

Maria Claudia Negret López

**Computação Bayesiana Aproximada
Monte Carlo Sequencial como ferramenta
para a inferência de parâmetros na rede
genética circadiana da *Arabidopsis
thaliana***

Rio Grande RS, Brasil

Abril 2015

Maria Claudia Negret López

**Computação Bayesiana Aproximada Monte Carlo
Sequencial como ferramenta para a inferência de
parâmetros na rede genética circadiana da
*Arabidopsis thaliana***

Dissertação apresentada à banca examinadora da Universidade Federal do Rio Grande, como exigência parcial para obtenção do título de Mestre em Modelagem Computacional pelo Programa de Pós-Graduação em Modelagem Computacional.

Universidade Federal do Rio Grande – FURG

Centro de Ciências Computacionais

Pós-Graduação em Modelagem Computacional

Orientador: Leonardo Ramos Emmendorfer

Coorientador: Adriano Velasque Werhli

Rio Grande RS, Brasil

Abril 2015

N285c Negret López, Maria Claudia.
Computação Bayesiana Aproximada Monte Carlo Sequencial como
ferramenta para a inferência de parâmetros na rede genética circadiana da
Arabidopsis thaliana / Maria Claudia Negret López. – 2015.
133 f.

Dissertação (mestrado) – Universidade Federal do Rio Grande – FURG,
Programa de Pós-graduação em Modelagem Computacional, Rio Grande/RS,
2015.
Orientador: Dr. Leonardo Ramos Emmendorfer.
Coorientador: Dr. Adriano Velasque Werhli.

1. Inferência Bayesiana 2. Monte Carlo Sequencial 3. Rede regulatória
4. Relógio biológico 5. Métodos matemáticos I. Emmendorfer, Leonardo
Ramos II. Werhli, Adriano Velasque III. Título.

CDU 519.245

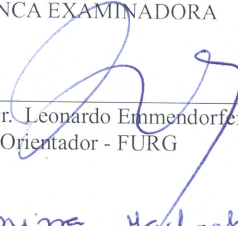
MARIA CLAUDIA NEGRET LÓPEZ

**“Computação Bayesiana Aproximada Monte Carlo Sequencial como
ferramenta para interferência de parâmetros na rede genética circadiana da
Arabidopsis thaliana”**

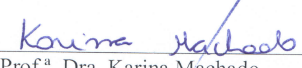
Dissertação apresentada ao Programa de Pós
Graduação em Modelagem Computacional da
Universidade Federal do Rio Grande -FURG,
como requisito parcial para obtenção do Grau
de Mestre. Área concentração: Modelagem
Computacional.

Aprovada em

BANCA EXAMINADORA



Prof. Dr. Leonardo Emmendorfer
Orientador - FURG



Prof.ª. Dra. Karina Machado
FURG



Prof.ª. Dr.ª Simone André da Costa Cavaleiro
UFPEL



Prof. Dr. Cleo Zanella Billa
FURG

Rio Grande -RS
2015

A mis padres

Agradecimentos

Não somente estes agradecimentos serão expressados pela elaboração deste trabalho mas também pela companhia, apoio e contribuição durante estes dois anos de novas experiências, emoções e sensações. Espero poder colocar o nome de cada um neste curto espaço e explicar o papel que desempenhou para o meu crescimento tanto pessoal como profissional.

Primeiramente, a Deus por abrir novas portas e rumos, com sua benção e ajuda e pelos anjos que me fizeram seguir o caminho correto para chegar até este ponto de minha vida; aos meus pais, por todo o amor e apoio incondicional durante estes anos e a paciência que tiveram comigo durante este desafio.

À OEA-GCUB e seu programa de bolsas de estudos no Brasil.

A meus professores na Colômbia Rafael Díaz, Rafael Ahumada Barrios, e Jesús Hernando Pérez (QED) pelo acompanhamento e confiança que depositaram em mim.

Aos professores Rosilene Clementine e Vilmar por seu acolhimento, apoio e paciência a cada vez que procurei por algo e por me ajudarem a encontrar uma solução.

Ao professor, Adriano Werhli, por seu apoio durante a minha co-orientação e por me ajudar com o desenvolvimento do trabalho. Ao professor Leonardo Emmendorfer por ser meu orientador, às professoras Karina Machado, e Diana Adamatti pelos sorrisos, pelas palavras de apoio e conselhos durante a minha estadia no Brasil.

A minhas brasileiras Núbia Prates e Miriam Blank Born por me acolherem desde o primeiro momento que cheguei à faculdade. Foram o primeiro grupo de trabalho que eu tive e as minhas primeiras e melhores amigas brasileiras, com as que compartilhei lindos momentos, me ajudando com meu português e me suportando nestes dois anos. A Nessa Gil, minha bela estrela, minha amiga de travessuras e risos.

A minha Yuni, que conheceu todos meus segredos, pensamentos e sentimentos nesta aventura, que suportou meu caráter e minhas lágrimas sem se importar, que me levantou a cada vez que caí e sempre teve uma palavra ou uma reação para cada ação, que me ensinou tantas coisas, verdadeiramente não tenho como lhe agradecer.

A Gabriela e Jose meus dois anjos que me acompanharam neste caminho me dando uma mão, uma palavra, uma companhia cada vez que precisei. A minha colega Ileana de grandes caminhadas sob a chuva, minha amiga esquisita que sempre teve uma palavra para me ajudar a abrir os olhos no pouco tempo que estivemos juntas e minha irmã de alma que sempre terei no meu coração.

A Lorena por suas loucuras, passeios pela FURG, risadas nas aulas de português, por suas palavras quando precisei e por sua companhia no dia a dia.

Aos meus amigos na Colômbia, Gloria, Claudia Locha, Claudia Polanía, Osman e Karen que não se importaram com meu estresse e tensão, sempre estiveram presentes para me apoiar e descontraír.

A todos muito obrigada.

*“Mathematics is the alphabet with
which God has written the universe.”
(Galileu Galilei)*

Resumo

A modelagem de sistemas complexos é um desafio para os pesquisadores nas diferentes áreas, já que cada vez deseja-se simular a natureza na forma mais precisa possível. Estes modelos, especialmente aqueles que representam sistemas biológicos, precisam de dados experimentais que, na maioria dos casos, são muito difíceis ou impossíveis de medir experimentalmente de forma direta, sendo que outra alternativa de procurá-los é por meio da estimação ou por inferência a partir dos dados observados. Neste contexto, os métodos de Computação Bayesiana Aproximada (ABC), e mais precisamente aqueles que incorporam os métodos Monte Carlo Sequencial (SMC), foram utilizados para aproximar a distribuição posteriori dos dados dadas as vantagens que apresentam estes métodos em relação aos anteriores. O sistema complexo considerado foi o correspondente à rede do ciclo circadiano da *Arabidopsis thaliana*, uma planta reconhecida pela comunidade científica para a realização de experimentos. Com isto, procurou-se aplicar o algoritmo ABC-SMC em uma rede regulatória biológica complexa, para ajustar a maior quantidade dos valores dos parâmetros relacionados especialmente com as proteínas LHY/CCA1 as quais envolvem os efeitos da luz na planta e que dos quais, segundo a literatura, tem-se menos confiabilidade de seus valores. Depois de ajustar cinco parâmetros, encontraram-se algumas limitações da metodologia utilizada com o pacote ABC-SysBio como o tamanho do conjunto de dados a trabalhar. Conclui-se que os valores dos parâmetros observados são confiáveis devido à semelhança que eles têm com aqueles que foram encontrados nas simulações.

Palavras-chaves: ABC-SMC. ABC-SysBio. Inferência Bayesiana. Rede regulatória. Relógio biológico. Modelo matemático.

Abstract

Complex system modeling is a challenge for the researchers in different fields, and the aim is to simulate the nature in the most precise way. These models, specially those that represent biological systems, need experimental data that, in the most cases, are difficult to be measured or it is almost impossible to measure them in a direct way. Thus, they must to be found by estimation or by inference from the observed data. In this context, Approximate Bayesian Computation (ABC) methods, and more precisely, those that incorporate Sequential Monte Carlo methods were used to approximate the posterior distribution of the data. In this way, the complex system considered was the circadian net in *Arabidopsis thaliana* a renamed plant in the scientific community, that it is recognized by their experimental advantages. With this aim the method ABC-SMC was applied to a biological complex net that represents the circadian circuit in *A. thaliana* to fit the most number of parameters values related specially with LHY/CCA1 proteins, those that involves light effects in the plant and those that have less confidence of their values. Fitted five parameters, it was found some limitations of the software ABC-SysBio that make the interpretation very difficult such the data set to be worked. It is concluded that the observed parameters could be considered reliable because the similarity between them.

Key-words: ABC-SMC. ABC-SysBio. Bayesian Inference. Regulatory Network. Biological clock. Mathematical model.

Lista de ilustrações

Figura 1 – <i>Arabidopsis thaliana</i>	32
Figura 2 – Rede circadiana da <i>A. thaliana</i>	33
Figura 3 – Algumas reações entre componentes da rede da <i>A. thaliana</i>	34
Figura 4 – Repressão de PRR9 pelo EC.	34
Figura 5 – Inibição da expressão dos genes LUX, ELF4, GI, TOC1 e PRR9 pelo EC.	34
Figura 6 – Inibição dos genes da noite.	35
Figura 7 – EC componente principal do ciclo noturno.	35
Figura 8 – Interações entre EC e LHY/CCA1 via PRRs.	36
Figura 9 – Uso do ABC desde sua aparição até 2008.	46
Figura 10 – Uso do ABC desde 1999.	47
Figura 11 – O algoritmo ABC-SMC.	52
Figura 12 – Procedimento para o uso de ABC-SysBio.	65
Figura 13 – Alguns arquivos de saída do ABC-SysBio. Taxas e gráfico dos dados.	66
Figura 14 – Alguns arquivos de saída do ABC-SysBio. Distâncias e trajetórias.	67
Figura 15 – Alguns arquivos de saída do ABC-SysBio. Pesos e dados populacionais.	68
Figura 16 – Alguns gráficos de saída do ABC-SysBio. Histogramas.	69
Figura 17 – Alguns gráficos de saída. Séries de tempo.	70
Figura 18 – Gráfico de dispersão e da distribuição posteriori para os parâmetros $n1, q1$	81
Figura 19 – Diagrama de caixas para $n1, q1$	82
Figura 20 – Gráfico de dispersão e da distribuição posteriori para os parâmetros $n1, g1, q1$	82
Figura 21 – Gráfico de dispersão e da distribuição posteriori para os parâmetros $n1, g1, g3, q1$	83
Figura 22 – Gráfico de dispersão e da distribuição posteriori para os parâmetros $n1, g1, g3, p1, q1$	83
Figura 23 – Diagrama de caixas para $n1$ e $q1$ da População 1	86

Figura 24 – Gráficos de dispersão por pares e por populações para os parâmetros $n1, g1, q1$	89
Figura 25 – Distribuição posteriori por populações para os parâmetros $n1, g1, q1$	90
Figura 26 – Séries de tempo por populações para os parâmetros $n1, g1, q1$	91
Figura 27 – Gráficos de dispersão por pares e por populações para os parâmetros $n1, g1, g3, q1$	92
Figura 28 – Distribuição Posteriori por populações para os parâmetros $n1, g1, g3, q1$	93
Figura 29 – Gráficos de dispersão por pares e por populações para os parâmetros $n1, g1, g3, p1, q1$	94
Figura 30 – Distribuição posteriori por populações para os parâmetros $n1, g1, g3, p1, q1$	95
Figura 31 – Diagrama de caixas para $n1, g1, g3, p1$ e $q1$ para (a) 500 e (b) 22 medições das variáveis ao longo do tempo	100

Lista de tabelas

Tabela 1 – Sumário das simulações 1 e 2	78
Tabela 2 – Aproximações para os valores dos parâmetros	80
Tabela 3 – Valores das médias e medianas dos parâmetros $n1$ e $q1$ do arquivo last_param.dat.	87
Tabela 4 – Aproximações para os valores dos parâmetros	88
Tabela 5 – Valores das médias e medianas dos parâmetros $n1$, $g1$, $g3$, $p1$ e $q1$ do arquivo last_param.dat.	100
Tabela 6 – Distribuições posterioris aproximadas para os parâmetros $n1$, $g1$, $g3$, $p1$ e $q1$	101

Lista de abreviaturas e siglas

ABC	Approximate Bayesian Computation
CCA1	Circadian Clock Associated1
COP1	Constitutive Photomorphogenic 1
EC	Evening component
ELF3	Early Flowering 3
ELF4	Early Flowering 4
GI	Gigantea
GPU	Graphics Processing Unit
LHY	Late Elongated Hypocotyl
LUX	Lux Arritmo
MCMC	Markov Chain Monte Carlo
mRNA	Messenger Ribonucleic Acid
NI	Night Inhibitor
ODE	Ordinary Differential Equation
OLCM	Optimal Local Covariance Matrix
PlaSMo	Plants Systems Modelling Repository
PMC	Populational Monte Carlo
PRR5	Pseudo-Response Regulator 5
PRR7	Pseudo-Response Regulator 7

PRR9	Pseudo-Response Regulator 9
SBML	Systems Biology Markup Language
SDE	Stochastic Differential Equations
SIR	Susceptible Infected Recovered
SIS	Sequential Importance Sampling
SMC	Sequential Monte Carlo
TOC1	Timing of Cab Expression1
ZTL	Zeitlupe

Lista de símbolos

$d(.,.)$	Distância
D	Conjunto de dados experimentais
D^*	Conjunto de dados simulados
ϵ	Tolerância
θ	Vetor de parâmetros amostrados da distribuição priori
θ^*	Vetor de parâmetros amostrado da distribuição anterior
θ^{**}	Vetor de parâmetros logo da perturbação
$\pi(.)$	Distribuição priori
$\pi(. .)$	Distribuição posteriori
$L(. .)$	Função de verossimilhança
$M(.,.)$	Modelo matemático
w	Peso

Sumário

1	INTRODUÇÃO	25
		25
1.1	Objetivo	27
1.1.1	Objetivos Específicos	27
1.2	Estrutura da Dissertação	28
I	REFERENCIAIS TEÓRICOS	29
2	REDE GENÉTICA CIRCADIANA NA <i>ARABIDOPSIS THALIANA</i>	31
2.1	<i>A Arabidopsis thaliana</i>	31
2.2	Ciclo circadiano em <i>Arabidopsis thaliana</i>	32
3	INFERÊNCIA BAYESIANA	37
4	COMPUTAÇÃO BAYESIANA APROXIMADA (ABC)	45
4.1	Computação Bayesiana Aproximada Monte Carlo Sequencial (ABC-SMC)	49
5	O MODELO MATEMÁTICO	57
6	O PACOTE ABC-SYSBIO	63
7	SIMULAÇÕES	71
II	RESULTADOS	75
8	RESULTADOS	77
9	SIMULAÇÕES 1	79

10	SIMULAÇÕES 2	85
III	DISCUSSÃO	97
	Conclusão	103
11	TRABALHOS FUTUROS	105
	Referências	107
	APÊNDICES	113
	APÊNDICE A – BASES DA REGULAÇÃO TRANSCRICIO- NAL PROCARIÓTICA	115
	ANEXOS	117
	ANEXO A – ESPÉCIES	119
	ANEXO B – PARÂMETROS GLOBAIS(120)	123
	ANEXO C – REAÇÕES	129

1 Introdução

O trabalho interdisciplinar tem ganhado muita importância nas últimas duas décadas e, na procura de resolver problemas que afetam a nossa forma de vida, vem sendo desenvolvidos modelos matemáticos para representar a dinâmica de diferentes sistemas evolutivos, por exemplo, os sistemas biológicos. Segundo (LIEPE et al., 2014) tanto os dados experimentais como os modelos matemáticos tem a mesma relevância. Os modelos matemáticos que representam sistemas biológicos tentam se assemelhar à natureza na forma mais precisa possível. Porém o modelo seja uma representação aproximada da realidade, dados numéricos são necessários para todos os parâmetros que descrevem o desenvolvimento do sistema matemático, de maneira a realizar uma ligação dos modelos com os dados.

Um dos objetivos da pesquisa em espécies cultivadas, é estudar como melhorar a produção de coletas em relativamente pouco tempo. Para realizar as experimentações com menor custo e investimento de tempo e conduzir os resultados obtidos por essas experimentações à prática, utiliza-se a planta *Arabidopsis thaliana*¹ como referência. Deste modo, as hipóteses obtidas em laboratório são levadas às plantas de importância econômica e cultural de forma rápida e eficiente.

Uma destas melhoras está relacionada com o crescimento das plantas, o que depende de sua resposta à luz. A luz é geralmente considerada como a fonte de energia mas também cumpre a função de acionador e modulador de processos complexos como o relógio circadiano (também chamado ciclo ou ritmo circadiano). Atua na produção da clorofila, o desenvolvimento das folhas, enquanto inibe o crescimento do caule em um processo chamado fotomorfogênese (The Arabidopsis Genome Initiative, 2000). Assim, os cientistas desejam encontrar a forma de converter essa planta em uma máquina fotossintética eficiente por meio de variações nas condições da luz. Alguns estudos

¹ A *Arabidopsis thaliana* é uma das espécies mais estudadas pela comunidade científica especialmente desde o sequenciamento completo do seu material genético no ano 2000 nas áreas da genética, a fisiologia e bioquímica (DELATORRE; SILVA, 2008). É uma planta pequena, com um ciclo de vida curto (aproximadamente de 6 semanas), produz uma grande quantidade de sementes, é geneticamente tratável e facilmente manipulável devido a seu tamanho e facilidade de cultivo (DELATORRE; SILVA, 2008). (The Arabidopsis Genome Initiative, 2000).

indicam que a luz ativa a expressão gênica das proteínas LHY (LATE ELONGATED HYPOCOTYL) e CCA1 (CIRCADIAN CLOCK ASSOCIATED1), que por sua vez estimulam genes que se expressam durante o dia enquanto também inibem genes que se expressam durante a noite, como TOC1 (Timing of CAB expression 1) (POKHILKO et al., 2012).

Neste contexto, os métodos de Computação Bayesiana Aproximada (ABC, pelo seu acrônimo em inglês), permitem realizar simulações que relacionam os dados com os modelos e vêm sendo empregado principalmente nos campos da biologia e da ecologia (CSILLÉRY et al., 2010), sendo que a sua primeira aplicação foi no campo da genética de populações (LOPES; BEAUMONT, 2010). Estão baseados na análise Bayesiana diferentemente da inferência frequentista, o qual trata os parâmetros como quantidades aleatórias juntamente com os dados e não como quantidades fixas dentro de um grupo de tentativas experimentais (TURNER; ZANDT, 2012). A inferência dos parâmetros dentro do contexto Bayesiano está baseada nas distribuições de probabilidade destes uma vez que alguns dados tenham sido observados o que é o mesmo que a distribuição posteriori e não, sobre o espaço amostral de resultados hipotéticos que são procurados pela execução do experimento várias vezes (TURNER; ZANDT, 2012). A característica mais importante do método é que não se precisa da especificação da função de verossimilhança², razão pela qual são chamados métodos livres de probabilidade (BARAGATTI; PUDLO, 2014) e são utilizados quando as funções de probabilidade são analiticamente ou computacionalmente intratáveis (SISSON; FAN; TANAKA, 2007).

Com ABC busca-se obter uma estimação da distribuição posteriori que melhor se ajuste ao modelo para esses parâmetros (BEAUMONT; ZHANG; BALDING, 2002). O ABC não precisa da especificação da função de verossimilhança para realizar a inferências sendo assim que, são métodos muito flexíveis na forma que admite quantidade de dados provenientes de modelos cada vez mais complexos (LOPES; BEAUMONT, 2010).

Das diferentes variações do método ABC, neste trabalho procurou-se utilizar o método ABC Monte Carlo Sequencial (ABC-SMC, pelo seu acrônimo em inglês), através do pacote ABC-SysBio, devido a não apresentar algumas das desvantagens das variantes de ABC anteriores, além que vem-se trabalhando nele desde o ano 2009 até a

² A verossimilhança é a função dos parâmetros que descreve em termos de probabilidade, que tão bem cada parâmetro explica os dados observados.

atualidade. A vantagem deste método é sua versatilidade, sendo que pode ser utilizado para a análise de sistemas complexos, e pode assumir uma grande quantidade de dados pela aplicação do conceito de sequência que está imerso em seu algoritmo (TONI et al., 2009). Este método já tem sido empregado para estimação de parâmetros de sistemas dinâmicos (TONI et al., 2009). Segundo (LIEPE et al., 2010), ABC-SysBio pode assumir modelos tanto determinísticos como estocásticos, ao contrário de outros pacotes como DIYABC ou BioBayes. Adicionalmente, está implementado em Python convertendo-o em um programa flexível.

O método ABC-SMC pode assumir modelos de alta complexidade e uma grande quantidade de dados, por isso neste trabalho será aplicado para ajustar os parâmetros de pouca confiabilidade ao modelo matemático do relógio genético circadiano na *Arabidopsis thaliana* (descrito em (POKHILKO et al., 2012)) já que atualmente deseja-se identificar os genes responsáveis da expressão das propriedades rítmicas da planta causadas pelo efeito da luz. Os valores dos parâmetros neste modelo foram procurados com base nos dados disponíveis ou foram ajustados a numerosos conjuntos de séries de tempo (POKHILKO et al., 2012).

1.1 Objetivo

O principal objetivo deste trabalho é aplicar o algoritmo ABC-SMC mediante o pacote ABC-SysBio em uma rede regulatória biológica complexa, neste caso aquela que representa o ciclo circadiano da *Arabidopsis thaliana*, determinada pelo modelo matemático descrito em (POKHILKO et al., 2012) com a finalidade de ajustar os valores dos parâmetros relacionados especialmente com as proteínas LHY/CCA1 e que, dos quais, tem-se menos confiabilidade de seus valores.

1.1.1 Objetivos Específicos

- Ajustar a maior quantidade de parâmetros ao modelo matemático em (POKHILKO et al., 2012);
- Contrastar a teoria conceptual do pacote ABC-SysBio em (LIEPE et al., 2010; LIEPE et al., 2014) com os resultados obtidos das simulações;

- Analisar os resultados obtidos das simulações;
- Interpretar estes resultados no contexto biológico que representa o ciclo circadiano da *Arabidopsis*.

1.2 Estrutura da Dissertação

O presente trabalho apresenta na introdução do trabalho, a importância das pesquisas de carácter interdisciplinar com o objetivo de contribuir no desenvolvimento das ciências. Sendo a *Arabidopsis thaliana* como objeto de estudo e a importância desta planta no desenvolvimento de melhoras nos cultivares, assim como alguns aspectos genéticos que controlam a dinâmica de seu relógio circadiano. Também é descrito o modelo matemático que representa esta dinâmica e o método de computação Bayesiana Aproximada Monte Carlo Sequencial como ferramenta para inferir parâmetros e ajustar os dados ao modelo, assim como os objetivos deste trabalho.

Na primeira parte, o **Capítulo 2** apresenta a rede genética do ciclo circadiano da *Arabidopsis thaliana*. No **Capítulo 3** descreve os princípios e importância da inferência Bayesiana. O **Capítulo 4** apresenta o método de Computação Bayesiana Aproximada, características e aplicações e expõe o método de Computação Bayesiana Aproximada Sequencial Monte Carlo as características e suas vantagens sobre outros métodos ABC. O **Capítulo 5** descreve o sistema de equações diferenciais que representa a rede do ciclo circadiano da *A. thaliana*. O **Capítulo 6**, apresenta o pacote ABC-SysBio com suas características. O **Capítulo 7** apresenta as configurações utilizadas pelo pacote ABC-SysBio, descrevendo-as em detalhe.

Na segunda parte, os **Capítulo 8**, **Capítulo 9** e **Capítulo 10** apresentam os resultados obtidos para a análise. Além destes resultados, uma pequena comparação entre duas formas de proceder é detalhada.

A terceira parte apresenta a discussão dos resultados apresentados na parte anterior e as conclusões do estudo. No **Capítulo 11**, são apresentadas sugestões para trabalhos futuros e as referências bibliográficas.

Parte I

Referenciais teóricos

2 Rede genética circadiana na *Ara-bidopsis thaliana*

Uma rede circadiana ou “relógio biológico” confere uma vantagem competitiva a um organismo, provavelmente, permitindo-lhe antecipar ciclos como luz/escuro no ambiente. Estas redes desempenham um papel muito importante em todos os seres vivos porque regulam cada processo biológico e incluem laços que geram expressões rítmicas dos genes chaves ([LOCKE; MILLAR; TURNER, 2005](#); [MCCLUNG; GUTIÉRREZ, 2010](#)). Para este trabalho os genes a considerar são principalmente CIRCADIAN CLOCK ASSOCIATED1 (CCA1) e LATE ELONGATED HYPOCOTYL (LHY) já que formam parte do laço central. Nos últimos anos a biologia de sistemas tem feito esforços para descrever os processos nas plantas por meio da modelagem da maquinaria da rede genética circadiana. Assim, no ano 2005 foi desenvolvido um modelo matemático simples do ciclo de realimentação contendo CCA1/LHY e TIMING OF CAB EXPRESSION1 (TOC1) utilizando dados moleculares e genéticos da literatura ([MCCLUNG; GUTIÉRREZ, 2010](#)), que foi melhorado depois no ano 2010.

2.1 A *Arabidopsis thaliana*

Arabidopsis thaliana ([Figura 1](#)) é considerada uma planta de grande importância para a comunidade científica desde o ano 1943 quando [Laibach \(1943\)](#) relatou o potencial desta espécie, a qual foi estabelecida como planta modelo em 1986.

Por outro lado, esta planta não possui importância econômica direta. Ela é muito valorosa na comunidade científica visto que é utilizada nas pesquisas nas áreas da genética, bioquímica e fisiologia. Foi a primeira planta com seu genoma totalmente sequenciado ([The Arabidopsis Genome Initiative, 2000](#)). Além do mais, é uma planta com tamanho pequeno, que produz uma grande quantidade de sementes, com um tempo de geração curto, fácil manipulação e pequeno genoma, características que fazem que esta planta seja considerada um organismo modelo ([DELATORRE; SILVA, 2008](#)), ([The Arabidopsis Genome Initiative, 2000](#)).

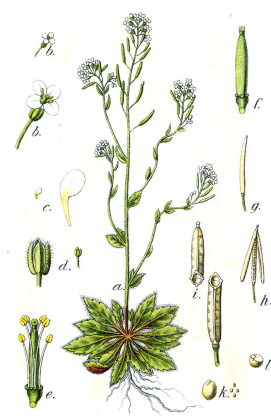


Figura 1 – *Arabidopsis thaliana*, pelo Johann Georg Sturm (Pintor: Jacob Sturm) (Domínio Público), via Wikimedia Commons.

A. thaliana tem sido estudada amplamente para o entendimento dos fenômenos biológicos, neste trabalho é considerado o ciclo circadiano para que os resultados destes estudos possam ser aplicados nos cultivos de outros organismos tornando-os mais produtivos, isto considerando que os processos biológicos que ocorrem nas espécies de importância econômica, acontecem também na *A. thaliana*. A afirmação anterior, está fundamentada em que uma quantidade importante das proteínas do arroz, por exemplo, mostram similaridade com *A. thaliana*. Espera-se que as funções celulares e bioquímicas dos genes de algumas culturas possam ser estabelecidas a partir de experimentos nesta planta modelo (DELATORRE; SILVA, 2008).

Os recentes estudos procuram elucidar a função de cada um dos genes da *A. thaliana* para esclarecer o funcionamento metabólico e suas relações com os processos fisiológicos da planta para desenvolver assim, cultivares mais produtivos. Desta forma, para este trabalho é analisado o ciclo circadiano da planta e mais precisamente como afeta a luz ao crescimento da planta.

2.2 Ciclo circadiano em *Arabidopsis thaliana*

Em *A. thaliana* o ciclo circadiano inclui três ciclos interconectados de realimentação de transição-translação na sua estrutura (USMAN; DAVIS, 2013). O ciclo diurno, o noturno e um ciclo central, que é chamado o oscilador nuclear e que está formado

por: CCA1, LHY, que são fatores de transcrição e pelo TOC1, o qual é um componente fundamental do circuito genético no relógio circadiano das plantas e que é um membro da família do gene regulador da pseudo-resposta (POKHILKO; MAS; MILLAR, 2013) (Figura 2 e Figura 3).

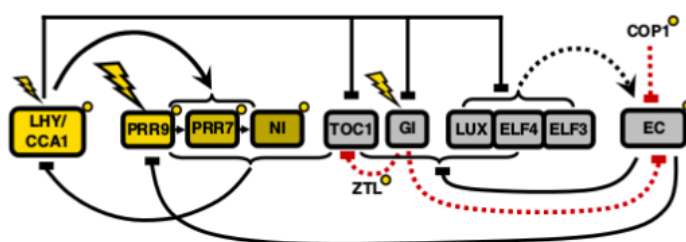


Figura 2 – Rede circadiana da *A. thaliana*.

Os retângulos representam proteínas. As flechas representam ativação, e as linhas perpendiculares representam repressão. Em amarelo estão representados os elementos que são ativados durante o dia e em cinza aqueles que são ativados durante a noite. As regulações transcricionais estão representados pelas linhas sólidas. A formação do complexo de proteínas está representado pelas linhas por traços pretas. As regulações translacionais estão representadas pelas linhas vermelhas. Por simplicidade estão representadas somente as proteínas para EC, ZTL e COP1. As respostas agudas à luz na transcrição de genes mostram-se com flashes. A regulação translacional pela luz esta denotada por círculos amarelos. Fonte: (POKHILKO et al., 2012).

Na rede proposta por Pokhilko et al. (2012), CCA1 e LHY são considerados como um componente individual (LHY/CCA1). Se incluem os genes COMPONENTE EVENING (EC) (que cumpre uma função importante no relógio)¹, junto com a regulação pós-translacional da proteína EARLY FLOWERING 3 (ELF3) pelo CONSTITUTIVE PHOTOMORPHOGENETIC 1 (COP1). Também, é considerada a conexão entre o ciclo da noite com o restante do relógio. Neste ponto, foi observada uma repressão de PSEUDO-RESPONSE REGULATOR 9 (PRR9) pelo EC (Figura 4), como também que a TOC1 regula a expressão dos genes LHY/CCA1. Esta rede integra conexões tanto

¹ EC é um repressor que pode ser essencial para os ritmos na expressão de genes em plantas.

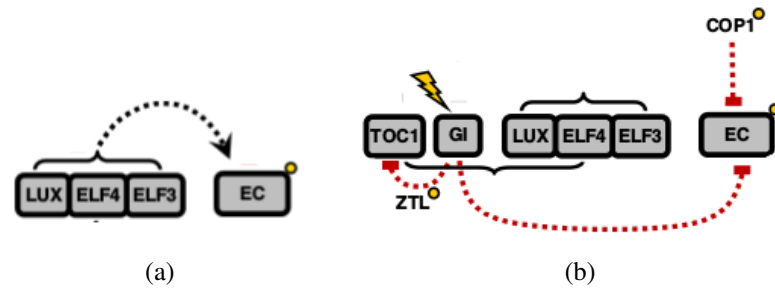


Figura 3 – Algumas reações entre componentes da rede da *A. thaliana*: (a) Formação do complexo de proteínas EC mostrado com a linha preta pontilhada, (b) Regulação translacional de TOC1 e EC por GI, ZTL e COP1 denotadas em vermelho.

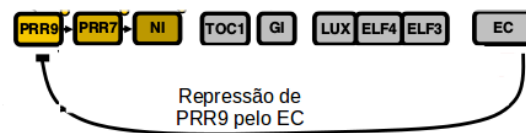


Figura 4 – Repressão de PRR9 pelo EC.

positivas como negativas em uma estrutura com múltiplos laços como também dados de tipo experimental que explicam as respostas às perturbações no ambiente e também de tipo genético e a resposta canônica aos impulsos de luz curtos em diferentes tempos.

O modelo mostra a formação da tripla de proteínas ELF3-ELF4-LUX, e assume que EC inibe a expressão dos genes LUX, ELF4, GI, TOC1 e PRR9 (Figura 5).

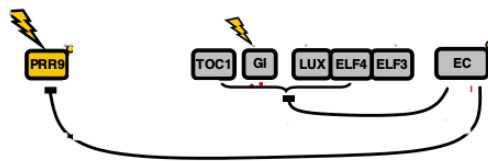


Figura 5 – Inibição da expressão dos genes LUX, ELF4, GI, TOC1 e PRR9 pelo EC.

A conexão entre o ciclo da noite e o ciclo da manhã acontece através da supressão de TOC1, LUX, ELF4, ELF3, e GI pelas proteínas CCA1 e LHY, isto inclui a inibição de PRR9 por TOC1, e à regulação negativa da expressão dos genes LHY/CCA1 pela

proteína TOC1. Segundo estudos na bioquímica é provável que a inibição da expressão de PRR9 seja causada por EC (Figura 6).

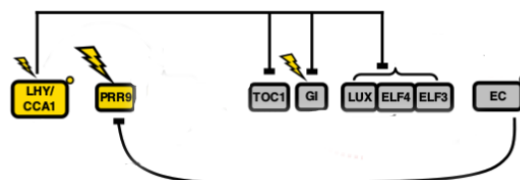


Figura 6 – Inibição de todos os genes da noite por CCA1 e LHY e de PRR9 pelo EC.

Depois de várias provas para analisar os efeitos da função repressiva de EC na dinâmica do sistema todo, afirmou-se que EC é o elemento principal no ciclo noturno, enquanto que as proteínas ELF3, ELF4 e LUX formam EC (Figura 3a) e as expressões de ELF4 e LUX são inibidas por EC criando um ciclo de retroalimentação negativa (Figura 7). Também considera-se o fato de que COP1 e GI regulam a atividade de EC (Figura 3b) e que TOC1 é um repressor do ciclo da manhã segundo múltiplas observações experimentais. No modelo, a importância de EC evidencia-se também



Figura 7 – EC componente principal do ciclo noturno.

porque EC controla as expressões de LHY e CCA1 através dos PPRs, enquanto apresenta-se uma retroalimentação duplo de LHY/CCA1 até os genes PPRs que está baseado em dados experimentais. As proteínas LHY e CCA1 inibem a expressão dos genes EC. Baseado em estudos genéticos, a repressão dos genes PRR por EC forma uma conexão dupla-negativa desde LHY e CCA1 até os PPRs a qual pode ser representada por uma regulação positiva da expressão de PRR por LHY/CCA1 (Figura 8) e as proteínas PRRs inibem a expressão de LHY e de CCA1 as quais inibem a expressão dos genes da noite.

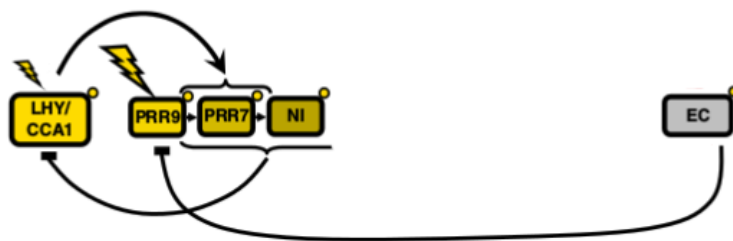


Figura 8 – Interações entre EC e LHY/CCA1 via PRRs.

3 Inferência Bayesiana

Após de cerca de um século de aproximações frequentistas das análises dos dados e ajuste de modelos, a “Revolução Bayesiana”, junto com a presença de computadores e algoritmos cada vez mais poderosos para ajustar modelos Bayesianos, têm permitido aos pesquisadores explorar os métodos Bayesianos no desenvolvimento de pesquisas. Estes métodos são importantes porque permitem realizar inferência estatística sem comprometer a teoria que motivou os experimentos que geraram os dados. Os processos de geração de dados dependem de um conjunto de parâmetros e o objetivo da inferência estatística é descrever a forma em que esses parâmetros mudam segundo alterações em condições experimentais (LOPES; BEAUMONT, 2010).

Uma variedade de aproximações, ou processos inferenciais, tem sido utilizados na prática para inferência de parâmetros a partir das observações disponíveis. Entre esses métodos, estão aqueles baseados em otimização que são concebidos para encontrar o melhor valor. Neste trabalho, busca-se encontrar o melhor conjunto de parâmetros. Um método comum de otimização, consiste em especificar a função objetivo que quantifica a discrepância entre os dados experimentais e as predições do modelo e assim procurar as diferentes combinações de parâmetros com o fim de minimizar essa discrepância. Algumas das desvantagens de utilizar estes métodos, segundo (LIEPE et al., 2014) são:

- a. Superajustar os dados, isto é, ajustar o ruído.
- b. Várias combinações de parâmetros podem se dar como ótimos locais, sendo necessário determinar se eles são ou não o melhor conjunto de parâmetros, o que é desafiante.
- c. Para a abordagem destes métodos deve-se contemplar a robustez da estimação dos parâmetros e a confiança depositada neles.

Também estão as aproximações pelo *bootstrapping*¹ ou pelo subamostragem que for-

¹ Processo de simular repetidamente novos conjuntos de dados que são inferidos dos dados observados e depois re-estimando os parâmetros destes conjuntos. Este processo é utilizado para obter intervalos de confiança (BEAUMONT; RANNALA, 2004).

necem uma série de métodos, computacionalmente intensivos, para quantificação da robustez gerando uma coleção de novos conjuntos desde o conjunto inicial de dados de observações para logo avaliar a variabilidade na estimação do parâmetro obtido nesta coleção (LIEPE et al., 2014).

Outros métodos utilizam a inferência estatística com a qual se procura obter os melhores estimadores para as taxas de reação assim como também de suas respectivas incertezas. O propósito da inferência Bayesiana é deduzir os valores dos parâmetros cuja probabilidade de gerar alguns dados observados é alta por isso é importante tanto encontrar o valor do parâmetro, como também avaliar a incerteza presente na estimação.

A primeira aplicação da estatística Bayesiana ocorreu nos anos 1990's para solucionar problemas de genética de populações. Sua eficiência foi reconhecida e começou-se a incrementar seu uso em diferentes campos como a filogenética, ecologia, conservação, evolução molecular e epidemiologia. Também ampliou-se seu uso nas áreas da biologia computacional e a genética devido a que oferecem um mecanismo explícito para incorporar informação prévia por meio das distribuições de probabilidade a priori. As resultantes distribuições posterioris são intuitivamente interpretadas em contraste aos resultados da estatística clássica (BEAUMONT; RANNALA, 2004; LOPES; BEAUMONT, 2010).

Logo, com o aparecimento do projeto do Genoma Humano (Human Genome Sequencing Consortium, 2001) e o HapMap (The International HapMap Consortium, 2003), na década de 2000 - 2010 houve um aumento na quantidade dos dados. Esses dados descrevem polimorfismos através das populações humanas. O anterior, permitiu o desenvolvimento de novos estudos de genes envolvidos em doenças humanas que requerem rigorosos métodos estatísticos os quais aplicam aproximações Bayesianas. Os pesquisadores enfrentaram diferentes desafios para obter conclusões certas e alcançar melhorias que tornou possível considerar interações em sistemas complexos (LOPES; BEAUMONT, 2010).

Os modelos usados em epidemiologia frequentemente tem muitos parâmetros a serem considerados, como por exemplo . Com a introdução dos métodos Bayesianos foi possível inferir tais valores de forma separada por meio do uso das distribuições marginais posterioris. Estes métodos têm sido evoluídos significativamente o que tem permitido análises de estruturas complexas e variadas. Os métodos Bayesianos também estão

desenhados para afrontar problemas com dados genéticos que são obtidos em diferentes tempos, o que tem sido utilizados em epidemiologia viral (LOPES; BEAUMONT, 2010).

A inferência Bayesiana ganhou atenção nos últimos anos por sua flexibilidade e forma coerente de aproximar-se ao problema de calibração de parâmetros. Nas aproximações Bayesianas é possível especificar as crenças a priori ou informações que se tem dos parâmetros desconhecidos que podem ser obtidos, por exemplo, de experimentações prévias. Adicionalmente, evita-se o problema de superajustamento e mede a incerteza que permanece nos parâmetros desconhecidos restantes (LIEPE et al., 2014).

O objetivo de utilizar inferência de parâmetros por métodos Bayesianos é encontrar a distribuição posterior a qual descreve a incerteza que permanece nos parâmetros depois de observar os dados. Isto é obtido por meio do Teorema de Bayes (Teorema 3.0.11) combinando as crenças prioris com as avaliações do ajuste fornecida aos dados observados.

O interessante da inferência Bayesiana é que não existe diferença lógica entre um modelo de parâmetros e os dados já que ambas são variáveis aleatórias com uma distribuição de probabilidade conjunta especificada por um modelo probabilístico. Neste contexto, definem-se os seguintes termos segundo (BEAUMONT; RANNALA, 2004) e (CSILLÉRY et al., 2010):

Definição 3.0.1 (Variável aleatória). Uma variável aleatória é uma função que atribui um número real a cada um dos resultados de um experimento aleatório.

Definição 3.0.2 (Dado). Variável aleatória observada.

Definição 3.0.3 (Parâmetro). Variável aleatória não observada.

Definição 3.0.4 (Distribuição de probabilidade). A distribuição de probabilidade descreve a probabilidade de que um evento ocorra no futuro. Indica todas os valores possíveis que uma variável aleatória pode receber como resultado de um experimento.

Definição 3.0.5 (Distribuição priori). Uma distribuição priori é uma distribuição de probabilidade dos valores dos parâmetros antes de observar os dados.

Definição 3.0.6 (Distribuição de probabilidade condicional). Uma distribuição de probabilidade condicional é a distribuição de uma ou mais variáveis aleatórias quando outras variáveis aleatórias de uma probabilidade conjunta são fixadas com valores particulares.

Definição 3.0.7 (Verossimilhança). A verossimilhança é a probabilidade condicional que especifica a probabilidade dos dados observados quando é fornecido qualquer conjunto particular de valores para os parâmetros e está baseado num modelo do processo subjacente. A verossimilhança é a função dos parâmetros que descreve em termos de probabilidade, que tão bem cada parâmetro explica os dados observados.

Formalmente, assumindo que as observações $\{D_1, D_2, \dots, D_n\}$ são independentes e identicamente distribuídas. A verossimilhança está definida como

$$L(\theta|D_1 = y_1, D_2 = y_2, \dots, D_n = y_n) = \prod_{i=1}^n f(y_i|\theta). \quad (3.1)$$

Definição 3.0.8 (Distribuição de probabilidade conjunta). A distribuição de probabilidade conjunta é a distribuição de probabilidade de todas as combinações de duas ou mais variáveis aleatórias. Também define-se como o produto entre a verossimilhança e a priori.

Definição 3.0.9 (Distribuição posteriori). A distribuição posteriori é a distribuição de probabilidade condicional dos parâmetros quando se tem os dados observados.

Definição 3.0.10 (Resumo estatístico). Um resumo estatístico é o resumo da localização de um valor de um parâmetro, neste contexto, é usualmente a média, moda, mediana ou o desvio padrão da distribuição posteriori.

Para calcular as distribuições posteriores temos que calcular a verossimilhança dos dados o que significa, dado um modelo com um conjunto de parâmetros θ , especifica-se a probabilidade de cada observação na amostra. Para modelos matemáticos isto é escrever a densidade de probabilidade $f(y|\theta)$ para qualquer observação $y \in D$. Se a distribuição posteriori não tem uma forma conveniente, devem-ser fazer muitos cálculos antes de que possa-se estimá-la ou obter amostras dela.

Também, é necessário para esse cálculo das distribuições posteriores, fornecer uma distribuição priori para θ que deve estar baseada no entendimento prévio de valores prováveis para θ .

Após a observação dos dados, a priori é atualizada por meio da verossimilhança para produzir a distribuição posteriori para θ . Utilizando o teorema de Bayes ([Teorema 3.0.11](#)) a probabilidade a posteriori é

$$\pi(\theta|D) = \frac{L(\theta|D)\pi(\theta)}{\pi(\theta)} = \frac{L(\theta|D)\pi(\theta)}{\int_{\Theta} L(\theta|D)\pi(\theta)d\theta}. \quad (3.2)$$

Tendo a distribuição posteriori de θ pode-se examinar o desenvolvimento aleatório de θ .

A avaliação da distribuição posteriori pode tornar-se muito difícil o que faz que somente se possa utilizar em uns poucos problemas. A dificuldade para avaliar a posteriori deve-se à presença da integral na [Equação 3.2](#) que, em termos realísticos, é intratável. Contudo esta integral é uma constante de normalização, é dizer, que a distribuição posteriori é proporcional à priori vezes a verossimilhança.

$$\pi(\theta|D) \propto L(\theta|D)\pi(\theta). \quad (3.3)$$

ou em palavras,

$$posterior \propto verossimilhança \times priori. \quad (3.4)$$

Se a forma da distribuição de $\pi(\theta|D)$ pode ser deduzida do produto da verossimilhança e a priori, somente precisa-se escrever sua forma de distribuição (por exemplo uma distribuição gama) para calcular as probabilidades posteriores de interesse.

O entusiasmo pelos métodos Bayesianos é devido a que é praticamente impossível escrever de forma analítica uma expressão para a distribuição posterior, possibilitando o surgimento de métodos de aproximação como os Métodos Cadeias de Markov Monte Carlo (MCMC) e Monte Carlo Sequencial (SMC), e da disponibilidade do poder dos computadores suficiente para implementar esses métodos para estimar a posteriori ([SECRIER; TONI; STUMPF, 2009](#); [AESCHBACHER; BEAUMONT; FUTSCHIK, 2012](#); [TURNER; ZANDT, 2012](#)).

Como consequência dos dados disponíveis e o desejo de utilizar sistemas cada vez mais complexos, se fizeram modificações aos métodos MCMC padrões em 2000 e 2003. Mas estes métodos muitas vezes apresentavam problemas de convergência na busca de uma amostra boa da distribuição posteriori dentro de um tempo computacional razoável ([LOPES; BEAUMONT, 2010](#)). Métodos alternativos chamados ABC (Computação Bayesiana Aproximada) foram criados pois estes são utilizados quando a função de verossimilhança é impossível de tratar ou para casos com um grande número de parâmetros que não são de interesse chamados parâmetros ruído. A primeira aplicação foi por ([PRITCHARD et al., 1999](#)) em diferentes estudos do cromossomo Y demonstraram sua competitividade com os correspondentes métodos de verossimilhança completa ([LOPES; BEAUMONT, 2010](#)).

Os métodos Monte Carlo utilizam uma distribuição proposta, como a Gaussiana para encontrar amostras de forma fácil que são filtradas de tal forma que elas são consistentes com a posteriori desejada são preservadas e as demais descartadas.

Quando um método Monte Carlo está implementado de forma apropriada, a teoria das cadeias de Markov assegura que, no limite com uma cadeia suficientemente grande de amostras, a distribuição das amostras filtradas aproxima-se à distribuição da posteriori $\pi(\theta|D)$.

A distribuição priori sempre está disponível porque é selecionada pelo pesquisador. No entanto existem muitos modelos para os quais a verossimilhança pode ser difícil ou até impossível de especificar matematicamente. Este problema é muito comum em modelos baseados em simulação ou modelos computacionais, os quais geram previsões simulando o mecanismo de geração de dados. Estes modelos são muito frequentes em ciências sociais, e psicologia cognitiva em particular (TURNER; ZANDT, 2012). Nestes casos, a aplicação de um método padrão de estimação Bayesiana, assim como estimação da clássica verossimilhança máxima não tem sido possível.

Teorema 3.0.11 (Teorema de Bayes). *Seja Ω um espaço amostral e $B_1, B_2, \dots, B_n$, uma coleção de eventos mutuamente exclusivos de tal forma que*

$$i. \bigcup_{i=1}^n B_i = \Omega$$

$$ii. P(B_i) > 0$$

para $i = 1, 2, \dots, n$, então para qualquer evento A de Ω

$$P(B_i/A) = \frac{P(B_i)P(A/B_i)}{\sum_{i=1}^n P(B_i)P(A/B_i)}$$

Assumimos que os dados surgem de um modelo estocástico de Markov que engloba dados identicamente distribuídos, parametrizados por um vetor θ o qual é o alvo da inferência e denotamos $\pi(x|\theta)$ a densidade de probabilidade dos dados quando um valor específico de θ é fornecido. A densidade posteriori está dada por $\pi(\theta|x)$ que é proporcional $\pi(x|\theta)\pi(\theta)$ (WHITE; KYPRAIOS; PRESTON, 2013)

A verossimilhança é usualmente definida por um modelo de probabilidade paramétrica, $P(D|\theta)$ como função dos parâmetros θ e com D o conjunto fixo de dados.

A diferença às aproximações de máxima verossimilhança, donde a função objetivo é a verossimilhança e utiliza aproximações na procura do melhor vetor de parâmetros que maximiza $P(D|\theta)$, as aproximações Bayesianas elucidam, ou pelo menos obtém amostras dos vetores de parâmetros da distribuição posteriori $P(D|\theta)$. Usualmente é impossível escrever uma expressão para a distribuição posterior de forma analítica, nestes casos as técnicas de Cadeias de Markov Monte Carlo são uma alternativa realizar aproximações computacionais destas distribuições.

Vale a pena reafirmar que a inferência Bayesiana tenta avaliar a probabilidade de que um parâmetro seja o correto tendo os dados. Isto implica também uma avaliação da incerteza da inferência, devido a que todos os valores dos parâmetros cuja probabilidade finita de ter gerado os dados, são o objetivo do processo de inferência. Esta incerteza pode ter um papel muito importante nas análises de um sistema dinâmico e ao fato de que o comportamento previsto pelo modelo é robusto a alterações aos parâmetros, especialmente quando a distribuição sobre as diferentes taxas de reação são consideradas em conjunto. Isto é, por exemplo, se um modelo tem dois parâmetros, o objetivo é encontrar a distribuição conjunta sobre ambos parâmetros. Frequentemente, a distribuição conjunta é diferente ao produto das probabilidades marginais dos parâmetros isolados, isto é um indicativo de que os parâmetros são estatisticamente dependentes. O anterior faz que o custo computacional seja mais alto, as vantagens obtidas pela consideração das distribuições conjuntas pode elevar de forma considerável os custos computacionais.

Ao fazer análises dos dados no contexto de um modelo matemático, sempre tem se que calibrar o modelo com os dados disponíveis isto é, estimar os parâmetros diretamente dos dados. Baseando-se somente nos valores de parâmetros obtidos de forma independente, como por exemplo da literatura, se podem ter muitos problemas potenciais, pelo fato de não terem sido considerados muitos fatores no modelo que pode causar mudanças no sistema.

4 Computação Bayesiana Aproximada (ABC)

Para modelar um sistema biológico é necessário saber sua estrutura e os parâmetros¹ que regulam sua dinâmica, entretanto para a maioria dos processos biológicos estes valores são desconhecidos ou muito difíceis de medir de forma direta por isso a necessidade de estimação dos dados observados. Assim, os métodos de Computação Bayesiana Aproximada (ABC), permitem melhor compreender o sistema dinâmico aproximando a distribuição posteriori sobre os parâmetros.

Os métodos ABC ou métodos livres de verossimilhança, foram desenvolvidos originalmente no ano 1999 por (PRITCHARD et al., 1999) para resolver problemas de genética de populações (BEAUMONT; ZHANG; BALDING, 2002; FILIPPI et al., 2012; LOPES; BEAUMONT, 2010; WHITE; KYPRAIOS; PRESTON, 2013), apesar das ideias relacionadas com ABC terem surgido com Donald Rubin em 1984, quando descreveu um mecanismo hipotético para amostragem. Esse mecanismo consiste em tomar uma amostra de uma distribuição posteriori e cujo mecanismo é igual ao esquema do ABC rejeição explicado na). Adicionalmente, Rubin apontou que em inferência Bayesiana era necessário considerar métodos computacionais para estimar a distribuição posteriori de interesse nos modelos analiticamente intratáveis (SUNNÅKER et al., 2013). Nesse mesmo ano, Diggle e Gratton (1984) introduziram pela primeira vez forma esquemática usando simulação para aproximar a função de verossimilhança e realizar inferência estatística em situações com verossimilhança intratáveis (SUNNÅKER et al., 2013). (TAVARÉ et al., 1997) propôs o primeiro algoritmo ABC, para inferir a distribuição posteriori. Com Mark Beaumont em 2002, o termo de *Approximate Bayesian Computation* foi formalmente estabelecido (BEAUMONT; ZHANG; BALDING, 2002) no seu trabalho de demografia humana (WHITE; KYPRAIOS; PRESTON, 2013).

Os métodos ABC aplicam-se a problemas que são um desafio para a inferência estatística convencional, utilizando o formalismo da inferência Bayesiana em situações

¹ Parâmetro é uma quantidade numérica calculada sobre uma população que resume os valores que esta toma em algum atributo

onde é impossível escrever uma expressão para a função de verossimilhança $L(\theta|D)$. Para evitar o uso das funções de verossimilhança, foram empregadas simulações Monte Carlo fazendo que estes métodos sejam muito flexíveis e permitindo assim, analisar uma grande quantidade de dados. Também, como consequência disto, podem ser aplicadas em uma variedade de problemas com verossimilhanças intratáveis, para analisar sistemas complexos decorrentes em genética de populações, epidemiologia e biologia de sistemas convertendo-o em um método muito popular (SECRIER; TONI; STUMPF, 2009; BEAUMONT, 2010; LIEPE et al., 2014; LOPES; BEAUMONT, 2010). A utilização destes métodos estendeu-se aos campos da filogenética, genética de conservação, evolução molecular, biologia de sistemas e epidemiologia de doenças infecciosas (Figura 9 e Figura 10). A seguir são descritas algumas aplicações do ABC:

Tanaka et al. (2006), o utilizou para estimar parâmetros típicos epidêmicos da transmis-

1997	Tavaré et al. Fu e Li
1998	Weiss e Haeseler
1999	Pritchard et al.
2000	
2001	Estoup et al.(4)
2002	Beaumont et al.(1)
2003	Marjoram e Tavaré(1); Plagnol e Tavaré, Estoup e Clegg.
2004	Tallmon et al.; Estoup et al.
2005	Excoffier et al.; Miller et al.(2); Hamilton et al.(2); Hickerson et al. Miller et al.(4); Wright et al.
2006	Thomton e Andolfatto; Hickerson et al.(2); Tanaka et al.(3); Shiner et al.(3); Chan et al.(4); Jakobsson et al.
2007	Rosenblum et al.; Ratman et al.; Sisson et al.(1); Fagundes et al.(2); Pascual et al.; Legras et al.(2);....
2008	Beaumont; Neuenschwander et al.; Bonhomme et al.; Grelaud et al.(2); Cornuet et al.; Wilson et al.(3);...

Figura 9 – Uso do ABC desde sua aparição até 2008, em campos particulares como: (1) filogenética, (2) epidemiologia, (3) teoria e (4) conservação. Fonte: (LOPES; BEAUMONT, 2010)

são da tuberculose utilizando dados do genótipo do *Mycobacterium tuberculosis* isolados dos pacientes em São Francisco. Shiner et al. (2006) estudou diversas hipóteses da evolução da diversidade genética do HIV-1 em um só indivíduo. Wilson et al. (2009) apresentou resultados sobre história evolutiva do *Campylobacter jejuni*, em quanto Toni et al. (2009) aplicou para um modelo SIR (Susceptible Infected Recovered) que descreve surtos de frio em Tristão da Cunha no Oceano Atlântico. Tallmon, Luikart e Beaumont

(2004) aplicou o método para estimar parâmetros demográficos, neste caso, o tamanho da população. Chan, Anderson e Hadly (2006), fez estimação de parâmetros demográficos históricos de um DNA antigo. Ross-Ibarra et al. (2008), para estimar parâmetros demográficos. McKinley, Cook e Deardon (2009) aplicou o método à epidemiologia. Guillemaud et al. (2010), estimou a probabilidade de rotas de introdução de populações invasivas baseadas em dados microssatélites. Drovandi e Pettitt (2011), estimou parâmetros para evolução de microparasitos. E Dean et al. (2011) utilizou para finanças entre outras aplicações.

2009	Patin, <i>et al.</i> ; Verdu <i>et al.</i> ; Touchon <i>et al.</i> ; Quachet <i>et al.</i> ; Gerbault <i>et al.</i> ; Itan <i>et al.</i> ; Wilkinson e Tavaré; Carnaval <i>et al.</i> ; Templeton, . . .
2010	Guillemaud, <i>et al.</i> ; Ray <i>et al.</i> ; Ghirotto <i>et al.</i> ; Lopes <i>et al.</i> ; Templeton
2011	Toni <i>et al.</i> ; Drovandi e Pettitt
2012	Beguerisse-Díaz <i>et al.</i> ; Liepe <i>et al.</i> ; Ratmann <i>et al.</i> ; Silk <i>et al.</i>
2013	
2014	Jou, Boys e Wilkinson,



Figura 10 – Uso do ABC desde 1999 nos campos de biologia de sistemas, ecologia e biologia evolutiva

O objetivo do método ABC é estimar a distribuição posteriori dos parâmetros em modelos baseados em simulação, aproximando a média e a curva de densidade da distribuição posteriori de um parâmetro sem precisar do cálculo explícito da verossimilhança (BEAUMONT; ZHANG; BALDING, 2002; SECRIER; TONI; STUMPF, 2009; TURNER; ZANDT, 2012). Para conseguir isto, simula-se o modelo produzindo um conjunto de dados artificial D^* gerados de uma distribuição de probabilidade, para logo reduzir o tamanho dos dados por meio de estatísticas resumo e compará-lo com o conjunto de dados que já foram observados D por meio de uma função distância em uma métrica (TURNER; ZANDT, 2012) a qual irá a comparar as estatísticas resumo de ambos conjuntos, reduzindo significativamente a quantidade de dados. (CSILLÉRY *et al.*, 2010; LOPES; BEAUMONT, 2010). O fato de que o ABC baseia-se nas estatísticas resumo para diminuir o tamanho dos conjuntos, é um indicador de que o algoritmo ABC tem suas raízes no algoritmo de rejeição que consiste em gerar amostras de uma distribuição de probabilidade que consta na simulação de grandes conjuntos de dados sob um cenário evolutivo (CSILLÉRY *et al.*, 2010).

O método tem sido validado por meio de exemplos onde a posteriori é conhecida realizando comparações entre as posteriores estimadas por ABC e as posteriores verdadeiras e verificando que os algoritmos recuperam as posteriores verdadeiras de forma eficiente (TURNER; ZANDT, 2012).

Existem muitas variações do método ABC, a aproximação mais simples de ABC é o ABC Rejeição Amostral que foi introduzido na área de genética de populações. As crenças prioris sobre θ estão expressadas por $\pi(\theta)$. O algoritmo é descrito a seguir (LIEPE et al., 2014),

1. Amostrar um vetor de parâmetros θ^* da distribuição priori;
2. simular um conjunto de dados D^* sintético;
3. utilizar uma função distância d para medir a discrepância entre D^* e o conjunto de dados observados D ,
4. aceite θ^* se $d(D, D^*)$ é menor que uma tolerância estabelecida ϵ .

O processo pode ser repetido várias vezes para construir um conjunto de vetores de parâmetros aceitos. É importante observar que se o conjunto de dados observados D possuem ruído, então para evitar algum viés, o conjunto de dados sintéticos D^* o qual para características conhecidas do ruído, pode ser incorporado em forma direta. Em alguns casos, como ao modelar utilizando equações diferenciais ordinárias (ODE) simplesmente especificando um modelo para o ruído, levará implícito uma verossimilhança (LIEPE et al., 2014).

Assim que o nível de tolerância ϵ , enquanto ϵ tende a zero, o conjunto de parâmetros aceitos representará uma amostra da distribuição posteriori $P(\theta|D)$. Na prática se ϵ é muito pequeno, a taxa de aceitação, isto é a proporção de vezes que temos $d(D, D^*) < \epsilon$, será inaceitavelmente baixa. O anterior, junto com suas implicações computacionais, incentivou a introdução de aproximações sequenciais, utilizando uma sucessão decrescente de valores para ϵ de tal forma que as aproximações gradualmente aproximam-se desde amostras da priori, quando ϵ é muito grande, até uma amostra da posteriori, quando ϵ tende a zero (LIEPE et al., 2014).

Como os modelos e o espaços de parâmetros tem tamanhos muito grandes, o algoritmo ABC estendeu-se ao algoritmo ABC-MCMC (Cadeias de Markov Monte Carlo) e ao SMC (Sequential Monte Carlo) e foi nesse momento que iniciaram-se as pesquisas que ajudaram ao avanço do método ABC e suas aplicações. A verossimilhança baseada em MCMC, ou em SMC e amostragem aninhadas também tem ajudado a melhorar o desenvolvimento dos sistemas dinâmicos estocásticos. Isto é importante nos casos onde a verossimilhança de um conjunto de realizações estocásticas de um sistema pode-se aproximar computacionalmente. Uma forma é por meio da aproximação linear do ruído ou o a generalização da série de tempo estocástica no sistema dinâmico (LIEPE et al., 2014).

4.1 Computação Bayesiana Aproximada Monte Carlo Sequencial (ABC-SMC)

O método ABC Monte Carlo Sequencial (ABC-SMC) foi proposto para fazer o algoritmo ABC mais eficiente porque oferece grande flexibilidade, e aplicabilidade já que a característica principal do ABC-SMC é a construção de uma série de distribuições intermédias que iniciam em uma distribuição priori adequada e vai incrementando a semelhança à uma distribuição posteriori desconhecida. Essas distribuições são construídas estimando distribuições intermediárias $P_{\epsilon_t}(\theta|x)$ da sucessão decrescente $\{\epsilon_t\}_{1 \leq t \leq T}$. Cada uma destas distribuições está descrita por uma amostra com peso, de vetores de parâmetros. A sucessão de distribuições é construída amostrando parâmetros da população anterior, fazendo uma perturbação sobre elas através da função kernel, $\tilde{\theta} \sim K_t(\cdot|\theta)$, gerando os dados simulados $y \sim f(\cdot|\tilde{\theta})$ e sob aceitação calculando os novos pesos correspondentes (FILIPPI et al., 2012). Esta função kernel, contribui na construção de distribuições que começam com uma adequada distribuição priori e que convergem para uma distribuição posteriori desconhecida.

A eficiência computacional não somente depende da complexidade do modelo, nem da quantidade de dados, mas também da sucessão $\{\epsilon_t\}_{1 \leq t \leq T}$, e na escolha dos kernels de perturbação.

Para modelos que demandam um alto custo computacional, como os sistemas di-

nâmicos, a escolha do kernel terá grande influência na eficiência no espaço de parâmetros é explorado e na estimação da distribuição posteriori.

Durante os últimos anos, tem-se realizado modificações para melhorar o algoritmo ABC. Nas diferentes propostas está a de (MARJORAM et al., 2003) que introduziram o método baseado em Cadeias de Markov Monte Carlo (ABC-MCMC). Estas consistem em construir uma cadeia de Markov com distribuição estacionária $p_\epsilon(\theta|x)$. Para este algoritmo a convergência está garantida, contudo é muito difícil avaliar quando a cadeia de Markov alcança um estado estacionário; além disso a cadeia pode ficar em um extremo local.

(SISSON; FAN; TANAKA, 2007), (TONI et al., 2009), (BEAUMONT et al., 2009), introduziram o amostragens SIS (Sequential Importance Sampling) e SMC (Sequential Monte Carlo) no ambiente ABC. O objetivo destes métodos é amostrar sequencialmente as distribuições $p_{\epsilon_t}(\theta|x)$ para uma sucessão decrescente $\{\epsilon_t\}_{1 \leq t \leq T}$. O algoritmo 1 descreve a implementação por (TONI et al., 2009) e (BEAUMONT et al., 2009).

(BEAUMONT et al., 2009) propôs o algoritmo ABC Monte Carlo Populacional (ABC-PMC de seu acrônimo em inglês) que apresenta muitas semelhanças com o algoritmo ABC-SMC com exceção da presença de um kernel de perturbação. A vantagem destes métodos sobre o ABC-MCMC é que as amostras da posteriori resultante são independentes e que devido a seu carácter sequencial os níveis de tolerância não devem especificar as prioris para as análises a menos que eles sejam ajustados de forma adaptativa (SUNNÅKER et al., 2013).

Segundo (BEAUMONT et al., 2009), quando se utilizam amostragens SMC, é preciso definir kernels para a frente e para trás, já que estes reduzem a complexidade computacional de $O(N^2)$ até $O(N)$ onde N é o número de partículas. Contudo, nas aplicações a parte mais custosa computacionalmente é a simulação dos dados, que apesar de ter uma complexidade de $O(N)$, a dominância é da $O(N^2)$ para qualquer valor de N .

Antes de ver os requerimentos teóricos para a convergência em ABC-SMC, enunciaremos alguns dos resultados mais importantes da estatística. Um deles a **Lei dos grandes números 4.1** é um dos vários teoremas que expressam a ideia de que em quanto o número de tentativas aumenta em um processo aleatório, ou percentagem da diferença

entre os valores esperados e os atuais tende ao zero. O outro é o teorema Central do Limite (Teorema 4.1.1):

Teorema 4.1.1 (Teorema central do limite). *Seja $\bar{X}_n$ a média amostral de uma amostra aleatória de tamanho n obtida de uma população com densidade $f_X(x)$, média μ e variância $\sigma^2 < \infty$. Seja Z_n a variável aleatória definida por $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$. Então Z_n é aproximadamente normal com média $\mu = 0$ e variância $\sigma^2 = 1$ quando n tende a ∞ .*

É construído um conjunto de distribuições intermediárias, chamadas populações, onde, para cada população t , todas as partículas aceitas geram o conjunto de dados simulados D^* . Estes dados simulados diferem dos verdadeiros dados experimentais D por no máximo uma distância $d(D, D^*) < \epsilon_t$. Esta aproximação precisa de uma sucessão decrescente de tolerâncias ϵ_t , com $t = 1, \dots, T$ especificando o último ajuste entre os dados reais e os simulados. O processo está resumido na Figura 11.

Uma sucessão de populações t , $t = 1, \dots, T$ vai-se gerando de tal forma que a população t é gerada pela população $t - 1$ se $t \neq 1$ ou pela priori quando $t = 1$. Com o propósito de garantir um esquema de amostragem da importância, as partículas são perturbadas utilizando um kernel de perturbação adequado para possibilitar que o espaço de parâmetros seja explorado suficientemente bem. Cada partícula aceita tem um peso associado.

Isto é, para cada população t ,

1. Amostre da proposta $\eta_t(\theta_t) = \int \pi_{t-1}(\theta_{t-1}) K_t(\theta_{t-1}, \theta_t) d\theta_{t-1}$, onde $K_t(\theta_{t-1}, \theta_t)$ é o kernel de Markov de perturbação.
2. A partícula é aceita se $d(D^*, D) < \epsilon_t$.
3. Vincular o peso à partícula.

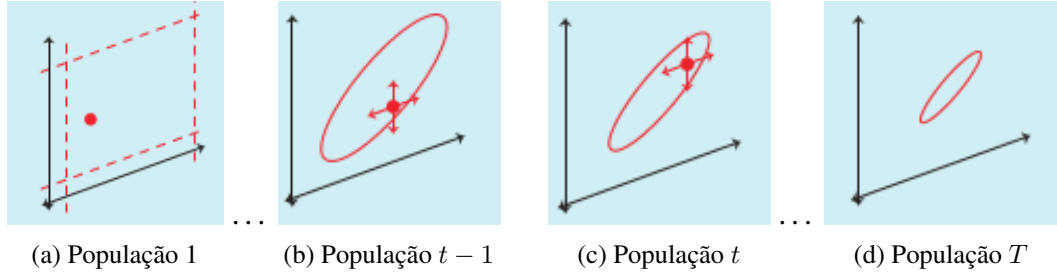


Figura 11 – O algoritmo ABC-SMC: Define-se o conjunto de distribuições intermediárias.

(a) Priori: $\pi(\theta)$. (b) $\pi_{t-1}(\theta|d(D^*, D) < \epsilon_{t-1})$. (c) $\pi_t(\theta|d(D^*, D) < \epsilon_t)$. (d) $\pi_T(\theta|d(D^*, D) < \epsilon_T)$.

Fonte: (LIEPE et al., 2014)

Na teoria para a convergência em ABC-SMC requer-se,

- Ter um suporte maior que a distribuição alvo, para garantir um não enviesamento assintótico da média empírica pela lei dos grandes números .
- Garantir variância finita dos pesos de importância para assegurar a normalidade assintótica e variância finita do estimador pelo teorema central do limite 4.1.1.

Perturbar uma partícula $\theta^{(j,t-1)}$ significa gerar uma nova partícula em conformidade a uma probabilidade parametrizada de $\theta^{(j,t-1)}$ e muitas vezes centrada em $\theta^{(j,t-1)}$. No ABC-SMC além de amostrar do kernel, deve-se ter confiança de calcular a densidade de transição $k_t(\theta^{(i,t)}|\theta^{(j,t-1)})$ para qualquer partículas $\theta^{(i,t)}$ e $\theta^{(j,t-1)}$. Então o espaço de todos os possíveis kernels é restringido a uma família paramétrica do qual é fácil amostrar e realizar otimizações. Podem ser utilizados diferentes modelos de probabilidade, sendo a distribuição uniforme, e a Gaussiana as mais comuns. (Para maiores detalhes ver (FILIPPI et al., 2012)).

Segundo (FILIPPI et al., 2012), para a escolha do kernel em ABC-SMC, tendo uma população multi-dimensional, pode ser mais eficiente considerar as correlações entre os diferentes elementos do vetor de parâmetros. Assim, fazendo uma perturbação seguindo uma distribuição normal multivariada a qual depende do kernel da população anterior. Quando se utiliza no kernel uma distribuição normal multivariada, poucas

partículas propostas são rejeitadas comparando com as distribuições por componentes normal e uniforme.

O kernel de perturbação normal multivariada se baseia na matriz de covariância que depende da população anterior. Para calcular matriz de covariância ótima utiliza-se a aproximação de minimização da divergência de Kullback-Leibler 4.1.2. Baseados na divergência de Kullback-Leibler 4.1.2, (PINTO, 2007), calcula-se a divergência entre uma distribuição e a distribuição de partículas perturbadas.

Definição 4.1.2 (A divergência de Kullback-Leibler). Sejam p_0, p_1 densidades de probabilidade sobre um conjunto finito X . A distância Kullback-Leibler D_{KL} é definida da seguinte forma

$$D_{KL}(p_1||p_0) = \sum_{x \in X} p_1(x) \log_2 \frac{p_1(x)}{p_0(x)}.$$

Esquema do método:

- Construir uma amostra de $P_{\epsilon_1}(\theta|x)$ com um valor para ϵ_1 suficientemente grande tal como o número de partículas aceitadas.
- Utilize o algoritmo ABC novamente para construir $P_{\epsilon_2}(\theta|x)$ com tolerância ϵ_2 , não amostrando da priori, mas do conjunto de partículas aceitas no passo anterior e perturbando tendo em consideração o kernel de perturbação adequado.
- Repita o mesmo processo até construir $P_{\epsilon_2}(\theta|x)$

O desenvolvimento do algoritmo depende principalmente na sucessão $\{\epsilon_t\}_{1 \leq t \leq T}$ assim, se a diferença entre ϵ_{t+1} e ϵ_t é pequena, então as correspondentes distribuições posteriores $p_{\epsilon_{t+1}}(\theta|x)$ e $p_{\epsilon_t}(\theta|x)$ são semelhantes e assim poucas simulações seriam necessárias para gerar N partículas da seguinte distribuição intermediária, $p_{\epsilon_{t+1}}(\theta|x)$ por meio do amostragem da população com peso $\{\theta^{(i,t-1,w^{i,t-1})}\}_{1 \leq t \leq T}$. Mas uma sucessão lenta de tolerâncias $\{\epsilon_t\}_{1 \leq t \leq T}$ produz uma grande quantidade de iterações se o T é grande para encontrar $\epsilon_T = \epsilon$. Na prática a sucessão de tolerâncias é realizada a mão, conforme ao modelo.

(MORAL; DOUCET; JASRA, 2011) e (DROVANDI; PETTITT, 2011) propõem um método de adaptação para a escolha de tal sucessão que consiste na seleção do

α -quantil das distâncias entre os dados simulados $y^{(i,t)}_{1 \leq t \leq T} \in D^*$ e o observado $x \in D$. Seleccionando a tolerância de forma adaptativa frequentemente melhora a eficiência do algoritmo ABC-SMC. Embora, a eficiência dependa fortemente da escolha do α e como em (SILK; FILIPPI; STUMPF, 2012) para alguns valores de α o algoritmo não converge, ou pode levar a obter uma distribuição posteriori diferente à posteriori verdadeira. (SILK; FILIPPI; STUMPF, 2012) propõem um novo método que experimentalmente alcança melhores resultados porque sempre converge e computacionalmente é muito mais efetivo que o método dos quantis. No entanto, o pacote ABC-SysBio tem implementado o método dos quantis.

E também a eficiência computacional do algoritmo depende da escolha dos kernels de perturbação $\{K_t(.|.)\}_{1 \leq t \leq T}$. Um kernel de perturbação local, causa um grande deslocamento das partículas e produz novas partículas que com a qualidade de ter alta probabilidade de ser aceitas se os valores da sucessão $\{\epsilon_t\}_{1 \leq t \leq T}$ são o suficientemente próximos. Por outro lado, se o kernel permite valores muito espalhados das tolerâncias, será explorar de forma mais completa o espaço de parâmetros, porém consegue isso com taxas de aceitação muito baixas.

Para um número suficientemente grande de partículas, a aproximação da população pode em princípio evitar o problema de ficar preso em áreas de baixa probabilidade como acontece nas aproximações MCMC (SECRIER; TONI; STUMPF, 2009).

Algoritmo 1: O algoritmo ABC-SMC**Entrada:** Uma tolerância ϵ **Saída:** Uma amostra com peso de partículas de $P_\epsilon(\theta|x)$ **início** $t \leftarrow 0;$ **repita** $t \leftarrow t + 1;$ determine a seguinte tolerância ϵ_t ;determine os parâmetros do kernel de perturbação $K_t(.|..)$; $i \leftarrow 1;$ **repita****se** $t = 1$ **então**| amostre $\tilde{\theta}$ da $\pi(\theta)$;**senão****fim**amostre D^* de $f(.|\tilde{\theta})$;amostre $\tilde{\theta}$ tal que $\pi(\tilde{\theta}) > 0$ amostre D^* de $f(.|\theta)$;**se** $d(D^*, D) \leq \epsilon_t$ **então**| $\theta_t \leftarrow \tilde{\theta}$;| $D_t^* \leftarrow D^*$;| $i \leftarrow i + 1$;**senão****até** $i = N + 1$;Calcule os pesos: para todo $1 \leq i \leq N$;**se** $t \neq 1$ **então**

$$w_t \leftarrow \frac{\pi(\theta_t)}{\sum_{j=1}^n w_{t-1}^j K_t(\theta_t | \theta_{t-1}^j)}$$

senão**fim** $w_1 \leftarrow 1$ normalize os pesos**até** $\epsilon_t \leq \epsilon$; $T \leftarrow t$ **fin**

5 O modelo matemático

Como está descrito em (POKHILKO et al., 2012), o modelo matemático que corresponde ao ciclo circadiano correspondente à Figura 2, é composto por 28 equações diferenciais com 104 parâmetros. Os valores de 43 deles foram baseados em dados disponíveis e o valor dos coeficientes de Hill¹ foi fixado em 2.

As equações para os componentes da manhã LHY, PRR9, PRR7 e NI(PRR5), são similares ao trabalho em (POKHILKO et al., 2010).

Na Equação 5.1 para LHY, TOC1 é um inibidor ao igual que os PRRs. Também a resposta à luz aguda em LHY foi assumida como livre de inibição similarmente que à respostas à luz aguda de PRR9 e GI.

Na Equação 5.15 se introduz a inibição da expressão de PRR9 por EC.

Na Equação 5.18, Equação 5.21 e Equação 5.26, ELF3, ELF4 e LUX são inibidos por LHY e CCA1, baseados em dados publicados por Dixon et al. (2011), Hazen et al. (2005), Li et al. (2011), Portolés e Más (2010). Além, EC inibe a expressão de LUX e ELF4.

Na Equação 5.22 e Equação 5.23 introduze-se as frações nucleares e citosólicas de COP1 para descrever a regulação diferencial da proteína ELF3 pelas frações nucleares e citosólicas de COP1.

Na Equação 5.19, Equação 5.23, Equação 5.27 são apresentados os termos para a ligação entre as proteínas ELF3, ELF4 e LUX, em quanto que as ligações presentes entre ELF3 e GI estão descritas em 5.4, 5.10, 5.22 e 5.23.

Os termos para degradação da proteína ELF3 e seus complexos por COP1 estão representados na Equação 5.4, Equação 5.13, Equação 5.22 e Equação 5.23. A equação Equação 5.13 para a EC também tem um termo que considera a aceleração da degradação

¹ O coeficiente de Hill descreve como uma substância ou ligante unirá-se a uma proteína receptora quando se têm outros ligantes. Este indicador estima o nível de cooperatividade no processo de união. Um coeficiente de 1 indica que a união é um processo independente da presença de outros ligantes. Os valores maiores a 1 indicam uma propensão positiva à união, enquanto que valores menores a 1 indicam uma propensão negativa.

de EC pelo a quantidade total de GI.

A equação [Equação 5.14](#) é a expressão de GI que mostra a inibição por EC.

Nas equações para ZTL e ZTL-GI ([Equação 5.3](#) e [Equação 5.12](#)), são como em ([POKHILKO et al., 2010](#)). A expressão de TOC1 na equação [Equação 5.6](#) sendo inibido por LHY/CCA1 e por EC. A equação para TOC1 [Equação 5.11](#).

$$\frac{dc_L^m}{dt} = q_1 L c_p + n_1 \frac{g_1^a}{g_1^a + (c_{p9} + c_{p7} + c_{NI} + c_T)^a} - (m_1 L + m_2 D) \cdot c_L^m \quad (5.1)$$

$$\frac{dc_p}{dt} = p_7 D \cdot (1 - c_p) - m_{11} c_p L \quad (5.2)$$

$$\frac{dc_{ZG}}{dt} = p_{12} L c_{ZTL} c_G - p_{13} c_{ZG} D - m_{21} c_{ZG} \quad (5.3)$$

$$\frac{dc_{EG_c}}{dt} = p_{17} c_{E3_c} c_{G_c} - m_9 c_{EG_c} c_{COP1_c} p_{18} c_{EG_c} + p_{31} c_{EG_m} \quad (5.4)$$

$$\frac{dc_L}{dt} = (p_2 + p_1 L) \cdot c_L^m - m_3 c_L - p_3 \frac{c_L^e}{c_L^e + g_3^e} \quad (5.5)$$

$$\frac{dc_T^m}{dt} = n_2 \cdot \frac{g_4}{g_4 + c_{Ec}} \cdot \frac{g_5^e}{g_5^e + c_L^e - m_5 c_T^m} \quad (5.6)$$

$$\frac{dc_{p9}}{dt} = p_8 c_{p9}^m - (m_{13} + m_{22} D) \cdot c_{p9} \quad (5.7)$$

$$\frac{dc_{NI}^m}{dt} = n_{10} \frac{c_{Lmod}^e}{g_{12}^e + c_{Lmod}^e} + n_{11} \frac{c_{p7}^b}{g_{13}^b + c_{p7}^b} - m_{16} c_{NI}^m \quad (5.8)$$

$$\frac{dc_{NI}}{dt} = p_{10} c_{NI}^m - (m_{17} + m_{24} D) \cdot c_{NI} \quad (5.9)$$

$$\frac{dc_{Gc}}{dt} = p_{11} c_G^m - p_{12} L c_{ZTL} c_{Gc} + p_{13} c_{ZG} D - m_{19} c_{Gc} - p_{17} c_{E3_c} c_{Gc} - p_{28} c_{Gc} + p_{29} c_{Gn} \quad (5.10)$$

$$\frac{dc_T^m}{dt} = p_4 c_T^m - (m_6 + m_7 D) \cdot c_T (c_{ZTL} \cdot p_5 + c_{ZG}) - m_8 c_T \quad (5.11)$$

$$\frac{dc_{ZTL}}{dt} = p_{14} - p_{12} L c_{ZTL} c_G + p_{13} c_{ZG} D - m_{20} c_{ZTL} \quad (5.12)$$

$$\begin{aligned} \frac{d_{EC}}{dt} = & p_{26} c_{LUX} c_{E34} - m_{36} c_{EC} c_{COP1_n} - m_{37} c_{EC} \cdot c_{COP1_d} + \\ & - m_{32} c_{EC} \left(1 + p_{24} \cdot L \cdot \frac{c_{Gn_tot}^d}{g_7^d + c_{Gn_tot}^d} \right) \end{aligned} \quad (5.13)$$

$$\frac{dc_G^m}{dt} = Lq_2c_P + n_{12}\frac{g_{14}}{g_{14} + c_{EC}} \cdot \frac{g_{15}^e}{g_{15}^e + c_L^e} - m_{18}c_G^m \quad (5.14)$$

$$\frac{dc_{p9}^m}{dt} = L \cdot q_3 \cdot c_P + \frac{g_8}{g_8 + C_{EC}} \left(n_4 + n_7 \cdot \frac{c_L^e}{g_9^e + c_L^e} \right) - m_{12}c_{p9}^m \quad (5.15)$$

$$\frac{dc_{p7}^m}{dt} = n_8 \frac{c_{Ltot}^e}{g_{10}^e + c_{Ltot}^e} + n_9 \frac{c_{p9}^f}{g_{11}^f + c_{p9}^f} - m_{14}c_{p7}^m \quad (5.16)$$

$$\frac{dc_{p7}}{dt} = p_9c_{p7}^m - (m_{15} + m_{23}D) \cdot c_{p7} \quad (5.17)$$

$$\frac{dc_{LUX}^m}{dt} = n_{13} \cdot \frac{g_2}{g_2 + c_{EC}} \cdot \frac{g_6^e}{g_6^e + c_L^e} - m_{34}c_{LUX}^m \quad (5.18)$$

$$\frac{dc_{E4}}{dt} = p_{23}c_{E4}^m - m_{35}c_{E4} - p_{25}c_{E4}c_{E3n} + p_{21}c_{E34} \quad (5.19)$$

$$\frac{dc_{Lmod}}{dt} = p_3 \frac{c_L^c}{c_L^c + g_3^c} - m_4c_{Lmod} \quad (5.20)$$

$$\frac{dc_{E3}^m}{dt} = n_3 \frac{g_{16}^e}{g_{16}^e + c_L^e} - m_{26}c_{E3}^m \quad (5.21)$$

$$\frac{dc_{E3c}}{dt} = p_{16}c_{E3}^m - m_9c_{E3c}c_{COP1c} - p_{17}c_{E3c}c_{Gc} - p_{19}c_{E3c} + p_{20}c_{E3n} \quad (5.22)$$

$$\begin{aligned} \frac{dc_{E3n}}{dt} = & p_{19}c_{E3c} - p_{20}c_{E3n} - p_{17}c_{E3n}c_{Gn} - m_{30}c_{E3n} \cdot c_{COP1d} + \\ & - m_{29}c_{E3n} \cdot c_{COP1n} + p_{21}c_{34} - p_{25}c_{E4}c_{E3n} \end{aligned} \quad (5.23)$$

$$\frac{dc_{COP1n}}{dt} = p_6c_{COP1c} - n_6L \cdot c_P \cdot c_{COP1n} - n_{14}c_{COP1n} - m_{27}c_{COP1n}(1 + p_{15}L) \quad (5.24)$$

$$\frac{dc_{COP1d}}{dt} = n_{14}c_{COP1n} + n_6L \cdot c_P \cdot c_{COP1n} - m_{31}(1 + m_{33}D) \cdot c_{COP1d} \quad (5.25)$$

$$\frac{dc_{E4}^m}{dt} = n_{13} \cdot \frac{g_2}{g_2 + c_{EC}} \cdot \frac{g_6^e}{g_6^e + c_L^e} - m_{34}c_{E4}^m \quad (5.26)$$

$$\frac{dc_{LUX}}{dt} = p_{27}c_{LUX}^m - m_{39}c_{LUX} - p_{26}c_{LUX} - p_{26}c_{LUX}c_{E34} \quad (5.27)$$

$$\frac{dc_{COP1c}}{dt} = n_5 - p_6c_{COP1c} - m_{27}c_{COP1c}(1 + p_{15}L) \quad (5.28)$$

$$c_{E34} = \frac{p_{25}c_{E4}c_{E3n}}{p_{26}c_{LUX} + p_{21} + m_{37}c_{COP1d} + m_{36}c_{COP1n}} \quad (5.29)$$

$$c_{EGn} = \frac{p_{18}c_{EGc} + p_{17}c_{E3n}c_{Gn}}{m_9c_{COP1n} + m_{10}c_{COP1d} + p_{31}} \quad (5.30)$$

$$c_{Gn} = \frac{p_{28}c_{Gc}}{p_{29} + m_{19} + p_{17}c_{E3n}} \quad (5.31)$$

$$c_{Gn_tot} = c_{Gn} + c_{EGn} \quad (5.32)$$

Onde c_i^m e c_i , são as concentrações adimensionais do mRNA e da proteína respectivamente, com i os componentes moleculares LHY/CCA1 mRNA o proteína (L). Os parâmetros m_j são constantes de taxa de degradação. Os parâmetros n_j são constantes de taxa de transcrição. Os parâmetros p_j são constantes de translação, modificação de proteínas, da formação do complexo de proteína, e translocação entre o núcleo e o citoplasma. O parâmetro g_j são as constantes Michaelis-Menten. Os parâmetros q_j são constantes de taxas de fotoativação aguda de transcrição (P-dependente). a, b, c, d, e, f são os coeficientes de Hill. A unidade de tempo t é a hora.

L_{mod} : proteína modificada LHY/CCA1.

L_{tot} : proteína LHY/CCA1 total.

Proteína L_{tot} : Quantidade total da proteína LHY/CCA1.

Proteína T : TOC1 mRNA.

$P9, P7$ mRNA e proteínas PRR9 e PPR7 respectivamente.

$E4$: ELF4.

$E3$: mRNA para ELF3.

G : GI.

$E3_c$: Proteína citoplasmática de ELF3.

G_c : Proteína citoplasmática de GI.

$COP1_c$: Proteína citoplasmática de COP1.

$E3_n$: Proteína nuclear ELF3.

G_n : Proteína nuclear GI.

COP_d : Proteína nuclear COP1 de dia.

COP_n : Proteína nuclear COP1 de noite.

EG_c : Complexos proteicos citoplasmáticos ELF3-GI.

ZG : Complexos proteicos citoplasmáticos GI-ZTL.

EG_n : Complexos de proteínas nucleares ELF3-GI.

$E34$: Complexos de proteínas nucleares ELF3-ELF4.

EC: Complexos de proteínas ELF3-ELF4-LUX.

Gn_tot: Quantidade total de proteínas GI no núcleo.²

² Para mais detalhes ver ([POKHILKO et al., 2012](#)).

6 O pacote ABC-SysBio

Como uma das características dos métodos ABC é o fato deste não precisar da especificação da verossimilhança, esta é substituída na implementação por simulações. Por essa razão, estes métodos apresentam muitas dificuldades de tipo numérico e técnicos como encontrar soluções numéricas para sistemas de equações diferenciais ordinárias (ODE) rígidos ou para sistemas estocásticos descritos por equações estocásticas (SDE) ou por uma equação mestre. Para enfrentar estes problemas, é necessário implementar os métodos ABC de forma adaptativa para assim, encaminhar cada modelo específico com seus problemas bioquímicos e sua dinâmica (LIEPE et al., 2010).

ABC-SysBio é um pacote computacional implementado na linguagem Python, que executa em Linux e no sistema operacional Mac OS X. Foi desenvolvido para resolver problemas com aplicações biológicas e realizar inferência de parâmetros e seleção de modelos para sistemas dinâmicos determinísticos ou estocásticos sob o contexto do ABC (LIEPE et al., 2010), e por conseguinte este é empregado para encontrar uma estimativa da distribuição posteriori verdadeira. Também pode ser usado para analisar e simular modelos sbml (modelos escritos em formato de troca de modelos SBML padrão), assim como aqueles definidos pelo usuário. Permite o uso de pacotes para solução de ODEs e SDEs como são os algoritmos Gillespie, Euler-Marujama e LSODA (LIEPE et al., 2014). Outra aplicação do ABC-SMC baseado em ABC-SysBio surge na biologia sintética, onde identificam-se redes de reação molecular que possuem alta probabilidade de cumprir um conjunto dado de objetivos. (LIEPE et al., 2014)

Para entender como funciona o pacote ABC-SysBio, define-se previamente os seguintes conceitos,

Definição 6.0.3 (Parâmetro). Um parâmetro é um vetor que contém um valor para o qual cada uma das taxas de reação é estimada.

Definição 6.0.4 (Partícula). Vetor de parâmetros.

Definição 6.0.5 (População). Conjunto de partículas.

Definição 6.0.6 (Tamanho da população). Número de partículas em uma população.

Definição 6.0.7 (Distância Euclideana). A distância entre os pontos x e y em um espaço euclidiano $\mathbb{R}^n$ está dada por

$$d = |x - y| = \sqrt{\sum_{i=1}^n |x_i - y_i|^2}.$$

O método utilizado no ABC-SysBio tem as seguintes propriedades,

- A extensão de uma partícula é igual ao número de parâmetros a serem estimados.
- Cada população t tem associado uma tolerância ϵ_t . As populações estão construídas tal que as partículas que formam a população produzem os dados simulados que diferem dos dados observados por no máximo uma tolerância ϵ dada.
- $\{\epsilon_t\}$, $t = 1, 2, \dots, T$ é uma sucessão decrescente, com ϵ_T tendendo a zero.
- O peso associado com a partícula é proporcional à probabilidade de que o vetor de parâmetros possa explicar os dados observados.

Para encontrar uma aproximação à distribuição posteriori, o algoritmo em ABC-SysBio tem implementada a função distância euclideana, mas o usuário pode implementar outras e também definir a distribuição da priori. Deve prover os kernels de perturbação que podem ser a uniforme, a Gaussiana, a lognormal, porém outras distribuições podem ser utilizadas facilmente. O conjunto de dados fornecido não necessariamente precisa dados para todas as espécies definidas no modelo, mas pode ser um subconjunto ou uma combinação delas (LIEPE et al., 2010).

Ao contrário à teoria exposta anteriormente, comparam-se as estatísticas resumo dos dados experimentais D com os dos dados simulados D^* , sendo que a dimensionalidade de D é muito alta ou vêm de uma estrutura muito complexa, como por exemplo uma rede, se produz perda de informação. Por este motivo, o ABC-SysBio faz comparações diretas entre os dados observados e os dados simulados (LIEPE et al., 2014).

Por meio de sequenciamentos, ABC-SysBio realiza inferências e seleção de modelos, para cada iteração do algoritmo, se encontra um conjunto de vetores de parâmetros.

Considerações

- Precisa-se um número fixo de partículas no ABC-SysBio para cada população. A velocidade de execução depende da escolha do kernel e da sucessão ϵ_t , $t = 1, \dots, T$.
- O tempo requerido para a execução do algoritmo é dependente do hardware do sistema. Também é dependente do tamanho do modelo, a complexidade dos dados, a dimensão do espaço dos parâmetros, o número de partículas, o kernel de perturbação, etc.
- Os seguintes pacotes são requeridos para ABC-SysBio: Numpy, Scipy, Matplotlib, Swig, libSBML e cuda-sim.

O procedimento descrito por (LIEPE et al., 2014) para o pacote ABC-SysBio está representado na [Figura 12](#).

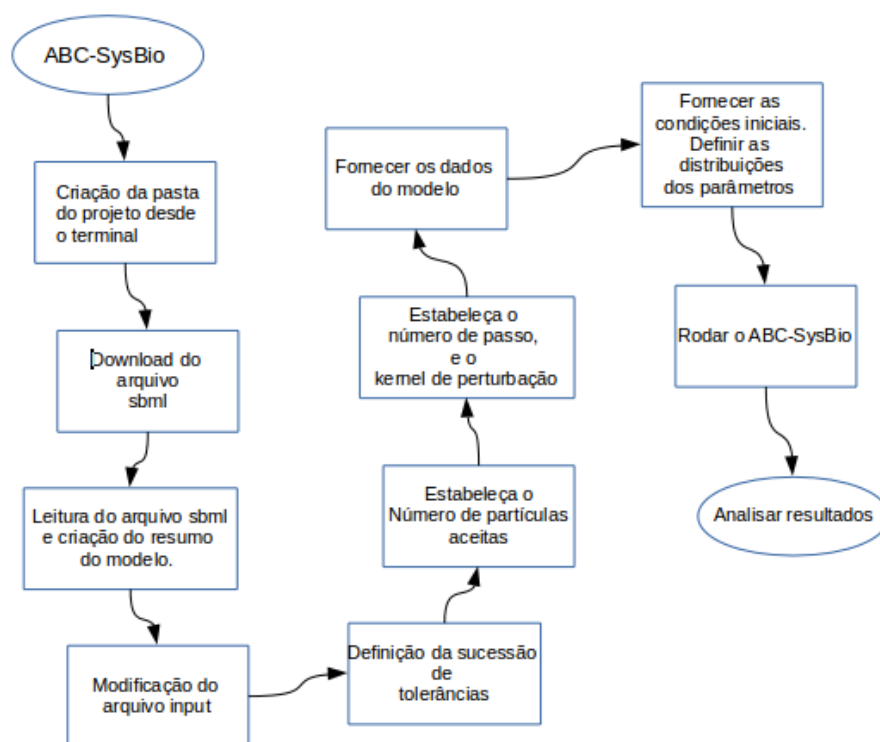


Figura 12 – Procedimento para o uso de ABC-SysBio.

Os resultados do programa ABC-SysBio são organizados em pastas, como se mostra nas figuras [Figura 13](#), [Figura 14](#), [Figura 15](#) e nos gráficos como das figuras [Figura 16](#) e [Figura 17](#).

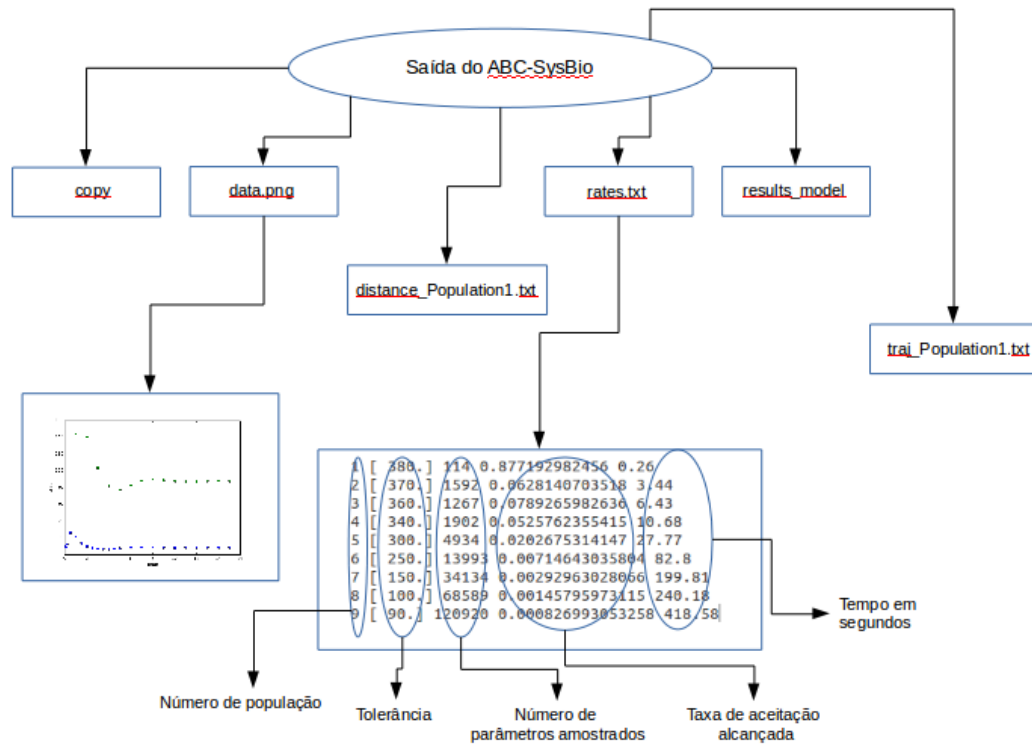


Figura 13 – Alguns gráficos de saída do ABC-SysBio. Taxas e gráfico dos dados. O gráfico data.png e o arquivo rates.txt da saída do ABC-SysBio.

Na figura [Figura 13](#), o arquivo rates.txt mostra dados em 5 colunas separadas por um espaço simples correspondentes a: número de população, valor da tolerância ϵ , número de combinações de parâmetros amostrados por população para conseguir uma população completa por ABC-SMC, a taxa de aceitação alcançada (i.e., a fração de simulações que foram necessárias para gerar os dados simulados que estão a uma distância determinada dos dados observados), e o tempo de geração da população em segundos, respectivamente. E o arquivo_data.png é um gráfico dos dados providenciados no arquivo input.

O pacote ABC-SysBio gera, como na figura [Figura 14](#), um arquivo distance_PopulationN.txt para cada população onde a informação está organizada por colunas,

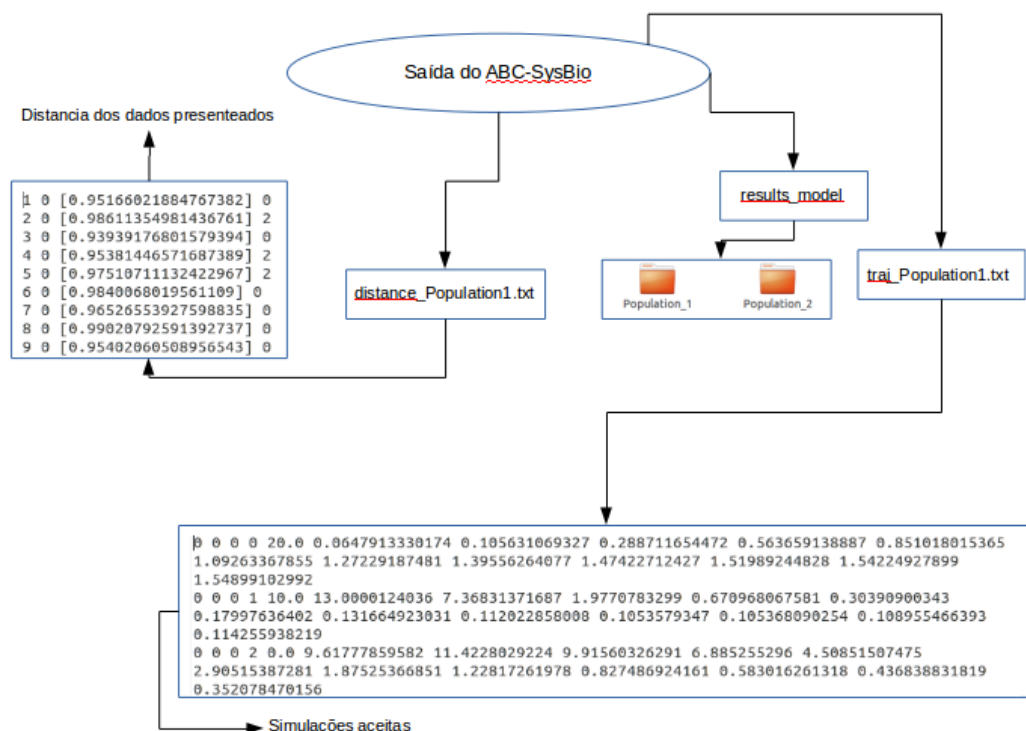


Figura 14 – Alguns arquivos de saída do ABC-SysBio. Distâncias e trajetórias. Os arquivos distance_Population1.txt, traj_Population1.txt e a pasta results_model da saída do ABC-SysBio.

sendo que a primeira corresponde ao número de partícula aceita, a segunda ao número de simulação de cada conjunto de parâmetros (especificado no tag <beta>) e a terceira coluna mostra a distância da partícula com os dados fornecidos. Existe também um arquivo traj_PopulationN.txt para cada população onde está especificado por colunas a condição inicial de cada partícula e sua trajetória por população. A pasta results_model contém uma pasta para cada população gerada e cada uma destas contém o arquivo data_PopulationN.txt e data_WeightsN.txt na figura [Figura 15](#); Scatter-Plots_PopulationN.png, Timeseries_PopulationN.png e weightedHistograms_PopulationN.png na figura [Figura 16](#).

Na [Figura 15](#), as pastas data_Weights1.txt e data_Population1.txt exibem os pesos estatísticos correspondentes as combinações dos parâmetros e as condições iniciais dos parâmetros, respectivamente.

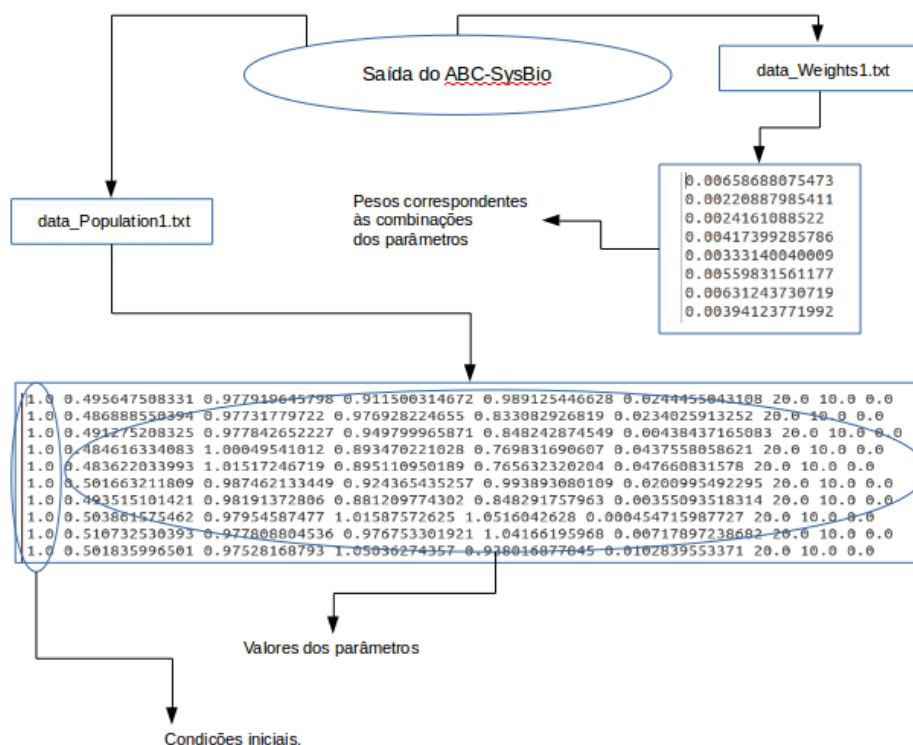


Figura 15 – Alguns arquivos de saída do ABC-SysBio. Pesos e dados populacionais. As pastas data_Weights1.txt e data_Population1.txt da saída do ABC-SysBio.

O gráfico `weightedHistograms_PopulationN.png` na Figura 16 mostra para cada população a dispersão por pares de parâmetros aceitos. Cada cor representa a informação para uma população diferente. Logo da primeira população nenhum parâmetro é inferido e as trajetórias não ajustam os dados, entretanto todos os parâmetros são inferidos por populações e todas as trajetórias concordam com os dados. Na diagonal está representada as distribuições posteriores da última população, as quais também encontram-se no gráfico `ScatterPlots_Population6.png`.

Depois de utilizar o pacote ABC-SysBio, obtém-se um conjunto de partículas com peso que resume a distribuição posteriori aproximada.

A distribuição marginal posteriori (i.e. a distribuição de probabilidade de cada taxa de reação considerada de forma independente) se obtém usando o histograma de pesos. O pacote ABC-SysBio providencia um arquivo de pesos para cada passo do

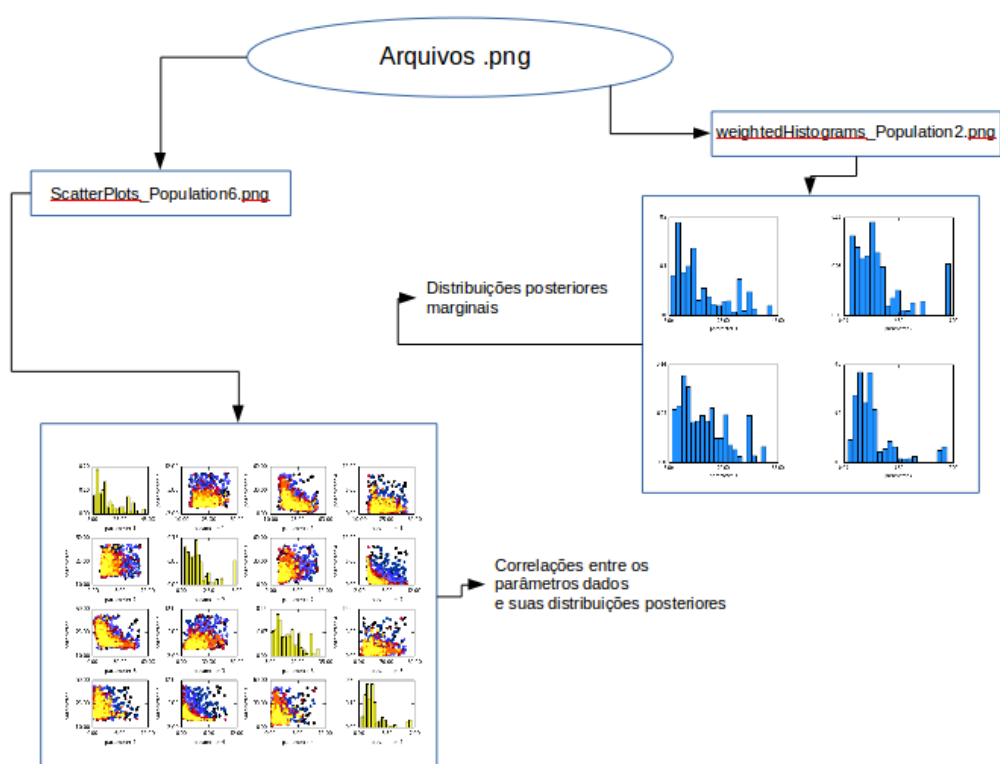
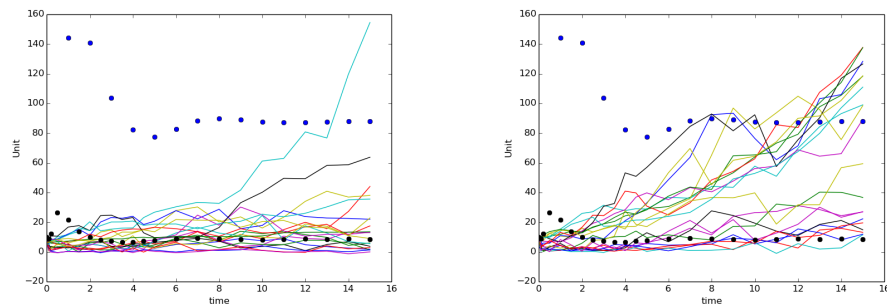


Figura 16 – Alguns gráficos de saída do ABC-SysBio. Histogramas. Os gráficos `weightedHistograms_Population2.png` e `ScatterPlots_Population6.png` da saída do ABC-SysBio.

algoritmo sequencial. Se a distribuição é muito pontiaguda em torno de um valor de um parâmetro, então a taxa de reação está bem inferida.

Em muitos dos sistemas biológicos, somente poucas taxas podem ser inferidas tendo um conjunto de dados observados e assim diferentes vetores de parâmetros podem explicar os dados observados igualmente bem. Tais questões são evidentes e importantes para considerar quando observa-se a distribuição de probabilidade conjunta sobre todas as taxas de reação.

É importante lembrar que os métodos ABC somente fornecem uma aproximação da distribuição posteriori. O algoritmo tem sido testado para exemplos onde a probabilidade posterior verdadeira é conhecida. Comparar os dados simulados com os dados obtidos é uma forma de comprovar a qualidade da posteriori obtida. Se os dados simulados são muito semelhantes aos observados, não se pode garantir que a posteriori obtida é



(a) Série de tempo 1 de uma população intermediária

(b) Série de tempo 2 da população final

Figura 17 – Alguns gráficos de saída. Séries de tempo. Gráficos de diagnóstico de uma população ABC-SMC intermediária e final. Estes permitem acompanhar o progresso da execução do algoritmo. Fonte: (LIEPE et al., 2014)

a verdadeira. Em particular, algumas regiões da distribuição posteriori podem não estar cobrindo que é devido a muito poucas partículas. Pelo anterior, (LIEPE et al., 2014) recomendam rodar o pacote várias vezes e comparar as distribuições obtidas.

A precisão da aproximação obtida da distribuição posteriori depende do valor de ϵ , do número de partículas por população, da forma da sucessão das tolerâncias, da função distância e dos kernels de perturbação.

Na atualidade, alguns dos aspectos computacionais de ABC estão em estudo e ABC-SysBio continuará incorporando esses desenvolvimentos que têm relação com a utilização de arquiteturas modernas computacionais ou com a programação em linguagens de baixo nível e com o desenvolvimento de simulação de sistemas dinâmicos estocásticos de forma efetiva. Para ABC-SysBio os autores em (LIEPE et al., 2014) recomendam utilizar as implementações GPU (Graphics Processing Unit) ou na linguagem C em vez de utilizar rotinas em Python. Outro aspecto de aperfeiçoamento está relacionado com os fundamentos do ABC. ABC é considerado como um formalismo inferencial distinto e não somente como uma aproximação à inferência Bayesiana convencional. ABC permite que os problemas podem ser abordados desde uma perspectiva Bayesiana e ABC-SysBio leva estas formulações ao alcance dos biólogos computacionais e de sistemas.

7 Simulações

Tendo descrito a rede regulatória do ciclo circadiano da *A. thaliana* (POKHILKO et al., 2012), trabalhou-se com o sistema de equações diferenciais que está apresentado como em (POKHILKO et al., 2012) e foi descrito no capítulo 4. Este sistema está disponível nos arquivos 97Model.m e 98mod_P2011.m no site PlaSMo (Plants Systems Modelling Repository) (<http://www.plasmo.ed.ac.uk/>)¹ e o qual foi resolvido utilizando o pacote MATLAB R2012a com o solver² ode23s (The MathWorks UK, Cambridge), nos tempos e com valores de parâmetros especificados no Anexo B, em um sistema 12 horas luz, 12 horas escuridão (12L:12D).

Também foi utilizado o arquivo BioMD0000000412.xml versão SBML L2 V4 curated, do site The BioModels database (NOVÈRE et al., 2006) <http://www.ebi.ac.uk/>. As simulações foram realizadas em um notebook TOSHIBA®Satellite S55 Series Intel®CORE™i7-4700MQ, CPU 2.46GHz x 8, com 7.7 GiB de memória, 633.0 GB de disco, Ostype 64B e sistema operacional Ubuntu 14.04.

Seguindo com os passos descritos na Figura 12:

1. Criou-se a pasta do projeto desde o terminal.
2. Fez-se a leitura do arquivo SBML BioMD0000000412.xml, para a criação automática do resumo do modelo e do arquivo *template* para ser utilizado como arquivo input das simulações.
3. Foi feita a escolha do número de partículas:
O número de partículas a amostrar foi escolhido em 1000, para obter mais informação sobre os parâmetros a estimar. `<particles> 1000 </particles>`, seguindo as observações dadas em (TURNER; ZANDT, 2012), mas também experimentou-se com uma população de 500 partículas.

¹ Acessado em 04/04/2015.

² Solver: Neste contexto, pacote desenhado para resolver equações diferenciais pelo MATLAB

4. Foi escolhida a tolerância:

Para as tolerâncias experimentou-se com um conjunto fornecido manualmente, assim como o conjunto fornecido automaticamente utilizando os quantis, com valores para $\alpha = 0.25$ e de $\alpha = 0.75$ dando ênfases neste último dado às recomendações em (FILIPPI et al., 2012) e em (SILK; FILIPPI; STUMPF, 2012).

5. Foi selecionado o kernel:

Baseados nos resultados obtidos por experimentação em (FILIPPI et al., 2012), escolheu-se o kernel de perturbação normal para otimizar o tempo de inferência dos parâmetros pelo ABC-SMC, porque está implementado no pacote e por ser mais eficiente que a distribuição uniforme. Assim, esta etiqueta fica `<kernel> normal </kernel>`.

6. Foram ajustados o tempo, as variáveis, as condições iniciais:

Para estabelecer os valores nas etiquetas `<time>` e `<variables><var1> </var1>, ..., <var28> </var28></variables>`, utilizam-se 500 medições das variáveis ao longo do tempo geradas ao executar os arquivos 97Model.m e 98mod_P2011.m com o fim de evidenciar o carácter oscilatório na rede e evitar perda da informação, isto com a ajuda das condições iniciais dadas no arquivo input. Também utilizaram-se 22 medições das variáveis ao longo do tempo ao executar os arquivos 97Model.m e 98mod_P2011.m para evidenciar o carácter sequencial do método ABC-SysBio. Os valores fornecidos pelo arquivo BIOMD0000000412 sobre as condições iniciais `<initial><ic1> </ic1>, ... , <ic28> </ic28></initial>` correspondem às espécies explicadas no Anexo A.

Neste trabalho só vamos considerar um modelo para ajuste de parâmetros assim que as seguintes etiquetas ficam assim: `<name> model1 </name>`, `<source> BIOMD0000000412.xml </source>`. A etiqueta `<fit>` corresponde a cada espécie utilizada, neste caso foram 28 e nenhum deles foi considerado uma combinação de outros mesmos dadas as condições do modelo em (POKHILKO et al., 2012), assim queda `<fit> species1 ... species28 </fit>`. O método utilizado para simular o modelo foi estabelecido assim: `<type> SDE </type>`

7. Foi realizado o ajuste dos parâmetros:

No arquivo input foi gerado automaticamente a lista de parâmetros com valores

constantes conforme o Anexo B. Os parâmetros de interesse neste trabalho são aqueles de CCA1/LHY, que segundo a [Equação 5.1](#), [5.5](#), [Equação 5.20](#) e [Equação 5.1](#) são: q_1 , n_1 , g_1 , g_3 , m_1 , m_2 , m_3 , m_4 , p_1 , p_2 , p_3 , a , c . Destes parâmetros, q_1 , n_1 , g_1 , g_3 , p_3 , c são os menos confiáveis da proteína CCA1/LHY, devido a que não foram medidos diretamente ou estimados de dados experimentais segundo (POKHILKO et al., 2012). Sendo assim, o trabalho foi realizado estimando primeiro dois parâmetros, e adicionou-se de a um parâmetro por vez para garantir o melhor ajuste. Assim, já tendo alguma referência dos valores (Anexo B), foi escolhida a distribuição normal para cada parâmetro com média e variância perto do valor referencial.

Parte II

Resultados

8 Resultados

Embora foram realizadas muitas simulações para o desenvolvimento deste trabalho, resume-se em duas formas de simulação. Na primeira ([Capítulo 9](#)), considerou-se 500 medições das variáveis ao longo do tempo para evidenciar a forma oscilante de cada variável. Neste contexto, o pacote ABC-SysBio aplicou o método de Rejeição amostral produzindo somente uma população na maioria dos casos e devido a que a exigência computacional é muito alta pela quantidade de dados a trabalhar, o método apresentou muitas vezes problemas de convergência sendo que convergia uma vez por cada quatro tentativas em média. E com o método do quantil, o algoritmo utilizou um valor para a tolerância de $\epsilon_1 = 10^{10}$ que é muito alto, o que faz que praticamente todas as partículas sejam aceitas na primeira população. Os resultados são explicados no [Capítulo 9](#).

Para a segunda simulação, cujos resultados são explicados no [Capítulo 10](#), consideraram-se 22 medições das variáveis ao longo do tempo para evidenciar o carácter sequencial do método. Este valor foi escolhido dado que apresenta convergência do 100% das simulações realizadas e proporciona de visualmente o progresso das simulações. Desta maneira, o ABC-SysBio proporcionou um conjunto de tolerâncias para as quais foi criada uma população por tolerância.

Na ??) encontra-se resumido as duas simulações consideradas.

Tabela 1 – Sumário das simulações: Resumo das diferentes forma de realizar as duas simulações evidenciando o método empregado pelo ABC-SysBio nos resultados

	Simulações 1	Simulações 2
Medições ao longo do tempo	500	22
Amostragem de uma população de	1000 partículas	1000 partículas
Sucessão de tolerâncias	0.25-quantil	0.75-quantil
Prioris	normal	normal
Kernel de perturbação	normal	normal
Método de simulação	SDE	SDE
Primeiro ajuste	n1, q1	n1, q1
Segundo ajuste	n1, g1, q1	n1, g1, q1
Terceiro ajuste	n1, g1, g3, q1	n1, g1, g3, q1
Quarto ajuste	n1, g1, g3, p1, q1	n1, g1, g3, q1, p1
Método realizado pelo ABC-SysBio	ABC Rejeição amostral	ABC-SMC

Fonte: Produzido pelo autor.

A medida que o número de parâmetros aumenta, os ajustes foram melhores dado que estes parâmetros estão relacionados como se viu na conformação da rede ([seção 2.2](#)).

9 Simulações 1

Inicialmente as simulações começaram ajustando o vetor de parâmetros $\theta=(n1,q1)$ os quais não evidenciam correlação como pode ser observado na [Figura 18a](#) tendo uma distribuição aproximada à posteriori normal $n(2, 4)$ e $n(0.2, 2.8)$ respectivamente. Ao analisar o arquivo `params_last.dat` gerado pela simulação para os últimos valores dos parâmetros $n1$ e $q1$ tem-se que o valor médio para $n1$ é $\mu_{n1} = 1.8121$ com uma variância $\sigma_{n1}^2 = 1.2950$ calculado para 30 valores de $n1$ estimado, entretanto a mediana é de 1.7527. O que indica que os dados não contém uma quantidade significativa de dados atípicos em torno da média, o que significa que este valor estimado para $n1$ pode ser considerado uma boa estimativa. Para o parâmetro $q1$ se tem que o valor médio é $\mu_{q1} = 1.8020$ com uma variância $\sigma_{q1} = 1.4898$ e uma mediana de 1.22, indicando como no caso anterior que os dados não estão muito espalhados e também pode ser considerado uma boa aproximação ([Figura 19](#)).

Para o vetor de parâmetros $\theta^*=(n1,g1,q1)$. Os gráficos indicam que a distribuição aproximada posteriori é normal. Assim, $n1 \sim n(2.5, 1.36)$, $g1 \sim n(0, 1.8)$ e $q1 \sim n(1.5, 1.8)$. O gráfico da distribuição posteriori ([Figura 20a](#)) indica que os dados estão simetricamente distribuídos em torno das médias o que dá lugar a que um valor aproximado para $n1$ de 2.5, para $g1$ de 0.0 e para $q1$ de 1.0. Os dados simulados podem ser considerados uma boa aproximação para os parâmetros.

Para o vetor de parâmetros $(n1,g1,g3,q1)$. Os parâmetros têm uma distribuição normal aproximada à posteriori sem evidencia de correlação entre os dados. $n1 \sim n(2.5, 1.36)$, $g1 \sim n(0.0, 1.7)$, $g3 \sim n(0.0, 7.0)$ e $q1 \sim n(1.0, 1.7)$. Neste caso, os parâmetros $n1$ e $g1$ e $q1$ com comportamento semelhante ao caso anterior ([Figura 21](#)). Com estes resultados, o valor para $g3$ não é um valor confiável.

Para o vetor de parâmetros $(n1,g1,g3,p1,q1)$. Os parâmetros têm uma distribuição normal aproximada à posteriori sem evidencia de correlação entre os dados. $n1 \sim n(2.5, 5.29)$, $g1 \sim n(0.5, 1.8)$, $g3 \sim n(0.0, 1.8)$, $p1 \sim n(0.0, 1)$ e $q1 \sim n(1.0, 2.8)$. Neste caso, os parâmetros $n1$ e $g1$ e $q1$ com comportamento semelhante ao caso anterior ([Figura 22](#)). Com estes resultados, o valor para $g3$ não é um valor confiável. O resumo

dos resultados apresentam-se na [Tabela 2](#).

Tabela 2 – Aproximações para os valores dos parâmetros

	Parâmetro				
	n1	g1	g3	p1	q1
Valores observados	2.6	0.1	0.6	0.13	1.2
Valores simulados aproximados	2.0				1.0
	2.5	0.0			1.0
	2.5	0.0	0.0		1.0
	2.5	0.5	0.0	0.0	1.0

Fonte: Produzido pelo autor.

Os dados na [Tabela 2](#), foram obtidos com as seguintes configurações no arquivo input:

- Para dois parâmetros: $n1 \sim n(2, 5)$ e $q1 \sim n(1, 5)$ com $\alpha = 0.25$.
- Para três parâmetros: $n1 \sim n(2, 1)$, $g1 \sim n(0, 1)$ e $q1 \sim n(1, 1)$ com $\alpha = 0.25$.
- Para quatro parâmetros: $n1 \sim n(2, 1)$, $g1 \sim n(0, 1)$, $g3 \sim n(0, 5)$ e $q1 \sim n(1, 1)$ com $\alpha = 0.25$.
- Para cinco parâmetros: $n1 \sim n(2.5, 1.5)$, $g1 \sim n(0, 1)$, $g3 \sim n(0.5, 1.5)$, $p1 \sim n(0, 1)$ e $q1 \sim n(1, 1)$ com $\alpha = 0.25$

Tanto nas simulações para dois, três, quatro e cinco parâmetros a distância modal entre os dados observados e os dados simulados foi de 65, com uma taxa de aceitação de 0.99. O tempo médio de simulação foi de 37 segundos. Embora a tolerância muito alta, os resultados de ajustar os cinco parâmetros $n1$, $g1$, $g3$, $p1$ e $q1$ podem considerar-se um bom ajuste dos dados ao modelo.

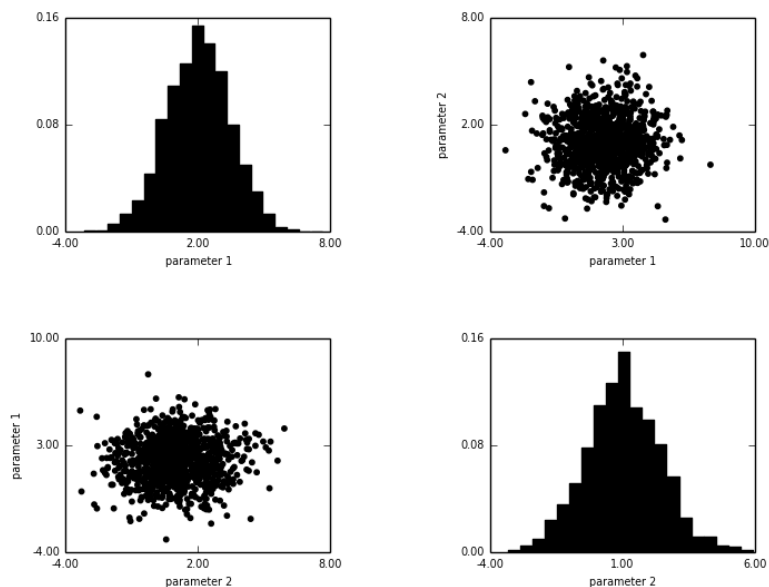
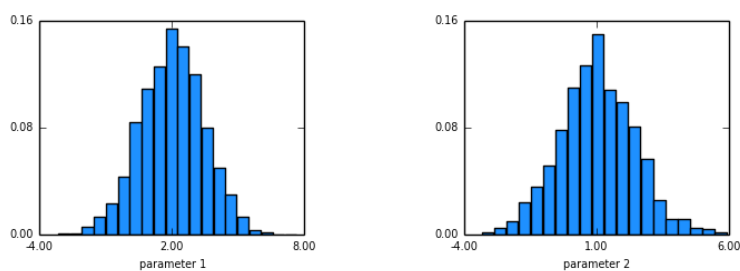
(a) ($n1, q1$)(b) ($n1, q1$)

Figura 18 – Gráfico de dispersão e da distribuição posteriori para os parâmetros $n1, q1$. (a) O parâmetro 1 é $n1$ e o parâmetro 2 corresponde ao $q1$. Os dois parâmetros não evidenciam correlação. A maioria dos dados estão localizados na parte central do diagrama de dispersão por pares de partículas. (b) Se mostram as distribuições posteriori normais para parâmetro 1 ($n1$) e parâmetro 2 ($q1$).

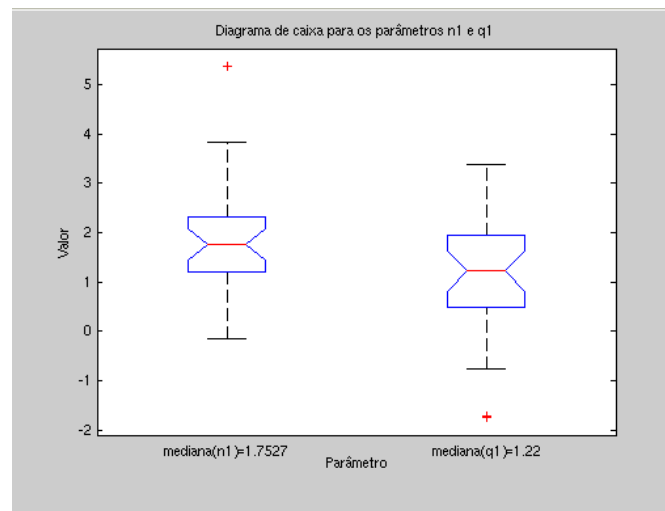


Figura 19 – Diagrama de caixas para $n1$, $q1$. Os dados estão distribuídos simetricamente.

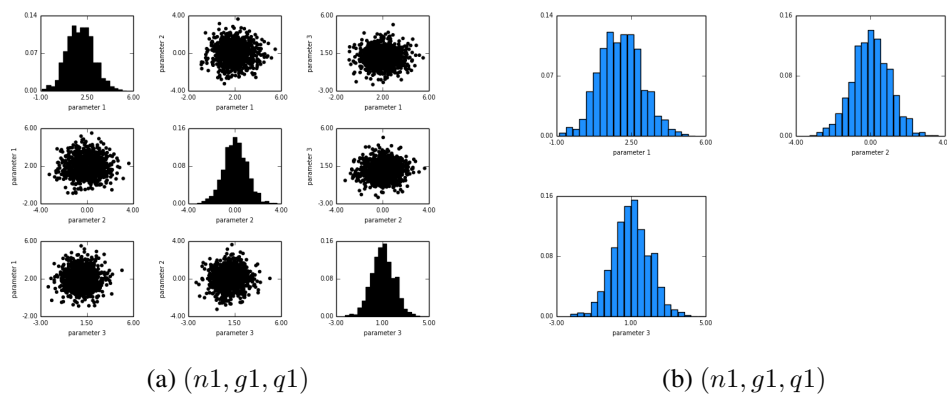


Figura 20 – Gráfico de dispersão e da distribuição posteriori para os parâmetros $n1$, $g1$, $q1$. (a) O parâmetro 1 ($n1$), parâmetro 2 ($g1$) e parâmetro 3 ($q1$). Os parâmetros não evidenciam correlação por pares. A maioria dos dados estão localizados na parte central do diagrama de dispersão. (b) $n1$, $g1$ e $q1$ se distribuem normalmente.

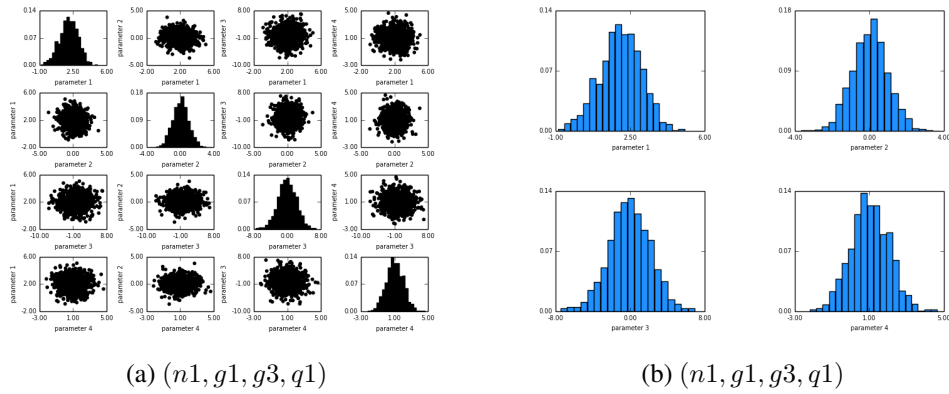


Figura 21 – Gráfico de dispersão e da distribuição posteriori para os parâmetros $n1$, $g1$, $g3$, $q1$. (a) O parâmetro 1 ($n1$), parâmetro 2 ($g1$), parâmetro 3 ($g3$) e parâmetro 4 ($q1$). Os parâmetros não evidenciam correlação por pares. A maioria dos dados estão localizados na parte central do diagrama de dispersão. (b) $n1$, $g1$, $g3$ e $q1$ se distribuem normalmente.

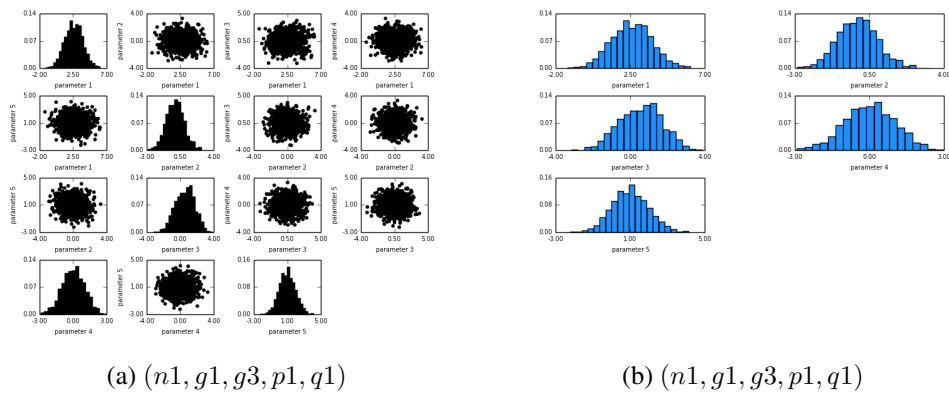


Figura 22 – Gráfico de dispersão e da distribuição posteriori para os parâmetros $n1$, $g1$, $g3$, $p1$, $q1$. (a) O parâmetro 1 ($n1$), parâmetro 2 ($g1$), parâmetro 3 ($g3$), parâmetro 4 ($q1$) e parâmetro 5 ($p1$). Os parâmetros não evidenciam correlação por pares. A maioria dos dados estão localizados na parte central do diagrama de dispersão. (b) $n1$, $g1$, $g3$, $p1$ e $q1$ se distribuem normalmente.

10 Simulações 2

Uma segunda forma para realizar as simulações, consistiu em considerar 22 medições das variáveis ao longo do tempo com a finalidade de procurar evidenciar os sequenciamentos dados pela sucessão de tolerâncias, assim como possibilitar a visualização das séries temporais nos gráficos de saída.

Em todos os casos as simulações convergiram e produziram-se mais de duas populações. Os tempos de simulação foram superiores à primeira forma de simulação.

Nas figuras, se apresentam alguns dos gráficos do processo de simulação para evidenciar seu comportamento. De maneira geral, em cada nova população foram obtidos umas distribuições com os dados mais concentrados em torno à média fazendo que a forma das distribuições posterioris sejam cada vez mais pontiagudas.

Foram gerados dois resultados para a estimação de dois parâmetros mudando as distribuições prioris dos parâmetros, no primeiro, o pacote gerou 12 populações correspondentes ao conjunto de tolerâncias $\{10^{10}, 9.2681, 8.804695, 8.5265, 8.1076, 7.5941, 7.1215, 6.6316, 6.2488, 5.9784, 5.7126, 5.5181\}$. Estabeleceram-se para o arquivo input os parâmetros $n1 \sim n(2, 5)$ e $q1 \sim n(1, 5)$ e $\alpha = 0.75$, alcançando uma taxa mínima de aceitação de 0.02 com 485778 parâmetros avaliados. Ao analisar o arquivo params_last.dat gerado pela simulação para os últimos valores dos parâmetros n1 e q1, pode-se dizer que esta aproximação aos parâmetros não é aceitável dado que para o parâmetro n1 o valor médio é $\mu_{n1} = 1.0028$ com variância $\sigma_{n1}^2 = 15.61$ e mediana 1.1474 que é muito alta. De forma semelhante tem-se que para o parâmetro q1 a média é $\mu_{q1} = -0.16$ com variância $\sigma_{q1}^2 = 8.99$ e mediana -0.498 fazendo que esta estimação seja rejeitada (Figura 23a). Por outro lado, com a outra confi-

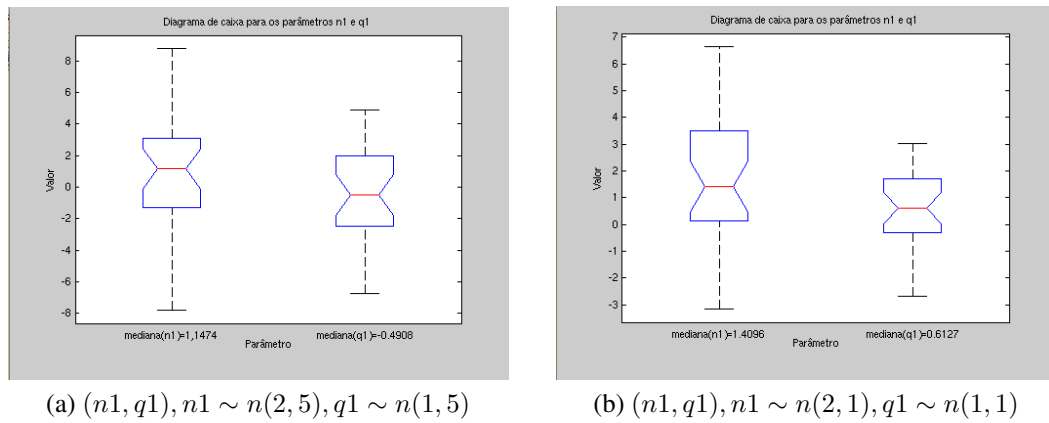


Figura 23 – (a) Diagrama de caixas para n1 e q1 da População 1. Os dados estão distribuídos simetricamente. Levando em consideração a mudança de escala nas gráficas, evidencia-se uma melhor aproximação.

guração, o pacote gerou 16 populações correspondentes ao conjunto de tolerâncias $\{10^{10}, 9.2681, 8.804695, 8.5153, 8.0801, 7.5049, 6.9672, 6.4582, 6.0904, 5.7828, 5.5855, 5.3812, 5.2382, 5.1027, 4.97995, 4.8853\}$. Estabeleceram-se para o arquivo input os parâmetros $n1 \sim n(2, 1)$ e $q1 \sim n(1, 1)$ e $\alpha = 0.75$, alcançando uma taxa mínima de aceitação de 0.007 com 135705 parâmetros avaliados. Esta aproximação para o parâmetro n1 o valor médio é $\mu_{n1} = 1.6852$ com variância $\sigma_{n1}^2 = 4.867$ e mediana 1.2909 que é alta. Enquanto que para o parâmetro q1 a média é $\mu_{q1} = 0.5953$ com variância $\sigma_{q1}^2 = 1.8676$

e mediana 0.6127 fazendo que a estimação não seja apropriada para o parâmetro $n1$ (Figura 23b).

A informação resume-se na seguinte Tabela 3: Para três parâmetros o melhor

Tabela 3 – Valores das médias e medianas dos parâmetros $n1$ e $q1$ do arquivo last_param.dat.

	$n1 \sim n(2, 5)$ e $q1 \sim n(1, 5)$			$n1 \sim n(2, 1)$ e $q1 \sim n(1, 1)$		
Parâmetro	μ	σ^2	Mediana	μ	σ^2	Mediana
$n1$	1.0028	15.61	1.1474	1.6852	4.867	1.2909
$q1$	-0.1653	8.9941	-0.498	0.5953	1.8676	0.6127

Fonte: Produzido pelo autor.

Nota: Note-se uma redução significativa no espalhamento dos dados ao tomar prioris normais com variância menor. Os As médias e as medianas são distantes dos valores observados.

ajuste correspondeu à seguinte configuração inicial: $n1 \sim n(2, 5)$, $g1 \sim n(0, 1)$, $q1 \sim n(1, 1)$ com $\alpha = 0.75$, onde o pacote gerou 16 populações correspondentes ao conjunto de tolerâncias $\{10^{10}, 9.2681, 8.804695, 8.52693, 8.1043, 7.6374, 7.1468, 6.6073, 6.2586, 5.9612, 5.7127, 5.5291, 5.3523, 5.1905, 5.0636, 4.9423\}$. O tempo de execução foi de aproximadamente 51 minutos, com uma taxa mínima de aceitação de 0.6% avaliando 143343 parâmetros (Figura 24, Figura 25).

Nos gráficos das distribuições posteriori todos parecem ser normais, as médias destas indicam uma aproximação aos valores dos parâmetros observados que podem ser aceitas (Figura 25). As séries de tempo nos gráficos na Figura 26 são de difícil interpretação e não ajudam no seguimento do processo realizado pelo algoritmo.

Para quatro parâmetros o ajuste encontrado corresponde à seguinte configuração inicial: $n1 \sim n(2, 1)$, $g1 \sim n(0, 1)$, $p1 \sim n(0, 5)$ e $q1 \sim n(1, 1)$ com $\alpha = 0.75$, onde o pacote gerou 12 populações correspondentes ao conjunto de tolerâncias $\{10^{10}, 9.2681, 8.804695, 8.5098, 8.0802, 7.5648, 6.57648, 6.2101, 5.9505, 5.7253, 5.5195\}$. O tempo de execução foi de aproximadamente 14 minutos, com uma taxa mínima de aceitação de 0.2% avaliando 43946 parâmetros (Figura 27, Figura 28).

Para cinco parâmetros o ajuste encontrado correspondeu seguinte configuração

inicial: $n1 \sim n(2.5, 1.5)$, $g1 \sim n(0, 1)$, $g3 \sim n(0.5, 1.5)$, $p1 \sim n(0, 1)$ e $q1 \sim n(1, 1)$ com $\alpha = 0.75$, onde o pacote gerou 10 populações correspondentes ao conjunto de tolerâncias $\{10^{10}, 9.2681, 8.804695, 8.5482, 7.315, 6.8913, 6.5062, 6.1983\}$. O tempo de execução foi de aproximadamente 13 minutos, com uma taxa mínima de aceitação de 0.3% avaliando 43946 parâmetros (Figura 29, Figura 30).

Os valores dos parâmetros estimados está resumido na Tabela 4

Tabela 4 – Aproximações para os valores dos parâmetros

	Parâmetro				
	n1	g1	g3	p1	q1
Valores observados	2.6	0.1	0.6	0.13	1.2
Valores simulados aproximados	2				1
	2.5	-0.5			1
	2	-0.5	1		1
	2.5	0.0	0.5	-0.5	1.5

Fonte: Produzido pelo autor.

Os dados na Tabela 4, foram obtidos com as seguintes configurações no arquivo input:

- Para dois parâmetros: $n1 \sim n(2, 5)$ e $q1 \sim n(1, 1)$ com $\alpha = 0.75$.
- Para três parâmetros: $n1 \sim n(2, 1)$, $g1 \sim n(0, 1)$ e $q1 \sim n(1, 1)$ com $\alpha = 0.75$.
- Para quatro parâmetros: $n1 \sim n(2, 1)$, $g1 \sim n(0, 1)$, $g3 \sim n(0, 5)$ e $q1 \sim n(1, 1)$ com $\alpha = 0.75$.
- Para cinco parâmetros: $n1 \sim n(2.5, 1.5)$, $g1 \sim n(0, 1)$, $g3 \sim n(0.5, 1.5)$, $p1 \sim n(0, 1)$ e $q1 \sim n(1, 1)$ com $\alpha = 0.75$

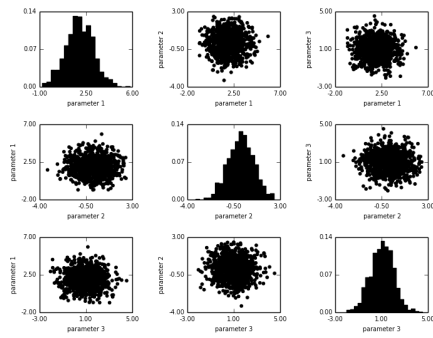
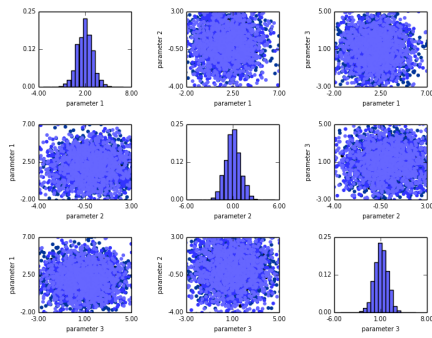
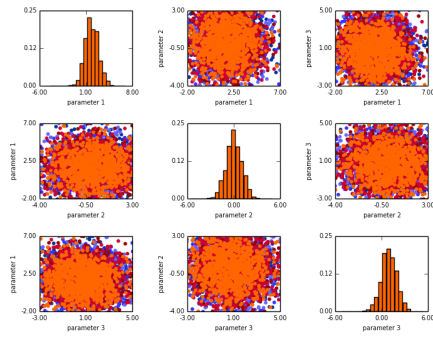
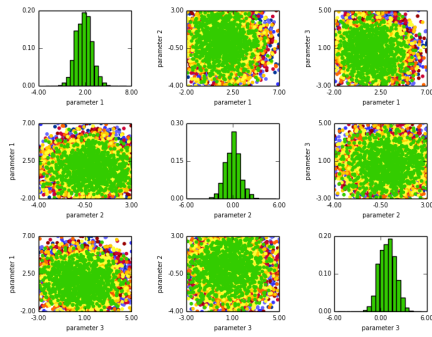
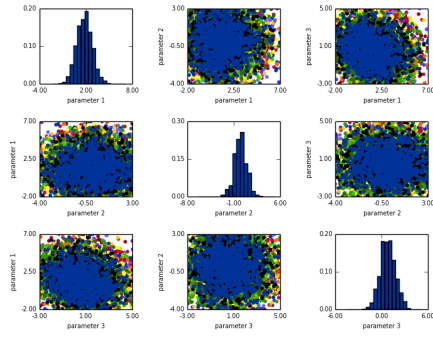
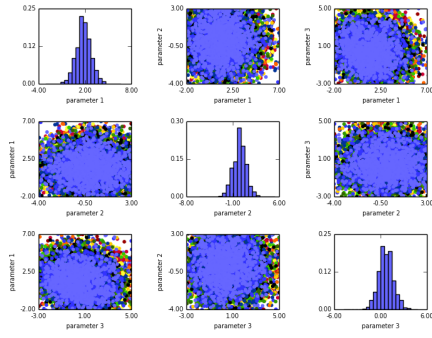
(a) $(n1, g1, q1)$ População 1(b) $(n1, g1, q1)$ População 4(c) $(n1, g1, q1)$ População 7(d) $(n1, g1, q1)$ População 10(e) $(n1, g1, q1)$ População 14(f) $(n1, g1, q1)$ População 16

Figura 24 – Gráficos de dispersão por pares e por populações para os parâmetros $n1$, $g1$, $q1$. (a) O parâmetro 1 ($n1$), parâmetro 2 ($g1$) e parâmetro 3 ($q1$). Os parâmetros não evidenciam correlação por pares. A maioria dos dados estão localizados na parte central do diagrama de dispersão. (b) $n1$, $g1$ e $q1$ se distribuem normalmente.

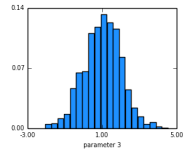
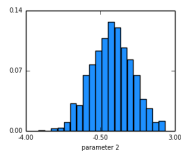
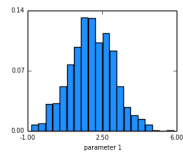
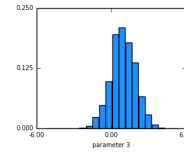
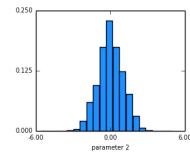
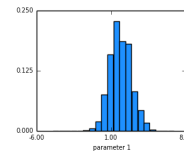
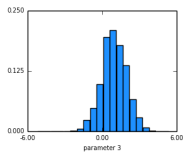
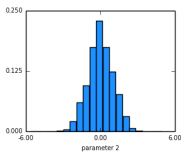
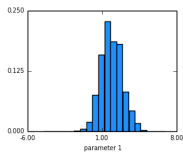
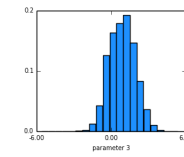
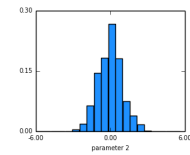
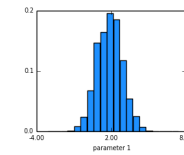
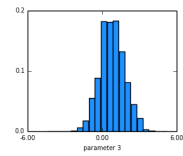
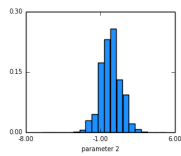
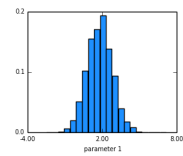
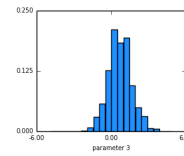
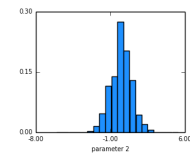
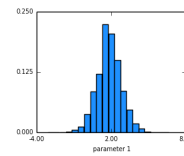
(a) $(n1, g1, q1)$ População 1(b) $(n1, g1, q1)$ População 4(c) $(n1, g1, q1)$ População 7(d) $(n1, g1, q1)$ População 10(e) $(n1, g1, q1)$ População 14(f) $(n1, g1, q1)$ População 16

Figura 25 – Distribuição posteriori por populações para os parâmetros $n1$, $g1$, $q1$. (a) O parâmetro 1 ($n1$), parâmetro 2 ($g1$) e parâmetro 3 ($q1$). $n1$, $g1$ e $q1$ se distribuem normalmente.

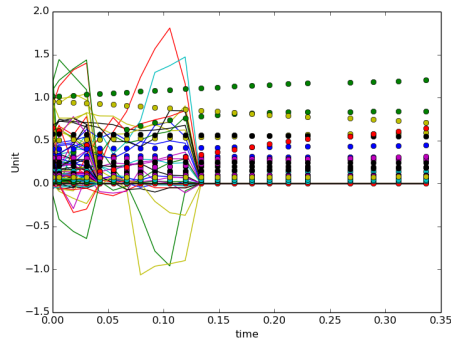
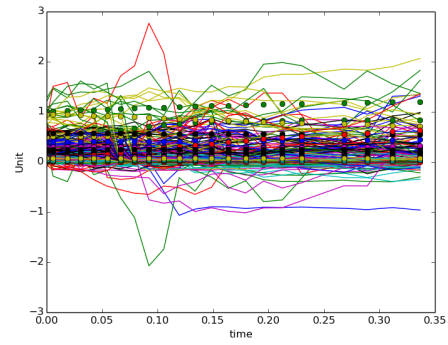
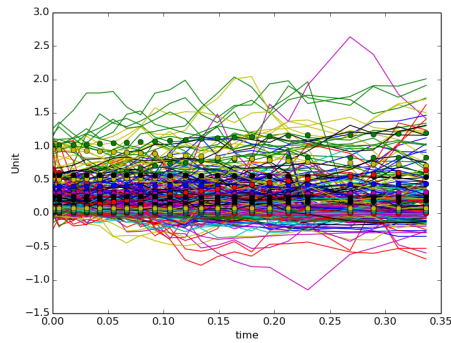
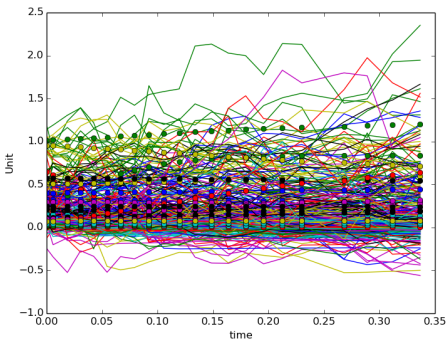
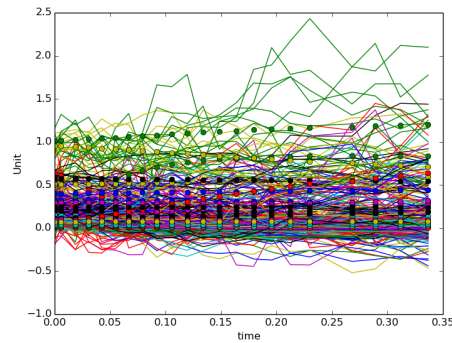
(a) $(n1, g1, q1)$ População 1(b) $(n1, g1, q1)$ População 4(c) $(n1, g1, q1)$ População 7(d) $(n1, g1, q1)$ População 10(e) $(n1, g1, q1)$ População 14

Figura 26 – Séries de tempo por populações para os parâmetros $n1$, $g1$, $q1$. Os pontos indicam as medições das variáveis no tempo, e as linhas mostram as trajetórias para as populações 1, 4, 7, 10 e 14. A Figura 26a mostra a trajetória para a primeira população onde o algoritmo somente convergiu nos primeiros dados, depois os gráficos não proporcionam informação que ajude a fazer o seguimento do algoritmo.

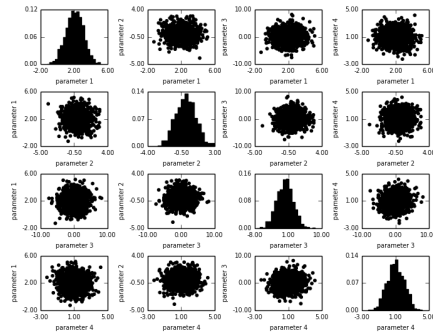
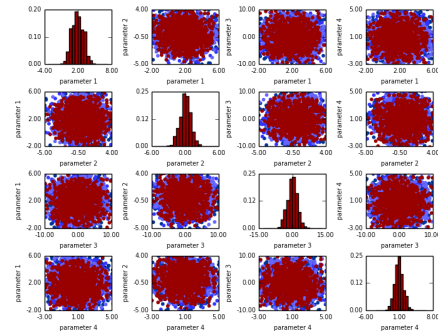
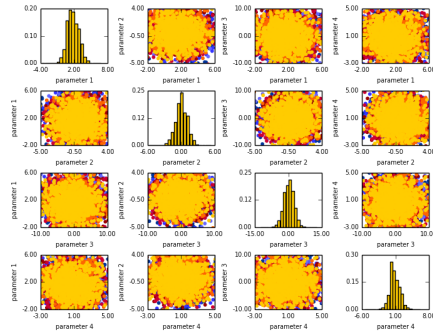
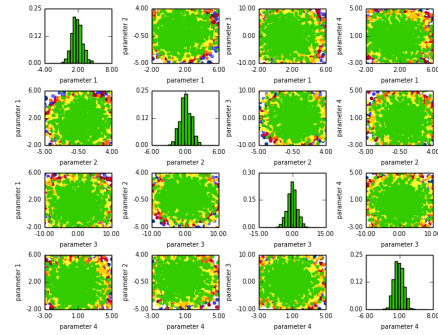
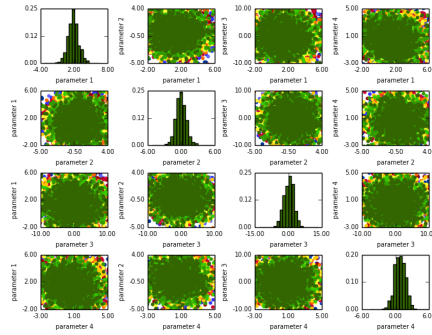
(a) $(n1, g1, g3, q1)$ População 1(b) $(n1, g1, g3, q1)$ População 5(c) $(n1, g1, g3, q1)$ População 8(d) $(n1, g1, g3, q1)$ População 10(e) $(n1, g1, g3, q1)$ População 12

Figura 27 – Gráficos de dispersão por pares e por populações para os parâmetros $n1$, $g1$, $g3$, $q1$. (a) O parâmetro 1 ($n1$), parâmetro 2 ($g1$), parâmetro 3 ($g3$) e parâmetro 4 ($q1$). Os parâmetros não evidenciam correlação por pares. A maioria dos dados estão localizados na parte central do diagrama de dispersão. (b) $n1$, $g1$, $g3$ e $q1$ se distribuem normalmente.

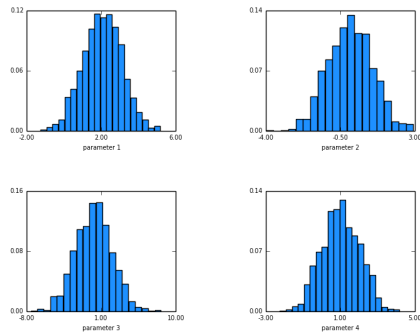
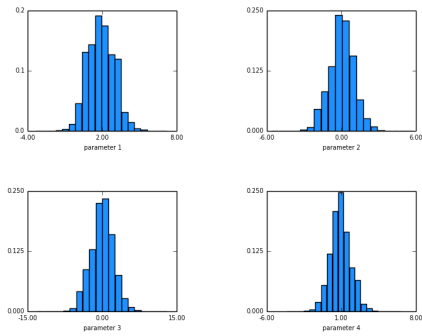
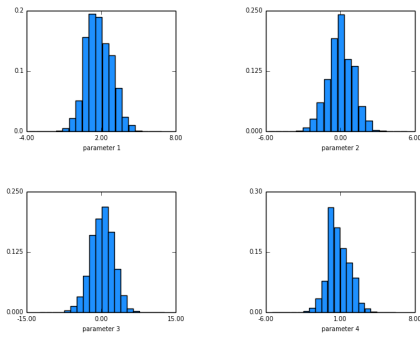
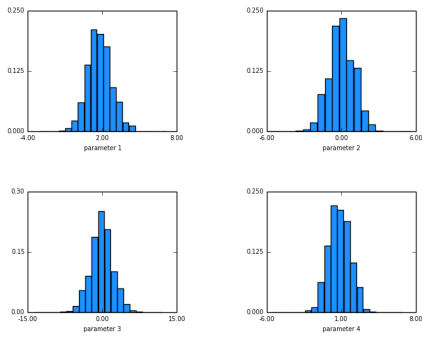
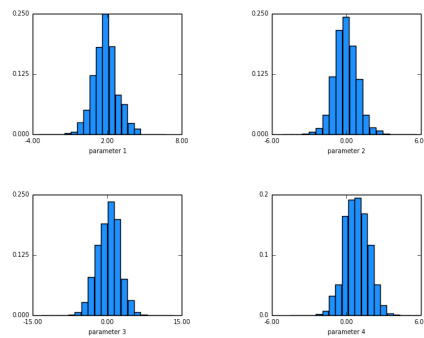
(a) $(n1, g1, g3, q1)$ População 1(b) $(n1, g1, g3, q1)$ População 5(c) $(n1, g1, g3, q1)$ População 8(d) $(n1, g1, g3, q1)$ População 10(e) $(n1, g1, g3, q1)$ População 12

Figura 28 – Distribuição Posteriori por populações para os parâmetros $n1$, $g1$, $g3$, $q1$. (a) O parâmetro 1 ($n1$), parâmetro 2 ($g1$), parâmetro 3 ($g3$) e parâmetro 4 ($q1$). $n1$, $g1$, $g3$ e $q1$ se distribuem normalmente.

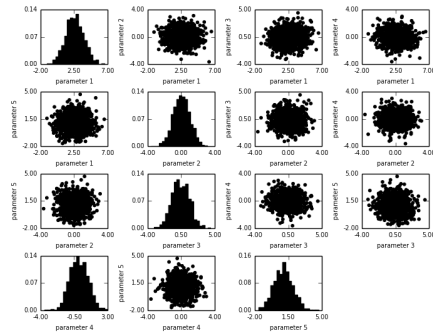
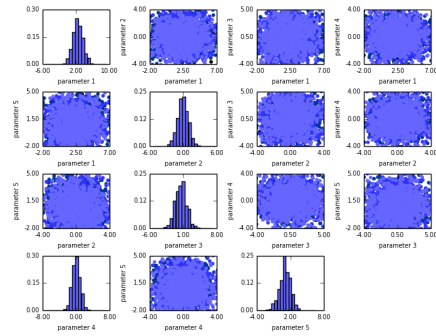
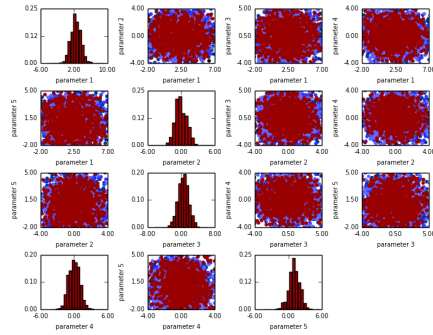
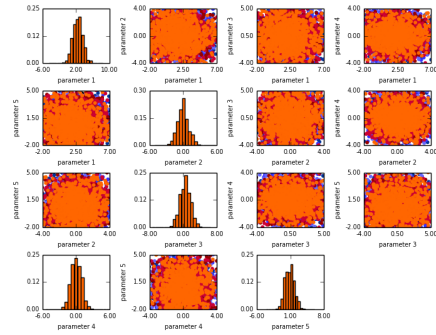
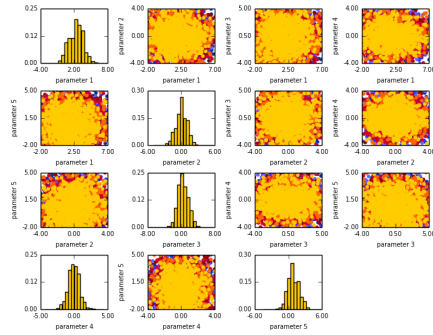
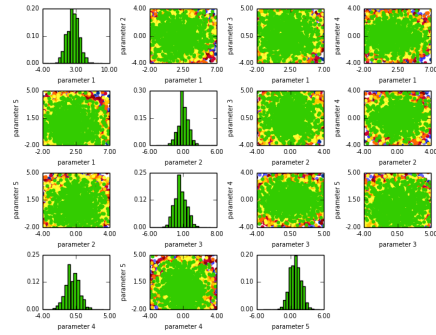
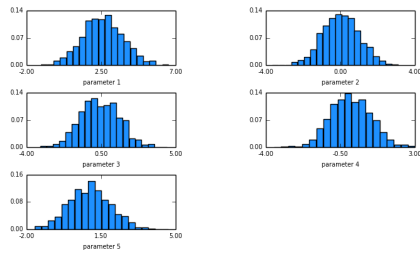
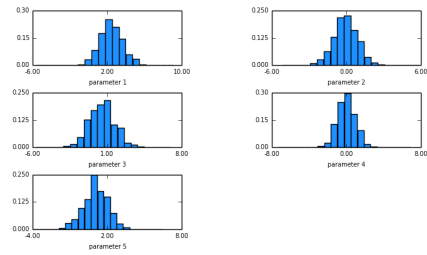
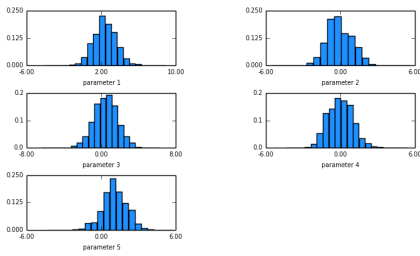
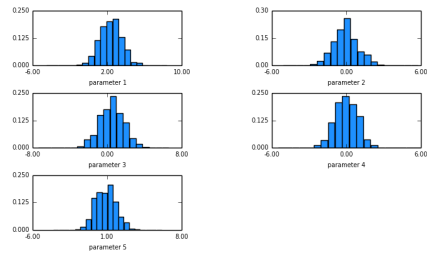
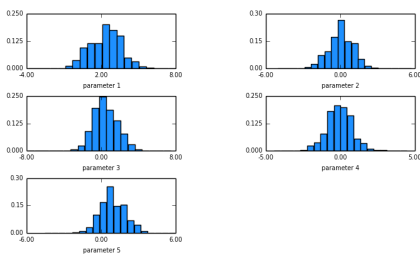
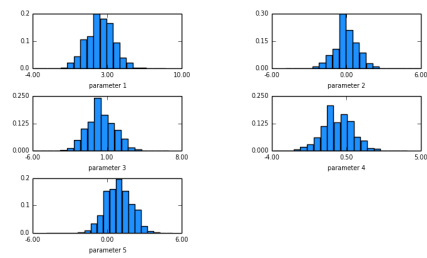
(a) $(n1, g1, g3, p1, q1)$ População 1(b) $(n1, g1, g3, p1, q1)$ População 4(c) $(n1, g1, g3, p1, q1)$ População 5(d) $(n1, g1, g3, p1, q1)$ População 7(e) $(n1, g1, g3, p1, q1)$ População 8(f) $(n1, g1, g3, p1, q1)$ População 10

Figura 29 – Gráficos de dispersão por pares e por populações para os parâmetros $n1$, $g1$, $g3$, $p1$, $q1$. (a) O parâmetro 1 ($n1$), parâmetro 2 ($g1$), parâmetro 3 ($g3$), parâmetro 4 ($p1$) e parâmetro 5 ($q1$). Os parâmetros não evidenciam correlação por pares. A maioria dos dados estão localizados na parte central do diagrama de dispersão. (b) $n1$, $g1$, $g3$, $p1$ e $q1$ se distribuem normalmente.

(a) $(n1, g1, g3, p1, q1)$ População 1(b) $(n1, g1, g3, p1, q1)$ População 4(c) $(n1, g1, g3, p1, q1)$ População 5(d) $(n1, g1, g3, p1, q1)$ População 7(e) $(n1, g1, g3, p1, q1)$ População 8(f) $(n1, g1, g3, p1, q1)$ População 10Figura 30 – Distribuição posteriori por populações para os parâmetros $n1, g1, g3, p1, q1$.

(a) O parâmetro 1 ($n1$), parâmetro 2 ($g1$), parâmetro 3 ($g3$), parâmetro 4 ($p1$) e parâmetro 5 ($q1$). $n1, g1, g3, p1$ e $q1$ se distribuem normalmente.

Parte III

Discussão

Discussão

Mesmo que o pacote forneça uma aplicação como o método ABC-SMC, o SysBio apresenta algumas limitações para necessidades reais, como neste caso de um sistema complexo conformado por 56 reações e 57 funções onde os parâmetros não são independentes dado a natureza do problema. Embora o pacote forneça visualização gráfica para seguir o comportamento das simulações, estes dificilmente geram informação clara do que está acontecendo. Assim, por exemplo, os gráficos correspondentes às séries de tempo contribuem com pouca informação devido à quantidade de dados manipulados.

Ainda que o pacote forneça como resultado em forma gráfica a distribuição posteriori dos dados simulados, que neste caso são normais como o esperado, indicando que estes possuem bom comportamento. Os gráficos gerados pelo pacote remetem a ideia de que o valor simulado do parâmetro está em um intervalo dado, entretanto não fornece um valor exato. O resumo dos dados estão na [Tabela 6](#), contudo os resultados podem garantir que os valores para os parâmetros que não eram supostamente confiáveis pelos próprios autores ([POKHILKO et al., 2012](#)), poderiam ser considerados devido a que os valores simulados estão perto dos observados e apresentam comportamento aproximadamente normal nas posterioris. Ao analisar o arquivo *param_last.dat* gerado pelas simulações para os últimos valores dos parâmetros $n1$, $g1$, $g3$, $p1$ e $q1$ tem-se os resultados resumidos na [Tabela 5](#), a qual apresenta os valores das médias, variâncias e medianas para 50 valores de parâmetros.

Evidencia-se nos diagramas de caixas ([Figura 31](#)), cujos valores estão resumidos na [Tabela 5](#), que com o método ABC rejeição amostral obtiveram aproximações semelhantes às dadas pelo ([POKHILKO et al., 2012](#)) com variâncias menores que quando o pacote utiliza ABC-SMC.

Os valores na [Tabela 6](#) foram estimados a partir dos gráficos da [Figura 22b](#), [Figura 30a](#) e [Figura 30f](#) em forma visual. Comparando com os resultados na [Tabela 5](#) pode-se dizer que independentemente como o objetivo do método é calcular a distribuição posteriori dos parâmetros simulados, também este é utilizado para inferir valores de parâmetros. Com isto, os valores mostrados nos gráficos produzidos durante as simu-

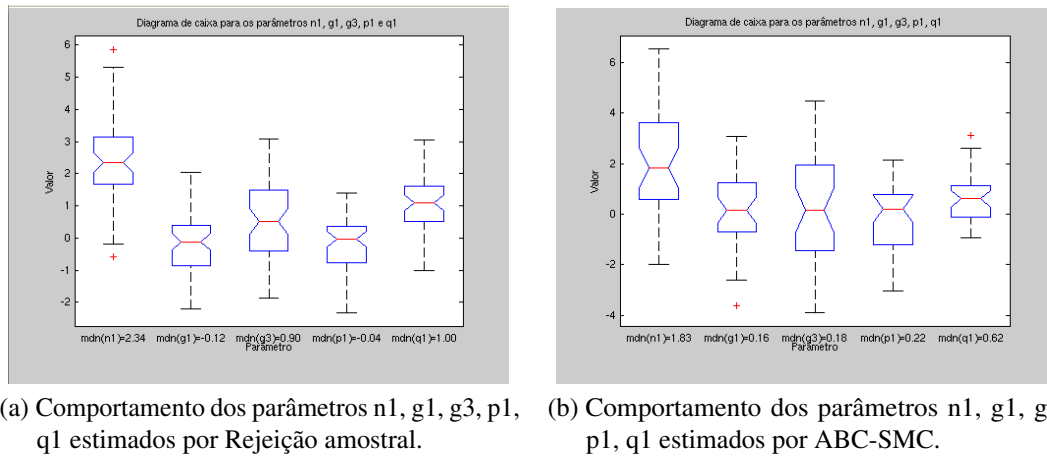


Figura 31 – Diagrama de caixas para $n1$, $g1$, $g3$, $p1$ e $q1$ para (a) 500 e (b) 22 medições das variáveis ao longo do tempo.

Tabela 5 – Valores das médias e medianas dos parâmetros $n1$, $g1$, $g3$, $p1$ e $q1$ do arquivo `last_param.dat`.

	Rejeição amostral			ABC-SMC			
Parâmetro	μ	σ^2	Mediana	μ	σ^2	Mediana	Valor observado
$n1$	2.4146	1.6288	2.3470	2.0310	4.7851	1.8293	2.6
$g1$	-0.2228	0.7756	-0.1206	0.1840	2.23	0.1646	0.1
$g3$	0.5752	1.4619	0.5084	0.3293	5.0463	0.1756	0.6
$p1$	-0.1763	0.7360	-0.0453	-0.1675	1.6962	0.2183	0.13
$q1$	1.0967	0.8860	1.003	0.7453	1.2548	0.6154	1.2

Fonte: Produzido pelo autor.

Nota: O método de Rejeição amostral foi realizado pelo pacote ao ter 500 medições das variáveis ao longo do tempo, enquanto que o método ABC-SMC foi utilizado pelo pacote ao ter 20 medições das variáveis ao longo do tempo.

lações (Figura 18, Figura 20, Figura 21, Figura 22, Figura 25, Figura 28 e Figura 30) somente sugerem a forma da distribuição e uma aproximação superficial do valor médio e a variância dos parâmetros simulados.

Identifica-se a limitação do pacote para abordar problemas complexos como o tratado neste trabalho. Devido a quantidade de dados e de operações necessárias para realizar as simulações apresentaram-se erros de *overflow* e de convergência. A primeira limitação corresponde à escolha da média e variâncias das priors já que estas tiveram que

Tabela 6 – Distribuições posteriores aproximadas para os parâmetros $n1$, $g1$, $g3$, $p1$ e $q1$.

Parâmetro	Figura 22b		Figura 30a		Figura 30f	
	μ	σ^2	μ	σ^2	μ	σ^2
$n1$	2.5	2.25	2.5	2.25	3	5.4
$g1$	0.0	1.8	0.5	1.36	0.0	4
$g3$	0.5	2.25	0.0	1.8	1.0	5.4
$p1$	-0.5	1.36	0.0	1.0	0.5	2.25
$q1$	1.5	2.25	1.0	1.8	0.0	4

Fonte: Produzido pelo autor.

estar perto dos valores observados, como se evidenciou na [Figura 23](#) e na [Tabela 3](#). Outra limitação corresponde à utilização do método dos quantis para estabelecer a sucessão de tolerâncias explicado por ([SILK; FILIPPI; STUMPF, 2012](#)) e no uso restrito de prioris como a uniforme e a normal que computacionalmente não oferecem um melhor desempenho, como na utilização da normal multivariada segundo ([FILIPPI et al., 2012](#)), devido à diminuição do tempo e do número de simulações requeridas para obter a posteriori que esta priori ofereceria. Outra limitação é a apresentação dos resultados, já que para saber com exatidão o valor dos dados tem-se de manipular os arquivos *.txt que o pacote gera para cada simulação e realizar tratamento dos dados.

Conforme os resultados poderia pensar-se que o valor para o parâmetro $q1$ não é de 1.2 como em ([POKHILKO et al., 2012](#)) se não como em [Pokhilko, Mas e Millar \(2013\)](#) o qual foi ajustado a 1.

Conclusão

No presente trabalho aplicou-se o algoritmo ABC-SMC utilizando o pacote ABC-SysBio para o sistema dinâmico dado pela rede regulatória genética do ciclo circadiano da *Arabidopsis thaliana* para ajustar os valores dos parâmetros com menos confiabilidade segundo (POKHILKO et al., 2012) e que estão relacionados com as proteínas LHY/CCA1: $n1$, $g1$, $g3$, $p1$ e $q1$, utilizando um conjunto de dados observados foi produzido pelo modelo desta rede e o conjunto de parâmetros encontrados na literatura.

Nas simulações realizadas para este trabalho, foi possível encontrar valores semelhantes aos cinco parâmetros observados, concluindo que o melhor ajuste encontrado ocorreu com cinco do que com dois e três. Isso sugere a existência de uma correlação entre eles a qual não foi evidenciada nos gráficos na [Figura 18](#), [Figura 20](#), [Figura 21](#), [Figura 22](#), [Figura 24](#), [Figura 27](#) e [Figura 29](#).

Sendo assim foi necessário realizar modificações para que o pacote pudesse efetuar o sequenciamento próprio do método ABC-SMC, estas foram na diminuição de 500 a 22 medições das variáveis ao longo do tempo. O anterior implicou perda de informação no comportamento das variáveis ao longo do tempo. O mesmo fato aconteceu quando foi utilizada inicialmente uma amostra de 1000 dados e foi diminuída a uma de 500 dados. Os melhores resultados obtiveram-se com a amostra de 1000 como está expressado na teoria. Nestes dois casos o pacote aplicou rejeição amostral. Adicionalmente, teve-se que diminuir a variâncias nas prioris utilizadas de 100 a 10, 5, 1.5 e 1, e utilizar médias com valores perto dos observados como sugere a literatura para encontrar valores simulados semelhantes aos observados o que implica ter um conhecimento priori dos valores. Também encontraram-se melhores resultados e desempenho computacional ao utilizar um valor para α de 0.75 do que um α de 0.25 também sugerido na teoria.

Os resultados obtidos com o pacote ABC-SysBio a nível dos gráficos de séries de tempo, não fornecem informação relevante para analisar a sequência das simulações devido à quantidade de dados. Os resultados fornecidos pelo pacote ABC-SysBio não representam resposta clara e de fácil interpretação, sendo assim necessária a manipulação dos dados fornecidos nos arquivos de saída para encontrar uma resposta clara. Os

valores encontrados nas simulações são semelhantes a aqueles que foram fornecidos em (POKHILKO et al., 2012) com uma variância pequena o qual poderia indicar que tais valores podem ser considerados em posteriores estudos.

No contexto de *Arabidopsis thaliana*, os resultados indicam que existe uma correlação entre os parâmetros considerados, pois o comportamento das reações é afetado pela presença de cada um destes e de todos os parâmetros considerados e de outras reações. Assim, modificar alguma destas poderia desencadear diferentes comportamentos nos outros parâmetros.

11 Trabalhos futuros

Como trabalhos futuros pretende-se:

- Visto que a eficiência computacional depende entre várias coisas do hardware utilizado, um trabalho interessante é realizar de novo as simulações feitas neste trabalho utilizando um servidor como os providenciados no laboratório LAMSA do Centro de Ciências Computacionais da Universidade Federal de Rio Grande e comparar os resultados obtidos.
- Como neste trabalho foram ajustados todos os parâmetros possíveis relacionados com a proteína CCA1/LHY é importante considerar também o ajuste dos parâmetros em TOC1 já que TOC1 e CCA1/LHY são proteínas fundamentais que regulam o comportamento da planta na presença ou ausência de luz.
- Utilizar o último modelo matemático descrito em [Pokhilko, Mas e Millar \(2013\)](#) e realizar as simulações para inferência de parâmetros para CCA1/LHY e TOC1.
- Implementar no pacote as propostas para estabelecimento de um kernel de perturbação que melhore computacionalmente como a normal multivariada, assim como também implementar formação da sucessão de tolerâncias $\{\epsilon_t\}_{1 \leq t \leq T}$ como foi proposto por ([SILK; FILIPPI; STUMPF, 2012](#)).
- Dado o carácter interdisciplinar deste trabalho, é necessário que em um grupo de trabalho interdisciplinar estes resultados tenham aplicação em plantas de importância econômica e /ou cultural.

Referências

AESCHBACHER, S.; BEAUMONT, M. A.; FUTSCHIK, A. A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics*, v. 192, n. November, p. 1027–1047, 2012. ISSN 00166731. Citado na página 41.

BARAGATTI, M.; PUDLO, P. An Overview On Approximate Bayesian Computation. *ESAIM: PROCEEDINGS*, v. 44, p. 291–299, 2014. Citado na página 26.

BEAUMONT, M. A. Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, v. 41, p. 379–406, 2010. ISSN 1543-592X. Citado na página 46.

BEAUMONT, M. A. et al. Adaptive approximate Bayesian computation. *Biometrika*, p. 1–8, 2009. ISSN 0006-3444. Disponível em: <<http://arxiv.org/abs/0805.2256>>. Citado na página 50.

BEAUMONT, M. A.; RANNALA, B. The bayesian revolution in Genetics. *Nature Reviews*, v. 5, p. 251–261, 2004. Citado 3 vezes nas páginas 37, 38 e 39.

BEAUMONT, M. A.; ZHANG, W.; BALDING, D. J. Approximate Bayesian Computation in Population Genetics. *Genetics*, v. 162, p. 2025–2035, 2002. Citado 3 vezes nas páginas 26, 45 e 47.

CHAN, Y. L.; ANDERSON, C. N. K.; HADLY, E. A. Bayesian estimation of the timing and severity of a population bottleneck from ancient DNA. *PLoS Genetics*, v. 2, n. 4, p. 451–460, 2006. ISSN 15537390. Citado na página 47.

CSILLÉRY, K. et al. Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, v. 25, n. 7, p. 410–418, 2010. Citado 3 vezes nas páginas 26, 39 e 47.

DEAN, B. T. A. et al. PARAMETER ESTIMATION FOR HIDDEN MARKOV MODELS WITH INTRACTABLE LIKELIHOODS. n. Marz, p. 1–49, 2011. Citado na página 47.

DELATORRE, C. A. A.; SILVA, A. A. d. Arabidopsis thaliana: uma pequena planta um grande papel. *Revista de Ciências Agrárias*, scielopt, v. 31, p. 58 – 67, 12 2008. ISSN 0871-018X. Disponível em: <http://www.scielo.gpeari.mctes.pt/scielo.php?script=sci_arttext&pid=S0871-018X2008000200007&nrm=iso>. Citado 3 vezes nas páginas 25, 31 e 32.

DIGGLE, P. J.; GRATTON, R. J. Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, Wiley for the Royal Statistical Society, v. 46, n. 2, p. pp. 193–227, 1984. ISSN 00359246. Disponível em: <<http://www.jstor.org/stable/2345504>>. Citado na página 45.

DIXON, L. E. et al. *Temporal repression of core circadian genes is mediated through EARLY FLOWERING 3 in Arabidopsis*. [S.l.], 2011. v. 21, 120–125 p. Disponível em: <<http://dx.doi.org/10.1016/j.cub.2010.12.013>>. Citado na página 57.

DROVANDI, C. C.; PETTITT, A. N. Estimation of Parameters for Macroparasite Population Evolution Using Approximate Bayesian Computation. *Biometrics*, v. 67, n. March, p. 225–233, 2011. ISSN 0006341X. Citado 2 vezes nas páginas 47 e 53.

FILIPPI, S. et al. On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. p. 1–24, 2012. Disponível em: <<http://www.degruyter.com/view/j/sagmb.2013.12.issue-1/sagmb-2012-0069/sagmb-2012-0069.xml>>. Citado 5 vezes nas páginas 45, 49, 52, 72 e 101.

GRIFFITHS, A. J. F. et al. *Introdução à genética*. [S.l.]: Guanabara Koogan, 2008. 303 p. Citado 2 vezes nas páginas 115 e 116.

GUILLEMAUD, T. et al. Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity*, Nature Publishing Group, v. 104, n. 1, p. 88–99, 2010. ISSN 0018-067X. Disponível em: <<http://dx.doi.org/10.1038/hdy.2009.92>>. Citado na página 47.

HAZEN, S. P. et al. LUX ARRHYTHMO encodes a Myb domain protein essential for circadian rhythms. *Proceedings of the National Academy of Sciences of the United States of America*, v. 102, n. 29, p. 10387–10392, 2005. ISSN 0027-8424. Citado na página 57.

LAIBACH, F. *Arabidopsis thaliana* (L.) Heynh. als object fur genetische und entwicklungsphysiologische untersuchun. *gen. Botanic Archives*, v. 44, p. 439–455, 1943. Citado na página 31.

LI, G. et al. Coordinated transcriptional regulation underlying the circadian clock in *Arabidopsis*. *Nature cell biology*, Nature Publishing Group, v. 13, n. 5, p. 616–622, 2011. ISSN 1465-7392. Disponível em: <<http://dx.doi.org/10.1038/ncb2219>>. Citado na página 57.

LIEPE, J. et al. ABC-SysBio-approximate bayesian computation in python with GPU support. *Bioinformatics*, v. 26, n. 14, p. 1797–1799, 2010. ISSN 13674803. Citado 3 vezes nas páginas 27, 63 e 64.

LIEPE, J. et al. *A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation*. [S.l.], 2014. v. 9, n. 2, 439–456 p. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/24457334>>. Citado 13 vezes nas páginas 25, 27, 37, 38, 39, 46, 48, 49, 52, 63, 64, 65 e 70.

LOCKE, J. C. W.; MILLAR, A. J.; TURNER, M. S. Modelling genetic networks with noisy and varied experimental data: The circadian clock in *Arabidopsis thaliana*. *Journal of Theoretical Biology*, v. 234, p. 383–393, 2005. ISSN 00225193. Citado na página 31.

LOPES, J. S.; BEAUMONT, M. A. ABC : A useful Bayesian tool for the analysis of population data. *"Infection, Genetics and Evolution"*, Elsevier B.V., v. 10, p. 826–833, 2010. ISSN 1567-1348. Disponível em: <<http://dx.doi.org/10.1016/j.meegid.2009.10.010>>. Citado 8 vezes nas páginas 26, 37, 38, 39, 41, 45, 46 e 47.

MARJORAM, P. et al. Markov chain Monte Carlo without likelihoods. *Proceedings Natural Academy Science U S A*, v. 100, n. 26, p. 15324–15328, 2003. ISSN 0027-8424. Citado na página 50.

MCCLUNG, C. R.; GUTIÉRREZ, R. A. Clocks, Network news: prime time for systems biology of the plant circadian clock truncated form of the title: Plant circadian. *Changes*, v. 20, n. 6, p. 588–598, 2010. Citado na página 31.

MCKINLEY, T.; COOK, A. R.; DEARDON, R. Inference in Epidemic Models without Likelihoods. *The International Journal of Biostatistics*, v. 5, n. 1, p. 1–40, 2009. ISSN 1557-4679. Citado na página 47.

MORAL, P. D.; DOUCET, A.; JASRA, A. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, v. 22, n. 1, p. 1009–1020, 2011. ISSN 0960-3174. Citado na página 53.

NOVÈRE, N. L. et al. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Research*, v. 34, n. Database issue, p. D689–D691, Jan 2006. Citado na página 71.

PINTO, D. *The Kullback-Leibler Distance*. 2007. Disponível em: <http://www.cs.buap.mx/~dpinto/research/CICLing07/_1/Pinto06c/node2.html>. Citado na página 53.

POKHILKO, A. et al. The clock gene circuit in *Arabidopsis* includes a repressilator with additional feedback loops. *Molecular Systems Biology*, v. 8, n. 574, p. 1–13, 2012. ISSN 1744-4292. Disponível em: <<http://dx.doi.org/10.1038/msb.2012.6>>. Citado 12 vezes nas páginas 26, 27, 33, 57, 61, 71, 72, 73, 99, 101, 103 e 104.

- POKHILKO, A. et al. Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model. *Molecular systems biology*, Nature Publishing Group, v. 6, n. 416, p. 1–10, 2010. ISSN 1744-4292. Disponível em: <http://dx.doi.org/10.1038/msb.2010.69>. Citado 2 vezes nas páginas 57 e 58.
- POKHILKO, A.; MAS, P.; MILLAR, A. J. Modelling the widespread effects of TOC1 signalling on the plant circadian clock and its outputs. *BMC Systems Biology*, BMC Systems Biology, v. 7, n. 23, p. 1–12, 2013. ISSN 1752-0509. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3614443&tool=pmcentrez&rendertype=abstract>. Citado 3 vezes nas páginas 33, 101 e 105.
- PORTOLÉS, S.; MÁ, S. P. The functional interplay between protein kinase ck2 and cca1 transcriptional activity is essential for clock temperature compensation in arabidopsis. *PLoS Genet*, Public Library of Science, v. 6, n. 11, p. e1001201, 11 2010. Disponível em: <http://dx.doi.org/10.1371%2Fjournal.pgen.1001201>. Citado na página 57.
- PRITCHARD, J. K. et al. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, v. 16, p. 1791–1798, 1999. ISSN 07374038. Citado 2 vezes nas páginas 41 e 45.
- ROSS-IBARRA, J. et al. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE*, Public Library of Science, v. 3, n. 6, p. e2411, 06 2008. Disponível em: <http://dx.plos.org/10.1371%2Fjournal.pone.0002411>. Citado na página 47.
- SECRIER, M.; TONI, T.; STUMPF, M. P. H. The ABC of reverse engineering biological signalling systems. *Molecular BioSystems*, v. 5, p. 1925–1935, 2009. ISSN 1742-206X. Citado 4 vezes nas páginas 41, 46, 47 e 54.
- SHRINER, D. et al. Evolution of intrahost HIV-1 genetic diversity during chronic infection. *Evolution*, v. 60, n. 6, p. 1165–1176, 2006. ISSN 0014-3820. Citado na página 46.
- SILK, D.; FILIPPI, S.; STUMPF, M. P. H. Optimizing Threshold – Schedules for Approximate Bayesian Computation Sequential Monte Carlo Samplers : Applications to Molecular Systems. p. 1–18, 2012. Citado 4 vezes nas páginas 54, 72, 101 e 105.
- SISSON, S. A.; FAN, Y.; TANAKA, M. M. Sequential Monte Carlo without likelihoods. *PNAS*, v. 104, n. 6, p. 1760–1765, 2007. ISSN 0027-8424. Citado 2 vezes nas páginas 26 e 50.
- SUNNÅKER, M. et al. Approximate Bayesian Computation. *PLoS Computational Biology*, v. 9, n. 1, p. 1–10, 2013. ISSN 1553734X. Citado 2 vezes nas páginas 45 e 50.

TALLMON, D. A.; LUIKART, G.; BEAUMONT, M. A. Comparative evaluation of a new effective population size estimator based on approximate Bayesian computation. *Genetics*, v. 167, p. 977–988, 2004. ISSN 00166731. Citado na página 47.

TANAKA, M. M. et al. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, v. 173, p. 1511–1520, 2006. ISSN 00166731. Citado na página 46.

TAVARÉ, S. et al. Inferring coalescence times from DNA sequence data. *Genetics*, v. 145, p. 505–518, 1997. ISSN 00166731. Citado na página 45.

The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, v. 408, n. December, p. 796–815, 2000. ISSN 0028-0836. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/11130711>>. Citado 2 vezes nas páginas 25 e 31.

TONI, T. et al. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, v. 6, p. 187–202, 2009. Citado 3 vezes nas páginas 27, 46 e 50.

TURNER, B. M.; ZANDT, T. V. A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, Elsevier Inc., v. 56, n. 2, p. 69–85, 2012. ISSN 00222496. Disponível em: <<http://dx.doi.org/10.1016/j.jmp.2012.02.005>>. Citado 6 vezes nas páginas 26, 41, 42, 47, 48 e 71.

USMAN, M. A.; DAVIS, S. J. An overview of natural variation studies in the *Arabidopsis thaliana* circadian clock. *Seminars in Cell and Developmental Biology*, Elsevier Ltd, v. 24, n. 5, p. 422–429, 2013. ISSN 10849521. Disponível em: <<http://dx.doi.org/10.1016/j.semcdb.2013.03.006>>. Citado na página 32.

WHITE, S. R.; KYPRAIOS, T.; PRESTON, S. P. Fast Approximate Bayesian Computation for discretely observed Markov models using a factorised posterior distribution. *arXiv*, p. 1–20, 2013. Citado 2 vezes nas páginas 42 e 45.

WILSON, D. J. et al. Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Molecular Biology and Evolution*, v. 26, n. 1993, p. 385–397, 2009. ISSN 07374038. Citado na página 46.

Apêndices

APÊNDICE A – Bases da regulação transcricional procariótica

Os organismos desenvolveram sistemas reguladores que acoplam a expressão dos produtos gênicos a sistemas sensores que detectam o composto relevante em um ambiente local (GRIFFITHS et al., 2008).

A célula tem um mecanismo que reprime a transcrição de todos os genes que codificam enzimas que não são necessárias em determinado tempo em quanto ativam esses genes que codificam enzimas agora necessárias. Assim, as células devem ser capazes de reconhecer condições ambientais como a presença ou ausência de luz, nas quais elas devem ativar ou reprimir a transcrição dos genes relevantes, devem ser capazes de ligar e desligar a transcrição de cada gene específico ou grupo de genes (GRIFFITHS et al., 2008).

A regulação transcricional depende das interações entre a proteína e o DNA que podem ser de duas formas: uma determina o lugar onde inicia a transcrição. Neste ponto um segmento de DNA chamado *promotor* interatua e que se liga a esse sítio é a proteína RNA polimerase. Quando isto ocorre, a transcrição pode começar algumas bases distantes do sítio promotor. Todos os genes devem ter um promotor para poder ser transcrito. A outra forma de interação decide se a transcrição pelo promotor ocorre. Os segmentos de DNA perto do promotor servem como pontos de ligação para proteínas reguladoras específicas de sequências chamadas *ativadores* e *repressores* (GRIFFITHS et al., 2008).

Para alguns genes, ocorre *regulação positiva* quando uma proteína ativadora deve ligar-se a seu sítio-alvo no DNA como um pré-requisito para a transcrição começar. Isto é porque a presença da proteína ligada é necessária para a transcrição. Para outros genes, ocorre a *regulação negativa* quando uma proteína repressora deve ser impedida de ligar-se a seu sítio-alvo como um pré-requisito para que a transcrição comece. Isto acontece quando a ausência do repressor ligado permite que a transcrição comece (GRIFFITHS et al., 2008).

Os ativadores e repressores regulam a transcrição quando uma proteína ativadora ligada ao DNA ajuda a levar RNA polimerase para seu promotor para que a polimerase possa comece a transcrição. Uma proteína repressora ligada ao DNA atua de duas formas: uma delas é interferindo com a ligação da RNA polimerase a seu promotor (bloqueando o início da transcrição) e a outra forma é impedindo o movimento da RNA polimerase ao longo da cadeia de DNA (bloqueando a transcrição). Juntas essas proteínas regulatórias e seus sítios de ligação constituem *interruptores genéticos* que controlam as mudanças eficientes na expressão genica dadas em resposta a condições ambientais (GRIFFITHS et al., 2008).

Anexos

ANEXO A – Espécies

<http://www.ebi.ac.uk/biomodels-main/BIOMD00000000412>

cCOP1c Initial Concentration: 0.3269

cCOP1d Initial Concentration: 0.2566

cCOP1n Initial Concentration: 0.65

cE3 Initial Concentration: 0.1503

cE3_m Initial Concentration: 0.2991

cE3n Initial Concentration: 0.0286

cE4 Initial Concentration: 0.207

cE4_m Initial Concentration: 0.1012

cEC Initial Concentration: 0.0709

cEG Initial Concentration: 0.0041

cG Initial Concentration: 0.0196

cG_m Initial Concentration: 0.1017

cL Initial Concentration: 0.506

cLUX Initial Concentration: 0.576

cLUX_m Initial Concentration: 0.1012

cL_m Initial Concentration: 1.0151

cLm Initial Concentration: 0.0788

cNI Initial Concentration: 0.0697

cNI_m Initial Concentration: 0.0731

cP Initial Concentration: 0.956

cP7 Initial Concentration: 0.1167

cP7_m Initial Concentration: 0.4016

cP9 Initial Concentration: 0.0238

cP9_m Initial Concentration: 0.0658

cT Initial Concentration: 0.0435

cT_m Initial Concentration: 0.0977

cZG Initial Concentration: 0.0755

cZTL Initial Concentration: 0.2505

s1 Initial Concentration: 0
s2 Initial Concentration: 0
s3 Initial Concentration: 0
s4 Initial Concentration: 0
s5 Initial Concentration: 0
s6 Initial Concentration: 0
s7 Initial Concentration: 0
s8 Initial Concentration: 0
s9 Initial Concentration: 0
s10 Initial Concentration: 0
s11 Initial Concentration: 0
s12 Initial Concentration: 0
s13 Initial Concentration: 0
s14 Initial Concentration: 0
s15 Initial Concentration: 0
s16 Initial Concentration: 0
s17 Initial Concentration: 0
s18 Initial Concentration: 0
s19 Initial Concentration: 0
s20 Initial Concentration: 0
s21 Initial Concentration: 0
s22 Initial Concentration: 0
s23 Initial Concentration: 0
s24 Initial Concentration: 0
s25 Initial Concentration: 0
s26 Initial Concentration: 0
s27 Initial Concentration: 0
s28 Initial Concentration: 0
s29 Initial Concentration: 0
s30 Initial Concentration: 0
s31 Initial Concentration: 0
s32 Initial Concentration: 0
s33 Initial Concentration: 0

s34 Initial Concentration: 0

s35 Initial Concentration: 0

s36 Initial Concentration: 0

s37 Initial Concentration: 0

s38 Initial Concentration: 0

s39 Initial Concentration: 0

s40 Initial Concentration: 0

s41 Initial Concentration: 0

s42 Initial Concentration: 0

s43 Initial Concentration: 0

s44 Initial Concentration: 0

s45 Initial Concentration: 0

s46 Initial Concentration: 0

s47 Initial Concentration: 0

s48 Initial Concentration: 0

s49 Initial Concentration: 0

s50 Initial Concentration: 0

s51 Initial Concentration: 0

ANEXO B – Parâmetros Globais(120)

n1 Value: 2.6 Constant
n2 Value: 0.64 Constant
n3 Value: 0.29 Constant
n4 Value: 0.07 Constant
n5 Value: 0.23 Constant
n6 Value: 20 Constant
n7 Value: 0.2 Constant
n8 Value: 0.5 Constant
n9 Value: 0.2 Constant
n10 Value: 0.4 Constant
n11 Value: 0.6 Constant
n12 Value: 12.5 Constant
n13 Value: 1.3 Constant
n14 Value: 0.1 Constant
g1 Value: 0.1 Constant
g2 Value: 0.01 Constant
g3 Value: 0.6 Constant
g4 Value: 0.01Constant
g5 Value: 0.15 Constant
g6 Value: 0.3 Constant
g7 Value: 0.6 Constant
g8 Value: 0.01 Constant
g9 Value: 0.3 Constant
g10 Value: 0.5 Constant
g11 Value: 0.7 Constant
g12 Value: 0.2 Constant
g13 Value: 1 Constant

g14 Value: 0.004 Constant
g15 Value: 0.4 Constant
g16 Value: 0.3 Constant
m1 Value: 0.54 Constant
m2 Value: 0.24 Constant
m3 Value: 0.2 Constant
m4 Value: 0.2 Constant
m5 Value: 0.3 Constant
m6 Value: 0.3 Constant
m7 Value: 0.7 Constant
m8 Value: 0.4 Constant
m9 Value: 1.1 Constant
m10 Value: 1 Constant
m11 Value: 1 Constant
m12 Value: 1 Constant
m13 Value: 0.32 Constant
m14 Value: 0.4 Constant
m15 Value: 0.7 Constant
m16 Value: 0.5 Constant
m17 Value: 0.5 Constant
m18 Value: 3.4 Constant
m19 Value: 0.2 Constant
m20 Value: 0.6 Constant
m21 Value: 0.08 Constant
m22 Value: 0.1 Constant
m23 Value: 1.8 Constant
m24 Value: 0.1 Constant
m25 Value: 1.8 Constant
m26 Value: 0.5 Constant
m27 Value: 0.1 Constant
m28 Value: 20 Constant
m29 Value: 5 Constant
m30 Value: 3 Constant

m31 Value: 0.3 Constant
m32 Value: 0.2 Constant
m33 Value: 13 Constant
m34 Value: 0.6 Constant
m35 Value: 0.3 Constant
m36 Value: 0.1 Constant
m37 Value: 0.8 Constant
m38 Value: 0.5 Constant
m39 Value: 0.3 Constant
a Value: 2 Constant
b Value: 2 Constant
c Value: 2 Constant
d Value: 2 Constant
e Value: 2 Constant
f Value: 2 Constant
p1 Value: 0.13 Constant
p2 Value: 0.27 Constant
p3 Value: 0.1 Constant
p4 Value: 0.56 Constant
p5 Value: 4 Constant
p6 Value: 0.6 Constant
p7 Value: 0.3 Constant
p8 Value: 0.6 Constant
p9 Value: 0.8 Constant
p10 Value: 0.54 Constant
p11 Value: 0.51 Constant
p12 Value: 3.4 Constant
p13 Value: 0.1 Constant
p14 Value: 0.14 Constant
p15 Value: 3 Constant
p16 Value: 0.62 Constant
p17 Value: 4.8 Constant
p18 Value: 4 Constant

p19 Value: 1 Constant
p20 Value: 0.1 Constant
p21 Value: 1 Constant
p22 Value: 0.5 Constant
p23 Value: 0.37 Constant
p24 Value: 10 Constant
p25 Value: 8 Constant
p26 Value: 0.3 Constant
p27 Value: 0.8 Constant
p28 Value: 2 Constant
p29 Value: 0.1 Constant
p30 Value: 0.9 Constant
p31 Value: 0.1 Constant
q1 Value: 1.2 Constant
q2 Value: 1.56 Constant
q3 Value: 2.8 Constant
L Value: 0.5
D Value: 0.5
E34 Value: 1 Constant
Gn Value: 1 Constant
EGn Value: 1 Constant
Lightstep Value: 0
offsetStep1 Value: 0
amplitudeStep1 Value: 1
phaseStep1 Value: 0
pulseDurationStep1 Value: 12
cyclePeriodStep1 Value: 24
rampDurationStep1 Value: 0.2

n15=0.4; n16=0.23; n17=0.1;

g17=0.03; g18=0.3; g19=0.5; g20=1.1; g21=0.5; g22=1.5; g23=0.01; g24=3;
g25=0.03; g26=0.3; g27=0.1; g28=0;

```

g=0; h=0;
q4=0;
period=24; dusk=12; daynumb = 5;
t = [0 daynumb * period];
options=odeset('stat','off');

y0=[1.0151 0.956 0.0755 0.0041 0.506 0.0977 0.0238 0.0731 0.0697 0.0196
0.0435 0.2505 0.0709 0.1017 0.0658 0.4016 0.1167 0.1012 0.207 0.0788 0.3102 0.0553
0.2991 0.1503 0.0286 0.65 0.2566 0.1012 0.576 0.3269];

```

Curator's comment: (updated: 04 Apr 2012 13:29:08 BST)

The plots corresponding to WT (Wild Type) in figure 2C of the reference publication has been reproduced here. The data were obtained by simulating the model using Copasi v4.8 (Build 35) and plotted using Gnuplot.

ANEXO C – Reações

$cL_m_trscr [s1] \rightarrow [cL_m]; cNI, cP, cP7, cP9, cT$

$cL_m_degr [cL_m] \rightarrow [s2];$

$cL_trsl [s3] \rightarrow [cL]; cL_m$

$cL_degr [cL] \rightarrow [s4];$

$cL_modif [s5] \rightarrow [cLm]; cL$

$cLm_degr [cLm] \rightarrow [s6];$

$cP_trsl [s7] \rightarrow [cP];$

$cP_degr [cP] \rightarrow [s8];$

$cP9_m_trscr [s9] \rightarrow [cP9_m]; cEC, cL, cP$

$cP9_m_degr [cP9_m] \rightarrow [s10];$

$cP9_trsl [s11] \rightarrow [cP9]; cP9_m$

$cP9_degr [cP9] \rightarrow [s12];$

$cP7_m_trscr [s13] \rightarrow [cP7_m]; cL, cLm, cP9$

$cP7_m_degr [cP7_m] \rightarrow [s14];$

$cP7_trsl [s15] \rightarrow [cP7]; cP7_m$

$cP7_degr [cP7] \rightarrow [s16];$

$cNI_m_trscr [s17] \rightarrow [cNI_m]; cLm, cP7$

$cNI_m_degr [cNI_m] \rightarrow [s18];$

$cNI_trsl [s19] \rightarrow [cNI]; cNI_m$

$cNI_degr [cNI] \rightarrow [s20];$

$cT_m_trscr [s21] \rightarrow [cT_m]; cEC, cL$

$cT_m_degr [cT_m] \rightarrow [s22];$

$cT_trsl [s23] \rightarrow [cT]; cT_m$
 $cT_degr [cT] \rightarrow [s24]; cZG, cZTL$
 $cE4_m_trscr [s25] \rightarrow [cE4_m]; cEC, cL$
 $cE4_m_degr [cE4_m] \rightarrow [s26];$
 $cE4_trsl [s27] \rightarrow [cE4]; cE4_m$
 $cE4_degr [cE4] \rightarrow [s28]; cCOP1d, cCOP1n, cE3n, cLUX$
 $cE3_m_trscr [s29] \rightarrow [cE3_m]; cL$
 $cE3_m_degr [cE3_m] \rightarrow [s30];$
 $cE3_trsl [s31] \rightarrow [cE3]; cE3_m$
 $cE3_degr [cE3] \rightarrow [s32]; cCOP1c$
 $cE3n_import [cE3] \rightarrow [cE3n];$
 $cE3n_degr [cE3n] \rightarrow [s33]; cCOP1d, cCOP1n, cE4, cG, cLUX$
 $cLUX_m_trscr [s34] \rightarrow [cLUX_m]; cEC, cL$
 $cLUX_m_degr [cLUX_m] \rightarrow [s35];$
 $cLUX_trsl [s36] \rightarrow [cLUX]; cLUX_m$
 $cLUX_degr [cLUX] \rightarrow [s37]; cCOP1d, cCOP1n, cE3n, cE4$
 $cCOP1c_trsl [s38] \rightarrow [cCOP1c];$
 $cCOP1c_degr [cCOP1c] \rightarrow [s39];$
 $cCOP1n_import [cCOP1c] \rightarrow [cCOP1n];$
 $cCOP1n_degr [cCOP1n] \rightarrow [s40];$
 $cCOP1d_activ [cCOP1n] \rightarrow [cCOP1d]; cP$
 $cCOP1d_degr [cCOP1d] \rightarrow [s41];$
 $cG_m_trscr [s42] \rightarrow [cG_m]; cEC, cL, cP$
 $cG_m_degr [cG_m] \rightarrow [s43];$
 $cG_trsl [s44] \rightarrow [cG]; cG_m$

$\text{cG_degr } [\text{cG}] \rightarrow [\text{s45}]; \text{cE3n}$

$\text{cG_cZTL_assoc } [\text{cG}] + [\text{cZTL}] \longleftrightarrow [\text{cZG}];$

$\text{cZTL_trsl } [\text{s46}] \rightarrow [\text{cZTL}];$

$\text{cZTL_degr } [\text{cZTL}] \rightarrow [\text{s47}];$

$\text{cZG_degr } [\text{cZG}] \rightarrow [\text{s48}];$

$\text{cG_cE3_assoc } [\text{cE3}] + [\text{cG}] \rightarrow [\text{cEG}];$

$\text{cEG_degr } [\text{cEG}] \rightarrow [\text{s49}]; \text{cCOP1c}, \text{cCOP1d}, \text{cCOP1n}, \text{cE3n}, \text{cG}$

$\text{cEC_form } [\text{s50}] \rightarrow [\text{cEC}]; \text{cCOP1d}, \text{cCOP1n}, \text{cE3n}, \text{cE4}, \text{cLUX}$

$\text{cEC_degr } [\text{cEC}] \rightarrow [\text{s51}]; \text{cCOP1d}, \text{cCOP1n}, \text{cE3n}, \text{cEG}, \text{cG}$