KIRMES

Kernel-based Identification of Regulatory Modules in Euchromatic Sequences

GCB 2008 Dresden September 10, 2008

Sebastian J. Schultheiss <sebi@tuebingen.mpg.de>
Wolfgang Busch, Jan U. Lohmann, Oliver Kohlbacher, and Gunnar Rätsch

Friedrich Miescher Laboratory
of the Max Planck Society

MAX-PLANCK-GESELLSCHAFT

Eberhard Karls
Universität
Tübingen

ZBIT

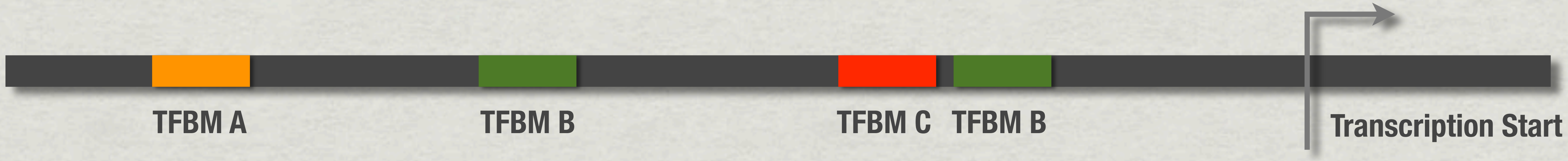
KIRMES Overview

- Combine motif finding with SVMs
- Model degenerate motifs and regulatory modules of promoters
- **Input:** sets of co-expressed genes
- **Output:** a function that classifies genes as co-regulated or not

First hit in Google Image Search for "KIRMES"
bocholt.de

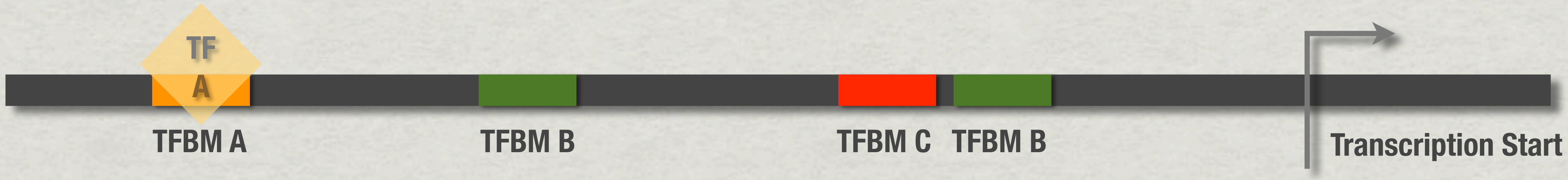
Transcriptional Regulation

- **Transcription factors** (TFs) are proteins regulating transcription (activate, repress, ...)
- TFs bind to euchromatic sequences, *e.g.* promoter regions, **regulatory regions** in introns, ...
- TFs recognize conserved **binding motifs** (TFBMs)
- TFs form complexes, **bind** to **modules** of TFBMs
- **Relative distance** of TFBMs to transcription start and among TFBMs is relevant



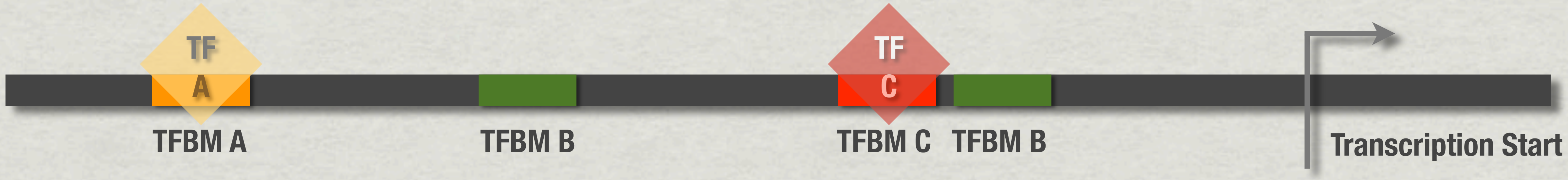
Transcriptional Regulation

- **Transcription factors** (TFs) are proteins regulating transcription (activate, repress, ...)
- TFs bind to euchromatic sequences, *e.g.* promoter regions, **regulatory regions** in introns, ...
- TFs recognize conserved **binding motifs** (TFBMs)
- TFs form complexes, **bind** to **modules** of TFBMs
- **Relative distance** of TFBMs to transcription start and among TFBMs is relevant



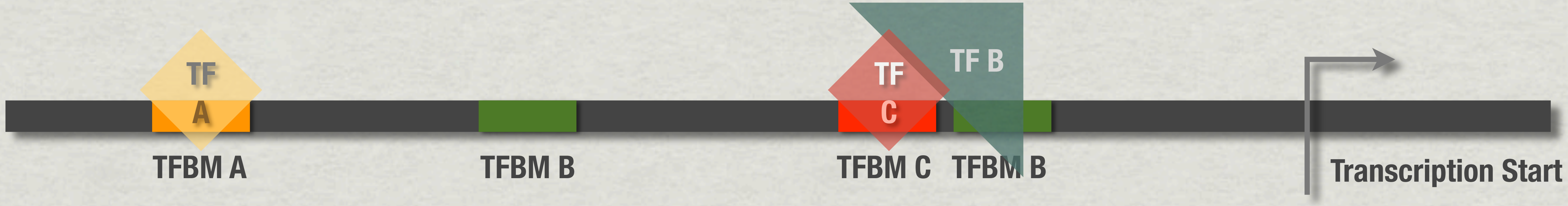
Transcriptional Regulation

- **Transcription factors** (TFs) are proteins regulating transcription (activate, repress, ...)
- TFs bind to euchromatic sequences, *e.g.* promoter regions, **regulatory regions** in introns, ...
- TFs recognize conserved **binding motifs** (TFBMs)
- TFs form complexes, **bind** to **modules** of TFBMs
- **Relative distance** of TFBMs to transcription start and among TFBMs is relevant



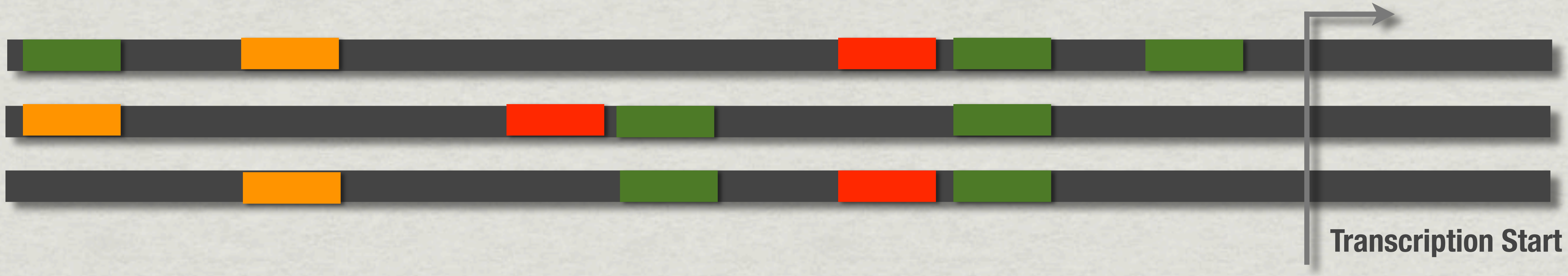
Transcriptional Regulation

- **Transcription factors** (TFs) are proteins regulating transcription (activate, repress, ...)
- TFs bind to euchromatic sequences, *e.g.* promoter regions, **regulatory regions** in introns, ...
- TFs recognize conserved **binding motifs** (TFBMs)
- TFs form complexes, **bind** to **modules** of TFBMs
- **Relative distance** of TFBMs to transcription start and among TFBMs is relevant



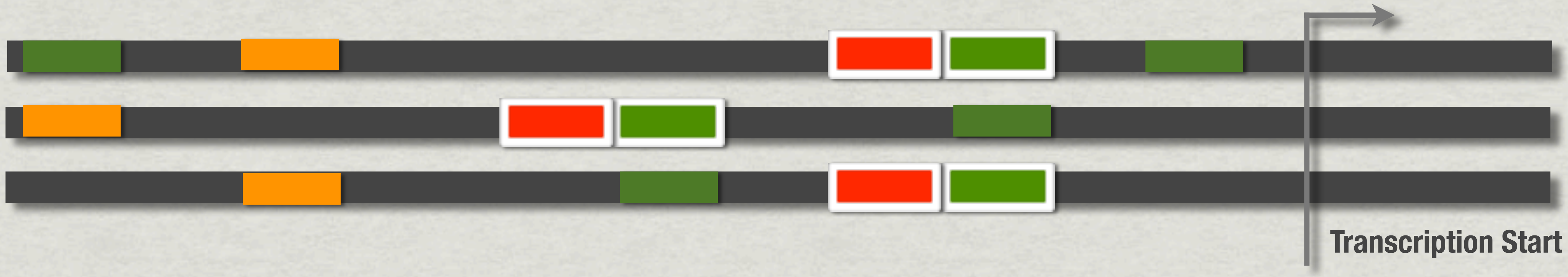
Machine Learning Method

- SVM for a **2-class problem**: genes are **co-regulated** by the same mechanism or not
- Compare region similarity
  - Straightforward: string kernel looks at whole regulatory region
  - Our method: windows around motif positions, relative distances, conservation



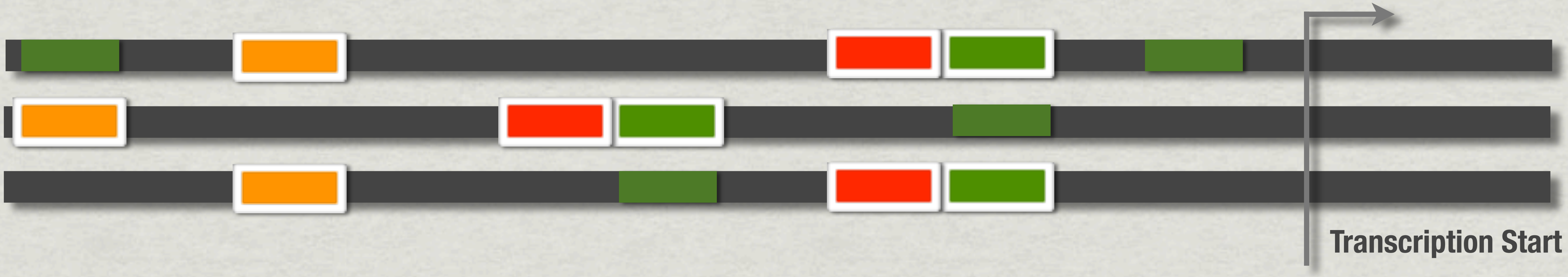
Machine Learning Method

- SVM for a **2-class problem**: genes are **co-regulated** by the same mechanism or not
- Compare region similarity
  - Straightforward: string kernel looks at whole regulatory region
  - Our method: windows around motif positions, relative distances, conservation



Machine Learning Method

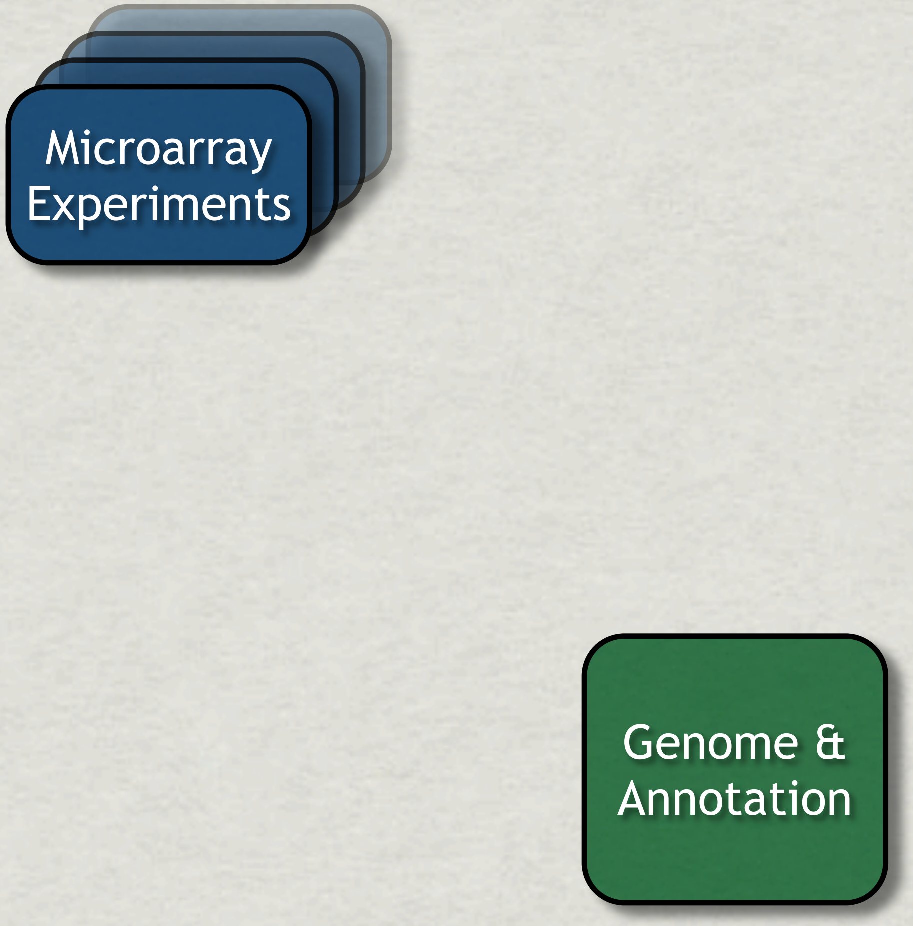
- SVM for a **2-class problem**: genes are **co-regulated** by the same mechanism or not
- Compare region similarity
    - Straightforward: string kernel looks at whole regulatory region
    - Our method: windows around motif positions, relative distances, conservation



Experimental Data

- Microarrays of *Arabidopsis thaliana*
- Statistical methods identify **co-expressed** genes in (several) experimental conditions
- Which are co-regulated?
  Unclear for many genes
  - not on chip
  - change below detection threshold
- **Classifier** needed, train on co-expressed genes
- **Predict co-regulation** for whole genome

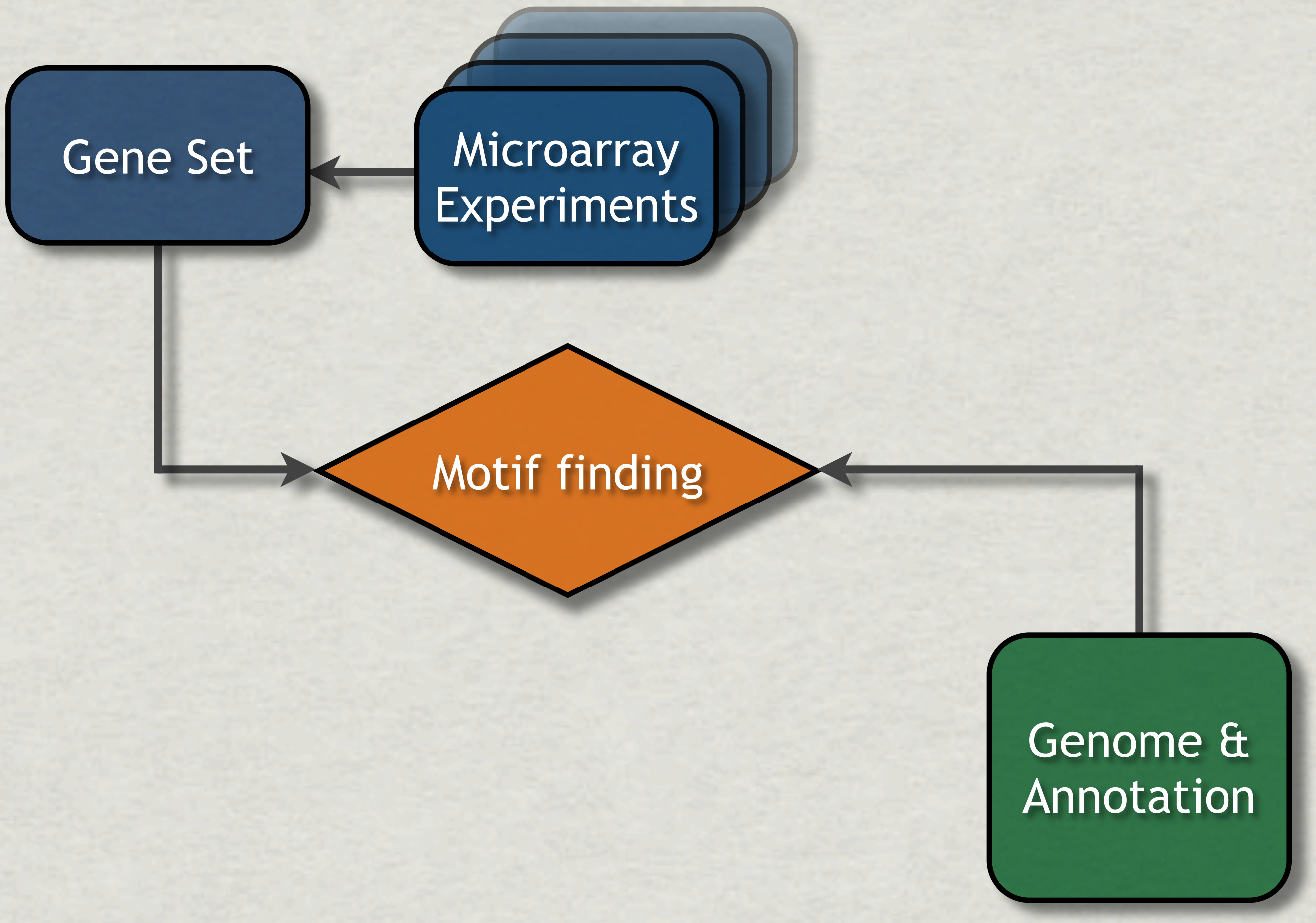
KIRMES Workflow



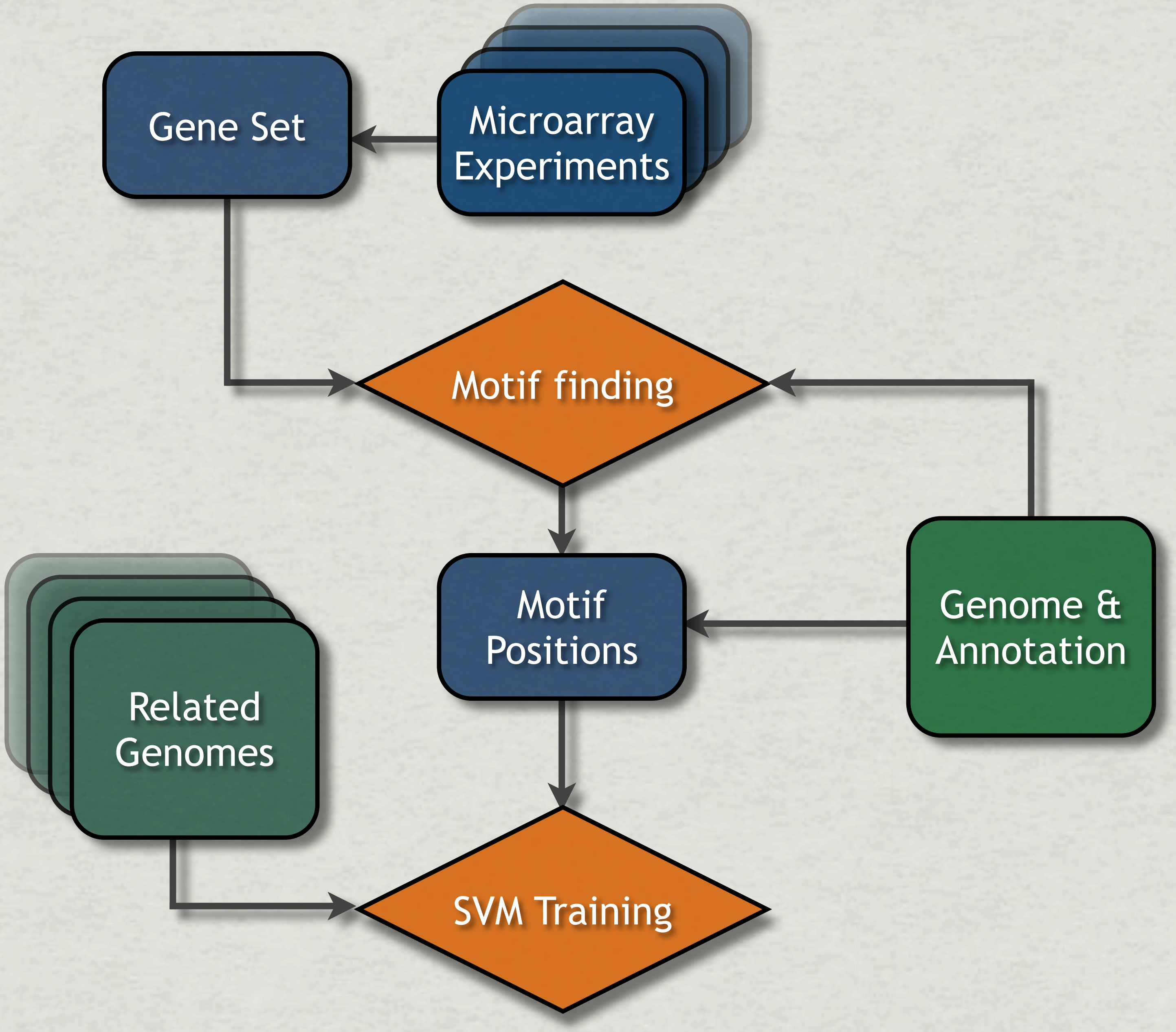
Microarray Experiments

Genome & Annotation

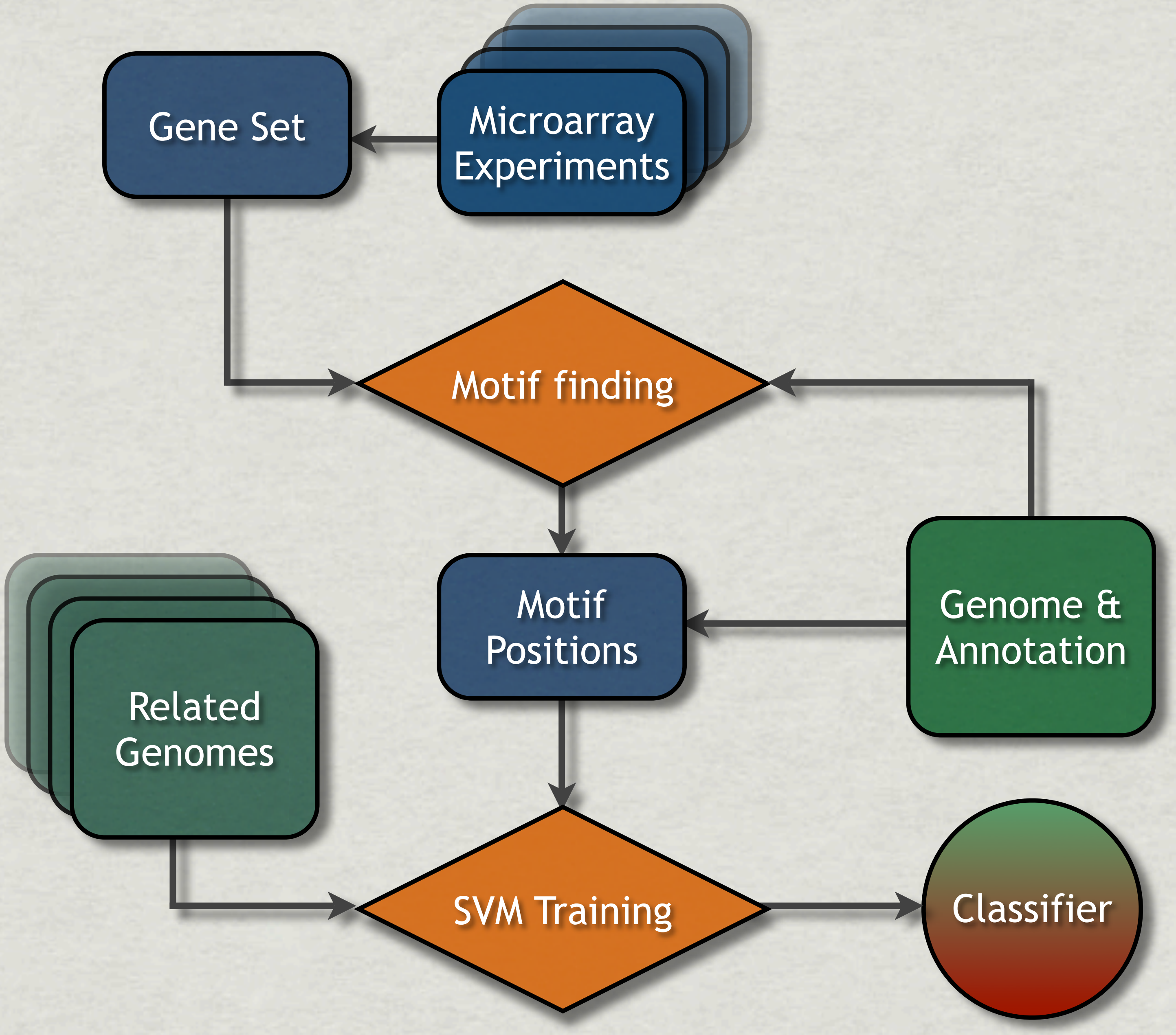
KIRMES Workflow



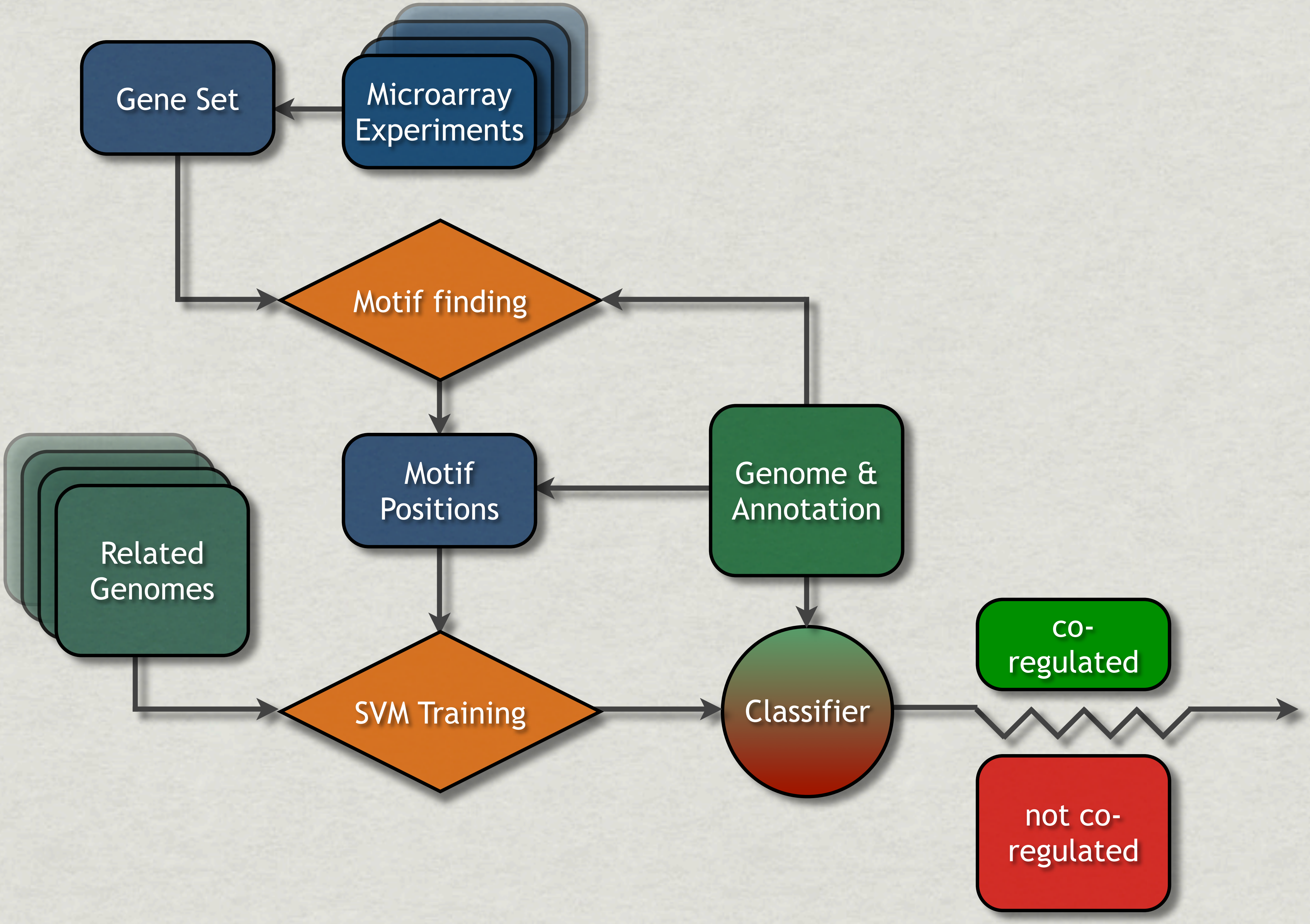
KIRMES Workflow



KIRMES Workflow



KIRMES Workflow



KIRMES Step by Step

1. Obtain sets of co-expressed genes
2. Select regulatory regions
3. Identify over-represented motifs
4. Learn to classify genes as co-regulated
5. Predict co-regulation for all genes

Co-Expressed Genes

1.

- Obtain sets from *e.g.* microarray experiments
- Genes can be up- or down-regulated
- Obtain **negative** control set: genes invariantly expressed at high levels

Co-Expressed Genes

1.

- Obtain sets from *e.g.* microarray experiments
- Genes can be up- or down-regulated
- Obtain **negative** control set: genes invariantly expressed at high levels

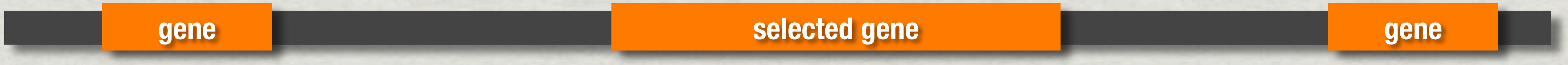
set of gene sequences

Regulatory Regions

- Core **promoter**: 400 bp upstream from transcription start site, up to 3000 bp
- **Untranslated** regions, up- and downstream
- 1st **intron**, all introns, 1st exon, all exons
- **Downstream** region, *e.g.* 500 bp

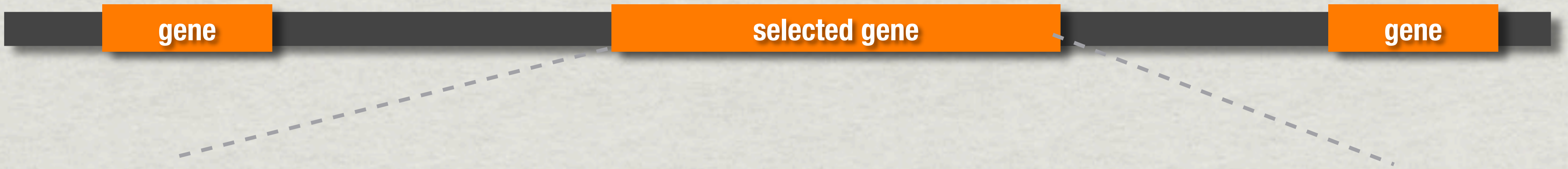
Regulatory Regions

- Core **promoter**: 400 bp upstream from transcription start site, up to 3000 bp
- **Untranslated** regions, up- and downstream
- 1st **intron**, all introns, 1st exon, all exons
- **Downstream** region, *e.g.* 500 bp



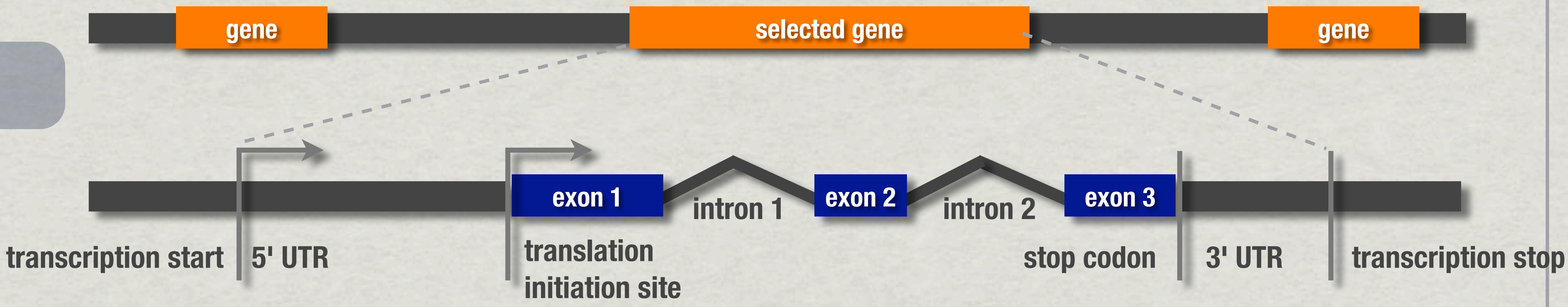
Regulatory Regions

- Core **promoter**: 400 bp upstream from transcription start site, up to 3000 bp
- **Untranslated** regions, up- and downstream
- 1st **intron**, all introns, 1st exon, all exons
- **Downstream** region, *e.g.* 500 bp



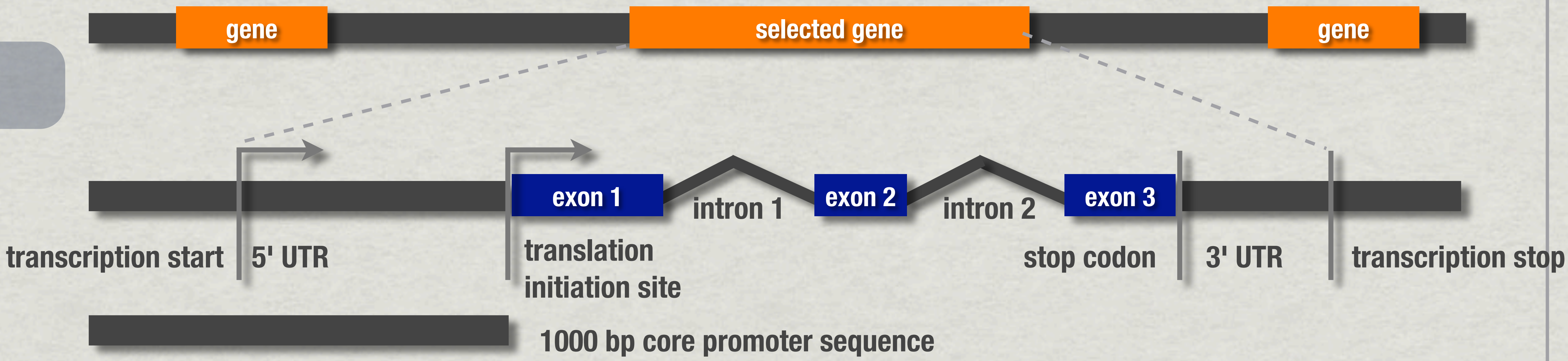
Regulatory Regions

- Core **promoter**: 400 bp upstream from transcription start site, up to 3000 bp
- **Untranslated** regions, up- and downstream
- 1st **intron**, all introns, 1st exon, all exons
- **Downstream** region, *e.g.* 500 bp



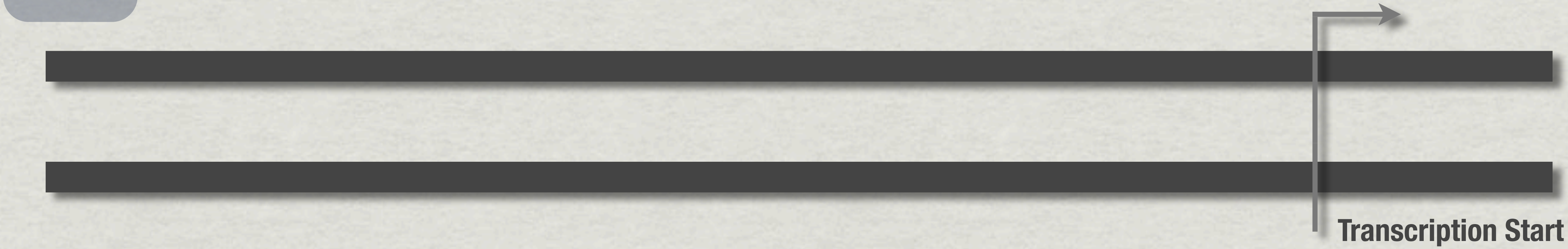
Regulatory Regions

- Core **promoter**: 400 bp upstream from transcription start site, up to 3000 bp
- **Untranslated** regions, up- and downstream
- 1st **intron**, all introns, 1st exon, all exons
- **Downstream** region, *e.g.* 500 bp



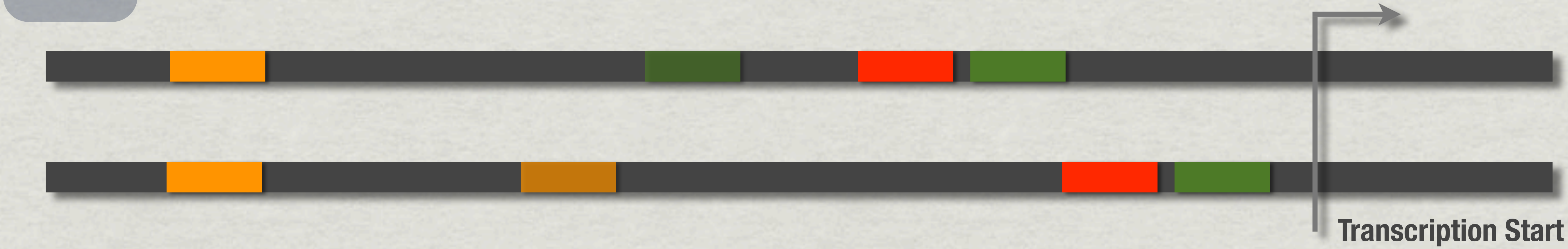
Identify over-represented motifs

- **Gibbs sampling** (Lawrence *et al.*, 1993)
  - identifies over-represented **weight matrices** that characterize TFBMs
- **Oligo counting** (faster)
  - count occurrences of all nucleotide sequences of length six
- Relative position, window around motifs constitute **SVM input vector**



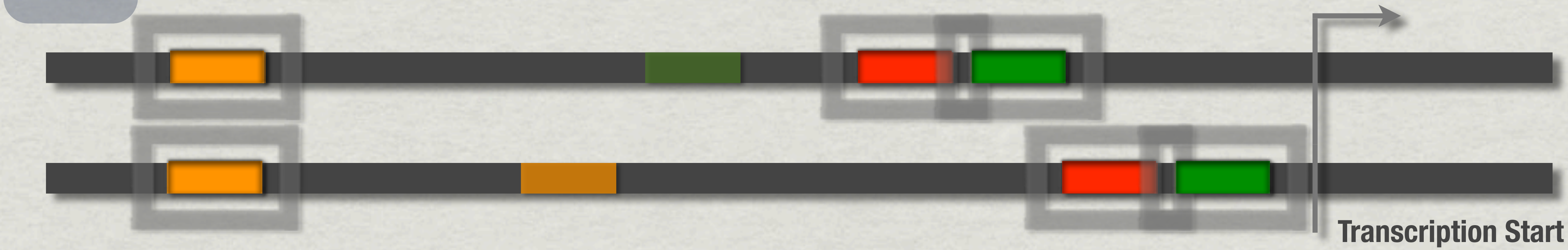
Identify over-represented motifs

- **Gibbs sampling** (Lawrence *et al.*, 1993)
  - identifies over-represented **weight matrices** that characterize TFBMs
- **Oligo counting** (faster)
  - count occurrences of all nucleotide sequences of length six
- Relative position, window around motifs constitute **SVM input vector**



Identify over-represented motifs

- **Gibbs sampling** (Lawrence *et al.*, 1993)
  - identifies over-represented **weight matrices** that characterize TFBMs
- **Oligo counting** (faster)
  - count occurrences of all nucleotide sequences of length six
- Relative position, window around motifs constitute **SVM input vector**

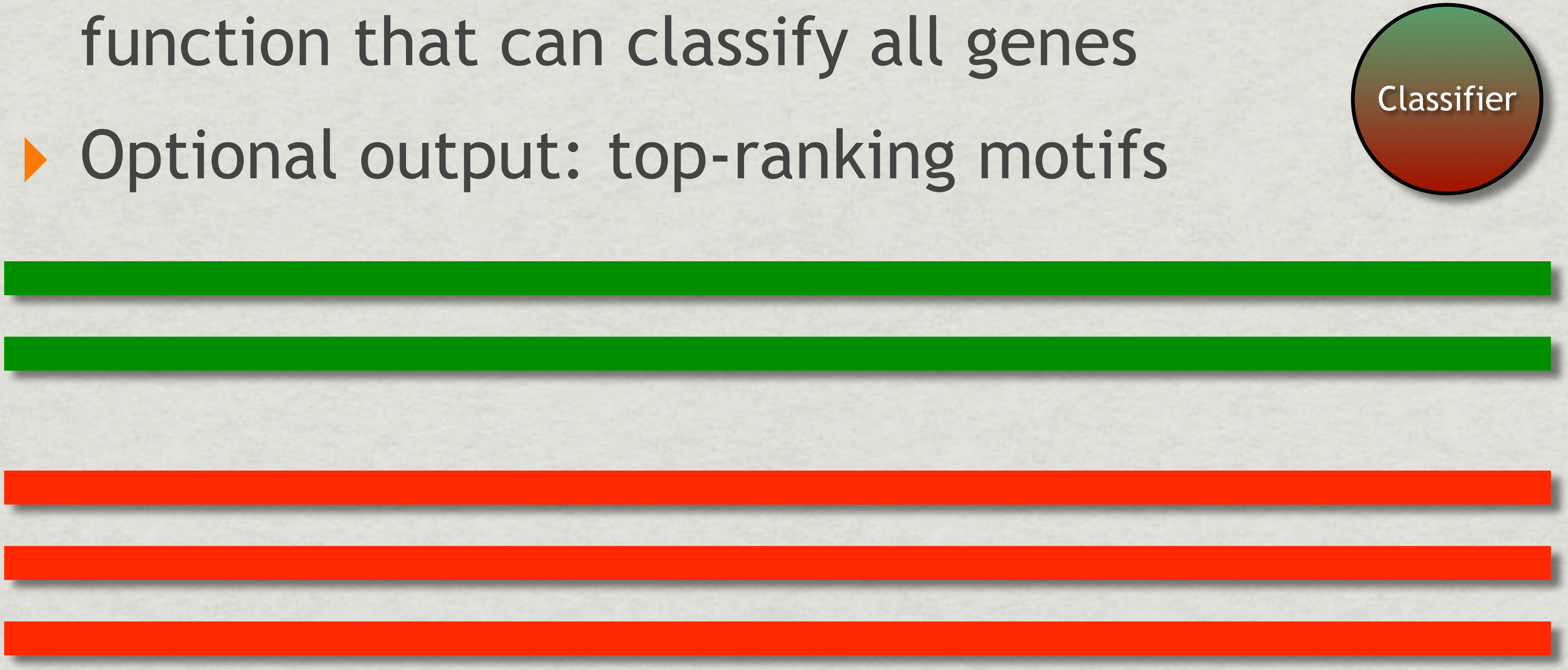


Learn to Classify Genes

- Labeled training data sets with genes **coexpressed** on microarrays
- **Classification** output:
  function that can classify all genes
- Optional output: top-ranking motifs

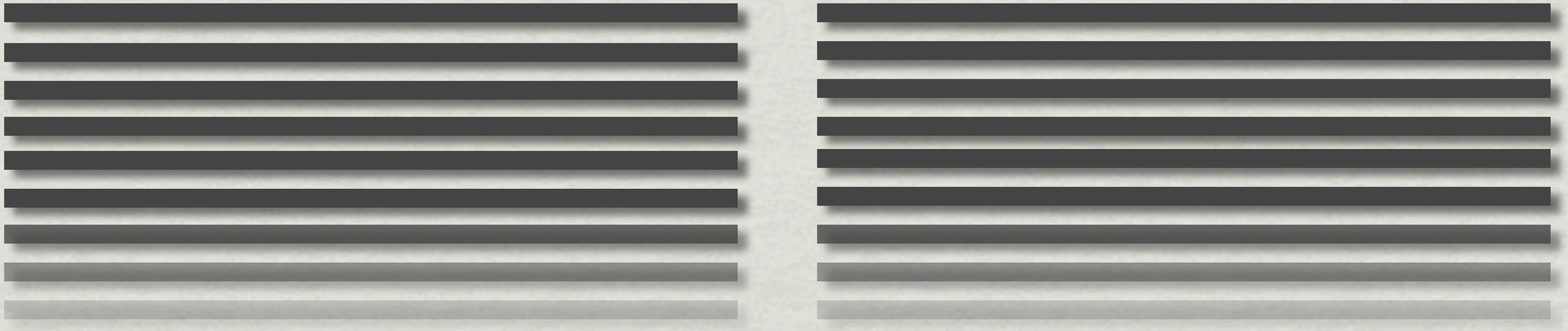
Learn to Classify Genes

- Labeled training data sets with genes **coexpressed** on microarrays
- **Classification** output: function that can classify all genes
- Optional output: top-ranking motifs



Predict Co-Regulation

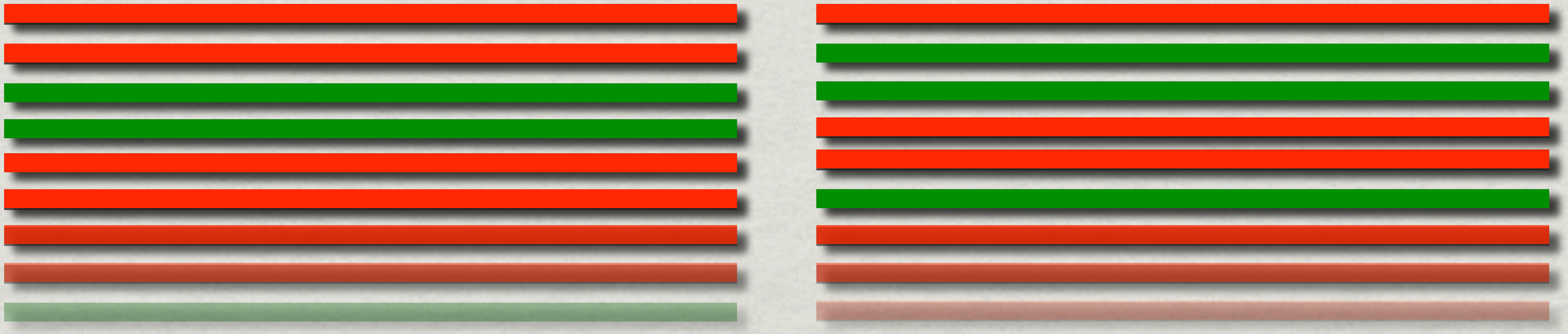
- Learning step creates classifier
- **Prediction of co-regulation**
  - All other genes, not on chip
  - Different regions than training data



Predict Co-Regulation

- Learning step creates classifier
- **Prediction of co-regulation**
  - All other genes, not on chip
  - Different regions than training data

Classifier



Machine Learning Methods

- String kernel: **Weighted Degree kernel**
- Uses whole regulatory region: can't identify modules, no positional interdependence

```
...AGTCAGATAGAGGACATCAGTAGACAGATTAAA...
       |||||||      ||  ||     |||
...TTATAGATAGACAAAGACATCAGTAGACTTATT...
```

Machine Learning Methods

- String kernel: **Weighted Degree kernel**
- Uses whole regulatory region: can't identify modules, no positional interdependence

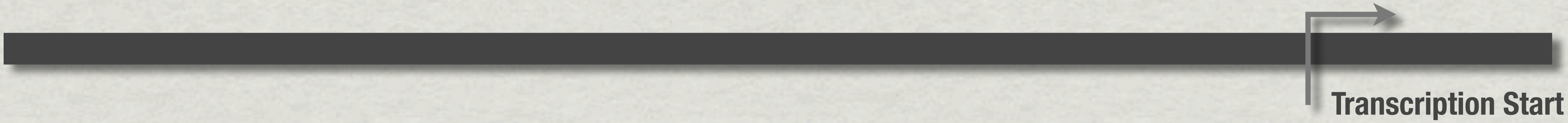
```
...AGTCAGATAGAGGACATCAGTAGACAGATTAAA...
       |||||||      ||  ||     |||
...TTATAGATAGACAAAGACATCAGTAGACTTATT...
```

- **With shifts:** still limited, no modules

```
...CGAACGCTACGTATTTTAGTCGGATTCGC...
    \\\\\        //////   ///
...TCGAACGAAAGGTTTTAGCCTGATGACGG...
```

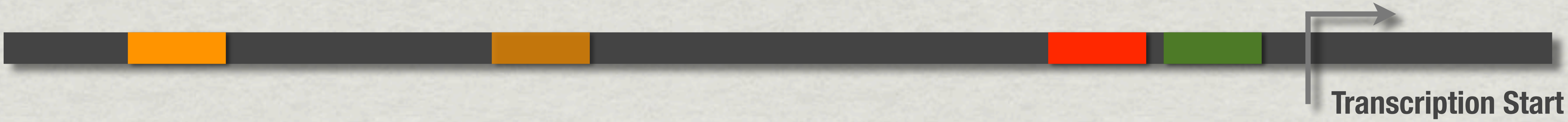
Novel SVM Kernels

- Idea: compare **characteristic** sequence
- **Motif finder** identifies TFBMs over background
- Novel kernels:
  - Relative **positional** information
  - **Window** around motif positions
  - Sequence **conservation**
- Sum up kernels (concatenates feature space)



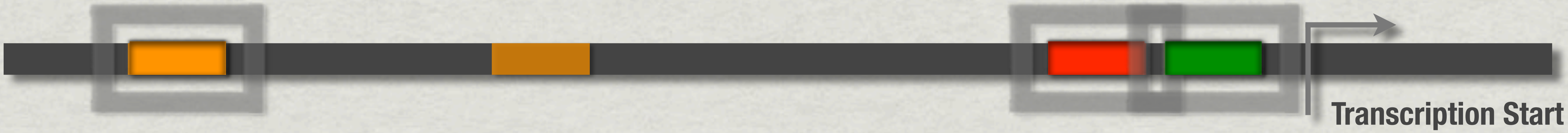
Novel SVM Kernels

- Idea: compare **characteristic** sequence
- **Motif finder** identifies TFBMs over background
- Novel kernels:
    - Relative **positional** information
    - **Window** around motif positions
    - Sequence **conservation**
- Sum up kernels (concatenates feature space)



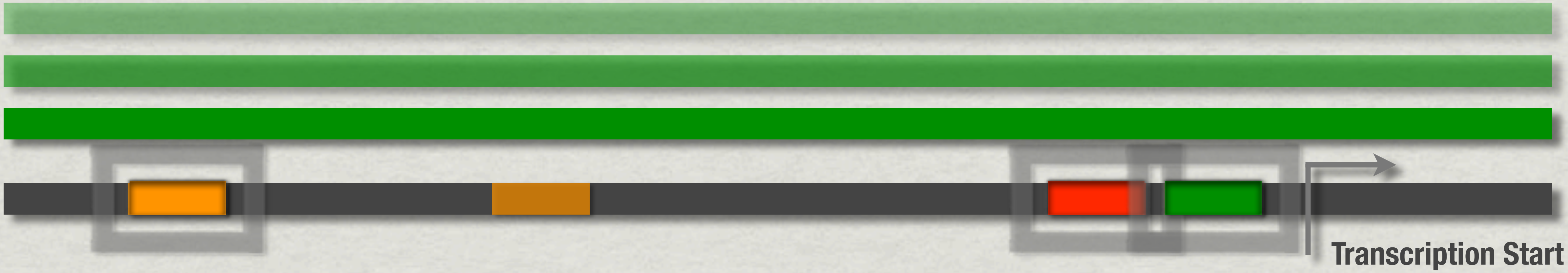
Novel SVM Kernels

- Idea: compare **characteristic** sequence
- **Motif finder** identifies TFBMs over background
- Novel kernels:
  - Relative **positional** information
  - **Window** around motif positions
  - Sequence **conservation**
- Sum up kernels (concatenates feature space)



Novel SVM Kernels

- Idea: compare **characteristic** sequence
- **Motif finder** identifies TFBMs over background
- Novel kernels:
  - Relative **positional** information
  - **Window** around motif positions
  - Sequence **conservation**
- Sum up kernels (concatenates feature space)

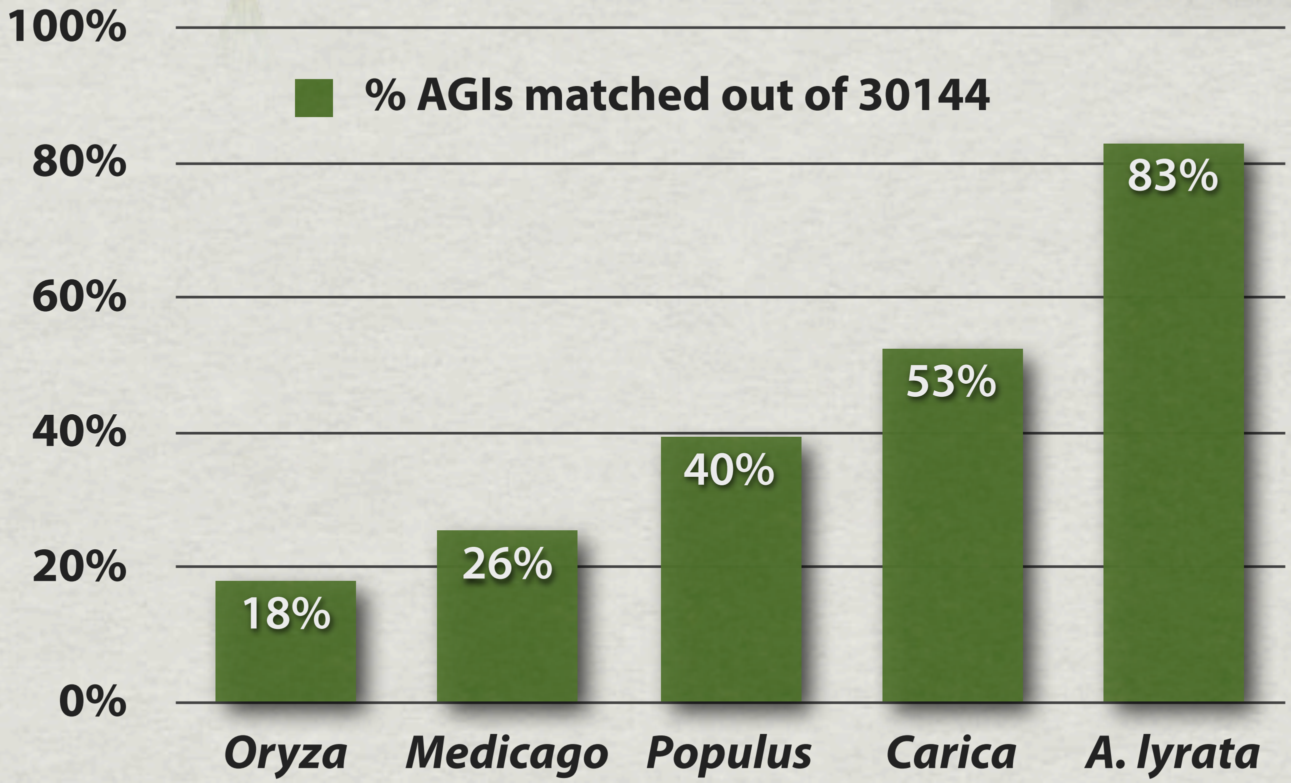


Kernel for Regulatory Modules

- **RBF kernel**
  - Relative position of motif matches to start and among each other
  - Pairwise distances between match positions of motifs
- **Weighted degree kernel with shifts**
  - Set of windows around best-matching motifs
  - Highest similarity: all motifs appear in both sequences

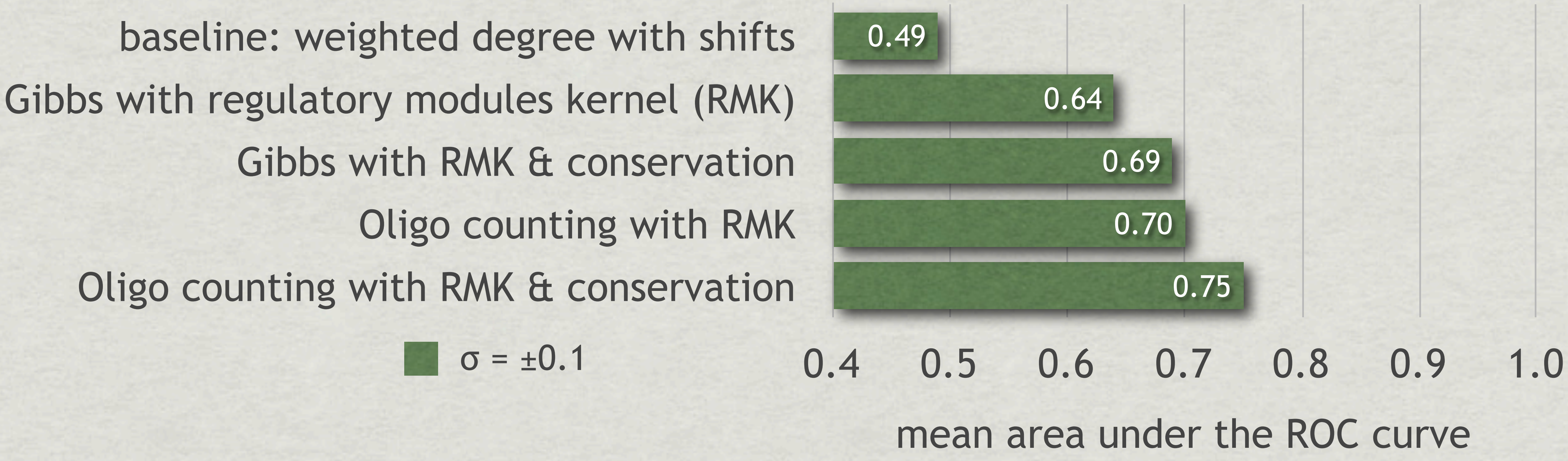
WDS Kernel with Conservation

- WDS kernel extended with conservation info
- Helpful, not essential
- *A. thaliana* AGI regions aligned with other plants
- How many of *A. thaliana*'s genes were aligned?



Results (Comparison)

- Sets of co-expressed genes from *A. thaliana* under different conditions, **auROC curve** evaluation
- Baseline method performs close to random guessing
- Oligo counting outperforms Gibbs



Results (Biology)

- **WUSCHEL**, a key player in *Arabidopsis* stem cell maintenance
- Binding site part of KIRMES output
- TFBM verified* (SELEX, gel shift)
- Regulatory network: ChIP on chip binding
  - **PWM**: 17% of genes
  - **KIRMES**: 64% of genes

A. thaliana wild type WUSCHEL overexpressing mutant

*W. Busch *et al.*, in preparation

Discussion

- Powerful approach
  - exploits relationships between motifs
  - uses **modules** for prediction
- **Conservation** is useful
- **Oligo counting** works surprisingly well

- Future directions:
  - Integrate binding data
  - Compare against established methods
  - **Visualize** modules

Acknowledgements

- **Gunnar Rätsch** and AG Rätsch/MLB at FML
- **Wolfgang Busch, Jan Lohmann** and the JLab
- **Oliver Kohlbacher**, Div. SBS at University of Tübingen
- Detlef Weigel and Dept. VI at MPI for Dev. Biology

Thank you for your attention!

Friedrich Miescher Laboratory
of the Max Planck Society

MAX-PLANCK-GESELLSCHAFT

Eberhard Karls
Universität
Tübingen

ZBIT