CMSC423

Chapter 2 – Motif finding/randomized algorithms

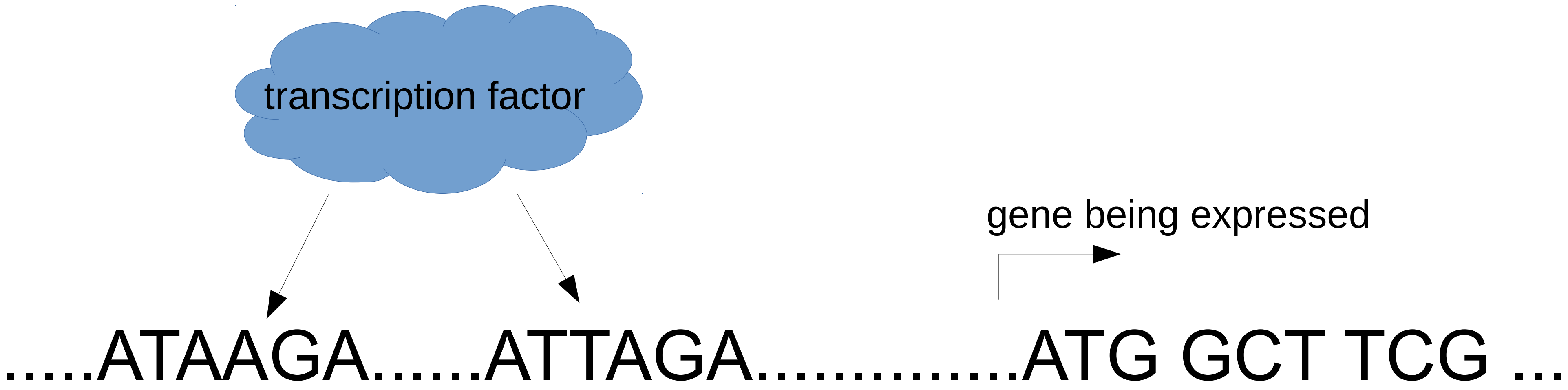
Recap

- Chapter 1 – look for "interesting" regions in a genome (regions where some patterns are frequent)

- Z+KMP – look for a specific pattern in a genome

This week

- Find something that's common to many pieces of DNA



transcription factor

gene being expressed

.....ATAAGA......ATTAGA.............ATG GCT TCG ...

Questions

- Group exercise in ELMS

Why earlier approaches don't work

- We don't know the motif (KMP and Z don't work)

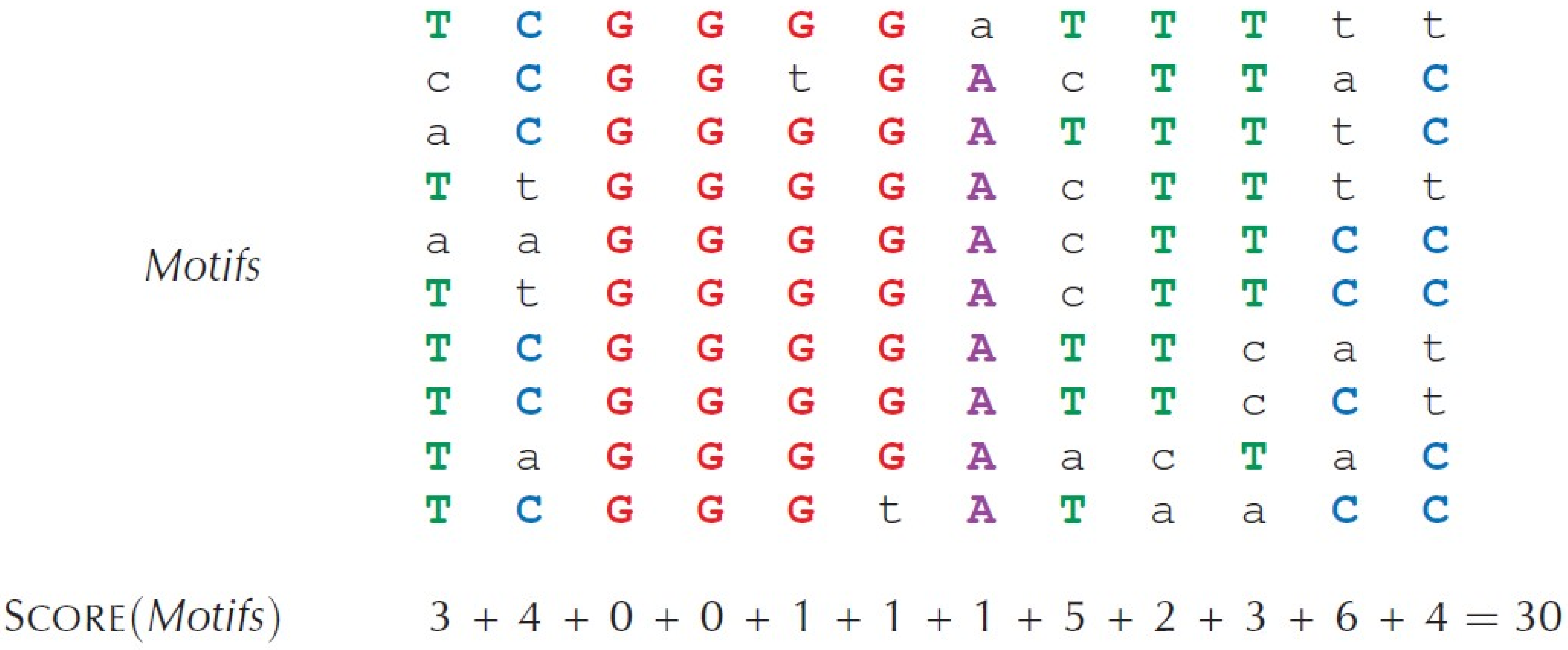
- Motifs are too inexact (frequency doesn't work)

The solution

- Randomized search:
  - take a random string from each upstream region
  - check the score of the profile
  - repeat until we find highest scoring profile

- What we need:
  - define the profile
  - define the score
  - come up with a search strategy

Option 1: count minority bases



Motifs	T	C	G	G	G	G	a	T	T	T	t	t
	c	C	G	G	t	G	A	c	T	T	a	C
	a	C	G	G	G	G	A	T	T	T	t	C
	T	t	G	G	G	G	A	c	T	T	t	t
	a	a	G	G	G	G	A	c	T	T	C	C
	T	t	G	G	G	G	A	c	T	T	C	C
	T	C	G	G	G	G	A	T	T	c	a	t
	T	C	G	G	G	G	A	T	T	c	C	t
	T	a	G	G	G	G	A	a	c	T	a	C
	T	C	G	G	G	t	A	T	a	a	C	C

SCORE(*Motifs*) $3 + 4 + 0 + 0 + 1 + 1 + 1 + 5 + 2 + 3 + 6 + 4 = 30$

Option 2: compute entropy

$$H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{N} p_i \log_2(p_i)$$

Motifs

T	C	G	G	G	G	a	T	T	T	t	t
c	C	G	G	t	G	A	c	T	T	a	C
a	C	G	G	G	G	A	T	T	T	t	C
T	t	G	G	G	G	A	c	T	T	t	t
a	a	G	G	G	G	A	c	T	T	C	C
T	t	G	G	G	G	A	c	T	T	C	C
T	C	G	G	G	G	A	T	T	c	a	t
T	C	G	G	G	G	A	T	T	c	C	t
T	a	G	G	G	G	A	a	c	T	a	C
T	C	G	G	G	t	A	T	a	a	C	C

PROFILE(*Motifs*)

A:	.2	.2	0	0	0	0	.9	.1	.1	.1	.3	0
C:	.1	.6	0	0	0	0	0	.4	.1	.2	.4	.6
G:	0	0	1	1	.9	.9	.1	0	0	0	0	0
T:	.7	.2	0	0	.1	.1	0	.5	.8	.7	.3	.4

Searching for motifs...deterministic

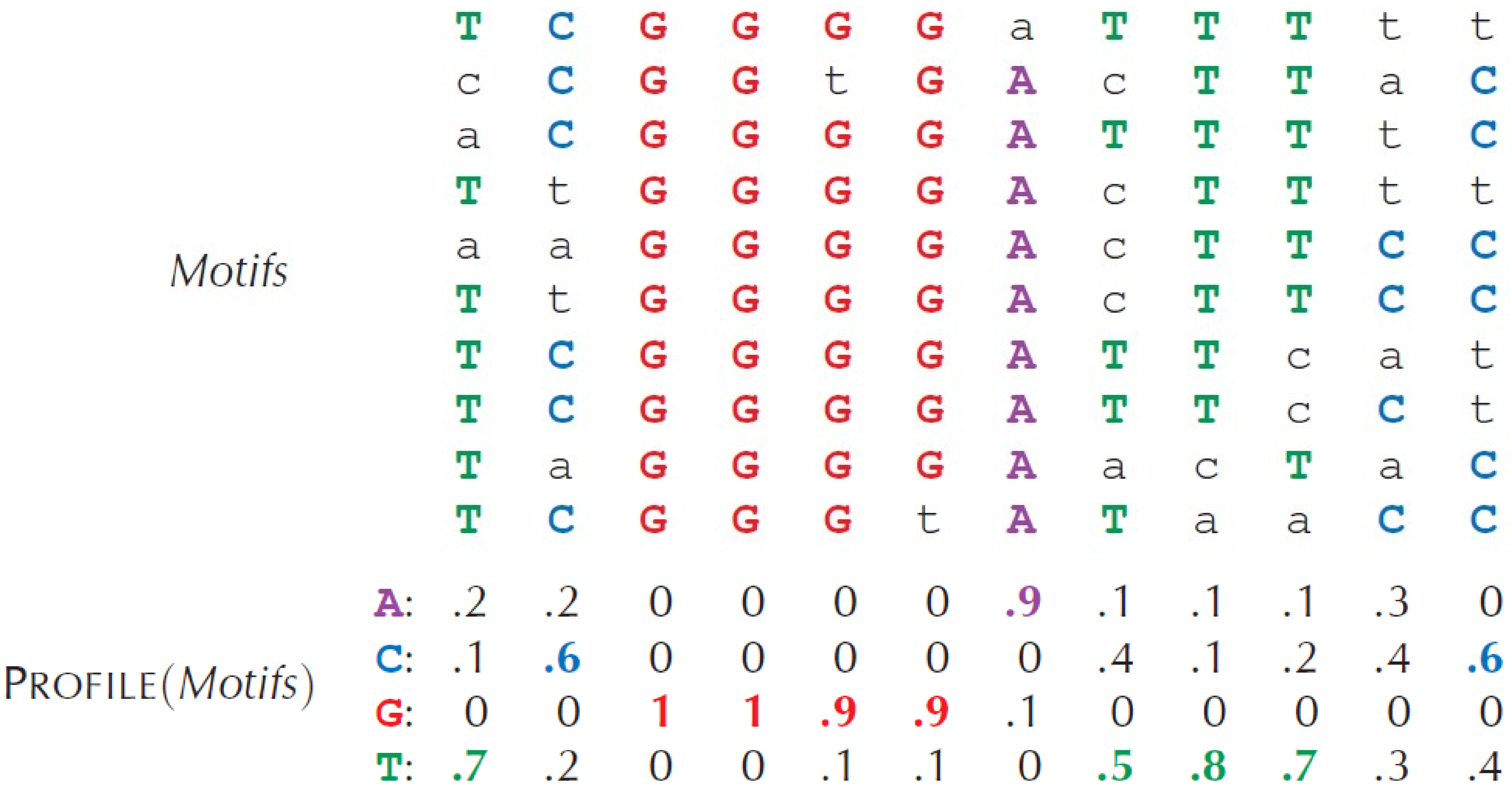
- Brute force – try all k-mers from all t strings

  runtime?

- Try all possible k-mers
  - from each string pick the one that is most similar to it

Searching for motifs...probabilistic

- Find the k-mer in each string that most probably fits the motif matrix/profile



Motifs	T	C	G	G	G	G	a	T	T	T	t	t
	c	C	G	G	t	G	A	c	T	T	a	C
	a	C	G	G	G	G	A	T	T	T	t	C
	T	t	G	G	G	G	A	c	T	T	t	t
	a	a	G	G	G	G	A	c	T	T	C	C
	T	t	G	G	G	G	A	c	T	T	C	C
	T	C	G	G	G	G	A	T	T	c	a	t
	T	C	G	G	G	G	A	T	T	c	C	t
	T	a	G	G	G	G	A	a	c	T	a	C
	T	C	G	G	G	t	A	T	a	a	C	C

PROFILE(*Motifs*)	A:	.2	.2	0	0	0	0	.9	.1	.1	.1	.3	0
	C:	.1	.6	0	0	0	0	0	.4	.1	.2	.4	.6
	G:	0	0	1	1	.9	.9	.1	0	0	0	0	0
	T:	.7	.2	0	0	.1	.1	0	.5	.8	.7	.3	.4

Probability-driven deterministic search

Profile

A:	.2	.2	.0	.0	.0	.0	.9	.1	.1	.1	.3	.0
C:	.1	.6	.0	.0	.0	.0	.0	.4	.1	.2	.4	.6
G:	.0	.0	1	1	.9	.9	.1	.0	.0	.0	.0	.0
T:	.7	.2	.0	.0	.1	.1	.0	.5	.8	.7	.3	.4

$$\Pr(\mathtt{ACGGGGATTACC}|\mathit{Profile}) = .2\cdot.6\cdot1\cdot1\cdot.9\cdot.9\cdot.9\cdot.5\cdot.8\cdot.1\cdot.4\cdot.6 = 0.000839808$$

Motifs

T	A	A	C
G	T	C	T
A	C	T	A
A	G	G	T

COUNT(*Motifs*)

A:	2	1	1	1
C:	0	1	1	1
G:	1	1	1	0
T:	1	1	1	2

PROFILE(*Motifs*)

2/4	1/4	1/4	1/4
0	1/4	1/4	1/4
1/4	1/4	1/4	0
1/4	1/4	1/4	2/4

Laplace's rule (avoids 0s with pseudocounts)

COUNT(*Motifs*)

A:	2+1	1+1	1+1	1+1
C:	0+1	1+1	1+1	1+1
G:	1+1	1+1	1+1	0+1
T:	1+1	1+1	1+1	2+1

PROFILE(*Motifs*)

3/8	2/8	2/8	2/8
1/8	2/8	2/8	2/8
2/8	2/8	2/8	1/8
2/8	2/8	2/8	3/8

Randomized search

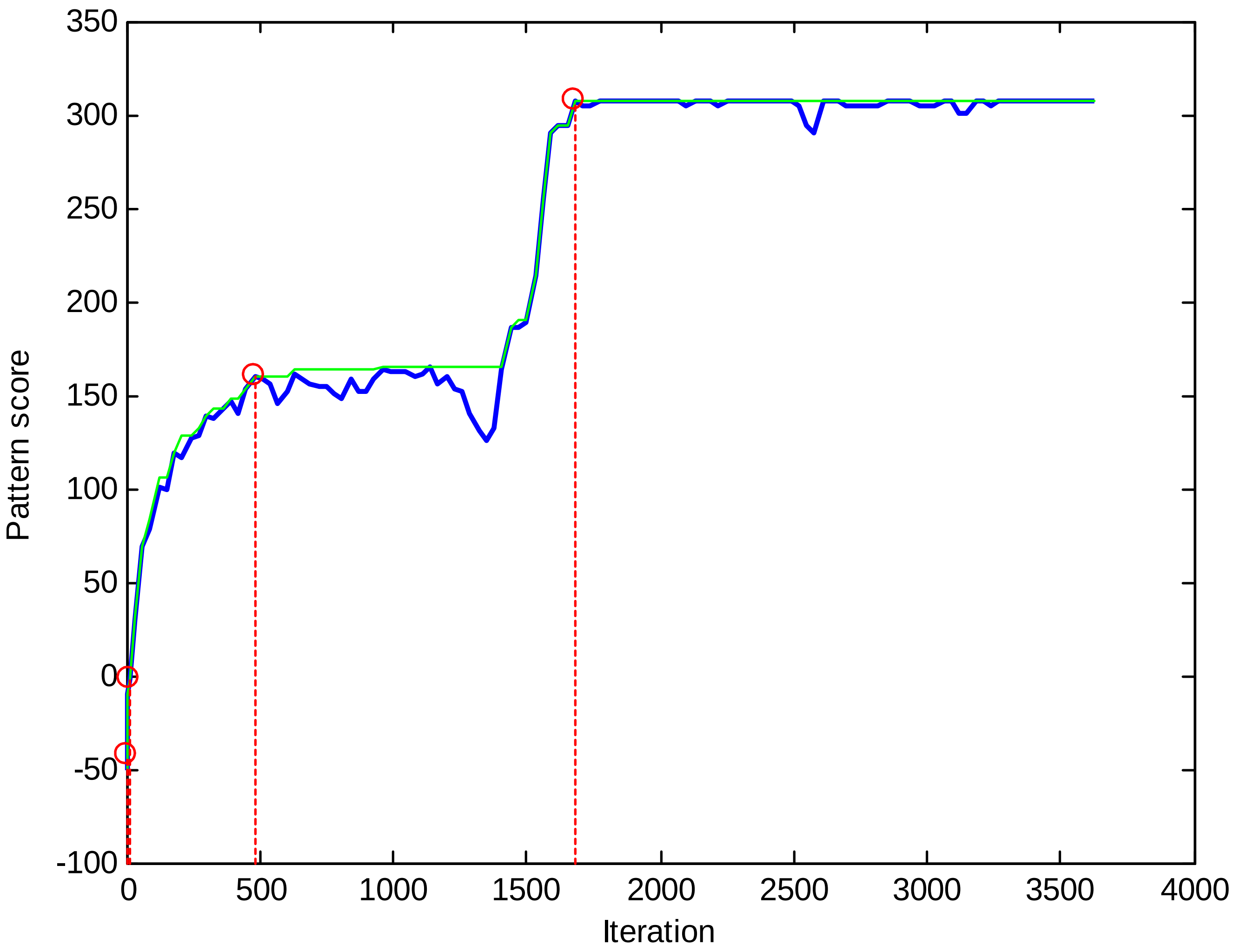
- We know:
  - how to create a profile of k-mers from each of the strings
  - calculate the score of the profile
  - find a string that most probably matches the profile
- Randomize motif search:
  - pick random starting k-mers
  - replace them with most probable k-mers from each string
  - repeat until best score found

Gibbs sampler

- Careful randomized search
  - pick random starting point
  - select ONE string
  - replace its k-mer with another random k-mer using a weighted die (biased towards the more probable k-mers)
  - repeat while score improves

Actual behavior of a Gibbs sampler

Behavior of the Objective Function



Input
30 sequences
k = 15

Search space
$\sim 10^{68}$

Time
< 1.0 seconds

The Evolving Multiple Alignment

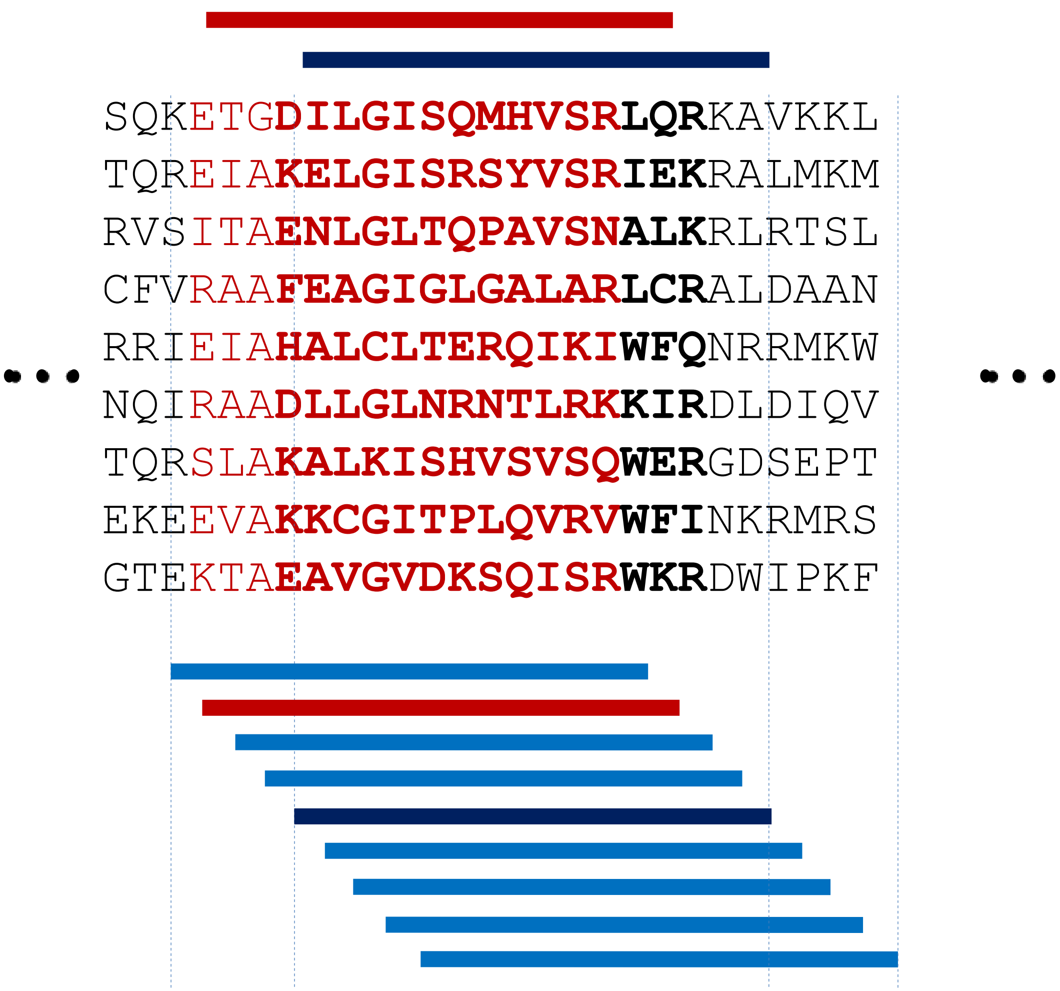
MTQPSKTTKLTKDEV	SKTTKLTKDEVDRLI	GISQMHVSRLQRKAV	ETGDILGISQMHVSR
MPPLFVMNNEILMHL	FVMNNEILMHLRALK	GISRSYVSRIEKRAL	EIAKELGISRSYVSR
VVFNQLLVDRRVSIT	QLLVDRRVSITAENL	TVRDSSMSLMQALQN	ITAENLGLTQPAVSN
WFQNRRMKWKKENKT	RYLTRRRRIEIAHAL	GVPQQQQQQQQQPSQ	EIAHALCLTERQIKI
SGTGKELVARALHDY	KELVARALHDYGRRR	KLDAQALERLKQHRW	RAADLLGLNRNTLRK
RIRYRRKNLKHTQRS	RRKNLKHTQRSLAKA	PESEQDTQLAEMRAR	SLAKALKISHVSVSQ
ALDAGVSVHIVRDYL	GVSVHIVRDYLLRGL	VLRQFVERRREALAN	RAAFEAGIGLGALAR
QLNGQDVNDLYELVL	QDVNDLYELVLAEVE	PLRDSVKQALKNYFA	RAALMMGINRGTLRK
LEIYHHIKKEKSPKG	HHIKKEKSPKGKSSI	FIMESNLTKVEQHTL	EVAKKCGITPLQVRV
SQISRWKRDWIPKFS	RWKRDWIPKFSMLLA	GVDKSQISRWKRDWI	KTAEAVGVDKSQISR
GSVAVLIKDEEGKEM	VLIKDEEGKEMILSY	RIAQTLLNLAKQPDA	EIGQIVGCSRETVGR
TINADGSVYAEEVKP	QTKTAKDLGVYQSAI	GVYQSAINKAIHAGR	KTAKDLGVYQSAINK
EIVTAGALKYQENAY	AGALKYQENAYRQAA	GISDAAVSQWKEVIP	AVAKALGISDAAVSQ
QLLLRRMEAINESLH	RRMEAINESLHPPMD	LLEQLLLRRMEAINE	SVAQHVCLSPSRLSH
DLSGKMPNLRQQMMR	QDMILLLSKKNAEER	NLRQQMMRLMSGEIK	DIGNYLGLTVETISR
GGLDSYIRAANAWPM	SYIRAANAWPMLSAD	RVRQLEKNAMKKLRA	ELADRYGVSAERVRQ
TRLAWPGNVRQLENT	RLARHFLQIAARELG	MLPDSWATLLGQWAD	EAARLLGWGRNTLTR
ETAATMKDVALKAKV	TMKDVALKAKVSTAT	KVSQATRNRVEKAAR	DVALKAKVSTATVSR
PRSASHYLLSDQKSR	LVEEKRRAAKLAATL	LLSDQKSRLVEEKRR	DAAALLGVSEMTIRR
YHNEQKERQAIEQLI	QKERQAIEQLIRHRC	KERQAIEQLIRHRCA	DVARLAGVSVATVSR
RLLQLSQGQAVKGNQ	AMLVANDQMALGAMR	ALADSLMQLARQVSR	DVAEYAGVSYQTVSR
TRPTEKQYETLENQL	KNKRALLDALAIEML	VLEDQEHQVAKEERE	KLAQKLGVEQPTLYW
SNSLKAAPVELRQWL	KAAPVELRQWLEEVL	YSAAMAEQRHQEWLR	ELKNELGAGIATITR
AFVKFNCAALPDNLL	FNCAALPDNLLESEL	LSRATEASKTLQEVL	KAARLLGMTPRQVAY
EQLNEREKQIMELRF	EREKQIMELRFGLVG	GISQSYISRLEKRII	DVADMMGISQSYISR
EDKISGTKSERPGLK	SGTKSERPGLKKLLR	MERELIVERTKAGLE	KVAIIYDVGVSTLYK
TIHQPKDSLGETAFN	PKDSLGETAFNMLLD	FEPESGYRAMQQILS	DVAKRANVSTTTVSH
FIGGEDEPGKADIRE	EDEPGKADIREVAFA	FSSSSGYELAKQMLA	DIAIEAGVSLATVSR
ARQQEVFDLIRDHIS	EVFDLIRDHISQTGM	HISQTGMPPTRAEIA	EIAQRLGFRSPNAAE
EDEELAELAKKVAHL	LAELAKKVAHLLTKE	GINESQISRWKGDFI	KVADALGINESQISR
1 iteration	10 iterations	480 iterations	1680 iterations

Phase Shifts

The Gibbs sampling algorithm may easily converge on a local optimum that is a “phase-shifted” version of the global optimum. Why?

Optimal solution:

Solution found:



```
SQKETGDILGISQMHVSRLQRKAVKKL
TQREIAKELGISRSYVSRIEKRALMKM
RVSITAENLGLTQPAVSNALKRLRTSL
CFVRAAFEAGIGLGALARLCRALDAAN
RRIEIAHALCLTERQIKIWFQNRRMKW
NQIRAADLLGLNRNTLRKKIRDLDIQV
TQRSLAKALKISHVSVSQWERGDSEPT
EKEEVAKKCGITPLQVRVWFINKRMRS
GTEKTAEAVGVDKSQISRWKRDWIPKF
```

One remedy is to add a separate “phase-shift sampling step”.

No segments are removed, but likelihoods are calculated for the current alignment and several phase-shifted alternatives. These alignments are then sampled among.

This can be understood as changing the topology, of definition of distance, on the underlying “alignment space.”