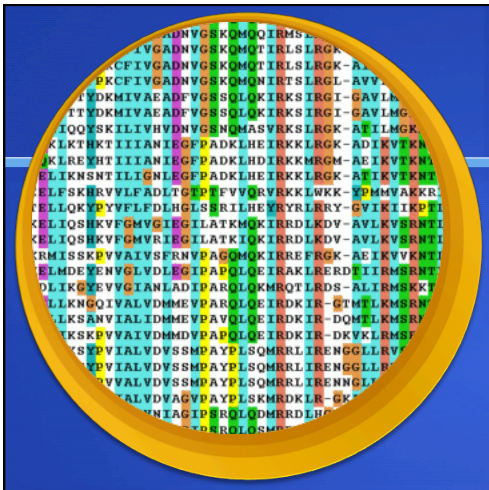
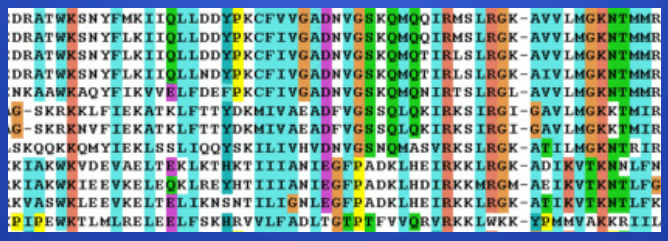


Examination of Pattern Discovery in Unaligned DNA Sequences

Sophie Saouma and Bryn Nisbet

http://en.wikipedia.org/wiki/Multiple_sequence_alignment

Motivation



http://www.psdgraphics.com/psd-icons/dna-strands-medical-icon-psd/

Problem: Motif Location

Given: Test packet containing 5 sequences

scxa_buteu	tatattgcggat
scx1_titse	ggctatccggtg
scx6_titse	tatccggcggat
scx1_cenno	tatctggtggat
six2_leiqu	ggctatattcgc

Goal: Find the motif locations

tatattgcggat
ggctatccggtg
tatccggcggat
tatctggtggat
ggctatattcgc

Return: Start position (a) of expected motif

0	1	2	3	4	5	6	7	8	9	10	11
G	G	C	T	A	T	C	C	G	G	T	G

a = 3

---tatattgcggat
ggctatccggtg---
---tatccggcggat
---tatctggtggat
ggctatattcgc---

Gibbs Sampling

Initialize with random alignment points.

While not converged:

Do steps 1 and 2 for each of N sequences :

1) Predictive update step : Calculate motif matrix and background using all sequences but the currently selected one.

2) Sampling step. Calculate $L_x = Q_x / P_x$ for all starting points x in this sequence. Choose one with probability proportional to L_x.

Results

Test Packet:

```
aacgcgtattgc------
---gcgtattgcgat---
gcgggctattgc------
------tattgcaaccgc
---gaaggctgcaac---
```

x100 runs per test packet

Packet Label	% Accuracy
A	72
B	68
C	69
D	70
E	69
F	71
G	62
H	73
I	67
J	70

$\bar{x} = 69.1\% \pm 3.1\%$