



DISCOVER: a feature-based discriminative method for motif search in complex genomes

Wenjie Fu ^{*}, Pradipta Ray ^{*}, Eric Xing

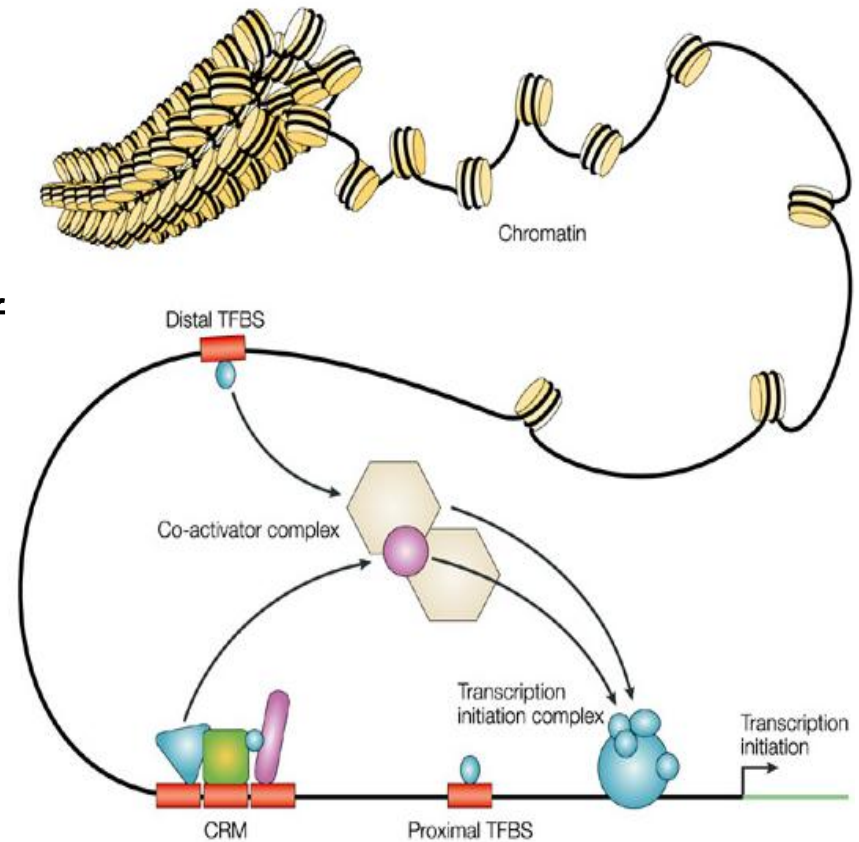
wenjief, pray, epxing@cs.cmu.edu

School of Computer Science
Carnegie Mellon University

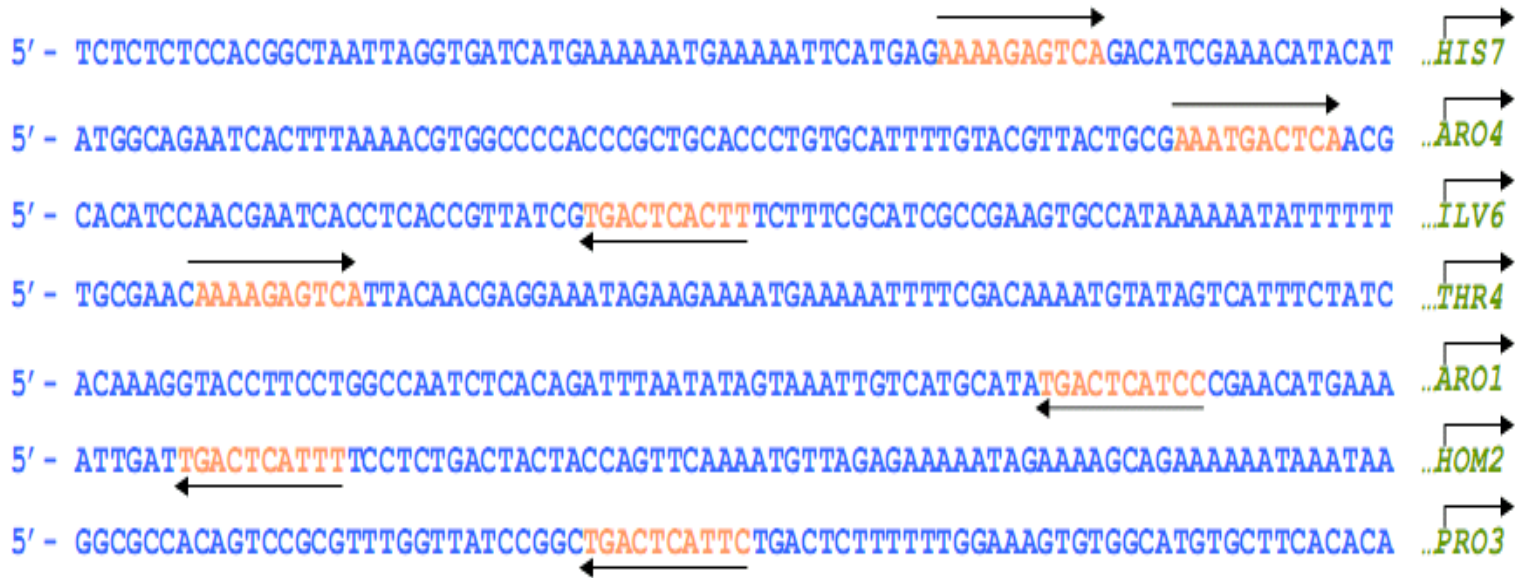
^{*} co-first authors

The problem

- **Transcription factors:** proteins which bind to the DNA at specific locations to aid or repress protein transcription
- **Learning patterns** indicative of binding sites and **predicting putative binding sites** using the learnt patterns
- Typically occur in clusters (**cis-regulatory modules** or CRM) and are noisy copies of each other.



The problem



Organization

Motivation

Schema

Empirical analysis

Results

Conclusion

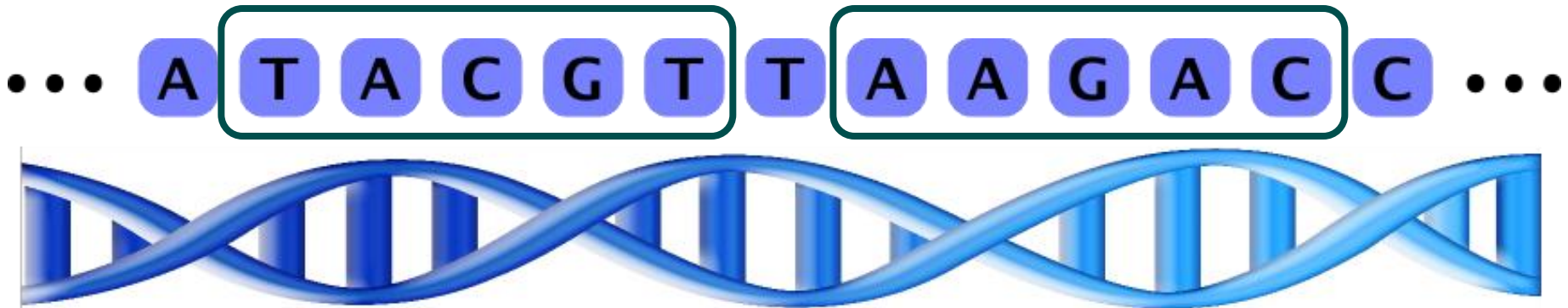
Computational formulation

- Given a sequence of nucleotides,



Computational formulation

- Given a sequence of nucleotides, identify subsequences where the transcription factor binds



Supervised vs unsupervised

- Supervised TFBS finding

REGULATORY DNA SEQUENCE

```
ATTGATGTACATTTTGAC  
TCTTGACCAGGTTGACTT  
GATTTGTGTCATTACTCC  
AGCCTGCACAATTTACGA
```

+

TRAINING BINDING SITE DATA

```
ATTGATGTACATTTTGAC  
TCTTGACCAGGTTGACTT  
GATTTGTGTCATTACTCC  
AGCCTGCACAATTTACGA
```

→

NEW BINDING SITE PREDICTIONS

```
ATTGATGTACATTTTGAC  
TCTTGACCAGGTTGACTT  
GATTTGTGTCATTACTCC  
AGCCTGCACAATTTACGA
```

- Unsupervised TFBS finding

REGULATORY DNA SEQUENCE

```
ATTGATGTACATTTTGAC  
TCTTGACCAGGTTGACTT  
GATTTGTGTCATTACTCC  
AGCCTGCACAATTTACGA
```

→

NEW BINDING SITE PREDICTIONS

```
ATTGATGTACATTTTGAC  
TCTTGACCAGGTTGACTT  
GATTTGTGTCATTACTCC  
AGCCTGCACAATTTACGA
```

Supervised TFBS prediction

Traditionally modelled as fixed width, table of position-specific frequencies by creating a multiple sequence



Supervised TFBS prediction

Each instance is called a **motif**, and the ordered frequency table is called the **Position Weight Matrix (PWM)**

	1	2	3	4	5	6	7	8	9	10
T	G	T	A	A	T	T	G	C	T	
C	C	T	A	A	T	T	G	T	G	
G	T	T	A	A	T	T	G	A	C	
T	T	T	A	A	T	T	G	A	C	
G	C	T	A	A	T	T	G	G	C	
T	T	T	A	A	T	G	G	C	C	
T	T	G	A	A	T	T	G	C	C	
T	C	G	A	A	T	T	G	T	C	
G	T	T	A	A	G	T	G	C	T	
G	T	T	A	A	G	T	G	C	C	
A	T	T	A	A	T	T	C	C	T	
T	C	T	A	A	T	T	A	G	C	
T	T	T	A	A	T	T	T	G	T	
C	A	T	A	A	T	T	T	T	T	
G	G	T	A	A	T	A	T	A	G	
G	C	T	A	A	T	A	A	A	A	
G	C	T	A	A	T	G	A	G	C	
G	T	T	A	A	T	G	A	T	C	
T	C	T	C	A	T	G	T	G	T	
T	T	T	T	A	T	G	A	C	C	
T	T	T	T	A	T	G	T	G	T	
G	G	T	A	A	A	A	G	A	T	

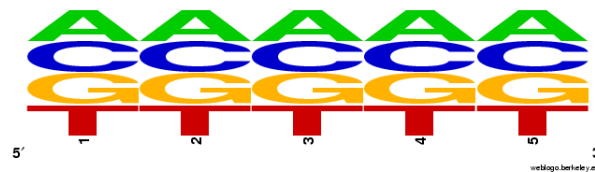


Traditional PWM scanning

$6 * 10^{-8}$ $2 * 10^{-4}$ $9 * 10^{-8}$



VS

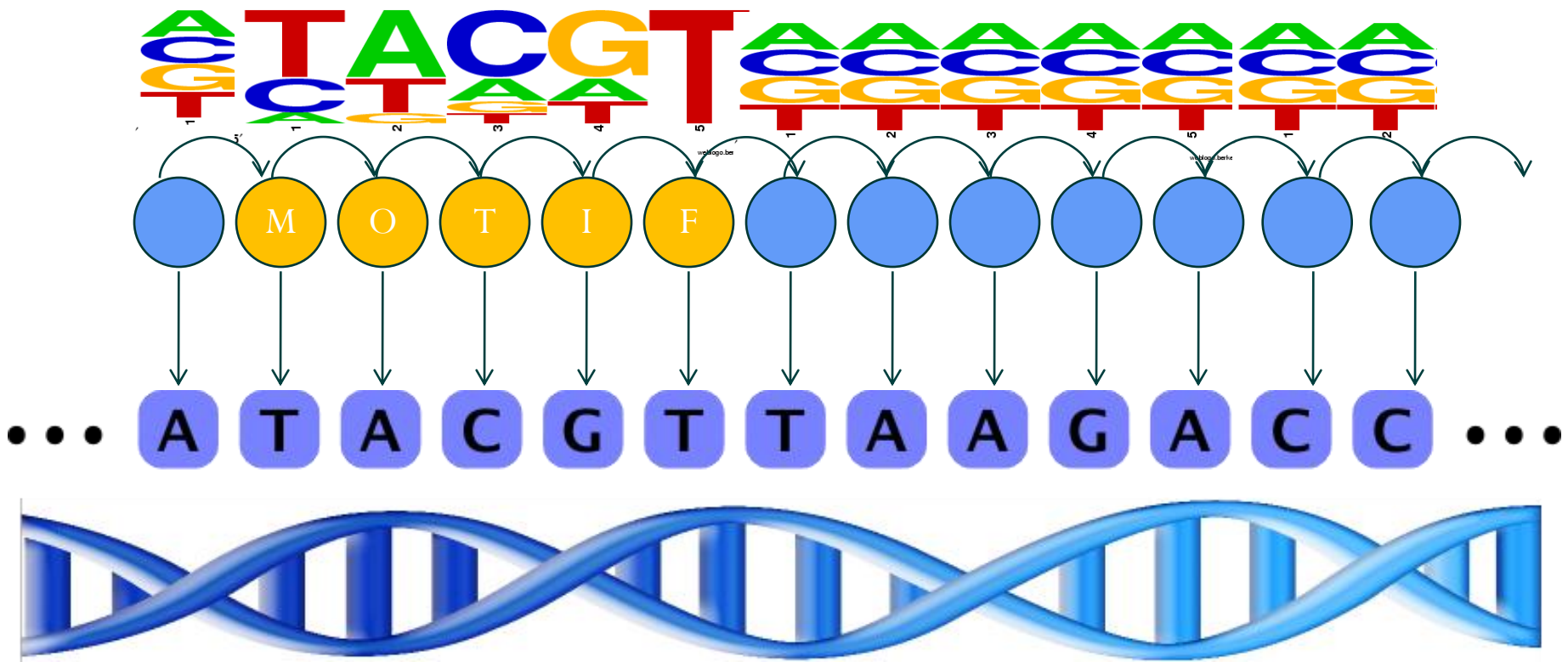


Traditional PWM scanning

- Typically generates a large number of false positives, as the motifs are noisy copies of each other



Generative models : HMMs



Generative models : HMMs

- May tune to the noise instead of the signal, for noisy data
- Performance saturation to some degree : sophisticated models like Baycis [Lin et al '08] have not improved on the state of the art significantly
- Difficult to include a variety of evidence, especially continuous ones

Diverse sources of evidence

- Multi-species phylogenetic motif finding [Loots et al 2002, Moses et al 2004, Ray et al 2008]
- Non PWM Genetic data
 - Distance from TSS [Sinha et al 2008]
 - Distance between TFBSs [Ray et al 2008]
- Epigenetic data
 - Nucleosome binding scores [Segal et al 2006, Narlikar et al 2007]
- Combining features [Sharon and Segal 2007, Ernst 2008]

DISCOVER

DIScriminative COnditional random field for motif recoVERY in metazoan genomes

- Conditional random fields : powerful **integrational device** where features (evidence like phylogeny, proximity to transcribed regions, nucleosome binding affinity score, etc) can be added at will
- Discriminative model
 - **Maximizes the conditional probability** of the label given the sequence, and not the joint probability of both
- Feature set : carefully chosen from the literature and modelled accordingly

Organization

Motivation

Schema and model

Features and empirical analysis

Results

Conclusion

Input & Output

- Estimation:

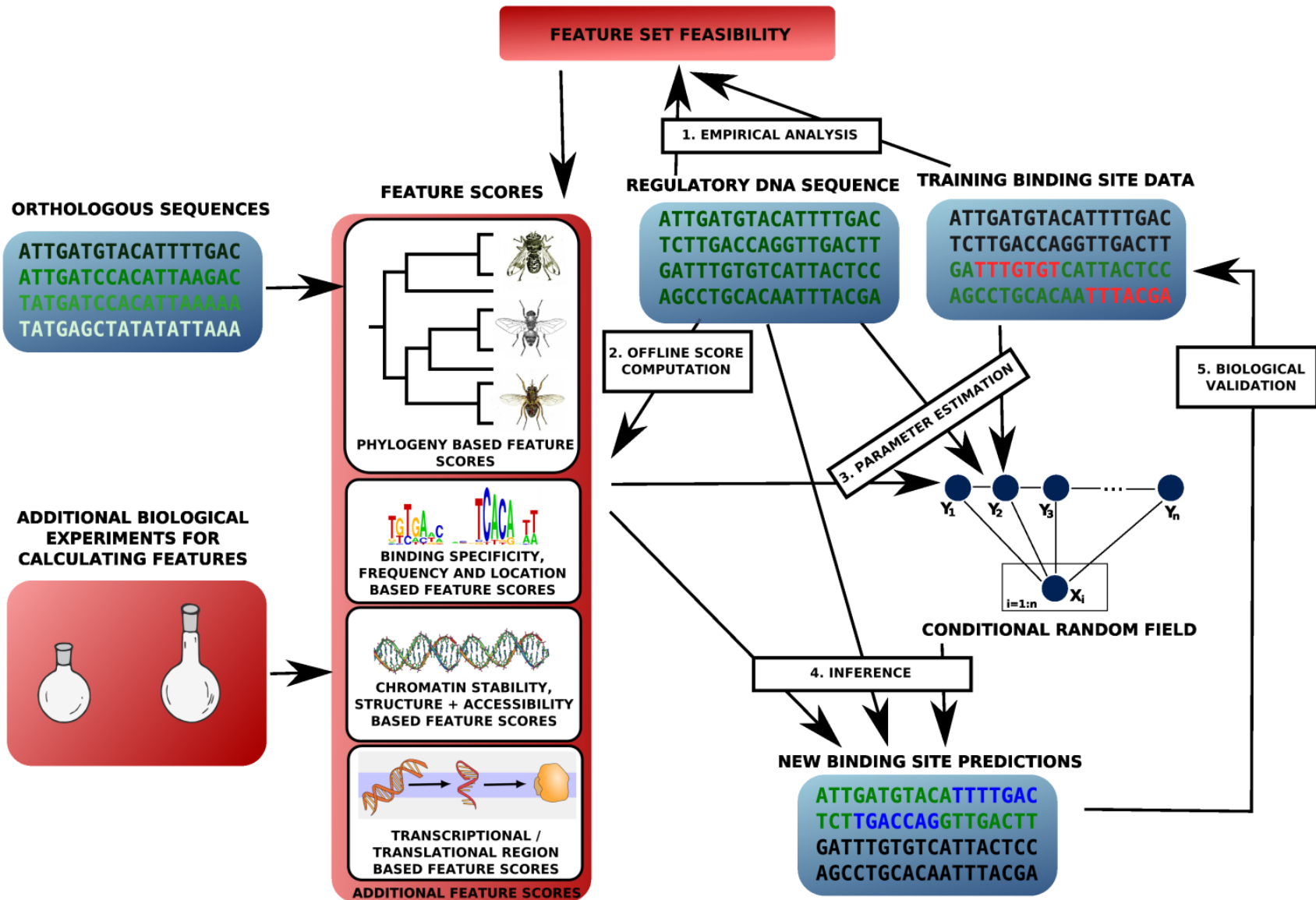
Input : A **set of sequences** which have corresponding feature values for each nucleotide and **positions of TFBS**

Output : Feature weights (ie the learnt conditional probability model)

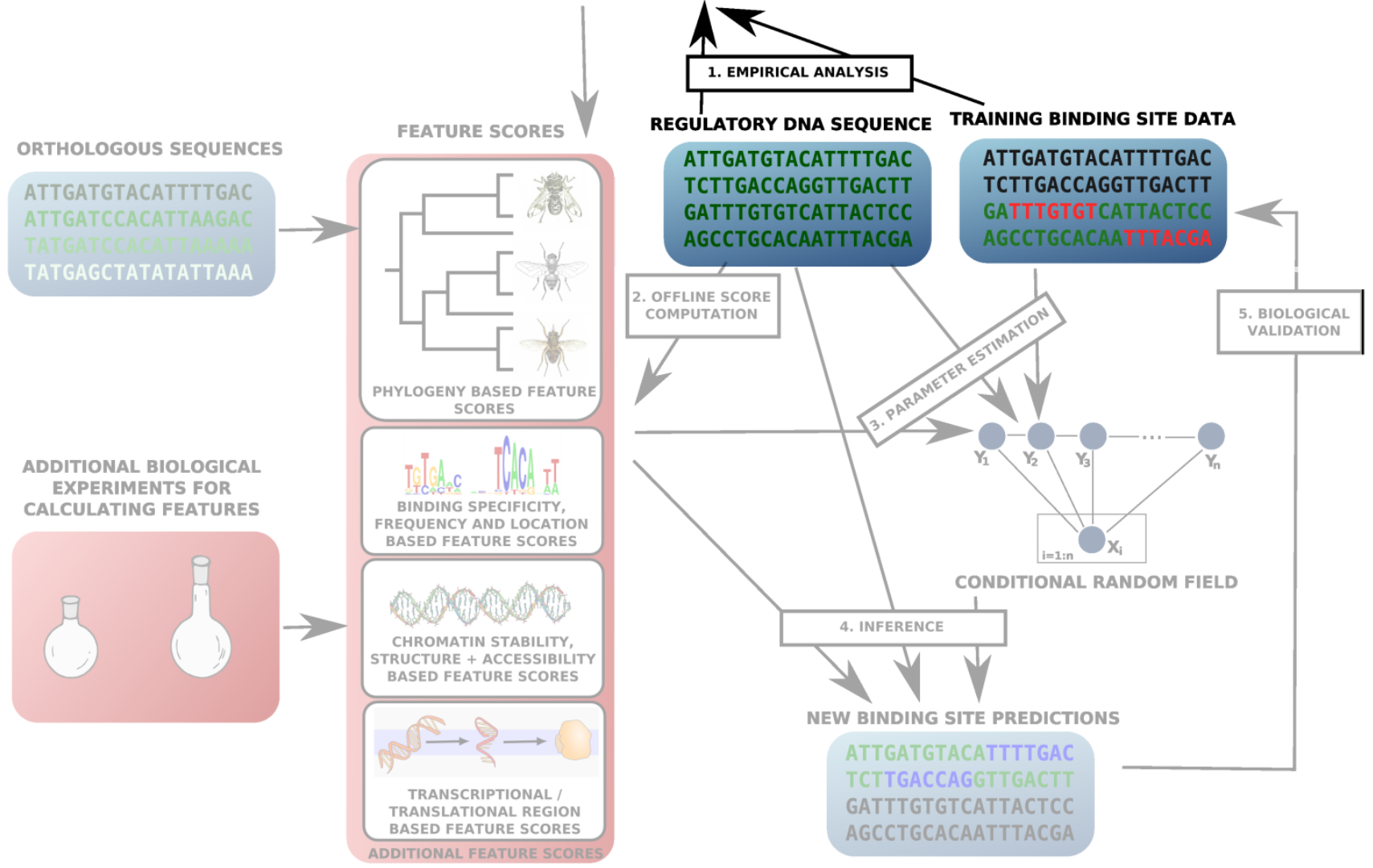
- Inference:

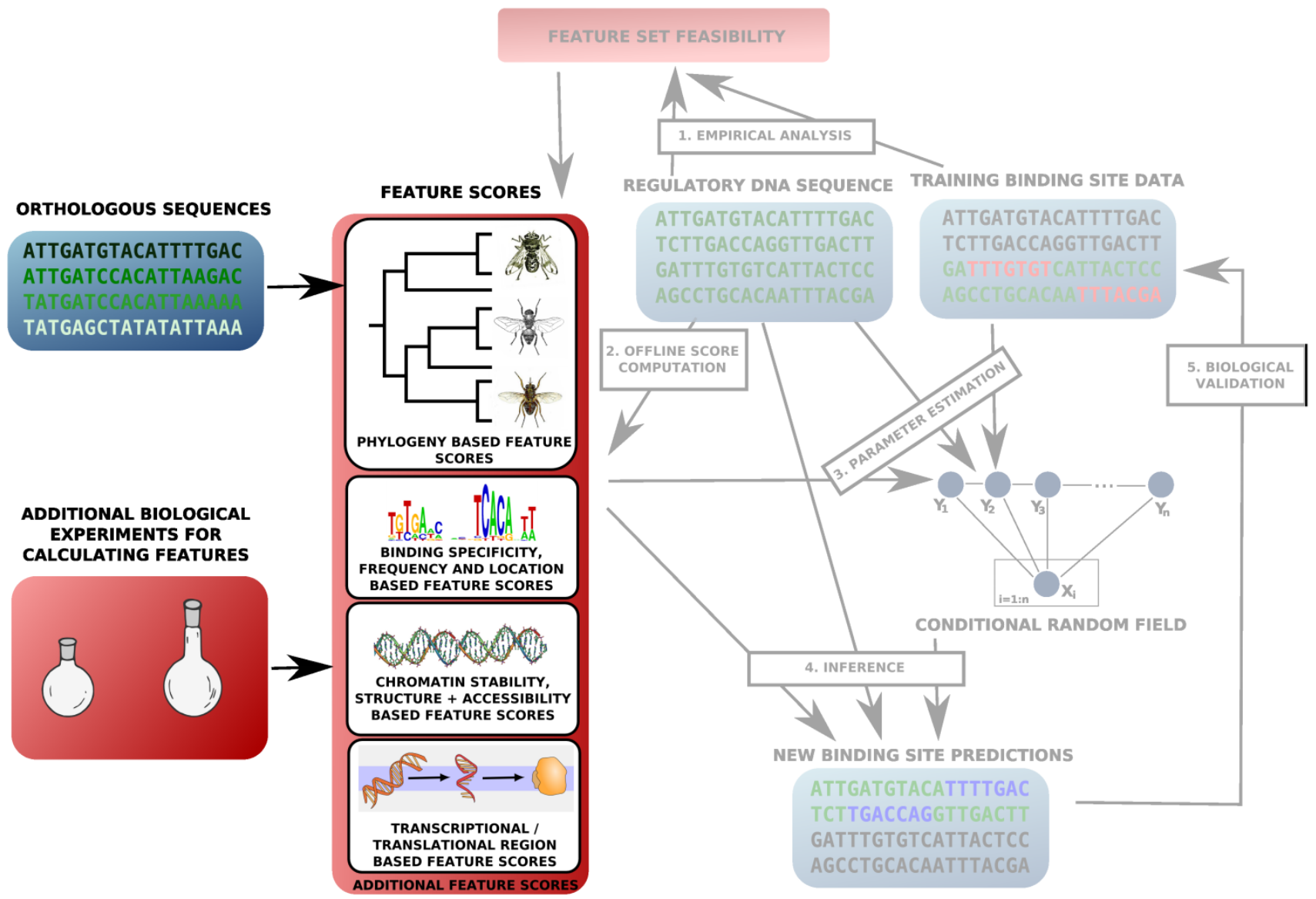
Input: A **set of sequences** which have corresponding feature values for each nucleotide and **learnt model**

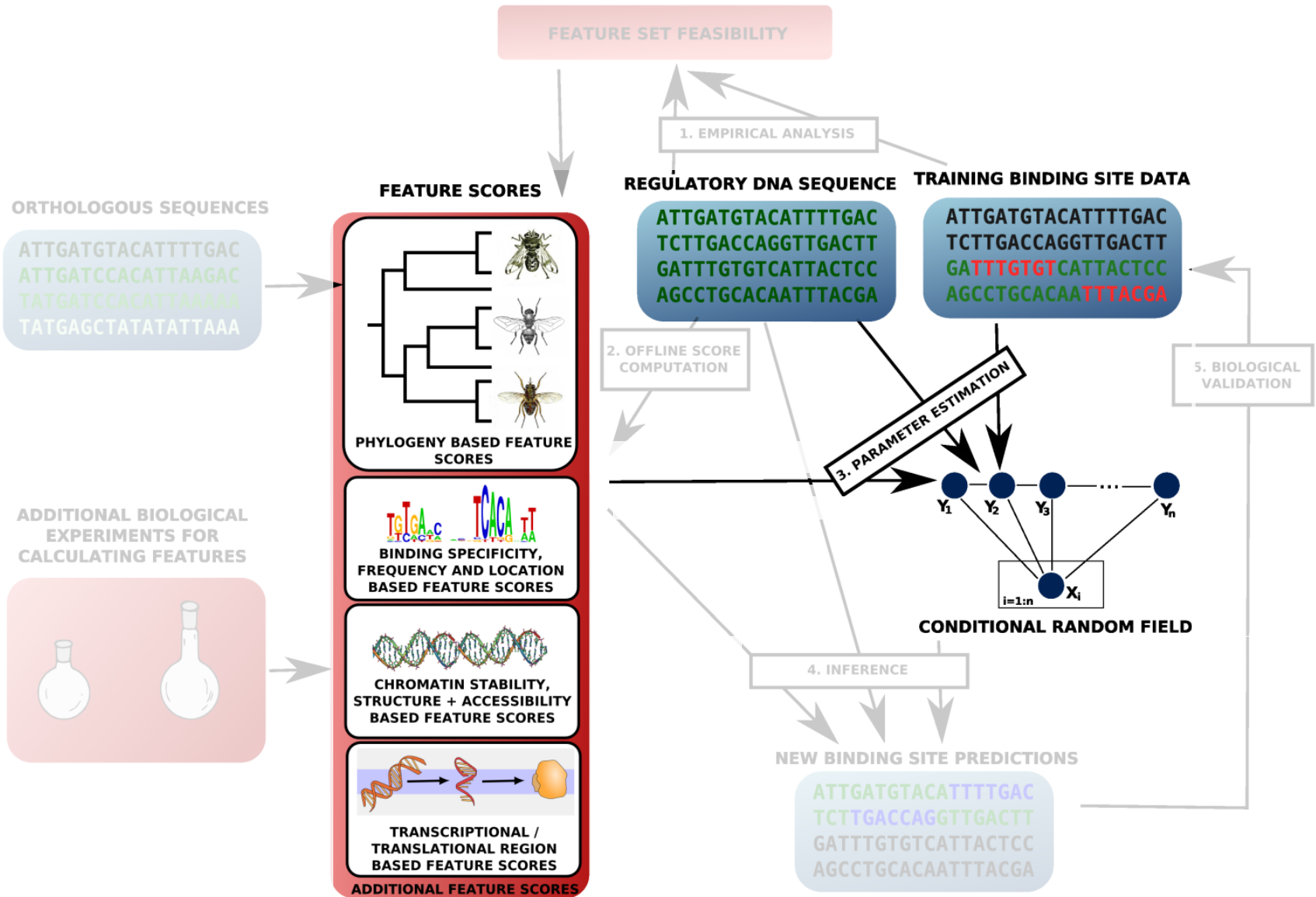
Output : A sequence of ML labels each corresponding to a nucleotide specifying its state : ie, binding site, CRM-background or background

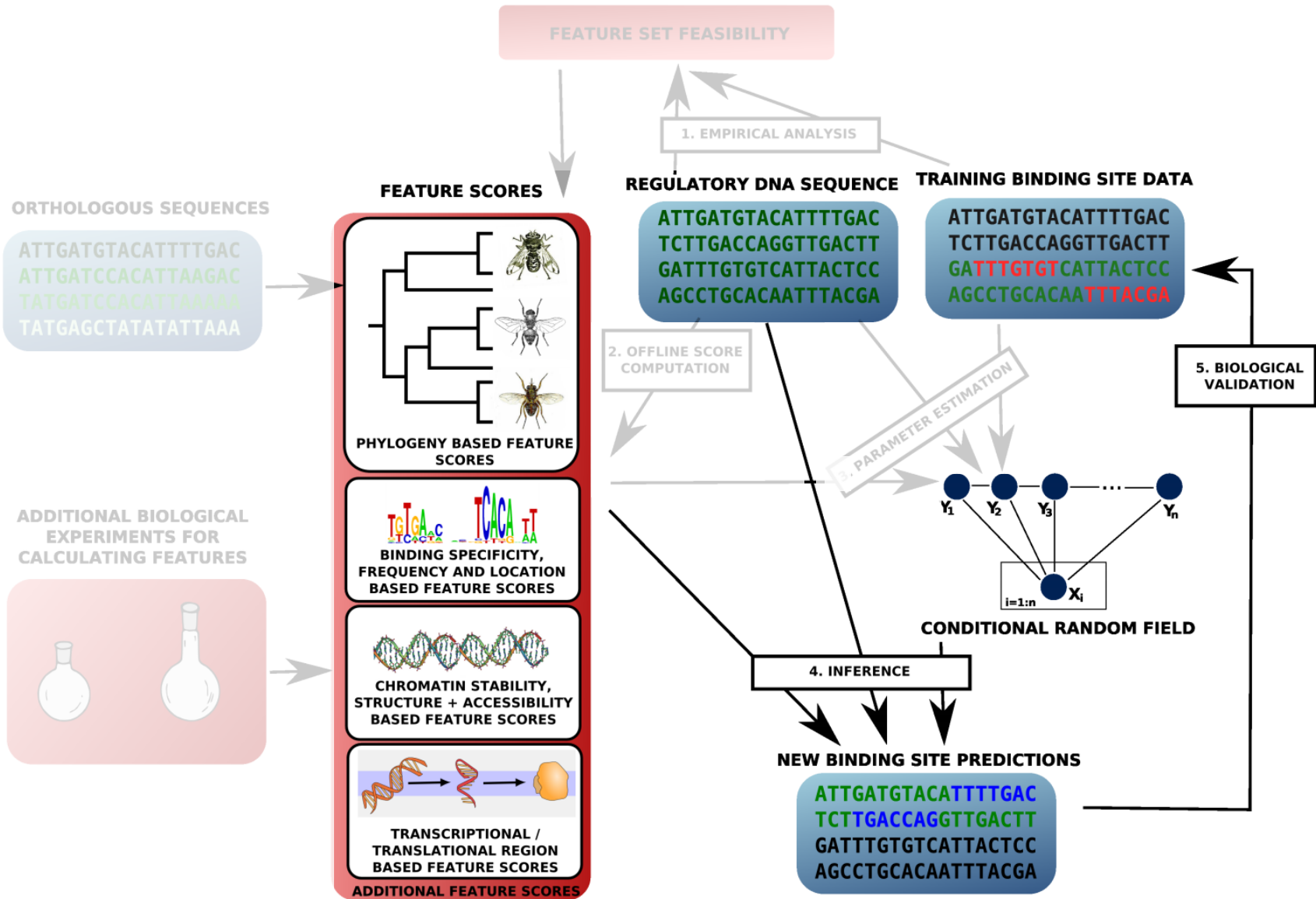


FEATURE SET FEASIBILITY









Estimation

- Conditional likelihood based framework for estimating the model parameters : the weights corresponding to each field
- Converted to a **convex optimization problem**
 \mathbf{x} : nucleotides, \mathbf{y} : state labels
- A quasi-Newton method is applied

$$\hat{\lambda} = \arg \max_{\lambda} L(\lambda | \mathbf{y}, \mathbf{x})$$

where $L(\lambda | \mathbf{y}, \mathbf{x}) = P(\mathbf{y} | \mathbf{x}, \lambda)$

Inference

- Use the parameters learnt in the estimation stage for predicting the binding sites
- Corresponding analogs of HMM inference algorithms
- We use marginal decoding, as it allows us to find overlapping motifs

$$\hat{y}_i = \arg \max_{y_i} P(y_i | \mathbf{x}, \lambda)$$

Organization

Motivation

Schema and model

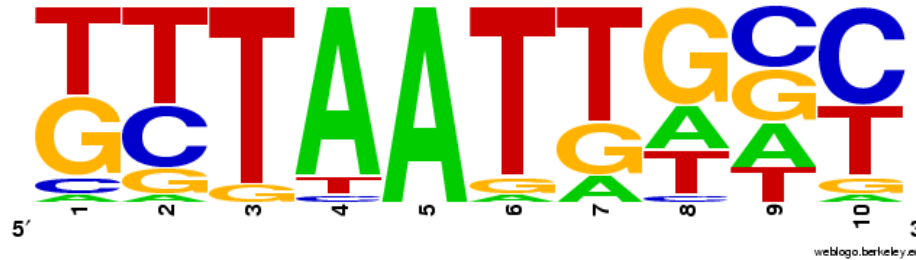
Features and empirical analysis

Results

Conclusion

Sequence-specificity

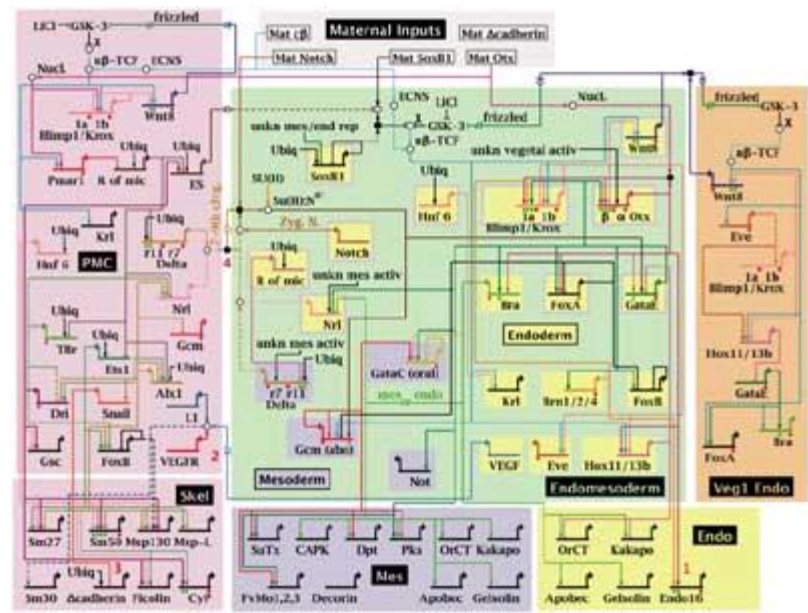
- PWM based score



- Self Complementarity
 - How similar is it to the reverse complement ?

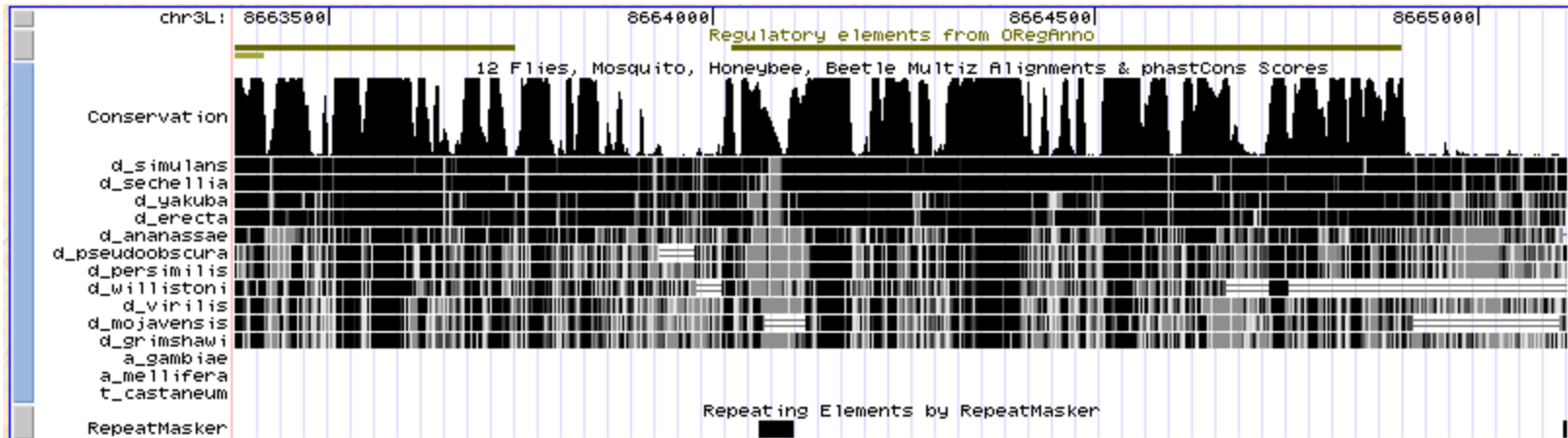
Cis-regulatory grammar

- Transition probabilities between states of the model : modelling the grammar
- Along with the PWM score, correspond to the traditional sources of evidence used by HMM based models



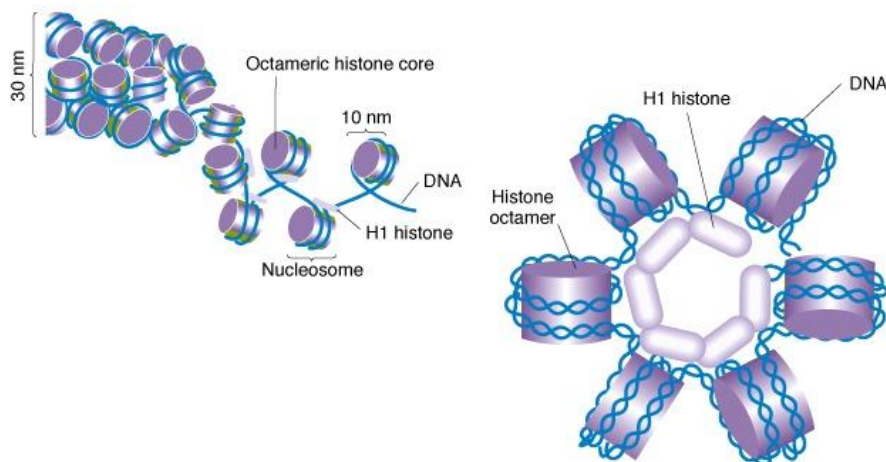
Evolutionary data

- Conservation : function implies conservation from the Neutral Theory
- Presence of repeats : duplicated binding sites or functional turnover ?



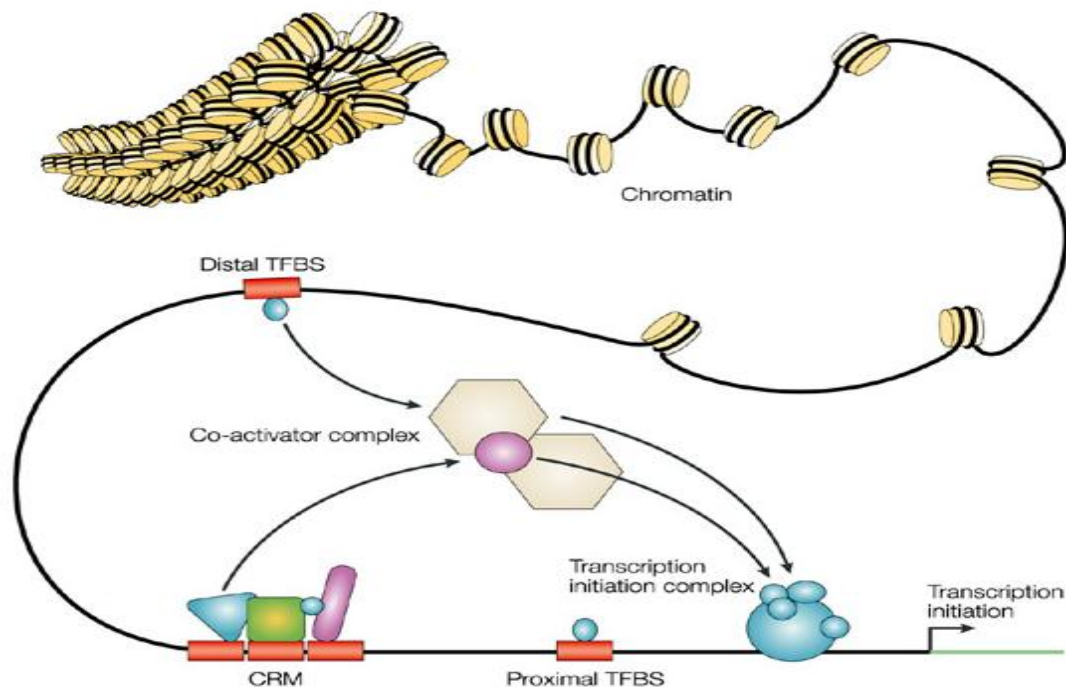
Chromatin stability & accessibility

- G-C content and melting temperature
 - Correlates with chromatin stability, which facilitates TFBS binding
- Nucleosome binding
 - Nucleosomes affect DNA wrapping and thus the accessibility of DNA to TFs

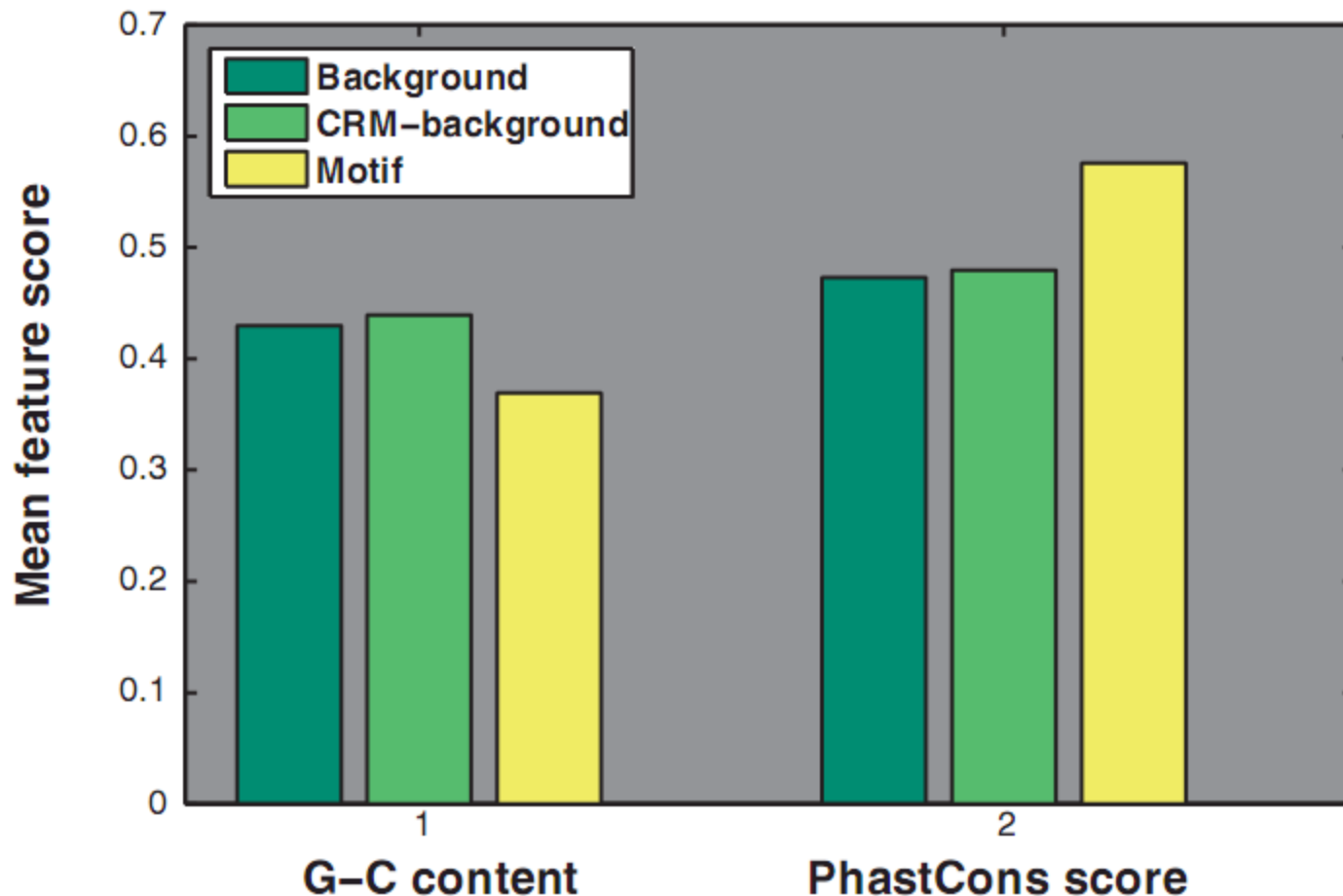


Distance and location based

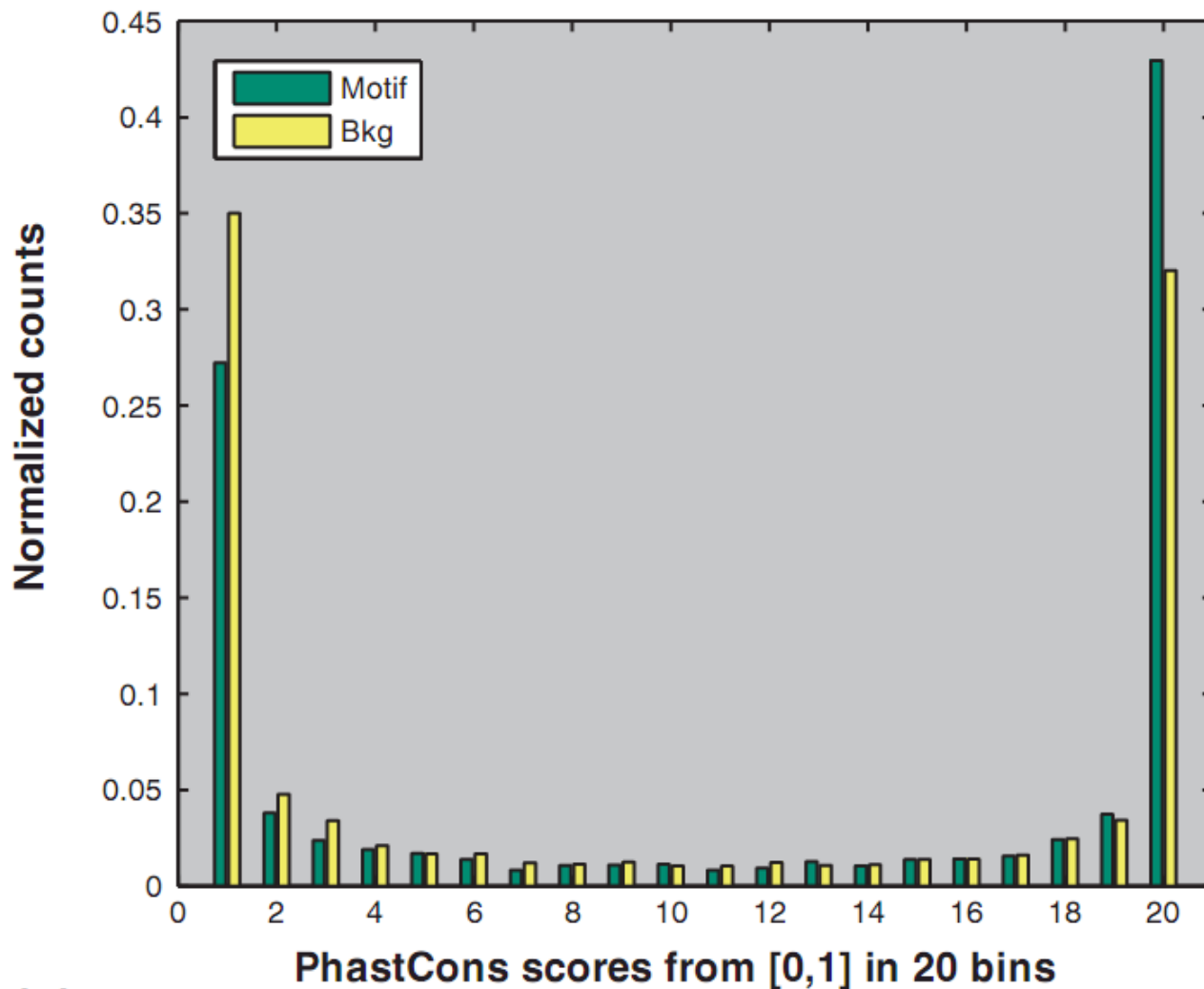
- Distance to TSS
- Presence in 5'UTR, or 3'UTR, or intron



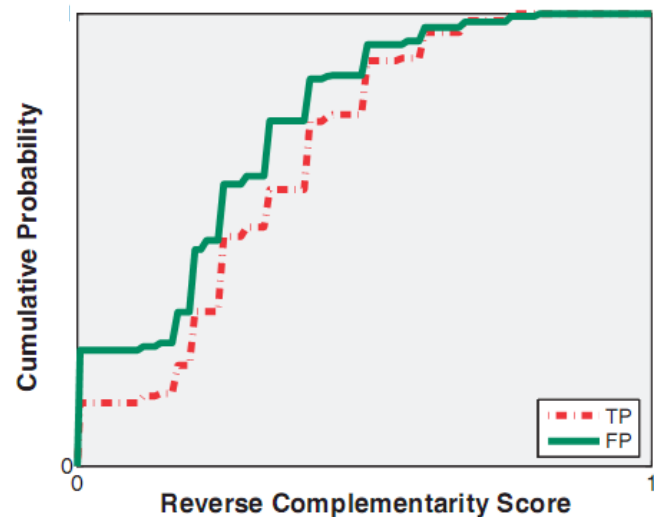
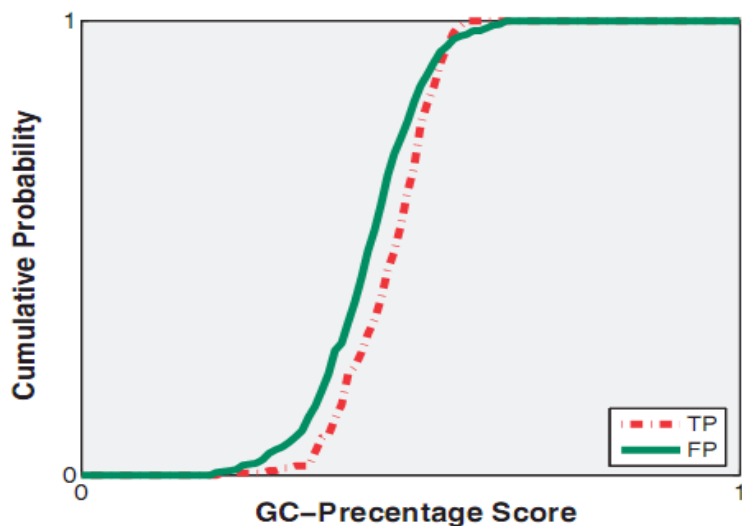
Is a feature discriminative ?



Could it be more discriminative?



How well do the features predict ?



Organization

Motivation

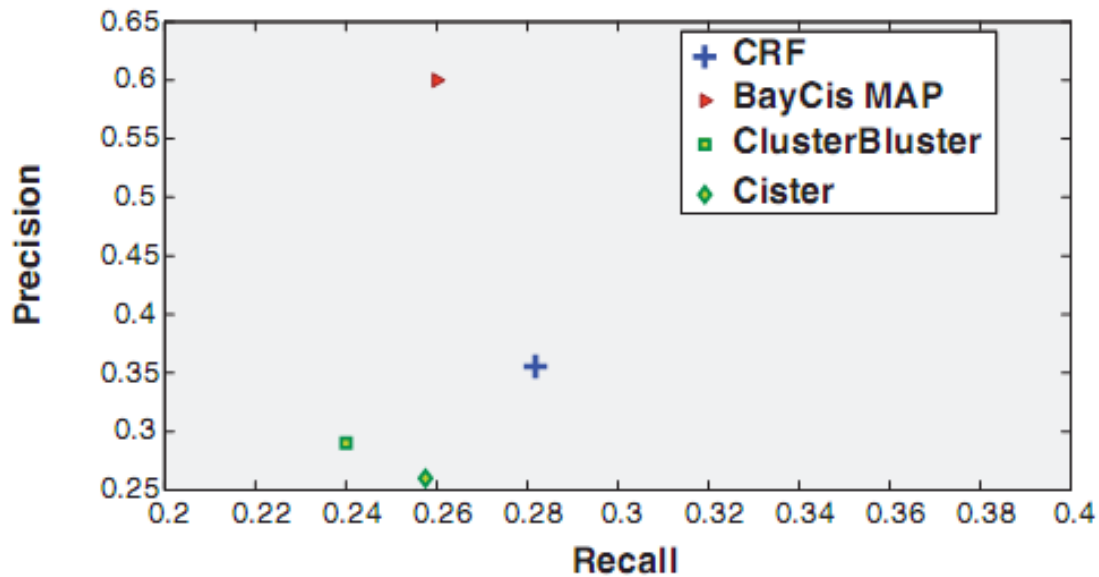
Schema and model

Features and empirical analysis

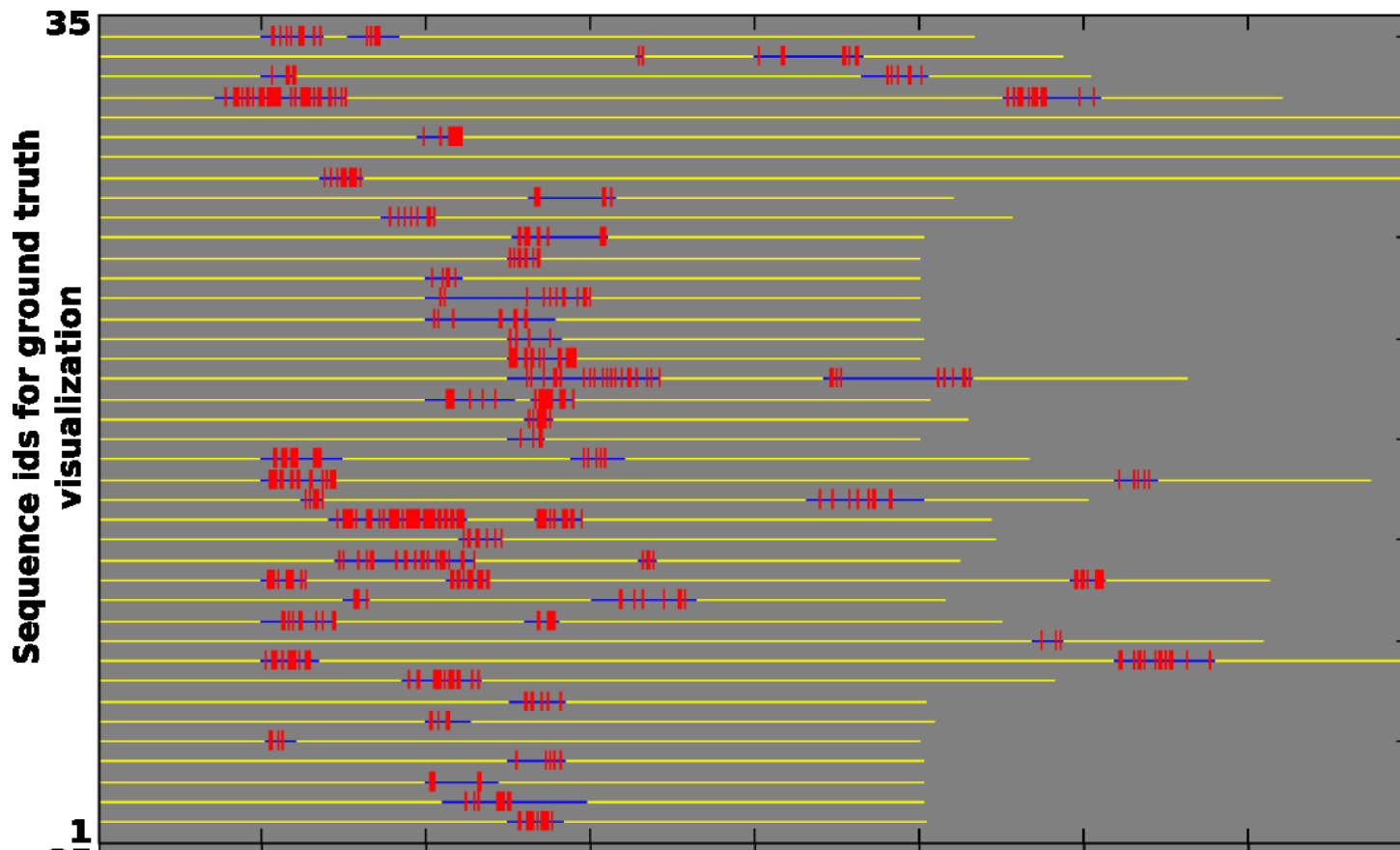
Results

Conclusion

Simulation Results



Datasets used



Legend :

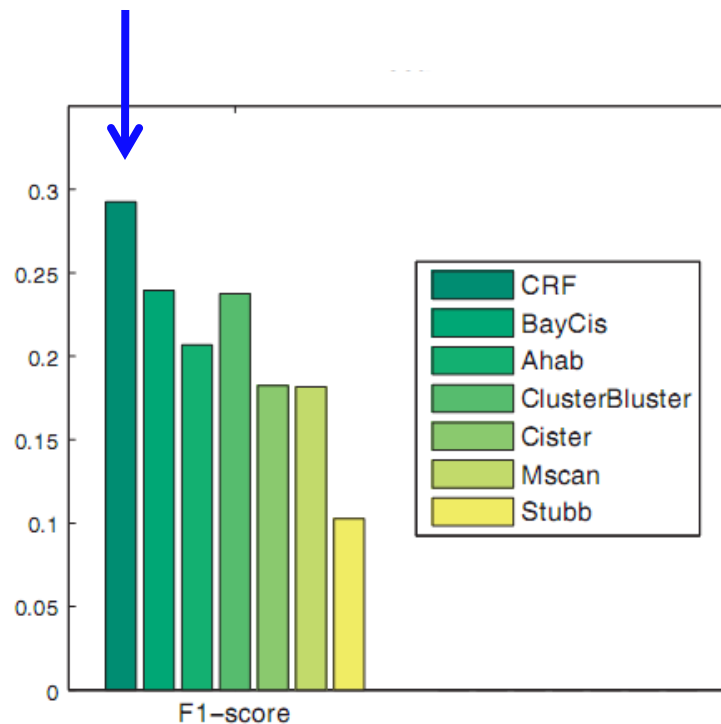
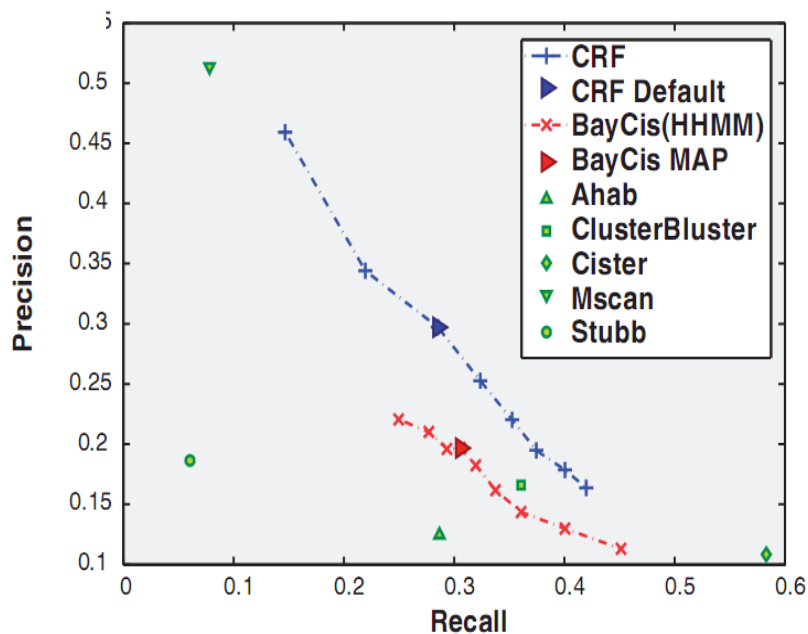


CIS REGULATORY MODULE



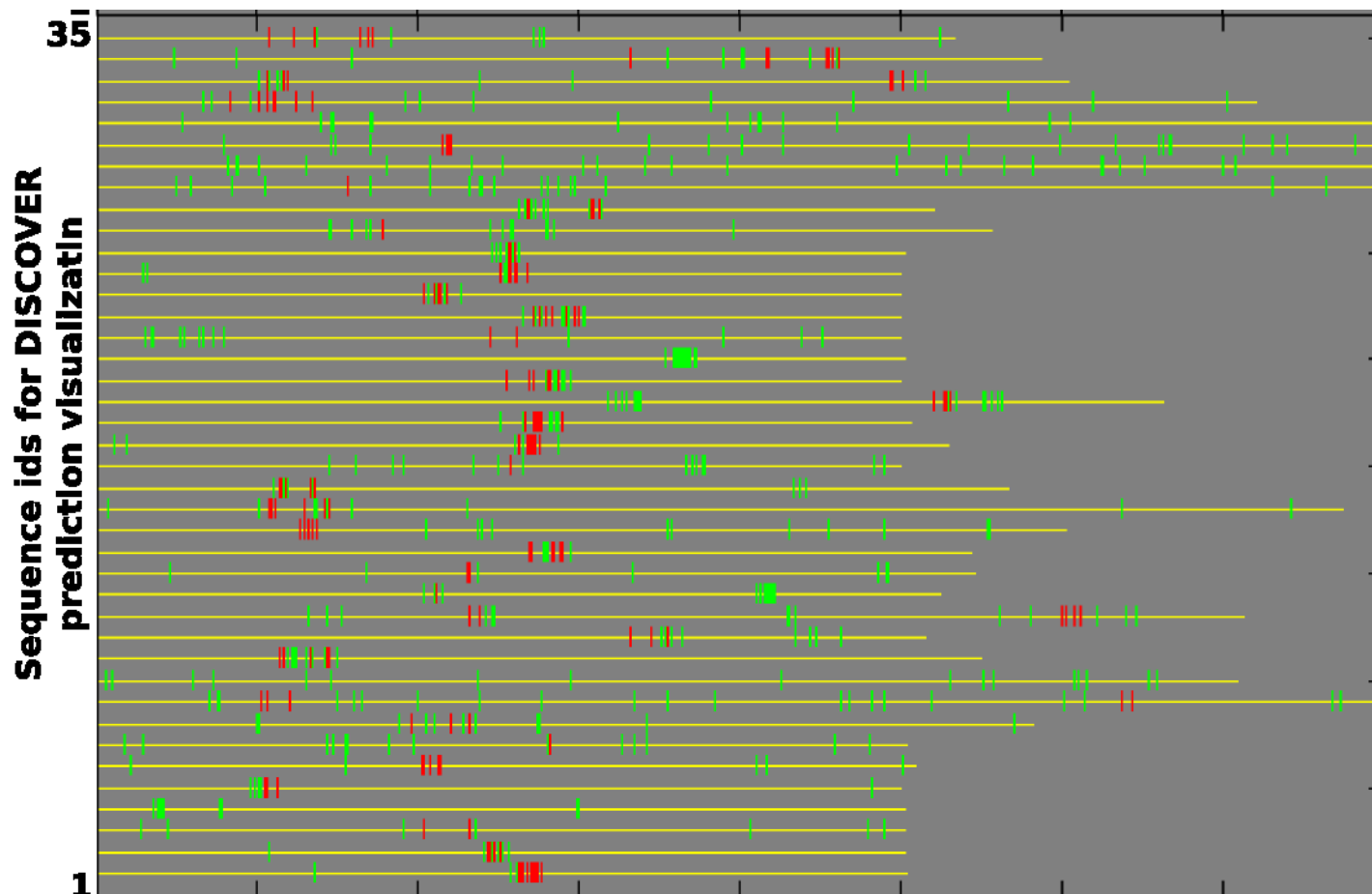
BINDING SITE

LOOCV Results : Drosophila data

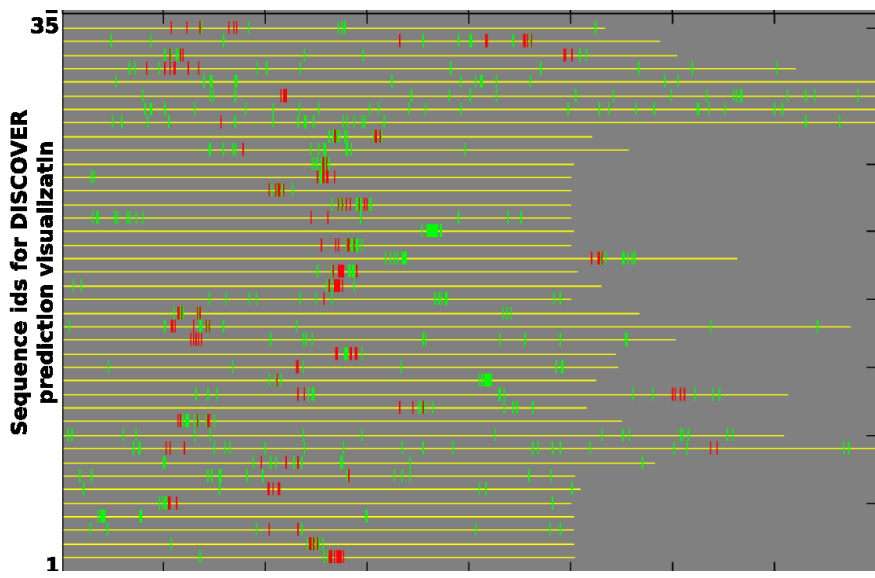


Precision = #TP / (#TP + #FP) , Recall = #TP / (#TP + #FN)
F1 = Harmonic mean of precision and recall

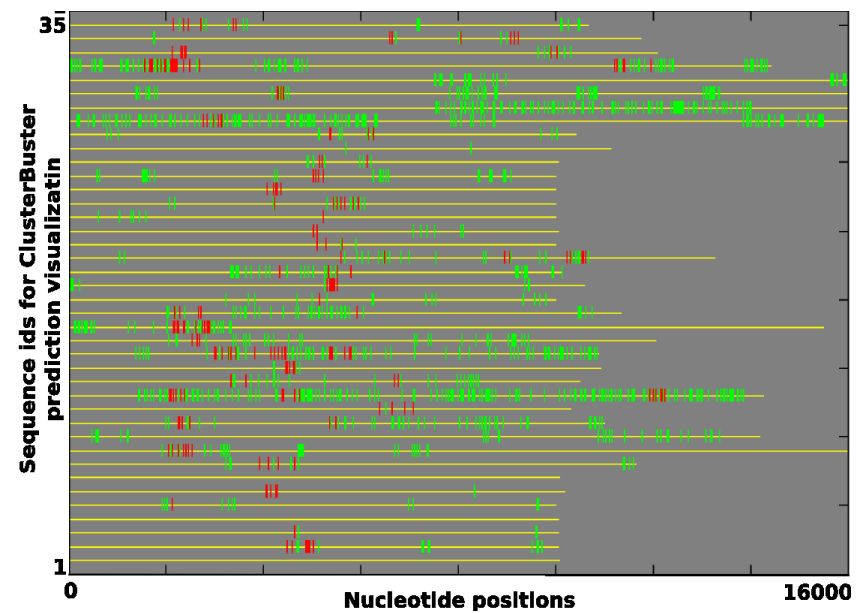
DISCOVER predictions



Cluster Buster predictions



DISCOVER predictions



ClusterBuster predictions

Organization

Motivation

Schema and model

Features and empirical analysis

Results

Conclusion

Future work

- Additional features : inter motif distances, composite motifs, etc
- Analysis of the assigned weights to different features and redundancy of different features wrt each other
- Unsupervised motif detection

Summary

- Generative models like HMM
 - saturated performance
 - risk of tuning to noise on noisy data
 - unclear about how to incorporate diverse evidence
- DISCOVER
 - **Discriminative model** which maximizes the conditional probabilities of the labels given the sequence
 - **Easy to integrate many kinds of evidence** into the prediction scheme
 - **Clean and abstracted estimation and inference techniques**, do not change on incorporating new features

<http://www.sailing.cs.cmu.edu/discover.html>



The screenshot shows the SAILING LAB website. At the top, there is a navigation bar with the SAILING LAB logo and the text "Laboratory for Statistical Artificial Intelligence & Integrative Genomics". Below this, a secondary navigation bar lists "Computer Science, Machine Learning, Language Technologies, Computational Biology | School of Computer Science | Carnegie Mellon". On the left side, there is a vertical menu with buttons for "Home", "People", "Research", "Project pages", "Baycis", "CSMET", "DISCOVER", "mStruct", "SPECTRUM", "IMM-PGE", "N/w modelling", "News", "Publications", "Calendar", and "Internal". The main content area features a large graphic of the word "DISCOVER" where the "S" is replaced by a DNA double helix and the "O" is a magnifying glass. Below this graphic, there is a text block describing the DISCOVER model and providing citation information.

DISCOVER

DISCOVER is a discriminative, conditional random field (CRF) based model used for supervised motif discovery on metazoan genomes. The code can be downloaded [here](#). To be cited as:
Wenjie Fu, Pradipta Ray and Eric P. Xing, DISCOVER: A feature-based discriminative method for motif search in complex genomes, Proceedings of the 16th International Conference on Intelligent Systems for Molecular Biology (ISMB 2009)
PDF can be found [here](#).

Acknowledgements

- My advisors Eric Xing and Veronica Hinman
- Co-author Wenjie Fu and SAILING Lab
- SAILING summer visitor Geir K. Sandve
- Travel Grant : Office of Science (BER), U.S. Department of Energy, Grant No. DE-FG02-06ED64270
- **Research Funding :**



Thanks!



SAILING LAB



Laboratory for Statistical Artificial Intelligence & Integrative Genomics

<http://www.sailing.cs.cmu.edu>

Normalizing the scores

