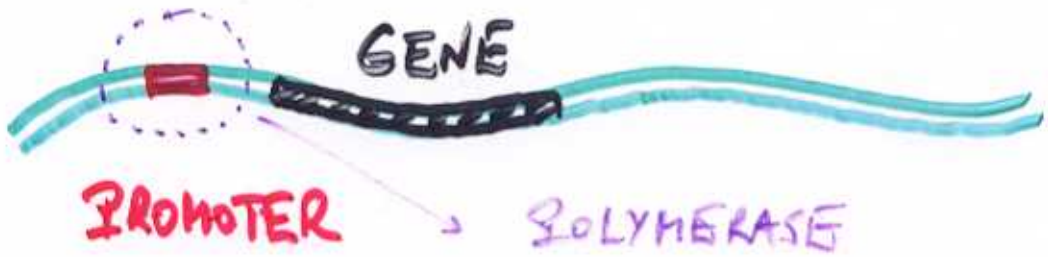
TRANSCRIPTION SIGNALS AND PROMOTER RECOGNITION

Q1: WHAT'S THE PROBLEM ?



Q2: WHY CARE ?

1. THEORETICAL:
   - UNDERSTAND GENE EXPRESSION
   - HOW DOES THE CELL FIND GENES
   - GENE REGULATION, CANCER, ...
2. PRACTICAL:
   - GENE FINDING
   - PREDICT GENE EXPRESSION
   - ...

Q3: WHAT DO WE DO ?

OUTLINE

I. MOLECULAR BIOLOGY (short)

II. METHODS (BIO INFORMATICS)

- CONSENSUS SEQUENCES
- REGULAR EXPRESSIONS
- POSITION WEIGHT MATRICES (PWM)
- HOW TO CONSTRUCT PWMs FROM SET OF ALIGNED SEQUENCES
- HOW TO ALIGN SEQUENCES TO PWM
- HOW TO FIND PWM AND ALIGNMENT FROM SET OF UN-ALIGNED SEQUENCES
  - DEFINITION OF GIBBS SAMPLER
  - PERFORMANCE OF GIBBS SAMPLER
- OUTLOOK

Q4 : WHERE DO PROTEINS COME FROM ?

→ TRANSCRIPTION

→ TRANSLATION

- PROMOTER = DNA BINDING SITE

  FOR DOCKING / ASSEMBLING TRANSCRIPT. MACHINERY

  ↳ RNA POLYMERASE AND OTHER FACTORS / COFACTORS

- -55 ... +20
- UPSTREAM ... -1 +1 ... DOWNSTREAM
- -10 SIGNAL ("TATA")
- -35 SIGNAL
- 7000 RNA POL. COPIES / CELL
- 40 nt / sec (20 × SLOWER THAN REPLIC.)

  ~ 15 aa / sec (TRANSLATION)

Q5: HOW DOES THE TRANSCRIPTION MACHINERY RECOGNIZE PROMOTERS ?

A5: ?

Q6: HOW CAN WE RECOGNIZE PROMOTERS ?

A6: 1. EXPERIMENTALLY

- TIME, $$$
- DO NOT REALLY UNDERSTAND BINDING OF 2 TO PROMOTER

2. THEORETICALLY (BIOPHYSICS)

- MODEL ALL MOLECULES (50 - 150)
- COMPUTE BINDING ENERGIES
- SOLVE EQUATIONS OF MOTION

→ CONFORMATION / BINDING

- WRONG POTENTIALS
- TOO COMPLEX !

3. THEORETICALLY (BIOINFORMATICS)

- SIMPLIFY MODELS TREMENDOUSLY
- NOT BIOPHYS. MODELS ANYMORE
- "SUPER" ABSTRACT
- ADVANTAGE: THEY WORK!

Q7: WHAT ARE CONSENSUS SEQUENCES ?

...AATCAT...
...TACATT...
...TACATT...
...TATGAT...
...TATTAT...

TATAAT

CONSENSUS SEQUENCE = SEQUENCE OF THE MOST FREQUENT NUCLEOTIDE AT EACH POSITION.

CONSENSUS SEQ $\neq$ MOST FREQUENT SEQ.

CONSENSUS SEQ DOES OFTEN NOT OCCUR AT ALL.

SHORTCOMINGS OF CONSENSUS SEQ.

1. ONLY INFORMATION ABOUT MOST-FREQUENT, NOT LEAST-FREQUENT NT.
2. NOT : HOW FREQUENT

→ REGULAR EXPRESSIONS (solve 1, not 2)

[AT]A[CT]*[AT]

→ GENERALIZED CONSENSUS SEQ (solve 2, not 1)

T_{20} A_{95} T_{45} A_{60} A_{50} T_{96} → FREQUENCY OF NT IN %

→ POSITION WEIGHT MATRICES (PWM)

SOLVE BOTH 1 AND 2

Q8 : WHAT ARE POSITION WEIGHT MATRICES ?

P_{ie} = PROBABILITY OF FINDING NUCLEOTIDE i IN POSITION e

i		1	2	3	4	5	6
1	A	0.2	1.0	0	0.4	0.6	0
2	C	0	0	0.4	0.2	0	0
3	G	0	0	0	0.2	0	0
I=4	T	0.8	0	0.6	0.2	0.4	1.0

1 2 . . . L=6

e

PWM CONTAINS ALL INFORMATION ABOUT CONSENSUS SEQ, GENERALIZED CONS. SEQ, REGULAR EXPRESS.

Q9: DO PWM's HAVE AN INTUITIVE MEANING?

PWM $\hat{=}$ "SUPER" SINGLE MODEL
(ISING MODEL, POTTS MODEL)

$E_{i\ell}$ = ENERGY IF POLYMERASE BINDS TO NUCLEOTIDE i AT POSITION ℓ

$\tilde{E}_{i\ell}$ = ENERGY IF ... DOES NOT BIND TO ...

$$\tilde{E} = \tilde{E}_{i_1 1} + \tilde{E}_{i_2 2} + \cdots + \tilde{E}_{i_L L}$$

$\Delta E_{i\ell} = E_{i\ell} - \tilde{E}_{i\ell}$ BINDING ENERGY

$\Delta E = E - \tilde{E}$ TOTAL BIND. ERERG.

Q 10 : HOW TO OBTAIN P_{WM} FROM SET OF ALIGNED SEQUENCES ?

$$\hat{p}_i = \frac{k_i}{N}$$

MAXIMUM LIKELIHOOD ESTIMATOR

$$\hat{p}_i = \frac{k_i + 1}{N + I}$$

MINIMUM VARIANCE ESTIMATOR
(BAYES)

PSEUDO COUNTS

Q12 : HOW TO ALIGN A SEQUENCE / A SET OF SEQUENCES TO A PWM ?

GIVEN : I_{ic} , Q_i , SEQUENCE

$$\text{SCORE} = \frac{I_{i_1 1}}{Q_{i_1}} \cdot \frac{I_{i_2 2}}{Q_{i_2}} \cdot \ldots \cdot \frac{I_{i_L L}}{Q_{i_L}}$$

...ACGTACCGTTTACCGT...

SCORE

- COMPUTE **SCORE** FOR EACH POSITION OF THE SEQUENCE

- CHOOSE THAT POSITION FOR WHICH **SCORE** = MAX.

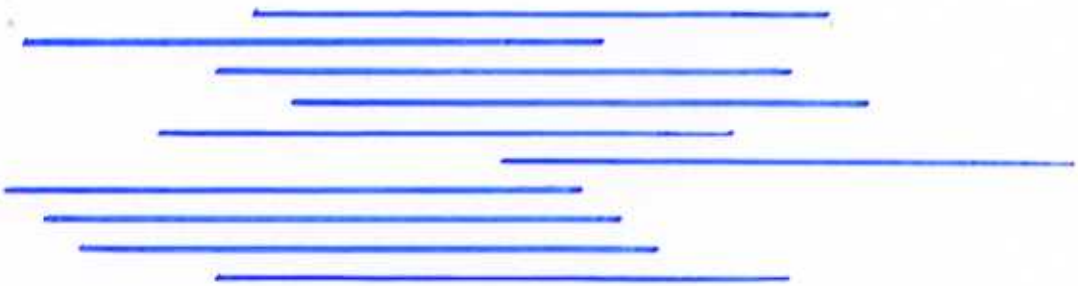
Q 13 : WHAT IF NEITHER PWM NOR ALIGNMENT IS KNOWN ?

- SEARCH FOR PATTERN, BUT DON'T KNOW WHICH PATTERN AND WHERE TO LOOK !

- WE ARE GIVEN A SET OF UN-ALIGNED SEQUENCES AND ARE TO SEARCH FOR PATTERNS
  ↑ UNKNOWN

- CONSTRUCT PWM AND ALIGNMENT SIMULTANEOUSLY

Q : CAN WE SOLVE THIS PROBLEM ?



gtctcggatcaccctccctagacgataagggcccgcacttgtagtggcgctctagcagat
cctccctagacgataagggcccgcacttgtactcgctgcttcaccgggtcttatcgattt
gtcacatgatactccctagacgataagggcccgcacttgtatactttcagcccaaaacga
accaatcaagatacactccctagacgataagggcccgcacttgtattctgtcctgtgggg
tgacgacgctccctagacgataagggcccgcacttgtaatagctccagcactacgctcgt
gatcgttctatgtcgcttatagacgcctccctagacgataagggcccgcacttgtaggtc
ctccctagacgataagggcccgcacttgtacccaatgagaatgttccccgcctaccgctc
ctctccctagacgataagggcccgcacttgtatctactctcgtccaaccagagccgatgt
ggtgctccctagacgataagggcccgcacttgtatcgagagctcctgacctgtctaagtt
ctcagtatgccctccctagacgataagggcccgcacttgtaccaggagttacttctcaga

gtctcggatcaccaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaagtggcgctctagcagat
caaaaaaaaaaaaaaaaaaaaaaaaaaaaaaactcgctgcttcaccgggtcttatcgattt
gtcacatgataaaaaaaaaaaaaaaaaaaaaaaaaaaaaatactttcagcccaaaacga
accaatcaagatacaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaattctgtcctgtgggg
tgacgacgaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatagctccagcactacgctcgt
gatcgttctatgtcgcttatagacgcaaaaaaaaaaaaaaaaaaaaaaaaaaaaaggtc
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaacccaatgagaatgttccccgcctaccgctc
ctaaaaaaaaaaaaaaaaaaaaaaaaaaaaatctactctcgtccaaccagagccgatgt
ggtgaaaaaaaaaaaaaaaaaaaaaaaaaaaaatcgagagctcctgacctgtctaagtt
ctcagtatgccaaaaaaaaaaaaaaaaaaaaaaaaaaaaaccaggagttacttctcaga

gtctcggatcaccCTCCCTAGACGATAAGGGCCCGCACTTGTAgtggcgctctagcagat
cCTCCCTAGACGATAAGGGCCCGCACTTGTActcgctgcttcaccgggtcttatcgattt
gtcacatgataCTCCCTAGACGATAAGGGCCCGCACTTGTAtactttcagcccaaaacga
accaatcaagatacaCTCCCTAGACGATAAGGGCCCGCACTTGTAttctgtcctgtgggg
tgacgacgCTCCCTAGACGATAAGGGCCCGCACTTGTAatagctccagcactacgctcgt
gatcgttctatgtcgcttatagacgcCTCCCTAGACGATAAGGGCCCGCACTTGTAggtc
CTCCCTAGACGATAAGGGCCCGCACTTGTAcccaatgagaatgttccccgcctaccgctc
ctCTCCCTAGACGATAAGGGCCCGCACTTGTAtctactctcgtccaaccagagccgatgt
ggtgCTCCCTAGACGATAAGGGCCCGCACTTGTAtcgagagctcctgacctgtctaagtt
ctcagtatgccCTCCCTAGACGATAAGGGCCCGCACTTGTAccaggagttacttctcaga

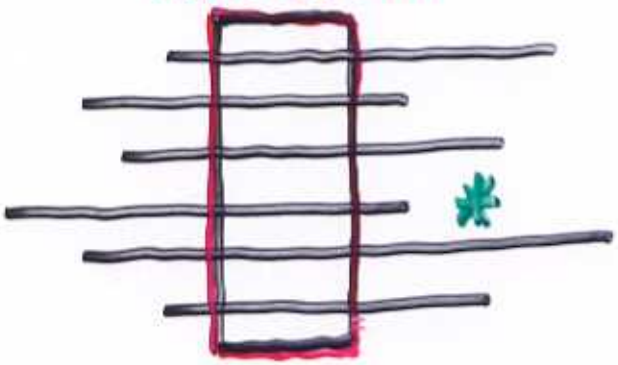
agactcttcattgaggaccatgttcacgcccgggaatagcagatccctcccgtatgactc
ttgaatctgtccagtcctcgagagctatgttcacgcccgggaatagcagatccctcgtgg
tgtagccgagaccaacctgctctcatctatgttcacgcccgggaatagcagatccctctc
ctcgccatccgccccttgtaagagtaacccatgttcacgcccgggaatagcagatccctc
ctggatgttcacgcccgggaatagcagatccctccgcagatattcgctgtatcttgctag
tgctcgcatcacgacctcgataatgttcacgcccgggaatagcagatccctcgcagcagt
ggggtgtcctgtcttatgttcacgcccgggaatagcagatccctcacatagaactaacca
agcaaaacccgactttcatatgttcacgcccgggaatagcagatccctcatagtacactg
tttagctattctgggccacttcgtcgctcatgttcacgcccgggaatagcagatccctcc
tagacgatctcgcggtgatgttcacgcccgggaatagcagatccctcccactaggctctg

Q14: HOW TO FIND PWM AND ALIGNMENT SIMULTANEOUSLY ?

IDEA: RANDOM ALIGNMENT ⇄ PWM

ITERATE !

ALIGNMENT



} N SEQS

1. CHOOSE 1 SEQ RANDOMLY (*)
   TAKE OUT *
2. COMPUTE PWM FROM N-1 SEQS
3. ALIGN SEQ * TO PWM
4. GOTO 1.

LAWRENCE, C.E., et al. SCIENCE 262: 208-214 (1993)

COMMENT:

(3.) IS NOT DETERMINISTIC

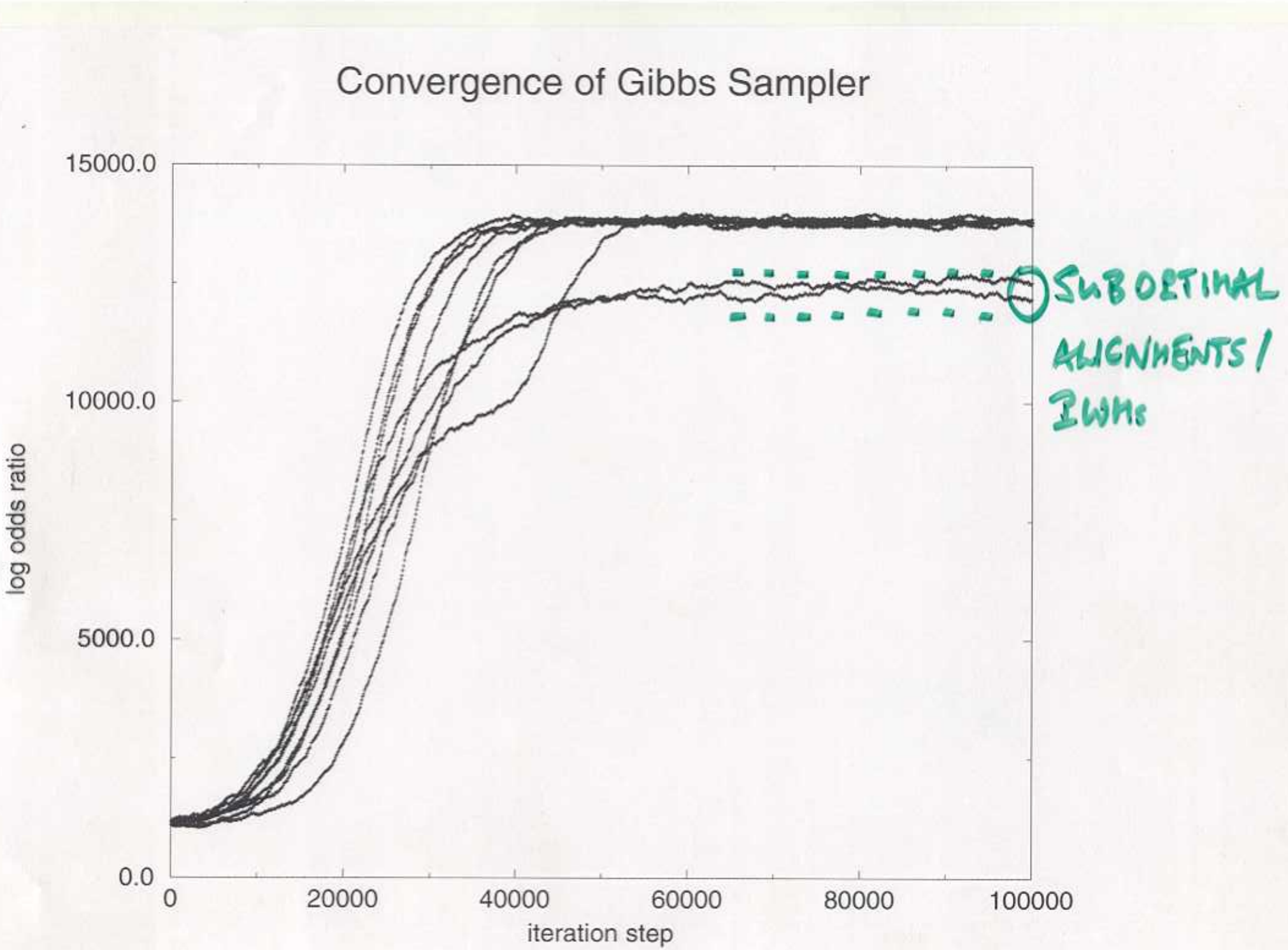
BUT STOCHASTIC

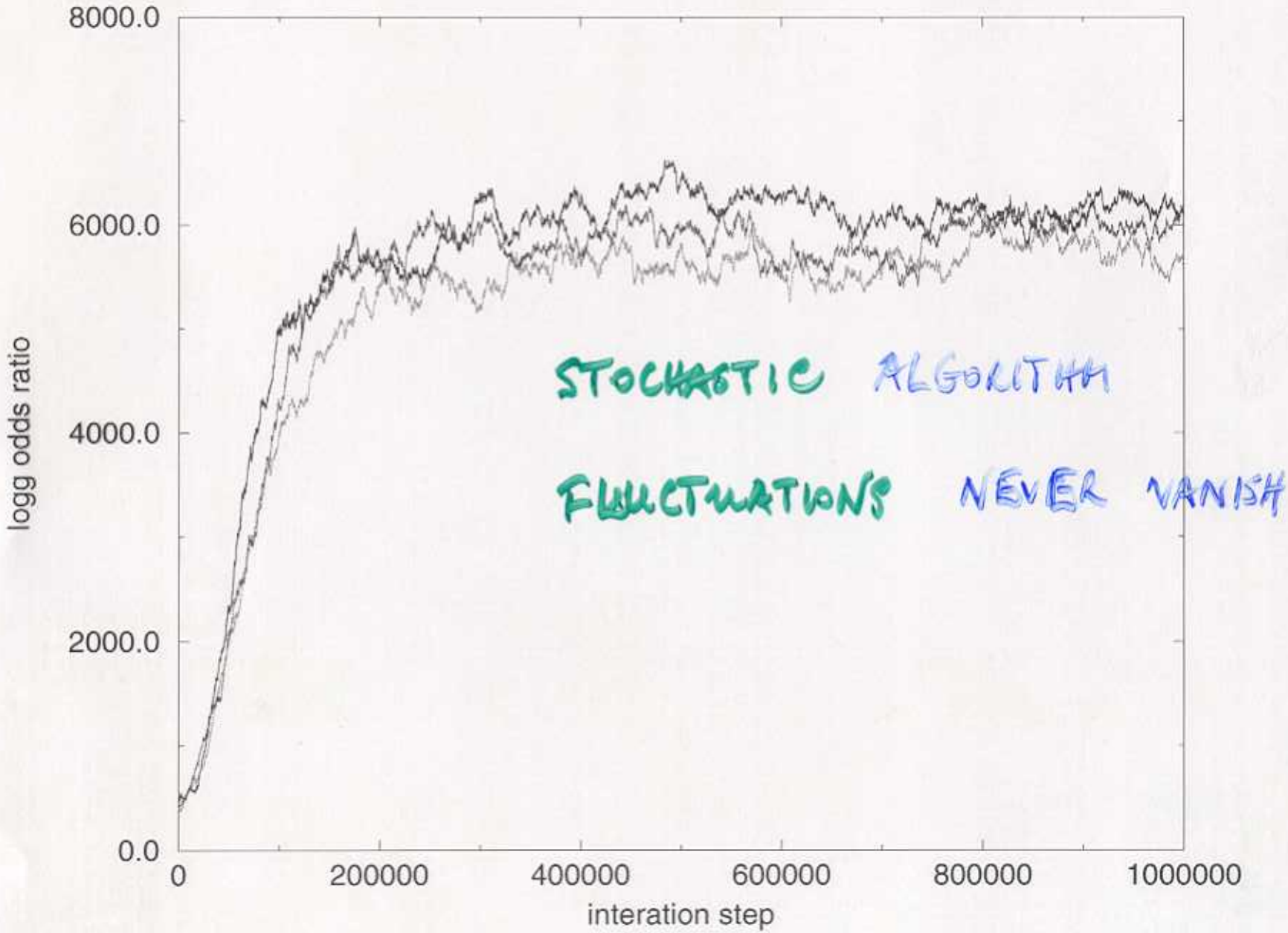
3.A: COMPUTE AT EACH POSITION

$$\text{SCORE} = \prod_{l=1}^{L} \frac{P_{i,l}}{Q_{i,l}}$$

3.B: CHOOSE POSITION RANDOMLY FROM DISTRIBUTION {SCORE (POS)}

(DO NOT CHOOSE POSITION AT WHICH SCORE = MAX.)


Convergence of Gibbs Sampler
15000.0
10000.0
5000.0
0.0
log odds ratio
0
20000
40000
60000
80000
100000
iteration step
SUBOPTIMAL ALIGNMENTS / IWHs


Convergence of Gibbs Sampler
8000.0
6000.0
4000.0
2000.0
0.0
logg odds ratio
0
200000
400000
600000
800000
1000000
interation step
STOCHASTIC ALGORITHM
FLUCTUATIONS NEVER VANISH

Q15: LIMITATIONS OF PWMs ?

1. HETEROGENEITY

- RNA POLYMERASE VERY COMPLEX MOLECULE (NOT ISING)
- RNA POL. BINDS TO MANY FACTORS

2. CORRELATIONS

- NEAREST NEIGHBORS
- LONG-RANGE CORR.

Q16: WHAT CAN WE DO ?

- PWM WITH DIMERS, TRIMERS, ...
- BETTER MODELS !!!
- GENE EXPRESSION DATA ANALYSIS !!!

Q$_{00}$: WHAT DO WE UNDERSTAND NOW THAT WE DID NOT PREVIOUSLY UNDERSTAND?

- MOL. BIOL., BIOPHYS., BIO INFO !!!
- PWM (NT ARE INDEPENDENT ≙ E IS ADDITIVE)
- PWM FROM UNALIGNED SEQs (GIBBS SAMPLER)
- PWM SIMPLE, BUT AS GOOD AS MOST-SOPHISTICATED METHODS
- NEVERTHELESS: LIMITATIONS! (WE ARE MISSING ESSENTIAL INFORMATION)
- HOPE FOR BETTER MODELS IN THE FUTURE
- GENE FINDING CANNOT YET RELY ON PROMOTER RECOGNITION