Scoring motifs

Hamming Distance

David Bernick

Finding motifs

Data

atgaccgggatactgatAgAAgAAAGGttGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatacAAtAAAAcGGcGGGa

tgagtatccctgggatgacttAAAAtAAtGGaGtGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgcAAAAAAAGGGattGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAtAAtAAAGaGGGGcttatag

gtcaatcatgttcttgtgaatggatttAAcAAtAAGGGctGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtAtAAAcAAGGaGGGccaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAtAGGGaGccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatActAAAAAGGaGcGGaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttActAAAAAGGaGcGGa

Motifs

A g A A g A A A G G t t G G G

c A A t A A A A c G G c G G G

A A A A t A A t G G a G t G G

c A A A A A A A G G G a t t G

A t A A t A A A G a G a G G G

A A c A A t A A G G G c t G G

A t A A A c A A G G a G G G c

A A A A A A t A G G G a G c c

A c t A A A A A G G a G c G G

A c t A A A A A G G a G c G G

- Each line has d=4 mutations with 15 positions. (15,4) motif

- for pairwise comparisons: *2d* possible mismatches (4 in each)

Counts (motifs)

A	8	5	7	9	7	8	9	9		1	4	3			
C	2	2	1			1			1			2	2	1	2
G		1			1				9	9	5	4	5	8	8
T		2	2	1	2	1	1	1			1	1	3	1	
	A	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**G**	**G**	**G**	**G**	**G**	**G**	**G**

Profile (motifs)

A	0.8	0.5	0.7	0.9	0.7	0.8	0.9	0.9		0.1	0.4	0.3			
C	0.2	0.2	0.1			0.1			0.1			0.2	0.2	0.1	0.2
G		0.1			0.1				0.9	0.9	0.5	0.4	0.5	0.8	0.8
T		0.2	0.2	0.1	0.2	0.1	0.1	0.1			0.1	0.1	0.3	0.1	
Consensus	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**G**	**G**	**G**	**G**	**G**	**G**	**G**

$$\text{Profile}_{\text{column=i}} = P(\, symbol_{column=i} \mid Motifs_{column=i}\,)$$

Scoring motifs given a consensus

Counts (motifs)

A	8	5	7	9	7	8	9	9		1	4	3			
C	2	2	1			1			1			2	2	1	2
G		1			1				9	9	5	4	5	8	8
T		2	2	1	2	1	1	1			1	1	3	1	
	A	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**G**	**G**	**G**	**G**	**G**	**G**	**G**

Profile (motifs)

A	0.8	0.5	0.7	0.9	0.7	0.8	0.9	0.9		0.1	0.4	0.3			
C	0.2	0.2	0.1			0.1			0.1			0.2	0.2	0.1	0.2
G		0.1			0.1				0.9	0.9	0.5	0.4	0.5	0.8	0.8
T		0.2	0.2	0.1	0.2	0.1	0.1	0.1			0.1	0.1	0.3	0.1	
Consensus		**A**	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**G**	**G**	**G**	**G**	**G**	**G**	**G**

Motifs

A	g	A	A	g	A	A	A	G	G	t	t	G	G	G		4
c	A	A	t	A	A	A	A	c	G	G	c	G	G	G		4
A	A	A	A	t	A	A	t	G	G	a	G	t	G	G		4
c	A	A	A	A	A	A	A	A	G	G	G	a	t	t	G	4
A	t	A	A	t	A	A	A	G	a	G	a	G	G	G		4
A	A	c	A	A	t	A	A	G	G	G	c	t	G	G		4
A	t	A	A	A	c	A	A	G	G	a	G	G	G	c		4
A	A	A	A	A	A	t	A	G	G	G	a	G	c	c		4
A	c	t	A	A	A	A	A	G	G	a	G	c	G	G		4
A	c	t	A	A	A	A	A	G	G	a	G	c	G	G		4
2	5	3	1	3	2	1	1	1	1	5	6	5	2	2		**40**

So.. in this case, the distance from each motif to consensus happens to be 4
We see that:

d(AAAAAAAAGGGGGGG, AgAAgAAAGGttGGG) = 4

consensus motif

We can also say:

score (Motifs) = 40

This scoring method measures the distance of a motif or set of motifs to a consensus motif

*this is the **Hamming distance***

Hamming

- assume: 10 DNA motifs of length 12, where we want to score that motif set (ACGT)

- what is the smallest possible Hamming distance ? How would that score happen?

- what is the largest possible Hamming distance? How would that score happen?

Selecting motifs

Profile methods

David Bernick

Given a profile, can we select a best 6-mer motif?

Profile

A	**1/2**	**7/8**	**3/8**	0	1/8	0
C	1/8	0	1/2	**5/8**	**3/8**	0
T	1/8	1/8	0	0	1/4	**7/8**
G	1/4	0	1/8	3/8	1/4	1/8

We are looking for:

P(motif|Profile)
in
Data

Data

CTATAAACCTTACAT	1/8 x 1/8 x 3/8 x 0 x 1/8 x 0	0
CTATAAACCTTACAT	1/2 x 7/8 x 0 x 0 x 1/8 x 0	0
CTATAAACCTTACAT	1/2 x 1/8 x 3/8 x 0 x 1/8 x 0	0
CTATAAACCTTACAT	1/8 x 7/8 x 3/8 x 0 x 3/8 x 0	0
CTATAAACCTTACAT	1/2 x 7/8 x 3/8 x 5/8 x 3/8 x 7/8	**0.0336**
CTATAAACCTTACAT	1/2 x 7/8 x 1/2 x 5/8 x 1/4 x 7/8	0.0299
CTATAAACCTTACAT	1/2 x 0 x 1/2 x 0 1/4 x 0	0
CTATAAACCTTACAT	1/8 x 0 x 0 x 0 x 0 x 1/8 x 0	0
CTATAAACCTTACAT	1/8 x 1/8 x 0 x 0 x 3/8 x 0	0
CTATAAACCTTACAT	1/8 x 1/8 x 3/8 x 5/8 x 1/8 x 7/8	0.0004

Given a Profile, can we select a set of Motifs?

Dna with implanted (4,1)-motif ACGT

```
ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT
```

Profile(*Motifs*)

A:	2/5	1/5	1/5	1/5
C:	1/5	2/5	1/5	1/5
G:	1/5	1/5	2/5	1/5
T:	1/5	1/5	1/5	2/5

.0016/ttAC .0016/tACC **.0128**/ACCT .0064/CCT**t** .0016/C**tta** .0016/T**taa** .0016/**taac**

.0016/gAT**G** **.0128**/AT**GT** .0016/T**GTc** .0032/**GTct** .0032/**Tct**g .0032/**ct**gt .0016/**t**gtc

.0064/**ccgG** .0036/**cgG**C .0016/**gG**CG **.0128**/**G**CGT .0032/CGTt .0016/Gtta .0016/Ttag

.0032/c**act** .0064/**acta** .0016/**cta**A .0016/**ta**AC .0032/**a**ACG **.0128**/ACGA .0016/CGAg

.0016/cgtc .0016/gtca .0016/tcag .0032/cag**A** .0032/ag**AG** .0032/g**AGG** **.0128**/**AGGT**

```
ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT
```

P(motifs|Profile)

EM

Dna with implanted (4,1)-motif ACGT

```
ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT
```

Motifs

```
 t a a c
G T c t
c c g G
a c t a
A G G T
```

Profile(***Motifs***)

A:	2/5	1/5	1/5	1/5
C:	1/5	2/5	1/5	1/5
G:	1/5	1/5	2/5	1/5
T:	1/5	1/5	1/5	2/5

$= f(\text{motifs}) = \Theta_{current}$

```
.0016/ttAC  .0016/tACC  .0128/ACCT  .0064/CCTt  .0016/Ctta  .0016/Ttaa  .0016/taac
.0016/gATG  .0128/ATGT  .0016/TGTc  .0032/GTct  .0032/Tctg  .0032/ctgt  .0016/tgtc
.0064/ccgG  .0036/cgGC  .0016/gGCG  .0128/GCGT  .0032/CGTt  .0016/Gtta  .0016/Ttag
.0032/cact  .0064/acta  .0016/ctaA  .0016/taAC  .0032/aACG  .0128/ACGA  .0016/CGAg
.0016/cgtc  .0016/gtca  .0016/tcag  .0032/cagA  .0032/agAG  .0032/gAGG  .0128/AGGT
```

maximum P(motifs|Profile, DNA) determines next set of motifs

next set of motifs

```
ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT
```

A	4/5			1/5
C		3/5		
G	1/5	1/5	4/5	
T		1/5	1/5	4/5

Θ_{new}

$$\Theta_{t=2} = Profile(\, motifs(\, \Theta_{t=1}, data\,)\,)$$

$$\Theta_{t=3} = Profile(\, motifs(\, \Theta_{t=2}, data\,)\,)$$

$$\ldots$$

What is $\boldsymbol{\theta}_{t=0}$?

How do we choose motifs($\boldsymbol{\theta}_{t=0}$) ?

When are we done ?

Is $\boldsymbol{\theta}_{last}$ the best?

If not … how do we get better ?

RandomizedMotifSearch

Dna with implanted (4,1)-motif ACGT	Randomly select *Motifs* (in **bold**)	*Motifs*	*Profile*(*Motifs*)
ttACCTtaac gATGTctgtc ccgGCGTtag cactaACGAg cgtcagAGGT	ttACCT**taac** gAT**GTct**gtc **ccgG**CGTtag c**acta**ACGAg cgtcag**AGGT**	**t a a c** **G T c t** **c c g G** **a c t a** **A G G T**	A: 2/5 1/5 1/5 1/5 C: 1/5 2/5 1/5 1/5 G: 1/5 1/5 2/5 1/5 T: 1/5 1/5 1/5 2/5

.0016/ttAC .0016/tACC **.0128**/ACCT .0064/CCT**t** .0016/C**tta** .0016/T**taa** .0016/**taac**

.0016/gAT**G** **.0128**/AT**GT** .0016/T**GTc** .0032/**GTct** .0032/**Tct**g .0032/**ct**gt .0016/**t**gtc

.0064/**ccgG** .0036/**cgG**C .0016/**gG**CG **.0128**/**G**CGT .0032/CGTt .0016/Gtta .0016/Ttag

.0032/c**act** .0064/**acta** .0016/**cta**A .0016/**ta**AC .0032/**a**ACG **.0128**/ACGA .0016/CGAg

.0016/cgtc .0016/gtca .0016/tcag .0032/cag**A** .0032/ag**AG** .0032/g**AGG** **.0128**/**AGGT**

ttACCTtaac
gATGTctgtc
ccgGCGTtag
cactaACGAg
cgtcagAGGT

Motifs (*Profile* (***Motifs***), *Dna*)

RandomizedMotifSearch

- With randomly selected k-mers, we expect an uninformative profile, initially with uniform(ish) probabilities at each position.

- A bias exists in {DNA} due to the common motif (if it exists) s.t. when we maximize the Pr(motif | profile), the parameters of our model Θ (derived from that new motif)

  begin to pick up that bias

- Notice that this search is fully deterministic after selection of the initial motifs

Profile scoring

Shannon’s entropy

David Bernick

Profile scoring

- hamming distance to consensus
- Shannon’s entropy, relative entropy, encoding cost

Hamming distance *d()*

	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
Profile(Motifs)	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
consensus		T	C	G	G	G	G	A	T	T	T	C	C
d(Profile) N=10		3	4	0	0	1	1	1	5	2	3	6	4

- *d(profile |N=10, columns = 12)* = 30
- minimum d(profile|*N=10, columns = 12*) = 0
- maximum d(profile|*N=10, columns = 12*) = 7 * 12 columns = 84

Entropy

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

	-0.46	-0.46	0.00	0.00	0.00	0.00	-0.14	-0.33	-0.33	-0.33	-0.52	0.00	
	-0.33	-0.44	0.00	0.00	0.00	0.00	0.00	-0.53	-0.33	-0.46	-0.53	-0.44	
	0.00	0.00	0.00	0.00	-0.14	-0.14	-0.33	0.00	0.00	0.00	0.00	0.00	
	-0.36	-0.46	0.00	0.00	-0.33	-0.33	0.00	-0.50	-0.26	-0.36	-0.52	-0.53	
H(x)	**1.16**	**1.37**	**0.00**	**0.00**	**0.47**	**0.47**	**0.47**	**1.36**	**0.92**	**1.16**	**1.57**	**0.97**	
													9.92

$$H(x) = -\sum_{i=1}^{n} P(x_i) log_2(P(x_i))$$

Entropy examples

$$H(x) = -\sum_{i=1}^{n} P(x_i) log_2(P(x_i))$$

define: 0 $log_2(0)$ = 0

A	0	0.25
C	1	0.25
G	0	0.25
T	0	0.25
H	**0**	**2**

- a column with constant data has no information
  a column with 4 characters of equal frequency has 2 “bits” of information

d(x) vs H(x)

N=120

	Case 1	Case 2	Case 3	Case 4
A	0	0.25	0.7	0.7
C	1	0.25		0.1
G	0	0.25	0.3	0.1
T	0	0.25		0.1
d(x)	0	90	36	36
H(x)	**0**	**2**	**0.88**	**1.36**

- Which better reflects the information in our columns?

Pseudocounts

Profiles that reflect our conviction

David Bernick

pseudo counts

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

- In this case, if we found a sequence with A at position 3,
  P(sequence | profile) = 0

- Maybe we believe that, or could this be caused by under sampling ?

- How does the profile change if we add some fixed amount
  to every symbol in every column ?

pseudo counts

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

with pseudo counts	A:	.24	.24	.20	.20	.20	.20	.38	.22	.22	.22	.26	.20
	C:	.22	.32	.20	.20	.20	.20	.20	.28	.22	.24	.28	.32
	G:	.20	.20	.40	.40	.38	.38	.22	.20	.20	.20	.20	.20
	T:	.34	.24	.20	.20	.22	.22	.20	.30	.36	.34	.26	.28

- N=10, with 10 pseudocounts

pseudo counts

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

with pseudo counts	A:	.24	.24	.20	.20	.20	.20	.38	.22	.22	.22	.26	.20
	C:	.22	.32	.20	.20	.20	.20	.20	.28	.22	.24	.28	.32
	G:	.20	.20	.40	.40	.38	.38	.22	.20	.20	.20	.20	.20
	T:	.34	.24	.20	.20	.22	.22	.20	.30	.36	.34	.26	.28

- N=10, with 10 pseudocounts
- What happened to the distribution?
- What happened to the consensus sequence?

Profile scoring

Interpreting scores

David Bernick

Entropy and Relative Entropy

$$H(x) = -\sum_{i \in [A,C,G,T]} P(x_i) log_2 P(x_i)$$

- Entropy - encoding cost

- Relative Entropy - a distance

$$D_{KL}(P||Q) = \sum_{i \in [A,C,G,T]} P(x_i) log_2 \frac{P(x_i)}{Q(x_i)}$$

- If we want distance between our profile and its base composition, Q is a distribution over our sequences

- *P describes our experimental model — $\boldsymbol{\Theta}_{experiment}$*

  Q describes our null model — $\boldsymbol{\Theta}_{null}$

Scores

- Scores are useful as comparisons
- When we score a final profile, we will always get a final, best score?
- What should we compare it to?

Sampling algorithm

Gibbs sampling

David Bernick

Sampling

- RandomizedMotifSearch will continue to improve a profile score until it can't. At that point, we stop it.

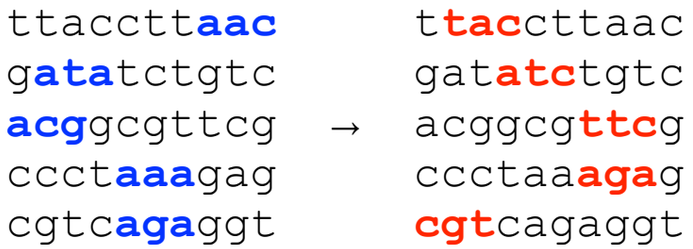
- Is this the best search algorithm?

- Why do we see improvements with new trajectories — new initial motif selections ?

- As we continue downhill, improving our score, can we hop up on rocks to get a better view?

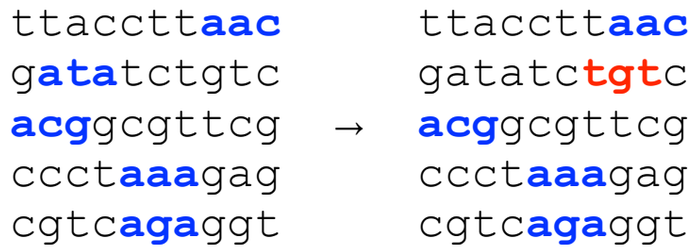
Gibbs Sampling

RandomizedMotifSearch may replace all *k*-mers in a single iteration and thus may potentially discard a nearly correct motif.



RandomizedMotifSearch
(may change **all** *k*-mers in 1 iteration)

Gibbs sampling replaces a single *k*-mer at each iteration and thus moves with more caution in the space of all motifs.



GibbsSampler
(changes a **single** *k*-mer in 1 iteration)

Gibbs Sampling

Randomly select ***Motifs*** (in **bold**)

ttACCT**taac**
gAT**GT**c**t**gtc
ccg**GCGT**tag
c**acta**ACGAg
cgtcag**AGGT**

Randomly remove a *k*-mer in ***Motifs***

ttACCT**taac**
gAT**GT**c**t**gtc
----
c**acta**ACGAg
cgtcag**AGGT**

Motifs matrix

t	**a**	**a**	**c**
G	**T**	**c**	**t**
----------			
a	**c**	**t**	**a**
A	**G**	**G**	**T**

Count matrix

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

Profile matrix

A:	2/4	1/4	1/4	1/4
C:	**0**	1/4	1/4	1/4
G:	1/4	1/4	1/4	**0**
T:	1/4	1/4	1/4	2/4

Choose a new starting position in the deleted sequence.
Weighted Random selection based on k-mer probabilities

Calculate the probabilities of all *k*-mers in the deleted string
ccgGCGTtag

0 (**ccgG**) **0** (**cgG**C) **0** (**gG**CG) **1/128** (**G**CGT) **0** (CGTt) **0** (GTta) **0** (Ttag)

Gibbs Sampler

1. Randomly choose one of selected *k*-mers (from *RemovedSequence)* and remove it from *Motifs*.
2. Create *Profile* from the remaining *k*-mers in *Motifs*.
3. For each *k*-mer in *RemovedSequence*, calculate *Pr(k*-mer|*Profile)* resulting in *n-k+1* probabilities:

   $p_1, p_2, \ldots, p_{n-k+1}$.
4. **Roll a die (with *n-k+1* sides) where probability of ending up at side *i* is proportional to p_i.**
5. Choose a new starting position based on rolling the die. Add the *k*-mer starting at this position in *RemovedSequence* to *Motifs*.
6. **Repeat steps 2-6.**

Notice that the Gibbs step(4) can move “backwards”

Gibbs Sampling with Pseudocounts

Randomly select ***Motifs*** (in **bold**)

ttACCT**taac**
gAT**GTct**gtc
ccg**GCGT**tag
c**acta**ACGAg
cgtcag**AGGT**

Randomly remove a *k*-mer in ***Motifs***

\- - - - -
gAT**GTct**gtc
ccg**GCGT**tag
c**acta**ACGAg
cgtcag**AGGT**

Motifs matrix

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

Count matrix

A:	**2+1**	**0+1**	**0+1**	1+**1**
C:	**0+1**	2+**1**	1+**1**	**0+1**
G:	2+**1**	1+**1**	2+**1**	**0+1**
T:	0+**1**	1+**1**	1+1	**3+1**

Profile matrix

A:	**3/8**	**1/8**	**1/8**	**2/8**
C:	**1/8**	**3/8**	**2/8**	**1/8**
G:	**3/8**	**2/8**	**3/8**	**1/8**
T:	**1/8**	**2/8**	**2/8**	**4/8**

Choose a new starting position in the deleted sequence.

Weighted Random selection based on k-mer probabilities

Calculate the probabilities of all *k*-mers in the deleted string ttACCT**taac**

2/4096 (ttAC) 2/4096 (tACC) 72/4096 (ACCT) 24/4096 (CCT**t**) 8/4096 (C**Tta**) 4/4096 (**Ttaa**) 1/4096 (**taac**)