Department of Energy Final Report for “Metabolic engineering of deinococcus radiodurans based on computational analysis and functional genomics” (DE-FG02-01ER63200)

PI: Jeremy Edwards

Objectives

The objective of our work is to develop novel computational tools to analyze the *Deinococcus radiodurans* DNA repair pathways and the influence of the metabolic flux distribution on DNA repair. These tools will be applied to provide insights for metabolic engineering of strains capable of growing under nutrient poor conditions similar to those found in mixed contaminant sites of interest to the DOE.

Over the entire grant period we accomplished all our specific aims and were also able to pursue new directions of research. Below, I will list the major accomplishments over the previous 3 years.

1. Performed Monte Carlo Simulations of RecA Mediated Pairing of Homologous DNA Molecules.
2. Developed a statistical approach to study the gene expression data from *D. radiodurans*. We have been studying the data from John Batista's.
3. Developed an expression profiling technology to generate very accurate and precise expression data. We followed up on results from John Batista's group using this approach.
4. Developed and put online a database for metabolic reconstructions.
5. We have developed and applied new Monte Carlo algorithms that are optimized for studying biological systems.
6. We developed a flux balance model for the *D. radiodurans* metabolic network.

In the next few pages I will summarize each of our major accomplishments. I will also attach copies of the papers for these works.

Approach & Results

Performed Monte Carlo Simulations of RecA Mediated Pairing of Homologous DNA Molecules.

Homologous recombination is a multi-step process orchestrated by DNA strand pairing and exchange enzymes, the *Escherichia coli* RecA protein being a classical example. The first step involves the alignment of homologous regions of the participating DNA molecules. The initial alignment is converted into a stable joint molecule followed by strand invasion, and finally, exchange. *In vitro* experiments intended for understanding the mechanism of homologous pairing monitor the formation of stable joint molecules, and therefore, do not allow the mechanism and determinants of the initial homologous alignment to be decoupled from those responsible for its conversion into a stable joint. In this study, the first step, namely, homologous alignment of DNA molecules, is investigated by performing Monte Carlo simulations. DNA molecules are modeled as bond-fluctuation chains localized on a three dimensional lattice with periodic boundaries. Heterologous and homologous interactions between RecA-coated single-stranded DNA (nucleoprotein filaments) and double-stranded DNA are incorporated into the simulation using the Metropolis algorithm. The simulation procedure is used to assess the effect of heterologous DNA and dilution on the rate of formation and yield of homologous alignments. The results show that the apparent first order kinetic behavior and the impact of heterologous DNA, reported in experimental literature, can be observed even if the conversion of the initially aligned molecules into a stable joint is not rate-determining. The present study is the first step towards developing rigorous computational models to describe the process of homologous recombination, and theoretical frameworks to retrieve biophysical parameters of strand pairing and exchange proteins from *in vitro* assays of joint molecule formation.

(See attached publication)

Developed a statistical approach to study the gene expression data from D. radiodurans. We have been studying the data from John Batista's.

We have developed an approach to study gene expression data that can attach 3 probabilities to each gene based on the microarray data. The probabilities are (i) probability the gene is upregulated, (ii) the probability the gene is down regulated and (iii) the probability the gene expression level is unchanged. We have looked at the data generated by John Batista's group at LSU in collaboration with TIGR. Our analysis has identified many genes that are significantly changed that were missed by other techniques. We have followed up on the leads with an accurate expression profiling technology.

(See attached manuscript)

We have another manuscript in preparation that looks directly at the results from *D. radiodurans*.

Developed an expression profiling technology to generate very accurate and precise expression data. We followed up on results from John Batista's group using this approach.

One of the primary goals of functional genomics is to provide a quantitative understanding of gene function. However, the success of this enterprise is dependent on the accuracy and precision of the functional genomic data upon which such understanding is based. Digital Analysis of Gene Expression (DAGE) described herein is a highly accurate and precise technology for measuring gene expression on an *absolute* scale by counting the number of transcripts of a gene being expressed at a given time. A rigorous statistical technique for analyzing the data has also been described. The result is a greatly improved technology sensitive enough for identifying and quantifying small (but biologically important) changes in gene expression. Fourteen genes involved in galactose metabolism in *S. cerevisiae* were analyzed for their expression levels in glucose and galactose minimal media. The quantitative expression results were characterized in terms of distributional and accuracy attributes; they were also in general agreement with corresponding results obtained using microarray technology.

We also analyzed 5 genes (DR0003, DR0070, DR0326, DR0423, DR2340) that were upregulated in *D. radiodurans* following ionizing radiation.

(Manuscript in review - See attached manuscript)

Developed and put online a database for metabolic reconstructions.

The Metabolic Reaction Analysis Database (MRAD) is a relational database based on the Entity-Relationship (ER) model which combines information about organisms, biochemical pathways, reactions, enzymes, substrates, products and genes. It describes 244596 genes in 79 organisms, 6552 enzymes, and 3552 reactions, 3100 substrates, 2866 products and 118 metabolic pathways. The MRAD graphical user interface allows for the identification of metabolic reactions which are similar and dissimilar in multiple organisms, reactions in a pathway which are missing in an organism and using any combination between one to six of the biological entities of organisms, genes, pathways, enzymes, substrates and products to determine metabolic reactions. MRAD provides a powerful and efficient tool for the construction of flux balance models for metabolic engineering applications.

(See attached manuscript)

We have used this database to construct an FBA model for *D. radiodurans*. (manuscript in preparation)

We have developed and applied new Monte Carlo algorithms that are optimized for studying biological systems.

Simulations of many biological processes require Monte Carlo methods that can account for the stochasticity due to a few copies of the molecules as well as spatial correlations. However, stochastic simulation of reaction-diffusion systems is computationally prohibitive. One of the major factors contributing to the computational cost arises from the low density of receptors in the membrane and intracellular molecules. We present an efficient coarse-grained hybrid null-event Monte Carlo algorithm which takes larger diffusion steps when the molecules are far apart and switches to small jumps when they are closer to capture the dynamics of interest due to interaction of molecules. Our algorithm is shown to predict the collision frequency of microscopic Monte Carlo simulations and achieving speed-up more than *five* orders of magnitude at low densities. Picking only the occupied sites results in major CPU savings, additionally coarse-graining results in an order of magnitude faster simulation.

(See attached manuscripts)

TIME ACCELERATED MONTE CARLO SIMULATIONS USING THE BINOMIAL τ -LEAP METHOD

Abhijit Chatterjee, Kapil Mayawala, Jeremy S. Edwards and Dionisios G. Vlachos*

Department of Chemical Engineering
University of Delaware, Newark, DE 19716, USA

ABSTRACT

Summary: Intracellular biochemical processes exhibit stochasticity that can potentially influence the phenotypic characteristics. However, stochastic simulations are prohibitive owing for most complex biological systems. The recently developed binomial τ -leap method (Chatterjee, Vlachos et al. 2005) is employed to carry out stochastic simulations of the epidermal growth factor (EGF) receptor induced mitogen activated protein (MAP) kinase cascade. Results indicate that the binomial τ -leap method is computationally 100-1000 times more efficient than the exact stochastic simulation algorithm of Gillespie. Furthermore, the new method avoids negative populations and accurately captures the disparity in populations of the reaction network along with their fluctuations.

Availability: Details about the binomial τ -leap algorithm, a software and manual are available at (Website).

Contact: vlachos@che.udel.edu

* To whom correspondence should be addressed.

INTRODUCTION

The role of stochasticity in biological systems is well established by several theoretical studies (Arkin, Ross et al. 1998; Morton-Firth and Bray 1998; Resat, Ewald et al. 2003; Meng, Somani et al. 2004) and experimental work (Elowitz, Levine et al. 2002). Stochasticity has been extensively studied using the exact stochastic simulation algorithm (SSA) of Gillespie (Gillespie 1976). However, SSA is limited to short times and to problems that do not exhibit a disparity in time scales. Biological networks exhibit multiple times scales, ranging from microseconds to days, and demand efficient stochastic simulation algorithms (Lok 2004). Such recently developed algorithms are reviewed in (Vlachos 2004; Turner, Schnell et al. 2004). The family of τ -leap methods, which have a mathematical foundation, is the only *coarse-grained* stochastic simulation method that can provide accuracy with nearly correct fluctuations.

Here the binomial τ -leap method (Chatterjee, Vlachos et al. 2005), a variant of the original Poisson τ -leap method of Gillespie (Gillespie 2001), is employed to study the signaling pathway of EGF receptor (EGFR) activated MAP kinase cascade. EGFRs belong to the receptor tyrosine kinase (RTK) family of receptors and play an important role in many physiological events in cell proliferation. In this work, we have used a recently proposed signaling pathway to study the surface and internalized EGFR activated MAP kinase cascade (Schoeberl, Eichler-Jonsson et al. 2002). The published model is based on ordinary differential equations (ODEs) of 94 signaling species among 296 reactions (referred to as events in this paper) in a well mixed environment. The complexity of the model renders the MAP kinase cascade an excellent benchmark problem for testing the applicability of the binomial τ -leap method for biological systems.

THE BINOMIAL τ -LEAP METHOD

The binomial τ -leap method follows the stochastic evolution of N species among M reactions as an approximate Markov process. The algorithm requires the initial population size of all species (i.e., the number of molecules) and the reaction kinetics as inputs. This information is used to compute transition probabilities per unit time of all events at time t . A ‘bundle or firings’ of events, sampled from a binomial distribution, are allowed to trigger in a time interval of size τ , and the time is incremented to $t+\tau$. This process is repeated as the algorithm leaps along the time axis. The maximum τ is chosen so it does not substantially change the populations (Gillespie 2001). This is automatically taken care of using a simple, inexpensive criterion (Chatterjee, Vlachos et al. 2005) according to which the user can control the speed-up by inputting the coarse-graining factor r of time stepping in comparison to the SSA. Previous experience indicated that $r \notin 0.2$ provides accuracy and speed-up.

The algorithm conserves mass at both the reaction and entire network levels by constraining the number of reaction firings, thereby eliminating for the first time negative populations in τ -leaping (Chatterjee, Vlachos et al. 2005). Numerical simulations using simple reaction networks (Chatterjee, Vlachos et al. 2005) demonstrated that the method is more accurate than the original Poisson τ -leap method of (Gillespie 2001).

RESULTS

Simulations spanning 15 minutes of real time were performed for the MAP kinase cascade using the binomial τ -leap method, with different levels of coarse-graining $r=0.1$ - 0.5 , and the SSA. Reaction dynamics and initial populations are specified in (Website). Intracellular populations are plotted against time in Figure 1 for selected species whose populations span from small to large ones. The binomial τ -leap method accurately captures the network kinetics for all times despite the large disparity in species population size in this complex network.

Figure 2 compares fluctuations of the two methods using the probability density functions obtained at time $t=1$ min (with ensemble averages over 500 trajectories for the SSA and 1000 for the binomial τ -leap method). The mean and standard deviation for representative species of small and large populations are tabulated in the inset of Figure 2. The excellent agreement obtained indicates that not only the correct mean and standard deviation but also nearly correct higher order moments can be obtained using the binomial τ -leap method.

The SSA required 3 days on a DEC alpha 833 MHz processor for the simulation depicted in Figure 1 as compared to 6–30 min taken using the τ -leap method. The binomial τ -leap is 100-1000 times more efficient than the SSA over the range of r values explored. The low computational requirements of the binomial τ -leap method enable accurate long time simulations that are beyond the reach of SSA (see example in (Website)). Due to its accuracy, robustness, simplicity and substantial speed-up, the binomial τ -leap method appears to be a particularly promising, approximate stochastic simulation method for computational biology.

ACKNOWLEDGEMENTS

This work was partially supported by the NSF through CTS-0312117.

REFERENCES

- Arkin, A. P., J. Ross, et al. (1998). "Stochastic kinetic analysis of developmental pathway bifurcation in phage λ -infected escherichia coli cells." *Genetics* **149**: 1633-1648.
- Chatterjee, A., D. G. Vlachos, et al. (2005). "Binomial distribution based τ -leap accelerated stochastic simulation." *Journal of Chemical Physics* **In press**.
- Elowitz, M. B., A. J. Levine, et al. (2002). "Stochastic gene expression in a single cell." *Science* **297**(16): 1183-1186.
- Gillespie, D. T. (1976). "A general method for numerically simulating the stochastic evolution of coupled chemical reactions." *Journal of Computational Physics* **22**: 403-434.
- Gillespie, D. T. (2001). "Approximate accelerated stochastic simulation of chemically reacting systems." *Journal of Chemical Physics* **115**(4): 1716-1733.
- Lok, L. (2004). "The need for speed in stochastic simulation." **22**(8): 964-965.
- Meng, T. C., S. Somani, et al. (2004). "Modeling and simulation of biological systems with stochasticity." *In Silico Biology* **4**: 0024.
- Morton-Firth, C. J. and D. Bray (1998). "Predicting Temporal Fluctuations in an Intracellular Signalling Pathway." *Journal of Theoretical Biology* **192**(1): 117-128.

- Resat, H., J. A. Ewald, et al. (2003). "An integrated model of epidermal growth factor receptor trafficking and signal transduction." Biophysical Journal **85**: 730-743.
- Schoeberl, B., C. Eichler-Jonsson, et al. (2002). "Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized receptors." Nature Biotechnology **20**: 370-375.
- Turner, T. E., S. Schnell, et al. (2004). "Stochastic approaches for modelling in vivo reactions." Computational Biology and Chemistry **28**: 165-178.
- Vlachos, D. G. (2004). "A review of multiscale analysis: Examples from systems biology, materials engineering, and other fluid-surface interacting systems." Adv. Chem. Eng.: submitted.
- Website "<http://www.che.udel.edu/>."

FIGURE CAPTIONS

Fig. 1 Time dependent populations using the stochastic simulation algorithm (SSA; thick solid line) and the binomial τ -leap method (thin solid line, $r=0.2$ and squares, $r=0.5$).

Fig. 2 Probability density functions (pdf) using the stochastic simulation algorithm (SSA, circles) and the binomial τ -leap method at time $t = 1$ min (squares, $r=0.2$; diamonds, $r=0.5$). The solid lines are curve fits for visual aid. Average populations and standard deviation are tabulated in the inset.

Figure 2

Heterogeneities in EGF Receptor Density at the Cell Surface can lead to Concave up Scatchard Plot of EGF binding

Kapil Mayawala, Dionisios G. Vlachos, Jeremy S. Edwards

Department of Chemical Engineering, University of Delaware, Newark, DE, USA

ABSTRACT The mechanism responsible for the concave up nature of the Scatchard plot of epidermal growth factor (EGF) binding on EGF receptor (EGFR) has been a controversial issue for more than a decade. Past efforts to mechanistically simulate the concave up nature of the Scatchard plot of EGF binding have shown that negative cooperativity in EGF binding on EGFR dimer and inclusion of external sites can describe this behavior. However, herein we show that heterogeneity in the density of EGFR on the cell surface, which has been experimentally reported, can also lead to concave up shape of the Scatchard plot of the EGF binding on EGFR.

Received for publication "Staff will complete" and in final form "Staff will complete"

Corresponding author: Jeremy S. Edwards. Tel: 302-831-8072. Fax: 302-831-1048. Email: edwards@che.udel.edu

The epidermal growth factor (EGF) receptor (EGFR) belongs to the ErbB family of receptor tyrosine kinases. These receptors trigger a rich network of signaling pathways and regulate cell functions such as proliferation, differentiation and migration (Yarden and Sliwkowski 2001). Since EGF binding represents the initial step for activating EGFR, considerable work has been devoted to elucidating the mechanisms of ligand binding and dimerization.

The Scatchard method has been extensively used to analyze the experimental data of equilibrium EGF binding on EGFR. In this method, the ratio of bound (B) to free ligand (L) concentrations is plotted as a function of bound ligand concentration, which results in a linear relation with a slope of $-K_a$ (association constant) and an x-intercept as R_T (concentration of total binding sites)

$$\frac{B}{L} = -K_a(B) + K_a(R_T).$$

However, experimentally reported Scatchard plot data of EGF binding on EGFR in intact cells exhibits a concave up shape (Klein and others 2004; Lemmon and others 1997; Wofsy and others 1992). The mechanism responsible for the concave up nature of

the Scatchard plot has been a controversial issue. The mathematically generated Scatchard plots that take into account experimentally reported differences in receptor affinity (monomeric receptors represent the low affinity population and dimeric or higher oligomeric receptors represent the high affinity population), i.e., $K_4 > K_5$ in Fig. 1, and EGF induced EGFR dimerization (and higher-mers formation) show concave down curvature (Klein and others 2004; Lemmon and others 1997; Wofsy and others 1992).

Wofsy *et al.* (Wofsy and others 1992) proposed that a *necessary* condition for the concave up Scatchard plot of EGF binding is that EGF binding to aggregated receptors should be negatively cooperative, i.e., ligand binding of a receptor to a dimer having one receptor already bound occurs with lower affinity than the first binding event (namely, $K_5 > K_6$ in Fig. 1). Chamberlin *et al.* (Chamberlin and Davies 1998) and Wofsy and Goldstein (Wofsy and Goldstein 1992), using this negative cooperativity, further showed that the receptor density is also an important aspect in determining the shape of the Scatchard plot. A concave up Scatchard plot has also been generated by considering the existence of molecules that trap a

small fraction of the receptors in the high-affinity state (Holbrook and others 2000).

K_1	400 M^{-1}	In the range suggested by (Lemmon and others 1997)
K_2	$1.2 \times 10^5 \text{ M}^{-1}$	Calculated based on equilibrium relations given in (Wofsy and others 1992)
K_3	$3.6 \times 10^7 \text{ M}^{-1}$	Calculated based on equilibrium relations given in (Wofsy and others 1992)
K_4	$4 \times 10^8 \text{ M}^{-1}$	In the range suggested by (Jorissen and others 2003; Kawamoto and others 1983; Lin and others 1986; Schlessinger 1986; Schlessinger 1988)
K_5, K_6	$1.2 \times 10^{11} \text{ M}^{-1}$	In the range suggested by (Jorissen and others 2003; Lin and others 1986; Schlessinger 1986; Schlessinger 1988)

Figure 1: Reactions considered in our model.

In a mathematical model based on structural studies, Klein *et al.* (Klein and others 2004) suggested that binding of ligand to a receptor dimer is independent of whether the other EGF binding site in the dimer is occupied. That study concluded that concave up plots can *only* be obtained by including an extra binding event, in which occupied EGFR dimers bind to an external site.

In this work, we show that heterogeneity in the surface density of EGFR on A-431 cells (total number of receptors $\sim 1.4 \times 10^6$) can lead to a concave up Scatchard plot without considering either negative cooperativity or inclusion of external sites. The model is shown in Fig. 1 along with the equilibrium constants (Lemmon and others 1997; Wofsy and

others 1992). We assume that there are two types of sites on the cell surface: low affinity sites and high affinity sites. This choice is motivated by strong evidence based on biochemical (see review by (Jorissen and others 2003)) and microscopy (Sako and others 2000) studies, which suggest that monomeric EGFR represents low affinity sites and dimerized EGFR exhibits high affinity sites. Furthermore, experimental evidence suggests localization of the EGFR in certain regions (lipid rafts and/or caveolae) of the plasma membrane (Jorissen and others 2003; Mineo and others 1999; Ringerike and others 2002; Waugh and others 1999). According to an experimental evidence, more than half of the unstimulated EGFRs on the cell surface are concentrated in ~ 5 -10% of the plasma membrane (Jorissen and others 2003; Mineo and others 1999). Herein, we consider a heterogeneous density model consisting of low and high density regions with several extents of localization in the suggested range. The binding at a certain density of receptors is calculated as described in (Klein and others 2004). Hence, the total bound ligand (B) is calculated as the sum of the binding, in terms of number of bound ligand molecules, in low density (Low-D) and high density (High-D) regions, i.e., $B = B_{\text{High-D}} + B_{\text{Low-D}}$. The concentration of bound ligand in each region is $RL + R_2L + 2R_2L_2$, where the amount of localized monomer (R) and dimerized (R_2) receptors depends *non-linearly* on the local density of receptors and free ligand concentration (L) (supplemental information). Hence, heterogeneities in receptor density change the amount of ligand binding.

Fig. 2 compares the results of our heterogeneous density model with a spatially homogeneous density model in terms of the Scatchard plot for several values of receptor heterogeneities. As can be observed in Fig. 2, heterogeneities in the density of the surface receptors can strongly modulate the curvature of the Scatchard plot and give rise to the concave up shape of the Scatchard plot of the EGF binding. The upper part of the curvature is generated by the binding in the high density regions and the lower part by the binding in low density regions of the cell surface. While the analysis was done for A-431 cells, the general approach of this study shows that heterogeneities in

the EGFR surface density can also be a potential factor controlling the shape of the Scatchard plot for other systems as well.

Figure 2: Scatchard plot of the EGF binding on the EGFR in A-431 cells with homogeneous and heterogeneous surface density of receptors.

SUPPLEMENTARY MATERIAL

An online supplement to this article can be found by visiting BJ Online at <http://www.biophysj.org>.

ACKNOWLEDGEMENTS

This work was supported by grant from the US Department of Energy and the National Science Foundation.

REFERENCES

- (1) Chamberlin SG, Davies DE. 1998. A unified model of c-erbB receptor homo- and heterodimerisation. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* 1384(2):223-232.
- (2) Holbrook MR, Slakey LL, Gross DJ. 2000. Thermodynamic mixing of molecular states of the epidermal growth factor receptor modulates macroscopic ligand binding affinity. *Biochemical Journal* 352:99-108.
- (3) Jorissen RN, Walker F, Pouliot N, Garrett TPJ, Ward CW, Burgess AW. 2003. Epidermal growth factor receptor: mechanisms of activation and signalling. *Experimental Cell Research* 284(1):31-53.
- (4) Kawamoto T, Sato JD, Le A, Polikoff J, Sato GH, Mendelsohn J. 1983. Growth stimulation of A431 cells by epidermal growth factor: identification of high-affinity receptors for epidermal growth factor by an anti-receptor monoclonal antibody. *PNAS* 80:1337-1341.
- (5) Klein P, Mattoon D, Lemmon MA, Schlessinger J. 2004. A structure-based model for ligand binding and dimerization of EGF receptors. *PNAS* 101(4):929-934.
- (6) Lemmon MA, Bu Z, Ladbury JE, Zhou M, Pinchasi D, Lax I, Engelman DM, Schlessinger J. 1997. Two EGF molecules contribute additively to stabilization of the EGFR dimer. *The EMBO Journal* 16:281-294.
- (7) Lin CR, Chen WS, Lazar CS, Carpenter CD, Gill GN, Evans RM, Rosenfeld MG. 1986. Protein kinase C phosphorylation at Thr 654 of the unoccupied EGF receptor and EGF binding regulate functional receptor loss by independent mechanisms. *Cell* 44(6):839-848.
- (8) Mineo C, Gill GN, Anderson RGW. 1999. Regulated Migration of Epidermal Growth Factor Receptor from Caveolae. *J. Biol. Chem.* 274(43):30636-30643.
- (9) Ringerike T, Blystad FD, Levy FO, Madhus IH, Stang E. 2002. Cholesterol is important in control of EGF receptor kinase activity but EGF receptors are not concentrated in caveolae. *J Cell Sci* 115(6):1331-1340.
- (10) Sako Y, Minoghchi S, Yanagida T. 2000. Single-molecule imaging of EGFR signalling on the surface of living cells. *Nature Cell Biology* 2(3):168-172.
- (11) Schlessinger J. 1986. Allosteric regulation of the epidermal growth factor receptor kinase. *J. Cell Biol.* 103(6):2067-2072.
- (12) Schlessinger J. 1988. The epidermal growth factor receptor as a multifunctional allosteric protein. *Biochemistry* 27:3119-3123.
- (13) Waugh MG, Lawson D, Hsuan JJ. 1999. Epidermal growth factor receptor activation is localized within low-buoyant density, non-caveolar membrane domains. *Biochemical Journal* 337:591-597.
- (14) Wofsy C, Goldstein B. 1992. Interpretation of Scatchard plots for aggregating receptor systems. *Mathematical Biosciences* 112(1):115-154.
- (15) Wofsy C, Goldstein B, Lund K, Wiley H. 1992. Implications of epidermal growth factor (EGF) induced egf receptor aggregation. *Biophys. J.* 63(1):98-110.
- (16) Yarden Y, Sliwkowski MX. 2001. Untangling the ErbB signalling network. *Nature Reviews Molecular Cell Biology* 2:127-137.

Fundamental probability models for microarray data distributions

*Babatunde A. Ogunnaike¹, Claudio A. Gelmi
and Jeremy S. Edwards*

Department of Chemical Engineering
and Center for Systems Biology
Delaware Biotechnology Institute
University of Delaware
Newark, DE, 19707

ABSTRACT

Motivation: Gene expression studies — as currently done with printed arrays or pre-fabricated gene chips — generate large quantities of data that are used to determine the *relative* expression levels of thousands of genes in a cell. The most compelling characteristic of these data sets is that the number of genes whose expression profiles are to be determined exceeds the number of replicates by several orders of magnitude. Standard spot-by-spot analysis seeks to extract useful information for each gene on the basis of the number of replicates available for the specific gene in question. As has become increasingly clear, this plays to the weakness rather than the strength of microarrays. On the other hand, by virtue of the sheer data volume alone, treating the entire data set as an ensemble, and developing fundamental theoretical distributions to represent these ensembles should provide us with a framework for efficient extraction of gene expression information that plays to the strength of microarrays. Relatively little attention has been paid to studying distributions of complete microarray data sets (see for example, Hoyle et al., 2002); and virtually all of the published studies are empirical approximations fitted to observed data.

Results: We present theoretical results that confirm what had previously been presented as speculations (e.g. Sebastiani, et al., 2003), or had been assumed for convenience (Newton, et al., 2001): that under very reasonable assumptions, the distribution of microarray intensities should follow the Gamma (not lognormal) model. Furthermore, the biological interpretations of the model parameters emerge naturally

from our derivation. We subsequently establish that a polar coordinate transformation of raw intensity data provides the basis for a technique in which each microarray data set is represented as a mixture of Beta densities for the fractional intensities (not intensity ratios), from which rigorous statistical inference may be drawn regarding differential gene expression. We illustrate the results and their application with experimental data from gene expression studies on *Deinococcus radiodurans* following DNA damage using cDNA microarrays.

Contact: ogunnaike@che.udel.edu

INTRODUCTION

Standard statistical analysis involves estimating (and drawing statistical inference about) p parameters using a data sample of size n , typically with $n \gg p$. The precision of the estimates, as well as the “power” of the accompanying statistical tests of significance, are well-known to improve with increasing n . The converse is also true: reducing sample size n reduces estimation precision and the power of statistical hypotheses (other things being equal). The problem with microarray data analysis is well-documented: p (in this context the expression levels of thousands of genes to be estimated) is very large, while n , the number of independent replicates, from which these estimates are to be determined, is quite small. Many techniques have been proposed for extracting information about gene expression levels from microarray data, but this intrinsic characteristic of the data set (large p , small n) remains a major impediment (cf. Nadon and Shoemaker, 2002).

To complicate matters further, genes are known to operate, not independently as individuals, but in networks, suggesting that microarray data sets represent the collective behavior of a population best studied jointly. Still, the inference problem of primary concern is the *quantitative* determination of the differential expression level of individual genes represented in the data set: which ones are over-expressed, which are under-expressed — and by how much — and which are unchanged. None of the preceding factors change this primary focus, but they motivate us to consider analytical techniques that take into con-

¹To whom correspondence should be addressed

sideration the intrinsic characteristics of microarray data we have highlighted here. Our perspective in this paper is that by examining the entire data set as an ensemble and characterizing it as such we are able to develop an analysis technique that actually plays to the strength of microarray technology.

The development of techniques for spot-by-spot analysis of microarray data has been the subject of much research effort; and many of the key results and the still unresolved issues have been discussed in several recent reviews (for example, Quackenbush, 2001; Nadon and Shoemaker, 2002; Sebastiani et al., 2003).

Far fewer attempts have been aimed at developing methods for distributional characterization of microarray data so that the data sets may be treated in a collective fashion rather than as unrelated individual gene-by-gene measurements. These efforts have involved empirical fits of various probability densities to intensity ratios (Ghosh and Chinnaiyan, 2002), or to raw intensities (Steinboff et al., 2003; Hoyle et al., 2002); little or no mechanistic interpretations of the resulting distributions are possible, and the issue of how these distributional representations may be used for determining differential expression levels were not addressed.

The objective of this paper is to present “first principles” theoretical (as opposed to merely empirical) results for representing and characterizing microarray data as an ensemble. The theoretical results and their application are illustrated with an example data set. Ultimately, these results will be assembled into a rigorous theoretical framework for (i) appropriately analyzing microarray data sets on the basis of this ensemble characterization; and (ii) drawing realistic inferences from the analysis. However, because of space constraints, these latter aspects are discussed in a subsequent paper.

A THEORETICAL MODEL OF MICROARRAY MEASUREMENTS

Intensity Measurements and Gene Expression levels: For each gene represented by a spot on a microarray, the expression level λ , corresponding to the number of cDNA molecules, is considered fixed,

but unknown. It is estimated by measurements of fluorescent dye intensity.

For a fixed expression level, repeated fluorescent signal intensity measurements taken n separate and independent times, $I_j; j = 1, 2, \dots, n$, will differ because of inherent variability. For fluorescent intensity measurements acquired from cDNA microarrays, it is typical to represent this fact as follows (Sebastiani et al., 2003):

$$I_j = a_j \lambda \quad (1)$$

where the uncertainty in the measurement is considered as having arisen primarily from a_j being a random proportionality constant, characteristic of the measurement device, relating the number of cDNA molecules to the observed intensity. In reality, it is more accurate to represent the *raw* intensity signal as:

$$I'_j = a_j \lambda + b_j \quad (2)$$

where b_j is another random component corresponding to “background noise”; i.e. the intensity measurement values one would obtain even when there are no cDNA molecules present on the spot (Rocke and Durbin, 2001). Because of the non-negativity of intensity measurements, b_j has a non-zero mean value that is typically “subtracted off” to obtain the so-called “background-corrected” intensity measurement. Different technologies employ different techniques for such background correction. We note here that for high and medium intensity measurements, the contributions from b_j are relatively insignificant; appropriately accounting for the presence of this component is critical however at very low intensity measurements.

In what follows, we assume that the intensity signal has been appropriately “background corrected” so that Equation (1) applies. Nevertheless, we will still pay due respect to the intensity magnitude.

Model Development: Let I represent the observed fluorescent intensity corresponding to a total of z cDNA molecules; furthermore, we consider this intensity to be cumulative in the following sense:

$$I = I^{(1)} + I^{(2)} + \dots + I^{(z)} \quad (3)$$

where $I^{(m)}$ is the intensity contribution from the m^{th} molecule. We now refer to $z(I)$ as the number of

molecules corresponding to the observed (background-corrected) *total* intensity measurement I . From this perspective, for a given observed I , z is an unknown random variable, an integer count.

The complementary element in this model development is the fluorescent intensity measurement obtained from k molecules, which, for purposes of this derivation, we will refer to as $y(k)$. From this intensity measurement perspective, given a fixed number k , y is an unknown (continuous) random variable.

Finally, let β represent the “mean intensity observed per molecule”, or conversely, its reciprocal, $\eta = 1/\beta$, which is the mean number of molecules per unit observed intensity. These quantities, characteristic of the measurement device (akin to a calibration factor), can be used to characterize the random behavior of both the continuous y and the discrete z and relate one to the other as we now show.

The problem at hand is to develop a probability model to characterize y as a function of k , the number of molecules responsible for the observed intensity signal measurement. We adopt a 2-step strategy: (i) begin with the complementary problem of characterizing $z(I)$, the discrete count of the number of molecules corresponding to the observed intensity, and (ii) convert the integer count to the continuous variable, y .

1. Characterizing $z(I)$. First-principles analysis and experimental confirmation (Thattai and Oudenaarden, 2001; Ozbudak, *et al.*, 2002) have established that the steady state statistical properties of mRNA expression levels results in a Poisson distribution for z , i.e.

$$f(z) = e^{-\theta} \frac{\theta^z}{z!} \quad (4)$$

where θ is the mean number of molecules. Within the context of our current formulation, in relation to the observed intensity I , and given the characteristic parameter η as described above, Eq (4) becomes:

$$f(z) = e^{-(\eta I)} \frac{(\eta I)^z}{z!} \quad (5)$$

which is now the probability model for characterizing z , the number of mRNA molecules in terms of I , the corresponding observed total intensity.

Remarks:

1. This provides an expression for computing the probability that z , the number of molecules responsible for generating the observed intensity I , takes on the integer values $0, 1, 2, \dots, \infty$ given the “population parameter” η ;
2. From this equation (Eq (5)) the theoretical expected value of z , $E(z) = \mu = \eta I$.

2. Convert count variable $z(I)$ to continuous variable y . Let $y(k)$ be the total observed intensity for k molecules, i.e.

$$y(k) = I^{(1)} + I^{(2)} + \dots + I^{(k)} \quad (6)$$

As we show in Appendix A, the probability distribution function for y given the probability distribution function for $z(I)$ is:

$$f(y; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-y/\beta} y^{(\alpha-1)} \quad (7)$$

with the following implication:

The total fluorescent intensity measurement for each spot on the microarray possesses a Gamma distribution with the *shape* parameter α related directly to the number of cDNA molecules responsible for the observed intensity, y ; the *scale* parameter β is the mean intensity per molecule.

It is important to note that Newton et al., 2001, invoked this model, strictly for convenience, but with a constant shape parameter. Also, Sebastiani et al., 2003, alluded to the fact that the Gamma distribution may be a more appropriate model for intensity data since a logarithmic transformation still leaves the data with unsatisfactory asymmetry.

A THEORETICAL MODEL OF DIFFERENTIAL EXPRESSION

At each spot on the microarray, the expression level of each represented gene is evaluated under two different conditions: 1 = “Test” conditions, labelled with Cy3 dye; 0 = “Reference” conditions, with Cy5 dye.

Let I_{i_1} and I_{i_0} be the fluorescent signal intensities measured from each spot i on a microarray, with I_{i_1} obtained from the gene in question under test conditions, and I_{i_0} from the gene under control conditions. Let λ_{i_1} and λ_{i_0} respectively represent the corresponding actual, but unknown, expression levels of the gene in question.

Then from Equation (1), we have:

$$I_{i_1} = a_{i_1} \lambda_{i_1} \quad (8)$$

$$I_{i_0} = a_{i_0} \lambda_{i_0} \quad (9)$$

For various well-documented reasons (Dudley et al., 2002; Sebastiani, et al., 2003) it is a near-impossible task to determine the absolute value of the expression levels $\lambda_{i_1}, \lambda_{i_0}$ strictly from the observed microarray intensity measurements. Even so, determining by how much λ_{i_1} differs from λ_{i_0} without necessarily knowing the absolute value of each number is still considered extremely valuable information about the gene in question. While methods for determining “absolute” gene expression levels are being perfected for lower cost and higher-throughput (for example, Brenner et al., 2000; Mikkilineni et al., 2004) “differential expression” determination via standard microarray technology remains an important aspect of gene expression studies.

The customary measure of differential expression is the so-called “fold change” statistic defined as the ratio

$$\rho_i = \frac{\lambda_{i_1}}{\lambda_{i_0}} \quad (10)$$

and the common practice is to base inference about ρ_i on the measured intensity signal ratio r_i defined as

$$r_i = I_{i_1} / I_{i_0} \quad (11)$$

It is important to observe from here however, that such a strategy is based on the (tacit) assumption that for both the green signal intensity as the with the red, the multiplicative factors are approximately the same. This may not be the case in general, especially since we have just shown that these multiplicative constants are related directly to the “mean intensity observed per molecule” a number that is clearly different for green and red fluorescent labels. (This is

Figure 1: Raw data from gene expression studies in *Deinococcus radiodurans* following DNA damage. Note the “conical” sector shape.

potentially one source of the well-documented “dye bias”.) Dye-swaps and data normalization are some of the tactics typically employed to minimize these problems.

Also, observe that if σ_a^2 is the variance associated with the random multipliers, then the observed variance in the signal intensity will be given by:

$$\sigma_I^2 = \lambda^2 \sigma_a^2 \quad (12)$$

so that the variance associated with signal intensity increases quadratically with expression level (and hence with intensity); or equivalently, the standard deviation varies linearly with intensity. This is in agreement with actual microarray data (Rocke and Durbin, 2001). For example, Figure 1 shows a plot of experimental data from gene expression studies in *Deinococcus radiodurans* following DNA damage (Battista, 2003) where the data variance is observed to increase with increasing intensity values. The implication for data analysis is that the statistical character of microarray measurements (especially measurement variance) is dependent on signal magnitude (Chen et al., 1997 and Newton et al., 2003).

The most commonly employed strategy for addressing the preceding issue is a logarithmic trans-

formation of the data to obtain:

$$\log I_{i_0} = \log a_{i_0} + \log \lambda_{i_0} \quad (13)$$

turning the expression into the more familiar additive error form:

$$Y_i = X_i + \epsilon_i \quad (14)$$

In most cases, it is now assumed that $\log a_{i_0}$ is normally distributed, and the analysis proceeds along these lines. However, it has been shown by various investigators (Rocke and Durbin, 2001; Sebastiani et al., 2003) that this assumption, even though convenient, is not necessarily valid. It would be valid if a lognormal density is an appropriate model for the intensity distribution. However, we have just established that the Gamma (not the lognormal) density is more appropriate.

An alternative transformation: The fan-shaped characteristics of raw cDNA microarray intensity data (illustrated by the example in Fig 1) suggests a strategy where the original data in the form of the ordered pair of *measured* signal intensities (y_{i_1}, y_{i_0}) for each spot i is converted from this “cartesian” representation to the corresponding “polar” form (R_i, x_i) where R_i , the vector magnitude defined by:

$$R_i = \sqrt{y_{i_1}^2 + y_{i_0}^2} \quad (15)$$

is the “intensity magnitude”, and x_i , the “fractional intensity” defined by:

$$x_i = \frac{y_{i_1}}{y_{i_1} + y_{i_0}} \quad (16)$$

is related to the vector angle.

The following are important attributes of this coordinate transformation:

1. It naturally compensates for the inherent heteroskedasticity of microarray data since pie-shaped regions in cartesian coordinates naturally map into rectangular regions in polar coordinates. (See Fig 2 which shows the effect of this transformation on the data in Fig 1);
2. It normalizes the data efficiently without losing any of the original information. Observe that

Figure 2: Raw data of Figure 1 after the “polar coordinates transformation”.

while x_i is dimensionless and naturally scaled between 0 and 1, R_i retains the information about intensity magnitude, thus preserving the two-dimensional character of the original data;

3. The characteristics of x_i are such that values less than 0.5 indicate down regulation; $x_i = 0.5$ indicates no change and $x_i > 0.5$ indicates up-regulation.

Even though the traditional “intensity ratio” r_i is related to x_i according to the expression:

$$r_i = \frac{y_{i_1}}{y_{i_0}} = \frac{x_i}{1 - x_i} \quad (17)$$

we recommend x_i as a more convenient variable on which to base inference about differential expression because it enables statistical rigor and endows the problem with analytical tractability. As we will now show, a probabilistic model can be derived for x_i from the model we presented earlier for raw intensity measurements, while R_i provides a rational metric for partitioning the data, (allowing us to separate low magnitude data from higher magnitude data) for more precise analysis.

Probability model for fractional intensity: If we represent as y_{i_1} , the *measured* intensities, corre-

sponding to gene i (on spot i) under “test conditions”, and y_{i_0} , the value obtained for the same gene i under “reference conditions”, then the implications of the results given above are that:

$$y_{i_1} \sim \text{Gamma}(\alpha_{i_1}, \beta_1) \quad (18)$$

$$y_{i_0} \sim \text{Gamma}(\alpha_{i_0}, \beta_0) \quad (19)$$

allowing the “equipment calibration” parameter (mean intensity per molecule) β to be dye dependent, but not gene dependent, so that we need only retain the subscripts 0 and 1 but not i . A well-known result from probability theory (e.g. Evans et al, 1993) is that if Eqs(18) and (19) are true then

$$y_{i_1}/\beta_1 \sim \text{Gamma}(\alpha_{i_1}, 1) \quad (20)$$

$$y_{i_0}/\beta_0 \sim \text{Gamma}(\alpha_{i_0}, 1) \quad (21)$$

and the fractional variable, x_i defined by

$$\begin{aligned} x_i &= \frac{y_{i_1}/\beta_1}{y_{i_1}/\beta_1 + y_{i_0}/\beta_0} \\ &= \frac{y_{i_1}}{y_{i_1} + w y_{i_0}} \end{aligned} \quad (22)$$

where

$$w = \frac{\beta_1}{\beta_0} \quad (23)$$

possesses a Beta distribution, $\text{Beta}(\alpha_{j_1}, \alpha_{j_0})$, with a probability density function given by:

$$f(x_i) = \frac{\Gamma(\alpha_{i_0} + \alpha_{i_1})}{\Gamma(\alpha_{i_1})\Gamma(\alpha_{i_0})} x^{\alpha_{i_1}-1} (1-x)^{\alpha_{i_0}-1} \quad (24)$$

Before discussing what this central result means for drawing statistical inference about differential expression, there are some important consequences for *practical* data analysis that we must first discuss.

Nominal Data Transformation: A subtle but important point about Equations (22) and (23) is that raw data consists of y_{i_1} and y_{i_0} pairs only; the parameters β_1 and β_0 (which we have established from our theoretical derivation as representing the “mean intensity per molecule”) are unknown “calibration parameters”. In fact this latter point is the very reason that absolute estimates of expression levels are not possible from microarray data. As noted earlier, the

rationale for using the intensity ratios as the basis for drawing inference about differential expression is predicated on the tacit assumption that:

$$\beta_0 = \beta_1 \quad (25)$$

We will designate these customary assumptions as representing what we refer to as “nominal” conditions. Note that in reality these conditions may in fact not hold. The implications of departure from these nominal conditions, how to detect and correct for such departures will be discussed shortly.

For now, note that under these conditions, $w = 1$ and Eq (22) indicates that the “nominal” fractional intensity defined by:

$$x_i = \frac{y_{i_1}}{y_{i_1} + y_{i_0}} \quad (26)$$

obtained directly from raw data, possesses a Beta distribution, so that for each gene, the probability model is as shown in Eq. (24).

When nominal conditions do not hold, then in reality the computed fractional intensity x_i ought to be “corrected” to obtain $\tilde{x}_i$ where

$$\tilde{x}_i = \frac{y_{i_1}}{y_{i_1} + w y_{i_0}} \quad (27)$$

and the relationship between this “corrected” variable and the nominal fractional intensity x_i is easily established as the following dual expression:

$$\tilde{x}_i = \frac{x_i}{x_i + w(1-x_i)} \quad (28)$$

and

$$x_i = \frac{w \tilde{x}_i}{w \tilde{x}_i + (1 - \tilde{x}_i)} \quad (29)$$

Detecting and Correcting for departure from nominal conditions: When nominal conditions do not hold, (i.e. $w \neq 1$) it is necessary to estimate w , and this is done as follows.

Under conditions when there is *no change* in gene expression, the expected value of the *nominal* fractional intensity data, $E(X) = \mu_0$, should be theoretically equal 0.5. However, when $w \neq 1$ in (22), even though there is no change in gene expression, $E(X) \neq 0.5$, and hence any significant deviation of

estimates of μ_0 from 0.5 provides an indication that nominal conditions do not hold. Given a value of $E(X) = \mu_0 \neq 0.5$, we are able to use Eq (29) to estimate the value of w required so that the expectation of the corrected intensity $\tilde{x}$ is 0.5. It is easy to show that under these conditions,

$$\mu_0 = \frac{w}{w+1} \quad (30)$$

so that

$$w = \frac{\mu_0}{1-\mu_0} \quad (31)$$

Thus, from an estimate of the expected value of the nominal x for those genes showing no change in gene expression, we are able (i) to determine if nominal conditions hold and then (ii) correct the nominal fractional intensity for the entire data set using the estimate of w given above when nominal conditions do not hold.

It remains to show how to determine μ_0 in practice, and in general apply the entire foregoing set of results for microarray data analysis.

APPLICATION TO MICROARRAY DATA ANALYSIS

The theoretical implications of the results presented in Eqs. (26) and (24) above are as follows: For a particular microarray dataset involving a total of N genes, each nominal fractional intensity $x_i, i = 1, 2, \dots, N$ is a random sample from its own (possibly unique) Beta distribution. Consequently, $f(x)$, the underlying theoretical distribution of (nominal) spot fractional intensities for the entire collection of genes, will be a mixture of overlapping Beta density functions consisting, *in principle*, of as many Beta densities as there are genes.

Because of the finite support $[0, 1]$ over which $f(x)$ is defined and the large number of genes involved, but most importantly as a result of the necessary “coalescence” of the distributions of genes with theoretically identical—or statistically similar—expression levels, in practice, the actual number of truly discernible Beta densities required to describe any specific collection of microarray data will be relatively few. Practically therefore, this suggests a characterization strat-

egy in which the histogram of the (nominal) fractional intensities is modelled as a mixture of just a few Beta densities.

In particular, the distributions of all genes showing *no* differential expression will coalesce into one clearly discernible distribution, $f_0(x)$. If, at the barest minimum, the distributions of genes showing lower differential expression also coalesce into a single discernible distribution and the same happens for the genes showing higher differential expression, we may thus characterize the histogram of fractional intensity data (at the barest minimum) as a mixture of (at least) three Beta distribution functions, $f_1(x)$ for the data from the collection of genes showing lower differential expression, $f_0(x)$ for those showing no differential expression, and $f_2(x)$ for those showing higher differential expression, i.e.

$$f(x) = \phi_1 f_1(x) + \phi_0 f_0(x) + \phi_2 f_2(x) \quad (32)$$

(with $\phi_0 + \phi_1 + \phi_2 = 1$).

Thus by fitting a mixture of Beta distributions to histogram data, we obtain a theoretical probabilistic characterization of the data where each contributing Beta distribution constitutes the population description of a category of genes with common differential expression attributes. Observe that the coefficients ϕ_0, ϕ_1 and ϕ_2 (which must equal $(1 - \phi_0 - \phi_1)$) indicate the fraction of the entire population that belongs in each respective indicated category. In particular, for the fraction ϕ_0 of genes for which there is truly no differential expression, x possesses a Beta distribution that is perfectly symmetric, with mean = mode = median = $x^* = 0.5$.

Data Analysis Procedure: We may now assemble these results into the following procedure for analyzing microarray data.

1. **Data Representation:** Convert raw data from cartesian representation (y_{i_1}, y_{i_0}) to polar form (R_i, x_i) for each gene i ;
2. **Data Partitioning:** Use R , the intensity magnitude, to partition the data into groups. (We recommend 3 groups corresponding to “Low”, “Medium”, and “High” intensity magnitude values, with each group containing an equal number of data points.)

3. **Data Characterization:** For each group, characterize the histogram of x as a mixture of Beta densities.
4. **Bias Detection and Correction:** On the basis of the characteristics of the Beta density $f_0(x)$, determine whether or not nominal conditions hold. If μ_0 , the mean value of $f_0(x)$ is not 0.5, then the nominal fractional intensity data is biased and must be corrected; estimate the correction factor w according to (31) and obtain corrected fractional intensity data $\tilde{x}_i$ according to (28).

The result will be a corrected probability model $f(x)$ as a mixture of Beta densities that may be used for statistical inference concerning each gene in the array.

How to carry out statistical inference using this probability model with and without replicate data will be the subject of a subsequent paper; for now, we illustrate the application of the theoretical results we have presented thus far with an example.

ILLUSTRATIVE EXAMPLE

The data set shown first in Fig 1 is from studies of ~ 3000 genes from *D. radiodurans* 0.5 hours after ionizing radiation (following DNA damage), using cDNA microarray.

Data representation and partitioning: The polar coordinate transformed data is shown in Fig 2 where the intensity magnitude R is plotted against nominal fractional intensity x . The entire range of intensity magnitude is from 5.3×10^4 to 1.42×10^7 . The data is partitioned into three groups containing an equal number of data points, ($5.3 \times 10^4 < R_{low} < 6.63 \times 10^5$); ($6.63 \times 10^5 < R_{med} < 1.27 \times 10^6$) and ($1.27 \times 10^6 < R_{high} < 1.42 \times 10^7$).

Nominal data characterization and model fits: A histogram of the data and the theoretical fit of mixture distributions are shown below. Figures 3 and 4 show the data and model fit for “Low R” data; Figures 5 and 6 for “Medium R”, and Figures 7 and 8 for “High R” data.

Figure 3: Histogram of Low R data and mixture model

Figure 4: Contributing Beta densities for mixture model fit to data in Fig 3: Green= $f_1(x)$; Black= $f_0(x)$; Red= $f_2(x)$

The probability model for the low R group is:

$$f(x) = 0.18f_1(x) + 0.79f_0(x) + 0.03f_2(x) \quad (33)$$

where the Beta density parameters are, respectively, $\alpha_{11} = 64.0, \alpha_{10} = 98.8$, for $f_1(x)$; $\alpha_{01} = 98.9, \alpha_{00} = 120.2$, for $f_0(x)$; and $\alpha_{21} = 187.8, \alpha_{20} = 104.4$ for $f_2(x)$. In general, the parameters associated with the contributing Beta density $f_j(x)$ (for the collection of genes sharing this probabilistic description) are α_{j1} , corresponding to the “effective” comparative number of cDNA molecules under test conditions for the collection, and α_{j0} , the counterpart value under reference conditions.

The corresponding results for medium and high R values are displayed in Appendix B. Note that for “high R ” data, more than 3 Beta densities are needed to capture the indicated data characteristics.

This latter observation underscores the importance of segregating the data by the intensity magnitudes. Not only are the qualitative characteristics of the data histograms in Figs 3, 5 and 7 visually different; quantitatively, the parameter values of the contributing Beta densities, the associated mixture weights, ϕ_j ’s, and, in the case of the high R group, the actual number of Beta densities, are all different. A less discriminating analysis that lumps all the data together into one mass grouping will fail to detect the subtle characteristics of the lower intensity magnitude data which would tend to be overwhelmed by the more pronounced characteristics of the higher intensity magnitude data. As will be discussed in greater detail in the forthcoming paper (Ogunnaike, et al., 2004), such considerations have major implications for our ability to draw reliable inferences about differential gene expression.

Bias Detection and Correction: Observe from each of the indicated $f_0(x)$ densities (either visually from the plots, or more quantitatively directly from the parameter estimates (α_{01}, α_{00}) that the mean values—by definition of the mean of a Beta random variable

$$\mu_0 = \frac{\alpha_{01}}{\alpha_{01} + \alpha_{00}} \quad (34)$$

deviate significantly from 0.5, with the immediate implication that nominal conditions do not hold for this

Figure 5: Histogram of Med R data and mixture model

Figure 6: Contributing Beta densities for mixture model fit to data in Fig 5: Green= $f_1(x)$; Black= $f_0(x)$; Red= $f_2(x)$

Figure 7: Histogram of Med R data and mixture model

Figure 8: Contributing Beta densities for mixture model fit to data in Fig 7: Green= $f_1(x)$; Black= $f_0(x)$; Pink= $f_2(x)$; Red = $f_3(x)$

Figure 9: Histogram of Bias-corrected Low R data and mixture model

data set. By estimating w for each data group, and using these to obtain bias-corrected fractional intensities $\tilde{x}$, a re-estimation of the mixture model now gives rise to the results shown in Figures 9–14. The resulting model parameters are shown in Appendix B. Note the “corrected” characteristics of each $f_0(x)$ corresponding to the distribution of genes showing no change in gene expression.

DISCUSSION AND CONCLUSIONS

The three main results of this paper are now summarized:

1. The distribution of spot intensity follows a Gamma model as indicated in Eq (7): the shape parameter, α , is related directly to the number of cDNA molecules responsible for the observed intensity, and the scale parameter, β , is the mean intensity per molecule—an “equipment calibration” parameter.
2. For modeling differential expression, a polar transformation of the raw spot intensity data (as in Eqs (15) and (16)) yields the intensity magnitude R_i , (to be used for rationally segregating the data) and the fractional intensity x_i , which possesses a Beta distribution as in Eq (24); the parameters, α_{i_1} and α_{i_0} are related directly to the number of cDNA molecules under test and reference conditions, respectively.

Figure 10: Contributing Beta densities for mixture model fit to corrected data in Fig 9: Green= $f_1(x)$; Black= $f_0(x)$; Red= $f_2(x)$

Figure 12: Contributing Beta densities for mixture model fit to corrected data in Fig 11: Green= $f_1(x)$; Black= $f_0(x)$; Red= $f_2(x)$

Figure 11: Histogram of Bias-corrected Med R data and mixture model

Figure 13: Histogram of Bias-corrected High R data and mixture model

Figure 14: Contributing Beta densities for mixture model fit to corrected data in Fig 13 : Green= $f_1(x)$; Black= $f_0(x)$; Pink= $f_2(x)$; Red = $f_3(x)$

3. The distribution of *entire* microarray fractional intensities is characterized, theoretically, by a mixture of overlapping Beta densities; in practice (and as demonstrated with an example) as a result of the coalescence of distributions of genes with similar expression levels, the discernible features of the histogram of real data may be captured by as few as three overlapping Beta densities, as shown in Eq (32).

This last result has the most direct impact on the quest for reliable determination of differential expression for thousands of genes from noisy, potentially biased, microarray data, with few or no replicates. It indicates that, $f(x)$, the underlying probability density function of the fractional intensity, is a linear combination of Beta density “basis functions,” providing the following important perspective of the data set: each contributing Beta density may be viewed as a “coalesced” population description of a category of genes with common differential expression attributes. The resulting coefficients of the component Beta densities therefore indicate what fraction of the entire data set belong to the category represented by the Beta density in question. For example, for the case shown in Figs 3 and 4, with the model fit shown in Eq (33),

the implications are the following: the contributions from the class of potentially “downregulated” genes (shown in the green curve), is about 18%; the contribution from the class of genes whose expression levels are potentially unchanged (shown in black) is 79%; the remaining 3% is the complement, the contribution from potentially “upregulated” genes.

Implications for Statistical inference: Just as classical statistical inference is based on theoretical reference distributions such as the t -distribution, the chi-squared distribution or the F -distribution, the mixture beta distribution, $f(x)$, developed above may now be used for drawing statistical inference regarding microarray fractional intensities. We have developed a methodology for doing precisely this—drawing statistical inference using the mixture distribution $f(x)$ as its theoretical basis: it consists of (i) a probability statement that a gene belongs to a category (the “down-regulated”, the “up-regulated”, or the “no-change”) and (ii) assigning to such statements a degree of confidence determined from the variability estimated from replicates, or else by propagation-of-error techniques when there are no replicates. The final outcome is an ordered triplet of results for each gene: a raw computed fractional (or relative) change in expression level, an associated probability that this number indicates lower, higher, or no differential expression (a category-membership probability) and a measure of confidence associated with the stated result. The details of the development of this probabilistic framework for analyzing microarray data and its application to example data sets will be discussed in a subsequent publication (Ogunnaike et al, 2004 (in preparation)).

References

- [1] Batista, J, (2003). Unpublished data
- [2] Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M. et al. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol*, 18, 630-634.

- [3] Chen, Y., E. R. Doherty, and M.L. Bottner. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics*, 2 364-374.
- [4] Dudley A.M., Aach J., Steffen M.A., and G.M. Church (2002). Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *PNAS*, 99, 11, 7554-7559.
- [5] Evans, M., N. Hastings, and B. Peacock. (1993). *Statistical Distributions*, 2nd Ed. Wiley-Interscience.
- [6] Ghosh, D. and A.M. Chinnaiyan (2002). Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, 18, 2, 275-286.
- [7] Hoyle, D. C., M. Rattay, R. Jupp, and A. Brass. (2002). Making sense of microarray data distributions. *Bioinformatics*, 18, 4 576-584.
- [8] Mikkilineni, V., Mitra, R. D., DiTonno, J. R., Merritt, J., Church, G. M., Ogunnaike, B. A. and J. S. Edwards, (2004). "Digital Quantitative Measurements of Gene Expression." *Biotech Bioeng.* 86,2,117-124.
- [9] Nadon, R. and J. Shoemaker, (2002). Statistical Issues with microarrays: processing and analysis. *Trends in Genetics*, 18, 5, 265-271.
- [10] Newton, M.N., C.M. Kendziorzski, C.S. Richmond, F.R. Blattner and K.W. Tsui (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8, 37-52.
- [11] Ogunnaike, B.A., Gelmi, C. A., and J.S. Edwards, (2004). A Probabilistic framework for microarray data analysis. (In Preparation).
- [12] Ozbudak, E. M., M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden (2002). Regulation of noise in the expression of a single gene. *Nature Genetics*, 31,69-73.
- [13] Thattai, M. and A. van Oudenaarden (2001). Intrinsic noise in gene regulatory networks, *PNAS*, 98, 15, 8614-8619.
- [14] Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Rev. Genet.* 2, 418-427
- [15] Rocke, D. M. and B. Durbin, (2001). A Model for measurement error for gene expression analysis. *Journal of Computational Biology*, 8, 557-569.
- [16] Sebastiani, P., E. Gussoni, I Kohane, and M. Ramoni. (2003). Statistical Challenges in Functional Genomics, *Statistical Science*, 18, 1, 35-70.
- [17] Steinhoff, C., T. Müller, Nuber U.A. and M Vingron (2003). Gaussian mixture density estimation applied to microarray data. *Advances In Intelligent Data Analysis V, Lecture Notes in Computer Science*, 2810, 418-429.

Appendix A: Derivation of the pdf in Eq (7)

To derive a probability distribution function for y given the probability distribution function for $z(i)$, we begin by observing that $z(i)$ and $y(k)$ are related according to the following expression:

$$Pr[z(i) < k] = Pr[y(k) > i] \quad (35)$$

which, from the definition of the variables z and y , translates in words to: "if the z -count is not yet up to k then the observed intensity i must be less than y ". In the language of probability, the event that z is less than k is equivalent to the event that the observed intensity i corresponding to z is less than y , and the probability of one event is identical to that of the other. From Eq (35) we obtain

$$Pr[z(i) < k] = Pr[z(i) \leq k-1] = \sum_{z=0}^{k-1} f(z) \quad (36)$$

which upon introducing Eq (5) gives

$$Pr[z(i) < k] = \sum_{z=0}^{k-1} e^{-\eta i} \frac{(\eta i)^z}{z!} \quad (37)$$

The key result at this point is the following identity for the RHS of Eq (37) above:

$$\sum_{z=0}^{k-1} e^{-\eta_i} \frac{(\eta_i)^z}{z!} = \frac{1}{(k-1)!} \int_{\eta_i}^{\infty} e^{-z} z^{k-1} dz \quad (38)$$

To establish this result, first let

$$\mathcal{J}(k) = \int_a^{\infty} e^{-z} z^{k-1} dz \quad (39)$$

from where it is easy to show via integration by parts, that

$$\mathcal{J}(k) = a^{k-1} e^{-a} + (k-1) \mathcal{J}(k-1) \quad (40)$$

a recursion equation that leads immediately, upon successive substitution and rearrangement, to

$$\mathcal{J}(k) = (k-1)! \sum_{m=0}^{k-1} e^{-a} \frac{a^m}{m!} \quad (41)$$

establishing the result in Eq (38).

We may now use this result in Eq (35) to obtain:

$$\begin{aligned} Pr[y(k) > i] &= 1 - Pr[y(k) \leq i] \\ &= \frac{1}{(k-1)!} \int_{\eta_i}^{\infty} e^{-z} z^{k-1} dz \end{aligned} \quad (42)$$

or, from the definition of the cumulative distribution function $F_y(i)$,

$$1 - F_y(i) = \frac{1}{(k-1)!} \int_{\eta_i}^{\infty} e^{-z} z^{k-1} dz \quad (43)$$

Differentiating with respect to i using Leibnitz's rule for differentiating under the integral, we obtain

$$\frac{\partial F_y}{\partial i} = f(y) = \frac{\eta^k}{\Gamma(k)} e^{\eta y} y^{(k-1)} \quad (44)$$

which may now be written in the familiar form:

$$f(y; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-y/\beta} y^{(\alpha-1)} \quad (45)$$

as shown in Eq(7).

Appendix B: Table of Complete Model Results

For each data group, segregated according to the intensity magnitude R , the parameters $(\alpha_{j1}, \alpha_{j0})$ for the contributing probability densities, $f_j(x)$, and the associated mixture coefficient weights, ϕ_j ($j = 0, 1, 2, 3$) are shown below for both “nominal” and “corrected” conditions. For each group, note the inequality of the “nominal” parameters $(\alpha_{01}, \alpha_{00})$ associated with $f_0(x)$; the indicated bias is eliminated upon correction.

Note also that bias correction has little or no effect on the mixture coefficient weights for low and medium R data; the influence of bias correction on these weights is more noticeable for the high R group.

	Low R	Med R	High R
$f_1; (\alpha_{11}, \alpha_{10})$	(64.0, 98.8)	(76.0, 126.1)	(15.2, 30.7)
Corrected	(69.3, 88.1)	(74.9, 100.3)	(16.9, 28.6)
$f_0; (\alpha_{01}, \alpha_{00})$	(98.8, 120.2)	(121.5, 148.8)	(31.1, 41.3)
Corrected	(106.9, 106.7)	(136.8, 136.1)	(30.1, 30.2)
$f_2; (\alpha_{21}, \alpha_{20})$	(137.8, 104.4)	(95.6, 72.5)	(29.4, 20.3)
Corrected	(150.3, 92.8)	(66.9, 42.4)	(50.3, 22.3)
$f_3; (\alpha_{31}, \alpha_{30})$			(61.6, 9.5)
Corrected			(96.3, 11.1)
ϕ_1	0.18	0.22	0.22
Corrected	0.18	0.24	0.16
ϕ_0	0.79	0.73	0.68
Corrected	0.79	0.71	0.77
ϕ_2	0.03	0.05	0.07
Corrected	0.03	0.05	0.05
ϕ_3			0.03
Corrected			0.02

Computational Modeling Reveals Molecular Details of Epidermal Growth Factor Binding

Kapil Mayawala, Dionisios G. Vlachos, and Jeremy S. Edwards

Authors' affiliation: Department of Chemical Engineering, 150 Academy Street,
University of Delaware, Newark, DE 19716

Corresponding authors:

Jeremy S. Edwards
Department of Chemical Engineering
150 Academy Street, Newark, DE 19716.
Phone: 302-831-8072. Fax: 302-831-1048. Email: edwards@che.udel.edu.

Dionisios G. Vlachos
Department of Chemical Engineering
150 Academy Street, Newark, DE 19716.
Phone: 302-831-2830. Fax: 302-831-1048. Email: vlachos@che.udel.edu.

Running title: Monte Carlo modeling of EGF binding on EGF receptors

Key words: Epidermal growth factor (EGF); EGF receptor; Kinetic Monte Carlo;
Receptor cell surface dynamics; Microscopy data; Spatiotemporal modeling

ABSTRACT

ErbB receptors are dysregulated in a high proportion of cancers, and signaling pathway of this receptor family is a critical target for anti-cancer drugs (like Erbitux[®], Herceptin[®], Iressa[®]). The initiation of signal transduction machinery activated by epidermal growth factor (EGF) occurs in the extracellular domain by the binding of a growth factor to the ErbB family of receptors. The early events in the signal transduction pathway involve ligand binding, receptor dimerization and autophosphorylation. Despite a plethora of biochemical studies and recent single particle tracking experiment (Sako and others 2000), the early molecular mechanisms involving EGF binding and EGF receptor dimerization are not well understood. Herein, we describe a spatially distributed Monte Carlo based simulation framework to enable the simulation of *in vivo* receptor diffusion and dimerization and to bridge the spatial and temporal resolution of various microscopy techniques. Our simulation results are in agreement with the data from single particle tracking and biochemical experiments on EGF receptors. Furthermore, the simulations reveal that the sequence of receptor-receptor and ligand-receptor reaction events depends on the ligand concentration and receptor density. Overall, we show that spatial simulation of receptor dynamics can be used to gain a mechanistic understanding of receptor activation which may be translated into improved cancer treatments in the future.

INTRODUCTION

Epidermal growth factor (EGF) is one of the most intensively studied growth factors and it activates the ErbB family of receptors. The activation of ErbB receptors elicits a wide range of responses leading to proliferation, differentiation and migration (Jorissen and others 2003; Yarden and Sliwkowski 2001). Due to the key role of the ErbB network in the development of solid tumors and the potential for its extracellular manipulation, ErbB has been acknowledged as a rational target for pharmacological intervention (Yarden and Sliwkowski 2001). In this work, we have studied one of the receptors of the ErbB family called epidermal growth factor receptor (EGFR). Several experimental studies suggest that EGFR is activated by ligand-induced receptor dimerization, resulting in autophosphorylation and phosphorylation of various cellular substrates (Jorissen and others 2003). Since EGF binding represents the initial step for activating EGFR, considerable work has been devoted to elucidating the mechanisms of ligand binding and dimerization (Goldman and others 2002; Resat and others 2003). However, molecular details of ligand-induced receptor dimerization are not as well understood.

Apart from several *in vitro* biochemical experiments to study mechanisms of EGFR activation (Jorissen and others 2003), recent developments in microscopy have made it possible to visualize protein dynamics in living cells (Weijer 2003). The current imaging methods either have a high spatial resolution, such as electron microscopy experiments using immunogold labeling (Vanbelzen and others 1988) and covalent linking to chemical conjugates like ferritin (Haigler and others 1979), or high temporal resolution, such as fluorescence confocal microscopy (Verveer and others 2000), single

particle tracking (Sako and others 2000) and more recently quantum dots based ligands (Lidke and others 2004). However, with the currently available imaging technologies, combined high temporal and spatial resolution is impossible.

Computational efforts devoted to understanding the extracellular mechanisms leading to EGFR activation are mostly equilibrium studies (Chamberlin and Davies 1998; Klein and others 2004; Wofsy and others 1992) or continuum reaction-diffusion models; see references in (Lauffenburger and Linderman 1993; Wiley and others 2003). While such studies have provided insights into EGFR behavior, they lack spatial correlations, known to arise from bimolecular reaction events (Ziff and others 1986), such as dimerization, and are incapable of describing spatiotemporal receptor dynamics with molecular level resolution. Monte Carlo (MC) techniques have proven powerful for systems biology modeling (Slepchenko and others 2002). In the past, lattice MC approaches have provided mechanistic understanding in other biological systems (Andrews and Bray 2004; Goldman and others 2004; Levin and others 2002; Mahama and Linderman 1994; Saxton 1995; Shea and others 2000; Shimizu and others 2003; Slepchenko and others 2002; Stiles and others 1998; Woolf and Linderman 2003; Woolf and Linderman 2004; Zhong and others 2003).

In this study, we have developed a general kinetic, lattice MC modeling framework to model the ligand (EGF) binding and dimerization of the EGFR. We compared our simulation results with single particle tracking experiments and analyzed the dominant mechanism of ligand binding and dimerization.

METHODS

(a) A hybrid null-event algorithm: A coarse, molecular level based computational framework that leaves atomistic details out, e.g., conformations and vibrations of proteins into potential energy surface minima, but still provides the sequence of molecular events at the receptor length scale is employed. Herein, we have utilized a kinetic, lattice MC method for simulating the EGFR dynamics. Microscopic events modeled include receptor dimerization and decomposition, ligand-receptor association and dissociation, and Brownian diffusion of receptors (see Figs. 1 and 2).

The existence of multiple timescales in the system and low surface density of receptors make lattice MC simulations computationally prohibitive. We have devised a hybrid between the continuous time MC method (Gillespie 1977) and the null-event MC method (see (Reese and others 2001) for an overview of MC algorithms) to increase the speed of the null-event algorithm but maintain its flexibility. In our hybrid null-event algorithm, only lattice sites filled with receptors are randomly selected, resulting in two to four orders of magnitude speedup, depending on receptor density, relative to a traditional null-event MC algorithm where all sites are randomly chosen (Mayawala *et al.*, in preparation). Additionally, operations involved in a continuous time MC method, such as summation of transition probabilities and searches over the entire lattice, are avoided, resulting in further simulation acceleration.

The spatial domain is represented using a two-dimensional square lattice that is initially randomly populated with a given density of receptors. Periodic boundary conditions are employed, which are appropriate for large systems, such as the entire plasma membrane (Jakobsson 1997). After initialization, an occupied site is randomly

picked and one of the microscopic events is possibly selected to occur based on probabilities described below.

(b) Transition probability of diffusion: The probability of diffusion in all four directions on a square lattice can be calculated using random walk theory from the diffusivity

$$\Gamma^d = \frac{4D}{a^2}, \quad (1)$$

where a is the microscopic lattice pixel dimension (taken here to be 2 nm) and D is the corresponding diffusivity of a receptor or a dimer. The transition probability of diffusion per unit time in moving from site i to site j is

$$\Gamma_{i \rightarrow j}^d = \frac{1}{4} \Gamma^d \sigma_i (1 - \sigma_j) \quad j \in B_i \quad (2)$$

where B_i denotes the set of sites to which diffusion from site i can occur. In our model, this set includes all 4 first-nearest neighboring sites. σ_i is the occupancy (discrete) function that is 1, if site i is filled, and 0, if site i is empty (a single index indicating the site is herein used to simplify notation). According to Eq. (2), the transition probability of diffusion per unit time along any direction on the square lattice can be 0 or $\frac{1}{4} \Gamma^d$ depending on the occupancy of the first-nearest neighboring site in the corresponding direction.

(c) Transition probability of reactions: The transition probability of a reaction can be obtained in terms of the macroscopic reaction rate constants, k . For a first-order (e.g., the decomposition of an EGFR dimer) or pseudo-first order (e.g., the EGF binding onto a

receptor because EGF is assumed at a constant concentration) reaction (see Fig. 2), one has

$$A \rightarrow C, \quad \Gamma_i^r = k\sigma_i \quad (3)$$

where Γ_i^r is the transition probability of reaction at site i . For a bimolecular reaction on a square lattice (e.g., receptor-receptor dimerization), the transition probability per unit time at selected site i is

$$\text{for the reaction: } A + B \rightarrow C, \quad \Gamma_{i \rightarrow j}^r = \frac{k}{4} \sigma_i \sigma_j, \quad (4)$$

$$\text{and for the reaction: } 2A \rightarrow C, \quad \Gamma_{i \rightarrow j}^r = \frac{k}{2} \sigma_i \sigma_j. \quad (5)$$

Here the reacting species (A and B or A and A) occupy adjacent sites i and j , and the units of k are $(\text{molecules/site})^{-1}\text{sec}^{-1}$. The factor of four in $k/4$ in Eq. (4) accounts for the fact that all four neighboring sites are randomly chosen for species B. Similar is the factor of 2 in Eq. 5 due to the degeneracy of species participating in the homodimerization (Mayawala *et al.*, in preparation).

(d) Event selection and time advancement: After an occupied site, say i , is selected, the transition probabilities of all possible events are computed. The probability for a certain event ‘ x ’ at site i is calculated as

$$p_i^x = \frac{\Gamma_i^x}{\Gamma_{\max}}, \quad (6)$$

where $\Gamma_{\max}$ is a normalization constant to ensure that the event selection probability is always less than or equal to 1 and Γ_i^x is the transition probability of event x (reaction or diffusion) at site i , as defined above. The maximum values of transition probabilities per unit time of events described by Eqs. (2)-(4) are Γ^d , k , $k/4$, and $k/2$, respectively. One convenient way to define the normalization constant is

$$\Gamma_{\max} = \Gamma^d + \max \left\{ \sum_{\text{all reaction events}} \Gamma^r \right\}. \text{ For the reaction events given in Fig. 2}$$

$$\Gamma_{\max} = \Gamma^d + \frac{k_{1f}}{2} + \frac{k_{2f}}{4} + \frac{k_{3f}}{2} + \sum_{i=4}^6 k_{if} + \sum_{i=1}^6 k_{ib}, \quad (7)$$

where subscripts f and b refer to the forward and backward reaction events, respectively. A random number ' r ' is finally chosen from a uniform distribution between 0 and 1. The events are randomly ranked. The smallest value of m satisfying the following criterion is chosen

$$\sum_{x=1}^m p_i^x > r. \quad (8)$$

If there is no value of m satisfying the above criterion, then no event is selected to occur (a null-event) and a new occupied site is again randomly picked. Otherwise, the m^{th} event is selected, the populations on the lattice are updated accordingly to reflect the stoichiometry of the reaction or diffusion process, and the real time is advanced based on the most frequently selected event as suggested by (Reese and others 2001). In our simulations, real time is advanced based on the diffusion of receptors according to which the average time step after each successful diffusion event is calculated as

$$\Delta t = \frac{1}{\sum_{i=1}^{\text{No. of sites}} \left(\sigma_i \sum_{j \in B_i} \Gamma_{i \rightarrow j}^d (1 - \sigma_j) \right)} = \frac{1}{\frac{1}{4} \Gamma^d \sum_{\text{occupied sites}} \left(\sum_{j \in B_i} (1 - \sigma_j) \right)} \quad . \quad (9)$$

The summation in the denominator is updated each time a successful event happens by subtracting the previous occupancy values affected by the selected event and adding the new ones. In this way, the summation is carried out only once (at the beginning of a simulation) over all occupied sites. Finally, a new site is again picked randomly until the desired real time is reached. The hybrid null-event MC algorithm explained above was implemented (for details on the algorithm see (Reese and others 2001)) using Fortran 90. For each data point, 10 simulations with different seeds of the random number generator were used to collect statistics.

(e) Simulation size and model parameters: The cell surface was represented using a 50x50 lattice with each pixel being $a=2$ nm in size. The simulation box represents a small fraction ($\sim 10^{-5}$) of the plasma membrane area of human epithelioid carcinoma cells (A-431 cells). Experimentally, the density of receptors has been shown to vary between 0.1% and 4%, with higher densities of localized receptors encountered in certain regions of the plasma membrane (Jorissen and others 2003; Mineo and others 1999; Ringerike and others 2002; Waugh and others 1999). The nominal simulated receptor density of 1.7% (55 receptors) chosen here is within the reported experimental range. Experimental studies have provided evidence of predimerized receptors to different extents (Gadella and Jovin 1995; Martin-Fernandez and others 2002; Moriki and others 2001; Vanbelzen and others 1988; Yu and others 2002). Consequently, a fraction ($\sim 82\%$) of receptors was initially placed at random locations as monomers and the remaining as dimers.

The model parameters are summarized in Table 1. We assume that there are two types of EGF binding on the cell surface: low affinity binding (on monomer EGFR) and high affinity binding (on dimerized EGFR). Experimental information suggests the existence of a high affinity receptor population for receptor-ligand association, most of which is present in the form of dimers; see review by (Jorissen and others 2003).

The receptor dimerization constants vary with ligand occupancy. Several experimental studies have shown that dimerization between unbound receptors occurs with lower affinity than that between one bound and one unbound receptor. Finally, dimerization between two ligand bound receptors occurs with the highest affinity (Lemmon and others 1997; Sherrill and Kyte 1996). Experimental studies (Rees and others 1984) suggest that the high affinity receptors, consisting of dimers, diffuse slower than the low affinity population, consisting of monomers. This observation is consistent with single particle tracking data (Sako and others 2000) that shows that the spots with higher intensity, representing dimers, diffuse slower than the spots with lower intensity representing monomers. Based on the aforementioned experimental evidence (Rees and others 1984), monomers were assumed to diffuse faster than dimers.

RESULTS AND DISCUSSION

Our computer simulations reveal the underlying mechanism on the plasma membrane leading to dimerized and ligand (EGF) bound receptors. The dynamics of the ligand binding events were compared with the recent single particle tracking experiment (Sako and others 2000) at an EGF concentration of 0.16 nM in the 0-60 sec time interval. To compare the simulation with the experiment, EGF was assumed to be associated with

Cy3 dye. A dimerized receptor with two EGF molecules was taken to fluoresce twice as intensely as a receptor (single or dimerized) bound to one EGF molecule. The predicted initial increase of low intensity spots (monomers plus dimers having one EGF bound) followed by a slower increase in high intensity spots is qualitatively consistent with the experimental data (see Fig. 3(a)). The rapid increase in the low intensity signal was due to the rapid binding of EGF on predimerized EGFR. Furthermore, the increase in the total density of Cy3-EGF spots (total bound EGF on all receptors) shown in Fig. 3(b) is also consistent with experimental data. The possible sequences of events leading to the formation of EGF bound dimerized EGFR are shown in Fig. 4(a). Sako *et al.* (Sako and others 2000) suggested sequence 1 as dominant. However, the experimental study alone cannot unambiguously determine the sequence due to its low spatial resolution and the fact that only ligand bound receptors can be tracked. Our simulations showed that 95-100% of the receptors follow sequence 1, 0-4.9% sequence 2, and the remaining receptors follow sequence 3 and are consistent with the hypothesis of Sako *et al.*

Single particle tracking experiments (Sako and others 2000) are typically limited to low ligand concentrations. Simulations were used to elucidate the influence of extracellular EGF concentration on EGFR dimerization (Fig. 4b). Our simulations indicate that the relative contributions of sequences 1-3 change with the concentration of ligand. At low ligand concentration, sequence 1 dominates, whereas at higher concentration, a significant fraction of dimers form via sequence 2. Furthermore, sequence 3 also occurs at appreciable extents at high concentration of EGF. At low ligand concentration, most of the ligand gets bound to dimerized receptors, which have a higher ligand affinity. There is a fraction of predimerized receptors in equilibrium with

monomer receptors. In the presence of ligand, dimerization is favored; however, the extent to which free EGFR dimerization can occur is limited. Hence, at higher ligand concentration, a significant fraction of ligand is attached to monomers and the coupling between ligand attached monomer and free or ligand attached monomer gives rise to dimers. Our conclusion about ligand binding driven dimerization (Fig. 4c) is consistent with several previous experimental studies (Gadella and Jovin 1995; Hurwitz and others 1991; Lax and others 1991).

Support for the suggested mechanisms also comes from biochemical studies. The experimental study of (Bellot and others 1990) supports our finding that at low doses of EGF, inhibition of high affinity binding by mAb108 can kill almost 50-100% of it, indicating that most of the early binding takes place by sequence 1 at low EGF concentration. However, this inhibition is overcome at higher concentration (~20-50 times) of EGF when there is substantial formation of EGF bound dimerized EGFR via sequence 2. A larger scale simulation with variable receptor densities in different regions of the plasma membrane will be developed in the future for quantitative comparison with such biochemical experiments.

Another important factor influencing ligand binding and dimerization is the receptor density. The receptor density can significantly influence the amount as well as the mechanism of EGF binding. Our simulations show that at 0.16 nM EGF, 75% and 89% reduction in receptor density (nominal value of 2.2%) leads to 55% and 71% reduction in the total amount of EGF binding, respectively. Furthermore, 75% and 89% reduction in receptor density makes sequence 2 relatively more significant occurring 0-10% and 0-30% of the times, respectively. However, a more detailed representation of

the plasma membrane, including various microdomains like clathrin coated pits, lipid rafts, and caveolae, is required for a realistic analysis of the effect of the receptor density, and will be the focus of future work. Future work will also address the molecular mechanisms leading to the formation of hetero-dimers with ErbB2, ErbB3 and ErbB3, and higher-mers.

CONCLUSION

In summary, we have developed a computational framework for studying cell surface receptor dynamics that can bridge biochemical data on one hand with various microscopy experiments on the other, which lack simultaneous high spatial and temporal resolution. This work provides an important step forward in the era of *in vivo* imaging based modeling approaches (Phair and Misteli 2001). This model can be used to further enhance our understanding of cellular signaling mechanisms. For example, in this work comparison of MC simulations with single particle tracking experiments reveals how the sequence of receptor-receptor and ligand-receptor reaction events is dependent on the ligand concentration and receptor density. Future advancements in medicine will ultimately also include the mathematical analysis and modeling of ErbB receptor diffusion, dimerization and clustering, which will increase our understanding of tumorigenesis and lead to medical advances, such as individualized therapy for heterogeneous cancers.

ACKNOWLEDGEMENTS

KM thanks Abhijit Chatterjee for useful discussions. The work was supported by grants from the US Department of Energy and the National Science Foundation (CTS-0312117).

REFERENCES

- Andrews SS, Bray D. 2004. Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. *Physical Biology* 1:137-151.
- Bellot F, Moolenaar W, Kris R, Mirakhur B, Verlaan I, Ullrich A, Schlessinger J, Felder S. 1990. High-affinity epidermal growth factor binding is specifically reduced by a monoclonal antibody, and appears necessary for early responses. *J. Cell Biol.* 110(2):491-502.
- Chamberlin SG, Davies DE. 1998. A unified model of c-erbB receptor homo- and heterodimerisation. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* 1384(2):223-232.
- Gadella T, Jr, Jovin T. 1995. Oligomerization of epidermal growth factor receptors on A431 cells studied by time-resolved fluorescence imaging microscopy. A stereochemical model for tyrosine kinase receptor activation. *J. Cell Biol.* 129(6):1543-1558.
- Gillespie DT. 1977. Exact stochastic simulation of coupled chemical reactions. *Journal of Chemical Physics* 81:2340-2361.
- Goldman J, Andrews S, Bray D. 2004. Size and composition of membrane protein clusters predicted by Monte Carlo analysis. *European Biophysics Journal* 33:506-512.
- Goldman JP, Gullick WJ, Bray D, Johnson CG. Individual-based simulation of the clustering behaviour of epidermal growth factor receptors. In: Lamont G, editor; 2002; Madrid, Spain. ACM Press.
- Haigler HT, Mckanna JA, Cohen S. 1979. Direct visualization of the binding and internalization of a ferritin conjugate of epidermal growth-factor in human carcinoma cells A-431. *Journal of Cell Biology* 81:382-395.
- Hendriks BS, Opresko LK, Wiley HS, Lauffenburger D. 2003. Quantitative Analysis of HER2-mediated Effects on HER2 and Epidermal Growth Factor Receptor Endocytosis: Distribution Of Homo- And Heterodimers Depends On Relative Her2 Levels. *J. Biol. Chem.* 278(26):23343-23351.
- Hurwitz D, Emanuel S, Nathan M, Sarver N, Ullrich A, Felder S, Lax I, Schlessinger J. 1991. EGF induces increased ligand binding affinity and dimerization of soluble epidermal growth factor (EGF) receptor extracellular domain. *J. Biol. Chem.* 266(32):22035-22043.
- Jakobsson E. 1997. Computer simulation studies of biological membranes: progress, promise and pitfalls. *Trends in Biochemical Sciences* 22:339-344.

- Jorissen RN, Walker F, Pouliot N, Garrett TPJ, Ward CW, Burgess AW. 2003. Epidermal growth factor receptor: mechanisms of activation and signalling. *Experimental Cell Research* 284(1):31-53.
- Kawamoto T, Sato JD, Le A, Polikoff J, Sato GH, Mendelsohn J. 1983. Growth stimulation of A431 cells by epidermal growth factor: identification of high-affinity receptors for epidermal growth factor by an anti-receptor monoclonal antibody. *PNAS* 80:1337-1341.
- Klein P, Mattoon D, Lemmon MA, Schlessinger J. 2004. A structure-based model for ligand binding and dimerization of EGF receptors. *PNAS* 101(4):929-934.
- Lauffenburger DA, Linderman JJ. 1993. *Receptors models for binding, trafficking, and signaling*. New York: Oxford University Press.
- Lax I, Mitra A, Ravera C, Hurwitz D, Rubinstein M, Ullrich A, Stroud R, Schlessinger J. 1991. Epidermal growth factor (EGF) induces oligomerization of soluble, extracellular, ligand-binding domain of EGF receptor. A low resolution projection structure of the ligand-binding domain. *J. Biol. Chem.* 266(21):13828-13833.
- Lemmon MA, Bu Z, Ladbury JE, Zhou M, Pinchasi D, Lax I, Engelman DM, Schlessinger J. 1997. Two EGF molecules contribute additively to stabilization of the EGFR dimer. *The EMBO Journal* 16:281-294.
- Levin MD, Shimizu TS, Bray D. 2002. Binding and Diffusion of CheR Molecules Within a Cluster of Membrane Receptors. *Biophys. J.* 82(4):1809-1817.
- Lidke DS, Nagy P, Heintzmann R, Arndt-Jovin DJ, Post JN, Grecco HE, Jares-Erijman EA, Jovin TM. 2004. Quantum dot ligands provide new insights into erbB/HER receptor-mediated signal transduction. *Nature Biotechnology* 22:198 - 203.
- Lin CR, Chen WS, Lazar CS, Carpenter CD, Gill GN, Evans RM, Rosenfeld MG. 1986. Protein kinase C phosphorylation at Thr 654 of the unoccupied EGF receptor and EGF binding regulate functional receptor loss by independent mechanisms. *Cell* 44(6):839-848.
- Mahama P, Linderman J. 1994. A Monte Carlo study of the dynamics of G-protein activation. *Biophys. J.* 67(3):1345-1357.
- Martin-Fernandez M, Clarke DT, Tobin MJ, Jones SV, Jones GR. 2002. Preformed Oligomeric Epidermal Growth Factor Receptors Undergo an Ectodomain Structure Change during Signaling. *Biophys. J.* 82(5):2415-2427.
- Mineo C, Gill GN, Anderson RGW. 1999. Regulated Migration of Epidermal Growth Factor Receptor from Caveolae. *J. Biol. Chem.* 274(43):30636-30643.
- Moriki T, Maruyama H, Maruyama IN. 2001. Activation of preformed EGF receptor dimers by ligand-induced rotation of the transmembrane domain1. *Journal of Molecular Biology* 311(5):1011-1026.
- Phair RD, Misteli T. 2001. Kinetic Modeling Approaches to In Vivo Imaging. *Nature Reviews Molecular Cell Biology* 2:898-907.
- Rees AR, Gregoriou M, Johnson P, Garland PB. 1984. High affinity epidermal growth factor receptors on the surface of A431 cells have restricted lateral diffusion. *The EMBO Journal* 3:1843-1847.
- Reese JS, Raimondeau S, Vlachos DG. 2001. Monte Carlo algorithms for complex surface reaction mechanisms: Efficiency and accuracy. *Journal of Computational Physics* 173:302-321.

- Resat H, Ewald JA, Dixon DA, Wiley HS. 2003. An Integrated Model of Epidermal Growth Factor Receptor Trafficking and Signal Transduction. *Biophys. J.* 85(2):730-743.
- Ringerike T, Blystad FD, Levy FO, Madshus IH, Stang E. 2002. Cholesterol is important in control of EGF receptor kinase activity but EGF receptors are not concentrated in caveolae. *J Cell Sci* 115(6):1331-1340.
- Sako Y, Minoghchi S, Yanagida T. 2000. Single-molecule imaging of EGFR signalling on the surface of living cells. *Nature Cell Biology* 2(3):168-172.
- Saxton M. 1995. Single-particle tracking: effects of corrals. *Biophys. J.* 69(2):389-398.
- Schlessinger J. 1986. Allosteric regulation of the epidermal growth factor receptor kinase. *J. Cell Biol.* 103(6):2067-2072.
- Schlessinger J. 1988. The epidermal growth factor receptor as a multifunctional allosteric protein. *Biochemistry* 27:3119-3123.
- Shea LD, Neubig RR, Linderman JJ. 2000. Timing is everything: The role of kinetics in G protein activation. *Life Sciences* 68(6):647-658.
- Sherrill JM, Kyte J. 1996. Activation of epidermal growth factor receptor by epidermal growth factor. *Biochemistry* 35:5705-5718.
- Shimizu TS, Aksenov SV, Bray D. 2003. A Spatially Extended Stochastic Model of the Bacterial Chemotaxis Signalling Pathway. *Journal of Molecular Biology* 329(2):291-309.
- Slepchenko BM, Schaff JC, Carson JH, Loew LM. 2002. Computational Cell Biology: Spatiotemporal Simulation of Cellular Events. *Annual Review of Biophysics and Biomolecular Structure* 31:423-441.
- Stiles JR, Jr TMB, Salpeter EE, Salpeter MM. 1998. Monte Carlo simulation of neurotransmitter release using MCell, a general simulator of cellular physiological processes. In: Bower JM, editor. *Computational Neuroscience*. New York: Plenum. p 279-284.
- Vanbelzen N, Rijken PJ, Hage WJ, Delaat SW, Verkleij AJ, Boonstra J. 1988. Direct visualization and quantitative-analysis of epidermal growth factor-induced receptor clustering. *Journal Of Cellular Physiology* 134:413-420.
- Verveer PJ, Wouters FS, Reynolds AR, Bastiaens PIH. 2000. Quantitative Imaging of Lateral ErbB1 Receptor Signal Propagation in the Plasma Membrane. *Science* 290(5496):1567-1570.
- Waugh MG, Lawson D, Hsuan JJ. 1999. Epidermal growth factor receptor activation is localized within low-buoyant density, non-caveolar membrane domains. *Biochemical Journal* 337:591-597.
- Weijer CJ. 2003. Visualizing signals moving in cells. *Science* 300:96-100.
- Wiley HS, Shvartsman SY, Lauffenburger DA. 2003. Computational modeling of the EGF-receptor system: a paradigm for systems biology. *Trends in Cell Biology* 13(1):43-50.
- Wofsy C, Goldstein B, Lund K, Wiley H. 1992. Implications of epidermal growth factor (EGF) induced egf receptor aggregation. *Biophys. J.* 63(1):98-110.
- Woolf PJ, Linderman JJ. 2003. Untangling Ligand Induced Activation and Desensitization of G-Protein-Coupled Receptors. *Biophys. J.* 84(1):3-13.

- Woolf PJ, Linderman JJ. 2004. An algebra of dimerization and its implications for G-protein coupled receptor signaling. *Journal of Theoretical Biology* 229(2):157-168.
- Yarden Y, Sliwkowski MX. 2001. Untangling the ErbB signalling network. *Nature Reviews Molecular Cell Biology* 2:127-137.
- Yu X, Sharma KD, Takahashi T, Iwamoto R, Mekada E. 2002. Ligand-independent Dimer Formation of Epidermal Growth Factor Receptor (EGFR) Is a Step Separable from Ligand-induced EGFR Signaling. *Mol. Biol. Cell* 13(7):2547-2557.
- Zhong H, Wade SM, Woolf PJ, Linderman JJ, Traynor JR, Neubig RR. 2003. A Spatial Focusing Model for G Protein Signals. *Regulator Of G Protein Signaling (Rgs) Protein-Mediated Kinetic Scaffolding. J. Biol. Chem.* 278(9):7278-7284.
- Ziff RM, Gulari E, Barshad Y. 1986. Kinetic phase transitions in an irreversible surface-reaction model. *Physical Review Letters* 56:2553-2556.

Figure 1: Schematic of simulated microscopic events. Each receptor can diffuse to an empty neighboring site, react with a neighboring receptor to form a dimer, and bind ligand. All events are reversible.

Figure 2: Reactions events considered in our model as given in (Wofsy and others 1992).

Figure 3: (a) Evolution of intensity of dimerized receptors with two ligands (high intensity spots) and of monomer plus dimerized receptors with a single ligand bound (low intensity spots) along with the data of single particle tracking experiments by Sako *et al.* over time intervals of 20 sec. The simulation intensity has been normalized with the experimental data. Error bars indicate 2 standard deviations obtained from 10 independent MC simulations. (b) Comparison of predicted density of Cy3-EGF spots with experimental data of Sako *et al.* The predicted density is normalized with the experimental one at 60 sec. Good agreement of simulations with experimental data is found. Error bars indicate 2 standard deviations obtained from 10 independent MC simulations.

Figure 4: (a) Sequence of reactions resulting in dimerized receptors with both receptors bound to ligand. All reactions are reversible. (b) Contributions of the different reaction mechanisms at different concentrations of EGF. With increasing concentration of EGF, sequence 2 occurs at the expense of sequence 1. At higher concentration of EGF, sequence 3 also occurs to appreciable extent. (c) Percentage of dimerized receptors vs. time for different concentrations of ligand.

Table1: Kinetic (reaction events given in Fig. 2) and transport (monomer and dimer diffusion) parameters.

Figure 1

Receptor-receptor reactions

Ligand-receptor reactions

Figure 2

Figure 3(a)

Figure 3(b)

Figure 4(a)

Figure 4(b)

Figure 4(c)

Table 1

Equilibrium constants		
K_1	13.3 (molecule/site) ⁻¹	In the range to have ~18% of the monomer EGFR as dimers in the absence of EGF (Gadella and Jovin 1995; Martin-Fernandez and others 2002; Moriki and others 2001; Vanbelzen and others 1988; Yu and others 2002)
K_2	4.0×10^3 (molecule/site) ⁻¹	Calculated based on equilibrium relations given in (Wofsy and others 1992)
K_3	1.2×10^6 (molecule/site) ⁻¹	Calculated based on equilibrium relations given in (Wofsy and others 1992)
K_4	$4 \times 10^8 \text{ M}^{-1}$	In the range suggested by (Jorissen and others 2003; Kawamoto and others 1983; Lin and others 1986; Schlessinger 1986; Schlessinger 1988)
K_5, K_6	$1.2 \times 10^{11} \text{ M}^{-1}$	In the range suggested by (Jorissen and others 2003; Lin and others 1986; Schlessinger 1986; Schlessinger 1988)

Kinetic parameters		
k_{1b}, k_{2b}	0.17 sec^{-1}	(Hendriks and others 2003)
k_{3b}	$1.7 \times 10^{-3} \text{ sec}^{-1}$	(Hendriks and others 2003)
k_{4b}, k_{5b}	$2.9 \times 10^{-3} \text{ sec}^{-1}$	(Bellot and others 1990)
k_{6b}	$5.8 \times 10^{-3} \text{ sec}^{-1}$	(Bellot and others 1990)

Transport parameters		
D_{monomer}	$2 \times 10^{-14} \text{ m}^2 \text{ sec}^{-1}$	(Rees and others 1984)
D_{dimer}	$4 \times 10^{-17} \text{ m}^2 \text{ sec}^{-1}$	(Rees and others 1984)

Digital Quantitative Measurements of Gene Expression

Venugopal Mikkilineni,¹ Robi D. Mitra,² Joshua Merritt,¹ Jason R. DiTonno,¹ George M. Church,³ Babatunde Ogunnaike,¹ Jeremy S. Edwards¹

¹Department of Chemical Engineering, University of Delaware, Newark, Delaware 19716; telephone: (302) 831-8072; fax: (302) 831-1048; e-mail: edwards@che.udel.edu

²Department of Genetics, Washington University, St. Louis, Missouri 63110

³Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115

Received 22 July 2003; accepted 14 November 2003

Published online 19 February 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/bit.20048

Abstract: One of the primary goals of functional genomics is to provide a quantitative understanding of gene function. However, the success of this enterprise is dependent on the accuracy and precision of the functional genomic data. A *novel* approach, digital analysis of gene expression (DAGE) described herein, is an accurate and precise technology for measuring digital gene expression on a *relative* or *absolute* scale by simply counting the number of transcripts of a gene being expressed at a given time. The result is a greatly improved technology sensitive enough for identifying and quantifying small (but biologically important and statistically relevant) changes in gene expression. Fourteen genes involved in galactose metabolism in *Saccharomyces cerevisiae* were analyzed for their expression levels in glucose and galactose minimal media. The quantitative expression results were characterized in terms of distributional and accuracy attributes; they were also in general agreement (in terms of direction of change) with corresponding results obtained using microarray technology. DAGE is likely to have profound implications in the field of functional genomics because the gene expression measurements are digital in nature and therefore more accurate than any other technologies. © 2004 Wiley Periodicals, Inc.

Keywords: yeast; gene expression; genomics; metabolism; colonies; bioinformatics

INTRODUCTION

The Human Genome Project sparked DNA sequencing initiatives for a number of model organisms from all the major kingdoms of life, and currently there are many completely sequenced and annotated genomes publicly available (<http://www.tigr.org/tldb/mdb/mdbcomplete.html>). Subsequent bioinformatics analysis has then brought us to the brink of having a complete molecular “parts catalogue” of many organisms. Such compositional and structural molecular information is of great value. However, a critical question has arisen (Hieter and Boguski, 1997; Koshland, 1998): how

can genomic data be used in tandem with bioinformatics to analyze, interpret, and predict the relationship between an organism’s genotype and its phenotype?

The molecular parts catalogue now needs to be translated into functional information, and this undertaking is likely more challenging than identifying the individual parts (Fields, 1997; Hieter and Boguski, 1997). The systematic study of cellular function by generation and collection of high-throughput data and computational tools has been called systems biology. However, broad applicability of systems biology is currently hindered by the limitations of the current measurement technologies. Thus, experimental and statistical tools need to be developed for generating high-quality, high-throughput data for efficiently extracting useful biological information.

A number of technologies have been developed for acquiring gene expression information on a *whole-transcriptome* level. The techniques currently in vogue are based on direct sequence analysis (Adams, 1996), differential display (Fisler, 1998), or specific hybridization of complex cDNA or mRNA probes to microarrays of oligonucleotides or cDNAs (Brown et al., 1998; Brown and Botstein, 1999; Elek et al., 2000; Ramsay, 1998). These approaches are all limited in the dynamic range of the measurements and by their inability to produce digital measurements (Brenner et al., 2000; Velculescu et al., 1995). Two high-throughput approaches have been developed to make “digital” measurements or measurements involving simple counting of gene expression, thus providing more accurate and more precise measurements (Brenner et al., 2000; Velculescu et al., 1995). The digital gene expression measurement technologies rely on counting the number of times a transcript occurs in a sample. Digital approaches are distinct from (and often more accurate than) other high-throughput approaches that rely on the fluorescent intensity for quantification. However, digital approaches are of limited practical utility because they are time consuming and expensive.

The premise of this paper is that, by using in-situ PCR, the accuracy and precision of “digital” approaches can be

Correspondence to: Jeremy S. Edwards

Contract grant sponsors: University of Delaware Research Foundation; NIH; U.S. DOE Office of Biological Research

maintained while greatly reducing the cost. The in-situ PCR method involves an amplification reaction in a polymerized acrylamide matrix containing standard PCR reagents and DNA template (Butz et al., 2003; Merritt et al., 2003; Mitra et al., 2003; Mitra and Church, 1999). As the products remain localized near their respective templates, an immobilized single DNA template molecule amplifies locally and gives rise to a PCR colony, or “polony”. Herein, we describe a transcript profiling technology using in-situ PCR as the digital readout called Digital Analysis of Gene Expression (DAGE). We have quantitatively measured the expression levels of 14 *Saccharomyces cerevisiae* genes involved in galactose regulation and metabolism. As the data obtained is digital in nature, rigorous statistical procedures have been developed for its analysis.

MATERIALS AND METHODS

Strains and Growth Conditions

Yeast FY4 cells were grown in glucose (2%) and galactose (2%) minimal media at 30°C until the cells reached steady state and were harvested at an optical density (600 nm) of 1.0. Yeast cells were collected as 10-mL aliquots, quick-chilled to 4°C in liquid nitrogen, and centrifuged at 4,000 rpm at 4°C, and the cell pellets were used for RNA extraction.

cDNA Preparation

Total RNA was isolated from *S. cerevisiae* FY4 strains grown as described above. The pellets were resuspended in RNawiz buffer (Ambion). Cells were lysed using a Mini-Bead Beater (Biospec), and RNA was isolated according to the manufacturer's protocol. Contaminating DNA was digested using DNase I according to the manufacturer's protocol

(Gibco). Total RNA levels were quantified using A_{260} measurements. cDNA was synthesized using Reverse Transcriptase-II (Gibco) according to the method outlined by the supplier with Oligo dT₁₈ as a primer. The cDNA was purified by RNaseH (Gibco) digestion and treated with 5 µL of Clontech enzyme removal resin and ethanol precipitated.

Gene Selection and Primer Design

Fourteen genes involved in galactose metabolism and regulation were selected for expression analysis. Primers used to amplify a portion of the 3' region of each gene were designed using Vector NT1 software (InfoMax) with the coding region as input. The entire genome was also input to ensure the specificity of primers for the gene of interest. Primers were designed to amplify regions between 250 and 350 bp. All primers used are listed online in the supplementary information, in Supplementary Table I.

Expression Profiling Using PCR Colonies

Purified first-stand cDNA (at the appropriate dilutions) was used as a template in the PCR reaction. Then based on this initial concentration the starting cDNA concentration was adjusted to create distinguishable non-overlapping polonies. Polony reactions were conducted as previously described (Butz et al., 2003; Merritt et al., 2003; Mitra et al., 2003; Mitra and Church, 1999)

Polony Gel Preparation

Preparation of Slides

Teflon coated slides (Erie Scientific) were treated with Bind Silane (Amersham, Inc.) per manufacturer's instructions.

Table I. Comparison of the gene expression of glucose and galactose.*

Gene	Log ξ_A (gluc)	Log ξ_B (gal)	95% confidence interval on δ	P value	Fold increase	
					DAGE data	Microarray data
GAL1	1.72	4.32	(2.47, 2.74)	0	398.66	42.14
GAL2	1.89	5.01	(2.97, 3.27)	0	1310.69	214.42
GAL3	1.97	2.63	(0.68, 1.08)	0	4.53	9.90
GAL4	2.10	1.74	(-0.47, -0.24)	0.001	0.44	1.21
GAL7	1.02	3.47	(2.31, 2.60)	0	285.69	2.43
GAL10	0.41	3.68	(2.87, 3.67)	0	1857.80	29.54
GAL80	2.15	2.84	(0.51, 0.87)	0	4.87	3.00
TUP1	2.39	2.71	(0.19, 0.44)	0.001	2.07	1.30
PGK1	4.19	3.95	(-0.51, 0.03)	0.074	Not significant	0.92
MIG1	1.05	1.40	(0.08, 0.63)	0.027	2.25	2.80
ADE13	2.42	2.65	(0.10, 0.36)	0.006	1.70	0.33
HXT1	0.53	0.52	(-0.40, 0.37)	0.932	Not significant	2.00
HXT10	0.42	0.36	(-0.33, 0.23)	0.627	Not significant	1.67
HXT14	0.10	0.58	(-0.03, 0.98)	0.057	Borderline	2.47

*Log ξ_A = Log₁₀ number of polonies per microgram of glucose first-strand cDNA. Log ξ_B = Log₁₀ number of polonies per microgram of galactose first-strand cDNA.

δ = Log ξ_A - Log ξ_B = Log (ξ_A/ξ_B); Log₁₀ (fold increase).

Step 1—Master Mix: A 250- μ L master mix was prepared by adding the following components: 1.5 $\times$ Jumpstart *Taq*, 10 $\times$ buffer, 0.52 μ M dNTPs, 0.25% BSA (Sigma), 0.25% Tween 20, 0.75 μ L of forward and reverse primers (100 μ M, and 12% degassed acrylamide/bis-acrylamide (19:1). **Step 2—Reaction Mix:** A 51.33- μ L aliquot of reaction mix was prepared using the following components: 1st strand RT - products (2 μ L), master mix (44.5 μ L), 16.65 U of Jumpstart *Taq* DNA Polymerase, 5 U/ μ L (Sigma), 0.85 μ L APS (5%), and 0.85 μ L *TEMED* (5%). **DNA Template:** An optimum concentration of DNA must be determined by conducting polony reaction with a series of dilutions for each primer being used. An optimum concentration was determined experimentally by creating non-overlapping colonies. The number of colonies or transcripts that can be detected with our scanner on a polony gel can range from 1 to $\sim$ 3000.

The reaction mix was thoroughly mixed, and 20 μ L was injected into the well. From each reaction mix, two slides were prepared. After the slide was loaded, the cover glass was adjusted to completely cover the well. Immediately after this step a hybridization chamber (Grace Biolabs) was adhered to the slide and the gel was polymerized for 10 min. Light mineral oil was injected into the open ports of the hybridization chamber, and the ports were sealed with sealants supplied by the manufacturer.

Thermal Cycling

Polony slides were thermal cycled in an in-situ PCR hybridization tower (MJ Research). The following cycling conditions were used: 96 $^{\circ}$ C for 2 min followed by 40 cycles of 96 $^{\circ}$ C for 30 s, 62 $^{\circ}$ C for 45 s, and 72 $^{\circ}$ C for 30 s. The final extension step was set for 72 $^{\circ}$ C for 2 min.

Staining and Visualization

After PCR amplification, the hybridization chambers were removed and the slides were immersed in hexane (Sigma) for 10 min. The cover glass was then removed. The gels were then stained in 2 $\times$ SYBR green for 10 minutes. Polonies were visualized using a Scanarray 5000 scanner.

Statistical Analysis

The primary objective of statistical analysis was to investigate the suitability of the proposed analysis technique and to characterize the inherent statistical distributional properties. Serial dilutions of the cDNA from the galactose minimal media culture were made at the rate of 50% of the previous dilution. The highest expressed gene *GAL2* was chosen to test each dilution for model adequacy and statistical distributional properties. After this characterization, the methodology was then used to determine the gene expression levels for 14 genes between the conditions of *S. cerevisiae* grown in galactose and glucose minimal media. A detailed description of the statistical model and explanation is provided in the online supplement.

RESULTS AND DISCUSSION

The basic procedure for DAGE is illustrated in Figures 1 and 2, and we used this approach to quantitatively measure the expression level of 14 genes in galactose and glucose minimal media at two dilutions per gene replicating each experiment twice. The results are summarized in Table I and II. The *GAL* genes, with the exception of *GAL 4*, were highly expressed in galactose and expressed at relatively low levels in glucose. This is in general agreement with the relative expression level changes determined from micro-array data (Dudley et al., 2002). Note also that the DAGE technique has a larger dynamic range than do other expression profiling techniques (Yuen et al., 2002). Namely, we were able to quantitatively measure gene expression at very low levels with similar precision as to measurements at very high expression levels; furthermore, the upper limit of detection is unbounded because one can always increase the dilution an unlimited number of times.

We obtained several key results which are summarized below and in Tables I and II: (1) We have provided absolute levels of expression for each gene in specific growth conditions the 95% confidence intervals for the absolute levels of expression and the differential expression (in the two media), and the *P*-value associated with the hypothesis that

Figure 1. Concept of DAGE. Total RNA is reverse transcribed into cDNA and PCR amplified with specific genes. The “red” cDNA molecule corresponds to the cDNA molecule of interest that we will quantitatively measure. The “black” cDNA molecules indicate all other transcripts.

Figure 2. Transcript profiling. Polony gel shows variable expression profiles of two different genes (red and green spots) for the same RNA condition.

the indicated differential expression is not significant. (2) The expression level of the housekeeping gene, PGK1, was not significantly different between the two conditions, which is consistent with the microarray data. (3) The expression levels of three hexose transporter genes (HXT1, HXT10, and HXT14) reported here demonstrate another important attribute of this technique. The HXT genes are homologs that exhibit a large degree of similarity at the DNA sequence level (HXT1 is 68% identical to HXT10 over 1,523, bp HXT1 is 56% identical to HXT14 over 608 bp); therefore, we expect that the accuracy and precision of the expression measurements using hybridization based methods may be in doubt. However, using DAGE we were able to target the unique regions with PCR primers to obtain accurate expression measurements. (4) GAL gene—GAL1, GAL2, GAL3, GAL7, GAL10, and GAL80—showed a significantly higher expression in galactose than in glucose. The differential expression of the GAL genes between the two growth conditions shows a much higher range of detection than obtained by the microarray hybridization, which is consistent with the literature (Yuen et al., 2002). This further validates the sensitivity and the range of detection of the DAGE approach.

CONCLUSIONS

The DAGE technique we have developed and discussed here is very well suited to substitute several applications for which traditional gene expression techniques have thus far been acceptable, i.e., Northern analysis and quantitative PCR. DAGE can be used to complement gene expression profiling methods such as SAGE and microarrays; for example, it is particularly useful for confirming transcript differences for low abundant SAGE tags. To achieve

Table II. Digital estimates and confidence intervals.*

Gene	Log ξ_A (gluc)	95% Confidence interval	Log ξ_B (gal)	95% Confidence interval
GAL1	1.72	(1.63, 1.82)	4.32	(4.22, 4.42)
GAL2	1.89	(1.78, 1.99)	5.01	(4.90, 5.12)
GAL3	1.97	(1.83, 2.11)	2.63	(2.49, 2.77)
GAL4	2.10	(2.02, 2.18)	1.74	(1.66, 1.82)
GAL7	1.02	(0.92, 1.12)	3.47	(3.37, 3.57)
GAL10	0.41	(0.13, 0.69)	3.68	(3.40, 3.96)
GAL80	2.15	(2.02, 2.28)	2.84	(2.71, 2.97)
TUP1	2.39	(2.30, 2.48)	2.71	(2.62, 2.80)
PGK1	4.19	(4.00, 4.38)	3.95	(3.76, 4.14)
MIG1	1.05	(0.86, 1.24)	1.40	(1.21, 1.59)
ADE13	2.42	(2.33, 2.51)	2.65	(2.56, 2.74)
HXT1	0.53	(0.26, 0.80)	0.52	(0.25, 0.79)
HXT10	0.42	(0.22, 0.62)	0.36	(0.16, 0.56)
HXT14	0.10	(-0.29, 0.49)	0.58	(0.19, 0.77)

*Log ξ_A = Log₁₀ number of polonies per microgram of glucose first-strand cDNA. Log ξ_B = Log₁₀ number of polonies per microgram of galactose first-strand cDNA.

statistically significant SAGE tag data of low expressed genes a great deal of sequencing is required thus incurring significant costs. DAGE can be substituted and can reduce the cost of sequencing significantly. DAGE can also be used after microarray analysis to confirm gene expression differences for genes that are too close to call for expression differences. Expression profiles of multigene families and splice variants cannot be distinguished using microarrays due to the problems of cross hybridization. In such cases, DAGE can be used to distinguish the gene expression because gene specific and sequence specific primers can be used to analyze gene expression.

DAGE is likely to have a profound impact on the field of functional genomics because of the digital nature of the data wherein the *absolute* or *relative* number of transcripts can be estimated more precisely and accurately. A key limitation of this method is that it is currently “low throughput” in nature. Therefore, efforts are underway to develop a high-throughput polony technology to simultaneously analyze thousands of genes on a single chip as a highly efficient tool for large-scale gene expression studies at the *whole-transcriptome* level.

The authors acknowledge James Butz for helpful discussions and for critically reviewing the manuscript. We also thank Betran Lemieu; and Chris Pridgen for their valuable help at various stages of the project. This work was supported by the University of Delaware Research Foundation (to JSE), a NIH COBRE grant (to JSE), and a U.S. Department of Energy Office of Biological Research Microbial Cell Project grant (to JSE and GMC).

References

- Adams MD. 1996. Serial analysis of gene expression: ESTs get smaller. *Bioessays* 18(4):261–262.
- Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 18(6):630–634.

- Brown PO, Bolstein D, Alizadeh A, Derisi J, Diehn M, Eisen M, Iyer V, Perou C, Pollack J, Ross D, et al. 1998. DNA microarrays as “microscopes” for watching a genome in action. *Mol Biol Cell* 9:2A.
- Brown PO, Botstein D. 1999. Exploring the new world of the genome with DNA microarrays. *Nat Genet* 21(1 Suppl):33–37.
- Butz J, Wickstrom E, Edwards JS. 2003. Characterization of mutations and LOH of p53 and K-ras2 in pancreatic cancer cell lines by immobilized PCR. *BMC Biotechnol* 3(1):11.
- Dudley AM, Aach J, Steffen MA, Church GM. 2002. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc Natl Acad Sci USA* 99(11):7554–7559.
- Elek J, Park KH, Narayanan R. 2000. Microarray-based expression profiling in prostate tumors. *In Vivo* 14(1):173–182.
- Fields S. 1997. The future is function. *Nat Genet* 15:325–327.
- Fislag R. 1998. Differential display approach to quantitation of environmental stimuli on bacterial gene expression. *Electrophoresis* 19(4):613–616.
- Hieter P, Boguski M. 1997. Functional genomics: it’s all how you read it. *Science* 278:601–602.
- Koshland DE Jr. 1998. The era of pathway quantification. *Science* 280(5365):852–853.
- Merritt J, DiTonno JR, Mitra RD, Church GM, Edwards JS. 2003. Functional characterization of mutant yeast PGK1 within the context of the whole cell. *Nucleic Acids Res* 31(15):e84.
- Mitra RD, Butty VL, Shendure J, Williams BR, Housman DE, Church GM. 2003. Digital genotyping with haplotyping with polymerase colonies. *Proc Natl Acad Sci USA* 100(10):5926–5931.
- Mitra RD, Church GM. 1999. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res* 27(24):e34.
- Ramsay G. 1998. DNA chips: state-of-the art. *Nat Biotechnol* 16(1):40–44.
- Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. Serial analysis of gene expression. *Science* 270(5235):484–487.
- Yuen T, Wurmback E, Pfeffer RL, Ebersole BJ, Sealfon SC. 2002. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res* 30(10):e48.

APPENDIX/ONLINE SUPPLEMENT

Statistical Framework for DAGE

One of the primary goals of functional genomics is to provide a quantitative understanding of gene function. Therefore, it is essential to generate high-quality gene expression data and to develop effective statistical tools for gene expression data analysis.

Many techniques have been developed for estimating the differential expression for the gene(s) of interest, each incorporating varying degrees of statistical rigor (Audic and Claverie, 1997; Chen et al., 2002; Ibrahim et al., 2002; Newton et al., 2001; Olshen and Jain, 2002; Rocke and Durbin, 2001). However, the fundamental problems associated with these techniques for generating gene expression data still remain; namely, there are many sources of systematic as well as random errors that ultimately limit the accuracy of the measurement. Hence, even with a prohibitively large number of carefully replicated experiments, the experimental data may be inadequate for estimating the desired expression levels with sufficient confidence. It is generally recognized therefore that effective design and analysis of gene expression data

remains a significant bottleneck. The data generated by the most accurate procedures are still extremely noisy by classic statistical data analysis standards creating significant statistical analysis challenges (Nadon and Shoemaker, 2002; Sebastiani et al., 2003). Developing a rigorous statistical framework for proper analysis of digital gene expression data from DAGE is therefore of prime importance.

The first set of experiments was conducted to establish a formal basis for the statistical analysis of DAGE data. The primary objective was to investigate the suitability of the proposed analysis technique and to characterize the inherent statistical distributional properties. For this purpose, a single gene (*GAL2*) was amplified from cDNA isolated from yeast grown in galactose minimal media. The original cDNA sample was sequentially diluted by 50% each step, and for each dilution, five replicated measurements of the amount of cDNA were performed. The *GAL2* gene was selected because it was very highly expressed in galactose minimal medium. After this characterization, the methodology was then used to determine the gene expression levels for fourteen genes between the conditions of *S. cerevisiae* grown in galactose and glucose minimal media from data gathered as described below.

The experimental sample consisted of first-strand cDNA products (from cells grown in galactose minimal medium) at a concentration of $\sim 5.3 \mu\text{g}/\mu\text{L}$. The gene expression level, denoted by N , is the total number of the cDNA molecules contained per microgram of total first-strand cDNA; this transcript number can, in principle be determined by directly counting the colonies. However, very high concentrations of cDNA do not always yield easily discernible colonies due to colony overlap. It was therefore often necessary to dilute (depending on the gene and the growth condition) the original sample down to an appropriate level at which the colonies could be distinguished and counted reliably. The number of colonies in the original undiluted sample can then be derived from the diluted sample count by accounting for the extent of dilution. To systematize this strategy, we employed the following experimental procedure: the original sample was first diluted to 50% of the initial concentration, giving rise to samples containing 50% fewer transcripts than the original undiluted sample. The samples resulting from this first dilution were then themselves subsequently diluted, again by 50%, to obtain second-generation samples. This was repeated sequentially for several more generations with each sequential generation diluted to 50% of the preceding one. Observe that theoretically each dilution generation should have 50% fewer transcripts than the previous one. For our characterization experiments, we employed five sequential cDNA dilutions beginning from the highest concentration that yielded discernible colonies. At each dilution, the experiment was replicated five times; and in every case, the experimental sample was divided into two sets, A and B, providing duplicate measurements of the transcript number for each replicate.

Framework for Statistical Analysis

The basic principle of the technique is that if N is the total number of cDNA molecules (per μg of total first strand cDNA) of the specific gene in the original sample, then at the i^{th} dilution in the sequence of dilutions, the observed polony count (per microgram of total first-strand cDNA), y_i is given by (assuming 100% efficiency of reverse transcription and PCR polony generation)

$$y_i = d_i^{-1}N \quad (\text{S1})$$

where d_i is the i^{th} dilution factor, the extent to which the original number of transcripts have been reduced via i sequential dilutions; its numerical value is obtained from

$$d_i = 2^{i-1}. \quad (\text{S2})$$

For example, for the original sample, $i = 1$, so that $d_i = 1$ and $y_i = N$; for the second dilution, where the analyzed sample has been diluted by 50%, $i = 2$, $d_i = 2$, and $y_i = \alpha_i N$.

Observe that, as a result of our experimental strategy, the dilution factor sequence d_i is *exponential* in i , the dilution

number [see Eq. (S2) above] motivating a logarithmic transformation of Eq. (S1) to obtain:

$$\log_{10} y_i = -\log_{10} d_i + N \log_{10}. \quad (\text{S3})$$

This is the underlying working model to use in determining the desired gene expression level, N , from the observed polony count data. If this model is adequate for the methodology, then a log–log plot of the data y_i versus the dilution factor d_i should yield a straight line with slope -1 and an intercept of N (the undiluted transcript count).

Model Adequacy

Starting with an original sample with a concentration of $5.26 \mu\text{g}$ of total cDNA per μL of solution, we collected experimental data as discussed above. The raw data consisted of polony counts that were then normalized to obtain N_i , the count per microgram of total first-strand cDNA.

Supplementary Figure 1 shows a log–log plot of the data y_i versus the dilution factor d_i , along with a summary of regression analysis results. First, linear regression analysis yielded estimates of the slope and intercept respectively as -0.995 and 5.066 . Subsequently, a hypothesis test that the

Characterization: Gal2-Galactose

$$\log y = 5.06632 - 0.994878 \log d$$

$$S = 0.165813 \quad R\text{-Sq} = 88.8 \% \quad R\text{-Sq}(\text{adj}) = 88.5 \%$$

Supplementary Figure 1. Characterization of Gal2 gene expression in galactose minimal medium. The graph shows the effect of dilutions on the polony count. Each data cluster shows normalized polony count obtained at the indicated dilution. Log d , $\log_{10}$ of the dilution factor; Log y , $\log_{10}$ of the normalized count per microgram of cDNA.

Histogram of $\log N^{\wedge}$, with Normal Curve

Supplementary Figure 2. Distributional characteristics of $\log_{10} \xi_i$ data. The P value for the normality test is 0.124, indicating that the data distribution is reasonably normal.

observed Slope is *significantly different* from the expected theoretical value of -1 yielded a P value of 0.924, indicating

that the estimated slope is not significantly different from the expected value of -1 and confirming the adequacy of the model in Eq. (3) for describing the methodology.

Supplementary Table I. Primers used in Digital Analysis of Gene Expression

Genes	Primer Sequences
GAL1YBR020W	GCACAATCCTTGAATTGTTCTCGCG CGGGAACCATATGATCCATTTGACA
GAL2YLR081W	TTCACCCATTATCACATCTGCC CTAGCATGGCCTTGTAACACGGTT
GAL3YDR009W	GGCGCTAAACTGTTACGTCGAGGA CCGAAAGAACCATTGCTAGGGCA
GAL4YPL248C	CATCCCTGTAGTGATTCCAAACGCG CATTAGTGCCACTGACCCCGTCTG
GAL7YBR081C	GTAGCTGATCTCAGTAAAGGTGGG TGTTCCATACTGGGCCATCTGGC
GAL10YBR019C	GGAATCGGGATGAAAAGCCTTGAC GCCATATGGAGACACTATTGAGGG
GAL80YML051W	ATGAGCGTGGTAACCGATTGGGC TGGCTAGCGGGAAGTCGTTTGC
TUP1YCR084C	CGTTATATTCTGGACCCAGCGGAGA GGACAAAGCGTTGTATCCGGCTCA
PGK1YCR012W	CGGTGACTCCATCTTCGACAAGGC TGACGGTGTACCAGCAGCAGAGC
MIG1YGL035C	CTACTTAGCATTGTCTGGGGCTGG TGAAACTGAACGCGTTATCGTCCC
ADE13YLR359W	AAACTGCATCCGTTCATGTTTCG CCTTGACTACTGCTGCGGCTTGAT
HXT1YHR094C	GACCATACCGACAGCACCCACAT TCGAAGAAATGAGAGCCGCTGGTA
HXT10YFL011W	ATCAATACAGTGGCGGTTCCCT CTGGCATTTCCAACAGCCCTTCT
HXT14YNL318C	TCCGTTTCTTCCAAATGTCCCC

It should be noted that the regression analysis also provides an estimate of the desired expression level, $\log_{10} N$, as 5.066. Additionally, the regression analysis identified two of the observations at $\log d = 4.52$ as outliers because of large standardized residuals; this will be of importance at the next stage of the analysis. We conclude from both the general qualitative indication of a linear relationship and the quantitative confirmation that the slope of -0.995 was not significantly different from -1.0 , that the model in Eq. (S3) is adequate.

Distributional Characteristics

The second part of our objective was to characterize the distributional properties of the data and the implied error statistics associated with the experimental technique and the analytical procedure. For this purpose we return to the model underlying the analytical procedure, as derived above in Eqs. (S1)–(S3). By defining

$$\xi_i = y_i d_i \quad (\text{S4})$$

(the expression level as determined directly from each individual observation, accounting for the extent of dilution), then from Eq. (S3), the associated experimental error, ε_i is given by

$$\log_{10} \xi_i = \log_{10} N + \varepsilon_i \quad (\text{S5})$$

the deviation of the individual “observed” quantity $\log_{10} \xi_i$ from the true—but unknown—expression level ($\log_{10} N$).

A histogram of the $\log_{10} \xi_i$ data with a normal curve superimposed is shown in Supplementary Figure 2. The summary statistics of the data are as follows: mean = 5.11, (95% confidence interval [5.09, 5.13]); standard deviation = 0.093, (95% confidence interval [0.077, 0.118]). With a p -value of 0.124 for the normality test, the implication is that the data distribution is not significantly different from a theoretical normal distribution with the indicated mean and standard deviation.

Furthermore, a one-way ANOVA of $\log_{10} \xi_i$ data versus dilution number yielded a p -value of 0.667 for data set A and a value of 0.570 for data set B, from which we conclude that there is no significant systematic effect of dilution number. A similar analysis of the data versus the replication number yielded a p -value of 0.549 for data set A and a p -value of 0.583 for data set B, from which we again conclude that there is no significant systematic effect of replication number.

References

- Audic S, Claverie JM. 1997. Detection of eukaryotic promoters using Markov transition matrices. *Comput Chem* 21(4):223–227.
- Chen Y, Kamat V, Dougherty ER, Bittner ML, Meltzer PS, Trent JM. 2002. Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* 18(9):1207–1215.
- Ibrahim JG, Chen M-H, Gray RJ. 2002. Bayesian models for gene expression with DNA microarray data. *J Am Stat Assoc (JASA)* 97: 88–99.
- Nadon R, Shoemaker J. 2002. Statistical issues with microarrays: processing and analysis. *Trends Genet* 18(5):265–271.
- Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. 2001. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol* 8(1):37–52.
- Olshen AB, Jain AN. 2002. Deriving quantitative conclusions from microarray expression data. *Bioinformatics* 18(7):961–970.
- Rocke DM, Durbin B. 2001. A model for measurement error for gene expression arrays. *J Comput Biol* 8(6):557–569.
- Sebastiani P, Gussani E, Kohane IS, Ramoni MF. 2003. Statistical challenges in functional genomics. *Stat Sci* (in press).

RecA mediated initial alignment of homologous DNA molecules displays apparent first order kinetics with little effect of heterology[☆]

Shwetal Patel^{*,1}, Jeremy S. Edwards

Department of Chemical Engineering, University of Delaware, Newark, DE 19716, USA

Received 4 August 2003; accepted 12 September 2003

Abstract

The mechanism and determinants of RecA mediated initial alignment of homologous DNA molecules were studied by performing Monte Carlo simulations of the dynamics of DNA molecules. The simulation procedure was used to assess the effect of heterologous DNA and dilution on the rate of formation and yield of homologous alignments. The results show that the apparent first order kinetic behavior and the impact of heterologous DNA, reported in literature [J. Biol. Chem. 261 (1986) 1025], can be observed even if the conversion of the initially aligned molecules into a stable joint is not rate-determining. The present study is the first step towards developing rigorous computational models to describe the process of homologous recombination, and theoretical frameworks to retrieve biophysical parameters of strand pairing and exchange proteins from in vitro assays of joint molecule formation.

© 2003 Elsevier B.V. All rights reserved.

Keywords: Homology-search; Recombination; Simulation; Nucleoprotein-dynamics; Bond-fluctuation Monte Carlo

1. Introduction

Early studies attempting to decipher the mechanism of RecA mediated pairing of homologous DNA molecules contended between two possibilities: RecA protein may move two DNA molecules relative to each other until homologous sequences come into alignment or pairing may result from random contacts between DNA molecules [1,2]. A systematic study to determine the kinetics of the strand pairing reaction revealed that a first order rate form could adequately describe it and that the rate constant did not change significantly upon two-fold dilution [3]. Moreover, the calculated first order rate constant and the final yield of joint molecules were found to be independent of the length of heterologous dsDNA tail as well as added heterologous dsDNA. These results were in stark contrast to earlier observations of Gonda

and Radding [4]. Julin et al. [3] mentioned that the differences in the results obtained in their study and those in earlier studies of Radding and co-workers may partly be due to the different reaction conditions, sources of DNA substrates and RecA protein, and, concentrations employed. However, in order to be conclusive one needs to know the sensitivity of the pairing reaction to changes in binding affinities of nucleoprotein filaments to homologous and heterologous duplex DNA and, the concentrations of the various participating species. Therefore, the first aim of this study is to elaborate on the effect of the nature of heterologous contacts and dilution on homologous pairing.

The formation of stable paranemic joints is a complex reaction involving RecA coated ssDNA and dsDNA molecules that can be represented by the following general reaction scheme:

Julin et al. [3] argued that the kinetics of this mechanism is first order if the equilibrium for complex formation lies well to the right, that is, if most of the added DNA substrates

[☆] Supplementary data associated with this article can be found at doi: 10.1016/j.dnarep.2003.09.009.

* Tel.: +1-214-648-4098; fax: +1-214-648-4067.

E-mail address: spate6@mednet.swmed.edu (S. Patel).

¹ Present address: Department of Pathology, The University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390, USA.

form this relatively stable complex rapidly and the reaction rate is that at which the complex forms a D-loop (denoted by J) with first order rate constant, k . The reaction should become second order at very low DNA concentrations. Two possible structures for the ssDNA:RecA:dsDNA intermediate can be envisaged. The DNA molecules could be held together non-specifically by the protein, the rate-determining step being the first-order conversion of the complex to a form in which the DNA molecules are aligned homologously in a configuration that is stable to, and thus detected by, the conditions of the nitrocellulose filter-binding assay. Alternatively, the DNA molecules in the complex might already be aligned at homologous sites, the rate-determining step being conversion to a stable paranemic joint. Formation of this unstable intermediate complex could occur by either random collisions or by a facilitated equivalent in which only stably paired structures are scored. Based on the observed invariance of the first order rate constant with increasing lengths of the heterologous tail on homologous dsDNA and the addition of heterologous dsDNA, Julin et al. [3] concluded that the reaction intermediate is comprised of homologously aligned molecules and that the observed reaction rate was determined solely by the conversion of this initial alignment into a form which is stable under the conditions employed in nitrocellulose assays. Because the in vitro experimental procedure employed in the above studies monitors the formation of stable joint molecules (J or D-loop), it does not allow the mechanism and determinants of the initial homologous alignment process to be decoupled from those responsible for its subsequent conversion into a stable joint. Therefore, the second aim of the present study is to investigate the process of initial homologous alignment, i.e., the formation of the ssDNA:RecA:dsDNA complex.

For the purpose of modeling the process of homologous alignment in vitro, the system is taken to be comprised of DNA molecules undergoing thermally induced random (Brownian) motion with energetic interactions between nucleoprotein filaments and nearby heterologous and homologous dsDNA molecules. The problem reduces to a tractable one that may be solved using methods of theoretical polymer physics. The formulation presented in this paper is the first step towards developing rigorous computational models to describe the process of homologous recombination and, theoretical frameworks to calculate biophysical parameters such as binding energetics of nucleoprotein filaments to heterologous duplex (ϵ_{het}) and homologous duplex (ϵ_{hom}) DNA and the kinetic rate parameter, k , of classical strand exchange proteins (RecA and Rad51) as well as novel ones (Rad51B and Rad51C) [5], from in vitro assays of D-loop formation.

2. Simulation of DNA dynamics

In this study, the bond-fluctuation Monte Carlo method is employed to describe DNA dynamics [6]. A DNA chain is

taken to consist of segments confined on a lattice. The size of each segment and hence that of a unit cell on the lattice is equal to the Kuhn's segment (persistence) length. Experimental studies have shown that nucleoprotein filaments formed by polymerization of RecA on ss or dsDNA display large flexibility with respect to bending, stretching and, twisting. The measured persistence length of RecA-dsDNA filaments has been estimated to be ~ 600 nm by electron microscopy [7], and either ~ 210 nm [8] or ~ 950 nm [9] using micromanipulation; about 4–15 times larger than that of dsDNA. Therefore, in our simulations, the size of the unit cell in the cubic lattice is taken to be the Kuhn's segment length of dsDNA, i.e. 50 nm.

There is now ample experimental evidence that RecA coated single-stranded DNA makes transient contacts with heterologous DNA leading to the formation of nucleoprotein networks [10]. Within these networks the search for homology is greatly facilitated and leads to the formation of strong homologous pairs that are subsequently converted into joint molecules. Transient interactions between nucleoprotein filaments and dsDNA involve the winding of dsDNA in the helical groove of the nucleoprotein filament. The close proximity of the bases of ssDNA and dsDNA within the RecA protein allows energetic interactions to take place. Because the size of each segment in the simulations represents approximately 150 bp, the interaction energies are the sum of individual contributions from the participating base-pairs. These interactions are incorporated into the simulation by defining interaction energies between the nucleoprotein filament and heterologous dsDNA, ϵ_{het} , and homologous dsDNA, ϵ_{hom} . Near-neighbor interactions between nucleoprotein filaments and dsDNA are taken to occur in all six directions. As a result, multiple homologous and heterologous contacts between a segment of a nucleoprotein filament and neighboring dsDNA are allowed and scored. The details of the simulation procedure and tables are given in the accompanying supplement.

3. Results

The simulations were performed on a cubic lattice with periodic boundaries. Two sizes of the simulation box were considered: $40 \times 40 \times 40$ (denoted by V_1) and $80 \times 80 \times 80$ (denoted by V_2). The simulations provide the number of joint molecules, $[J]$, formed as a function of the simulation time, Monte Carlo steps (attempts) per monomer (MCS/ N).

To test the hypothesis that the process of initial homologous alignment displays a first order rate law, the simulated data of $[J]$ versus MCS/ N were fitted (using the routine *lmdr* of MINPACK [11]) to a mono-exponentially increasing function,

$$[J] = [J]_1 - [J]_2 e^{-k(\text{MCS}/N)} \quad (2)$$

where for first order kinetics, $[J]_1 = [J]_2$, the number of joint molecules formed at infinite time.

Fig. 1. The number of homologous alignments (joint molecules), $[J]$, as a function of the simulation time, MCS/N (corresponding estimates of the real time, t (minutes), are shown on separate axes) for case I (control; see text) for cubic box sizes, 40 (V_1 ; top) and 80 (V_2 ; bottom). The statistics were computed from the results of 30 independent trials. Eq. (2) (smooth lines) adequately describes the data and, moreover, the formation of homologous alignments displays apparent first order kinetics, i.e., $[J]_1 \approx [J]_2$. The rate of formation as well as the yield of joint molecules increases significantly with an increase in the homologous interaction energy, ϵ_{hom} . Dilution (V_2) results in lower rates of formation and yields of joint molecules.

The process of initial homologous alignment (and also the formation of joint molecules in this study) was simulated by assigning finite values to the heterologous and homologous interaction energies. In the simulations, the degree of homology between nucleoprotein filaments and dsDNA is limited to the first two segments (294 bp). The results obtained from independent initial configurations were fitted to Eq. (2) to obtain $[J]_1$, $[J]_2$, and, k . Because of the inherent randomness in the simulations, not all outcomes could be adequately fitted and a visual inspection of the data revealed obvious outliers that were not included in computing the statistics.

Simulations of the process of homologous alignment were performed for three cases. I (control): 10 molecules of 40 segment nucleoprotein filaments + 10 molecules of 20 seg-

ment partially homologous dsDNA. II (increased length of heterologous tail on dsDNA): 10 molecules of 40 segment nucleoprotein filaments + 10 molecules of partially homologous 40 segment dsDNA. III (added heterologous dsDNA): 10 molecules of 40 segment nucleoprotein filaments + 10 molecules of 20 segment partially homologous dsDNA + 10 molecules of 40 segment heterologous dsDNA. Ten DNA chains within a cubic simulation box of size 40 corresponds to a concentration of ~ 2000 pM and for a box of size 80, ~ 250 pM. For the purpose of comparison, the concentrations of individual species (ssDNA and dsDNA) employed in *in vitro* studies are ~ 500 – 800 pM [3] and ~ 50 – 80 pM [4]. Due to limitations in the lattice Monte Carlo technique employed, equivalence of concentrations does not imply equivalence of volume fractions.

Fig. 2. The number of joint molecules, $[J]$, (top) and apparent first order rate constant, k (N/MCS), (bottom) for a cubic simulation box of size 40 (V_1) as a function of the heterologous interaction energy, ϵ_{het} (units of $k_B T$) for two values of the homologous interaction energy, ϵ_{hom} . The statistics were computed from the results of 30 independent trials. Case I (control), case II (increased length of heterologous tail on dsDNA) and case III (added heterologous dsDNA) (see text).

A typical plot of the formation of homologous alignments (joint molecules), $[J]$, as a function of the number of Monte Carlo attempts per monomer (simulation time), MCS/N , is shown in Fig. 1 for case I (control). For concentrated solutions (V_1) (Fig. 1: top panel), complete pairing is achieved in 2.5×10^5 MCS/N or 20 min (1 $MCS/N \sim 5$ ms; see supplement), comparable to the times for complete pairing required in experimental studies. In the study reported by Julin et al. [3], the pairing reaction was completed in about 5 min. Note that the rate of formation and yield of paired molecules increases in direct proportion to the extent of homology between nucleoprotein filaments and dsDNA providing justification for the difference between the observed times. The figure also shows the adequacy of Eq. (2) in describing the data. Moreover, the curves display apparent first order behavior, i.e. $[J]_1 \approx [J]_2$. For dilute solutions (V_2) (Fig. 1: bottom panel), the DNA chains have to diffuse longer distances before making fruitful contacts resulting not only in slower kinetics but lower yields of aligned molecules.

For concentrated solutions, heterologous interaction energy (ϵ_{het}), length of heterologous tail on dsDNA, and, addition of heterologous dsDNA do not have a significant impact on the rate of formation and yield of homologous alignments (Fig. 2). These results are in qualitative agree-

ment with those reported by Julin et al. [3]. Based on the general reaction scheme for the pairing reaction (Eq. (1)), Julin et al. [3] argued that their observations could be explained only if the conversion of the initial homologous alignment into stable joint molecules was rate limiting. Because in our simulations the process of initial alignment is of interest, we assumed that its conversion into a stable joint molecule is instantaneous. Therefore, the apparent first order kinetics and the effect of heterology on joint molecule formation, reported in experimental literature, can be observed even if the conversion of the initial homologous alignment into a stable joint molecule is not rate-determining or slow. As expected, an increase in the energy of homologous interactions, ϵ_{hom} , has a significant effect on the rate of formation and yield of homologous alignments.

The qualitative behavior of the initial alignment reaction was found to remain the same after an eight-fold dilution (Fig. 3). A quantitative comparison of the results for concentrated (Fig. 2) and dilute (Fig. 3) solutions reveals that the first order rate constant as well as the yield of homologous alignments decreases significantly upon dilution. The small change in the kinetic parameter upon two-fold dilution observed by Julin et al. [3] may have resulted from the

Fig. 3. The number of joint molecules, $[J]$, (top) and apparent first order rate constant, k (N/MCS), (bottom) for a cubic simulation box of size 80 (V_2) as a function of the heterologous interaction energy, ϵ_{het} (units of $k_B T$) for two values of the homologous interaction energy, ϵ_{hom} . The statistics were computed from the results of 30 independent trials. Case I (control), case II (increased length of heterologous tail on dsDNA) and case III (added heterologous dsDNA) (see text).

insignificant impact of the low extent of dilution on the rate of initial alignment of joint molecules.

The insensitivity of the apparent first order rate constant, k , to homologous and heterologous interaction energies and the decrease in its value upon dilution strongly suggests that the process of initial alignment is dominated by diffusion of the DNA segments until they are in close proximity. Once in close proximity, the balance between the interaction energy between the segments and random thermal (diffusive) motion determines their permanence; strong homologous interactions are favored. Therefore, the yield of joint molecules increases with an increase in the homologous interaction energy.

In the above simulations the nucleoprotein filaments were taken to be linear. The classical in vitro model for homologous recombination, employed in the experimental studies mentioned above, involves the exchange of strands between circular single-stranded DNA and its homologous duplex. Further simulations show that the mechanism and determinants of the process of initial homologous alignment for circular nucleoprotein filaments are not different from those for linear nucleoprotein filaments (see supplement).

4. Discussion

To keep the analysis simple, only excluded volume interactions were accounted for, i.e., the DNA chains were comprised of hard segments. All long-range molecular, electrostatic, and, hydrodynamic forces were neglected. Moreover, the bond-stretching and torsion energies were not considered. Because only the role of heterologous and homologous interactions between nucleoprotein filaments and double-stranded DNA was of interest in this study, the above assumptions do not have serious implications on the results and may be easily incorporated (except hydrodynamic interactions), if required.

According to the commonly accepted model for the interaction of nucleoprotein filaments with dsDNA, RecA aids in partially unwinding the duplex and samples the bases by binding them to a second site in the protein (the first site is bound to single-stranded DNA). In the simulations, near-neighbor interactions with dsDNA occur in all six directions. As a result, multiple homologous alignments between a nucleoprotein filament and double-stranded DNA are allowed and scored. In reality, only one dsDNA segment may bind to a nucleoprotein filament segment. However, to avoid introducing a bias into the study and also keeping in mind the statistical nature of the simulation, allowing fictitious interactions in all directions is not a severe limitation. Because the resolution of the current procedure is limited by the simulation cell size, the length of the triplex structures formed during the sampling process is equal to the Kuhn's segment length of dsDNA. To account for shorter triplex structures, the size of the simulation cell may be reduced to the appropriate size.

In conclusion, the present study shows that simulations of the dynamics of interacting DNA chains is a useful tool to study the process of RecA mediated pairing of homologous DNA molecules. The results of the simulations performed in this study elucidate the mechanism and determinants of the process of initial homologous alignment. This study is the first step towards developing rigorous computational models to describe the process of homologous recombination and, theoretical frameworks to retrieve biophysical parameters of strand pairing and exchange proteins from in vitro assays of joint molecule formation.

Acknowledgements

The author would like to thank Dr. Norman J. Wagner and Dr. Dionisios G. Vlachos for useful discussions on the Monte Carlo simulation technique. The generous support from US. Department of Energy's Office of Biological and Environmental Research to Dr. Jeremy S. Edwards is greatly appreciated.

References

- [1] D.K. Gonda, C.M. Radding, By searching processively RecA protein pairs DNA molecules that share a limited stretch of homology, *Cell* 34 (1983) 647–654.
- [2] P. Howard-Flanders, S.C. West, A. Stasiak, Role of RecA protein spiral filaments in genetic recombination, *Nature* 309 (1984) 215–220.
- [3] D.A. Julin, P.W. Riddles, I.R. Lehman, On the mechanism of pairing of single- and double-stranded DNA molecule by the RecA and single-stranded DNA-binding proteins of *Escherichia coli*, *J. Biol. Chem.* 261 (1986) 1025–1030.
- [4] D.K. Gonda, C.M. Radding, The mechanism of the search for homology promoted by RecA protein, *J. Biol. Chem.* 261 (1986) 13087–13096.
- [5] Y.C. Lio, A.V. Mazin, S.C. Kowalczykowski, D.J. Chen, Complex formation by the human Rad51B and Rad51C DNA repair proteins and their activities in vitro, *J. Biol. Chem.* 24 (2003) 2469–2478.
- [6] I. Carmesin, K. Kremer, The bond fluctuation method: a new effective algorithm for the dynamics of polymers in all spatial dimensions, *Macromolecules* 21 (1988) 2819–2823.
- [7] E.H. Egelman, A. Stasiak, The structure of helical RecA-DNA complexes. I. Complexes formed in the presence of ATP- γ -S or ATP, *J. Mol. Biol.* 191 (1986) 677–697.
- [8] G.V. Shivshankar, M. Feingold, O. Krichevsky, A. Libchaber, RecA polymerization on double-stranded DNA by using single-molecule manipulation: the role of ATP hydrolysis, *Proc. Natl. Acad. Sci. USA* 96 (1999) 7916–7921.
- [9] M. Hegner, S.B. Smith, C. Bustamante, Polymerization and mechanical properties of single RecA-DNA filaments, *Proc. Natl. Acad. Sci. USA* 96 (1999) 10109–10114.
- [10] C.M. Radding, Helical interactions in homologous pairing and strand exchange driven by RecA protein, *J. Biol. Chem.* 266 (1991) 5355–5358.
- [11] J.J. More, B.S. Garbow, K.E. Hillstrom, User Guide for MINPACK-I. Argonne National Laboratory Report ANL-80-74, Argonne, IL, 1980.

Brief Report

An integrated mechanistic model for transcription-coupled nucleotide excision repair[☆]Shwetal Patel^{a,*,1}, K.V. Venkatesh^b, Jeremy S. Edwards^a^a Department of Chemical Engineering, University of Delaware, Newark, DE 19716, USA^b Department of Chemical Engineering, Indian Institute of Technology, Bombay, Powai, Mumbai-400076, India

Received 9 October 2003; accepted 21 November 2003

Abstract

Preferential repair of the transcribed strand of active genes is usually attributed to a coupling protein that dislodges RNA polymerase stalled at a damage site and recruits repair enzymes. Experimental observations of the effect of transcription on preferential repair in *Escherichia coli* are contradictory and inexplicable by this model. In this study, it is argued that the multiple conformations displayed by a stalled RNA polymerase result in two sub-pathways for repair: Mfd coupled and direct. Together with the fact that RNA polymerase recruits the repair enzymes in a promoter dependent manner, an integrated mechanistic model is proposed that is capable of explaining the effect of transcription on preferential repair reported in literature. The quantitative behavior of the model is illustrated by describing the various reactions using a biochemical network. The implications of the model on the mechanism for transcription-coupled repair in higher organisms are briefly discussed.

© 2003 Elsevier B.V. All rights reserved.

Keywords: Cockayne syndrome; In silico biology; Hypothesis; UvrABCD

1. Introduction

Transcription-coupled repair, the rapid and preferential mending of actively transcribing genes, is an important mode of DNA excision repair [1–7] defects in which are responsible for the symptoms of Cockayne syndrome [8,9]. Preferential repair of actively transcribing genes in *Escherichia coli* is brought about by Mfd protein that dislodges a stalled RNAP and recruits repair enzymes (UvrA₂B) [10]. The molecular mechanism of this pathway is remarkably conserved exhibiting mechanistic similarity across the living kingdom. Experiments in *E. coli* have revealed two important properties of this pathway that are contradictory and inexplicable by the current model: (i) for the *lacZ* gene, Mfd is required at low levels of transcription but not at elevated levels of transcription [11]; and (ii) for the *tryT* gene, induction by the transcriptional activator Fis suppresses preferential repair

[12]. In this report, an integrated mechanistic model that accounts for events at the promoter as well as at a stalled RNA polymerase is proposed that is capable of explaining the above observations. Because the model involves complex interactions between several factors, the quantitative behavior of the model is illustrated by using a biochemical network.

2. Model

Factors responsible for preferential repair of the transcribed strand arise during transcription initiation. DNA supercoiling and footprint analysis have revealed that the binding of RNAP to λP_L promoter provides UvrA₂B complex with a preferred binding site on the non-transcribed strand resulting in the activation of 5′–3′ helicase activity [13]. Moreover, crosslinked immunoprecipitates of isolated membrane-associated nucleoids from *E. coli* show that UvrA is specifically associated with the β subunit of RNAP [14]. Based on these results and further studies on dimer excision, Ahn and Grossman suggested that upon recruitment of UvrA₂B to the transcription bubble at the promoter region, a competent nucleoprotein complex for helicase action is formed that translocates along the non-transcribed

[☆] Supplementary data associated with this article can be found at doi: 10.1016/j.dnarep.2003.11.010.

* Tel.: +1-214-648-4098; fax: +1-214-648-4067.

E-mail address: spate6@mednet.swmed.edu (S. Patel).

¹ Present address: Department of Pathology, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, TX 75390, USA.

Fig. 1. The presence of a dimer on the transcribed strand causes RNAP to stall. Subsequently, RNAP alternates between multiple backtracked conformations some of which are substrates for Mfd (stalled) while others allow direct binding of UvrA₂B to dimer (free), i.e., there are two sub-pathways for repair. Upon binding to DNA and stalled RNAP, Mfd protein pushes RNAP molecule off the DNA and recruits the UvrA₂B complex to initiate nucleotide excision repair. The free conformation allows UvrA₂B to bind directly to the dimer. The number of UvrA₂B molecules available at a dimer depends on its recruitment to the transcription bubble at the promoter through physical interaction with RNAP.

strand in an ATP hydrolysis-dependent reaction [13] (leading the actively transcribing RNAP) and targets upstream dimers on the transcribed strand. However, the nature of the interaction between RNAP and UvrA₂B is yet unclear, particularly in the regions away from the promoter.

In general, the presence of transcriptional activator or repressor bound to the promoter may alter the ability of RNAP to recruit UvrA₂B to the transcription bubble by either physically preventing its access to promoter regions (e.g., repressor) or by altering the topology of the promoter (e.g., activator). Upon encountering a lesion (here dimer) on the transcribed strand, RNAP stalls and adopts multiple backtracked conformations (simple and arrested) [15]. These conformations are reversible and it is thought that RNAP undergoes one-dimensional “diffusion-like” motion along a restricted length of DNA bounded by the dimer on one end [16]. Reversibility of some backtracked conformations may involve simple isomerization while for others, transcript cleavage and re-initiation by accessory proteins such as GreB. However, even if the backed off RNAP reattempts an energetic forward motion to resume RNA synthesis, it will be rendered inactive upon encountering

the dimer. Therefore, the stalled RNAP alternates between the various conformations. Some of the backtracked conformations are substrates for Mfd that pushes RNAP forward resulting in activation or detachment [17]. Here, two sets of conformations are considered: stalled and free.

Upon binding to the stalled conformation, Mfd pushes the RNAP molecule off the DNA and recruits the complex UvrA₂B through physical association [10]. Here, it is hypothesized that the free conformations allow the dimer to be bound directly by UvrA₂B. The number of UvrA₂B complexes available at the dimer that has stalled a RNAP is determined by their recruitment to the transcription bubble at the promoter through physical association with RNAP. Therefore, there are two sub-pathways for repair and the choice between the pathways is determined by events at the promoter as well as the competing events occurring at a dimer (Fig. 1).

3. Results

To illustrate the quantitative behavior of the model, a mathematical description of the biochemical reactions in-

here with the recruitment of UvrA₂B, i.e., for low values of β , repair is predominantly through the Mfd pathway (compare left two panels and right two panels of Fig. 3): repair of the transcribed strand is much slower than that of the non-transcribed strand in mutant strains. Moreover, an increase in transcription has an insignificant effect on the rate of repair (compare top two panels and bottom two panels of Fig. 3). Promoters with higher values of β require Mfd at low levels of transcriptional activation (left two panels of Fig. 3). For a value of β in this range, an increase in the levels of transcription has the potential to obviate the requirement of Mfd (compare right two panels of Fig. 3). It is likely that most genes operate with values of β in this range. Simulations of the time-course of repair for a value of β in this range compare well with experimental data for the *lacZ* gene reported in literature (Fig. 4). In the parametric study reported above, the value of β was a constant. In reality, β is a function of the level of transcriptional activation. This notion opens the possibility of studying the process of recruitment of the repair enzymes to promoters with a variety of repressors and activators in detail using experimental as well as computational tools [19]. Furthermore, β is likely to

Fig. 3. Performance of the biochemical network (Fig. 2) with the recruitment factor, β , as a parameter. Four cases are considered comprising a combination of uninduced (Un) and induced (In) gene in the wild-type (mfd^+) and mutant (mfd^-) strains. The fractional repair is computed as its mean value along the length of the transcribed strand. The fractional repair on the non-transcribed strand is 0.56 for all cases. There is a window of values of the recruitment parameter, β , for which comparable repair of the wild-type and mutant strains may be achieved in the induced state but not in the uninduced state. For these values of β , the fractional repair shows the strongest dependence on the level of transcriptional activation. In realistic promoters, the recruitment factor is a function of the level of transcriptional activation. For the *lac* promoter, induction removes bound repressor and, therefore, leads to an increase in the value of β , resulting in faster repair by the direct pathway in the induced state. For the *tryT* promoter, induction by the binding of Fis reduces the value of β , resulting in slower repair in the induced state via the Mfd pathway. Initial dimer concentration: $[d]_0 = 0.005$ nM/bp (thin lines) and $[d]_0 = 0.05$ nM/bp (thick lines). (Solid lines) $K^* = 10^2$ nM $^{-1}$; (dashed lines) $K^* = 10^4$ nM $^{-1}$; $t = 20$ min; 0.05 nM/bp ≈ 1 dimer in every 80 nucleotides.

decrease as RNAP moves along the gene due to detachment. Together with the reduction in recruitment of the repair enzymes to the transcription bubble at the promoter upon binding of the transcription activator Fis, the observed slower repair in regions downstream of the transcription start site upon transcriptional activation of the *tryT* gene [12] can be explained by the proposed model.

4. Discussion

The proposed model can be rigorously validated through well-designed experiments. For example, controlled expression of UvrA using plasmid constructs can increase the concentration of UvrA₂B (UvrB is in excess of UvrA) leading to an increase in the recruitment of this complex to the

transcription bubble at the promoter as well as availability of this complex to dimers through non-specific DNA binding. Another interesting possibility is to mutate RNAP or UvrA so as to alter their affinity for each other and thereby affect the efficiency with which the complex is recruited to the transcription bubble. The above model also suggests that the mechanism of transcription-coupled repair may be determined by the properties of the promoter and the manner in which transcription is regulated. Two promoters, resulting in different mechanisms of repair were mentioned above, namely, *lac* and *tryT*. Further studies in other promoters will help elucidate the role of the promoter in determining the mechanism of transcription-coupled repair.

The existence of an alternative pathway for preferential repair is mentioned in the works of Ahn and Grossman [13] and Grossman and co-workers [14]. This pathway relies on

Fig. 4. Comparison of model calculations with experimental data for transcription-coupled repair of the *lacZ* gene reported in literature. Model calculations: fractional repair vs. time (min) on the non-transcribed (thin lines) and transcribed (thick lines) strands. Experimental data ((*) non-transcribed strand; (+) transcribed strand [11]): Region B, extending 150bp after the mRNA start site. A value of β for which the fractional repair is most sensitive to transcriptional activation was chosen, i.e., $\beta = 10 \text{ nM}^{-1}$. The values of other parameters are: $K^* = 10^4 \text{ nM}^{-1}$; $[d]_0 = 0.025 \text{ nM/bp}$. The model clearly captures the fact that Mfd is required in the uninduced state (left two panels) but not required for preferential repair in the induced state (right two panels).

the enhanced delivery of UvrA₂B due to interaction and translocation with RNAP but still requires Mfd to dislodge the stalled RNAP.

The recent observation that conformation modulating sub-units (Rpb4 and Rpb9) of RNA Pol II in *Saccharomyces cerevisiae* are capable of inducing a switch between Rad26 (Mfd pathway) and direct repair pathways [22] can be explained by the above model and unknown proteins need not be invoked, i.e., it is likely that these conformation modulating sub-units alter the distribution of the stalled and free states. The recruitment of the repair enzyme complex to the transcription bubble at the promoter region through physical association with RNAP appears to have a strong analogy with the loading of TFIIH (a basal transcription factor and an important NER protein [23,24]). TFIIH is usually jettisoned after about 30–60 nucleotides downstream from the transcription start site [6] and, is therefore, implicated in preferential repair in the immediate vicinity of the start site that does not require Rad26 (Mfd) [25]. A similar phenomenon has been noted in vitro using human nucleotide excision repair proteins [26] although no preferential repair

was observed in the absence of CSB. Moreover, footprint analysis revealed that RNA Pol II stalled at a thymine dimer covers $\sim 40 \text{ nt}$ nearly symmetrical around the dimer. As static footprints capture only the average of many dynamic conformations, they do not provide adequate insights into the extent of backtracking and the underlying mechanism of repair. These examples provide ample motivation to investigate the interplay between RNAP conformations, delivery of repair enzymes, and the choice between the sub-pathways for preferential repair in eukaryotes within the framework of the model proposed above.

Methods

The mathematical representation of the biochemical network (Fig. 2) is described in detail in supplementary information. The parameter values were taken from literature and reasonable values were assigned to those whose values are not reported. The model equations were integrated in time using the function `ode15s` available in MATLAB[®] (The Mathworks, Natick, MA).

Acknowledgements

The generous support from the US Department of Energy's Office of Biological and Environmental Research to Dr. Jeremy S. Edwards is greatly appreciated.

References

- [1] V.A. Bohr, C.A. Smith, D.S. Okumoto, P.C. Hanawalt, DNA repair in an active gene: removal of pyrimidine dimers from the DHFR gene of CHO cells is much more efficient than the genome overall, *Cell* 40 (1985) 359–369.
- [2] I. Mellon, V.A. Bohr, A. Smith, P.C. Hanawalt, Preferential DNA repair of an active gene in human cells, *Proc. Natl. Acad. Sci. U.S.A.* 83 (1986) 8878–8882.
- [3] I. Mellon, P.C. Hanawalt, Induction of the *Escherichia coli* lactose operon selectively increases repair of its transcribed DNA strand, *Nature* 342 (1989) 95–98.
- [4] E.C. Friedberg, Relationships between DNA repair and transcription, *Ann. Rev. Biochem.* 65 (1996) 15–42.
- [5] A. Sancar, DNA excision repair, *Ann. Rev. Biochem.* 65 (1986) 43–81.
- [6] J.Q. Svejstrup, Mechanism of transcription-coupled DNA repair, *Nat. Rev. Mol. Cell. Biol.* 3 (2002) 21–29.
- [7] J.Q. Svejstrup, Rescue of arrested RNA polymerase II complexes, *J. Cell. Sci.* 116 (2003) 447–451.
- [8] J. Venema, L.H.F. Mullenders, A.T. Natarajan, A.A. van Zeeland, L.V. Mayne, The genetic defects of Cockayne syndrome is associated with a defect in repair of UV-induced DNA damage in transcriptionally active DNA, *Proc. Natl. Acad. Sci. U.S.A.* 87 (1990) 4707–4711.
- [9] P.C. Hanawalt, The bases for Cockayne syndrome, *Nature* 405 (2000) 415–416.
- [10] C.P. Selby, A. Sancar, Molecular mechanism of transcription-repair coupling, *Science* 260 (1993) 53–58.
- [11] S. Kunal, D.E. Brash, Intragenic domains of strand-specific repair in *Escherichia coli*, *J. Mol. Biol.* 246 (1995) 264–272.
- [12] S. Li, R. Waters, Induction and repair of cyclobutane pyrimidine dimers in the *Escherichia coli* tRNA gene *tyrT*: Fis protein affects dimer induction in the control region and suppresses preferential repair in the coding region of the transcribed strand, except in a short region near the transcription start site, *J. Mol. Biol.* 271 (1997) 31–46.
- [13] B. Ahn, L. Grossman, RNA polymerase signals UvrAB landing sites, *J. Biol. Chem.* 271 (1996) 21453–21461.
- [14] C.G. Lin, O. Kovalsky, L. Grossman, Transcription coupled nucleotide excision repair by isolated *Escherichia coli* membrane-associated nucleoids, *Nucl. Acids Res.* 26 (1998) 1466–1472.
- [15] D.A. Erie, The many conformational states of RNA polymerase elongation complexes and their roles in the regulation of transcription, *Biochim. Biophys. Acta* 1577 (2002) 224–239.
- [16] P.H. von Hippel, An integrated model of the transcription complex in elongation, termination, and editing, *Science* 281 (1998) 660–665.
- [17] J.S. Park, M.T. Marr, J.W. Roberts, *E. coli* transcription repair coupling factor (Mfd protein) rescues arrested complexes by promoting forward translocation, *Cell* 109 (2002) 757–767.
- [18] I.H. Segel, *Enzyme Kinetics*, Wiley, New York, 1975.
- [19] V. Mekler, E. Kortkhonja, J. Mukhopadhyay, J. Knight, A. Revyakin, A.N. Kapanidis, W. Niu, Y.W. Ebricht, R. Levy, R.H. Ebricht, Structural organization of bacterial RNA polymerase holoenzyme and the RNA polymerase-promoter open complex, *Cell* 108 (2002) 599–614.
- [20] E.C. Friedberg, G.C. Walker, W. Siede, *DNA Repair and Mutagenesis*, American Society for Microbiology Press, Washington, DC, 1995.
- [21] Y. Zou, B. van Houten, Strand opening by the UvrA₂B complex allows dynamic recognition of DNA damage, *EMBO J.* 18 (1999) 4889–4901.
- [22] S. Li, M.J. Smerdon, Rbp4 and Rpb9 mediate subpathways of transcription-coupled DNA repair in *Saccharomyces cerevisiae*, *EMBO J.* 21 (2002) 5921–5929.
- [23] Z. Wang, J.Q. Svejstrup, W.J. Feaver, X. Wu, R.D. Kornberg, E.C. Friedberg, Transcription factor b (TFIIH) is required during nucleotide-excision repair in yeast, *Nature* 368 (1994) 74–76.
- [24] Z. Wang, S. Buratowski, J.Q. Svejstrup, W.J. Feaver, X. Wu, R.D. Kornberg, T.F. Donahue, E.C. Friedberg, The yeast *TFB1* and *SSL1* genes, which encode subunits of transcription factor IIH, are required for nucleotide excision repair and RNA polymerase transcription, *Mol. Cell. Biol.* 15 (1995) 2288–2293.
- [25] M. Tijsterman, R.A. Verhage, P. van de Putte, J.G. Tasseron-de Jong, J. Brouwer, Transitions in the coupling of transcription and nucleotide excision repair within RNA polymerase II-transcribed genes of *Saccharomyces cerevisiae*, *Proc. Natl. Acad. Sci. U.S.A.* 94 (1997) 8027–8032.
- [26] C.P. Selby, R. Drapkin, D. Reinberg, A. Sancar, RNA polymerase II stalled at a thymine dimer: footprint and effect on excision repair, *Nucl. Acids Res.* 25 (1997) 787–793.