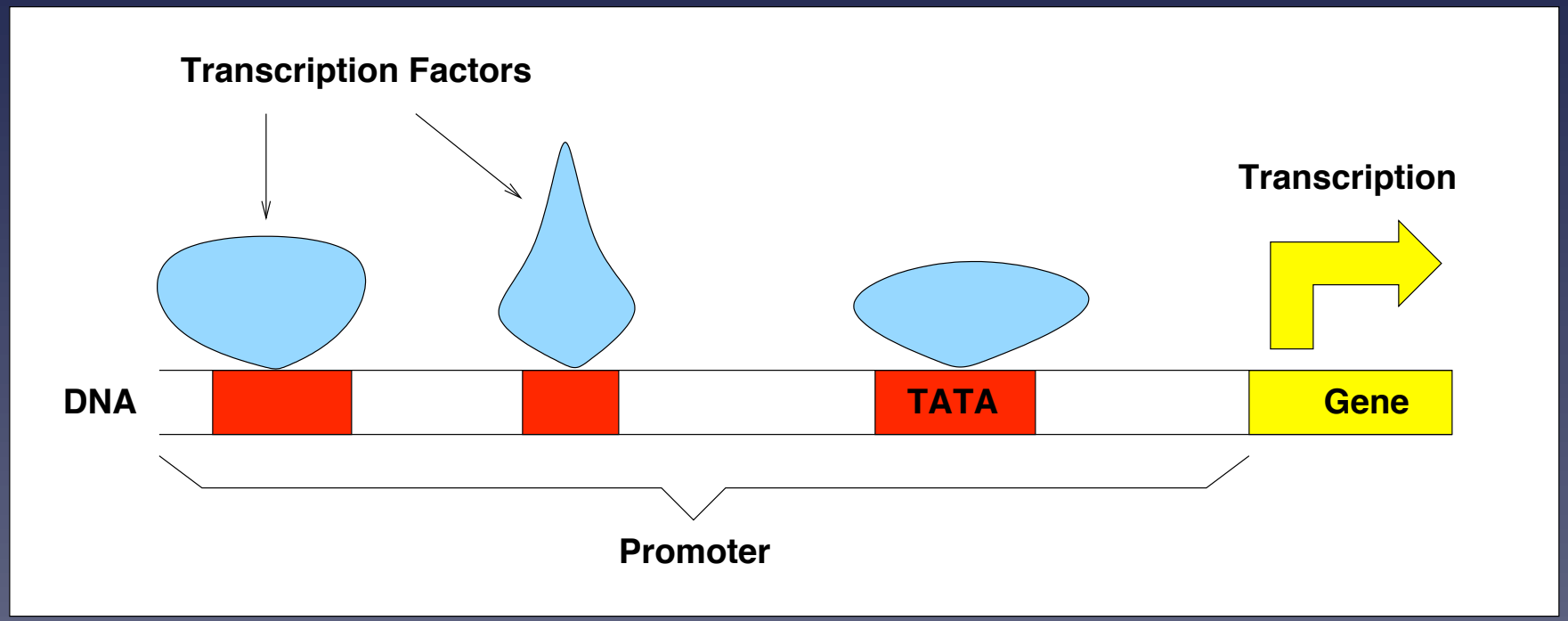
Motif finding in promoter regions with support vector machines

Jean-Philippe Vert
Ecole des Mines de Paris, Centre for Computational Biology
Jean-Philippe.Vert@ensmp.fr

Robert Thurman, William S. Noble
University of Washington, Genome Science
{rthurman,noble}@gs.washington.edu

Journées thématiques Ouest Génopole "Promoteurs", IRISA/INRIA, Rennes, France, January 11th, 2007.

Background


Transcription Factors
Transcription
DNA
TATA
Gene
Promoter

Motivation

- How to find functional motifs in promoter sequences?
- How to assign them a function?

Motivation

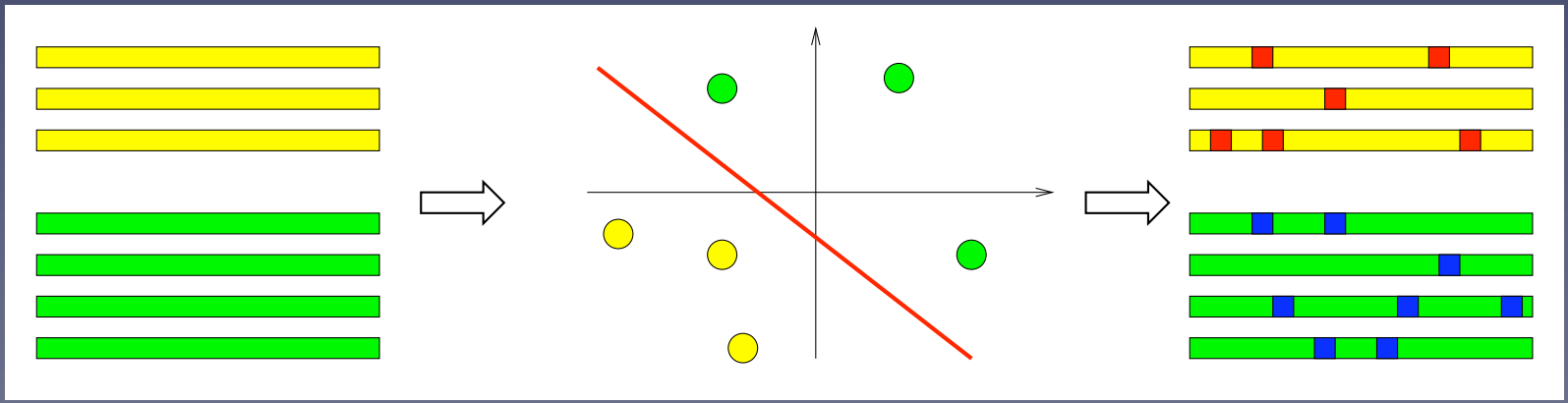
- How to find functional motifs in promoter sequences?
- How to assign them a function?

Related work

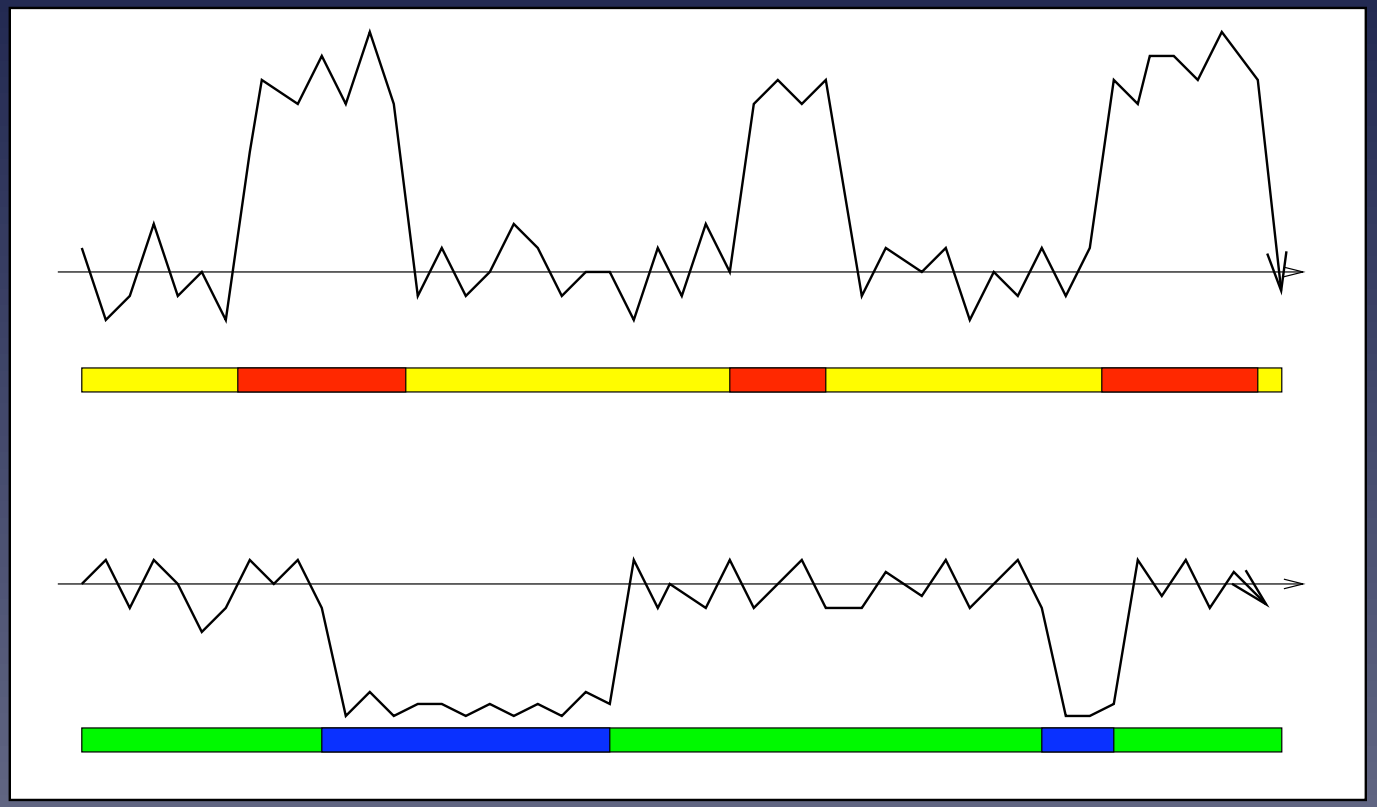
- Motif detection by EM, Gibbs sampling,...
- Detection of over- or under-represented oligomers

Overview of our approach

1. Take 2 sets of genes with different properties (e.g., expressed at 2 different times in the cell cycle)
2. Create a discriminative model using the promoter regions
3. Extract discriminative motifs from the model



From discriminative models to motifs



We need a discriminative score additive along the sequence.

Additive scores

- Chose a Euclidean feature space $\mathcal{H} = \mathbb{R}^N$
- For a given sequence x, map each position i to a vector $\phi(x, i)$ in $\mathcal{H}$
- Represent x by the sum $\phi(x) = \sum_i \phi(x, i) \in \mathcal{H}$
- Create a linear discriminative models in $\mathcal{H}$:

$$f(x) = w \cdot \phi(x) + b = \sum_i \underbrace{(w \cdot \phi(x, i))}_{s(i)} + b$$

Main contributions

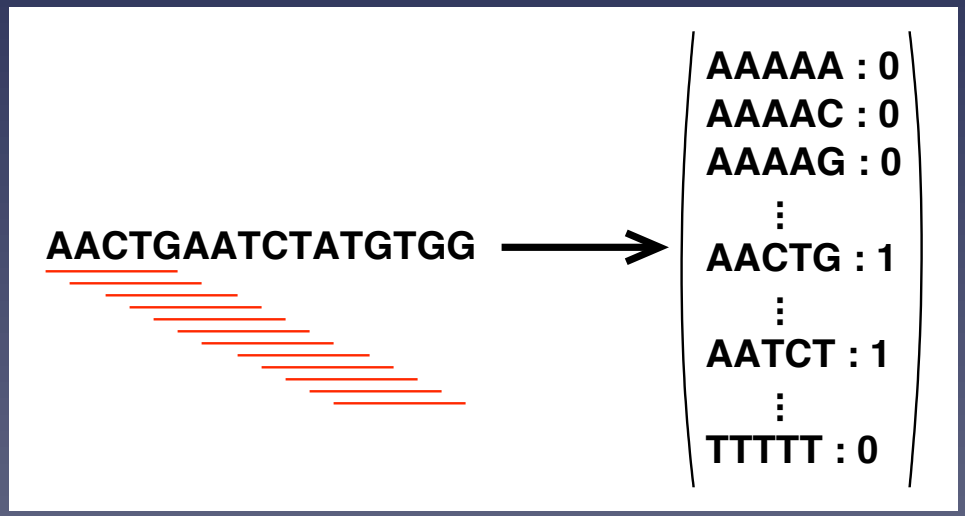
- Chose $\mathcal{H} = \mathbb{R}^{1024}$, indexed by all 5-mers of nucleotides (AAAAA, AAAAC, ... , TTTTT)
- Use a linear SVM as a discriminative model
- Test several mappings $\phi(x, i)$, including more and more prior knownledge
- Compare the representations in terms of classification accuracy
- Extract motifs (ongoing work)

Different mappings

1. spectrum mapping
2. multi-spectrum mapping
3. marginalized spectrum mapping
4. marginalized phylogenetic footprinting mapping
5. Markov marginalized phylogenetic footprinting mapping

Spectrum mapping

Map the i-th position to the basis vector corresponding to the 5-mer $x_i \ldots x_{i+4}$:



The result for a sequence is the feature space of the spectrum kernel (Leslie et al., 2002).

Spectrum mapping : limitations

- Only few discriminative motifs are expected : very hard problem in this feature space
- The functional motifs themselves may slightly vary between sequences : this would correspond to difference k-mers...
- Can we use information from promoter sequences of orthologous genes?

Promoters of orthologs

- For each gene, let us retrieve its orthologs in evolutionary close species
- They should contain the same functional motifs
- Nonfunctional nucleotides should mutate more than functional ones
- Slight variations in functional motifs should represent allowed variations

Multi-spectrum mapping

Represent a gene by the sum of the spectrum vectors of its orthologs:

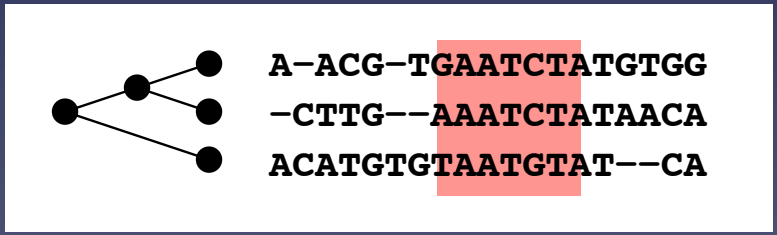
$$\phi(x) = \sum_{g \text{ orthologs}} \phi(x_g)$$

Multi-spectrum mapping : limitations

- Does not take into account the relative similarities between species (e.g., if two species are similar then their spectrum would be counted twice)
- Does not cover all possible variations in functional motifs (more restricted than, e.g., position-specific score matrices)

Promoter multiple alignment

- We assume that a multiple alignment over q species is available for each promoter, e.g.:



- The alignment uses a phylogenetic tree as stochastic model
- At each position i, we can estimate the distribution of the common ancestor given the column : $P(h_i \mid c_i)$.

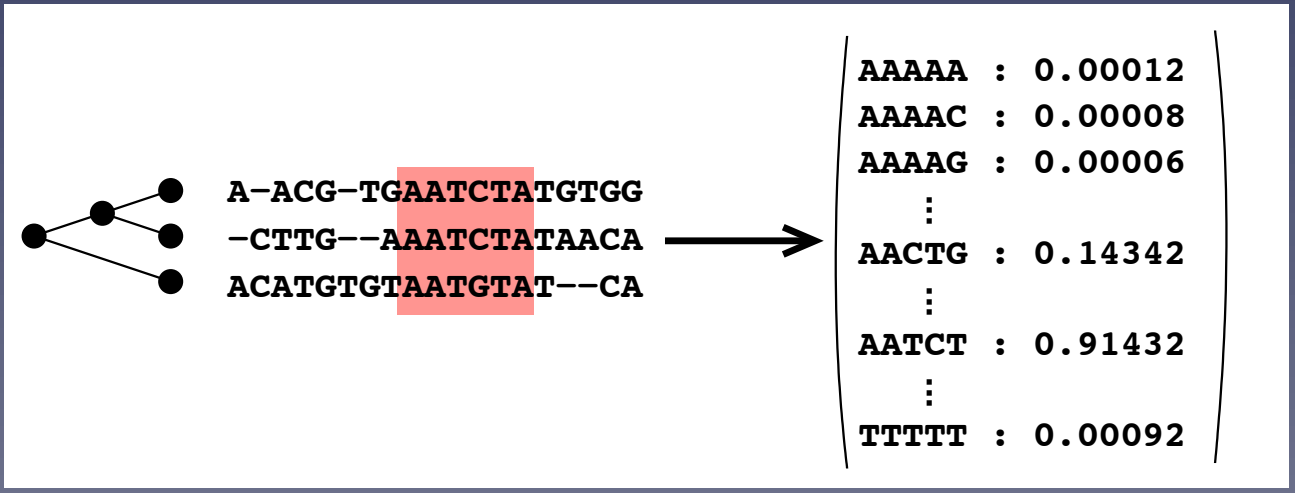
Basic probabilistic model of k-mers

We consider basic probabilistic models assuming independence of different positions. They correspond to a natural mapping to the feature space:

$$\begin{matrix} p(A) \\ p(C) \\ p(G) \\ p(T) \end{matrix} \begin{bmatrix} 0.2 & 0.1 & 0.8 & 0.1 & 0.8 \\ 0.1 & 0.7 & 0.1 & 0 & 0.1 \\ 0.6 & 0.1 & 0 & 0.6 & 0 \\ 0.1 & 0.1 & 0.1 & 0.3 & 0.1 \end{bmatrix} \rightarrow \begin{pmatrix} p(AAAAA) = 0.00128 \\ \vdots \\ p(GCAGA) = 0.16128 \\ \vdots \\ p(TTTTT) = 0.00003 \end{pmatrix}$$

Marginalized spectrum mapping

1. Do a multiple alignment of each promoter over the species
2. At each position i, compute $P(h_i|c_i)$
3. Map each k-mer to the feature space through the PSSM



Property of the marginalized spectrum mapping (1)

The mapping is the average spectrum mapping of the common ancestor (w.r.t to the conditional probability of the common ancestor given the observed sequences):

$$\begin{aligned}
\phi(x) &= \sum_i \phi(x,i) \\
&= \sum_i \left(\sum_h P(h \mid c) \phi_{spectrum}(h,i) \right) \\
&= \sum_h P(h \mid c) \phi_{spectrum}(h) \\
&= E_h \left[\phi_{spectrum}(h) \right].
\end{aligned}$$

Property of the marginalized spectrum mapping (2)

The resulting kernel is a marginalized kernel (Tsuda et al. 2002):

$$\begin{aligned}\phi(x) \cdot \phi(x') &= \sum_{h,h'} \phi_{spectrum}(h) \cdot \phi_{spectrum}(h') P(h \,|\, c) P(h' \,|\, c') \\ &= \sum_{h,h'} K_{spectrum}(h, h') P(h \,|\, c) P(h' \,|\, c').\end{aligned}$$

Limitation of the marginalized spectrum mapping

- Only a few motifs, i.e., a few k-mers are expected to be discriminant: we are searching a needle in a haystack.
- Additional hypothesis : functional positions are more conserved than non-functional ones.
- How to modify the mapping to emphasize conserved positions?

Indicator of functional position

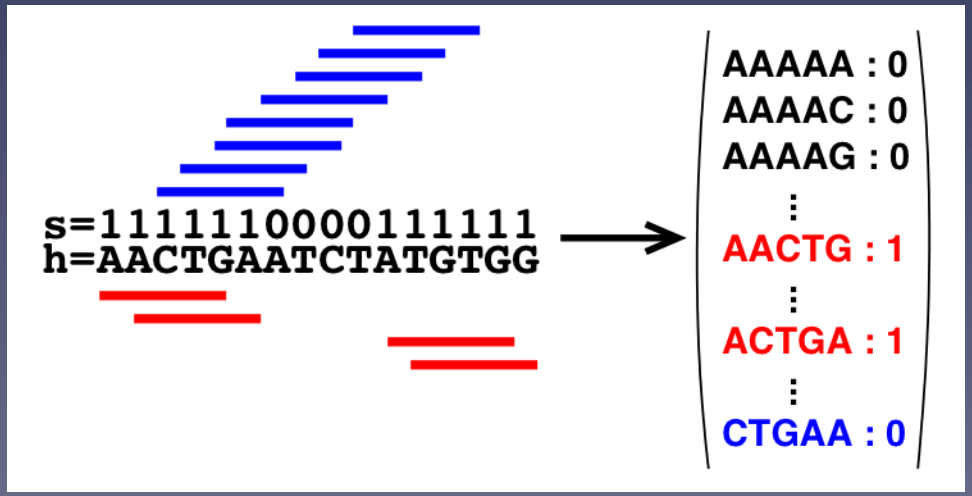
- At each position i, we add a hidden variable s_i that describes whether this position is functional or not ($s_i = 1$ is the position is functional, 0 otherwise).
- Of course we do not know s. But we can estimate it...

Phylogenetic shadowing (Boffelli et al., 2002)

- For each column c of the multiple alignment, create two phylogenetic tree models : $P(c \mid h, s = 0)$ and $P(c \mid h, s = 1)$
- The model for $s = 0$ has faster mutation rates than for $s = 1$
- Chose a prior probability of being functional P(s), and prior distributions over the ancestor nucleotide in each case $P(h \mid c = 0)$ and $P(h \mid c = 1)$
- By Bayes'rule, we can then estimate $P(s = 1 \mid c)$, and $P(h, s = 1 \mid c)$.

Marginalized phylogenetic footprinting mapping (1)

If s and h are known along the sequence, we define the mapping $\phi_{functional}(h, s)$ as the spectrum mapping of h restricted to the functional positions.



Marginalized phylogenetic footprinting mapping (2)

- We then define:

$$\phi_{shadow}(x) = \sum_{h,s} \phi_{functional}(h,s) P(h,s \mid c).$$

- It can be computed as a sum of features along the sequence.

Incorporating Markov dependencies

- Prior knownledge : functional positions form short islands
- We can therefore use a Markov model for the variable s:

$$P_{Markov}(s) = P(s_1) \prod_{i=2}^{N} P(s_i \,|\, s_{i-1})$$

- We can then marginalize

$$\phi_{shadow}(x) = \sum_{h,s} \phi_{functional}(h, s) P_{Markov}(h, s \,|\, c).$$

Mapping summary

We have created a family of mappings of increasing complexity, including more and more prior knowledge about:

- Conservation of motifs across evolution
- Slow rate of mutation of functional positions
- Structure of the promoter with short islands of functional positions.

We will now compare them experimentally.

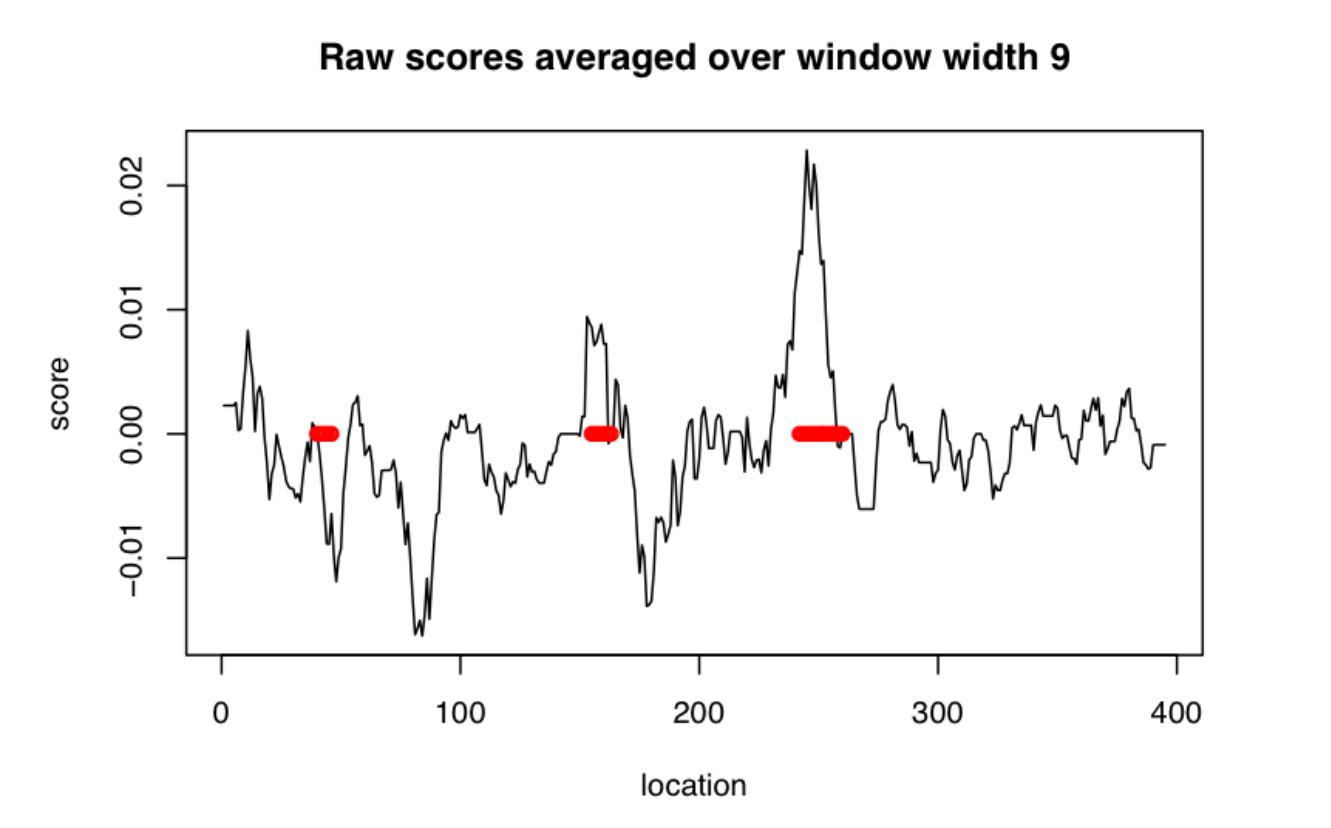
Experiment : supervised classification

- Collect promoter regions from 5 closely related yeast species
- 3591 promoter regions aligned with ClutalW accross the 5 genomes
- Create 8 classes of genes, obtained by clustering cell cycle expression data (Eisen et al., 1998)
- Use SVM, no parameter optimization, one-vs-all (8 experiments)

Classification results (ROC50)

Kernel	ATP 15	DNA 5	Glyc 17	Ribo 22	Prot 27	Spin 11	Splic 14	TCA 16	Mean
single	0.711	0.777	0.814	0.743	0.735	0.716	0.683	0.684	0.733
concatenation	0.773	0.768	0.824	0.750	0.763	0.756	0.739	0.740	0.764
marginalized	0.799	0.805	0.833	0.729	0.748	0.721	0.676	0.673	0.748
shadow 2	0.881	0.929	**0.928**	0.840	0.867	**0.827**	0.787	**0.770**	**0.854**
shadow 5	**0.889**	**0.935**	0.927	0.819	0.849	0.821	0.766	0.752	0.845
Markov 2 90/90	0.848	0.891	0.908	0.830	0.853	0.801	0.773	0.758	0.833
Markov 2 90/99	0.868	0.911	0.915	0.826	0.850	0.782	0.752	0.735	0.830
Markov 2 99/99	0.869	0.910	0.912	0.816	0.840	0.773	0.737	0.724	0.823
Markov 5 90/90	0.875	0.922	0.924	**0.844**	**0.868**	0.814	**0.788**	0.769	0.851
Markov 5 90/99	0.872	0.916	0.920	0.834	0.858	0.794	0.774	0.755	0.840
Markov 5 99/99	0.868	0.917	0.921	0.830	0.853	0.774	0.751	0.733	0.831

Motif detection example : YKR010C



Discriminative motif extraction (in progress)

- For each SVM trained, extract the 20 k-mers with largest weight (linear classifier).
- Take the union of extracted k-mers over 15 runs of cross-validation.
- Gold standard $=$ for each class find the top 3 motifs of the JASPER database (using MONKEY), extract corresponding 5-mers.
- Compare the gold standard with the k-mers found by SVM.

Motif detection : results

	ATP	DNA	Glyc	Ribo	Prot	Spin	Splic	TCA
SVM	46	40	55	50	49	43	48	50
Motif	180	68	227	38	148	152	52	104
Class	1006	839	967	973	1001	891	881	995
Inter	24	8	23	18	23	19	14	21
Expect	8.23	3.24	12.91	1.95	7.25	7.34	2.83	5.23
p-value	6.19e-8	1.15e-2	1.44e-3	3.88e-15	3.24e-8	1.74e-5	1.15e-7	2.00e-9

Conclusion

- An attempt to find motifs in a discriminative setting
- A new family of mappings/kernels incorporating various biological knowledge
- Positive effect on classification performance
- Promising results on motif extraction but method to be improved and automatized
- Reference : J.-P. Vert, R. Thurman and W. S. Noble, Kernels for gene regulatory regions, *NIPS 2005*.