제한 논리 프로그래밍 언어에서 DCG를 이용한 생물학적 서열의 구조 검색†

이근우[0*], 이수현*, 이명준**
*창원대학교 전자계산학과,
**울산대학교 컴퓨터정보통신공학부
*hermes@pl.changwon.ac.kr

Structure Searching of Biological Sequence using DCG in Constraint Logic Programming Language

Geun-Woo Lee[0*], Su-Hyun Lee*, Myung-Joon Lee**
*Dept. of Computer Science, Changwon National University
**School of Computer Engineering & Information Technology, University of Ulsan

요 약

생물학적 서열의 구조 검색은 생물학적 특성을 예측하는데 많은 도움을 주며, 서열에서 나타나는 구조의 패턴은 촘스키의 형식 언어로 기술 가능하다. 본 논문에서는 문맥무관문법의 확장된 표기법인 DCG를 이용하여 구조 검색을 위한 구조 패턴의 생성 규칙을 정의하였다. 또한 구조 검색의 효율향상을 위하여 구조와 관련한 제한(constraint)을 정의하였고 이를 제한 논리 프로그래밍 언어로 구현하였다. 구현된 구조 검색 엔진은 웹 인터페이스를 통하여 접근할 수 있다.

1. 서 론

DNA, RNA, 단백질 등 생물학적 서열 정보의 많은 데이터베이스들이 구축됨에 따라 효과적인 서열 정보의 분석 기술이 중요하게 되었다. 생물학적 서열은 진화적인 관계나 물리 화학적인 제약 때문에 두 서열 사이의 유사성 또는 구조적 유사성은 생물학적 특성을 예측하는데 많은 도움을 준다.

RNA, DNA 서열은 a, c, g, u(또는 t)의 4개의 염기, 단백질 서열은 20개의 아미노산으로 쓰여진 기호 문자열이다. 기호 문자열은 생물학적 의미를 가지고 있으며, 특별한 의미를 갖는 구조로 이루어져 있다. 그리고 구조의 생성 규칙은 촘스키의 형식 언어로 기술 가능하다[1]. DCG(Definite-Clause Grammar)는 형식 문법의 하나인 문맥무관문법(CFG)의 확장된 표기법으로 CFG의 생성규칙을 그대로 사용할 수 있다. 생물학적 서열에서 나타나는 stem-loop 같은 구조 또는 palindrome, repeat 같은 특별한 순서 패턴은 CFG의 생성 규칙으로 표현될 수 있고 푸시-다운 오토마타에 의해 해결이 가능하다. 그리고 DCG의 표현은 논리언어인 Prolog 문법 표현으로 쉽게 변환이 가능하고 Prolog 문법보다 이해하기 쉬운 형태로 이루어져 있어 많은 Prolog 언어에서 제공하고 있다.

본 논문에서는 하나의 기호 문자열인 생물학적 서열의 구조 또는 특별한 순서 패턴의 생성 규칙을 제한(constraint)을 포함하는 DCG로 정의하였다. 그리고 이것을 이용해 제한 논리 프로그래밍 언어인 GNU-prolog[2]에서 생물학적 서열의 구조 검색 엔진을 구현하였다.

2. DCG와 제한 논리 프로그래밍

생물학적 서열 정보는 촘스키의 형식 언어로 기술할 수 있다. 촘스키의 형식 언어는 정규표현, 문맥무관문법, 문맥민감문법, 무제한문법 순으로 4 단계로 구분되고 각 단계는 이전 단계의 생성규칙을 포함한다. 생물학적 서열 정보에서 repeat 같은 패턴은 정규 표현으로 나타낼 수 있고, *palindrome*, *stem-loop* 등은 CFG로 나타낼 수 있다. 그리고 pseudo-knot은 CSG로 표현이 가능하다. 형식 언어를 사용한 예로 단백질 Motif 검색 데이터베이스로 유명한 PROSITE는 정규표현을 이용하고 있다[3].

DCG는 CFG의 확장된 표기법으로 자연어 처리에 많이 사용되었고 생물학 분야에서 D. Searls에 의해 DNA 서열의 구조를 나타내는 문법으로 사용되었다[4].

생물학적 서열에서 CFG에 의해 정의될 수 있는 stem-loop, palindrome, repeat 등은 DCG로 생성 규칙을 표현할 수 있다. 다음의 문법은 palindrome 구조를 가지는 임의의 RNA 서열 "aggcuucgga"의 생성규칙을 CFG와 DCG로 표현한 예이다. (CFG에서 비단말 기호는 대문자로 표기하고 단말기호는 소문자로 표기한다. DCG의 생성규칙

† 본 연구는 한국과학재단 목적기초연구(R01-2001-00535) 지원으로 수행되었음.

은 "-->" 표기를 사용하고 단말 기호는 리스트 형태로 나타낸다.)

CFG	DCG
S → aSa | cSc | gSg | uSu | λ	S --> [a], S, [a] | [c], S, [c] | [g], S, [g] | [u], S, [u] | []

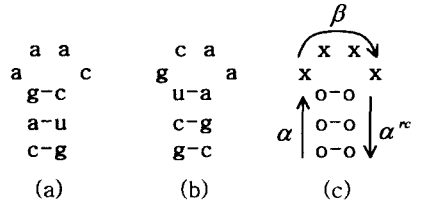
제한 논리 프로그래밍은 두 가지 선언적 패러다임인 논리 프로그래밍과 제한 풀이(constraint solving)의 결합이다[5]. 제한 논리 프로그래밍은 기존 논리 프로그래밍의 " 생성 후 검사(generate and test)" 알고리즘의 수행력을 향상시키기 위해 제한 기술을 도입하여 " 제한 후 생성(constraint and generate)" 알고리즘을 제시하였다. " 제한 후 생성" 알고리즘은 논리 프로그래밍 수행에 핵심적인 탐색 방법에 있어서 탐색할 도메인을 제한 풀이를 통해 줄임으로써 보다 효율적으로 수행이 가능하도록 한다.

논리 프로그래밍 언어의 기본문은 Horn 절의 세가지 형태인 규칙(rule), 질의(query), 사실(fact)로 표현된다. DCG로 생성규칙 표기법은 Horn 절의 한 형태인 규칙(rule)으로 표현 된 것이므로 논리 프로그래밍 언어로 쉽게 변환이 가능하다.

3. 생물학적 서열 정보의 구조 검색

David Gilbert는 제한 논리 언어를 이용한 생물학적 서열의 구조 검색에 관한 연구에서 구조의 패턴 기술을 위한 새로운 표현법을 정의하여 사용하였다[7]. 정의된 표기법은 패턴 기술에는 뛰어나지만 일반적이지 못하다. 따라서 본 논문에서는 형식언어 CFG의 확장된 형태인 DCG를 이용해 구조 패턴의 생성 규칙을 기술하는 방법을 제안하고, 이것을 이용해 RNA 서열에 대한 구조 검색엔진을 구현한다. 검색엔진 구현은 유한 도메인상의 제한을 지원하는 논리 프로그래밍 언어인 GNU Prolog를 사용한다[2].

3.1 생물학적 서열 정보의 구조 패턴



(그림 1) stem-loop 구조

(그림 1)은 RNA의 기본 이차 구조인 stem-loop를 나타낸 것이다. (그림 1)의 (a), (b)는 다른 서열들로 이루어져 있지만 같은 이차 구조로 나타날 수 있다. (a), (b)의 생성규칙은 DCG로 다음과 같이 나타낼 수 있다.

$$stem\_loop --> [a], seq1, [u] | [c], seq1, [g] | [g], seq1, [c] | [u], seq1, [a].$$
$$seq1 --> [a], seq2, [u] | [c], seq2, [g] | [g], seq2, [c] | [u], seq2, [a].$$
$$seq2 --> [a], seq3, [u] | [c], seq3, [g] | [g], seq3, [c] | [u], seq3, [a].$$
$$seq3 --> [g,a,a,a] | [g,c,a,a].$$

Stem-loop의 일반적인 구조는 (c)와 같은 패턴으로 표현 할 수 있다. (c)에서 x는 임의의 염기를 나타내고 o는 서로 상보적인 관계를 가지는 염기를 나타낸다. Stem-loop 구조는 "$\alpha\beta\alpha^{rc}$" 형태의 패턴으로 나타낸다. α, β, α^{rc} 는 각각 서열에 매치되는 서브 스트링을 나타낸 것으로 "$\alpha\beta\alpha^{rc}$"는 세 개의 서브 스트링이 순서대로 연결된 서열을 나타낸다. 그리고 α^{rc}는 α 와 역순으로 상보적 관계를 나타내는 서브 스트링이다.

"$\alpha\beta\alpha^{rc}$" 패턴으로 나타나는 stem-loop 구조의 생성규칙은 DCG로 다음과 같이 나타낼 수 있다.

$$stem\_loop --> seq(A), seq(B), seq\_complement(C), \{reverse(A,C)\}.$$
$$seq([a|X]) --> [a], seq(X).$$
$$seq([c|X]) --> [c], seq(X).$$
$$seq([g|X]) --> [g], seq(X).$$
$$seq([u|X]) --> [u], seq(X).$$
$$seq([]) --> [].$$
$$seq\_complement([a|X]) --> [u], seq\_complement(X).$$
$$seq\_complement([c|X]) --> [g], seq\_complement(X).$$
$$seq\_complement([g|X]) --> [c], seq\_complement(X).$$
$$seq\_complement([u|X]) --> [a], seq\_complement(X).$$
$$seq\_complement([]) --> [].$$

생물학적 서열의 여러 구조는 위의 DCG와 유사한 방법으로 나타낼 수 있다. 다음은 서열의 구조에 나타날 수 있는 패턴의 예이다[6].

구조 종류	패턴형태	예
Tandem repeat	$\alpha\alpha$	acg acg
Simple repeat	$\alpha\beta\alpha$	acg aa acg
Multiple repeat	$\alpha\beta\alpha\beta_1\alpha$	acg aa acg uu acg
Stem-loop	$\alpha\beta\alpha^{rc}$	acg aa cgu
Palindrome,even	$\alpha\alpha^{r}$	acg gca
Palindrome,odd	$\alpha x\alpha^{r}$	Acg a gca
Pseudoknot	$\alpha_1\beta\alpha_2\beta_1\alpha_1^{rc}$ $\beta_2\alpha_1^{rc}$	acg aa ucg gc cgu aua aga

3.2 구조 패턴의 제한

DCG로 정의된 RNA 서열의 구조 패턴은 구조에 대한 length, distance, contents, position, correlation와 같은

5가지의 제한을 포함할 수 있다.

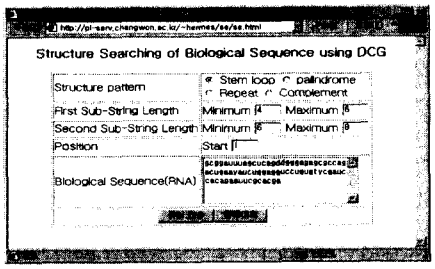
- **length constrint**: 서브 스트링들의 길이 및 최소, 최대 길이 등을 제한. 예, max_length(*S*, *n*)은 스트링 *S*의 최대 길이를 *n*까지로 제한한다.
- **distance constraint**: 두 서브 스트링 사이의 거리 제한. 예, distence(*S1*, *S2*, *n*)은 두 서브 스트링 *S1*과 *S2*사이의 거리를 *n*으로 제한한다.
- **Contents**: 서브 스트링의 특정 위치에 나타나는 염기를 제한. 예, Contents(*S*, *n*, *c*)은 *S*의 *n*번째 염기를 *c*에 나타난 염기로 제한한다.
- **Position**: 서브 스트링이 나타나는 위치의 범위를 제한. 예, start_position(*S*, *n*)은 *S*의 시작 위치를 제한한다.
- **Correlation**: 두 서브 스트링 사이의 관계를 제한. 일치, 역순, 염기의 상보 관계 등을 제한할 수 있다. 예, reverse(*S1*, *S2*)는 *S1*과 *S2*사이의 관계가 서로 역순으로 나타나도록 제한한다.

구조에 대한 제한은 GNU Prolog에서 제공하는 기본 술어(predicate)와 유한 산술 도메인상의 제한 풀이를 위한 술어 그리고 DCG를 함께 사용하여 구현된다. stem의 길이가 4 이상이고 loop의 길이가 5와 10 사이인 stem-loop 구조는 다음과 같이 나타낼 수 있다.

```
max_length(S, N) - - > { L# =< N, length(S, L) }.
min_length(S, N) - - > { L# >= N, length(S, L) }.
stem _ loop - - > seq(A), seq(B), seq _ complement(C)
                ,{reverse(A,C)}
                ,min_length(A,5)
                ,max_length(B,10), min_length(B,5).
```

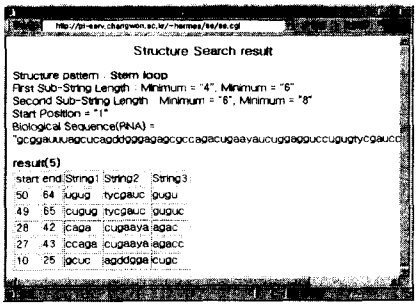
3.3 구조 검색 엔진

다음은 RNA 서열에 대하여 stem-loop, palindrome, repeate, complement(inverse) 구조의 검색 엔진을 구현한 것이다. (그림 2)와 (그림 3)은 구조검색을 위한 질의를 입력하는 화면과 결과 보여주는 화면이다.



(그림 2) 구조 검색을 위한 Web 인터페이스

(그림 2)에서는 입력한 RNA 서열에 대해 stem의 길이가 4와 6사이이고 loop의 길이가 6과 8사이인 stem-loop 구조의 검색에 대한 질의를 입력하였다. (그림 3)은 (그림 2)에 대한 질의 결과 5개의 stem-loop가 검색된 것을 보여준다.



(그림 3) stem-loop 구조의 검색 결과

4. 결론 및 향후 연구방향

본 논문은 생물학적 서열을 형식언어 CFG의 확장된 표기법인 DCG를 사용하여 기호 문자열의 생성 규칙을 기술하는 방법으로 구조 검색이 가능하도록 하였다. 그리고 구조의 특징을 제한 기술을 이용해 제한 함으로써 효율적 검색이 가능하도록 하였다. 그러나 구조 제한에 있어서 완전한 제한 풀이 기술의 특성을 이용하지 못하였고, 많은 생물학적 특성을 고려하지 않고 문자 기호로써의 일부 특성만을 이용하였다. 따라서 향후 서열 구조에 대한 제한 기술의 체계적 연구와 실제 생물학적 서열의 구조 분석에 적용하기 위한 연구가 필요하다.

참고문헌

[1] D. Searls, "The Computational Linguistics of biological sequence," In Larry Hunter, editor, *Artificial Intelligence and Molecular Biology*, chapter 2, pp 47-120, AAAI Press, 1993.

[2] D. Diaz, *A Native Prolog Compiler with Constraint Solving over Finite Domains Edition 1.4, for GNU Prolog 1.2.1*, 2000.

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis*, Cambridge university press, 1998.

[4] D. Searls, "Investigating the linguistics of DNA with definite clause grammars," *Logic Programming: Proceedings of the North American Conference*, pp198-208, 1989.

[5] 신동하, 창병모, " 제한 논리 프로그래밍언어," *정보과학회지* 제15권, 제1호, pp.29-35,1997.

[6] I. Eidhammer, D. Gilbert, I. Jonassen, M. Ratnayake, and S. H. Grindhaug, " *A Constraint Based Structure Description Language for Biosequences*," In submitted to CP97, 1997.

[7] D. Gilbert, I. Eidhammer, and I. Jonassen, " StructWeb: Biosequence structure searching on the Web using clp(FD)," *Workshop on Constraint Reasoning on the Internet*, 1997.