

# **Reducing Complexity in Biological and Genetic Data**

*Wentian Li*

Lab of Statistical Genetics  
Rockefeller University

November 30, 2000

Thursday 4-5pm

# Presentation Outline

## Background and Motivation

**model selection** using Akaike and Bayesian information criterion (AIC/BIC)

## Applications

**epidemiology analysis:** assessing risk factors

**linkage analysis:** selecting markers

**microarray data analysis:** distinguishing cancer types via gene expression profiles

**segmenting DNA sequences:** finding borders of homogeneous domains

## Future Works

# Background and Motivations

In post-genomic era, we will have large amount of information including (human genome alone):

DNA sequences ( $3 \times 10^9$  bases),

gene annotations (20,000-60,000)

gene products/proteins

gene expressions

gene functions

population polymorphism/variations

phenotypes (12,000 in OMIM)

# Proposed Solution for Data Utilization

*since we are only interested in a specific biological problem or a specific human disease, we do not need all information.*

*use **less** information, while keeping the information as **relevant** as possible.*

*In this way, we will **not** be overwhelmed by the vast amount of information in post-genomic era.*

Question:

**How To Do This?**

Proposed Answer:

**Model/Variable  
Selection**

## Model/Variable Selection

The purpose of a model selection is to find a model, with the appropriate chosen parameter values, that best fits the available data, *after applying a correction or a penalty for the model complexity.*

In contrast, the traditional “maximum likelihood” framework is only concerned with fitting the data *without* regard to model complexity.

## Why consider the model complexity?

Models with more [free, adjustable] parameters will always fit a given data set better than [no worse than] the model with less parameters [with one parameter removed]. So it is fundamentally unfair to compare models with different number of parameters by its goodness-of-fit alone.

A model with enough number of parameters will fit any given data perfectly: “saturation model”, “overfitting model”. But such good fit is misleading, because it also fits noise in the data set [or, mainly fits the noise].

## Balancing the two factors:

1. *How good does the model fit the data ?*

example: likelihood (probability of observing the data given a model with some free parameters)

2. *How complex is the model ?*

example: number of free parameters in the model

## Balancing the two factors by Akaike Information Criterion (AIC)

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

$L(\theta)$  = likelihood ( $\theta$  is a collection of free parameters)

$\hat{L} = L(\hat{\theta})$  = maximized likelihood

$K$  = number of free parameters in the model

The smaller the AIC, the better the model.

## Balancing the two factors by Bayesian Information Criterion (BIC)

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

$N$  = sample size.

The smaller the BIC, the better the model.

## Where does AIC come from?

A model  $g$  is judged by how close it is to the “true model”  $f$  (defined on sample space  $x$ ). The closeness is measured by the Kullback-Liebler distance:

$$I(f, g) = \int f(x) \log \frac{f(x)}{g(x|\theta)} dx$$

We average it over data  $y$  (the maximized parameter value depends on the data):

$$\hat{I}(f, g) = \int f(y) \int f(x) \log \frac{f(x)}{g(x|\hat{\theta}(y))} dx dy$$

AIC is an approximation of  $2\hat{I}$  up to a constant term.

Smaller AIC also implies a smaller distance between the model and the “truth”.

# Where does BIC come from?

In Bayesian framework, a model can also be assigned a probability:  $P(M)$ . Posterior probability of a model after observing the data  $D$  is  $P(M|D)$ . When two models are compared, we select the model with the larger posterior probability:

$$\frac{P(M_2|D)}{P(M_1|D)} = \frac{P(D|M_2)}{P(D|M_1)} \times \frac{P(M_2)}{P(M_1)}$$

for the middle term, an integral over the parameter (model is fixed) can be introduced (called “integrated likelihood”):

$$P(D|M_i) = \int P(D|\theta_i, M_i)P(\theta_i|M_i)d\theta_i$$

BIC is an approximation of the  $-2 \log P(D|M)$ . When the prior probability of all models is the same, the smaller the BIC, the larger the posterior probability.

A quote from Akaike in his interview in 1995:

“Many had such a dogmatic attitude that they did not even question the use of a maximum likelihood estimate when nothing was known about the ‘true’ form of the distribution, yet they criticized the use of a minimum AIC estimate. People with serious statistical problems, like applied engineers, could easily appreciate the contribution of AIC simply by getting useful answers to problems which could not be handled by a conventional statistical approach”. (*Statistical Science*, 10:104-117 (1995) )

## Properties of AIC and BIC

1. Both AIC and BIC are *approximations* of reasonable theoretical quantities ( [2] expected Kullback-Liebler distance, and  $[-2\log]$  integrated likelihood) that evaluate a model, and high-order terms (with respect to sample size  $N$ ) are ignored in the above definition
2. BIC is dimensional-consistent whereas AIC is not, meaning the AIC- selected model tends to become more complicated as the sample size is increased, whereas BIC- selected model does not.

### References:

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

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

**Variable selection is a  
special form of model  
selection**

## why?

typically, variables are combined in a model (e.g. linear combination  $a_1x_1 + a_2x_2 + \dots$ ), and for that purpose, a coefficient is assigned to a variable. removing that variable also removes the coefficient, thus reducing the number of parameters, and, model complexity.

in this context, models with more variables are more complex than those with less number of variables.

# **general procedures for reducing the complexity**

- use a reasonable type of models
- start with the saturate situation with a lot of variables (overfitting). maximize the likelihood.
- gradually remove variables one at the time, and re-maximize the likelihood to see whether the loss in likelihood is small enough to be compensated by the gain in model simplicity (i.e. whether AIC or BIC is reduced).
- stop when the smallest number of variable are attained with respect to the smallest AIC or BIC.
- may repeat for another type of models

# Applications

## I. Epidemiology analysis

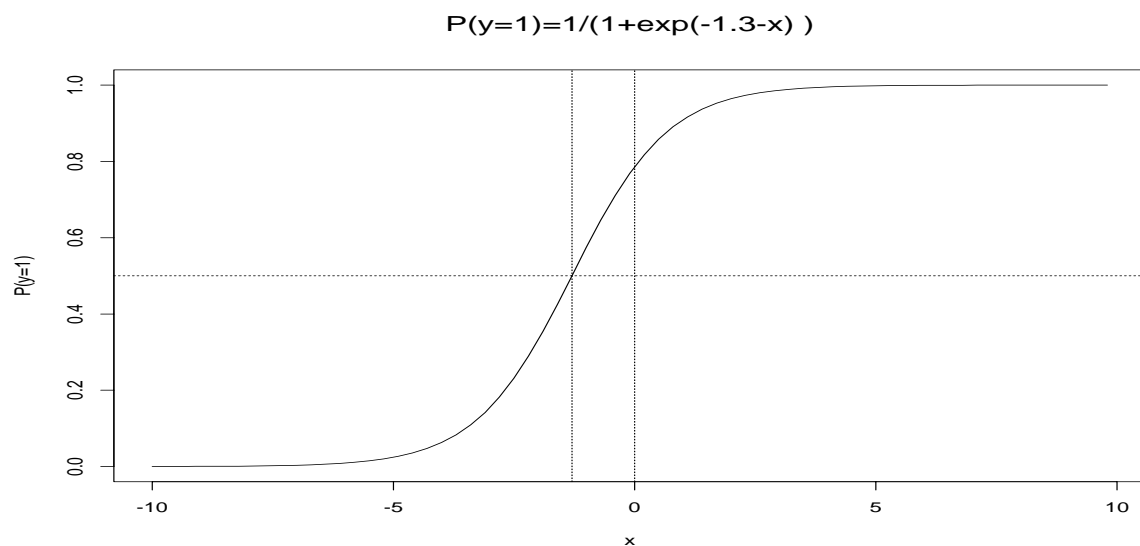
Risk is the probability that an individual becomes newly diseased given certain attributes – these attributes are risk factors.

Traditionally, risk factors are assessed one by one. Then model selection is not necessary. However, if we consider a joint action by several risk factors, complicated models will be compared with simpler models.

# A Standard Framework: Logistic Regression

y: output (0 and 1 only), x's: inputs

$$P(y = 1 | \{x_i\}, \{a_i\}) = \frac{1}{1 + e^{-a_0 - \sum_i a_i x_i}}$$



linear combination of variables (hyperplane)

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

$a_0$  determines the threshold value,  $\{a_i\}$  determines the “soft margin” region (undecided)

a reasonable middle ground between traditional statistics and machine learning.

similar names: prediction, discrimination, decision, classification

# Accessing Risk Factors with Logistic Regression (LR) Framework

- one-risk-factor LR
- two-risk-factor LR
- as many risk factors as possible (saturation model)
- stepwise variable selection from the saturation model
  
- pick the model with the smallest AIC or BIC
- allowing a range of small AIC/BIC (model uncertainty)

## Data: Child Asthma

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. 24 factors are used in this analysis.

$y = 1$  for infant wheeze (affected). 1316 children are affected (18%).

## 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

## Child Asthma: Single Risk Factor

$$N=7318 \text{ (}\log(N)=8.898\text{)}$$

Logistic regression (LR)

model/var	-2log(L)	K	AIC	$\Delta$ AIC	BIC	$\Delta$ BIC
LR: $x_{14}$	6810.95	2	6814.95	-82.60	6828.75	-75.70
LR: $x_6$	6849.08	2	6853.08	-44.47	6866.88	-37.57
LR: $x_1$	6855.49	2	6859.49	-38.06	6873.29	-31.16
LR: $x_7$	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.

$$-2 \log(L) = -2 * \log(p_0^{N_0} p_1^{N_1}) = 6895.55$$

-  $x_{14}$  is whether the child has elder sibling or not

-  $x_6$  is whether breast fed or not

-  $x_1$  is gender

these results should be comparable and similar to traditional approaches

the worst single-factor LR has similar likelihood as the fixed probability model, but its AIC/BIC are worse.

## Child Asthma: Two Risk Factors

Logistic regression (LR)

model/var	-2log(L)	K	AIC	$\Delta$ AIC	BIC	$\Delta$ BIC
LR: $x_5 + x_{14}$ +int	6755.01	4	6763.01	-134.54	6790.60	-113.85
LR: $x_5 + x_{14}$	6761.91	3	6767.91	-129.64	6788.60	-115.85
LR: $x_1 * x_{14}$	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 not

- $x_5$  is the age of the mother

adding the product interaction may not improve AIC/BIC.

## Child Asthma: More Factors

Logistic regression (LR)

model/var	-2log(L)	K	AIC	$\Delta$ AIC	BIC	$\Delta$ BIC
LR: core17	6513.82	18	6549.82	-347.73	6673.99	-230.46
LR: core10	6547.44	11	6569.44	-328.11	6645.32	<b>-259.13</b>
LR: core9	6565.22	10	6585.22	-312.33	6654.20	-250.25
LR: full	6510.15	25	6560.15	-337.40	6732.60	-171.85
LR: 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)

## Child Asthma: Neural Networks

Neural networks (NN)

model/var	-2log(L)	K	AIC	$\Delta$ AIC	BIC	$\Delta$ BIC
NN17	5060.83	443	5946.83	<b>-950.72</b>	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

-clear difference between AIC- based and BIC- based model selection. all NNs here are bad models by BIC, but good by AIC.

# Applications

## II. Linkage Analysis

A genetic marker is linked to a disease (or its underlying disease gene) if the two co-segregate during the meiosis.

A simple test of this is to see whether two affected siblings share some chromosomal regions in common more often than randomly (identity-by-descent, sharing). Also, whether one affected and one unaffected sib do not share that region in common more often than randomly.

Usually, this test is carried out one region (one marker) at the time. No possible joint action is considered.

# Sibpair Analysis with Logistic Regression

Each sibpair is assigned a label  $y = 1$  for affected-affected and  $y = 0$  for affected-unaffected.

The identity-by-descent derived from both parents are average at each marker  $\{x_i\}$

Again, the classification of the two types of sibpairs is assumed to be carried out by a logistic regression over the mean identity-by-descent on all markers.

## Data: Asthma

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

Total 333 genetic markers are typed on 22 autosomal chromosomes.

415 individuals in 97 families that lead to 152 sibpairs

109 are affected-affected pairs (72%) and 43 are affected-unaffected pairs (28%).

## Linkage: Single Marker

logistic regression (N=152)

model	K	$-2\log(\hat{L})$	AIC	$\Delta$ AIC	BIC	$\Delta$ 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. it is a typical situation in linkage analysis of complex human diseases

-marker 121 is D6S276 located on chromosome 6

-marker 7 is D1S197 located on chromosome 1

## Linkage: Two Markers

logistic regression (N=152)

model	K	$-2\log(\hat{L})$	AIC	$\Delta$ AIC	BIC	$\Delta$ 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

-within the sample classification rate is still bad

-the evidence for product interaction is weak.

## Linkage: More Markers

logistic regression (N=152)

model	K	$-2\log(\hat{L})$	AIC	$\Delta$ AIC	BIC	$\Delta$ BIC	$\rho_{within}$
PA100.AIC.25	26	0.68	52.68	<b>-130.4</b>	131.30	<b>-54.8</b>	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

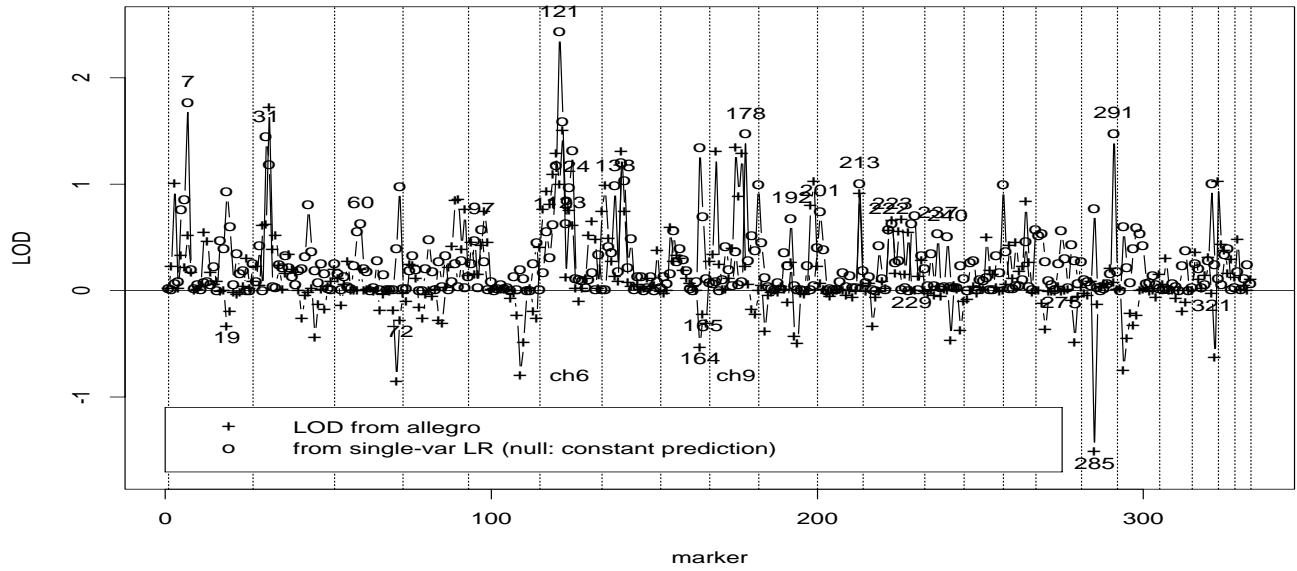
T50.AIC.26: stepwise variable selection (via AIC) starting from the top 50 markers, ending at 26 markers

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

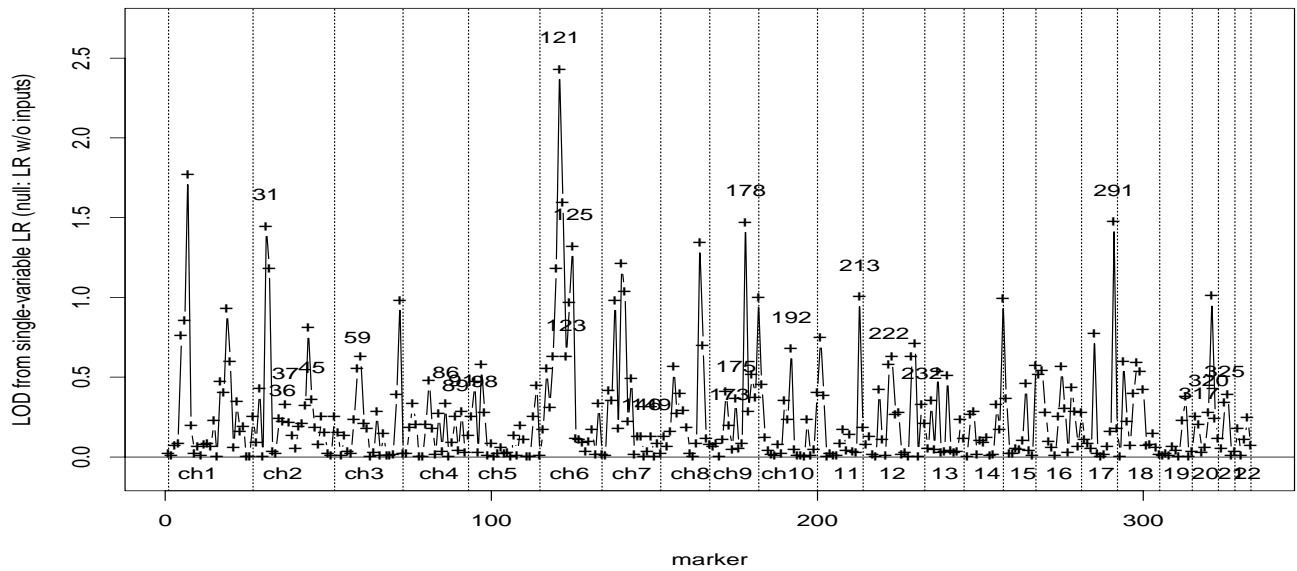
perfect within-the-sample classification is possible, with at little as 25 markers.

these markers are not identical to the top 25 markers according to single-variable LR.

T50.AIC.26



PA100.AIC.25



# Applications

## III. Microarray Data Analysis

Gene expression profile is used to distinguish diseased and normal tissues, different cancer types, etc. The number of genes in a profile can be thousands, whereas the human sample can be hard to obtained, so these are limited.

Can we use less number of genes in microarray data for a cancer classification? How many genes are really needed?

## 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 38 samples for training, with extra 34 samples for testing (not necessary from bone marrow)

number of genes is 7129

## Leukemia: Many Genes

logistic regression results (N=38, log(N)=3.637568)

type	K	$-2\log(\hat{L})$	AIC	$\Delta$ 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	$\Delta$ 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

almost always perfect classification on training set, even with only two genes.

single gene should also classify well.

## Leukemia: Single Genes

type	K	$-2\log(\hat{L})$	AIC	$\Delta$ AIC
#1 g4847 (zyxin)	2	$\approx 0$	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	$\Delta$ BIC	$p_{train}$	$p_{test}$
#1 g4847 (zyxin)	7.275	0	<b>38/38</b>	31/34
#2 g1882 (CST3 cystatin C)	14.248	6.973	36/38	<b>32/34</b>
#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	<b>32/34</b>
#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

indeed, single gene can classify the cancer class quite well

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

In the original Golub et al *Science* paper, 50 genes are used to classify cancer types. With only 38 samples, is it a case of overfitting?

Interestingly, the answer is *no*.

Golub et al was doing *model averaging*, not model selection. Each model involves only a single gene.

## How Many Models Should be Included in a Model Averaging?

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

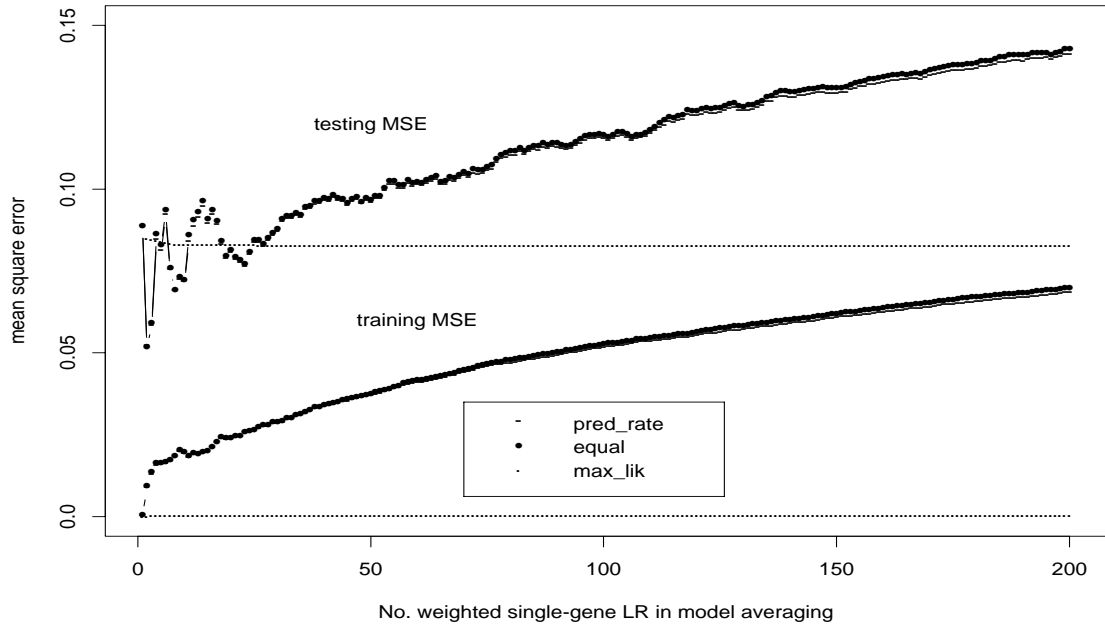
We have tried these weighting schemes:

(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 all models are single-gene classifiers with the same number of parameters.

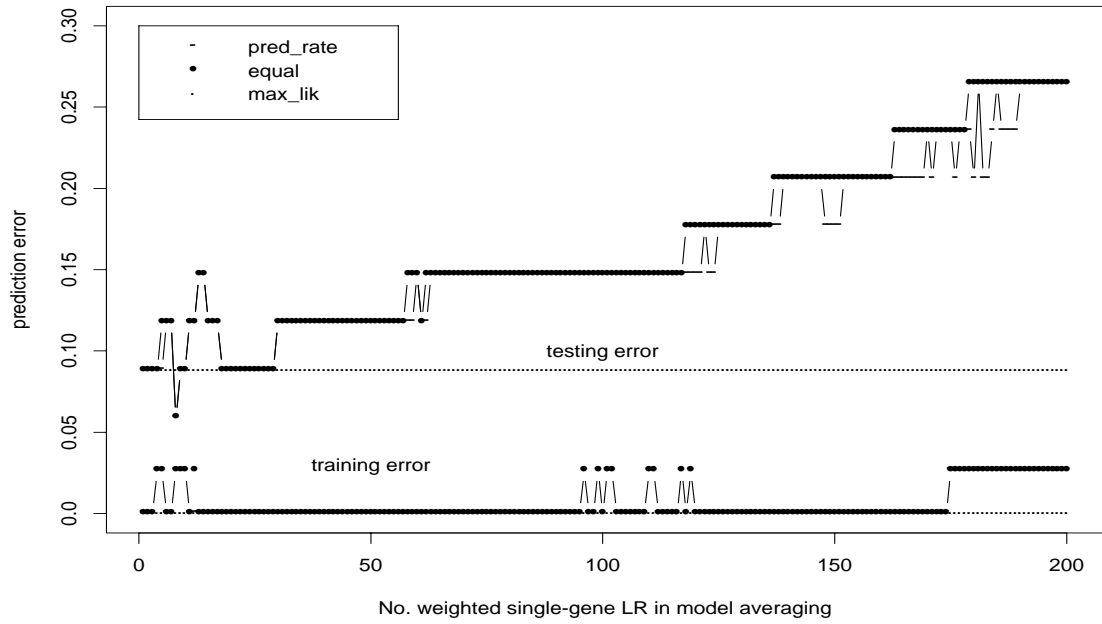
(2)  $w_j \propto p_{train}$ :  $p_{train}$  contains information only on classification rate, not on classification confidence.

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

### mean square error



### prediction error



## Observations on Model Averaging

With  $w_j \propto \hat{L}$  weight, the best model (using zyxin gene only) dominates all others, and the performance on both training and testing set does not change (slightly better with more genes).

With other weight, the prediction performance is getting worse as more models (single-gene predictors) are included.

In mean-square-error (MSE) vs number of models (genes) plot the performance on testing set fluctuates when the number of genes is smaller than  $\sim 25$ .

# Applications IV. Segmentation

## Analysis of DNA Sequences

DNA sequences are not homogeneous, there are all kinds of domain structures including isochores (C+G base composition), CpG island (5'-CG-3' dinucleotide), coding-noncoding regions (periodicity-of-3), etc.

Partitioning (segmenting) a sequence into two subsequences is operationally simple: find the position that maximizes the base composition difference between the two subsequences. This process can be continued recursively.

The question is when to stop? Are two subsequences better than one whole sequence?

# DNA Segmentation as a Model Selection

the model before the partitioning: one random sequence ( $K=3$ : 4 base composition, but 1 constraint).

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

the model after the partitioning: two random subsequences each with a different base composition ( $K=3+3+1 = 7$ , one more parameter for the partitioning point)

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

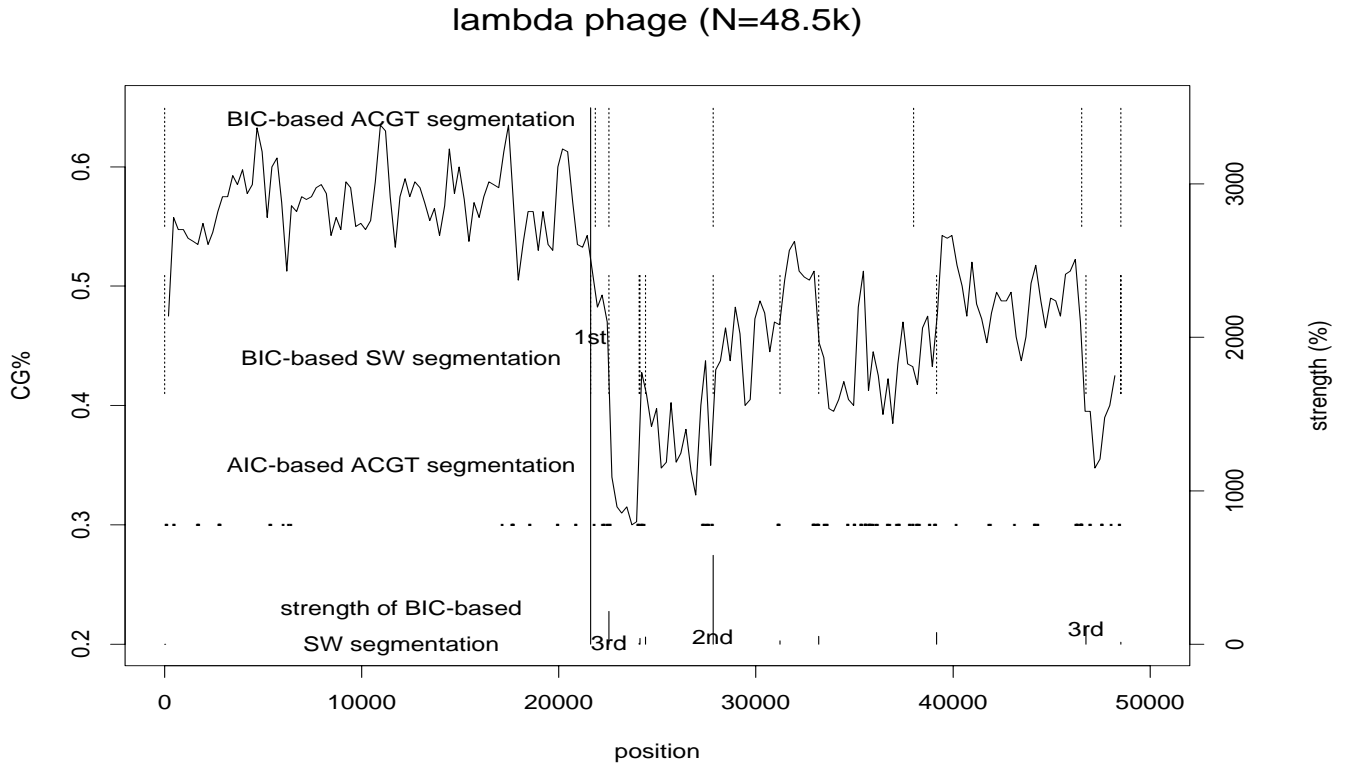
$AIC_2 < AIC_1$  (or  $BIC_2 < BIC_1$ ) leads to

$$2N \left( E - \frac{i}{N} E_l - \frac{N-i}{N} E_r \right) > 8 \text{ (or } 4 \log(N) \text{)}$$

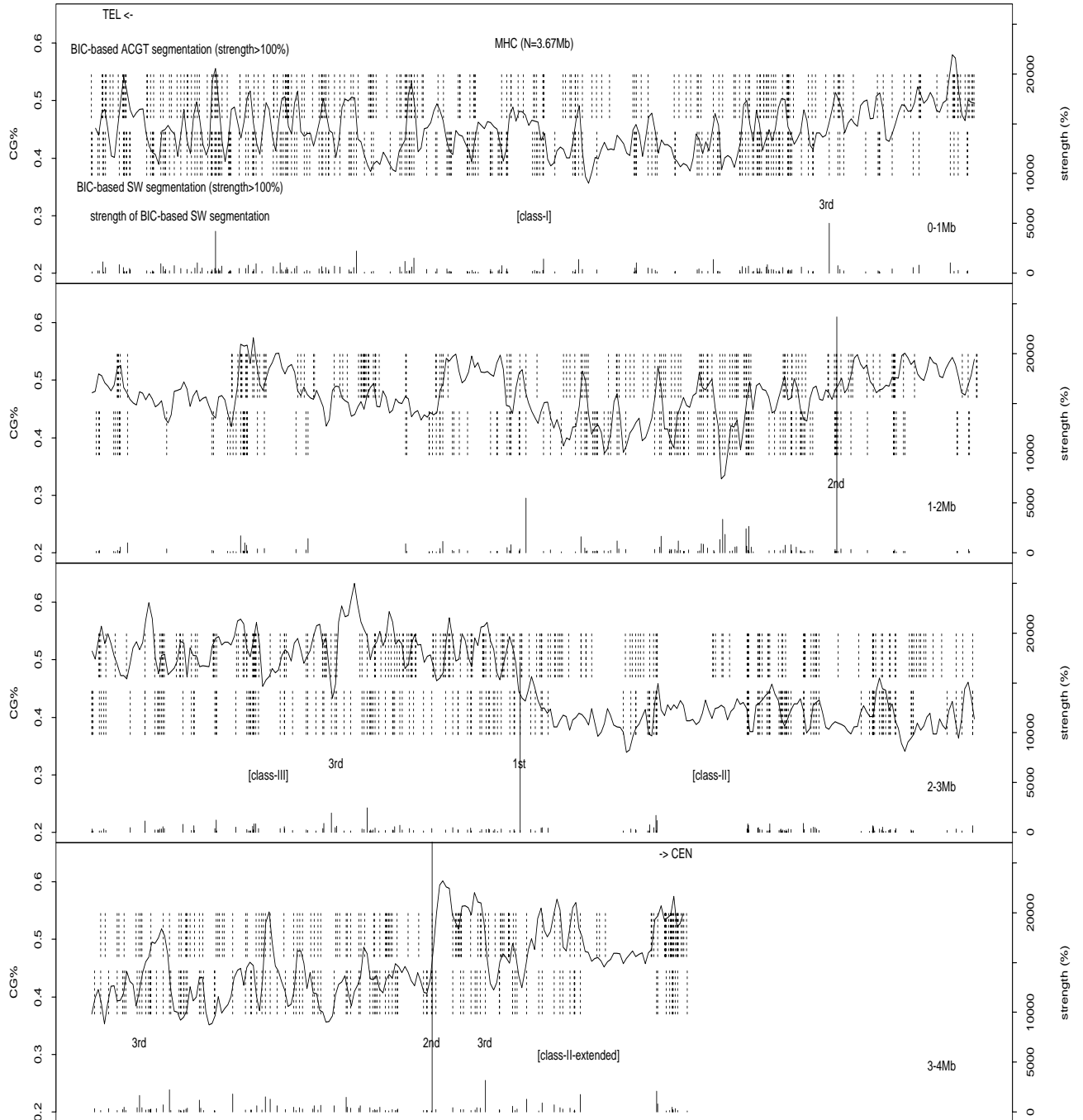
where  $E, E_l, E_r$  are entropies of the whole, left, and right sequence,  $N$  the sequence length. the quantity in the parenthesis is called Jensen-Shannon distance  $D_{JS}$ .

In practice, AIC-based stopping criterion is too relaxed. BIC-based stopping criterion is much more stringent, but may still not be stringent enough if one is interested in very large domains.

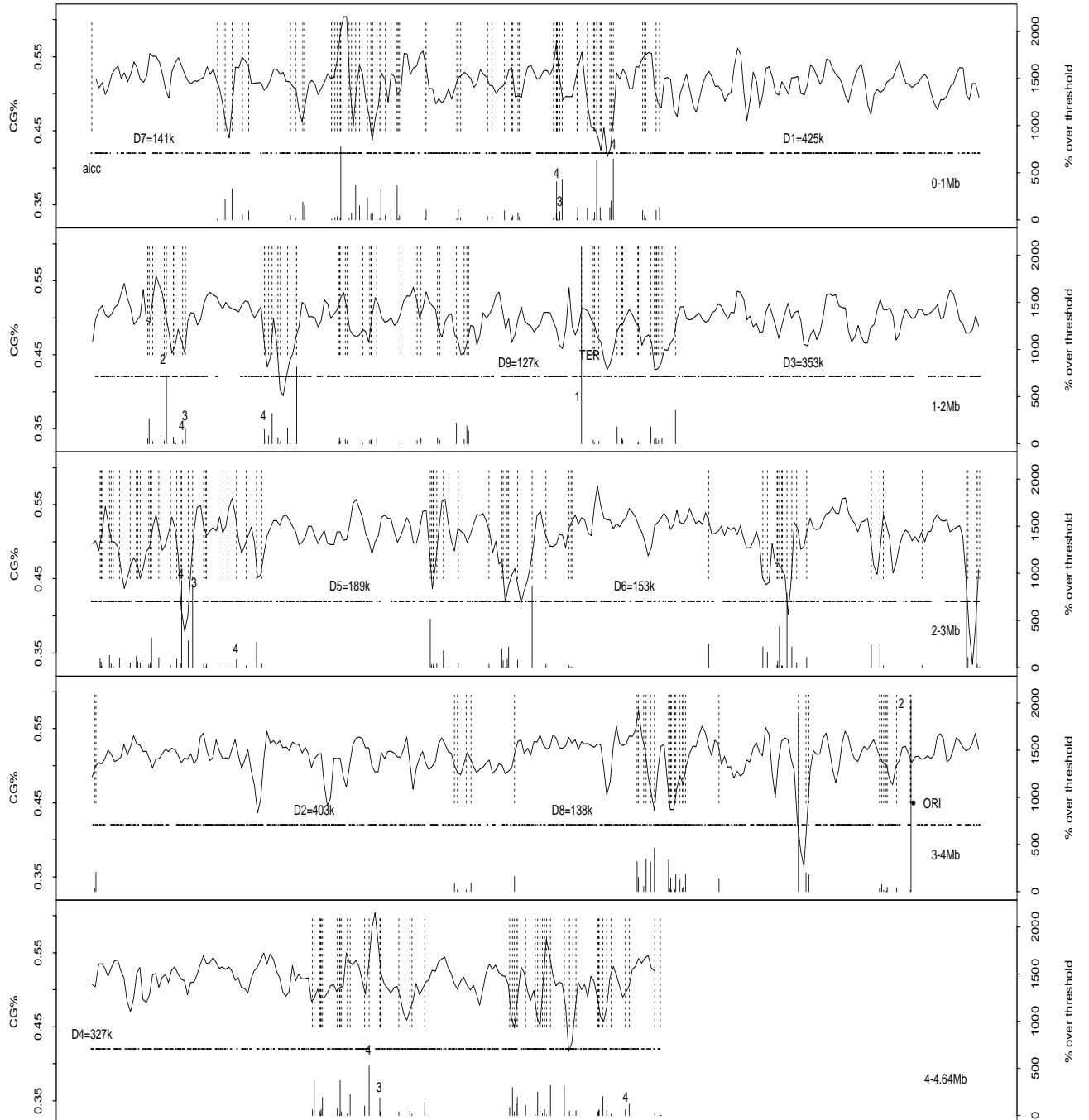
lambda bacteriophage: 2 domains can be easily seen



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



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



To raise the stringency of the stopping criterion even further, we can define the “segmentation strength” (BIC- based version)

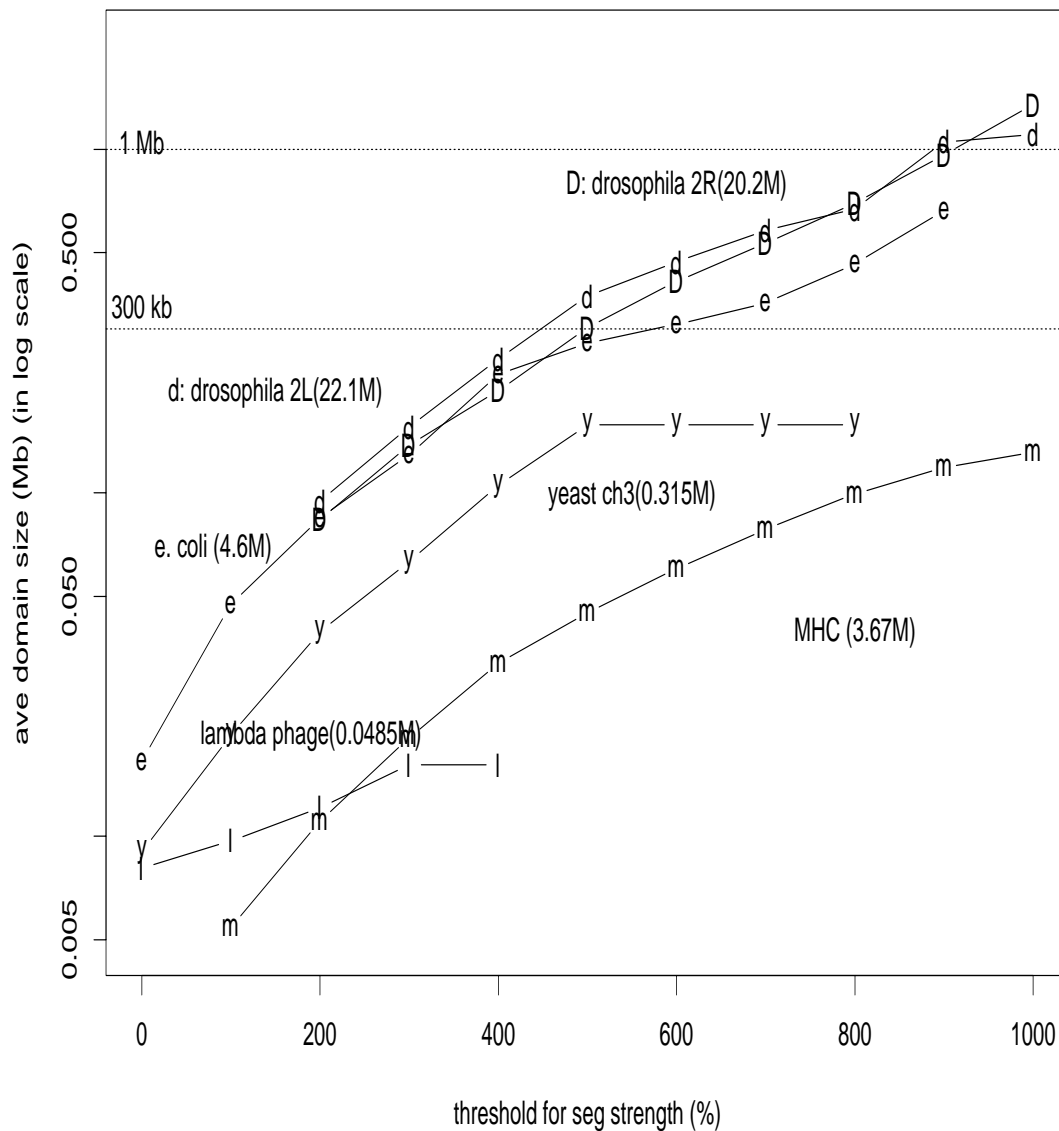
$$strength = \frac{2ND_{JS} - 4 \log(N)}{4 \log(N)}$$

where  $D_{JS}$  is the Jensen-Shannon entropy

Here, the two-subsequence model should not only be better than the one-sequence model, but it should be “much much better” (arbitrarily determined by a threshold on the strength).

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



In all four applications, model/variable selection is able to simplify the data needed: selective number of risk factors, limited number of genetic markers, fewer genes (or even one gene!), an appropriate number of DNA domains depending on one's interest.

# Future Works

**better understanding of AIC/BIC** cross-validation and generalization ability

**other models** tree-based models (if-else situation), relationship between heterogeneity and model selection.

**model averaging** better understanding. boosting(?)

**new data and new applications**

**extensions of the segmentation method** CpG island identification, coding-noncoding borders, protein domains(?)

## Acknowledgments

Andrea Sherriff: child asthma data

Yaning Yang: s-plus

Dale Nyholt: linkage analysis

many others for discussions