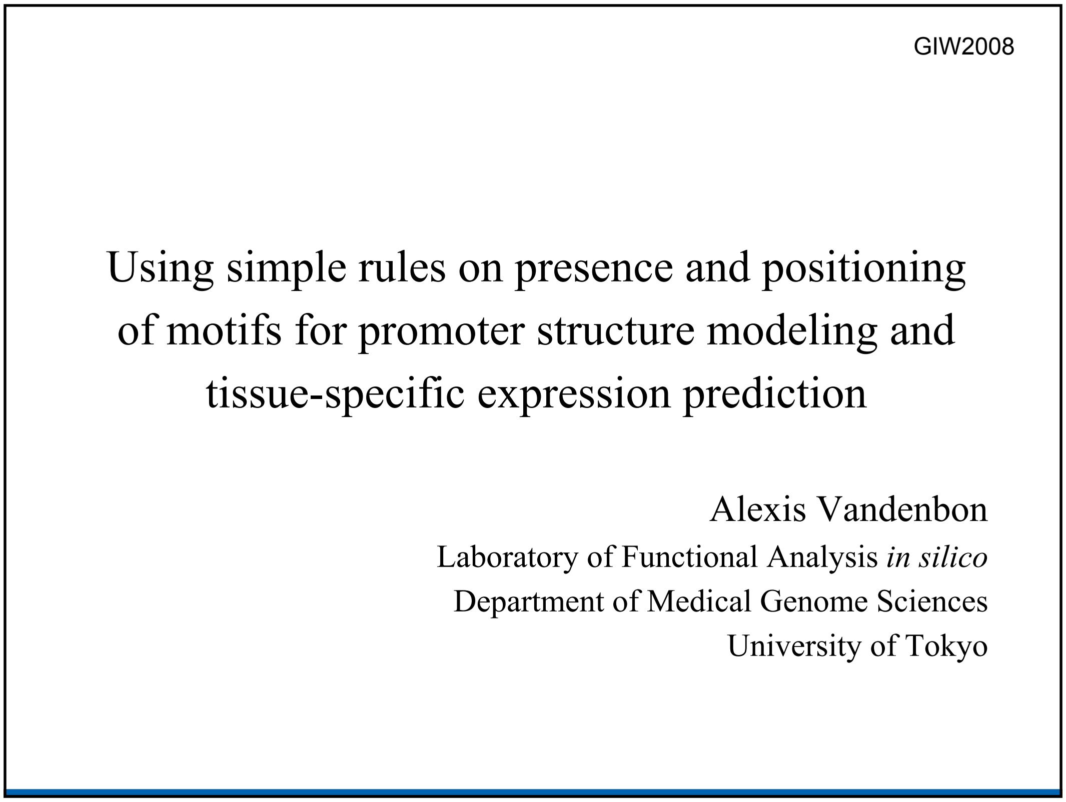
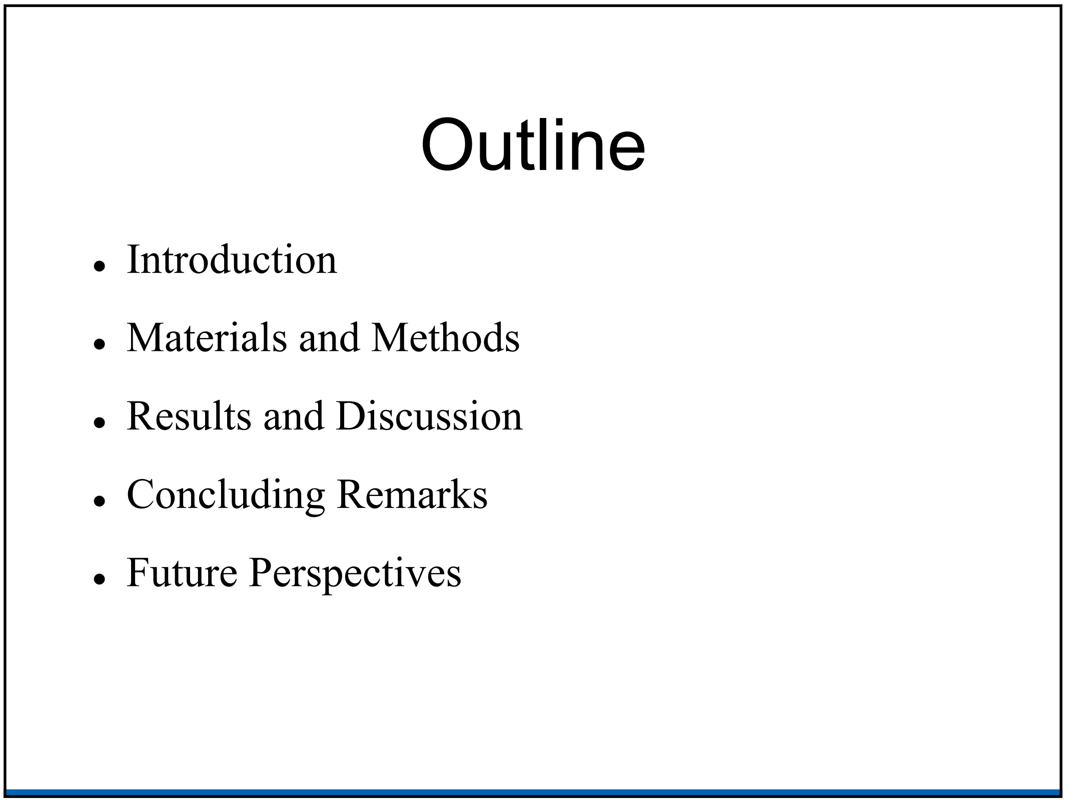

GIW2008

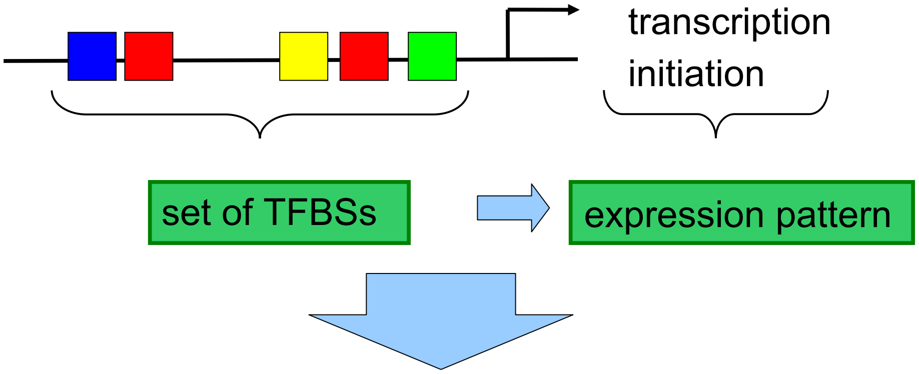
Using simple rules on presence and positioning of motifs for promoter structure modeling and tissue-specific expression prediction

Alexis Vandenbon
Laboratory of Functional Analysis *in silico*
Department of Medical Genome Sciences
University of Tokyo


Outline

- Introduction
- Materials and Methods
- Results and Discussion
- Concluding Remarks
- Future Perspectives

Regulation of transcription



transcription initiation

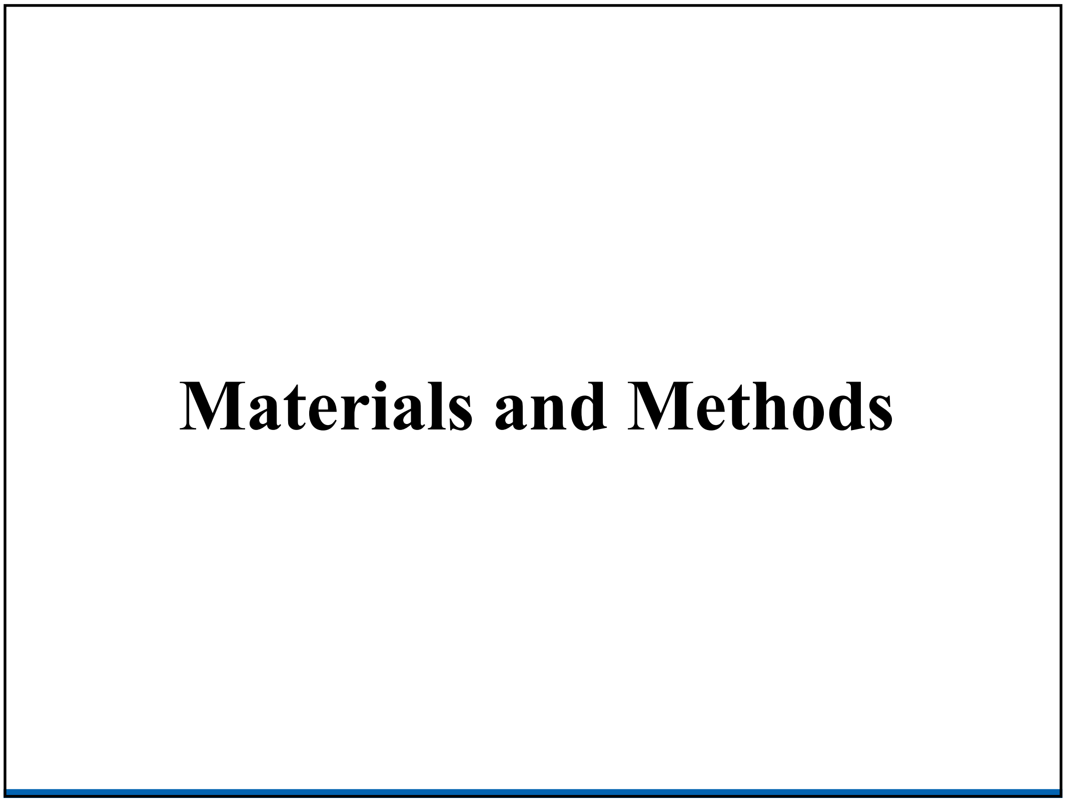
set of TFBSs

expression pattern

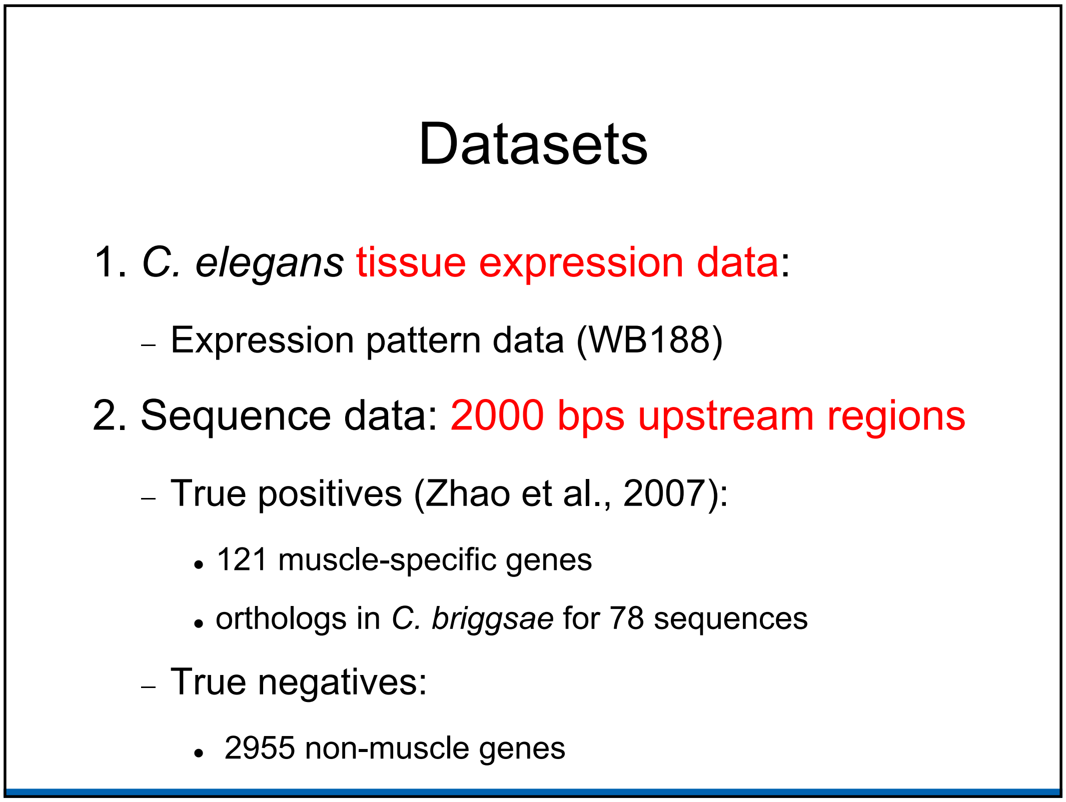
Can we predict expression patterns from the promoter sequence architecture?

C. elegans muscle tissue

- Extensive tissue expression data
- For muscle
  - Muscle-specific genes, regulatory regions
  - Candidate *cis*-regulatory motifs



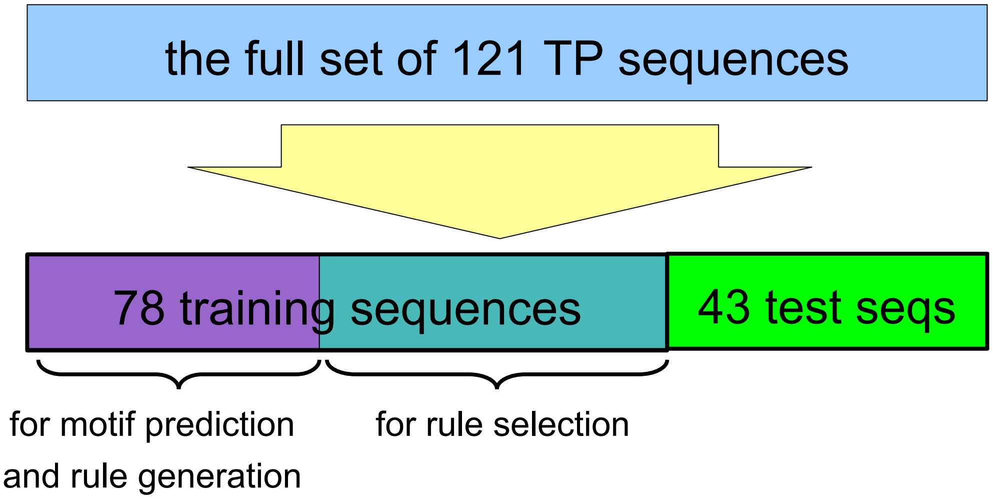
Materials and Methods



Datasets

1. *C. elegans* tissue expression data:
   - Expression pattern data (WB188)
2. Sequence data: 2000 bps upstream regions
   - True positives (Zhao et al., 2007):
     - 121 muscle-specific genes
     - orthologs in *C. briggsae* for 78 sequences
   - True negatives:
     - 2955 non-muscle genes

True Positive Datasets



the full set of 121 TP sequences

78 training sequences

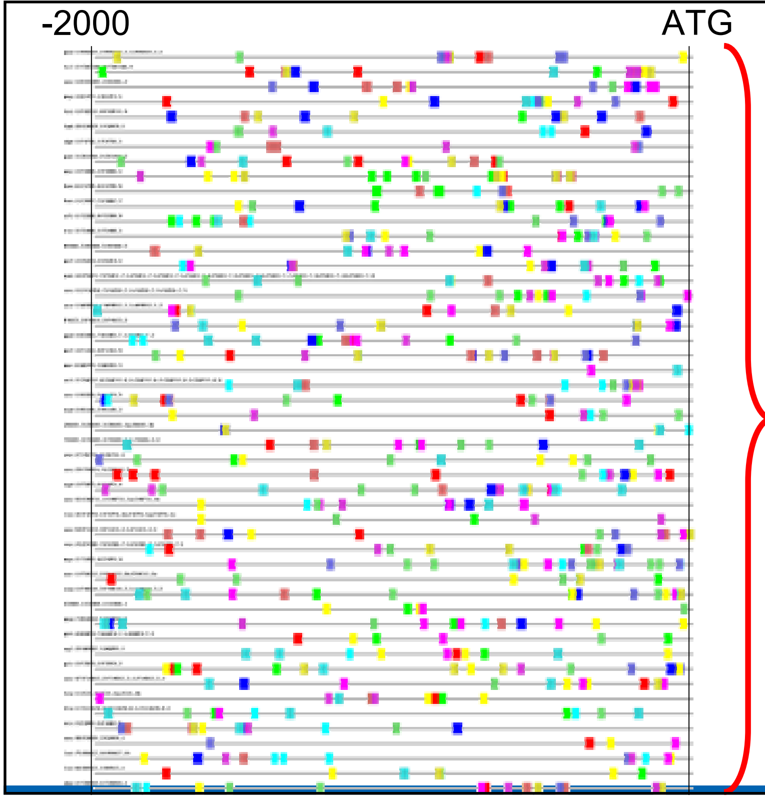
43 test seqs

for motif prediction and rule generation

for rule selection

De novo motif prediction

- Predictions
  - Using conservation in *C. briggsae*
  - In the entire sequences, and in sub-regions of different lengths
  - Using MEME, AlignACE, MotifSampler, Weeder
- Get an over-representation measure for each motif
- Remove redundancies

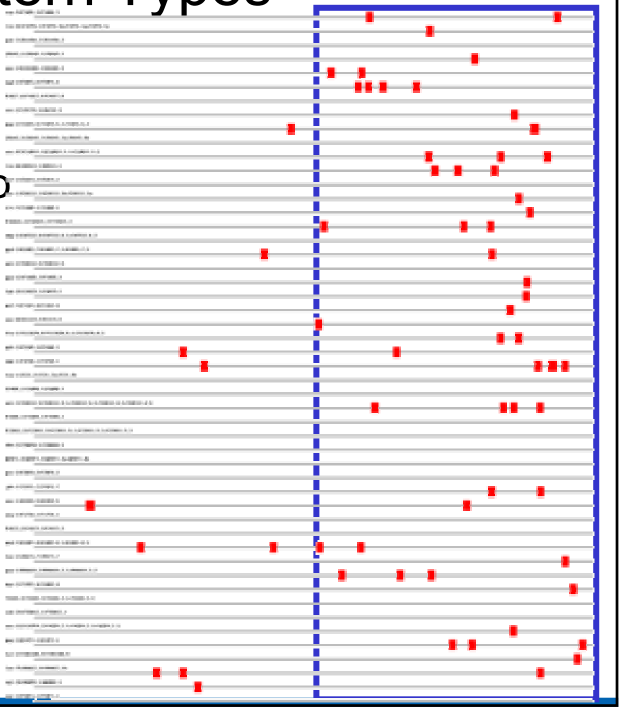


About 10 predicted motifs

Are there any significant patterns??

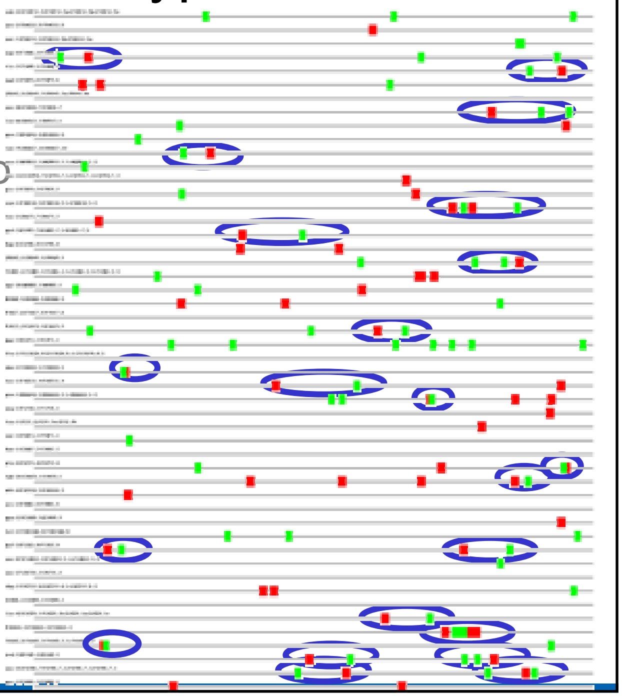
Pattern Types

- positioning relative to the translation start site



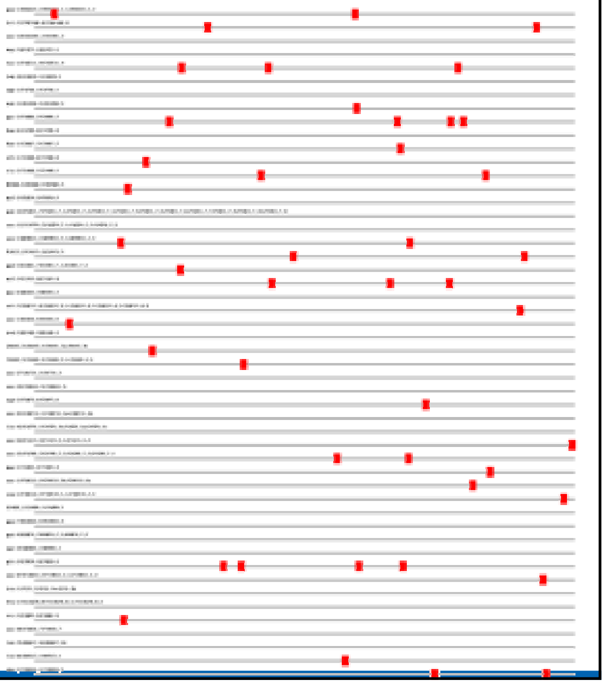
Pattern Types

- positioning relative to the translation start site
- relative positioning of pairs of motifs



Pattern Types

- positioning relative to the translation start site
- relative positioning of pairs of motifs
- presence of a motif



What are Useful Patterns?

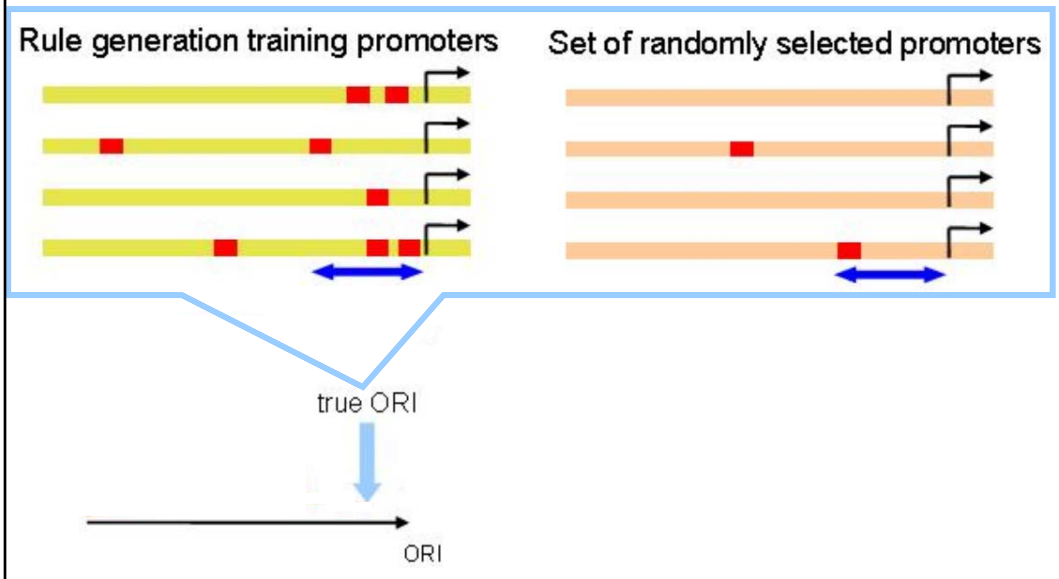
- Patterns that are **over-represented** in the input sequences vs non-inputs
  - in **total number of occurrences**, and
  - in the number of sequences containing **at least 1 occurrence**

Over-Representation Index

$$\mathrm{ORI}_i = \underbrace{\frac{\mathrm{density}_{TP,i}}{\mathrm{density}_{genomic,i}}}_{\text{Total number of occurrences of pattern i}} \times \underbrace{\frac{\mathrm{proportion}_{TP,i}}{\mathrm{proportion}_{genomic,i}}}_{\text{At least 1 occurrence of pattern i}}$$

Bajic et al., 2004

Rule generation



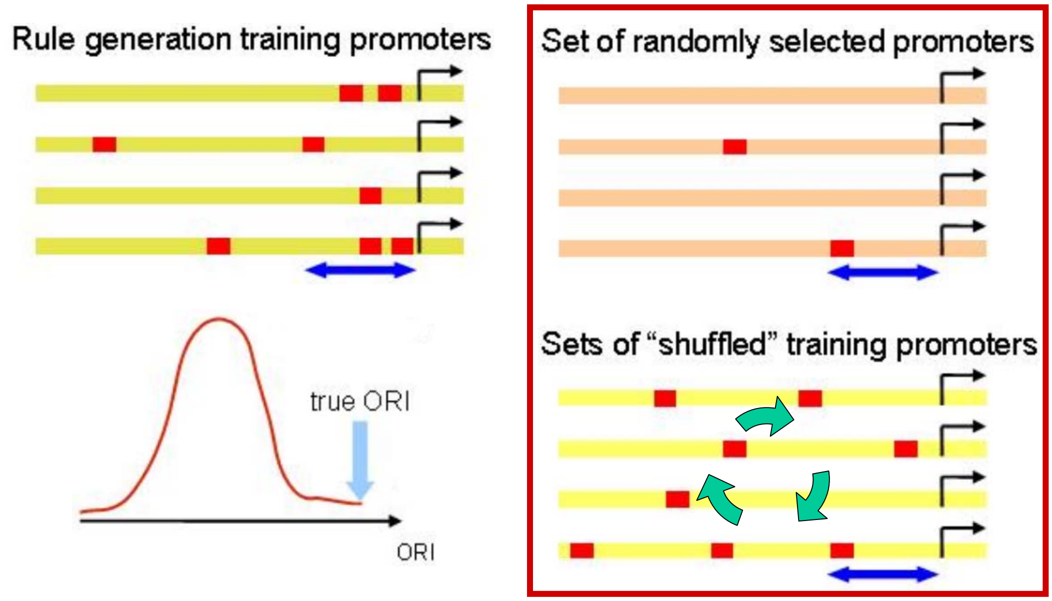
Rule generation training promoters

Set of randomly selected promoters

true ORI

ORI

Rule generation



Rule generation training promoters

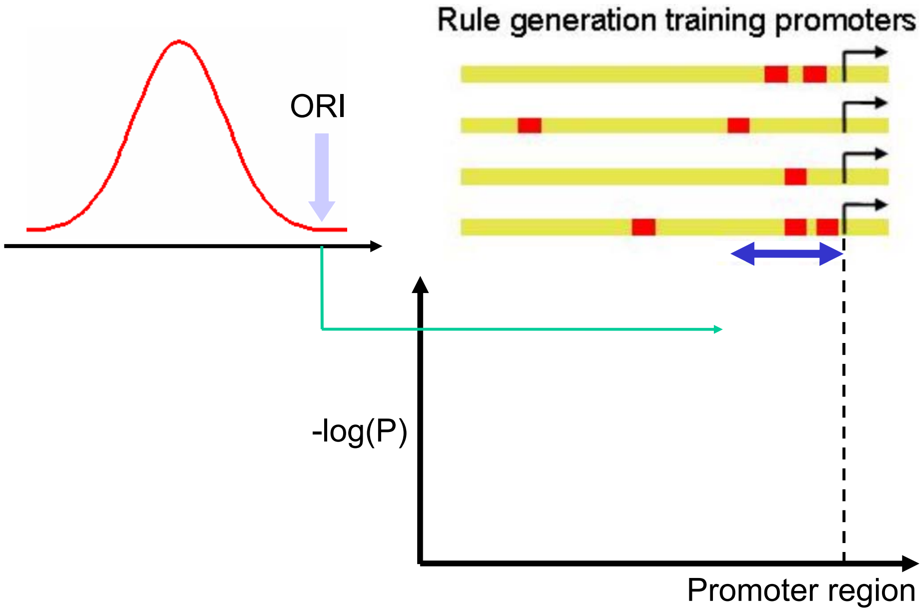
Set of randomly selected promoters

Sets of “shuffled” training promoters

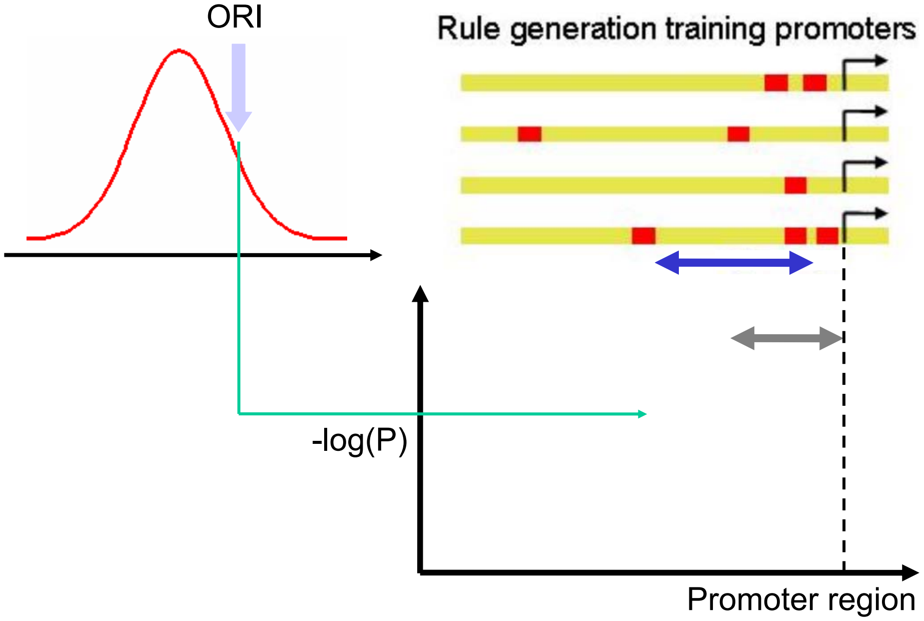
true ORI

ORI

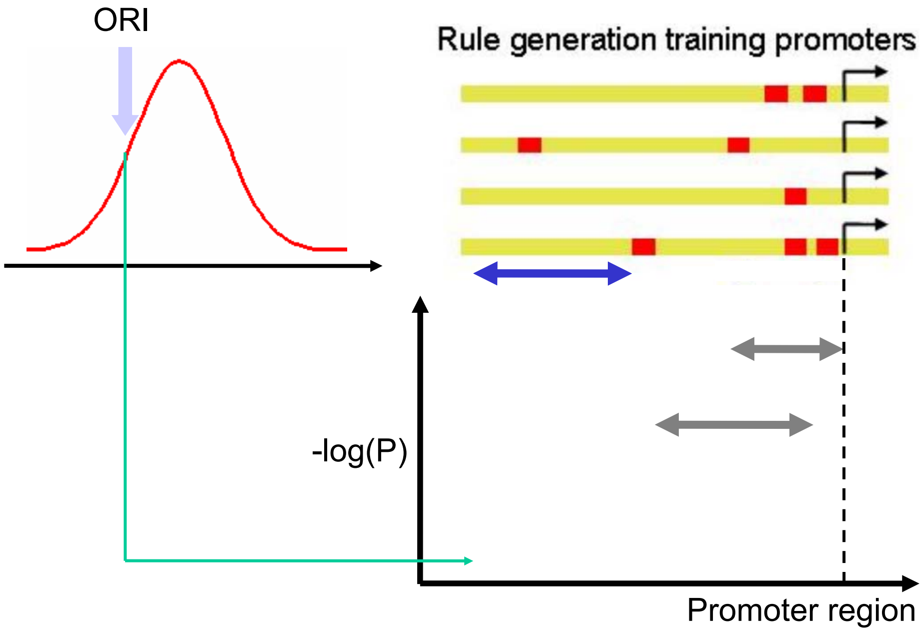
Rule generation



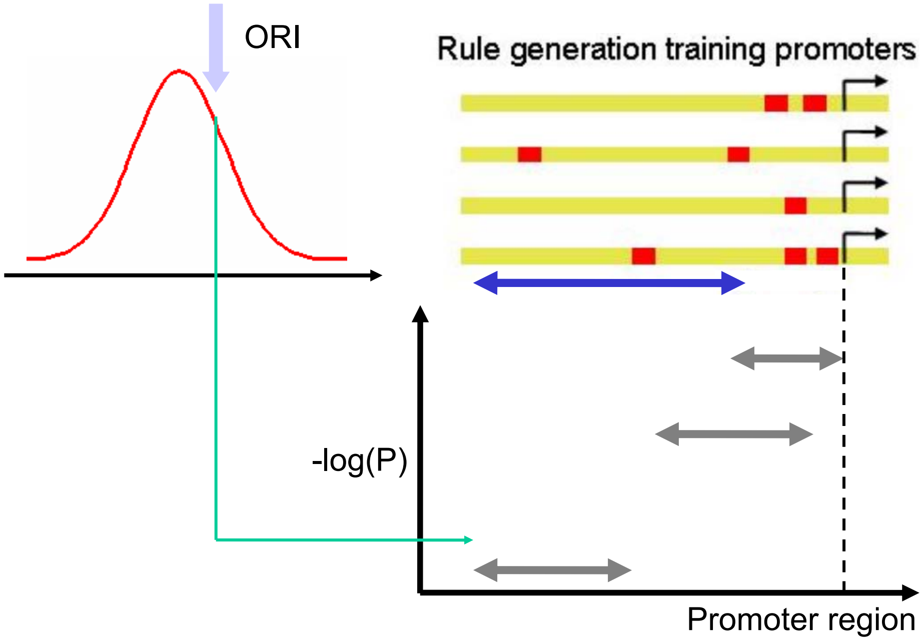
Rule generation



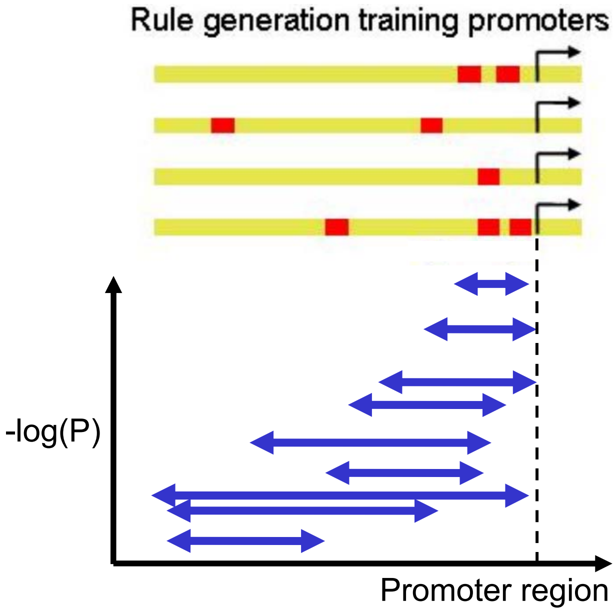
Rule generation



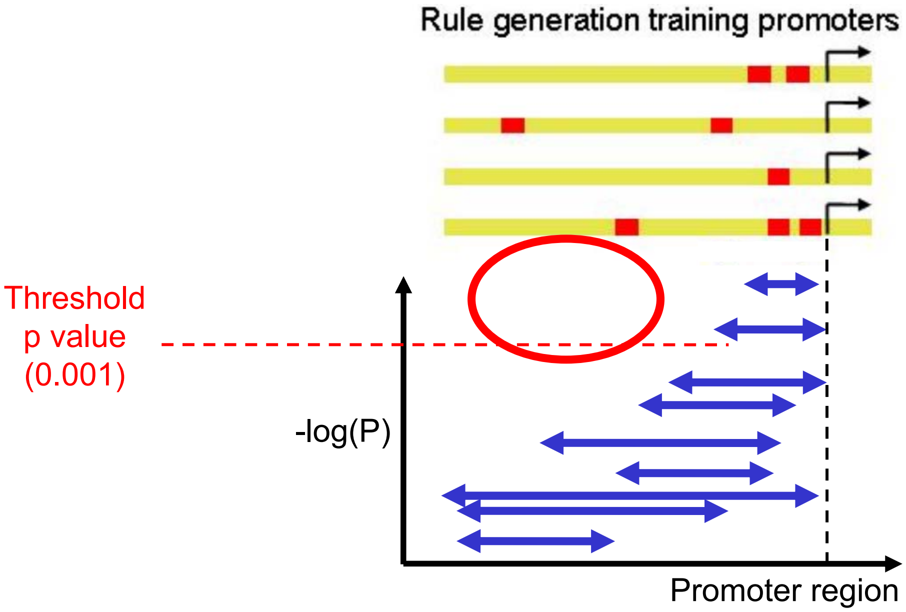
Rule generation



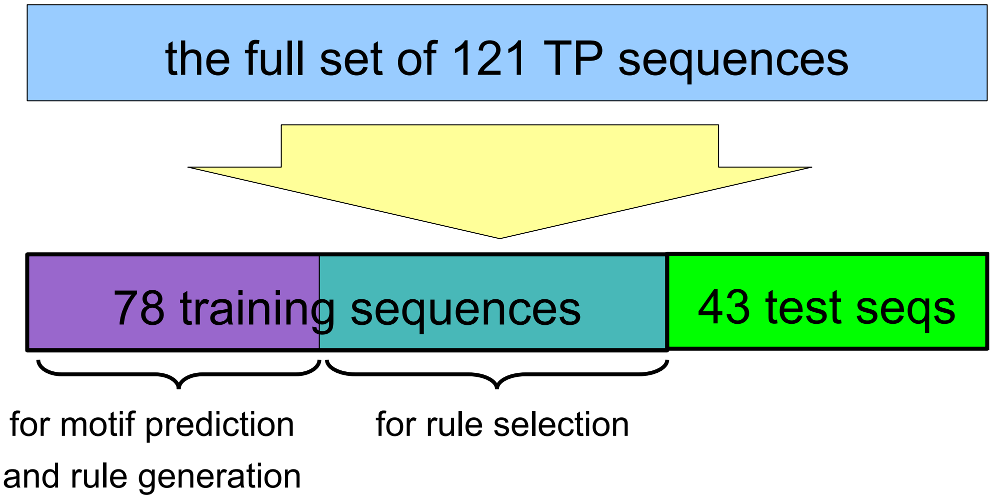
Rule generation



Rule generation



True Positive Datasets


the full set of 121 TP sequences

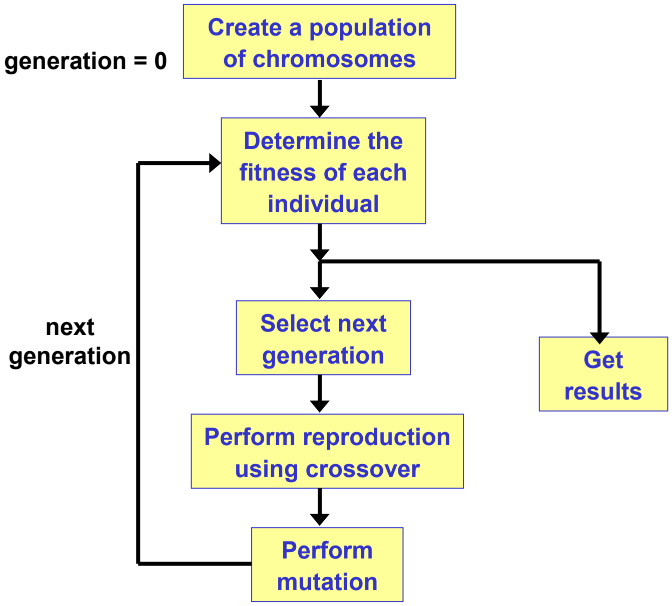
78 training sequences

43 test seqs

for motif prediction and rule generation

for rule selection

Genetic Algorithm (GA)


generation = 0

Create a population of chromosomes

Determine the fitness of each individual

Select next generation

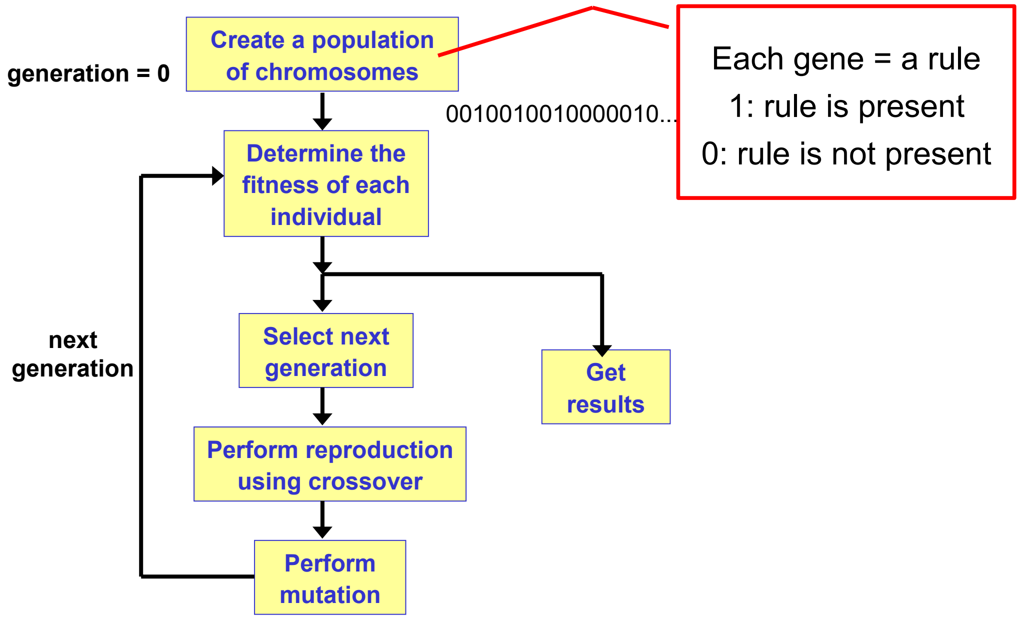
Get results

Perform reproduction using crossover

Perform mutation

next generation

Genetic Algorithm (GA)



generation = 0

Create a population of chromosomes

0010010010000010...

Each gene = a rule

1: rule is present

0: rule is not present

Determine the fitness of each individual

next generation

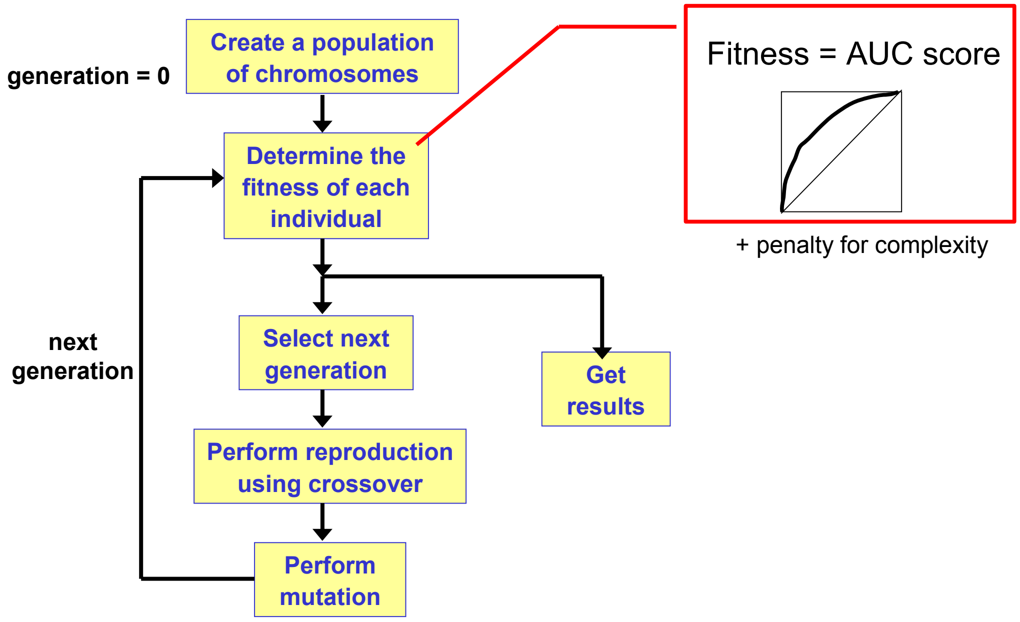
Select next generation

Get results

Perform reproduction using crossover

Perform mutation

Genetic Algorithm (GA)



generation = 0

Create a population of chromosomes

Determine the fitness of each individual

Fitness = AUC score

+ penalty for complexity

next generation

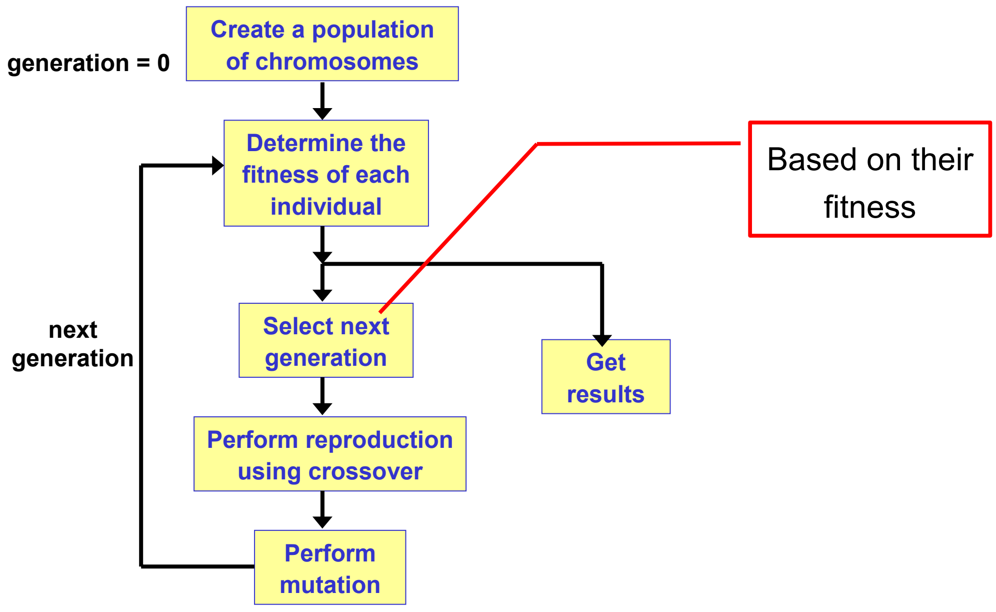
Select next generation

Get results

Perform reproduction using crossover

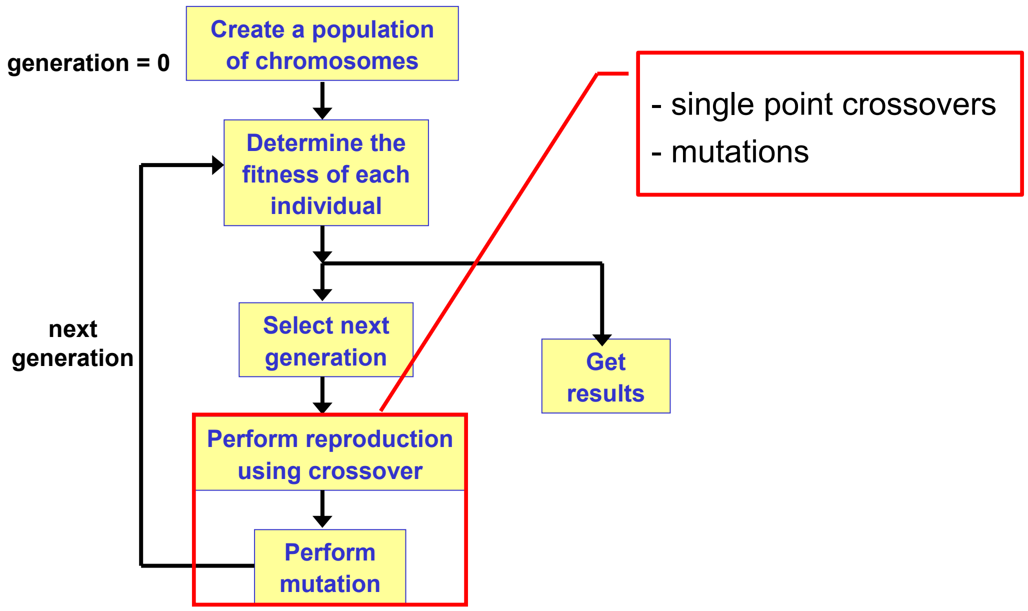
Perform mutation

Genetic Algorithm (GA)


generation = 0
Create a population of chromosomes
Determine the fitness of each individual
next generation
Select next generation
Get results
Perform reproduction using crossover
Perform mutation

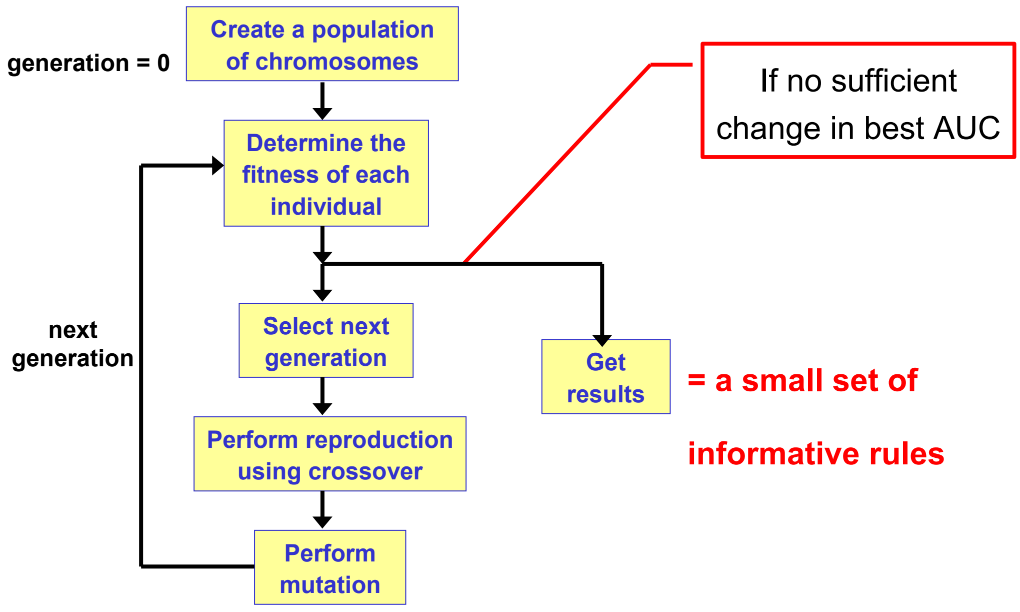
Based on their fitness

Genetic Algorithm (GA)


generation = 0
Create a population of chromosomes
Determine the fitness of each individual
next generation
Select next generation
Get results
Perform reproduction using crossover
Perform mutation

- single point crossovers
- mutations

Genetic Algorithm (GA)



generation = 0

Create a population of chromosomes

Determine the fitness of each individual

Select next generation

Perform reproduction using crossover

Perform mutation

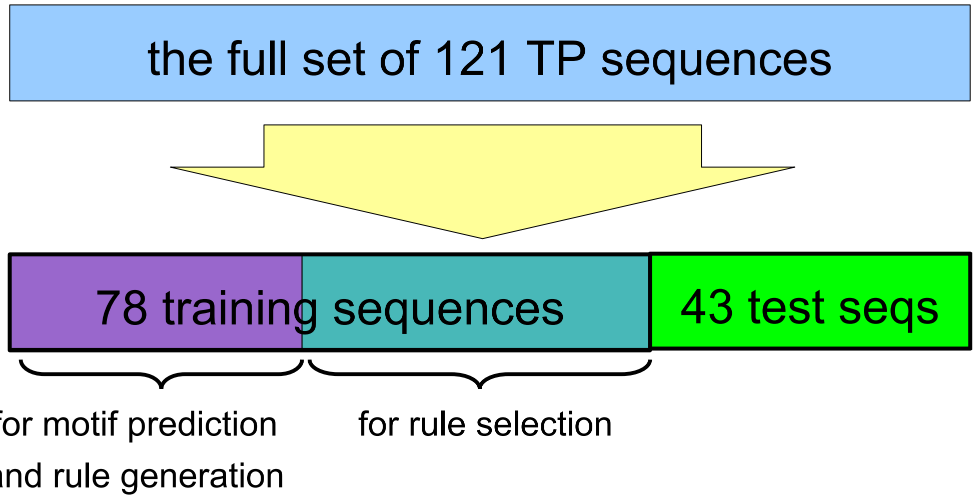
next generation

Get results

If no sufficient change in best AUC

= a small set of informative rules

True Positive Datasets



the full set of 121 TP sequences

78 training sequences

43 test seqs

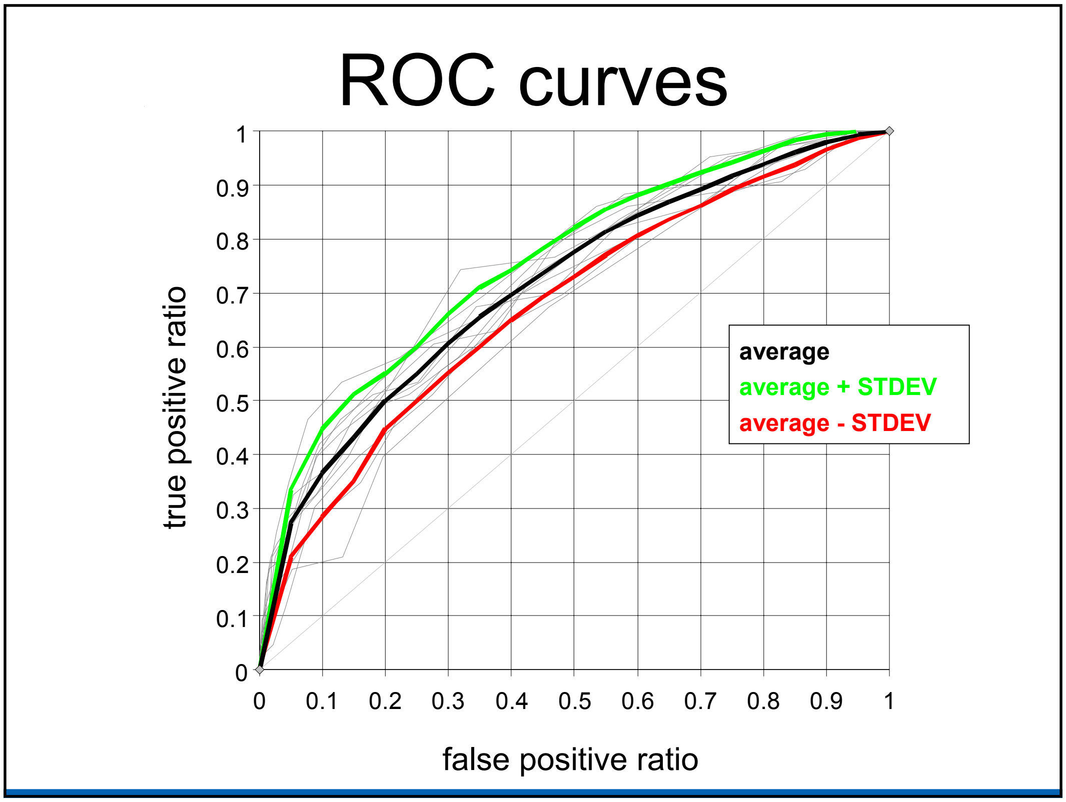
for motif prediction and rule generation

for rule selection

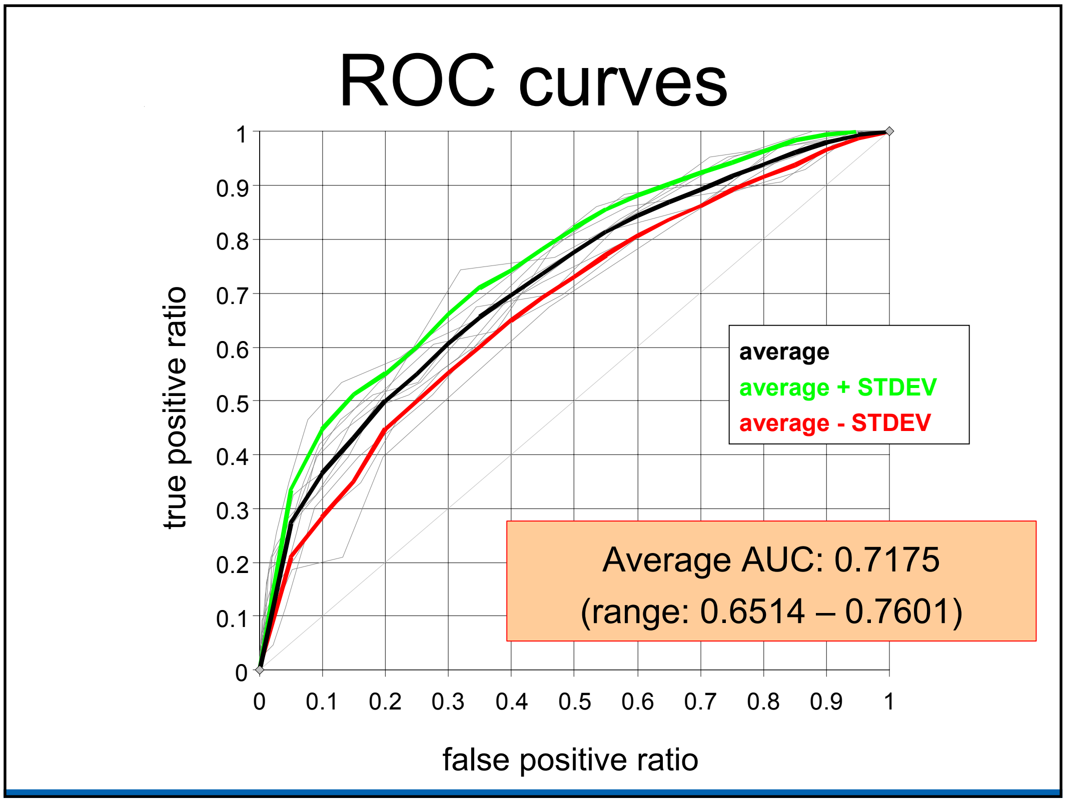
Final Performance Evaluation

- True positives: test set (remaining 43 seqs)
- “True negatives” (2955 seqs with no reported expression in muscle)
- Measures of performance:
  - ROC curve AUC
  - enrichment of test seqs in top scoring seqs
  - etc

Results and Discussion

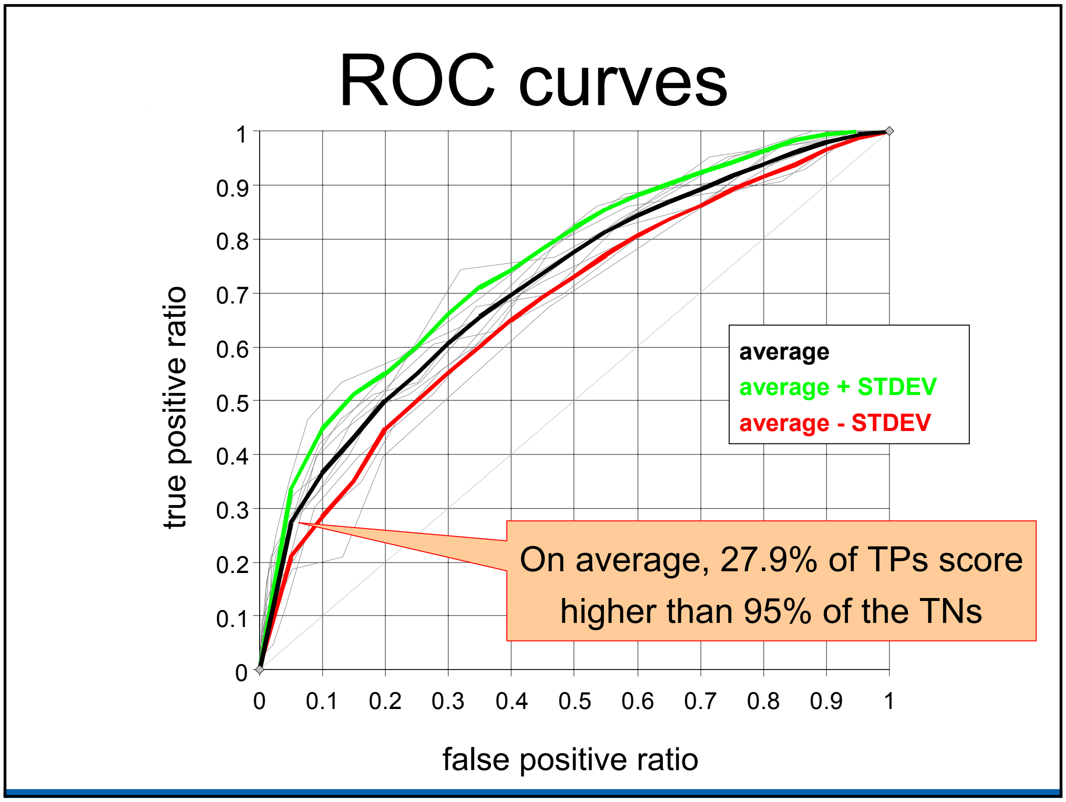


ROC curves



ROC curves

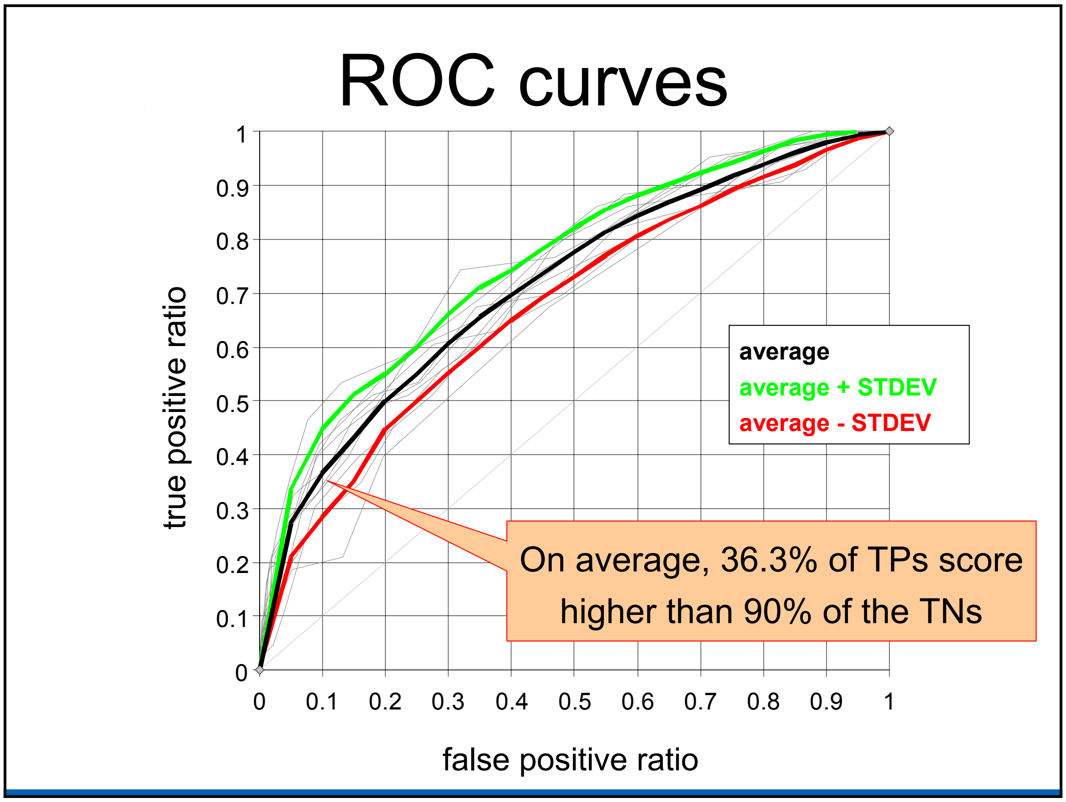
Average AUC: 0.7175
(range: 0.6514 – 0.7601)



ROC curves

average
average + STDEV
average - STDEV

On average, 27.9% of TPs score higher than 95% of the TNs

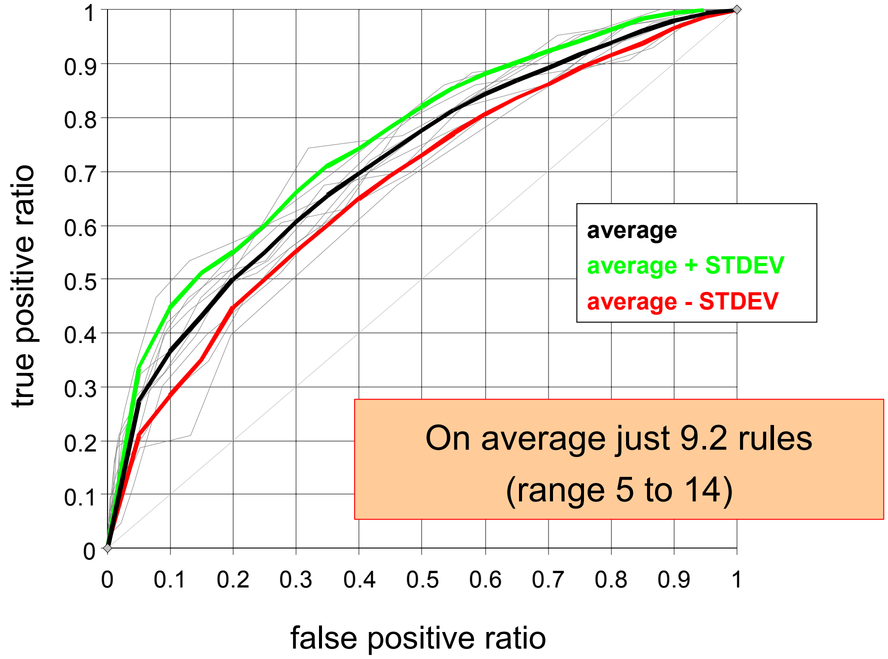


ROC curves

average
average + STDEV
average - STDEV

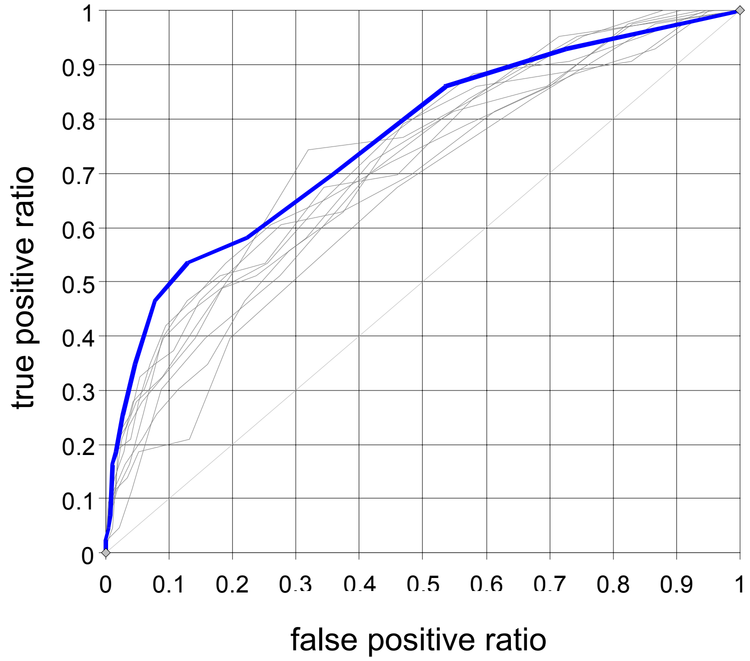
On average, 36.3% of TPs score higher than 90% of the TNs

ROC curves

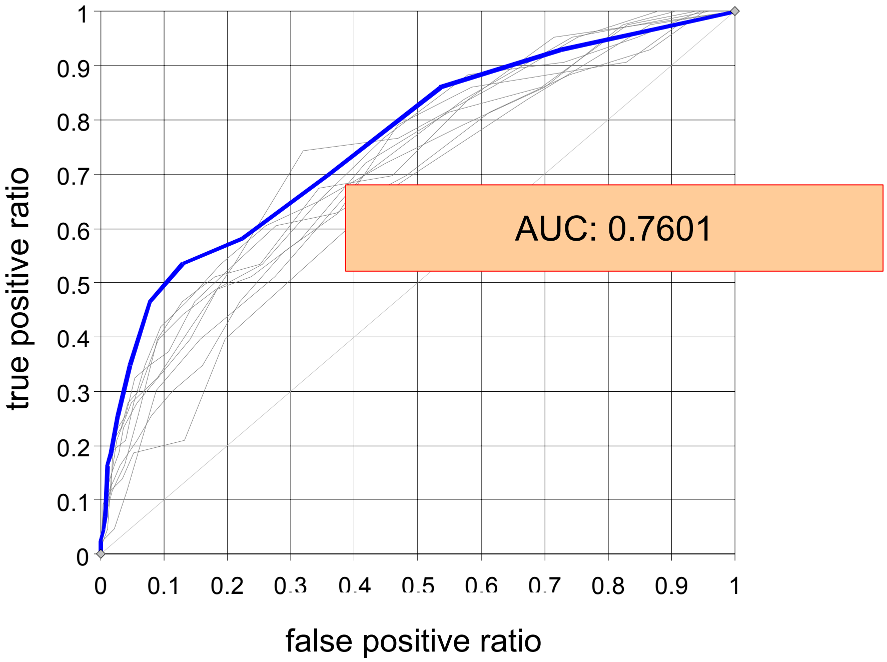


On average just 9.2 rules
(range 5 to 14)

Focus on the Best Run

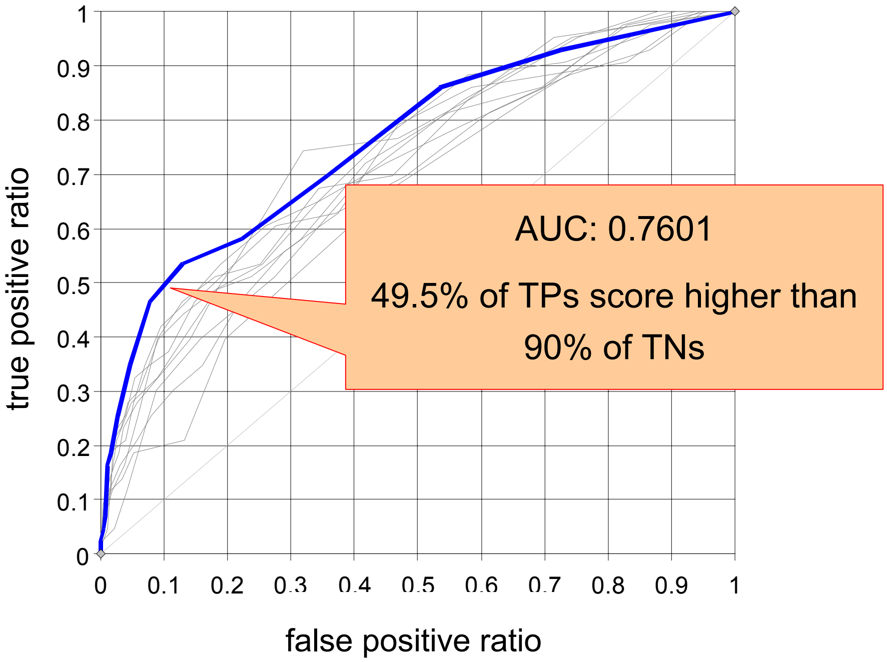


Focus on the Best Run



AUC: 0.7601

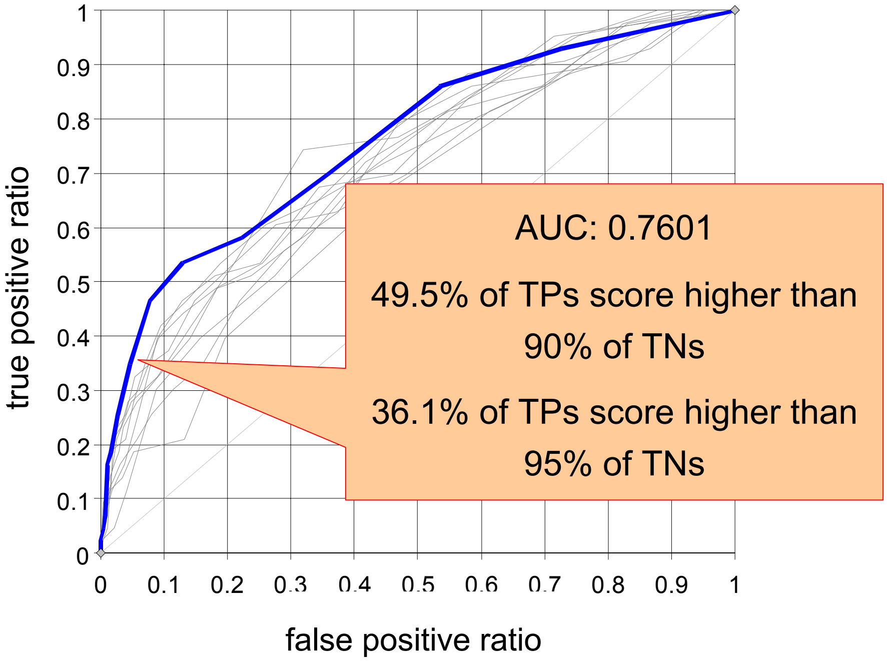
Focus on the Best Run



AUC: 0.7601

49.5% of TPs score higher than 90% of TNs

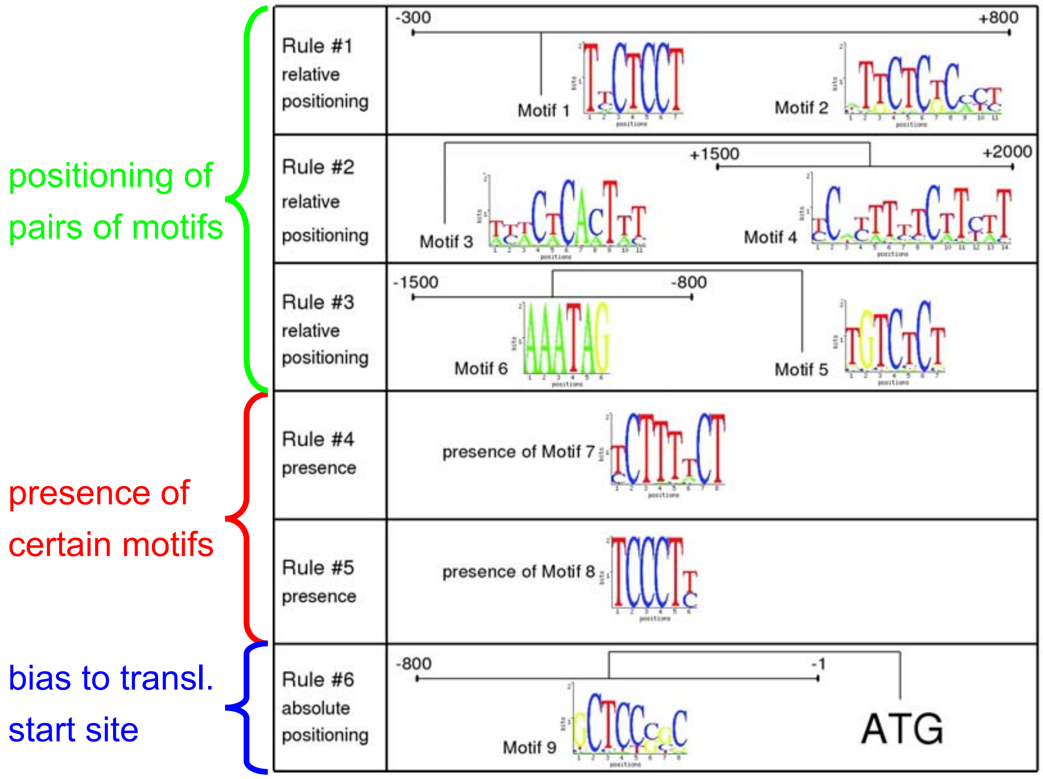
Focus on the Best Run



AUC: 0.7601

49.5% of TPs score higher than 90% of TNs

36.1% of TPs score higher than 95% of TNs



positioning of pairs of motifs

Rule #1 relative positioning

Rule #2 relative positioning

Rule #3 relative positioning

presence of certain motifs

Rule #4 presence

presence of Motif 7

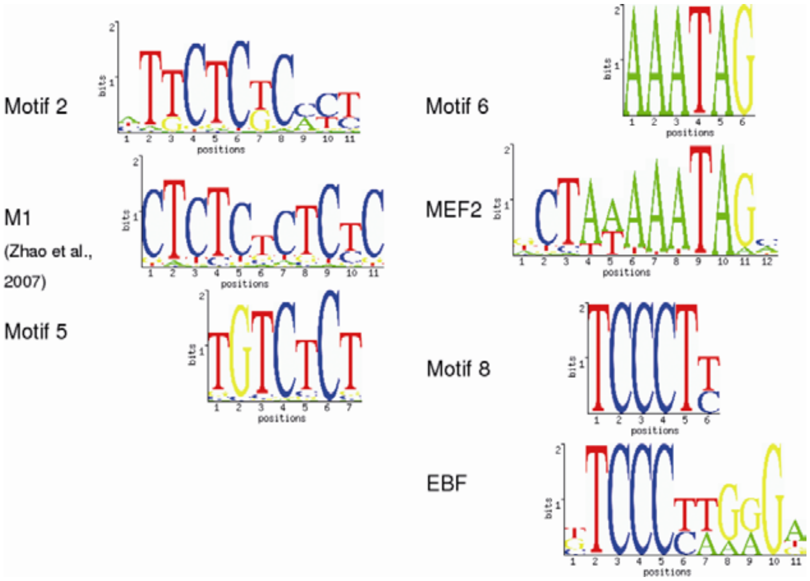
Rule #5 presence

presence of Motif 8

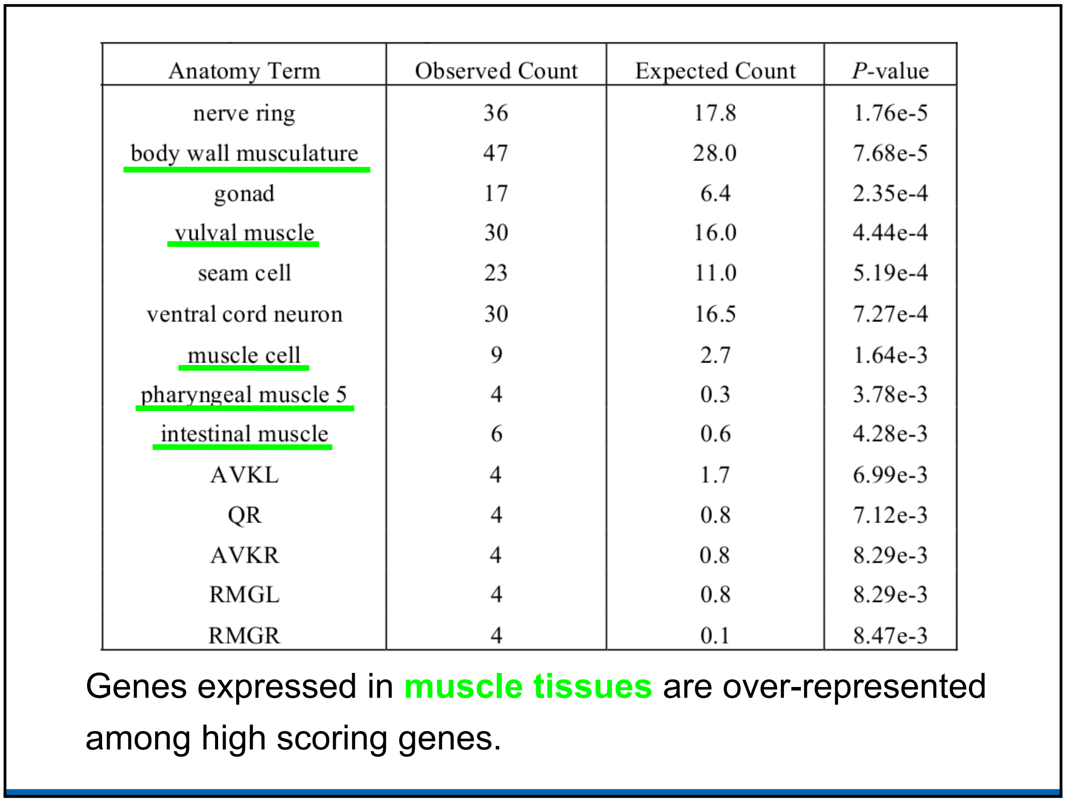
bias to transl. start site

Rule #6 absolute positioning

Similarities to Known Motifs

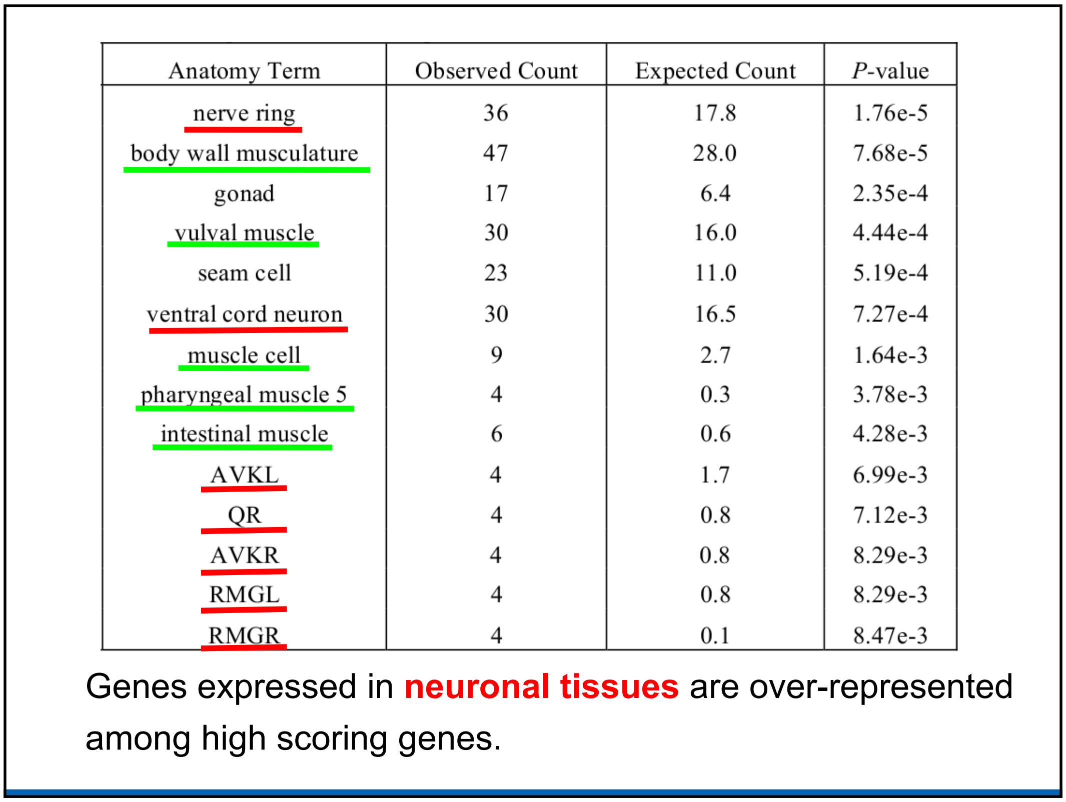


Anatomy Term	Observed Count	Expected Count	*P*-value
nerve ring	36	17.8	1.76e-5
body wall musculature	47	28.0	7.68e-5
gonad	17	6.4	2.35e-4
vulval muscle	30	16.0	4.44e-4
seam cell	23	11.0	5.19e-4
ventral cord neuron	30	16.5	7.27e-4
muscle cell	9	2.7	1.64e-3
pharyngeal muscle 5	4	0.3	3.78e-3
intestinal muscle	6	0.6	4.28e-3
AVKL	4	1.7	6.99e-3
QR	4	0.8	7.12e-3
AVKR	4	0.8	8.29e-3
RMGL	4	0.8	8.29e-3
RMGR	4	0.1	8.47e-3



Anatomy Term	Observed Count	Expected Count	*P*-value
nerve ring	36	17.8	1.76e-5
body wall musculature	47	28.0	7.68e-5
gonad	17	6.4	2.35e-4
vulval muscle	30	16.0	4.44e-4
seam cell	23	11.0	5.19e-4
ventral cord neuron	30	16.5	7.27e-4
muscle cell	9	2.7	1.64e-3
pharyngeal muscle 5	4	0.3	3.78e-3
intestinal muscle	6	0.6	4.28e-3
AVKL	4	1.7	6.99e-3
QR	4	0.8	7.12e-3
AVKR	4	0.8	8.29e-3
RMGL	4	0.8	8.29e-3
RMGR	4	0.1	8.47e-3

Genes expressed in **muscle tissues** are over-represented among high scoring genes.

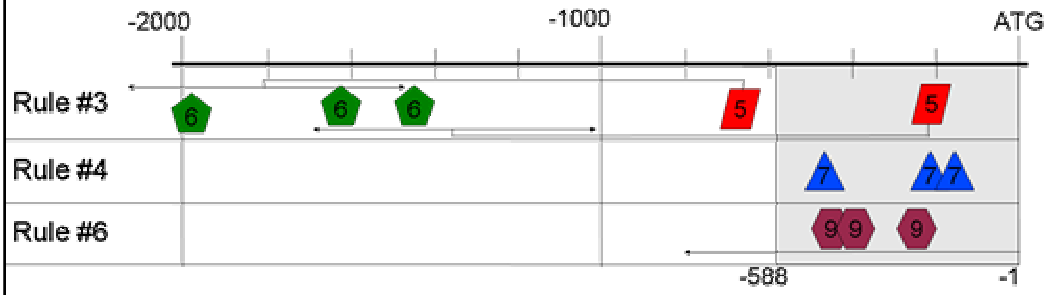


Anatomy Term	Observed Count	Expected Count	*P*-value
nerve ring	36	17.8	1.76e-5
body wall musculature	47	28.0	7.68e-5
gonad	17	6.4	2.35e-4
vulval muscle	30	16.0	4.44e-4
seam cell	23	11.0	5.19e-4
ventral cord neuron	30	16.5	7.27e-4
muscle cell	9	2.7	1.64e-3
pharyngeal muscle 5	4	0.3	3.78e-3
intestinal muscle	6	0.6	4.28e-3
AVKL	4	1.7	6.99e-3
QR	4	0.8	7.12e-3
AVKR	4	0.8	8.29e-3
RMGL	4	0.8	8.29e-3
RMGR	4	0.1	8.47e-3

Genes expressed in **neuronal tissues** are over-represented among high scoring genes.

Match to Experimentally Verified Regions

Sites fitting the rules show a tendency to be present in **experimentally verified regulatory regions** (p-value 0.0017).



GuhaThakurta et al., 2004

unc-89 *(UNCoordinated)* is expressed in body wall muscle, pharyngeal muscle and some cells in the tail.

Concluding Remarks (1)

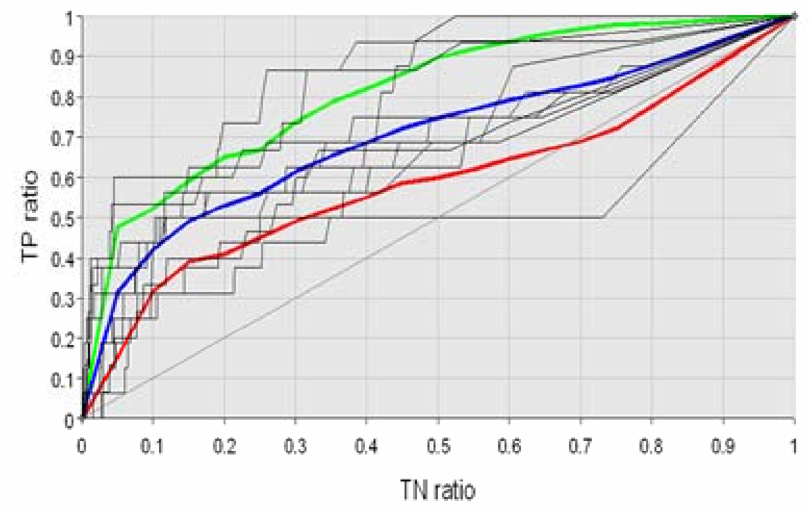
- Simple, easy to understand, model
- Good performance
  - high AUC values
  - strong enrichment of TPs among high scoring seqs
  - high scoring genes tend to be expressed in muscle tissues
  - similarity to known motifs
  - match to experimentally verified regions

Concluding Remarks (2)

- Models more than just clustering of sites
  - positioning to TSS/translation start site
  - proximal positioning of pairs of sites
  - distal positioning of pairs of sites
- Generally applicable
  - no species specific or tissue specific information is used

Perspectives

Application in other organisms and tissues



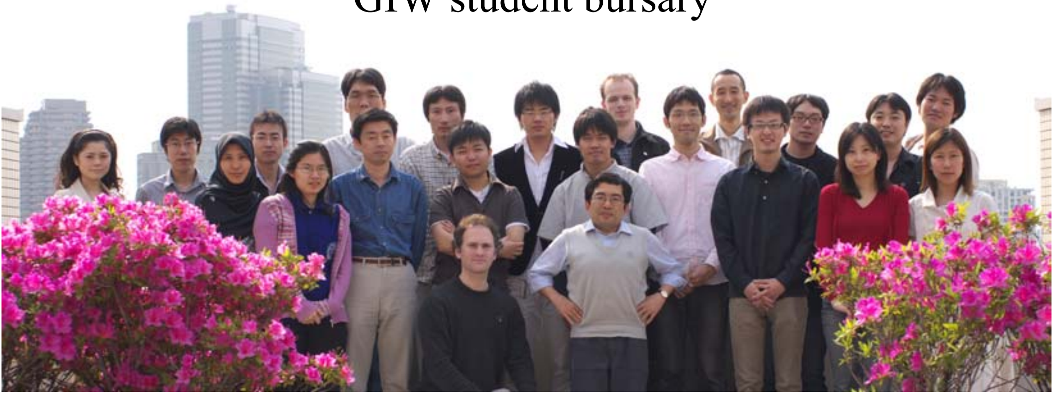
Acknowledgments

Prof. Kenta Nakai

The members of the Nakai Lab

Japanese Government (MEXT Scholarship)

GIW student bursary



GIW2008

Using simple rules on presence and positioning of motifs for promoter structure modeling and tissue-specific expression prediction

Alexis Vandenbon
Laboratory of Functional Analysis *in silico*
Department of Medical Genome Sciences
University of Tokyo