

**Selecting Mathematical Models of
DNA Sequences by Akaike
Information Criterion**

Wentian Li

Lab of Statistical Genetics
Rockefeller University

July 2000

Describing the Symbolic Sequence of DNA (DNA Text)

- Random sequence
- Random sequence with a constraint
($P_c = P_g, P_a = P_t$)
- First-order Markov chain
- Second-order and higher-order Markov chains
- Hidden Markov chains
- Any complicated descriptions one comes up with

The Main Question:

Which model is a better (best?) model of a given DNA sequence?

The Naive Answer:

Whatever the model that fits the data better is the better model.

The Problem with This Naive Answer:

A more complicated model will *always* fit a *given* data set better than a simpler model. Listing all items one by one, for example, will describe the given data set perfectly.

The Correct Answer:

A cost related to model complexity should be deducted from a “goodness of fit” of the data.

A better model is a model that fits a given data well *and* has a lower model complexity.

The emphasis on simple models can also be called “Ockham/Occam’s Razor”

“frustra fit per plura quod potest fieri per pauciora” (it is vain to do with more what can be done with fewer)

More quantitatively: how good a model fits the data?

Answer: Likelihood

Prob(data| model): as a function of the data is called probability; as a function of the model is called likelihood.

Even when the model is chosen, the parameters in the model must be determined by the data. The typical approach is the

Maximum Likelihood Estimation

(Other approaches include “least square”, “Bayesian estimation”).

Example: Random Model

The sequence is:

aaccgt

Maximum likelihood estimation of the four parameters (three are independent): $P_a = 1/3$, $P_c = 1/3$, $P_g = 1/6$, $P_t = 1/6$.

Likelihood of the sequence is:
 $(1/3)^2(1/3)^2(1/6)(1/6) = 1/324$

In general,

$$L = P_a^{N_a} P_c^{N_c} P_g^{N_g} P_t^{N_t} = \prod_{i=1}^4 P_i^{N_i}$$

Its minus log version is

$$-\log(L) = -\sum_{i=1}^4 N_i \log(P_i) = N \cdot E$$

where E is entropy. (note 1: $-2 \log(L)$ is $2NE$. note 2: maximizing likelihood is parallel to minimizing entropy.)

A new proposal for maximizing the goodness-of-fit with consideration of the model complexity cost: Akeike Information Criterion

Rather than minimizing $-2 \log(L)$, we minimize $-2 \log(L) + 2K$, where K is the number of parameters to be estimated in the model.

This quantity is called *AIC*.

Proposed by the Japanese applied mathematician Hirotugu Akaike (1971, though published in 1973-74).

The factor of 2 is due to the historical reason ($-2 \log(L)$ is closely related to the χ^2 when likelihood is normally distributed)

The Theoretical Foundation of AIC is the Kullback Distance (Relative Entropy) in Information Theory

The goal of a model selection is to minimize the difference between the *true model* ($T(x)$) and your approximation model ($M(x)$).

Such difference can be measured by the Kullback distance:

$$\begin{aligned} I &= \int T(x) \log \frac{T(x)}{M(x)} dx \\ &= C - \int T(x) \log M(x) dx = C - E_x[\log(M(x))] \end{aligned}$$

Model is a function of the data (y) because the parameters θ in the model are estimated by the data. We write $M(x|\hat{\theta}(y))$.

Theoretical Foundation (cont.)

Akaike's contribution was to *average the Kullback distance over $\hat{\theta}$* :

$$\begin{aligned} E_{\hat{\theta}}[I] &= \int T(y) \int T(x) \log \frac{T(x)}{M(x|\hat{\theta}(y))} dx dy \\ &= C - E_{\hat{\theta}(y)} E_x [\log(M(x|\hat{\theta}(y)))] \end{aligned}$$

And under certain assumption, the second term is approximately equal to $\log(L) - K$ (L is the maximum likelihood: $L(\hat{\theta}|y)$).

Or

$$E_{\hat{\theta}}[I] \approx C + \frac{AIC}{2}$$

Back to our random model:

The number of parameters to be estimated is $K=3$.

$$AIC = -2 \log(\hat{L}) + 2K = 2N \cdot E + 6$$

Using strand symmetry to reduce K

$$K=1 \quad (P_{gc}, P_g = P_c = P_{gc}/2 \text{ and } P_a = P_t = (1 - P_{gc})/2)$$

$$L = \left(\frac{P_{at}}{2}\right)^{N_a} \left(\frac{P_{at}}{2}\right)^{N_t} \left(\frac{P_{gc}}{2}\right)^{N_g} \left(\frac{P_{gc}}{2}\right)^{N_c} = \left(\frac{P_{at}}{2}\right)^{N_{at}} \left(\frac{P_{gc}}{2}\right)^{N_{gc}}$$

$$-2 \log(L) = -2 \sum_{i=s,w} N_i \log\left(\frac{P_i}{2}\right) = 2N \cdot E + 2N \log(2)$$

(note: E is the two-symbol entropy)

$$AIC = 2N \cdot E + 2N \log(2) + 2$$

(note: AIC can only be used to compare models on the *same* data set. A 4-symbol sequence and a 2-symbol sequence are not considered to be the same.)

First-order Markov model

K=15 (3 first position base probabilities, plus 12 transition probabilities) (16 joint dinucleotide probabilities minus 1 normalization condition).

$$\begin{aligned} L &= P(x_1)P(x_2|x_1) \cdots P(x_N|x_{N-1}) \\ &= P(x_1) \prod_{i,j=(a,c,g,t)} P_{i \rightarrow j}^{N_{ij}} \end{aligned}$$

(note: $\sum_{i,j=(a,c,g,t)} N_{ij} = N - 1$)

$$\begin{aligned} -2 \log(L) &= -2 \log(P(x_1)) - 2 \sum_{ij} N_{ij} \log\left(\frac{p_{ij}}{p_i}\right) \\ &= -2 \log(P(x_1)) + 2(N - 1)E_2 - 2(N - 1)E_1 \end{aligned}$$

(notes: $p_{i \rightarrow j} = p_{ij}/p_i$; E_2 and E_1 are entropies based on dinucleotides and single-base, respectively)

$$AIC = -2 \log(P(x_1)) + 2(N - 1)E_2 - 2(N - 1)E_1 + 30$$

(note: if a periodic boundary condition is used, the first term is dropped, and N-1 becomes N).

Mth-order Markov model

$$K = 4^{M+1} - 1$$

$$\begin{aligned} L &= P(x_1, \dots, x_M) P(x_{M+1} | (x_1, \dots, x_M)) \cdots P(x_N | (x_{N-M}, \dots, x_{N-1})) \\ &= P(x_1, \dots, x_M) \prod_{i_1, i_2, \dots, i_M, j} P_{(i_1, i_2, \dots, i_M) \rightarrow j}^{N(i_1, i_2, \dots, i_M, j)} \end{aligned}$$

$$\begin{aligned} -2 \log L &= -2 \log P(x_1, \dots, x_M) \\ &\quad -2 \sum N(i_1, i_2, \dots, i_M, j) \log \left(\frac{P(i_1, i_2, \dots, i_M, j)}{P(i_1, i_2, \dots, i_M)} \right) \end{aligned}$$

$$\begin{aligned} AIC &= -2 \log P(x_1, \dots, x_M) \\ &\quad 2(N - M)E_{M+1} - 2(N - M)E_M + 2(4^{M+1} - 1) \end{aligned}$$

note: E_{M+1} and E_M are entropies based on (M+1)-block and M-block, respectively.

Cyclic Markov model (inhomogeneous Markov model)

$$K=15 \times 3 = 45.$$

A model for codon structure: the transition probabilities depend on codon position (1,2,3).

Suppose the periodic boundary condition is used...

$$\begin{aligned} L &= P(x_1|x_N)P(x_2|x_1) \cdots P(x_N|x_{N-1}) \\ &= \prod_{i,j=(a,c,g,t)} P_{1,i \rightarrow j}^{N(1,ij)} P_{2,i \rightarrow j}^{N(2,ij)} P_{3,i \rightarrow j}^{N(3,ij)} \end{aligned} \quad (1)$$

$$AIC = \sum_{k=1,2,3} (2N_k E_{k,2} - 2N_k E_{k,1}) + 90$$

notes: N_k are the number of bases in codon position k . $E_{k,1}$'s are the single-base entropy calculated from codon position k ...

Hidden Markov Models

Suppose there are two hidden variables, $K=9$ (3 for the hidden variable Markov chain, 6 emission probabilities). (?)

Hidden variable transition probabilities: $p(\alpha \rightarrow \beta)$.

Emission probabilities: $e(\alpha \rightarrow (a, c, g, t))$

For one (hidden-variable) path:

$$L(X, S) = p(s_1)e(s_1 \rightarrow x_1)p(s_1 \rightarrow s_2)e(s_2 \rightarrow x_2) \cdots$$

Sum over all possible paths, we have the likelihood:

$$L \equiv L(X) = \sum_S L(X, S)$$

L can be determined by an estimation-maximization (EM) algorithm (Baum-Welch training) (for details, see e.g. Durbin, Eddy, Krogh, Mitchison, *Biological Sequence Analysis* (1998)).

A real sequence: yeast chromosome III

Sequence length N=315,341

model	$-2 \log(L)$	K	AIC	AIC-AIC _r
3rd MC	849452.60	255	849962.60	-7567.73*
4th MC	848077.53	1023	850123.53	-7406.80
2nd MC	851680.44	63	851806.44	-5723.89
5th MC	845080.99	4095	853270.99	-4259.34
cyclic MC	853261.98	45	853351.98	-4178.35
1st MC	853474.63	15	853504.63	-4025.70
HMM	854889.64(?)	9	854907.64	-2622.69
ran	857524.33	3	857530.33	0
AT/GC	857619.96	1	857621.96	91.63
AG/CT	874310.89	1	874312.89	16782.56

note: the (two-hidden-variable) HMM result is preliminary

Application of AIC to Segmentation of DNA sequences

The purpose of a segmentation is to delineate regions (domains) with distinct (different) base compositions.

E.g. GC-rich/GC-poor isochores. They may match the chromosome staining bands, and gene-rich/gene-poor regions.

The 1 → 2 Segmentation by Jensen-Shannon divergence

In [Bernaola-Galván, Román-Roldán, Oliver, Phys Rev E, 53:5181-5189 (1996)], it was suggested to pick position i which maximizes the Jensen-Shannon divergence:

$$D(i) = E_{total} - \frac{i}{N}E_{left} - \frac{N-i}{N}E_{right}$$

where E_{total} , E_{left} , E_{right} are the entropies obtained from the complete sequence, subsequence from position 1 to i , and subsequence from position $i+1$ to N .

Recursive 1 → 2 Segmentation

The 1 → 2 segmentation is repeatedly applied to smaller and smaller domains.

A New Perspective of the 1 → 2 Segmentation

Before the segmentation, **the sequence is modelled as one random sequence** (with 3 parameters).

After the segmentation, **the sequence is modelled as two domains each as a random sequence (but different base compositions)**. The number of parameters in this model is 7 (3 for left base composition, 3 for right base composition, and 1 for partition point).

Which model is better can be determined by AIC. And it provides a stopping criterion for the recursive segmentation!

Before the segmentation, $AIC = 2NE_{total} + 6$.

After the segmentation,

$$AIC = 2N_{left}E_{left} + 2N_{right}E_{right} + 14$$

Decrease of AIC:

$$\begin{aligned} & 2(NE_{total} - N_{left}E_{left} - N_{right}E_{right}) - 8 \\ = & 2ND_{JS} - 8 \end{aligned}$$

To make sure AIC decreases, D_{JS} should be larger than $4/N$: this is the stopping criterion in the AIC framework

A Small Complication: AIC_c

When the sample size N is small, it is shown that the following quantity is a better approximation of AIC:

$$AIC_c = -2 \log(\hat{L}) + 2K + \frac{2K(K + 1)}{N - K - 1}$$

AIC_c should always be used when the sample size is smaller (relative to the K).

Another Competing Quantity: Bayesian Information Criterion (BIC)

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

BIC assumes that the underlying model (true model) does not increase its dimensionality as the sample size is increased. It tends to select under-fit (simpler) models than AIC.

Whether AIC or BIC should be used is under debate (and the crucial consideration is whether the data is “dimensionally consistent”).

AIC/BIC should be useful to other sequence analysis projects as long as a training or model selection is an issue. For example, the selection of a hidden Markov model in multiple-sequence alignment.