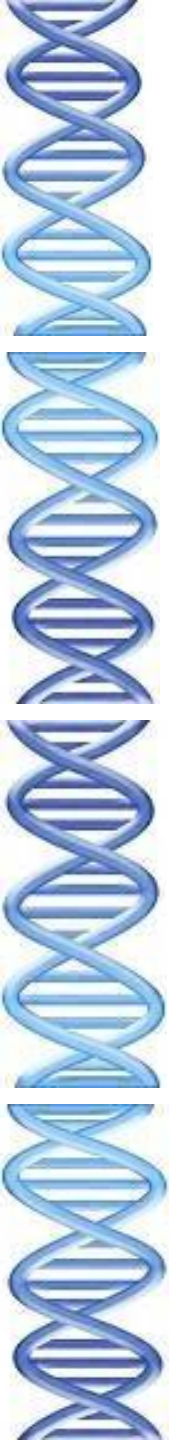


A survey of motif discovery methods in an integrated framework

(Biology Direct, 2006)

Authors: Geir Kjetil Sandve and Finn Drabløs

Talk by: Maarika Traat

About authors

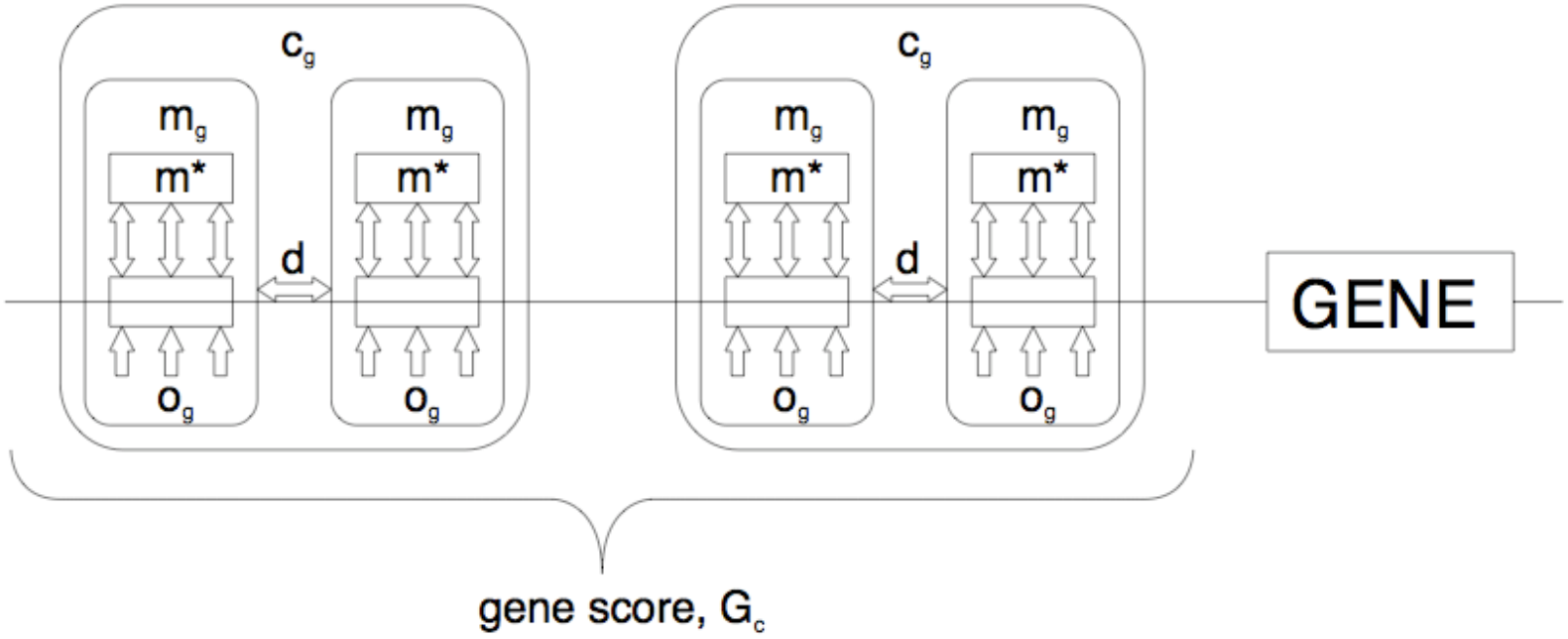
- Geir Kjetil Sandve
  - Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim (Algorithms, HPC, and Graphics group)
    homepage, publications
- Finn Drabløs
  - Department of Cancer Reseach and Molecular Medicine Norwegian University of Science and Technology, Trondheim
    homepage

An integrated framework

- Level 1: Single motif models
  - give a distinct score for each sequence segment in a regulatory system
  - score: match btw the segment and a motif consensus model + prior belief (occurrence score)
- Level 2: Composite motif models
  - modules: clusters of TFs that bind to DNA in proximity to each other (set of single motifs), certain flexibility about distance btw binding sites.
  - score: score of single motifs + inter-motif distances
- Level 3: Gene level models
  - several modules act together to determine the regulation of a single gene
  - gene score: ∑composite motif scores
- Level 4: Genome level models

An integrated framework (contd.)

- Level 4: Genome level models
  - sets of modules act on sets of genes
  - scores at this level used for evaluation and ranking of *de novo* discovered motifs



Single motif models (Level 1)

- TFs bind to specific short segments of DNA: TF binding sites (e.g. TATAAAA, TATATAT)
  - Match model: gives the degree of match btw the substring beginning at position *p* and and underlying consensus model
  - The occurrence prior: gives the prior belief that position *p* represents a regulatory element for gene *g*

Match models

- Match model *m*(p)* is a fn that gives a distinct score for any given substring
- Deterministic match model
  - binary score (0,1): hit/no hit
  - oligos (TATAAAA)
  - regular expressions: exact, ambiguous symbols and fixed/flexible gaps (TATAA*)
  - mismatch expressions

  (hamming dist 1: TATAAAA, matches also TATATAA)
- Probabilistic match models
- PMs more expressive than DMs, but DMs allow exhaustive discovery of optimal motifs

Probabilistic match models

- weighted score
- position weight matrix (PWM, or PSSM), assumes iid

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.33	A	-0.79	0.13	-0.23	-2.20	**1.05**	-2.20	-2.20	-2.20	-2.20	-2.20	-0.79	-0.23
0.18	C	0.32	0.32	**0.70**	**1.65**	-2.20	**1.65**	-2.20	-2.20	-2.20	0.32	-2.20	0.32
0.18	G	-0.29	0.32	**0.70**	-2.20	-2.20	-2.20	**1.65**	-2.20	**1.19**	**0.97**	**1.19**	0.32
0.33	T	0.39	-0.79	-2.20	-2.20	-2.20	-2.20	-2.20	**1.05**	0.13	-0.23	-0.23	-0.23
1	Sum	-0.37	-0.02	-1.02	-4.94	-5.55	-4.94	-4.94	-5.55	-3.08	-1.13	-2.03	0.19

- extensions: positional dependencies
  - PWM with pairs of correlated positions
  - mixture model (motif occurs as a stochastic prototype)
  - n'th order Markov chains
    - ✓ a variable-length Markov Model (VLMM): if the relative importance of dependencies varies within a motif
    - ✓ for long range (not neighbouring positions) dependencies:
      - permutation of positions in the Markov chain + VLMM
  - Bayesian networks

Occurrence priors

- Prior belief that an (unspecified) regulatory element is located at the given position *p*
- concerning genetic context → important for activity (distance to TSS, DNA structure, presence of CpG-islands)
  - spatial distribution of binding sites: most known regulatory element immediately upstream from TSS
  - conservation in orthologous sequences: phylogenetic footprinting - only search for motifs in highly conserved sequence parts
  - DNA structure: 3D structure of the DNA: bendability of a region, position in DNA loops
  - nucleotide distribution high CG content, presence of CpG islands

Composite motif models (Level 2)

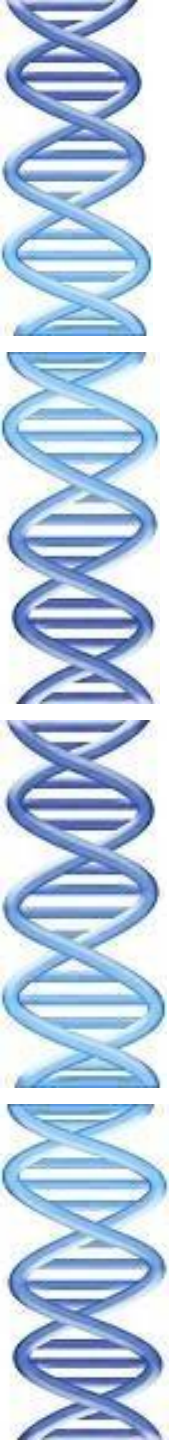
- Clusters of binding sites for cooperating TFs - modules
  - both sequential order and distance of independent binding site can be important
  - score: the sum or product of individual motifs and the distance scores
  - distance fns: constraints (fixed dist, dist below threshold, dist within interval, window of certain length, non-binary score fns: increase linearly with distance; also conservation fns of inter-motif distances), usually in base pairs
  - combining single motifs:
    - ✓ deterministic models: intersection of single motif scores (exceptional case: m out of n single motif scores required to be 1; also motif count can be used directly)
    - ✓ if single motif scores non-binary: calculate sum of motif scores and distance scores (variations)

Gene level models (Level 3)

- gene score: calculated from composite motif scores across the regulatory region of gene *g*
- gene level score:
  - some methods define it as the max motif score (assumes exactly 1 relevant occurrence of motif in the regulatory region)
  - most sum the (log-)scores of all motifs in the area, or the ones above a certain threshold
    - ✓ variations: *p*-value of the observed set of motif scores;
    - ✓ use logistic regression, ANN, etc.
  - calculate the score from the scores of composite motifs

Genome level models (Level 4)

- Motif scores at the genome level are generally used for significance evaluation of *de novo* motifs
- genome score usually based on either overrepresentation of the motif, or on the correspondence between gene scores and experimental data



Genome level models (Contd.)

- Motif overrepresentation
- Correspondence with experimental data
- Some Algorithmic concerns

Comparison of methods

Table 1: Overview of methods. The match model is the consensus representation of a single motif, motif combination is how the component scores of a composite motif are combined, and distance score is how the conservation of inter-motif distances within a composite motif is modeled.

ALGORITHM NAME	MATCH MODEL	MOTIF COMBINATION	DISTANCE SCORE
Weeder [42]	mismatch	-	-
Dyad analysis [35]	oligos	dyad[1]	constraint
MCAST [71]	PWM	sum	gap penalty
REDUCE [67]	PWM	dyad	constraint[2]
MDScan [87]	PWM	-	-
Gibbs sampler [97]	PWM	intersection[3]	uniform
MEME [98]	PWM	-	-
LOGOS [73]	DM	HMM	distribution
Motif regressor [89]	PWM	-	-
ModuleSearcher [70]	PWM	sum	window[4]
Stubb [48]	PWM	HMM	window
GANN [60]	flexible	ANN[5]	window
ANN-Spec [86]	PWM	-	-
(Wasserman) [58]	PWM	Logistic regr.	window
CoBind [68]	PWM	sum	window
Cister [72]	PWM	HMM	distribution
SeSiMCMC [122]	PWM	-	-
SMILE [40, 123]	mismatch	intersection	constraint
BioProspector [49]	PWM	sum	constraint
(Segal) [94]	PWM	-	-
(Sinha) [33]	reg.exp	dyad	constraint
ConsecID [56]	PWM	intersection	window
SCORE [69]	IUPAC	intersection	window
Gibbs recursive [52]	PWM	mixture model	distribution
(Hong) [95]	PWM	-	-
AlignACE [124]	PWM	-	-
Improbizer [117]	PWM	-	-
CisModule [119]	PWM	mixture model	mixture model
(Thompson) [66]	PWM	Markov model	constraint