UNIVERSIDADE ESTADUAL DE SANTA CRUZ
PRO-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL
EM CIÊNCIA E TECNOLOGIA

FREDERICO CHAVES CARVALHO

MÉTODO ESTOCÁSTICO PARA SIMULAÇÃO INTEGRADA DE MODELOS COMPUTACIONAIS HETEROGÊNEOS DE SISTEMAS BIOLÓGICOS

PPGMC – UESC

ILHÉUS-BA
2020

FREDERICO CHAVES CARVALHO

MÉTODO ESTOCÁSTICO PARA SIMULAÇÃO INTEGRADA DE MODELOS COMPUTACIONAIS HETEROGÊNEOS DE SISTEMAS BIOLÓGICOS

PPGMC – UESC

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional em Ciência e Tecnologia da Universidade Estadual de Santa Cruz, como parte das exigências para obtenção do título de Mestre em Modelagem Computacional em Ciência e Tecnologia.

Orientador: Prof. Dr. Paulo Eduardo Ambrosio

Coorientador: Prof. Dr. Leandro Lopes Loguércio

ILHÉUS-BA
2020

FREDERICO CHAVES CARVALHO

MÉTODO ESTOCÁSTICO PARA SIMULAÇÃO INTEGRADA DE MODELOS COMPUTACIONAIS HETEROGÊNEOS DE SISTEMAS BIOLÓGICOS

PPGMC – UESC

Ilhéus-BA, 02/04/2020

Comissão Examinadora

Prof. Dr. Paulo Eduardo Ambrosio
UESC
(Orientador)

Prof. Dr. Leandro Lopes Loguércio
UESC
(Coorientador)

Profa. Dra. Fernanda Amato Gaiotto
UESC

Profa. Dra. Raquel Cardoso de Melo Minardi
UFMG

Dedico este trabalho à minha família e a todos aqueles que, um dia, poderão se beneficiar dos avanços tecnológicos que a modelagem computacional ajudará a criar.

Agradecimentos

- Minha gratidão maior é primeiramente aos meus pais, Maria da Pena e Allan Carvalho, à minha irmã Joanna Isis e aos meus avós Carmem, João (Zito), Adão (in memoriam) e Teodora (in memoriam). O suporte, incentivo e apoio incondicional em todas as etapas de minha vida foram essenciais para eu chegar até aqui. Obrigado por me ensinarem o valor da determinação, do foco e do trabalho duro para conquistar os meus objetivos. Com vocês aprendi a buscar crescer cada vez mais, a sonhar e expandir meus horizontes.
- Agradeço aos meus orientadores Prof. Dr. Paulo Eduardo Ambrósio e Prof. Dr. Leandro Lopes Loguércio por terem embarcado comigo nesta jornada de quase dois anos, me ajudando sempre com orientações, sugestões e ensinamentos. Obrigado também pela confiança, companheirismo e paciência que demonstraram do início ao fim desta trajetória.
- Sou grato também ao Prof. Dr. Gildemar Carneiro dos Santos, à Profª. Drª. Claudiani Waiandt, à Profª. Drª. Paula Frassinetti Cavalante, ao Prof. Dr. Esbel Tomás Vallero Orellana e à Profª. Drª. Fernanda Amato Gaioto pelo apoio e motivação em diversos momentos da minha trajetória acadêmica antes e/ou durante o mestrado, e pelo incentivo a seguir adiante nas próximas etapas. Professores como vocês e meus orientadores são uma inspiração para todos os estudantes que aspiram à carreira acadêmica.
- Obrigado aos colegas do mestrado, em especial ao Luís Augusto Franco, à Maíra Kersul, à Natália Jordana, ao Luiz Vinícius Soglia e ao Marcus Vinícius Sodré por compartilharem conhecimentos, bons momentos e caronas. A convivência com vocês tornou minha estadia na UESC mais agradável e produtiva.
- Gratidão também ao Programa de Pós Graduação em Modelagem Computacional da UESC, por oferecer a infraestrutura necessária à condução de diversas etapas desta pesquisa, e à toda equipe do programa, em especial a Ellen Pitombo, pelo suporte em nas questões administrativas e burocráticas, e pelas boas conversas e risadas.
- Por fim, meus sinceros agradecimentos à Fundação de Amparo à Pesquisa no Estado da Bahia (FAPESB) pelo apoio financeiro em forma de bolsa durante todo o período da pesquisa.

‘”I thought there couldn’t be anything as complicated as the universe, until I started reading about the cell.” Eric de Silva

Método estocástico para simulação integrada de modelos computacionais heterogêneos de sistemas biológicos

PPGMC – UESC

Resumo

Modelos computacionais de sistemas biológicos são uma das principais ferramentas utilizadas pela Biologia de Sistemas, seja para estudar detalhadamente fenômenos biológicos, testar hipóteses ou até mesmo para realizar experimentos *in silico*. Apesar dos avanços tecnológicos das últimas décadas, a exemplo das técnicas de sequenciamento de alto rendimento, e da atual abundância de dados, a construção de modelos mais detalhados e precisos de sistemas biológicos permanece uma tarefa desafiadora. Dentre os desafios está a escassez de técnicas computacionais que permitam criar modelos de grande porte de maneira escalável e reutilizável, a exemplo de modelos de célula completa, que buscam representar todos os processos celulares de forma a conseguir reproduzir *in silico* os mesmos comportamentos verificados experimentalmente. A necessidade de utilização de diferentes formalismos matemáticos para representar diferentes processos, a dificuldade em realizar simulações multi-algorítmicas de maneira coerente, e a escassez de softwares de fácil utilização para a construção e simulação de modelos de grande porte são algumas das atuais adversidades. Neste trabalho, propõe-se uma metodologia que permite o processamento simultâneo de modelos computacionais que sigam diferentes formalismos matemáticos, efetivamente integrando-os em uma simulação multi-algorítmica. Para facilitar a aplicação da metodologia a novos modelos, e modelos já existentes, desenvolveu-se também um software com interface gráfica amigável, que permite construir modelos computacionais graficamente, editá-los e simulá-los utilizando algoritmos de distintos formalismos para cada um de seus submodelos componentes. Com o auxílio do software desenvolvido, a metodologia foi aplicada a dois modelos: um deles representando o uma reação enzimática em duas etapas e um segundo modelo híbrido descrevendo o funcionamento do Operon Lac num sistema onde novas moléculas de lactose são constantemente injetadas. Os resultados dos testes realizados demonstraram boa precisão e viabilidade de aplicação da metodologia proposta para modelos de variadas complexidades e escalas. As funcionalidades do software desenvolvido também se mostraram satisfatórias em seu propósito de facilitar tanto a criação quanto modificação de modelos computacionais de sistemas biológicos. Entretanto, os resultados dos testes indicaram que o custo computacional da aplicação da metodologia é seu maior fator limitante.

Palavras-chave: Integração de modelos. Simulação multi-algorítmica. Modelagem de sistemas biológicos. Software para Biologia de Sistemas.

Stochastic method for integrated simulation of heterogeneous computer models of biological systems

PPGMC - UESC

Abstract

Computational models of biological systems are one of the main tools used by Systems Biology, whether to study biological phenomena in detail, test hypotheses or even to perform *in silico* experiments. Despite technological advances in recent decades, such as high-throughput sequencing techniques, and the current abundance of data, building more detailed and accurate models of biological systems remains a challenging task. Among the challenges is the scarcity of computational techniques that can be used to create large-scale models in a scalable and reusable way, such as whole-cell models, which seek to represent all cellular processes in order to be able to reproduce *in silico* experimentally verified behaviors. The need to use different mathematical formalisms to represent different processes, the difficulty in performing multi-algorithmic simulations in a coherent way, and the scarcity of user-friendly software for the construction and simulation of large models are some of the current adversities. Here we propose a methodology for simultaneous processing of computational models that follow different mathematical formalisms, effectively integrating them in a multi-algorithmic simulation. A software with user-friendly graphical user inteface was also created to facilitate the application of the methodology to new models and existing models, allowing the user to build and edit models graphically and simulate them using algorithms of different formalisms for each of its component submodels. With the aid of this software, the methodology was applied to two models: one representing a two-stage enzymatic reaction and a second hybrid model describing the operation of the LAC Operon in a system where new lactose molecules are constantly injected. The results of the tests performed corroborate the good precision and feasibility of the proposed methodology for models of varying complexities and scales. The functionalities of software were also satisfactory in its purpose of facilitating both the creation and modification of computational models of biological systems. However, the test results indicate that the computational cost of applying the methodology is a major limiting factor.

Keywords: Model integration. Multi-algorithmic simulation. Modelling of biological systems. Systems Biology software.

Lista de figuras

Lista de tabelas

Lista de quadros

Lista de abreviaturas e siglas

CME	Equação Mestra da Química (Chemical Master Equation)
DNA	Ácido Desoxirribonucleico
EDO	Equações diferenciais ordinárias
FBA	Flux Balance Analysis
KEGG	Kyoto Enciclopedia of Genes and Genomes
RNA	Ácido Ribonucléico
SBGN	Systems Biology Graphical Notation
SBML	Systems Biology Markup Language
SSA	Algorítmo de simulação estocástica (stochastic simulation algorithm)

Sumário

1 Introdução

Desde sua concepção, os computadores têm dado suporte a avanços científicos significativos em diversas áreas do conhecimento, e com a Biologia não foi diferente. A capacidade de realizar cálculos complexos e com alta repetitividade promoveu avanços científicos sem precedentes. Um desses avanços foi a possibilidade de desenvolvimento de modelos matemáticos mais complexos permitindo também a recriação melhorada de modelos já existentes.

Os primeiros modelos matemáticos aplicados à Biologia dos quais se tem registro datam de cerca de 300 anos atrás (ASHRAFIAN, 2013). Tratavam-se de modelos simplificados de fenômenos e aspectos como a movimentação, relações entre proporções anatômicas e, principalmente, descrições estatísticas de observações. Nos séculos seguintes, com a evolução da matemática, da física e da química, novas aplicações se tornaram possíveis. Todavia, ainda sob o predomínio das aplicações estatísticas(LEFÈVRE et al., 2014). Somente em meados dos anos 1940, com os primeiros avanços de computadores programáveis, que modelos mais complexos passaram a ser criados, incluindo modelos aplicados à medicina buscando representar fenômenos a exemplo da circulação sanguínea, da condução de impulsos elétricos pelos nervos, do ciclo de vida de bactérias, etc. (BRADHAM, 1964). O advento de computadores digitais e a sua popularização em universidades e centros de pesquisa foram fatores importantes para a disseminação da prática da modelagem de sistemas biológicos, resultando em um crescimento expressivo de modelos mais novos e avançados(BAIANU, 1986).

O progresso da computação também teve papel importante na evolução das ferramentas e técnicas de aquisição e tratamento de dados e, eventualmente, deu origem aos métodos classificados como métodos de alto rendimento, que são capazes de realizar análises extensivas, gerando grandes volumes de dados. A disponibilidade desse novo tipo de tecnologia permitiu a realização de projetos em escala inédita, a exemplo do projeto genoma humano, além do sequenciamento dos genomas de diversos outros organismos, como o da bactéria *Mycoplasma genitalium* (FRASER et al., 1995). O sucesso desses empreendimentos e a contínua redução dos custos para sequenciamento de genomas contribuíram para avanços importantes na área da Genômica (COLLINS et al., 2003). Como resultado, a quantidade de dados biológicos experimentais (especialmente genômicos) gerados tem apresentado crescimento exponencial, e estima-se que até 2025 a genômica gere anualmente um volume de dados superior aos gerados pela astronomia ou pelas redes sociais, que são hoje os maiores campos da Big Data (STEPHENS et al., 2015).

Além dos dados genômicos, dados biológicos experimentais de diversas nature-

zas também tem se multiplicado em similar escala, exigindo constante atualização da capacidade de armazenamento de dados (COOK et al., 2016). Como exemplo, o European Bioinformatics Institute (EBI), em seus relatórios anuais, ressalta o crescimento de diversos tipos de dados armazenados em seus servidores, com destaque para o volume ocupado por sequências de nucleotídeos (9,24 Petabytes) e o European Genome-phenome Archive (5,85 Petabytes), que armazena dados relacionais de genótipos e fenótipos resultantes de pesquisas biomédicas (EBML, 2017).

Entretanto, a maior parte dos dados gerados permanece subutilizada. A alta complexidade das análises necessárias, o custo elevado de ferramentas computacionais com capacidade para análise e utilização desses dados (STEPHENS et al., 2015), bem como a inadequação dos currículos acadêmicos atuais no tocante ao ensino e treinamento matemático e computacional oferecido aos estudantes das áreas que mais se beneficiariam com o uso de tais dados(SCHATZ, 2012) são apontados como algumas das principais causas desta subutilização de dados.

Em face do crescimento exponencial no volume de dados, a comunidade científica tem se empenhado em desenvolver novas técnicas e recursos capazes de processar as novas informações, extraindo delas novos conhecimentos e hipóteses testáveis(LARRAÑAGA et al., 2006). Neste quesito, duas aplicações distintas da modelagem computacional têm alcançado êxito no aproveitamento eficaz destes dados para geração de novos conhecimentos: a criação modelos estatísticos voltados ao aprendizado de máquina (*machine learning*) e a criação de modelos mecanísticos para estudo do comportamento dos sistemas(BAKER et al., 2018).

Os modelos voltados ao aprendizado de máquina visam analisar grandes volumes de dados para encontrar padrões relacionando as entradas às saídas, por isso essa abordagem tem se destacado ao oferecer soluções para análises aprofundadas de conjuntos massivos de dados (ANGERMUELLER et al., 2016). Aplicações notórias das técnicas de aprendizado de máquina incluem o uso de informações clínicas para identificação da segurança e efetividade de medicamentos para tratamento de doenças diversas (MCMAHON et al., 2020; CAMACHO et al., 2018) e, mais recentemente, análise de dados de expressão gênica para inferir fatores que podem provocar alterações na regulação gênica (TAREEN; KINNEY, 2019). Contudo, vale ressaltar que, além de precisar de uma quantidade muito grande de dados para ser eficaz, modelos de aprendizado de máquina oferecem predições estatísticas cuja aplicação muitas vezes é limitada a fenômenos diretamente relacionados aos dados utilizados para treinar os modelos.

Já os modelos mecanísticos, por sua vez, buscam estabelecer relações de causalidade entre os dados de entrada e os de saída, focando nos processos ou mecanismos por trás dos fenômenos estudados. Este tipo de modelo requer uma quantidade menor de dados, porém mais específicos que os dados para modelos de aprendizado de má-

quina. Comumente os modelos mecanísticos buscam replicar, através de simulações, comportamentos observados em organismos vivos (ex.: metabolismo) ou envolvendo estes organismos (ex.: teias alimentares), mas não se limitam a reproduzir resultados experimentais (KLIPP et al., 2016). Uma vez construídos, calibrados e validados, modelos mecanísticos possuem a capacidade de integrar conhecimentos biológicos permitindo que cientistas testem hipóteses, prevejam comportamentos e executem experimentos *in silico* (NEAL et al., 2014).

Historicamente, a modelagem mecanística de sistemas biológicos apresenta a tendência de se recorrer ao reducionismo para a criação de modelos como forma de contornar os desafios atuais, sejam eles de natureza científica (ex.: falta de dados específicos) ou tecnológicas (ex.: inadequação dos recursos computacionais) (STEUER, 2007). Tal abordagem se limita a retratar os sistemas em isolamento e falha em capturar os efeitos das propriedades emergentes das interações entre distintos sistemas (SIMEONOV et al., 2013). Além disso, as abordagens, procedimentos e tecnologias disponíveis atualmente frequentemente resultam na criação de modelos com baixa possibilidade de reutilização e dificuldade de reprodução dos resultados. Para solucionar este problema e facilitar o uso de modelos atuais para a construção de modelos cada vez mais abrangentes, novos métodos, ferramentas computacionais e padrões são necessários (MEDLEY et al., 2016).

Nas últimas décadas, o interesse na elaboração de modelos mais abrangentes tem crescido significativamente. Esta mudança de paradigma, apesar de latente, já tem resultado na criação de alguns modelos capazes de melhor capturar comportamentos novos resultantes da interação entre sistemas distintos, além de possuírem aplicabilidade mais ampla. Um dos exemplos mais notáveis é o modelo de célula completa, que busca elucidar todos os processos metabólicos celulares e efeitos dos genes conhecidos, com o objetivo de prever fenótipos através do genótipo de uma determinada célula (KARR et al., 2012). As potencialidades de um modelo de célula completa capaz de reproduzir com boa precisão resultados experimentais são inúmeras e capazes de revolucionar a biologia atual. Por exemplo, a redução significativa no custo de pesquisas e no tempo necessário para sua execução, bem como a possibilidade de projetar e desenvolver organismos sintéticos capazes de realizar funções de interesse médico (ex.:biossensores para diagnóstico) ou econômico (ex.: organismos que produzam biocombustíveis a baixo custo) (CARRERA; COVERT, 2015; PURCELL et al., 2013).

Entretanto, com o aumento da escala dos eventos representados pelos modelos, aumentam também os desafios e dificuldades na coleta de dados, na construção, na simulação e na análise dos resultados. Primeiramente, os dados necessários se tornam mais numerosos e mais difíceis de se obter, dada a sua especificidade. Soma-se a isso a falta de ferramentas computacionais apropriadas para construção e simulação de modelos maiores, o que torna a complexidade e custo computacional fatores proibitivos. Por

fim, a inexistência de técnicas de integração e simulação que ofereçam a possibilidade de reutilizar modelos prontos como subunidades de um modelo mais abrangente torna ainda mais onerosa esta missão (MACKLIN et al., 2014). Tal nível de complexidade posiciona a criação de modelos precisos de célula completa como um dos "grandes desafios do século XXI" (TOMITA, 2001).

No que se refere ao desafio da integração de modelos, verifica-se que um dos pontos que torna tal atividade mais complicada é o fato de que nenhum formalismo per si é adequado para descrever de forma eficaz e eficiente todos os processos celulares, resultando na necessidade de utilização de distintas abordagens para modelar diferentes aspectos do funcionamento de uma célula (WALTEMATH; WOLKENHAUER, 2016). Por exemplo, vias metabólicas podem ser representadas por sistemas de equações diferenciais ordinárias, fenômenos como a transcrição e tradução podem utilizar algoritmo de simulação estocástica, o crescimento celular pode ser representado por fluxos de matéria orgânica, enquanto a regulação da expressão gênica pode ser simplificada como uma rede booleana (MACHADO et al., 2011).

Em face dos desafios supracitados, e considerando a necessidade de recorrer-se a distintos padrões matemáticos para a criação de novos e mais precisos modelos de célula completa, levanta-se o seguinte questionamento: como integrar modelos matematicamente heterogêneos de maneira a capturar as propriedades emergentes e oferecendo possibilidade de reutilização de modelos já existentes?

Buscando uma solução para este questionamento, este estudo tem o objetivo geral de desenvolver um método para simulação integrada que permita o processamento de modelos, mesmo que estes utilizem distintos algoritmos de simulação (determinísticos ou estocásticos), possibilitando também a reutilização de modelos existentes. Como objetivos específicos, buscou-se aplicar a metodologia para construir uma versão modificada de um processo biológico e elaborar uma ferramenta computacional que permita a aplicação simplificada da metodologia a modelos, sejam eles novos ou existentes. O primeiro desses objetivos específicos visou demonstrar a aplicabilidade da técnica para a criação de modelos mais complexos, enquanto o segundo objetivo tornou a reaplicação da técnica mais simples e expressa.

Diante do exposto, estruturou-se este trabalho em seis seções, incluindo a presente introdução. As seções e seus conteúdos são os que seguem:

A seção 2 traz uma análise do estado da arte da prática de construção de modelos computacionais mecanísticos de sistemas biológicos, enfatizando os modelos de célula completa e os atuais desafios, limitações e tecnologias relacionados a esses modelos. Em seguida as diferentes abordagens que podem ser utilizadas para a criação e simulação de modelos computacionais de sistemas biológicos são introduzidas e explicadas, dando ênfase às duas abordagens dinâmicas mais populares para este fim: os sistemas de

equações diferenciais ordinárias (EDOs) e o algoritmo de simulação estocástica (SSA). Em cada uma das partes, apresenta-se os dados necessários para a construção do tipo de modelo explicado, com exemplificações didáticas.

A seção 3 apresenta a metodologia adotada para o condução deste trabalho, bem como o processo de elaboração do método proposto para simulação integrada de modelos, elucidando as considerações e aproximações realizadas, os procedimentos seguidos para validação do método e os recursos e estratégias empregados na elaboração do software para aplicação expressa e reprodutível do método proposto.

A seção 4 expõe e detalha os resultados obtidos, iniciando-se pela validação do método proposto neste estudo através de sua aplicação a modelos selecionados para verificar sua viabilidade e desempenho computacional. Em seguida, as funcionalidades oferecidas pela ferramenta computacional desenvolvida são apresentadas e exemplificadas através da construção de um modelo simples.

Na seção 5, é realizado um breve apanhado geral dos resultados da pesquisa, seguido de uma discussão onde são ressaltadas as vantagens, desvantagens e limitações observadas nos testes realizados com a aplicação do método e com os testes de usabilidade do software desenvolvido.

Por fim, a seção 6 traz as conclusões e considerações finais do projeto, apresentando também as potencialidades de aplicação e sugestões para futuros trabalhos.

2 Revisão Bibliográfica

2.1 Estado da arte, desafios e limitações da modelagem de sistemas biológicos

2.1.1 Biologia de Sistemas

O século XX foi marcado pela criação de técnicas e tecnologias que revolucionaram o fazer científico e ofereceram suporte a avanços nos conhecimentos em diversas áreas. Na Biologia, o desenvolvimento das primeiras técnicas de sequenciamento de nucleotídeos, e a posterior criação dos sequenciadores automáticos representaram o início de uma revolução que culminou no surgimento das ciências ômicas. A habilidade de "ler o DNA"permitiu que cientistas tivessem acesso a dados que possibilitaram o desenvolvimento de uma série de novas aplicações. Entre elas, estão a utilização de dados de sequenciamento de genomas para reconstrução de árvores filogenéticas, identificação da função dos genes, e a criação de modelos computacionais englobando todo o metabolismo de uma célula.

Um dos maiores expoentes dessa revolução tenha sido o projeto Genoma Humano. Iniciado em 1990, esse projeto tinha objetivos que iam além do puro sequenciamento do genoma humano. Nos 14 anos do projeto, parte dos mais de três bilhões de dólares investidos foi utilizada para o aprimoramento tecnológico necessário ao avanço do projeto, incluindo o melhoramento das técnicas e instrumentos de sequenciamento e o desenvolvimento de técnicas e ferramentas computacionais para dar suporte às análises dos dados genômicos (ENERGY, 2013). O projeto Genoma Humano é um exemplo que ilustra bem a relação simbiótica que se estabeleceu entre a Biologia, a Computação e outras tecnologias, capturada nas palavras de Alan Aderem:

> "A Biologia dita quais novas tecnologias e ferramentas computacionais devem ser desenvolvidas. Essas novas ferramentas abrem novas fronteiras a serem exploradas na Biologia. Dessa forma, a Biologia impulsiona a tecnologia e a Computação, e por sua vez, a tecnologia e a Computação revolucionam a Biologia."[1] (ZOU; LAUBICHLER, 2018)

Das colaborações entre Biologia e Computação nasceram a Bioinformática e a Biologia Computacional. A Bioinformática é definida como o campo que foca na aplicação de ferramentas computacionais e analíticas para aquisição e interpretação de dados biológicos (BAYAT, 2002). Já a Biologia Computacional é a área voltada à

[1] Tradução nossa.

aplicação de métodos matemáticos e computacionais para a criação de modelos de sistemas biológicos, partindo do conhecimento biológico e de dados experimentais (MURPHY, 2016). A junção dos esforços dessas duas áreas resulta na Biologia de Sistemas, que nas palavras de Ron Germain pode ser definida como:

> "[...] a abordagem científica que combina princípios de Engenharia, Matemática, Física e Ciências da Computação com extensivos dados experimentais para desenvolver um entendimento quantitativo, bem como profundamente conceitual, dos fenômenos biológicos, permitindo a predição e simulação precisa de comportamentos biológicos complexos (e emergentes)."[2] (CHRISTOPHER WANJEK, 2011)

Nas últimas décadas, a Biologia de Sistemas tem ganhado bastante destaque por oferecer infraestrutura e soluções tecnológicas importantes que tem permitido traduzir a recente abundância de dados ômicos em novos novos conhecimentos. Por exemplo, ferramentas de comparação de sequências, como o Basic Local Alignment Search Tool, conhecido como BLAST (ALTSCHUL et al., 1990), tem tornado possível a identificação de novos genes e mutações. Tais instrumentos, aliados ao crescimento de bancos de dados como o Kyoto Encyclopedia of Genes and Genomes, ou KEGG (KANEHISA et al., 2017), conferem maior agilidade às pesquisas além de fornecerem dados para novos caminhos de investigação.

O rápido desenvolvimento experimentado pela Biologia de Sistemas e suas contribuições para novas tecnologias que impactam diversas outras áreas a coloca como um dos campos do conhecimento mais promissores da atualidade. Dentre as maiores promessas da Biologia de Sistemas, destacam-se a possibilidade substituir experimentos em laboratórios por experimentos virtuais (FREDDOLINO; TAVAZOIE, 2012), ferramentas que permitam projetar microrganismos sintéticos para funções específicas (PURCELL et al., 2013) e até mesmo a criação de medicamentos personalizados ao genoma de cada pessoa (SZIGETI et al., 2018). No centro de todas essas promessas estão os modelos de célula completa.

2.1.2 Modelos de célula completa: O estado da arte da Biologia de Sistemas

A Biologia de Sistemas moderna ainda se encontra em sua infância (ZOU; LAUBICHLER, 2018) e as fronteiras atuais desta área são tão diversas quanto suas aplicações. Do ponto de vista da modelagem de sistemas biológicos, os modelos de célula completa podem ser considerados como o próximo objetivo a ser alcançado (SZIGETI et al., 2018). Propostos ainda na década de 1980 (MOROWITZ, 1984), este tipo de modelo começou

[2] Tradução nossa.

a se tornar possível graças à identificação e sequenciamento do organismo com o menor genoma conhecido (FRASER et al., 1995).

O sequenciamento da bactéria *Mycoplasma genitalium* representou um passo fundamental para que o modelo proposto por Morowitz (1984) fosse criado. A capacidade prevista do modelo de reproduzir *in silico* experimentos laboratoriais seria uma forma de mensurar a completude do conhecimento da biologia molecular. Logo após o sequenciamento da *M. genitalium*, iniciou-se o projeto E-CELL, representando os primeiros esforços para a criação do primeiro modelo computacional de uma célula completa. Concluída a primeira etapa do projeto, a equipe foi capaz de criar um modelo simplificado da bactéria contendo 127 dos seus 525 genes (TOMITA et al., 1999).

O modelo desenvolvido pelo projeto E-CELL representa 4 vias metabólicas: glicólise, biossíntese de fosfolipídios, transcrição e tradução, além de representar o meio extra-celular e o transporte de substâncias para dentro (glicose, ácidos graxos e glicerol) e para fora (lactato) da célula. Todos os processos supracitados foram modelados utilizando sistemas de equações diferenciais ordinárias, conferindo ao modelo um caráter determinístico. A decisão por utilizar uma única abordagem tornou possível que o modelo assumisse uma estrutura monolítica, ou seja, sem particionamento em módulos. A simplicidade do modelo permitiu que as simulações levassem somente 5% do tempo do ciclo celular esperado para a *M. genitalium* (9 horas) (TOMITA, 2001).

Por definição, um modelo de célula completa é um modelo computacional que "considera a função integrada de todos os genes e moléculas presentes em uma célula"[3] (CARRERA; COVERT, 2015), e cujo objetivo é possibilitar "predições de fenômenos biológicos complexos e integrados"[3] (SANGHVI et al., 2013). Neste aspecto, o modelo construído pelo projeto E-CELL não poderia ser classificado como um modelo de célula completa propriamente dito, uma vez em que falha em representar a ação de todos os genes da célula da bactéria *M. genitalium*. Entretanto, os esforços empreendidos na missão de criar um modelo de célula completa foram importantes para revelar limitações tecnológicas e no conhecimento, além de estabelecer uma metodologia que serviria de base para novas tentativas de criação de um modelo do tipo.

A construção do primeiro modelo de uma célula completa pode ser atribuída à equipe liderada pelo professor Markus Covert, da Universidade de Stanford, nos Estados Unidos (HAYES, 2013). O modelo da bactéria *M. genitalium* leva em consideração o efeito de todos os 525 genes da bactéria, além de ser capaz de reproduzir, em simulação, uma vasta quantidade de resultados experimentais (KARR et al., 2012). Para a construção do modelo foram utilizados dados genômicos, proteômicos, transcriptômicos e metabolômicos da bactéria retratada, coletados através da análise de mais de 900 artigos e organizados num banco de dados próprio (KARR et al., 2013). Diferentemente

[3]Tradução nossa.

do seu antecessor, este modelo consiste de 28 submodelos que utilizam 4 distintas abordagens: sistemas de equações diferenciais ordinárias (EDO), *flux balance analysis* (FBA), algorítmo de simulação estocástica (SSA) e redes booleanas (WALTEMATH et al., 2016). O método utilizado para coordenar e integrar a simulação dos módulos baseou-se na suposição de que, para curtos intervalos de tempo (1 segundo, neste caso), cada submodelo se comportaria de maneira independente. Dessa maneira, a cada passo de 1 segundo, todos os modelos eram simulados separadamente e tinham seus resultados sincronizados antes do próximo passo temporal. O novo estado de cada uma das variáveis do sistema era verificado, e caso fossem matematicamente possíveis (ex.: não houvesse concentrações negativas), o procedimento era repetido sucessivamente até que o critério de parada fosse alcançado (KARR et al., 2012; GOLDBERG et al., 2016).

Apesar do último modelo ser bem sucedido em sua missão de reproduzir *in silico* resultados experimentais com razoável precisão, alguns dos objetivos da construção de modelos de célula completa permanecem inalcançados. O modelo publicado em 2012 não foi capaz de prever novos comportamentos (SANGHVI et al., 2013), o que indica a inaptidão do modelo de substituir, mesmo que parcialmente a realização de experimentos ou mesmo sugerir novos experimentos. Além disso, o tempo de simulação do modelo mostrou ser um de seus aspectos mais negativos, sendo superior ao tempo do ciclo celular da bactéria modelada (GOLDBERG et al., 2016). A complexidade, o grande volume de dados utilizados para modelagem e opção por um algoritmo de simulação sequencial foram os fatores que mais influenciaram para esta limitação.

Mais de três décadas após ser inicialmente proposto, a construção de um modelo de célula completa abrangente e preciso ainda permanece sendo "um dos grandes desafios do século XXI"(TOMITA, 2001). Entretanto, uma vez superado este desafio, estes modelos serão ferramentas de grande valor tanto para o avanço de pesquisas científicas quanto para aplicações comerciais. A possibilidade de utilizar simulações *in silico* para estudar comportamentos biológicos representaria grandes reduções de custos e maior agilidade para pesquisas, enquanto a combinação dessas mesmas simulações com técnicas de edição genética, como a CRISPR-CAS9, possibilitaria verificar os efeitos das modificações com antecedência. Estas e mais potencialidades tem levado a um crescimento contínuo do número de cientistas engajados em pesquisas relacionadas a estes modelos nos últimos anos, especialmente (SZIGETI et al., 2018; KARR et al., 2017).

2.1.3 Barreiras e limitações atuais

A criação de um modelo computacional de célula completa mostrou que desenvolver modelos abrangentes de sistemas biológicos com alto nível de detalhamento é factível. Tal feito também serviu para estabelecer as bases metodológicas a serem seguidas para se construir novos modelos do tipo. Entretanto, diversas barreiras e

limitações de natureza técnica e científica precisam ser superadas para que a elaboração de novos modelos seja viável e proveitosa.

Do ponto de vista científico, a complexidade inata do processo de criação de modelos em nível celular se traduz na necessidade de dados e parâmetros precisos e consistentes. Muito além do código genético de um organismo, modelar uma célula em computador requer identificação de todas as espécies químicas e reações que ocorrem no sistema, além da caracterização seus parâmetros e de seus componentes, tais como taxas de reação, constantes cinéticas e afinidades de ligação das proteínas (ISALAN, 2012). A dificuldade e custos envolvidos na elaboração e condução de experimentos para obtenção desses dados resulta, muitas vezes na escassez dessas informações, que são necessárias não só para a construção, como também para a calibração e validação dos modelos.

Outro fator que impacta sensivelmente o processo de construção de modelos de célula completa é a heterogeneidade dos dados experimentais disponíveis. A natureza aparentemente desordenada e caótica da biologia em seus diversos níveis gera variações nas medições empíricas. Tais variações são muitas vezes exacerbadas por oscilações nas condições experimentais e no meio de cultura utilizados, podendo originar dados significativamente divergentes (PALSSON; ZENGLER, 2010). Adicionalmente, o uso de distintas técnicas, metodologias e/ou equipamentos pode ocasionar resultados experimentais discordantes. Por esta razão, um tratamento rigoroso dos dados e seu minucioso estudo são essenciais para garantir a qualidade do modelo final (ANDRES; EILIS, 2006).

Já sob a ótica dos aspectos técnicos, destacam-se dentre as principais barreiras: a diversidade de abordagens aplicáveis à construção de modelos, a carência de ferramentas adequadas para a construção e simulação de modelos holísticos (MACKLIN et al., 2014; TAKAHASHI et al., 2002). Além de prejudicar o processo de criação de modelos precisos, tais barreiras tornam o processo de colaboração para criação de modelos mais difícil.

Cada aspecto do comportamento celular pode ser modelado seguindo-se diferentes abordagens matemático-computacionais. As vias metabólicas de um organismo unicelular, por exemplo, podem ser modeladas e simuladas utilizando-se equações diferenciais ordinárias, equações diferenciais parciais, equações diferenciais estocásticas (ou o algoritmo de simulação estocástico), técnicas de modelagem baseadas em regras ou análise de balanço de fluxo (TAKAHASHI et al., 2002). Se por um lado tal diversidade oferece flexibilidade para a criação de modelos, por outro resulta na dificuldade em sua reutilização e melhoramento, principalmente devido à utilização de diferentes variáveis, que podem ser inclusive de distintas naturezas (concentrações, contagens de moléculas e valores booleanos, por exemplo) (MACKLIN et al., 2014). Além disso, a complexidade

da tarefa de integrar modelos se torna ainda maior com a utilização de formalismos distintos.

Por fim, a escala e nível de detalhamento de um modelo computacional resulta não só na necessidade de maior quantidade e diversidade de dados, como também na necessidade de poder computacional para sua simulação. Atualmente, poucas ferramentas de suporte à modelagem oferecem possibilidade de otimização dos modelos e opções para uso de técnicas de computação paralela (WALTEMATH et al., 2016). Além disso, muitas da ferramentas atuais falham em oferecer suporte aos padrões estabelecidos atuais e opções de distintos formalismos para construção de modelos (WALTEMATH; WOLKENHAUER, 2016). No tocante às funcionalidades oferecidas, cabe ainda destacar que as ferramentas atuais para criação e simulação e modelos computacionais falham, em sua maioria, em oferecer conveniências que simplifiquem o processo de desenvolvimento, a exemplo da possibilidade de construir modelos graficamente (MYERS et al., 2009).

A deficiência de ferramentas para construção e simulação de modelos computacionalmente eficientes torna inconveniente a escalabilidade de modelos existentes, uma vez que os tempos de simulação de modelos de grande porte é proibitivo. O modelo atual de célula completa, por exemplo, necessita de aproximadamente 9 horas para ser simulado, tempo que supera a média do ciclo celular da bactéria representada. Estima-se que, seguindo os mesmos padrões utilizados para modelar a bactéria *M. genitalium*, o modelo de uma célula humana precisaria levaria de 10^4 a 10^5 dias para ser simulado (GOLDBERG et al., 2016), tempo que inviabiliza a construção e simulação de tais modelos.

2.1.4 Tecnologias e soluções emergentes

Em face dos obstáculos que precisam ser superados para a criação de modelos computacionais robustos e precisos, diversas inovações vem sendo desenvolvidas em diversas frentes para viabilizar a criação de modelos computacionais de células completas. Na última década, notou-se uma melhora sensível na infraestrutura da Biologia de Sistemas, sob a forma de novos e melhorados bancos de dados, consolidação de padrões e surgimento de novas e melhoradas ferramentas para modelagem (SZIGETI et al., 2018; STANFORD et al., 2015). Tais avanços tem contribuído para criação de soluções para as atuais limitações da área.

Os avanços nas tecnologias de medição e o constante barateamento dos custos relacionados à obtenção de grandes volumes de dados tem motivado os bancos de dados atuais a se reinventar. Bancos de dados já bem estabelecidos, a exemplo do KEGG(<kegg.jp/>), recentemente introduziram novas ferramentas e melhoraram a organização de seus dados com vistas a facilitar o acesso e utilização. Além disso, novas

utilidades visam facilitar a criação de softwares e plataformas com acesso direto ao conteúdo dos bancos de dados.

Com o crescente interesse em modelar sistemas cada vez mais complexos de células específicas, a importância de bancos de dados especializados em fornecer informações sobre todos os aspectos de um organismo tem aumentado. Conhecidos como Pathway/Genome databases, bancos de dados como BioCyC (KARP et al., 2017) e, mais recentemente, WholeCellKB (KARR et al., 2013) oferecem a vantagem de reunir, em um único lugar, informações de grande interesse para a modelagem de organismos específicos, e são um excelente ponto de partida, uma vez que possibilitam a realização de análises preliminares e identificação de entidades químicas que desempenham funções essenciais para o sistema, por exemplo.

Ao passo em que a comunidade dedicada à modelagem de sistemas biológicos cresce, iniciativas como COMBINE (<co.mbine.org/>) e o Center for Reproducible Biomedical Modeling (<reproduciblebiomodels.org>) têm se destacado na missão de consolidar padrões e procedimentos metodológicos através da realização de conferências, publicações, cursos e elaboração de materiais de apoio. Ademais, a contínua evolução dos padrões e formatos de arquivos estabelecidos (HUCKA et al., 2018) os torna cada vez mais abrangentes e adequados às funções e recursos necessários à modelagem.

Outra frente que tem apresentado avanços recentes se refere às ferramentas de construção e simulação de modelos. Apesar das atuais limitações, verifica-se o surgimento de esforços em desenvolver novas ferramentas e melhoramento das ferramentas já existentes e consolidadas. Uma das tendências recentes na área é a transformação de softwares em plataformas online que auxiliam nas diversas etapas da modelagem, desde a coleta de dados até a análise dos resultados (GHOSH et al., 2011). Concomitantemente, novos algoritmos vem sendo desenvolvidos para facilitar o suporte aos formatos mais populares de arquivo para compartilhamento de modelos (SOMOGYI et al., 2015), o que por sua vez ajuda a reforçar o estabelecimento desses formatos como padrão na área. Em 2017, por exemplo, 296 softwares e ferramentas computacionais já ofereciam suporte ao formato SBML (SBML, 2017). O número representa um avanço na quantidade de softwares quando comparado aos 230 existentes em 2011 (HUCKA et al., 2011).

Em suma, a despeito de todos os desafios e limitações atuais, diversos esforços tem sido empreendidos no sentido de fomentar a criação e compartilhamento de modelos computacionais, bem como a inclusão desses recursos em pesquisas cientificas. Entretanto, para viabilizar a criação de modelos cada vez mais complexos é necessário coordenação dos esforços atuais e futuros no sentido de estabelecer métodos e procedimentos padronizados que permitam a criação de modelos reutilizáveis, que possam

servir de componentes para modelos futuros, mais abrangentes. Dessa forma, a colaboração científica também será facilitada, tornando a construção de modelos precisos e preditivos de célula completa de diversos organismos algo possível e viável.

2.2 Classificação dos modelos de sistemas biológicos e seus algoritmos de simulação

Modelos computacionais mecanísticos podem ser construídos para cumprir diversas funções, dentre as quais podemos destacar a necessidade de explicar observações experimentais, relacionar observações distintas ou fazer predições sobre comportamento de um determinado sistema (KLIPP et al., 2016). A dimensão do sistema representado, as aproximações realizadas e o nível de detalhamento objetivado são alguns dos fatores que podem influenciar significativamente o modelo final (BRODLAND, 2015).

Dadas as distintas necessidades e finalidades da modelagem, diversos abordagens matemáticas e computacionais foram desenvolvidas e empregadas na construção de modelos mecanísticos de sistemas biológicos (GLONT et al., 2018). Dentre as mais populares destacam-se a utilização de equações diferenciais para representar o comportamento dinâmico dos sistemas, a modelagem baseada em restrições e a modelos baseados em redes de Petri. Entretanto, outros tipos de modelo, a exemplo dos modelos baseados em regras e modelos lógicos tem aumentado nos últimos anos (MALIK-SHERIFF et al., 2019). Essas diferentes abordagens geram modelos que podem ser classificados de acordo com sua natureza (quantitativos ou qualitativos), nível de detalhamento e número de componentes (STEUER, 2007).

Alternativamente, os modelos podem ser classificados de acordo com sua topologia e a metodologia utilizada para criá-los. Segundo estes critérios, os modelos de sistemas biológicos podem pertencer a um de três grandes grupos: modelos baseados em redes; modelos baseados em regras; e modelos estatísticos (KLIPP et al., 2016). Cada um desses grupos requer diferentes tipos de dados de entrada e podem gerar dados de saída igualmente variados, além de apresentarem mecanismos de funcionamento distintos.

Os modelos baseados em redes são apropriados quando os sistemas sendo modelados podem ser descritos como um conjunto de nós interconectados por relações bem definidas. As vias metabólicas são um exemplo desse tipo de modelo, pois possuem metabólitos como nós, e as reações químicas como conexões entre os nós (HOU et al., 2016). Tais modelos podem ser submetidos a simulações tanto estáticas (ex.:algoritmo de simulação lógica, flux balanace analysis, etc.) quanto dinâmicas (ex.: pela utilização de equações diferenciais ordinárias, equações diferenciais parciais, algoritmos de simulação estocástica, etc.).

O segundo grupo de modelos são os modelos baseados em regras, ou modelos baseados em agentes. Assim como os primeiros, estes possuem elementos claramente definidos, porém, a relação entre eles ocorre através de um conjunto de regras, e não reações ou interações que podem ser descritas por equações (HARRIS et al., 2016). Esse tipo de modelo simplifica todas as possíveis interações entre moléculas, e é de grande utilidade quando se quer modelar sistemas em que o estado de cada um dos componentes é relevante para as análises, podendo-se levar em consideração ou não dados temporais e / ou espaciais (DANOS et al., 2007). Um exemplo de sistema que pode ser representado segundo esta abordagem são as vias de sinalização celular, onde em um determinado momento, os complexos moleculares podem assumir um estado entre diversas opções (fosforilado, metilado, ubiquitinado, etc.).

Por fim, os modelos estatísticos, incluídos no último grupo, são modelos mais apropriados para trabalhar com modelos criados a partir de grande volume de dados experimentais. Esses modelos tem se tornado mais comuns e importantes à medida que o volume de dados se torna maior, já que através deste tipo de modelagem é possível investigar relações de causalidade e frequências de ocorrência de determinados eventos dentro de um conjunto de dados. Um exemplo é a utilização de métodos como a inferência Bayesiana e a máxima verossimilhança para reconstruir árvores filogenéticas, identificando relações evolucionárias entre espécies baseado em seus genomas (Da Silva et al., 2017).

É importante ressaltar que apesar de estarem proximamente relacionados, os modelos e suas simulações geralmente são independentes. Modelos buscam representar um determinado sistema seja de forma gráfica, matemática, meio de regras, etc. Já as simulações são o processamento dos modelos com vistas a obter informações sobre seu comportamento em determinadas condições. Muitas vezes, um modelo construído seguindo-se uma abordagem poderá ser processado por distintos algoritmos de simulação. Como exemplo, pode-se citar um modelo metabólico baseado em redes, o qual podem ser simuladas de maneira determinística ou estocástica sem necessidade de alterações significativas nos dados utilizados. Para tanto, é necessário que todos os dados necessários à aplicação de um algoritmo alternativo sejam conhecidos.

O quadro 1 sumariza e compara as principais características de seis das abordagens mais populares para desenvolvimento de modelos computacionais de sistemas biológicos.

Neste trabalho, o foco é a construção de modelos cinéticos baseados em redes, com destaque aos modelos capazes de ser simulados por equações diferenciais ordinárias, que são determinísticos, e pelo algoritmo de simulação estocástica, que levam em consideração as aleatoriedades nas reações químicas, buscando reproduzi-las na simulação. A escolha dessas abordagens baseou-se em dois fatores principais: sua po-

pularidade na comunidade acadêmica dedicada à modelagem de sistemas biológicos, e a possibilidade de modelar e simular eventos em diversas escalas espaciais e temporais através dessas abordagens.

Quadro 1 – Comparação dos principais tipos de simulação de modelos.

	Natureza		Resultados		Variablidade		Tempo		Dados utilizados						Aplicabilidade		
Tipo de simulação	Estático	Dinâmico	Quantitativo	Qualitativo	Determinístico	Estocástico	Discreto	Contínuo	Topologia	Estequiometria	Parâmetros cinéticos	Concentrações	Nº de moléculas	Volume celular	Sinalização celular	Regulação gênica	Vias metabólicas
Sistemas de EDOs		x	x		x			x		x	x	x		•	x	x	x
Simulação estocástica (SSA)		x	x			x	x			x	x		x	•	x	x	x
Flux balance analysis (FBA)	x			x	x			•	x	x					•	•	x
Redes de Petri	x		x	x	x	•	•		x						x	•	x
Redes Booleanas	x			x	x		•		x						x	x	
Modelos baseados em regras	x	•	•	x	x	•	•	•		x	•	•	•	•	x		x

Legenda: **x** = Característica intrínseca • = Requer adaptações

Fonte: Elaborado pelo autor baseado em dados e informações de (BARDINI et al., 2017; MACHADO et al., 2011; FISHER; HENZINGER, 2007) .

2.2.1 Simulações baseados em equações diferenciais ordinárias

Uma das abordagens mais antigas, e ainda bastante utilizadas para a simulação de modelos computacionais de sistemas biológicos é a descrição matemática dos fenômenos sob a forma de equações diferenciais ordinárias (EDOs). Esta abordagem gera modelos dinâmicos e contínuos que simulam o comportamento de componentes do sistema modelado, dando origem a uma série temporal que representa a variação da concentração e/ou número de moléculas de cada uma das espécies.

Modelos simulados segundo essa abordagem consideram que as espécies que compõem o sistema estão homogeneamente distribuídas por todo o volume, sendo mais adequados para descrever sistemas em que o número de moléculas envolvidas é muito grande. Além disso, este tipo de simulação leva em consideração que possíveis variações na temperatura e na pressão do sistema são desprezíveis, não influenciando na constante de velocidade de reação (KLIPP et al., 2016).

A descrição de um sistema bioquímico em termos de EDOs pode ser obtida pela aplicação da Lei de ação das massas (WAAGE; GULBERG, 1864) ao sistema, ou pela utilização de uma das derivações desta lei, como é o caso das equações de Michaelis-Menten (PLUNKETT; GEMMILL, 1951).

A lei de ação das massas afirma que a taxa de reação é proporcional a uma constante multiplicada pelo produto das concentrações dos reagentes. A constante à qual a lei se refere recebe o nome de constante cinética, e é representada pela letra k.

A determinação do valor desta constante pode ser feito experimentalmente (CHEN et al., 2010), utilizando medições das concentrações de substratos e produtos ao longo de uma reação. Esta constante é convencionalmente definida como sendo a taxa média de reações ocorrendo em um determinado volume, dividida pelo produto da densidade média dos reagentes (GILLESPIE, 1976). Alternativamente, técnicas de estimativas de parâmetros podem ser utilizadas para obtenção de um valor aproximado para esta constante (ZHANG et al., 2015). Em ambos os casos, os valores publicados para estes parâmetros podem ser encontrados através de busca na literatura ou em bancos de dados especializados.

Conhecidas as reações que ocorrem no sistema, é possível iniciar a criação de um modelo matemático através da aplicação da lei da ação das massas. Vejamos como exemplo a equação 1, que representa uma reação enzimática em duas etapas, na qual a enzima E se liga ao substrato S na primeira reação, formando o complexo ES. A partir deste ponto, dois caminhos são possíveis: o composto ES pode se dissociar sem conversão de S (reação 2), ou S pode ser convertido no produto P (reação 3), o que causa a dissociação do complexo ES em E + P.

$$\mathrm{E + S} \underset{\mathrm{k\_2}}{\overset{\mathrm{k\_1}}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{\mathrm{k3}} \mathrm{E + P} \tag{1}$$

Utilizando a lei de ação das massas para descrever a reação matematicamente, obtemos o sistema de equações diferenciais ordinárias composto pelas equações 2, 3, 4, 5 que descrevem o comportamento dinâmico das espécies E, S, ES e P, respectivamente.

$$\frac{d[E]}{dt} = k_2[ES] - k_1[E][S] + k_3[ES] \tag{2}$$

$$\frac{d[S]}{dt} = k_2[ES] - k_1[E][S] \tag{3}$$

$$\frac{d[ES]}{dt} = k_1[E][S] - k_2[ES] - k_3[ES] \tag{4}$$

$$\frac{d[P]}{dt} = k_3[ES] \tag{5}$$

Onde [E] representa a concentração da enzima E, [S] represente a concentração da espécie S, [ES] representa a concentração do complexo ES e [P] representa a concentração da espécie P. Temos então um sistema de 4 equações, com 4 variáveis, que pode ser resolvido analiticamente ou numericamente, sendo também possível a utilização de algoritmos computacionais para este último caso. Uma vez resolvido o sistema, obtém-se

as expressões matemáticas que descrevem o comportamento de cada uma das espécies em função do tempo.

Considerando os valores fornecidos na tabela 1 para realização dos cálculos, é possível resolver o sistema de equações diferenciais apresentado acima, o que possibilita a obtenção do gráfico representado na figura 1, que retrata a evolução das concentrações em função do tempo.

Considerando os valores $k_{1+} = k_{1-} = k_2 = 1$ e as concentrações iniciais $[A]_0 = [B]_0 = 10, [AB]_0 = [C]_0 = 0$, chegamos ao gráfico mostrado na figura 1, que representa a evolução das concentrações em função do tempo.

Tabela 1 – Parâmetros e concentrações iniciais para simulação utilizando EDOs

Parâmetro	Valor	Unidade
k_1	$1 \cdot 10^6$	$M^{-1}s^{-1}$
k_2	$1 \cdot 10^{-4}$	s^{-1}
k_3	$0,1$	s^{-1}
$[E]_0$	$2 \cdot 10^{-7}$	M
$[S]_0$	$5 \cdot 10^{-7}$	M
$[ES]_0$	0	M
$[P]_0$	0	M

Fonte: Adaptado de (WILKINSON, 2018)

Da análise do gráfico mostrado na figura 1 e das equações 2, 3, 4 e 5, conclui-se que, fixados os valores das concentrações iniciais e das constantes cinéticas, o comportamento obtido será sempre o mesmo. Por este motivo, os modelos criados utilizando equações diferenciais ordinárias são denominados modelos determinísticos, pois desconsideram a possibilidade de variações aparentemente aleatórias e intrínsecas dos sistemas bioquímicos, gerando resultados aproximados, sendo portanto mais apropriado a sistemas onde a quantidade de moléculas envolvidas é grande o suficiente para que a ocorrência de reações possa ser considerada como um evento contínuo, e com taxa constante.

2.2.2 Modelos estocásticos

A utilização de equações diferenciais ordinárias para representar sistemas bioquímicos, apesar de conveniente e muitas vezes apropriada, falha em representar com razoável precisão alguns casos. Na prática, essa inadequação da utilização das EDOs ocorre principalmente em sistemas onde a reduzida quantidade de moléculas faz com que o sistema se torne mais sensível a pequenas flutuações naturais, como a necessidade de colisão entre moléculas para que haja reação e a disponibilidade de enzimas livres suficientes para catalisar as reações.

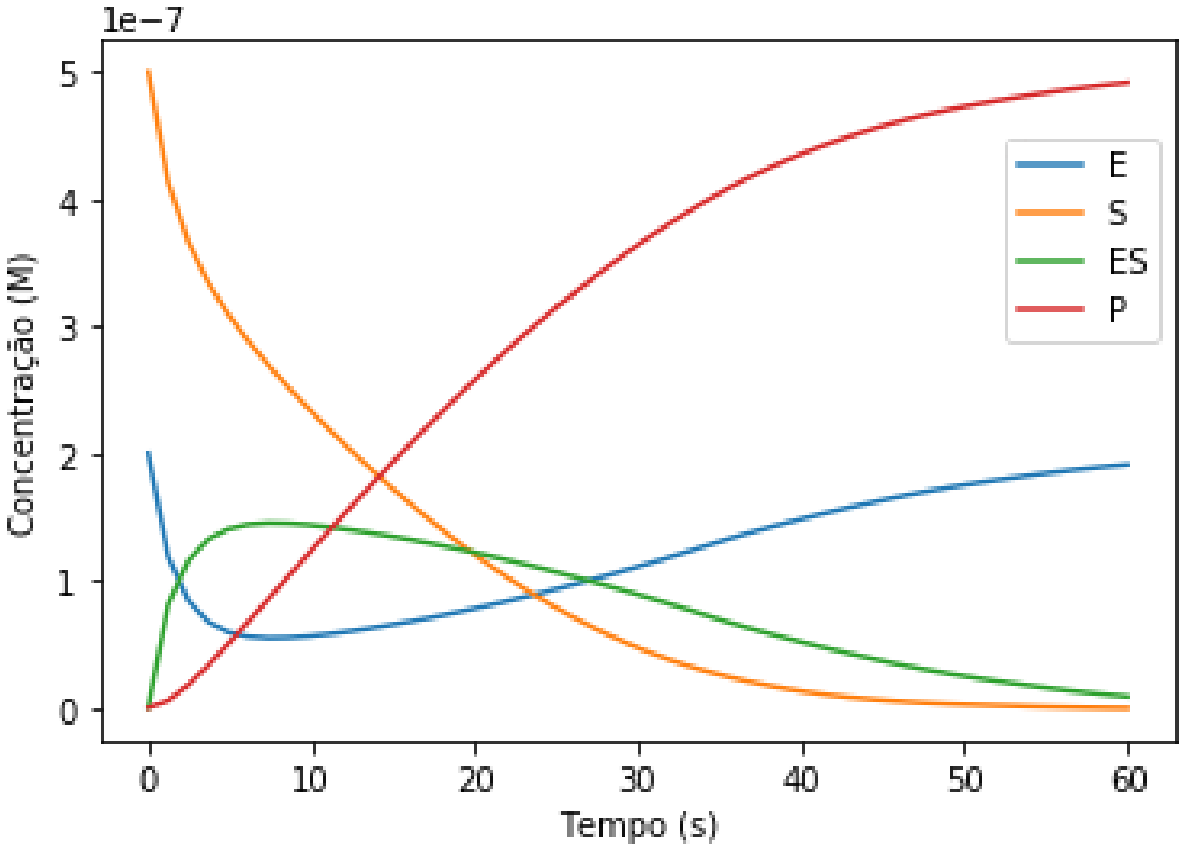


Figura 1 – Evolução temporal das concentrações das espécies E, S, ES e P

Nesses casos, a abordagem mais apropriada é a utilização da chamada equação mestra da química (CME, do inglês chemical master equation). A CME é uma equação diferencial que descreve as probabilidades do sistema se encontrar em um determinado estado (ou seja, possuir uma determinada quantidade de moléculas para cada um de seus componentes) em função do tempo (KLIPP et al., 2016). O estado do sistema é representado pelo vetor x, no qual cada linha representa o número de moléculas de cada espécie do sistema. Para um sistema com m espécies o estado do sistema pode ser representado como:

$$\mathrm{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_m \end{bmatrix}$$

É importante ressaltar, que a quantidade de moléculas de um composto na linha i x_i é uma função do tempo, dependendo também das reações que tenham ocorrido.

Assumindo que o sistema possua n distintas reações, podemos também definir a matriz estequiométrica N do sistema como sendo:

$$N = \begin{array}{c} \begin{array}{ccccc} R_1 & R_2 & R_3 & \cdots & R_n \end{array} \\ \begin{bmatrix} n_{11} & n_{12} & n_{13} & \cdots & n_{1n} \\ n_{21} & n_{22} & n_{23} & \cdots & n_{2n} \\ n_{31} & n_{32} & n_{33} & \cdots & n_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_{m1} & n_{m2} & n_{m3} & \cdots & n_{mn} \end{bmatrix} \end{array}$$

Cada valor n_{ij} equivale ao coeficiente estequiométrico da espécie x_i na reação R_j. Quando uma determinada espécie não participa da reação, então seu coeficiente será 0. Para a construção da matriz estequiométrica é necessário considerar que todas as reações do sistema são irreversíveis. Dessa maneira, reações reversíveis devem ser decompostas em duas reações irreversíveis em direções opostas, o que significa dizer que cada reação reversível gera duas colunas com sinais opostos na matriz estequiométrica do sistema.

Tomando como exemplo o sistema descrito pela equação 1, podemos construir sua matriz estequiométrica, que nomearemos de $N_{(1)}$.

$$N_{(1)} = \begin{array}{c} \begin{array}{ccc} R_{1+} & R_{1-} & R_2 \end{array} \\ \begin{bmatrix} -1 & 1 & 1 \\ -1 & 1 & 0 \\ 1 & -1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{c} A \\ B \\ AB \\ C \end{array}$$

Definidos o vetor estado e a matriz estequiométrica do sistema, podemos escrever a CME como sendo:

$$\frac{dp(x,t)}{dt} = \sum_j a_j(x - n_j) \cdot p(x - n_j, t) - \sum_j a_j(x) \cdot p(x,t) \tag{6}$$

Onde *p(x,t)* é a distribuição de probabilidades do sistema encontrar-se no estado *x* no tempo *t*, e n_j é a coluna equivalente à reação R_j na matriz estequiométrica. Cada termo $a_j(x)$ é chamado de propensão, e representa a probabilidade, dado o estado $x(t)$ da reação R_j ocorrer em algum momento no próximo intervalo infinitesimal de tempo $[t, t + dt)$ (GILLESPIE, 2007).

A propensão é uma função do número de moléculas que participam da reação R_j, e depende da natureza da reação. Para reações unimoleculares do tipo $S_1 \longrightarrow$ P, a propensão a_j pode ser calculada como sendo o produto da constante cinética estocástica c_j pelo número x_{S_1} de moléculas da espécie S_1 presentes no sistema, como mostrado

na equação 7a. Para reações entre duas moléculas de espécies distintas a exemplo de $S_1 + S_2 \longrightarrow P$ a propensão pode ser calculada pela equação 7b. Já quando duas moléculas de S_1 são consumidas numa mesma reação, a equação 7c deve ser utilizada para calcular a propensão.

$$a_j(x) = c_j \cdot x_{S_1} \tag{7a}$$

$$a_j(x) = c_j \cdot x_{S_1} \cdot x_{S_2} \tag{7b}$$

$$a_j(x) = c_j \cdot \frac{x_{S_1} \cdot (x_{S_1} - 1)}{2} \tag{7c}$$

Analisando as distintas fórmulas que podem ser assumidas pela propensão, percebe-se que seu cálculo pode ser generalizada como sendo uma constante multiplicada pelo número de combinações distintas das moléculas envolvidas nas reações. Assim, para toda reação multi-molecular envolvendo m espécies distintas e na qual r_i moléculas da espécie S_i são consumidas, a propensão é calculada pela equação 8.

$$a_j(x) = c_j \cdot \prod_{i=1}^{m} {}^{x_i}C_{r_i} = \frac{x_i!}{r_i!(x_i - r_i)!} \tag{8}$$

Dada a complexidade da CME (equação 6), tanto sua solução analítica quanto integração numérica por computador são praticamente intratáveis. Uma das soluções mais bem sucedidas foi proposta pelo físico Daniel T. Gillespie, em 1976 (GILLESPIE, 1976), sob a forma de um algoritmo capaz de chegar a uma solução exata para a CME de uma maneira computacionalmente eficiente. Para isso, o autor utiliza de distribuições de probabilidade para simplificar os cálculos, sem comprometer a exatidão dos resultados. O algoritmo foi nomeado de algoritmo de Gillespie (em homenagem ao seu criador), ou algorítmico de simulação estocástica (SSA, do inglês stochastic simulation algorithm), e sua versão original encontra-se reproduzida no anexo A, na linguagem FORTRAN. Já a figura 2 mostra a representação do algoritmo de Gillespie em forma de fluxograma.

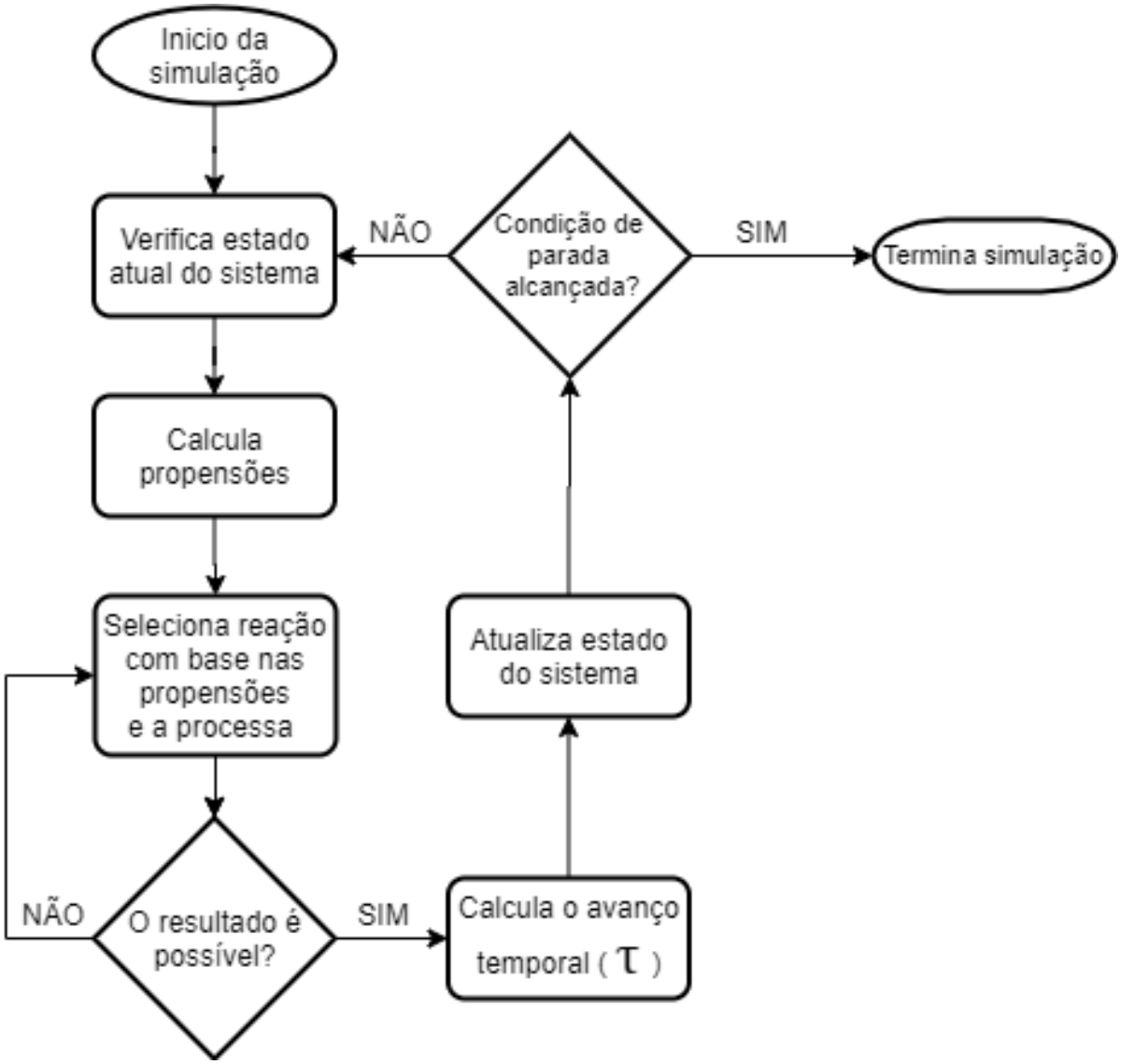


Figura 2 – Fluxograma do algoritmo de simulação estocástica

O caráter exato do algoritmo de Gillespie pode ser verificado através da análise da simulação da mesma reação enzimática utilizada na seção 2.2.1. Para aplicá-lo ao mesmo sistema, as medidas e parâmetros devem antes ser convertidos de maneira que as concentrações sejam transformadas em contagens de moléculas de cada espécie no sistema e as constantes cinéticas determinísticas sejam convertidas em suas equivalentes estocásticas. A tabela 2 mostra os valores já adaptados dos parâmetros e contagens de moléculas, levando-se em consideração um volume celular de 1 femtolitro ($10^{-15}L$).

Da análise dos distintos resultados obtidos com 5 execuções do SSA (figura 4) para o mesmo sistema, verifica-se que os efeitos das variações naturais do sistema são simulados pelo SSA, o que não ocorre em simulações determinísticas. Essas variações tem efeitos importantes em sistemas onde o número de moléculas é baixo, tornando-se menos significativos à medida que o número de moléculas no sistema aumenta.

Entretanto, o custo computacional de uma simulação por SSA é maior.

Tabela 2 – Parâmetros e condições iniciais para simulação utilizando SSA

Parâmetro	Valor	Unidade
c_1	$1,66 \cdot 10^{-3}$	moléculas$^{-1}s^{-1}$
c_2	$1 \cdot 10^{-4}$	s^{-1}
c_3	$0,1$	s^{-1}
$X(E)_0$	120	moléculas
$X(S)_0$	301	moléculas
$X(ES)_0$	0	moléculas
$X(P)_0$	0	moléculas

Fonte: Adaptado de (WILKINSON, 2018)

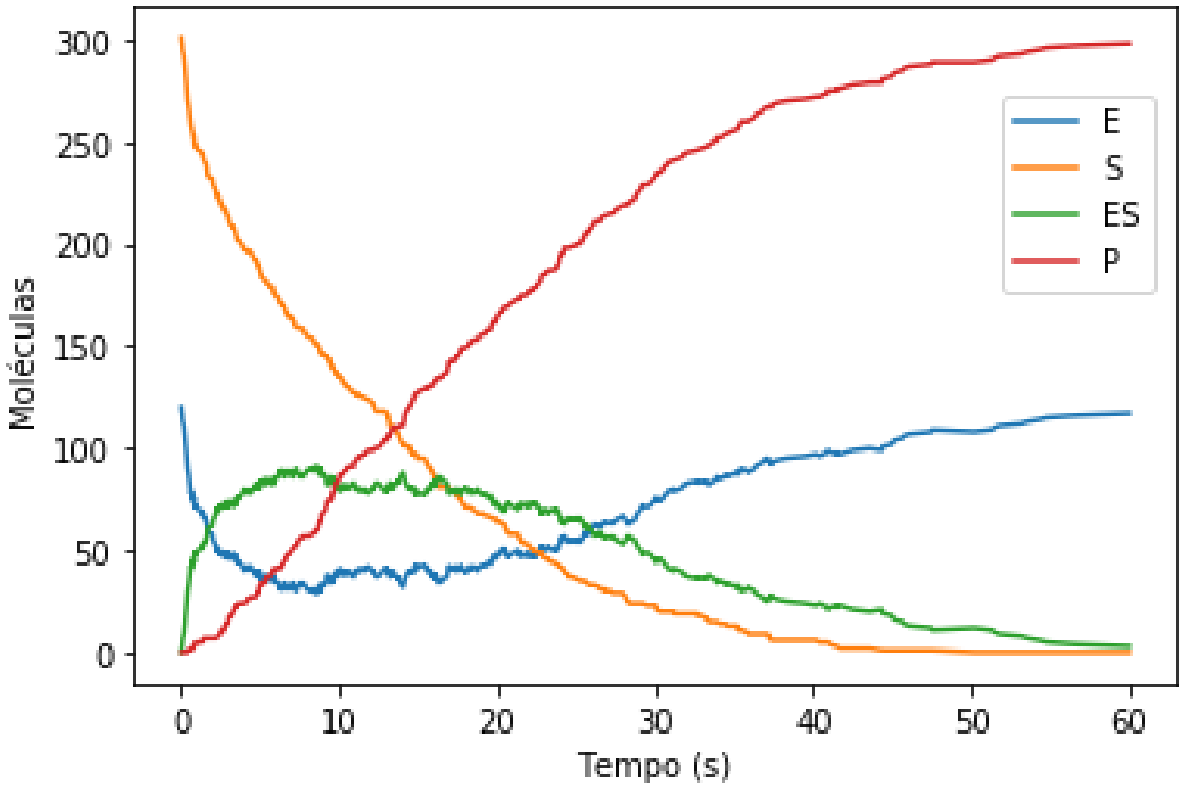


Figura 3 – Evolução temporal das concentrações das espécies E, S, ES e P

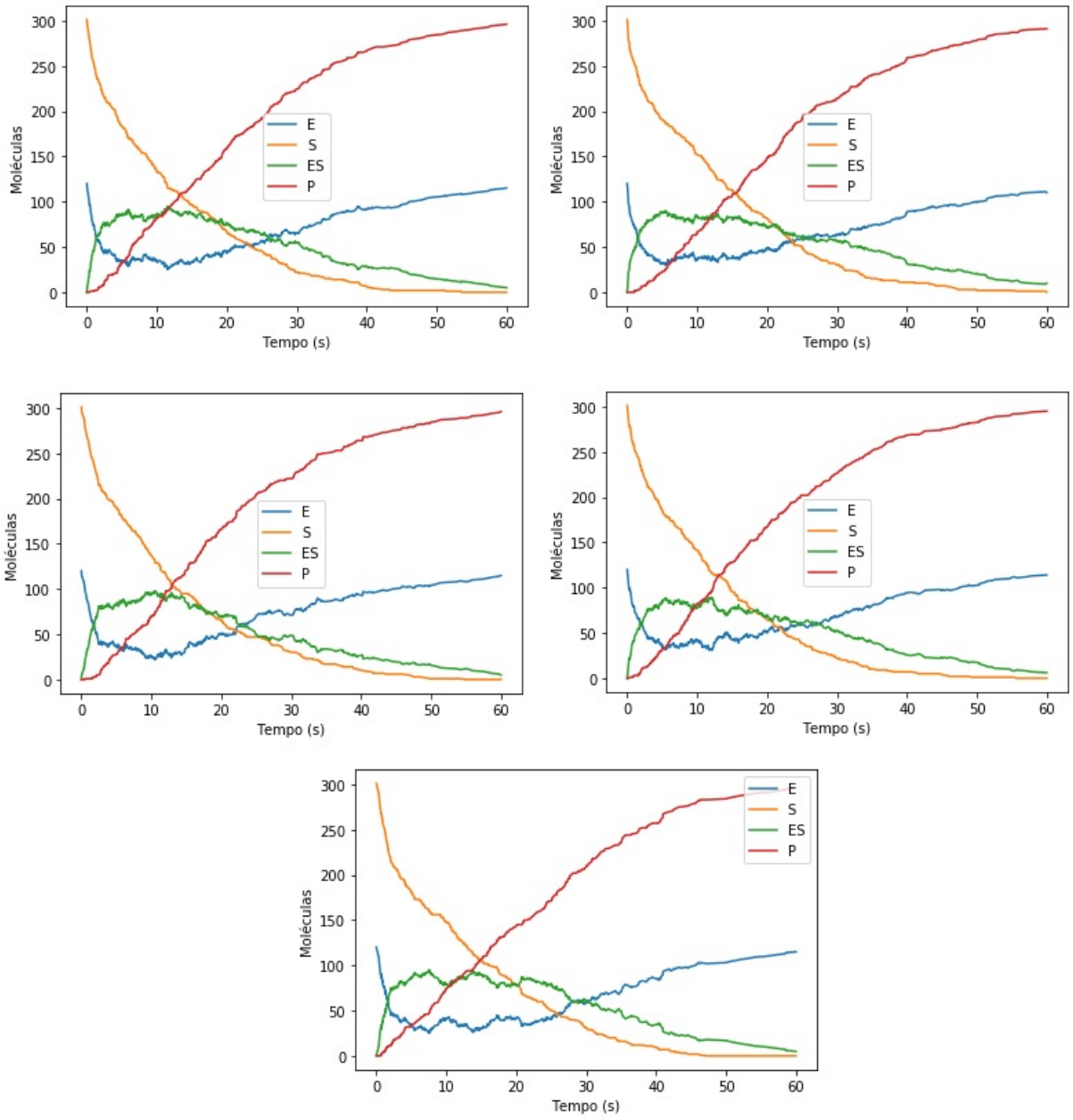


Figura 4 – Variações observadas em cinco simulações seguidas utilizando SSA

2.2.3 Equivalência entre variáveis e parâmetros dos modelos determinísticos e estocásticos

Quando um modelo determinístico é construído e simulado, a série temporal resultante traça a variação das concentrações de cada uma das espécies químicas do sistema. Já quando um modelo estocástico é submetido ao mesmo processo, o resultado obtido é a variação do número de moléculas de cada uma das espécies com o tempo. Dessa maneira, para garantir a homogeneidade dimensional da simulação integrada, ambos os modelos devem expressar as quantidades das espécies em contagem de

moléculas ou em concentração. Aqui, opta-se pela contagem de moléculas, convertendo os valores de concentração em número de moléculas para os modelos determinísticos. Entretanto, as conversões só são realizadas após finalizado cada passo da simulação, de maneira a evitar disrupções nos algoritmos originais de simulação.

O procedimento utilizado para a conversão consiste em primeiramente multiplicar a concentração (N, em M/L) pela constante de Avogadro (6,022 x 10^{23}), obtendo-se então o número de moléculas (X) por unidade de volume (v, em litros). Em seguida, multiplica-se o valor encontrado pelo volume do compartimento onde a espécie se encontra, já convertido para litro. Temos então a equação 9, que

$$X = N \cdot 6,022 \cdot 10^{23} \cdot v \tag{9}$$

A segunda conversão necessária se refere ao cálculo da constante cinética estocástica (c), partindo-se da constante cinética determinística (k). Para realizar tal conversão, podemos utilizar como ponto de partida a análise dimensional de cada uma das constantes. A constante cinética estocástica é medida em moléculas por segundo (X/s), enquanto a unidade de medida da constante cinética determinística varia com a ordem da reação (WILKINSON, 2018).

Para reações de ordem 0 (equação 11), a constante cinética possui dimensão Ms^{-1}. Dessa forma, a equivalência entre as constantes seria dada pela equação 12.

$$\varnothing \longrightarrow \text{A} \tag{10}$$

$$c = 6,022 \cdot 10^{23} \cdot v \cdot k \tag{11}$$

Para reações de ordem 1 (equação 13), a constante cinética determinística assume a dimensão s^{-1}, e portanto possui valor numérico equivalente ao de c, como representado na equação 14.

$$\text{A} \longrightarrow \text{B} \tag{12}$$

$$c = k \tag{13}$$

Para reações de ordem 2 (equação 15), a constante cinética determinística passa a ter dimensão $M^{-1}s^{-1}$, e portanto sua conversão para c é representada pela equação 16.

$$\text{A} + \text{B} \longrightarrow \text{C} \tag{14}$$

$$c = \frac{k}{6,022 \cdot 10^{23} \cdot v} \tag{15}$$

Generalizando, podemos assumir que a equivalência entre as constantes k e c é uma função da ordem o da equação química, que obedece a equação 17.

$$c = \frac{k}{(6,022 \cdot 10^{23} \cdot v)^{o-1}} \tag{16}$$

Aplicando-se as equações 9 e 16, é possível converter os dados apresentados na tabela 1, obtendo-se os valores informados na tabela 2. Para isso, consideraremos o volume do sistema como sendo 1 femtolitro (10^{-15}L). Além disso, como a ordem reação R_1 é 2 ($o = 2$), e as reações R_2 e R_3 possuem ordem 1 ($o = 1$), somente o valor de c_1 será diferente de k1.

$$X[S] = [S] \cdot 6,022 \cdot 10^{23} \cdot v = 5 \cdot 10^{-7} \cdot 6,022 \cdot 10^{23} \cdot 10^{-15} = 301 \text{ moléculas} \tag{17}$$

$$X[E] = [E] \cdot 6,022 \cdot 10^{23} \cdot v = 3 \cdot 10^{-7} \cdot 6,022 \cdot 10^{23} \cdot 10^{-15} = 120 \text{ moléculas} \tag{18}$$

$$c_1 = \frac{k_1}{(6,022 * 10^{23} \cdot v)^{o-1}} = \frac{1 \cdot 10^{6}}{(6,022 \cdot 10^{23} \cdot 10^{-15})^{o-1}} = 1,66 \cdot 10^{-3} \text{ moléculas}^{-1} s^{-1} \tag{19}$$

Por fim, é importante ressaltar que os valores da contagem de moléculas deve ser sempre inteiro, e portanto, a norma NBR 5891 é aplicada para fins de conversão dos valores quando necessário.

3 Metodologia

Com vistas a desenvolver uma técnica inovadora que permita a simulação multi-algorítmica de modelos de sistemas biológicos, esta pesquisa adota uma abordagem quali-quantitativa, analisando os métodos, ferramentas e padrões vigentes para a elaboração deste tipo de modelo, ao mesmo tempo que mensura o desempenho da nova proposta e o compara com outras abordagens. Neste aspecto, as estratégias exploratórias e explicativas são as mais adequadas para a condução deste estudo, cujo êxito requer tanto o aprofundamento teórico quanto a identificação e detalhamento dos mecanismos e princípios que regem o funcionamento de modelos computacionais de sistemas biológicos.

Devido à natureza interdisciplinar da proposta, tomou-se como ponto de partida a pesquisa bibliográfica abrangendo tópicos na seara da Química, da Biologia, da Matemática e da Computação. Em particular, buscou-se analisar as técnicas que permitiram a criação e processamento do único modelo de célula completa atual, o qual consiste de 28 submodelos integrados. Sabe-se que apesar de efetiva na missão de coordenar a simulação respeitando os aspectos físicos, químicos e biológicos, as técnicas então empregadas apresentam importantes limitações nos quesitos eficiência computacional, escalabilidade e possibilidade de melhoramento através da substituição de submodelos ou parâmetros (GLONT et al., 2018; FREDDOLINO; TAVAZOIE, 2012). Identificadas suas limitações e aspectos que precisam ser melhorados, buscou-se propor uma alternativa que seja capaz de suprir essas deficiências.

Concluída as fases iniciais do aprofundamento teórico, procedeu-se à caracterização dos requisitos para a criação de modelos de de processos biológicos, cuja simulação é um dos focos desta pesquisa. Todos os dados necessários foram extraídos da literatura examinada, tendo em vista que a validação do método depende do conhecimento prévio do comportamento esperado do sistema.

Com os resultados das análises em mãos, definida a abordagem e as fontes dos dados para testes e validação, procedeu-se ao desenvolvimento do método para integração de modelos em três etapas: proposição de um método alternativo para possibilitar a simulação multi-algorítmica de modelos, validação do método proposto e desenvolvimento de ferramenta para sua aplicação expressa.

3.1 Do desenvolvimento do método estocástico de simulação multi-algorítmica

Visando oferecer uma alternativa para solucionar um dos maiores desafios técnicos da construção e simulação de modelos em escala celular, aqui propomos um método baseado no algoritmo de simulação estocástica para permitir simulação de modelos construídos seguindo distintos formalismos. O método busca viabilizar a construção de modelos de forma seccionada e sua posterior integração, possibilitando a seleção dos formalismos mais apropriados para descrever cada um dos distintos processos celulares com o nível de detalhamento limitado apenas pela disponibilidade de dados experimentais.

Como implicações práticas da alternativa proposta, destacam-se a oportunidade de reutilizar modelos já construídos como partes componentes de modelos mais abrangentes (realizadas as adaptações necessárias) e a facilitação de colaborações científicas entre distintos grupos com o objetivo comum de desenvolver modelos que representem eventos ocorrendo em múltiplas escalas espaciais (ex.: moleculares e interações entre células). Ademais, o tratamento de cada submodelo como um módulo independente permitirá a substituição de modelos por versões mais atualizadas, o que facilitará a escalabilidade e melhoramento futuro dos modelos criados e simulados com o uso desta metodologia.

A seleção do algoritmo de simulação estocástica como inspiração e base para desenvolvimento deste método se deu devido a dois fatores principais. O primeiro deles é a versatilidade oferecida pelo SSA. Como exposto na seção 2.2, a utilização do SSA é uma das abordagens válidas para a construção de modelos que representam o comportamento dinâmico de sistemas biológicos. Com efeito, este formalismo pode ser utilizado para representar quase a totalidade dos comportamentos celulares, incluindo metabolismo, sinalização celular, regulação gênica, transporte celular e o próprio ciclo celular (WILKINSON, 2018). Segundo, a possibilidade de aplicação de técnicas de otimização e paralelização é uma característica desejável para tornar as simulações mais rápidas e eficientes.

O ponto de partida para a elaboração do método aqui proposto é a suposição de que seja possível a criação de um modelo monolítico (sem subdivisões) capaz de ser simulado com o SSA e que seja capaz de representar, com razoável precisão, uma célula completa. E de fato, da análise dos resultados alcançados pelo projeto E-CELL (vide seção 2.1.2), depreende-se que tal suposição é razoável já que com um modelo monolítico simulado através do uso de equações diferenciais ordinárias, esse projeto foi capaz de reproduzir alguns resultados experimentais. Além disso, como destacado na seção 2.2.3, modelos simulados com EDOs podem também ser facilmente adaptados

para serem simulados com SSA, obtendo-se resultados comparáveis.

Na hipótese de um modelo monolítico simulado por SSA, em cada intervalo de tempo τ, uma determinada reação seria selecionada e processada, com chance determinada pela sua propensão, assim como detalhado na seção 2.2.2. Como mostrado na figura 5, podemos bipartir o modelo monolítico inicial de maneira que reações presentes em um dos modelos resultantes (MI e MII) não estejam no outro. Desta maneira, como o número total de reações do sistema não foi alterado, as propensões de cada uma das reações também não foram. Considerando que ativar uma reação que esteja no modelo MI signifique ativar também o modelo MI, a propensão de ativação de um modelo será equivalente à soma das propensões de todas as reações nele incluídas.

No exemplo mostrado na figura 5, cada nível representa uma bipartição do(s) modelo(s) no nível anterior, e as setas em verde indicam quais modelos resultam da bipartição dos modelos no nível anterior. Após o primeiro seccionamento (segundo nível na figura), a propensão do modelo MI ser ativado seria o equivalente à soma das propensões das reações R1, R2, R3 e R5, enquanto que a propensão do modelo MII seria dada pela soma das propensões das reações R4, R6, R7 e R8. Já após o segundo seccionamento (terceiro nível na figura), a propensão do $MIII$ ser ativado seria a soma das propensões para as reações R1 e R2 somente. No último nível inferior, a propensão de cada um dos modelos equivaleria à propensão de uma única reação, assim como no modelo monolítico (situado no topo da hierarquia representada). Conclui-se então que a aplicação de sucessivas partições aos modelos eventualmente resultaria na obtenção de vários submodelos contendo uma única reação cada, e cuja simulação seria equivalente à simulação de um modelo estocástico monolítico.

Como destacado na seção 2.2, modelos são independentes de seus algoritmos de simulação. Considerando que todas as informações necessárias estejam disponíveis, o algoritmo de simulação de cada um dos submodelos poderia ser alterado sem comprometer significativamente a precisão dos resultados, desde que o algoritmo selecionado seja apropriado para descrever os eventos representados pelo submodelo.

Em suma, o método aqui proposto age coordenando as simulações dos submodelos e intermediando a comunicação ao atualizar do estado do sistema ao término de cada período τ atribuído para a simulação de um dos submodelos. Dessa maneira, torna-se possível integrar modelos de distintas naturezas e que utilizem diferentes algoritmos de simulação de forma a capturar os efeitos das propriedades emergentes do sistema sem deixar de levar em consideração a presença de variações aleatórias existentes. O fluxograma apresentado na figura 6 representa o funcionamento do método proposto.

Entretanto, ao substituir o processamento de reações (que envolvem somente operações de soma e subtração) por processamento de modelos (que envolvem opera-

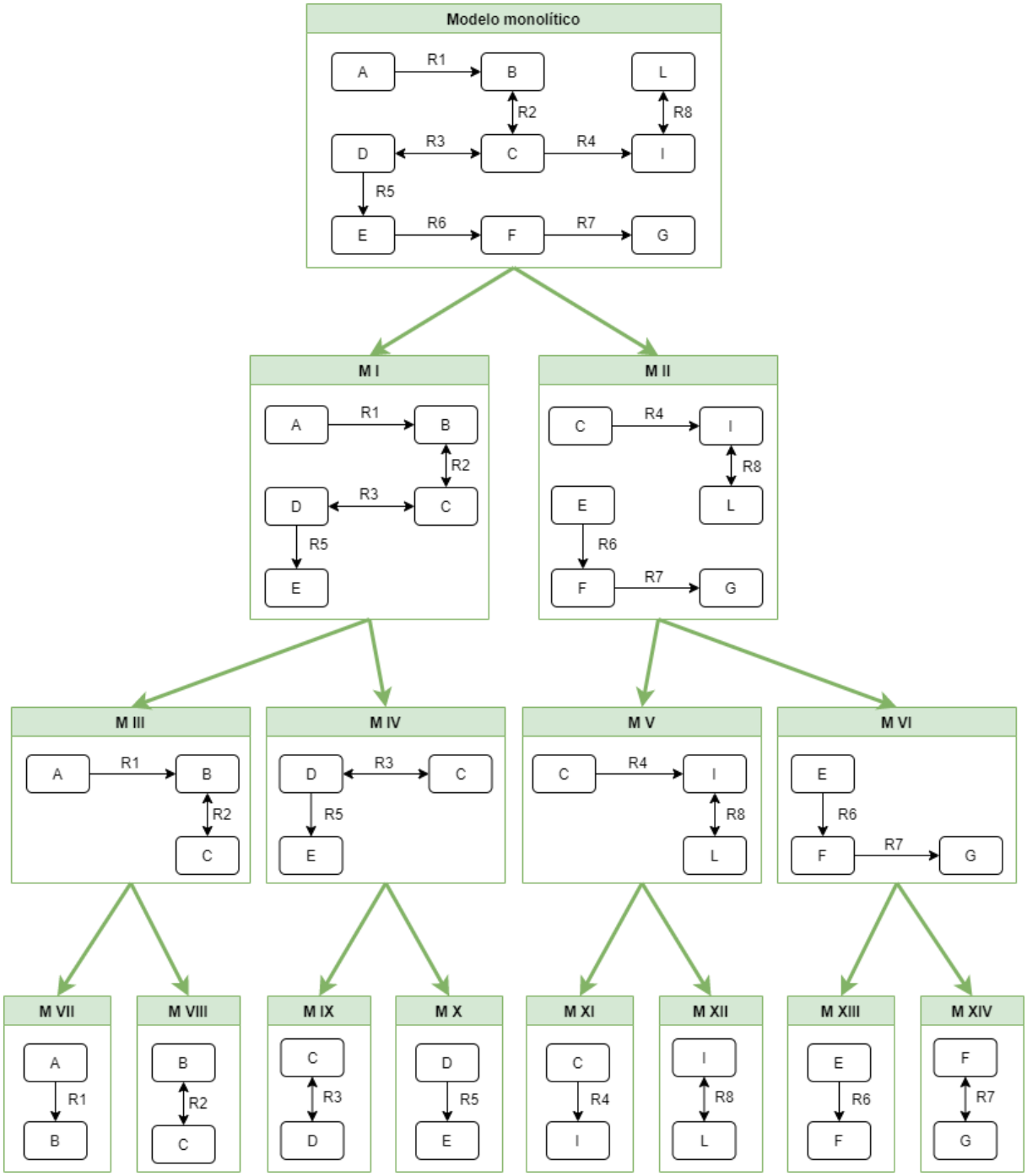


Figura 5 – Múltiplos seccionamentos de um modelo monolítico

ções mais complexas, como integração), a complexidade computacional da simulação integrada torna-se significativamente maior, traduzindo-se em maior tempo necessário para a simulação do modelo. Uma maneira de mitigar este problema seria adotar uma estratégia similar à técnica conhecida como "τ-*leaping*"(GILLESPIE, 2001), que consiste em estabelecer um passo temporal maior, de tal forma que as propensões não sejam significativamente alteradas, tornando o método estocástico mais eficiente do ponto de vista computacional.

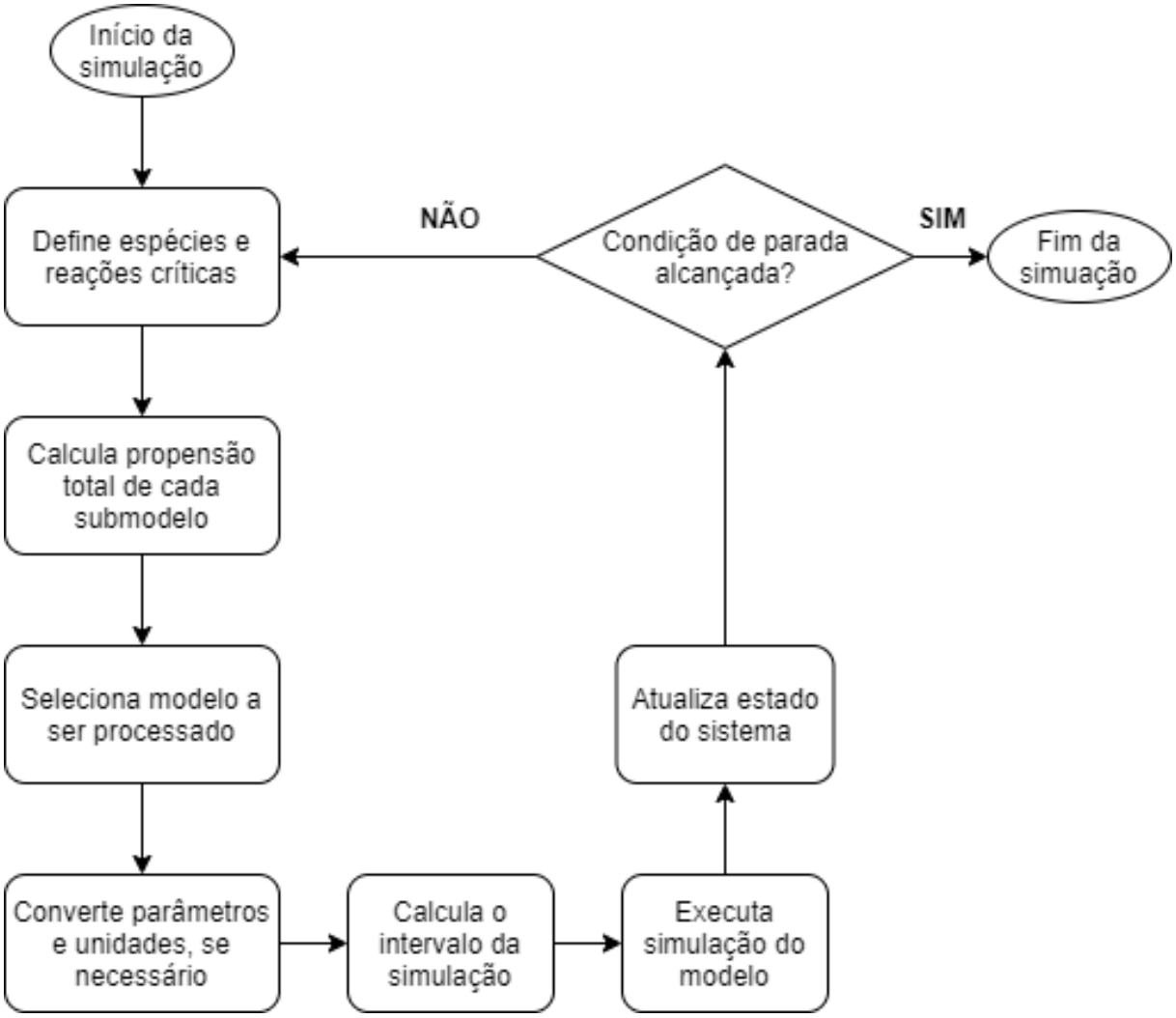


Figura 6 – Fluxograma do método estocástico simulação multi-algorítmica

Para fins de cálculo do passo temporal τ, optou-se pela metodologia de cálculo proposta por CAO et al. (2006), que gera uma aproximação mais precisa, e com menor custo computacional que a proposta original (GILLESPIE, 2001). A validade desta abordagem se deve ao fato de que, para valores não muito altos de τ, a variação nas propensões das reações é baixa, permitindo que várias reações sejam processadas em um único passo temporal. Com efeito, esta aproximação concorda com as premissas assumidas pela método aqui proposto, uma vez que ao simular um modelo inteiro em determinado tempo, várias reações estariam sendo processadas. Destarte o passo temporal pode ser calculado pela equação 20 (CAO et al., 2006), na qual $0 < \epsilon \ll 1$ é o parâmetro de controle de erro, e equivale à variação percentual aceitável da soma das propensões, x_i é a quantidade de moléculas da espécie S_i no tempo atual, os parâmetros $\mu_i(x)$, $\sigma_i^2(x)$, e g_i podem ser calculado pelas equações 21, 22 e 23 (CAO et al., 2006),

respectivamente.

$$\tau = min_{i \in I_{ncr}} \left\{ \frac{max\{\epsilon x_i / g_i, 1\}}{|\mu_i(x)|}, \frac{max\{\epsilon x_i / g_i, 1\}^2}{|\sigma_i^2(x)|} \right\} \tag{20}$$

$$\mu_i(x) = \sum_{j \in J_{ncr}} v_{ij} a_j(x), \forall i \in I_{rs} \tag{21}$$

$$\sigma_i^2(x) = \sum_{j \in J_{ncr}} v_{ij}^2 a_j(x), \forall i \in I_{rs} \tag{22}$$

$$g_i = 1 \tag{23a}$$

$$g_i = 2 \tag{23b}$$

$$g_i = \left(2 + \frac{1}{x_i - 1}\right) \tag{23c}$$

$$g_i = 3 \tag{23d}$$

$$g_i = \frac{3}{2}\left(2 + \frac{1}{x_i - 1}\right) \tag{23e}$$

$$g_i = \left(3 + \frac{1}{x_i - 1}\frac{2}{x_i - 2}\right) \tag{23f}$$

O termo I_{ncr} refere-se à lista de reagentes das equações não-críticas, já o termo J_{ncr} se refere à lista de reações não-críticas. Reações não-críticas são aquelas reações cujos reagentes estão presentes em quantidade suficiente para evitar que a possibilidade de ocorrência de contagem negativa de moléculas após processadas as reações. Os termos v_{ij} e a_j representam a variação da contagem de moléculas da espécies S_i na reação R_j e a propensão da reação R_j, respectivamente. Por fim, a seleção da equação para cálculo de g obedece ao seguinte critério:

- se, dentre as reações nas quais S_i é reagente, a maior ordem de reação for 1 deve-se selecionar a equação 23a
- se, dentre as reações nas quais S_i é reagente, a maior ordem de reação for 2 deve-se selecionar a equação 23b
- se, dentre as reações nas quais S_i é reagente, a maior ordem de reação for 2 e pelo menos uma reação de ordem 2 requeira duas moléculas de S_i deve-se selecionar a equação 23c
- se, dentre as reações nas quais S_i é reagente, a maior ordem de reação for 3 deve-se selecionar a equação 23d

- se, dentre as reações nas quais S_i é reagente, a maior ordem de reação for 3 e pelo menos uma reação de ordem 2 requeira duas moléculas de S_i deve-se selecionar a equação 23e
- se, dentre as reações nas quais S_i é reagente, a maior ordem de reação for 3 e pelo menos uma reação de ordem 2 requeira três moléculas de S_i deve-se selecionar a equação 23f

3.2 Da validação do método proposto

A validação do método proposto foi feita em duas etapas: a primeira consistiu em comprovar que as hipóteses consideradas em sua elaboração são verdadeiras e a segunda etapa consistiu em testes para verificar o desempenho da técnica quando comparada à simulação de um modelo estocástico monolítico que represente o mesmo sistema.

Na primeira etapa utilizou-se um modelo simples que representa uma reação enzimática em duas etapas, envolvendo poucas reações e espécies, o que permite um maior controle do experimento e clareza nos resultados. Este modelo, introduzido inicialmente na seção 2.2.1, pode ser descrito conforme mostrado na equação 1, reproduzida abaixo. O mecanismo de funcionamento do modelo considera que num sistema espacialmente homogêneo têm-se inicialmente uma determinada quantidade de enzimas (E) e substratos (S) que, ao chocar-se, formam o complexo ES, que pode dissociar-se retornando o substrato e a enzima (caso a reação não proceda), ou converter o substrato em produto (P), que é então liberado da enzima.

$$\mathrm{E} + \mathrm{S} \underset{\mathrm{k\_2}}{\overset{\mathrm{k\_1}}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{\mathrm{k3}} \mathrm{E} + \mathrm{P}$$

A seleção deste modelo confere maior simplicidade e agilidade à aplicação do método e, como seus resultados já são conhecidos, possibilita uma análise mais precisa. Além disso, por apresentar duas reações (uma reversível e outra não), sua bipartição se torna mais natural e lógica. Através da utilização deste modelo, buscou-se avaliar a precisão dos resultados produzidos pela aplicação do método proposto, bem como os efeitos da utilização de distintos algoritmos de simulação para cada um dos modelos utilizados.

Para realização dos testes, duas versões do modelo foram criadas: uma monolítica (sem subdivisões), e uma bipartida. Cada uma das versões foi submetida a dois tipos distintos de simulação. A versão monolítica foi simulada primeiramente de forma determinística (intitulada simulação Det) e em seguida de maneira estocástica

(intitulada simulação Est). Já a versão bipartida foi simulada inicialmente utilizando-se o algoritmo determinístico para simular ambos os submodelos (intitulada simulação Bip) e, posteriormente, com cada um dos submodelos utilizando algoritmos distintos (intitulada simulação Hib). Na figura 7 é mostrado como foi realizada a bipartição do modelo, e as simulações utilizadas. Os dados iniciais utilizados para a simulação do modelo foram os mesmos apresentados nas tabelas 1 e 2.

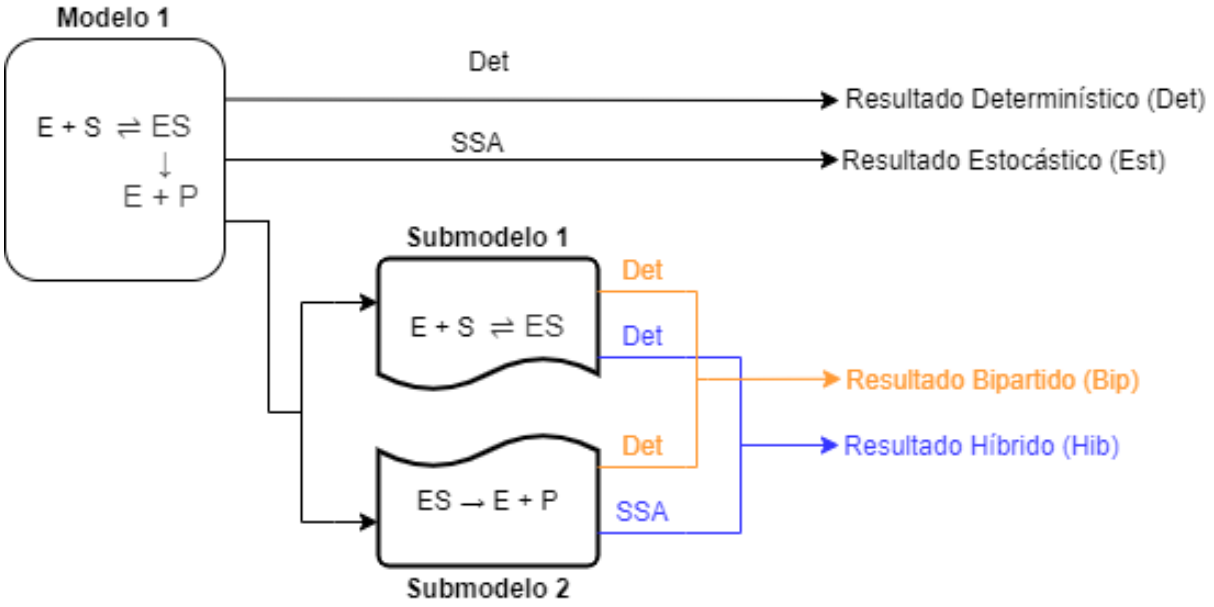


Figura 7 – Esquema dos modelos e simulações utilizadas para os testes da primeira etapa de validação

A simulação do modelo monolítico pelo algoritmo determinístico foi executada uma única vez, haja vista que esta abordagem não apresenta variabilidade nos resultados obtidos. Já para cada uma das demais abordagens, executou-se 100 repetições da simulação e registrou-se os resultados obtidos. Para cada um dos 301 resultados obtidos, realizou-se uma interpolação linear simples para obtenção dos valores aproximados da quantidade de moléculas nos tempos 2, 5, 8, 11, 14, 17, 20, 30 e 50 segundos, através da utilização dos valores obtidos para os tempos imediatamente inferior e imediatamente superior a cada um dos pontos de interesse. Os 8 primeiros pontos foram selecionados por estarem em regiões do gráfico onde ocorrem as maiores variações entre os resultados dos modelos, enquanto que o último ponto (50s) foi selecionado por situar-se no ponto em que a maioria das simulações já alcançou o equilíbrio.

Para cada um dos pontos selecionados, calculou-se o valor médio do número de moléculas de cada espécie presentes no sistema, os valores obtidos para os modelos bipartidos foram comparados aos resultados provenientes da simulação determinística e às médias dos resultados das simulações estocásticas do modelo monolítico. O estudo

das divergências observadas permitiu avaliar a precisão do modelo, respondendo-se os questionamentos postos.

Na segunda etapa de teste, buscou-se verificar o impacto da aplicação do método no tempo de simulação total. Para isso, utilizou-se como base um modelo de maior porte que o anterior, com 11 espécies e 13 reações (3 das quais são reversíveis), representando o funcionamento do Operon Lac, que faz parte do metabolismo da lactose em bactérias como a *Escherichia coli*. Além da maior complexidade do novo modelo em relação ao anterior, sua seleção se deu por outros dois fatores: a possibilidade do modelo ser seccionado ou incrementado pela adição de um novo modelo e pelo modelo não apresentar um estado de equilíbrio químico, o que permite que sua simulação seja prolongada por tempos indefinidos. O detalhamento do modelo e seu princípio de funcionamento encontram-se no apêndice A.

Para avaliação do desempenho do método proposto, utilizou-se dois modelos: o primeiro monolítico, representando o Operon Lac, e um segundo modelo representando a injeção linear de lactose no sistema a uma taxa de 1 molécula a cada 10 segundos (c = 0.1). Duas simulações foram então elaboradas utilizando estes modelo, de acordo com o representado na figura 8. A primeira delas (linha preta, denominada simulação monolítica) trata-se da simulação do modelo monolítico utilizando o algoritmo de simulação estocástica (SSA), enquanto a segunda simulação (linha vermelha, denominada simulação híbrida) processa o modelo gerado pela combinação dos outros dois. Na simulação híbrida, o modelo representando o Operon Lac foi simulado utilizando SSA, enquanto o modelo de injeção de Lactose no sistema foi simulado utilizando equações diferenciais ordinárias (determinístico).

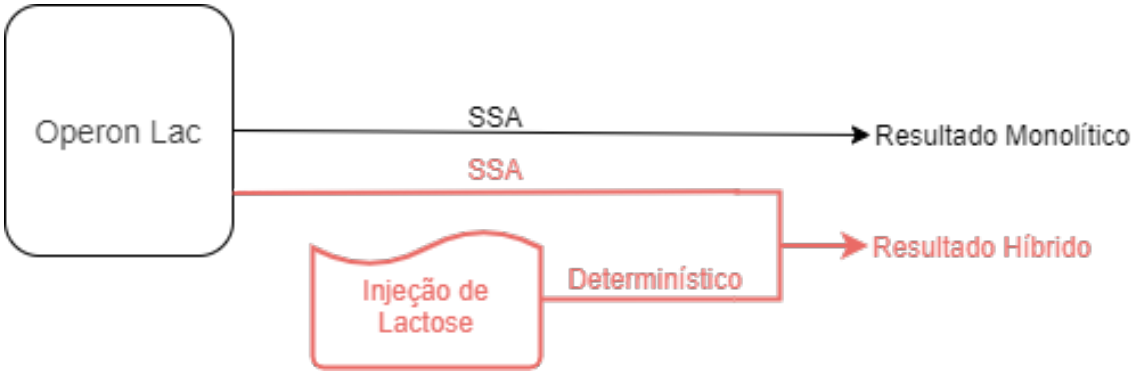


Figura 8 – Esquema dos modelos e simulações utilizadas para os testes da segunda etapa de validação

Em seguida criou-se cinco variantes de cada uma das simulações, representando o comportamento do sistema por 120, 600, 1200, 6000 e 12000 segundos. Cada variante das simulações foi executada 50 vezes e o tempo necessário para completar cada uma

delas foi registrado. Com os resultados em mãos, comparou-se então as médias, valores mínimos e máximos necessários para executar as simulações em cada caso.

Por fim, como verificação adicional da correção do funcionamento do método proposto, analisou-se os comportamentos médios obtidos com a simulação das duas variantes do sistema modelado (monolítico e híbrido).

3.3 Da ferramenta computacional desenvolvida: Cell Lab

Um dos aspectos mais importantes para a difusão de novas técnicas e conhecimentos é a possibilidade de sua aplicação prática a estudos similares e a reprodutibilidade dos resultados obtidos. Para alcançar este objetivo, é essencial que a técnica seja acessível aos cientistas que dela se beneficiarão. No caso específico da metodologia aqui desenvolvida (detalhada na seção 3.1), sua aplicação requer que o usuário possua conhecimentos avançados de programação, além de sólida compreensão matemática dos princípios de modelagem de sistemas biológicos. Tais características constituem uma barreira para a reaplicação da metodologia por cientistas que não possuam tais conhecimentos, mas que poderiam se beneficiar com a construção de modelos híbridos em suas pesquisas.

Desta maneira, considerando a inexistência de softwares com capacidade de realizar simulações que integrem distintos modelos e a importância de facilitar a aplicação do método proposto e torná-lo mais acessível a cientistas que dele se beneficiariam, foi desenvolvida uma ferramenta computacional com interface gráfica amigável para automatizar o processo de criação e simulação dos modelos seguindo a metodologia anteriormente desenvolvida. Denominado de Cell Lab, o software desenvolvido no presente trabalho tem dentre suas principais funções:

- Auxiliar na aquisição e organização de dados para construção de modelos
- Assistir a criação de modelos computacionais de sistemas biológicos sem a necessidade de conhecimentos de programação
- Permitir simular os modelos construídos, seguindo a metodologia aqui desenvolvida
- Visualizar os resultados e verificar o comportamento dos sistemas simulados

Para se alcançar o resultado desejado, e permitir que o Cell Lab seja um software de fácil utilização, foi desenvolvida uma interface gráfica leve e amigável. A interface da ferramenta foi projetada para guiar o usuário pelas etapas necessárias para criação e simulação dos modelos. Por isso, o Cell Lab apresenta interface gráfica dividida em três módulos principais:

- Módulo de construção de modelos
- Módulo de gerenciamento de dados e anotações
- Módulo de simulação e visualização de resultados

Com finalidade de exemplificar as funcionalidades do framework, construiu-se um modelo determinístico simples das primeiras etapas da glicólise. Os dados para construção do modelo retratado nas figuras que seguem foram extraídos de (KLIPP et al., 2016, p. 18-19).

3.3.1 Tecnologias e convenções utilizadas

O Cell Lab foi desenvolvido utilizando a linguagem de programação Python, destacadamente pela simplicidade e disponibilidade de recursos que esta linguagem oferece. Além disso, a versatilidade e possui grande popularidade da linguagem selecionada permite estruturar o código fonte de maneira que permita futuras modificações de maneira simplificada. Este conjunto de características é bastante favorável para permitir que usuários e membros da comunidade de modelagem computacional de sistemas biológicos possam compreender e colaborar futuramente no melhoramento do framework.

A linguagem Python conta também com ferramentas e pacotes que fornecem funcionalidades bastante convenientes, como as bibliotecas PyQt, Matplotlib, Numpy, Scipy e Pandas. A biblioteca PyQt permite a criação de interfaces gráficas leves, funcionais e que suportam todas as funcionalidades necessárias ao Cell Lab. Por se tratar de um framework para criação de modelos matemáticos-computacionais, o uso das bibliotecas Numpy e Scipy é bastante conveniente para permitir tanto a criação quanto o processamento dos modelos. Além disso, as bibliotecas Pandas e Matplotlib permitem oferecer ao usuário alternativas de organização e visualização de dados, respectivamente.

Por fim, a utilização de *open source*, como a libchebipy e a pubchempy, permite o acréscimo de funções de grande conveniência ao framework, sem a necessidade de criação de código próprio adicional para as funções secundárias ou de suporte, permitindo que o processo de desenvolvimento seja acelerado.

Para facilitar a utilização do Cell Lab por parte da comunidade de modelagem computacional, os modelos podem ser construídos graficamente através da criação de diagramas que aderem ao padrão Systems Biology Graphical Notation (SBGN), já bastante estabelecido entre os membros da comunidade, e que torna mais simples o processo de criação e compartilhamento dos diagramas através de publicações.

4 Resultados

A aplicação dos testes apresentados na seção anterior produziu resultados que permitiram avaliar não somente a validade da metodologia proposta, mas também sua eficiência e viabilidade de aplicação. Nesta seção estão detalhados cada um desses resultados, acompanhados das devidas interpretações e análises do significado de cada um deles. Cabe aqui ressaltar que, devido ao grande volume de dados gerados pelos testes, as tabelas e figuras da presente seção trazem os valores médios das métricas mais importantes às avaliações.

4.1 Validação do método proposto

4.1.1 Teste 1 - Verificação da correção das hipóteses assumidas

Após a realização de uma simulação determinística com o modelo monolítico e 100 repetições de cada uma das demais simulações, os resultados obtidos foram convertidos e uniformizados para refletirem a quantidade de moléculas no sistema através do tempo. Estes foram então comparados em 9 pontos temporais selecionados para análise. Na figura 9, estão apresentados os resultados médios obtidos através de cada tipo de simulação para cada uma das espécies.

Os valores utilizados para a elaboração dos gráficos estão discriminados na tabela 3 onde, para cada uma das espécies, são apresentadas as contagens médias e arredondadas das quantidades de moléculas para cada uma das espécies do sistema para os modelos determinístico monolítico (coluna D), estocástico monolítico (coluna E), determinístico bipartido (coluna B) e híbrido bipartido (coluna H).

Tabela 3 – Contagem média de moléculas para 100 simulações para o teste 1

	Enzima (nE)				Substrato (nS)				Complexo (nES)				Produto (nP)			
T	**D**	**E**	**B**	**H**	**D**	**E**	**B**	**H**	**D**	**E**	**B**	**H**	**D**	**E**	**B**	**H**
2	56	54	57	50	229	223	229	219	64	66	63	70	8	12	9	12
5	35	37	35	40	185	178	185	197	85	83	85	80	31	40	31	24
8	33	35	34	37	157	146	158	165	87	85	86	83	57	70	57	53
11	35	39	35	38	133	129	132	149	85	81	85	82	83	91	84	70
14	38	44	40	44	111	111	112	122	82	76	80	76	108	114	109	103
17	42	43	41	47	91	88	93	100	78	77	79	73	132	136	129	128
20	47	47	46	60	73	68	72	76	73	73	74	60	155	160	155	165
30	67	70	66	71	28	26	28	28	54	50	55	49	219	225	218	224
50	105	102	105	108	2	1	3	3	15	18	15	12	284	282	283	286

Tomando-se como referência os valores obtidos pelo modelo determinístico, e comprando-os com os valores encontrados para as demais simulações, foi possível realizar uma estimativa da precisão dos resultados obtidos com a aplicação do método. Os gráficos apresentados na figura 10 mostram as diferenças absolutas obtidas, em número de moléculas, para cada um dos tipos de simulação utilizada.

Partindo da análise das figuras 9 e 10 e dos dados numéricos apresentados na tabela 3, é possível concluir que a precisão dos resultados é aceitável, o que se traduz na confirmação da adequação da metodologia para cada um dos casos no qual ela foi aplicada.

Comparando-se os resultados da aplicação do método proposto para simular um modelo bipartido, quando mantido algoritmo de simulação original em ambos os submodelos, com os resultados do modelo original simulado deterministicamente, percebe-se o impacto do processamento sequencial e alternado nos resultados finais é baixo. Além disso, pela análise das discrepâncias ocorridas em cada um dos modelos, verificou-se que a utilização desta abordagem resulta em um modelo aproximadamente determinístico, cuja variabilidade dos resultados é muito pequena.

Já quando são analisados os resultados do modelo híbrido, verificou-se que o impacto da aplicação da metodologia é maior. Mesmo neste caso, a variabilidade apresentada é comparável à ocorrida com a simulação de um modelo monolítico estocástico, indicando que são variações aceitáveis. Dois fatores apresentaram contribuições importantes para a existência das discrepâncias observadas: a presença de um submodelo estocástico e as conversões de unidade (de concentração para moléculas e vice-versa) durante a simulação. Inspecionando os resultados nos modelos que apresentam maior discrepância média, observou-se que o efeito das conversões e aproximações constantes pode ocasionar pequenos erros na contagem total de moléculas do sistema, os quais podem se multiplicar e gerar impactos nos resultados das simulações. Entretanto, esta ocorrência foi observada somente em 4 das 100 simulações realizadas.

Desse modo, percebeu-se que a aplicação do método para coordenar simulações de múltiplos modelos contendo espécies em comum é apropriada, sendo possível verificar as interferências de um modelo sobre o outro, como teorizado inicialmente. Deve-se ressaltar que, nos casos em que os tipos dos submodelos exijam conversões constantes de unidade, a possibilidade de erros torna-se mais elevada. Entretanto, mesmo nos casos em que os erros estavam presentes, os resultados mantiveram-se comparáveis aos observados nas simulações através do uso do algoritmo de simulação estocástica.

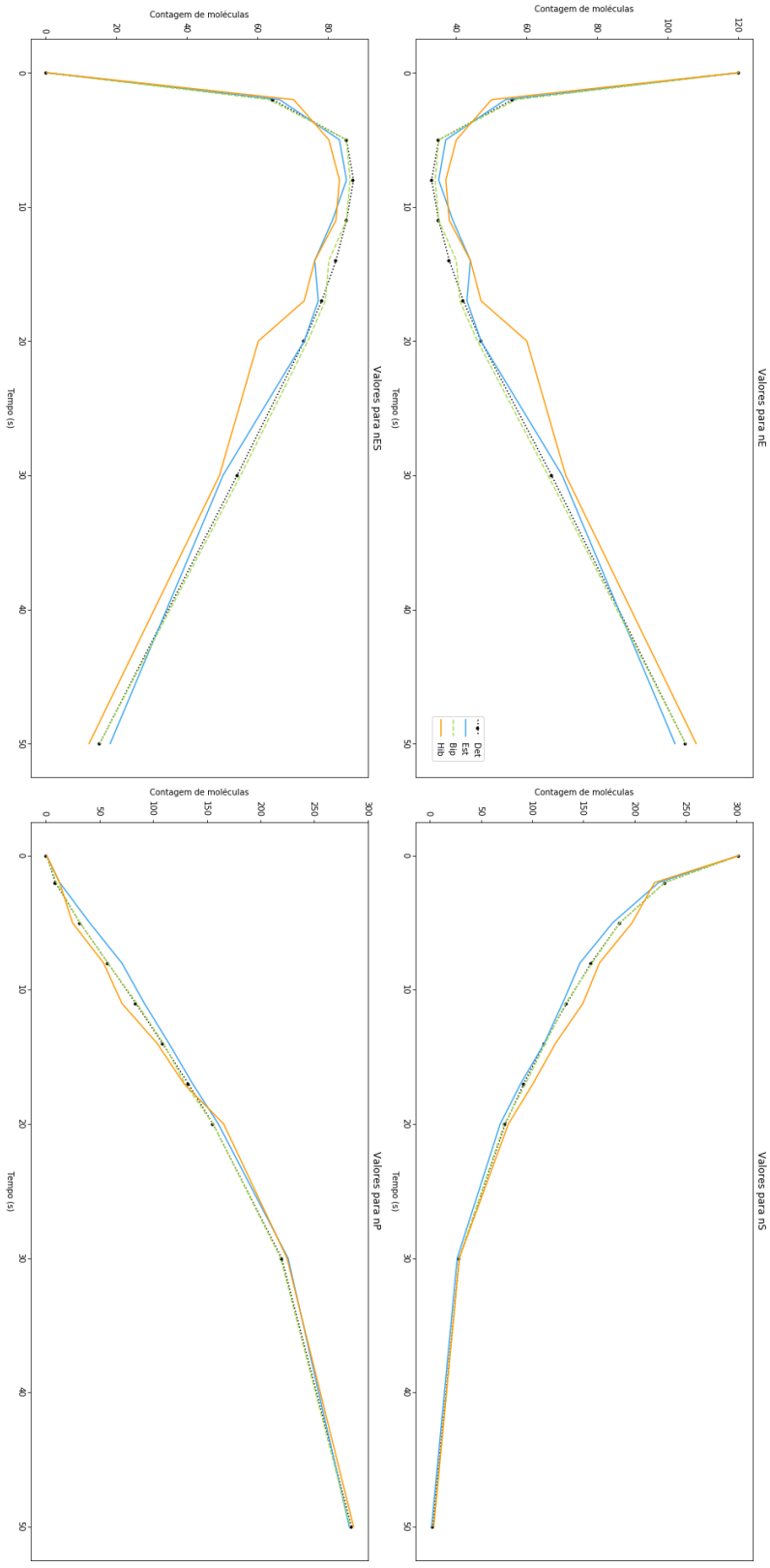


Figura 9 – Comparação dos resultados alcançados com as 4 abordagens utilizadas

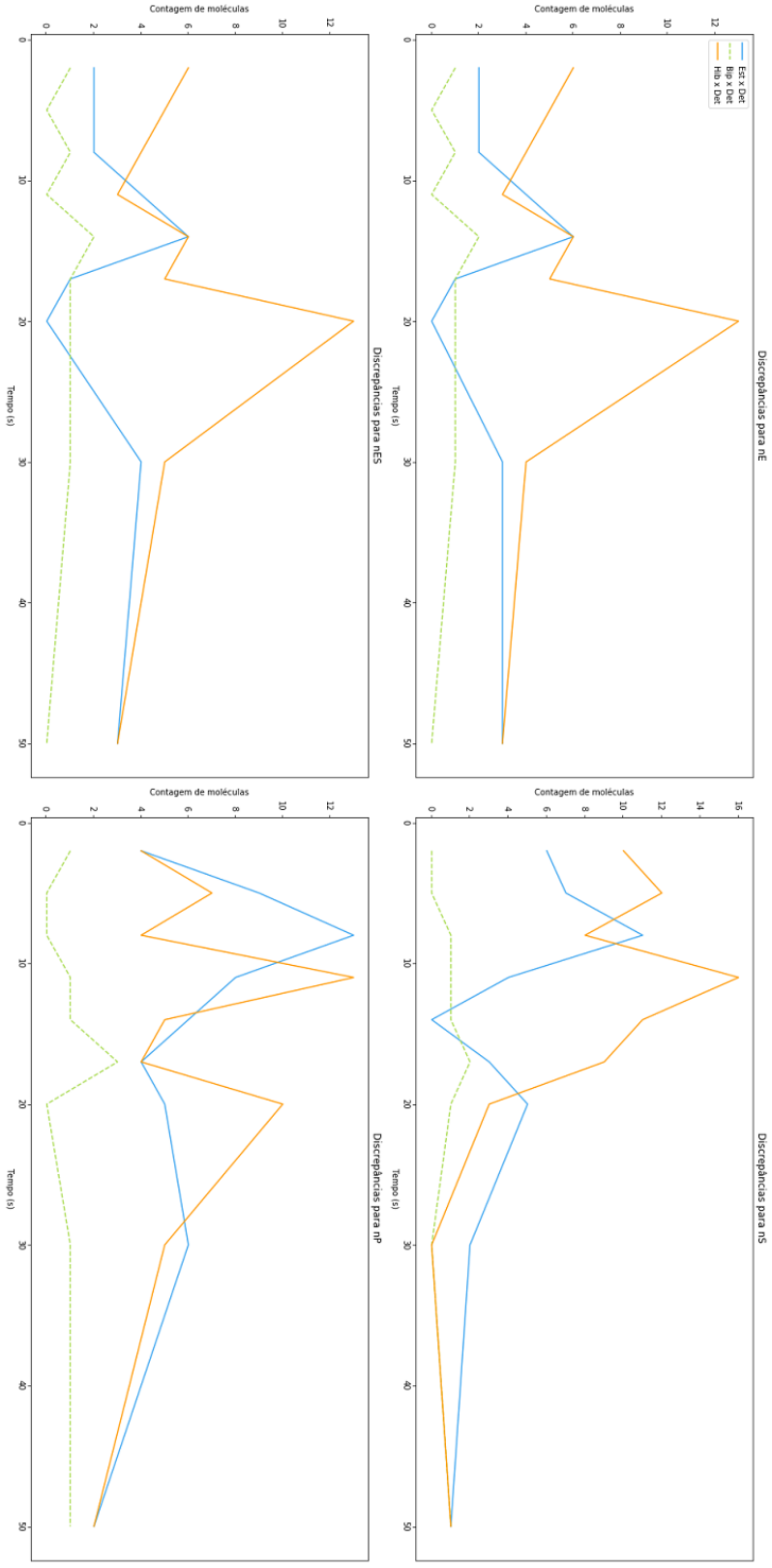


Figura 10 – Diferenças médias entre as abordagens e a simulação determinística

4.1.2 Teste 2 - Eficiência computacional do método proposto

Concluídas as simulações da etapa 2, as durações auferidas foram resumidas na tabela 5 e encontram-se representados graficamente nas figuras 11 e 12. A primeira delas mostra os resultados obtidos quando não há injeção de lactose no sistema. Verifica-se que nesta situação, a produção de inibidor continua (linha laranja), enquanto os inibidores existentes se ligam às moléculas de lactose (linha verde), resultando no inibidor ligado (linha azul). Dessa maneira, eventualmente algumas moléculas da enzima Z (linha roxa) são produzidas, e metabolizam as moléculas livres de lactose.

Tabela 4 – Durações de processamento auferidas durante o teste 2

	Monolítico				Híbrido			
T simulado	Min(s)	Méd(s)	Máx(s)	P	Min(s)	Méd(s)	Máx(s)	P
120s	0.02	0.05	0.08	354	0.89	1.00	1.23	865
600s	0.10	0.20	0.32	2164	4.99	6.08	10.42	4304
1200s	0.21	0.43	0.58	3965	10.21	13.42	18.47	8974
6000s	1.86	2.50	3.09	22031	68.61	83.80	103.12	43538
12000s	4.27	5.74	8.79	39119	147.90	180.99	221.60	84055

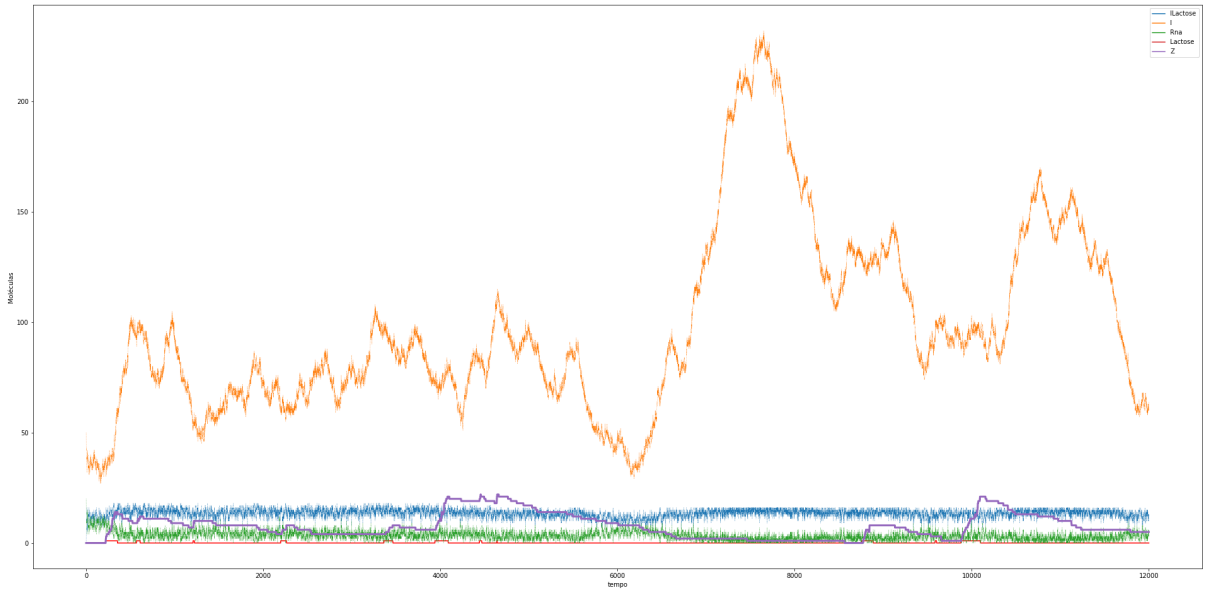


Figura 11 – Resultado da simulação do modelo original do Operon Lac

A figura 12 mostra o efeito da adição do modelo de injeção constante de lactose a uma taxa de 0.1 molécula/s. A quantidade de inibidor ligado às moléculas de lactose aumenta, resultando em maior produção da enzima Z, que tem como consequência a redução da quantidade de lactose livre no sistema.

Comprovado o comportamento esperado do sistema sob as novas condições, verificou-se também os tempos necessários para completar cada uma das simulações.

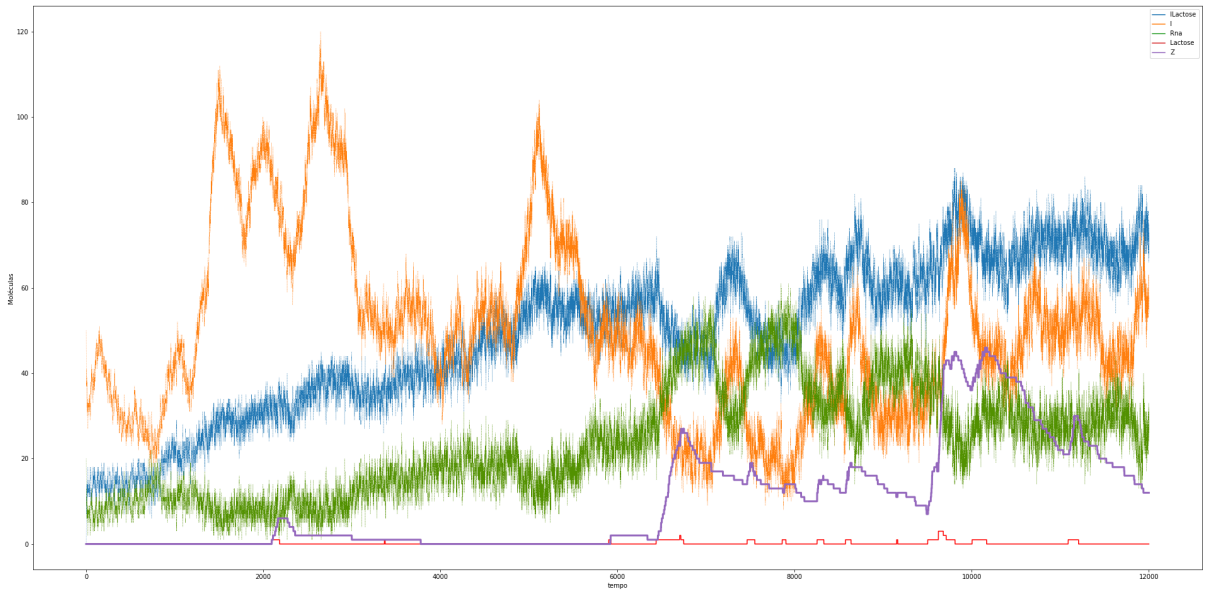


Figura 12 – Resultado da simulação integrada: Operon Lac + injeção de lactose

A tabela 4 traz os valores do menor tempo (min), do maior tempo (máx) e da média dos tempos (méd) das 50 simulações realizadas para cada tempo final selecionado. Os valores da tabela foram então utilizados para elaboração do gráfico representado na figura 13.

Da análise dos resultados obtidos, depreende-se que a aplicação da metodologia para realização da simulação multi-algorítmica resulta num aumento significativo do tempo necessário para a conclusão da simulação. Isso é observado mesmo quando o modelo adicionado representa uma simples reação.

Observando o comportamento descrito pelo gráfico, nota-se que o aumento do custo computacional segue o mesmo padrão em ambos os modelos, indicando que a aplicação da metodologia não resulta em mudança na forma como o número de passos afeta o custo computacional. Além disso, a forma assumida pela curva na escala logarítmica-linear mostra que o custo computacional escala de maneira aproximadamente linear com o aumento do número de passos.

Através dos dados apresentados na tabela 4 é possível inferir qual é a natureza da causa da queda considerável na eficiência computacional do método proposto quando comparado ao modelo monolítico. Dividindo-se o número médio de passos (P) executados por cada um dos modelos, constata-se que a simulação do modelo híbrido requer aproximadamente o dobro de passos que a simulação do modelo monolítico. Entretanto, dividindo-se o tempo médio necessário para a conclusão de cada simulação, percebe-se que, em média, o modelo híbrido requer 29,54 vezes o tempo que o modelo monolítico leva para completar a mesma simulação. Isso sugere que os fatores responsáveis por

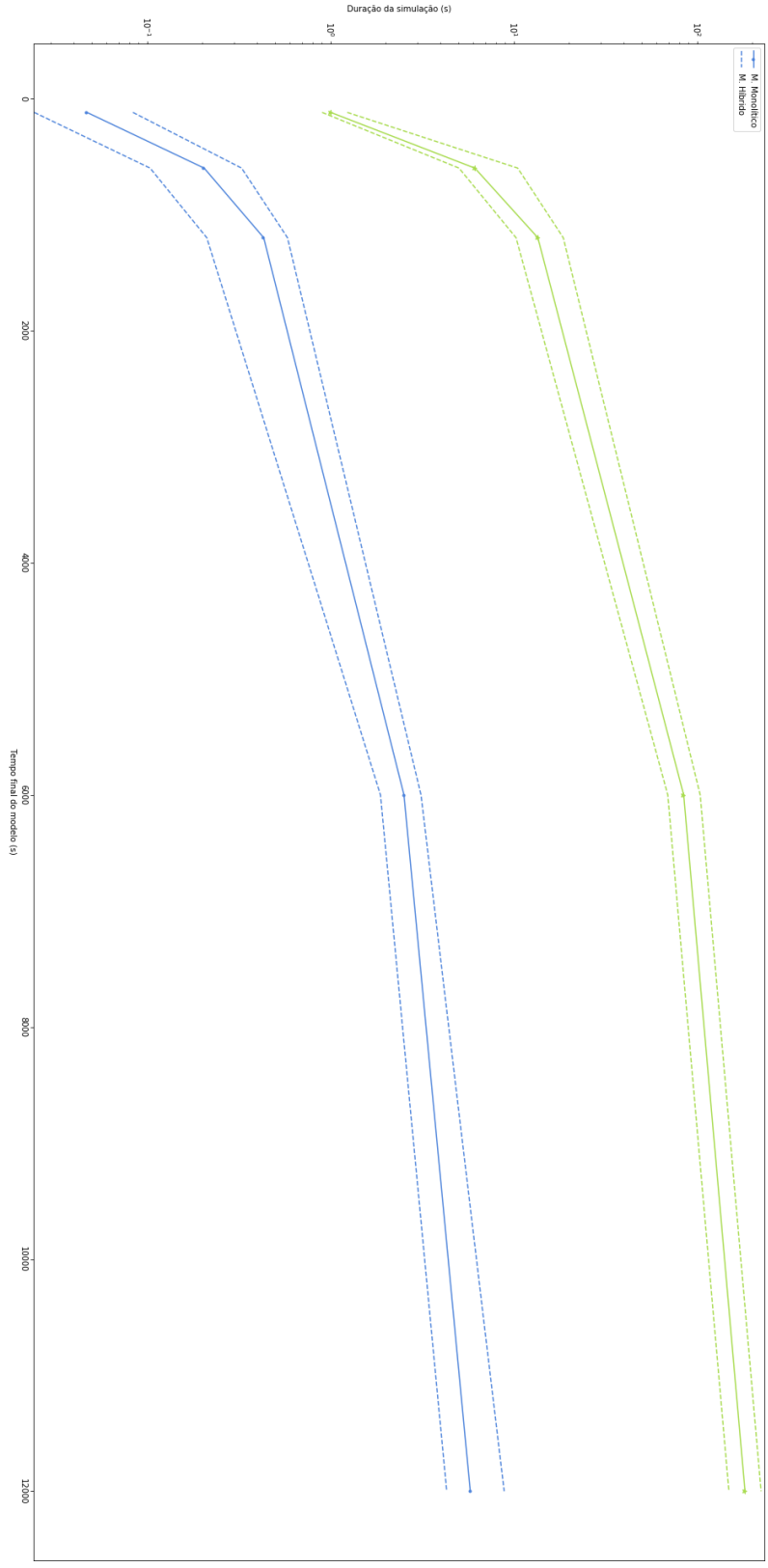


Figura 13 – Gráfico log-lin da evolução da duração das simulações

parte significativa da perda de eficiência são os cálculos adicionais realizados para coordenação das simulações dos modelos.

Com isso, conclui-se que o aumento da eficiência computacional do método estocástico de simulação multi-algorítmica pode ser alcançado através da otimização dos cálculos e conversões necessários para a coordenação dos modelos durante a simulação integrada.

4.2 Funcionamento do Cell Lab e seus módulos

Concebido como uma alternativa tornar a aplicação da metodologia de integração mais simples e expressa, o Cell Lab demonstrou-se efetivo no cumprimento de suas funções. Nesta seção destacamos as principais funcionalidades dos software, elencando, para cada módulo, as principais características e ações ao alcance do usuário.

É importante destacar que os módulos desenvolvidos, apesar de complementares, não são estritamente sequenciais, sendo permitido ao usuário ter a liberdade de alternar entre eles durante a construção dos modelos, adequando o fluxo de trabalho às suas necessidades e preferências. Entretanto, dada a natureza complementar de cada um dos módulos, algumas funcionalidades só estarão ativas após determinadas condições sejam atendidas. Por exemplo, o usuário não poderá acessar o módulo de visualização de resultados até que um modelo válido tenha sido simulado.

4.2.1 Módulo de construção de modelos

O primeiro módulo com o qual o usuário tem contato é com o módulo de construção de modelos, que contém funcionalidades projetadas para possibilitar a construção gráfica de modelos computacionais de sistemas biológicos em escala molecular. Nesta etapa da modelagem, o usuário tem acesso a funções básicas e janelas auxiliares, que permitem definir em detalhes todos as espécies químicas presentes no sistema bem como as interações entre elas, na forma de reações químicas.

Este é o módulo principal do Cell Lab, e é nele onde o usuário definirá os modelos, as espécies que o compõem, e as reações que envolvem as espécies. A figura 14 mostra detalhes da interface gráfica deste módulo.

O módulo é constituído por três elementos principais: a área de construção gráfica de modelos, a barra de ferramentas e a tabela-resumo do modelo. Todos esses elementos funcionam em conjunto, para permitir que o usuário possa criar o modelo, e conferir os dados entrados ou importados à medida que o modelo é construído. Na barra de ferramentas, o usuário possui acesso às seguintes funções:

- Adicionar novo modelo
- Selecionar modelo a ser visualizado / editado
- Deletar modelo
- Adicionar espécie
- Adicionar reação
- Deletar espécie ou reação selecionada
- Editar espécie ou reação selecionada
- Executar simulação do modelo

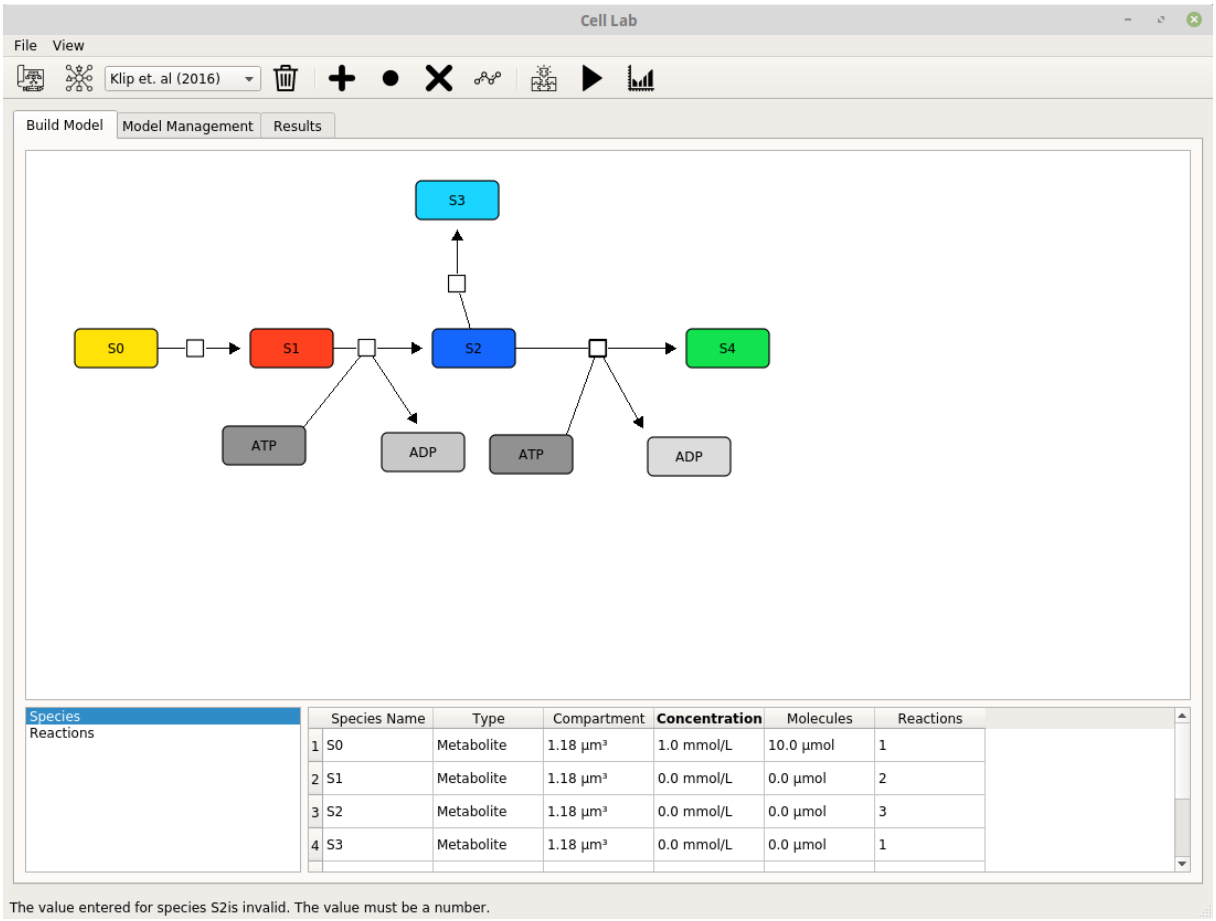


Figura 14 – Interface gráfica do módulo de construção de modelos

Ao adicionar um novo modelo, o usuário deverá especificar um nome, o algoritmo de simulação que deseja usar para o modelo e terá a opção de adicionar anotações ou observações iniciais sobre o modelo. Todas essas ações poderão ser executadas através da janela de criação de novo modelo, que é apresentada ao usuário ao clicar na ferramenta: "Adicionar novo modelo".

Criado um modelo em branco, o usuário poderá proceder à adição de espécies químicas que compõem o sistema. Assim como no caso dos modelos, cada espécie deverá receber um nome único, que poderá ser ou não o mesmo que espécies presentes em outros modelos. No momento da inserção de uma nova espécie no sistema, o usuário pode optar por importar diretamente os dados da espécie de um dos dois bancos de dados suportados pelo Cell Lab: PubChem ou ChEBI. Caso essa seja a opção do usuário, a nova espécie definida contará com informações importadas diretamente de uma das plataformas, tais como identificador IUPAC, fórmula química (versão SMILES), estrutura molecular e anotações.

Além de espécies, é possível também inserir no sistema pontos de importação e exportação, permitindo indicar taxas de produção ou consumo de determinadas espécies. Dessa forma, torna-se possível simplificar alguns fenômenos, como nos casos em que a mesma espécie participa de diversos processos celulares, a exemplo das moléculas carregadoras de energia.

Por fim, uma vez definidas espécies ou pontos de importação e exportação de espécies do sistema, é possível o usuário especificar reações químicas, que transformam os componentes do sistema. Uma reação válida no Cell Lab possui reagentes (ou ponto de importação), produtos (ou ponto de exportação), um identificador único e um coeficiente de taxa de reação (k).

A notação gráfica utilizada pelo Cell Lab segue o padrão SGBN, que é um dos padrões mais utilizados na área. A qualquer momento durante o processo de construção dos modelo o usuário poderá mover os blocos representando reações ou espécies, adicionar novos blocos, deletá-los ou editar as informações neles contidas. Por conveniência, também é possível alterar cor das formas ou linhas e espessura das linhas, permitindo ao usuário destacar determinadas reações ou entidades.

Na parte inferior da tela principal, encontra-se a tabela-resumo do modelo, onde o usuário poderá conferir as informações entradas para espécies ou reações, bem como checar as equações do sistema ou a matriz estequiométrica que representa aquele sistema. O usuário também pode editar as informações contida nas tabelas de reações ou espécie.

Outras funções acessíveis pela tela principal do Cell Lab incluem: Salvar modelos, carregar modelos, salvar figura do modelo, salvar figura da tabela estequiométrica ou alternar entre os modelos.

4.2.2 Módulo de gerenciamento de dados e anotações

O módulo de gerenciamento de dados e anotações permite ao usuário verificar de maneira detalhada todos os dados e anotações utilizados para a construção do

modelo. A conveniência de visualizar todos os dados organizados em um módulo foi pensada como uma forma da facilitar a conferência do modelo e ajuste de parâmetros quando necessário. A figura 15 mostra detalhes da interface gráfica deste módulo, mais especificamente na visualização de espécies.

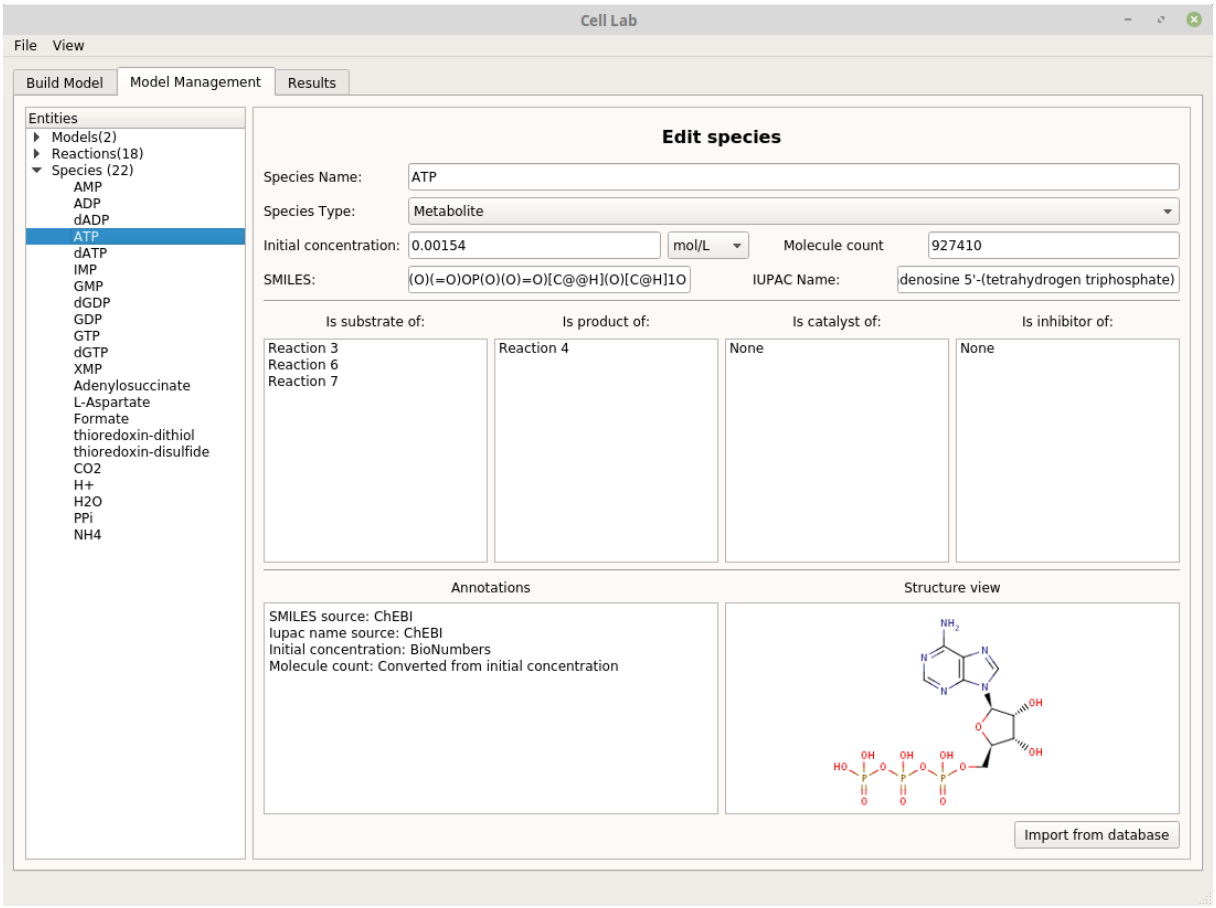


Figura 15 – Módulo de gerenciamento de dados e anotações

Além de poder editar manualmente todas as informações referentes à entidade visualizada (espécies, reações ou modelos), o usuário também tem acesso à ferramenta de importação de dados, a qual oferece integração com dois dos principais bancos de dados de informações sobre espécies químicas: ChEBI e PubChem. Na figura 16, é possível verificar as principais funcionalidades oferecidas pela ferramenta de importação de dados. Dentre os dados importados estão:

- Nome formal da espécie química segundo IUPAC
- Representação SMILES da espécie química
- Peso molecular
- Fórmula molecular

- Estrutura 2D da espécie

A possibilidade do usuário acessar este módulo durante ou após a criação do modelo permite que anotações sejam adicionadas à medida que o modelo é criado, ou que dados sobre as espécies químicas sejam importados do banco de dados somente após a adição de todas as espécies químicas, a critério do usuário.

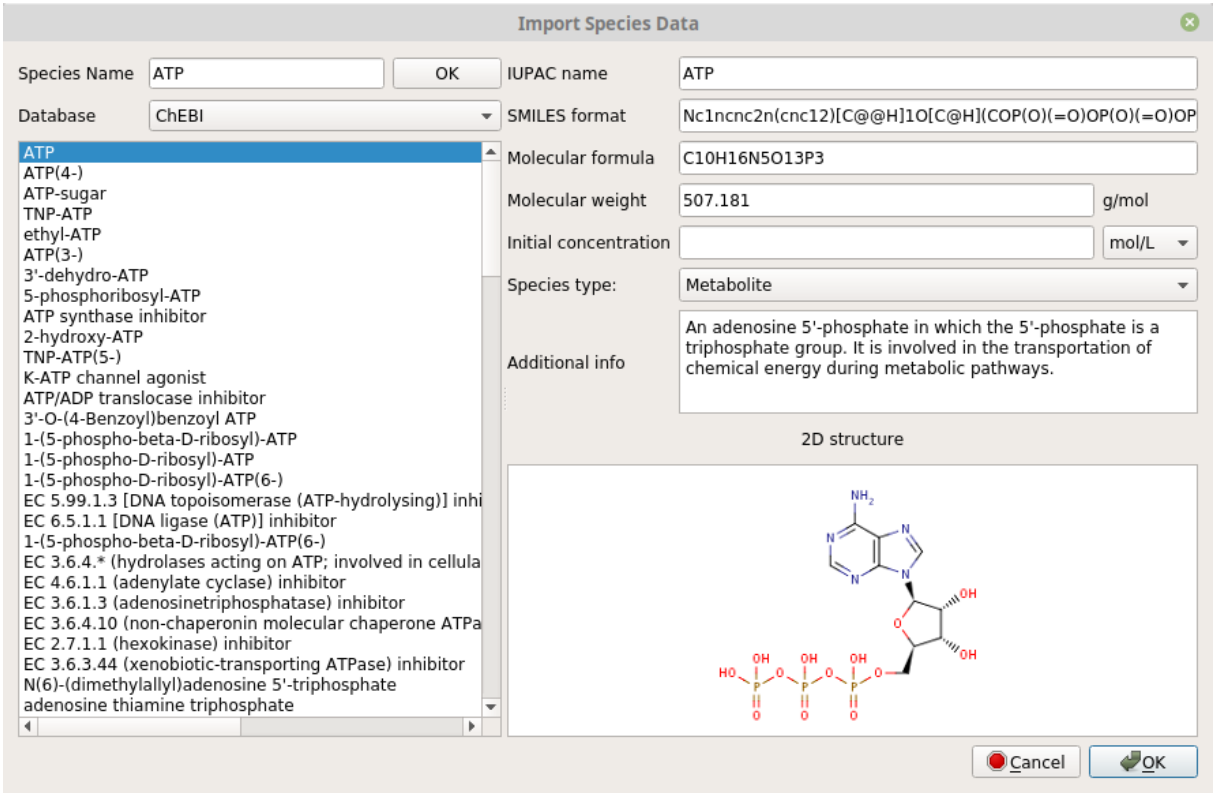


Figura 16 – Interface gráfica da ferramenta de importação de dados

Por fim, na opção de edição de modelo oferecida neste módulo, o usuário tem acesso às configurações da simulação de cada um dos modelos, podendo especificar ou modificar os algoritmos que serão utilizados para simular cada um dos modelos, ou definir manualmente a duração da simulação de cada um dos modelos ou da simulação integrada.

4.2.3 Módulo de simulação e visualização de resultados

Concluída a construção do(s) modelo(s), o usuário é direcionado a este módulo automaticamente, onde ele poderá acessar os resultados da simulação. Três visualizações distintas são oferecidas ao usuário: Uma série temporal, uma tabela de resultados e o gráfico de ativação de modelos. Nas figuras 17, 18 e 19 são mostradas as três visualizações disponíveis.

A primeira visualização retrata a evolução das concentrações ou contagem de moléculas de cada um dos componentes do sistema ao longo do tempo. Através dela, o usuário pode verificar visualmente a dinâmica dos diversos componentes, identificando padrões e pontos a partir dos quais o equilíbrio é alcançado, por exemplo. Neste módulo, é permitido ao usuário selecionar as espécies que deseja visualizar, bem como o estilo de cada uma das linhas que as representa, facilitando a visualização e análise dos resultados da simulação dos modelos desenvolvidos.

A tabela de resultados mostra os valores numéricos assumidos pelas variáveis em cada um dos tempos. Através dessa visualização, o usuário pode verificar numericamente os valores de cada uma das variáveis através do tempo, permitindo, por exemplo, identificar qual a concentração de determinada espécie no sistema no momento em que o equilíbrio é alcançado.

Já o gráfico de ativação dos modelos permite ao usuário avaliar a ordem e o tempo de ativação de cada um dos modelos, o que permite avaliar o efeito da ativação de cada um dos modelos sobre a dinâmica geral de cada uma das espécies no sistema combinado.

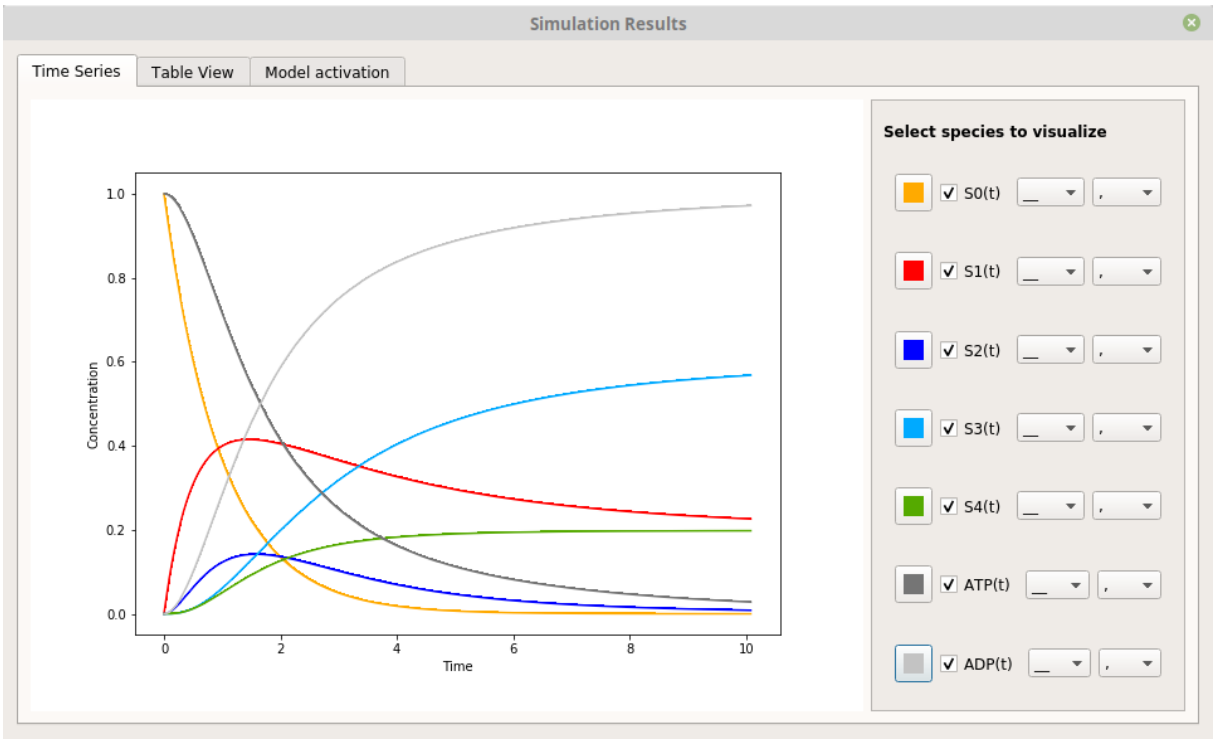
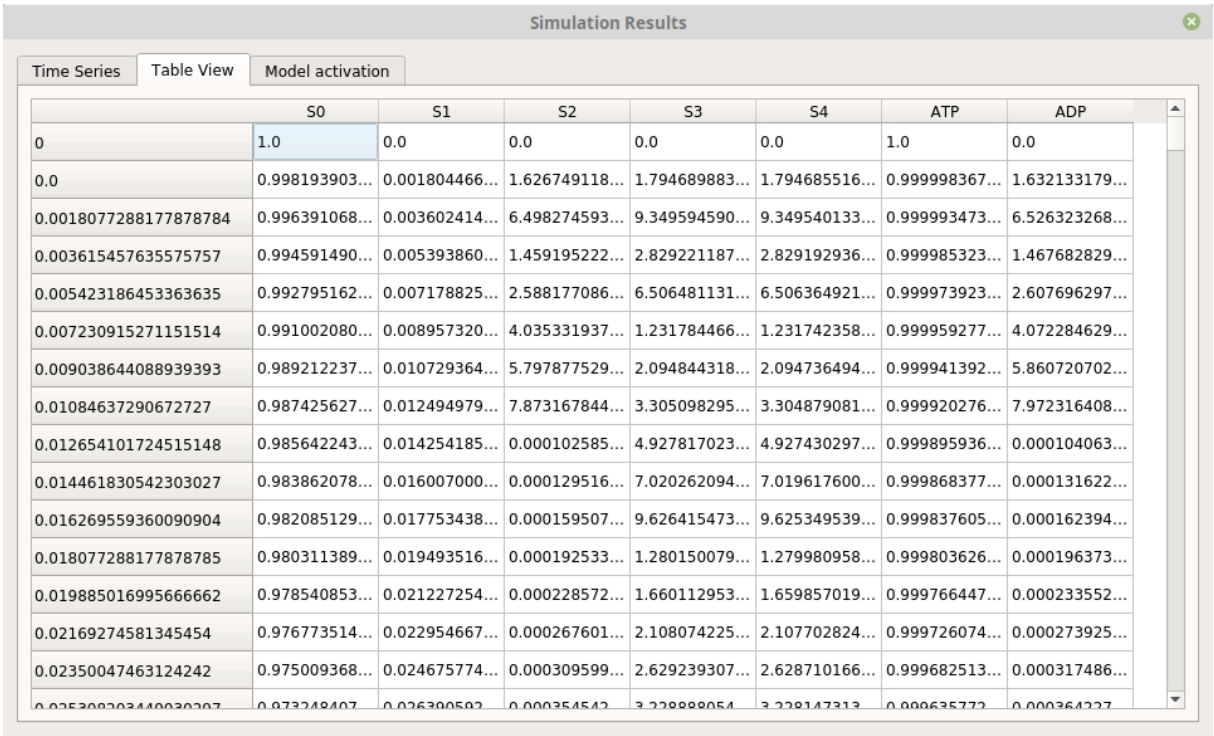


Figura 17 – Visualização da série temporal indicando a dinâmica de cada uma das espécies no sistema



	S0	S1	S2	S3	S4	ATP	ADP
0	1.0	0.0	0.0	0.0	0.0	1.0	0.0
0.0	0.998193903...	0.001804466...	1.626749118...	1.794689883...	1.794685516...	0.999998367...	1.632133179...
0.0018077288177878784	0.996391068...	0.003602414...	6.498274593...	9.349594590...	9.349540133...	0.999993473...	6.526323268...
0.003615457635575757	0.994591490...	0.005393860...	1.459195222...	2.829221187...	2.829192936...	0.999985323...	1.467682829...
0.005423186453363635	0.992795162...	0.007178825...	2.588177086...	6.506481131...	6.506364921...	0.999973923...	2.607696297...
0.007230915271151514	0.991002080...	0.008957320...	4.035331937...	1.231784466...	1.231742358...	0.999959277...	4.072284629...
0.009038644088939393	0.989212237...	0.010729364...	5.797877529...	2.094844318...	2.094736494...	0.999941392...	5.860720702...
0.01084637290672727	0.987425627...	0.012494979...	7.873167844...	3.305098295...	3.304879081...	0.999920276...	7.972316408...
0.012654101724515148	0.985642243...	0.014254185...	0.000102585...	4.927817023...	4.927430297...	0.999895936...	0.000104063...
0.014461830542303027	0.983862078...	0.016007000...	0.000129516...	7.020262094...	7.019617600...	0.999868377...	0.000131622...
0.016269559360090904	0.982085129...	0.017753438...	0.000159507...	9.626415473...	9.625349539...	0.999837605...	0.000162394...
0.018077288177878785	0.980311389...	0.019493516...	0.000192533...	1.280150079...	1.279980958...	0.999803626...	0.000196373...
0.019885016995666662	0.978540853...	0.021227254...	0.000228572...	1.660112953...	1.659857019...	0.999766447...	0.000233552...
0.02169274581345454	0.976773514...	0.022954667...	0.000267601...	2.108074225...	2.107702824...	0.999726074...	0.000273925...
0.02350047463124242	0.975009368...	0.024675774...	0.000309599...	2.629239307...	2.628710166...	0.999682513...	0.000317486...
0.025308203440930307	0.973248407...	0.026390592...	0.000354542...	3.228888054...	3.228147313...	0.999635772...	0.000364227...

Figura 18 – Visualização numérica das concentrações ao longo do tempo

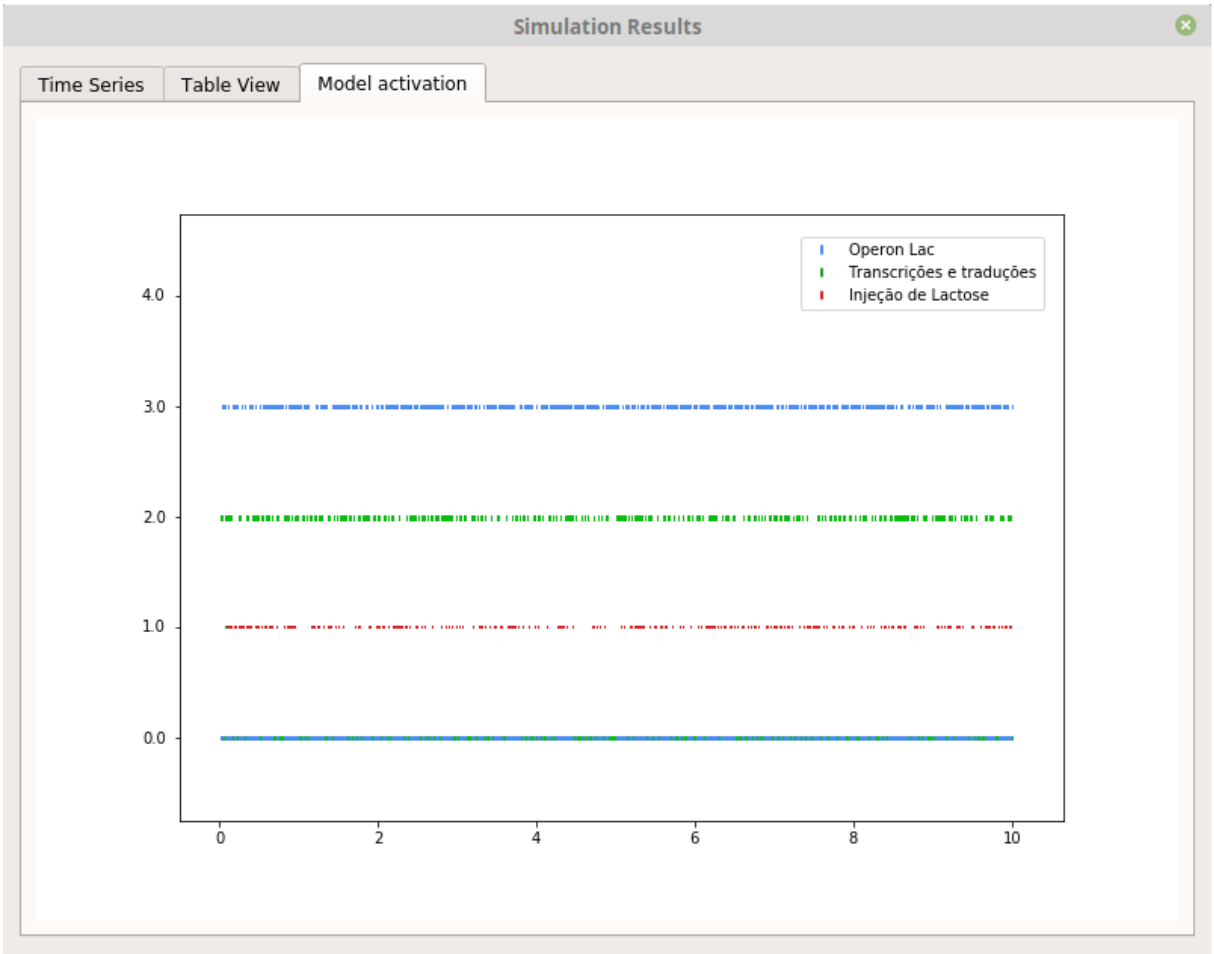


Figura 19 – Gráfico de ativação de modelos simulados utilizando o Cell Lab

5 Discussão

Neste trabalho, foi detalhado o desenvolvimento de uma alternativa para executar simulações multi-algorítmicas tendo como inspiração o algoritmo de simulação estocástica. O método estocástico de simulação multi-algorítmica visa viabilizar a realização de simulações integradas utilizando modelos que sigam distintos formalismos matemáticos e/ou utilizem diferentes algoritmos para simulação.

Através da aplicação de testes com o intuito de validar o método proposto foi possível constatar que os resultados alcançados com o método são comparáveis àqueles obtidos através da aplicação do SSA a um modelo monolítico que represente o mesmo sistema total. Isso demonstra que as hipóteses utilizadas na formulação do método são válidas, o que se traduz na possibilidade de sua aplicação para o desenvolvimento de modelos compostos e mais abrangentes, a exemplo de um modelo de célula completa.

Dentre as vantagens oferecidas pelo método proposto destaca-se a possibilidade de utilização de distintas metodologias, o que torna possível a descrição de um modelo com distintas granularidades e, por conseguinte, facilita a descrição do sistema baseado nos dados e informações disponíveis. Por sua vez, esta possibilidade faz com que o método habilite a reutilização modelos existentes, simulando-os conjuntamente com modelos de sistemas adjacentes. Isso se traduz na capacidade de integrar modelos de uma forma "*plug-and-play*", no sentido em que as partes (submodelos) podem ser facilmente substituídas ou adicionadas com o objetivo de melhorar ou ampliar a simulação. Deve-se ressaltar, no entanto, que a nomenclatura utilizada em todos os modelos deve ser consistente, resultando na necessidade de adaptações.

Além da flexibilidade conferida pelo método proposto, este apresenta possibilidade de ser paralelizado, uma vez que esta é uma característica herdada do SSA, que foi o ponto de partida para sua formulação. Neste cenário, em cada passo temporal, um ou mais modelos poderiam ser ativados e simulados separadamente. Entretanto, uma coordenação cuidadosa do número de moléculas ou concentração de cada um dos reagentes deveria ser realizada, de maneira a evitar a ocorrência de quantidades negativas, o que contrariaria os princípios da física.

A desvantagem mais importante da simulação de modelos através do método estocástico de simulação multi-algorítmica é o aumento considerável do custo computacional para processamento dos modelos. Os testes realizados mostraram que, apesar do ganho em eficiência alcançado pela seleção de um salto temporal maior, a aplicação da metodologia (em sua versão sequencial, como proposta aqui) é mais demorada que a utilização do SSA convencional para a simulação de um modelo monolítico de um

sistema equivalente.

Como limitação, destaca-se a incapacidade da técnica ser utilizada para permitir a simulação de modelos cuja simulação seja realizada através de algoritmos estáticos, a exemplo dos modelos baseados em restrições, das redes booleanas e das redes de Petri. Entretanto, assume-se que aplicação desta metodologia às versões dinâmicas dos algoritmos utilizados para simular estes modelos (ex.: FBA dinâmico e redes de Petri temporizadas) seja possível.

No tocante à ferramenta Cell Lab, esta mostrou-se um recurso conveniente tanto para construção de modelos quanto para aplicação da metodologia. Seu uso permite que modelos sejam construídos e salvos em formato próprio, além de oferecer funcionalidades que visam facilitar a conferência dos modelos, reduzindo as chances de erros. A possibilidade de criar modelos diagramaticamente é um auxílio importante para permitir seu uso por cientistas sem conhecimentos extensivos de programação. Além disso, a facilidade da aplicação expressa do método aqui proposto através do uso do framework faz com que a replicação dos resultados aqui obtidos seja simplificada.

Entretanto, cabe ressaltar que apesar das utilidades conferidas pela ferramenta, sua aplicação a modelos já publicados ainda é inviabilizada pela falta de suporte aos formatos de arquivo mais comumente utilizados pela comunidade para publicação de modelos, a exemplo do SBML. Neste aspecto, a capacidade não só de interpretar tais padrões, como também salvar os modelos construídos utilizando esta ferramenta nos padrões atuais é uma característica essencial para sua popularização.

6 Conclusão

Perante o exposto nas seções anteriores, conclui-se que a metodologia aqui proposta é uma resposta satisfatória ao principal questionamento que o presente trabalho buscou solucionar. Sua aplicação aos modelos selecionados mostrou que através das modificações e generalizações realizadas no método de simulação estocástica foi possível combinar modelos matematicamente dissimilares em uma única simulação, mantendo seus algoritmos de simulação originais.

No entanto, dadas as vantagens, desvantagens e limitações identificadas, percebe-se que esta não é uma solução final e definitiva, tampouco única, para o problema levantado. A metodologia aqui proposta deve ser vista como um primeiro passo numa longa jornada de viabilizar a construção de modelos computacionais de sistemas biológicos que representem um conjunto mais abrangente de fenômenos. Para o caso específico dos modelos de célula completa, nota-se que os resultados alcançados permitem vislumbrar a aplicação desta metodologia para a criação de futuros modelos do tipo, possibilitando a reutilização de modelos já publicados de vias metabólicas independentes, por exemplo. Contudo, em face da baixa eficiência computacional alcançada com a metodologia, melhorias seriam necessárias para que um tempo razoável de simulação fosse obtido.

Levando-se todos esses aspectos em consideração, entende-se que trabalhos futuros podem aprimorar a metodologia proposta, tornando sua aplicação mais vantajosa não só do ponto de vista da eficiência computacional, como também da fidelidade aos eventos representados (múltiplos modelos sendo processados simultaneamente). Além disso, a implementação dos melhoramentos propostos ao Cell Lab aumentaria ainda mais as potencialidades e aplicabilidade do método.

Uma das principais sugestões para futuros esforços refere-se à busca pelo aumento da eficiência do método estocástico de simulação multi-algorítmica, dada a possibilidade de paralelização que este método possui. Para este fim, tecnologias como o MPI (CLARKE et al., 1994), OpenMP (DAGUM; MENON, 1998), Nvidia CUDA (NICKOLLS et al., 2008) ou OpenCL (Stone et al., 2010) podem ser utilizadas para alcançar bons resultados. A paralelização da técnica pode ser feita em dois níveis: no sistema (simulando modelos paralelamente) e nos modelos (processando reações paralelamente). O ganho em eficiência promovido por esta melhoria é importante para aprimorar os resultados alcançados através da aplicação desta metodologia à simulação de modelos, especialmente os que englobam muitos processos e possuem maior complexidade.

No tocante ao software Cell Lab, verifica-se a oportunidade de ampliar seu propósito, tornando-o uma ferramenta capaz de atender a todos os objetivos de modelagem

de sistemas biológicos. Para isso, alguns requisitos devem ser cumpridos, a exemplo de:

- Adicionar suporte a distintos algoritmos de simulação (FBA, Redes de Petri, algoritmos lógicos, etc.).
- Implementar suporte a modelos em formatos padrão, como SBML e SEDML.
- Oferecer novas opções de visualização de resultados.
- Introduzir novas funcionalidades para a modelagem (ex.: Definição de múltiplos compartimentos).
- Expandir capacidade de acesso a bancos de dados de interesse (ex.: BioNumbers, Rhea e Sabio-RK).
- Incluir algoritmos de estimativa de parâmetros.
- Permitir trabalhar com equações personalizadas pelo usuário.

Referências

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, n. 3, p. 403 – 410, 1990. ISSN 0022-2836. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0022283605803602>.

ANDRES, K.; EILIS, R. Computational System Biology. p. 1–409, 2006.

ANGERMUELLER, C.; PÄRNAMAA, T.; PARTS, L.; STEGLE, O. Deep learning for computational biology. **Molecular Systems Biology**, v. 12, n. 7, p. 878, 2016. ISSN 1744-4292.

ASHRAFIAN, H. Mathematics in medicine: the 300-year legacy of iatromathematics. **Lancet**, v. 382, n. 9907, p. 1780, 2013. ISSN 1474547X.

BAIANU, I. C. Computer Models and Automata Theory in Biology and Medicine. **Mathematical Modelling**, v. 7, p. 1513–1577, 1986.

BAKER, R. E.; PEÑA, J. M.; JAYAMOHAN, J.; JÉRUSALEM, A. Mechanistic models versus machine learning, a fight worth fighting for the biological community? **Biology Letters**, v. 14, n. 5, p. 1–4, 2018. ISSN 1744957X.

BARDINI, R.; POLITANO, G.; BENSO, A.; Di Carlo, S. Multi-level and hybrid modelling approaches for systems biology. **Computational and Structural Biotechnology Journal**, Elsevier B.V., v. 15, p. 396–402, 2017. ISSN 20010370. Disponível em: <https://doi.org/10.1016/j.csbj.2017.07.005>.

BAYAT, A. Clinical review Science, medicine, and the future Bioinformatics. **British Medical Journal**, v. 324, n. April, p. 1018–1022, 2002.

BRADHAM, G. B. COMPUTERS IN MEDICINEl.2 By JAMES V. MALONEY, JR., M. D., AND GILBERT B. BRADHAM, M. D. n. January, p. 239–253, 1964.

BRODLAND, G. W. Seminars in Cell & Developmental Biology How computational models can help unlock biological systems. Elsevier Ltd, v. 48, p. 62–73, 2015.

CAMACHO, D. M.; COLLINS, K. M.; POWERS, R. K.; COSTELLO, J. C.; COLLINS, J. J. Next-Generation Machine Learning for Biological Networks. **Cell**, Elsevier Inc., v. 173, n. 7, p. 1581–1592, 2018. ISSN 10974172. Disponível em: <https://doi.org/10.1016/j.cell.2018.05.015>.

CAO, Y.; GILLESPIE, D. T.; PETZOLD, L. R. Efficient step size selection for the tau-leaping simulation method. **Journal of Chemical Physics**, v. 124, n. 4, 2006. ISSN 00219606.

CARRERA, J.; COVERT, M. W. Why Build Whole-Cell Models? **Trends in Cell Biology**, Elsevier Ltd, v. 25, n. 12, p. 719–722, 2015. ISSN 18793088. Disponível em: <http://dx.doi.org/10.1016/j.tcb.2015.09.004>.

CHEN, W. W.; NIEPEL, M.; SORGER, P. K. Classic and contemporary approaches to modeling biochemical reactions. **Genes and Development**, v. 24, n. 17, p. 1861–1875, 2010. ISSN 08909369.

CHRISTOPHER WANJEK. Systems Biology as Defined by NIH An Intellectual Resource for Integrative Biology. v. 19, n. 6, p. 1–20, 2011.

CLARKE, L.; GLENDINNING, I.; HEMPEL, R. The mpi message passing interface standard. In: DECKER, K. M.; REHMANN, R. M. (Ed.). **Programming Environments for Massively Parallel Distributed Systems**. Basel: Birkhäuser Basel, 1994. p. 213–218. ISBN 978-3-0348-8534-8.

COLLINS, F. S.; MORGAN, M.; PATRINOS, A. The Human Genome Project: Lessons from large-scale biology. **Science**, v. 300, n. 5617, p. 286–290, 2003. ISSN 00368075.

COOK, C. E.; BERGMAN, M. T.; FINN, R. D.; COCHRANE, G.; BIRNEY, E.; APWEILER, R. The European Bioinformatics Institute in 2016: Data growth and integration. **Nucleic Acids Research**, v. 44, n. D1, p. D20–D26, 2016. ISSN 13624962.

Da Silva, J. O.; ORELLANA, E. T. V.; DELGADO, M. X. T. Development of a parallel version of PhyML 3.0 using shared memory. **IEEE Latin America Transactions**, v. 15, n. 5, p. 959–967, 2017. ISSN 15480992.

DAGUM, L.; MENON, R. Openmp: An industry-standard api for shared-memory programming. **IEEE Comput. Sci. Eng.**, IEEE Computer Society Press, Los Alamitos, CA, USA, v. 5, n. 1, p. 46–55, jan. 1998. ISSN 1070-9924. Disponível em: <https://doi.org/10.1109/99.660313>.

DANOS, V.; FERET, J.; FONTANA, W.; KRIVINE, J. Scalable simulation of cellular signaling networks. In: SHAO, Z. (Ed.). **Programming Languages and Systems**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 139–157. ISBN 978-3-540-76637-7.

EBML. **Scientific Report 2017**. Hixton - Cambrige, 2017. 92 p.

ENERGY, U. D. of. **Human Genome Project Information Archive1990–2003**. 2013. Disponível em: <https://web.ornl.gov/sci/techresources/Human_Genome/project/budget.shtml>.

FISHER, J.; HENZINGER, T. A. Executable cell biology. **Nature Biotechnology**, v. 25, n. 11, p. 1239–1249, 2007. ISSN 10870156.

FRASER, C. M.; GOCAYNE, J. D.; WHITE, O.; ADAMS, M. D.; CLAYTON, R. A.; FLEISCHMANN, R. D.; BULT, C. J.; KERLAVAGE, A. R.; SUTTON, G.; KELLEY, J. M.; FRITCHMAN, J. L.; WEIDMAN, J. F.; SMALL, K. V.; SANDUSKY, M.; FUHRMANN, J.; NGUYEN, D.; UTTERBACK, T. R.; SAUDEK, D. M.; PHILLIPS, C. A.; MERRICK, J. M.; TOMB, J.-F.; DOUGHERTY, B. A.; BOTT, K. F.; HU, P.-C.; LUCIER, T. S. The Minimal Gene Complement of Mycoplasma genitalium. **Science**, v. 270, n. 5235, p. 397–404, 1995. ISSN 0036-8075. Disponível em: <http://www.sciencemag.org/cgi/doi/10.1126/science.270.5235.397>.

FREDDOLINO, P. L.; TAVAZOIE, S. The dawn of virtual cell biology. **Cell**, v. 150, n. 2, p. 248–250, 2012. ISSN 00928674.

GHOSH, S.; MATSUOKA, Y.; ASAI, Y.; HSIN, K. Y.; KITANO, H. Software for systems biology: From tools to integrated platforms. **Nature Reviews Genetics**, Nature Publishing Group, v. 12, n. 12, p. 821–832, 2011. ISSN 14710056. Disponível em: <http://dx.doi.org/10.1038/nrg3096>.

GILLESPIE, D. Approximate accelerated stochastic simulation of chemically reacting systems. **Journal of Chemical Physics**, v. 115, p. 1716–1733, 07 2001.

GILLESPIE, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. **Journal of Computational Physics**, v. 22, n. 4, p. 403–434, 1976. ISSN 10902716.

GILLESPIE, D. T. Stochastic Simulation of Chemical Kinetics. **Annual Review of Physical Chemistry**, v. 58, n. 1, p. 35–55, 2007. ISSN 0066-426X.

GLONT, M.; NGUYEN, T. V.; GRAESSLIN, M.; HÄLKE, R.; ALI, R.; SCHRAMM, J.; WIMALARATNE, S. M.; KOTHAMACHU, V. B.; RODRIGUEZ, N.; SWAT, M. J.; EILS, J.; EILS, R.; LAIBE, C.; MALIK-SHERIFF, R. S.; CHELLIAH, V.; Le Novère, N.; HERMJAKOB, H. BioModels: Expanding horizons to include more modelling approaches and formats. **Nucleic Acids Research**, v. 46, n. D1, p. D1248–D1253, 2018. ISSN 13624962.

GOLDBERG, A. P.; CHEW, Y. H.; KARR, J. R. Toward Scalable Whole-Cell Modeling of Human Cells. **Proceedings of the 2016 annual ACM Conference on SIGSIM Principles of Advanced Discrete Simulation - SIGSIM-PADS '16**, p. 259–262, 2016. Disponível em: <http://dl.acm.org/citation.cfm?doid=2901378.2901402>.

HARRIS, L. A.; HOGG, J. S.; TAPIA, J. J.; SEKAR, J. A.; GUPTA, S.; KORSUNSKY, I.; ARORA, A.; BARUA, D.; SHEEHAN, R. P.; FAEDER, J. R. BioNetGen 2.2: Advances in rule-based modeling. **Bioinformatics**, v. 32, n. 21, p. 3366–3368, 2016. ISSN 14602059.

HAYES, B. Imitation of life. **American Scientist**, v. 101, n. 1, p. 10–15, 2013. ISSN 00030996.

HOU, J.; ACHARYA, L.; ZHU, D.; CHENG, J. An overview of bioinformatics methods for modeling biological pathways in yeast. **Briefings in Functional Genomics**, v. 15, n. 2, p. 95–108, 2016. ISSN 20412657.

HUCKA, M.; BERGMANN, F. T.; DRÄGER, A.; HOOPS, S.; KEATING, S. M.; Le Novère, N.; MYERS, C. J.; OLIVIER, B. G.; SAHLE, S.; SCHAFF, J. C.; SMITH, L. P.; WALTEMATH, D.; WILKINSON, D. J. **The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core**. [S.l.: s.n.], 2018. v. 15. ISSN 16134516. ISBN 0000000252935.

HUCKA, M.; BERGMANN, F. T.; KEATING, S. M.; SMITH, L. P. A profile of today's SBML-compatible software. **Proceedings - 7th IEEE International Conference on e-Science Workshops, eScienceW 2011**, n. May 2014, p. 143–150, 2011.

ISALAN, M. A cell in a computer. **Nature**, v. 488, n. 7409, p. 40–41, 2012. ISSN 0028-0836.

KANEHISA, M.; FURUMICHI, M.; TANABE, M.; SATO, Y.; MORISHIMA, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. **Nucleic Acids Research**, v. 45, n. D1, p. D353–D361, 2017. ISSN 13624962.

KARP, P. D.; BILLINGTON, R.; CASPI, R.; FULCHER, C. A.; LATENDRESSE, M.; KOTHARI, A.; KESELER, I. M.; KRUMMENACKER, M.; MIDFORD, P. E.; ONG, Q.; ONG, W. K.; PALEY, S. M.; SUBHRAVETI, P. The BioCyc collection of microbial genomes and metabolic pathways. **Briefings in Bioinformatics**, v. 20, n. 4, p. 1085–1093, 08 2017. ISSN 1477-4054. Disponível em: <https://doi.org/10.1093/bib/bbx085>.

KARR, J. R.; LLUCH-SENAR, M.; SERRANO, L.; KASTELIC, D. The 2017 Whole-Cell Modeling Summer School. p. 2–6, 2017.

KARR, J. R.; SANGHVI, J. C.; MACKLIN, D. N.; GUTSCHOW, M. V.; JACOBS, J. M.; BOLIVAL, B.; ASSAD-GARCIA, N.; GLASS, J. I.; COVERT, M. W. A whole-cell computational model predicts phenotype from genotype. **Cell**, v. 150, n. 2, p. 389–401, 2012. ISSN 00928674.

KARR, J. R.; SANGHVI, J. C.; MACKLIN, D. N.; ARORA, A.; COVERT, M. W. WholeCellKB: Model organism databases for comprehensive whole-cell models. **Nucleic Acids Research**, v. 41, n. D1, p. 787–792, 2013. ISSN 03051048.

KLIPP, E.; Liebermeister, Wolfram Wierling, C.; AXEL, K. **Systems Biology**. [S.l.: s.n.], 2016. v. 2. 1–488 p. ISSN 1098-6596. ISBN 9788578110796.

LARRAÑAGA, P.; CALVO, B.; SANTANA, R.; BIELZA, C.; GALDIANO, J.; INZA, I.; LOZANO, J. A.; ARMAÑANZAS, R.; SANTAFÉ, G.; PÉREZ, A.; ROBLES, V. Machine learning in bioinformatics. **Briefings in Bioinformatics**, v. 7, n. 1, p. 86–112, 2006. ISSN 14675463.

LEFÈVRE, T.; STINDEL, E.; ANSART, S.; ROUX, C. Mathematics in medicine: beyond iatromathematics. **The Lancet**, v. 383, n. 9916, p. 513, 2014. ISSN 01406736.

MACHADO, D.; COSTA, R. S.; ROCHA, M.; FERREIRA, E. C.; TIDOR, B.; ROCHA, I. Modeling formalisms in Systems Biology. **AMB Express**, v. 1, p. 45, dec 2011. ISSN 2191-0855. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/22141422http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3285092>.

MACKLIN, D. N.; RUGGERO, N. A.; COVERT, M. W. The future of whole-cell modeling. **Current Opinion in Biotechnology**, Elsevier Ltd, v. 28, p. 111–115, 2014. ISSN 09581669. Disponível em: <http://dx.doi.org/10.1016/j.copbio.2014.01.012>.

MALIK-SHERIFF, R. S.; GLONT, M.; NGUYEN, T. V. N.; TIWARI, K.; ROBERTS, M. G.; XAVIER, A.; VU, M. T.; MEN, J.; MAIRE, M.; KANANATHAN, S.; FAIRBANKS, E. L.; MEYER, J. P.; ARANKALLE, C.; VARUSAI, T. M.; KNIGHT-SCHRIJVER, V.; LI, L.; DUEñAS-ROCA, C.; DASS, G.; KEATING, S. M.; PARK, Y. M.; BUSO, N.; RODRIGUEZ, N.; HUCKA, M.; HERMJAKOB, H. BioModels—15 years of sharing computational models in life science. **Nucleic Acids Research**, 11 2019. ISSN 0305-1048. Gkz1055. Disponível em: <https://doi.org/10.1093/nar/gkz1055>.

MCMAHON, A. W.; COOPER, W. O.; BROWN, J. S.; CARLETON, B.; DOSHI-VELEZ, F.; KOHANE, I.; GOLDMAN, J. L.; HOFFMAN, M. A.; KAMALESWARAN, R.; SAKIYAMA, M.; SEKINE, S.; STURKENBOOM, M. C.; TURNER, M. A.; CALIFF, R. M. Big data in the assessment of pediatric medication safety. **Pediatrics**, American Academy of Pediatrics, v. 145, n. 2, 2020. ISSN 0031-4005. Disponível em: <https://pediatrics.aappublications.org/content/145/2/e20190562>.

MEDLEY, J. K.; GOLDBERG, A. P.; KARR, J. R. Guidelines for Reproducibly Building and Simulating Systems Biology Models. **IEEE Transactions on Biomedical Engineering**, IEEE, v. 63, n. 10, p. 2015–2020, 2016. ISSN 15582531.

MOROWITZ, H. J. The completeness of molecular biology. **Israel journal of medical sciences**, v. 20, n. 9, p. 750–3, sep 1984. ISSN 0021-2180. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/6511349>.

MURPHY, R. F. **What is Computational Biology?: Computational Biology Department**. Carnegie Mellon University, 2016. Disponível em: <http://www.cbd.cmu.edu/about-us/what-is-computational-biology/>.

MYERS, C. J.; BARKER, N.; JONES, K.; KUWAHARA, H.; MADSEN, C.; NGUYEN, N. P. D. iBioSim: A tool for the analysis and design of genetic circuits. **Bioinformatics**, v. 25, n. 21, p. 2848–2849, 2009. ISSN 13674803.

NEAL, M. L.; COOLING, M. T.; SMITH, L. P.; THOMPSON, C. T.; SAURO, H. M.; CARLSON, B. E.; COOK, D. L.; GENNARI, J. H. A Reappraisal of How to Build Modular, Reusable Models of Biological Systems. **PLoS Computational Biology**, v. 10, n. 10, p. e1003849, 2014.

NICKOLLS, J.; BUCK, I.; GARLAND, M.; SKADRON, K. Scalable parallel programming with CUDA. **Queue**, v. 6, n. 2, p. 40–53, 2008. ISSN 15427730.

PALSSON, B.; ZENGLER, K. The challenges of integrating multi-omic data sets. **Nature Chemical Biology**, Nature Publishing Group, v. 6, n. 11, p. 787–789, 2010. ISSN 15524469. Disponível em: <http://dx.doi.org/10.1038/nchembio.441>.

PLUNKETT, R. L.; GEMMILL, C. L. The kinetics of invertase action. **The Bulletin of Mathematical Biophysics**, v. 13, n. 4, p. 303–312, 1951. ISSN 00074985.

PURCELL, O.; JAIN, B.; KARR, J. R.; COVERT, M. W.; LU, T. K. Towards a whole-cell modeling approach for synthetic biology. **Chaos**, v. 23, n. 2, p. 1–8, 2013. ISSN 10541500.

SANGHVI, J. C.; REGOT, S.; CARRASCO, S.; KARR, J. R.; GUTSCHOW, M. V.; BOLIVAL, B.; COVERT, M. W. Accelerated discovery via a whole-cell model. **Nature Methods**, Nature Publishing Group, v. 10, n. 12, p. 1192–1195, 2013. ISSN 15487091. Disponível em: <http://dx.doi.org/10.1038/nmeth.2724>.

SBML, T. **SBML Software Guide**. 2017. "<http://sbml.org/SBML_Software_Guide>".

SCHATZ, M. C. Computational thinking in the era of big data biology. **Genome Biology**, v. 13, n. 11, p. 177, 2012. ISSN 1465-6906.

SIMEONOV, P. L.; GOMEZ-RAMIREZ, J.; SIREGAR, P. On some recent insights in Integral Biomathics. **Progress in Biophysics and Molecular Biology**, Elsevier, v. 113, n. 1, p. 216–228, 2013. ISSN 00796107. Disponível em: <http://dx.doi.org/10.1016/j.pbiomolbio.2013.06.001>.

SOMOGYI, E. T.; BOUTEILLER, J. M.; GLAZIER, J. A.; KÖNIG, M.; MEDLEY, J. K.; SWAT, M. H.; SAURO, H. M. LibRoadRunner: A high performance SBML simulation and analysis library. **Bioinformatics**, v. 31, n. 20, p. 3315–3321, 2015. ISSN 14602059.

STANFORD, N. J.; WOLSTENCROFT, K.; GOLEBIEWSKI, M.; KANIA, R.; JUTY, N.; TOMLINSON, C.; OWEN, S.; BUTCHER, S.; HERMJAKOB, H.; NOVèRE, N. L.; MUELLER, W.; SNOEP, J.; GOBLE, C. The evolution of standards and data management practices in systems biology. **Molecular Systems Biology**, v. 11, n. 12, p. 851, 2015. Disponível em: <https://www.embopress.org/doi/abs/10.15252/msb.20156053>.

STEPHENS, Z. D.; LEE, S. Y.; FAGHRI, F.; CAMPBELL, R. H.; ZHAI, C.; EFRON, M. J.; IYER, R.; SCHATZ, M. C.; SINHA, S.; ROBINSON, G. E. Big data: Astronomical or genomical? **PLoS Biology**, v. 13, n. 7, p. 1–11, 2015. ISSN 15457885.

STEUER, R. Computational approaches to the topology, stability and dynamics of metabolic networks. **Phytochemistry**, v. 68, n. 16-18, p. 2139–2151, 2007. ISSN 00319422.

Stone, J. E.; Gohara, D.; Shi, G. Opencl: A parallel programming standard for heterogeneous computing systems. **Computing in Science Engineering**, v. 12, n. 3, p. 66–73, May 2010. ISSN 1558-366X.

SZIGETI, B.; ROTH, Y. D.; SEKAR, J. A.; GOLDBERG, A. P.; POCHIRAJU, S. C.; KARR, J. R. A blueprint for human whole-cell modeling. **Current Opinion in Systems Biology**, Elsevier Ltd, v. 7, p. 8–15, 2018. ISSN 24523100. Disponível em: <https://doi.org/10.1016/j.coisb.2017.10.005>.

TAKAHASHI, K.; YUGI, K.; HASHIMOTO, K.; YAMADA, Y.; TOMITA, M.; PICKETT, C. J. Computational Challenges in Cell Simulation: A Software Engineering Approach. **IEEE Intelligent Systems**, v. 17, n. 5, p. 64–71, 2002. ISSN 15411672.

TAREEN, A.; KINNEY, J. B. Biophysical models of cis-regulation as interpretable neural networks. **bioRxiv**, n. Mlcb, p. 835942, 2019. Disponível em: <https://www.biorxiv.org/content/10.1101/835942v2>.

TOMITA, M. Whole-cell simulation: A grand challenge of the 21st century. **Trends in Biotechnology**, v. 19, n. 6, p. 205–210, 2001. ISSN 01677799.

TOMITA, M.; HASHIMOTO, K.; TAKAHASHI, K.; SHIMIZU, T. S.; MATSUZAKI, Y.; MIYOSHI, F.; SAITO, K.; TANIDA, S.; YUGI, K.; VENTER, J. C.; HUTCHISON, C. A. E-CELL: Software environment for whole-cell simulation. **Bioinformatics**, v. 15, n. 1, p. 72–84, 1999. ISSN 13674803.

WAAGE, P.; GULBERG, C. M. Studies concerning affinity. **Journal of Chemical Education**, v. 63, n. 12, p. 1044–1047, 1864. ISSN 00219584.

WALTEMATH, D.; KARR, J. R.; BERGMANN, F. T.; CHELLIAH, V.; HUCKA, M.; KRANTZ, M.; LIEBERMEISTER, W.; MENDES, P.; MYERS, C. J.; PIR, P.; ALAYBEYOGLU, B.; ARANGANATHAN, N. K.; BAGHALIAN, K.; BITTIG, A. T.; BURKE, P. E.; CANTARELLI, M.; CHEW, Y. H.; COSTA, R. S.; CURSONS, J.; CZAUDERNA, T.; GOLDBERG, A. P.; GOMEZ, H. F.; HAHN, J.; HAMERI, T.; GARDIOL, D. F.; KAZAKIEWICZ, D.; KISELEV, I.; KNIGHT-SCHRIJVER, V.; KNUPFER, C.; KONIG, M.; LEE, D.; LLORET-VILLAS, A.; MANDRIK, N.; MEDLEY, J. K.; MOREAU, B.; NADERI-MESHKIN, H.; PALANIAPPAN, S. K.; PRIEGO-ESPINOSA, D.; SCHARM, M.; SHARMA, M.; SMALLBONE, K.; STANFORD, N. J.; SONG, J. H.; THEILE, T.; TOKIC, M.; TOMAR, N.; TOURE, V.; UHLENDORF, J.; VARUSAI, T. M.; WATANABE, L. H.; WENDLAND, F.; WOLFIEN, M.; YURKOVICH, J. T.; ZHU, Y.; ZARDILIS, A.; ZHUKOVA, A.; SCHREIBER, F. Toward

Community Standards and Software for Whole-Cell Modeling. **IEEE Transactions on Biomedical Engineering**, v. 63, n. 10, p. 2007–2014, 2016. ISSN 15582531.

WALTEMATH, D.; WOLKENHAUER, O. How Modeling Standards, Software, and Initiatives Support Reproducibility in Systems Biology and Systems Medicine. **IEEE Transactions on Biomedical Engineering**, IEEE, v. 63, n. 10, p. 1999–2006, 2016. ISSN 15582531.

WILKINSON, D. J. **Stochastic modelling for systems biology**. 3. ed. [S.l.]: CRC press, 2018.

ZHANG, W.; KOLTE, R.; DILL, D. L. Towards in vivo estimation of reaction kinetics using high-throughput metabolomics data: A maximum likelihood approach. **BMC Systems Biology**, v. 9, n. 1, p. 1–9, 2015. ISSN 17520509.

ZOU, Y.; LAUBICHLER, M. D. From systems to biology: A computational analysis of the research articles on systems biology from 1992 to 2013. **PLoS ONE**, v. 13, n. 7, p. 1–16, 2018. ISSN 19326203.

Apêndices

APÊNDICE A – Detalhamento do funcionamento e modelo do Operon Lac

O princípio de funcionamento do Operon Lac é bastante simples e previsível. Na ausência de lactose no sistema, a enzima RNA polimerase atua constantemente na transcrição do repressor lac, o qual se liga ao operon, e impede que a RNA polimerase transcreva as enzimas atuantes no metabolismo da Lactose. Na ausência de glicose e presença de lactose no sistema, o repressor se liga à lactose, desligando-se do operon, o que faz com que a RNA polimerase possa transcrever as enzimas que dão início ao metabolismo da lactose, eliminando-a do sistema.

As espécies do sistema, bem como seus valores iniciais e nomenclatura utilizada no modelo estão detalhados na tabela 5. As equações 24 a 36 representam cada uma das reações do sistema. Já os parâmetros cinéticos do sistema encontram-se especificados na tabela 6.

Tabela 5 – Espécies e contagens iniciais para o modelo do funcionamento do Operon Lac

Espécie	Nomenclatura	Contagem inicial
DNA do Repressor Lac	Idna	1
RNAm do Repressor Lac	Irna	0
Repressor Lac	I	50
RNA Polimerase	Rnap	100
Operon Lac	Op	1
RNAm das enzimas Z	Rna	0
Enzimas Z	Z	0
Lactose	Lactose	20
Repressor ligado à lactose	ILactose	0
Repressor ligado ao Operon	IOp	0
RNA polimerase ligada ao Repressor	RnaOp	0

Fonte: (WILKINSON, 2018)

R1: Transcrição do gene que codifica Repressor Lac

$$\text{Idna} \longrightarrow \text{Idna} + \text{Irna} \tag{24}$$

R2: Tradução do RNAm que codifica o Repressor Lac

$$\text{Irna} \longrightarrow \text{Irna} + \text{I} \tag{25}$$

R3: Formação(+) e dissociação(-) do complexo Repressor Lac - Lactose

$$\text{I} + \text{Lactose} \rightleftharpoons \text{ILactose} \tag{26}$$

R4: Ligamento(+) e desligamento(-) do Repressor Lac ao Operon

$$I + Op \rightleftharpoons IOp \tag{27}$$

R5: Ligamento(+) e desligamento(-) da RNA polimerase ao Operon

$$Rnap + Op \rightleftharpoons RnapOp \tag{28}$$

R6: Transcrição do Operon dando origem ao RNAm que codifica a enzima Z

$$RnapOp \longrightarrow Op + Rnap + Rna \tag{29}$$

R7: Tradução do RNAm que codifica a enzima Z

$$Rna \longrightarrow Rna + Z \tag{30}$$

R8: Metabolização da Lactose pela enzima Z

$$Z + Lactose \longrightarrow Z \tag{31}$$

R9: Degradação do RNAm que codifica o Repressor Lac

$$Irna \longrightarrow \tag{32}$$

R10: Degradação do Repressor Lac

$$I \longrightarrow \tag{33}$$

R11: Degradação do complexo Repressor Lac - Lactose

$$ILactose \longrightarrow \tag{34}$$

R12: Degradação do RNAm que codifica a enzima Z

$$Rna \longrightarrow \tag{35}$$

R13: Degradação da enzima Z

$$Z \longrightarrow \tag{36}$$

Tabela 6 – Parâmetros utilizados no modelo do funcionamento do Operon Lac

Parâmetro	Nomenclatura	Valor
Coeficiente cinético estocástico da reação 1	c1	0.02
Coeficiente cinético estocástico da reação 2	c2	0.1
Coeficiente cinético estocástico da reação 3 direta	c3	0.005
Coeficiente cinético estocástico da reação 3 reversa	c4	0.1
Coeficiente cinético estocástico da reação 4 direta	c5	1
Coeficiente cinético estocástico da reação 4 reversa	c6	0.01
Coeficiente cinético estocástico da reação 5 direta	c7	0.1
Coeficiente cinético estocástico da reação 5 reversa	c8	0.01
Coeficiente cinético estocástico da reação 6	c9	0.03
Coeficiente cinético estocástico da reação 7	c10	0.1
Coeficiente cinético estocástico da reação 8	c11	10^{-5}
Coeficiente cinético estocástico da reação 9	c12	0.01
Coeficiente cinético estocástico da reação 10	c13	0.002
Coeficiente cinético estocástico da reação 11	c14	0.002
Coeficiente cinético estocástico da reação 12	c15	0.01
Coeficiente cinético estocástico da reação 13	c16	0.001
Volume do sistema	v	10^{-15} L

Fonte: (WILKINSON, 2018)

Anexos

ANEXO A – Algoritmo de simulação estocástica original (GILLESPIE, 1976)

Dada as reações:

$$X \underset{c2}{\overset{c1}{\rightleftharpoons}} Y \tag{37}$$

$$_2X \underset{c4}{\overset{c3}{\rightleftharpoons}} Z \tag{38}$$

$$W + X \rightleftharpoons [c_5][c_6]_2X \tag{39}$$

Considerando que os 6 parâmetros c_1, c_2, $\cdots$, c_6 sejam conhecidos, bem como as concentrações iniciais W_0, X_0, Y_0 e Z_0, temos que o seguinte algoritmo pode ser utilizado para simular a evolução do número de moléculas de cada uma das espécies em relação ao tempo.

```
      DIMENSION C(6), A(6)
1     READ ((C(MU), MU = 1,6), T, W, X, Y, Z, T2, TINT)
      TPRINT = T
10    A(1) = C(1) * X
      A(2) = C(2) * Y
      A(3) = C(3) * X * (X-1.)/2.
      A(4) = C(4) * Z
      A(5) = C(5) * W * X
      A(6) = C(6) * X * (X-1.)/2.
      A0 = A(1) + A(2) + A(3) + A(4) + A(5) + A(6)
20    CALL URN (R1, R2)
21    T = T + ALOG(1./R1)/A0
22    IF (T . LT . TPRINT) GO TO 25
23    PRINT(TPRINT, W, X, Y , Z)
      TPRINT = TPRINT + TINT
      GO TO 22
25    R2A0 = R2 * A0
      SUM = 0
      DO 29 NU = 1,6
      MU = NU
      SUM = SUM + A(NU)
      IF (SUM . GE . R2A0) GO TO 30
29    CONTINUE
30    GO TO (31, 32, 33, 34, 35, 36), MU
31    X = X -1
      Y = Y + 1
      GO TO 40
32    X = X + 1
      Y = Y - 1
      GO TO 40
33    X = X - 2
      Z = Z + 1
      GO TO 40
34    X = X + 2
      Z = Z - 1
      GO TO 40
35    X = X + 1
      W = W - 1
      GO TO 40
36    X = X - 1
      W = W + 1
40    IF (T . LT . T2) GO TO 10
      STOP
      END
```

Algorithm 1: programa para simular as reações 37, 38, 39