Regressão simbólica sobre séries temporais de dados meterológicos utilizando programação genética

Roberto Oliveira Santos[1], **Heitor Silvério Lopes**[2]

[1]Programa de Pós-graduação em Computação Aplicada - PPGCA
Universidade Tecnológica Federal do Paraná - UTFPR
Curitiba – PR – Brasil

py5gol@gmail.com, hslopes@utfpr.edu.br

Abstract. *This paper presents a heuristic method for time series modeling. The method is based on genetic programming, an evolutionary computation technique. The text explains the method, the treatment applied to the experiment data and the results achieved. The experiment was carried out with temperature data over a period of 14 years divided into Decentius. The performance was measured by the correlation coefficient, the coefficient of variation and mean square error. The results show that the heuristic method applied is suitable to capture the average behavior of the time series with a good level of performance.*

Keywords – Evolutionary Computation, Genetic Programming, Time Series, Weather data, Symbolic Regression.

Resumo. *Este artigo apresenta um método de heurística para modelagem de séries temporais. O método é baseado na programação genética, uma técnica de computação evolucionária. No texto explica-se o método, o tratamento aplicado aos dados do experimento e os resultados alcançados. O experimento foi realizado com dados de temperatura média em um período de 14 anos agrupados em decêndios. A performance foi medida através do coeficiente de correlação, o coeficiente de variação e o erro médio quadrático. Os resultados mostram que o método de heurística aplicado é apropriado para capturar o comportamento médio da série temporal com bom nível de performance.*

Palavras-chave – Computação Evolucionária, Programação Genética, Séries Temporais, Dados Meteorológicos, Regressão Simbólica

1. Introdução

Séries temporais são sequências de observações, medidas tipicamente em intervalos uniformes. A análise de séries temporais pode incluir diversos métodos estatísticos que pretendem entender os dados através da construção de um modelo. Modelos, por sua vez, são utilizados para entender o comportamento físico do processo. Eles são criados para capturar as características significantes e representa-las normalmente através de equações matemáticas. Utilizando um modelo apropriado, é possível predizer eventos futuros baseado nos eventos passados. Sob determinado aspectos, a modelagem de uma série temporal pode ser entendida com um problema de identificação, ou mais genericamente, um problema de regressão simbólica.

O problema de encontrar uma função que melhor se ajuste ao conjunto de observações é conhecido como regressão. Em sua forma mais comum, a estrutura de tal função é pré-fixada pelo analista e o problema se reduz a encontrar parâmetros numéricos que completem sua definição. Uma medida da discrepância entre a resposta prevista pelo modelo e aquela realmente observada é então minimizada por um algoritmo de otimização conveniente. Na regressão simbólica pretende-se resolver o mesmo problema sem se fixar a *priori* a forma ou os parâmetros da função. A técnica de regressão simbólica pode ser usada para resolver diferentes problemas como: descoberta de identidades trigonométricas, indução de seqüências, solução de equações diferenciais e integrais de forma simbólica, identificação de sistemas além de séries temporais, entre outros.

A utilização de dados climatológicos em estudos exige séries de dados meteorológicos com longos períodos de observação. Entretanto, é comum a existência de falhas nestas séries observadas devido a problemas com sensores de estações meteorológicas, problemas no armazenamento e/ou transmissão dos dados ou erro humano na coleta. Para estas situações, a regressão simbólica apresenta-se como uma alternativa viável e eficiente para complementar os dados faltantes com pequenas margens de erro.

A literatura apresenta uma quantidade expressiva de trabalhos que visam resolver problemas de regressão simbólica com a utilização de técnicas de computação evolucionária, especificamente programação genética.

Este artigo esta estruturado da seguinte forma: primeiro apresenta-se o problema de modelagem de séries temporais. Em seguida, aborda-se o método de programação genética, a metodologia utilizada no experimento e os resultados obtidos. Por fim, descreve-se as conclusões.

2. Fundamentação Teórica

Computação Evolucionária - CE - Charles Darwin (1859) defende em seu trabalho "Sobre a origem das espécies por meio de seleção natural" que a evolução dos seres vivos é baseada no Princípio da Seleção Natural onde "Indivíduos mais fortes e mais bem-adaptados ao meio ambiente têm maior chance de sobrevivência e de dar continuidade à sua espécie". A seleção natural atua sobre os indivíduos de uma população de forma probabilística e seu resultado a longo prazo é a evolução da população ou espécie.

Ao invés de população ou espécies de seres vivos, tem-se na CE um conjunto de indivíduos que representam potenciais soluções para determinado problema.

O objetivo das diversas técnicas de CE é obter uma boa solução visto não ser possível na maior parte do problemas identificar a melhor solução. Para isto, inicia-se com uma população aleatória que representa um conjunto inicial de soluções. Em seguida, geração após geração, aplicam-se os operadores genéticos para simular o processo evolutivo até que um determinado critério de término seja satisfeito [Ashlock 2006].

A CE abrange ou está associada a um grande número de algoritmos computacionais entre as quais pode-se citar: Algoritmos Genéticos (AG), Programação Genética (PG), Programação de Expressão Genética (PEG), Evolução Diferencial (ED), Colônia de Formigas - Ant Colony optimization (ACO), Enxame de Partículas - Particle Swarm Optimization (PSO), Colônia de Abelhas - Artificial Bee Colony (ABC), entre outros. Todos estes possuem uma característica de refinamentos sucessivos com comportamento

descrito na lista abaixo.

1. Gere soluções para os problemas;
2. Avalie as soluções;
3. Se a melhor solução satisfaz, então pare;
4. Selecione as melhores soluções;
5. Construa novas soluções utilizando partes das melhores soluções;
6. Eventualmente modifique as soluções;
7. Retorne para o passo 2;

Programação Genética - PG - É uma abordagem para geração automática de programas de computador desenvolvida por John Koza [Koza 1992] com base nos trabalhos de John Holland em Algoritmos Genéticos [Holland 1975]. Os programas possuem representação baseada em árvores sintáticas, através da combinação de funções e terminais adequados ao problema a ser solucionado [Koza 1992]. Os programas possuem tamanhos variados e virtualmente uma infinidade de combinações entre funções e terminais, gerando um espaço de busca de soluções infinito.

A PG permite que computadores resolvam problemas para os quais não foram previamente programados. Este processo de "ensinar" os computadores a programar é baseado em um conjunto de especificações de comportamento. A especificação de comportamento é definida, normalmente, por um conjunto de valores de entrada-saída, denominados casos de *fitness*, que representam o conjunto de treinamento. Com base neste conjunto, a PG visa evoluir um programa que produza primeiramente, de forma não trivial, as saídas corretas para cada entrada fornecida (casos de *fitness*) e, em segundo lugar, calcule as saídas de tal forma que, se as entradas forem representativamente escolhidas, o programa terá capacidade de obter saídas corretas para entradas não cobertas pelo conjunto de treinamento [OREILLY et al. 2005].

Pelo fato da PG manipular programas diretamente, uma estrutura de representação relativamente complexa e variável acompanha este paradigma. Nos modelos tradicionais, esta estrutura é uma árvore de sintaxe abstrata composta por funções em seus nós internos e por terminais em seus nós folhas. A especificação do domínio do problema é estabelecida pela definição dos conjuntos de funções e de terminais [Koza 1992].

Os conjuntos de funções e de terminais devem satisfazer às propriedades de fechamento e suficiência. A propriedade de fechamento estabelece que cada uma das funções do conjunto de funções seja capaz de aceitar, como seus argumentos, qualquer valor e tipo de dado que possa ser retornado por qualquer função do conjunto de funções, e qualquer valor de qualquer tipo que possa ser assumido por qualquer terminal do conjunto de terminais. A propriedade da suficiência requer que o conjunto de terminais e o conjunto de funções sejam capazes de expressar a solução do problema.

Segundo Rodrigues [Rodrigues 2007], parte-se de dois conjuntos: F como sendo o conjunto de funções e T como o conjunto de terminais. O conjunto F pode conter operadores aritméticos $(+, -, *, etc)$, funções matemáticas (seno, log, etc), operadores lógicos (E, OU, etc) dentre outros. Cada $f \in F$ tem associada uma aridade (número de argumentos) superior a zero. O conjunto T é composto pelas variáveis, constantes e funções de aridade zero (sem argumentos).

Cálculo do valor de $fitness$ - Segundo Koza [Koza 1992], a PG utiliza normalmente quatro medidas de *fitness* para avaliar quão adaptado está um indivíduo ao seu ambiente: *fitness* cru, *fitness* padronizado, *fitness* ajustado e *fitness* normalizado.

O *fitness* cru é uma medida que expressa a terminologia natural do próprio problema. O *fitness* padronizado expressa o valor do *fitness* cru de modo que o menor valor numérico seja sempre o melhor valor. O *fitness* ajustado é computado em função do *fitness* padronizado. O *fitness* normalizado possui três características: gera valores num intervalo de 0 a 1; é maior para os melhores indivíduos da população e a soma do *fitness* normalizado para todos os indivíduos de uma geração é igual a 1.

Métodos de seleção - A PG utiliza uma metodologia de seleção beaseada no valor de *fitness* dos indivíduos. Os indivíduos selecionados estarão sujeitos à ação dos operadores genéticos nos passos seguintes do método.

A seleção influencia a velocidade do processo evolucionário e esta pode, freqüentemente, levar o algoritmo a uma convergência prematura, ou seja, atingir uma solução que não satisfaz o problema. Esta convergência prematura pode ser causada por uma pressão seletiva excessiva. A introdução de pressão seletiva alta provoca invariavelmente um aumento da taxa de decrescimento da diversidade.

Dentre os vários métodos de seleção existentes, utiliza-se normalmente o da roleta (ou seleção proporcional) e o por torneio [Koza 1992].

O método da roleta consiste em dar uma porção de uma "roleta virtual' proporcional ao *fitness* de cada indivíduo da população, ou seja, cada indivíduo possui uma chance de ser escolhido proporcional ao seu valor de *fitness*. Desta forma os melhores indivíduos da população tendem a serem selecionados para uma posterior aplicação dos operadores genéticos, o que caracteriza o processo de evolução, ou seja, os indivíduos mais bem adaptados tendem a sobreviver.

A seleção por torneio é feita escolhendo-se aleatoriamente k indivíduos de uma população (k representa o tamanho do torneio e normalmente utiliza-se um tamanho equivalente a 10% do número de indivíduos que compõem a população). Então, o indivíduo de maior *fitness* dentre os k escolhidos é selecionado.

Operadores genéticos - A PG possui em sua forma mais básica, 3 operadores genéticos: reprodução, recombinação e mutação. O operador de reprodução seleciona um indivíduo dentro da população, de acordo com algum método de seleção, e copia-o para a próxima geração sem que este sofra nenhuma alteração em seu material genético ou sua estrutura. O operador de recombinação cria diversidade na população, pela produção de uma descendência que consiste de partes retiradas de cada indivíduo previamente selecionado. A operação de recombinação começa com dois indivíduos-pais que trocam material genético originando dois novos indivíduos. O operador de mutação não é muito utilizado pois sua utilização não traz benefícios significativos à PG [Koza 1992].

3. Metodologia

Para avaliar a utilização da PG em problemas de regressão símbolica foram selecionados os dados de temperatura média da estação meteorológica do Instituto Tecnológico SIMEPAR localizada na cidade de Palotina, extremo oeste do Estado do Paraná. Os dados de temperatura são coletados pelos sensores da estação em intervalos de 15 (quinze)

minutos e transmitidos para o servidor de banco de dados via tecnologia GPRS. Os dados abrangem o período entre julho de 1997 a março de 2011.

O roteiro seguido para aplicação da PG seguiu os passos abaixo:

1. Definição do conjunto de terminais (T);
2. Definição do conjunto de funções (F);
3. Definição dos casos de *fitness*;
4. Definição das medidas de *fitness*;
5. Definição dos parâmetros de controle e variáveis qualitativas;
6. Definição do critério de parada;
7. Especificação do resultado.

Definição do conjunto de terminais (T) - Este conjunto foi definido utilizando os seguintes elementos:

$$T = \{[0,1], 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, \pi, x, x_{-1}, x_{-2}, x_{-3}\} \tag{1}$$

onde:

- $[0,1]$ representa o intervalo de valores decimais entre 0 e 1;
- $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, \pi$ representam constantes;
- $x, x_{-1}, x_{-2}, x_{-3}$ representam respectivamente os casos de *fitness* no momento atual e anteriores.

O intervalo entre $x, x_{-1}, x_{-2}, x_{-3}$ foi definido visando induzir a PG a encontrar o comportamento sazonal do dado coletado, portanto, a PG poderia fazer uso da informação da temperatura dos períodos anteriores para estimar a temperatura do decêndio atual).

Definição do conjunto de funções (F) - Após a realização de testes experimentais, identificou-se o seguinte conjunto de funções:

$$F = \{+, -, *, /, seno, sigmoide, sinc\} \tag{2}$$

onde:

- +, -, * representam as operações básicas;
- / representa a divisão protegida com numerador $>= 1$;
- seno representa a função trigonométrica tradicional;
- sigmóide definida por $sigm(x) = \frac{1}{e^{-x}}$;
- sinc definida por $sinc(x) = \frac{sin(x\pi)}{x\pi}$

Definição dos casos de *fitness*; - Os dados coletados foram utilizados como especificação do comportamente (casos de *fitness*) para o programa a ser gerado pela PG. Neste caso em específico, regressão símbolica, o programa a ser gerado será uma expressão matemática que tenta representar da melhor forma possível o comportamento informado como entrada da PG. Visando utilizar um intervalo de tempo bastante comum na atividade de agricultura, agrupou-se os casos de *fitness* em intervalos de 10 dias (decêndios), totalizando 449 casos de *fitness*. Os casos de *fitness* estão disponíveis na figura 1.

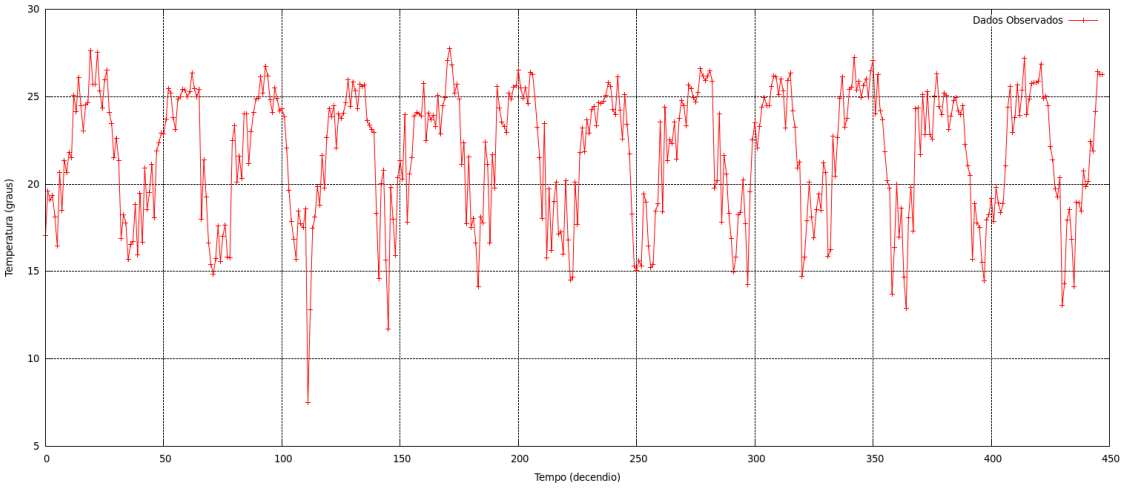

Figura 1. Dados observados de temperatura média no período entre julho de 1997 e março de 2011

Em seguida, iniciou-se a definição do conjunto de funções e terminais a serem utilizados no problema. Como visto na revisão bibliográfica, o espaço de busca de uma PG é infinito e a inclusão de novos elementos no conjunto F de funções ou T de terminais aumenta a complexidade do problema na qual a PG atua. Entretanto, a exclusão de funções ou terminais pode piorar a qualidade das soluções visto que a PG trabalharia com um conjunto reduzidos de elementos para elaborar a árvore sintática que cada indivíduo da população representa. Portanto, a definição destes conjuntos deve ser realizada com parcimônia, sempre que possível baseado nas características do problema. Neste trabalho utilizou-se o comportamento apresentado pelo casos de *fitness* que pode ser visto na figura 1.

Definição das medidas de *fitness* - Segundo Lopes [Lopes and Weinert 2004], existem diversas formas de avaliar a performance do modelos obtidos mas nenhum consenso sobre o assunto. Neste trabalho os modelos foram avaliados utilizando três medidas: coeficiente de correlação de Pearson (R), definido pela equação (3); o coeficiente de variação (CV), definido pela equação (4); e o erro médio quadrático normalizado (*normalized mean square error* - NMSE), definido pela equação (5).

$$R = \frac{N * \sum_{i=1}^{N}(x_i \bar{x}_i) - (\sum_{i=1}^{N} x_i) * (\sum_{i=1}^{N} \bar{x}_i)}{\sqrt{[N * \sum_{i=1}^{N}(x_i)^2 - (\sum_{i=1}^{N} x_i)^2] * [N * \sum_{i=1}^{N}(\bar{x}_i)^2 - (\sum_{i=1}^{N} \bar{x}_i)^2]}} \tag{3}$$

$$CV = \frac{1}{x}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}_i)^2} \tag{4}$$

$$NMSE = \frac{1}{\sigma^2}[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}_i)^2] \tag{5}$$

O coeficiente de correlação mensura o ajuste entre o modelo gerado pela PG e os casos de *fitness*, variando de -1 a 1. Um valor positivo significa uma correlação linear positiva; um valor negativo significa o oposto. Valores de R próximos de 0 significam uma correlação ruim ou nenhuma. O coeficiente de variação mensura o espalhamento do dado em relação a média e portanto, quanto mais próximo de zero, melhor. O NMSE é um método para comparar a média de uma série contra os valores gerados pelo modelo.

Definição dos parâmetros de controle e variáveis qualitativas - O experimento apresentado foi executado com os seguintes parâmetros: geração da população inicial = ramped half-and-half, seleção = torneio, tamanho do torneio = 7, número de populações = 1, tamanho da população = 50000, operadores genéticos: clonagem, recombinação e mutação, função de *fitness* = soma dos erros absolutos, profundidade máxima da árvore de solução: 10, Probabilidade de recombinação = 0,9 e Probabilidade de reprodução = 0,1.

Definição do critério de parada - Foi definido como critério de parada o número máximo de gerações estipulado em 100 + 1.

4. Resultados

Na figura 2 é possível visualizar a série observada em sobreposição com a série gerada pelo modelo da PG. Neste experimento a PG conseguiu reproduzir fielmente o comportamento médio da série, entretanto, não foi possível recuperar as mudanças bruscas, em especial, nas baixas temperaturas.

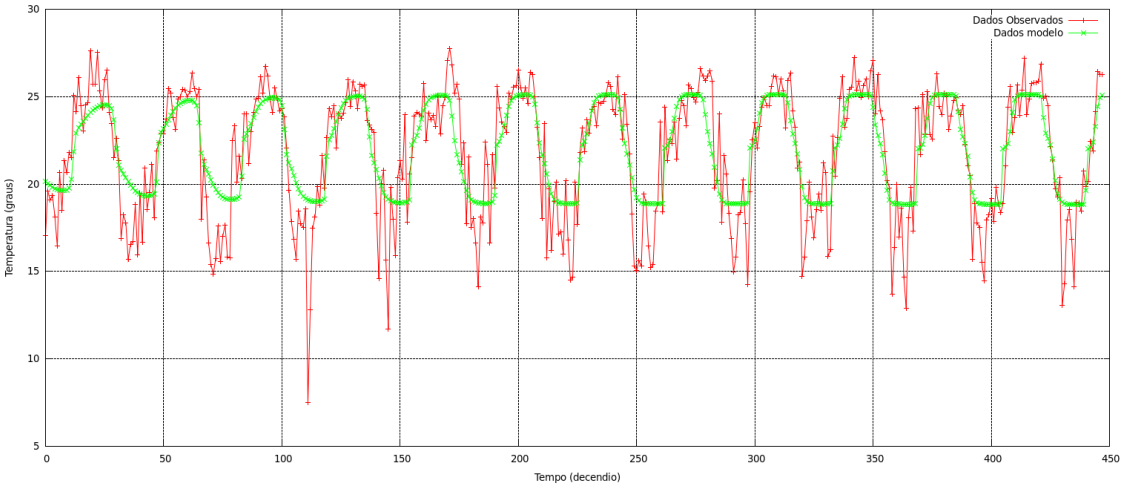

Figura 2. Dados observados x Dados do modelo de temperatura média no período entre julho de 1997 e março de 2011

Na figura 3 é possível observar o comportamento da valor de *fitness* do melhor indivíduo de cada geração durante a execução da PG.

Foram obtidos valores de R = 0,838803, CV = 0,095568 e NMSE=0,344196. A representação S do modelo gerado pela programação genética: *(* π (+ (+ (+ (- x x3) (sigm (sinc (+ (* x 0.056) (* π 0.971))))) (- x x3)) (sigm (* (* (+ (+ π (+ π π)) (+ (* (* x 0.056) 0.056) (* (- x x3) 0.971))) (sigm (* 0.537] (sinc (* 0.054 x))))) (sinc (- (sigm (+ (- x x3) (- x x3))) (+ (+ π π) (* x 0.056))))))))*

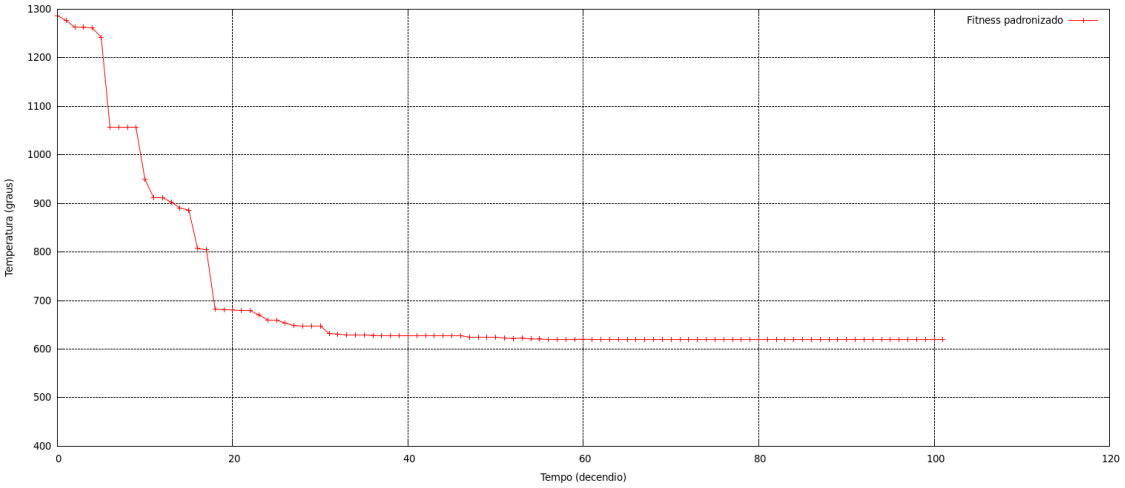

Figura 3. Valor de *fitness* padronizado do modelo gerado pela programação genética

5. Conclusão

Neste artigo apresentou-se a aplicação da programação genética em problemas de regressão simbólica. Foram utilizados dados de temperatura média que representam um desafio, visto que séries univariadas não incluem explicitamente informações sobre o processo físico que gerou a série. Os resultados foram interessantes, pois é possível avaliar visualmente que a PG "capturou" o comportamento médio da série. Os resultados dos coeficientes R, CV e NMSE também corroboram este entendimento, visto que os valores se aproximam da média em ambos os testes. A programação genética apresentou-se como uma ferramenta adequada ao problema de regressão simbólica devido a facilidade de entendimento e de configuração/execução, gerando modelos compatíveis com os dados observados. Os autores agradecem ao SIMEPAR pelo fornecimento dos dados utilizados neste trabalho.

Referências

Ashlock, D. (2006). *Evolutionary Computation for Modeling and Optimization*. Springer, Ontario.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor.

Koza, J. R. (1992). *Genetic Programming: on the programming of computers by means of natural selection*. MIT Press, Cambridge.

Lopes, H. S. and Weinert, W. (2004). A gene expression programming system for time series modeling. *Proceedins of XXV Iberian Latin American Congress on Computational Methods in Engineering (CILAMCE)*, CD-ROM.

OREILLY, U., YU., T., RIOLO, R., and WORZEL, B., editors (2005). *Genetic Programming Theory and Practice II*. Springer, Boston.

Rodrigues, E. L. M. (2007). *Inferência de gramáticas formais livres de contexto utilizando computação evolucionária com aplicação em bioinformática*. Dissertation, Universidade Tecnológica Federal do Paraná, Curitiba.