

Equilibrium Model of DNA Chip Hybridization Error

John A. Rose¹

johnrose@is.s.u-tokyo.ac.jp

Masami Hagiya¹

hagiya@is.s.u-tokyo.ac.jp

Akira Suyama²

suyama@dna.c.u-tokyo.ac.jp

¹ Dept. of Computer Science, UPBSB, University of Tokyo, Japan, and JST-CREST

² Institute of Physics, University of Tokyo, Japan, and JST-CREST

Keywords: DNA Microarray, Tag-Antitag System, Hybridization error, Nearest-neighbor model

1 Introduction

Existing methods for designing DNA microarrays and Tag-Antitag (TAT) systems do not address the inverse problem of prediction of an error-rate with which to associate designed sequences, and also neglect effects due to process coupling. A coupled equilibrium model is therefore under development for estimating the mean error-probability per chip-hybridized input-strand, or *computational incoherence*, ϵ [1, 2]. Computationally tractable cases include the error-response due to: (1) a (simple) single-tag input at arbitrary concentration (ϵ_i), useful for investigating the mean-case behavior ($\langle \log_{10} \epsilon_i \rangle$), bounding behaviors, and concentration-dependence of ϵ ; and, (2) a complex (multi-tag) dilute input (ϵ_d), considered to approximate target operating conditions. Together, these quantities may be applied to obtain a good picture of system error-response. In this brief treatment, attention is focused on the first formalism. Although couched for convenience in the language of the TAT system, the derived results apply equally to gene-specific microarrays, which involve the same fundamental principles.

2 Method: The Computational Incoherence

Consider a TAT system, in which each tag species, i is the reverse-complement of a single antitag species, $i^* \in \{j^*\}$, and measurements are taken only over hybridized pairs, ij^* . At experimental resolution, target hybridizations include the set of dsDNA conformations between matching TAT pairs $\{i, i^*\}$, irrespective of alignment, so that an error occurs for any hybrid $i, j^* \neq i^*$. At equilibrium, the probability of hybridization error/input tag, upon input of a single tag species, i is given by:

$$\epsilon_i = \frac{\sum_{j^* \neq i^*} C_{ij^*}}{\sum_{j^*} C_{ij^*}} = \left(1 + \frac{C_{i^*} K_{ii^*}}{\alpha_i} \right)^{-1}, \quad (1)$$

where $\alpha_i = \sum_{j^* \neq i^*} C_{j^*} K_{ij^*}$. Re-expression in terms of equilibrium constants and total concentrations requires solution of the accompanying $|j^*| + 1$ strand conservation equations, comprised of: $C_i^o \approx C_i(1 + K_i^{hp} + C_{i^*} K_{ii^*} + \alpha_i)$ for tag i ; and, an equation, $C_a^o = C_{j^*}(1 + K_{j^*}^{hp} + C_i K_{ij^*})$ for each antitag, j^* . Although this procedure strictly requires numerical solution of this coupled set of non-linear equations, for all but the worst encoding sets, the equation for C_i may be approximately decoupled from $|j^*| - 1$ of the remaining conservation equations by noting that $C_i K_{ij^*} \ll 1, \forall j^* \neq i^*$. In this case, $\alpha_i \approx \sum_{j^* \neq i^*} C_a^o K_{ij^*} / (1 + K_{j^*}^{hp})$, so that the equation for C_i remains coupled only to that of target antitag, C_{i^*} . Conceptually, this condition requires that encoding and reaction conditions be appropriate to avoid a sizable saturation of any untargeted antitag, so that $C_{j^*} \approx C_a^o / (1 + K_{j^*}^{hp}), \forall j^* \neq i^*$. Failure of this approximation overestimates ϵ_i . Solution for C_{i^*} proceeds by combining conservation equations for C_i and C_{i^*} , yielding quadratic equation, $a_2 C_{i^*}^2 + a_1 C_{i^*} + a_0 = 0$, with coefficients: $a_2 = K_{ii^*} (1 + K_{i^*}^{hp})$,

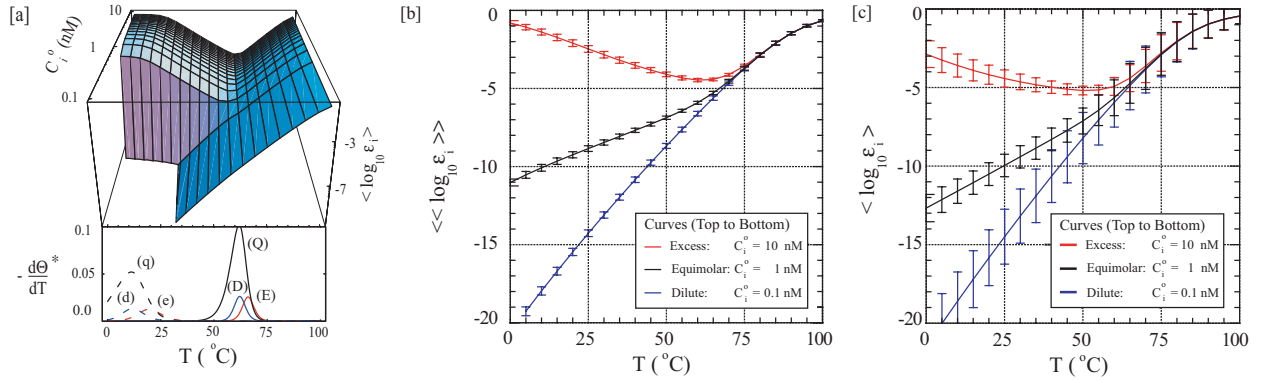


Figure 1: Simulations of TAT system error-response.

$a_1 = (1 + K_i^{hp})(1 + K_i^{hp} + \alpha_i) + K_{ii^*}(C_i^o - C_a^o)$, and $a_0 = -(1 + K_i^{hp} + \alpha_i)C_a^o$, and physical solution, $C_i^* \approx -A + \sqrt{A^2 - B}$, where $A \equiv \frac{a_1}{2a_2}$, and $B \equiv \frac{a_0}{a_1}$. This solution may be combined with α_i to yield ϵ_i . The mean value averaged over all tag species, $\langle \log_{10} \epsilon_i \rangle$ provides the first of the two tractable measures of mean-case TAT system error-response described above. The extrema of the set $\{\log_{10} \epsilon_i\}$ (the *error spectrum*) under dilute and excess conditions also serve to bound the error-response, ϵ .

3 Results and Discussion

Fig. 1[a] illustrates $\langle \log_{10} \epsilon_i \rangle$ for the simplest TAT system, in which an input tag may participate in an error duplex (10/20 bps) or a full-length target duplex (20/20 bps). Equilibrium constants were estimated via an all-or-none, Watson-Crick model, using mean stacking energetics [3], as implemented by *Mathematica*TM. For comparison, the lower panel depicts differential melting behavior for the planned and error duplexes (solid curves/upper-case letters and dotted curves/lower-case letters, respectively), predicted *in isolation* (*i.e.*, as uncoupled equilibria), under dilute (blue curves; (D), (d)), equimolar (black; (Q), (q)), and excess (red; (E), (e)) input, relative to the total concentration of each antitag species, $C_a^o = 1$ nM. Fig. 1[b] shows population-averaged values of the mean single-tag input error-response, $\langle \langle \log_{10} \epsilon_i \rangle \rangle$ for 10^4 randomly-encoded 100 strand TAT sets (strand-length, 20 bases), for dilute, equimolar, and excess inputs. Equilibrium constants were estimated using a mismatched zipper model implemented by *NucleicPark* [1], and parameters in [3]. Points and brackets denote mean values and standard deviations. The principal feature of each simulation is the concentration-dependence, which indicates a sharp transition to low-error behavior for dilute inputs, an effect not expected via consideration of the uncoupled equilibria, verifying the necessity of accounting for process coupling. Fig. 1[c] shows $\langle \log_{10} \epsilon_i \rangle$ for a custom-designed, 100 strand TAT system (not shown) evolved for minimized error-response, using a standard GA implemented by *NucleicPark*. The large improvement predicted over the mean encoding ($\sim 9\sigma$) suggests the utility of the model for design.

References

- [1] Rose, J.A., Hagiya, M., and Suyama, A., The Fidelity of the Tag-Antitag System 2: Reconciliation with the Stringency Picture, *Proc. Congress on Evolutionary Computation*, 2740–7, 2003.
- [2] Rose, J.A., Deaton, R.J., and Suyama, A., Statistical Thermodynamic Analysis and Design of DNA-based Computers, *Natural Computing*, *in press*, 2004.
- [3] SantaLucia, J. Jr. and Hicks, D., The Thermodynamics of DNA Structural Motifs, *Annu. Rev. Biophys. Biomolec. Struct.*, 33, 415–40, 2004.