UNIVERSIDADE FEDERAL DE JUIZ DE FORA
PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

ESTRATÉGIAS NUMÉRICAS E DE OTIMIZAÇÃO PARA INFERÊNCIA DA DINÂMICA DE REDES BIOQUÍMICAS

Carlos Roberto Lima Ladeira

Juiz de Fora
Fevereiro de 2014

Carlos Roberto Lima Ladeira

Estratégias Numéricas e de Otimização para Inferência da Dinâmica de Redes Bioquímicas

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Mestre em Modelagem Computacional.

Orientador: Prof. D.Sc. Carlos Cristiano Hasenclever Borges

Juiz de Fora

2014

Ladeira, Carlos Roberto Lima

Estratégias Numéricas e de Otimização para Inferência da Dinâmica de Redes Bioquímicas/Carlos Roberto Lima Ladeira. – Juiz de Fora: UFJF/MMC, 2014.

IX, 95 p.: il.; 29, 7cm.

Orientador: Carlos Cristiano Hasenclever Borges

Dissertação (mestrado) – UFJF/MMC/Programa de Modelagem Computacional, 2014.

Referências Bibliográficas: p. 89 – 95.

1. Sistemas bioquímicos. 2. Sistema S. 3. Problema inverso. 4. Método da Entropia Cruzada. 5. Método do Passo Complexo. I. Borges, Carlos Cristiano Hasenclever. II. Universidade Federal de Juiz de Fora, MMC, Programa de Modelagem Computacional.

Carlos Roberto Lima Ladeira

Estratégias Numéricas e de Otimização para Inferência da Dinâmica de Redes Bioquímicas

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Mestre em Modelagem Computacional.

Aprovada em 28 de Fevereiro de 2014.

BANCA EXAMINADORA

Prof. D.Sc. Carlos Cristiano Hasenclever Borges - Orientador
Universidade Federal de Juiz de Fora

D.Sc. Helio José Corrêa Barbosa
Universidade Federal de Juiz de Fora

D.Sc. Wagner Antônio Arbex
Empresa Brasileira de Pesquisa Agropecuária

AGRADECIMENTOS

À minha família, pela educação e conforto que sempre me deram.

À minha namorada, Cinara, pela paciência e incentivo (mais uma vez!).

A meus professores, pelo seu conhecimento passado com sabedoria e eficiência.

A meu orientador, Carlos Cristiano, pela boa sugestão de um tema interessante e motivante, pela confiança, pela paciência e por seu conhecimento compartilhado durante a orientação.

A todos os amigos do Programa de Pós-graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora pelo companheirismo sempre presente. Durante minha caminhada acabei não pertencendo a nenhuma turma específica e fiz minhas amizades em cada uma das disciplinas. Agradeço especialmente a Bruno Zonovelli (amigo de todos e sempre prestativo), Ricardo Campos, Fabrízzio Condé, Francisco Manfrini, Anderson da Paz, Guilherme Guilhermino, Vinícius Schmitz e Érica Carvalho.

Agradeço também, a todos que direta ou indiretamente me ajudaram nessa conquista.

RESUMO

Estimar parâmetros de modelos dinâmicos de sistemas biológicos usando séries temporais é cada vez mais importante, pois uma quantidade imensa de dados experimentais está sendo mensurados pela biologia molecular moderna. Uma abordagem de resolução baseada em problemas inversos pode ser utilizada na solução deste tipo de problema. A escolha do modelo matemático é uma tarefa importante, pois vários modelos podem ser utilizados, apresentando níveis diversos de precisão em suas representações.

A Teoria dos Sistemas Bioquímicos (TSB) faz uso de equações diferenciais ordinárias e expansões de séries de potências para representar processos bioquímicos. O Sistema S é um dos modelos usados pela TSB que permite a transformação do sistema original de equações diferenciais em um sistema algébrico desacoplado, facilitando a solução do problema inverso. Essa transformação pode comprometer a qualidade da resposta se o valor das derivadas nos pontos das séries temporais não for obtidos com precisão. Para estimar as derivadas pretende-se explorar o método do passo complexo, que apresenta vantagens em relação ao método das diferenças finitas, mais conhecido e utilizado.

A partir daí pode então ser realizada a busca pelas variáveis que definirão as equações do sistema. O método da Regressão Alternada é um dos mais rápidos para esse tipo de problema, mas a escolha inicial dos parâmetros possui influência em seu resultado, que pode até mesmo não ser encontrado. Pretende-se avaliar o método da Entropia Cruzada, que possui a vantagem de realizar buscas globais e talvez por esse motivo a escolha dos parâmetros inicias não cause tanta influência nos resultados. Além disso, será avaliado um método híbrido que fará uso das principais vantagens do método da Regressão Alternada e do Entropia Cruzada para resolver o problema.

Experimentos numéricos sistematizados serão realizados tanto para a etapa de estimativa das derivadas quanto para a etapa de otimização para obtenção dos parâmetros das equações do sistema.

Palavras-chave: Sistemas bioquímicos. Sistema S. Problema inverso. Método da Entropia Cruzada. Método do Passo Complexo.

ABSTRACT

Estimating parameters of dynamic models of biological systems using time series is becoming very important because a huge amount of experimental data is being measured by modern molecular biology. A resolution-based approach on inverse problems can be used in solving this type of problem. The choice of the mathematical model is an important task, since many models can be used, with varying levels of accuracy in their representations.

The Biochemical Systems Theory (BST) makes use of ordinary differential equations and power series expansions to represent biochemical processes. The S-system is one of the models used by BST that allows the transformation of the original system of differential equations in a decoupled system of algebric equations, favouring the solution of the inverse problem. This transformation can compromise the quality of the response if the value of the derivatives at points of time series are not obtained accurately. To estimate the derivatives we intend to explore the complex-step method, which has advantages over the finite difference method, best known and used .

So the search for the variables that define the equations of the system can be performed. The Alternating Regression method is one of the fastest for this type of problem, but the initial choice of parameters has influence on its performance, which may not even be found. We intend to evaluate the Cross-entropy method, which has the advantage of performing global searches and for this reason the choice of the initial search parameters does not cause as much influence on the results. Also, will be assessed a hybrid method that makes use of the main advantages of Alternating Regression and Cross-entropy to solve the problem.

Systematic numerical experiments will be conducted for both the step of estimating derivatives as for the optimization step to estimate the variables of the equations of the system.

Keywords: Biochemical systems. S-system. Inverse problem. Cross-entropy Method. Complex Step Method.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

1 INTRODUÇÃO

O estudo das ciências biológicas possui fundamental importância para a sociedade, pois sua aplicação se estende em diversas áreas de interesse, como botânica, zoologia, ecologia, genética e fisiologia humana. Vários temas com discussões recorrentes são fundamentados em conceitos biológicos, como pesquisas com células-tronco, alimentos transgênicos, poluição do meio ambiente, entre outros. Mais recentemente, a biotecnologia e as aplicações em medicina estão recebendo bastante atenção em pesquisas e aplicações práticas.

Assim como em qualquer outra ciência, métodos de representação e aplicação de modelos matemáticos são comumente utilizados com o objetivo de tentar analisar, entender e realizar previsões sobre determinado fenômeno. A biologia possui grande diversidade de modelos matemáticos que são utilizados para representar os mais variados fenômenos, com diferentes escalas de tamanho e tempo, além de quantidade de componentes variáveis (KLIPP et al., 2009). No entanto, existem críticas relacionadas ao fato de que vários modelos aplicados à biologia vieram de analogias feitas com outras ciências, como a física e a química, e que portanto as aproximações não são adequadas para representar os fenômenos biológicos estudados (KONOPKA, 2007)(WOLKENHAUER, 2001).

A ideia de estudar a biologia a partir de uma abordagem sistêmica foi apresentada há bastante tempo, mas não foi imediatamente adotada (WIENER, 1948)(BERTALANFFY, 1969). A biologia de sistemas, que também propõe essa abordagem, não tem o objetivo de substituir os modelos já usados para representar os sistemas biológicos por um outro mais sofisticado, mas se preocupa em considerar o relacionamento dos componentes desses sistemas de uma forma mais abrangente, levando em conta níveis de organização diferentes, como o molecular, o celular, órgãos e organismos. Além disso, a busca de modelos que sejam capazes não só de se ajustar a dados experimentais mas também de representar de forma intuitiva as estruturas e as funções de componentes de sistemas biológicos é outro fator considerado na biologia de sistemas (KLIPP et al., 2009).

Pelo fato descrito acima de a biologia de sistemas não ser um modelo específico mas sim um conjunto de conceitos sobre a abordagem sistêmica dos fenômenos biológicos, as aplicações, mesmo já sendo amplas, ainda podem se estender mais ao ser realizada a integração entre os vários níveis de organização existentes nos sistemas biológicos. Áreas

como descoberta de fármacos e medicina estão utilizando conceitos da biologia de sistemas em suas pesquisas e alcançando resultados relevantes. Outras duas áreas nas quais a biologia de sistemas vem encontrando bastante aplicação são redes metabólicas e expressão gênica, ambas importantes para o estudo em biologia (VOIT, 2013).

Além disso, com a evolução constante de técnicas de mensuração de alta performance, principalmente na biologia molecular, uma grande quantidade de dados de boa qualidade está sendo produzida. Esses dados, geralmente se apresentam na forma de séries temporais e possuem informações sobre o fenômeno biológico ao qual estão relacionados, fazendo com que seu processamento e interpretação sejam de grande importância. Portanto, a escolha do modelo matemático e dos métodos computacionais que estarão envolvidos em sua resolução afetará diretamente os resultados obtidos e a consequente interpretação do fenômeno biológico estudado (BORGES, 2007).

Assim, a tarefa de estimar parâmetros de modelos dinâmicos de sistemas biológicos utilizando dados de séries temporais experimentais tem se tornando altamente relevante e bastante utilizada para sistemas biológicos complexos tais como redes de regulação gênicas, redes metabólicas entre outros que apresentam comportamento não linear, tipicamente relacionados a sistemas biológicos. Esses comportamentos devem ser capturados por modelos matemáticos adequados e ajustados utilizando dados experimentais confiáveis por meio de técnicas de resolução de problemas inversos (VOIT, 2013)(BORGES, 2007).

Para realizar essa tarefa de estimativa são necessários cuidados tais como a seleção de um modelo matemático que seja capaz de representar adequadamente o fenômeno biológico estudado, realizar o pré-processamento dos dados experimentais (séries temporais) de forma cuidadosa para melhorar a busca dos parâmetros que será realizada a seguir, e escolher uma estratégia eficiente e eficaz para resolver o problema inverso de busca dos parâmetros do modelo (BORGES, 2007).

Este trabalho propõe estudos que abordam esses três aspectos. A seleção do modelo matemático, fazendo uso do Sistema S (S-system), baseado em leis de potências (power-law) e que são capazes de capturar praticamente qualquer não-linearidade, incluindo oscilações complexas, procura fazer uso de um modelo que pode se ajustar aos dados da melhor maneira possível e conseguir dessa forma retirar o máximo de informação do fenômeno biológico. Na utilização do Sistema S, uma estratégia de resolução que tem

se tornado padrão é a aproximação do sistema de equações diferenciais, que representam o sistema biológico, por um sistema de equações algébricas desacopladas, onde substitui-se a diferencial pela derivada. Essa estratégia será utilizada neste trabalho. Em relação ao pré-processamento dos dados experimentais, geralmente aplica-se uma estratégia de filtragem de ruídos e suavização dos dados. Pretende-se nessa fase, implementar e avaliar o método do passo complexo (Complex-step) para diferenciação para obtenção das derivadas, visto que esta técnica apresenta resultados com erro na ordem da precisão de máquina do computador. No que tange a resolução do problema inverso, tentativas de resolução utilizando algoritmos evolutivos tem se apresentado como uma estratégia razoável na obtenção dos parâmetros do problema inverso. Objetiva-se aqui, implementar e avaliar o método da Entropia Cruzada (Cross Entropy), técnica adaptativa para estimativa de probabilidade de eventos raros que envolve minimização de variância. Esta técnica pode ser adaptada para a otimização estocástica de problemas altamente multimodais e será aplicada na resolução do problema inverso.

Portanto, considerando o contexto de aplicação de modelos matemáticos para análise e interpretação de fenômenos biológicos, esse trabalho tem o objetivo de utilizar um entre os vários modelos existentes para utilização em um problema inverso, que se inicia com os dados de séries temporais relacionados a determinado sistema biológico e que através de uma estratégia resolução específica tem o objetivo de encontrar parâmetros para o modelo que sejam capazes de se ajustar da melhor maneira possível aos dados experimentais e dessa forma explicar o comportamento do sistema de maneira adequada. O modelo considerado para uso neste trabalho foi especificamente desenvolvido para aplicação em sistemas bioquímicos e possui características que o tornam apropriado de ser estudado em uma perspectiva de biologia de sistemas.

A seguir serão apresentados alguns conceitos biológicos com o objetivo de contextualizar a aplicação do problema estudado neste trabalho.

1.1 Biologia de Sistemas

A biologia de sistemas tenta utilizar uma abordagem a partir de uma perspectiva sistêmica, não reducionista, para fenômenos biológicos. De acordo com Kitano (2002b), o entendimento de genes e proteínas é importante na biologia, mas um sistema biológico

não é somente um conjunto de genes e proteínas, pois mapeando sua estrutura e conexões será formado apenas um diagrama estático. O que deve ser compreendido são os padrões que emergem desses relacionamentos e como eles podem ser controlados. Segundo Konopka (2007), os componentes biológicos, em vários níveis de organização, interagem uns com os outros. Essa interação pode gerar respostas ou propriedades que não podem ser explicadas por somente um dos componentes que participaram da interação e nem mesmo em somente um dos níveis de organização. Portanto, a compreensão de como essas "propriedades emergentes" surgem, porque elas se manifestam somente em células ou no organismo como um todo e quais são as consequências de suas modificações podem ser considerados como as ideias principais a serem entendidas pela biologia de sistemas.

É interessante notar que a biologia é uma ciência que já possui sólido conhecimento em áreas distintas tais como biologia molecular, sequenciamento de genomas, proteínas, biologia celular, tecidos, órgãos, organismos e ecologia (KITANO, 2002a). No entanto, uma abordagem tipicamente reducionista fez com que o conhecimento adquirido nessas áreas fosse muito específico, não levando em consideração o relacionamento entre elas. Para entender essa abordagem reducionista pode-se considerar o exemplo do estudo das células, que pode ser considerado altamente detalhado, mas não é capaz de fornecer previsões com razoável grau de certeza em relação a comportamentos de um organismo completo (KONOPKA, 2007). Assim, segundo Kitano (2002a), é necessário o desenvolvimento de um conhecimento sistêmico que seja capaz de integrar o conhecimento das várias áreas da biologia de uma forma coerente.

Os componentes biológicos e suas propriedades podem ser considerados em diversas escalas de tamanho e tempo. Considerando um exemplo para as diferenças de escala no tamanho, existem fenômenos que podem ser estudados considerando populações, indivíduos, tecidos, órgãos, células e moléculas, variando desde de metros até micrômetros. Em relação à escala de tempo, existem processos que podem durar anos até processos que duram alguns segundos ou até menos. Assim, uma visão sistêmica precisa considerar modelos genéricos que sejam capazes de integrar a representação de diversos níveis de organização (KLIPP et al., 2009).

Além disso, considerando a complexidade de grande parte dos sistemas biológicos e a quantidade de dados disponíveis atualmente, a integração dos dados experimentais com métodos computacionais deverá ser capaz de gerar inovações que poderão ser aplicadas

em várias áreas, como medicina e descoberta de fármacos (KITANO, 2002a).

1.1.1 Origem e Desenvolvimento

Apesar do conceito de biologia de sistemas estar ganhando força mais recentemente, alguns de seus conceitos e seu foco principal, que é estudar a biologia a partir de uma perspectiva sistêmica, não são novidades, pois Wiener (1948) e Bertalanffy (1969) já haviam apresentado ideias que podiam ser aplicáveis à biologia da mesma forma que é proposta pela biologia de sistemas atual. No entanto, um dos grandes problemas existentes na época era a quantidade de dados disponíveis para análise, pois áreas como a biologia molecular só foram desenvolvidas após a adoção de métodos experimentais de alta capacidade que proporcionaram a geração de uma grande quantidade de dados de boa qualidade. O desenvolvimento desses métodos experimentais e a consequente geração de grande quantidade de dados foi um dos fatores que proporcionou a oportunidade da análise da biologia a partir de uma perspectiva sistêmica (KITANO, 2002b). O aumento da capacidade computacional dos computadores modernos é outro fator que contribui para o estudo da biologia de sistemas através de métodos computacionais aplicados no processamento de dados experimentais (KLIPP et al., 2009).

Em (KONOPKA, 2007), ao escrever um dos capítulos, Eberhard O. Voit apresenta outros motivos para os conceitos que já eram existentes não serem utilizados na biologia. O forte paradigma reducionista seria um deles, fazendo com que a maioria dos cientistas continuasse a aplicar métodos já conhecidos e aproximações que faziam analogias com disciplinas da engenharia, que já eram dominadas e amplamente utilizadas na época. Outro motivo apresentado por Voit é a complexidade envolvida nos relacionamentos de componentes biológicos, que possuem interações que raramente são lineares e que quase sempre exigiam modelos mais complexos, não-lineares, para serem estudados. Em (WOLKENHAUER, 2001) também são feitas considerações similares às de Voit sobre o uso de conceitos de disciplinas como física e engenharia aplicados à biologia. O desenvolvimento de modelos e métodos que sejam capazes de representar não somente os dados observados em experimentações mas também conceitos específicos da biologia é defendido. Além disso, é realizada uma discussão sobre o trabalho de Robert Rosen (ROSEN, 1978)(ROSEN, 1985)(ROSEN, 1991)(ROSEN, 2000), que apresenta o mesmo tipo de críticas. No entanto, apesar de haver a necessidade de desenvolver modelos mais representativos para

a biologia, vários modelos derivados de outras disciplinas, como a física, conseguem representar de forma adequada aspectos de fenômenos biológicos (KLIPP et al., 2009).

Portanto, é perceptível a necessidade de busca por modelos alternativos que sejam capazes de fornecer métodos de análise matemática tão bons quanto os já existentes mas que também possam representar conceitos relacionados à fenômenos biológicos de maneira mais intuitiva, permitindo aos pesquisadores que a partir da análise dos dados mensurados experimentalmente possam ser feitas interpretações relacionados ao fenômeno biológico diretamente a partir dos parâmetros observados no modelo. Além disso, a capacidade de representar comportamentos não-lineares, característica inata de sistemas biológicos, é outra propriedade interessante a qualquer modelo que tenha o objetivo de ser aplicado à área de biologia.

1.1.2 Desafios e Dificuldades

A grande variedade de modelos matemáticos existentes para descrever diferentes fenômenos biológicos é um problema para uma abordagem sistêmica da biologia. Oscilações glicolíticas podem ser representadas por equações diferenciais ordinárias, expressões gênicas podem ser representadas por redes booleanas. Diferentes modelos e algoritmos podem ser escolhidos para representar um único fenômeno biológico, podendo cada um deles mostrar diferentes aspectos do sistema descrito. Um exemplo pode ser uma rede de reações bioquímicas, que pode ser representadas por um gráfico no qual pontos representam os metabólitos envolvidos e setas representam as reações, permitindo uma visualização de sua estrutura e relacionamento entre partes; essa mesma rede pode ser representada por um sistema de equações diferenciais que permite sua simulação e predição de seus comportamentos dinâmicos. Apesar dessas opções terem vantagens, como fornecer visões diferentes do mesmo sistema, a tarefa de integrar os modelos ou generalizá-los em um modelo mais completo fica mais difícil (KLIPP et al., 2009).

Outro problema é a construção de modelos que, além de predizer os dados de saída de um sistema a partir dos dados de entrada, sejam capazes de fornecer informações sobre a estrutura do sistema estudado e dos relacionamentos existentes entre suas partes da forma mais próxima da realidade possível. Essa tarefa não é trivial, pois geralmente os modelos matemáticos são feitos de maneira mais simples de forma que que possa ser facilmente compreendido e implementado mas não consegue realizar uma boa aproximação da rea-

lidade, ou então podem ser bastante realísticos mas altamente complexos. Nenhuma das características descritas faz um modelo bom ou ruim, mas o tornam apropriado ou não para resolver um determinado problema (KLIPP et al., 2009). Por esses motivos, a avaliação de novos modelos alternativos que permitem análise e interpretações diferentes dos modelos já existentes deve ser considerada relevante sempre que seja bem fundamentada.

Um dos grandes desafios atuais da biologia de sistemas é a integração de dados. A grande quantidade de dados gerada pelas novas tecnologias de alta capacidade precisa criar padrões de representação para facilitar a comunicação e além disso criar bons esquemas de armazenamento e compartilhamento de dados. Além disso, em um nível mais complexo, é necessário realizar o correlacionamento de dados de bases de dados diversas, por exemplo vinculando dados clínicos, genéticos, comportamentais e ambientais com tipos de informação fenotípica molecular e identificar suas associações (KLIPP et al., 2009).

Assim, apesar de ser possível a modelagem e simulação de sistemas biológicos em diferentes contextos, como células e bactérias, a construção e integração de modelos mais complexos que considerem órgãos inteiros e sua interação com componentes em outros níveis organizacionais pode ser considerada uma tarefa bem mais difícil (KITANO, 2002a) mas que se encontra em desenvolvimento (KARR et al., 2012)(JOYCE; PALSSON, 2006).

1.1.3 Aplicações

Entre as diversas aplicações da biologia de sistemas, uma das que chamam mais atenção é o estudo de doenças, pelo fato de o assunto ser de extremo interesse na área de medicina. Em (HOOD et al., 2004) é feita uma discussão que considera a doença em humanos a partir de uma abordagem sistêmica. Inicialmente são feitas considerações sobre o estudo de processos metabólicos em levedura, que forneceram informações de como aplicar conceitos de biologia de sistemas para doenças em humanos. A partir daí os conceitos foram aplicados no estudo de diversos estados de um tipo de câncer de próstata e foram apresentados bons resultados. A medicina personalizada é outro importante novo desafio para a pesquisa biomédica e que pode ser alcançada fazendo uso de conceitos de biologia de sistemas (KLIPP et al., 2009).

A expressão gênica é tradicionalmente vista como processos bioquímicos independentes, mas análises genéticas e bioquímicas mais recentes mostraram que os diversos processos envolvidos na expressão gênica podem ser altamente acoplados, influenciando uma

visão que aborda de forma sistêmica o funcionamento das células a partir da regulação gênica (KOMILI; SILVER, 2008). Em (PETER; DAVIDSON, 2011) é realizada uma discussão sobre a importância da regulação gênica, como as informações em nível de DNA podem influenciar o desenvolvimento em outros níveis de organização, como o plano corporal de um organismo, e realiza uma abordagem baseada em uma perspectiva de biologia de sistemas.

Outra área de aplicação da biologia de sistemas bastante ampla são as redes metabólicas (FILIPP, 2013)(ADIAMAH; SCHWARTZ, 2012)(BERKHOUT; BRUGGEMAN; TEUSINK., 2012)(EICHER; SNOEP; ROHWER, 2012), na qual entre diversos estudos, existem pesquisas que consideram vários níveis de organização de componentes biológicos para avaliar o metabolismo do câncer, abrindo novas oportunidades terapêuticas (FILIPP, 2013) e outras que estudam funções biológicas análogas que ocorrem em diferentes organismos e supõem que sejam implementadas por redes metabólicas também análogas, emergindo por forças relacionadas a pressão seletiva e seleção natural (BERKHOUT; BRUGGEMAN; TEUSINK., 2012).

Em (KONOPKA, 2007) e principalmente em (KLIPP et al., 2009) são encontradas várias outras aplicações da biologia de sistemas.

Iyengar (2013), além de descrever algumas aplicações, também fala sobre duas classes de modelos considerados importantes para a biologia de sistemas, sendo elas: modelos capazes de representar o comportamento do sistema através de dinâmica de sistemas, geralmente através de sistemas de equações diferenciais; e modelos capazes de representar a topologia de sistemas, geralmente através de gráficos de redes. No entanto, poucos modelos seriam capazes de fornecer as duas características ao mesmo tempo, ou seja, representar a topologia dos sistemas estudados e ao mesmo tempo permitir que seja realizada a dinâmica de sistema para o análise de seu comportamento.

É nesse ponto que a apresentação da teoria conhecida como Teoria dos Sistemas Bioquímicos (Biochemical System Theory) (SAVAGEAU, 1970) (SAVAGEAU, 1969b) (SAVAGEAU, 1969a) se torna relevante, pois uma das características mais importantes dos modelos que implementam os conceitos da TSB (Teoria dos Sistemas Bioquímicos) é exatamente a capacidade de representar a estrutura dos sistemas e ao mesmo tempo determinar seu comportamento (CHOU; VOIT, 2009)(VOIT, 2013). O formato proposto para os modelos são baseados em equações diferenciais ordinárias e fazem uso de uma

característica específica da TSB, que é o uso de leis de potência (VOIT, 2013). Como será mostrado no Capítulo 2, as leis de potência desempenham papel essencial nos modelos, pois deriva daí a capacidade de representar a estrutura do sistema a partir das equações diferenciais, que por sua vez, são capazes de definir seu comportamento. Outra característica da TSB que pode ser capaz de implementar conceitos relacionados à biologia de sistemas recebe a terminologia específica de "propriedade telescópica", termo usado somente nessa área, mas que é conhecida em outras disciplinas como "invariância de escala" (KENDAL; JØRGENSEN, 2011) (LESNE; LAGUËS, 2011), e que seria capaz de representar e relacionar sistemas em níveis de organização diferentes (SAVAGEAU, 1979), apesar de poucos estudos práticos terem sido realizados sobre o assunto (VOIT, 2013).

Portanto, considerando *(i)* a crescente importância da biologia de sistemas para a representação e entendimento de fenômenos biológicos, *(ii)* os vários modelos disponíveis para representar seus diversos componentes em diversos níveis de organização diferentes, *(iii)* a necessidade de integração dos modelos para proporcionar uma visão sistêmica do fenômeno estudado, *(iv)* a heterogeneidade na capacidade de representação e análise de cada um dos modelos e *(v)* a relevância das representações topológicas e de dinâmica de sistemas para um determinado modelo, é que se resolveu adotar o modelo Sistema S, derivado da TSB para a realização de estudos numéricos na resolução de problemas relacionados a sistemas bioquímicos, pois entre as características desse modelo estão a capacidade de representação da estrutura e do comportamento do sistema (CHOU; VOIT, 2009) e a possibilidade de realizar integração de sistemas (SAVAGEAU, 1979), ambas compatíveis com os conceitos utilizados na biologia de sistemas.

O Capítulo 2 fará uma apresentação mais detalhada da TSB e do modelo Sistema S para que sua estrutura e seu contexto de aplicação em biologia de sistemas possa ser compreendido, abrindo caminho para a posterior explicação dos métodos que serão utilizados na estratégia de resolução de problema que é utilizada neste trabalho.

1.2 Objetivos do Trabalho

A partir de séries temporais mensuradas experimentalmente para determinado sistema biológico para o qual não se conhece o sistema de equações que o representa, diversas

estratégias podem ser adotadas para resolver esse problema inverso, ou seja, encontrar um modelo (sistema de equações) capaz de representar o sistema biológico a partir dos dados das séries temporais. Considerando a escolha do modelo Sistema S para representação do processo biológico, pretende-se também adotar aqui uma estratégia de resolução utilizada em diversos outros trabalhos (CHOU; MARTENS; VOIT, 2006) (VILELA et al., 2008) (VILELA et al., 2007) (BEYER, 2008) na qual é realizado um desacoplamento do sistema de equações diferenciais para um sistema de equações algébricas (VOIT; ALMEIDA, 2004). Essa estratégia diminui o custo computacional do problema mas cria a necessidade de se realizar boas estimativas das derivadas para cada ponto das séries temporais, pois a qualidade da estimativa das derivadas afeta diretamente os resultados da etapa de otimização, utilizada mais a frente para estimar os parâmetros do sistema de equação que será encontrado para representar o processo biológico.

Sendo assim, este trabalho pretende avaliar tanto métodos que realizam a estimativa das derivadas das séries temporais com a maior precisão possível, para que a qualidade das estimativas não influenciem de modo negativo a etapa de otimização dos parâmetros do sistema, quanto métodos que realizam a otimização dos parâmetros com boa precisão e com baixo custo computacional. Especificamente nesta etapa de otimização, serão comparados três métodos, sendo um deles um método híbrido, que fará a combinação dos outros dois métodos, usando as melhores características de cada um deles, tentando gerar resultados que nenhum dos dois geraria sozinho.

1.3 Organização do Trabalho

Este trabalho é dividido em seis capítulos, sendo que o primeiro contém uma introdução e a apresentação de alguns conceitos biológicos que servem para contextualizar o problema. Os demais capítulos de dividem a maneira apresentada a seguir.

O Capítulo 2 mostra a Teoria dos Sistemas Bioquímicos (Biochemical System Theory), ou TSB, mostrando detalhadamente o desenvolvimento dessa teoria e os dois modelos derivados de seus conceitos (Sistema S e GMA). Mostra ainda que entre as várias aplicações possíveis destacam-se as redes metabólicas e as redes de regulação gênica, duas áreas de estudo importantes na biologia e que podem ter conceitos de biologia de sistemas aplicados. É explicado também porque, entre os dois modelos da TSB, o modelo Sistema S é

mais utilizado que o modelo GMA, e ainda apresenta alguns exemplos de aplicação.

No Capítulo 3 são mostradas as etapas da estratégia adotada neste trabalho para resolução de problemas inversos para sistemas que são modelados com Sistema S. Os métodos utilizados nas etapas de estimação de derivadas e otimização para busca dos parâmetros das equações também são explicados.

O Capítulo 4 apresenta os sistemas que serão utilizados neste trabalho. Esse sistemas, apesar de sintéticos, possuem propriedades semelhantes aos sistemas reais e são suficientes para uso no estudo em questão. Seu uso em trabalhos importantes na literatura de referência é uma garantia de sua eficácia como um sistema de teste. Neste capítulo também são apresentadas as implementações realizadas, assim como a descrição dos experimentos computacionais utilizados.

Os resultados obtidos nos experimentos são apresentados no Capítulo 5, juntamente com discussões relevantes.

No Capítulo 6 são feitas considerações e conclusões em relação ao trabalho desenvolvido.

2 TEORIA DOS SISTEMAS BIOQUÍMICOS

2.1 Conceitos

Como discutido no capítulo anterior, apesar de existirem modelos que são capazes de representar de forma satisfatória alguns fenômenos biológicos em contextos específicos, existem várias características desejáveis que não são encontradas na maioria dos modelos utilizados. A descrição da estrutura e do comportamento de um sistema, a capacidade de integrar sistemas em diferentes níveis de organização e a representação de comportamentos não-lineares são algumas características desejáveis em modelos que procuram ser utilizados em biologia. Por isso, novas propostas continuam a ser apresentadas e testadas.

A Teoria dos Sistemas Bioquímicos foi proposta com o objetivo de fornecer modelos que tivessem a capacidade de fornecer análises matemáticas e ao mesmo tempo fosse capaz de permitir simulações computacionais de sistemas biológicos. Apesar de inicialmente ter sido desenvolvida apenas para aplicação em bioquímica, aos poucos passou a ser utilizada em vários outros contextos relacionados à biologia (VOIT, 2013). A TSB não é um modelo propriamente dito, mas um conjunto de conceitos que podem ter um ou mais modelos relacionados, desde que sejam implementados de acordo com os conceitos que a teoria indica.

Segundo Konopka (2007), os conceitos da TSB, propostos inicialmente por Savageau (SAVAGEAU, 1970)(SAVAGEAU, 1969b)(SAVAGEAU, 1969a), possibilitam que os modelos utilizados tenham características de sistemas lineares mas ao mesmo tempo possuam a habilidade de capturar fenômenos não-lineares. Considerando o exposto no capítulo anterior em relação ao comportamento de fenômenos biológicos e a busca de modelos que sejam adequados para sua representação, essas propriedades da TSB tornam o estudo de seus modelos em aplicações relacionadas à biologia relevante para avaliação da representação de tais fenômenos.

Uma das principais características da TSB é a representação de todos os processos de determinado sistema modelado através de lei de potências (VOIT, 2013), que já mostrou

bons resultados na aplicação dos modelos para representar diversos sistemas biológicos tais como vias metabólicas, redes de regulação gênica, redes imunológicas, sinalização celular e outros (CHOU; VOIT, 2009). Uma das grandes vantagens das representações através do formalismo das leis de potências é a sua capacidade de realizar a aproximação de fenômenos não-lineares, essencial para modelagem de sistemas biológicos (SRINATH; GUNAWAN, 2010). As leis de potência serão mais exploradas na Seção 2.3.

Outra característica extremamente relevante dos modelos implementados de acordo com a TSB é que a estrutura dos sistemas biológicos estudados pode ser representada diretamente através dos parâmetros do modelo (CHOU; VOIT, 2009). Essa é outra vantagem proporcionada pela utilização da representação através de leis de potência, que permite a identificação da estrutura e das propriedades cinéticas do modelo sejam realizadas a partir somente de uma única busca pelos parâmetros do problema estudado (SRINATH; GUNAWAN, 2010).

Alguns dos modelos que são implementados de acordo com a TSB são chamados de canônicos, o que significa que a sua construção e as subsequentes análises que podem ser realizadas por eles seguem regras extremamente bem definidas (VOIT, 2013).O modelo GMA (Generalized Mass Action) é a forma de representação canônica que pode ser considerada como a mais genérica utilizada para implementar os conceitos da TSB. Todos os outros modelos podem ser considerados representações especializadas se comparados ao GMA (KONOPKA, 2007). No entanto, o modelo mais implementado em pesquisas que utilizam a TSB é o Sistema S (S-system), por possuir vantagens em relação à implementação e aplicação de métodos de otimização (CHOU; VOIT, 2009). Outras formas de implementação de modelos que utilizam os conceitos da TSB são: sistemas lineares (linear systems), sistemas Lotka-Volterra (Lotka-Volterra systems), sistemas Lotka-Volterra generalizados (generalized Lotka-Volterra systems), half-systems (KONOPKA, 2007), que não serão abordados neste trabalho.

Todas as características descritas acima relacionadas à TSB e aos principais modelos utilizados que a implementam, GMA e Sistema S, serão detalhadas ao longo desse capítulo.

2.2 Representação de Reações

Todos os elementos que fazem parte de reações bioquímicas podem ser utilizados em várias vias metabólicas que são formadas e ligadas através de reações, formando redes bioquímicas complexas. Para que esses fenômenos possam ser representados adequadamente através de modelos matemáticos é necessário adotar métodos que façam uma boa transformação dessas redes de reações bioquímicas em sistemas de equações diferencias que representem o comportamento do sistema bioquímico com a melhor aproximação possível (KONOPKA, 2007).

Sendo uma reação discreta que a partir do consumo de uma única molécula de um substrato S seja capaz de produzir também uma única molécula de um produto P, se for considerada uma quantidade grande de moléculas então um processo contínuo (em vez de discreto) que seja baseado em "taxas" pode ser formulado para essa reação, descrevendo o relacionamento de S e P através de equações diferenciais ordinárias. Para uma taxa de v conversões por segundo para a reação, as taxas de consumo e produção de S e P também serão v (KONOPKA, 2007).

Para os casos de reações que envolvem mais de dois metabólitos, considere que uma molécula do substrato A e uma molécula do substrato B sejam utilizadas para produzir uma molécula do substrato C. Para v conversões por segundo, a taxa de mudança de A e B é -v e de C é v (KONOPKA, 2007).

Agora, para a mesma reação de A, B e C descrita acima, considere a situação na qual mais de uma molécula de cada metabólito envolvido na reação seja consumido ou produzido. Tomando o exemplo da produção de uma única molécula do produto C a partir do consumo de duas moléculas do substrato A e de três moléculas do substrato B a uma taxa de v conversões por segundo, então o consumo de A nessa reação será de $2v$ por segundo, o consumo de B será de $3v$ por segundo e a produção de C será de $1v$ por segundo. No caso dessa situação na qual existe o consumo e/ou produção de mais de uma molécula de determinado metabólito em uma reação, a proporção específica na qual cada reagente ou produto (metabólitos) participa na reação é determinada estequiometria (KONOPKA, 2007).

Portanto, a partir das descrições acima é possível a criação de modelos de equações para esses tipos de reações que sejam capazes de descrever a taxa de variação dos metabólitos a partir de sua estequiometria e de suas taxas de reação (KONOPKA, 2007).

Para exemplificar o uso de equações diferenciais na modelagem de reações bioquímicas vamos considerar a rede da Figura 2.1, que possui 5 metabólitos e três reações. A substância A é consumida para que a substância B seja produzida e que por sua vez pode ser consumida para a produção de C ou D. No caso da substância B ser consumida para a produção da substância C, ainda são usadas 2 moléculas da substância E nessa reação, que são transformadas em duas moléculas da substância F. As três taxas de reação são $v_{A,B}$ (consumo de A e produção de B), $v_{B,C}$(consumo de B e produção de C) e $v_{B,D}$ (consumo de B e produção de D) (KONOPKA, 2007).

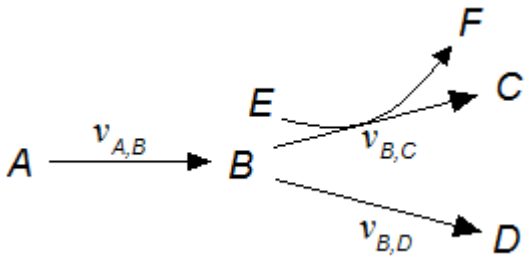


Figura 2.1: Exemplo de uma reação (KONOPKA, 2007)

Para modelar o comportamento do sistema bioquímico da Figura 2.1 são considerados a taxa de reação e a estequiometria da rede de reações, que são utilizados para gerar as equações diferenciais que representam o sistema. Se o metabólito envolvido na reação for consumido então o termo que irá representar o produto de sua estequiometria e sua taxa de reação será negativo. Caso contrário, ou seja, se o metabólito for produzido na reação, então o termo será positivo (KONOPKA, 2007). O sistema de equações diferenciais a seguir mostra a representação matemática da rede de reações bioquímicas apresentada na Figura 2.1:

$$\begin{aligned}
\frac{dA}{dt} &= -1v_{A,B} \\
\frac{dB}{dt} &= +1v_{A,B} - 1v_{B,C} - 1v_{B,D} \\
\frac{dC}{dt} &= +1v_{B,C} \\
\frac{dD}{dt} &= +1v_{B,D} \\
\frac{dE}{dt} &= -2v_{B,C} \\
\frac{dF}{dt} &= +2v_{B,C}
\end{aligned} \tag{2.1}$$

Esse sistema pode ser escrito em forma matricial de modo que as taxas de reação são usadas através de combinação lineares para representar a taxa de variação de cada metabólito. Os coeficientes dessa combinação linear são os valores da estequiometria das reações. Assim, X representa o vetor com a taxa de variação em relação ao tempo de todos os metabólitos envolvidos na reação, enquanto S é a matriz que representa a estequiometria do sistema e v é o vetor das taxas de reação (KONOPKA, 2007). A seguir está a representação do sistema baseada em álgebra linear:

$$\begin{bmatrix} dA/dt \\ dB/dt \\ dC/dt \\ dD/dt \\ dE/dt \\ dF/dt \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ +1 & -1 & -1 \\ 0 & +1 & 0 \\ 0 & 0 & +1 \\ 0 & -2 & 0 \\ 0 & +2 & 0 \end{bmatrix} \cdot \begin{bmatrix} v_{A,B} \\ v_{B,C} \\ v_{B,D} \end{bmatrix} \tag{2.2}$$

Essa matriz pode se escrita como $\mathrm{dX/dt} = \mathrm{S}.v$.

Em relação às taxas de reação de um sistema bioquímico, pode-se considerar que trata-se de um fator altamente relevante para o estudo do comportamento do sistema, pois além de estarem sempre presentes em qualquer modelo, são fundamentais na determinação de sua dinâmica (KONOPKA, 2007). Assim, um estudo um pouco mais detalhado será realizado, mas que apesar de existirem diversas formas funcionais sugeridas para representar as taxas de reação, será focado na representação que é utilizada na TSB, que será adotada neste trabalho.

As reações que ocorrem em sistemas bioquímicos normalmente envolvem diversos metabólitos que podem ser inibidos ou ativados por outros metabólitos ou proteínas; consequentemente a taxa de determinada reação dependerá, além do substrato que será consumido, da quantidade de todos as outras espécies que estarão envolvidas na reação. Essas taxas de reações e suas relações com todos os metabólitos envolvidos pode ser representado de diversas formas. De maneira genérica, pode-se representar a taxa de reação da i-ésima reação como $v_i = V_i(\mathbf{X}) = V_i(X_1, X_2, X_3, \ldots, X_n)$. Entre as formas funcionais propostas para representar as taxas de reação, a cinética de ação de massas, a cinética enzimática e as leis de potência são as principais (KONOPKA, 2007).

A representação das taxas de reação baseada em cinética de ação de massas possui sua

estequiometria na forma de um produto de funções de potência que de forma geral pode ser definido como $v = k \cdot {X_1}^{g_1} \cdot {X_2}^{g_2} \ldots {X_n}^{g_n}$ onde k é uma constante positiva chamada de constante de taxa e os g_i são valores inteiros positivos que refletem a estequiometria da reação. Considerando a reação $5A+3B \rightarrow 2C$, então a taxa de reação seria definida como $v = k \cdot A^5 \cdot B^3$. A vantagem desse tipo de representação é que o modelo pode ser determinado diretamente das reações elementares e sua estequiometria, mas sua desvantagem é que as constantes de taxa devem determinadas mas nem sempre podem ser observadas experimentalmente (KONOPKA, 2007).

O modelo que demonstra maior capacidade para representar as taxas de reações quando essas são catalizadas por enzimas (cinética enzimática) é o Michaelis-Menten (MICHAELIS; MENTEN, 1913) que pode ser derivado do modelo de ação de massas se forem feitas considerações sobre o comportamento em estado estacionário e a razão entre substratos e enzimas. Apesar do modelo fornecer informações sobre a taxa de saturação da reação e a quantidade de substrato existente quando é atingida metade do máximo dessa taxa de saturação, a maioria dos experimentos nos quais ele se aplica são *in vitro*, não prevendo situações como efeitos inibitórios ou reversibilidade, comuns em sistemas bioquímicos *in vivo* (KONOPKA, 2007).

As leis de potência representam outro modelo possível de se aplicar nas taxas de reações de sistemas bioquímicos, que por ser a representação adotada na Teoria dos Sistemas Bioquímicos, considerada neste trabalho, será um pouco mais detalhada a seguir. Essas três formas funcionais de representação das taxas de reações (cinética de ação de massas, cinética enzimática e leis de potência) são as mais utilizadas em meio a uma grande variedade de outras formas disponíveis (KONOPKA, 2007).

2.3 Lei de potências

Apesar do paradigma dos sistemas lineares possuir grande quantidade de métodos matemáticos consolidados que estão disponíveis para realização de análises, além de ter obtido sucesso na aplicação desses métodos em uma grande variedade de aplicações em diversas áreas, não é possível sua aplicação em sistemas complexos que exibem características altamente não-lineares, como no caso da maioria dos sistemas biológicos. Apesar disso, por ser consolidado e amplamente utilizado, o paradigma dos sistemas lineares pode servir

de referência na busca de técnicas não-lineares análogas, além de poder fornecer métodos que podem ser utilizados para análise em determinadas situações, mesmo não sendo a melhor representação para determinado fenômeno (SAVAGEAU, 1988). Desse modo, para permitir análises a partir de métodos já conhecidos para os modelos representados através do formalismo das leis de potência, é realizada uma linearização logarítmica nos eixos, ou seja, o modelo se torna linear em um gráfico log-log (MARIN-SANGUINO et al., 2009).

A origem da ideia para a aplicação das leis de potência vem inicialmente do trabalho de (BODE, 1945), que mostrou que funções racionais podem ser aproximadas por retas em um gráfico log-log. O que levou Savageau (SAVAGEAU, 1969b)(SAVAGEAU, 1969a) sugerir o formalismo das leis de potência para a aplicação em sistemas bioquímicos (SAVAGEAU, 1988). Apesar de ser praticamente impossível definir qual representação é melhor e mais precisa para aplicação em determinado contexto devido ao fato de não existir uma resposta exata para o balanço entre precisão e estrutura na aproximação de um modelo, o desenvolvimento das ideias de Savageau foram utilizados em vários tipos de sistemas durante praticamente 40 anos, nos quais grande quantidade de técnicas foram desenvolvidas para permitir a análise do modelo em suas aplicações (MARIN-SANGUINO et al., 2009).

Como descrito em Konopka (2007):

> A maioria das taxas de reações biológicas, quando são comparadas com as concentrações dos substratos em um gráfico com dimensões logarítmicas, apresentam amplas regiões lineares. Assim, realizar aproximações lineares para essas funções em dimensões logarítmicas fornece uma relação adequada da forma $log(v) \approx g \cdot log(X) + a$ na qual g representa a inclinação da reta na região linear e a representa a interseção com o eixo y. Quando é feita a transformação de volta para o espaço cartesiano, a relação segue uma lei de potência dada por $v = \alpha \cdot X^g$ na qual a inclinação g é o expoente do substrato e $\alpha = e^a$ é um fator de escala para o termo. De forma realística, a taxa normalmente é uma função de mais de uma espécie, e portanto é importante que a aproximação possa ser estendida para várias variáveis. Nesse caso, é necessário a aproximação de uma superfície em coordenadas logarítmicas, que é dada por um plano (no caso de suas variáveis) ou por um hiperplano (para mais de duas variáveis) e tem a forma matemática $log(v) \approx g_1 \cdot log(X_1) + g_2 \cdot log(X_2) + \cdots + g_n \cdot log(X_n) + a$. Fazendo a transformação para coordenadas cartesianas chega-se a uma lei de potências na

forma $v = \alpha \cdot {X_1}^{g_1} \cdot {X_2}^{g_2} \ldots {X_n}^{g_n}$. Essa fórmula pode ser entendida como a representação de reações usando a cinética de ação de massas na qual α é a constante de taxa e os g_i são os coeficientes cinéticos. Mas na representação de lei de potências utilizada na TSB os coeficientes cinéticos podem ter valores fracionários ou negativos, e portanto os g_i são mais conhecidos como coeficientes cinéticos aparentes. No entanto, quando é evidente na literatura que a discussão ocorre no contexto de modelos baseados em leis de potência então os g_i são chamados simplesmente de coeficientes cinéticos.

2.4 Modelos Canônicos

Modelos canônicos permitem formulações que podem ser aplicadas em determinadas classes de problema nas quais grande variedade de situações podem ser representadas. Isso é uma grande vantagem, pois ao invés de limitar a aplicação dos modelos em situações muito específicas que são definidas em relação a determinadas condições, os modelos canônicos, com sua representação bem definida, são genéricos e podem ser aplicados em diversos contextos. Sendo assim, modelos canônicos são considerados extremamente vantajosos quando o objetivo é o desenvolvimento de teorias matemáticas e métodos analíticos e computacionais, como é o caso da TSB (KONOPKA, 2007).

Além disso, é mais fácil desenvolver técnicas e métodos para análise e estimativa de parâmetros em modelos canônicos. Os dois modelos canônicos baseados em lei de potências e que são mais utilizados na Teoria dos Sistemas Bioquímicos são o GMA (Generalized Mass Action - Ação de Massas Generalizada) e Sistema S (S-system) (CHOU; VOIT, 2009).

2.4.1 GMA

Considerando novamente a forma $dX/dt = S \cdot v$ que foi descrita anteriormente, pode-se definir uma função de taxa de reação como $v_k = V_k(X_1, X_2, \ldots, X_n) = \theta_k \cdot {X_1}^{f_{k,1}} \cdot {X_2}^{f_{k,2}} \cdot \cdots \cdot {X_n}^{f_{k_n}}$, onde θ_k é a constante de taxa e $f_{k,1}, f_{k,2}, \ldots, f_{k,n}$ são os coeficientes cinéticos para as espécies $X_1, \ldots, X_n$ na k-ésima reação. Além disso, a matriz de estequiometria S pode ser dividida em duas, sendo que uma delas agrupa todos os valores positivos que representam a produção das espécies da reação (S^+) e outra agrupa todos os valores negativos que representam o consumo das espécies da reação (S^-) (KONOPKA, 2007).

Assim, em uma rede de p reações para as X_i espécies, a equação diferencial a seguir pode ser utilizada:

$$\begin{aligned}\frac{dX_i}{dt} &= \sum_{k=1}^{p} s_{i,k}^{+} \cdot v_k - \sum_{k=1}^{p} s_{i,k}^{-} \cdot v_k \\ &= \sum_{k=1}^{p} s_{i,k}^{+} \cdot \theta_k \prod_{j=1}^{n} {X_j}^{f_{k,j}} - \sum_{k=1}^{p} s_{i,k}^{-} \cdot \theta_k \prod_{j=1}^{n} {X_j}^{f_{k,j}}\end{aligned} \tag{2.3}$$

Se $s_{i,k}$ e θ_k forem combinados em um único termo $\gamma_{i,k}$ então é alcançada a forma geral mais utilizada para representar um modelo GMA, no qual $\gamma_{i,k}$ pode ser positivo em caso de produção ou negativo em caso de consumo de X_i (KONOPKA, 2007):

$$\frac{dX_i}{dt} = \sum_{k=1}^{p} \gamma_{i,k} \prod_{j=1}^{n} {X_j}^{f_{k,j}} \tag{2.4}$$

No modelo GMA, tanto a produção quanto o consumo de cada metabólito envolvido são representados em somente um termo baseado em leis de potência (CHOU; VOIT, 2009). A seguir será mostrado que o modelo Sistema S representa a produção e o consumo em termos separados.

Apesar do modelo GMA ser uma aproximação mais intuitiva em relação às reações bioquímicas, não permite cálculos algébricos de estados estacionários, que é uma condição na qual várias análises importantes (como consistência, robustez e estabilidade estrutural) relacionadas à características comuns em sistemas bioquímicos são realizadas com o objetivo de validação dos parâmetros encontrados para o modelo (VOIT, 2013). Já o modelo Sistema S possui facilidades em relação a isso. Portanto, a escolha do modelo a ser utilizado deve ser feita com cuidado e considerando os objetivos a serem alcançados no estudo e os dados disponíveis a serem analisados (CHOU; VOIT, 2009).

2.4.2 Sistema S

O outro modelo baseado nos conceitos da TSB, o Sistema S (o "S" vem de "fenômeno sinérgico e saturável" (VOIT, 2013)), possui uma particularidade em relação ao GMA: todos os fluxos de produção são agregados em um único termo que é representado através

de lei de potência e o mesmo é feito em relação aos fluxos de consumo, que também são agregados por um único termo, também representado através de lei de potência. Assim, ao contrário do GMA, que representa cada fluxo de reação separadamente em um termo, qualquer equação baseada no modelo Sistema S terá dois termos, um de produção e um de consumo, ambos baseados em lei de potência. A grande vantagem dessa forma é que os estados estacionários podem ser representados por equações algébricas, consequentemente podendo fazer uso de uma grande variedade de métodos analíticos já existentes para serem utilizados nas soluções (KONOPKA, 2007). A forma matemática do modelo Sistema S é:

$$\begin{aligned}\frac{dX_i}{dt} &= V_i^{+}(X_1, X_2, \ldots, X_n) - V_i^{-}(X_1, X_2, \ldots, X_n) \\ &= \alpha_i \prod_{j=1}^{n} {X_j}^{g_{ij}} - \beta_i \prod_{j=1}^{n} {X_j}^{h_{ij}}\end{aligned} \tag{2.5}$$

Os parâmetros α_i e β_i, que são valores não negativos, são chamados de constantes de taxa e referem-se as taxas de produção e consumo, respectivamente. Os parâmetros $g_{i,j}$ e $h_{i,j}$, que podem ser representados por valores reais, são os coeficientes cinéticos que refletem a ação de determinado metabólito X_j possuem no termo de produção ou consumo. Valores positivos significam um efeito catalizador exercido por X_j, enquanto valores negativos significam um efeito inibitório. Valores iguais a zero significam que X_j não exerce nenhum efeito (CHOU; VOIT, 2009).

A relação entre o sinal dos coeficientes cinéticos e a influência de um metabólito na produção ou consumo de outros é uma característica valiosa dos modelos GMA e Sistema S, que são baseados na TSB, pois através de uma rápida avaliação dos valores pode-se perceber a influência que os metabólitos exercem uns sobre os outros, tanto em termos de produção quanto em termos de consumo (KONOPKA, 2007).

2.4.3 Comparação da representação dos dois modelos

O exemplo a seguir serve para mostrar a diferença de representação dos dois modelos que implementam os conceitos da TSB. As Equações 2.6 e 2.7 são relacionadas aos modelos GMA e Sistema S, respectivamente, e representam o mapa da Figura 2.2. Pode-se perceber em ambos os modelos que a produção de X_1 é uma função tanto de X_6 (que é um substrato) quanto de X_3 (que é um inibidor da reação). Os parâmetros f_{13} e g_{13} deverão ter um valor

negativo em ambos os modelos pelo fato de X_3 ser um inibidor da reação (KONOPKA, 2007).

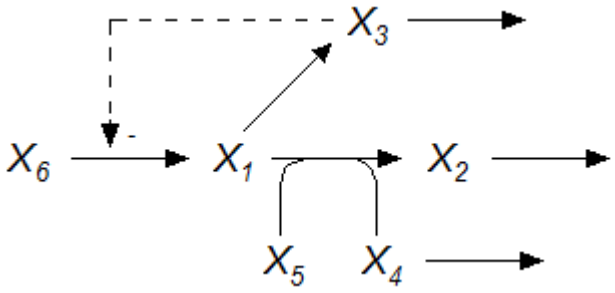


Figura 2.2: Reação para comparar modelos GMA e Sistema S (KONOPKA, 2007)

Modelo GMA:

$$\begin{aligned}
\frac{dX_1}{dt} &= \gamma_1 {X_6}^{f_{16}} {X_3}^{f_{13}} - \gamma_2 {X_1}^{f_{21}} {X_5}^{f_{25}} - \gamma_3 {X_1}^{f_{31}} \\
\frac{dX_2}{dt} &= \gamma_2 {X_1}^{f_{21}} {X_5}^{f_{25}} - \gamma_5 {X_2}^{f_{52}} \\
\frac{dX_3}{dt} &= \gamma_3 {X_1}^{f_{31}} - \gamma_6 {X_3}^{f_{46}} \\
\frac{dX_4}{dt} &= \gamma_2 {X_1}^{f_{21}} {X_5}^{f_{25}} - \gamma_4 {X_4}^{f_{64}}
\end{aligned} \tag{2.6}$$

Modelo Sistema S:

$$\begin{aligned}
\frac{dX_1}{dt} &= \alpha_1 {X_6}^{g_{16}} {X_3}^{g_{13}} - \beta_1 {X_1}^{h_{11}} {X_5}^{h_{15}} \\
\frac{dX_2}{dt} &= \alpha_2 {X_1}^{g_{21}} {X_5}^{g_{25}} - \beta_2 {X_2}^{h_{22}} \\
\frac{dX_3}{dt} &= \alpha_3 {X_1}^{g_{31}} - \beta_3 {X_3}^{h_{33}} \\
\frac{dX_4}{dt} &= \alpha_4 {X_1}^{g_{41}} {X_5}^{g_{45}} - \beta_4 {X_4}^{h_{44}}
\end{aligned} \tag{2.7}$$

Além da representação dos coeficientes cinéticos, os modelos GMA e Sistema S possuem somente uma diferença nos termos de consumo para o metabólito X_1, que mostra a agregação dos fluxos de consumo em somente um termo, conforme mostrado nos modelos teóricos (KONOPKA, 2007).

2.4.4 Vantagens dos modelos baseados em TSB

A representação em duas diferentes formas, GMA e Sistema S, para os conceitos da TSB possuem a vantagem de permitir que os modelos sejam usados de acordo com os interesses principais dos estudos realizados. No caso do GMA, sua forma é mais aproximada em relação à realidade, pois cada processo é modelado individualmente e se torna mais fácil de identificar na equação e permite maior intuição por parte do pesquisador. Para o Sistema S, as vantagens são mais relacionadas a razões matemáticas, pois o processamento e a análise dos sistemas que se encontram em estado estacionário são mais simples. Além disso, o formato do Sistema S é muito mais adequado para tarefas de otimização e identificação de estruturas a partir de dados experimentais em forma de séries temporais (KONOPKA, 2007), que faz parte da estratégia que será utilizada mais a frente neste trabalho.

Outras vantagens são comuns a ambos os modelos GMA e Sistema S, pois são mais relacionadas aos conceitos da TSB. Entre elas estão *(i)* a capacidade de capturar comportamentos não-lineares, incluindo oscilações complexas, *(ii)* a não necessidade de depender de parâmetros empíricos específicos relacionados ao modelo estudado, *(iii)* a possibilidade de realizar análises relacionadas a estados estacionários, sensitividade e a características dinâmicas do sistema e principalmente *(iv)* a capacidade de realizar o mapeamento um-para-um entre parâmetros e características estruturais do sistema, ou seja, a identificação de um parâmetro pode ser imediatamente interpretada em termos de propriedades estruturais do sistema (CHOU; VOIT, 2009).

2.5 MCA

Outra teoria, denominada Análise de Controle Metabólico (Metabolic Control Analysis - MCA), a partir da qual modelos matemáticos em outros formatos são implementados, se desenvolveu paralelamente à TSB. Apesar do formalismo matemático ser diferente, ambas estão bastante relacionadas e às vezes até mesmo complementares (MARIN-SANGUINO et al., 2009). Apesar deste trabalho utilizar apenas conceitos e modelos da TSB, será feita uma breve apresentação da MCA, pois além de sua aplicação estar sendo ampla, existem trabalhos que fazem comparações entre MCA e TSB (KONOPKA, 2007).

Kacser e Burns (1973) e Heinrich e Rapoport (1974) realizaram trabalhos independen-

tes que basicamente focava o estudo de análise de sensitividade em reações metabólicas. Suas ideias foram mais tarde unificadas e formalizadas no que se tornou a MCA, que apesar de ter surgido a partir de diferentes linhas de pensamento, possui soluções matemáticas similares às utilizadas na TSB (KONOPKA, 2007). A MCA é baseada em relações de matrizes que relacionam sensitividades locais (denominadas elasticidades) e sensitividades globais (denominados coeficientes de controle e resposta) e sua formulação permite uma interpretação biológica direta a partir das equações, que possuem variáveis que representam propriedades biológicas (como enzimas, metabólitos, efetores, etc). No entanto, o formalismo não se preocupa com a dinâmica comportamental dos sistemas biológicos estudados e quando as suposições sobre esse sistema são alteradas, a forma básica das equações também precisa ser modificada (MARIN-SANGUINO et al., 2009).

Em comparação, o formalismo relacionado aos conceitos da TSB são altamente direcionados à análise dos comportamentos dinâmicos do sistema a partir da observação do modelo. A classificação das variáveis, por sua vez, é mais direcionada às suas funções matemáticas. Tomando como exemplo as enzimas, que na MCA possui variáveis específicas, na TSB elas podem ser variáveis independentes (se forem constantes) ou variáveis dependentes (se forem variáveis, considerando sua síntese ou degradação). Finalmente, a formulação básica dos modelos na TSB também permanece consistente mesmo com mudanças feitas nas suposições sobre o sistema em estudo (MARIN-SANGUINO et al., 2009).

A MCA surgiu a partir da necessidade de complementar as explicações de observações experimentais, principalmente relacionadas ao controle realizado por enzimas em sistemas bioquímicos. Parâmetros foram definidos como propriedades que poderiam ser mensuradas em laboratório, sendo formulado portanto, um relacionamento muito bem definido entre teoria e aplicação. Em contraste com o desenvolvimento prático da MCA, os conceitos da TSB surgem a partir da necessidade de desenvolvimento de métodos quantitativos de análise para exploração do comportamento estático e dinâmico, análise de modelos, definição de princípios relacionados a sistemas biológicos e desenvolvimento de métodos de otimização. Portanto, apesar de as formulações matemáticas da MCA e da TSB poderem ser relacionadas e complementares, as questões levantadas por ambas geralmente são diferenciadas (KONOPKA, 2007).

2.6 Aplicações

Atualmente, os modelos que implementam os conceitos da TSB já são utilizados em vários contextos diferentes na biologia, mas entre duas categorias distintas de aplicação que podem ser diferenciadas uma é a análise de modelos, que tem o objetivo de descobrir novas interpretações específicas de um sistema estudado, e a outra é o desenvolvimento de métodos genéricos que podem ser utilizados na análise de vários tipos de sistemas biológicos (VOIT, 2013).

Um grande número de estudos que utilizam TSB tem focado na análise de dados experimentais para avaliar propriedades cinéticas ou concentrações metabólicas, fluxos ou outras características quantitativas do sistema em estudo (VOIT, 2013). A seguir são apresentadas algumas aplicações da TSB em diversas áreas da biologia, incluindo vários níveis de organização diferentes. Todos os exemplos de aplicações a seguir foram retiradas de (VOIT, 2013).

2.6.1 Microbiologia

As primeiras análises numéricas realizadas utilizando a BST são relacionadas à fermentação de levedura (VOIT; SAVAGEAU, 1982a), que tiveram seus modelos detalhados rapidamente por novos trabalhos. O modo como a cinética enzimática tradicional pode ser convertida para modelos baseados em lei de potência que oferecem novas informações para análise foi um dos resultados dos estudos.

A bactéria Lactococcus lactis recebeu a atenção de vários trabalhos (SRINATH; GUNAWAN, 2010)(VOIT et al., 2006)(CRAMPIN; MCSHARRY; SCHNELL, 2007)(GOEL; CHOU; VOIT, 2008)(VOIT et al., 2009)(VOIT; NEVES; SANTOS, 2006) que realizaram análises de séries temporais derivadas de estudos que usaram Ressonância Magnética Nuclear (Nuclear magnetic resonance - NMR) no organismo *in vivo*, além da estimativa de parâmetros para modelagem de reações metabólicas. A bactéria recebeu atenção por ser utilizada comumente na indústria para produção de iogurtes, queijos e outros produtos alimentícios.

A TSB também tem sido amplamente utilizada em uma grande variedade de sistemas de transdução de sinais, que existe em praticamente qualquer organismo (VERA et al., 2007a)(NIKOLOV et al., 2010)(VERA et al., 2008)(VERA; WOLKENHAUER,

2009)(VERA et al., 2008)(VERA et al., 2010)(LAI et al., 2009)(NIKOLOV et al., 2009).

2.6.2 Plantas e animais

Modelos baseados em Sistema S já foram utilizados com sucesso para simular florestas (MARTIN, 1997), otimização do regime de alimentação de polvos em cativeiro (HORMIGA et al., 2010) e simulação vias glicolíticas e glicogenolítica em fígado de ratos (TORRES, 1994). Modelos de fluxo de magnésio em florestas tropicais, considerando plantas, animais e ingredientes do solo como variáveis também já foram realizados (TORRES, 1996).

2.6.3 Fisiologia humana e doenças

O metabolismo humano tem recebido cada vez mais atenção, principalmente devido ao numero cada vez maior de dados disponíveis para análise. Modelos de dinâmica metabólica de células vermelhas do sangue, metabolismo e respostas de sinalização em células vasculares, entro outros, são algumas das aplicações (NI; SAVAGEAU, 1996b)(NI; SAVAGEAU, 1996a).

Outro estudo relacionado à metabolismo, mas que pode ser aplicado não somente em humanos mas em mamíferos de uma forma geral, é a aplicação do modelo Sistema S no metabolismo de purinas, que são substâncias relacionadas a algumas doenças mentais (CURTO et al., 1998)(CURTO; VOIT; CASCANTE, 1998)(CURTO et al., 1997). Na mesma linha, também podendo se aplicar a mamíferos, modelos que descrevem a dinâmica do metabolismo de dopaminas também foram desenvolvidos (VOIT; QI; KIKUCHI, 2002) (WU; QI; VOIT, 2010)(MARIN-SANGUINO; MENDOZA, 2008) (QI et al., 2011) (QI; MILLER; VOIT, 2008b) (QI; MILLER; VOIT, 2008a) (QI; MILLER; VOIT, 2009) (QI; MILLER; VOIT, 2010a) (QI; MILLER; VOIT, 2010b) (DULAM-BANAWA; MARIN-SANGUINO; MENDOZA, 2010) (MARIN-SANGUINO; ROSARIO; MENDOZA, 2009). A dopamina é um neurotransmissor envolvido no sistema de recompensa de mamíferos e cujas concentrações, caso alteradas, podem provocar esquizofrenia, desordens no sono, déficit de atenção e até doença de Parkinson.

Outras doenças também estudadas através de aplicação da TSB são esclerose múltipla, para a qual foram utilizados modelos para identificar possíveis gatilhos para a doença (BROOME; COLEMAN, 2011), e câncer, para o qual foi estudado a dinâmica regulatória

da via de sinalização do p53, que é crucial na supressão de tumores (LIU et al., 2010). No último caso, a simulações permitiram a identificação de moléculas importantes na via de sinalização estudada.

Doenças infecciosas também já foram estudadas através da aplicação da TSB. A eficácia antimicrobiana de drogas em humanos já foi modelada utilizando Sistema S (BERG; VOIT; WHITE, 1996), assim como identificação de alvos potenciais de droga em casos de disfunção enzimática (VERA et al., 2007b), a dinâmica da leishmaniose (LANGER et al., 2012) e modelos para a tuberculose (MAGOMBEDZE; MULDER, 2013).

2.7 Considerações finais

Os sistemas biológicos são altamente modulares e organizados em diferentes níveis que possuem características específicas e são controlados de forma hierárquica. O entendimento de módulos funcionais específicos está altamente relacionado com o entendimento do sistema como um todo. A área de pesquisa de biologia de sistemas está em constante expansão e entre seus objetivos encontram-se *(i)* a criação de modelos que podem ser aplicados em células ou organismos inteiros, podendo ser utilizados em diferentes áreas de pesquisa relacionadas à biologia, como metabolismo, desenvolvimento de fármacos e desenvolvimento de simulações personalizadas para doenças, e também *(ii)* a descoberta de princípios operacionais que explicam determinadas estruturas ou processos de fenômenos biológicos (VOIT, 2013).

Se avaliado de forma cuidadosa, pode-se perceber que as metas da biologia de sistemas não diferem muito das metas da TSB de "fornecer explicações para o comportamento de sistemas em larga escala" (VOIT, 2013). Na verdade, tanto a biologia de sistemas quanto a TSB estão associadas ao desenvolvimento de modelos de sistemas biológicos (KONOPKA, 2007). Mas enquanto a biologia de sistemas é mais voltada para conceitos gerais como organização em vários níveis, integração de módulos de sistemas, interpretação intuitiva de modelos e não possui modelos específicos, a TSB foi originalmente idealizada para sistemas bioquímicos e possui um formalismo matemático que define modelos específicos que devem ser utilizados.

Como foi mostrado no último capítulo, os modelos baseados em TSB são capazes de implementar características relacionadas a alguns conceitos da biologia de sistemas, como

representar a estrutura do sistema de forma mais intuitiva e que possa se relacionada ao fenômeno biológico diretamente através da observação do modelo, capacidade de realizar análise do comportamento dinâmico do sistema (VOIT, 2013)(CHOU; VOIT, 2009), e a possibilidade de relacionar sistemas em níveis de organização diferentes (SAVAGEAU, 1979).

Portanto, a TSB será utilizada neste trabalho, através do modelo Sistema S, em uma estratégia de resolução de problema inverso que se inicia a partir de dados experimentais de séries temporais e faz uso do modelo para estimativa de seus parâmetros, para encontrar o sistema de equações que melhor representa o sistema biológico em estudo e cujos dados estão nas séries temporais. O capítulo seguinte fará uma explicação detalhada da estratégia de resolução e como o Sistema S é utilizado para se chegar ao melhor resultado possível para os parâmetros do sistema e a representação do sistema estudado.

3 ESTRATÉGIA DE RESOLUÇÃO DO PROBLEMA INVERSO

O Capítulo 1, Seção 1.1 apresentou a biologia de sistemas e serviu para contextualizar a aplicação da matemática a fenômenos biológicos, mostrando que vários modelos podem ser utilizados para representar diferentes tipos de sistemas. O Capítulo 2 apresentou a Teoria dos Sistemas Bioquímicos e seus modelos derivados na Seção 2.4, entre eles o Sistema S, que através do uso de lei de potências pode representar características não-lineares de sistemas biológicos complexos, além de possuir a grande vantagem de poder identificar a estrutura do sistema estudado a partir dos parâmetros do modelo. Assim, tem-se o contexto da aplicação da matemática a fenômenos biológicos e o uso de um dos possíveis modelos de representação, o Sistema S, derivado da TSB, para entendimento de estrutura e comportamento de sistemas complexos não-lineares.

Agora pretende-se entender como, a partir de dados de séries temporais, pode ser adotada uma estratégia de resolução de um problema inverso que pode ser computacionalmente simples de ser processada e que pode obter os parâmetros relacionados ao modelo Sistema S, definindo as equações que representam o sistema estudado.

3.1 Descrição do Problema Inverso

Na estratégia de resolução do problema inverso, os seguintes tópicos são de grande importância e precisam ser bem entendidos:

- uso de séries temporais
- filtragem dos dados das séries temporais
- utilização do modelo Sistema S para representar o fenômeno estudado
- desacoplamento do sistema de equações diferenciais ordinárias para um sistema de equações algébricas e estimativa das derivadas em cada ponto de uma série temporal

- estimativa dos parâmetros do modelo Sistema S a partir das equações algébricas utilizando métodos de otimização

Como já dito anteriormente, uma grande quantidade de dados está sendo gerada atualmente por causa dos avanços de métodos experimentais tais como espectrometria de massa e ressonância magnética nuclear. A evolução desses métodos está permitindo que esses dados sejam obtidos não somente em grande quantidade mas também com boa qualidade. Grande parte desses dados podem ser visualizados a partir de séries temporais, principalmente relacionados a reações metabólicas e à proteômica, e geralmente são chamados de perfis metabólicos, que consistem basicamente de mensurações de vários metabólitos que são feitas em intervalos de tempo regulares em determinado contexto biológico, como uma célula ou um organismo. Esses perfis metabólicos, por terem sido retirados da mensuração de determinado fenômeno biológico, possuem informações sobre sua dinâmica e sua estrutura. Para que essas informações passem a ser conhecidas, é necessário utilizar modelos matemáticos adequados para representar o fenômeno e métodos computacionais eficazes no processamento dos dados envolvidos (VOIT; ALMEIDA, 2004).

A Figura 3.1 mostra quatro perfis metabólicos para determinado fenômeno biológico. Para que informações relevantes possam ser apropriadamente conhecidas, é necessário determinar os parâmetros corretos para o modelo matemático utilizado para representar esse sistema biológico.

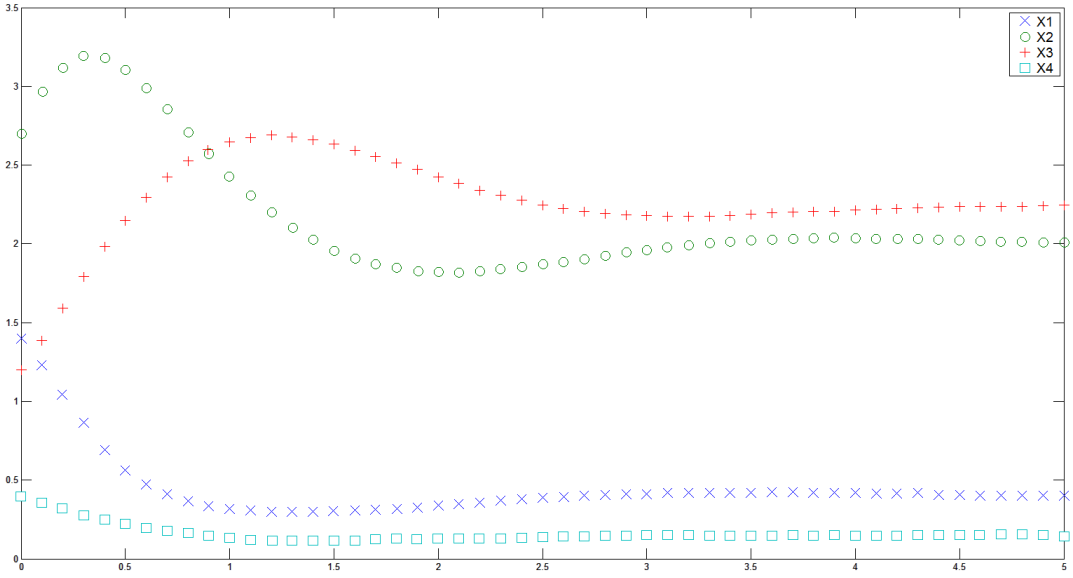


Figura 3.1: Séries temporais para um sistema biológico com quatro metabólitos e 50 mensurações realizadas em intervalos regulares

O problema inverso pode ser descrito usando como base os dados da Figura 3.1, pois a obtenção de dados de séries temporais é o ponto de partida para a estratégia de resolução adotada neste trabalho. A partir desses dados, pode-se fazer a seguinte pergunta: qual equação ou sistema de equações é capaz de descrever com precisão esses dados? Ou seja, qual modelo matemático, cujos parâmetros sejam adequadamente definidos, é capaz de representar os dados originais das séries temporais, fornecendo a oportunidade adicional de obter outras informações sobre esse fenômeno? Encontrar uma representação matemática para os dados da Figura 3.1 é a ideia principal da resolução desse problema inverso. As etapas seguintes na estratégia de resolução desse problema serão detalhadas a seguir.

Uma dessas etapas, a filtragem dos dados dessas séries temporais, é um importante aspecto na resolução do problema inverso, pois realiza a diminuição do ruído existente e permite a obtenção de curvas mais suaves para se ajustarem aos dados. Portanto, o pré-processamento dos dados das séries temporais pode ser considerado altamente relevante na estratégia de resolução escolhida para o problema inverso, e como será visto mais adiante, possui grande influencia nos resultados devido à necessidade do uso de estimativas para as derivadas em cada ponto, o que depende da qualidade do filtro realizado nas séries temporais (BORGES, 2007).

Com os dados das séries temporais filtrados e tendo conhecimento que o modelo utilizado para representar o fenômeno biológico será o Sistema S, já descrito na Seção 2.4.2, o problema inverso pode ser resolvido, mas não pode ser considerado um problema trivial por se tratar de funções não-lineares e ser necessária a otimização iterativa com uso de integração numérica das equações diferenciais, o que torna o custo computacional da resolução do problema extremamente alto (VOIT; ALMEIDA, 2004). Para os dados mostrados na Figura 3.1, o sistema de equações diferenciais (baseado no modelo Sistema S) capaz de representar esses dados estaria de acordo com:

$$
\begin{aligned}
\frac{dX_1}{dt} &= \alpha_1 {X_1}^{g_{11}} {X_2}^{g_{12}} {X_3}^{g_{13}} {X_4}^{g_{14}} - \beta_1 {X_1}^{h_{11}} {X_2}^{h_{12}} {X_3}^{h_{13}} {X_4}^{h_{14}} \\
\frac{dX_2}{dt} &= \alpha_2 {X_1}^{g_{21}} {X_2}^{g_{22}} {X_3}^{g_{23}} {X_4}^{g_{24}} - \beta_2 {X_1}^{h_{21}} {X_2}^{h_{22}} {X_3}^{h_{23}} {X_4}^{h_{24}} \\
\frac{dX_3}{dt} &= \alpha_3 {X_1}^{g_{31}} {X_2}^{g_{32}} {X_3}^{g_{33}} {X_4}^{g_{34}} - \beta_3 {X_1}^{h_{31}} {X_2}^{h_{32}} {X_3}^{h_{33}} {X_4}^{h_{34}} \\
\frac{dX_4}{dt} &= \alpha_4 {X_1}^{g_{41}} {X_2}^{g_{42}} {X_3}^{g_{43}} {X_4}^{g_{44}} - \beta_4 {X_1}^{h_{41}} {X_2}^{h_{42}} {X_3}^{h_{43}} {X_4}^{h_{44}}
\end{aligned}
\tag{3.1}
$$

No entanto, é possível substituir o sistema de equações diferenciais acopladas por um sistema de equações algébricas desacopladas se forem conhecidas (ou estimadas) as derivadas em cada um dos pontos das séries temporais (VOIT; ALMEIDA, 2004)(VOIT; SAVAGEAU, 1982b). Assim, os termos à esquerda da Equação 3.1 seriam conhecidos em cada um dos pontos e o problema passa a ser encontrar a estimativa dos parâmetros dos termos à direita em cada uma das equações. A etapa de desacoplamento do sistema será mostrada na Seção 3.2 enquanto a etapa de estimativa das derivadas é mostrada na Seção 3.3.

Voltando à questão da filtragem dos dados, agora fica claro que, como dito anteriormente, a qualidade do filtro realizado nas séries temporais possui grande influencia nos resultados obtidos (BORGES, 2007). Considerando que na estratégia de resolução adotada, o desacoplamento do sistema de equações diferenciais ordinárias acopladas para um sistema de equações algébricas desacopladas depende do conhecimento das derivadas em cada ponto das séries temporais, considerando também que a obtenção da estimativa das derivadas depende da configuração dos pontos das séries temporais, então a filtragem que será realizada no pré-processamento dos dados (que modifica a configuração/disposição dos pontos) pode alterar os resultados obtidos na estimativa das derivadas. Consequentemente, a filtragem dos dados se torna extremamente importante para a definição da precisão dos resultados finais que serão obtidos na solução do problema. Portanto, como indicado em Chou, Martens e Voit (2006), o desacoplamento torna a estimativa das derivadas uma etapa crucial na estratégia de resolução do problema inverso.

Nesse ponto, tendo os dados das séries temporais filtrados, é feita a escolha do modelo Sistema S para representar os dados e as derivadas estimadas em cada ponto das séries temporais gerando um sistema de equações algébricas desacopladas.

Finalizando a estratégia, o sistema de equações algébricas desacoplado pode ser utilizado para realizar a estimativa dos parâmetros do modelo Sistema S que é usado para representar o fenômeno biológico (Equação 3.1), ou seja, os parâmetros α_i, g_{ij}, β_i e h_{ij} devem ser encontrados para que a equação fique em uma configuração similar à da Equação 2.7, que representa o sistema da Figura 2.2. Isso é feito a partir das séries temporais, considerando cada equação separadamente (ou seja, para cada metabólito), em cada instante de tempo t_k em que se tem medições dos metabólitos. São considerados o valor estimado da derivada em cada ponto e as concentrações de cada metabólito no instante

t_k para estimar os parâmetros α_i, g_{ij}, β_i e h_{ij} para a equação do metabólito (CHOU; MARTENS; VOIT, 2006). As seções 3.2 e 3.4 apresentam mais detalhes em relação ao desacoplamento e a otimização para estimativa das derivadas.

Uma grande diversidade de algoritmos pode ser utilizada para realizar essa estimativa dos parâmetros do sistema. Entre as opções adotadas na literatura existem algoritmos baseados em gradiente e vários algoritmos de busca estocástica, como algoritmos genéticos, simulated annealing, colônia de formigas e enxame de partículas (CHOU; VOIT, 2009).

Portanto, para a estratégia de solução adotada para o problema inverso, considerando um perfil metabólico com várias séries temporais (uma para cada metabólito), e o modelo Sistema S para representar o fenômeno biológico, pode-se usar a ideia do desacoplamento de um sistema de equações diferenciais para um sistema de equações algébricas, realizando a filtragem dos dados e a estimativa dos parâmetros para se chegar a solução final do problema, representada por um sistema de equações diferenciais ordinárias, sendo uma equação diferencial para cada um dos metabólitos envolvidos no sistema.

A seguir, serão mostrados os métodos utilizados em cada uma das etapas da resolução proposta para o problema inverso. Além disso, será descrita a utilização de um método Híbrido para a etapa de otimização que apresenta características de um método de busca global que vem sendo apresentado como alternativa em otimização recentemente (Entropia Cruzada (RUBINSTEIN; KROESE, 2004)) com um método de busca local específico considerado um dos melhores na literatura para aplicação em Sistema S (Regressão Alternada (CHOU; MARTENS; VOIT, 2006)).

3.2 Desacoplamento do sistema de equações diferenciais

Considerando $\frac{dX}{dt} = f(X)$, e $X(t_0) = X_0$, onde X é um vetor de variáveis e X_0 são as condições iniciais, então a derivada em um instante t_k ($k = 0,..,N$ sendo N o tempo final da medição) é igual a $f(X)$ em t_k, ou seja, $\frac{dX}{dt}(t_k) = f[X(t_k)]$. Essa equação diferencial pode ser substituída por um conjunto de equações em cada um dos pontos t_k, que podem ser consideradas as derivadas $S(t_k)$ como (VOIT; ALMEIDA, 2004):

$$S(t_k) \approx \frac{dX}{dt}|_{t_k} = f[X(t_k)] \tag{3.2}$$

Assim, um sistema com n metabólitos com medições em N pontos terá n x N derivadas $S_i(t_k)$. Portanto, o problema inverso que inicialmente era representado por n equações diferenciais agora é representado por um sistema de n x N equações algébricas na forma a seguir (VOIT; ALMEIDA, 2004):

$$\begin{aligned}
S_i(t_1) &\approx \alpha_i \prod_{j=1}^{n} X_j(t_1)^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j(t_1)^{h_{ij}}, \\
S_i(t_2) &\approx \alpha_i \prod_{j=1}^{n} X_j(t_2)^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j(t_2)^{h_{ij}}, \\
&\bullet \\
&\bullet \\
&\bullet \\
S_i(t_k) &\approx \alpha_i \prod_{j=1}^{n} X_j(t_k)^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j(t_k)^{h_{ij}}, \\
&\bullet \\
&\bullet \\
&\bullet \\
S_i(t_N) &\approx \alpha_i \prod_{j=1}^{n} X_j(t_N)^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j(t_N)^{h_{ij}}
\end{aligned} \tag{3.3}$$

O desacoplamento realizado nessa estratégia de resolução do problema inverso permite que as variáveis $X_i(t_k)$, que são as concentrações de cada metabólito em determinado instante de tempo, e $S_i(t_k)$, que são as derivadas de cada metabólito em cada instante de tempo (que devem ser estimadas a partir dos dados das séries temporais), sejam conhecidas para um conjunto finito de pontos discretos. Portanto, os parâmetros α_i, g_{ij}, β_i e h_{ij} passam a ser os valores que devem ser encontrados na Equação 3.3 (VOIT; ALMEIDA, 2004) e vários métodos de otimização podem ser utilizados para esse propósito.

As Figuras 3.2 e 3.3 mostram a interpretação gráfica para o desacoplamento do sistema de equações diferenciais. Nesse exemplo serão considerados 4 metabólitos e 10 instantes de tempo. A interpretação gráfica para o desacoplamento será feita considerando somente o metabólito X_2 (em verde).

Sendo o exemplo da Figura 3.2 um sistema biológico com quatro metabólitos para o

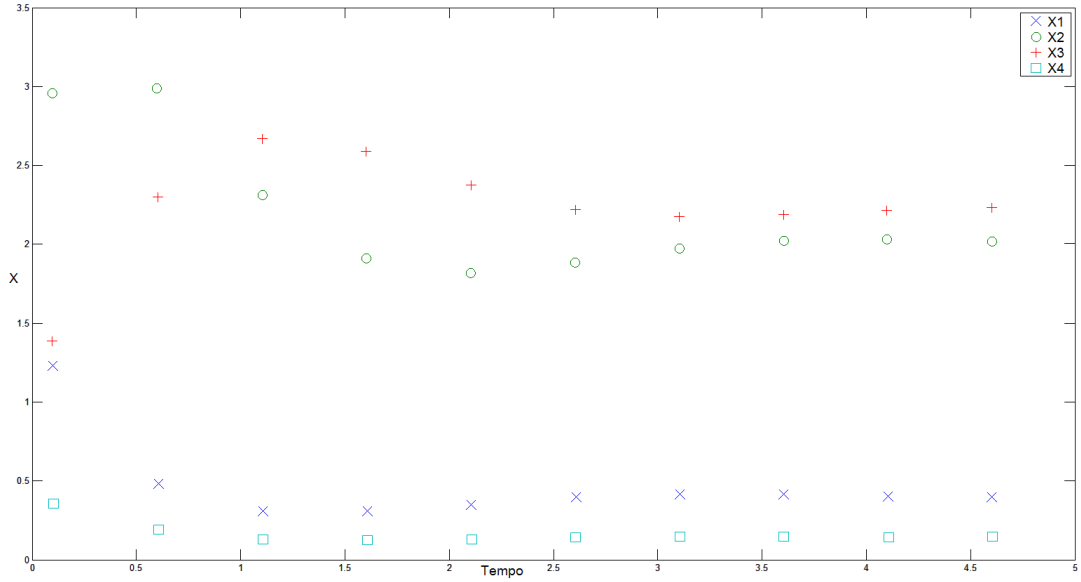


Figura 3.2: Séries temporais de quatro metabólitos de um sistema biológico

qual foram feitas mensurações experimentais que resultaram em quatro séries temporais (uma para cada metabólito), então as séries temporais são consideradas separadamente para cada equação e é realizada a estimativa da derivada em cada ponto (instante de tempo) para cada série. As derivadas podem ser estimadas através do uso de métodos numéricos.

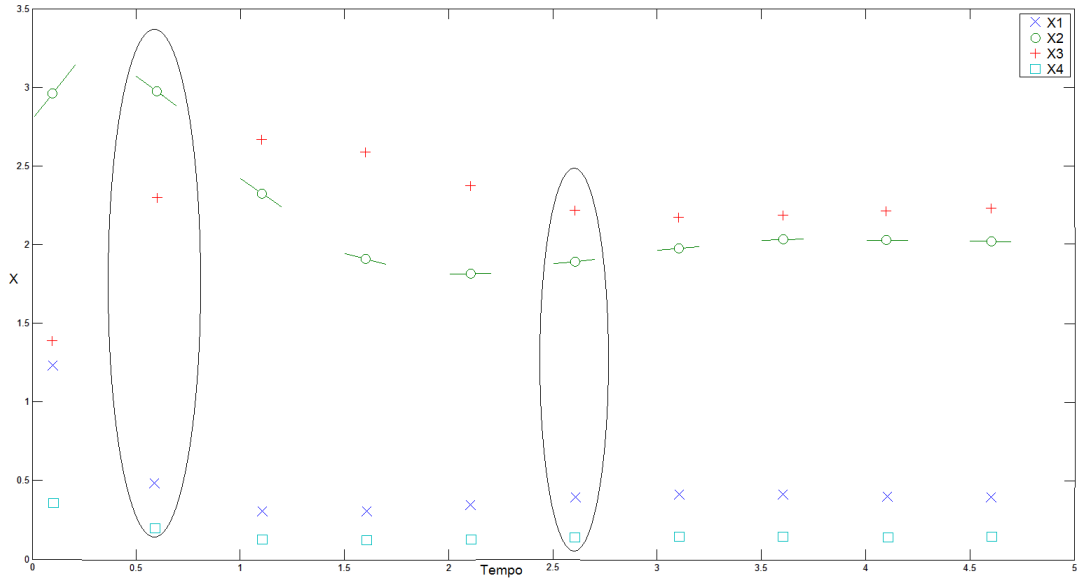


Figura 3.3: Estimativa das derivadas para cada ponto da série temporal do metabólito X2.

Após ser realizada a estimativa das derivadas em cada ponto para o metabólito X_2, o desacoplamento pode ser visualizado na Figura 3.3 e faz uso da Equação 3.2 em cada

instante de tempo. Por exemplo: para a estimativa da derivada do metabólito X_2 em t_1 e considerando as concentrações de todos os metabólitos no instante t_1 (circundados pela primeira elipse), teremos:

$$S_2(t_1) \approx \alpha_2 X_1(t_1)^{g_{21}} X_2(t_1)^{g_{22}} X_3(t_1)^{g_{23}} X_4(t_1)^{g_{24}} - \beta_2 X_1(t_1)^{h_{21}} X_2(t_1)^{h_{22}} X_3(t_1)^{h_{23}} X_4(t_1)^{h_{24}} \tag{3.4}$$

Da mesma forma, para a estimativa da derivada do metabólito X_2 em t_5 e considerando as concentrações de todos os metabólitos no instante t_5 (circundados pela segunda elipse) então teremos:

$$S_2(t_5) \approx \alpha_2 X_1(t_5)^{g_{21}} X_2(t_5)^{g_{22}} X_3(t_5)^{g_{23}} X_4(t_5)^{g_{24}} - \beta_2 X_1(t_5)^{h_{21}} X_2(t_5)^{h_{22}} X_3(t_5)^{h_{23}} X_4(t_5)^{h_{24}} \tag{3.5}$$

Como todos os valores de concentrações são conhecidos para cada um dos metabólitos em cada um dos instantes ($X_i(t_k)$) e foi realizado a estimativa das derivadas para o metabólito X_2 em cada um dos instantes ($S_2(t_k)$), sendo $k = 0,..,\ N$, então trata-se de encontrar os valores adequados para os parâmetros α_2, g_{21}, g_{22}, g_{23}, g_{24}, β_2, h_{21}, h_{22}, h_{23} e h_{24} para o metabólito X_2. Considerando todos os pontos da série temporal de determinado metabólito, pode-se usar um algoritmo de otimização direta para determinar os valores dos parâmetros e se chegar à forma final da equação para o metabólito.

Os mesmos conceitos se aplicam para encontrar os valores dos parâmetros dos outros três metabólitos do sistema da Figura 3.2.

3.3 Filtragem de dados e estimativa das derivadas

Conforme discutido anteriormente, o modelo conhecido como Sistema S possui características que podem ser consideradas vantajosas para a representação de sistemas bioquímicos, podendo permitir ao pesquisador realizar análises que relacionem os parâmetros existentes nas equações matemáticas às características do sistema em estudo. A representação do sistema através desse modelo, como já mostrado, é feita através de um sistema de equações diferenciais, sendo que existem estratégias (como a mostrada na Seção anterior) que proporcionam, para cada metabólito envolvido, a substituição da sua equação diferencial

pela estimativa da derivada em cada ponto para os dados que foram gerados a partir das observações experimentais. Dessa forma, a equação diferencial para o metabólito é substituído por uma equação algébrica equivalente, e consequentemente, o sistema de equações diferenciais é substituído por um conjunto de equações algébricas desacoplado, o que proporciona vantagens na resolução do problema original (VOIT; ALMEIDA, 2004).

Sendo assim, para a resolução de sistemas bioquímicos utilizando Sistema S, podemos considerar duas tarefas importantes para a fase de pré-processamento dos dados experimentais: a primeira é a filtragem dos dados, minimizando dessa forma os ruídos existentes na obtenção das séries. A segunda tarefa importante é a utilização de métodos numéricos adequados para se conseguir com a maior precisão possível a estimativa da derivada em cada ponto para os dados existentes, sendo a qualidade dessa estimativa fundamental no uso de Sistema S com sistemas bioquímicos (CHOU; MARTENS; VOIT, 2006).

Em relação à filtragem dos dados, será utilizado o filtro de Savitzky-Golay, amplamente conhecido e utilizado com sucesso em várias áreas (SAVITZKY; GOLAY, 1964). Para a estimativa das derivadas, existem duas maneiras possíveis, sendo uma delas a geração de equações analíticas que se ajustam aos dados (e o consequente cálculo da derivada) e a outra o uso de métodos numéricos que fazem uso dos dados para estimar a derivada em cada ponto. As duas estratégias (analítica e numérica) serão utilizadas, viabilizando uma comparação dos resultados para determinar as vantagens no uso de cada um dos modelos. Além disso, como serão utilizados dois métodos numéricos (método da diferença finita central e método do passo complexo), então será feita uma comparação entre os dois para, da mesma forma, determinar as vantagens no uso de cada um deles em conjunto com a geração analítica na filtragem dos dados.

3.3.1 O filtro Savitzky-Golay

O método de Savitzky-Golay faz simultaneamente a filtragem e a regressão de dados que originalmente contenham ruídos (SAVITZKY; GOLAY, 1964). Em relação à filtragem que é realizado pelo algoritmo, as principais vantagens são relacionadas ao fato de se evitar a ocorrência de atrasos, não ocorrendo deslocamento de picos (máximos), e a facilidade de lidar com dados que possuam pontos faltando (BORGES, 2007). O algoritmo realiza o ajuste de um polinômio de ordem p em uma janela local w dos dados experimentais, sendo que a janela w deve possuir pontos suficientes para a realização do ajuste do polinômio.

O conceito de janela que se move é utilizado na implementação do algoritmo, sendo que a primeira janela escolhida possui os pontos iniciais e a última janela possui os pontos finais. O algoritmo realiza o deslocamento continuo da janela (que altera os pontos utilizados a cada momento) dos primeiros aos últimos pontos dos dados experimentais (BORGES, 2007). Se o número de pontos na janela for maior que o número mínimo necessário para o ajuste do polinômio então a solução é obtida a partir de uma aproximação por mínimos quadrados (BROMBA; ZIEGLER, 1981).

O filtro é desenvolvido escolhendo uma janela com w pontos (adotando-se sempre w ímpar) e usar coordenadas locais e simétricas para os pontos da janela definida como:

$$i = -n_e, ..., -2, -1, 0, 1, 2, ..., n_d \tag{3.6}$$

que é relacionado aos pontos da série que pertencem à janela $y_{n_e}, ..., y_{n_d}$, onde $w = n_e + n_d + 1$ e n_e e n_d são respectivamente a distância à esquerda e à direita do centro da janela. Normalmente n_e é igual a n_d, ou seja, a janela possui mesmo tamanho para os dois lados. Portanto, o algoritmo ajusta em $i = 0$ os pontos $y_{n_e}, ..., y_{n_d}$ ao polinômio de ordem p dado por:

$$a_0 + a_1 i + a_2 i^2 + ... + a_p i^p \tag{3.7}$$

gerando uma matriz A dada por:

$$A_{ij} = i^j, \text{ para } i = -n_e, ..., n_d \text{ e } j = 0, ..., p \tag{3.8}$$

Os coeficientes do polinômio de ordem p são obtidos minimizando o erro dos mínimos quadrados:

$$min(Aa - y)^2 \tag{3.9}$$

gerando a equação dada por:

$$(A^T A)a = A^T y \tag{3.10}$$

que quando resolvida produz os valores de coeficientes polinomiais:

$$a = (A^T A)^{-1}(A^T y) \tag{3.11}$$

A derivada dos dados filtrados pode ser obtida facilmente através da derivação do polinômio ajustado para a janela corrente. A principal desvantagem do método Savitzky-Golay é a dificuldade de trabalhar com os valores limítrofes (iniciais e finais) dos dados experimentais, onde ocorre um declínio de qualidade do ajuste (BORGES, 2007).

Pode-se considerar que o algoritmo utiliza uma abordagem de solução local, pois o filtro é realizado apenas em pequenas regiões de cada vez, com o uso de janelas, ao contrário de outros tipos de algoritmos que usam uma abordagem global, considerando todos os pontos existentes para realizar a filtragem dos dados. Outra classificação possível para o método Savitzky-Golay é a que o define com controle implícito, pois não existe nenhum parâmetro que possa ser configurado para definir o nível de suavização a ser aplicado na regressão a ser realizada pelo polinômio. Se existisse esse parâmetro, como em outros métodos, a classificação seria de controle explícito.

3.3.2 Método das Diferenças Finitas

Em relação aos métodos numéricos, o método das diferenças finitas é extremamente conhecido e utilizado para diferenciação. Através da utilização de uma série de Taylor, pode-se encontrar o valor da derivada de primeira ordem em um ponto x. A partir daí, a implementação do método pode ser realizada de três formas distintas, implicando erros de truncamento com ordens de grandeza diferentes. O método da diferença finita progressiva e o método da diferença finita regressiva, consideram intervalos após um ponto x ou antes de um ponto x, respectivamente. Esses dois métodos apresentam erro de truncamento com ordem de grandeza $O(h)$ (sendo h o tamanho do intervalo), ou seja, uma aproximação de primeira ordem. Já o método das diferenças finitas centrais, que basicamente é uma combinação dos dois métodos anteriores, considera dois intervalos de tamanho h para o ponto x, um antes e outro depois, de mesmo tamanho. O erro de truncamento para esse método possui ordem de grandeza $O(h^2)$, ou seja, apresenta uma aproximação de segunda ordem, melhor do que os dois métodos anteriores (LYNESS; MOLER, 1967).

Para o método das diferença finita progressiva, considerando uma expansão na série

de Taylor, temos:

$$f(x+h) = f(x) + h\frac{df}{dx} + \frac{h^2}{2!}\frac{d^2f}{dx^2} + \frac{h^3}{3!}\frac{d^3f}{dx^3} + ... \tag{3.12}$$

pode-se então isolar o termo relacionado à derivada, obtendo:

$$\frac{df}{dx} = \frac{f(x+h) - f(x)}{h} + \frac{h}{2!}\frac{d^2f}{dx^2} + ... \tag{3.13}$$

considerando o erro de truncamento como:

$$O(h) = \frac{h}{2!}\frac{d^2f}{dx^2} + ... \tag{3.14}$$

então temos a derivada de primeira ordem para o método da diferença finita progressiva:

$$\frac{df}{dx} = \frac{f(x+h) - f(x)}{h} + O(h) \tag{3.15}$$

para a função escalar $f(x+h)$.

De maneira similar, para a função $f(x-h)$, é feita a expansão em série de Taylor:

$$f(x-h) = f(x) + (-h)\frac{df}{dx} + \frac{(-h)^2}{2!}\frac{d^2f}{dx^2} + \frac{(-h)^3}{3!}\frac{d^3f}{dx^3} + ... \tag{3.16}$$

realizando a mesma dedução, podemos encontrar o valor da derivada para o método da diferença finita regressiva:

$$\frac{df}{dx} = \frac{f(x) - f(x-h)}{h} + O(h) \tag{3.17}$$

Portanto, podemos ver que a derivada para os dois métodos possui erro de truncamento de mesma ordem de grandeza, resultando em uma aproximação de mesma ordem.

Finalmente, para a diferença finita central, subtraindo a Equação 3.17 da Equação 3.15 e isolando o termo correspondente à derivada chegamos em uma aproximação de segunda ordem, na qual o erro de truncamento é $O(h^2)$:

$$\frac{df}{dx} = \frac{f(x+h) - f(x-h)}{2h} + O(h^2) \tag{3.18}$$

onde:

$$O(h^2) = \frac{(-h)^2}{3!}\frac{d^3f}{dx^3} + ... \tag{3.19}$$

Apesar da equação das diferenças finitas centrais apresentar uma aproximação de segunda ordem, melhor do que os métodos das diferenças finitas progressiva e regressiva, todos os três métodos possuem o mesmo tipo de problema para que seja feita a escolha de um tamanho adequado para h, que é o problema do cancelamento subtrativo. Ele ocorre porque quanto mais for diminuído o tamanho de h, mais esse valor se aproximará de zero, consequentemente $f(x+h)$ e $f(x-h)$ terão valores mais próximos de $f(x)$ (SQUIRE; TRAPP, 1998) e o resultado das Equações 3.15, 3.17 e 3.18 tendem a zero. Assim, esses métodos, apesar de muito usados, se tornam ineficientes quando é necessário utilizar h com valores muito pequenos.

3.3.3 Método do Passo Complexo

Outro modo de se obter uma aproximação para derivadas de primeira ordem de maneira relativamente simples pode ser feito através de cálculo complexo. A aproximação de derivadas usando o método do passo complexo não está sujeito ao erro de cancelamento subtrativo, simplesmente porque não envolve uma operação de diferenças. Em relação aos métodos de aproximação por diferenças finitas, isso se torna uma enorme vantagem (MARTINS; STURDZA; ALONSO, 2003).

A utilização de variáveis complexas na resolução de problemas numéricos, especificamente diferenciação, foi apresentada por Lyness e Moler (1967) ao descrever um método para o cálculo de derivadas de qualquer função analítica. Pouco desenvolvimento foi realizado até que Squire e Trapp (1998) apresentaram o método do passo complexo (Complex Step Method), utilizando o número i ($i^2 = -1$), puramente imaginário, para se calcular a primeira e segunda derivadas de funções reais (ABREU; STICH; MORALES, 2013).

Da mesma forma que os métodos anteriores, o método do passo complexo pode ser derivado de uma série de Taylor a partir da função $f(x+ih)$:

$$f(x+ih) = f(x) + (ih)\frac{df}{dx} + \frac{(ih)^2}{2!}\frac{d^2f}{dx^2} + \frac{(ih)^3}{3!}\frac{d^3f}{dx^3} + ... \tag{3.20}$$

como $i^2 = -1$ então:

$$f(x+ih) = f(x) + (ih)\frac{df}{dx} - \frac{h^2}{2!}\frac{d^2f}{dx^2} - \frac{ih^3}{3!}\frac{d^3f}{dx^3} + ... \tag{3.21}$$

como para um número complexo $c = a + bi$, $Im[c] = b$, então tirando a parte imaginária de ambos os termos:

$$Im\left[f(x+ih)\right] = Im\left[f(x)\right] + Im\left[(ih)\frac{df}{dx}\right] - Im\left[\frac{h^2}{2!}\frac{d^2f}{dx^2}\right] - Im\left[\frac{ih^3}{3!}\frac{d^3f}{dx^3}\right] + ... \tag{3.22}$$

considerando que $f(x)$ e $\frac{h^2}{2!}\frac{d^2f}{dx^2}$ são reais, portanto $Im\left[f(x)\right]$ e $Im\left[\frac{h^2}{2!}\frac{d^2f}{dx^2}\right]$ são iguais a zero, então temos:

$$Im\left[f(x+ih)\right] = h\frac{df}{dx} - \frac{h^3}{3!}\frac{d^3f}{dx^3} + ... \tag{3.23}$$

dividindo por h:

$$\frac{Im\left[f(x+ih)\right]}{h} = \frac{df}{dx} - \frac{h^2}{3!}\frac{d^3f}{dx^3} + ... \tag{3.24}$$

como $\frac{h^2}{3!}\frac{d^3f}{dx^3} = O(h^2)$, então:

$$\frac{df}{dx} = \frac{Im\left[f(x+ih)\right]}{h} + O(h^2) \tag{3.25}$$

Portanto, pode-se notar que o método do passo complexo, assim como o método da diferença finita central, apresenta erro de truncamento de ordem $O(h^2)$, ou seja, ambos os métodos apresentam uma aproximação de mesma ordem. No entanto, o método do passo complexo não possui nenhuma operação de subtração na definição do cálculo da derivada, o que evita a ocorrência do cancelamento subtrativo, apesar de continuar dependendo da função ou de uma estimativa da função, assim como o método da diferença finita central.

Sendo assim, o primeiro objetivo do presente trabalho é avaliar os resultados, principalmente a precisão, do uso do método do passo complexo na estimativa das derivadas das séries temporais para determinado sistema biológico em estudo, pois como já visto no início deste capítulo, a estimativa das derivadas torna-se importante quando é adotada a estratégia de desacoplamento do sistema de equações diferenciais.

Deve-se considerar que o valor de h no caso da aplicação em problemas de sistemas bi-

ológicos é dado pela distância entre os intervalos de tempo em que são obtidas as medições experimentais. Usualmente, estes intervalos não costumam ser pequenos o suficiente para que se apresente o cancelamento subtrativo. Pelo contrário, o maior problema é a perda de precisão com a aplicação de diferenças finitas pelo tamanho do intervalo temporal. Desta forma, para que se tenha ganhos com a aplicação do método do passo complexo em sistemas biológicos é necessário que se determine, de forma artificial, intervalos para o cálculo da derivada que sejam inferiores ao intervalo temporal de referência. Esta estratégia será apresentada em seções posteriores.

3.4 Otimização direta do problema inverso

A estimativa dos parâmetros pode ser formulada como um problema de otimização direto cujo objetivo é minimizar uma função objetivo que mede a diferença entre os dados experimentais e as predições do modelo, geralmente com um critério baseado em erro de mínimos quadrados. As funções objetivo mais comumente utilizadas são baseadas em erro nas concentrações ou erro nas derivadas. As funções objetivo baseadas em erro nas concentrações geralmente são computacionalmente mais caras, por exigir métodos de integração numérica que resolvem as equações diferenciais, gerando os perfis metabólicos que serão comparados com os dados originais. Em comparação, as funções objetivo baseadas em erro nas derivadas fazem uso da informação das derivadas estimadas e as comparam com as derivadas preditas pelo método de resolução. Essa segunda opção depende da utilização da técnica de desacoplamento descrita anteriormente, passando o sistema de equações diferenciais para um sistema de equações algébricas (CHOU; VOIT, 2009) gerando um procedimento de otimização computacionalmente mais barato pelo fato de não ser necessário resolver o sistema de equações diferenciais, e é a opção que foi adotada neste trabalho.

Basicamente, a estimativa de parâmetros a partir de séries temporais pode ser agrupada em três classes distintas de métodos que são encontrados na literatura para solução de sistemas biológicos que fazem uso do modelo Sistema S, a saber: métodos baseado em gradiente, algoritmos de busca estocástica e outros métodos específicos para esse problema (não pertencentes a nenhuma das duas classes anteriores) (CHOU; VOIT, 2009).

Dois métodos serão implementados neste trabalho. Um deles é o método da Regressão

Alternada (*Alternating Regression*) (CHOU; MARTENS; VOIT, 2006) que é um método direcionado exclusivamente para a aplicação em TSB, pois utiliza o fato de que as funções baseadas em leis de potência são lineares em escalas logarítmicas (CHOU; VOIT, 2009). O método da RA (Regressão Alternada) é um método específico que não pertence a nenhuma das duas primeiras classes apresentadas anteriormente. O outro método estudado será o método da Entropia Cruzada - EC - (*Cross-entropy*) (RUBINSTEIN; KROESE, 2004), um método de otimização estocástica que ainda não foi aplicado ao contexto de resolução de sistemas biológicos com Sistema S. Portanto, a avaliação de sua utilização na solução do problema abordado neste trabalho será de extremo interesse para avaliação de seu potencial.

Outra opção de solução que será apresentada será a utilização de um método Híbrido, que fará uso tanto do RA quanto do EC para a resolução do problema de otimização. A descrição do método Híbrido e suas vantagens em relação aos outros dois métodos serão apresentadas na Seção 3.5.

3.4.1 Método da Regressão Alternada

Aplicado à estimativa de parâmetros em modelos de Sistema S, o algoritmo de RA funciona alternando entre fases de uma regressão linear múltipla, estimando os parâmetros do termo de produção e do termo de degradação (um em cada fase) para a equação de cada um dos metabólitos. São realizadas iterações entre as fases até que um critério de parada seja satisfeito (CHOU; MARTENS; VOIT, 2006). O algoritmo de RA é conhecido por ser mais rápido do que qualquer outro algoritmo que realiza estimativas diretas para sistemas de equações diferenciais não-lineares (BEYER, 2008). No entanto a velocidade de convergência do algoritmo de RA depende de vários fatores, entre eles a escolha dos valores iniciais dos parâmetros do termo de degradação do sistema (CHOU; MARTENS; VOIT, 2006).

Considerando n metabólitos e N instantes de tempo, para a i-ésima equação diferencial de determinado metabólito, o algoritmo de RA executa os seguintes passos:

(1) Calcule L_p, que é uma matriz (n+1) x N, utilizada para determinar os valores dos

parâmetros do termo de produção a partir dos regressores X_i:

$$L_p = \begin{bmatrix} 1 & log(X_1(t_1)) & ... & log(X_i(t_1)) & ... & log(X_n(t_1)) \\ 1 & log(X_1(t_2)) & ... & log(X_i(t_2)) & ... & log(X_n(t_2)) \\ \vdots & \vdots & ... & \vdots & ... & \vdots \\ 1 & log(X_1(t_k)) & ... & log(X_i(t_k)) & ... & log(X_n(t_k)) \\ \vdots & \vdots & ... & \vdots & ... & \vdots \\ 1 & log(X_1(t_N)) & ... & log(X_i(t_N)) & ... & log(X_n(t_N)) \end{bmatrix} \tag{3.26}$$

(2) Calcule L_d, que é uma matriz (n+1) x N, utilizada para determinar os valores dos parâmetros do termo de degradação e é análoga à L_p.

(3) Calcule as matrizes C_p e C_d, que são constantes durante todo o processo de iteração:

$$C_p = ({L_p}^T L_p)^{-1} {L_p}^T \tag{3.27}$$

$$C_d = ({L_d}^T L_d)^{-1} {L_d}^T \tag{3.28}$$

(4) Escolha valores iniciais para β_i e h_{ij} (parâmetros do termo de degradação).

(5) (INÍCIO DA ITERAÇÃO) Usando os valores de $X_j(t_k)$ da série temporal, calcule $\beta_i \prod_{j=1}^{n} {X_j}^{h_{ij}}$ para todo t_k, k = 1, 2, ..., N.

(6) Calcule o vetor N-dimensional y_d:

$$y_d = log\left(S_i(t_k) + \beta_i \prod_{j=1}^{n} {X_j}^{h_{ij}}(t_k)\right) k = 1, 2, \ldots, N \tag{3.29}$$

onde $S_i(t_k)$ é a derivada estimada no instante de tempo para o metabólito X_i.

(7) Considerando o modelo de regressão linear múltipla

$$y_d = L_p b_p + \epsilon_p \tag{3.30}$$

estime os parâmetros do termo de produção calculando o vetor de coeficientes b_p:

$$b_p = [\hat{\alpha}_i, \hat{g}_{ij}, j = 1, 2, \ldots, n]' = C_p y_d \tag{3.31}$$

(8) Usando os valores de $X_j(t_k)$ da série temporal, calcule $\alpha_i \prod_{j=1}^{n} {X_j}^{g_{ij}}$ para todo t_k, k = 1, 2, ..., N.

(9) Calcule o vetor N-dimensional y_p:

$$y_p = log\left(\alpha_i \prod_{j=1}^{n} {X_j}^{g_{ij}}(t_k) - S_i(t_k)\right) k = 1, 2, \ldots, N \tag{3.32}$$

(10) Considerando o modelo de regressão linear múltipla

$$y_p = L_d b_d + \epsilon_d \tag{3.33}$$

estime os parâmetros do termo de degradação calculando o vetor de coeficientes b_d:

$$b_d = [\hat{\beta}_i, \hat{h}_{ij}, j = 1, 2, \ldots, n]' = C_d y_p \tag{3.34}$$

(11) Calcule o logaritmo da soma dos erros quadrados:

$$log(erro) = log\left(\sum_{k=1}^{N} (y_k - \hat{y}_k)^2\right) \tag{3.35}$$

onde $\hat{y} = L \times b, L$ igual L_p ou L_d e b é o vetor solução b_p ou b_d.

(12) Continue as iterações dos passos (5)-(11) até alcançar um critério de parada (log(erro) menor do que um valor ou número de iterações maior do que um valor).

O algoritmo acima, da maneira que foi proposto em (CHOU; MARTENS; VOIT, 2006), deve ser executado de acordo com os passos mostrados anteriormente para cada uma das equações relacionadas a cada um dos metabólitos do sistema, uma de cada vez, até completar a estimativa de todos os parâmetros do sistema.

3.4.2 Método da Entropia Cruzada

O método da EC (Entropia Cruzada) foi idealizado por Rubinstein (1996) e inicialmente proposto para estimativa de probabilidades de eventos raros utilizando minimização de variância. Trata-se de um algoritmo adaptativo que através de algumas modificações passou a ser usado também na solução de problemas de otimização combinatória (RUBINSTEIN, 1999)(RUBINSTEIN, 2001). Pode ser classificado na mesma categoria de

outros métodos usados na solução de problemas de otimização combinatória, como simulated annealing, tabu search, algoritmos genéticos e colônia de formigas (BOER et al., 2004). É desenvolvido como um algoritmo de estimação de distribuição, que se diferencia dos algoritmos genéticos pelo fato de que a evolução entre as gerações é realizada através da estimativa da distribuição de probabilidade dos indivíduos mais aptos em uma geração e essa informação é utilizada para criar a geração seguinte, o que elimina a necessidade de se utilizar os operadores de recombinação e mutação, usados nos Algoritmos Genéticos (BENGOETXEA et al., 2001).

Como a proposta do método é direcionada para a estimativa de eventos raros então a solução dos problemas de otimização combinatória precisam sofrer uma transformação de um problema de otimização "determinístico" (eventos raros) para um problema de otimização "estocástico" (otimização combinatória) no qual serão utilizadas técnicas de simulação similares às definidas em (RUBINSTEIN, 1996). O problema que é resultado dessa transformação para que sejam utilizadas as técnicas de solução propostas é conhecido como "problema estocástico associado" (associated stochastic problem - ASP) (BOER et al., 2004).

De maneira genérica, o método da EC realiza iterações que passam por duas fases (BOER et al., 2004):

1. Geração de uma amostra de dados aleatória (criação dos indivíduos) de acordo com um mecanismo específico;

2. Atualização dos parâmetros do mecanismo aleatório tomando como base a classificação dos dados (ranqueamento dos indivíduos) para que sejam produzidos amostras melhores na próxima iteração.

O método tem sido utilizado em diversas publicações e aplicado em ampla variedade de contextos como alinhamento de sequência de DNA, roteamento de veículos, aprendizado por reforço, gerenciamento de projetos, confiabilidade de redes, controle e navegação, entre outros (BOER et al., 2004). Segundo Keith e Kroese (2002) o método possui potencial para ser utilizado em diversas classes de problemas de otimização combinatória, incluindo ciências biológicas, devido a sua versatilidade, simplicidade e tratabilidade matemática.

Como a aplicação do método da EC neste trabalho será restrita à solução de um problema de otimização combinatória então será apresentada somente a fundamentação

relacionada à solução deste tipo de problema. A fundamentação e apresentação de exemplos para a simulação de eventos raros pode ser encontrada em (BOER et al., 2004) e (RUBINSTEIN, 1996).

3.4.2.1 Fundamentação para problemas de otimização combinatória

Considerando χ um conjunto finito de *estados* e S uma *função de custo* em χ na qual o resultado é um valor real, deve-se encontrar o maior valor de S em χ e o seu estado correspondente. Chamando o maior valor de γ^*, então (BOER et al., 2004):

$$S(x^*) = \gamma^* = \max_{x \in \chi} S(x) \tag{3.36}$$

O método da EC associa um problema de otimização ao problema de estimativa da Equação 3.36. Para isso é definida uma coleção de funções indicadoras $\{I_{\{S(x) \geq \gamma\}}\}$ em χ para vários limiares ou níveis $\gamma \in \mathbb{R}$. A seguir, considere $\{f(\cdot; \mathbf{v}), \mathbf{v} \in \nu\}$ uma família de funções de distribuição de probabilidades em χ, parametrizadas por um parâmetro $\mathbf{v}$ (vetor). Para um determinado $\mathbf{u} \in \nu$ é associado à Equação 3.36 o problema de estimativa (BOER et al., 2004):

$$l(\gamma) = \mathbb{P}_{\mathbf{u}}(S(\mathbf{X}) \geq \gamma) = \sum_{\mathbf{x}} I_{\{S(\mathbf{x}) \geq \gamma\}} f(\mathbf{x}; \mathbf{u}) = \mathbb{E}_{\mathbf{u}} I_{\{S(\mathbf{X}) \geq \gamma\}} \tag{3.37}$$

no qual $\mathbb{P}_{\mathbf{u}}$ é a medida de probabilidade sobre a qual o estado aleatório $\mathbf{X}$ tem a função de distribuição de probabilidade $f(\cdot; \mathbf{u})$ e $\mathbb{E}_{\mathbf{u}}$ representa o operador esperança correspondente. O problema representado na Equação 3.37 é conhecido como problema estocástico associado (associated stochastic problem - ASP) (BOER et al., 2004).

3.4.3 Exemplo para o método da EC

Considerando um vetor binário $\mathbf{y} = (y_1, \ldots, y_n)$, suponha que não se saiba qual componente de $\mathbf{y}$ seja 0 ou 1. Pode ser definida uma função de custo (BOER et al., 2004):

$$S(\mathbf{x}) = n - \sum_{j=1}^{n} |x_j - y_j| \tag{3.38}$$

que aceita um vetor de entrada $\mathbf{x} = (x_1, \ldots, x_n)$ e cujo resultado representa o número de acertos entre os elementos de $\mathbf{x}$ e $\mathbf{y}$. Portanto, deve ser apresentado um algoritmo de

busca aleatória que encontre o vetor $\mathbf{y}$ maximizando a função $S(\mathbf{x})$ (BOER et al., 2004).

Uma maneira de encontrar $\mathbf{y}$ é gerar vetores binários $\mathbf{X} = (X_1, \ldots, X_n)$ de tal forma que $X_1, \ldots, X_n$ são variáveis aleatórias de Bernoulli, ou seja, $\mathbf{X} \sim \mathbf{p} = (p_1, \ldots, p_n)$. Se $\mathbf{p} = \mathbf{y}$ então $S(\mathbf{X}) = n$, $\mathbf{X} = \mathbf{y}$ e a busca termina com a solução ótima (BOER et al., 2004).

No entanto, na aplicação do método da EC para otimização combinatória, o problema é transformado em uma simulação de evento raro conforme a Equação 3.37 e uma sequência de vetores de parâmetros $\hat{\mathbf{p}}_0, \hat{\mathbf{p}}_1, \ldots$ e níveis $\hat{\gamma}_1, \hat{\gamma}_2, \ldots$, de tal forma que a sequência $\hat{\gamma}_1, \hat{\gamma}_2, \ldots$ possa convergir para o resultado ótimo, $S(\mathbf{x}) = n$, e a sequência $\hat{\mathbf{p}}_0, \hat{\mathbf{p}}_1, \ldots$ possa convergir para o vetor de parâmetros ótimo ($\mathbf{y}$). O algoritmo é descrito a seguir (BOER et al., 2004):

(1) Inicie $\hat{\boldsymbol{p}}_0$ (por exemplo, $\hat{\boldsymbol{p}}_0 = (1/2, 1/2, \ldots, 1/2)$) e $t := 0$.

(2) Instancie a população (amostragem) $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ (onde N é o tamanho da população, que deve ser pré-definido) de vetores baseado no vetor de probabilidade $\hat{\boldsymbol{p}}_{t-1}$. Calcule a função de custo $S(\boldsymbol{X}_i)$ para todo i e ordene como $S_{(1)} \leq \cdots \leq S_{(N)}$. Seja $\hat{\gamma}_t$ o quantil $(1-\rho)$ da amostragem com performance $\hat{\gamma}_t = S_{(\lceil (1-\rho)N \rceil)}$.

(3) Use a população para calcular $\hat{\boldsymbol{p}} = (\hat{p}_{t,1}, \ldots, \hat{p}_{t,n})$ através de:

$$\hat{p}_{t,j} = \frac{\sum_{i=1}^{N} I_{\{S(\boldsymbol{X}_i) \geq \hat{\gamma}_t\}} I_{\{X_{ij}=1\}}}{\sum_{i=1}^{N} I_{\{S(\boldsymbol{X}_i) \geq \hat{\gamma}_t\}}} \tag{3.39}$$

onde $j = 1, \ldots, n$ e $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{in})$.

(4) Se determinado critério de parada for alcançado então pare. Senão faça $t := t + 1$ e inicie uma nova iteração a partir do passo 2.

Dois possíveis critérios de parada a serem utilizados são: (i) $\hat{\gamma}_t$ não se altera durante um número determinado de iterações; ou (ii) o vetor $\hat{\boldsymbol{p}}_t$ converge para o vetor de parâmetros ótimos (BOER et al., 2004).

A interpretação da Equação 3.39 pode ser feita como: para atualizar o j-ésimo elemento, são contados quantos vetores da última amostragem (população) $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ possuem o resultado da função de custo maior ou igual a $\hat{\gamma}_t$ e possuem a j-ésima coordenada igual a 1. Então é feita a divisão (normalização) pelo número de vetores que possuem o resultado da função de custo maior ou igual a $\hat{\gamma}_t$ (BOER et al., 2004).

Considerando o caso no qual $n = 10$, $\boldsymbol{y} = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$, o tamanho da população é $N = 50$, o parâmetro $\rho = 0,1$ e o vetor de parâmetros inicial é $\hat{\boldsymbol{p}}_0 = (1/2, 1/2, \ldots, 1/2)$, então a Tabela 3.1 mostra os resultados para 5 iterações (BOER et al., 2004).

Tabela 3.1: Convergência do vetor de parâmetros (BOER et al., 2004)

t	$\hat{\gamma}_t$	$\hat{\boldsymbol{p}}_t$									
0		0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50
1	7	0,60	0,40	0,80	0,40	1,00	0,00	0,20	0,40	0,00	0,00
2	9	0,80	0,80	1,00	0,80	1,00	0,00	0,00	0,40	0,00	0,00
3	10	1,00	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00	0,00
4	10	1,00	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00	0,00

Pode-se observar na Tabela 3.1 que a convergência dos parâmetros $\hat{\boldsymbol{p}}_t$ e $\hat{\gamma}_t$ é extremamente rápida em direção aos parâmetros ótimos $\boldsymbol{p}^* = \boldsymbol{y}$ e $\gamma^* = n$, respectivamente (BOER et al., 2004).

3.4.4 Interpretação gráfica para o método da EC aplicado em otimização

Devido ao fato de que os conceitos matemáticos envolvidos na elaboração do método da EC não trazem uma clareza completa de seu funcionamento na prática, serão mostrados dois exemplos gráficos para tentar facilitar a intuição em relação ao funcionamento do algoritmo.

Inicialmente será considerado um exemplo em somente uma dimensão com o objetivo de explicar de forma didática os conceitos. A seguir será apresentado um exemplo em duas dimensões com o objetivo de generalizar o entendimento e aplicação dos conceitos na resolução do algoritmo.

A Figura 3.4 mostra a função na qual será realizada a busca e o ponto de valor mínimo (ponto O) a ser obtido na otimização. A população (amostragem) que será instanciada corresponde a valores no eixo horizontal (eixo x), e a função de custo corresponde à função $f(x)$, que será calculada para cada indivíduo da população.

A Figura 3.5 mostra no gráfico superior os pontos que representam o resultado do cálculo da função de custo para a população selecionada (a população selecionada são os valores do eixo x para cada um dos pontos). Pode ser observado que a instanciação da

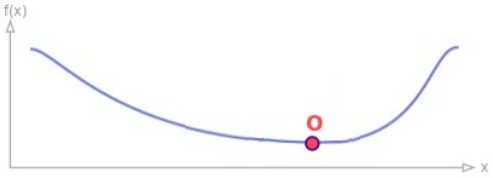


Figura 3.4: Função inicial para minimização (LARRANAGA; LOZANO, 2002)

população na primeira iteração foi realizada baseando-se em uma função de distribuição uniforme (linha amarela).

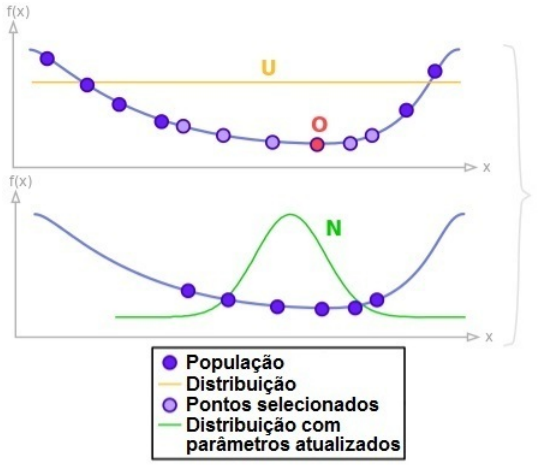


Figura 3.5: Primeira iteração do algoritmo para a função em uma dimensão (LARRANAGA; LOZANO, 2002)

Ainda na Figura 3.5, no gráfico inferior, são mostrados os pontos que foram selecionados como os "melhores indivíduos", ou seja, os valores mais próximos do ponto mínimo (note que no gráfico superior esses pontos aparecem em uma cor mais clara). A função de distribuição de probabilidade mostrada em verde é uma fator essencial para o entendimento do funcionamento do algoritmo. Trata-se da função de distribuição de probabilidade calculada a partir desses pontos que foram selecionados como os melhores da iteração. Essa função será utilizada para instanciar a população da iteração seguinte.

Conforme pode ser visto no gráfico superior da Figura 3.6, a função de distribuição de probabilidade (em amarelo) utilizada para instanciação da população é a mesma função da Figura 3.5, da parte inferior (em verde). Portanto, pode-se observar aqui uma ideia central utilizada no método da EC, que é o uso da função de distribuição de probabilidade calculada para os indivíduos selecionados em uma iteração para realizar a instanciação da população da iteração seguinte. Dessa forma, através da atualização da média e da variância da distribuição de probabilidade, o algoritmo é capaz de instanciar uma população

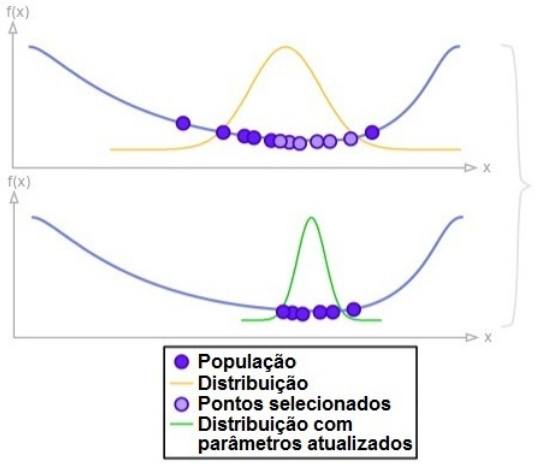


Figura 3.6: Segunda iteração do algoritmo para a função em uma dimensão (LARRANAGA; LOZANO, 2002)

que a cada iteração converge para a minimização (ou maximização) da função objetivo.

Continuando a observação do exemplo, o gráfico inferior da Figura 3.6 mostra o cálculo da nova função de distribuição de probabilidade (em verde) baseada nos indivíduos selecionados (em cor mais clara no gráfico superior). Essa nova função, novamente será utilizada para a instanciação da população da iteração seguinte.

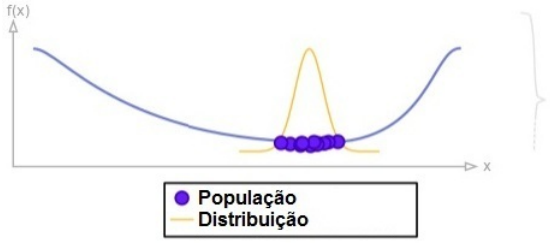
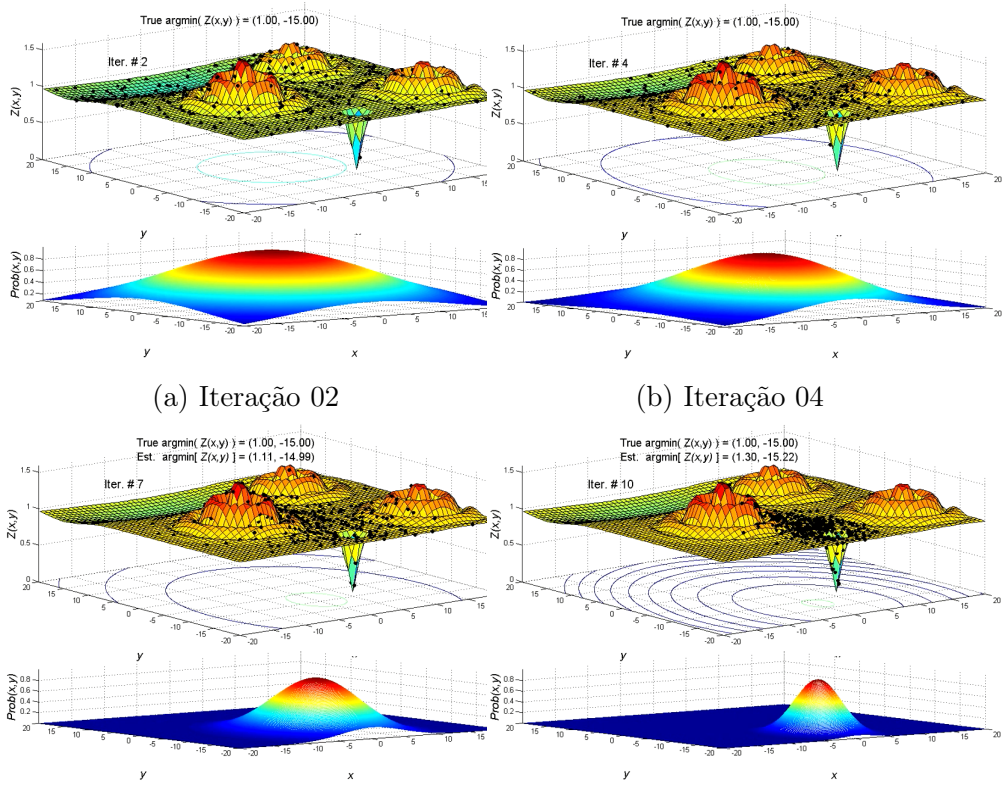


Figura 3.7: Terceira iteração do algoritmo para a função em uma dimensão (LARRANAGA; LOZANO, 2002)

A Figura 3.7 mostra a função de distribuição de probabilidade (em amarelo) utilizada para realizar a instanciação da população para a terceira iteração do algoritmo. Repare que a função é a mesma da iteração anterior que for resultado do cálculo para os indivíduos selecionados (em verde, no gráfico inferior da Figura 3.6). A convergência dos resultados da função de custo para a população pode ser claramente percebida, assim como as alterações da média e variância das funções de distribuição de probabilidades calculadas.

A Figura 3.8 mostra um exemplo de aplicação do método da EC em duas dimensões, disponibilizado por Oklander (2012). Trata-se da minimização de uma função que apresenta diversos ótimos locais e um ótimo global.



(a) Iteração 02 (b) Iteração 04

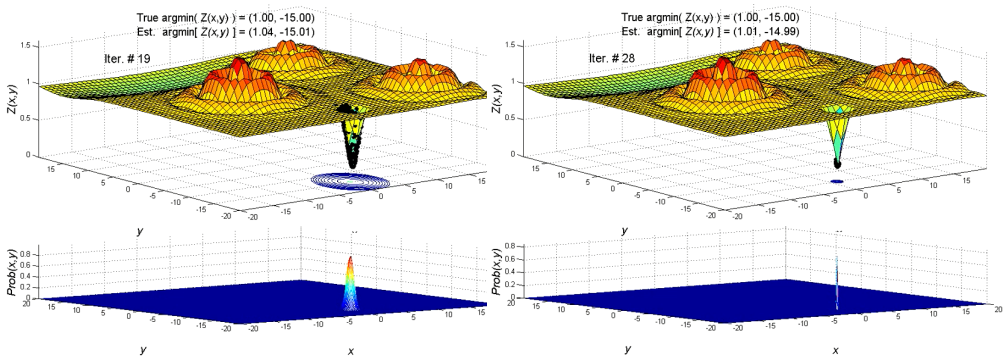
(c) Iteração 07 (d) Iteração 10

Figura 3.8: Primeira iterações do EC para duas dimensões (OKLANDER, 2012)

As quatro iterações visualizadas na Figura 3.8 (iterações 02, 04, 07 e 10) mostram a evolução da função de distribuição de probabilidade, que se move em direção ao ponto de mínimo. Da mesma forma que no exemplo anterior, essa função é encontrada a partir dos indivíduos selecionados como sendo os melhores, depois que a função de custo é calculada. Na iteração seguinte, essa mesma função de distribuição de probabilidade é usada para instanciar a nova população.

A Figura 3.9 mostra algumas das últimas iterações do método para o exemplo em duas dimensões. Chama a atenção o valor da variância da função de distribuição de probabilidade quando a população se aproxima do valor mínimo, pois o gráfico da função se torna extremamente fino.

Os exemplos gráficos mostrados acima tiveram o objetivo de proporcionar uma visualização mais intuitiva do funcionamento do método da EC, para auxiliar o entendimento da formalização matemática apresentada anteriormente.



(a) Iteração 19 (b) Iteração 28

Figura 3.9: Últimas iterações do EC para duas dimensões (OKLANDER, 2012)

3.5 Otimização com estratégia híbrida

A ideia de se utilizar um método Híbrido para resolver o problema de otimização para estimativa dos parâmetros vem da constatação de que grande parte dos algoritmos são capazes de realizar de maneira eficiente somente um tipo de busca: global ou local. Ou seja, determinados algoritmos são muito bons para encontrar os parâmetros corretos em um domínio multimodal (globalmente), mas de forma geral possuem alto custo computacional em sua execução. Outros algoritmos são extremamente rápidos em sua execução, mas no entanto conseguem chegar a resultados corretos somente se os valores iniciais dos parâmetros buscados estiverem próximos (localmente) aos resultados procurados.

Cada um dos dois algoritmos apresentados aqui neste trabalho pertence a um grupo desses. O modelo de EC é um método de busca global, que mesmo considerando um domínio multimodal para o valor da função, espera-se que a solução seja encontrada. No entanto, o custo computacional pode ser alto, principalmente se os valores iniciais da busca estiverem muito longe do ótimo global.

Por sua vez, o método da RA é capaz de realizar uma busca extremamente rápida, mas somente para valores iniciais de busca que estejam próximos ao ótimo procurado, ou seja, esse algoritmo é altamente dependente de uma escolha favorável dos valores iniciais dos parâmetros de busca, caso contrário podem aumentar muito seu tempo de processamento ou encontrar valores ótimos locais (CHOU; MARTENS; VOIT, 2006).

O método Híbrido que será utilizado neste trabalho procura fazer uso das características mais vantajosas de cada um dos dois algoritmos. A busca se iniciará com o método da EC, considerando uma faixa ampla de valores para os parâmetros buscados. Dessa forma, a busca realizada pelo algoritmo tenderá a diminuir o domínio e aproximar os valores dos

parâmetros de seu resultado ótimo. A partir de determinado ponto, ou seja, um critério de parada que será definido e explicado na Seção 4.2, o método da EC termina sua execução e passa os valores dos parâmetros para o método da RA. Por sua vez, este último algoritmo, que por ser efetivo em buscas locais, apresentará uma maior probabilidade de sucesso, pois o valor inicial dos parâmetros foi determinado pelo EC de forma a ficar em uma região mais adequada para a busca do ótimo global.

O algoritmo relacionado ao método Híbrido pode ser descrito na forma:

(1) Iniciar a população inicial do método da EC com valores aleatórios (usando algoritmo baseado no que foi apresentado nas seções 3.4.2 e 3.4.3).

(2) Ao encontrar um dos critérios de parada definidos (valor do erro mínimo ou número máximo de iterações) terminar o processamento e salvar o valor das variáveis para o termo de degradação (β_i e $h'_{ij}s$) do melhor indivíduo da população.

(3) Passar as variáveis do termo de degradação como parâmetros para o método da RA.

(4) Iniciar o método da RA com as variáveis recebidas como parâmetro (usando o mesmo algoritmo apresentado na Seção 3.4.1).

(5) Ao encontrar um dos critérios de parada definidos (valor do erro ou número de iterações) terminar o processamento.

Assim, fazendo uso das principais vantagens de cada um dos dois algoritmos, ou seja, busca global feita pelo método da EC (que permite encontrar os resultados mesmo para amplos domínios das variáveis) e busca local pelo método da RA (com grande velocidade na execução), pretende-se avaliar a eficácia desse método Híbrido na resolução de problemas de otimização na busca dos valores das variáveis para sistemas bioquímicos representados através de Sistema S.

3.5.1 Considerações

Ressalta-se, nesta Seção, que a montagem do modelo de otimização híbrido foi desenvolvido da maneira mais simples e direta para o acoplamento das duas estratégias de otimização envolvidas. Isto porque pretende-se avaliar, principalmente, a possibilidade de um algoritmo de busca global, a saber o método da EC, poder trazer ganhos em termos de qualidade de resposta quando utilizado em conjunto com um algoritmo especifico para resolução de problemas em TSB, como é o caso do método da RA. A questão é verificar se

a grande efetividade do método da RA como é altamente ressaltado na literatura possa ter um acréscimo em desempenho quando se tem valores mais adequados para os parâmetros iniciais de degradação.

Na construção do modelo Híbrido, com certeza a consideração de características especificas de cada método pode gerar um modelo mais eficiente e efetivo para a obtenção dos parâmetros do Sistema S. Por exemplo, a consideração que o método da RA necessita somente dos parâmetros de degradação pode levar a construção de um método da EC mais especifico para estes parâmetros. Da mesma forma, uma possível utilização de parâmetros de produção inicializados podem levar ao desenvolvimento de uma estratégia mais adequada aos dados provenientes do método da EC.

Questões como a utilização de mais de um indivíduo para a inicialização do método da RA também podem trazer desempenho mais eficientes para a qualidade final de um modelo hibrido.

4 EXPERIMENTOS NUMÉRICOS

Os experimentos serão apresentados separadamente para cada uma das etapas estudadas, pois apesar de fazerem parte da mesma estratégia de resolução, os sistemas utilizados para estudo em cada uma delas foram diferentes. O motivo é que o interesse principal neste trabalho não foi a análise global da estratégia de resolução, mas o estudo de métodos utilizados em cada uma dessas etapas para alcançar melhorias específicas em cada uma delas, consequentemente melhorando a estratégia completa para determinação de Sistemas S associado a dados metabólicos experimentais.

Todos os experimentos foram realizados em um Vostro 3550, com processador Intel(R) Core(TM) i7-2620M (4M de memória cache e 2.70 GHz), memória RAM DDR3 de 6 GB e Windows 7 Professional 64 Bits.

4.1 Estimativa das derivadas

Para avaliar a precisão do cálculo da derivada para cada um dos métodos será utilizado um modelo de sistema Sistema S sintético, que possui características semelhantes a sistemas reais. Serão geradas duas séries temporais a partir desse sistema sintético que terão um nível de 5% de ruído adicionado a partir de uma distribuição normal com média zero e desvio padrão igual a 1. Cada uma dessas séries terá um número diferente de pontos para um mesmo intervalo de tempo, o que torna o processo de filtragem mais difícil e menos preciso para a série com menos pontos, servindo como mais um parâmetro para avaliação dos resultados que serão obtidos pelos métodos.

O sistema sintético utilizado é baseado em sistemas de redes metabólicas e foi definido

por Voit (2000) da seguinte forma:

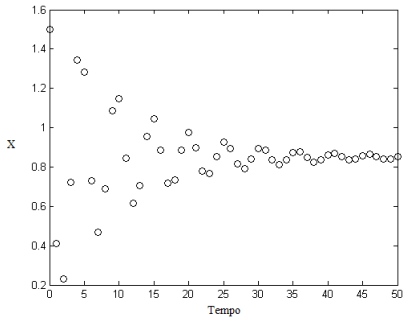
$$\begin{aligned}
&\tfrac{dX_1}{dt} = {X_2}^{-2}X_3 - {X_1}^{0.5}X_2,\\
&\tfrac{dX_2}{dt} = {X_1}^{0.5}X_2 - {X_2}^{0.5},\\
&X_3 = 1.5\\
&\text{sujeitas aos seguintes valores iniciais:}\\
&X_1(0) = 1.5,\\
&X_2(0) = 1.5
\end{aligned} \tag{4.1}$$

A partir do sistema da Equação 4.1 serão geradas duas séries temporais, considerando a concentração de X_2 e sua variação no tempo (ou seja, $\frac{dX_2}{dt}$). A seguir, adiciona-se ruído nos dados das séries geradas. Como resultado, tem-se duas séries temporais com ruído, ambas relacionadas à X_2 da Equação 4.1, que serão usadas para os cálculos utilizando os métodos estudados e posterior análise de seus resultados. Não é preciso uma estratégia de validação cruzada, pois o filtro foi aplicado em uma série temporal conhecida e com ruído controlado.

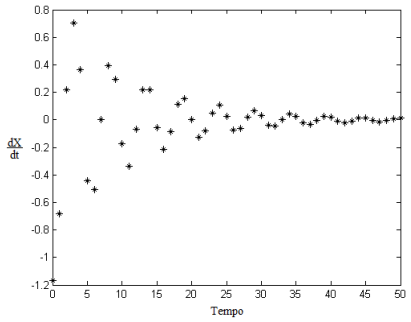
A Figura 4.1a mostra os 50 pontos gerados para a primeira série, com intervalo $h = 1$ entre os pontos. A Figura 4.1c mostra os pontos da segunda série temporal utilizada, com intervalo $h = 2$ entre os 25 pontos. Note que apesar de não serem utilizadas unidades bem definidas, os intervalos podem ser considerados unidades de tempo, ou seja, os 50 pontos são espaçados por uma unidade de tempo entre eles enquanto os 25 pontos serão espaçados por duas unidades de tempo. Ainda não foi adicionado ruído às séries mostradas nessas figuras. Os valores das derivadas também são mostrados para a primeira série, de 50 pontos, na Figura 4.1b e para a segunda série, de 25 pontos, na Figura 4.1d. Repare que os gráficos não apresentam a mesma escala para o eixo y, mas o eixo x é apresentado de zero até 50 para todos os gráficos.

Com as séries temporais geradas e com o ruído adicionado, será usado o algoritmo de Savitzky-Golay visando a filtragem e suavização dos dados, ou seja, é feita a tentativa de eliminação do ruído da melhor forma possível. A partir daí, para cada ponto na série, a resolução do problema de obtenção da diferenciação numérica será determinada de duas maneiras distintas, cada uma utilizando um método diferente, que terão seus resultados comparados depois.

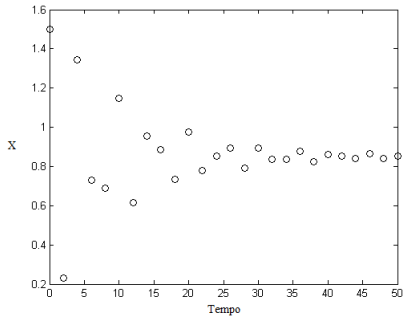
A primeira estratégia usa o resultado do próprio filtro de Savitzky-Golay, baseada



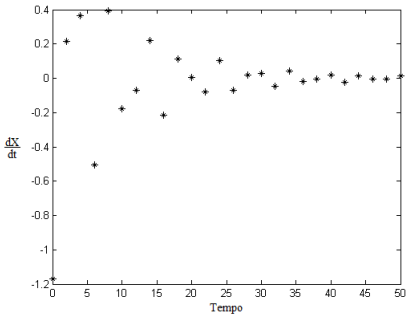
(a) Dados originais - 50 pontos



(b) Derivadas originais - 50 pontos



(c) Dados originais - 25 pontos



(d) Derivadas originais - 25 pontos

Figura 4.1: Dados e derivadas das séries temporais

na técnica da janela que se move, apresentada anteriormente na Seção 3.3.1, que realiza o ajuste de um polinômio para cada uma dessas janelas (cada janela terá 5 pontos e cada polinômio ajustado terá grau 3) e em seguida obtém o cálculo da derivada de forma analítica. Assim, cada ponto na série temporal terá um valor de derivada calculado analiticamente através do polinômio do filtro de Savitzky-Golay. A parte seguinte se refere ao cálculo utilizando os métodos numéricos baseados em aproximação de Taylor. Para cada ponto, serão utilizados vários valores de h (intervalo entre os pontos), que decrescerão continuamente até 10^{-12}. Essa diminuição do valor de h será fundamental para a análise dos métodos utilizados neste trabalho. Para cada valor de h será feito o cálculo da derivada numericamente através do método da diferença finita central, usando a Equação 3.18. A seguir, é feito outro cálculo numérico da derivada, dessa vez através do método do passo complexo, usando a Equação 3.25. Assim, para cada ponto, existirá um valor de derivada calculado de forma analítica polinomial (usando Savitzky-Golay) e diversos valores calculados (considerando os vários tamanhos de h) de forma numérica para cada um dos dois métodos. Esses valores servirão de base para a análise a ser realizada em relação aos métodos estudados. Vale ressaltar que o cálculo dos dois métodos numéricos e baseado no polinômio encontrado pelo filtro de Savitzky-Golay, apenas realizando a

diminuição no tamanho do intervalo h.

Como já explicado, o cálculo da derivada para os pontos experimentais da série temporal é de fundamental importância para que a estratégia adotada de desacoplamento do sistema de equações diferenciais ordinárias através do uso de Sistema S seja bem sucedida, portanto a principal métrica utilizada neste trabalho é o cálculo da estimativa da derivada em cada ponto. Assim, para avaliar a qualidade das derivadas encontradas em todos os pontos de uma determinada série, é definido um erro quadrático médio para o cálculo das derivadas:

$$MSE_{derivada} = \sqrt{\sum_{i=1}^{n}(y'_i - z'_i)^2} \tag{4.2}$$

onde y'_i é a derivada real e z'_i é a derivada estimada em cada ponto, sendo que a série temporal possui n pontos.

São executados 30 experimentos e é considerado o erro quadrático médio (MSE) e desvio padrão de todos os experimentos para cada um dos métodos, sendo que para os métodos numéricos são considerados os vários valores de h. Também é realizado um cálculo que considera esse MSE para os resultados dos métodos das diferenças finitas e passo complexo e é feita a diferença em relação ao MSE dos resultados para o método analítico.

4.2 Otimização para estimativa de parâmetros

Os experimentos da etapa de otimização possuem dois objetivos:

1. avaliar a eficácia do Método da Entropia Cruzada para a estimativa de parâmetros do modelo Sistema S;

2. avaliar a eficácia de um método Híbrido que utiliza tanto o Entropia Cruzada quanto o Regressão Alternada para a estimativa de parâmetros do modelo Sistema S.

O primeiro objetivo é avaliar a eficácia do método da EC, que é um algoritmo de otimização combinatória ainda não aplicado na solução de problemas de estimação de parâmetros para sistemas biológicos a partir de séries temporais. Outros algoritmos de

otimização combinatória, como algoritmos genéticos, colônia de formigas, enxame de partículas, entre outros, já foram utilizados com graus variados de sucesso (BOER et al., 2004).

O segundo objetivo é a avaliação do método Híbrido, no qual as características de cada um dos dois algoritmos serão combinadas para tentar alcançar resultados melhores que cada um deles alcança separadamente. A ideia principal do método Híbrido é usar a característica de busca global do método da EC para encontrar variáveis que estejam em uma região que possa facilitar a busca pelo método da RA, o qual por sua vez possui a desvantagem de não conseguir encontrar bons resultados para variáveis iniciais em qualquer região, por ser um algoritmo de busca local.

O método da EC possui níveis de erro com queda rápida nas primeiras iterações e uma convergência mais lenta quando o erro já é pequeno e próximo do resultado. A principal vantagem desse algoritmo é que a busca geralmente funciona mesmo para funções multimodais, inclusive quando as variáveis independentes possuem domínios com faixas de busca bem amplas. As principais desvantagens é que em algumas situações este algoritmo de busca global pode ser lento e a convergência do erro quando os valores estão próximos do resultado passa a acontecer mais devagar.

Já o método da RA, que é um algoritmo de busca local, possui níveis de erro com variações pequenas nas primeiras iterações e queda brusca no momento em que o algoritmo encontra uma combinação boa entre as variáveis viabilizando a convergência rápida para o resultado. A principal vantagem desse algoritmo é sua rapidez quando as variáveis iniciais se encontram com uma configuração de valores favorável (VILELA et al., 2008). A principal desvantagem, de forma contrária, é quando as variáveis iniciais se encontram com uma configuração de valores desfavorável ele pode demorar muito a convergir ou então pode até mesmo divergir e encontrar resultados errados (BEYER, 2008) (CHOU; MARTENS; VOIT, 2006).

O sistema que será usado, mostrado nas Equações 4.3, apesar de também ser sintético (assim como nos testes de estimativa das derivadas), possui características similares a sistemas reais mesmo com maior número de variáveis e já foi utilizado em diversos trabalhos (VOIT; ALMEIDA, 2004) (CHOU; MARTENS; VOIT, 2006) (VILELA et al., 2008)

(BEYER, 2008).

$$
\begin{aligned}
&\frac{dX_1}{dt} = 12{X_3}^{-0.8} - 10{X_1}^{0.5},\\
&\frac{dX_2}{dt} = 8{X_1}^{0.5} - 3{X_2}^{0.75},\\
&\frac{dX_3}{dt} = 3{X_2}^{0.75} - 5{X_3}^{0.5}{X_4}^{0.2},\\
&\frac{dX_4}{dt} = 2{X_1}^{0.5} - 6{X_4}^{0.8}\\
&\text{com as condições iniciais:}\\
&X_1(t_0) = 1.4,\\
&X_2(t_0) = 2.7,\\
&X_3(t_0) = 1.2,\\
&X_4(t_0) = 0.4
\end{aligned}
\tag{4.3}
$$

A Figura 4.2 mostra a representação da topologia do sistema acima. Pode-se observar as influências que cada metabólito exerce nos outros, como por exemplo a ativação da degradação de X_3 realizada por X_4 e a inibição que X_3 exerce em X_1 (BEYER, 2008).

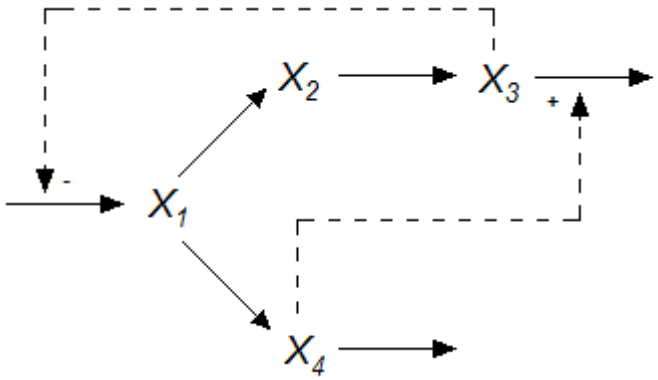
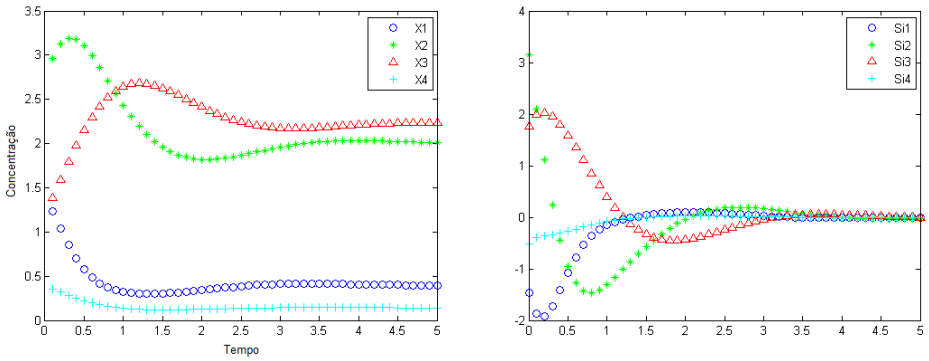


Figura 4.2: Topologia do sistema

Esse sistema é utilizado para gerar séries temporais para cada uma das quatro equações. A Figura 4.3 mostra os gráficos gerados para as concentrações (equações de $\frac{dX1}{dt}$ até $\frac{dX4}{dt}$) e para as derivadas em cada ponto para cada uma das equações. Como apresentado nas seções 3.1 e 3.2, a estratégia de resolução adotada neste trabalho realiza o desacoplamento do sistema de equações diferenciais, com as concentrações para cada uma das variáveis sendo obtidas através da resolução do sistema sintético associado, podendo-se adotar a inclusão de ruído para avaliação da robustez dos métodos avaliados. Da mesma forma, os valores das derivadas do sistema sintético também podem ser obtidas para posterior verificação da qualidade dos modelos adotados.



(a) Concentrações (b) Derivadas reais em cada ponto

Figura 4.3: Concentrações e derivadas das séries temporais para metabólitos do sistema

É importante notar que existem três possíveis cenários para realizar a busca dos parâmetros das equações do sistema, pois os dados existentes nas séries temporais podem fornecer informações relacionadas tanto à estrutura quanto à dinâmica do sistema bioquímico que os gerou (BEYER, 2008).

Sendo assim, no primeiro cenário a estrutura do sistema bioquímico é conhecida, então sabe-se quais parâmetros das equações do Sistema S são iguais a zero e quais são positivos ou negativos (CHOU; MARTENS; VOIT, 2006). A busca passa a ser relacionada somente às informações dinâmicas do sistema, ou seja, os valores dos parâmetros de cada equação que são diferentes de zero. É baseado neste tipo de cenário, também utilizado por Chou, Martens e Voit (2006) e Beyer (2008), que serão realizados os experimentos numéricos deste trabalho.

No segundo cenário, pelo menos alguma parte da estrutura do sistema é conhecida. Caso sejam conhecidos, por exemplo, efeitos inibitórios ou de ativação em alguma reação então os valores dos parâmetros relacionados podem ser restritos a valores negativos ou positivos, respectivamente, o que pode melhorar o tempo de execução da busca (CHOU; MARTENS; VOIT, 2006).

O terceiro cenário envolve a situação extrema na qual não é conhecida nenhuma informação sobre a estrutura do sistema biológico e nenhum dos parâmetros de nenhuma das equações possui restrições específicas (VILELA et al., 2008).

As séries temporais das concentrações para cada um dos metabólitos, assim como as séries temporais das derivadas, serão geradas no intervalo [0;5] utilizando 51 pontos distribuídos em intervalos iguais. A partir daí a estratégia de resolução é a apresentada na Seção 3.1, ou seja, é considerado o desacoplamento do sistema, feito a estimativa das

derivadas e realizada a otimização para se encontrar a estimativa dos parâmetros das equações do sistema.

Serão obtidas soluções para cada uma das quatro equações com três diferentes métodos:

- Regressão Alternada
- Entropia Cruzada
- Híbrido

e em três diferentes configurações para uso dos dados experimentais simulados para as concentrações e suas derivadas, a saber:

1. dados sem ruído e derivadas reais;
2. dados sem ruído e derivadas estimadas;
3. dados com ruído e derivadas estimadas.

Assim, tem-se um total de 36 experimentos, de acordo com a Tabela 4.1. Para cada uma dessas situações serão realizadas 30 execuções.

Tabela 4.1: Experimentos realizados

Equações		X1	X2	X3	X4
Dados sem ruído e Derivadas reais	RA	30 execuções	30 execuções	30 execuções	30 execuções
	EC	30 execuções	30 execuções	30 execuções	30 execuções
	Híbrido	30 execuções	30 execuções	30 execuções	30 execuções
Dados sem ruído e Derivadas estimadas	RA	30 execuções	30 execuções	30 execuções	30 execuções
	EC	30 execuções	30 execuções	30 execuções	30 execuções
	Híbrido	30 execuções	30 execuções	30 execuções	30 execuções
Dados com ruído e Derivadas estimadas	RA	30 execuções	30 execuções	30 execuções	30 execuções
	EC	30 execuções	30 execuções	30 execuções	30 execuções
	Híbrido	30 execuções	30 execuções	30 execuções	30 execuções

Foi realizado um sorteio aleatório para o valor inicial das variáveis para cada uma das equações a serem utilizados em cada um dos 30 experimentos. Por exemplo, para a equação de X_1 foi realizado o sorteio de 30 combinações das variáveis α_1, g_{13}, β_1 e h_{11}, cujos valores corretos são respectivamente 12, -0,8, 10 e 0,5 (de acordo com a Equação 4.3). Essas 30 combinações foram usadas nos experimentos desse metabólito, ou seja, nos três casos (dados sem ruído e derivadas reais, dados sem ruído e derivadas estimadas, dados com ruído e derivadas estimadas) e em cada um dos métodos (Regressão Alternada, Entropia Cruzada e Híbrido). O mesmo procedimento é adotado para as outras

três equações. O sorteio aleatório das variáveis utilizou distribuição uniforme e foi feito considerando o domínio de 1 até 60 para as variáveis α e β e o domínio -2 até 4 para as variáveis $g's$ e $h's$, domínios esses também considerados em outros trabalhos (CHOU; MARTENS; VOIT, 2006).

De acordo com a Tabela 4.1, serão realizados os experimentos para o primeiro caso, nos quais serão utilizadas as séries temporais sem ruído com os valores das derivadas reais (exatos), ou seja, não foi utilizado nenhum método numérico para realizar a estimativa das derivadas como foi apresentado na Seção 3.3. Apesar do uso dos valores exatos para as derivadas não corresponder à estratégia de resolução completa utilizada neste trabalho, serve como experimento adicional para avaliação dos resultados, já que cada um dos algoritmos será testado nessa condição (em 30 execuções).

Um segundo caso a ser considerado nos experimentos fará uso das séries temporais sem ruído mas dessa vez com as derivadas estimadas. No entanto, como a série temporal das concentrações dos metabólitos desse caso não possui ruído então não foi utilizado o método de Savitzky-Golay, apresentado na Seção 3.3.1, pois esse método realiza a filtragem dos dados, tarefa não necessária neste caso. A estimativa das derivadas desse segundo caso foi realizada usando uma função polinomial através de interpolação de splines cúbicas.

No terceiro caso os experimentos utilizarão séries temporais que desta vez terão ruído adicionado. Neste caso, é aplicado o método de Savitzky-Golay para realizar a filtragem dos dados com ruído e a seguir é feita a estimativa das derivadas respectivamente pela derivação do polinômio obtido pelo método de Savitzky-Golay e pelo método do Passo-complexo, apresentados na Seção 3.3. A Figura 4.4 mostra os dados originais e os dados com ruído de 2% de uma das equações.

Em todos os três casos de experimentos descritos acima, os algoritmos que serão testados (RA, EC e Híbrido) recebem como entrada os dados das séries temporais e suas derivadas. Se os dados serão filtrados ou não e se as derivadas serão reais (exatas) ou estimadas, depende de cada caso de experimentos. Dois exemplos seguem abaixo:

1. Primeiro caso de experimentos numéricos e equação de $\frac{dX1}{dt}$: são dados de entrada a série temporal das concentrações sem ruído de $\frac{dX1}{dt}$ e as derivadas reais em cada ponto. Serão feitas 30 execuções com cada um dos algoritmos (RA, EC e Híbrido) e usados diferentes valores para os parâmetros iniciais em cada execução.

2. Terceiro caso de experimentos numéricos e equação $\frac{dX3}{dt}$: são dados de entrada a série

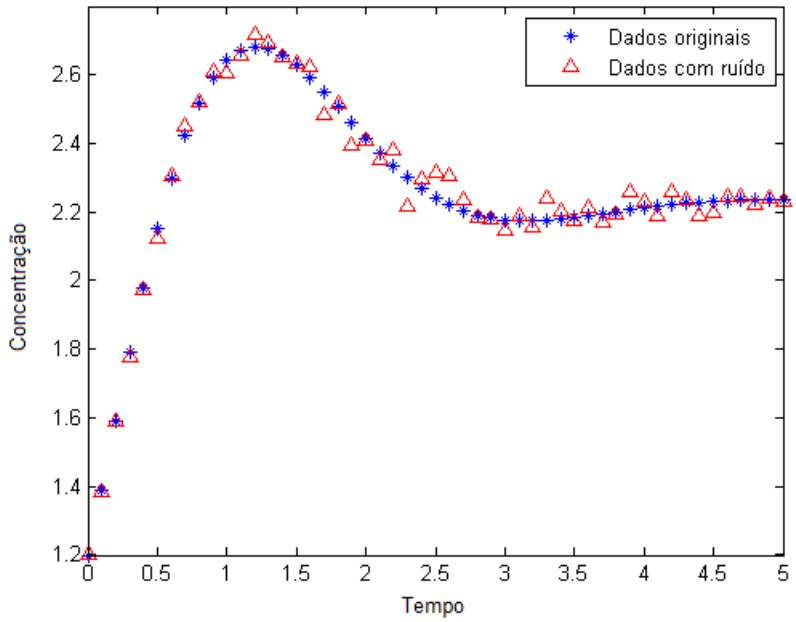


Figura 4.4: Dados originais e com ruído de uma das equações

temporal das concentrações (que originalmente possuía ruído e que já se encontra filtrada com o filtro Savitzky-Golay) para a equação $\frac{dX3}{dt}$ e as derivadas estimadas em cada ponto com o método do passo-complexo. Serão feitas 30 execuções com cada um dos algoritmos (RA, EC e Híbrido) e usados diferentes valores para os parâmetros iniciais em cada execução.

O algoritmo da RA foi executado usando a implementação do paper (BEYER, 2008), que é basicamente a mesma do paper original que apresentou o algoritmo (CHOU; MARTENS; VOIT, 2006). Já o algoritmo da EC foi implementado tomando como base o exemplo Rosenbrock Visualisation (METHOD, 2013) do site original do algoritmo e realizando modificações na função de custo e nas saídas de dados para realização das análises posteriores. A função de custo fez uso do MSE, que será explicado a seguir, para que os indivíduos em uma geração pudessem ser ordenados.

O algoritmo Híbrido foi feito através da combinação dos dois algoritmos anteriores. A única diferença é a definição do critério de parada para o algoritmo da EC, que passa os dados para o algoritmo da RA continuar o processamento. A intenção é que o algoritmo da EC faça uma busca inicial, melhorando os valores dos parâmetros, e portanto faça uso de sua principal vantagem. A partir de certo momento (critério de parada), o algoritmo da EC termina seu processamento e passa os valores otimizados dos parâmetros para

o algoritmo da RA, que por ser uma busca local irá realizar um processamento rápido para refinar os resultados. Dessa forma, o algoritmo da EC não demora a terminar o processamento e o algoritmo da RA diminui a probabilidade de não-convergência, pois os parâmetros obtidos em uma região possivelmente mais adequada pelo algoritmo da EC. A Seção 3.5 apresenta mais informações sobre o método Híbrido.

O critério de parada utilizado neste trabalho para realizar a passagem dos parâmetros do método da EC para o método da RA é o valor do MSE calculado na função de custo. Mas a ordem de grandeza do erro escolhida deve ser diferente da ordem de grandeza definida como critério de parada para as execuções realizadas exclusivamente com o método da EC, pois o objetivo aqui é somente realizar uma busca inicial que melhore os valores dos parâmetros para o método da RA. Em trabalhos futuros ou outros que pretendam adotar esse método Híbrido, um estudo mais sistemático da definição do critério de parada para passagem dos parâmetros deve ser levado em consideração.

Os principais indicadores que serão utilizados para analisar os resultados dos experimentos numéricos no Capítulo 5 serão: tempo de processamento e ordem de grandeza do erro.

O cálculo da função de custo utilizada no método da EC para a avaliação de cada indivíduo da população envolveu os seguintes passos:

1. Considerar apenas uma equação do sistema de cada vez
2. Selecionar todos os pontos da série temporal da equação escolhida
3. Para cada ponto da série temporal, obter o valor da derivada ($S_i(t_k)$, real ou estimado, dependendo do caso)
4. Selecionar o valor de cada uma das variáveis do indivíduo (α, $g's$, β e $h's$)
5. Selecionar o valor das concentrações de cada um dos metabólitos no ponto ($X_i(t_k)$)
6. Substituir o valor das variáveis e o valor das concentrações na equação considerada e encontrar o valor da derivada no ponto
7. Encontrar a diferença (erro) entre o valor da derivada $S_i(t_k)$ e o valor da derivada encontrada no passo anterior

8. Ao final do cálculo do erro para todos os pontos da série temporal, utilizá-los para determinar o MSE

A Seção 3.2 pode ser útil para entender mais facilmente o cálculo da função de custo acima, que é feita considerando o desacoplamento do sistema. Apesar de o método da RA não precisar desse cálculo, foi colocada uma condição para realizá-lo em determinados intervalos de iterações para observar a convergência do mesmo tipo de erro usado no método da EC, facilitando a comparação entre as duas estratégias. O método Híbrido também faz uso desse cálculo quando usa o método da EC e quando usa o método da RA. O MSE é calculado de acordo com a Equação 4.4, sendo y'_j a derivada estimada $S_i(t_k)$, z'_j a derivada calculada no item 6 (a partir dos parâmetros do indivíduo) e N o número de pontos da série temporal.

$$MSE_{individuo} = \frac{1}{N}\sqrt{\sum_{j=1}^{N}(y'_j - z'_j)^2} \tag{4.4}$$

Como o método da EC pode demorar muito a encontrar as variáveis quando os valores iniciais estiverem muito longe dos resultados e o método da RA pode divergir nessa mesma situação, então foram adotados critérios de parada baseados em valor do erro e número de iterações para limitar o processamento durante os experimentos numéricos. Esses critérios de parada serão explicados mais detalhadamente na Seção 5.2.

No Capítulo 5 os resultados dos experimentos aqui descritos serão analisados para avaliar sua adequação aos objetivos propostos.

5 RESULTADOS

Os resultados dos experimentos serão apresentados separadamente para cada uma das etapas estudadas, pois apesar de ambas fazerem parte da mesma estratégia de resolução, os sistemas utilizados para estudo em cada uma delas foram diferentes. O motivo é que o interesse principal neste trabalho não foi a análise completa da estratégia de resolução, mas o estudo de métodos utilizados em cada uma dessas etapas para alcançar melhorias específicas em cada uma delas, consequentemente melhorando a estratégia global.

5.1 Estimativa das derivadas

Os resultados, relacionados aos dados da concentração de X_2 da Equação 4.1, serão apresentados em quatro tabelas, sendo que a Tabela 5.1 e a Tabela 5.2 apresentam os resultados detalhados do erro quadrático médio (da estimativa das derivadas) de acordo com a Equação 4.2 e o desvio padrão, além da diferença entre o erro quadrático médio de cada um dos métodos numéricos utilizados para a diferenciação numérica e o erro quadrático médio para o método analítico polinomial usado no filtro de Savitzky-Golay. Já a Tabela 5.3 e a Tabela 5.4 apresentam, em ordem de grandeza, a diferença do erro quadrático médio de cada método numérico em relação ao erro quadrático médio para o método analítico polinomial. O objetivo do uso da Tabela 5.3 e da Tabela 5.4 é facilitar a interpretação dos resultados, pois a Tabela 5.1 e a Tabela 5.2 apresentam dados numéricos muito extensos e de visualização mais difícil. Por outro lado, essas duas últimas tabelas não poderiam deixar de ser mostradas por apresentarem os resultados detalhados dos experimentos.

Para um melhor esclarecimento, a Tabela 5.3 pode ter a primeira e segunda linhas entendidas como a sexta e nona linhas da Tabela 5.1 (“MSE polinomial - MSE diferença central” e “MSE polinomial - MSE passo complexo”), mas apresentada em ordem de grandeza. Similarmente, a Tabela 5.4 pode ter a primeira e segunda linhas entendidas como a sexta e nona linhas da Tabela 5.2, mas apresentada em ordem de grandeza. A apresentação dos resultados nas Tabelas 5.3 e 5.4, proporciona uma interpretação mais intuitiva do que os resultados das Tabelas 5.1 e 5.2.

A Tabela 5.1 mostra os resultados detalhados para o MSE e desvio padrão das derivadas de todos os pontos da série temporal de 50 pontos (entre os 30 experimentos realizados para cada um dos valores de h) para cada um dos métodos numéricos. Também é mostrado o valor do MSE e desvio padrão para as derivadas de todos os pontos para o método analítico polinomial (primeira e segunda linhas). Nesse último caso, o valor de h não causa nenhuma interferência, pois ele não é utilizado no cálculo. É por esse motivo que é mostrado somente um resultado tanto para o MSE quanto para o desvio padrão.

Tabela 5.1: Resultados para polinomial (Savitzky-Golay), Diferença Central e Passo Complexo - 50 pontos

Polinomial (Savitzky-Golay)	MSE	0.584807983426077						
	Desvio Padrão	0.043823353641649						
Tamanho h		$10^0 = 1$	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-10}	10^{-12}
Diferença Central	MSE	0.699241825049795	0.584815391138832	0.584807984165636	0.584807983746055	0.584807980732690	0.584808247903889	0.584965350943582
	Desvio Padrão	0.046238310855525	0.043823855971677	0.043823353691915	0.043823353651852	0.043823352982358	0.043823828659719	0.043910621114674
	MSE polinomial - MSE diferença central	0.114433841623718	0.000007407712755	0.000000000739559	0.000000000319978	0.000000002693387	0.000000265577812	0.000157367517505
Passo complexo	MSE	0.561808382214613	0.584800576650996	0.584807982685352	0.584807983426003	0.584807983426077	0.584807983426077	0.584807983426077
	Desvio Padrão	0.036007459859114	0.043822851249861	0.043823353591412	0.043823353641643	0.043823353641649	0.043823353641648	0.043823353641649
	MSE polinomial - MSE passo complexo	0.022999601211464	0.000007406775081	0.000000000740725	0.000000000000074	0	0	0

A Tabela 5.2 mostra os mesmos resultados que a Tabela 5.1, mas para a segunda série temporal, com 25 pontos (com espaçamento maior entre cada ponto). Também nesse caso, o tamanho de h não influencia o método analítico, que possui somente um valor. A sexta linha das Tabelas 5.1 e 5.2, que mostra a diferença do MSE do método polinomial em relação ao MSE do método da diferença finita central é mostrada em módulo.

Tabela 5.2: Resultados para polinomial (Savitzky-Golay), Diferença Central e Passo Complexo - 25 pontos

Polinomial (Savitzky-Golay)	MSE	0.645925894524810						
	Desvio Padrão	0.011736663880745						
Tamanho h		2	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-10}	10^{-12}
Diferença Central	MSE	0.670353147319329	0.645928214855420	0.645925894757624	0.645925894679855	0.645925875475002	0.645926578037808	0.645894485338634
	Desvio Padrão	0.010700857479395	0.011736563020267	0.011736663871136	0.011736663840470	0.011736661045356	0.011736568453148	0.011701797368580
	MSE polinomial - MSE diferença central	0.024427252794519	0.000002320330610	0.000000000232814	0.000000000155045	0.000000019049808	0.000000683512998	0.000031409186176
Passo Complexo	MSE	0.624038050955194	0.645923574219597	0.645925894292778	0.645925894524787	0.645925894524810	0.645925894524810	0.645925894524810
	Desvio Padrão	0.012710836030675	0.011736764740606	0.011736663890831	0.011736663880746	0.011736663880745	0.011736663880745	0.011736663880745
	MSE polinomial - MSE passo complexo	0.021887843569616	0.000002320305213	0.000000000232032	0.000000000000023	0	0	0

Dois resultados podem ser observados claramente na Tabela 5.3, relacionados à primeira série temporal. O primeiro é a ocorrência do cancelamento subtrativo para o método da diferença finita central a partir de $h = 10^{-6}$, sendo que além de não apresentar resultados melhores, a precisão passa a piorar a partir daí, quanto menor fica o tamanho de h. Enquanto isso, o método do passo complexo continua melhorando seus resultados e diminuindo o erro quadrático médio. Isso evidencia a superioridade do segundo método em relação ao primeiro para valores muito pequenos de h, ou seja, em problemas discretos nos quais o espaçamento entre pontos consecutivos seja muito pequeno o método do passo complexo consegue continuar melhorando sua precisão cada vez mais, ao contrário do método das diferenças finitas, que a partir de determinado ponto começa a apresentar o problema do cancelamento subtrativo.

Tabela 5.3: Ordem de grandeza do MSE de cada método em relação ao MSE da derivada analítica polinomial - 50 pontos

Tamanho h	$10^0 = 1$	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-10}	10^{-12}
Diferença central	10^{-1}	10^{-6}	10^{-10}	10^{-10}	10^{-9}	10^{-7}	10^{-4}
Passo complexo	10^{-2}	10^{-6}	10^{-10}	10^{-14}	0	0	0

Tabela 5.4: Ordem de grandeza do MSE de cada método em relação ao MSE da derivada analítica polinomial - 25 pontos

Tamanho h	2	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-10}	10^{-12}
Diferença central	10^{-2}	10^{-6}	10^{-10}	10^{-10}	10^{-8}	10^{-7}	10^{-5}
Passo complexo	10^{-2}	10^{-6}	10^{-10}	10^{-14}	0	0	0

O outro resultado que pode ser observado na Tabela 5.3 é que apesar do método do passo complexo melhorar continuamente a medida que o tamanho de h diminui, a aproximação nunca se torna melhor que a aproximação calculada analiticamente. Pode-se observar que a partir de valores de h iguais ou menores a 10^{-8}, os resultados obtidos pelo método são tão bons quanto os resultados obtidos com o método analítico polinomial. No entanto, mesmo para valores tão pequenos de h quanto 10^{-12} os resultados continuam os mesmos. Como já dito anteriormente, os resultados da Tabela 5.3 estão detalhados na Tabela 5.1.

Em relação à segunda série temporal, apesar de numericamente os valores serem diferentes, as observações são exatamente as mesmas realizadas para a primeira série. O maior espaçamento entre os pontos só causou uma piora da precisão do cálculo das deri-

vadas (que pode ser observado com mais detalhe comparando as Tabelas 5.1 e 5.2), mas o comportamento foi exatamente o mesmo em todas as situações.

É possível também reparar que existe diferença numérica nos valores apresentados na Tabela 5.1 e na Tabela 5.2, mas na Tabela 5.3 e na Tabela 5.4, que apresentam a diferença entre o MSE do método analítico polinomial em relação ao MSE dos métodos numéricos, as ordens de grandeza entre as tabelas são próximas. O método das diferenças finitas deixa de apresentar melhora na precisão da derivada a partir de 10^{-6}. Já o método do passo complexo apresenta praticamente os mesmos resultados, melhorando a precisão a medida que h diminui e praticamente se igualando à aproximação da derivada analítica polinomial a partir de 10^{-8}.

Assim, apesar de observado a grande vantagem apresentada pelo método do passo complexo em relação ao método das diferenças finitas para pequenos valores de h, não foi possível observar cálculos mais precisos em relação a derivada calculada pelo método analítico polinomial, por nenhum dos dois métodos. No entanto, é interessante notar que o método do passo complexo consegue uma precisão igual ao método analítico polinomial do filtro de Savitzky-Golay, que faz uso de um polinômio e sua equação derivada.

5.2 Otimização para estimativa de parâmetros

Os experimentos numéricos da etapa de otimização serão analisados separadamente para cada um dos três casos. Inicialmente serão analisados os resultados para o caso que considerou os dados sem ruído e as derivadas reais.

5.2.1 Primeiro caso de experimentos numéricos

Foram adotados critérios de parada para os métodos da EC e da RA baseados nos erros calculados em ambos os algoritmos e no número de iterações em cada um deles. Para o método da EC que faz o cálculo do MSE conforme explicado na Seção 4.2 (Equação 4.4), o critério de parada definido foi um máximo de 2.000 iterações ou MSE na ordem de 10^{-8}. A população utilizada foi de 2.000 indivíduos. Para o método da RA que faz o cálculo do erro conforme Seção 3.4.1, o critério de parada definido foi um máximo de 4.000.000 iterações ou log(erro) menor do que -15 (que gerava um MSE na ordem de 10^{-7}).

A Tabela 5.5 mostra a quantidade de rodadas no primeiro caso, das 30 realizadas,

nas quais o erro alcançou valores da ordem de 10^{-6} ou menos. Essa ordem de grandeza do erro apresenta valores que já se aproximam bastante dos valores reais das variáveis, como será mostrado mais adiante. Essa tabela mostra a quantidade de resultados que alcançaram essa ordem de grandeza no erro na busca dos parâmetros para cada uma das equações com cada um dos métodos.

A linha com os resultados para o método da RA mostra que o método alcançou mais ou menos a mesma quantidade de resultados para cada uma das equações, mas com maior quantidade para a equação da concentração de X_1. Os arquivos de saída com os resultados detalhados de cada rodada mostram que quando a convergência não acontece para esse método os motivos foram a divergência nos resultados, ou seja, foram encontrados resultados errados, e o fim da execução pelo critério de parada de número máximo de iterações com valores encontrados longe dos resultados reais, o que também ocorreu em Beyer (2008) e Chou, Martens e Voit (2006) e foi descrito na Seção 4.2. Isso foi observado para todas as quatro equações que o método da RA executou.

Tabela 5.5: Resultados dos experimentos para o primeiro caso

Equações		X1	X2	X3	X4
Erro $\leqslant 10^{-6}$	RA	27	24	21	24
	EC	30	19	01	11
	Híbrido	30	30	30	30

Já para o método da EC, os resultados mostram que para a equação da concentração de X_1 todos os parâmetros foram encontrados com o erro alcançando os valores dos critérios de parada definidos. Na verdade, grande parte das rodadas desse caso alcançou o erro na ordem de grandeza de 10^{-8} antes do número máximo de iterações. No entanto, para as outras equações houveram resultados que não alcançaram a ordem de grandeza de 10^{-6} até chegar ao critério de parada do número de iterações. Nas equações das concentrações de X_2 e X_4, o motivo principal observado para que isso ocorresse foi o fato de que a velocidade de convergência do erro diminui gradativamente a medida que o número de iterações aumenta, dessa forma tornando a convergência mais lenta e fazendo com que o algoritmo finalize pelo motivo do critério de parada que define o número máximo de 2.000 iterações. A Figura 5.1 mostra um gráfico típico de convergência de erro para o método da EC. O erro não chega ao valor 0, mas após grande queda nas iterações iniciais torna-se pequeno e passa a convergir mais lentamente.

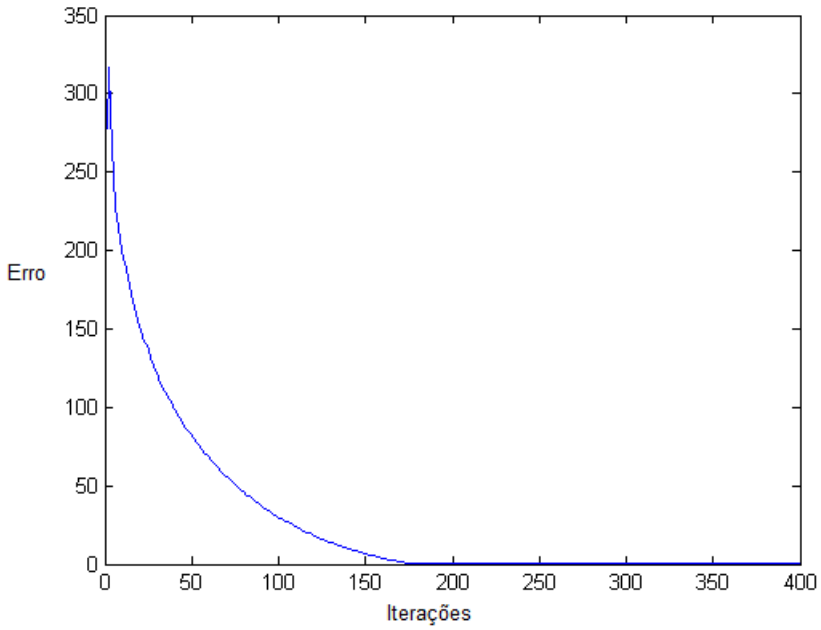


Figura 5.1: Erro do método da EC em uma das rodadas (convergente)

Em relação à equação da concentração de X_3, que obteve apenas 1 resultado que alcançou o erro na ordem de 10^{-6} para o método da EC, foi observada outra situação, além de casos como o descrito no parágrafo anterior. Em aproximadamente metade das 30 rodadas ocorreu uma parada na diminuição do erro da função de custo, que passou a oscilar ligeiramente para mais ou para menos, mas deixou de diminuir continuamente e consequentemente impediu que os valores buscados se aproximassem dos resultados reais.

Ao observar o comportamento descrito acima surgiu a preocupação em relação à eficácia do método da EC aplicado ao contexto de busca de parâmetros para equações representadas por Sistema S. Ao encontrar o artigo de Chan e Kroese (2012) no qual são descritos problemas que poderiam ocorrer com o método em algumas situações com grande número de variáveis se tornou necessário realizar alguns testes para confirmar a eficácia do método.

Primeiramente foi desenvolvido um algoritmo no qual poderia ser configurado em seu início qualquer número de dimensões e qualquer domínio para os valores das variáveis nas dimensões que seriam usadas. A função de custo utilizada foi a função quadrática, cujo resultado é sempre o vetor nulo, independente do número de dimensões. Ou seja, se fossem definidas 6 dimensões então o algoritmo deveria achar o resultado (0,0,0,0,0,0). Diversos testes foram realizados e em todos eles os resultados foram encontrados com sucesso. Um dos testes chegou a definir 100 variáveis e mesmo assim o resultado correto

foi encontrado e não foi observada parada na convergência do erro.

Para confirmar que o método não possui problema com o número de variáveis ao se utilizar Sistema S foram então implementados dois outros sistemas encontrados em Vilela et al. (2008) que também possuem 5 variáveis em suas equações. Um dos sistemas possui duas equações:

$$\begin{aligned} \frac{dX_1}{dt} &= 3{X_2}^{-2} - {X_1}^{0.5}X_2, \\ \frac{dX_2}{dt} &= {X_1}^{0.5}X_2 - {X_2}^{0.5} \end{aligned} \tag{5.1}$$

E o outro sistema possui cinco equações:

$$\begin{aligned} \frac{dX_1}{dt} &= 5X_3{X_5}^{-1} - 10{X_1}^2, \\ \frac{dX_2}{dt} &= 10{X_1}^2 - 10{X_2}^2, \\ \frac{dX_3}{dt} &= 10{X_2}^{-1} - 10{X_2}^{-1}{X_3}^2, \\ \frac{dX_4}{dt} &= 8{X_3}^2{X_5}^{-1} - 10{X_4}^2 \\ \frac{dX_5}{dt} &= 10{X_4}^2 - 10{X_5}^2 \end{aligned} \tag{5.2}$$

Usando as condições iniciais informadas em Vilela et al. (2008) e testando os dois sistemas, foi observado que os resultados foram encontrados na maioria das execuções pelo algoritmo mesmo para 5 variáveis para a realização da busca, ou seja, as duas do sistema de duas equações e as concentrações de X_1, X_3 e X_4 do sistema de cinco equações.

Outros testes foram feitos com um sistema de 30 equações (30 elementos bioquímicos) que é descrito por Kimura et al. (2005) e Maki et al. (2001). O sistema pode ser visto na Equação 5.3. Usando o mesmo cenário que considera a estrutura do sistema conhecida e a busca dos parâmetros para conhecer a dinâmica do sistema bioquímico, foram avaliadas várias equações do sistema, sendo algumas delas com 5 variáveis e outras com 6 variáveis (por exemplo as concentrações de X_{11}, X_{24} e X_{27}). Várias rodadas foram executadas e na maioria delas o algoritmo convergiu em direção aos resultados corretos.

Portanto, o método da EC parece funcionar na resolução de equações representadas por Sistema S mas especificamente para a equação de $\frac{dX_3}{dt}$ do sistema estudado neste trabalho o algoritmo encontrou dificuldades e apresentou estagnação na convergência do erro. Sendo assim, foi feito um estudo do algoritmo que foi implementado e dos arquivos de saída detalhados das 30 rodadas do experimento que executou o método da EC para

essa equação no primeiro caso de experimentos numéricos.

$$
\begin{aligned}
\frac{dX_1}{dt} &= 1{X_{14}}^{-0.1} - 1{X_1}^1, & \frac{dX_{16}}{dt} &= 1{X_{11}}^{0.5}{X_{12}}^{-0.2} - 1{X_{16}}^1 \\
\frac{dX_2}{dt} &= 1 - 1{X_2}^1, & \frac{dX_{17}}{dt} &= 1{X_{13}}^{0.5} - 1{X_{17}}^1 \\
\frac{dX_3}{dt} &= 1 - 1{X_3}^1, & \frac{dX_{18}}{dt} &= 1 - 1{X_{18}}^1 \\
\frac{dX_4}{dt} &= 1 - 1{X_4}^1 & \frac{dX_{19}}{dt} &= 1{X_{14}}^{0.1} - 1{X_{19}}^1 \\
\frac{dX_5}{dt} &= 1{X_1}^1 - 1{X_5}^1 & \frac{dX_{20}}{dt} &= 1{X_{15}}^{0.7}{X_{26}}^{0.3} - 1{X_{20}}^1 \\
\frac{dX_6}{dt} &= 1{X_1}^1 - 1{X_6}^1 & \frac{dX_{21}}{dt} &= 1{X_{16}}^{0.6} - 1{X_{21}}^1 \\
\frac{dX_7}{dt} &= 1{X_2}^{0.5}{X_3}^{0.4} - 1{X_7}^1 & \frac{dX_{22}}{dt} &= 1{X_{16}}^{0.5} - 1{X_{22}}^1 \\
\frac{dX_8}{dt} &= 1{X_4}^{0.2}{X_{17}}^{-0.2} - 1{X_8}^1 & \frac{dX_{23}}{dt} &= 1{X_{17}}^{0.2} - 1{X_{23}}^1 \\
\frac{dX_9}{dt} &= 1{X_5}^1{X_6}^{-0.1} - 1{X_9}^1 & \frac{dX_{24}}{dt} &= 1{X_{15}}^{0.2}{X_{18}}^{-0.1}{X_{19}}^{0.3} - 1{X_{24}}^1 \\
\frac{dX_{10}}{dt} &= 1{X_7}^{0.3} - 1{X_{10}}^1 & \frac{dX_{25}}{dt} &= 1{X_{20}}^{0.4} - 1{X_{25}}^1 \\
\frac{dX_{11}}{dt} &= 1{X_4}^{0.4}{X_7}^{-0.2}{X_{22}}^{0.4} - 1{X_{11}}^1 & \frac{dX_{26}}{dt} &= 1{X_{21}}^{-0.2}{X_{28}}^{0.1} - 1{X_{26}}^1 \\
\frac{dX_{12}}{dt} &= 1{X_{23}}^{0.1} - 1{X_{12}}^1 & \frac{dX_{27}}{dt} &= 1{X_{24}}^{0.6}{X_{25}}^{0.3}{X_{30}}^{-0.2} - 1{X_{27}}^1 \\
\frac{dX_{13}}{dt} &= 1{X_8}^{0.6} - 1{X_{13}}^1 & \frac{dX_{28}}{dt} &= 1{X_{25}}^{0.5} - 1{X_{28}}^1 \\
\frac{dX_{14}}{dt} &= 1{X_9}^1 - 1{X_{14}}^1 & \frac{dX_{29}}{dt} &= 1{X_{26}}^{0.4} - 1{X_{29}}^1 \\
\frac{dX_{15}}{dt} &= 1{X_{10}}^{0.2} - 1{X_{15}}^1 & \frac{dX_{30}}{dt} &= 1{X_{27}}^{0.6} - 1{X_{30}}^1
\end{aligned}
\tag{5.3}
$$

O algoritmo EC implementado possui diversos parâmetros que controlam por exemplo o número de indivíduos da população, percentual de indivíduos melhores classificados utilizados para gerar a população seguinte, controle de atualização de médias, controle de atualização de variâncias e controle da diminuição da amplitude da função de distribuição de probabilidade. No entanto, não foi realizado um estudo sistemático em relação aos melhores valores para cada parâmetro. Também não houve mudança nos parâmetros para nenhuma execução nos testes.

Para avaliar a influência desses parâmetros foram realizados alguns testes com alterações aleatórias nos valores dos parâmetros. Os resultados mostraram influência tanto na velocidade de convergência do erro quanto na ordem de grandeza alcançada até o critério

de parada definido (2000 iterações).

Assim, concluiu-se que os valores dos parâmetros utilizados nos experimentos numéricos com o método da EC não foram favoráveis para a realização da busca das variáveis da equação da concentração de X_3 do sistema estudado, já que para as outras equações desse sistema e para as equações dos outros dois sistemas implementados não houve problema para encontrar os resultados. Além disso, ao alterar os valores dos parâmetros em testes aleatórios os resultados para essa mesma equação foram diferentes e geralmente melhores.

Portanto, considerando os resultados da Tabela 5.5 para o método da EC para todas as equações (principalmente para a concentração de X_3), os testes adicionais realizados e o estudo dos parâmetros utilizados na implementação do algoritmo, podemos sugerir que uma estratégia que utilize parâmetros adaptativos, se alterando em situações como estagnação na convergência do erro, entre outras, pode ser mais eficiente para a aplicação do EC nesse tipo de sistema. É importante notar que caso essa alteração seja vantajosa, também terá influência nas soluções do método Híbrido, que faz uso do EC em seu início.

Por sua vez, o método Híbrido mostrou os melhores resultados e confirmou as expectativas em relação à sua eficácia. Foram encontrados todos os resultados em todas as rodadas para todas as quatro equações do sistema. O motivo disso é que mesmo quando o método da EC não é capaz de encontrar o resultado correto dos valores das variáveis de determinada equação, ele faz o trabalho de aproximar o valor de cada uma das variáveis para o seu resultado correto, permitindo que quando esses valores sejam passados para o método da RA, ele então consiga encontrar os resultados de maneira adequada. Mesmo nas situações nas quais o método da EC não obteve sucesso sozinho em grande número de execuções, como no caso da equação da concentração de X_3, durante a estratégia híbrida ele funciona como um direcionador do valor das variáveis, que chegam para o RA com valores favoráveis para que sua busca seja bem sucedida.

A Tabela 5.6 mostra os valores corretos de cada uma das variáveis do sistema e a seguir os valores encontrados por cada um dos algoritmos em uma execução escolhida aleatoriamente e que tenha obtido bons resultados nesse primeiro caso.

Em relação ao tempo de execução, a Tabela 5.7 mostra a média para as rodadas bem sucedidas de cada experimento numérico, ou seja, a média do tempo para as rodadas que fazem parte dos resultados mostrados na Tabela 5.5. Conforme esperado, o método da RA apresentou as médias de tempo mais baixas, mas é importante lembrar que nem todas as

Tabela 5.6: Resultados de execuções específicas de cada método - Primeiro caso

		α_i	g_{i1}	g_{i2}	g_{i3}	g_{i4}	β_i	h_{i1}	h_{i2}	h_{i3}	h_{i4}
Valores corretos	X_1	12	0	0	−0.8	0	10	0.5	0	0	0
	X_2	8	0.5	0	0	0	3	0	0.75	0	0
	X_3	3	0	0.75	0	0	5	0	0	0.5	0.2
	X_4	2	0.5	0	0	0	6	0	0	0	0.8
RA	X_1	12.03568	0	0	−0.79369	0	10.04818	0.49646	0	0	0
	X_2	8.01474	0.49873	0	0	0	3.01326	0	0.74795	0	0
	X_3	2.96045	0	0.75773	0	0	4.97968	0	0	0.50806	0.20528
	X_4	2.00628	0.49654	0	0	0	5.98289	0	0	0	0.79527
EC	X_1	12.01044	0	0	−0.80270	0	9.99711	0.50121	0	0	0
	X_2	7.98943	0.50068	0	0	0	2.99222	0	0.75106	0	0
	X_3	2.83112	0	0.77810	0	0	4.82884	0	0	0.52328	0.21137
	X_4	2.01186	0.49313	0	0	0	5.96554	0	0	0	0.79076
Híbrido	X_1	11.96561	0	0	−0.80622	0	9.95320	0.50348	0	0	0
	X_2	8.01474	0.49873	0	0	0	3.01327	0	0.74795	0	0
	X_3	3.04100	0	0.74215	0	0	5.02168	0	0	0.49187	0.19470
	X_4	2.00628	0.49654	0	0	0	5.98289	0	0	0	0.79527

rodadas encontraram o resultado. Pode ser observado que a ordem de grandeza do tempo para o método da EC é quase sempre a mesma que a ordem de grandeza do método da RA, sendo a única exceção a equação de $\frac{dX_3}{dt}$, que é uma ordem de grandeza superior. Além disso, o método Híbrido se mostrou mais rápido do que o método da EC em todas as equações, e se manteve na mesma ordem de grandeza do método da RA também para todas as equações. Em especial, a equação de $\frac{dX_1}{dt}$ apresentou resultados para o método Híbrido que ficaram muito próximos do método da RA.

Tabela 5.7: Tempo médio para as rodadas bem sucedidas (em segundos)

Equações		X1	X2	X3	X4
Experimentos do primeiro caso	RA	13.32	21.56	13.48	10.66
	EC	83.76	81.59	161.27	47.93
	Híbrido	18.34	44.07	99.45	41.59

Na Tabela 5.8 são apresentados os tempos médios das rodadas para o método Híbrido, mas de forma separada para o método da EC e para o método da RA. A soma dos tempos apresentados nesta tabela para cada uma das equações é igual ao valor mostrado na Tabela 5.7 para o método Híbrido. Pode-se perceber que a média do tempo de processamento para o método da EC é menor do que a média do tempo para o mesmo método que é apresentada na Tabela 5.7, que tenta resolver o problema. Isso se deve ao fato de que o valor do erro considerado para o critério de passagem do método da EC para o método da RA no método Híbrido é da ordem de 10^{-3}, pois o objetivo é conseguir que o valor das variáveis passado pelo método da EC fique favorável ao processamento pelo método da RA. Ao contrário, na resolução pelo método da EC o objetivo é encontrar os resultados corretos, por isso o critério de parada (além do número máximo de iterações) é o erro na

ordem de 10^{-8}, tornando a média do tempo maior.

Tabela 5.8: Tempo médio (em segundos)

Equações		X1	X2	X3	X4
Híbrido no primeiro caso	EC	11.16	43.45	72.02	17.73
	RA	7.18	0.62	27.43	23.86

Outra informação que pode ser observada na Tabela 5.8 é que a média dos tempos do método da RA para as equações das concentrações de X_1 e X_2 são menores do que a média dos tempos para as equações das concentrações de X_3 e X_4. Ao fazer uma análise amostral de alguns arquivos de saída das rodadas para os experimentos numéricos fica evidente que isso se deve ao fato de o método da EC chegar ao critério de parada para a passagem das variáveis do método da EC para o método da RA com valores favoráveis para as duas primeiras equações, o que torna o processamento do segundo método mais rápido. De fato, se for observada a Tabela 5.5 para o método da EC, as duas primeiras equações conseguiram um número bem maior de resultados favoráveis do que as duas últimas, o que acabou se repetindo nos experimentos do método Híbrido e causou um maior tempo de processamento para o método da RA para as duas últimas equações após o recebimento das variáveis.

Dessa forma, considerando esse primeiro caso de experimentos numéricos, com dados sem ruído e derivadas reais, o método da EC mostrou-se eficaz na busca dos parâmetros do sistema, assim como a estratégia híbrida se mostrou também bem sucedida e mais eficiente do que cada um dos outros dois métodos isoladamente.

5.2.2 Segundo caso de experimentos numéricos

Para o segundo caso de experimentos numéricos, com dados sem ruído e derivadas estimadas através de splines cúbicas (pois não existe ruído), a Tabela 5.9 mostra os resultados para o número de execuções corretas para cada método.

Nesse caso fica evidente a influência do desacoplamento nesse tipo de estratégia de resolução, pois como o valor das derivadas é estimado então os algoritmos não conseguem encontrar os valores corretos das variáveis para o critério de convergência adotado no primeiro caso, ou seja, MSE de ordem 10^{-8}, o que acarretou a consequência de que praticamente todas as rodadas para todos os algoritmos terminassem pelo critério de parada

do número de iterações. Por isso, a Tabela 5.9 mostra os resultados que apresentaram MSE igual ou menor a 10^{-4}.

Tabela 5.9: Resultados dos experimentos para o segundo caso

Equações		X1	X2	X3	X4
Erro $\leqslant 10^{-4}$	RA	28	23	21	24
	EC	30	04	02	07
	Híbrido	30	30	30	27

Todos os resultados para o algoritmo da RA são parecidos aos do primeiro caso e ao estudar os arquivos de saída das rodadas pode-se observar que os motivos dos resultados também são os mesmos, ou seja, quando o resultado não é alcançado ocorreu divergência no resultado ou o critério de parada de número máximo de iterações foi alcançado.

Para o método da EC foi observado o mesmo comportamento para a equação da concentração de X_3, mas as equações das concentrações de X_2 e X_4 tiveram poucos resultados bons. Observando os arquivos de saída percebe-se que o principal motivo foi o critério de parada de 2.000 iterações e a diminuição da velocidade de convergência do erro, pois o algoritmo estava caminhando na direção de bons resultados na maioria das execuções, ou seja, mesmo comportamento observado no primeiro caso na Tabela 5.5.

O método Híbrido mais uma vez foi eficaz (considerando a nova ordem de grandeza do erro) pois da mesma forma que no primeiro caso, fez uso do método da EC para colocar os valores das variáveis favoráveis ao método da RA, que termina a execução. A exceção foi a equação da concentração de X_4, pois três resultados não convergiram. Isso se deveu ao fato de o método da EC ter apresentado comportamento similar a algumas execuções da equação da concentração de X_3, ou seja, a convergência do erro ficou estagnada durante a execução, o que causou a parada pelo critério de iterações e impediu que a fossem passados parâmetros favoráveis para o método da RA.

Pode ser observado na Tabela 5.10 que os resultados encontrados para o método da RA e para o método Híbrido são iguais. Isso se deve ao fato de que o método Híbrido sempre termina com a execução do método da RA, e esse algoritmo, quando converge, quase sempre encontra os mesmos valores. Resultados semelhantes foram observados na Tabela 5.6.

Mesmo considerando as três execuções que não foram bem sucedidas para a equação da concentração de X_4, a maioria dos resultados mostra que o método Híbrido é eficiente

Tabela 5.10: Resultados de execuções específicas de cada método - Segundo caso

		α_i	g_{i1}	g_{i2}	g_{i3}	g_{i4}	β_i	h_{i1}	h_{i2}	h_{i3}	h_{i4}
Valores corretos	X_1	12	0	0	−0.8	0	10	0.5	0	0	0
	X_2	8	0.5	0	0	0	3	0	0.75	0	0
	X_3	3	0	0.75	0	0	5	0	0	0.5	0.2
	X_4	2	0.5	0	0	0	6	0	0	0	0.8
RA	X_1	11.40442	0	0	−0.87252	0	9.35664	0.54619	0	0	0
	X_2	8.09491	0.47969	0	0	0	3.15175	0	0.72344	0	0
	X_3	3.44587	0	0.66881	0	0	5.16494	0	0	0.42542	0.14394
	X_4	2.00258	0.51118	0	0	0	6.05698	0	0	0	0.8096
EC	X_1	13.86100	0	0	−0.55104	0	12.46395	0.36404	0	0	0
	X_2	7.44661	0.52431	0	0	0	2.65226	0	0.79413	0	0
	X_3	3.82009	0	0.60578	0	0	5.27375	0	0	0.35304	0.09428
	X_4	2.42589	0.35991	0	0	0	5.57318	0	0	0	0.59649
Híbrido	X_1	11.40442	0	0	−0.87252	0	9.35664	0.54619	0	0	0
	X_2	8.09491	0.47969	0	0	0	3.15175	0	0.72344	0	0
	X_3	3.44587	0	0.66881	0	0	5.16494	0	0	0.42542	0.14394
	X_4	2.00258	0.51118	0	0	0	6.05698	0	0	0	0.80962

na resolução desse tipo de problema. Portanto, nesse segundo caso de experimentos numéricos, os resultados parecem confirmar a eficácia dos métodos da EC e Híbrido para busca de parâmetros de equações representadas por Sistema S. No entanto, os resultados sugerem uma forte influência da etapa de desacoplamento (que faz parte da estratégia de resolução adotada) nos resultado, pois a estimativa das derivadas, que são utilizadas na etapa de otimização causou mudanças não somente nos resultados mas também na convergência do erro, que foi consideravelmente mais lenta que no primeiro caso dos experimentos, com derivadas reais.

5.2.3 Terceiro caso de experimentos numéricos

Para o terceiro caso de experimentos, dados com ruído de 2% e derivadas estimadas, foi feito o uso do filtro de Savitzky-Golay para suavizar os dados e do método do passo complexo (com intervalo $h = 10^{-10}$) para estimar as derivadas. Os resultados são apresentados na Tabela 5.11 e mostram que o uso de derivadas estimadas e dados com ruído influenciaram os resultados alcançados por todos os métodos. Especialmente as execuções do método da RA para a equação da concentração de X_3 apresentaram resultados que divergiram em praticamente todas as rodadas, apesar do erro terminar dentro da ordem de 10^{-4}. O mesmo comportamento sendo observado para o método Híbrido, que termina sua execução com o método da RA.

Em outras situações, como nos experimentos do método da EC com as equações das concentrações de X_2 e X_3 foi observado que apesar dos resultados não alcançarem o erro na ordem de 10^{-4}, algumas rodadas se aproximaram bastante dos resultados quando o

Tabela 5.11: Resultados dos experimentos para o terceiro caso

Equações		X1	X2	X3	X4
Erro $\leqslant 10^{-4}$	RA	26	00	23	27
	EC	30	00	00	30
	Híbrido	30	00	29	28

erro ainda estava na ordem de 10^{-3}. Nesses casos, para cada uma das equações, foi observado que tanto os resultados que se aproximaram dos valores corretos quanto os que não se aproximaram ficaram com o erro na ordem de 10^{-3} e tiveram a aproximação dos resultados definidas por ordens de grandezas menores que essas. As Tabelas 5.12 e 5.13 mostram essa situação.

Tabela 5.12: Resultados com valores diferentes para o método da EC - Terceiro caso

		α_i	g_{i1}	β_i	h_{i2}
Valores corretos	X_2	8	0.5	3	0.75
Erro: 0.002911038375273	X_2	8.08076	0.49298	3.07406	0.73858
Erro: 0.003188788715785	X_2	9.10085	0.42690	3.95375	0.63367

Para os resultados do método da RA e do método Híbrido para a equação da concentração de X_2 foi observado o mesmo tipo de comportamento mostrado nas Tabelas 5.12 e 5.13, ou seja, o erro não alcançou a ordem de 10^{-4} mas apresentou aproximações razoáveis para erros na ordem de 10^{-3}.

Tabela 5.13: Resultados com valores diferentes para o método da EC - Terceiro caso

		α_i	g_{i2}	β_i	h_{i3}	h_{i4}
Valores corretos	X_3	3	0.75	5	0.5	0.2
Erro: 0.003772114783522	X_3	3.02118	0.73119	4.79400	0.46656	0.16768
Erro: 0.004048045410152	X_3	10.53101	0.27829	12.00662	0.13789	0.02464

Além disso, os resultados usando o método da EC para a equação da concentração de X_4 não mostraram valores com boas aproximações quando o erro estava com ordem 10^{-4}, mas os resultados melhoraram ao chegar em 10^{-5}, o que pode ser observado na Tabela 5.14.

Os resultados mostram que o método da RA apresenta sensibilidade para dados que originalmente possuíam ruído e foram filtrados, o que foi observado também em outros trabalhos (BEYER, 2008). Nos experimentos do terceiro caso, onde além dos dados com ruído, foram utilizadas as derivadas estimadas, provavelmente esses dois fatores afetaram o resultado do algoritmo, encontrando valores que não foram muito bons principalmente

Tabela 5.14: Valores para erros em ordens de grandeza diferentes para o método da EC - Terceiro caso

		α_i	g_{i1}	β_i	h_{i4}
Valores corretos	X_4	2	0.5	6	0.8
Erro: 0.000148030	X_4	7.51197	0.08555	9.48508	0.15969
Erro: 0.000021665	X_4	2.16104	0.42924	5.70493	0.70074

para as equações das concentrações de X_2 e de X_3, como pode ser observado na Tabela 5.15, que mostra o resultado de algumas execuções selecionadas aleatoriamente para cada um dos algoritmos para o terceiro caso.

Portanto, para esse terceiro caso de experimentos numéricos, os resultados mostraram comportamentos com um maior nível de variação em comparação com os dois primeiros casos. Boas aproximações para os valores das variáveis em cada uma das equações foram encontradas em diferentes ordens de grandeza para o erro, o que dificultou um pouco as análises e tornou a Tabela 5.11 não tão intuitiva quanto as Tabelas 5.5 e 5.9. As diferenças de resultados exibidos nas Tabelas 5.12, 5.13 e 5.14 mostram o grande grau de variação encontrado nesse terceiro caso para os experimentos numéricos. No entanto, de forma geral, o método da EC apresentou comportamento similar aos dois primeiros casos, com o erro sendo minimizado e as variáveis convergindo em direção aos resultados corretos, apesar de isso ter acontecido em diferentes ordens de grandeza em cada uma das equações. Por sua vez, o método da RA mostrou dificuldade em trabalhar com os dados que originalmente possuíam ruído e foram filtrados e com as derivadas estimadas, o que já havia sido observado em outros trabalhos (BEYER, 2008) refletindo, também, nos resultados do método Híbrido.

Tabela 5.15: Resultados de execuções específicas de cada método - Terceiro caso

		α_i	g_{i1}	g_{i2}	g_{i3}	g_{i4}	β_i	h_{i1}	h_{i2}	h_{i3}	h_{i4}
Valores corretos	X_1	12	0	0	−0.8	0	10	0.5	0	0	0
	X_2	8	0.5	0	0	0	3	0	0.75	0	0
	X_3	3	0	0.75	0	0	5	0	0	0.5	0.2
	X_4	2	0.5	0	0	0	6	0	0	0	0.8
RA	X_1	11.59258	0	0	−0.85682	0	9.52692	0.53449	0	0	0
	X_2	6.95608	0.61402	0	0	0	2.06599	0	0.93682	0	0
	X_3	−0.36043	0	0.49273	0	0	0.45456	0	0	−0.30117	−0.66487
	X_4	1.88274	0.58081	0	0	0	6.45492	0	0	0	0.90713
EC	X_1	11.96738	0	0	−0.76795	0	10.11054	0.48641	0	0	0
	X_2	8.08076	0.49298	0	0	0	3.07406	0	0.48641	0	0
	X_3	3.02118	0	0.73119	0	0	4.79400	0	0	0.46656	0.16768
	X_4	2.16176	0.42582	0	0	0	5.67491	0	0	0	0.69613
Híbrido	X_1	11.59258	0	0	−0.85682	0	9.52692	0.53449	0	0	0
	X_2	6.95608	0.61402	0	0	0	2.06599	0	0.93682	0	0
	X_3	−0.36043	0	0.49273	0	0	0.45456	0	0	−0.30117	−0.66487
	X_4	1.88274	0.58081	0	0	0	6.45492	0	0	0	0.90713

5.3 Considerações

Na etapa de estimativa das derivadas, foi estudado o comportamento do método do passo complexo, que pode ser utilizado para realizar aproximações numéricas de derivadas. Foi feita uma comparação com o método das diferenças finitas, que apresenta o problema do cancelamento subtrativo em suas aproximações quando são usados intervalos muito pequenos entre os pontos. O método do passo complexo, além de simples implementação, não possui o problema do cancelamento subtrativo, podendo encontrar derivadas com qualidade consideravelmente superior. A qualidade da derivada estimada pelos métodos é extremamente importante para a estratégia global de resolução do problema inverso, pois influencia o resultado da busca na etapa de otimização. Por esse motivo, foi interessante avaliar a superioridade do método do passo complexo em relação ao método das diferenças finitas para pequenos intervalos entre os pontos, o que é evidente pelos resultados obtidos. A possibilidade de se utilizar um método numérico de grande precisão, simples implementação e que evita o problema do cancelamento subtrativo faz com que o método do passo complexo seja uma excelente ferramenta, que apesar de pouco conhecida pode ser mais explorada em estudos numéricos.

Em relação à etapa de otimização para estimativa dos parâmetros das equações, visando avaliar o desempenho do método da Entropia Cruzada e do método Híbrido, optou-se por utilizar um modelo específico para obtenção de parâmetros de um sistema S conhecido como método da Regressão Alternada, método este reconhecido como bastante eficaz no trato desta classe de problemas. Os resultados do método da RA mostraram uma excelente eficácia quando adota-se valores iniciais adequados para as variáveis do termo de degradação, visto que executa uma busca local na otimização.

Foi observado que o tempo de processamento de todos os métodos aumentou com a sequência dos casos avaliados, chegando ao ponto de todas as execuções terminarem no terceiro caso (dados com ruído e derivadas estimadas pelo método do passo complexo) por causa do critério de parada de limitação no número de iterações. Para o algoritmo da EC, assim como nos dois primeiros casos, foi observado no arquivo de saída de cada execução que os valores das variáveis convergiam em direção ao resultado, mas nunca terminaram com erros na ordem de 10^{-8}, mostrando que o algoritmo é eficiente para a solução desse tipo de problema mas terminou devido ao critério de parada.

Em algumas situações nas quais houve um desempenho de menor eficiência para o

método da EC, com a estagnação da convergência do erro, a configuração dos parâmetros podem ter influenciado os resultados, por isso uma implementação do método em conjunto com uma estratégia adaptativa para alguns parâmetros do algoritmo, apresenta-se como uma possível forma de se aumentar o desempenho e a robustez do método em relação a convergência. Ressalta-se que, entre futuros desenvolvimentos, pretende-se apresentar e implementar tais modelos adaptativos.

Um dos principais desenvolvimentos deste trabalho foi o modelo denominado como método Híbrido, visando associar as propriedades de busca global inerente ao método da EC com as características de busca local do método da RA. Algumas modificações, como definição do critério de parada para passagem de parâmetros de um algoritmo para o outro, podem ser feitas de forma mais sistemática do que a adotada neste trabalho, o que pode melhorar mais ainda seus resultados.

Um dos problemas observados, que veio confirmar descrições já encontradas na literatura (CHOU; MARTENS; VOIT, 2006), é a influência nos resultados da estimativa das derivadas, realizada na fase de desacoplamento do sistema. Na evolução dos testes realizados nos três diferentes casos de experimentos numéricos ficou evidente sua influência nos resultados, que pioraram sensivelmente mesmo quando os dados não tinham ruído (segundo caso) e chegaram a não encontrar resultados com boa aproximação por causa do critério de parada de limitação do número de iterações no terceiro caso, quando a estimativa foi realizada em dados com ruído. Assim, por mostrar uma grande influência nos resultados, sugere-se que estudos adicionais sejam necessários visando aprimorar o uso dessa estratégia com desacoplamento e estimativa das derivadas.

Considerando os resultados dos três casos de experimentos numéricos realizados, o método da EC demonstra bons resultados na busca dos parâmetros de sistemas representados por Sistema S, mostrando sua eficácia nesse tipo de problema. O método Híbrido também se mostrou eficaz e eficiente nos dois primeiros casos, apresentando resultados muito superiores aos dois outros métodos individuais que serviram de base para sua construção. No entanto, o terceiro caso, com condições de dados com ruído e derivada estimada, mostrou resultados que tornaram sua avaliação inconclusiva, pois os problemas encontrados com o método da RA fizeram seus resultados apresentarem uma perda de qualidade. Durante a evolução dos três casos de experimentos ficou evidente a influência da estimativa das derivadas na etapa de otimização e consequentemente nos resultados, confirmando o que

já havia sido descrito na Seção 3.1 e em Chou, Martens e Voit (2006). Portanto, apesar dos objetivos deste trabalho, apresentados na Seção 4.2, terem sido, de certa forma, contemplados, a etapa de desacoplamento e estimativa das derivadas, assim como o uso de dados com presença de ruído, se mostraram extremamente sensíveis e causaram grande influência nos resultados da solução do problema merecendo mais atenção em estudos futuros.

6 CONCLUSÕES

Duas etapas importantes da estratégia de solução do problema inverso associado à representação delineada por modelos de sistemas biológicos foram focadas neste trabalho: a estimativa das derivadas a partir dos pontos das séries temporais para que o desacoplamento do sistema pudesse ser realizado e a otimização para a determinação dos parâmetros das equações do sistema.

Comparando a estimativa da derivada realizada em cada ponto por métodos numéricos e pelo método analítico polinomial usado no filtro de Savitzky-Golay, chama a atenção a mesma precisão conseguida pelo método do passo complexo em relação ao método analítico polinomial. Assim, apesar de não ter sido observado melhora na precisão do cálculo das derivadas a partir dos métodos numéricos utilizados, os resultados obtidos pelo método do passo complexo fazem dele uma excelente ferramenta a ser considerada, pois além de demonstrar superioridade em relação ao método das diferenças finitas no cálculo de derivadas, conseguiu alcançar precisão tão boa quanto o método analítico polinomial, podendo então, através de simples implementação, ser de grande utilidade ao ser usado em estratégias que não possuem soluções analíticas, viabilizando a obtenção de derivadas numéricas com qualidade. Borges (2007) apresenta uma situação em que o método do passo complexo poderá substituir com eficiência o método da diferenças finitas, onde, a utilização com o filtro de Whittaker, que mostrou um ótimo desempenho em um estudo comparativo entre filtros, pode melhorar a qualidade das derivadas estimadas, cujo desempenho não foi adequado.

Em relação à otimização para a determinação dos parâmetros, considerando os dois principais objetivos descritos na Seção 4.2, pode-se concluir em relação ao primeiro (avaliar a eficácia do método da Entropia Cruzada para a estimativa de parâmetros do modelo Sistema S) que o método da Entropia Cruzada é eficaz para se aplicar na resolução de problemas com busca de variáveis de sistemas representados por Sistema S, pois os resultados encontrados mostraram-se satisfatórios na maioria das situações. Para o segundo objetivo (avaliar a eficácia de um método Híbrido para a estimativa de parâmetros do modelo Sistema S) pode-se concluir, de maneira similar, que o método Híbrido também é eficaz na resolução de problemas com busca de variáveis de sistemas representados por

Sistema S, pois apresentou resultados muito superiores nas duas primeiras fases de experimentos numéricos da Seção 5.2, sugerindo que sua eficiência pode ser até melhor que cada um dos outros dois métodos estudados.

Portanto, a avaliação foi positiva para os três métodos apresentados e avaliados neste trabalho para a obtenção de parâmetros de Sistemas S em suas várias etapas. Em relação a estimativa de derivadas, o método do passo complexo mostrou-se efetivo como opção à estratégia de diferenças finitas. No caso do processo de otimização, o método da Entropia Cruzada indicou efetividade no que tange a uma busca global dos parâmetros bem como o método Híbrido conseguiu um equilíbrio interessante entre a busca global e local que caracterizam os algoritmos envolvidos.

Modificações como implementação de parâmetros adaptativos e definição sistemática de critérios de parada (no caso do método da Entropia Cruzada e do Híbrido) podem ser feitas para melhorar a eficiência dos métodos, mas de forma geral eles funcionaram bem nos experimentos numéricos realizados. Um dos grandes problemas encontrados e que merece estudos mais detalhados em trabalhos futuros é a etapa de desacoplamento e estimativa das derivadas, que causa grande influência nos resultados finais da resolução do problema. A aplicação desses métodos em sistemas reais e com o objetivo de identificar a estrutura do sistema estudado, e não somente sua dinâmica (como discutido na Seção 4.2), consequentemente fazendo a busca por um maior número de parâmetros em cada equação, também serve como sugestão a ser aplicada em trabalhos futuros.

REFERÊNCIAS

ABREU, R.; STICH, D.; MORALES, J. On the generalization of the complex step method. *Jornal of Computational and Applied Mathematics*, 2013. v. 241, p. 84–102, 2013.

ADIAMAH, D. A.; SCHWARTZ, J.-M. Construction of a genome-scale kinetic model of mycobacterium tuberculosis using generic rate equations. *Metabolites*, 2012. v. 2, n. 3, p. 382–397, 2012.

BENGOETXEA, E. et al. Estimation of distribution algorithms: A new evolutionary computation approach for graph matching problems. In: FIGUEIREDO, M.; ZERUBIA, J.; JAIN, A. (Ed.). *Energy Minimization Methods in Computer Vision and Pattern Recognition*. [S.l.]: Springer Berlin Heidelberg, 2001. v. 2134, p. 454–469.

BERG, P. H.; VOIT, E. O.; WHITE, R. L. A pharmacodynamic model for the action of the antibiotic imipenem on pseudomonas aeruginosa populations in vitro. *Bulletin of Mathematical Biology*, 1996. v. 58, n. 5, p. 923–938, 1996.

BERKHOUT, J.; BRUGGEMAN, F. J.; TEUSINK., B. Optimality principles in the regulation of metabolic networks. *Metabolites*, 2012. v. 2, n. 3, p. 529–552, 2012.

BERTALANFFY, L. von. *General Systemstheory: Foundations, Development, Applications*. [S.l.]: George Braziller Inc., 1969. 296 p. ISBN 0807604534.

BEYER, W. *Combining Autosmooth and Alternating Regression for the Estimation of S-System Parameters*. [S.l.], 2008. Acessado em: 24/11/2013 20:34. Disponível em: <http://kdbio.inesc-id.pt/~svinga/dynamo/beyer/Beyer_TecRep29-2008.pdf>.

BODE, H. W. *Network Analysis and Feedback Amplifier Design*. [S.l.]: Van Nostrand, 1945.

BOER, P.-T. de et al. A tutorial on the cross-entropy method. *ANNALS OF OPERATIONS RESEARCH*, 2004. v. 134, 2004.

BORGES, C. C. H. A comparison of regression strategies applied on experimental biochemical data. *XXVIII Iberian Latin American Congress on Computatonal Methods in Engineering (CILAMCE)*, 2007. 2007.

BROMBA, M. U. A.; ZIEGLER, H. Applications hints for savitzky-golay digital smoothing filters. *Anal. Chem.*, 1981. v. 53, n. 11, p. 1583–1586, 1981.

BROOME, T. M.; COLEMAN, R. A. A mathematical model of cell death in multiple sclerosis. *Journal of Neuroscience Methods*, 2011. v. 2017, p. 420–425, 2011.

CHAN, J. C. C.; KROESE, D. P. Improved cross-entropy method for estimation. *Statistics and Computing*, 2012. v. 22, n. 5, p. 1031–1040, 2012.

CHOU, I.-C.; MARTENS, H.; VOIT, E. O. Parameter estimation in biochemical systems models with alternating regression. *Theoretical Biology and Medical Modelling*, 2006. v. 3, n. 1, p. 1–11, 2006.

CHOU, I.-C.; VOIT, E. O. Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Mathematical Biosciences*, 2009. v. 219, n. 2, p. 57–83, 2009.

CRAMPIN, J. S. adn E. J.; MCSHARRY, P. E.; SCHNELL, S. Reconstructing biochemical pathways from time course data. *Proteomics*, 2007. v. 7, n. 6, p. 828–838, 2007.

CURTO, R.; VOIT, E. O.; CASCANTE, M. Analysis of abnormalities in purine metabolism leading to gout and to neurological dysfunctions in man. *Biochemical Journal*, 1998. v. 329, n. 3, p. 477–487, 1998.

CURTO, R. et al. Validation and steady-state analysis of a power-law model of purine metabolism in man. *Biochemical Journal*, 1997. v. 324, n. 3, p. 761–775, 1997.

CURTO, R. et al. Mathematical models of purine metabolism in man. *Mathematical Biosciences*, 1998. v. 151, n. 1, p. 1–49, 1998.

DULAM-BANAWA, B.; MARIN-SANGUINO, A.; MENDOZA, E. The evolution of synapse models from numbers to networks to spaces. *Pharmacopsychiatry*, 2010. v. 43, n. 1, p. S42–S49, 2010.

EICHER, J. J.; SNOEP, J. L.; ROHWER, J. M. Determining enzyme kinetics for systems biology with nuclear magnetic resonance spectroscopy. *Metabolites*, 2012. v. 2, n. 4, p. 818–843, 2012.

FILIPP, F. V. Cancer metabolism meets systems biology: Pyruvate kinase isoform pkm2 is a metabolic master regulator. *Journal of Carcinogenesis*, 2013. v. 12, n. 1, p. 14, 2013.

GOEL, G.; CHOU, I. C.; VOIT, E. O. System estimation from metabolic time-series data. *Bioinformatics*, 2008. v. 24, n. 21, p. 2505–2511, 2008.

HEINRICH, R.; RAPOPORT, T. A linear steady-state treatment of enzymatic chains. general properties, control and effector strength. *Eur J Biochem*, 1974. v. 42, n. 1, p. 89–95, 1974.

HOOD, L. et al. Systems biology and new technologies enable predictive and preventative medicine. *Science*, 2004. v. 306, n. 306, p. 640–643, 2004.

HORMIGA, J. A. et al. Model based optimization of feeding regimens in aquaculture: application to the improvement of octopus vulgaris viability in captivity. *Journal of Biotechnology*, 2010. v. 149, n. 3, p. 209–214, 2010.

IYENGAR, R. *Introduction to Systems Biology*. 2013. [Online; acessado em 22/09/2013]. Disponível em: <https://www.coursera.org/course/sysbio>.

JOYCE, A. R.; PALSSON, B. O. The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology*, 2006. v. 7, n. 3, p. 198–210, 2006.

KACSER, H.; BURNS, J. The control of flux. *Symp Soc Exp Biol*, 1973. v. 27, p. 65–104, 1973.

KARR, J. R. et al. A whole-cell computational model predicts phenotype from genotype. *Cell*, 2012. v. 150, n. 2, p. 389–401, 2012.

KEITH, J.; KROESE, D. P. Sequence alignment by rare event simulation. In: *Simulation Conference, 2002. Proceedings of the Winter*. [S.l.: s.n.], 2002. v. 1, p. 320–327 vol.1.

KENDAL, W. S.; JØRGENSEN, B. Taylor's power law and fluctuation scaling explained by a central-limit-like convergence. *Phys. Rev. E*, 2011. American Physical Society, v. 83, 2011.

KIMURA, S. et al. Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm. *Bioinformatics*, 2005. v. 21, n. 7, p. 1154–1163, 2005.

KITANO, H. Computational systems biology. *Nature*, 2002. v. 420, n. 6912, p. 206–210, 2002.

KITANO, H. Systems biology: A brief overview. *Science*, 2002. v. 295, n. 5560, p. 1662–1664, 2002.

KLIPP, E. et al. *Systems Biology: A Textbook*. [S.l.]: John Wiley and Sons, 2009. 592 p.

KOMILI, S.; SILVER, P. A. Coupling and coordination in gene expression processes: a systems biology view. *Nature Reviews Genetics*, 2008. v. 9, p. 38–48, 2008.

KONOPKA, A. K. *Systems Biology: Principles, Methods, and Concepts*. 1. ed. [S.l.]: CRC Press, 2007. ISBN 0824725204.

LAI, X. et al. A multi-level model accounting for the effects of jak2-stat5 signal modulation in erythropoiesis. *Computational Biology and Chemistry*, 2009. v. 33, n. 4, p. 312–324, 2009.

LANGER, B. M. et al. Modeling of leishmaniasis infection dynamics: novel application to the design of effective therapies. *BMC Systems Biology*, 2012. v. 6, n. 1, 2012.

LARRANAGA, P.; LOZANO, J. A. *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation (Genetic Algorithms and Evolutionary Computation)*. [S.l.]: Kluwer Academic, 2002.

LESNE, A.; LAGUËS, M. *Scale Invariance: From Phase Transitions to Turbulence*. [S.l.]: Springer, 2011.

LIU, S. et al. Modeling of p53 signaling pathway based on s-system equations. *Journal of Biomedical Engineering*, 2010. v. 27, p. 505–510, 2010.

LYNESS, J.; MOLER, C. Numerical differentiation of analytic functions. *Journal of Numerical Analysis*, 1967. v. 4, p. 202–210, 1967.

MAGOMBEDZE, G.; MULDER, N. Understanding tb latency using computational and dynamic modelling procedures. *Infection, Genetics and Evolution*, 2013. v. 13, n. 0, p. 267–283, 2013.

MAKI, Y. et al. Development of a system for the inference of large scale genetic networks. *Pacific Symposium on Biocomputing*, 2001. n. 6, p. 446–458, 2001.

MARIN-SANGUINO, A.; MENDOZA, E. R. Hybrid modeling in computational neuropsychiatry. *Pharmacopsychiatry*, 2008. v. 41, n. 1, p. S85–S88, 2008.

MARIN-SANGUINO, A.; ROSARIO, R. C. del; MENDOZA, E. R. Concept maps and canonical models in neuropsychiatry. *Pharmacopsychiatry*, 2009. v. 42, p. S110–S117, 2009.

MARIN-SANGUINO, A. et al. Metabolic engineering with power-law and linear-logarithmic systems. *Mathematical Biosciences*, 2009. v. 218, n. 1, p. 50–58, 2009.

MARTIN, P. G. The use of canonical s-system modelling for condensation of complex dynamic models. *Ecological Modelling*, 1997. v. 103, n. 1, p. 43–70, 1997.

MARTINS, J. R. R. A.; STURDZA, P.; ALONSO, J. J. The complex-step derivative approximation. *ACM Trans. Math. Softw.*, 2003. v. 29, p. 245–262, 2003.

METHOD, T. C.-E. *Rosenbrock Visualisation*. 2013. [Online; acessado em 07/12/2013]. Disponível em: <http://iew3.technion.ac.il/CE/soft1.php?8>.

MICHAELIS, L.; MENTEN, M. L. Die kinetic der invertinwirkung. *Biochem Zeitschrift*, 1913. v. 49, p. 333–369, 1913.

NI, T. C.; SAVAGEAU, M. A. Application of biochemical systems theory to metabolism in human red blood cells: signal propagation and accuracy of representation. *Journal of Biological Chemistry*, 1996. v. 271, n. 14, p. 7927–7941, 1996.

NI, T. C.; SAVAGEAU, M. A. Model assessment and refinement using strategies from biochemical systems theory: application to metabolism in human red blood cells. *Journal of Theoretical Biology*, 1996. v. 179, n. 4, p. 329–354, 1996.

NIKOLOV, S. et al. Integration of sensitivity and bifurcation analysis to detect critical processes in a model combining signalling and cell population dynamics. *International Journal of Systems Science*, 2010. v. 41, n. 1, p. 81–105, 2010.

NIKOLOV, S. et al. Role of inhibitory proteins as modulators of oscillations in nfb signalling. *IET Systems Biology*, 2009. v. 3, n. 2, p. 59–76, 2009.

OKLANDER, B. *Stochastic Optimization - Cross Entropy Visualization*. 2012. [Online; acessado em 07/12/2013]. Disponível em: <http://www.youtube.com-/watch?v=tNAIHEse7Ms>.

PETER, I. S.; DAVIDSON, E. H. Evolution of gene regulatory networks controlling body plan development. *Cell*, 2011. v. 144, n. 6, p. 970–985, 2011.

QI, Z. et al. Effects of dopamine and glutamate on synaptic plasticity: a computational modeling approach for drug abuse as comorbidity in mood disorders. *Pharmacopsychiatry*, 2011. v. 44, n. 1, p. S62–S75, 2011.

QI, Z.; MILLER, G. W.; VOIT, E. O. Computational systems analysis of dopamine metabolism. *PLoS ONE*, 2008. v. 3, n. 6, 2008.

QI, Z.; MILLER, G. W.; VOIT, E. O. A mathematical model of presynaptic dopamine homeostasis: implications for schizophrenia. *Pharmacopsychiatry*, 2008. v. 41, n. 1, p. S89–S98, 2008.

QI, Z.; MILLER, G. W.; VOIT, E. O. Computational analysis of determinants of dopamine (da) dysfunction in da nerve terminals. *Synapse*, 2009. v. 63, n. 12, p. 1133–1142, 2009.

QI, Z.; MILLER, G. W.; VOIT, E. O. Computational modeling of synaptic neurotransmission as a tool for assessing dopamine hypotheses of schizophrenia. *Pharmacopsychiatry*, 2010. v. 43, n. 1, p. S50–S60, 2010.

QI, Z.; MILLER, G. W.; VOIT, E. O. The internal state of medium spiny neurons varies in response to different input signals. *BMC Systems Biology*, 2010. v. 4, n. 26, 2010.

ROSEN, R. *Fundamentals of Measurement and Representation of Natural Systems*. [S.l.]: North-Holland, 1978. 221 p.

ROSEN, R. *Anticipatory systems: philosophical, mathematical, and methodological foundations*. [S.l.]: Pergamon Press, 1985. 436 p.

ROSEN, R. *Life Itself: A Comprehensive Inquiry Into the Nature, Origin, and Fabrication of LifeAnticipatory systems: philosophical, mathematical, and methodological foundations*. [S.l.]: Columbia University Press, 1991. 285 p.

ROSEN, R. *Essays on Life Itself*. [S.l.]: Columbia University Press, 2000. 285 p.

RUBINSTEIN, R. Combinatorial optimization, cross-entropy, ants and rare events. In: URYASEV, S.; PARDALOS, P. (Ed.). *Stochastic Optimization: Algorithms and Applications*. [S.l.]: Springer US, 2001, (Applied Optimization, v. 54). p. 303–363.

RUBINSTEIN, R. Y. Optimization of computer simulation models with rare events. *European Journal of Operations Research*, 1996. v. 99, p. 89–112, 1996.

RUBINSTEIN, R. Y. The simulated entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 1999. v. 2, p. 127–190, 1999.

RUBINSTEIN, R. Y.; KROESE, D. P. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. 1st. ed. [S.l.]: Springer, 2004.

SAVAGEAU, M. A. Biochemical systems analysis I. some mathematical properties of the rate law for the component enzymatic reactions. *Journal of Theoretical Biology*, 1969. v. 25, n. 3, p. 365–369, 1969.

SAVAGEAU, M. A. Biochemical systems analysis II. the steady-state solutions for an n-pool system using a power-law approximation. *Journal of Theoretical Biology*, 1969. v. 25, n. 3, p. 370–379, 1969.

SAVAGEAU, M. A. Biochemical systems analysis III. dynamic solutions using a power-law approximation. *Journal of Theoretical Biology*, 1970. v. 26, n. 2, p. 215–226, 1970.

SAVAGEAU, M. A. Growth of complex systems can be related to the properties of their underlying determinants. *Proceedings of the National Academy of Sciences of the United States of America*, 1979. v. 76, n. 11, p. 5413–5417, 1979.

SAVAGEAU, M. A. Introduction to s-systems and the underlying power-law formalism. *Mathematical and Computer Modelling*, 1988. v. 11, p. 546–551, 1988.

SAVITZKY, A.; GOLAY, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, 1964. v. 36, n. 8, p. 1627–1639, 1964.

SQUIRE, W.; TRAPP, G. Using complex variables to estimate derivatives of real functions. *SIAM Review*, 1998. v. 40, n. 1, p. 110–112, 1998.

SRINATH, S.; GUNAWAN, R. Parameter identifiability of power-law biochemical system models. *Journal of Biotechnology*, 2010. v. 149, n. 3, p. 132–140, 2010.

TORRES, N. V. Modelization and experimental studies on the control of the glycolytic-glycogenolytic pathway in rat liver. *Molecular and Cellular Biochemistry*, 1994. v. 132, n. 2, p. 117–126, 1994.

TORRES, N. V. S-system modelling approach to ecosystem: application to a study of magnesium flow in a tropical forest. *Ecological Modelling*, 1996. v. 89, n. 1-3, p. 109–120, 1996.

VERA, J. et al. A systems biology approach to analyse amplification in the jak2-stat5 signalling pathway. *BMC Systems Biology*, 2008. v. 2, n. 38, p. 81–105, 2008.

VERA, J. et al. Power-law models of signal transduction pathways. *Cellular Signalling*, 2007. v. 19, n. 7, p. 1531–1541, 2007.

VERA, J. et al. Detection of potential enzyme targets by metabolic modelling and optimization: application to a simple enzymopathy. *Bioinformatics*, 2007. v. 23, n. 17, p. 2281–2289, 2007.

VERA, J. et al. Dynamics of receptor and protein transducer homodimerisation. *BMC Systems Biology*, 2008. v. 2, n. 92, 2008.

VERA, J. et al. Investigating dynamics of inhibitory and feedback loops in erk signalling using power-law models. *Molecular BioSystems*, 2010. v. 6, n. 11, p. 2174–2191, 2010.

VERA, J.; WOLKENHAUER, O. A system biology approach to understand functional activity of cell communication systems. *Methods in Cell Biology*, 2009. v. 90, p. 99–415, 2009.

VILELA, M. et al. Automated smoother for the numerical decoupling of dynamics models. *BMC Bioinformatics*, 2007. BioMed Central, v. 8, n. 1, p. 1–8, 2007.

VILELA, M. et al. Parameter optimization in s-system models. *BMC Systems Biology*, 2008. BioMed Central, v. 2, n. 1, p. 1–13, 2008.

VOIT, E.; NEVES, A. R.; SANTOS, H. The intricate side of systems biology. *Proceedings of the National Academy of Sciences of the United States of America*, 2006. v. 103, n. 25, p. 9452–9457, 2006.

VOIT, E. O. *Computational analysis of biochemical systems*. [S.l.]: Cambridge University Press, 2000.

VOIT, E. O. Biochemical systems theory: A review. *ISRN Biomathematics*, 2013. v. 2013, p. 53, 2013.

VOIT, E. O.; ALMEIDA, J. Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics*, 2004. v. 20, n. 11, p. 1670–1681, 2004.

VOIT, E. O. et al. Regulation of glycolysis in lactococcus lactis: an unfinished systems biological case study. *Systems Biology, IEE Proceedings*, 2006. v. 153, n. 4, p. 286–298, 2006.

VOIT, E. O. et al. Estimation of metabolic pathway systems from different data sources. *IET Systems Biology*, 2009. v. 3, n. 6, p. 513–522, 2009.

VOIT, E. O.; QI, Z.; KIKUCHI, S. Mesoscopic models of neurotransmission as intermediates between disease simulators and tools for discovering design principles. *Pharmacopsychiatry*, 2002. v. 45, n. 1, p. S22–S30, 2002.

VOIT, E. O.; SAVAGEAU, M. A. Power-law approach to modeling biological systems? II. application to ethanol production. *Journal of Fermentation Technology*, 1982. v. 60, p. 229–232, 1982.

VOIT, E. O.; SAVAGEAU, M. A. Power law approach to modeling biological systems; III. methods of analysis. *J. Ferment. Technol*, 1982. v. 60, p. 233–241, 1982.

WIENER, N. *Cybernetics*. 1. ed. [S.l.]: MIT Press, 1948.

WOLKENHAUER, O. Systems biology: The reincarnation of systems theory applied in biology? *Briefings in Bioinformatics*, 2001. v. 2, n. 3, p. 258–270, 2001.

WU, J.; QI, Z.; VOIT, E. O. Investigation of delays and noise in dopamine signaling with hybrid functional petri nets. *In Silico Biology*, 2010. v. 10, 2010.