Modelling along the DNA template in the Sanger method: inhibition through competition and form

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 have used basic models of the Sanger reaction chemistry and machine learning techniques. These have enabled us to make base calls only 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 require 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

DNA sequencing is achieved using two main methods. The Sanger method [1] was invented in the late 70s, improved with fluorescent rather than radioactive instrumentation a decade later [2], and commonly reads of the order 1000 bases per sample. The Pyrosequencing approach [3] 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 [4].

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 a novel method [5] which involves 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. Hypotheses remaining after the competition are then regarded as plausible interpretations of the data. This process requires a

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

model which can predict the trace data expected from a sample of DNA with any given base sequence. In this paper, we focus on the Sanger reaction, but note that the pyrosequencing reaction will be susceptible to similar but somewhat simpler analysis, and we already have a basic model provided by Svantesson [6] which mimics Pyrograms.

Our Sanger models to date, while operable in validation experiments, have been disappointingly approximate. Recent developments 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. We are particularly interested in the developing approach of Calder, Gilmore and Hillston, of which we see an example in [7]. In this approach using the PEPA [8] language and associated tools, dual viewpoints on the system are formulated, allowing some freedom in manipulation, which 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 [7] has highlighted the representation of inhibition of enzyme action as an interesting issue. The Sanger reaction exhibits substrate substitution and sequestering, which are the competitive elements of inhbition, and allosteric modulation of enzyme activity, which is inhibition due to detail in the form of the species involved.

2 The sanger method

The Sanger method [1] 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 primer sequence which is complementary to (*i.e.* sticks to) the location on the template at which we wish to start copying. The peaks in the trace data arise from imaging the distribution 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 sequence of peaks of four colours as shown.

3 Sanger sequencing trace data

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.

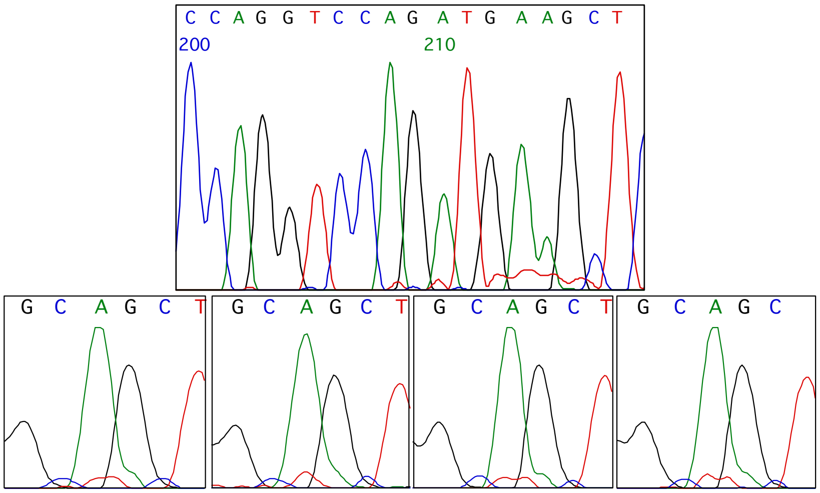

Figure 1: Sequencing trace data excerpts. Each peak indicates a base in the sequence. Samples with the same base sequence appear substantially identical.

This is the goal of our modeling research.

4 Abduction sequencing

We first suggested performing base calling by hypothesis rejection in 1997 [4], and formulated various simple models to predict trace data from a hypothesis of base composition. The most successful 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 under which that hypothesis generates 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.

5 Enzyme kinetics

We are fortunate to have a general model of polymerase behaviour in copying DNA, provided by Keller and Brozik [9]. The model focusses on a single DNA strand with an associated polymerase molecule, and tracks teh changes in internal configuration of this complex and its interactions with substrate. We refer the reader to the second figure in [9], 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.

The progress of copying DNA is stochastic, and is described in [9] as obeying an approximately Poissonian process because the distribution of times taken to copy a given length of DNA loo00ks 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.

5.1 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. The strongest source of sequence dependent inhibition is 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 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.

The framework model also suggests that the polymerase is apt to wiggle its thumb between open and closed when the opportunity arises. This can impede progress around the incorporation cycle, and will interact with the other forms of inhibition to generate more complex behaviour. The behaviour of this system as a whole will be revealed through experimentation with a model in simulation, through integration of the corresponding ODEs and subsequent calibration of its rates, and model checking.

5.2 Calibration

The natural reaction of a numerical analyst faced with the task of calibrating such a model is to formulate an expression for the error between the output of the 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. For the present model, we believe that selection of that training data is crucial, since the parameter set is large and the interactions complex.

We consider it likely that the model will exhibit resonance modes, as we might reasonably expect from a large set 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.

6 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 healthcare.

References

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

[2] 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.

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

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

[5] International Patent Application WO96/20286 July 4, 1996, European Patent EP0799320 Mar. 7 2001 and US Patent 6,090,550, Jul. 18, 2000

[6] 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

[7] 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.

[8] 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.

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