Bayesian modeling and MCMC algorithms for analysis of bacterial ~~mRNA expression data and~~ promoter sequences

Pierre Nicolas

Mathématiques et Informatique Appliquées du Génome à l'Environnement
INRA Jouy-en-Josas, Université Paris-Saclay

AppliBUGS – AgroParisTech – 13 décembre 2018

Outline

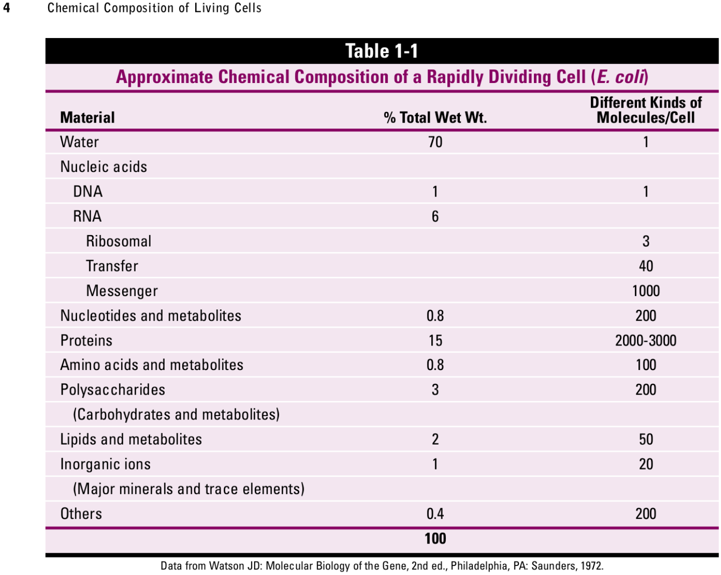
1. Biological context and motivation
2. Model and algorithm for sigma factor binding sites
3. Model and algorithm for transcription factor binding sites

1. Biological context and motivation

2. Model and algorithm for sigma factor binding sites

3. Model and algorithm for transcription factor binding sites

Molecular composition of a bacterial cell



4 Chemical Composition of Living Cells

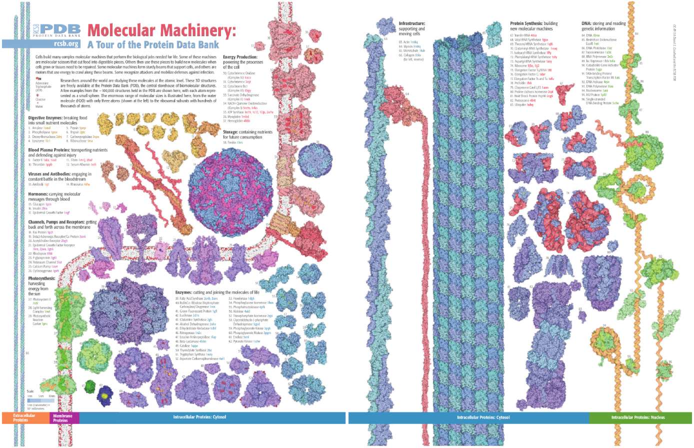
Table 1-1

Approximate Chemical Composition of a Rapidly Dividing Cell (*E. coli*)

Material	% Total Wet Wt.	Different Kinds of Molecules/Cell
Water	70	1
Nucleic acids		
DNA	1	1
RNA	6	
Ribosomal		3
Transfer		40
Messenger		1000
Nucleotides and metabolites	0.8	200
Proteins	15	2000-3000
Amino acids and metabolites	0.8	100
Polysaccharides	3	200
(Carbohydrates and metabolites)		
Lipids and metabolites	2	50
Inorganic ions	1	20
(Major minerals and trace elements)		
Others	0.4	200
	100	

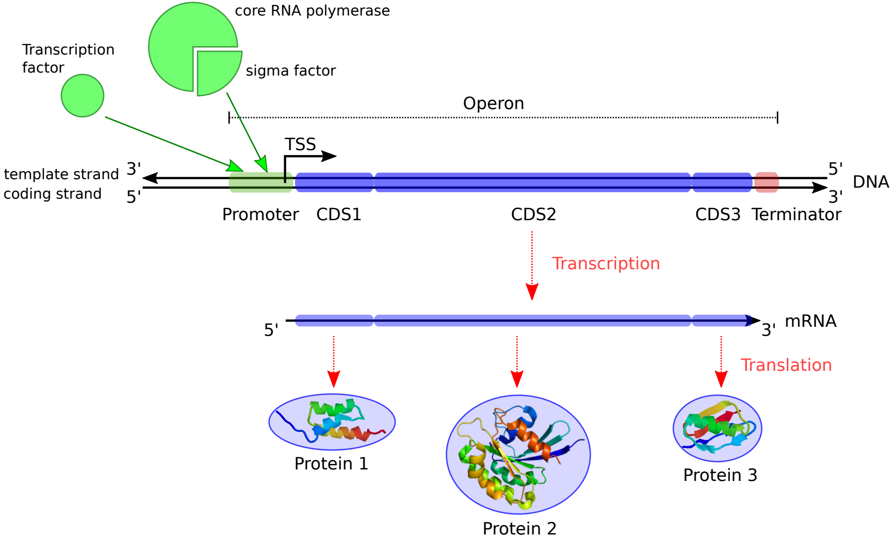
Data from Watson JD: Molecular Biology of the Gene, 2nd ed., Philadelphia, PA: Saunders, 1972.

Proteins: catalytic, signaling, structural roles

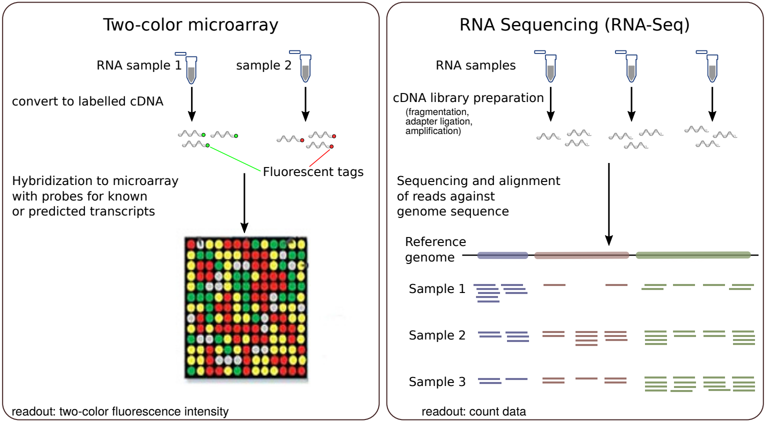

RCSB PDB PROTEIN DATA BANK
rcsb.org
Molecular Machinery:
A Tour of the Protein Data Bank
Extracellular Proteins
Membrane Proteins
Intracellular Proteins: Cytosol
Intracellular Proteins: Cytosol
Intracellular Proteins: Nucleus

Multiply and survive through the different environmental conditions → express the genetic information by producing ribonucleic acid (RNA) molecules in the right time and right amount.

Transcriptional activity in bacteria

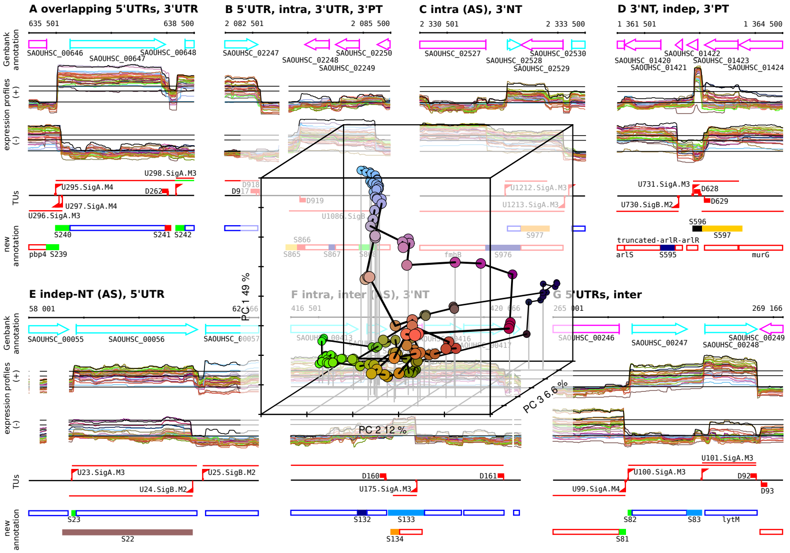


Transcriptomics: from two-color microarrays to RNA-Seq



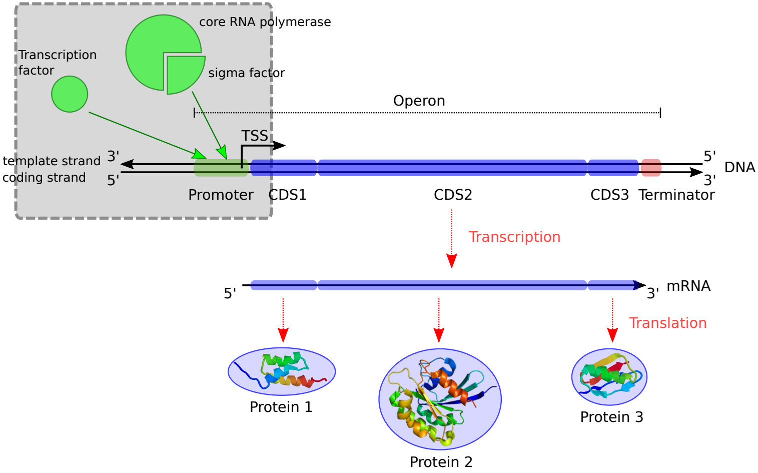
- Despite considerable improvements in the microarray technology until ≈2010, RNA-Seq progressively replaced the microarrays during the last ten years with the development of high-throughput sequencing.
- Dedicated RNA-Seq protocols can be used to map TSS at 1-bp resolution.

Data on condition-dependent transcriptomes



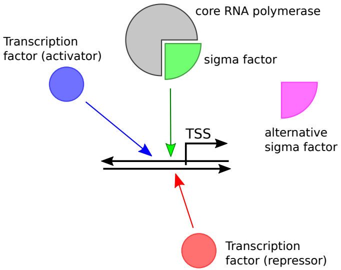
Data from Mäder *et al.*, 2016

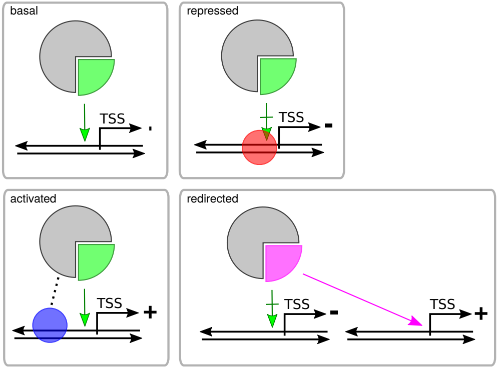
The focus of this work: analysis of promoter regions



Much of the regulation of transcription takes place in the promoter region and involves the binding of proteins to DNA.

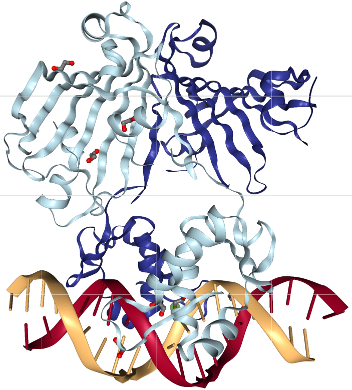
Transcriptional regulation mediated by protein-DNA interactions





When present and/or activated sigma factors and transcription factors bind to specific sites of the DNA sequence and modulate transcription initiation rate.

Protein-DNA interaction involves recognition of sequence motifs



5'-CAGTGGTCTAGACCACTG-3'
3'-GTCACCAGATCTGGTGAC-5'

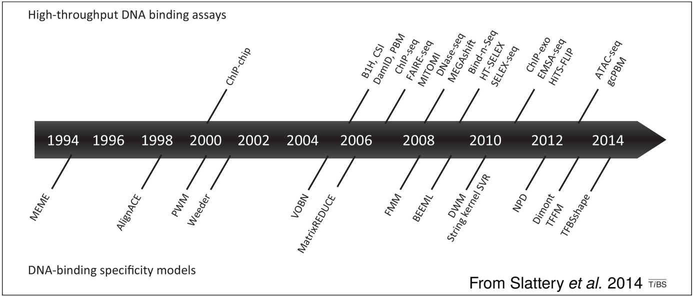
Sigma factors and transcription factors recognize sequence patterns (motifs) that we would like to discover based on statistical over-representation in DNA sequences.

Methodological approaches to regulatory motif discovery

One of the oldest and most studied problem in computational biology . . .

- Motif representation : words, regular expressions, position weight matrices
- Statistical methodology : null-hypothesis testing vs. modeling of motif occurrences
- Unsupervised vs. supervised approaches (discriminative learning based on positive and negative data sets)

Much effort put in the use of data from high-throughput DNA binding assays



From Slattery *et al.* 2014 TIBS

Our goal and methodological framework

Goal: develop new approaches to discover regulatory motifs in bacterial promoters by making full use of

- statistical properties of the sequence composition (over-representation wrt background model)
- expression data sets exploring the diversity of lifestyles (and possibly mutants)
- knowledge of the precise position of the TSSs

Integrative probabilistic modelling

$$\begin{array}{ccc} \text{promoter DNA sequence} & & \text{mRNA levels} \\ X & \rightarrow & Y \\ X & \leftarrow & Y \end{array}$$

Our methodological framework

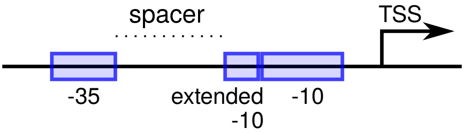
- Modelling $\pi(X|Y)$ instead of $\pi(Y|X)$ or $\pi(X, Y)$
- Bayesian inference and MCMC algorithms

1. Biological context and motivation

2. Model and algorithm for sigma factor binding sites

3. Model and algorithm for transcription factor binding sites

Sigma factor binding sites



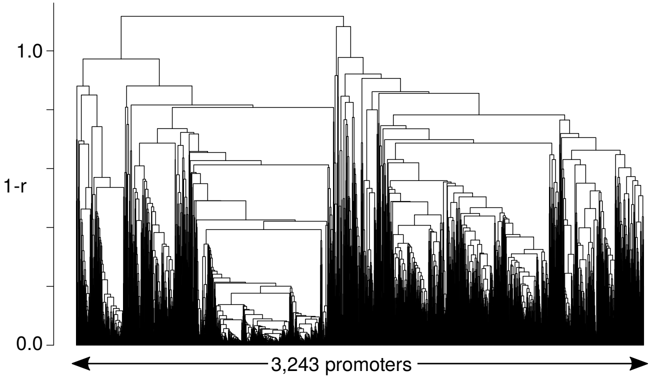
Motifs

- Sigma factors recognize degenerate motifs located directly upstream the TSS (-10 and -35 boxes)
- Each promoter contains “by definition” one Sigma factor binding site
- Sigma factors partition the promoter space $\leftrightarrow$ first level of regulation

Data (Nicolas *et al.*, 2012)

- Sequences: 3,243 promoter regions of length $L = 101$ *bp*
  aligned wrt TSS (-60,+40) as determined by upshifts in expression signal along the chromosome
- Expression matrix: 3,243 promoters × 269 conditions

Expression data across conditions summarized in correlation tree



Relevant for the search for sigma factor binding sites since first level of regulation

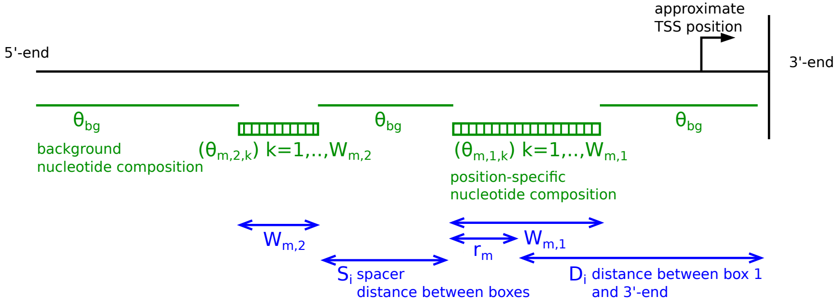
Statistical model aims to combine

- hierchical classification of correlations between expression profiles
- DNA sequence information

Sequence model

Mixture model for sequence data (adapted from Nicolas *et al.*, 2006)

- Each type of motif has its own probability $\pi(M_i = m) = \alpha_m$, $\sum_m \alpha_{m=1}^{M} = 1$
- $\pi(X_i \mid M_i = m)$, proba. of sequence $X_i \in \{\text{a,c,g,t}\}^L$ given the presence of a motif of type $M_i = m$
- given motif type M_i and position (S_i, D_i), X_i is modeled as four-state inhomogeneous Markov chain



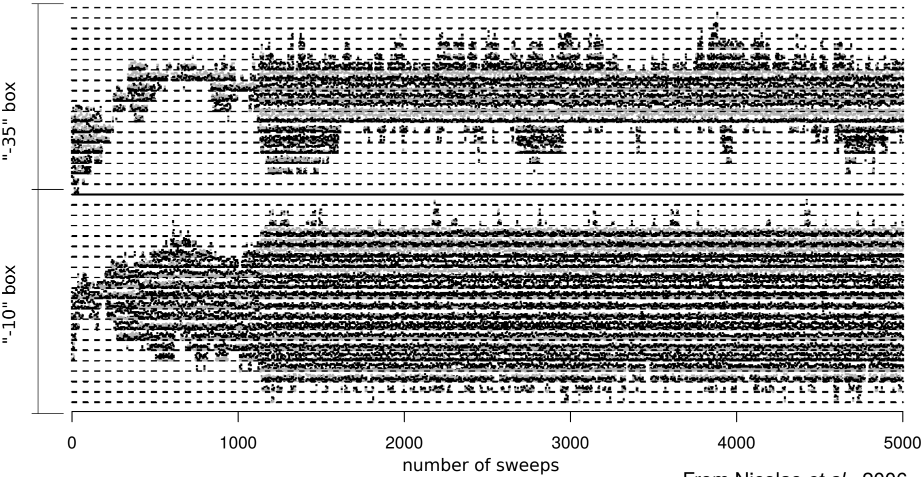
approximate TSS position
5'-end
3'-end
θ_{bg}
background nucleotide composition
(θ_{m,2,k}) k=1,..,W_{m,2}
θ_bg
(θ_{m,1,k}) k=1,..,W_{m,1}
position-specific nucleotide composition
θ_bg
W_{m,2}
W_{m,1}
r_m
S_i spacer distance between boxes
D_i distance between box 1 and 3'-end

Motif discovery

- unsupervised estimation of model parameters (transdimensional MCMC)
- computation of $\pi(M_i = m \mid x_i) \propto \pi(x_i \mid M_i = m)\alpha_m$

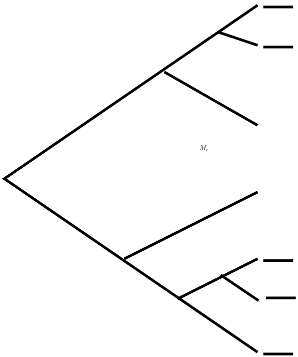
Estimation of box width using MCMC (one type of motif)

Joint sampling of motif occurrence positions $(S_i, D_i, M_i)_{i=1 \dots I}$ and model parameters $(\alpha, W_{\text{box1}}, W_{\text{box2}}, \theta_{\text{box1}}, \theta_{\text{box2}}, \theta_{\text{bg}}, \dots)$



From Nicolas *et al.*, 2006

Incorporating information from promoter correlation tree



M_1

M_2

λ rate of motif switches per unit of branch length

δ probability of motif switch at the leaf level

$\alpha=(\alpha_\varnothing,\alpha_1,\ldots,\alpha_M)$ probability of each motif after a switch

$M_i\in\{\varnothing,1,\ldots,M\}$ motif associated with sequence i

Motif allocation $(M_1, M_2, \ldots, M_I)$ is no longer iid. Instead it results from an “evolution” process along the branches of the tree.

Remarks

- add only two parameters wrt classical iid mixture (λ et δ)
- degenerate to iid mixture when $\lambda \to +\infty$ or $\delta \to 1$
  $\hookrightarrow$ model for motif discovery with or without expression data

MCMC algorithm

In brief

- all parameters estimated simultaneously
- parameters and latent variables updated by blocks (gibbs-type and MH-type moves)
- transdimensional moves (Reversible Jump MH moves) are implemented to accommodate incremental changes of dimension (eg width of position weight matrices).

Inward-outward (or upward-downward) recursion in tree to update simultaneously $(M_i)_{i=1\dots I}$, where I is the number of leafs (promoters).

- Sort by height the $I-1$ internal nodes of the tree.
- Node i has
  - height h_i
  - left and right children $l(i)$ and $r(i)$
  - parent $p(i)$
  - vector of indexes for the leafs of the subtree $s(i)$
- Latent variables $(\tilde{M}_i)_{i=1:(2I-1)}$ record the hidden type associated to each node i (leaves and internal nodes).

Inward recursion

Inward (upward) recursion consists of computing $\pi(x_{s(i)} \mid \tilde{m}_i)$ for $\tilde{m}_i \in \{1, \ldots, \mathcal{M}\}$ et i de 1 à $2I - 1$.

In practice

- For $i = 1 \ldots I$ (ie the leaves for which $s(i) = i$), compute

$$\pi(x_i \mid \tilde{m}_i) \quad = \quad (1-\delta)\pi(x_i \mid M_i = \tilde{m}_i) + \delta \sum_m \alpha_m \pi(x_i \mid M_i = m)\,,$$

  where $\pi(x_i \mid M_i = m) = \sum_{d,s} \pi(x_i, D_i = d, S_i = s \mid M_i = m)$ has been computed for all possibles motif positions (D_i, S_i)

- For $i = I + 1 \ldots 2I - 1$ (internal nodes), compute

$$\begin{aligned} &\pi(x_{s(i)} \mid \tilde{m}_i) \\ &\quad = \prod_{j \in \{l(i), r(i)\}} \Big\{ e^{-(h_i - h_j)\lambda} \pi(x_{s(j)} \mid \tilde{M}_j = \tilde{m}_i) \\ &\qquad\qquad + (1 - e^{-(h_i - h_j)\lambda}) \sum_m \alpha_m \pi(x_{s(j)} \mid \tilde{M}_j = m) \Big\}\,, \end{aligned}$$

Outward recursion

At root node $r = 2I - 1$, start the outward recursion that consists in sampling from the joint distribution of $(\tilde{M}_i)_{i=1:2I-1}, (M_i)_{i=1:I})$ given $x = (x_i)_{i=1:I}$ $(x = x_{s(r)})$.

Conditional independence properties makes that it is enough for this to sample $\tilde{M}_i$ given $\tilde{M}_{p(i)}$ and $x_{s(i)}$.

In practice

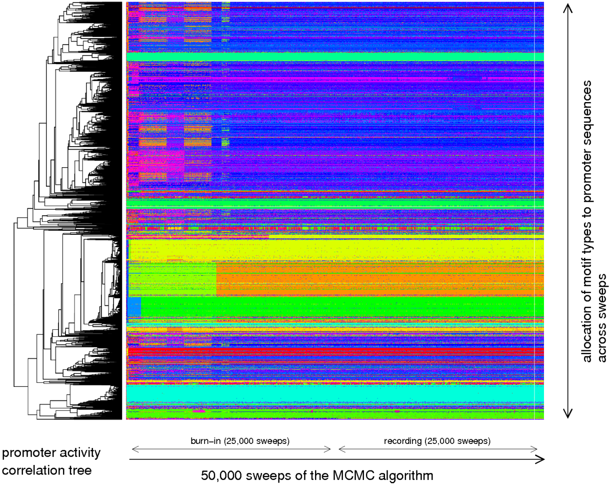
- For $i = r = 2I - 1$, draw $\tilde{m}_r$ from $\pi(\tilde{m}_r \mid x) \propto \alpha_m \pi(x_{s(r)} \mid \tilde{m}_r)$.
- For $i = 2I - 2 \dots 1$ draw $\tilde{m}_i$ frow

$$\begin{aligned} &\pi(\tilde{m}_i \mid x, \tilde{m}_{p(i)}) \\ &\quad \propto \quad \pi(x_{s(i)} \mid \tilde{m}_i) \times \left[e^{-(h_{p(i)}-h_i)\lambda} \mathbb{I}\{\tilde{m}_{p(i)} = \tilde{m}_i\} + \alpha_m(1 - e^{-(h_{p(i)}-h_i)\lambda})\right]. \end{aligned}$$

- For $i = I \dots 1$ draw m_i from $\pi(m_i \mid x, \tilde{m}_i) \propto \pi(x_i \mid m_i) \times \left[(1-\delta)\mathbb{I}\{\tilde{m}_i = m\} + \alpha_m \delta\right]$.

The values of the rv D_i and S_i that indicates the exact position of the motif M_i in the promoter region i are then drawn given m_i.

Behaviour of the MCMC algorithm, $M = 20$ motifs



promoter activity correlation tree

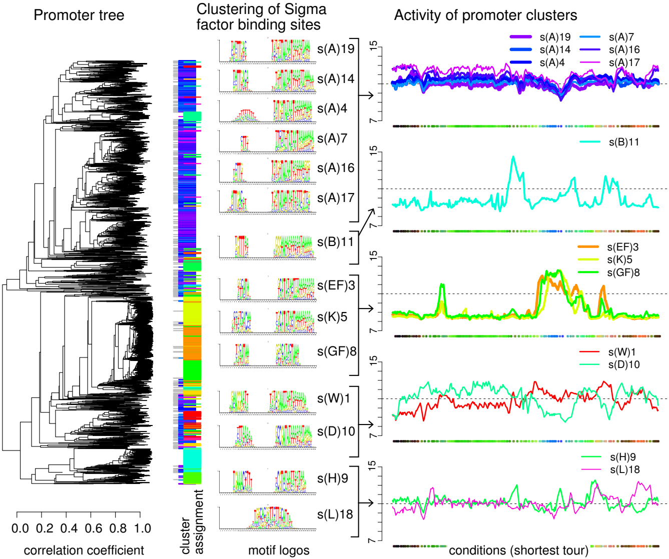
burn–in (25,000 sweeps)

recording (25,000 sweeps)

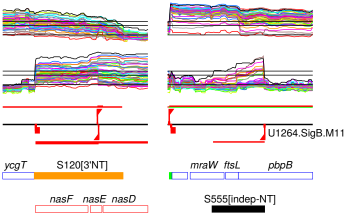
50,000 sweeps of the MCMC algorithm

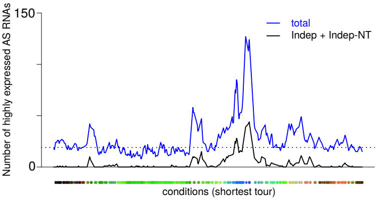
allocation of motif types to promoter sequences across sweeps

Back to biology: activity of sigma factors across conditions



Back to biology: origin of antisense RNAs





On a total of 423 antisense RNAs (*B. subtilis*), 82% linked either to activity of alternative sigma factors or to incomplete terminations:

- 48% under control of alternative sigma factors (up to 77% for those with their own promoters),
- 62% in contexts on incomplete termination.

↪ Hypothesis (Nicolas *et al.*, 2012) : transcriptional noise

sites recognized by alternative sigma factors appear and sites of transcription termination disappear at random during evolution

Hypothesis that received further support from results obtained on another bacterium (*S. aureus*, Mäder *et al.*, 2016):
less alternative sigma factors → less antisense RNAs with their own promoters

1. Biological context and motivation

2. Model and algorithm for sigma factor binding sites

3. Model and algorithm for transcription factor binding sites

Transcription factor (TF) vs. sigma factor binding sites

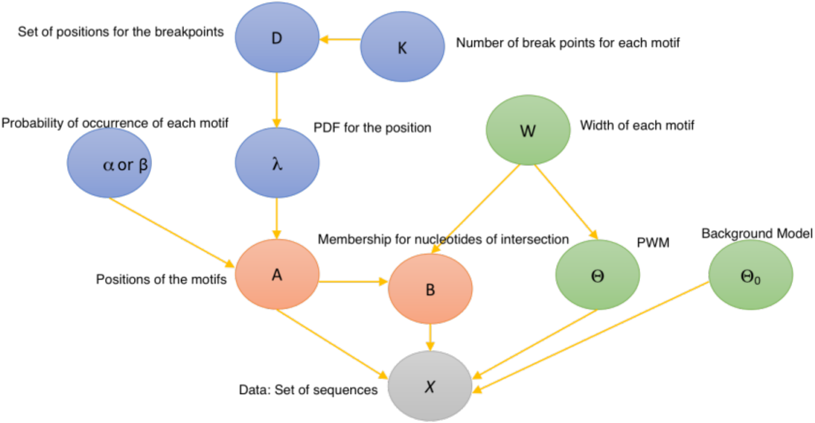
Same aim as for Sigma factor binding sites: discover motifs recognized by the different TF by simultaneous analysis of all the promoter regions, making full use of sequence data, expression data, exact position of the TSSs.

- in each promoter region any number of TF binding sites (0 → many) vs. one sigma factor binding site
- TF regulons (set of regulated genes) partially overlap → no clear hierarchy
- TF binding sites can overlap
- TF binding sites can exhibit any type of positional preference wrt TSS
- impact of TF on expression levels is more subtle than Sigma factors (can be activator or repressor)
- TF binding motifs can be represented by a single box (in first approx.), they are often palindromic

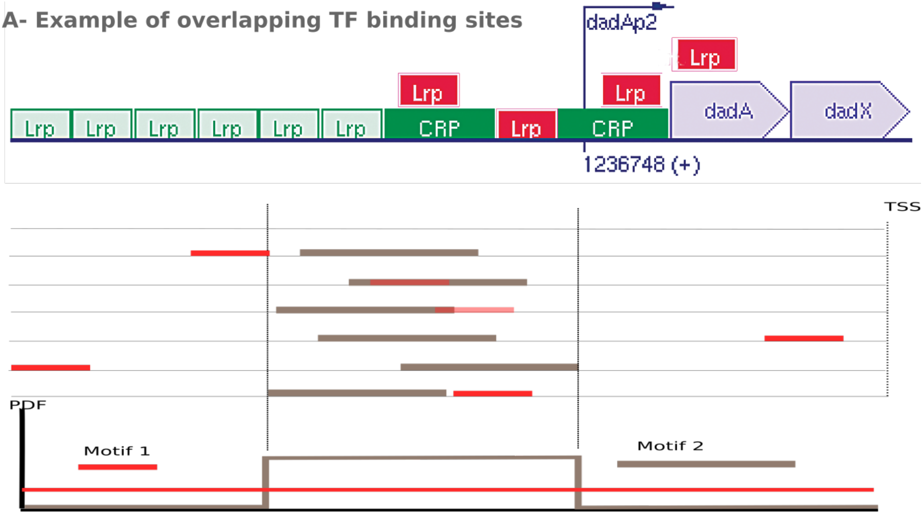
↪ Need for a different statistical model for the

- sequence given the presence of some motifs
- occurrence of the different motifs given expression data

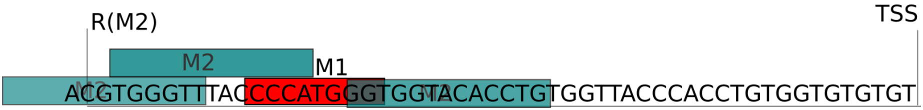
Sequence model for promoter regions with TF binding motifs


Set of positions for the breakpoints
D
K
Number of break points for each motif
Probability of occurrence of each motif
PDF for the position
W
Width of each motif
α or β
λ
Membership for nucleotides of intersection
PWM
Background Model
Positions of the motifs
A
B
Θ
Θ_0
Data: Set of sequences
X

Motif occurrences are allowed to overlap



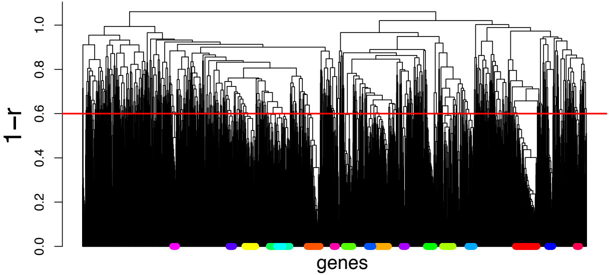
Sequence model with overlapping motif occurrences

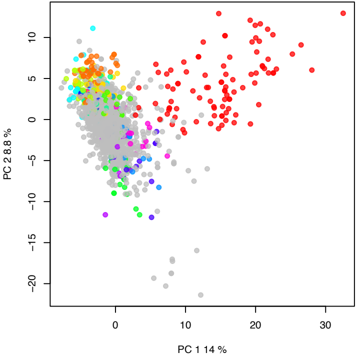


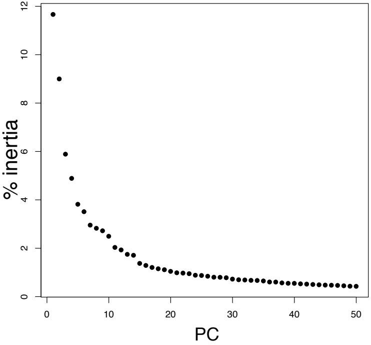
$$\pi(x_n|A,\theta,\theta_0,r,W)$$
$$= \left[\prod_{\{l\in M_n\}} \theta_{m,l-a_{m,n}+r,x_{n,l}}\right]\left[\prod_{\{l\in Bg\}} \theta_{0,x_{n,l}}\right]\left[\prod_{\{l\in O\}} \frac{1}{|O(l)|}\sum_{m\in O(l)} \theta_{m,l-a_{m,n}+r,x_{n,i}}\right]$$

- realistic from a biological perspective? (not necessarily)
- simple
- possible to model motif occurrences as independent random variables
- avoid hard constraints on positions of occurrences
  → easier for updating motif positions and motif width (no collision)

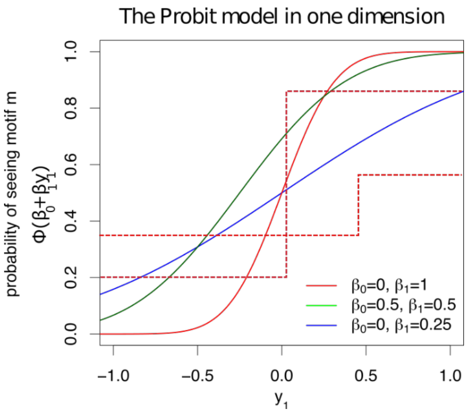
Summarizing expression data (here 1,512 promoters × 165 conditions)







Modified probit regression framework to incorporate expression data *y*

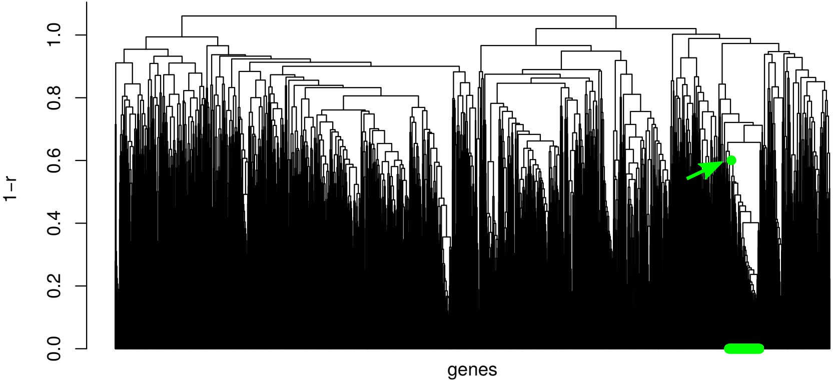


Probability of occurrence of motif m in sequence n, $\pi(a_{m,n} > 0 | y, t, \beta, b, c)$, written

$$\Phi\Big(\beta_{m,0} + \sum_{c} \beta_{m,c}\mathbb{I}_{\{t_{m,c}=1\}}\big[y_{n,c}\mathbb{I}_{\{b_{m,c}=0\}} + \mathbb{I}_{\{b_{m,c}=1\}}[(\gamma_{m,c}-1)\mathbb{I}_{\{y_{m,c}\le y_{[c_{m,c}],c}\}} + \gamma_{m,c}\mathbb{I}_{\{y_{m,c}>y_{[c_{m,c}],c}\}}]\big]\Big)$$

Data augmentation for Bayesian inference of probit model (Gaussian rv $Z_{m,n}$)

Modified probit regression on a tree



$$
\begin{aligned}
&\pi(a_{m,n} > 0 | y, t, \beta, b, c) \\
&\quad = \quad \Phi\Big(\beta_{m,0} + \sum_{c} \beta_{m,c}\mathbb{I}_{\{t_{m,c}=1\}}\big[y_{n,c}\mathbb{I}_{\{b_{m,c}=0\}} \\
&\qquad\qquad + \mathbb{I}_{\{b_{m,c}=1\}}[(\gamma_{m,c} - 1)\mathbb{I}_{\{y_{m,c} \le y_{[c_{m,c}],c}\}} + \gamma_{m,c}\mathbb{I}_{\{y_{m,c} > y_{[c_{m,c}],c}\}}]\big]\Big)
\end{aligned}
$$

Importantly: algorithm in $O(N)$ to sample the position of the cut in the tree given $Z_{m,\cdot}$.

Application on real data (set-up)

Data set

- *Listeria monocytogenes*
- Number of promoter sequences N = 1,512
- Original dimension of the transcriptome data-set 1,512×165 (Bécavin *et al.*, 2017)
- C = 50 covariates were defined using PCA, ICA, and hierarchical clustering

Model and algorithm

- Number of motifs searched for is M = 75
- Model and algorithm allow soft transitions between normal and palindromic motif
- MCMC algorithm was run 10 times for 50,000 sweeps (burn-in 10,000)
- Postprocessing (clustering) to identify stable motifs

Example of motif identified with the algorithm

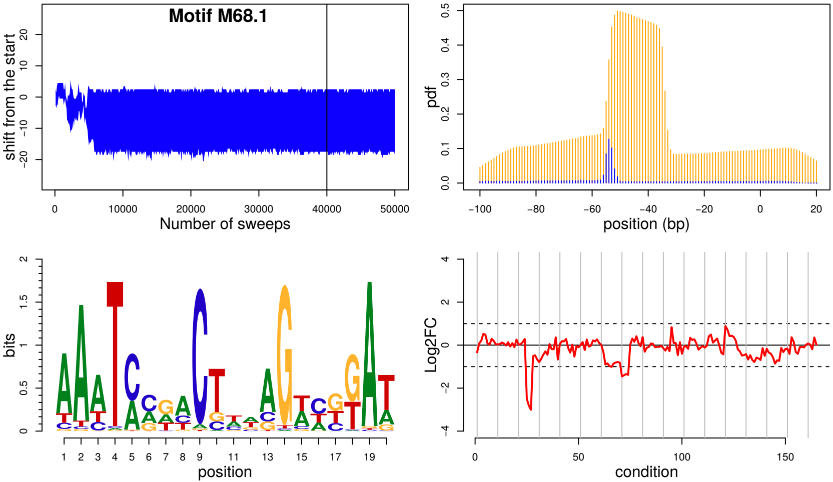
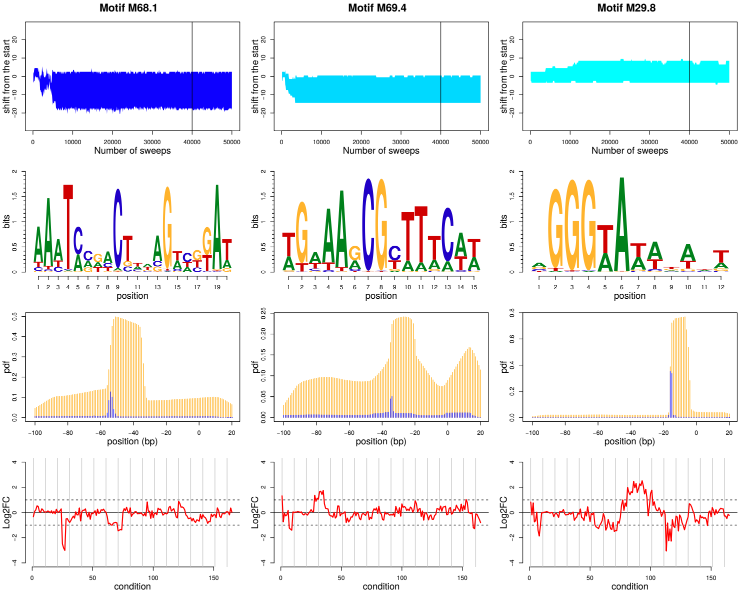


Illustration of the diversity of motifs



Acknowledgments

Ibrahim Sultan who developed the model for TF binding site discovery based on probit regression (co-supervised by Sophie Schbath).

Main collaborators involved over the years in experimental aspects linked to this work

- INRA Micalis: Elena Bidnenko, Étienne Dervyn, Philippe Noirot (*Bacillus subtilis*)
- INRA VIM: Tatiana Rochat (*Flavobacterium psychrophilum*)
- U. Greifswald: Ulrike Mäder (*Staphylococcus aureus* and *Bacillus subtilis*)

Funding: EU Basysbio, EU List_Maps