

# Simulating cellular dynamics through a coupled transcription, translation, metabolic model

Elizabeth L. Weitzke, Peter J. Ortoleva\*

*Department of Chemistry, Indiana University, 800 East Kirkwood Ave, Bloomington, IN 47405, USA*

Received 29 July 2003; received in revised form 25 August 2003; accepted 27 August 2003

## Abstract

In order to predict cell behavior in response to changes in its surroundings or to modifications of its genetic code, the dynamics of a cell are modeled using equations of metabolism, transport, transcription and translation implemented in the Karyote software. Our methodology accounts for the organelles of eukaryotes and the specialized zones in prokaryotes by dividing the volume of the cell into discrete compartments. Each compartment exchanges mass with others either through membrane transport or with a time delay effect associated with molecular migration. Metabolic and macromolecular reactions take place in user-specified compartments. Coupling among processes are accounted for and multiple scale techniques allow for the computation of processes that occur on a wide range of time scales. Our model is implemented to simulate the evolution of concentrations for a user-specifiable set of molecules and reactions that participate in cellular activity. The underlying equations integrate metabolic, transcription and translation reaction networks and provide a framework for simulating whole cells given a user-specified set of reactions. A rate equation formulation is used to simulate transcription from an input DNA sequence while the resulting mRNA is used via ribosome-mediated polymerization kinetics to accomplish translation. Feedback associated with the creation of species necessary for metabolism by the mRNA and protein synthesis modifies the rates of production of factors (e.g. nucleotides and amino acids) that affect the dynamics of transcription and translation. The concentrations of predicted proteins are compared with time series or steady state experiments. The expression and sequence of the predicted proteins are compared with experimental data via the construction of synthetic tryptic digests and associated mass spectra. We present the mathematical model showing the coupling of transcription, translation and metabolism in Karyote and illustrate some of its unique characteristics.

© 2003 Elsevier Ltd. All rights reserved.

*Keywords:* Cell modeling; Cellular dynamics; Multi-scale kinetics; Stoichiometric analysis; Bio-polymerization kinetics

## 1. Introduction

A model for predicting the response of a cell to changes in host medium (e.g. pH, salinity, and concentrations of oxygen, nutrients and waste products) or to modifications of its genetic code (e.g. gene deletion and mutation) is presented. While many individual cellular processes have been investigated, the coupling among them must be accounted for in order to understand the interplay among them and the associated nonlinear dynamics. Nonlinear phenomena including multiple steady states, periodic or chaotic temporal evolution and self-organization can be supported by the dynamical cellular system since the rate laws are nonlin-

ear in the descriptive variables and the system is maintained far from equilibrium. A classic example of such behavior is the metabolic temporal oscillation that arises through the coupling of biochemical reactions (Field and Burger, 1985; Goldbeter, 1996; Hess and Boiteux, 1973). In order to capture such phenomenon a cell model must be fully dynamical (e.g. not limited to steady state behavior). The growing availability of genomic, proteomic, biochemical and other data have facilitated the progress in cell modeling.

In our approach, the cell is understood in terms of a physical chemical model that includes a wide range of processes. Our model simulates the dynamics of a cell by dividing its volume into discrete compartments representing the organelles in eukaryotes and specialized zones in prokaryotes. Each compartment is assigned a set of geometric variables to describe their connectivity, size and transport parameters for the exchange of molecules between compartments and with the host medium. Linear, nonlinear, passive and active

\* Corresponding author. Tel.: +1-812-855-2717; fax: +1-812-855-8300.

*E-mail addresses:* [eweitzke@indiana.edu](mailto:eweitzke@indiana.edu) (E.L. Weitzke), [ortoleva@indiana.edu](mailto:ortoleva@indiana.edu) (P.J. Ortoleva).

processes can be used to describe the exchange of molecules between compartments and with the surroundings. Within each compartment, metabolic and macromolecular reactions take place as defined by the user. This provides an integration of compartment localized and inter-compartmentalized processes. All of the equations that underlie our model are coupled through non-linearities that describe the creation, annihilation, and exchange of species involved so as to capture key processes of a cell, notably transcription, translation and metabolic reactions. Our model accounts for the wide separation of time scales common in cellular activity through multiple time scale techniques (Ortoleva, 1992; Ortoleva and Ross, 1975). We allow for a user-specified set of fast (equilibrated and steady state) and slow (finite rate) reaction stoichiometries. The algebra of the elimination of the fast reactions is automated, and mass balance discrepancies encountered by the inconsistent application of steady state or Michaelis–Menton rate laws is avoided (see Section 2.1). Due to the hierarchical structure of our implementation (i.e. one may have compartment within compartment), Karyote can also be used to simulate tissues and other multi-cellular systems.

One of the unique features of our implementation is that gene sequence data is used as input. Through the input gene sequence, the dynamics of the pool of nucleotides and amino acids is controlled by the rate of the specific sequential addition to the elongating mRNAs and proteins (Fig. 1). Thus the rate of transcription and translation is controlled by the kinetics of nucleotide and amino acid formation through metabolic reactions. The reciprocal relation between transcription and translation with nucleotide and amino acid production and availability is accounted for. For convenience, polypeptides resulting from translation are denoted proteins, as post-translational modifications have not been accounted for yet.

The development of cell models dates back to the 1950s (Progogine and Lefever, 1968; Rashevsky, 1960; Turing, 1952). Recent developments in computational science and the availability of mass quantities of experimental data have triggered the development of many cell models, though we only list a few (Arkin, 1998; Bartol et al., 1997, 2001; Hines, 1989; Hucka et al., 2003; Loew and Schaff, 2001; Mannella et al., 2001; Mendes, 1993, 1997; Mendes and Kell, 1998, 2001; Mendes et al., 2001; Sauro, 1993; Sauro et al., 1994; Schaff and Loew, 1999; Schaff et al., 2000, 2001a,b; Slepchenko et al., 2002; Tomita, 2001; Tomita et al., 1999, 2000; Wilson and Bower, 1991). A variety of authors have investigated the influence of fluctuations in biochemical dynamics (Frith and Bray, 2000; Gillespie, 1976, 1977; Larter and Ortoleva, 1981, 1982; LeNovere and Shimizu, 2001; McAdams and Arkin, 1997; Shimizu and Bray, 2001), an important issue in some phenomenon wherein the number of molecules (e.g. certain enzymes) is small so that macroscopic rate laws do not apply.

The parameters used in the model are calibrated using a variety of types of experimental data. A representative sub-

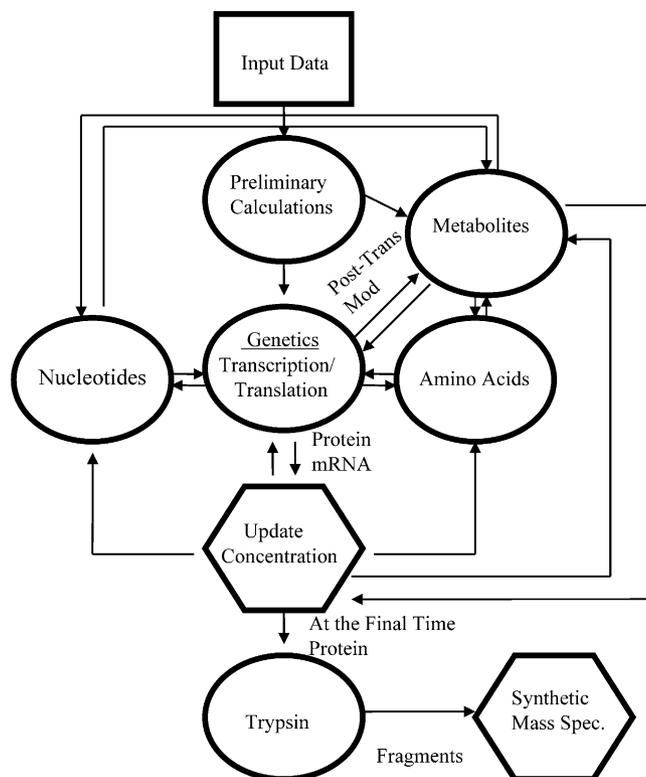


Fig. 1. A brief schematic diagram of the processes implemented in Karyote. Once the initial data are read in, the matrix equation (see methods section) is constructed. The metabolic and genetic subroutines solve for concentrations of metabolites, transcription species (enzymes, DNA) and translation species (mRNA, proteins, ribosomes). These two subroutines are coupled in that the metabolism routine creates nucleotides/amino acids, which are in turn used to construct mRNA/proteins. Also enzymes are produced and assembled which are necessary for certain reactions in metabolism. For each time step, all species concentrations are updated which take into account their use or production in all subroutines. Once proteins are made, we, for testing purposes, allow trypsin to act on the nascent protein to produce a synthetic mass spectrum.

set of our input data is seen in Table 1. As the model accounts for a broad range of reaction and transport processes, Karyote can predict a full array of measurable parameters

Table 1  
Representative subset of input data used to run Karyote

Input data
Each compartment's volume, area and connectivity to other compartments
Total number of species
Number of fast (majority/minority designations) and slow reactions
Sequence for each gene
Promotor/terminator sequences for each gene
Ribosome Binding Site (RBS)/termination sequence for each gene
Stoichiometric matrix for fast/slow reactions
Initial concentrations of all species
Composition of host medium
Membrane transport parameters
Rate constants for slow and fast reactions
Equilibrium constants (fast/slow)

that can be used for calibration and testing. Advances in cell modeling are facilitated through a methodology for calibrating the many physical and chemical rate and equilibrium parameters and accessing the uncertainties in these values. To do so requires an automated procedure such as that based on information theory as discussed in Sayyed-Ahmad et al. (2003).

## 2. Methods

There are two general classes of processes accounted for in our model. One class includes metabolic cycles and the exchange of molecules between compartments and with the cell's surroundings. The other includes mRNA and protein synthesis. These two classes are coupled through the production of nucleotides and amino acids in metabolic cycles (consumed through transcription and translation respectively) and the control of metabolite flux by the enzymes created from the products of translation.

### 2.1. Compartmentalized reaction-transport metabolic model

In our model, the cell is divided into  $N_c$  compartments labeled  $p = 1, 2, \dots, N_c$ . In compartment  $p$  there are molecular species  $j = 1, 2, \dots, N$  described by their concentrations  $c_j^p(t)$ . Conservation of mass implies

$$V^p \frac{dc_j^p}{dt} = \sum_{p' \neq p} A^{p'p} J_j^{p'p,m} + V^p R_j^{p,slow} + V^p \sum_{k=1}^{N_f} v_{jk}^{fast} \frac{W_k^{p,fast}}{\epsilon} \quad (1)$$

where  $V^p$  is the volume of compartment  $p$ ,  $c_j^p$  is the concentration of species  $j$  in compartment  $p$ ,  $A^{p'p}$  is the area of membrane separating compartments  $p'$  and  $p$ ,  $J_j^{p'p,m}$  is the net membrane flux of species  $j$  from compartment  $p'$  to  $p$ ,  $R_j^{p,slow}$  is the net reaction rate for slow reactions affecting species  $j$  in compartment  $p$ ,  $v_{jk}^{fast}$  is the stoichiometric coefficient for fast reaction  $k$  affecting species  $j$  (assumed same for all compartments),  $W_k^{p,fast}$  is the net rate of the  $k$ th fast reaction in compartment  $p$ , and  $\epsilon$  is the ratio of the short to the long characteristic time.

Bio-chemical reactions proceed on a wide range of time-scales. By definition, the time scales of the fast reactions can be many orders of magnitude smaller than those of the slow ones. Therefore, it is difficult to simulate the evolution of a cell on the time scales of interest (e.g. those of slow processes). Our multiple time scale approach is based on modifying Eq. (1) to project the sets of fast reactions on the slow manifold such that the differential equations can be solved on the time scales of interest (Ortoleva, 1992, 1994). Thus, for practical and conceptual reasons we di-

vide reactions into a fast and a slow group. For example, consider the reactions



where  $E$  denotes an enzyme,  $S$  a substrate,  $ES$  an enzyme–substrate complex, and  $P$  the product. Fast reactions are further divided into two groups; one that involves only majority species (i.e. the species of high concentration), and the other refers to minority species (notably enzymes) of low concentration. The low concentration species play a major role in the dynamics of majority species as they may provide a kinetic bottleneck but which may also accelerate a metabolic cycle due to associated large rate coefficients. In our model, chemical reactions can be formulated with general stoichiometry, for example, enzymes complexing with user-specified stoichiometries of factors, substrates, and products for each metabolic sub-network. The general stoichiometry of the reaction network (as in Eqs. (2) and (3)) allows one to include reactions such as  $E + X = EX$  for any factor  $X$ , so that if  $EX$  is the active form,  $X$  is a promoter or if  $EX$  is the inactive form,  $X$  is a repressor. This generality allows for enzyme modification by metabolites through inhibition or activation of the enzyme. This is but one example of how our model allows for the control of enzymes of other processes. We have accounted for the two extreme cases (fast and slow reactions) explicitly. One may simulate processes on intermediate time scales by identifying them as slow (finite rate) reactions. The compromise is that the faster these intermediate processes are, the smaller the time step required and therefore the increase in cpu time. Through the use of our slow manifold projection methodology, the algebra is automated for equilibrated and steady state fast processes with a wide range of complexities. In this method, mass is conserved so that the total amount of low concentration co-factors or other molecules and their complexes with the enzyme are accounted for.

The slow manifold method allows one to solve Eq. (1) in the limit  $\epsilon \rightarrow 0$ . In this way, the system reaches equilibrium and maintains complex steady state cycles (Fig. 2) on a

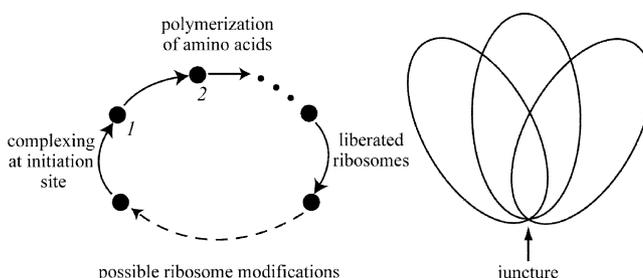


Fig. 2. (a) In a complex biochemical network an enzyme or other factors (here a ribosome) can undergo a many-step repeating (cyclic) sequence of interactions. (b) The kinetic loops can have a multi-lobed structure depending on the order of substrate binding in the metabolic networks.

time-scale that is short relative to those of interest. In this limit, equilibrium conditions imply,

$$W_k^{p,f_m} = 0, \quad k = 1, 2, \dots, \hat{N}_f. \quad (4)$$

where  $W_k^{p,f_m}$  is the net reaction rate for the  $k$ th fast majority reaction in compartment  $p$ , where the total number of fast reactions,  $N_f$  is  $N_f = \hat{N}_f + N_f^*$ , where  $N_f^*$  is the number of minority fast reactions. This yields  $N_c \times \hat{N}_f$  of the  $N_c \times N_c$  equations needed to determine the concentration of the  $N_c \times N_c$  species (except for special cases where it is necessary to account for redundant reactions (Ortoleva, 1992)). As  $\varepsilon \rightarrow 0$ , equations for steady state cycles appear in our formulation as linear combinations of  $W_l^{p,f^*}$ 's dictated by the structure of the rectangular stoichiometric matrix for the fast reactions yielding  $N_f^*$  equations, where  $W_l^{p,f^*}$  is the net rate for the  $l$ th fast minority reaction in compartment  $p$ , and  $l = \hat{N}_f + 1, \dots, N_f$ . In order to eliminate the secular  $\varepsilon^{-1}$  terms, and thereby capture the slow behavior, we introduce  $\sigma_{\alpha j}$  such that:

$$\sum_{j=1}^N \sigma_{\alpha j} \cdot v_{jk}^{\text{fast}} = 0 \quad (5)$$

where  $\sigma_{\alpha}$  is one of the  $\tilde{N}$  row vectors, that is orthogonal to the independent columns of the stoichiometric matrix for fast processes. There are  $N_c \times \tilde{N}$  fast equilibrium or steady state relations, so that the above formulation provides a sufficient number of equations to determine  $N_c \times N_c$  unknowns. The  $\sigma_{\alpha}$  row vectors are constructed using the singular value decomposition method (Press et al., 1992). The  $N_c \times N_c - \tilde{N}$  equations needed to complete this description are found by multiplying Eq. (1) recast in vector form by the  $\tilde{N} \sigma_{\alpha}$  vectors, to obtain

$$\sum_{j=1}^N \sigma_{\alpha j} \frac{dc_j^p}{dt} = \sum_{p' \neq p}^{N_c} \frac{A^{p'p}}{V^p} \sum_{j=1}^N \sigma_{\alpha j} J_j^{p'p,m} + \sum_{j=1}^N \sigma_{\alpha j} R_j^{p,\text{slow}}, \quad (6)$$

where

$$\sum_{j=1}^N \sigma_{\alpha j} \sum_{k=1}^{N_f} v_{jk}^{\text{fast}} \frac{W_k^{p,\text{fast}}}{\varepsilon} = 0. \quad (7)$$

With this approach we solve these equations for the concentrations of  $N$  species in each of the  $N_c$  compartments in our model. The differential equations are solved using the adaptive time step Runge–Kutta–Fehlberg method (Press et al., 1992), a method well suited for the present formulation as the fast behavior has been projected out. The equilibrium/steady state relations take the general form  $W=0$ , which we transform to differential equations with the form  $dW/dc=0$ , so that the entire problem is reduced to a set of coupled nonlinear differential equations.

## 2.2. Prokaryotic macromolecular synthesis

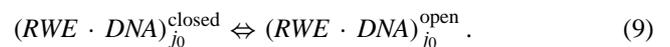
The dynamics of transcription and translation are accounted for by computing the temporal evolution of the populations of DNA, RNA, proteins and their various complexes within the cell. The model reads and transfers nucleotide and amino acid sequences through a polymerization kinetic model. The rapidly expanding genomic and proteomic databases can thereby be utilized for model development and calibration. We now illustrate how this is accomplished by considering the kinetics of transcription and translation in prokaryotes. Key aspects of the synthesis of mRNA, proteins and other macromolecules are the role of a template molecule (e.g. mRNA for proteins) and the mediation by enzymes in controlling the polymerization. Our chemical kinetic formalism is as follows.

A read/write/edit (*RWE*) complex associates with the template strand and advances along it, reading its information in search of an initiation sequence where upon the *RWE* forms a closed complex. *RWE* incorporates RNA polymerase II with transcription factors like regulatory proteins including repressors and activators. Also, the specificity is built in to account for multiple sigma factors that make up the whole RNA polymerase II enzyme (Lewin, 1997). The different sigma subunits allow for multiple affinities for the promoter region on the DNA, thereby controlling the frequency of transcription. This is accomplished by casting this process as a chemical reactions and writing the associated mass action rate laws. The individual regulatory proteins are apart of the *RWE* complex and are controlled by the rate of *RWE* binding to the promoter region on the DNA. Below, a demonstration is given for repressor and activator proteins and their role in the overall transcription and translation network. Through enzyme-induced isomerization, an open complex is formed in order for the single strand of DNA to be read for transcription (Lewin, 1997). Whereupon, polymerization occurs, building the mRNA nucleotide sequence dictated by the template DNA. Auxiliary molecules complex with *RWE* to modify its rules of reading the templating strand, wherein initiation, elongation and termination can be altered. In this way, the model captures the evolving synthesis and destruction of proteins. The reactions below describe the coupled transcription and translation process performed in a simulation of a prokaryotic cell. All the following equations depend on which gene is being transcribed and translated, although the gene labeling has been omitted for simplicity here.

The multi-step transcription/translation typical of prokaryotes is as follows (Niedhardt et al., 1996). The *RWE* joins the DNA strand for a given gene at the initiation site, labeled  $j_0$ , creating a closed complex:



Through enzyme-induced isomerization, an open conformation is adopted (Lewin, 1997):



A short mRNA strand of length  $\ell$  is transcribed (in most species  $7 \leq \ell \leq 12$ ): the addition of  $ntp(j_0)$ , the appropriate transcript for site  $j_0$  on the DNA, is written

$$(RWE \cdot DNA)_{j_0}^{\text{open}} + ntp(j_0) \Leftrightarrow (RWE \cdot DNA \cdot mRNA)_{j_0+1}, \quad (10)$$

while the last  $\ell$ th monomer addition for this short transcript takes the form

$$(RWE \cdot DNA \cdot mRNA)_{j_0+\ell-1} + ntp(j_0 + \ell - 1) \Leftrightarrow (RWE \cdot DNA \cdot mRNA)_{j_0+\ell}. \quad (11)$$

Next either an abortive mRNA of length  $\ell$  is released or the  $\sigma$ -subunit of  $RWE$  is shed; the latter is written

$$(RWE \cdot DNA \cdot mRNA)_{j_0+\ell} + ntp(j_0 + \ell) \rightarrow (RWE^* \cdot DNA \cdot mRNA)_{j_0+\ell+1} + \sigma, \quad (12)$$

where  $RWE^*$  is  $RWE$  devoid of  $\sigma$  and ready for efficient further polymerization through nonspecific binding. Polymerization proceeds via monomer addition:

$$(RWE^* \cdot DNA \cdot mRNA)_j + ntp(j) \Leftrightarrow (RWE^* \cdot DNA \cdot mRNA)_{j+1}, \quad (13)$$

where  $j$  indicates the position of  $RWE^*$  along the DNA. Process (13) proceeds until  $RWE^*$  reaches its termination site  $j_f$ . The species on the RHS involves an mRNA of length  $(j_f - j_0 + 1)$ , with addition of  $ntp(j_f)$ .

It is a complication of the prokaryote genome that transcription and translation can take place simultaneously. During polymerization, nascent mRNA is exposed, enabling ribosomes to attach at the ribosome-binding site (labeled  $k_o$ )

$$(RWE^* \cdot DNA \cdot mRNA)_j + 30s + 50s \Leftrightarrow (RWE^* \cdot DNA \cdot mRNA \cdot Rib)_{j,k_o}. \quad (14)$$

The species on the RHS consists of  $RWE^* \cdot DNA \cdot mRNA$  complexed with a ribosome at the ribosome-binding site on the mRNA; thus the two subscripts on this complex indicate the location of  $RWE^*$  and the ribosomes on the DNA and mRNA, respectively. A further complication is that multiple ribosomes can attach simultaneously, a process not made explicit here for simplicity. Another complexity is seen if the RBS on the mRNA is not present, meaning transcription needs to proceed further to expose the RBS, or if the ribosome just never attaches, hence only transcription occurs.

The general transcription step for the single ribosome complex is written

$$(RWE^* \cdot DNA \cdot mRNA \cdot Rib \cdot protein)_{j,k} + ntp(j) \Leftrightarrow (RWE^* \cdot DNA \cdot mRNA \cdot Rib \cdot protein)_{j+1,k}, \quad (15)$$

where  $j$  indicates the position of  $RWE^*$  along the DNA while  $k$  indicates the position of the ribosome along the mRNA.

In the simplest process wherein no ribosome attaches, only transcription occurs and the mRNA is length  $(j_f - j_0 + 1)$ . Polymerization of proteins proceeds via

$$\begin{aligned} & (RWE^* \cdot DNA \cdot mRNA \cdot Rib \cdot protein)_{j,k} \\ & + aa(k) \cdot tRNA(aa(k)) \\ \Leftrightarrow & (RWE^* \cdot DNA \cdot mRNA \cdot Rib \cdot protein)_{j,k+3} \\ & + tRNA(aa(k)), \end{aligned} \quad (16)$$

for  $k = k_o, k_o + 3, \dots, k_f - 3$ . Here  $aa(k)$  is the amino acid scheduled for addition when the codon (e.g. mRNA nucleotides at positions  $k, k + 1, k + 2$ ) is read by the ribosome. The specific tRNA appropriate for the transfer of  $aa(k)$  is denoted  $tRNA(aa(k))$ . At this stage, the protein is of length  $(k - k_o)/3$ . The process described by Eq. (16) proceeds until one codon before the ribosome resides at the final codon  $(k_f - 2, k_f - 1, k_f)$ . Similarly, transcription proceeds for  $RWE^*$  on the DNA sequence at sites from  $j = j_0$  to  $j = j_f$ , the latter being the point at which mRNA is released upon addition of the final  $ntp$  (see Eq. (18)).

The ribosome could reach its termination sequence before the DNA is fully transcribed whereupon it dissociates to add to  $(RWE^* \cdot DNA \cdot mRNA)_j$ :

$$\begin{aligned} & (RWE^* \cdot DNA \cdot mRNA \cdot Rib \cdot protein)_{j,k_f} \\ & + aa(k_f) \cdot tRNA(aa(k_f)) \\ \rightarrow & (RWE^* \cdot DNA \cdot mRNA)_j + 30s + 50s \\ & + protein + tRNA(aa(k_f)). \end{aligned} \quad (17)$$

$RWE^*$  may reach its termination  $j = j_f$  and release the fully transcribed mRNA, while the ribosome is still attached and building proteins via the free-floating mRNA according to

$$\begin{aligned} & (RWE^* \cdot DNA \cdot mRNA \cdot Rib \cdot protein)_{j_f,k} + ntp(j_f) \\ \rightarrow & (mRNA \cdot Rib \cdot protein)_k + RWE^* + DNA. \end{aligned} \quad (18)$$

In order that  $RWE^*$  revert to  $RWE$  to complete the cycle, it must add a  $\sigma$  unit:

$$RWE^* + \sigma \Leftrightarrow RWE. \quad (19)$$

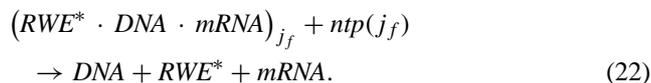
Although  $(mRNA \cdot Rib \cdot protein)_k$  is free-floating, the ribosome is still reading the template mRNA and protein building continues:

$$\begin{aligned} & (mRNA \cdot Rib \cdot protein)_k + aa(k) \cdot tRNA(aa(k)) \\ \Leftrightarrow & (mRNA \cdot Rib \cdot protein)_{k+3}, \end{aligned} \quad (20)$$

where  $k = k_o$  until  $k_f$  in steps of 3. Once the ribosome reaches its termination codon  $(k_f - 2, k_f - 1, k_f)$ , the final amino acid is added and then, the mRNA and protein are liberated and the ribosome breaks up into its subunits (30s, 50s):

$$\begin{aligned} & (mRNA \cdot Rib \cdot protein)_{k_f} + aa(k_f) \cdot tRNA(aa(k_f)) \\ \rightarrow & 30s + 50s + mRNA + protein. \end{aligned} \quad (21)$$

Complete mRNA can be released without an attached ribosome either before a ribosome has attached or after having been released:



This completes the prokaryote co-evolving transcription/translation dynamic for a single gene. Our model allows for a user-specified set of genes; it is being generalized to allow for multiple simultaneously translating ribosomes on each gene and multiple control factors on *RWE* and the ribosomes (a preliminary result is seen in Fig. 7).

The above complex transcription and translation network is simulated using chemical kinetic rate laws. The result is a set of ordinary differential equations for the evolution of the populations of *RWE*, *RWE*<sup>\*</sup>, *proteins*, *mRNA* and other species and their complexes appearing in Eqs. (8)–(22). These equations of mRNA and protein synthesis are coupled to the metabolic equations via consumptions of nucleotides and amino acids.

### 3. Implementation and discussion

Our program predicts the time evolution of user-specified species and reactions that participate in the metabolics, transcription and translation networks. Karyote is written in FORTRAN77 and runs on UNIX, LINUX and Windows operating systems. Typical cpu requirements for our method using a Dell Optiplex GX240 with 2.0 GHz Pentium processor and 1.0 GB SDRAM is as follows. A simulation of multi-compartmentalized *Trypanosoma brucei* glycolysis (Navid and Ortoleva, 2003) involving 59 species, 11 slow reactions, 28 fast reactions in three compartments for 3600 s of biological time takes 16 min and 13 s cpu time. A simulation of coupled transcription and translation for two genes from *Caulobacter crescentus* (cc1750 with 1953 nucleotides and cc1461 with 822 nucleotides) involves over 2 million species; when run for a biological time of 2500 s, the simulation took approximately 15 h. An example simulation result is seen in Fig. 9 depicted as a mass spectrum of the translated protein cc1750 (at which steady state has been achieved). In ongoing work, we are implementing an optimization scheme decreasing memory size and cpu time by more than 1000. In the following subsections, examples are discussed to illustrate some of Karyote's features.

#### 3.1. Transcription and translation

Consider the case where only transcription is operating. This models the experimental result wherein only DNA, enzyme and nucleotides are involved. The system studied, in vitro, was bacteriophage T7 RNA polymerase in conjunction with an *E. coli* gene fragment (Weston et al., 1997). The T7 family of DNA-dependent RNA polymerases repre-

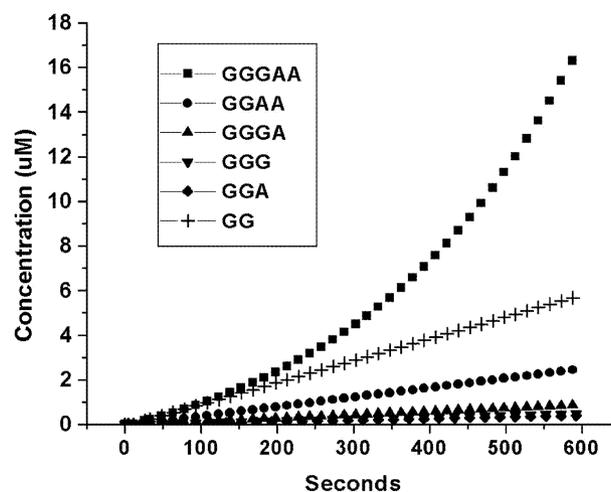


Fig. 3. Simulation using Karyote for transcription performed by bacteriophage T7 RNA polymerase system in conjunction with a gene fragment from *E. coli*. The simulated concentration evolution agrees with that observed, see Table 2.

sents a simple model system for the study of fundamental aspects of transcription, because T7 RNA polymerases do not require any helper proteins and exist as single subunits. These single-subunit RNA polymerases are highly specific for an approximately 20 base pair, nonsymmetric promoter sequence. The mutation feature of our model is illustrated as follows; there is one major transcript mRNA sequence, *GGGAA*, and five error copies, which arise from misinitiation or premature termination and were used to calibrate the rates of error copies generated from our simulation (Fig. 3). The agreement seen illustrates the validity of our formulation (Table 2).

#### 3.2. Coupled transcription and translation in prokaryotes

Consider an example that illustrates the more complete transcription and translation coupled dynamics for prokaryotes. In Figs. 4–6, the evolving concentrations of several representative species are shown to illustrate the coupled dynamics seen in our model, using the DNA sequence *TACTTT-TAGGGG* and initial data as described in Fig. 4.

It is seen that mRNA and protein synthesis depletes nucleotide and amino acid concentrations, illustrating the

Table 2

Comparison between predicted and experimental data from an in vitro RNA synthesis shows the sequence and concentration ( $\mu\text{M}$ ) after 10 min of evolution (experimental data from Weston et al., 1997)

Sequence	Predicted concentration ( $\mu\text{M}$ )	Experimental concentration ( $\mu\text{M}$ )
GGGAA	17.1	17.0
GGAA	2.49	2.6
GGGA	0.869	0.9
GGG	0.478	0.5
GGA	0.388	0.4
GG	5.8	6.7

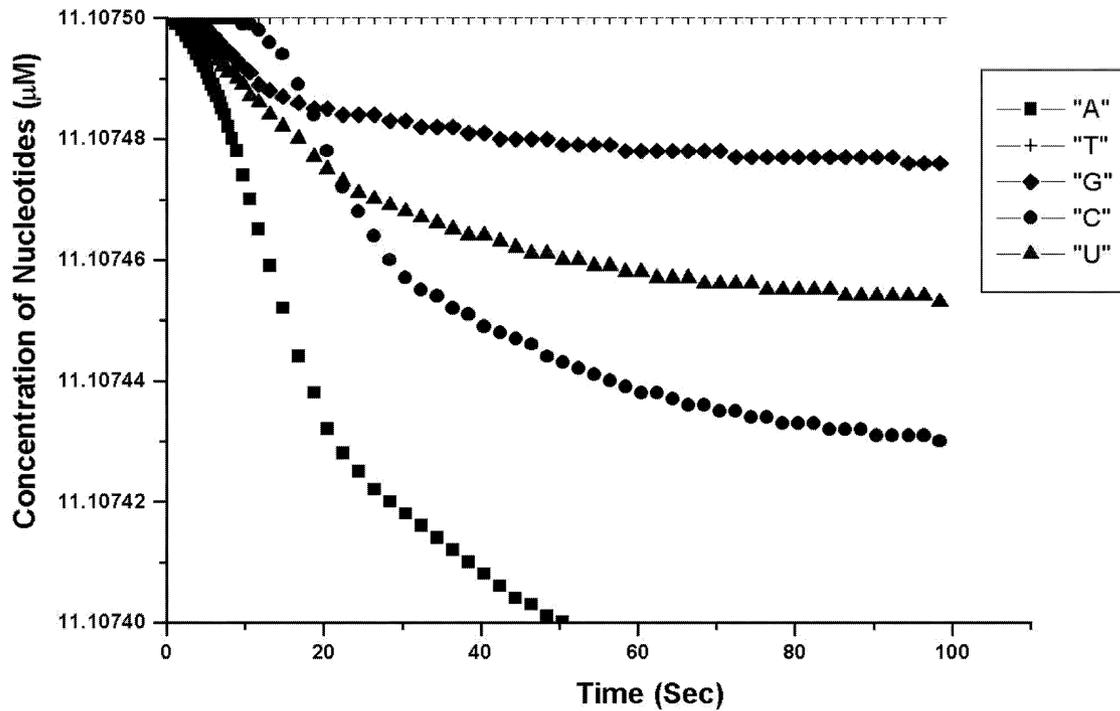


Fig. 4. Karyote dependence of predicted nucleotide concentration on time during transcription is shown for the transcribed gene *TACTTTTAGGGG*. As nucleotides are depleted, mRNA synthesis slows down, illustrating an important feature of coupled cellular dynamics since nucleotides are created through metabolic pathways. The system initially contained:  $9.2563 \times 10^{-12}$  M of gene, enzyme, ribosome;  $1.11075 \times 10^{-3}$  M of tRNA; and  $1.11075 \times 10^{-5}$  M of each nucleotide and amino acid, all other species were initially set to zero.

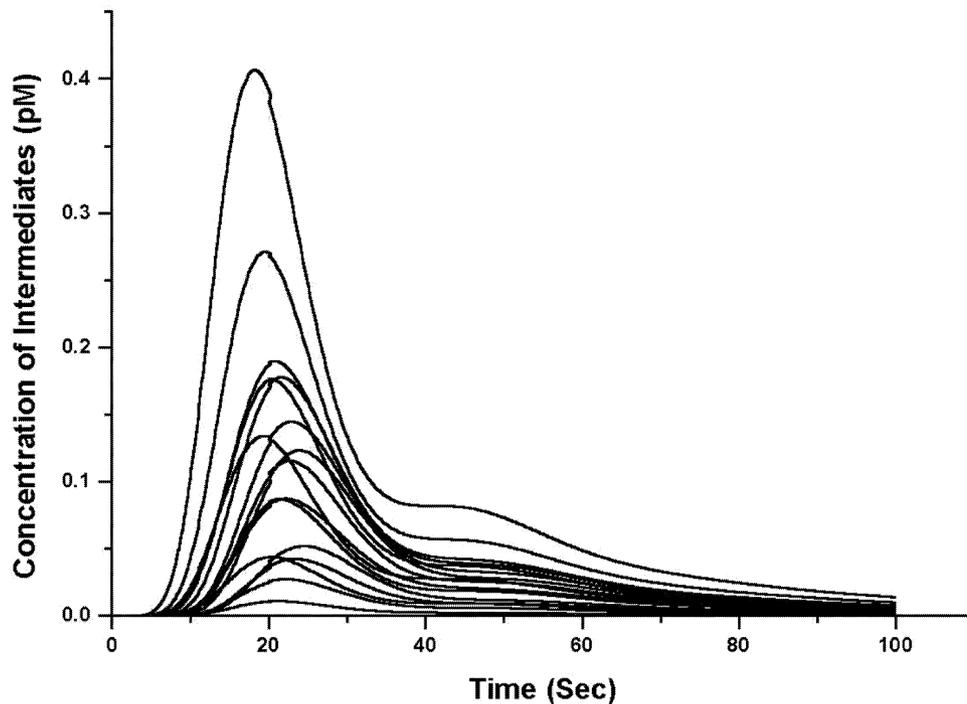


Fig. 5. The dependence of concentration of all the second class of intermediates created during coupled transcription/translation for the same simulation as described in Fig. 4. Each curve represents the different complexes involved in adding a nucleotide for transcription and an amino acid for translation. The intermediates represent complexes composed of enzyme, DNA, mRNA, ribosome and protein.

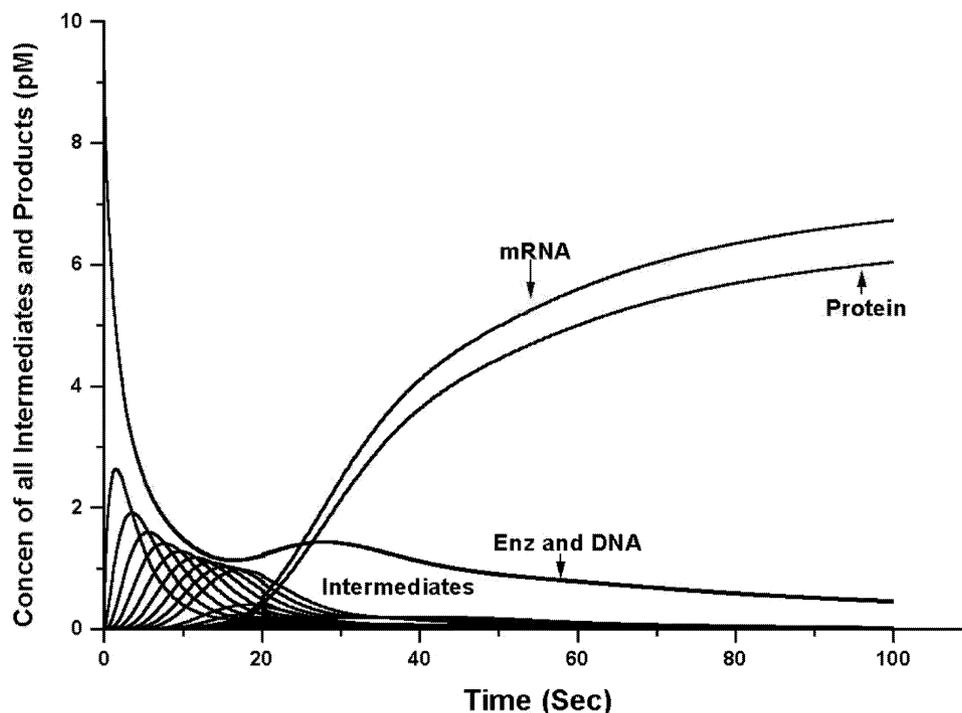


Fig. 6. The time evolution of the DNA, RNA polymerase, all of the first class of intermediates involved in transcription/translation and the mRNA and protein created from gene with sequence *TACTTTAGGGG*. The intermediates include complexes of enzyme, DNA and mRNA.

coupling of mRNA and protein synthesis through the evolution of precursor metabolites. As nucleotides or amino acids are depleted, mRNA or protein synthesis slows down, illustrating one of the many feed back loops captured by our method. In Karyote, all nucleotides and amino acids used in transcription and translation have been tracked during the simulation. As seen in Fig. 4, transcription of the DNA sequence *TACTTTAGGGG* results in an mRNA product of *AUGAAAUC*, which is reflected by consumption of nucleotides with a decreasing rate at approximately 20 s. Nucleotide concentration evolves to a transient steady state (Fig. 4), during which the enzyme is re-complexing at the initiation site (Fig. 6). Just before the transient steady state begins, the total mRNA concentration rises corresponding to the liberation of the complete mRNA (Fig. 6). Similar behavior is seen for the liberated protein with respect to the transient steady state and notably amino acid synthesis, an important element of the overall coupling of transcription and translation with metabolic pathways.

In our model, there are two classes of macromolecular complexes that mediate the production of mRNA and protein. The first class (seen in Fig. 6) includes complexes of all the intermediates from the initiation of transcription to the loss of the sigma subunit from RNA polymerase seen in the mechanisms of Eqs. (8)–(12).

The second class of intermediates, in Fig. 5, includes all complexes after the sigma subunit is released from the enzyme until the liberation of the products mRNA and protein (seen in the mechanisms of Eqs. (13)–(22)). Each concentration profile shows a wave-like behavior. As one intermediate

increases, its predecessor decreases. This wave-like behavior, arising from the polymerization dynamics, presented in our model captures the evolution of the intermediates and the release of the final mRNAs and proteins.

Other key genomic species that are simulated by our model are DNA and RNA polymerase II (see Fig. 6). As expected, the simulation shows the time-independent state of the system at long times due to the depletion of the nucleotides and amino acids in this closed system. As seen in Fig. 6, the enzyme concentration initially decreases, signifying the binding of enzyme to DNA. After some time, there is an increase in enzyme concentration that represents its liberation as the full mRNA is released. The free enzyme concentration ultimately reaches a steady state in this closed system.

The products of transcription and translation are seen in Fig. 6. The liberation of mRNA is coordinated with that of free enzyme as they are created simultaneously.

### 3.3. Control of gene expression

Another feature built into Karyote is gene control through protein activation or repression. This process is formulated as kinetic equations built into the transcription (*RWE* factor) and translation model. To demonstrate this feature, we choose a model where protein one activates gene two, which through transcription and translation creates protein two. Protein two represses gene one through binding, thereby limiting the expression of gene one (see Fig. 7). Fig. 7 shows a transient period terminating in a steady state balance of repression and activation.

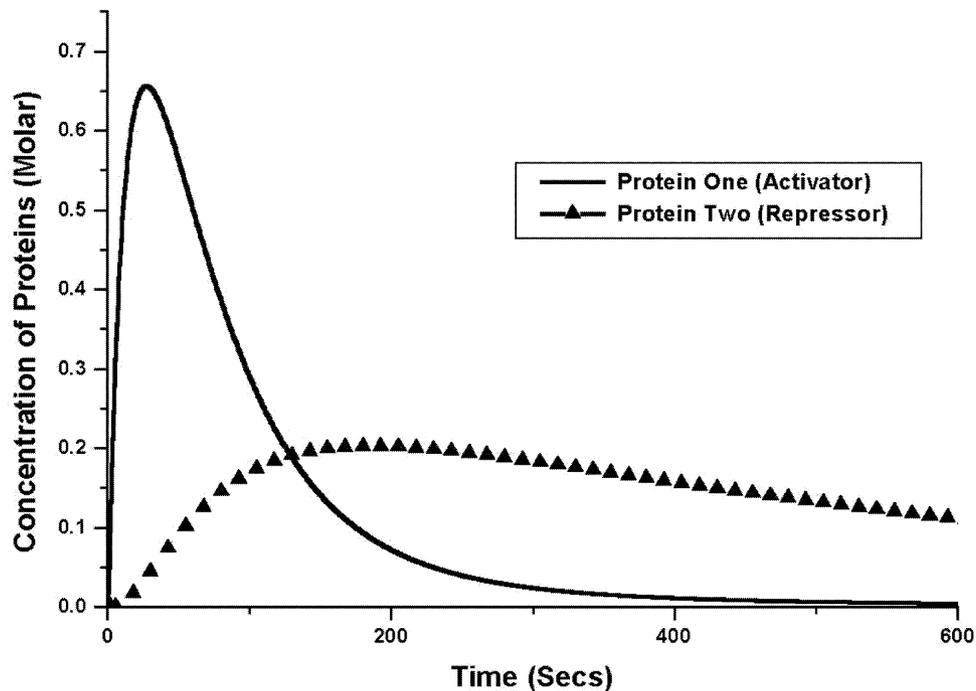


Fig. 7. Concentration profiles of protein one and two during gene controlled coupled transcription and translation. Gene one is active producing protein one, whereby protein one activates gene two producing protein two, which, in turn, inactivates gene one, decreasing production of protein one.

#### 3.4. Tryptic fragment/mass spectra of Karyote predicted proteins

In this section, we illustrate how predicted amino acid sequences are analyzed using a synthetic tryptic digest, resulting in a computed qualitative mass spectrum. The lat-

ter can then be compared with the observed mass spectra for identification purposes. Synthetic tryptic digestion was accomplished using the rules as in ([http://us.expasy.org/tools/peptidecutter/peptidecutter\\_enzymes.html](http://us.expasy.org/tools/peptidecutter/peptidecutter_enzymes.html)) and Barrett et al. (1998). Since trypsin does not cleave with 100% efficiency, these rules are often reformulated as probabilities.

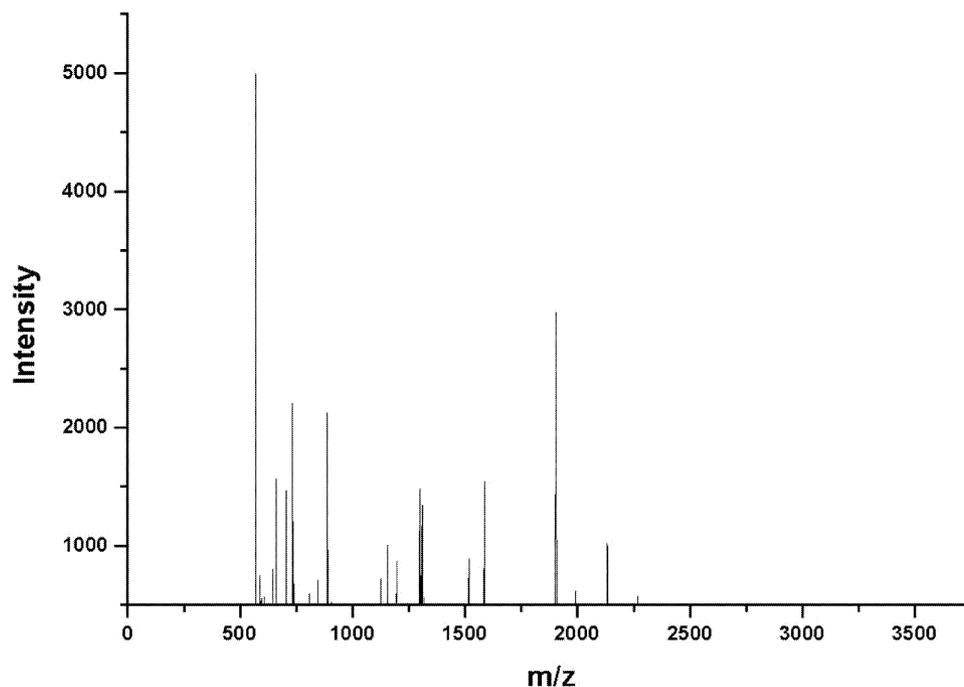


Fig. 8. Observed mass spectrum of the tryptically digested protein cc1750 from *C. crescentus* (Karty et al., 2002).

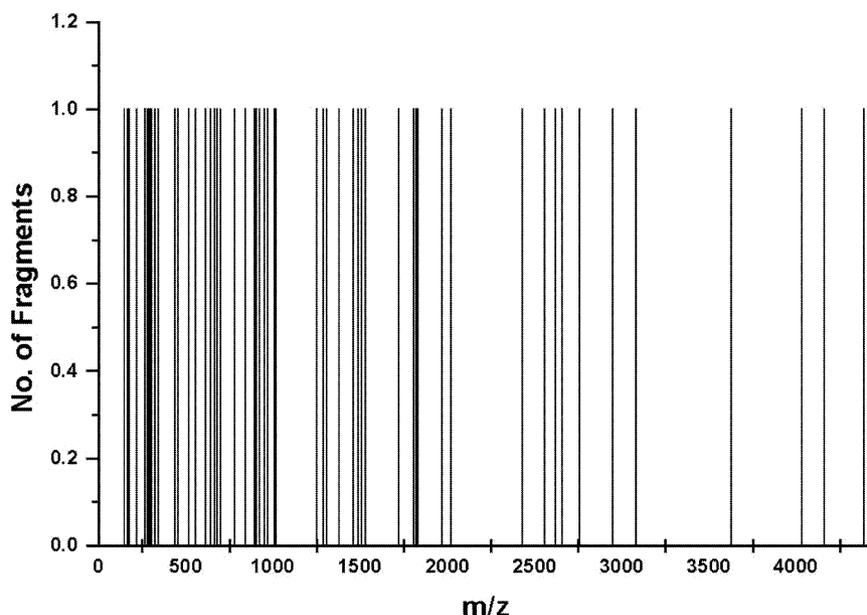


Fig. 9. Mass spectrum generated from Karyote for the tryptically digested protein cc1750 from *C. crescentus*. Comparison with Fig. 8 shows that the predicted spectrum does not necessarily agree with that observed due to the inefficiency of trypsin digestion, uncertainty in the fragments complexing with cations ( $\text{Na}^+$ ,  $\text{K}^+$ , etc.), and inherent inefficiencies of the MALDI technique.

However, in the present study, 100% efficiency is assumed for demonstration purposes.

The prediction of mass spectra is difficult because many factors influence the spectral profile of the matrix-assisted laser desorption ionization (MALDI) technique. These include inherent inefficiencies in the MALDI technique, pH dependent ion charges, uncertainty in ionization of the fragments, complexation of cations or the matrix to the analyte and the propensity of certain amino acids to oxi-

dize. Another difficulty lies in predicting the integrated line intensity, as it is not directly correlated with concentration, but would be in a simulation. Therefore, only qualitative information on line-integrated intensity is sought through a simulation for identification purposes, and not for quantitative information. The integrated line intensity will change with concentration of a given protein, giving us qualitative information about cell state through protein extract analysis.

Table 3

Comparison between the observed and synthetic mass spectra of the tryptically digested protein cc1750 from *C. crescentus*<sup>a</sup>

Observed mass (Daltons)	Observed intensity	Synthetic mass (Daltons)	Mass difference
705.29	2788	683.18514	22.10486
705.25	2541	683.18514	22.06486
705.33	2537	683.18514	22.14486
705.21	2521	683.18514	22.02486
705.37	2337	683.18514	22.18486
705.17	2325	683.18514	21.98486
705.41	2149	683.18514	22.22486
705.44	2128	683.18514	22.25486
1303.73	1971	1287.16314	16.56686
705.48	1945	683.18514	22.29486
705.52	1903	683.18514	22.33486
1309.75	1884	1287.16314	22.58686
1303.68	1862	1287.16314	16.51686
1309.7	1801	1287.16314	22.53686
1303.78	1754	1287.16314	16.61686
705.56	1545	683.18514	22.37486
644.1	1490	645.08217	0.98217
706.23	1477	669.15068	37.07932
705.13	1468	683.18514	21.94486
705.13	1468	666.15687	38.97313

<sup>a</sup> This result shows that the predicted spectrum agrees with that observed taking into account the possibility of the fragments complexing with cations ( $\text{Na}^+$ ,  $\text{K}^+$ , etc.) as well as the possible oxidation of certain amino acids.

In Fig. 8, the experimentally determined mass spectrum of protein *cc1750* from *C. crescentus* is given (Karty et al., 2002; gene sequence found on <http://www.ncbi.nlm.nih.gov/>). Fig. 9 represents the predicted mass spectrum generated from a simulation for the same protein. The gene with 1953 nucleotides, which encodes for protein *cc1750*, was used as input; through Karyote transcription and translation kinetics the protein is translated and then digested with trypsin mathematically and these fragments populations were plotted assuming a charge of  $+1$  ( $M + H^+$ , as observed in the experiments). Table 3 shows the comparison between the experimental and synthetic mass spectra. There are 111 coincidences between the observed and synthetic spectra (with 610 data points) and the 20 most intense peaks in the observed spectrum were matches (Table 3). This result shows that the predicted spectrum agrees with that observed taking into account the small mass differences due to fragments complexing with cations ( $Na^+$ ,  $K^+$ , etc.) as well as the possible oxidation of certain amino acids. In the comparison between the experimental and predicted mass spectra, there were 38% hits involving the 75% highest peak intensities. Our number of coincidences is low in relation to what is usually assumed to be required for an experimentally determined identification of a protein sequence. However, this serves as a suggestive, be it preliminary, result.

This demonstrates the potential predictive utility of our model/predicted mass spectra. Given the sequence for the active genes, the protein mass spectrum can be generated. This will allow for help in identification of proteins expressed at a given stage of the cell cycle and the genes generating them. While the demonstration was for one gene, in a multiple gene case where a time series measurement is taken, the spectrum will change due to the changing relative populations of the expressed proteins. Difficulties with our approach may arise due to inefficiencies of trypsin digestion, uncertainty in the fragments complexing with cations ( $Na^+$ ,  $K^+$ , etc.), inefficiency of ionization and inherent error in the MALDI technique. Improvements in the predictive power of our approach are underdevelopment by accounting for a wider spectrum of complexing ions and a new approach for the rules of trypsin digestion to aid us in acquiring more accurate  $m/z$  ratios.

#### 4. Conclusion

Advanced cellular modeling holds great promise for use in medical sciences and bio-technical applications. With further development, our approach may be used to identify the vulnerabilities in abnormal versus normal cells, predict possible pathways by which a cell targeted by a drug molecule can thwart its effect, and identify metabolic signatures of abnormal cells for medical diagnostics. We have shown how one may embed nucleotide and amino acid polymerization kinetics in a compartmentalized metabolic model by formulating the entire problem in a chemical kinetic framework.

Our model accounts for the wide separation of time scales common in cellular activity through our multiple scale technique, which is fully automated in our implementation. Our implementation of our slow manifold projection methodology automates the underlying algebra for equilibrated and steady state (complex cycles) fast processes. Mass is conserved so that the total amount of low concentration co-factors or other molecules and their complexes with enzymes are accounted for.

Gene sequence data is used as input and resulting mRNAs and proteins are produced mathematically while mass conservation of the nucleotides and amino acids is maintained through the coupling of DNA, mRNA, protein and metabolite dynamics. As our model accounts for a broad range of reaction and transport processes, it can predict a full array of measurable parameters that can be used to calibrate and test it based on information theory (Sayyed-Ahmad et al., 2003).

#### Acknowledgements

This work was supported in part by a grant from the US Department of Energy (DE-FG02-01ER25498), Indiana Genomics Consortium, and Indiana 21st Century Science and Technology Fund. We greatly appreciate conversations with Professors J. Richardson; J. Karty and Professor J. Reilly for use of their mass spectral data from *Caulobacter crescentus*. We thank the referees for a number of constructive criticisms, which have been addressed in the revised manuscript.

#### References

- Arkin, A.P., 1998. Stochastic kinetic analysis of pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* 149 (4), 1633–1648.
- Barrett, A., Rawlings, N., Woessner, F., 1998. Handbook of Proteolytic Enzymes. Academic Press, San Diego.
- Bartol, T., Stiles, J.R., Sejnowski, T., Salpeter, M., Salpeter, E., 1997. Mcell is: A General Monte Carlo Simulator of Cellular Microphysiology. Found on website <http://www.mcell.cnl.salk.edu/>
- Bartol, T., Stiles, J.R., Sejnowski, T., Salpeter, M., Salpeter, E., 2001. Synaptic Variability: New Insights from Reconstructions and Monte Carlo Simulations with Mcell Synapses. John Hopkins University Press, Baltimore, MD, pp. 681–731.
- Field, R.J., Burger, M., 1985. Oscillations and Traveling Waves in Chemical Systems. John Wiley and Sons, New York, NY.
- Frith, C.A.J.M., Bray, D., 2000. Stochastic simulation of cell signaling pathways. In: Bower, J.M. Bolouri, H. (Eds.), Computational Modeling of Genetic and Biochemical Networks. The MIT Press, Cambridge, MA pp. 263–286.
- Goldbeter, A., 1996. Biochemical Oscillations and Cellular Rhythms: The Molecular Basis of Periodic and Chaotic Behavior. Cambridge University Press, Cambridge.
- Gillespie, D.T., 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* 22, 403–434.
- Gillespie, D.T., 1977. Exact stochastic simulation of coupled reactions. *J. Phys. Chem.* 81, 2340–2361.

- Hess, B., Boiteux, A., 1973. Substrate control of glycolytic oscillations. In: Chance, B., Pye, E.K., Ghosh, A., Hess, B. (Eds.), *Biological and Biochemical Oscillations*. Academic Press, New York, pp. 229–241.
- Hines, M., 1989. A program for simulation of nerve equations with branching geometries. *Int. J. Biomed. Comput.* 24, 55–68.
- Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., and the rest of the SBML Forum: Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgman, T.C., Hofmeyr, J.-H., Hunter P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U., Le Novère, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang J., 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models, *Bioinformatics* 19, 524–531.
- Karty, J.A., Ireland, M.E., Brun, Y.V., Reilly, J.P., 2002. Defining absolute confidence limits in the identification of *Caulobacter* proteins by peptide mass mapping. *J. Proteome Res.* 1 (4), 325–335.
- Larter, R., Ortoleva, P.J., 1981. A theoretical basis for self-electrophoresis. *J. Theor. Biol.* 88, 599–630.
- Larter, R., Ortoleva, P.J., 1982. A study of instability to electrical symmetry breaking in unicellular systems. *J. Theor. Biol.* 96, 175–200.
- LeNovere, N., Shimizu, T.S., 2001. StochSim: modeling of stochastic biomolecular processes. *Bioinformatics* 17, 575–576.
- Lewin, B., 1997. *Genes VI* (Chapter 11). Oxford University Press, New York, NY (Chapter 11).
- Loew, L.M., Schaff, J.C., 2001. The virtual cell: a software environment for computational cell biology. *Trends Biotechnol.* 19 (10), 401–406.
- Mannella, C.A., Pfeiffer, D.R., Bradshaw, P.C., Moraru, I.I., Slepchenko, B., Loew, L.M., Hsieh, C.E., Buttle, K., Marko, M., 2001. Topology of the mitochondrial inner membrane: dynamics and bioenergetic implications. *IUBMB Life* 52 (3,4,5), 93–100.
- McAdams, H.H., Arkin, A., 1997. Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA* 94, 814–819.
- Mendes, P., 1993. Gepasi—a software package for modeling the dynamics, steady-states and control of biochemical and other systems. *Comput. Appl. Biosci.* 9 (5), 563–571.
- Mendes, P., 1997. Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* 22 (9), 361–363.
- Mendes, P., Kell, D.B., 1998. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* 14 (10), 869–883.
- Mendes, P., Kell, D.B., 2001. MEG (Model Extender for Gepasi): a program for the modeling of complex, heterogeneous, cellular systems. *Bioinformatics* 17 (3), 288–289.
- Mendes, P., Martin, A.M., Cordeiro, C., Freire, A.P., 2001. In situ kinetic analysis of glyoxalase II in *Saccharomyces cerevisiae*. *Eur. J. Biochem.* 268 (14), 3930–3936.
- Navid, A., Ortoleva, P., 2003. Simulated nonlinear dynamics of simulated glycolysis in the protozoan parasite *Trypanosoma brucei*. *J. Theor. Biol.* submitted for publication.
- Niedhardt, F.C., Curtiss, R., Ingraham, J.C., Lin, E.C.C., Low, K.B., Magasanik, B., Rexnikoff, W.S., Riley, M., Schaechter, M., Umbarger, H.E., 1996. *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology, second edition. ASM Press, Washington, DC.
- Ortoleva, P., 1992. *Nonlinear Chemical Waves*. John Wiley and Sons, New York.
- Ortoleva, P., 1994. *Geochemical Self-Organization*. Oxford University Press, New York.
- Ortoleva, P.J., Ross, J., 1975. Studies in dissipative phenomena with biological application. Membranes, dissipative structures and evolution. In: Nicolis, G., Lefever, R. (Eds.), *Advances in Chemical Physics*, vol. 29. Wiley, New York, pp. 49–61.
- Press, W., Teukolsky, S., Vetterling, W., Flannery, B., 1992. *Numerical Recipes in C*, second edition. Cambridge University Press, Cambridge.
- Progogine, I., Lefever, R., 1968. Symmetry Breaking Instabilities in Dissipative Systems II. *J. Chem. Phys.* 48 (4), 1695–1700.
- Rashevsky, N., 1960. *Mathematical Biophysics Physico-Mathematical Foundations of Biology*, vol. 1 and 2, third edition. Dover Publications, New York, New York.
- Sauro, H.M., 1993. SCAMP: a general purpose simulator and metabolic control analysis program. *CABIOS* 9 (4), 441–450.
- Sauro, H.M., Kholodenko, B.N., Westerhoff, H., 1994. Metabolic control analysis of linked moiety-conserved cycles. Responses to perturbations of internal variables and conservation totals. *Eur. J. Biochem.* 225, 179–186.
- Sayyed-Ahmad, A., Tuncay, K., Ortoleva, P., 2003. Towards automated cell development through information theory. *J. Phys. Chem.*, in press.
- Schaff, J.C., Loew, L.M., 1999. The Virtual Cell. *Pacific Symposium on Biocomputing*, 4, 228–239. <http://www.nrcam.uchc.edu/>
- Schaff, J.C., Slepchenko, B.M., Loew, L.M., 2000. Physiological modeling with virtual cell framework. *Method Enzymol.* 321, 1–23.
- Schaff, J.C., Slepchenko, B.M., Choi, Y.S., Wagner, J., Resasco, D., Loew, L.M., 2001a. Analysis of nonlinear dynamics on arbitrary geometries with the virtual cell. *Chaos* 11 (1), 115–131.
- Schaff, J.C., Moraru, I.I., Slepchenko, B.M., Lucio, D.A., Means, S.A., Wagner, J.M., Loew, L.M., 2001b. Improvements to the virtual cell modeling environment. *Biophys. J.* 82 (1), 2310.
- Shimizu, T.S., Bray, D., 2001. In: Kitano, H. (Ed.), *Computational Cell Biology—The Stochastic Approach*. Foundations of Systems Biology. The MIT Press, Cambridge, MA.
- Slepchenko, B.M., Schaff, J.C., Carson, J.H., Loew, L.M., 2002. Computational cell biology: spatiotemporal simulation of cellular events. *Annu. Rev. Bioph. Biom.* 31, 423–441.
- Tomita, M.K., 2001. Whole-cell simulation: a grand challenge of the 21st century. *Trends Biotechnol.* 19 (6), 205–210.
- Tomita, M.K., Hashimoto, K., Takahashi, K., Shimizu, T.S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J.C., Hutchinson, C.A., 1999. E-CELL: software environment for whole-cell simulation. *Bioinformatics* 15 (1), 72–84.
- Tomita, M.K., Hashimoto, K., Takahashi, K., Matsuzaki, Y., Matsushima, R., Saito, K., Yugi, K., Miyoshi, F., Nakano, H., Tanida, S., Saito, Y., Kawase, A., Watanabe, N., Shimizu, T.S., Nakayama, Y., 2000. The E-cell project: towards integrative simulation of cellular processes. *New Generat. Comput.* 18, 1–12.
- Turing, A.M., 1952. The chemical basis of morphogenesis. *Phil. Trans. R. Soc. B London* 237, 37–72.
- Weston, B., Kuzmine, I., Martin, C., 1997. Positioning of the start site in the initiation of transcription by bacteriophage T7 RNA polymerase. *J. Mol. Biol.* 272 (1), 21–30.
- Wilson, M., Bower, J., 1991. A computer simulation of oscillatory behavior in primary visual cerebral cortex. *Neural Computation* 3, 498–509.