A Parametric Joint Model of DNA-Protein Binding, Gene Expression and DNA Sequence Data to Detect Target Genes of a Transcription Factor

Wei Pan

(joint work with Peng Wei, Arkady Khodursky)

Division of Biostatistics, SPH

University of Minnesota

PSB 2008, Big Island, Hawaii

Jan 7, 2008

Problem

- Goal: discovery of regulatory targets of a TF.
  Example: LexA in *E. coli*.
- Data: DNA-protein binding; gene expression and DNA sequence data.
  Each type of the data has been *separately* used; a large literature.
  Here, use all three *simultaneously*!
- Why joint modeling?
  high noise level with high-throughput data: combining information;
  different strengths of the data.
- Approach: (mostly) unsupervised learning (only few known target); mixture model with empirical Bayes.

Model

- Statistical problem: testing $H_{0,i}$ vs $H_{1,i}$ for each gene i.
  $H_{0,i}$: gene i is not a target of the TF;
  $H_{1,i}$: opposite to $H_{0,i}$.

- $T_i = 0$: $H_{0,i}$ is true;
  X_i: a statistic measuring relative abundance of the TF bound to gene i; binding data.
  Y_i: a statistic for differential expression for gene i; expression data.
  Z_i: a score measuring the degree to which one of its subsequences matches a known motif for the TF; DNA seq data.

- Mixture distribution for (X_i, Y_i, Z_i):

$$f(x, y, z) = \pi f_1(x, y, z) + (1 - \pi) f_0(x, y, z),$$

$\pi = Pr(H_{1,i})$;

f_0 and f_1: for genes with $H_{0,i}$ and $H_{1,i}$ being true respectively.

- Conditional independence:

$$f(x,y,z) = \pi f_{11}(x;\theta_{11}) f_{12}(y;\theta_{12}) f_{13}(z;\theta_{13}) + (1-\pi) f_{01}(x;\theta_{01}) f_{02}(y;\theta_{02}) f_{03}(z;\theta_{03}),$$

  θ_{jk}: unknown parameters for f_{jk}.

- Further assume each f_{jk} to be Normal.
  $\theta_{jk} = (\mu_{jk}, \sigma_{jk})$.

- Parameter estimation: EM algorithm.

- Inference: use posterior probabilities

$$Pr(H_{1,i}|X_i, Y_i, Z_i) = \frac{\pi f_{11}(X_i;\theta_{11}) f_{12}(Y_i;\theta_{12}) f_{13}(Z_i;\theta_{13})}{f(X_i, Y_i, Z_i)}.$$

  Rank the genes by their $Pr(H_{1,i}|X_i, Y_i, Z_i)$...

E coli data

- Binding data of Wade et al (2005, Genes & Devel.):
  4 Affy arrays;
  Largely followed the authors to obtain $X_i =$ *peak* log intensity ratio (LIR) for gene i.
- Expression data of Courcelle et al (2001, Genetics):
  4 cDNA arrays with a common control sample: wild-type before and 20-min after UV treatment; lexA mutatnt before and 20-min after UV treatment.
  $\Longrightarrow M_{1i}$,...,M_{4i}, normalized log intensity ratios for gene i on the four arrays;
  $\Longrightarrow Y_i = (M_{1i} - M_{2i}) - (M_{3i} - M_{4i})$.
- DNA data:
  Why relevant? LexA is known to be a repressor of some "SOS response" genes.

Extracted DNA seq in July 2006;
Downloaded 10 known binding sites of LexA from regulonDB (v4,0); 9 genes, one with 2 binding sites;
Input all or only a half of them to MEME (Bailey and Elkan 1995, ML) to find a top motif;
Used scanACE (Roth et al, 1998, Nat Genet) to scan the genome with a very low threshold, yielding at least one matching subseq for most genes;
Z_i = highest matching score for gene i.

- Combining the data and deleting genes with missing data, we obtained $G = 3779$ genes.

Results

Known/putative targets listed in regulonDB (v5.5) (Salgado et al 2006), downloaded in November 2006:

Gene	Bind	Expr	B+E	Seq1	Seq2	B+E+S1	B+E+S2
polB	156	114	135	153	1593	127	146
phrB	1346	1826	2083	530*	81**	1516	452
uvrB	48	172	92	31*	6**	78	46
dinG	96	448	213	138	143	169	171
ftsK	75	3757	223	127	303	173	199
sulA	11	12	1	17*	728	1	1
umuD	31	29	1	19	8	1	1
umuC	192	12	1	3454	3652	34	37
ydjM	30	111	53	70	74	49	44
ruvB	2780	313	509	1471	2966	645	708
ruvA	127	147	141	10	38	94	108
uvrC	3015	3104	3646	3008	796	3377	2692
uvrY	3538	3473	3679	3008	796	3384	2685

Gene	Bind	Expr	B+E	Seq1	Seq2	B+E+S1	B+E+S2
recN	7	5	1	33	36	1	1
oraA	82	50	54	1220	871	61	59
recA	12	15	1	23*	4**	1	1
rpsU	464	1214	766	1097*	304	896	572
dnaG	2906	3621	3451	782	177	2620	954
rpoD	2906	3749	3455	782	177	2621	953
t150	2121	175	262	50	76	176	178
uvrD	263	245	274	4*	50	106	160
lexA	15	61	1	7*	1**	1	1
dinF	2549	217	323	7	1	118	77
uvrA	41	169	77	14*	114	58	72
ssb	41	143	74	14*	114	54	68

Putative targets with a common motif (Class II, first 5) and without any common motif (Class III) identified based on only the binding data by Wade et al.:

Gene	Bind	Expr	B+E	Seq1	Seq2	B+E+S1	B+E+S2
fadE	147	1578	399	1563	293	456	348
mmuP	137	125	136	981	1031	150	143
clpX	160	3252	473	432	796	412	517
ybbJ	196	2024	553	2281	1018	593	455
ybbK	128	1391	349	3303	1269	378	391
intD	195	2532	564	3463	3458	704	719
ybeR	94	3070	285	413	108	258	195
ybeS	1634	1771	2290	2378	3354	2926	2707
ybeX	58	3220	168	3496	3016	182	182
ybjK	34	2457	88	2378	2958	90	93
minC	20	360	58	35	50	46	48
ydeQ	81	2253	227	1042	350	234	204

Gene	Bind	Expr	B+E	Seq1	Seq2	B+E+S1	B+E+S2
ydfJ	136	1706	393	2967	1166	452	442
hdhA	175	2395	505	3394	2251	612	680
malI	269	1609	703	2502	2251	933	970
ydjF	55	203	111	873	1111	115	118
yoaC	99	2781	293	2301	1474	315	330
yebN	98	972	265	2361	781	279	281
otsA	3246	2645	3327	296	728	1322	2389
otsB	49	1450	133	266	295	129	132
idi	66	592	163	3394	2505	175	180
b3776	106	130	129	848	2902	125	138
polA	46	2961	122	1422	452	124	121
cadA	83	2456	245	1028	2003	248	282
cadB	71	3109	205	1028	2003	208	229

Discussion

- Extension of Wang et al (2002, PNAS): from 2 types of data to 3 types of data.
- Parametric vs nonparametric (Pan et al, in press, Statistica Sinica).
- Posterior probabilities can be used to estimate FDR. depends on the correctness of the model.
- Extension: a mixture for each component.
- Extension to semi-supervised learning.
- Extension: incorporation of operon info, gene functions and gene networks (Pan 2006; Wei and Pan, in press, Bioinformatics).....

Acknowledgement: This research was supported by NIH and a UM AHC Faculty Research Development grant.

You can download our papers from
http://www.biostat.umn.edu/rrs.php

Thank you!