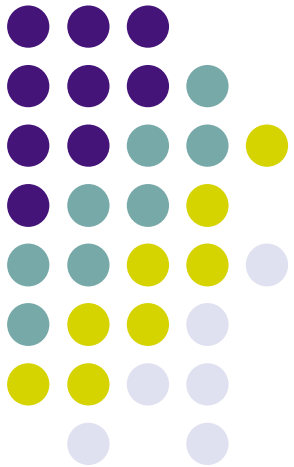
Graduate Computational Genomics

02-710 / 10-810 & MSCBIO2070

Computational motif discovery

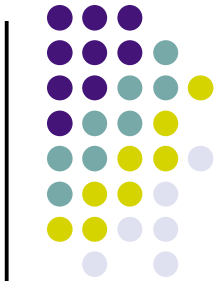


Takis Benos

Lecture #11, February 20, 2007

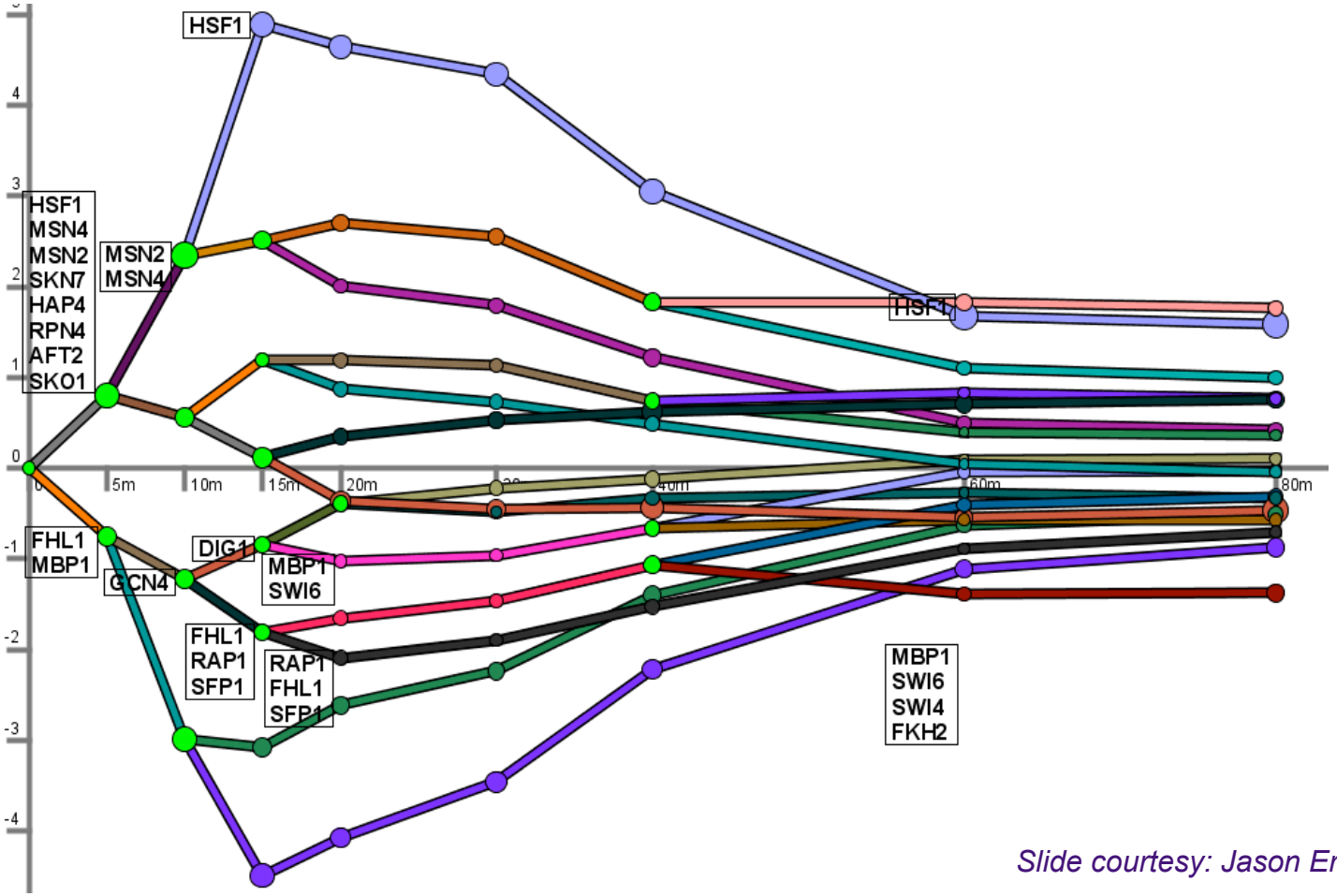
Reading: handouts & papers

Outline



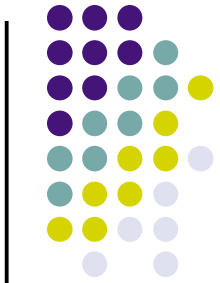
- ✓ The problem
- ✓ Pattern representation
- ✓ Pattern matching
- Pattern discovery
  - A greedy algorithm
  - Expectation-Maximization
  - Gibbs sampling
  - Self-organizing maps

Time-course microarray data



Slide courtesy: Jason Ernst

De novo motif finding



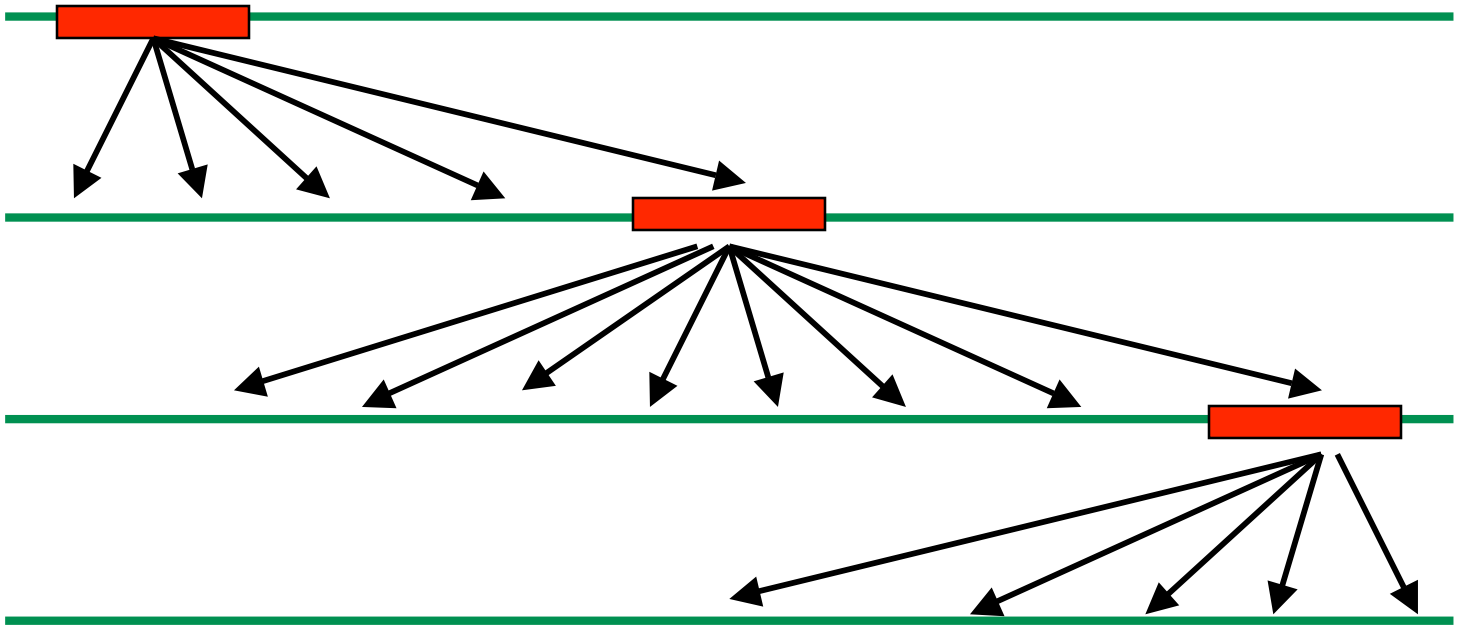
GIVEN

- A set of unaligned sequences that contain instances of a motif
- The expected motif length
- A set of background sequences
- An objective function

FIND

- The motif that maximizes the objective function

Motif finding in unaligned sequences

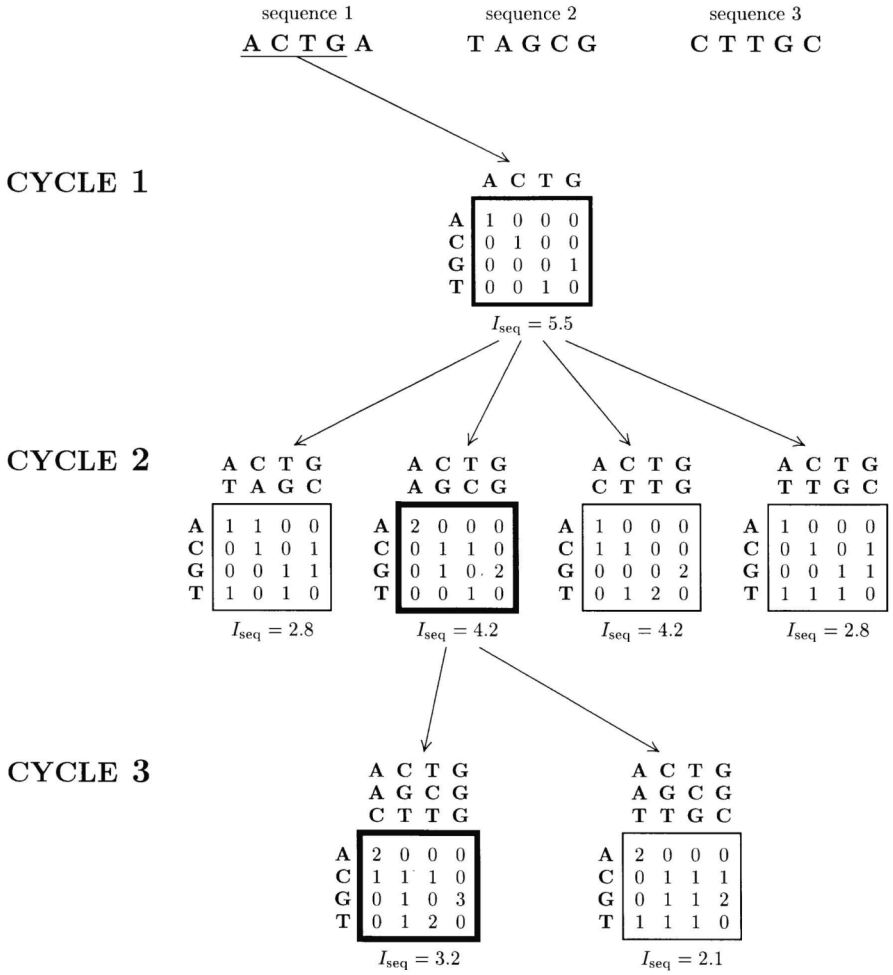


CONSENSUS, WConsensus

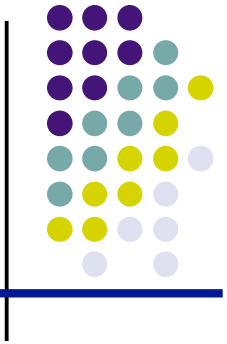
- Objective function
  - *Relative entropy*

$$RH_{PSSM} = \sum_{i=1}^{N} \sum_{b=A}^{T} f_b(i) \cdot \ln \frac{f_b(i)}{P_{ref}(b)}$$

- Optimization method
  - *Greedy algorithm* (brute force)



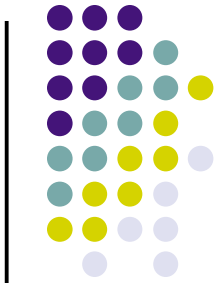
CONSENSUS, WConsensus



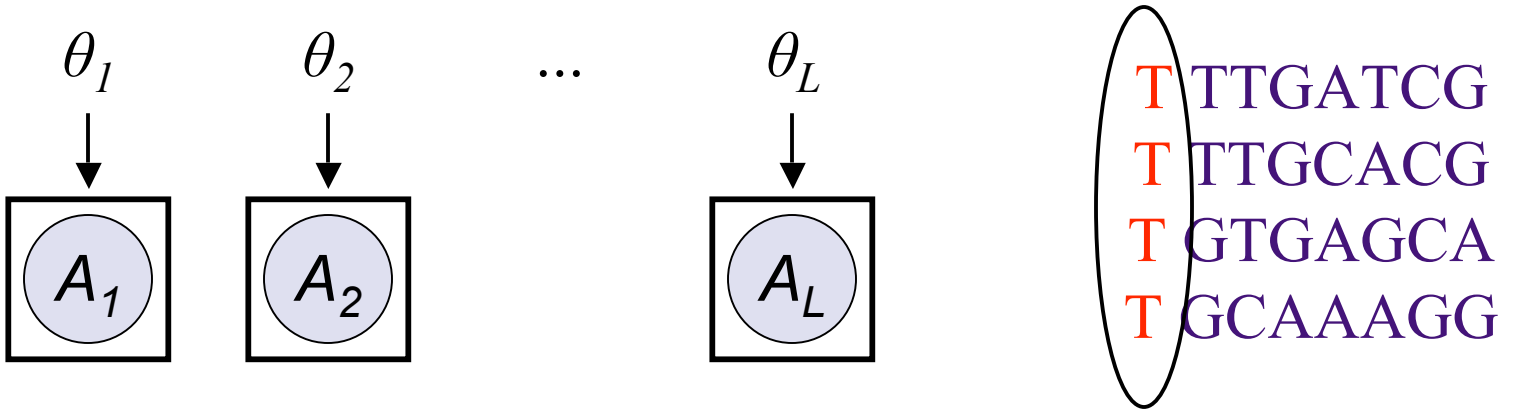
- No need for positive set (background estimated from the genome).
- In each round only a number of matrices is retained.
- For *CONSENSUS*, the user specifies the length of the pattern; for *WConsensus* the user -alternatively- determines the number of SD.
- Programs can predict zero or more target sites in the examined promoters.
- *WConsensus* permits *p-value* calculations of a PSSM.

- Testing
  - 18 CRP-regulated genes
  - 105 bp long promoters
  - 24 putative CRP binding sites
  - CRP binds as dimer
  - *WConsensus* predicted 19 of the 24 (S_n=79%) plus 3 additional sites (S_p=89%).

Product multinomial (PM) model [Lawrence *et al.*, *Science*, 1993]

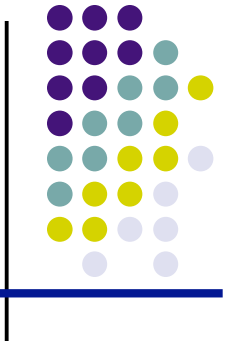


- Position-specific multinomial distribution: $\theta_i = [\theta_{iA}, \theta_{iC}, \theta_{iG}, \theta_{iT}]^T$



- Position-Specific Scoring Matrix (PSSM): θ
  - Assumes nucleotide distribution in one position is independent of distributions in other positions

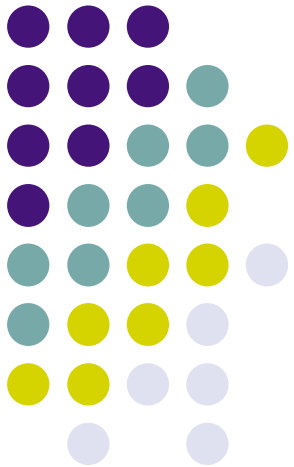
Alternating approach



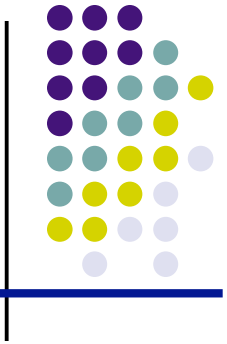
1. Guess an initial weight matrix
2. Use weight matrix to predict instances in the input sequences
3. Use instances to predict a weight matrix
4. Repeat 2 & 3 until no change.

Examples: MEME (expectation maximization / Bailey, Elkan)
Gibbs sampler (Lawrence et al.)
ANN-Spec (neural net / Workman, Stormo)

Expectation - Maximization



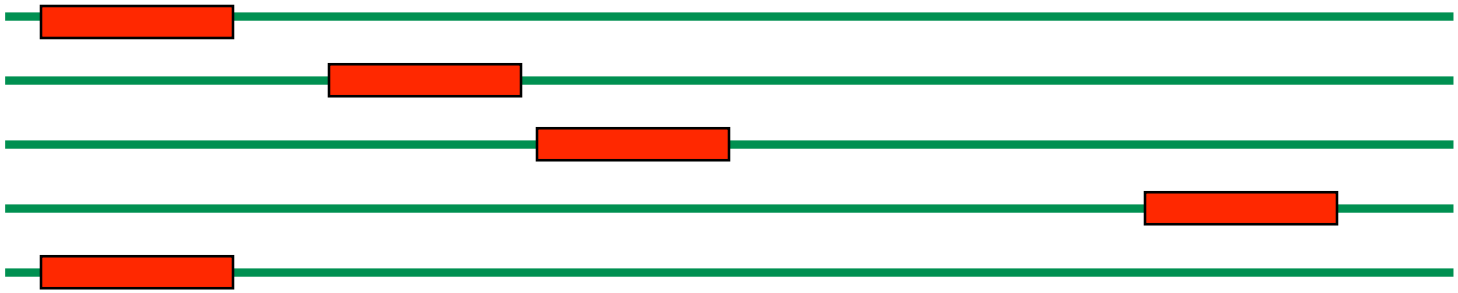
Expectation-maximization



EM

```
foreach subsequence of width W
   convert subsequence to a matrix
   do {
      re-estimate motif occurrences from matrix
      re-estimate matrix model from motif occurrences
   } until (matrix model stops changing)
end
select matrix with highest score
```

Expectation-maximization



DNA “signal”

TGACCTCT
TGATCTTA
GGACCCTA
TGATCCGT
TGACCCTT
GGACCCTT
TGACCTCT
TGACCTTA

PSSM

A	-1.1	-1.1	**+1.1**	-1.1	-1.1	-1.1	-1.1	**+.29**
C	-1.1	-1.1	-1.1	**+.85**	**+1.1**	**+.51**	0.0	-1.1
G	0.0	**+1.1**	-1.1	-1.1	-1.1	-1.1	-.40	-1.1
T	**+.85**	-1.1	-1.1	0.0	-1.1	**+.51**	**+.69**	**+.69**

Expectation-maximization

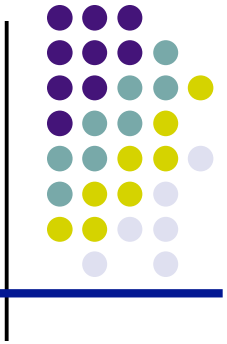
DNA “signal”

TGACCTTT
TGATCTTA
TGACCCTA
TGATCCGT
TGACTCTT
GGACCCTT
TGACCTCT
TGACCTTA

PSSM (2)

A	-1.1	-1.1	**+1.1**	-1.1	-1.1	-1.1	-1.1	**+.29**
C	-1.1	-1.1	-1.1	**+.85**	**+1.1**	**+.51**	-.40	-1.1
G	-.40	**+1.1**	-1.1	-1.1	-1.1	-1.1	-.40	-1.1
T	**+.98**	-1.1	-1.1	+0.0	-1.1	**+.51**	**+.85**	**+.69**

Scoring each subsequence

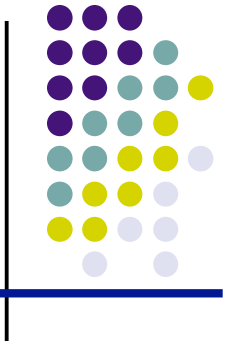


Sequence: TGTGACCTTTTTTGTGGCATCGGGCGAGAATA

Subsequences	Score
TGTGACCT	-0.42
GTGACCTG	-2.34
TGCTGGTT	+5.93
GCTGGTTT	-2.42
	...

Select from each sequence the subsequence with maximal score.

Re-estimating motif matrix

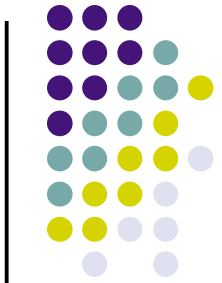


Occurrences	Counts	+ *pseudo*
TTTGATCG	**A** 00013201	**A** 11124312
TTTGCACG	**C** 00101030	**C** 11212141
TGTGAGCA	**G** 02030113	**G** 13141224
TGCAAAGG	**T** 42300100	**T** 53411211

Convert to frequencies

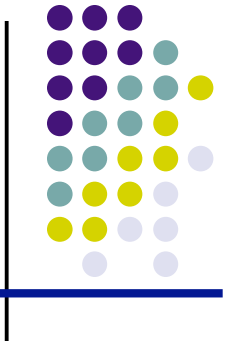
A	0.13	0.13	0.13	0.25	0.50	...
C	0.13	0.13	0.25	0.13	0.25	
G	0.13	0.38	0.13	0.50	0.13	
T	0.63	0.38	0.50	0.13	0.13	

Possible problems



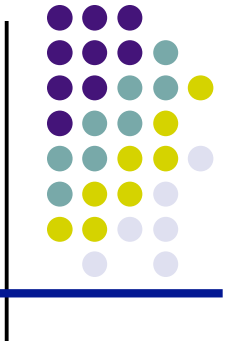
- **Problem**: How do we estimate counts accurately when we have only a few examples?
  - **Solution**: Use Dirichlet mixture priors.
- **Problem**: Too many possible starting points.
  - **Solution**: Save time by running only one iteration of EM.
- **Problem**: Too many possible widths.
  - **Solution**: Consider widths that vary by $\sqrt{2}$ and adjust motifs afterwards.
- **Problem**: The EM algorithm finds only one motif.
  - **Solution**: Probabilistically erase the motif from the data set, and repeat.

Possible problems (cntd)



- **Problem**: The motif model is too simplistic.
  - **Solution**: Use a two-component mixture model that captures the background distribution. Allow the background model to be more complex.
- **Problem**: The EM algorithm does not tell you how many motifs there are.
  - **Solution**: Compute statistical significance of motifs and stop when they are no longer significant.
- **Problem:**
  - This procedure doesn't allow the motifs to move around very much. Taking the max is too brittle.
- **Solution:**
  - Associate with each start site a probability of motif occurrence.

What's the idea behind it?



- Assume we have a set of observed quantities x (e.g., *binding sites*), determined from a set of missing or hidden data y (e.g., PSSM model).
- Assume we have set of estimated parameters θ^t.
- We want to calculate a new θ^{t+1} with:

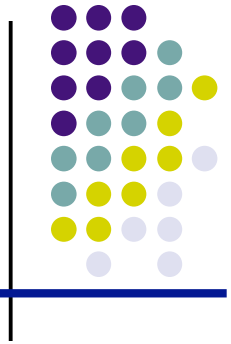
$$\log P(x \mid \theta^{t+1}) > \log P(x \mid \theta^t)$$

- We define/calculate:

$$Q(\theta \mid \theta^t) = \sum_y P(y \mid x, \theta^t) \cdot \log P(x, y \mid \theta)$$

E-step

What's the idea behind it?



$$P(x,y \mid \theta) = P(y \mid x,\theta) \cdot P(x \mid \theta) \Rightarrow$$

$$\log P(x \mid \theta) = \log P(x,y \mid \theta) - \log P(y \mid x,\theta)$$

$$= \sum_y P(y \mid x,\theta^t) \cdot \log P(x,y \mid \theta) - \sum_y P(y \mid x,\theta^t) \cdot \log P(y \mid x,\theta)$$

multiply $P(y|x,\theta^t)$ and sum over y

$$= Q(\theta \mid \theta^t) - \sum_y P(y \mid x,\theta^t) \cdot \log P(y \mid x,\theta) \Rightarrow$$

$$\log P(x \mid \theta) - \log P(x \mid \theta^t) =$$

Jensen's inequality

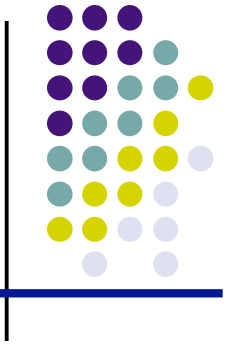
$$= Q(\theta \mid \theta^t) - Q(\theta^t \mid \theta^t) + \sum_y P(y \mid x,\theta^t) \cdot \log \frac{P(y \mid x,\theta^t)}{P(y \mid x,\theta)} \geq Q(\theta \mid \theta^t) - Q(\theta^t \mid \theta^t)$$

So, choose:

$$\theta^{t+1} = \arg\max_\theta Q(\theta \mid \theta^t)$$

M-step

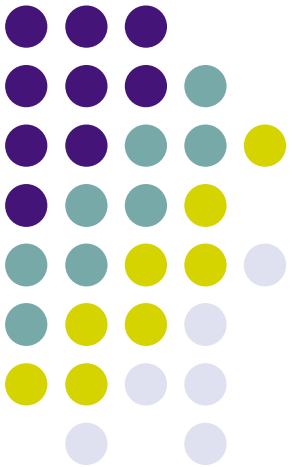
MEME algorithm



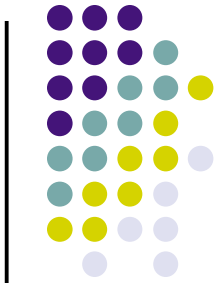
- Bailey & Elkan 1995

```
do
    for (width = min; width *= √2; width < max)
        foreach possible starting point
            run 1 iteration of EM
        select candidate starting points
        foreach candidate
            run EM to convergence
    select best motif
    "erase" motif occurrences
until (E-value of found motif > threshold)
```

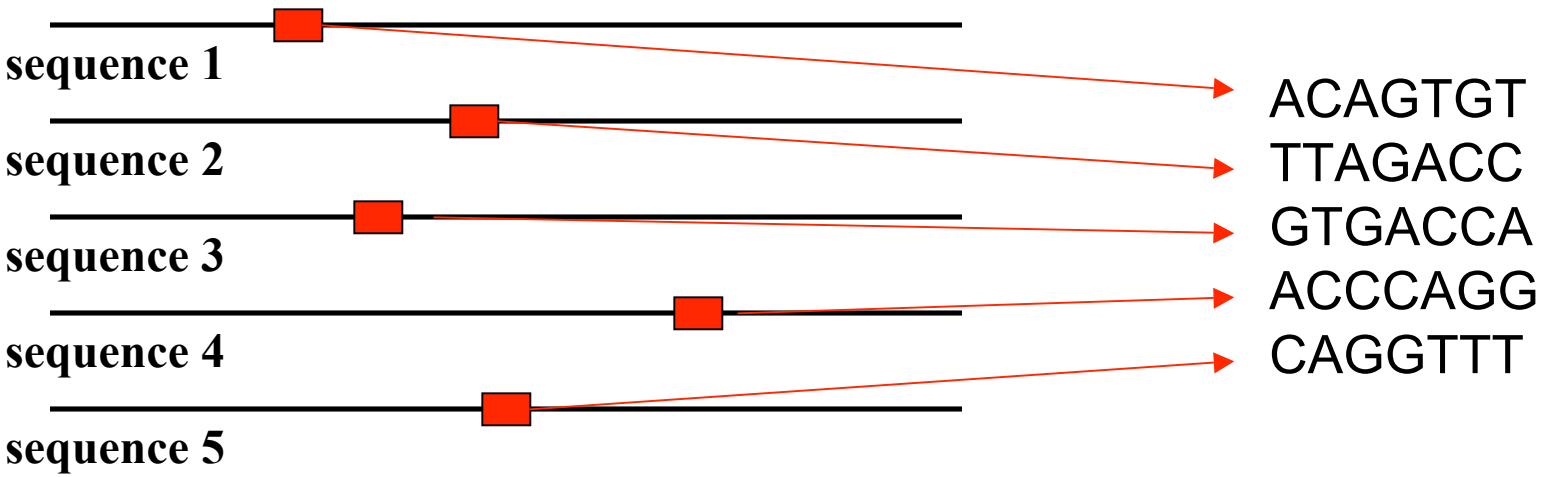
Gibbs sampling



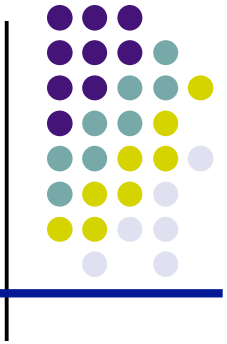
Initialization



- Randomly select an instance s_i from each of t input sequences $\{S_1, \ldots, S_t\}$.

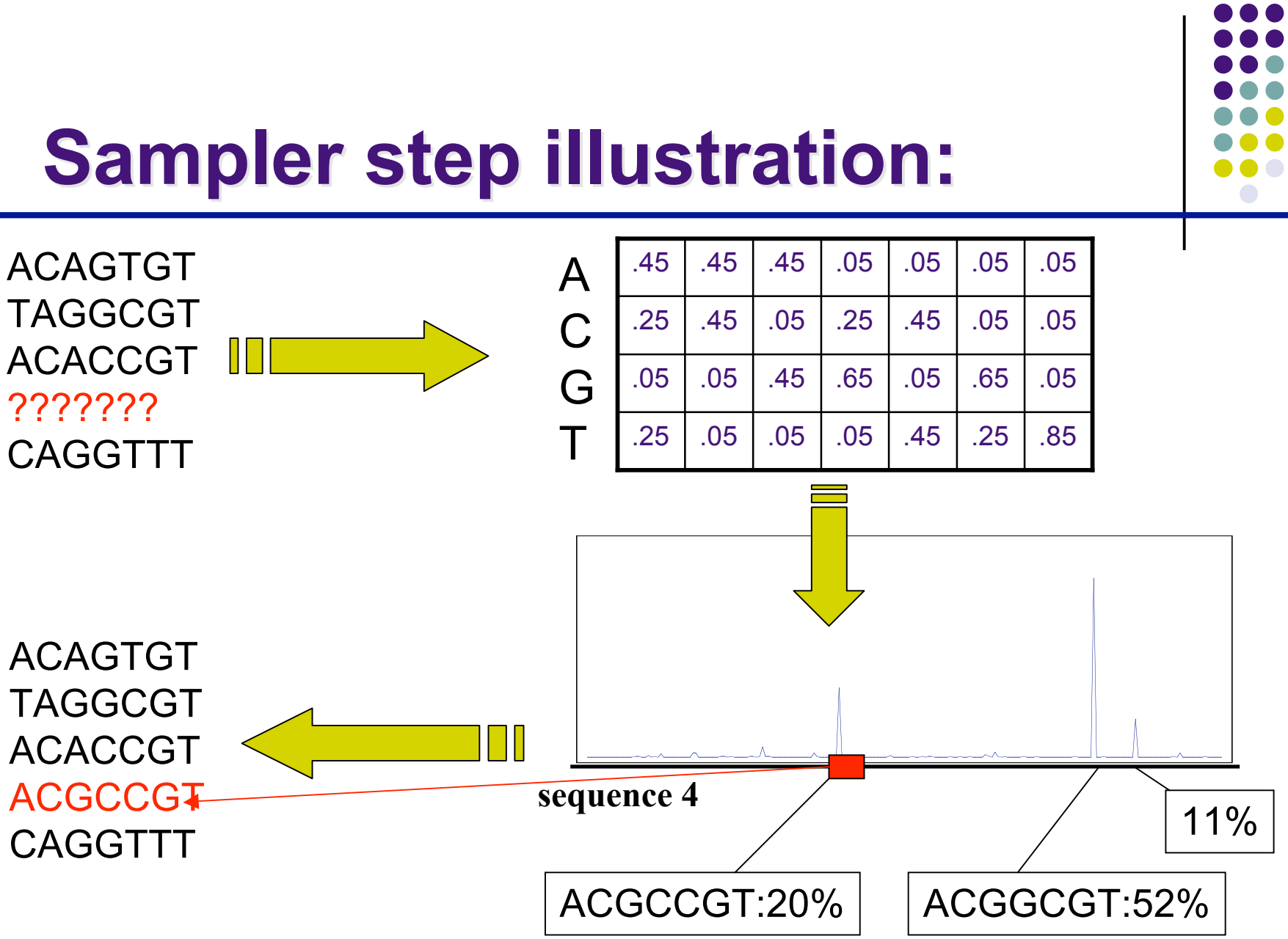


Gibbs sampler

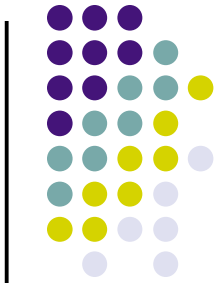


- Initially: randomly guess an instance s_i from each of t input sequences $\{S_1, ..., S_t\}$.
- Steps 2 & 3 (search):
  - Throw away an instance s_i: remaining $(t - 1)$ instances define weight matrix.
  - Weight matrix defines instance probability at each position of input string S_i
  - Pick new s_i according to probability distribution
- Return highest-scoring motif seen

Sampler step illustration:

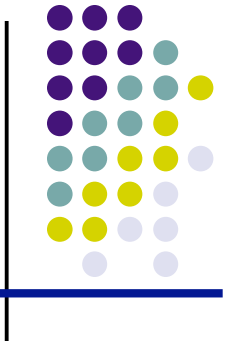


EM & Gibbs sampling: take home…



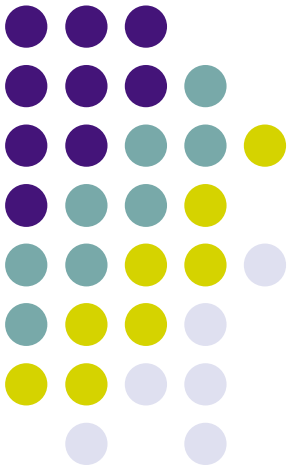
- Motifs are represented by weight matrices.
- Motif quality is measured by relative entropy.
- Motif occurrences are scored using log likelihood ratios.
- EM and the Gibbs sampler attempt to find a motif with maximal relative entropy.
- Both algorithms alternate between predicting instances and predicting the weight matrix.

Disadvantages…

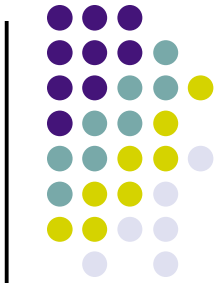


- The transcription factors that bind to the identified motifs are not known
- Sometimes, a dominant motif and its variants are reported
- Prior knowledge can be used on a case-by-case basis

Self-Organizing Maps

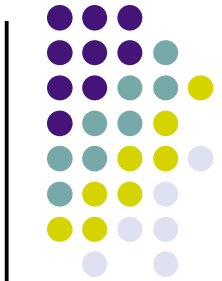


Self-Organizing Maps (SOM)



- SOM is a type of unsupervised neural network (Teuvo Kohonen, 1982).
- It reduces the dimensionality of the data while keeping the topological relations of the vectors.
- It was initially used for visualization of high-dimensional data.
- Applied in motif finding for fitst time in 2005 (Mahony *et al.*).

SOM: the problem



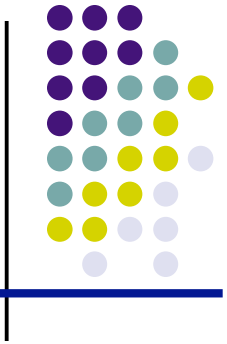
GIVEN

- An initial set of models (nodes) on a grid
- A set of test data (vectors)
- A distance measure between a node and a vector

FIND

- A set of nodes that best represents the associations between the data

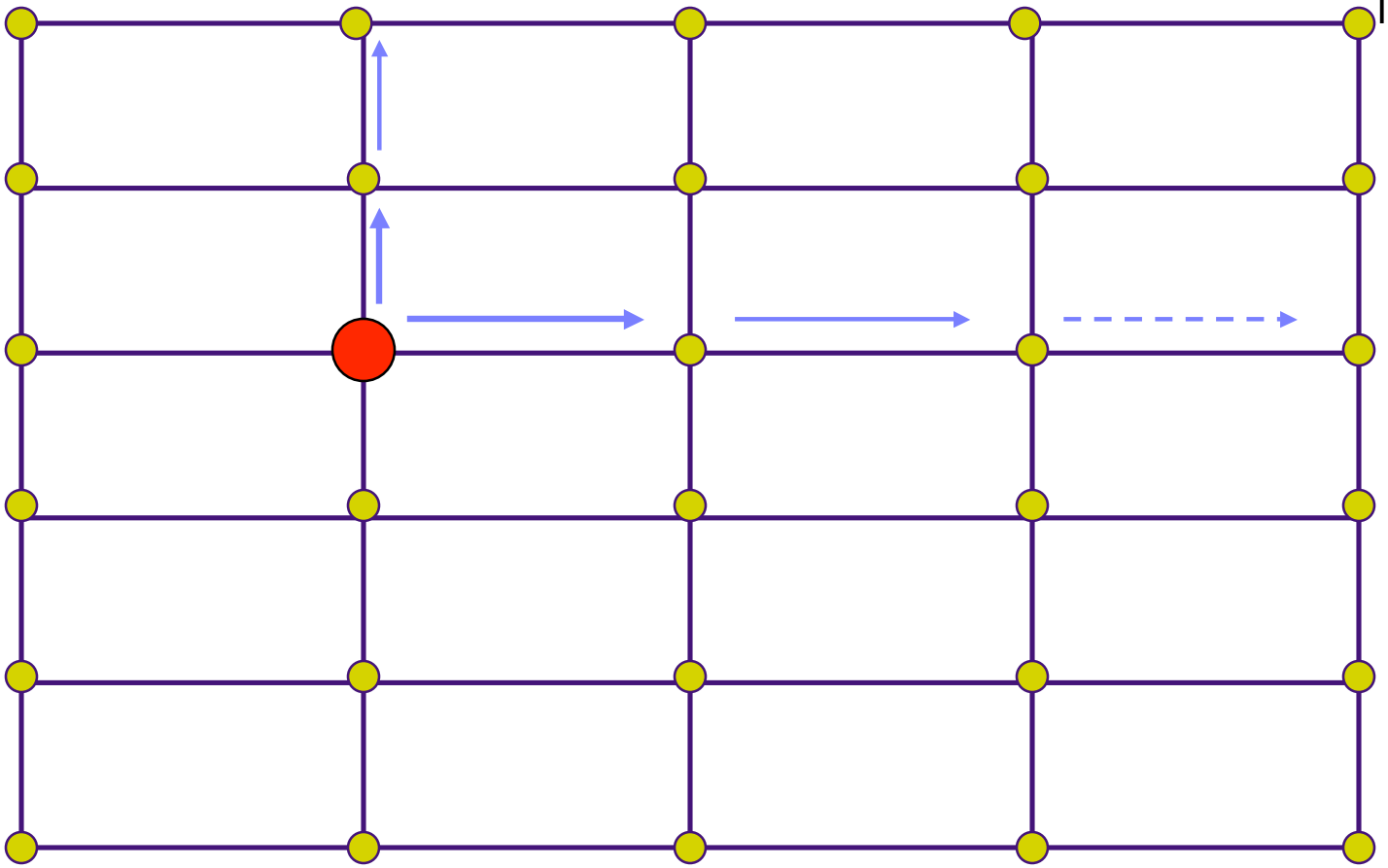
SOM: the algorithm



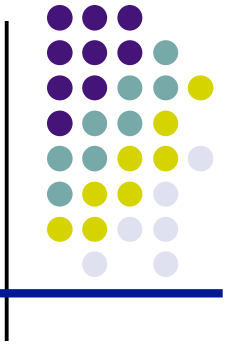
SOM training

foreach subsequence, x, of width W
  do {
    find the model (node), M_i, with $mindist(x,M_i)$
    adjust M_i to incorporate x
    propagate adjustment to neighbouring models
  } until (models stop changing)
end
report all models with Z-score above threshold

Self-Organizing Maps



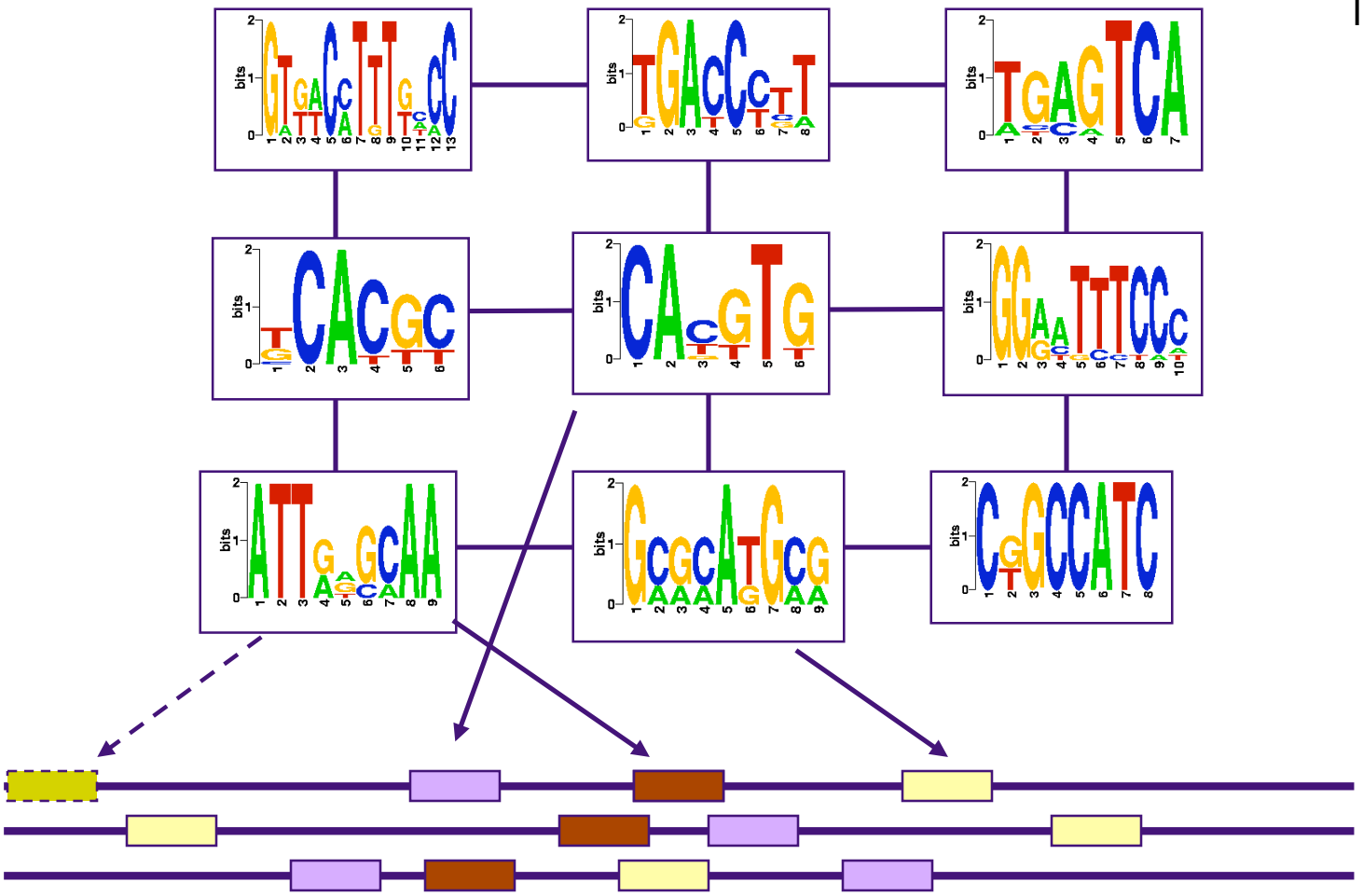
Methods: SOMBRERO



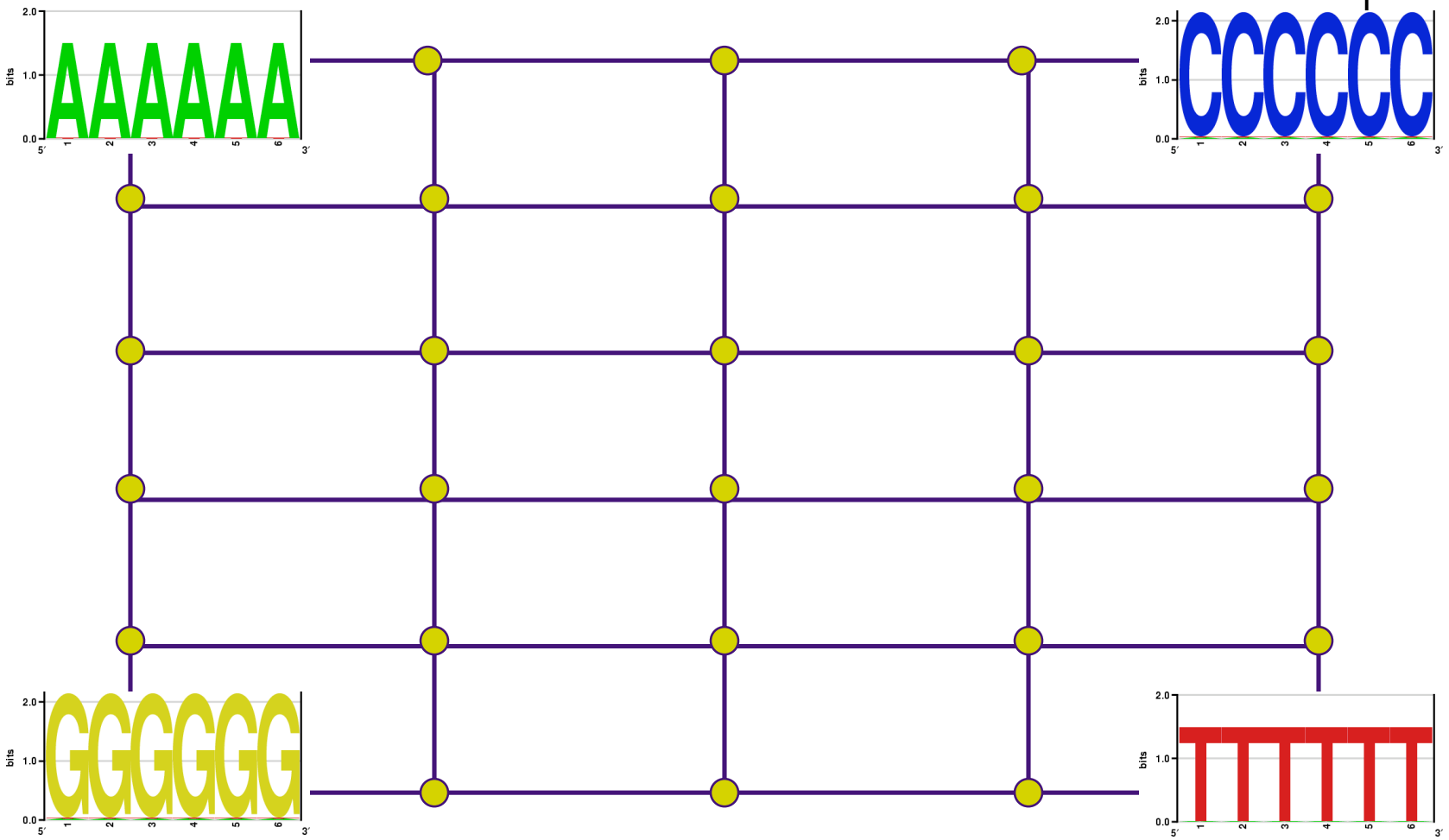
Characteristics.

1. Uses *Self-Organizing Map* to detect motifs and their relations (*unsupervised learning*)
2. Calculates significance of a motif
3. Identifies *multiple* motifs in a sequence set
4. Sets of known PSSMs can be used as priors

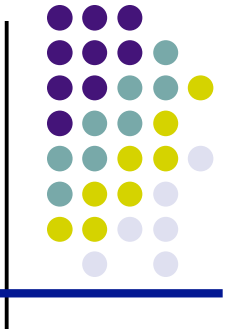
SOMBRERO (cntd)



SOMBRERO: initialization



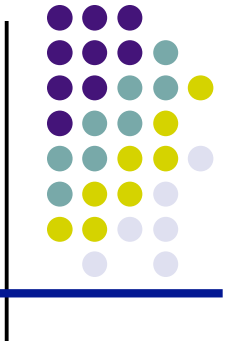
SOMBRERO (cntd)



				SOMBRERO				MEME			AlignACE		
Name	**seq**	**sites**	**bp**	**SOM**	**NP**	**FN**	**FP**	**NP**	**FN**	**FP**	**NP**	**FN**	**FP**
abf1	19	20	8600	40x20	25	**0.450**	0.560	11	0.550	0.182	16	0.500	0.375
csre	4	4	2550	20x10	11	**0.250**	0.727	6	0.500	0.667	17	**0.250**	0.824
gal4	4	14	3100	20x10	17	**0.071**	0.235	12	0.286	0.167	12	0.214	0.083
gcn4	9	25	4500	30x15	14	**0.600**	0.286	10	0.920	0.800	18	**0.600**	0.444
gcr1	6	9	3350	30x15	29	**0.222**	0.690	9	0.444	0.444	16	0.333	0.625
hstf	6	9	3400	30x15	21	**0.111**	0.571	24	0.333	0.750	18	**0.111**	0.556
mat	7	13	3500	30x15	12	0.308	0.250	15	**0.154**	0.267	9	0.308	0.000
mcb	6	12	3150	20x10	31	**0.083**	0.645	12	0.250	0.250	12	**0.083**	0.083
mig1	9	10	4500	30x15	25	**0.200**	0.680	0	1.000	1.000	11	0.900	0.909
pho2	3	6	2350	20x10	33	**0.500**	0.909	0	1.000	1.000	0	1.000	1.000

Table 2. Comparison of motif detectors on 10 yeast promoter sequence datasets. For each method, the number of predictions (NP), false negative (FN), and false positive (FP) rates are shown. The best FN rate in each dataset is highlighted in **bold**.

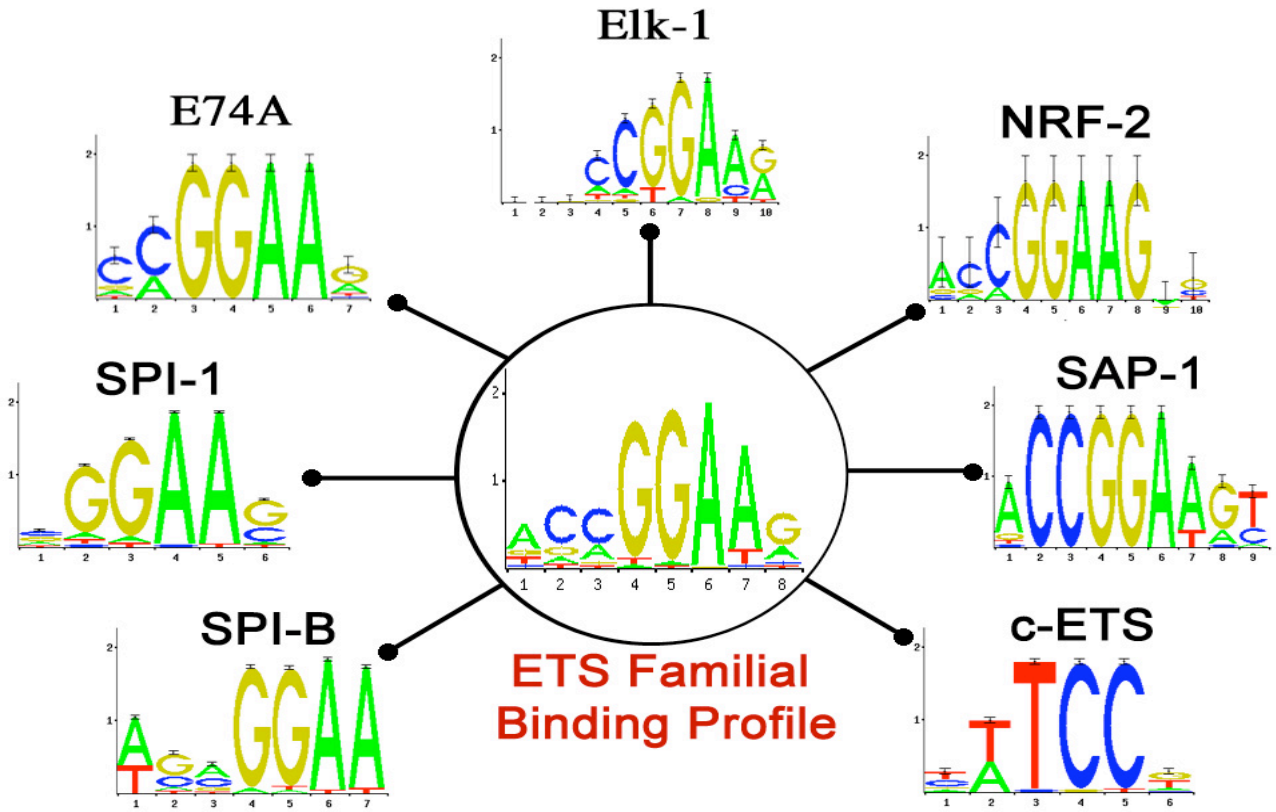
SOMBRERO (cntd)



		SOMBRERO		MEME		AlignACE	
	sites	FN	FP	FN	FP	FN	FP
bcd	23	**0.57**	0.80	0.87	0.93	0.78	0.83
cad	63	**0.43**	0.46	0.75	0.43	0.78	0.67
hb	119	**0.35**	0.40	0.82	0.21	0.77	0.37
kni	24	**0.76**	0.94	0.88	0.82	0.88	0.93
Kr	61	0.61	0.59	0.64	0.46	**0.52**	0.25

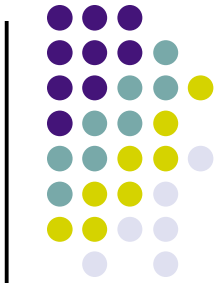
Table 3. Comparison of motif detectors on 19 Drosophila regulatory sequences that contain instances of 5 regulatory binding sites. The best FN rate for each motif is highlighted in **bold**.

Familial binding profiles



Source: Sandelin & Wasserman (2004) J. Mol. Biol. 338, 207-215

Binding Profile SOM

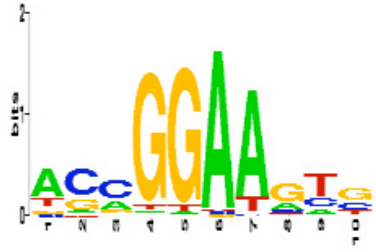


- Same methodology as SOMBRERO.
  - Except… motifs are clustered in each bin instead of sequences.
  - Motifs compared to one another using Pearson correlation coefficient and Smith-Waterman alignments.

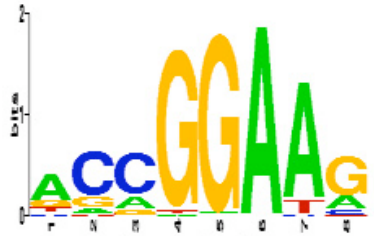
Binding Profile SOM (BP-SOM)

bHLH Tal1beta-E47S bHLH Thing1-E47	NUCLEAR Androgen HMG HMG-1	REL Dorsal_1	TRP-CLUSTER MYB.ph3 TRP-CLUSTER c-MYB_1 TRP-CLUSTER GAMYB
FORKHEAD HFH-2 FORKHEAD HNF-3b FORKHEAD HFH-1 FORKHEAD HFH-3 FORKHEAD FREAC-2 FORKHEAD FREAC-4	MADS AGL3 MADS MEF2 FORKHEAD FREAC-7	NUCLEAR RORalfa-1 NUCLEAR RORalfa-2	ETS SAP-1 ETS SPI-B ETS NRF-2 ETS SPI-B ETS E74A ETS c-ETS ETS Elk-1
MADS Agamous	MADS SRF	HMG SOX-9 HMG Sox-5 HMG SOX17 HMG SRY	HOMEO Nkx HOMEO S8
bZIP HLF bZIP E4BP4 bZIP cEBP bZIP CREB	bHLH-ZIP Max bHLH-ZIP USF bHLH-ZIP n-MYC bHLH-ZIP Myc-Max bHLH ARNT bZIP bZIP911 bZIP TCF11 bZIP bZIP910	TRP_CLUSTER Irf-1 TRP_CLUSTER Irf-2	bHLH Hen-1 bHLH Ahr-ARNT FORKHEAD FREAC-3
MADS SQUA HOMEO HNF-1 HMG HMG-IY	bHLH Myf	NUCLEAR CFI-USP NUCLEAR COUP-TF NUCLEAR RXR-VDR NUCLEAR PPARgamma NUCLEAR PPARgamma-rxr	REL p65 REL NF-kB REL p50 REL Dorsal_2 REL c-REL HOMEO EN-1 bZIP Chop-cEBP

ETS



BP-SOM

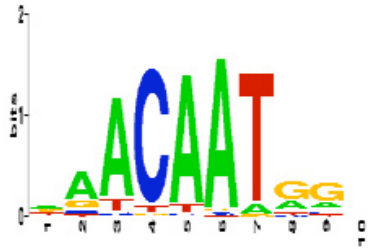


Sandelin & Wasserman

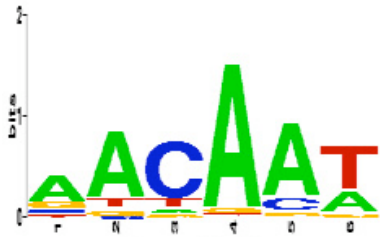
Binding Profile SOM (BP-SOM)

bHLH Tal1beta-E47S bHLH Thing1-E47	NUCLEAR Androgen HMG HMG-1	REL Dorsal_1	TRP-CLUSTER MYB.ph3 TRP-CLUSTER c-MYB_1 TRP-CLUSTER GAMYB
FORKHEAD HFH-2 FORKHEAD HNF-3b FORKHEAD HFH-1 FORKHEAD HFH-3 FORKHEAD FREAC-2 FORKHEAD FREAC-4	MADS AGL3 MADS MEF2 FORKHEAD FREAC-7	NUCLEAR RORalfa-1 NUCLEAR RORalfa-2	ETS SAP-1 ETS SPI-B ETS NRF-2 ETS SPI-B ETS E74A ETS c-ETS ETS Elk-1
MADS Agamous	MADS SRF	HMG SOX-9 HMG Sox-5 HMG SOX17 HMG SRY	HOMEO Nkx HOMEO S8
bZIP HLF bZIP E4BP4 bZIP cEBP bZIP CREB	bHLH-ZIP Max bHLH-ZIP USF bHLH-ZIP n-MYC bHLH-ZIP Myc-Max bHLH ARNT bZIP bZIP911 bZIP TCF11 bZIP bZIP910	TRP_CLUSTER Irf-1 TRP_CLUSTER Irf-2	bHLH Hen-1 bHLH Ahr-ARNT FORKHEAD FREAC-3
MADS SQUA HOMEO HNF-1 HMG HMG-IY	bHLH Myf	NUCLEAR CFI-USP NUCLEAR COUP-TF NUCLEAR RXR-VDR NUCLEAR PPARgamma NUCLEAR PPARgamma-rxr	REL p65 REL NF-kB REL p50 REL Dorsal_2 REL c-REL HOMEO EN-1 bZIP Chop-cEBP

HMG



BP-SOM

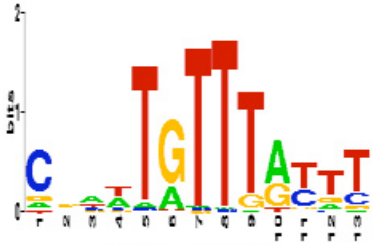


Sandelin
&
Wasserman

Binding Profile SOM (BP-SOM)

bHLH Tal1beta-E47S bHLH Thing1-E47	NUCLEAR Androgen HMG HMG-1	REL Dorsal_1	TRP-CLUSTER MYB.ph3 TRP-CLUSTER c-MYB_1 TRP-CLUSTER GAMYB
FORKHEAD HFH-2 FORKHEAD HNF-3b FORKHEAD HFH-1 FORKHEAD HFH-3 FORKHEAD FREAC-2 FORKHEAD FREAC-4	MADS AGL3 MADS MEF2 FORKHEAD FREAC-7	NUCLEAR RORalfa-1 NUCLEAR RORalfa-2	ETS SAP-1 ETS SPI-B ETS NRF-2 ETS SPI-B ETS E74A ETS c-ETS ETS Elk-1
MADS Agamous	MADS SRF	HMG SOX-9 HMG Sox-5 HMG SOX17 HMG SRY	HOMEO Nkx HOMEO S8
bZIP HLF bZIP E4BP4 bZIP cEBP bZIP CREB	bHLH-ZIP Max bHLH-ZIP USF bHLH-ZIP n-MYC bHLH-ZIP Myc-Max bHLH ARNT bZIP bZIP911 bZIP TCF11 bZIP bZIP910	TRP_CLUSTER Irf-1 TRP_CLUSTER Irf-2	bHLH Hen-1 bHLH Ahr-ARNT FORKHEAD FREAC-3
MADS SQUA HOMEO HNF-1 HMG HMG-IY	bHLH Myf	NUCLEAR CFI-USP NUCLEAR COUP-TF NUCLEAR RXR-VDR NUCLEAR PPARgamma NUCLEAR PPARgamma-rxr	REL p65 REL NF-kB REL p50 REL Dorsal_2 REL c-REL HOMEO EN-1 bZIP Chop-cEBP

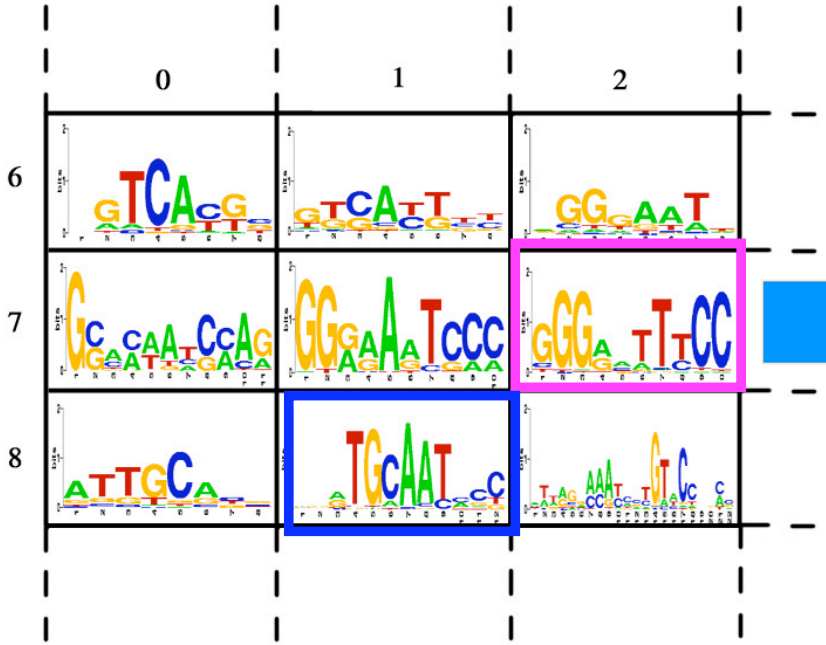
Forkhead

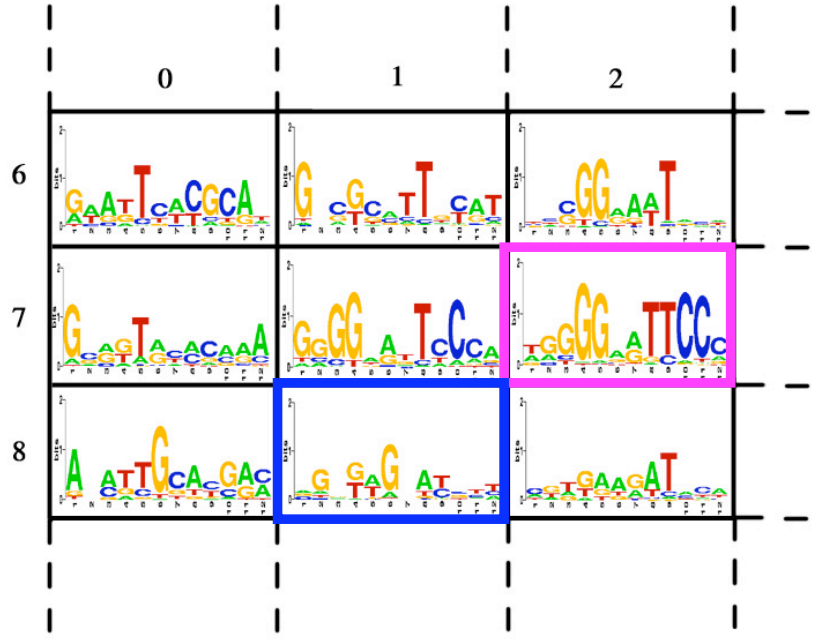


BP-SOM

Sandelin & Wasserman

SOMBRERO & FBP Priors

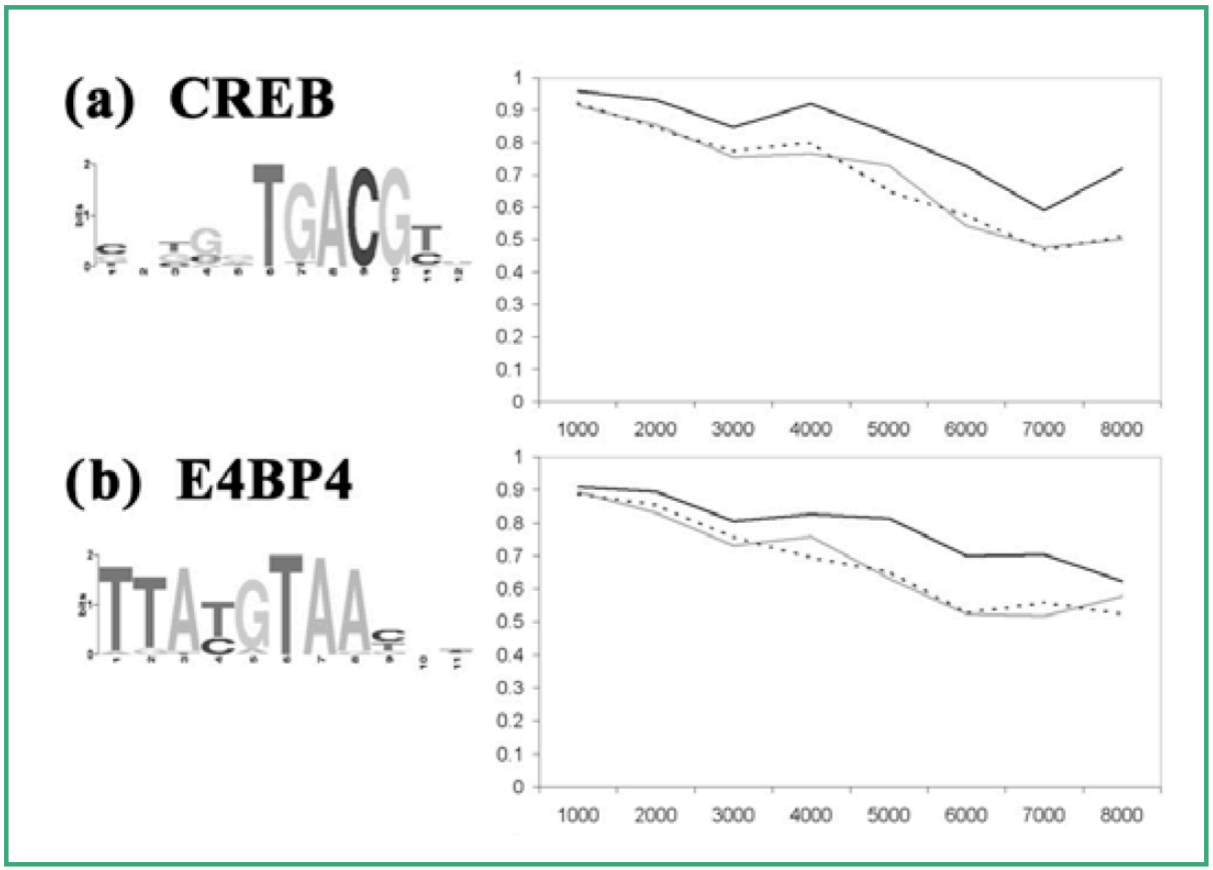




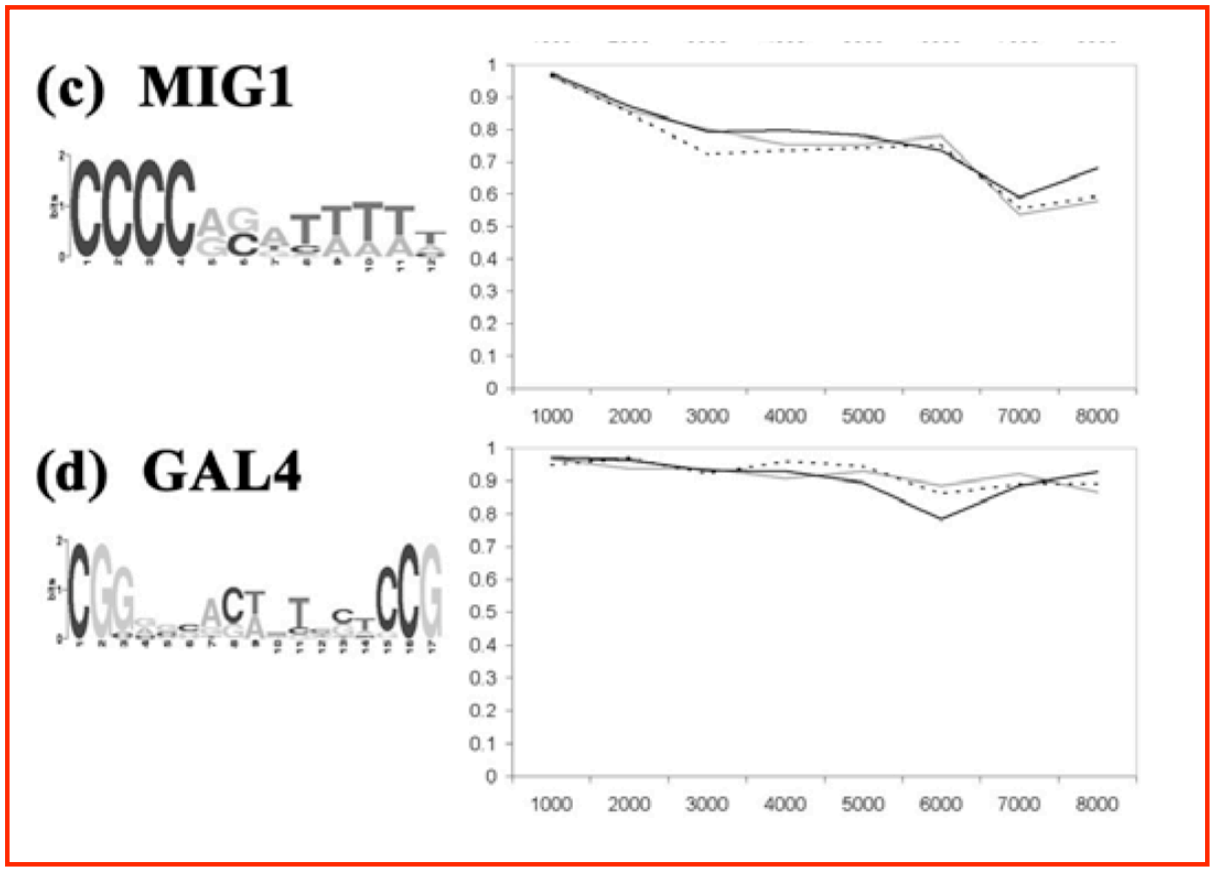
Start: BP-SOM trained on 257 mammalian PSSMs (incl. REL)

Finish: SOMBRERO used to find motifs of NF-κB (REL-class).

SOMBRERO: results



SOMBRERO: results (cntd)



SOMBRERO: results (cntd)

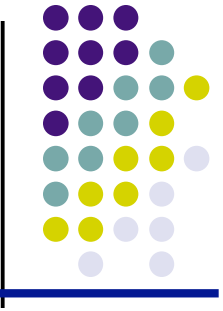


Table 1. Comparison of motif detectors on 10 yeast promoter sequence datasets. The best F score in each dataset is highlighted in **bold**.

			SOMBRERO (original initialisation)			SOMBRERO (with prior)			MEME			AlignACE		
	bp	sites	FN	FP	F	FN	FP	F	FN	FP	F	FN	FP	F
abf1	8600	20	0.45	0.56	*0.489*	0.40	0.29	***0.649***	0.55	0.18	*0.581*	0.50	0.38	*0.556*
csre	2550	4	0.25	0.73	***0.400***	0.00	0.75	***0.400***	0.50	0.67	***0.400***	0.25	0.82	*0.286*
gal4	3100	14	0.07	0.24	*0.839*	0.07	0.07	***0.929***	0.29	0.17	*0.769*	0.21	0.08	*0.846*
gcn1	4500	25	0.60	0.29	*0.513*	0.44	0.33	***0.609***	0.92	0.80	*0.114*	0.60	0.44	*0.465*
gcr1	3350	9	0.22	0.69	*0.389*	0.00	0.41	***0.720***	0.44	0.44	*0.556*	0.33	0.63	*0.480*
hstf	3400	9	0.11	0.57	*0.552*	0.11	0.53	***0.615***	0.33	0.75	*0.364*	0.11	0.56	*0.593*
mat	3500	13	0.31	0.25	*0.720*	0.15	0.27	*0.801*	0.15	0.27	*0.786*	0.31	0.00	***0.818***
mcb	3150	12	0.08	0.65	*0.512*	0.08	0.31	*0.786*	0.25	0.25	*0.750*	0.08	0.08	***0.917***
mig1	4500	10	0.20	0.68	*0.457*	0.10	0.47	***0.667***	1.00	1.00	*0.000*	0.90	0.91	*0.095*
pho2	2350	6	0.50	0.91	*0.154*	0.33	0.80	***0.364***	1.00	1.00	*0.000*	1.00	1.00	*0.000*
Avg			0.32	0.61	*0.493*	0.22	0.42	***0.663***	0.56	0.45	*0.489*	0.43	0.47	*0.550*

- 10 yeast datasets (test set)
- SOMBRERO trained on SCPD models

SOMBRERO: results (cntd)

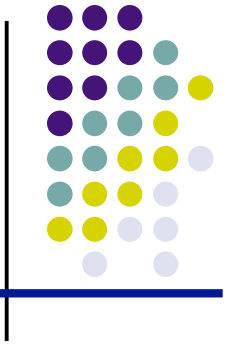
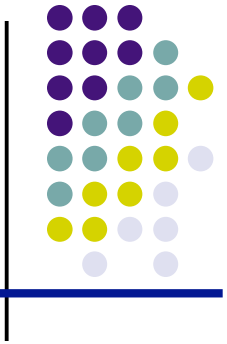


Table 2. Comparison of motif detectors on 19 *Drosophila* regulatory sequences that contain instances of 5 regulatory binding sites. The best F score for each motif is highlighted in **bold**.

		SOMBRERO (original initialisation)			SOMBRERO (with prior)			MEME			AlignACE		
	sites	FN	FP	F	FN	FP	F	FN	FP	F	FN	FP	F
bcd	23	0.57	0.80	*0.274*	0.43	0.73	***0.366***	0.87	0.93	*0.094*	0.78	0.83	*0.189*
cad	63	0.43	0.46	*0.554*	0.43	0.45	***0.562***	0.75	0.43	*0.352*	0.78	0.67	*0.264*
hb	119	0.35	0.40	*0.621*	0.50	0.16	***0.624***	0.82	0.21	*0.299*	0.77	0.37	*0.333*
kni	24	0.76	0.94	*0.095*	0.76	0.89	***0.150***	0.88	0.82	*0.146*	0.88	0.93	*0.086*
Kr	61	0.61	0.59	*0.400*	0.30	0.33	***0.683***	0.64	0.46	*0.431*	0.52	0.25	*0.581*

- 1 *Drosophila* dataset (>22 kb) with 5 motifs (test set)
- SOMBRERO trained on 75 *Drosophila* motifs

SOMBRERO: (dis)advantages



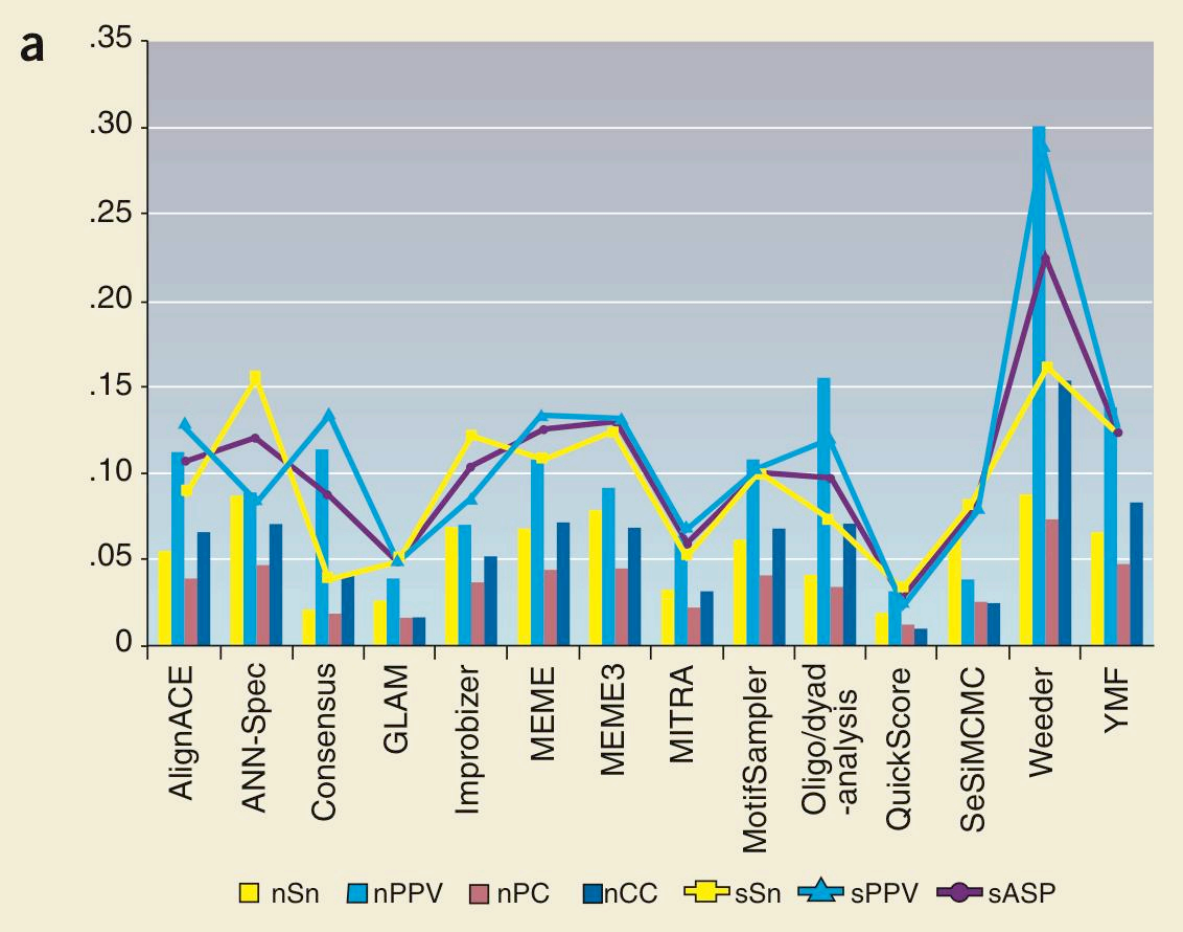
Advantages

- The transcription *family* that bind to the identified motifs can be predicted (even for newly sequenced organisms)
- Multiple motifs are reported
- Method compatible with phylogenetic footprinting
- Method compatible with alternative PSSM models

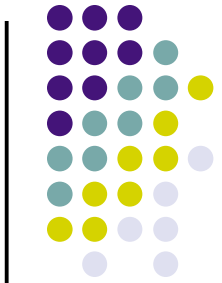
Disadvantages

- Computational cost
- Use of heuristics

Assessing performance



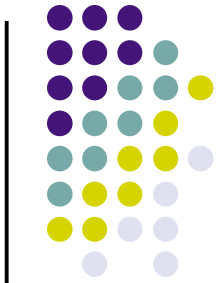
For some reading



- CONSENSUS: Hertz & Stormo, "*Identifying DNA and protein patterns with statistically significant alignments of multiple sequences*" (1999) *Bioinformatics* **15:** 563-577
- MEME: Bailey, Elkan, "*Fitting a mixture model by expectation maximization to discover motifs in biopolymers*" (1994) *Proc ISMB* **2:** 28-36
- Gibbs Sampler: Lawrence *et al.*, "*Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment*" (1993) *Science* **262:** 208-214
- SOMBRERO (with FBP initialization): Mahony, Golden, Smith, Benos, "*Improved detection of DNA motifs using self-organized clustering of* familial *binding profiles*" (2005) *Bioinformatics* **21** ***(Suppl 1)*****:** i283-i291

- Methods' comparison: Tompa *et al.*, "*Assessing computational tools for the discovery of transcription factor binding sites*" (2005) *Nat Biotechnol* **23:** 137-144

Acknowledgements



Some of the slides used in this lecture are adapted or modified slides from lectures of:

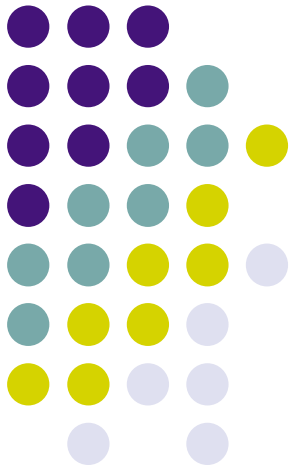
- Serafim Batzoglou, Stanford University
- Bill Noble, University of Washington, Seattle
- Eric Xing, Carnegie-Mellon University

Theory and examples from the following:

- R. Durbin, S. Eddy, A. Krogh, G. Mitchison, "Biological Sequence Analysis", 1998, Cambridge University Press
- T. Kohonen, "Self-Organizing Maps", 2001, Springer-Verlag

Graduate Computational Genomics

02-710 / 10-810 & MSCBIO2070



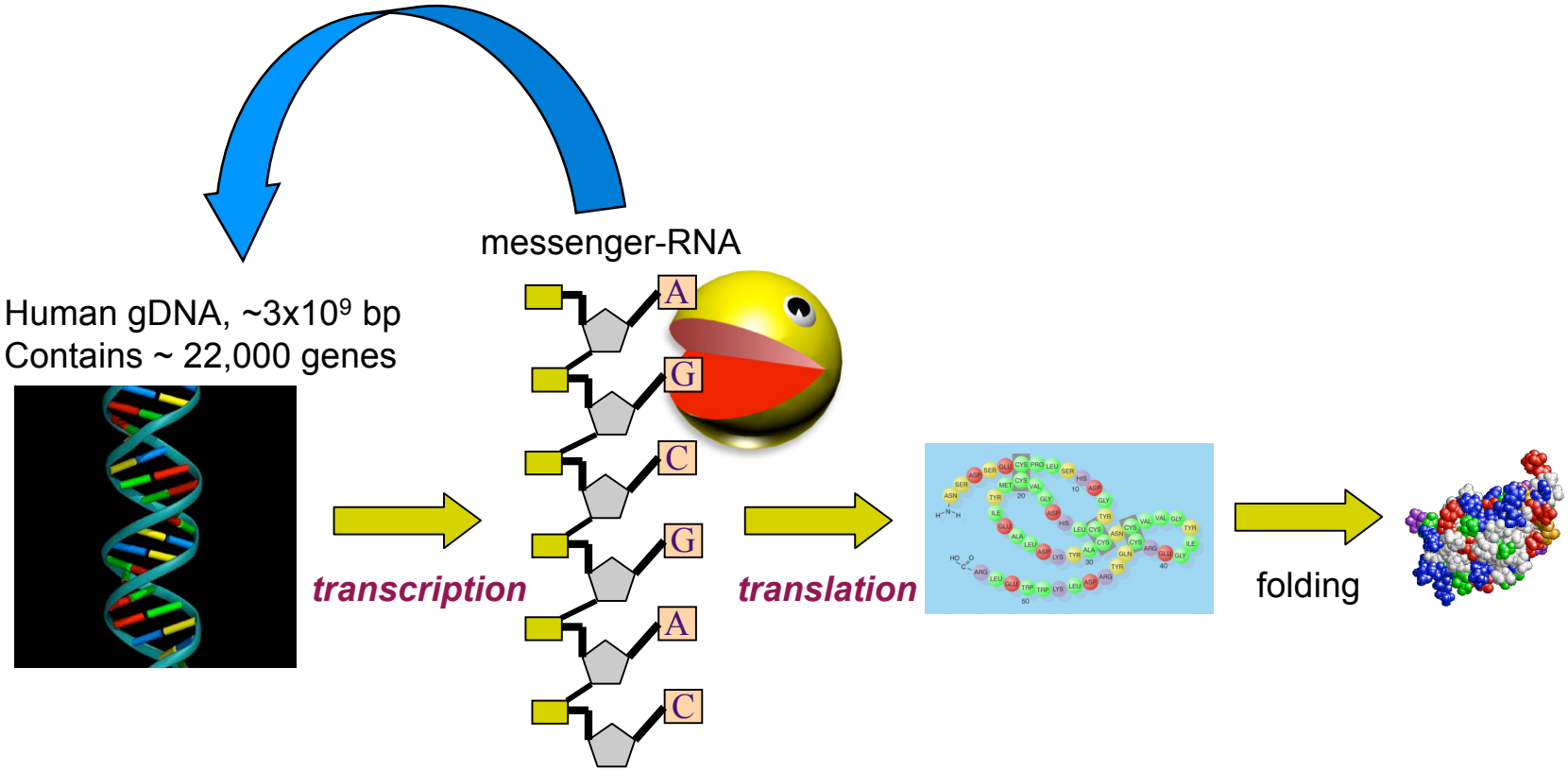
microRNA

Takis Benos

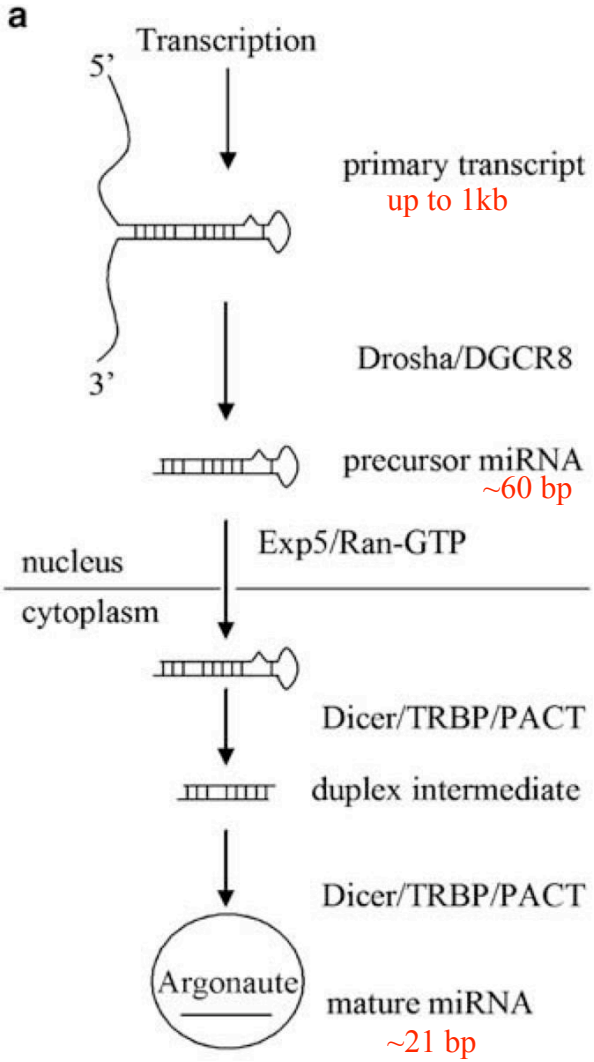
Lecture #11b, February 20, 2007

Reading: handouts & papers

Central Dogma

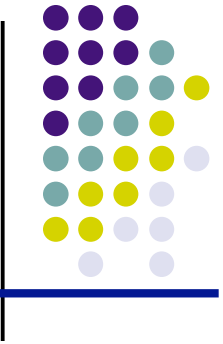


miRNA processing



- Size
  - 60-80bp pre-miRNA
  - 20-24 nucleotides mature miRNA
- Location: intergenic or intronic
- Regulation: *pol II* (mostly)
- Role: translation regulation, cancer diagnosis
- Biochemical identification
  - Low-efficiency, high cost, time consuming

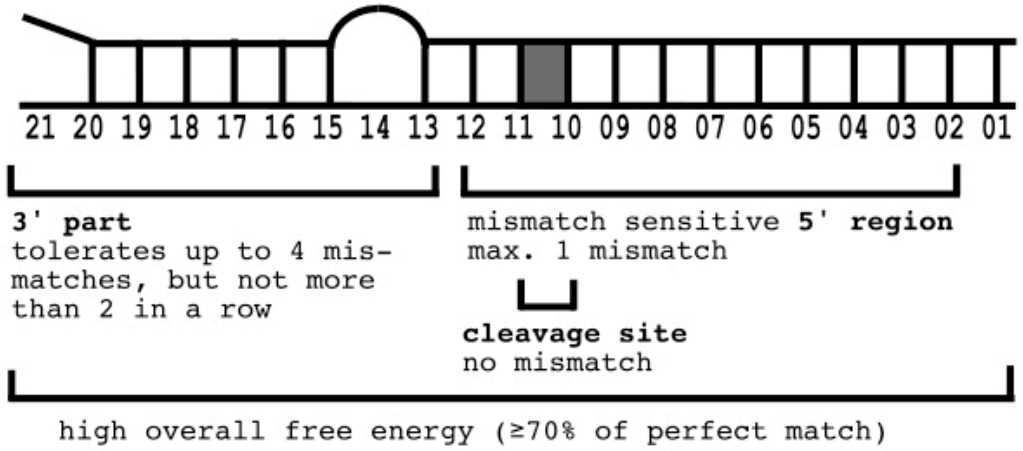
miRNA gene prediction/ identification



- Computational prediction :
  - Evolutionary conservation
    - Homology search (*direct BLAST searches*)
  - Stem loop secondary structure (hairpin)
    - Neighbor stem loop searches (*identify closely located stem loops*)
  - Folding free energy
    - Gene-finding (*identify conserved genomic regions, then run Mfold*)
- The numbers
  - ~4,000 miRNA genes from 45 species (most of them predictions) in *miRBase* (release 8.2, July 2006)

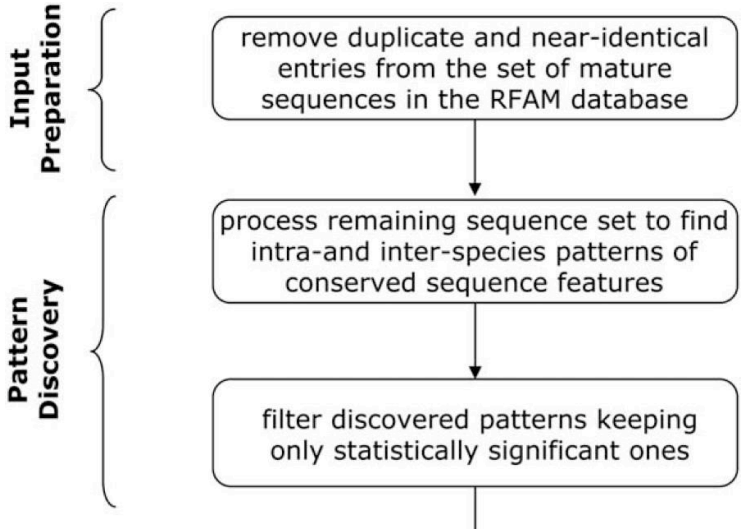
miRNA target prediction

- This is a more difficult task
  - High variability, small size ⇒ many false positives
- Potential binding
  - 3' UTR or in the ORF



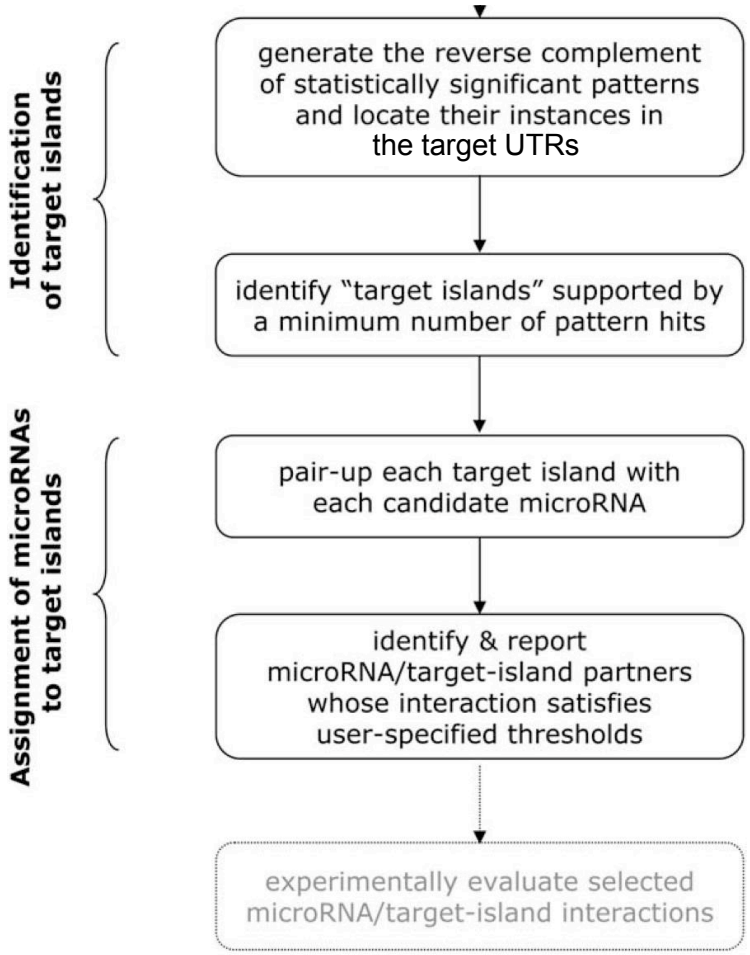
- Solution
  - Rely on evolutionary conservation (co-evolution)

Rna22: a different strategy



- Start: 644 mature miRNA sequences
- End: 354 sequences with ≤90% identity (training set)
- Pattern identification: *Teiresias* (on the training set)
- Significance: compare to a 2nd order Markov from the genome
- E.g.: [AT][CG].TTTTT[CG]G..[AT]

rna22 (cntd)

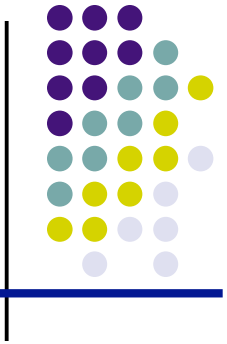


- Target islands: "hot spots" with ≥30 statistically significant mature miRNA patterns
- Results: *rna22* identifies correctly 17/21 "new" full-length sites

rna22 (cntd)


C. elegans
(lin-28) [328,342] 527
(lin-41) [689,710] 1162
(lin-41) [738,757] 1162
(cog-1) [267,285] 394
D. melanogaster
(hid) [874,896] 2269
(hid) [1711,1731] 2269
(hairy) [441,465] 816
(grim) [63,86] 964
(sickle) [34,58] 945
(bagpipe) [87,109] 165
H. sapiens
(hoxb8) [441, 432] 856
(mtpn) [3126,3141] 3165
(lin-28) [890,913] 2751
(mapk-14) [652,669] 2311
(fbxwb) [2393,2415] 2575
(laminin) [210,223] 1504
(dmtf-1) [130,146] 1187
(clock) [215,233] 2922
(cgi-38) [295,311] 363
(smc-1) [74,91] 5945
(bm3b) [103,109] 1663
(bm3b) [464,470] 1663
(enx-1) [60,66] 263
(enx-1) [115,121] 263
(bdnf) [221,227] 2925
(bdnf) [1323,1329] 2925
(g6pd) [98,104] 609
(g6pd) [434,440] 609
(smad) [47,54] 141
(smad) [104,110] 141
(nmyc) [495,501] 904
(sdf-1) [1387,1394] 1570
(sdf-1) [1452,1459] 1570
(mcsf-1) [800,807] 901
M. musculus
(limk1) [180,201] 1170

For some reading



- Baohong Zhang *et al.*, "*Computational idnetification of microRNAs and their targets*" (2006) *Computational Biology and Chemistry* **30:** 395-407

  and references therein…

- *rna22*: Miranda *et al.*, "*A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes*" (2006) *Cell* **126:** 1203-1217

Acknowledgements

Some of the slides used in this lecture are adapted or modified slides from lectures of:

- Serafim Batzoglou, Stanford University
- David Corcoran, University of Pittsburgh

Pictures from the following:

- http://wmd.weigelworld.org/bin/mirnatools.pl?page=6