

Applications of Akaike and Bayesian Information Criteria

**in epidemiology, linkage analysis,
microarray data analysis, and DNA
sequence analysis**

Wentian Li, Ph.D

Lab of Statistical Genetics
Rockefeller University

The current genetic analysis is mainly confined to two theoretical frameworks:

1. maximum likelihood

2. hypothesis testing

Example: Morton's LOD score approach. Pedigree likelihood is maximized by adjusting the recombination fraction θ under a single-gene Mendelian model, then the null hypothesis $\theta = 0.5$ is tested by the likelihood ratio test.

These two frameworks serve us well for single-gene Mendelian diseases. With more and more non-Mendelian (complex) diseases to study, we want to reexamine the two frameworks themselves.

re-examine the maximum likelihood framework

Likelihood is the probability of observing the data, *give a model*: $P(y|\theta, M)$

(y is the data at hand, M is the model, θ is the collection of all parameters in the model)

If the model M is not selected appropriately, even the best (maximized) likelihood does not fit the data well. There is a need to explore and to expand to other models. Maximum likelihood framework does not address this issue.

*re-examine the hypothesis testing
framework*

Rejecting the null does not provide a guidance on which alternative is correct. There is no appropriate way to judge different alternative models.

Also, The artificial dichotomy between $\theta = 0$ and $\theta > 0$ creates difficulties in interpreting a test result.

An arbitrary significance level (e.g. 0.01) is introduced.

Likelihood ratio test is not “dimensionally consistent” (when sample size goes to ∞ , false positive rate remains nonzero).

“The maximum likelihood principle has mainly been utilized in two different branches of statistical theories. The first is the **estimation theory**, where the method of maximum likelihood has been used extensively, and the second is the **test theory** where the log-likelihood ratio statistics is playing a very important role. ... these two problems should be combined into a **single problem of statistical decision**. Thus instead of considering a **single estimate** of θ , we consider **estimates corresponding to various possible restrictions of the distribution**, and instead of treating the problem as a **multiple decision** or a **test between hypotheses**, we treat it as a **problem of general estimation procedure based on the decision theoretic consideration**.” [Hirotogu Akaike, 1973]

To combine the maximum likelihood (data-fitting) and the choice of model, we would like to penalize the (log) maximum likelihood with a term related to the model complexity.

A typical penalty is like this:

$$2 \log(\hat{L}) - \alpha K$$

where K is the number of free parameters in the model.

$$\alpha = 2$$

Akaike Information Criterion

With a minus sign, maximizing the penalized likelihood becomes a minimizing AIC:

$$AIC = -2 \log(\hat{L}) + 2K$$

Note

$$e^{-\frac{1}{2}AIC} = \hat{L}e^{-K}$$

So the likelihood is modified by a factor of $\exp(-K)$

For pairwise nested model comparison, AIC is similar, but not identical, to the likelihood ratio test:

likelihood-ratio test:

$$2 \log \left(\frac{L_2}{L_1} \right) \xrightarrow{N \rightarrow \infty} \chi_{K_2 - K_1}^2$$

AIC model selection ($AIC_2 < AIC_1$ to reject M_1):

$$2 \log \left(\frac{L_2}{L_1} \right) > 2(K_2 - K_1)$$

it is equivalent to a likelihood-ratio test with a significance level which **changes with $K_2 - K_1$ (ΔK)**. e.g. significance levels are 0.157 ($\Delta K = 1$), 0.075 ($\Delta K = 5$), 0.029 ($\Delta K = 10$), ... 1.37×10^{-8} ($\Delta K = 100$).

i.e., in AIC model selection, it is more and more difficult (more stringent condition) to select M_2 when M_2 becomes more complex. in contrast, the likelihood-ratio test with a fixed significance level set the rejection condition too relax.

Even in $N \rightarrow \infty$ limit, false positive rate

$P(\text{accept } M_2 | M_1 \text{ true})$ does not approach 0 (dimensionally inconsistent).

$$\alpha = \log(N)$$

Bayesian Information Criterion

$$BIC = -2 \log(\hat{L}) + \log(N)K$$

Note

$$e^{-\frac{1}{2}BIC} = \hat{L}(\sqrt{N})^{-K}$$

e is replaced by \sqrt{N} . When the sample size $N = 7.389$ (quite small!), AIC and BIC are the same.

BIC vs likelihood-ratio test in a pairwise nested model comparison:

likelihood-ratio test:

$$2 \log \left(\frac{L_2}{L_1} \right) \xrightarrow[N \rightarrow \infty]{} \chi_{K_2 - K_1}^2$$

BIC model selection ($\text{BIC}_2 < \text{BIC}_1$ to reject M_1):

$$2 \log \left(\frac{L_2}{L_1} \right) > \log(N)(K_2 - K_1)$$

it is equivalent to a likelihood-ratio test with a significance level which **changes with both ΔK and sample size N** .

significance levels based on BIC

ΔK	N			
	7.389	10	100	1000
1	0.157	0.129	0.032	0.0086
5	0.075	0.042	3.3×10^{-4}	1.9×10^{-6}
10	0.029	0.011	1.4×10^{-6}	6.7×10^{-11}
100	$\sim 10^{-7}$	$\sim 10^{-12}$	~ 0	~ 0

in the $N \rightarrow \infty$ limit, the significance level is approaching 0.

When AIC/BIC is cast in the likelihood-ratio test framework, the significance level is not fixed: it changes with the difference of model complexity in two models. AIC/BIC tends to make it more difficult to reject the null when the alternative is too complex.

AIC and likelihood-ratio test with a fixed significance level are not consistent (or “dimensionally consistent”, false positive rate does not approach zero in the $N \rightarrow$ limit). On the other hand, BIC is consistent (false positive rate approaches 0 in the $N \rightarrow$ limit).

Derivation of AIC

Suppose y is the data available, \tilde{y} is the future data, M is a given model. The Kullbak-Liebler distance between the true distribution of \tilde{y} and the **posterior predictive distribution evaluated at the maximum likelihood estimation of the parameter value**:

$$I = \int_{d\tilde{y}} p(\tilde{y}) \log \frac{p(\tilde{y})}{p(\tilde{y}|\hat{\theta}(y), M)} > 0$$

The smaller the distance, the better the predictive distribution. If we assume that the data available is sampled from a distribution itself, we can average over all possible available data sets:

$$E_y[I] = E_{\hat{\theta}(y)}[I] = \int_{dy} p(y) \int_{d\tilde{y}} p(\tilde{y}) \log \frac{p(\tilde{y})}{p(\tilde{y}|\hat{\theta}(y), M)}$$

Akaike derived an approximation for $E_y[I]$:

$\text{const} - \log(L(\hat{\theta}|y)) + K = \text{const} + \text{AIC}/2$. The key is that $E_y[I]$ does not depend on the $p(\tilde{y})$.

Minimizing AIC is equivalent to minimizing the (averaged) Kullbak-Liebler distance.

Derivation of BIC

The posterior probability of model M is:

$$p(M|y) = \frac{p(y|M)p(M)}{p(y)}$$
$$\propto \int p(y|\theta, M)p(\theta|M)d\theta \cdot p(M)$$

posterior \propto (integrated likelihood) \times prior

BIC is an approximation of the $-2 \log$ (integrated likelihood)

Laplace method of integral is used (a Taylor expansion of $\exp(f(x))$))

Minimizing BIC is equivalent to maximizing integrated likelihood, which is equivalent to maximizing posterior probability of a model when the priors are all equal.

AIC/BIC based analysis is one approach for data fitting when the model uncertainty is taken into consideration. It is easy to use because it is simply a maximum likelihood approach plus a model correction.

Remember that AIC/BIC selects one model from many alternatives. A further extension is to combine many alternatives rather than selecting one (model averaging).

example 1: assessing risk factors in epidemiology analysis

$$y = f(x_1, x_2, x_3, \dots, x_p)$$

y: affection status (0,1)

$\{x_i\}$: risk factors.

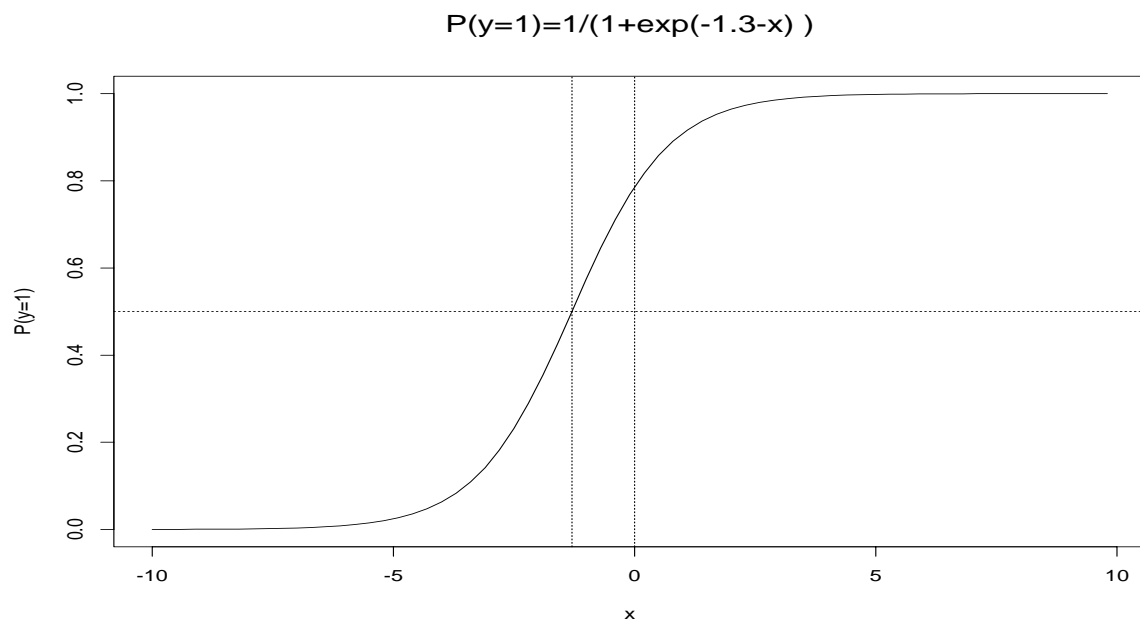
model: logistic regression

$$p(y=1) = 1/(1 + \exp(-a_0 - \sum_{i=1}^p a_i x_i))$$

how to use AIC/BIC: select a subset of risk factors that minimize AIC/BIC. here the model selection is a variable selection. each variable is accompanied by a coefficient (a free parameter).

Logistic Regression:

“prediction”, “discrimination”,
“decision”, “classification”



linear combination of variables (hyperplane - neural networks use nonlinear hypersurface)

likelihood is easy to calculate (product of these probabilities for individual data point)

$-a_0 \cdot \vec{a}^{-1}$ determines the threshold value, $2 \vec{a}^{-1}$ determines the “soft margin” region (undecided)

a reasonable middle ground between traditional statistics and machine learning.

Data: Child Asthma

$N=7318$ children in the data set

mothers complete questionnaires during pregnancy and post-natally (after 6 and 42 months)

information on many risk factors – putative, perinatal, social, parental, atopic, and environmental – are collected. $p=24$ factors are used in this analysis.

$y=1$ for infant wheeze (affected). $N_1=1316$ children are affected (18%), and $N_0=6002$ are unaffected (82%).

Child Asthma: Risk Factors

The meaning of risk factors (L for leveled categorical variable, D for dummy variable with more than two states, others are binary):

x_1 : gender

x_2 : whether mother has asthma or not

x_3 : whether has father or not

x_4 : whether father has asthma or not

x_5 : age of mother (L)

x_6 : breast fed or not

x_7 : whether low birth weight or not

x_8 : whether preterm delivery or not

x_9, x_{10}, x_{11} : season of birth (D)

x_{12} : whether mother has education or not

x_{13} : mother's education level (L)

x_{14} : whether has elder siblings or not

x_{15} : whether has more than 1 elder sibling

$x_{16}, x_{17}, x_{18}, x_{19}$: rash type 1,2 at 6,42 months

x_{20}, x_{21}, x_{22} : housing type (D)

x_{22} : tobacco smoke exposure level (L)

x_{23} : whether has pet or not

x_{24} : whether has cats/dogs or not

single risk factor

model/var	-2log(L)	K	AIC	Δ AIC	BIC	Δ BIC
x_{14} (elder sib)	6810.95	2	6814.95	-82.60	6828.75	-75.70
x_6 (breast fed)	6849.08	2	6853.08	-44.47	6866.88	-37.57
x_1 (gender)	6855.49	2	6859.49	-38.06	6873.29	-31.16
x_7	6895.32	2	6899.32	1.77	6913.12	8.67
fixed	6895.55	1	6897.55	0	6904.45	0
random	10144.90	0	10144.90	3245.58	10144.90	3231.78

* random guess: $-2 \log(L) = -2 * \log(0.5^N) = 10144.90$

* fixed: LR without using any variable.

$$-2 \log(L) = -2 * \log[p_0^{N_0} p_1^{N_1}] = -2 * \log[(6002/7318)^{6002} (1316/7318)^{1316}] = 6895.55$$

two risk factors

model/var	-2log(L)	K	AIC	Δ AIC	BIC	Δ BIC
$x_5 + x_{14} + \text{int}$	6755.01	4	6763.01	-134.54	6790.60	-113.85
$x_5 + x_{14}$	6761.91	3	6767.91	-129.64	6788.60	-115.85
$x_1 * x_{14}$	6800.87	2	6804.87	-92.68	6818.67	-85.78
fixed	6895.55	1	6897.55	0	6904.45	0

* x_{14} is whether the child has elder sibling or no

* x_5 is the age of the mother

* adding the product interaction may not improve AIC/BIC.

many risk factors

model/var	-2log(L)	K	AIC	Δ AIC	BIC	Δ BIC
core17	6513.82	18	6549.82	-347.73	6673.99	-230.46
core10	6547.44	11	6569.44	-328.11	6645.32	-259.13
core9	6565.22	10	6585.22	-312.33	6654.20	-250.25
full	6510.15	25	6560.15	-337.40	6732.60	-171.85
full+int	6165.22	289	6743.22	-154.34	8736.76	1832.31
fixed	6895.55	1	6897.55	0	6904.45	0

* full: LR using all 24 factors

* full+int: LR using all factors plus all product terms.

* core17: risk factors by stepwise variable selection via AIC (1,2,3,4,5,6,8,9,11,14,16,17,18,19,20,22,23)

* core10: risk factors by stepwise variable selection via BIC (1,2,4,5,6,9,11,14,19,23)

* core9: risk factors 1,2,4,5,6,9,11,14,23 (remove 19 from core10)

* classification is still bad. for core10, there are 1311 classification errors even within the sample (as compared to the 1316 errors for the fixed probability rate)

neural networks

model/var	-2log(L)	K	AIC	Δ AIC	BIC	Δ BIC
NN17	5060.83	443	5946.83	-950.72	9002.68	2098.23
NN10	5721.40	261	6243.40	-654.15	8043.80	1139.35
NN4	6244.52	105	6454.52	-443.03	7178.82	274.37
NN1	6511.93	27	6565.93	-331.62	6752.18	-152.27
fixed	6895.55	1	6897.55	0	6904.45	0

* NN1,4,10...: neural network with one, four, ten, ... hidden units

observations

this data set: whether the child has elder sibling or not is important (need further clarification)... the prediction rate is low (knowing the value of these risk factors couldn't predict the affection status)... a few (e.g. 10) risk factors one may focus attention on...

the method: the debate is not between using one risk factor vs using all factors, but using an appropriate number of risk factors... other models can also be used (beyond logistic regression)...neural network is selected by AIC, but soundly rejected by BIC...

example 2: selecting markers in linkage analysis

$$y = f(x_1, x_2, x_3, \dots, x_p)$$

y : sibpair affection status (0,1 for AA,AU)

$\{x_i\}$: mean IBD at each marker.

model: logistic regression

how to use AIC/BIC: select a subset of markers that minimize AIC/BIC.

Data: Asthma

German data set from *Genetic Analysis Workshop 12* (GAW12), as part of the *Collaborative Study on the Genetics of Asthma* (CSGA)

Total $p=333$ genetic markers are typed on 22 autosomal chromosomes.

$N=152$ sibpairs (obtained from 415 individuals in 97 families)

$N_0=109$ are affected-affected pairs (72%) and $N_1=43$ are affected-unaffected pairs (28%).

single marker

model	K	$-2\log(\hat{L})$	AIC	Δ AIC	BIC	Δ BIC	p_{within}
m_{121}	2	169.90	173.90	-9.18	179.95	-6.16	109/152
m_7	2	172.94	176.94	-6.14	182.99	-3.12	109/152
m_{293}	2	181.08	185.08	2	191.13	5.02	109/152
fixed	1	181.08	183.08	0	186.11	0	109/152
random	0	210.72	210.72	27.64	210.72	24.61	76/152

* the result can be compared to other traditional linkage analysis (ALLEGRO, GENEHUNTER programs)

* classification rate, even within the sample, is bad.

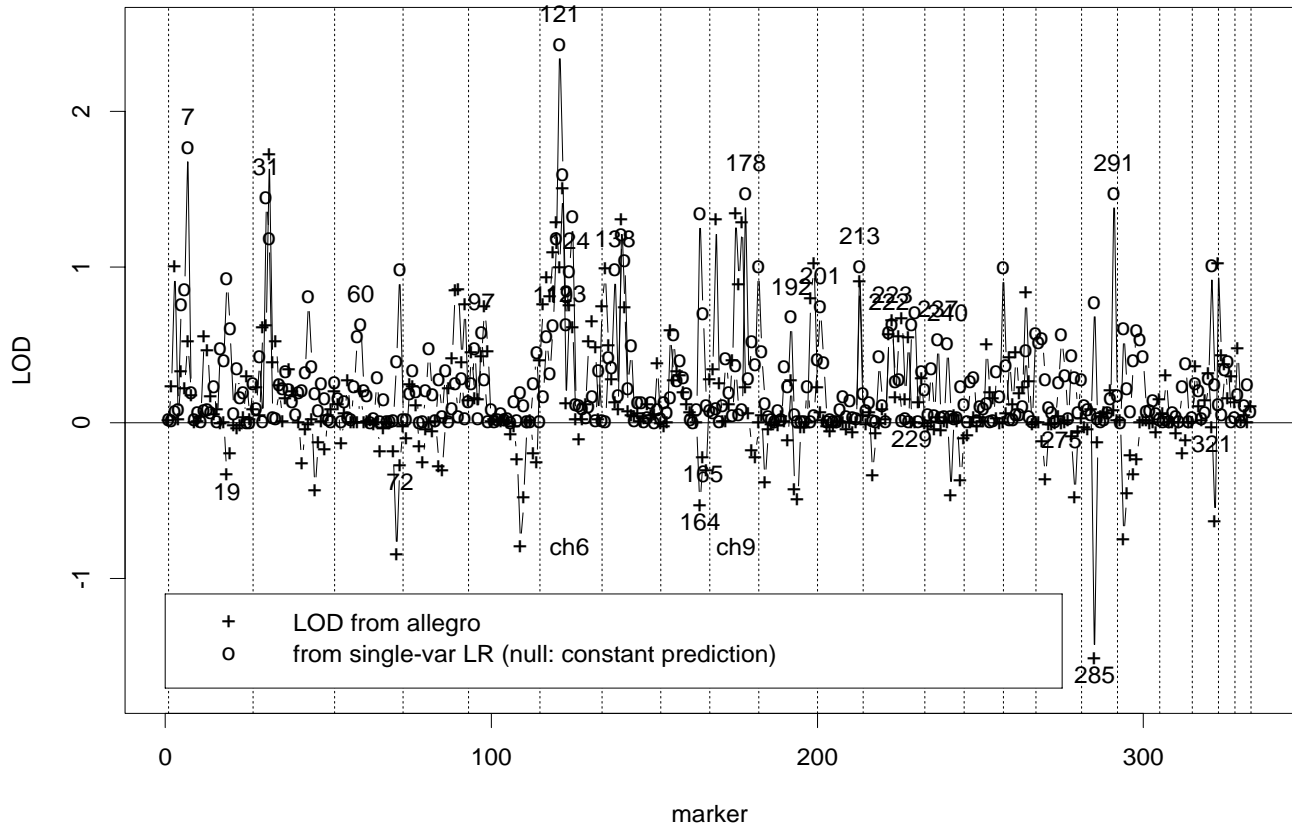
* marker 121 is D6S276 located on chromosome 6

* marker 7 is D1S197 located on chromosome 1

two markers

model	K	$-2\log(\hat{L})$	AIC	Δ AIC	BIC	Δ BIC	p_{within}
121+7	3	161.48	167.48	-15.60	176.56	-9.55	112/152
121+178	3	162.18	168.18	-14.90	177.26	-8.85	109/152
121 · 7	2	164.64	168.64	-14.44	174.69	-11.42	109/152
121+7+int	4	161.46	169.46	-13.62	181.55	-4.56	112/152
fixed	1	181.08	183.08	0	186.11	0	109/152

T50.AIC.26



LOD from ALLEGRO: $\log_{10} L(\hat{Z}|y)/L(Z_0)$

where $L(Z) = \sum_{i=0}^2 z_i P(y|IBD = i)$, $Z_0 = (1/4, 1/2, 1/4)$, and the sign is positive when the average IBD is larger than 1 in AA pairs, and negative otherwise.

LOD from logistic regression (discrimination): $\log_{10} \hat{L}/L_0$

where L_0 is the fixed prediction model, or $(AIC_0 - AIC_1 + K - 1)/2$.

more markers

model	K	$-2\log(\hat{L})$	AIC	Δ AIC	BIC	Δ BIC	ρ_{within}
PA100.AIC.25	26	0.68	52.68	-130.4	131.30	-54.8	152/152
T50.AIC.26	27	0.99	54.99	-128.1	136.64	-49.5	152/152
T50.BIC.25	26	3.91	55.91	-127.2	134.53	-51.6	152/152
PA50.AIC.46	47	6.53	100.5	-82.5	242.65	56.5	152/152
T9	10	134.70	154.7	-28.4	184.94	-1.2	118/152
T25	26	104.62	156.6	-26.5	235.24	49.1	129/152
fixed	1	181.08	183.1	0	186.11	0	109/152

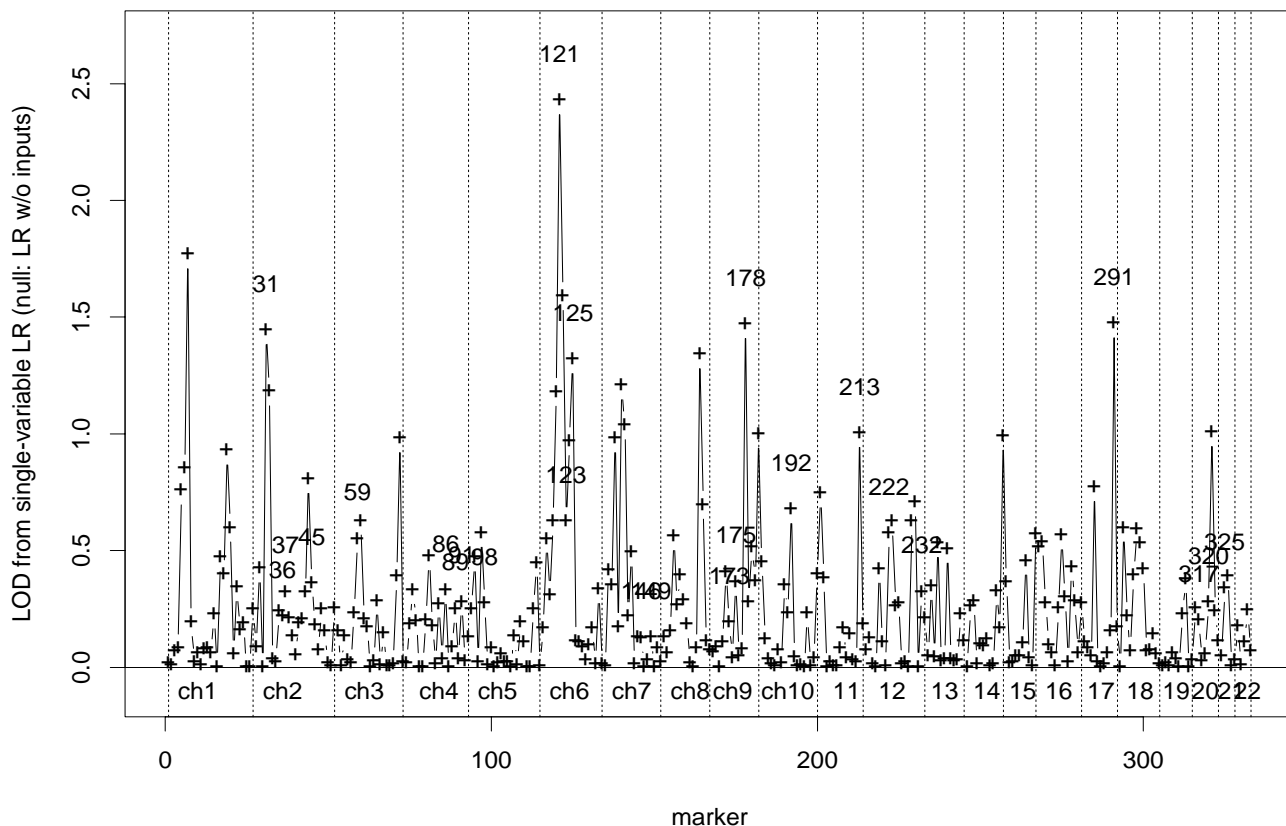
* T9: using top 9 markers with the highest single-variable LR likelihood

* T50.AIC.26: stepwise variable selection (via AIC) starting from the top 50 markers, ending at 26 markers

* PA100.AIC.25: stepwise variable selection (via AIC) starting from the top 100 markers where higher IBD in affected-affected pair

* perfect within-the-sample classification is possible, with at little as 25 markers.

PA100.AIC.25



PA100.AIC.25 is a set of 25 markers that are selected stepwisely from the top 100 markers with the best single-variable LR likelihoods plus a positive ALLEGRO LOD score (max iteration=10): 31(D2S2374), 36(D2S2216), 37(D2S160), 45(D2S116), 59(D3S1300), 86(D4S393), 89(D4S1535), 91(D4S426), 98(D5S418), 121(D6S276), 123(D6S426), 125(D6S455), 146(D7S2446), 149 (D7S684), 173(D9S175), 175(D9S283), 178(D9S195), 192(D10S537), 213 (D11S968), 222(D12S355), 232(D12S97), 291(D17S928), 317(D20S186), 320 (D20S891), and 325(D21S263).

observations

this data set: chromosome 6 contains strong linkage signal... the selected 20 or so markers by AIC/BIC are not identical to those with the best performing single-marker logistic regression... average IBD implies additive affect of two parents...

the method: the number of variable (333) is larger than the number of sample (152) so a two-step variable selection is needed... prediction rate becomes perfect (overfitting?)

example 3: discriminant microarray data analysis

$$y = f(x_1, x_2, x_3, \dots, x_p)$$

y: cancer type

$\{x_i\}$: (log) mRNA expression level of each gene.

model: logistic regression

how to use AIC/BIC: select a subset of genes that are best for discriminating two types of cancers

Data: Leukemia

from MIT's group (Golub, et al. *Science*, 286:531-537 (1999))

samples were derived from bone marrow

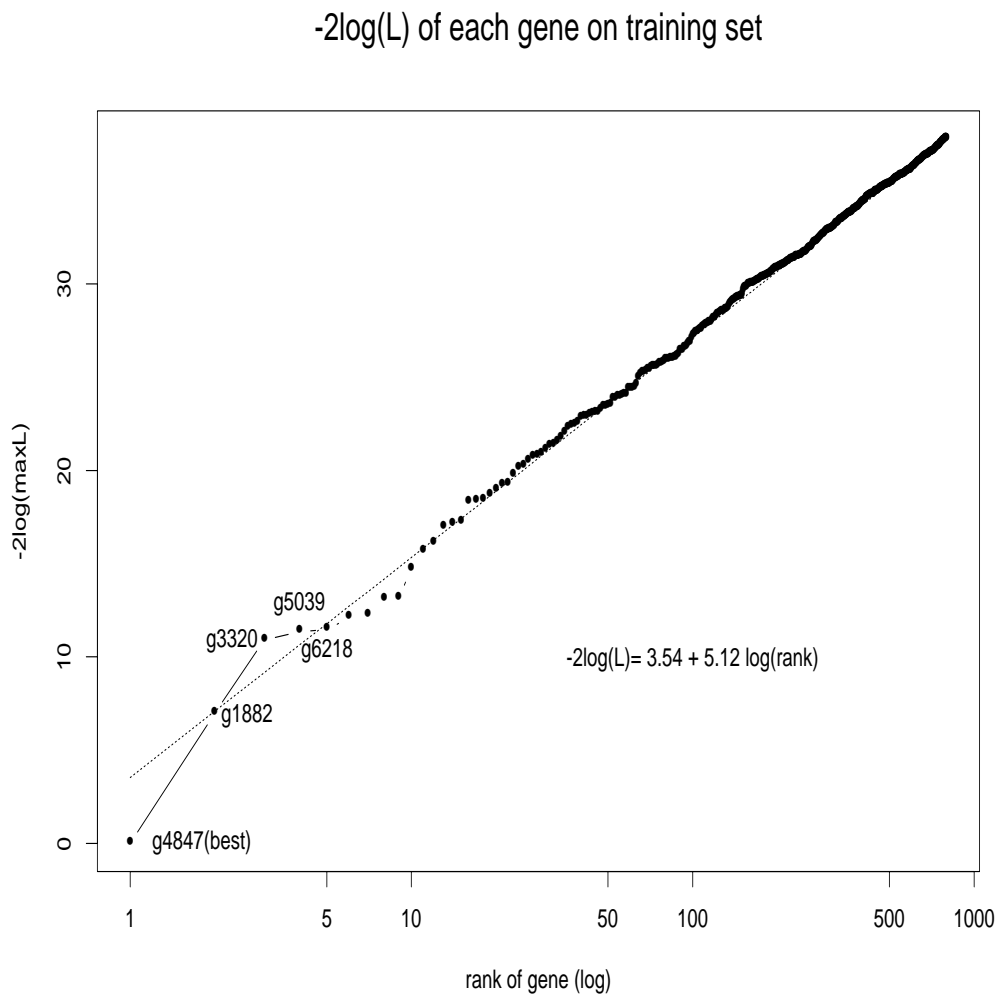
two types of leukemia (acute myeloid leukemia (AML, $y=1$), acute lymphoblastic leukemia (ALL, $y=0$).

originally $N=38$ samples for training, with extra 34 samples for testing (not necessary from bone marrow)

number of genes is $p=7129$

Since the number of variables (7129) is far larger than the number of samples (38), we are forced to use less number of variables in a model.

Single-variable logistic regression is carried out for all genes. They are ranked ($-2\log(L)$ vs rank):



No separation between the relevant and irrelevant genes.

single-gene

type	K	$-2\log(\hat{L})$	AIC	Δ AIC
#1 g4847 (zyxin)	2	≈ 0	4.000	0
#2 g1882 (CST3 cystatin C)	2	6.973	10.973	6.973
#3 g3320 (leukotriene c4 synthase)	2	10.914	14.914	10.914
#4 g5039 (LEPR leptin receptor)	2	11.355	15.355	11.355
#5 g6218 (ELA2 elastatse 2)	2	11.459	15.459	11.459
#6 g2020 (FAH ..)	2	12.103	16.103	12.103
#7 g1834 (CD33 antigen)	2	12.226	16.226	12.226
#8 g760 (cystatin A)	2	13.104	17.104	13.104
#9 g1745 (LYN v-yes-1..)	2	13.151	17.151	13.15
#10 g5772 (c-myb)	2	14.723	18.723	14.723
#100 g2833(AF1q)	2	27.215	31.215	27.21
#200 g3312(protein kinase ATR)	2	30.841	34.841	30.841
fixed	1	45.728	47.728	43.728

type	BIC	Δ BIC	p_{train}	p_{test}
#1 g4847 (zyxin)	7.275	0	38/38	31/34
#2 g1882 (CST3 cystatin C)	14.248	6.973	36/38	32/34
#3 g3320 (leukotriene c4 synthase)	18.190	10.915	35/38	27/34
#4 g5039 (LEPR leptin receptor)	18.630	11.355	36/38	22/34
#5 g6218 (ELA2 elastatse 2)	18.734	11.459	34/38	22/34
#6 g2020 (FAH ..)	19.378	12.103	36/38	25/34
#7 g1834 (CD33 antigen)	19.501	12.226	35/38	31/34
#8 g760 (cystatin A)	20.379	13.104	35/38	32/34
#9 g1745 (LYN v-yes-1..)	20.426	13.151	33/38	28/34
#10 g5772 (c-myb)	21.998	14.723	35/38	27/34
#100 g2833(AF1q)	34.490	27.215	30/38	28/34
#200 g3312(protein kinase ATR)	38.117	30.842	29/38	21/34
fixed	49.365	42.090	27/38	20/34

best gene: g4847 (zyxin) (“a component of adhesion plaques that has been suggested to perform regulatory functions at these specialized regions of the plasma membrane”)

many genes

type	K	$-2\log(\hat{L})$	AIC	Δ AIC
g1834+g2267	3	0.004	6.004	2.004
g5039+g5772	3	0.008	6.008	2.008
sum of top 2 (g4847+g1882)	3	0.029	6.029	2.029
sum of top 5	6	0.011	12.011	8.011
sum of top 10	11	0.002	22.002	18.002
sum of top 22	23	0.001	46.001	42.001
sum of top 37	38	0.001	76.001	72.001
fixed probability	1	45.728	47.728	43.728
random guess	0	52.679	52.679	48.679

type	BIC	Δ BIC	p_{train}	p_{test}
g1834+g2267	10.917	3.642	38/38	22/34
g5039+g5772	10.921	3.646	38/38	26/34
sum of top 2 (g4847+g1882)	10.942	3.667	38/38	32/34
sum of top 5	21.837	14.562	38/38	24/34
sum of top 10	40.016	32.741	38/38	31/34
sum of top 22	83.666	76.391	38/38	27/34
sum of top 37	138.229	130.954	38/38	21/34
fixed probability	49.365	42.090	27/38	20/34
random guess	52.679	45.404	19/38	17/34

since single-gene logistic regression has already reached the perfect prediction, there is no room for improvement by making the model more complex!

observations

this data set: a single gene is enough... many genes are good for a single-gene model...

the method: stepwise variable selection fails to find the best solution because it is a local-optimization procedure... the problem with perfect fitting: the a_0 parameter cannot be determined uniquely, but a_0/a_1 seems to be stable...

Re-reading Golub et al's paper

They used 50 genes in the classifier (predictor), but the sample size is only 38. Is it possible???

Model Selection versus Model Averaging

Model selection picks one out of many alternative models. Models not selected are not used.

Model averaging keeps all models, each assigned a weight. But many included models may not contribute much because their weights are close to zero. (Most natural in a Bayesian framework.)

There is no restriction on the number of models included in a model averaging (can be larger than the sample size!).

In Golub et al.'s paper, 50 single-variable models are averaged in a classifier.

Can we determine the number of models in a model averaging? It is determined by the choice of weighting scheme.

Three Weighting Schemes

$$P(y = 1) = \sum_j w_j \left(\frac{1}{1 + e^{-a_0 - a_j x_j}} \right)$$

(1) $w_j \propto \hat{L}$: for this example, it is identical to the use of Akaike weight $w_j \propto \exp(-AIC_j/2)$ and the Bayesian weight $w_j \propto \exp(-BIC_j/2)$, because models being averaged are all single-gene classifiers with the same number of parameters.

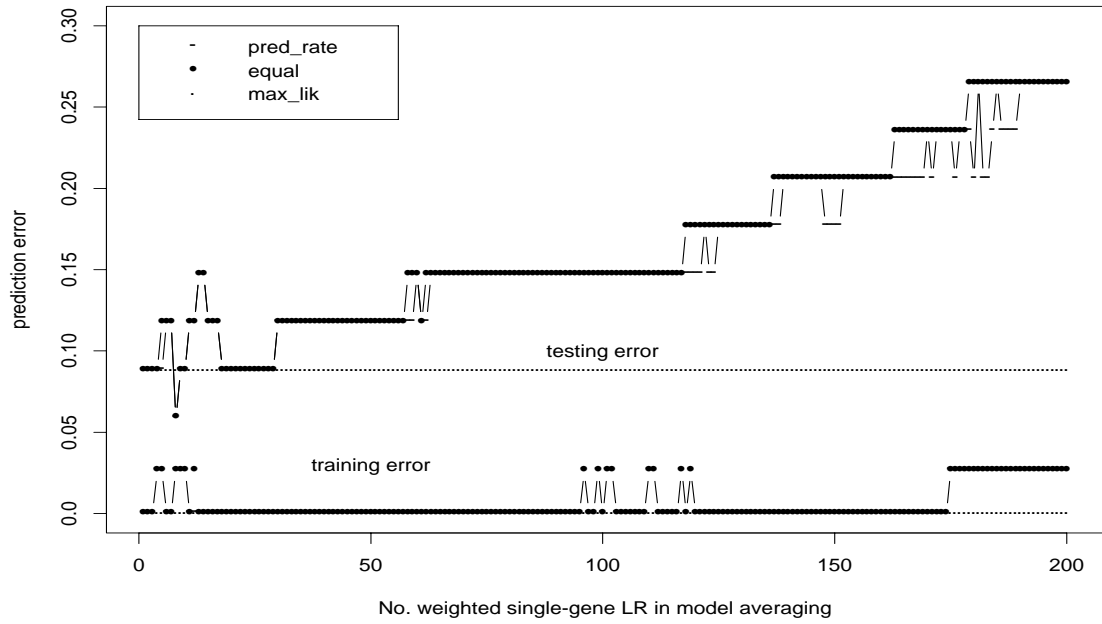
(2) $w_j \propto p_{train}$: p_{train} contains information only on classification rate, not on classification confidence.

(3) $w_j \propto 1$: equal weight

mean square error



prediction error



example 4: segmenting DNA sequences

DNA sequences are not homogeneous. Partitioning (segmenting) a sequence into two subsequences is a model selection process which compares: (1) modelling the DNA as a random sequence with 3 free parameters (p_A, p_C, p_G , with $p_T = 1 - p_A - p_C - p_G$); (2) modelling the DNA as two random subsequences with 7 free parameters (3 base compositions for each side, and the partitioning point).

This 1-to-2 partition can be carried out recursively, until no further partition is necessary.

how to use AIC/BIC: determine the stopping criterion for the recursive partitioning.

Likelihoods before and after a 1-to-2 partition:

$$L_1(\{p_\alpha\}) = \prod_\alpha p_\alpha^{N_\alpha}$$

$$L_2(\{p_{\alpha,l}\}, \{p_{\alpha,r}\}, i) = \prod_\alpha p_{\alpha,l}^{N_{\alpha,l}} \prod_\alpha p_{\alpha,r}^{N_{\alpha,r}}$$

$$AIC_1 = -2N \sum_\alpha \hat{p}_\alpha \log \hat{p}_\alpha + 6$$

$$AIC_2 = -2N_l \sum_\alpha \hat{p}_{\alpha,l} \log \hat{p}_{\alpha,l} - 2N_r \sum_\alpha \hat{p}_{\alpha,r} \log \hat{p}_{\alpha,r} + 14$$

Requiring $AIC_2 < AIC_1$, we have

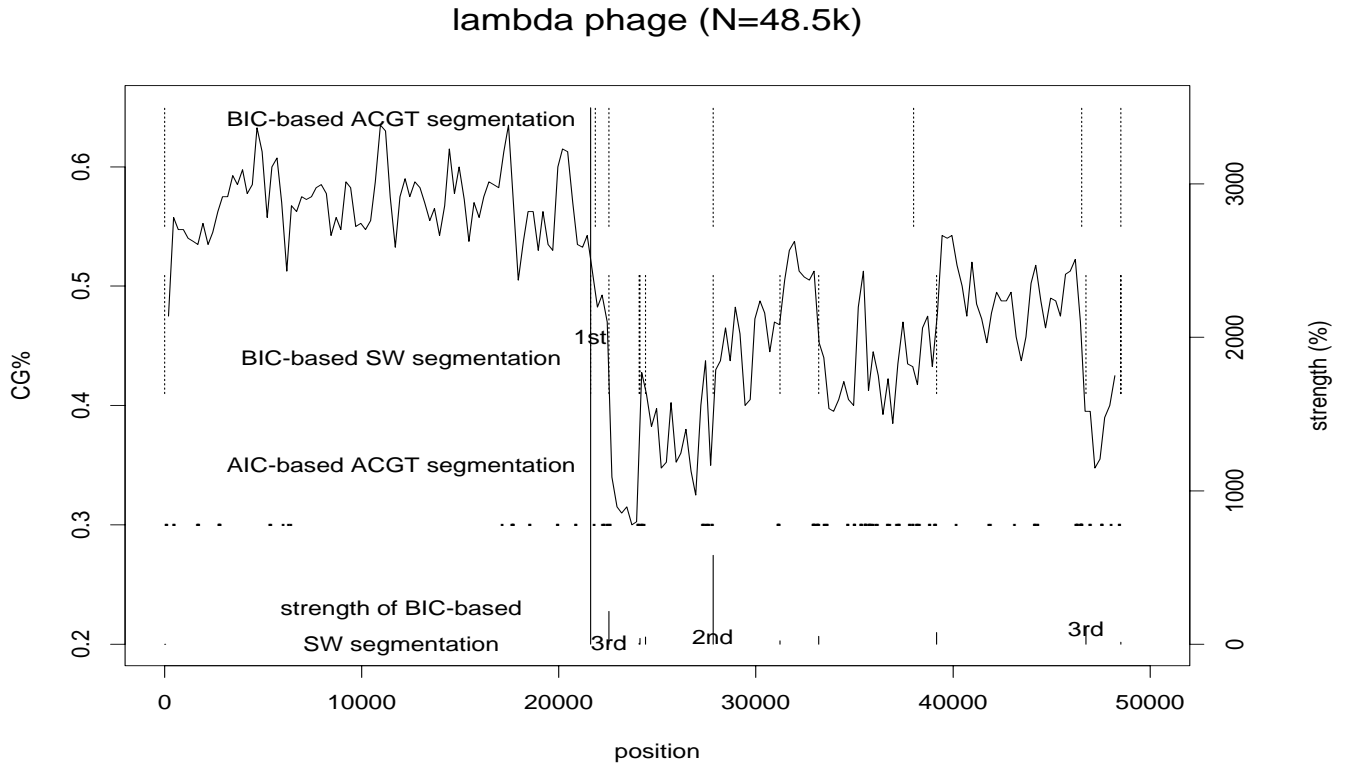
$2N\hat{D}_{JS} > 8$ where \hat{D}_{JS} is the Jensen-Shanon divergence (a distance measure based on entropy).

Similarly, Requiring $BIC_2 < BIC_1$, we have

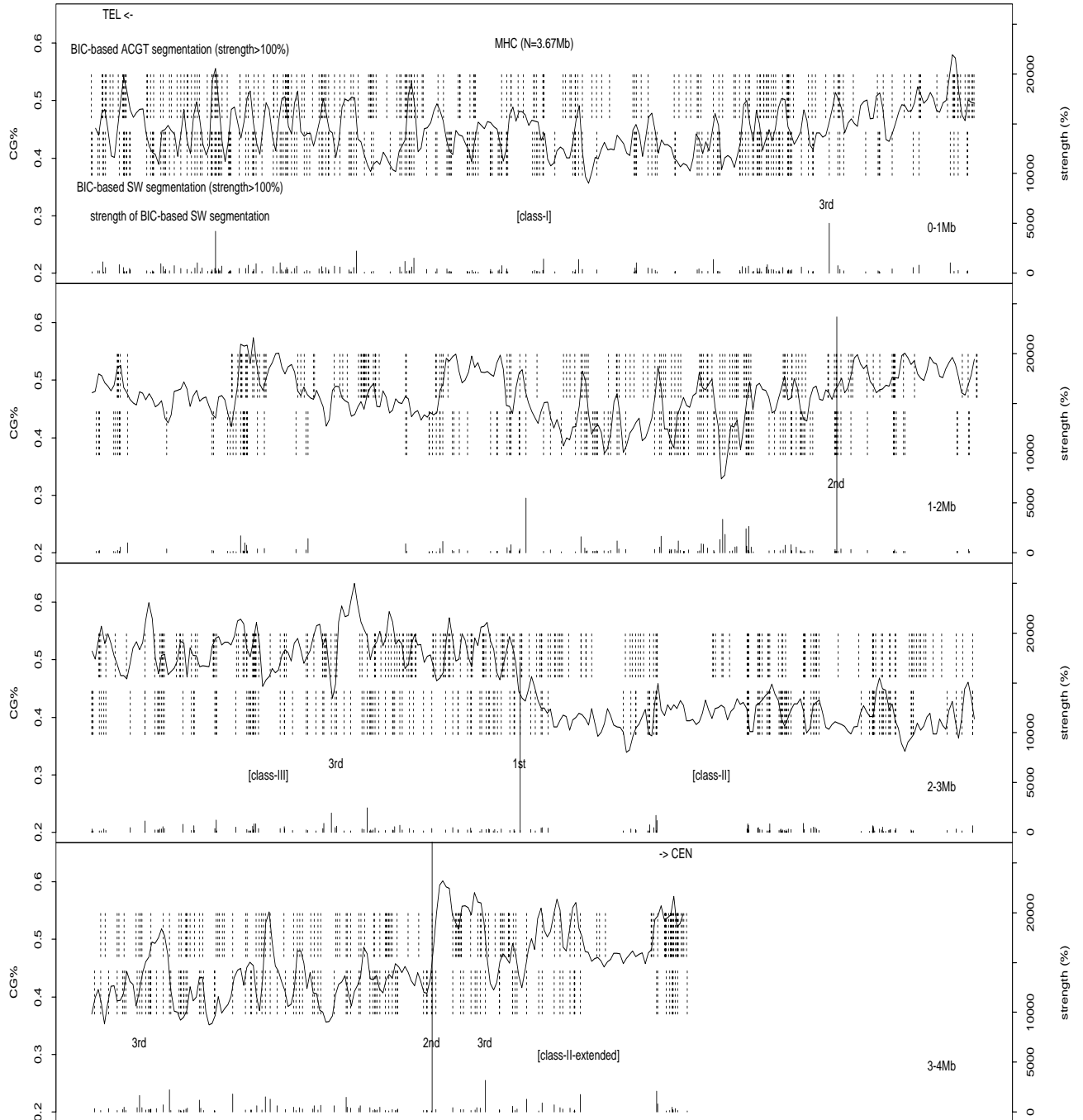
$$2N\hat{D}_{JS} > 4 \log(N)$$

The BIC-based stopping criterion is different from previous approaches because it depends on sequence length N . On the other hand, the AIC-based stopping criterion is similar to the previous approach with a fixed significance level.

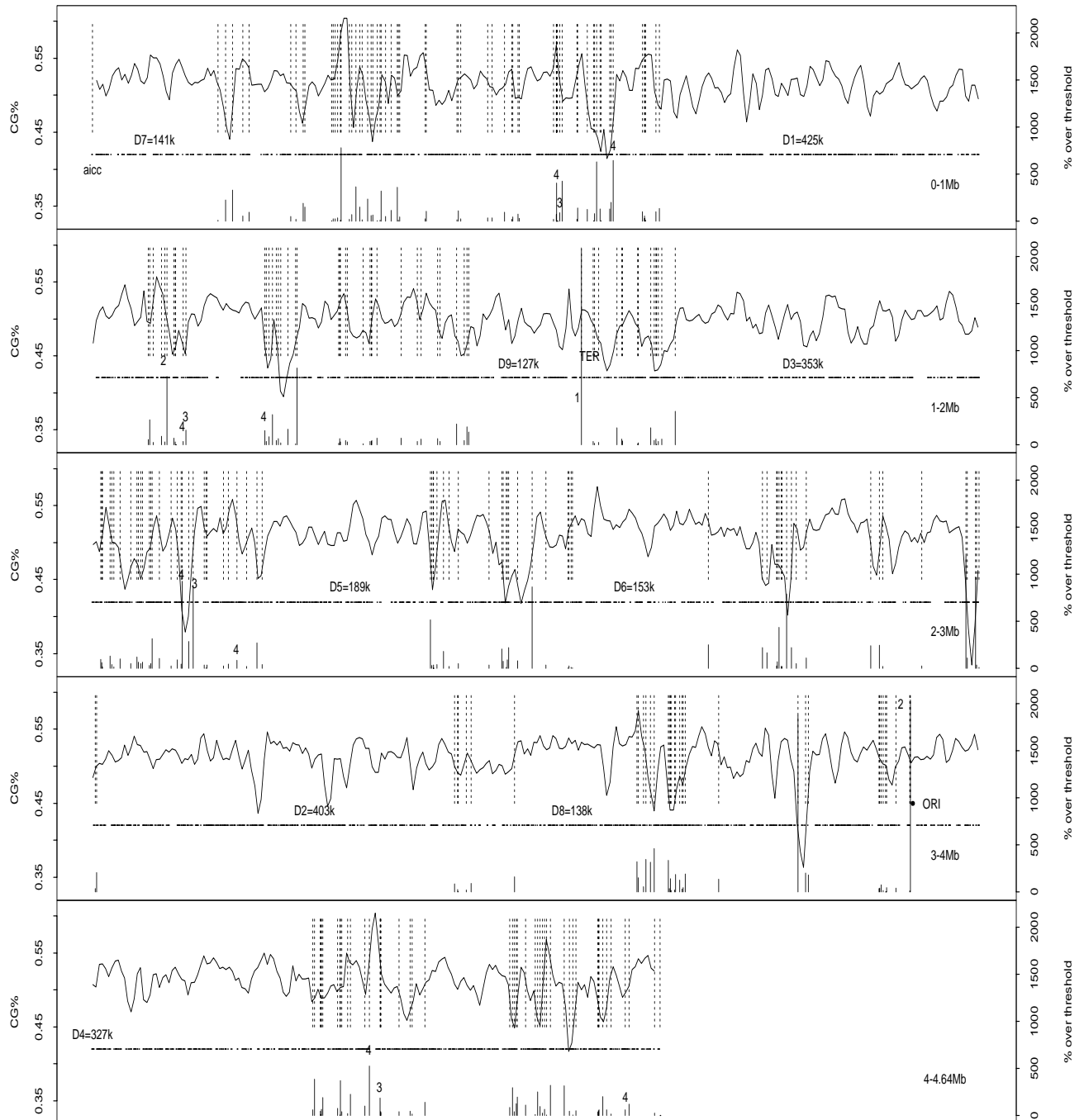
lambda bacteriophage: 2 domains can be easily seen



human major histocompatibility complex (MHC): 4 isochores are identified



e.coli complete sequence: replication origin/terminus are identified



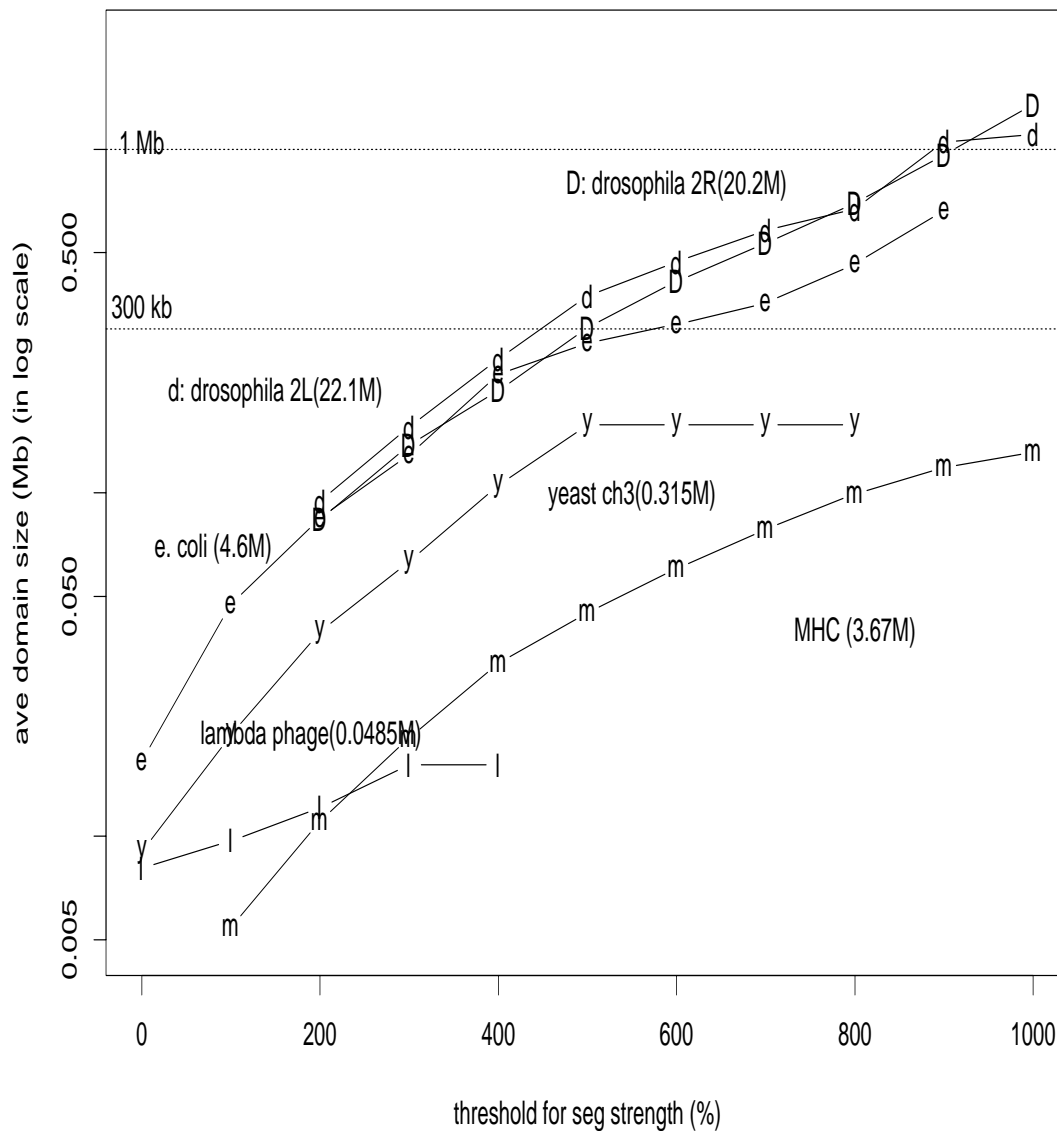
Sometimes, the BIC stopping criterion is still not stringent enough. We further increase the difficulty in segmentation by requiring that “strength”, defined as

$$strength = \frac{2ND_{JS} - 4 \log(N)}{4 \log(N)}$$

(where D_{JS} is the Jensen-Shannon entropy), should be larger than some value (for BIC-based stopping criterion, the strength should be larger than 0).

when a more stringent condition is applied, the resulting domain sizes are larger. here is how average domain size changes with the segmentation stringency:

average domain size vs threshold for segmentation strength



Conclusion

We are facing more complicated data sets in the study of complex diseases and genomics. Tools that are more appropriate to the problem, more flexible to use, providing a better description, should be adopted. Model/variable selection by AIC and BIC is one of these tools.

Acknowledgments

Andrea Sherriff: child asthma data

Yaning Yang: s-plus

Dale Nyholt: linkage analysis

many others for discussions

general reference

Burnham, Anderson, *Model Selection and Inference* (Springer, 1998).

Parzen, Tanabe, Kitagawa, *Selected Papers of Hirotugu Akaike* (Springer, 1998).

epidemiology

W Li, A Sherriff, X Liu (2000), "Assessing risk factors of complex diseases by Akaike and Bayesian information criterion" (abstract), *American Journal of Human Genetics* (suppl), s67:222.

linkage

W Li, D Nyholt (2000), "Marker selection by Akaike and Bayesian information criteria", *Proceedings of Genetic Analysis Workshop 12*, vol 1:194-198.

microarray

W Li, Y Yang (2000), "How many genes are needed for a discriminant microarray data analysis?" *Assessment of Techniques for Microarray Data Mining Workshop' 2000 (CAMDA'00)*, accepted.

DNA segmentation

W Li (2001), "DNA segmentation as a model selection process", *Proceedings of Research in Computational Molecular Biology Workshop' 2001 (RECOMB'01)*, accepted.

W Li, "New stopping criteria for segmenting DNA sequences", submitted.