Trace modelling for abduction basecalling

D.J.Thornley*

Abstract

DNA sequencing using the fluoresence based Sanger method comprises interpretation of a sequence of signal peaks of varying size whose colour indicates the presence of a base. We have established that the ability to predict the variations effectively makes available novel error correction information which will improve sequencing efficacy. Our experiments so far have used basic models of the Sanger reaction chemistry and machine learning techniques. These have enabled us to make base calls just using context information, specfically ignoring the peak data at the base calling position. The 80% success rate of our blind experiments is striking, and will be improved by a more accurate model of trace behaviour. To this end, and to integrate the information into mainstream basecalling, we wish to develop an enzyme kinetics model susceptible to calibration of its component rates such that trace data can be accurately predicted. We describe DNA sequencing trace data, outline the trace prediction problem requirements on the model, and discuss model construction and calibration issues.

1 Introduction

This technical report extends the brief working paper presented to the PASTA 2006 workshop [1] with more detail and initial sketches of the modelling structures which we intend to explore as part of the next phase of research into Sanger and Pyrosequencing modelling.

DNA sequencing is achieved using two main methods. The Sanger method [2] was invented in the late 70s, improved with fluorescent rather than radioactive instrumentation a decade later [3], and commonly reads of the order 1000 bases per sample. The Pyrosequencing approach [4] is more recent, with enormous throughput, but shorter read lengths currently of the order 100 bases. We have proposed an approach to interpreting DNA sequencing data which improves accuracy and read length by leveraging a unique source of information encoded in the behaviour of the signals [5].

Signal intensity in both methods varies in a repeatable, sequence-dependent manner. This leads to base calling errors later in the data where noise levels are higher and separation less clear. We suggest abduction of the base sequence through hypothesis of sequence composition for subsequent rejection if the predictable data does not agree with the target data as well as other hypotheses.

*D.J.Thornley, Dept. of Computing, Imperial College London, UK. Funded by EPSRC grant GR/S60266/01 Email: djt@doc.ic.ac.uk

Hypotheses remaining after the competition are then regarded as plausible interpretations of the data. This process requires a model which can predict the trace data expected from a sample of DNA with any given base sequence. Modelleing trace data behaviour in sufficient detail is an on-going research issue. We have performed machine learning work in which the behaviour is exploted through a classifier, resulting in a clear demonstration of the presence of contextual information by calling bases using the context without reference to the signal at the calling position [6]. In this paper, we focus on the Sanger reaction, but note that the pyrosequencing reaction will be susceptible to similar but conveniently simpler analysis. Svantesson provides a basic model and initial parameterizations [8] which approximately simulates Pyrograms as a system of ordinary differential equations produced from a simplified model of polymerase action.

Our Sanger models to date, while operable in validation experiments, have been approximate, and mimic rather than simluate the system. That is, the model is designed to reproduce the observable characteristics of the data, rather than fundmentals of the system producing it. Analysis of DNA polymerases in the biochemistry literature has reached a level of maturity and coherence which we can realistically work to adopt into a computation framework. In particular, a framework model for DNA polymerase activity recently produced [12] can be rewritten as a Markov chain or Petri net, or in a stochastic process algebra for subsequent manipulation.

Recentdevelopments in the modeling of biochemical systems by members of the process algebra research community offer a means for pursuing the next stage of modeling research in which the true biochemical interactions and structural dynamics will be modelled in a formal context. We are particularly interested in the developing approach of Calder, Gilmore and Hillston, of which we see an example in [9]. In this approach using the PEPA [10] language and associated tools, dual viewpoints on the system are formulated, allowing some freedom in manipulation, which will assist the analysis of some of the parameters we need to calibrate.

This working paper introduces the motivation for our model of the Sanger reaction kinetics, and some of the main issues which will influence the form of the model. Personal communication with the authors of [9] has highlighted the representation of inhibition of enzyme action as an interesting issue. The Sanger reaction exhibits substrate substitution and sequestering, and allosteric modulation of enzyme activity, which correspond to inhibition.

2 The sanger method

The Sanger method [2] allows us to identify the base sequence of a sample of DNA by copying all the DNA molecules in the sample starting at the same location, but ending at stochastically selected locations with a label indicating that final base. When we electrophorese these fragments, they are sorted into order of size, and imaging them allows the sequence to be read off according to the labels. In figure 1 we see a sequence of clear peaks which are read off to give the base sequence shown as letters over the trace.

DNA is copied by using a DNA polymerase and providing it with nucleotides to add to a complementary primer sequence. The peaks in the trace data arise

from imaging the bodies of fragments resulting from copying the DNA template by adding deoxynucleotides (or dNTP) from a given position, but terminating when a modified terminator or dideoxynucleotide (ddNTP) is incorporated. The terminators are labelled according to which type of base they represent, and when the fragments are sorted by size through electrophoresis, these labels can be imaged to give a trace of which an example excerpt is shown in figure 1.

3 Sanger sequencing trace data

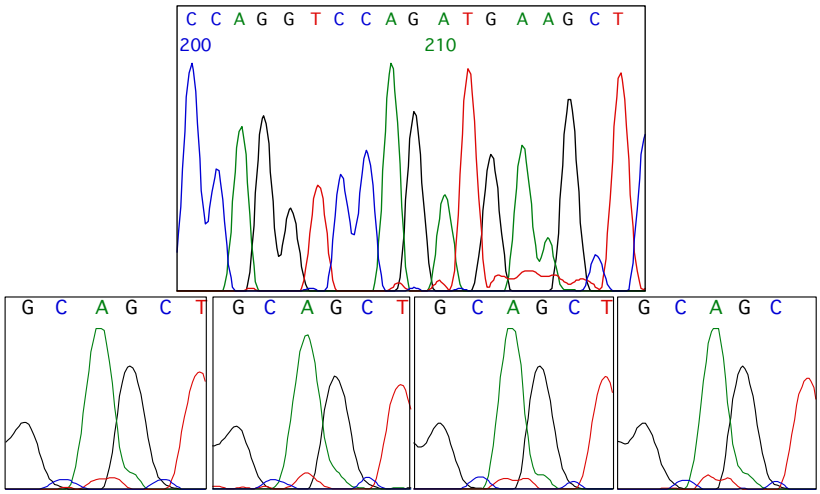

Figure 1: Sequencing trace data excerpts. Each peak indicates a base in the sequence. Samples with the same base sequence appear substantially identical.

The peak patterns from one sample of DNA appear substantially identical to those from another sample with the same base sequence. Figure 1 indicates that the peak heights vary widely. Later in the data, noise levels worsen, peak spacing becomes more erratic, and small peaks are sometimes submerged. If we know what size the peaks should be, then we can detect the features of interest. This is the goal of our modeling research.

4 An early modelling approach - the bulk sequential model

Begninning with a similar premise to [13], we wrote down a description of the production of extension fragments which expresses sequence dependent action of the polymerase[1]. We refer the reader to [5] for details.

Each polymerase begins by copying at the first base position after the end of the primer. The probability of termination at that point is the probability

[1]Which we propose independently in [5], but recently learn was suggested – but not modelled or leveraged – in [7]

of incorporating a terminator. Constant factors indexed on the DNA sequence forming the footprint of the polymerase express the dependency of the extent to which the polymerase discriminates against terminators. We suggest that this is an *allosteric*, or shape dependent effect. Terminators are unnatural, and hard to incorporate as part of the copying process. This is an important point which we will discuss later when summarizing the detailed enzyme kinetic model we wish to build.

This model gives an overall exponential decay, with local detail in the peak heights provided by the sequence dependent terminator discrimination fators R. It can predict in the region of 80% of the variation in peak sizes, which is enough to apply abductive basecalling. The sequence footprint affecting polymerase activity covers about three positions to the left, and one to the right [7], so five peak heights are directly affected by each base position, and many more are indirectly affected by the change in number of template molecules still copying after them.

We have tried a simplified representation of the biochemistry, and machine learning methods, and each approach has demonstrated that the data encodes redundant information in the peak heights and spacing. This enables error correction, and a range of interpretive capabilities previously thought impossible using standard sequencing equipment. Currently, we capture approximately 80% of the variation in peak heights with either of these approaches. We believe that the variation we have missed results from inhibition of processes in the Sanger reaction through allosteric effects, sequestering of substrate, and subsitution of incorrect substrate. Ssubstrate availability effectively links the otherwise independent polymerase activity on the template DNA molecules.

4.1 Calibration

The bulk sequential model requires values for the preference of the polymerase for normal nucleotides over terminators. In [5] we derive a basic numerical analysis approach which seeks to minimize the error between predictied ratios of neighbouring peak heights against ratios in real trace data. This places constraints on the sequence composition of the trace data, which led us to build bespoke DNA molecules for the purpose. While the ability to swiftly calibrate the models from a small set of DNA molecules is desirable, practical barriers have motivated the formulation of a description of the fit which allows calibration from existing data by applying an information theoretic approach. Essentially, the error function to be minimized comprises the variance in estimates of a given sequence dependent terminator discrimination factor (R_k in section 4) as observed at different locations in the calibration data. We refer the reader to [5] for details.

5 Abduction sequencing

Using the bulk sequential model, if we ignore the peak size at the position we're trying to call, and hypothesize A, C, G or T for the basecall, the hypothesis which leads to the best fit peaks for the neighbourhood is the correct basecall on 78% of occasions. We can beat this by a small margin using machine learning techniques as we describe below in section 6, and if our model were perfect,

we might expect close to 100%. This theme of base calling through hypothesis testing, first suggested in [5] motivates the development of predictive models of peak size in trace data.

The most successful model so far predicts about 80% of the detailed variation in peak sizes. We have used this in a peculiar blind basecalling experiment in which we predict the peaks in the context of the hypothesized basecall, but ignore the data at the calling position. This is succeeds for just under 78% of base calls, which compares to the approximately 25% we would expect from random guessing. In the particular data set we used, we could have predicted the base on almost 30% of occasions looking at the surrounding sequence because of a biased base composition. We found this as part of classification experiments in which neural nets responding to contextual bases, peak sizes and spacing achieved a blind calling rate of just under 80%. We expect the abductive blind calling rate to approach 100% in traditionally "good" data with a full model of the system. This will allow implementation of a basecaller to sequence traditionally "bad" data by using all the information available.

Predicting traces involves interaction with the target data, since some parameters are not known *a priori*, most importantly the terminator fraction, or relative concentration of ddNTP molecules in the reaction. For example, if we use too high a value for the terminator fraction in our prediction, the trace peaks will die away too quickly. When we have the footprint modulated terminator discrimination factors, we can numerically estimate the required value with, for example, the Levenberg marquardt approach, with the four terminator fractions as free variables.

We therefore propose a sequence composition, find the reaction conditions which make that hypothesis generate traces which best fit the target data, and measure the degree of fit. If we propose a set of hypotheses which includes the correct interpretation, we find our answer as that which fits best.

6 Machine learning

We trained a classification tree taking as its inputs the base sequence and trace peak sizes around a basecall, to output the base call as its class. When the inputs include the trace peaks at the basecall position, the classifier did not include any other measurements in its decision process and produced a base calling error rate similar to a traditional basecaller. When the peak sizes at the base calling position were excluded (as in the example we described in section 5), this gave an error rate of just under 30%. The classification tree made decisions swinging on peak heights, peak spacing and base composition. These measurements were used to train a range of neural nets for bagging, which resulted in an error rate of less than 20%, which we compare with an error rate of 68% using just the base composition. If the base sequence were truly random, this would be 75%.

7 Enzyme kinetics

We are fortunate to have a general model of polymerase behaviour in copying DNA, provided by Keller and Brozik [12]. The model focusses on a single DNA strand with an associated polymerase molecule, and tracks this complex's

internal configuration changes and interactions with substrate. We refer the reader to the second figure in [12], which gives a summary of the incorporation cycle. Briefly, the polymerase resembles a right hand cradling the DNA template with the sticky end of the complementary strand at the crook of the thumb and forefinger. An addition cycle involves the arrival of a nucleotide, the closing of the fingers, chemical cleaving of the pyrophosphate, opening of the fingers, and then escape of the pyrophosphate. There are a couple of alternative routes in this process in which stacking of the template base occurs before or after the arrival of the nucleotide. In addition, there are some loops of activity off the main cycle which do not contribute to progress, and may interact with the main forms of inhibition which we describe below.

To model the Sanger reaction using this description of polymerase activity, we must explicitly model incomplete interactions with incorrect bases, and processing of terminators. We consider it likely that interactions with incorrect terminators at various stages in the framework model generates some of the behaviour which cannot be attributed to footprint DNA directly.

The progress of copying DNA is stochastic, and is described in [12] as obeying an approximately Poissonian process because the distribution of times taken to copy a given length of DNA loooks approximately normal. We suggest that it more closely resembles the phase-type distribution often used in queueing theory, since the polymerase goes through a number of steps to achieve a nucleotide incorporation. Each step involves the crossing of an energy barrier or conjunction of two species, which are commonly regarded as Poisson processes for the purposes of kinetic modeling.

We include an intial sketch of the states and transitions we will use to model the polymerase for the Sanger and Pyrosequencing processes. First we show the same model as given in [12] as a more regular structure for clarity in figure 2. This is exactly the same model, including the labelling of states according to z) the position of the polymerase with respect to the free end of the extending strand (up/down), a) the state of the thumb (open/closed), n) the occupancy of the active regions of the polymerase (nucleotide,pyrophosphate,empty) and f) the template base (stacked/unstacked). For the Sanger model, this must be augmented with terminator incorporation, which results in an absorbing state per position on the template as shown in figure 3. The distraction states represented by thumb oscillations are omitted for clarity. This Sanger reaction component, which represents the states corresponding to polymerase activity at a given position on the template can then be composed to follow extension of a DNA strand, as sketched in figure 4.

7.1 Inhibition

The likelihood of incorporation of a terminator is much lower than for a normal nucleotide. This is intrinsic to the chemical entity, but we are more interested in what causes such interactions to vary with sequence. A strong source of sequence dependent inhibition may be allosteric: the polymerase is distorted differently by each possible base sequence in its footprint.

Other factors which inhibit the progress of the polymerase in its tasks of copying DNA include restriction of the availability of substrate, and distraction of the polymerase by incorrect substrate. Consider a single polymerase molecule which is associated with a DNA molecule with an unoccupied position for an A

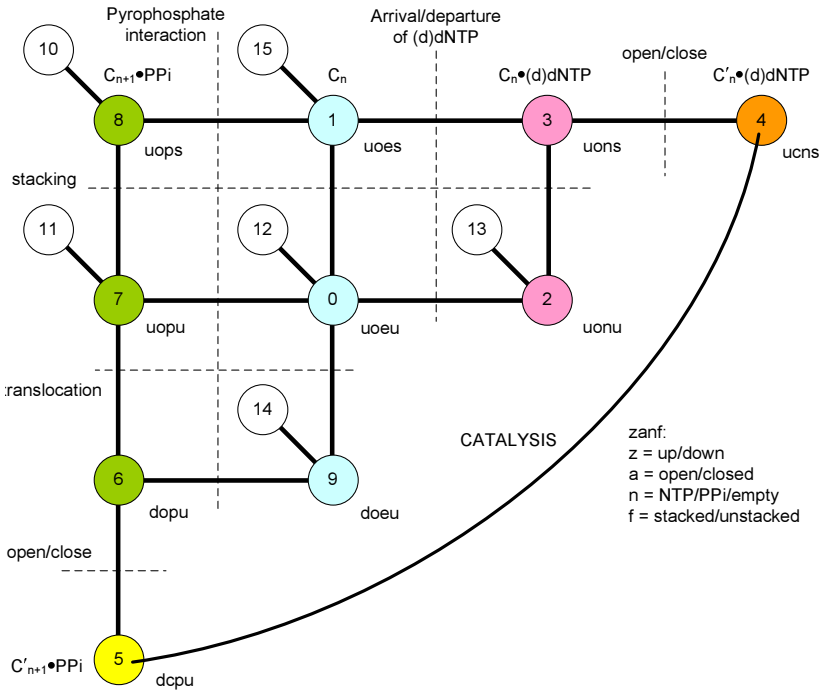

Figure 2: Keller and Brozik's model laid out for clarity as a regular structure with common transitions between variant stages in the copying process highlighted with dotted lines.

nucleotide or terminator. If an A terminator finds its way to the incorporation site, it will remain there until it is either incorporated, or it dissociates from the complex. During this period, it is not available to other complexes, which therefore experience a lower terminator fraction. This effectively inhibits the takeup of terminators by other complexes.

As well as enzymes sequestering substrate, we also see substrate sequestering enzymes. This happens in the Sanger reaction, because all four dNTPs and ddNTPs are made available, and each of these is free to associate with the polymerase/DNA complex. If an incorrect nucleotide enters the site, this blocks other incoming material, thus inhibiting the polymerase copying process. This could be referred to as transient substitution.

These various inhibitory effects involving common substrate interactions give rise to cross terms in the set of ODEs describing the system behaviour, which generates higher order derivatives, and this is a classic opportunity for oscillation. Our best explanation for the oscillation-like patterns in runs of the same base are a combination of polymerase footprint effects and emergent dynamic behaviour.

The framework model suggests that the polymerase alternates its thumb between open and closed if it is free to do so. This affects progress around the incorporation cycle, and will interact with the other forms of inhibition to generate more complex behaviour. This additional behaviour is modelled by in-

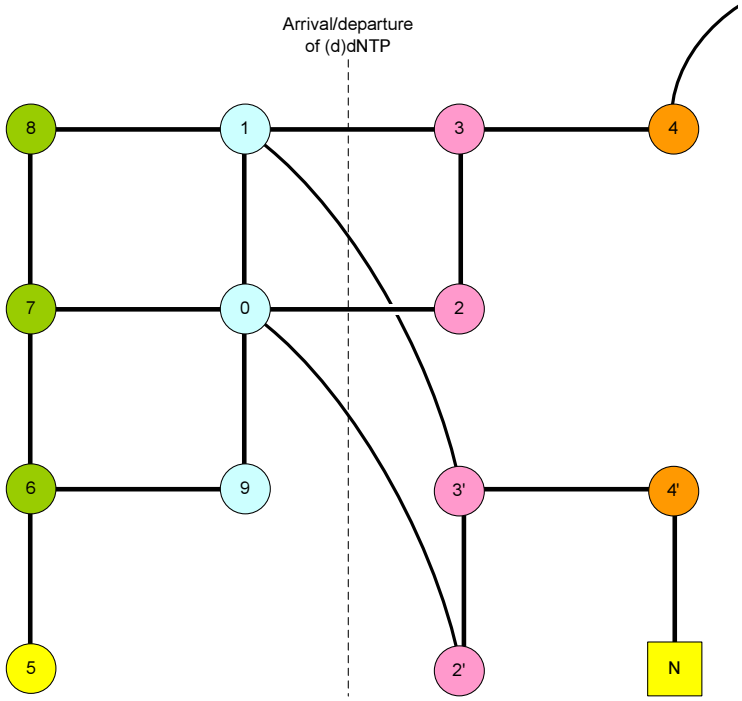

Figure 3: The framework model with an additional termination structure. The absorbing state is shown as a square. The catalysis transition is shown leading to another position on the template, since this model is intended to map onto particular base positions.

cluding the uncoloured states omitted between figures 2 and 3. The behaviour of this system as a whole will be revealed through experimentation with the models in simulation, through integration of the corresponding ODEs and subsequent calibration of rate constants, and model checking.

7.2 Calibration

The natural reaction of a numerical analyst faced with the task of calibrating a non-linear model is to formulate an expression for the error between a measurable output of the model model and some training data. We then use an iterative optimization approach – probably Levenberg Marquardt or a close relation – to explore this error space to find a minimum. However, in this case, we are dealing with an oscillatory system, so techniques from the system identification literature are indicated. It is straightforward enough to calibrate a simple harmonic model, but there may be a number of modes of oscillation in any given run of bases, so we will have to do some work in the frequency domain.

Our bulk sequential replication model has 1024 parameters to be calibrated, which are accessible through a relatively simple error expression. The framework model has potentially many more, but the dominance of the base immediately before the incorporation position suggests we may be able to begin estima-

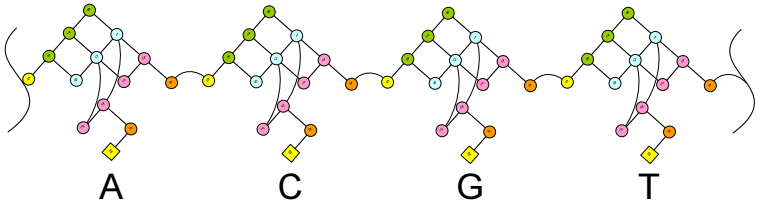

Figure 4: Sanger components composed to follow four base positions (left to right) on a template with sequence TGCA. The terminating fragments end with complementary bases ACGT respectively

tion using an approximated parameter space. The model will exhibit resonance modes, as is common in large sets of interacting ODEs. These are necessary to describe the oscillatory behaviour in homogeneous base runs, *e.g.* more than five As, in which the peak sizes oscillate, or drop suddently then seem to oscillate to an asymptote. This behaviour varies widely with the length of the homogeneous run, but is strongly repeatable for the same length of run. Without such interactions, we would expect essentially identical peak heights three bases in to the region, continuing up to a base from the end. We will therefore calculate relationships between rate constants in the model using the sharing of substrate in long runs of single bases to constrain the numerical system.

The complement of calibration of the model is profiling of the free variables for assessment of particular hypotheses in our abduction sequencing approach. We need to be able to predict peak sizes in the middle of the trace without explicitly modeling the initial conditions of the reaction. This will require the approximation of activity during this period, to be polished by reference to the target data.

8 Conclusions

We have some plausible descriptions of how substrate titre variation creates interactions between reacting complexes at different positions on the template, and can affect product titres in the Sanger reaction. The research issues to be addressed include integration of the DNA polymerase framework model into a structure which accurately reflects activity in the Sanger reaction, and calibrating the kinetic rates in that model. This calibration must ensure that local distractions and inter-complex activity dependencies are expressed with sufficient accuracy to predict the behaviour in DNA sequencing traces.

This will be pursued through the construction of a process algebraic representation of the enzyme and substrate interactions, exploration of this model's behaviour through translation to ODEs, with intial approximate calibration of the model, or calibration of an approximate model, and model checking to examine the potential for expression of certain behaviours. This research aims to take fundamental biochemistry results, extend them to an application which can only be solved using computation principles, leveraging a developing technology in computer science to achieve outcomes which feed back to biochemistry, and enable a DNA sequencing method which will benefit genetic research and

healthcare.

References

[1] David Thornley Modelling along the DNA template in the Sanger method: inhibition through competition and form Process Algebra and Stochastically Timed Activities, June 2006, London

[2] F. Sanger, S. Nicklen, and A.R. Coulson, Chain Sequencing with Chain-Terminating Inhibitors, Proc. Nat. Acad. Sci. USA 74, 1977, 5463.

[3] James M. Prober, George L. Trainor, Rudy J. Dam, Frank W. Hobbs, Charles W. Robertson, Robert J. Zagursky, Anthony J. Cocuzza, Mark A. Jensen and Kirk Baumeister. A System for Rapid DNA Sequencing with Fluorescent Chain Terminating Dideoxynucleotides Science 1987 238, 336-341.

[4] Ahmadian A, Ehn M, Hober S. Pyrosequencing: History, biochemistry and future. Clin Chim Acta, Sep 2005.

[5] D.J.Thornley Analysis of trace data from fluorescence based Sanger sequencing,, PhD thesis 1997, Department of Computng, Imperial College London

[6] David Thornley, Stavros Petridis Machine Learning in Basecalling - Decoding trace peak behaviour IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, September 2006, Toronto

[7] Alan Philippe Blanchard Sequence Specific Effects on the Incorporation of Dideoxynucleotides by a Modified T7 Polymerase Ph.D. Computation and Neural Systems, 1993 Caltech Thesis

[8] Svantesson A, Westermark PO, Kotaleski JH, Gharizadeh B, Lansner A, Nyren P, A mathematical model of the Pyrosequencing reaction system, Biophysical Chemistry 110 (1-2): 129-145 JUL 1 2004

[9] M. Calder, S. Gilmore and J. Hillston. Modelling the influence of RKIP on the ERK signalling pathway using the stochastic process algebra PEPA Transactions on Computational Systems Biology, Springer, to appear.

[10] J. Hillston A Compositional Approach to Performance Modelling, Vol. 12 of Distinguished Dissertations in Computer Science, Cambridge University Press. (1996) ISBN 0 521 57189 8.

[11] Walther D, Bartha G, Morris M Basecalling with LifeTrace, Genome Research 11 (5): 875-888 MAY 2001

[12] Keller DJ, Brozik JA., Framework model for DNA polymerases. Biochemistry, 2005 May 10;44(18):6877-88.

[13] Linda G. Lee, Charles R. Connell, Sam L. Woo, Richard D. Cheng, Bernard F. McArdle, Carl W. Fuller, Nicolette D. Halloran and Richard K. Wilson. DNA sequencing with dye-labelled terminators and T7 DNA polymerase: effect of dyes and dNTPs on incorporation of dye-terminators and probability analysis of termination fragments Nucleic Acids Research 20, 2471-2483.