Maximizing the Information Rate from Single-Molecule, Motion-Based DNA Sequencing Using RNA Polymerase

William H. Press

May 2, 2006

1 Introduction

Abbondanzieri et al. [1] have observed single base-pair stepping by RNA polymerase along DNA as the polymerase produces a new RNA transcript. At each step, the RNAP binds a coding-appropriate NTP (that is, one of ATP, CTP, GTP, and UTP) and incorporates it into the RNA. Varying the concentration of the individual NTPs results in waiting times for particular steps that are coding specific [2]. For example, if a particular NTP is starved, then there are relatively long pauses before the steps that add that base pair. Starving each NTP in turn results in a four-pass method for doing full sequencing at the single molecule level, as actually accomplished by Greenleaf and Block.[2]

One immediately thinks of reducing the four-pass method to a single-pass method by adjusting the NTP concentrations so as to make the mean waiting times widely separated on a logarithmic scale, for example 1, 10, 10^2, and 10^3 time units. Then the observation of a single waiting time is informative (if with some residual ambiguity) about all four base-pair possibilities. However, 25% of the time, one must wait a long time, $\sim 10^3$ time units, for the answer. So, while the information content of a single observed step may be increased, the information *rate* has become very low. In practice, of course, 10^3 would be such a long time that other, confounding, sources of noise would dominate.

What about decreasing the logarithmic separation, which, while reducing the information content per measurement, might give a larger information rate? Or what about a different pattern of mean waiting times entirely?

In this note we imagine the fanciful or futuristic scenario in which large-scale sequencing is performed by observing the times of single-molecule steps, and in which high accuracy is achieved by multiple resequencing. In such a scenario, we want to maximize the information rate or throughput associated with each measurement. We will show how to quantify this, and we will solve for values of the mean waiting times (which can be mapped to NTP concentrations) that achieve the maximum information throughput.

OK, so you just want to know the answer? It turns out to in fact be optimal to do one base pair at a time and *not* try to get information that separates the other three. Maximum information throughput is obtained when three (non-informative) bases have equal mean waiting times, and the (informative) fourth base has a mean waiting time about 6.66 times longer. The rest of this note derives this result.

2 Assumptions and Setup

For each nucleotide $i = A, C, G, U$, we assume an exponential distribution of waiting times with rate λ_i, adjustable by the experimenter by varying NTP concentrations. For simplicity, we assume that the distribution of bases on the DNA template is uniform and base-to-base independent. If $p(t, i)$ is the joint probability of seeing a waiting time t resulting in the addition of nucleotide i, then we have,

$$p(t, i) = \tfrac{1}{4}\lambda_i \exp(-\lambda_i t) \tag{1}$$

Note that

$$\int_0^\infty \sum_i p(t, i) dt = 1 \tag{2}$$

Also of interest is the unconditional probability of a waiting time t, denoted $p(t)$, and the conditional probability of seeing nucleotide i given t, denoted $p(i|t)$. These are given by

$$p(t) = \tfrac{1}{4} \sum_i \lambda_i \exp(-\lambda_i t) \tag{3}$$

and

$$p(i|t) = \frac{\lambda_i \exp(-\lambda_i t)}{\sum_j \lambda_j \exp(-\lambda_j t)} \tag{4}$$

From a single observed waiting time t, we infer a probability distribution for the four possible values of i, namely $p(i|t)$, $i = A, C, G, U$. The figure of merit for this distribution—its information content in bits—is

$$I(t) \equiv 2 - H(t) = 2 + \sum_i p(i|t) \log_2 p(i|t) \tag{5}$$

a quantity with value between 0 and 2 (bits). This can be viewed as either the Shannon *neg*entropy (that is, the entropy deficit from a uniform, noninformative, 2-bit distribution), or, similarly, as the Kullback-Leibler distance from a uniform prior distribution. (See, e.g., [3].) It is well known, and we will not derive here, that the mean number of independent measurements necessary to achieve a desired certainty of classification, as measured by a fixed log-odds threshold, varies inversely with the information content defined by equation (5). The value $I(t) = 2$ implies a perfect classification, that is, an error-free choice among A, C, G, U. The value $I(t) = 0$ implies no information.

Equation (5) is conditioned on a particular value of t, which we do not control. The expectation over the values of t that we expect to observe is

$$\begin{aligned} E[I(t)] &= \int_0^\infty p(t) I(t) dt \\ &= 2 + \tfrac{1}{4} \int_0^\infty \sum_j \lambda_j e^{-\lambda_j t} \left[\log_2 \left(\lambda_j e^{-\lambda_j t} \right) - \log_2 \left(\sum_i \lambda_i e^{-\lambda_i t} \right) \right] dt \end{aligned} \tag{6}$$

where some algebra has been performed.

While some further algebraic simplification of $E[I(t)]$ is possible, the core expression,

$$\int_0^\infty \left(\sum_j \lambda_j e^{-\lambda_j t} \right) \ln \left(\sum_j \lambda_j e^{-\lambda_j t} \right) dt \tag{7}$$

which is a function of the four λ_i's, requires numerical integration. However, the single-exponential convergence of the integrand admits very efficient integration methods, such as the DE rule.

The mean rate at which the RNAP steps is the harmonic mean of the four λ_i's,

$$\lambda_m = 4 \Big/ \sum_j \lambda_j^{-1} \tag{8}$$

The final figure of merit that we want to maximize over all choices of λ_i's is then

$$\text{F.M.} = \lambda_m E[I(t)] \tag{9}$$

in units of information bits per unit time. As one might expect, this goes to infinity if all the λ_i's are increased simultaneously, corresponding to infinite stepping rates. The interesting case is to set one of the λ_i's to unity, defined as the most rapid mean rate at which individual steps can be measured, and then maximize over the other three λ_i's, with the constraint that they be all less than unity. This we now do.

3 Results

We accomplish the numerical maximization by standard hill-climbing methods. Although we cannot offer a proof that equation (9) has only a single maximum, we obtain the same maximum (up to permutations of the λ_i's) from many different starting points. Interestingly, the answer is not in the interior region (with four separated values) but is rather

$$\begin{aligned} \lambda_0 &= 1 \qquad \text{(defined)} \\ \lambda_1 &\approx 1 \\ \lambda_2 &\approx 1 \\ \lambda_3 &\approx 0.150 \\ \text{F.M.} &\approx 0.1397 \end{aligned} \tag{10}$$

That is, two of the three free λ_i's are always driven to the constraint limit of 1. The reciprocal of 0.150 is the value 6.66 that was quoted in the introduction (for a waiting time, reciprocal to a rate). While 0.14 bits may not seem like a lot, note that it is obtained (on average) per waiting time for the fast (noninformative) bases. In fact, with the assumptions made, this is the best one can do.

References

[1] Abbondanzieri, E.A., Greenleaf, W.J., Shaevitz, J.W., Landick, R., and Block, S.M. (2005) Nature, 438, 460–465.

[2] Greenleaf, W.J., and Block, S.M. (2006) Science, 313, 801.

[3] MacKay, D.J.C. (2003) *Information Theory, Inference, and Learning Algorithms* (Cambridge, UK: Cambridge University Press).