Using biological prior information in motif discovery

Statistical Modeling and Machine Learning in Computational Systems Biology June 22-26, 2009, Tampere, Finland

Ole Winther

Technical University of Denmark (DTU) & University of Copenhagen (KU)

June 25, 2009

All lectures

1. Introduction to graphical models and Bayesian networks
2. Estimating the size of the transcriptome
3. Using biological prior information in motif discovery
4. Learning linear Bayes networks with sparse Bayesian models

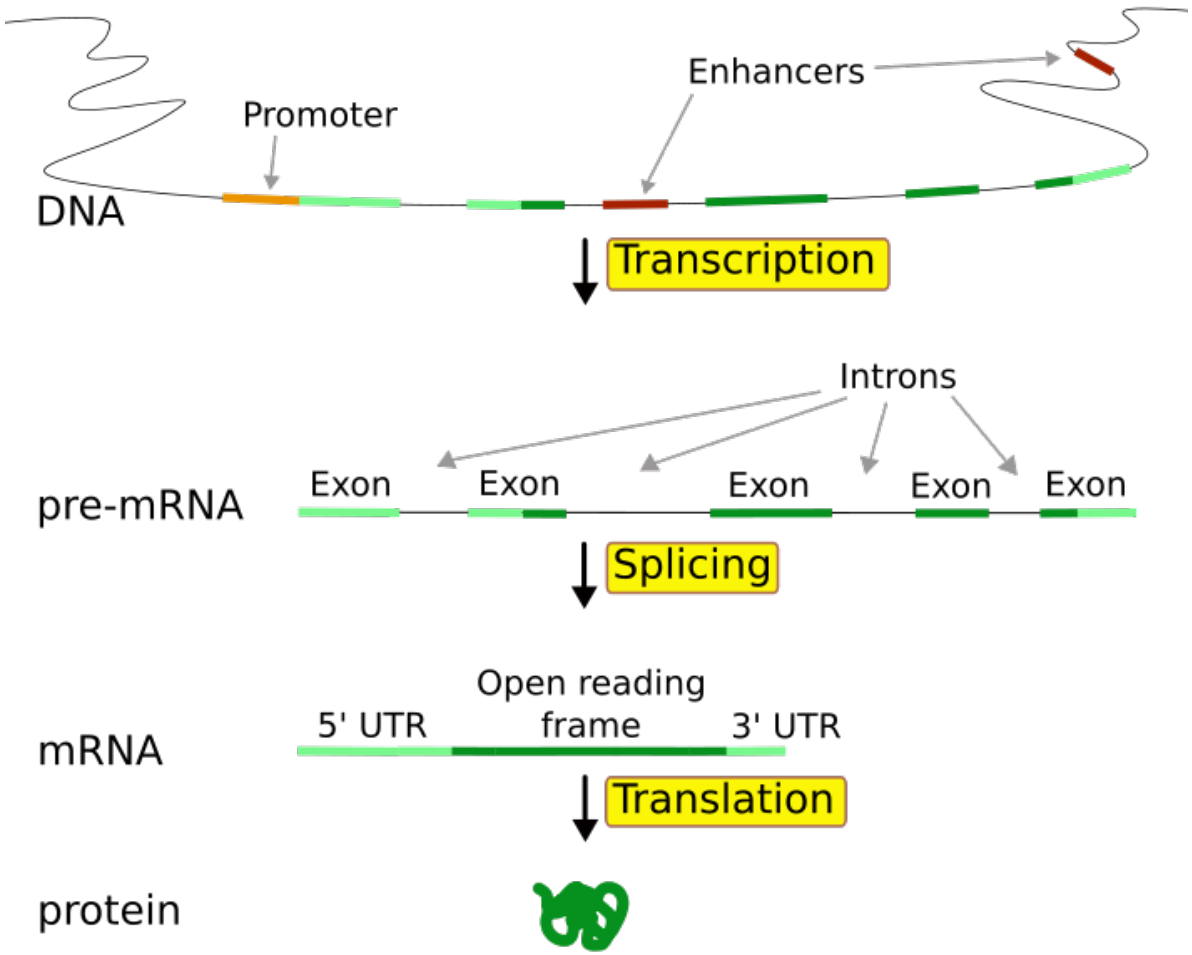
Common theme:

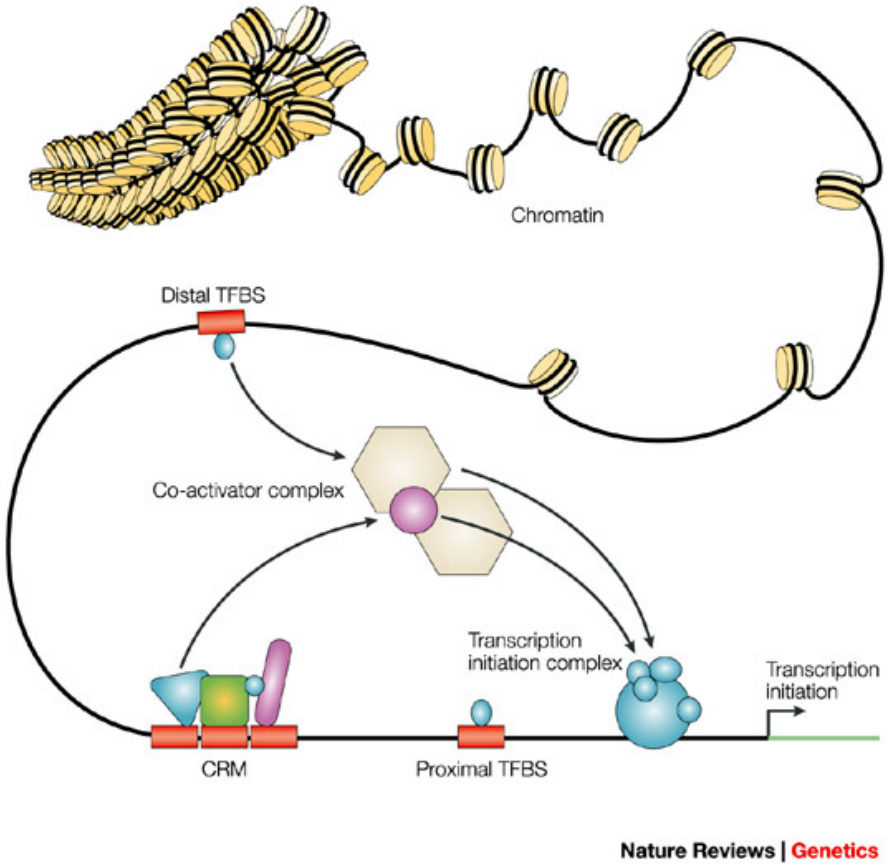
- Complex Bayesian model building possible and advantageous
- Model checking – prediction, marginal- and test-likelihood

Lecture 3

- Transcriptional regulation
- Motif discovery
- Biological prior knowledge
- BayesMD

Gene – from DNA to protein





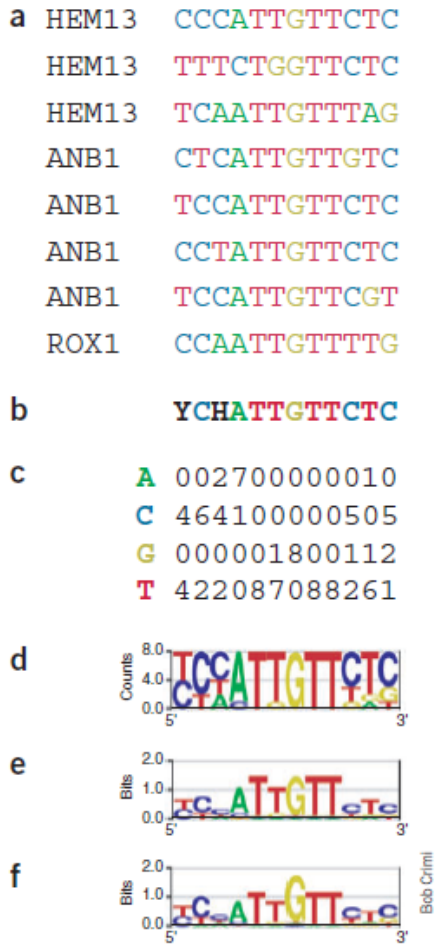
Sandelin and Wasserman, 2004

- Motif discovery typical set-up:
  1. Collect set of co-regulated genes
  2. Extract promoter sequences from these genes
  3. Search for over-represented motifs in a see of background signal
- A motif is a short, 6-20, word.
- This word may represent a transcription factor binding site (TFBS) for a specific TF.
- Motif finding - scanning promoter sequences with position weight matrices (PWMs) for known motifs.
- Many false positives - need more in vivo constraints!

- Motif logos - visualize information content:
- Information content:

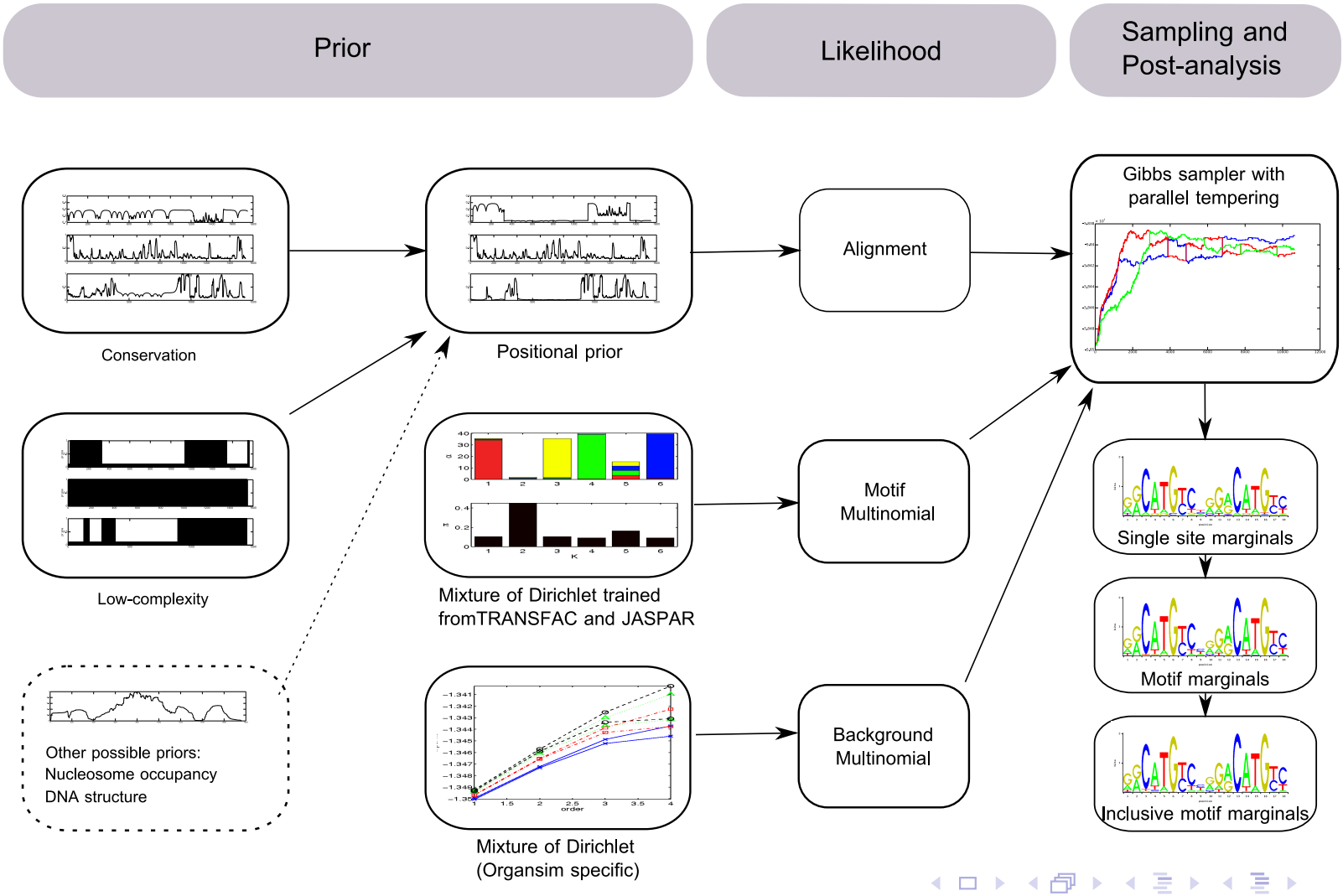
$$I = -\sum_i f_i \log_2 \frac{f_i}{(q_i)}$$

- f_i: empirical frequency
- q_i: background frequency
- From D'haeseleer, Nat. Biotech. 2006



- Sources of biological a priori knowledge:
  - Motif – what are the typical statistics of motif? We have this kind of information in databases like Jaspar and Transfac.
  - Background – organism-specific higher order Markov dependencies - train on all promoter sequences of organism in question.
  - Positional – conservation, low complexity, nucleosome occupancy, DNA structure. We have predictions for this!
- Our approach probabilistic with Gibbs sampling search
- Weeder enumeration quite successful!

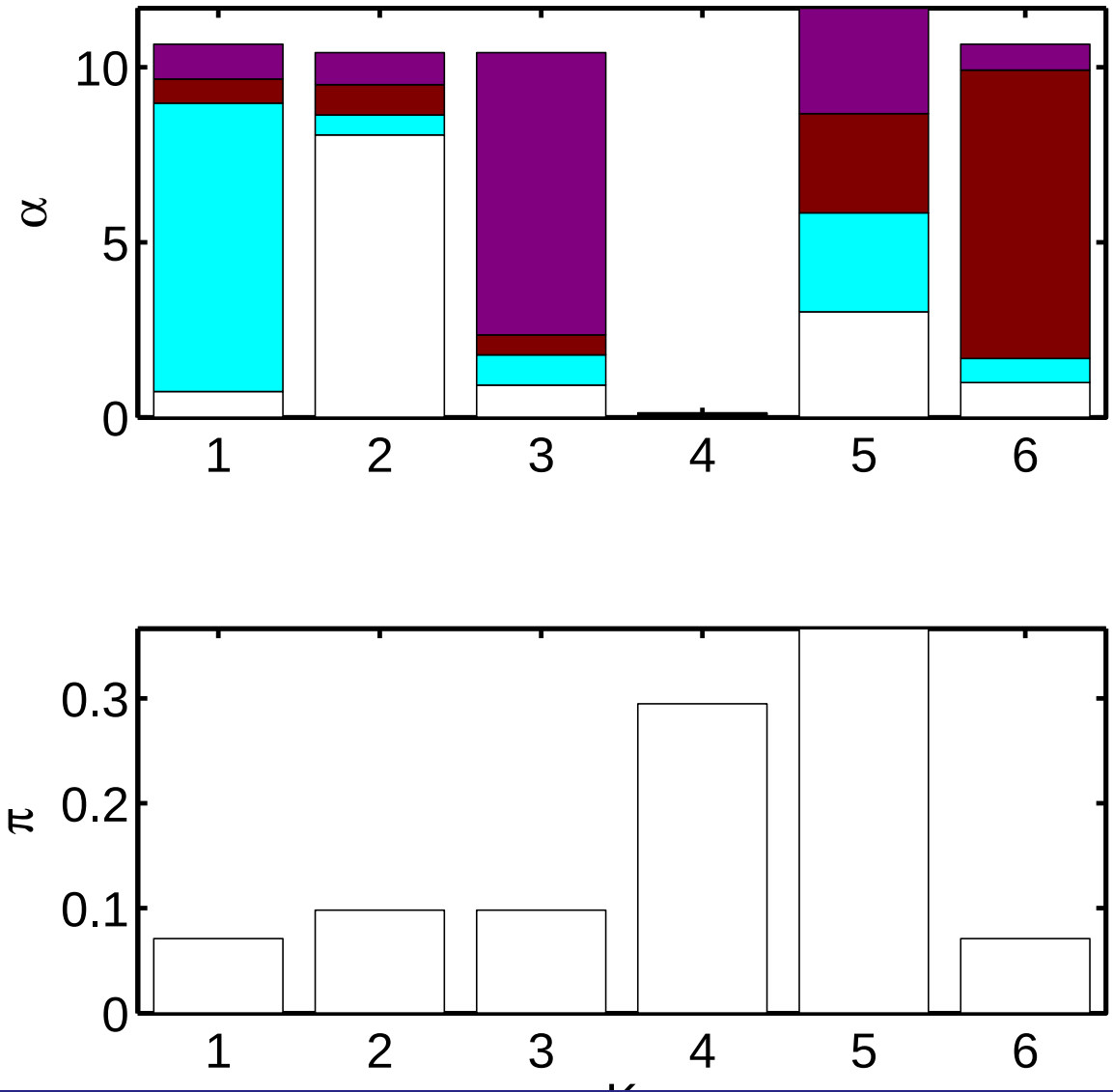
Figure 1


Prior
Likelihood
Sampling and Post-analysis
Conservation
Positional prior
Alignment
Gibbs sampler with parallel tempering
Low-complexity
Mixture of Dirichlet trained fromTRANSFAC and JASPAR
Motif Multinomial
Single site marginals
Motif marginals
Inclusive motif marginals
Other possible priors:
Nucleosome occupancy
DNA structure
Mixture of Dirichlet (Organsim specific)
Background Multinomial

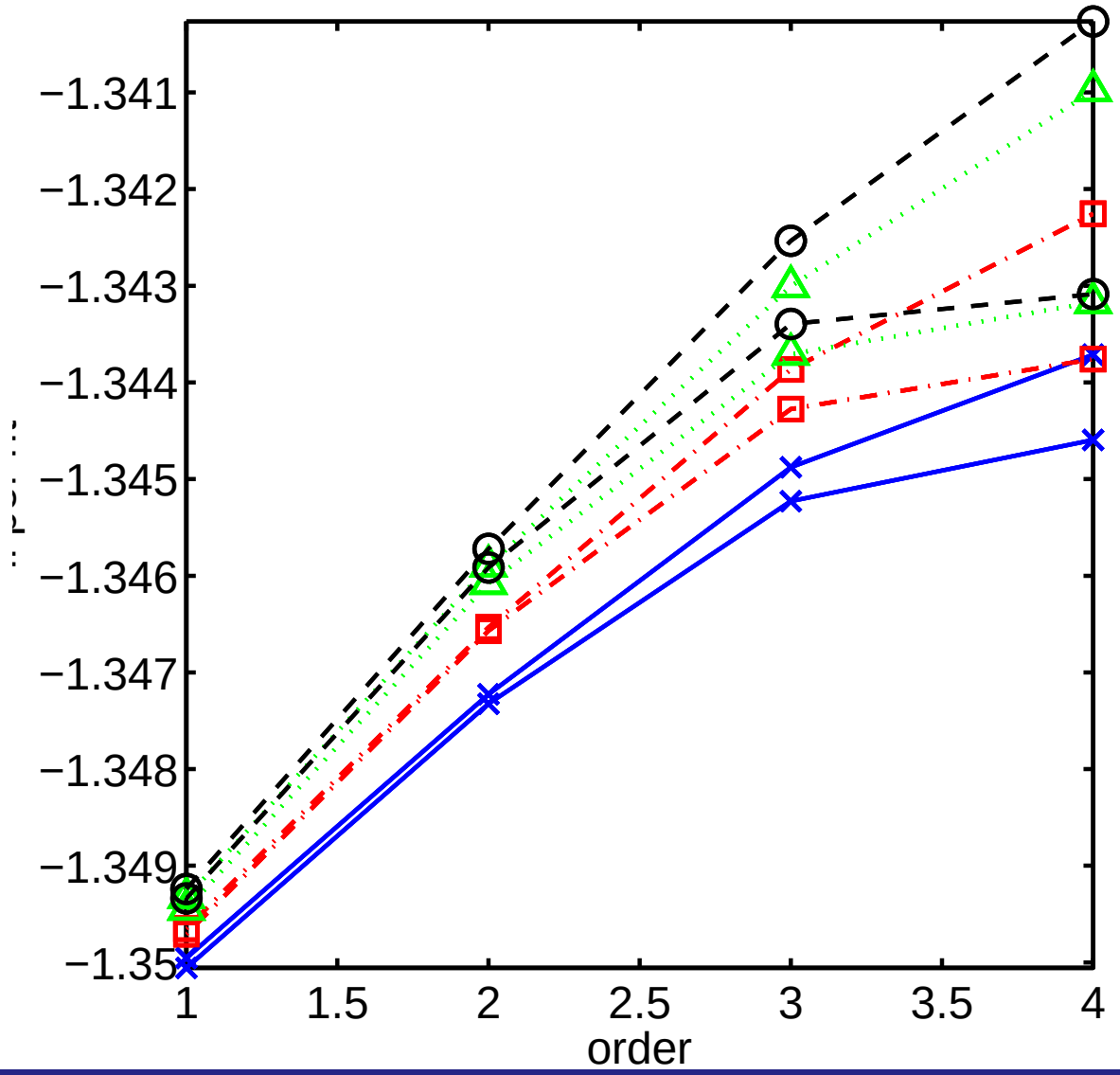
$$P(\mathbf{A}, \mathbf{S}|\mathbf{B}) = \prod_m P_m(\mathbf{S}(\mathbf{A}_m)|\mathbf{A}_m, \mathbf{B}_m) P_{\text{bg}}(\mathbf{S}_{\text{bg}}|\mathbf{A}, \mathbf{B}_{\text{bg}}) P(\mathbf{A}|\mathbf{B}_{\text{align}})$$

- Sequences $\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$
- Alignment tensor $\mathbf{A}$ (element a_{mnr}
- Starting position of the rth occurrence of the mth motif in the nth sequence
- P_m is the distribution for motif m
- $\mathbf{S}(\mathbf{A}_m)$ is shorthand for the sequences contained in motif m
- P_{bg} is the background distribution for
- sequences not in motifs $\mathbf{S}_{\text{bg}} = \mathbf{S} \backslash \{\mathbf{S}(\mathbf{A}_m)\}$.

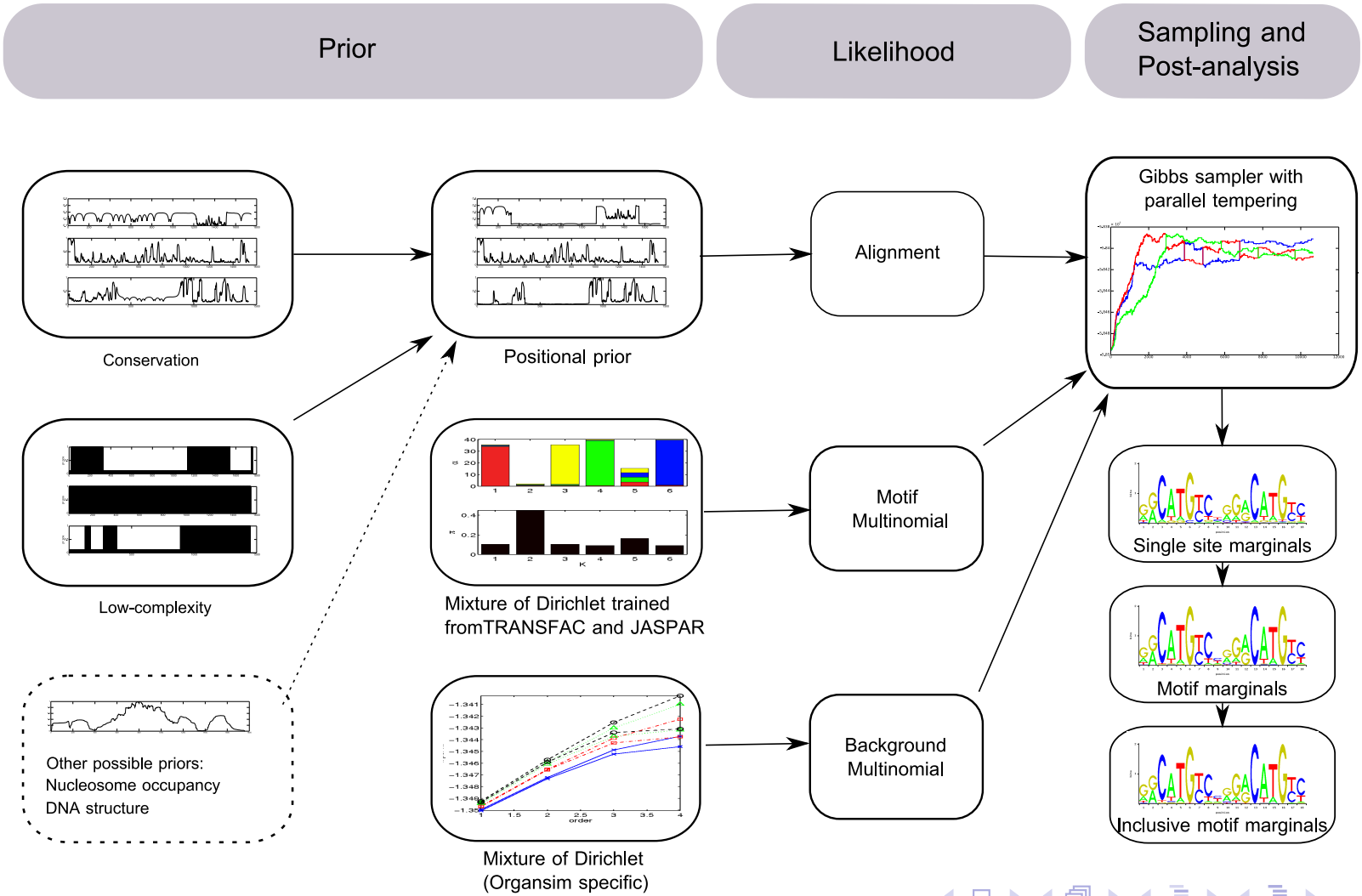
Learning the priors



Learning the priors

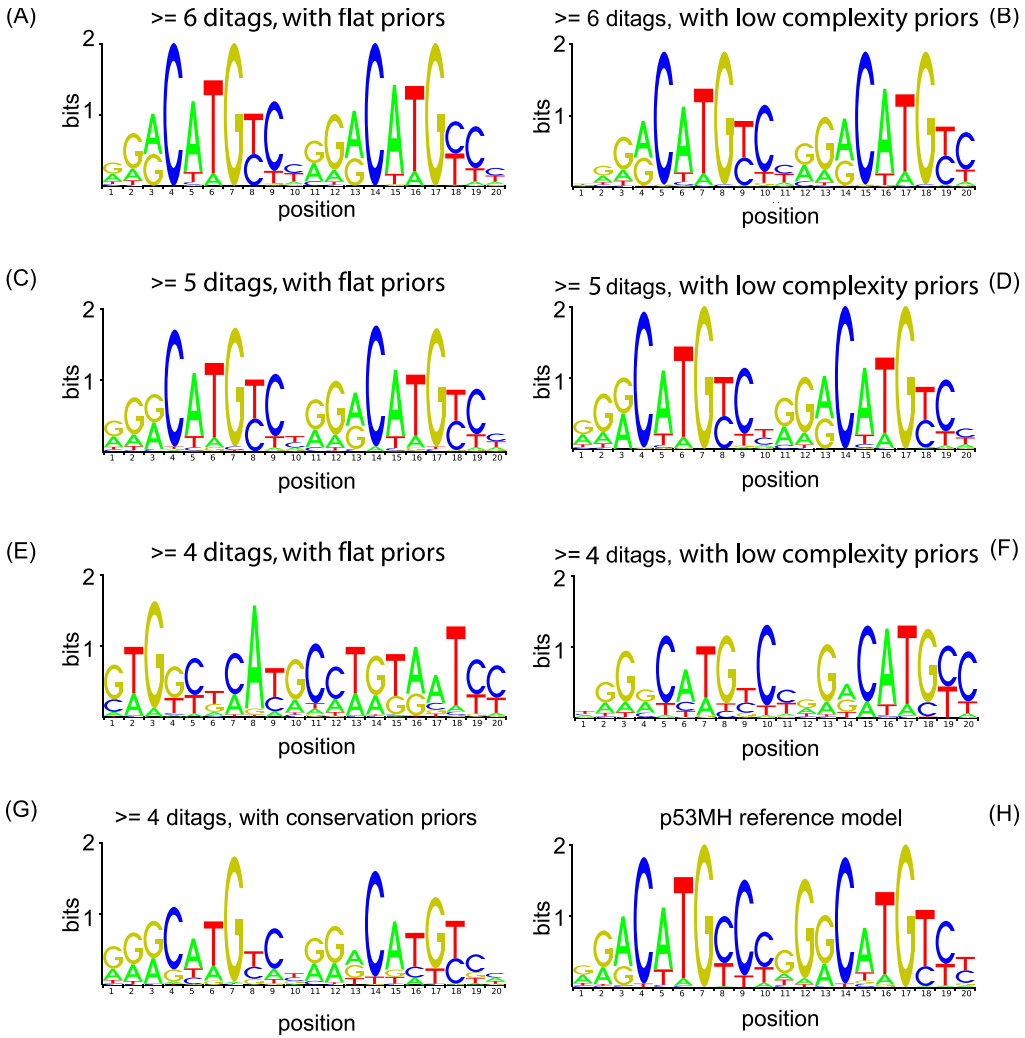


Learning the priors



- Did a decent job in (non-blind) Tompa assessment. Better than other probabilistic approaches but worse than Weeder.
- Did much better than NestedMICA in decoy test proposed in NestedMICA paper.
- Next slide illustrates the use of positional prior.
- No real de-novo successes on data sets provided by collaborators. :-(

P53


(A)
>= 6 ditags, with flat priors
(B)
>= 6 ditags, with low complexity priors
(C)
>= 5 ditags, with flat priors
(D)
>= 5 ditags, with low complexity priors
(E)
>= 4 ditags, with flat priors
(F)
>= 4 ditags, with low complexity priors
(G)
>= 4 ditags, with conservation priors
p53MH reference model
(H)
bits
position

- Motif discovery and finding well-established methodology, 15 year+ old.
- Low success rate in real tasks.
- More a priori filtering, higher precision data and better understanding on thermodynamics of binding needed.
- Reference: Man-Hung Eric Tang, Anders Krogh and OW, BayesMD: Flexible Biological Modeling for Motif Discovery, Journal of Computational Biology, 15, 1347-1363, 2008.
- Many references to related work see paper.