Outline

What do you need to know?

- I understand you have a biological background
- You will likely not develop new algorithms from core
- You will mainly use methods, possibly design
- Know about, select between, use practically

What do you need to know?

- I understand you have a biological background
- You will likely not develop new algorithms from core
- You will mainly use methods, possibly design
- Know about, select between, use practically

What do you need to know?

- I understand you have a biological background
- You will likely not develop new algorithms from core
- You will mainly use methods, possibly design
- Know about, select between, use practically

What do you need to know?

- I understand you have a biological background
- You will likely not develop new algorithms from core
- You will mainly use methods, possibly design
- Know about, select between, use practically

What do you need to know?

- I understand you have a biological background
- You will likely not develop new algorithms from core
- You will mainly use methods, possibly design
- Know about, select between, use practically

So, why understand a nice "black box"

- Always good to learn !?
- Understand and communicate with CS people
- May be easier to learn workings than all characteristics:
  - Select appropriate to task
  - Understand limitations

So, why understand a nice "black box"

- Always good to learn !?
- Understand and communicate with CS people
- May be easier to learn workings than all characteristics:
  - Select appropriate to task
  - Understand limitations

So, why understand a nice "black box"

- Always good to learn !?
- Understand and communicate with CS people
- May be easier to learn workings than all characteristics:
  - Select appropriate to task
  - Understand limitations

So, why understand a nice "black box"

- Always good to learn !?
- Understand and communicate with CS people
- May be easier to learn workings than all characteristics:
  - Select appropriate to task
  - Understand limitations

So, why understand a nice "black box"

- Always good to learn !?
- Understand and communicate with CS people
- May be easier to learn workings than all characteristics:
  - Select appropriate to task
  - Understand limitations

So, why understand a nice "black box"

- Always good to learn !?
- Understand and communicate with CS people
- May be easier to learn workings than all characteristics:
  - Select appropriate to task
  - Understand limitations

My perspective today

- Mainly focus on the core problems and solutions
- First in a purely abstract (CS) way, then tying up
- Practical lab tomorrow
- You: Focus on intuitions, not technical details!

My perspective today

- Mainly focus on the core problems and solutions
- First in a purely abstract (CS) way, then tying up
- Practical lab tomorrow
- You: Focus on intuitions, not technical details!

My perspective today

- Mainly focus on the core problems and solutions
- First in a purely abstract (CS) way, then tying up
- Practical lab tomorrow
- You: Focus on intuitions, not technical details!

My perspective today

- Mainly focus on the core problems and solutions
- First in a purely abstract (CS) way, then tying up
- Practical lab tomorrow
- You: Focus on intuitions, not technical details!

My perspective today

- Mainly focus on the core problems and solutions
- First in a purely abstract (CS) way, then tying up
- Practical lab tomorrow
- You: Focus on intuitions, not technical details!

Outline

The problem

- Given some examples, e.g: ACGT, ACCT, AGAT
- Where are other occurrences in a large sequence, e.g.: AGGGCCTGAGGTAAACA
- In reality: tens of examples, each 5-15 long, with several large sequences 1k-10k long
- Think, suggest and discuss!

The problem

- Given some examples, e.g: ACGT, ACCT, AGAT
- Where are other occurrences in a large sequence, e.g.: AGGGCCTGAGGTAAACA
- In reality: tens of examples, each 5-15 long, with several large sequences 1k-10k long
- Think, suggest and discuss!

The problem

- Given some examples, e.g: ACGT, ACCT, AGAT
- Where are other occurrences in a large sequence, e.g.: AGGGCCTGAGGTAAACA
- In reality: tens of examples, each 5-15 long, with several large sequences 1k-10k long
- Think, suggest and discuss!

The problem

- Given some examples, e.g: ACGT, ACCT, AGAT
- Where are other occurrences in a large sequence, e.g.: AGGGCCTGAGGTAAACA
- In reality: tens of examples, each 5-15 long, with several large sequences 1k-10k long
- Think, suggest and discuss!

Divide into two subproblems

- Known sites (examples), unknown motif: construct motif
- Unknown sites, known motif: determine hits

Divide into two subproblems

- Known sites (examples), unknown motif: construct motif
- Unknown sites, known motif: determine hits

Step 1: Building a motif representation

1. . . . CGCGCCGTGCGTTGACCCGCC. . .
2. . . . TACCATTGACACGGCTACCAG. . .
3. . . . ACCATACAGTTAACGATCCGA. . .

Step 1: Building a motif representation

1. ...CGCGCCGTGCG**TTGAC**CCGCC...
2. ...TACCA**TTGAC**ACGGCTACCAG...
3. ...ACCATACAG**TTAAC**GATCCGA...

Step 1: Building a motif representation

1. . . . CGCGCCGTGCG**TTGAC**CCGCC. . .
2. . . . TACCA**TTGAC**ACGGCTACCAG. . .
3. . . . ACCATACAG**TTAAC**GATCCGA. . .

1. TTGAC, allowing 1 mismatch
2. or maybe TT[GA]AC
3. Or other consensus model, e.g. PWM . . .

Step 1: Building a motif representation

1. . . . CGCGCCGTGCG**TTGAC**CCGCC. . .
2. . . . TACCA**TTGAC**ACGGCTACCAG. . .
3. . . . ACCATACAG**TTAAC**GATCCGA. . .

1. TTGAC, allowing 1 mismatch
2. or maybe TT[GA]AC
3. Or other consensus model, e.g. PWM . . .

Step 1: Building a motif representation

1. . . . CGCGCCGTGCG**TTGAC**CCGCC. . .
2. . . . TACCA**TTGAC**ACGGCTACCAG. . .
3. . . . ACCATACAG**TTAAC**GATCCGA. . .

1. TTGAC, allowing 1 mismatch
2. or maybe TT[GA]AC
3. Or other consensus model, e.g. PWM . . .

Step 1: Building a motif representation

1. . . . CGCGCCGTGCG**TTGAC**CCGCC. . .
2. . . . TACCA**TTGAC**ACGGCTACCAG. . .
3. . . . ACCATACAG**TTAAC**GATCCGA. . .

1. TTGAC, allowing 1 mismatch
2. or maybe TT[GA]AC
3. Or other consensus model, e.g. PWM . . .

PWM

- Explained well by others..
- (I will use a presentation made by Jacques Van Helden)

Regulatory sequence analysis

Matrix-based pattern matching

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Regulatory sites : matrix description

Alignment matrix

Pos Base	1	2	3	4	5	6	7	8	9	10	11	12
A	1	3	2	0	8	0	0	0	0	0	1	2
C	2	2	3	8	0	8	0	0	0	2	0	2
G	1	2	3	0	0	0	8	0	5	4	5	2
T	4	1	0	0	0	0	0	8	3	2	2	2
			V	C	A	C	G	T	K	B		

Binding site for the yeast Pho4p transcription factor

(Source : Transfac matrix F$PHO4_01)

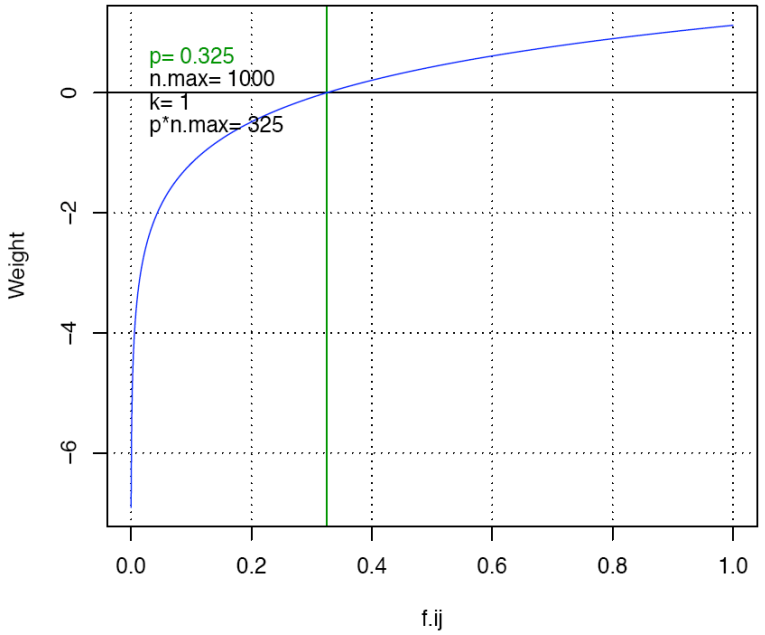
Position-weight matrix

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.33	A	-0.79	0.13	-0.23	-2.20	**1.05**	-2.20	-2.20	-2.20	-2.20	-2.20	-0.79	-0.23
0.18	C	0.32	0.32	**0.70**	**1.65**	-2.20	**1.65**	-2.20	-2.20	-2.20	0.32	-2.20	0.32
0.18	G	-0.29	0.32	**0.70**	-2.20	-2.20	-2.20	**1.65**	-2.20	**1.19**	**0.97**	**1.19**	0.32
0.33	T	0.39	-0.79	-2.20	-2.20	-2.20	-2.20	-2.20	**1.05**	0.13	-0.23	-0.23	-0.23
1	Sum	-0.37	-0.02	-1.02	-4.94	-5.55	-4.94	-4.94	-5.55	-3.08	-1.13	-2.03	0.19

$$W_{i,j} = \ln\left(\frac{f'_{i,j}}{p_i}\right)$$

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k} \qquad \sum_{i=1}^{A} f'_{i,j} = 1$$

A	alphabet size (=4)
p_i	prior residue probability for residue i
$f_{i,j}$	relative frequency of residue i at position j
k	pseudo weight (arbitrary, 1 in this case)
$f'_{i,j}$	corrected frequency of residue i at position j



Information content

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.33	A	-0.12	0.05	-0.06	-0.08	**0.97**	-0.08	-0.08	-0.08	-0.08	-0.08	-0.12	-0.06
0.18	C	0.08	0.08	**0.25**	**1.50**	-0.04	**1.50**	-0.04	-0.04	-0.04	0.08	-0.04	0.08
0.18	G	-0.04	0.08	**0.25**	-0.04	-0.04	-0.04	**1.50**	-0.04	**0.68**	**0.45**	**0.68**	0.08
0.33	T	0.19	-0.12	-0.08	-0.08	-0.08	-0.08	-0.08	**0.97**	0.05	-0.06	-0.06	-0.06
1	Sum	0.11	0.09	**0.36**	**1.29**	**0.80**	**1.29**	**1.29**	**0.80**	**0.61**	**0.39**	**0.47**	0.04

$$I_{i,j} = f'_{i,j} \ln\left(\frac{f'_{i,j}}{p_i}\right)$$

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k}$$

$$I_j = \sum_{i=1}^{A} I_{i,j}$$

$$I_{matrix} = \sum_{j=1}^{w} \sum_{i=1}^{A} I_{i,j}$$

A	*alphabet size (=4 for DNA)*
w	*matrix width (=12)*
p_i	*prior residue probability for residue i*
$f_{i,j}$	*relative frequency of residue i at position j*
k	*pseudo weight (arbitrary, 1 in this case)*
$f'_{i,j}$	*corrected frequency of residue i at position j*

Scanning a sequence with a profile matrix

- The weight matrix is successively aligned to each position of the sequence, and the score is the sum of weights for the letters aligned at each position (Hertz & Stormo, 1999).

Ex: sequence . . . . . **GCTGCACGTGGCCC . .**

Weight matrix

	1	2	3	4	5	6	7	8	9	10	11	12
A	-0.8	0.1	-0.2	-2.2	1.0	-2.2	-2.2	-2.2	-2.2	-2.2	-0.8	-0.2
C	0.3	0.3	0.7	1.6	-2.2	1.6	-2.2	-2.2	-2.2	0.3	-2.2	0.3
G	-0.3	0.3	0.7	-2.2	-2.2	-2.2	1.6	-2.2	1.2	1.0	1.2	0.3
T	0.4	-0.8	-2.2	-2.2	-2.2	-2.2	-2.2	1.0	0.1	-0.2	-0.2	-0.2

Scanning

1	**SUM**	**G**	**C**	**T**	**G**	**C**	**A**	**C**	**G**	**T**	**G**	**G**	**C**	**C**	**C**
	-10.54	**-0.3**	**0.3**	**-2.2**	**-2.2**	**-2.2**	**-2.2**	**-2.2**	**-2.2**	**0.1**	**1.0**	**1.2**	**0.3**		
2		**C**	**T**	**G**	**C**	**A**	**C**	**G**	**T**	**G**	**G**	**C**	**C**	**C**	
	7.55	**0.3**	**-0.8**	**0.7**	**1.6**	**1.0**	**1.6**	**1.6**	**1.0**	**1.2**	**1.0**	**-2.2**	**0.3**		
3		**T**	**G**	**C**	**A**	**C**	**G**	**T**	**G**	**G**	**C**	**C**	**C**		
	-9.93	**0.4**	**0.3**	**0.7**	**-2.2**	**-2.2**	**-2.2**	**-2.2**	**-2.2**	**1.2**	**0.3**	**-2.2**	**0.3**		

Interpretation of the matching weight

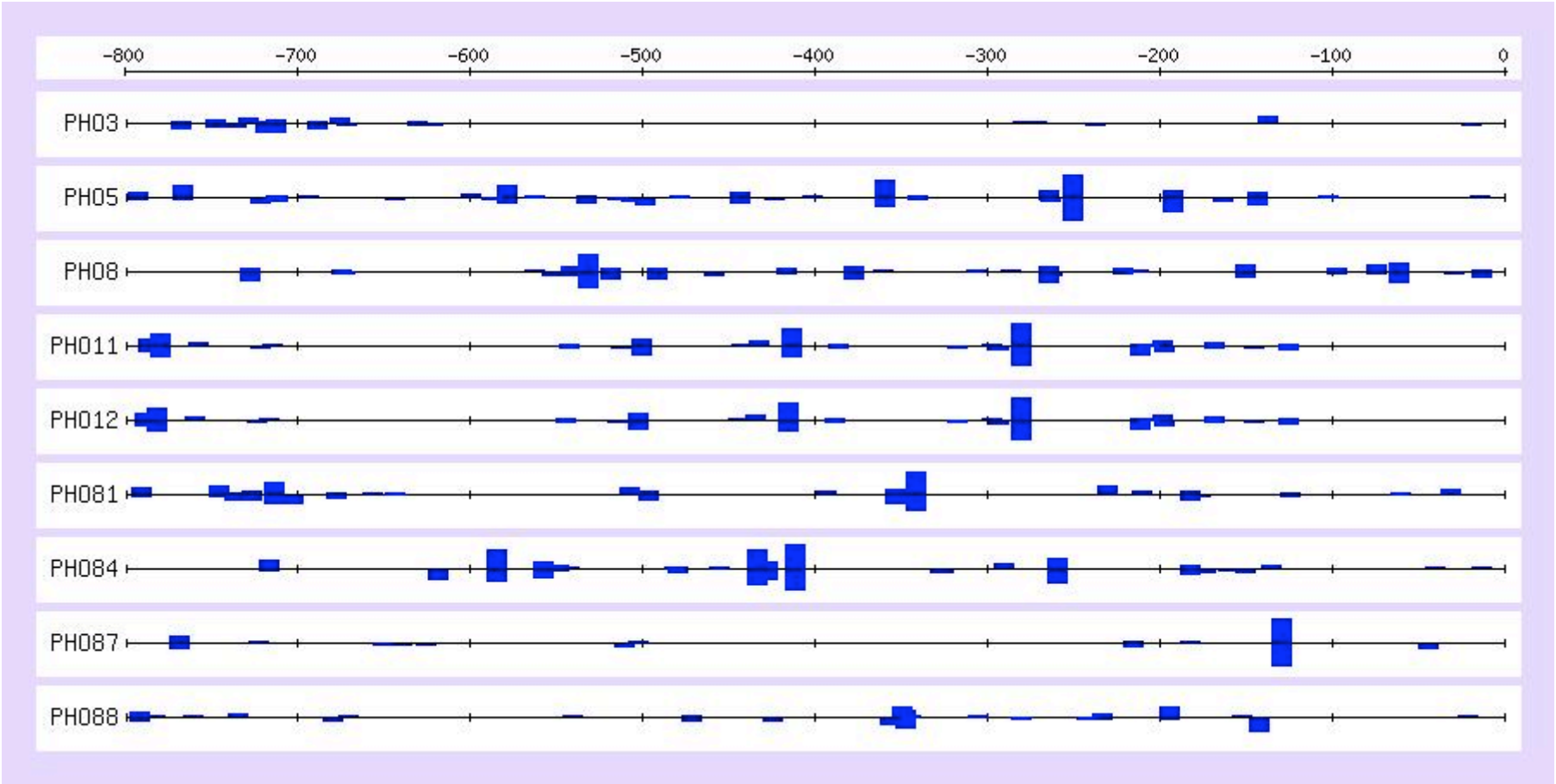
$$W_S = \sum_{k=1}^{w} W_{r_k k} = \sum_{k=1}^{w} \ln\left(\frac{f'_{r_k k}}{p_{r_k}}\right) = \ln\left(\prod_{k=1}^{w} \frac{f'_{r_k k}}{p_{r_k}}\right) = \ln\left(\frac{\prod_{k=1}^{w} f'_{r_k k}}{\prod_{k=1}^{w} p_{r_k}}\right) = \ln\left(\frac{P(S \mid M)}{P(S \mid B)}\right)$$

- The matching between a matrix and a segment of sequence is the sum of weights of the aligned residues.
- This is equivalent to the logarithm of the ratio between
  - (1) the product of the matrix frequencies, and
  - (2) the product of the prior probabilities for the residues found in the sequence segment.
- The term (1) is the probability to observe the sequence segment within the motif described by the PSSM. The term (2) is the probability to observe the sequence segment within the background.
- In other terms, the segment weight is the log likelihood ratio of its probability to be found within/without the motif. It indicates the likelihood to be within a motif when we observe that sequence segment.

Ws	weight of sequence segment S
k	position within the alignment
r_k	residue at position k of the sequence segment
p_{rk}	prior probability of residue r_k
f'_{rkk}	probability of residue r_k at positio k of the matric
$P(S\|M)$	probability of the sequence segment, given the matrix
$P(S\|B)$	probability of the sequence segment, given the background

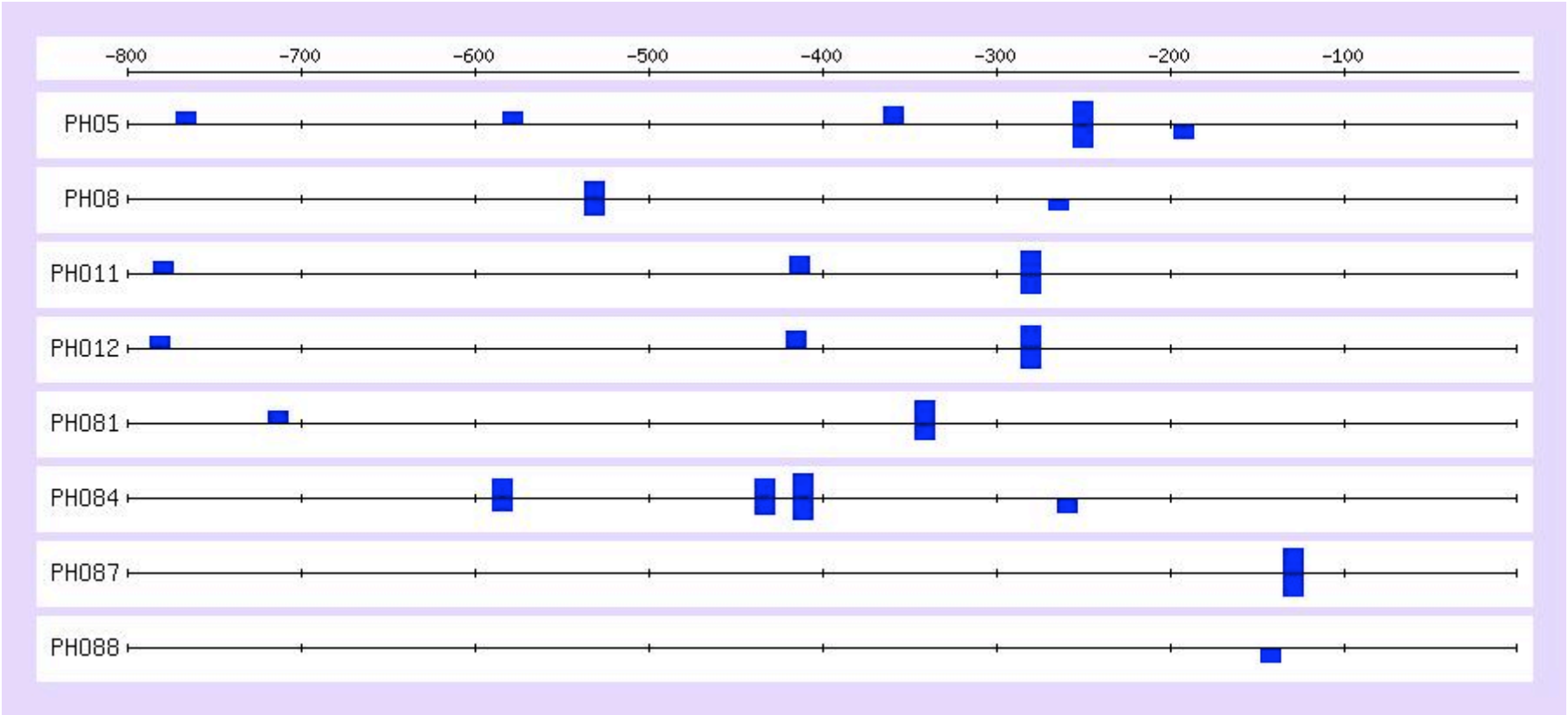
Matrix search : matching positions

- Matrix-based pattern matching is more sensitive than string-based pattern matching.
- How to choose the threshold ?

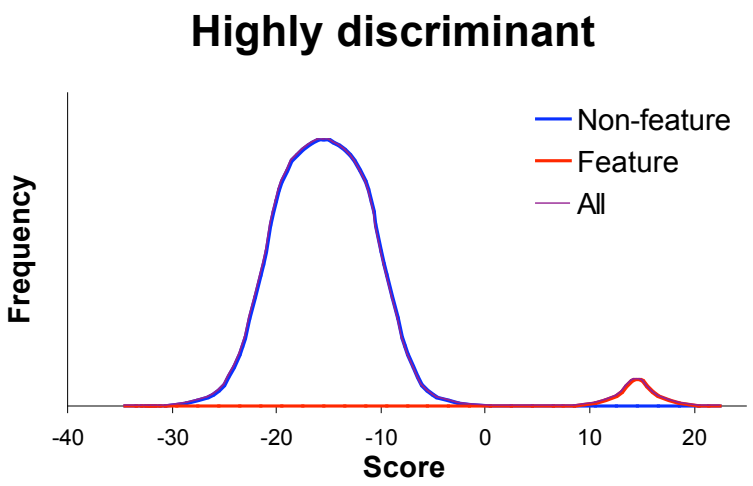


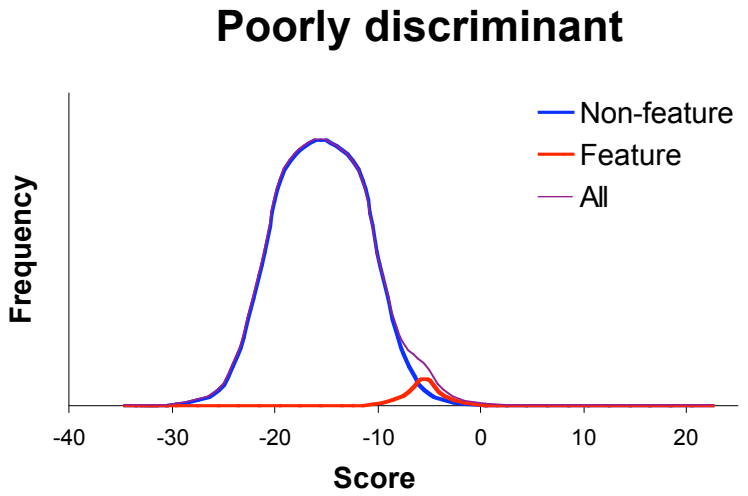
Matrix search : threshold selection

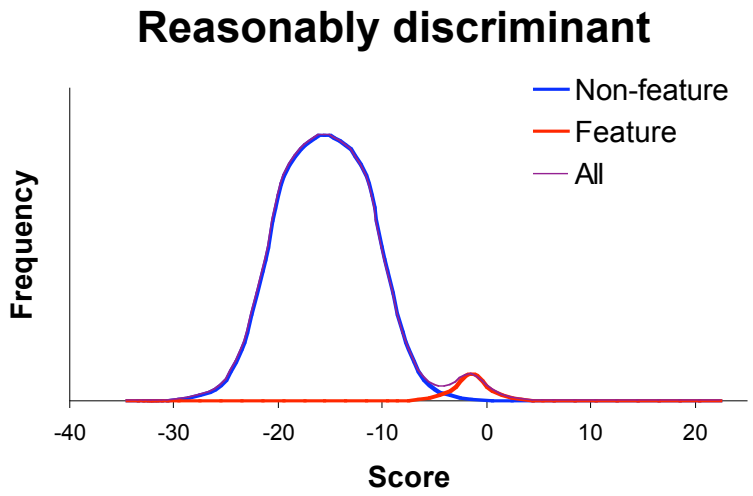
- Patser includes an option to automatically select a threshold on the basis of
  - the information content of the matrix
  - the length of the sequence to be scanned
- Note : the gene PHO3 is not displayed because there was not a single match. This gene is indeed not regulated by phosphate.
- Another approach would be to select the threshold on the basis of scores returned when the matrix is used to scan known binding sites for the factor.



Discrimination power of a matrix







Matrix search

- The sequence is scanned with the matrix, and a score is assigned to each position.
- The highest score reflects the highest probability of having a functional site.
- How to define the threshold ? There is a trade :
  - high selectivity ⇔ low sensitivity
  - high confidence in the predicted sites, but many real sites are missed
  - low selectivity ⇔ high sensitivity
    the real sites are drawn in a see of false positive

Step 2: Determine hits

- Mismatches/Reg.exp: simple scanning
- PWM: Score threshold
- PWM: Motif prob. vs background prob

Step 2: Determine hits

- Mismatches/Reg.exp: simple scanning
- PWM: Score threshold
- PWM: Motif prob. vs background prob

Step 2: Determine hits

- Mismatches/Reg.exp: simple scanning
- PWM: Score threshold
- PWM: Motif prob. vs background prob

Outline

The problem

- Unknown sites, unknown motif: find both!
- Where are the occurrences in a large sequence, e.g.: AGGGCCTGAGGTAAACA
- In reality several large sequences 1k-10k long
- Think, write notes and discuss!

The problem

- Unknown sites, unknown motif: find both!
- Where are the occurrences in a large sequence, e.g.: AGGGCCTGAGGTAAACA
- In reality several large sequences 1k-10k long
- Think, write notes and discuss!

The problem

- Unknown sites, unknown motif: find both!
- Where are the occurrences in a large sequence, e.g.: AGGGCCTGAGGTAAACA
- In reality several large sequences 1k-10k long
- Think, write notes and discuss!

The problem

- Unknown sites, unknown motif: find both!
- Where are the occurrences in a large sequence, e.g.: AGGGCCTGAGGTAAACA
- In reality several large sequences 1k-10k long
- Think, write notes and discuss!

Pretend examples are known!

- Choose or create some random substrings
- Train one kind of model from the random examples
- Determine hits of the model

Pretend examples are known!

- Choose or create some random substrings
- Train one kind of model from the random examples
- Determine hits of the model

Pretend examples are known!

- Choose or create some random substrings
- Train one kind of model from the random examples
- Determine hits of the model

Pretend examples are known!

- Choose or create some random substrings
- Train one kind of model from the random examples
- Determine hits of the model

Over-representation!

- Does the hypothesized motif occur suspiciously often?
- Could generate many sequences and count motif hits. Does it seldom occur as often as in the original sequences?
- What is the calculated probability of seeing this many hits in random sequences?
- If unlikely in random sequences, this could be something functional!

Over-representation!

- Does the hypothesized motif occur suspiciously often?
- Could generate many sequences and count motif hits. Does it seldom occur as often as in the original sequences?
- What is the calculated probability of seeing this many hits in random sequences?
- If unlikely in random sequences, this could be something functional!

Over-representation!

- Does the hypothesized motif occur suspiciously often?
- Could generate many sequences and count motif hits. Does it seldom occur as often as in the original sequences?
- What is the calculated probability of seeing this many hits in random sequences?
- If unlikely in random sequences, this could be something functional!

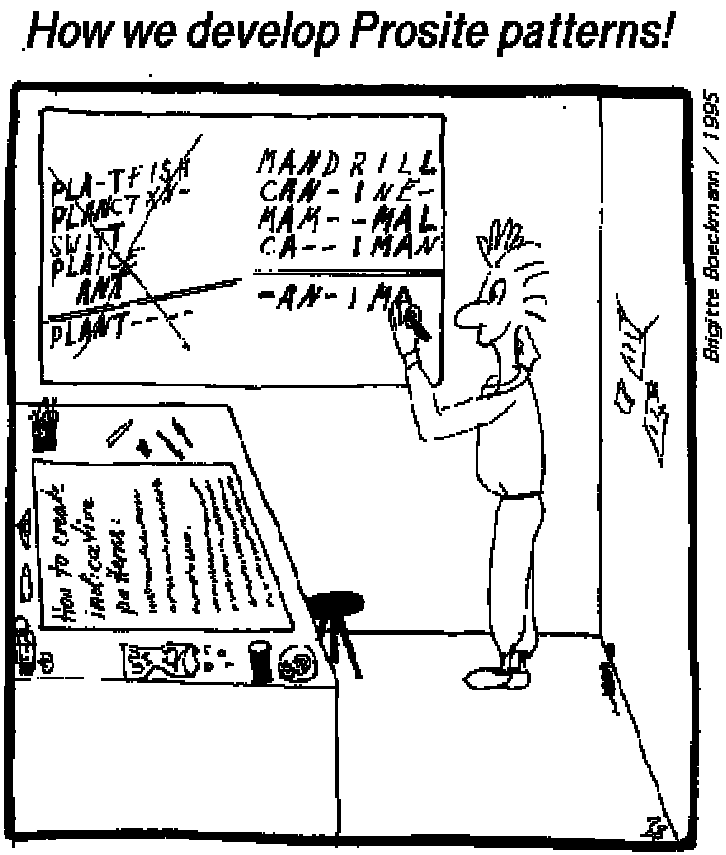
Over-representation!

- Does the hypothesized motif occur suspiciously often?
- Could generate many sequences and count motif hits. Does it seldom occur as often as in the original sequences?
- What is the calculated probability of seeing this many hits in random sequences?
- If unlikely in random sequences, this could be something functional!

Over-representation!

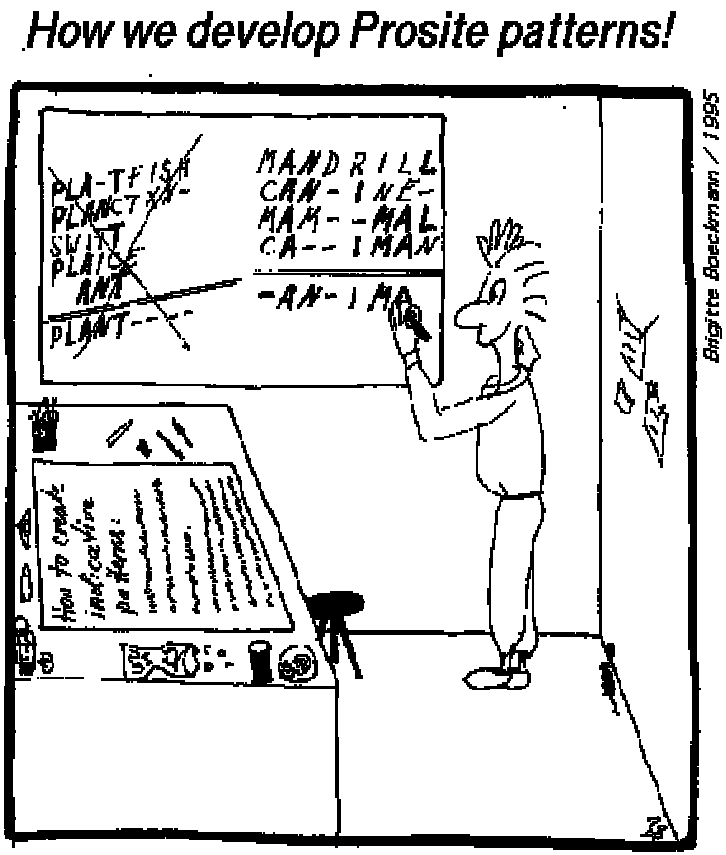
- Does the hypothesized motif occur suspiciously often?
- Could generate many sequences and count motif hits. Does it seldom occur as often as in the original sequences?
- What is the calculated probability of seeing this many hits in random sequences?
- If unlikely in random sequences, this could be something functional!

Enumerate motif candidates



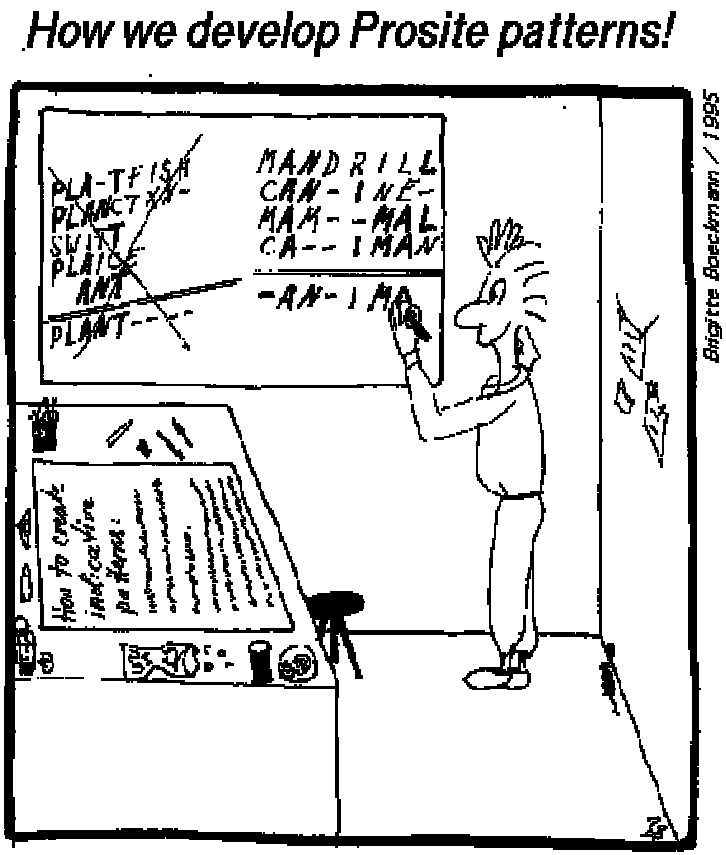
- Could systematically try different possibilities
- Too many possible motifs to exhaustively enumerate
- Search in direction of what looks promising
- PWM: Iterative optimization (EM), MCMC (gibbs sampling)

Enumerate motif candidates



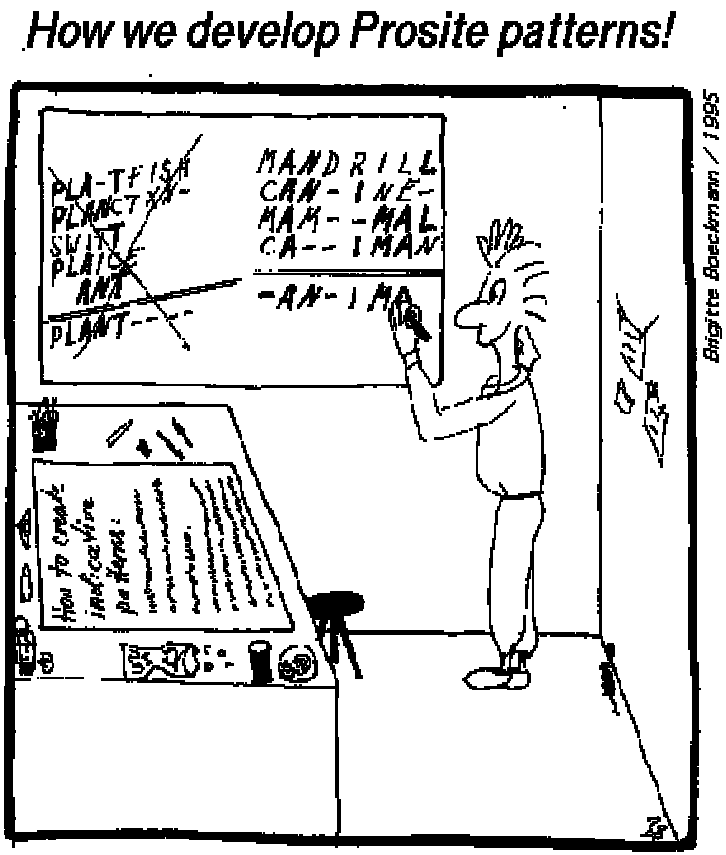
- Could systematically try different possibilities
- Too many possible motifs to exhaustively enumerate
- Search in direction of what looks promising
- PWM: Iterative optimization (EM), MCMC (gibbs sampling)

Enumerate motif candidates



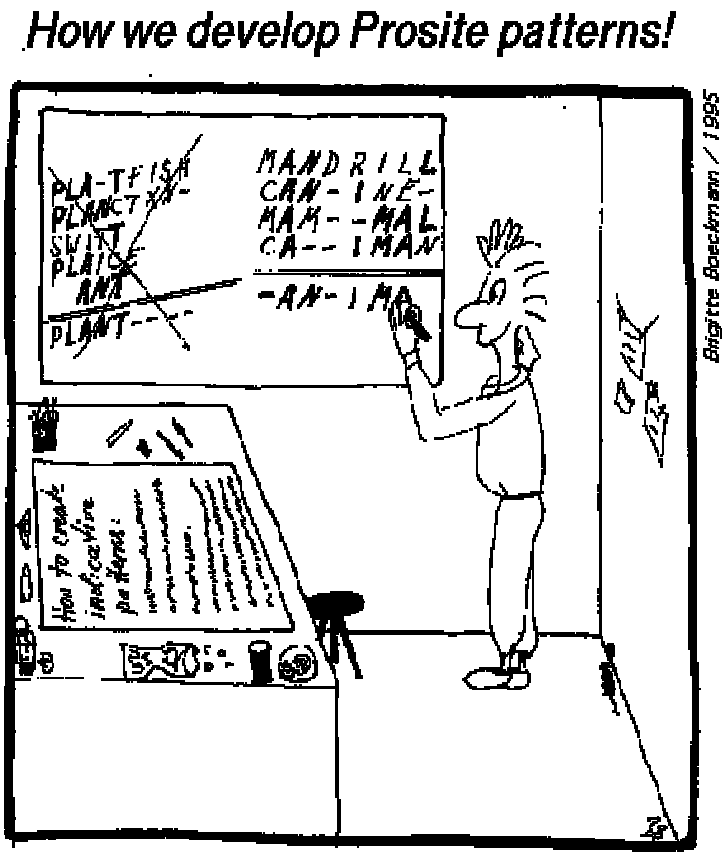
- Could systematically try different possibilities
- Too many possible motifs to exhaustively enumerate
- Search in direction of what looks promising
- PWM: Iterative optimization (EM), MCMC (gibbs sampling)

Enumerate motif candidates



- Could systematically try different possibilities
- Too many possible motifs to exhaustively enumerate
- Search in direction of what looks promising
- PWM: Iterative optimization (EM), MCMC (gibbs sampling)

Enumerate motif candidates



- Could systematically try different possibilities
- Too many possible motifs to exhaustively enumerate
- Search in direction of what looks promising
- PWM: Iterative optimization (EM), MCMC (gibbs sampling)

What is (TFBS) motif discovery?

Computer science perspective:

- Discovery of frequent patterns
- Unknown consensus and positions
- Patterns can be represented in many ways
- Overrepresentation can be calculated in many ways

Biological perspective:

- Discovery of binding sites for transcription factors
- Are often conserved in evolution because of importance
- A single TF bind to similar sites in several regulatory regions

What is (TFBS) motif discovery?

Computer science perspective:

- Discovery of frequent patterns
- Unknown consensus and positions
- Patterns can be represented in many ways
- Overrepresentation can be calculated in many ways

Biological perspective:

- Discovery of binding sites for transcription factors
- Are often conserved in evolution because of importance
- A single TF bind to similar sites in several regulatory regions

What is (TFBS) motif discovery?

Computer science perspective:

- Discovery of frequent patterns
- Unknown consensus and positions
- Patterns can be represented in many ways
- Overrepresentation can be calculated in many ways

Biological perspective:

- Discovery of binding sites for transcription factors
- Are often conserved in evolution because of importance
- A single TF bind to similar sites in several regulatory regions

What is (TFBS) motif discovery?

Computer science perspective:

- Discovery of frequent patterns
- Unknown consensus and positions
- Patterns can be represented in many ways
- Overrepresentation can be calculated in many ways

Biological perspective:

- Discovery of binding sites for transcription factors
- Are often conserved in evolution because of importance
- A single TF bind to similar sites in several regulatory regions

What is (TFBS) motif discovery?

Computer science perspective:

- Discovery of frequent patterns
- Unknown consensus and positions
- Patterns can be represented in many ways
- Overrepresentation can be calculated in many ways

Biological perspective:

- Discovery of binding sites for transcription factors
- Are often conserved in evolution because of importance
- A single TF bind to similar sites in several regulatory regions

What is (TFBS) motif discovery?

Computer science perspective:

- Discovery of frequent patterns
- Unknown consensus and positions
- Patterns can be represented in many ways
- Overrepresentation can be calculated in many ways

Biological perspective:

- Discovery of binding sites for transcription factors
- Are often conserved in evolution because of importance
- A single TF bind to similar sites in several regulatory regions

What is (TFBS) motif discovery?

Computer science perspective:

- Discovery of frequent patterns
- Unknown consensus and positions
- Patterns can be represented in many ways
- Overrepresentation can be calculated in many ways

Biological perspective:

- Discovery of binding sites for transcription factors
- Are often conserved in evolution because of importance
- A single TF bind to similar sites in several regulatory regions

What is (TFBS) motif discovery?

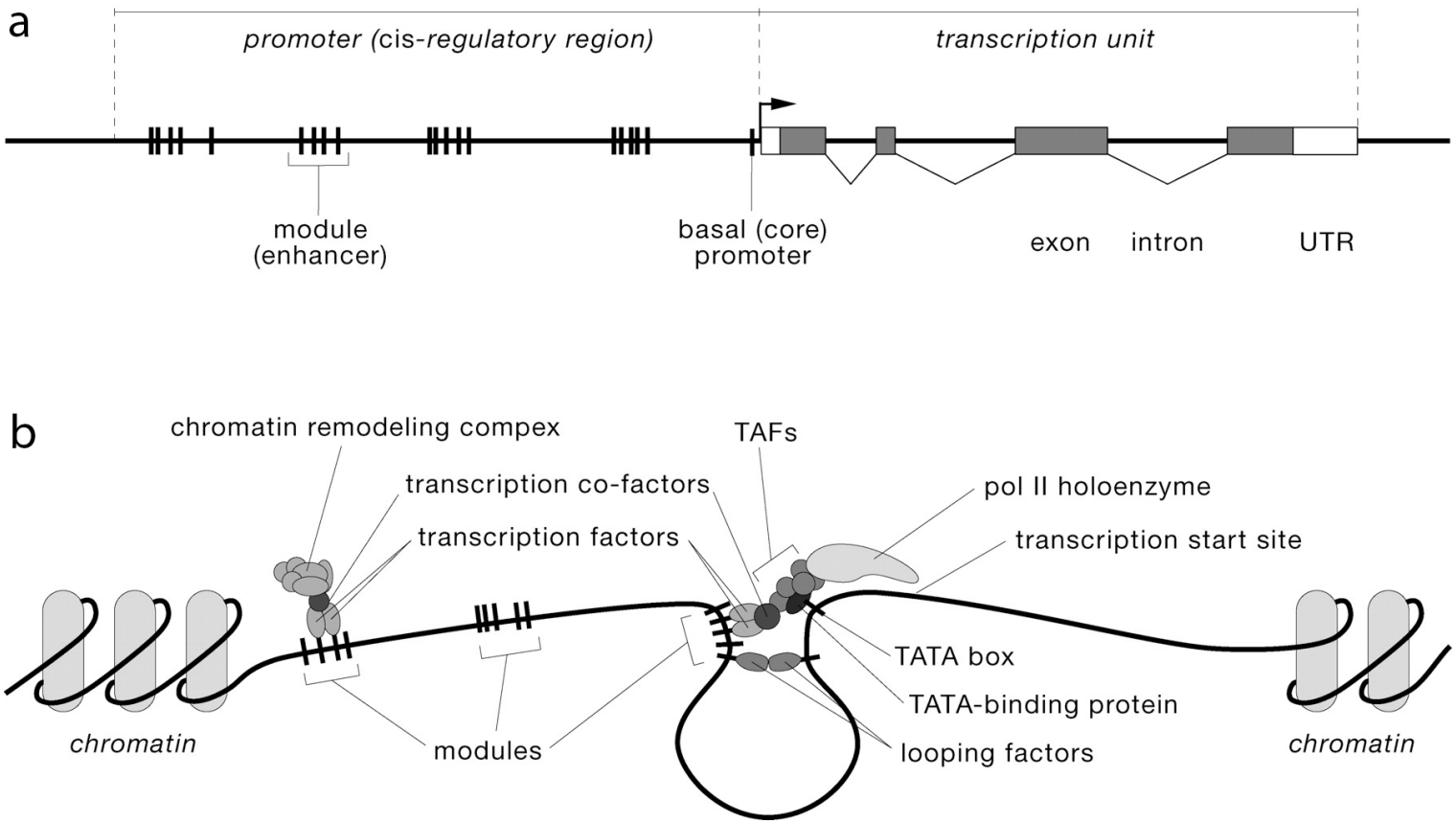
Computer science perspective:

- Discovery of frequent patterns
- Unknown consensus and positions
- Patterns can be represented in many ways
- Overrepresentation can be calculated in many ways

Biological perspective:

- Discovery of binding sites for transcription factors
- Are often conserved in evolution because of importance
- A single TF bind to similar sites in several regulatory regions

Regulatory regions



Wray et al. (2003)

Use cases

- **TF of interest, find target genes:**
  Library profile -> PWM scanning
- **Co-expressed genes, find regulators:**
  All library profiles -> PWM scanning
- **Co-expressed genes, novel regulators:**
  Motif discovery
- **Co-expressed genes, find co-regulated subset:**
  Motif discovery

Use cases

- **TF of interest, find target genes:**
  Library profile -> PWM scanning
- **Co-expressed genes, find regulators:**
  All library profiles -> PWM scanning
- **Co-expressed genes, novel regulators:**
  Motif discovery
- **Co-expressed genes, find co-regulated subset:**
  Motif discovery

Use cases

- **TF of interest, find target genes:**
  Library profile -> PWM scanning
- **Co-expressed genes, find regulators:**
  All library profiles -> PWM scanning
- **Co-expressed genes, novel regulators:**
  Motif discovery
- **Co-expressed genes, find co-regulated subset:**
  Motif discovery

Use cases

- **TF of interest, find target genes:**
  Library profile -> PWM scanning
- **Co-expressed genes, find regulators:**
  All library profiles -> PWM scanning
- **Co-expressed genes, novel regulators:**
  Motif discovery
- **Co-expressed genes, find co-regulated subset:**
  Motif discovery

Use cases

- **TF of interest, find target genes:**
  Library profile -> PWM scanning
- **Co-expressed genes, find regulators:**
  All library profiles -> PWM scanning
- **Co-expressed genes, novel regulators:**
  Motif discovery
- **Co-expressed genes, find co-regulated subset:**
  Motif discovery

Use cases

- **ChIP-data, find precise locations:**
  Motif discovery
- **ChIP-data, complement findings:**
  Build PWM -> G.W. PWM scanning
- **ChIP-data, find co-regulators:**
  CRM discovery (not discussed here..)
- **HCNR-regions, find regulatory regions:**
  CRM discovery

Use cases

- **ChIP-data, find precise locations:**
  Motif discovery
- **ChIP-data, complement findings:**
  Build PWM -> G.W. PWM scanning
- **ChIP-data, find co-regulators:**
  CRM discovery (not discussed here..)
- **HCNR-regions, find regulatory regions:**
  CRM discovery

Use cases

- **ChIP-data, find precise locations:**
  Motif discovery
- **ChIP-data, complement findings:**
  Build PWM -> G.W. PWM scanning
- **ChIP-data, find co-regulators:**
  CRM discovery (not discussed here..)
- **HCNR-regions, find regulatory regions:**
  CRM discovery

Use cases

- **ChIP-data, find precise locations:**
  Motif discovery
- **ChIP-data, complement findings:**
  Build PWM -> G.W. PWM scanning
- **ChIP-data, find co-regulators:**
  CRM discovery (not discussed here..)
- **HCNR-regions, find regulatory regions:**
  CRM discovery

Use cases

- **ChIP-data, find precise locations:**
  Motif discovery
- **ChIP-data, complement findings:**
  Build PWM -> G.W. PWM scanning
- **ChIP-data, find co-regulators:**
  CRM discovery (not discussed here..)
- **HCNR-regions, find regulatory regions:**
  CRM discovery