RECIIS
Revista Eletrônica de Comunicação Informação & Inovação em Saúde
[www.reciis.cict.fiocruz.br]
ISSN 1981-6278

SUPLEMENTO – BIOINFORMÁTICA E SAÚDE

Artigos originais

Inferência gramatical aplicada à modelagem lingüística de redes de regulação biológicas[1]

DOI: 10.3395/reciis.v1i2.Sup.104pt

Elias Bareinboim
Programa de Engenharia de Sistemas e Computação, COPPE, UFRJ / Laboratório de Bioinformática, LNCC, Rio de Janeiro, Brasil
eliasb@ufrj.br

Ana Tereza R. Vasconcelos
Laboratório de Bioinformática, LNCC, Rio de Janeiro, Brasil
atrv@lncc.br

João C. P. da Silva
Departamento de Ciência da Computação, Instituto de Matemática, UFRJ, Rio de Janeiro, Brasil
jcps@dcc.ufrj.br

Resumo
Apresentamos uma metodologia baseada em algoritmos de inferência gramatical aplicada à modelagem lingüística de redes de regulação biológicas. A abordagem lingüística para o problema de redes de regulação foi proposta por Collado-Vides, que provou e formalizou a necessidade de utilização de linguagens sensíveis ao contexto para representar tais redes. A aprendizagem de linguagens sensíveis ao contexto é uma tarefa difícil; nossa metodologia propõe descrever tal classe a partir de linguagens de natureza mais simples, que possam ser aprendidas por algoritmos de inferência gramatical já consolidados. Além da metodologia proposta, sugerimos direções para esta pesquisa que nos parecem promissoras.

Palavras-chave
Regulação gênica, modelagem lingüística, linguagens sensíveis ao contexto, expressão regular aumentada, inferência gramatical

O objetivo deste trabalho é apresentar uma proposta de refinamento da abordagem lingüística definida por Collado-Vides (COLLADO-VIDES, 1989; 1991; 1992; 1993a; b) para a modelagem de redes de regulação gênica utilizando algoritmos de inferência gramatical. Após um levantamento extenso sobre os promotores sigma 70 da bactéria *E. coli* (COLLADO-VIDES, 1993c), o autor construiu uma gramática que gerava uma linguagem contendo todas as seqüências regulatórias conhecidas até então.

A gramática resultante satisfazia algumas propriedades que foram estabelecidas como relevantes (COLLADO-VIDES, 1993b; c). Por exemplo, os nucleotídeos individualmente, em pares ou em tripletos, não deveriam ser utilizados como a menor unidade de informação passível de manipulação no sistema, ou seja, não deveriam ser o alfabeto da linguagem. Para este papel, foram consideradas como relevantes três tipos de categorias: os promotores (Pr), os operadores (Op) e os ativadores (I).

Uma vez que estas categorias foram estabelecidas, COLLADO-VIDES (1993b; c) procurou selecionar propriedades simples e relevantes que caracterizassem melhor cada uma delas, e que permitissem a definição da gramática. As propriedades consideradas foram: (i) existência de um sítio proximal (dentro do intervalo de -60 a +20 nucleotídeos) relativo a um promotor; (ii) obediência ao *principio de precedência proximal*, que estabelece que os operadores proximais são representados à direita do promotor, e ativadores proximais são representados à esquerda do promotor; (iii) obediência ao *principio da precedência posicional*, que estabelece que dados os sítios A e B, cuja posição relativa na fita de DNA é, repectivamente, c_1 e c_2 , se $c_1 < c_2$, então dizemos que A precede B; (iv) outros sítios, que não são classificados como proximais, são considerados sítios opcionais e remotos; (v) identificação, através de um atributo, de quais proteínas podem se ligar a sítios proximais e remotos; (vi) identificação de dois tipos de coordenadas: as *coordenadas c*, que servem para explicitar as distâncias absolutas dos sítios até o promotor; e as *coordenadas d*, que servem para definir as distâncias relativas entre os sítios remotos e os sítios proximais homólogos.

A linguagem gerada por esta gramática pertence à chamada classe de linguagens sensíveis ao contexto (COLLADO-VIDES, 1991). Dentro da Hierarquia de Chomsky (CHOMSKY, 1959; ULLMAN et al., 2001), esta classe é uma das mais complexas do ponto de vista computacional, estando acima das classes das linguagens livres de contexto e das linguagens regulares, sendo esta última a mais simples da hierarquia.

Em seu trabalho, Collado-Vides gerou sua gramática manualmente a partir de alguns exemplos de seqüências regulatórias e de algumas das propriedades identificadas como representativas. Nosso objetivo é gerar tal gramática automaticamente aplicando algoritmos de aprendizado de máquina ao conjunto de exemplos das seqüências regulatórias.

Inicialmente, buscamos na literatura algoritmos de aprendizado de máquina aplicados à classe de linguagens que fossem sensíveis ao contexto. Em geral, a aprendizagem de linguagens sensíveis ao contexto é uma tarefa árdua. Tal fato decorre basicamente da grande complexidade de tal classe de linguagens, seja do ponto de vista teórico assim como computacional. Notada esta dificuldade técnica de se trabalhar com tais linguagens, buscamos formas alternativas para abordar este problema.

Em ALQUÉZAR et al. (1995; 1997) foi proposta uma abordagem para aprendizagem de linguagens sensíveis ao contexto através da aproximação por linguagens regulares. Ou seja, inicialmente aprende-se uma linguagem mais simples (regular), e então refina-se, restringindo esta linguagem e gerando uma linguagem sensível ao contexto associada. A classe de linguagens sensíveis ao contexto gerada não é completa, porém vai além das linguagens consideradas triviais (FU, 1982).

As expressões regulares aumentadas (AREs) (ALQUÉZAR et al., 1995; 1997) são utilizadas para descrever, reconhecer e aprender classes de linguagens sensíveis ao contexto. Elas associam o poder descritivo das expressões regulares, utilizadas para denotar linguagens regulares, a um conjunto de restrições lineares que correlacionam os símbolos da expressão.

Resumidamente, a manipulação das AREs é feita em duas fases, a saber, aprendizado e reconhecimento. O método de aprendizado que irá induzir a ARE pode ser descrito por três passos básicos.

Passo 1. Inferência da Gramática Regular

Entrada: Conjunto de exemplos positivos e negativos.

Saída: Autômato finito que reconhece uma linguagem regular que contém todos os exemplos positivos e nenhum negativo.

Passo 2. Tradução

Entrada: Autômato obtido no passo anterior.

Saída: Expressão regular equivalente ao autômato da entrada.

Passo 3. Processo de Indução das Restrições

Entrada: Expressão regular obtida no passo anterior.

Saída: Linguagem sensível ao contexto que aproxima o conjunto alvo desejado, aceitando os exemplos positivos e rejeitando os exemplos negativos.

O método de reconhecimento pode ser apresentado em dois passos.

Passo 1. Verificação preliminar da expressão, no qual se averigua se a expressão, independentemente das restrições, é consistente em relação à expressão regular original, sendo uma expressão no formato do passo três do processo de aprendizado.

Passo 2. Verificação das restrições, no qual, depois da avaliação da expressão e sucesso na confirmação da mesma, deve-se averiguar se o conjunto de restrições não está sendo violado.

Dado o exposto, nossa proposta é adaptar a metodologia das AREs ao problema de refinamento da representação lingüística das expressões regulatórias propostas por COLLADO-VIDES (1991; 1992) e ROSENBLUETH (1996). Em última instância, esperamos obter como resultado um conjunto com o maior número possível de seqüências regulatórias, além de novas configurações candidatas que possam ser comprovadas experimentalmente.

Os principais passos da metodologia proposta podem ser descritos da seguinte maneira:

• **algoritmo A_0**: faz a tradução automática das informações (fator de transcrição, promotor, posições iniciais e finais da seqüência, tipo e posição central) presentes no banco de dados RegulonDB (HUERTA et al., 1998) para cláusulas Prolog (BRATKO, 2000) que serão utilizadas para calcular os possíveis candidatos a seqüências regulatórias.

As melhores seqüências candidatas são selecionadas (ROSENBLUETH, 1996; HERTZ, 1999), e então cons-

truímos o conjunto de exemplos positivos. Este conjunto é utilizado como entrada para o algoritmo de inferência gramatical A_1;

• **algoritmo A_1**: utiliza-se um algoritmo de inferência gramatical para linguagens regulares, gerando um autômato finito. Não existe restrição inicial em relação a tal algoritmo, sendo possível a utilização de qualquer algoritmo de inferência gramatical já disponível na literatura. Se o autômato gerado for não determinístico, utiliza-se o algoritmo A_2. Caso contrário, utiliza-se diretamente o algoritmo A_3;

• **algoritmo A_2**: transforma um autômato finito não-determinístico em um autômato finito determinístico (ULLMAN et al., 2001);

• **algoritmo A_3**: transforma um autômato finito determinístico na expressão regular correspondente (ULLMAN et al., 2001);

• **algoritmo A_4**: a partir da expressão regular gerada pelo algoritmo A_3 e o conjunto de exemplos gerado pelo algoritmo A_0, construímos um conjunto de restrições lineares (ALQUÉZAR et al., 1995).

A expressão resultante do algoritmo A_4 em conjunto com a expressão regular de A_3 caracteriza a linguagem sensível ao contexto desejada e é chamada Expressão Regular Aumentada (*Augmented Regular Expression*, ARE).

Observamos que para um funcionamento adequado de nossa proposta, devemos fazer uma adaptação na representação das seqüências regulatórias. Diferentemente da proposta de COLLADO-VIDES (1993b), as distâncias deverão ser representadas como uma nova categoria e não mais como uma propriedade das categorias existentes, permitindo que as distâncias sejam expressas através das restrições geradas pela ARE.

Todos os algoritmos descritos anteriormente já foram implementados. O algoritmo de aprendizado de linguagens regulares A_1 que estamos utilizando e avaliando, é o *K-tail* (FU, 1982).

Optamos para o desenvolvimento do trabalho nas seguintes direções, a saber, utilização de outros algoritmos de aprendizagem gramatical, ARE não linear, classificação utilizando contexto, gramáticas probabilísticas e outras propostas pontuais.

Mais especificamente, estamos pesquisando outros algoritmos de aprendizagem gramatical de linguagens regulares em substituição ao algoritmo *K-tail*, e pretendemos testá-los. No problema em questão, o algoritmo *K-tail* permite a restrição da linguagem regular até um determinado limite baseada em um parâmetro k, que representa o tamanho da cauda de uma palavra. Tal restrição implica que ainda existe uma quantidade de palavras bastante significativa que gostaríamos de reduzir. A busca por outros algoritmos de aprendizagem tem como objetivo contornar este tipo de restrição.

A abordagem das AREs atualmente utiliza como restrições somente funções lineares, que impõem certa rigidez ao tratamento das distâncias entre as categorias (Pr, Op, I). Para contornar tal fato, devemos atribuir distribuições de probabilidades a estas distâncias e depois modelar uma estrutura de dependências entre as variáveis presentes nas restrições lineares. Baseados nesta modelagem, obteríamos as distribuições de probabilidades, possivelmente condicionais, em relação às distâncias entre os sítios proximais.

Na área de classificação textual, os trabalhos de COHEN et al. (1996) e FREUND et al. (1997) apresentam dois interessantes algoritmos que contemplam a questão do contexto, chamados de *Sleeping Experts* e *RIPPER*. Ambos os algoritmos são para classificação e não geração de palavras.

Esses algoritmos são atraentes para problemas de classificação de textos de grande porte, sendo eficientes e robustos, tolerantes a ruídos, com execução em tempo linear ou quase linear. Ambos permitem que o contexto de uma palavra influencie como será feita a classificação da sentença como um todo. Tal classificação é internamente tratada através de classificadores lineares e não lineares, respectivamente.

Pretendemos verificar a possibilidade de utilização desses algoritmos em duas situações: (i) substituindo todo o arcabouço das AREs; (ii) utilizando de forma combinada, concatenado à ARE, classificando as palavras que são geradas pela ARE de acordo com os exemplos positivos e negativos originais utilizados no processo de aprendizagem.

Quanto às gramáticas probabilísticas, em uma construção típica de uma gramática determinística geramos várias regras de transição, sendo que algumas são aplicadas um grande número de vezes enquanto outras somente a alguns poucos exemplos isolados. Caso eliminássemos as regras pouco utilizadas, a gramática não seria completa. Esta distinção entre regras muito ou pouco utilizadas não é contemplada na descrição de uma gramática típica.

As gramáticas probabilísticas são definidas como gramáticas que têm associadas a cada regra uma probabilidade (STERGOS et al., 2001). O resultado é que quando o gerador sintático retorna uma nova palavra, ele associa à mesma sua probabilidade de ser gerada. Essa probabilidade pode posteriormente ser utilizada para determinar qual entre todas as palavras geradas é a mais provável de ser obtida. Com isso, podemos ignorar palavras peculiares e pouco prováveis de serem obtidas na realidade. Uma outra possibilidade é utilizar *Modelos de Markov Ocultos (Hidden Markov Models,* HMMs*)*, que são um tipo específico de autômato com probabilidades associadas aos seus estados.

Podemos sugerir algumas outras propostas pontuais como:

• Comparar a *performance* entre as abordagens ARE, Collado-Vides, categorização textual, ARE não linear, gramáticas probabilísticas;

• Utilizar o conhecimento implícito adquirido com a bactéria *E. coli* através da execução de tal metodologia para inferir seqüências em outros organismos, filogeneticamente similares (transferência de conhecimento);

• Baseado no modelo de analisador sintático estatístico (COLLINS et al., 1997), usar uma aproximação através de gramáticas livres de contexto, ao invés de regulares, para então se aproximar a gramática alvo. Tal abordagem também é utilizada em problemas de classificação textual.

Através do exposto neste trabalho, e de forma mais detalhada em BAREINBOIM (2005), podemos considerar a metodologia proposta como sendo um bom ponto de partida para um melhor entendimento das redes de regulação biológicas. Além disso, tais avanços teóricos em conjunto com as propostas detalhadas de atuação, de como desenvolver tal trabalho, são bastante desafiadoras e interessantes como objeto de estudo futuro.

Notas

1. Este projeto foi financiado através de apoio especializado de projetos multi-tarefas do Laboratório de Bioinformática (LABINFO) do Laboratório Nacional de Computação Científica/Ministério da Ciência e Tecnologia [Processo Nº 506321/2004-5].

Referências bibliográficas

ALQUÉZAR, R.; SANFELIU, A. Augmented Regular Expressions: A formalism to describe, recognize and learn a class of context-sensitive languages. **Relatório Técnico (12-06)** – Politécnica de Cataluna, 1995. Disponível em: http://www.lsi.upc.es/dept/techreps/techreps.html. Acesso em: 11 out. 2007.

ALQUÉZAR, R.; SANFELIU, A. Recognition and learning of a class of context-sensitive language described by augmented regular expressions. **Pattern Recognition**, v.30, p.163-182, 1997.

BAREINBOIM, E. **Técnicas de inteligência artificial aplicadas ao problema das redes de regulação biológicas**. 2005. 75f. Monografia (Graduação em Ciência da Computação) – Departamento da Ciência da Computação, Instituto de Matemática, Universidade Federal do Rio de Janeiro.

BRATKO, I. **Prolog**: Programming for Artificial Intelligence. Addison Wesley. 3rd edition, 2000.

CHOMSKY, N. On certain formal properties of grammars. **Information and Control**, v.2, p. 91-112, 1959.

COHEN, W.; SINGER Y. Context-Sensitive Learning Methods for Text Categorization. In: ACM International Conference on Research and Development in Information Retrieval, 19., 1996, Nova Iorque. **Proceedings...** Nova Iorque: ACM Press, 1996. p.307-315.

COLLADO-VIDES, J. A Transformational-Grammar Approach to the Study of The Regulation of Gene Expression. **Journal of Theoretical Biology**, v.136, p.403-425, 1989.

COLLADO-VIDES J. The Search of a Grammatical Theory of Gene Regulation is Formally Justified by Showing the Inadequacy of Context-free Grammars. **Computer Applications in the Bioscience *(CABIOS)*** v.7, p. 321-326, 1991.

COLLADO-VIDES, J. Grammatical model of the regulation of gene expression. **Proceedings of National Academy of Science, USA**, v.89, p. 9405-9409, 1992.

COLLADO-VIDES, J. The Elements for a Classification of Units of Genetic Information with a Combinatorial Component. **Journal of Theoretical Biology**, v. 163, p.527-548, 1993.

COLLADO-VIDES, J. A Linguistic Representation of the Regulation of Transcription Initiation: I. An Ordered Array of Complex Symbols with Distinctive Features. **Biosystems**,v.29, p.87-104, 1993.

COLLADO-VIDES, J. A Linguistic Representation of the Regulation of Transcription Initiation: II. Distinctive Features of Sigma 70 Promoters and their Regulatory Binding Sites. **Biosystems**,v.29, p.105-128, 1993.

COLLINS, M. Three Generative, Lexicalised Models for Statistical Parsing. In: Annual Meeting of the Association for Computational Linguistics, 35. and Conference of the European Chapter of the Association for Computational Linguistics, 8., 1997, Nova Jersey. **Proceedings...** Nova Jersey: Association for Computational Linguistics, 1997. p.16-23. Disponível em: www.aclweb.org/anthology/P97-1003.pdf. Acesso em: 11 out. 2007.

FREUND, Y. et al. Using and Combining Predictors That Specialize. In: Annual ACM Symposium on Theory of Computing, 29., El Paso, Texas, USA. **Proceedings...** 1997, p. 334-343.

HERTZ, G.Z.; STORMO, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. **Bioinformatics**, v.15, n.7-8, p.563-77, 1999.

HUERTA, A. M. et al. RegulonDB: A Database on Transcription Regulation in Escherichia coli. **Nucleic Acids Research**, v. 26, p. 55-60, 1998.

FU, K. S. **Syntactic pattern recognition and applications**. Nova Jersey: Prentice Hall, 1982.

ROSENBLUETH, D.A. et al. Syntactic recognition of regulatory regions in Escherichia coli. **Computer Applications in the Bioscience**, v.12, p. 415-422, 1996.

STERGOS, A. D. On Grammars – The Chomsky Hierarchy and Probabilistic Grammars. Disponível em http://citeseer.ist.psu.edu/443342.html. Acesso em: 11 de out. de 2007.

ULLMAN, J. D.; HOPCROFT, R. **Introduction to Automata Theory, Languages, and Computation**. Addison-Wesley, 2001.

Sobre os autores

Elias Bareinboim

Possui grau de bacharel em Ciência da Computação pelo Instituto de Matemática/ Universidade Federal do Rio de Janeiro (UFRJ) e de mestre em Engenharia de Sistemas e Computação pela COPPE/Universidade Federal do Rio de Janeiro (UFRJ). Durante seu mestrado trabalhou na modelagem de propriedades teóricas e com simulações computacionais sobre Sistemas Complexos, tendo como principal resultado a caracterização e interpretação de uma importante propriedade das redes complexas. Tem experiência na área de Ciência da Computação, com ênfase em modelagem de Sistemas Complexos e Inteligência Artificial. Entre as suas atividades, desenvolveu softwares e recebeu quatro prêmios, sendo dois de iniciação cientifica. Seus interesses atuais incluem Sistemas Complexos, Inteligência Artificial e Bioinformática. Atualmente colabora com o Laboratório de Bioinformática no Laboratório Nacional de Computação Científica (LNCC) e o Programa Engenharia de Sistemas e Computação da COPPE/UFRJ.

Ana Tereza Ribeiro de Vasconcelos

Possui graduação em Ciências Biológicas pela Universidade do Estado do Rio de Janeiro (1983), mestrado em Ciências Biológicas (Biofísica) pela Universidade Federal do Rio de Janeiro (1995) e doutorado em Ciências Biológicas (Genética) pela Universidade Federal do Rio de Janeiro (2000). Atualmente é pesquisadora e orientadora do Laboratório Nacional de Computação Científica e credenciada como orientadora na Universidade Federal do Rio de Janeiro. Tem experiência na área de Genética, com ênfase em Bioinformática, atuando principalmente nos seguintes temas: genoma, bioinformática, bioinformática, bactérias e anotação de genomas. Foi presidente da Associação Brasileira de bioinformática e Biologia Computacional -AB3C.