Kor. J. Appl. Microbiol. Biotechnol.
Vol. 24, No. 3, 371-375 (1996)

TFSCAN 검색 프로그램 TFSCAN의 개발

이병욱[1,2] · 박기정[1] · 김기봉[1,2] · 박 완[2] · 박용하[1]*
[1]한국과학기술연구원 생명공학연구소 유전자은행, [2]경북대학교 미생물학과

Development of TFSCAN as a Program to Search for DNA Signals of TFSCAN. **Byung-Uk Lee[1,2], Kie-Jung Park[1], Ki-Bong Kim[1,2], Wan Park[2] and Yong-Ha Park[1]*.** *[1]KCTC, Korea Research Institute Of Bioscience and Biotechnology, KIST, P.O. Box 115, Yusong, Taejon 305-600, Korea, [2]Department of Microbiology, Kyungpook National University, Taegu 702-701, Korea* – TFD is a transcription factor database which consists of short functional DNA sequences called as signals and their references. SIGNAL SCAN, developed by Dan S. Prestridge, is used to determine what signals of TFD may exist in a DNA sequence. This program searches TFD database by using a simple algorithm for character string comparison. We developed TFSCAN that aims at searching for signals in an input DNA sequence more efficently than SIGNAL SCAN. Our algorithms consist of two parts, one constructs an automata by scanning sequences of TFD, the other searches for signals through this automata. Searching for signal-related references is radically improved in time by using an indexing method. Usage of TFSCAN is very simple and its output is obvious. We developed and installed a TFSCAN input form and a CGI program in GINet Web server, to use TFSCAN. The algorithm applying automata showed drastical results in improvement of computing time. This approach may apply to recognizing several biological patterns. We have been developing our algorithm to optimize the automata and to search more sensitively for signals.

진핵 생물의 유전자 조절에 관한 연구가 활발해지면서, transcription에 연관된 많은 sequence 정보들이 알려지게 되었는데, 이 중 기능이 알려진 짧은 DNA sequence들을 'signal'이라 하며, 이들은 promoter나 enhancer처럼 DNA 상에서 주로 단백질 결합 부위의 역할을 한다. 미국의 NIH에서는 이들 signal들과 그와 관련된 문헌들의 자료를 GENBANK 데이타베이스(database)에서 추출하여 TFD(Transcription Factor Database)라는 데이타베이스를 구축하였다(1-3). 이 데이타베이스는 정기적으로 갱신되며, 현재 TFD 7.4(1995년 1월 9일 갱신)가 가장 최근 버전이다. TFD는 연구자들이 사용하기 용이하도록 signal sequence를 대상 생물체 별(포유류, 양서류, 조류, 곤충류, 식물, 효모, 그 외의 진핵생물 그리고 원핵생물)로 분류하여 각각을 화일로 구성하였다.

현재까지 DNA sequence를 TFD와 검색하여 signal sequence를 찾는데 가장 많이 사용되는 프로그램으로는 Dan S. Prestidge가 개발한 SIGNAL SCAN이 알려져 있는데(현재의 버젼은 4.0), 이 프로그램은 간단한 문자열 비교 알고리즘(algorithm)을 사용하여 signal 부위를 찾아낸다(4).

전산학에서의 automata 이론은(5) 여러 수준(level)의 언어나 패턴(pattern)을 인식하기 위한 자동 인식 기계(소프트웨어적인)를 제작하고, 응용하기 위한 이론을 다루는 것으로, 성격이 정의된 패턴의 인식에서 광범위하게 활용되고 있다. 생물학 sequence 패턴은 소수의 구성 원소로 이루어져 있어 이러한 automata를 사용하여 시간적으로 매우 효율적인 결과를 얻을 수 있다(6).

본 연구팀에서는 계산상 보다 효율적인 검색을 위해서, TFD의 DNA signal을 검색하기 위한 automata를 구성하고, 이 automata에 따라서 입력된 DNA sequence에서 signal을 검색하는 알고리즘을 개발하여, TFSCAN이라는 프로그램으로 구현하였다. 이 프로그램에 사용되는 입력 sequence는 fasta 양식(format)이나 sequence로만으로 구성된 화일이면 되고, 염기는 대·소문자 모두 처리되도록 하였다. SIGNAL SCAN에서는 signal에 관한 문헌 정보를 모두 보기 위해 참고 문헌 번호를 입력하여 이를 검색하는 과정을 반복해야 하는 불편한 점이 있는데, TFSCAN에서는 TFD 검색 후 관련 문헌을 모두 출력하도록 하였고, 검색 결과도 실용적인 양식으로 일목요연하게 보여주도록 하였다. 문헌 검색의 경우, SIGNAL SCAN에서는 문헌 화일을 직렬 검색(serial search)하는 반면, TFSCAN에서는 인덱싱에 의해 검색할 수 있는 알고리즘을 채택하여 검색 시간을 개선하였다. World Wide Web을 통해서도 이 프로그램을 활용할 수 있도록 생명공학연구소의 GINet(Genome Information Network of Korea) Web 서버인(7) grcsys1에 TFSCAN 입력용 form과 CGI(Common Gateway Interface) 프로그램을 개발·설치하였다(8). 데이타베이스의 크기에 대한 signal 검색을 위한 문자 비교 횟수를 조사한 결과, 크기가 증가할수록 TFSCAN의 비교 횟

*Corresponding author.
Key words: TFD, automata, TFSCAN, WWW, CGI program, SIGNAL SCAN

수가 SIGNAL SCAN에 비해 급격히 감소함을 볼 수 있었다. 현재 이 automata를 최적화하여 이 비교 횟수를 더욱 급격히 감소시키는 알고리즘을 시험하고 있으며, 보다 민감한(sensitive) 검색을 위해, signal과 일정 퍼센트 내의 상동성을 보이는 sequence도 검색하는 automata를 구성하고 이를 구동하는 알고리즘도 고안하고 있다.

재료 및 방법

TFSCAN 환경

TFSCAN은 C언어로 짜여진 3개의 소스 코드로 구성되었으며, 생명공학연구소의 UNIX 워크스테이션인 grcsys1(INDIGO 2)에서 개발하였다. 본 프로그램은 표준 C 라이브러리만을 사용하였으므로 UNIX를 운영체계로 사용하는 컴퓨터로 이식이 용이하며, 일반 PC에서도 LAN을 통해 사용할 수 있다.

프로그램 구성

DNA sequence는 A, C, G, T의 4개의 문자로만 구성된 매우 단순한 구조의 문자열이기 때문에 signal sequence는 서로 중복 부분을 많이 가진다. 이런 구조의 패턴은 automata를 만들어(Fig. 1) 검색할 경우, 검색 빈도가 급격히 감소한다. 본 연구의 automata 관련 알고리즘은, automata를 구성하는 알고리즘과 구성된 automata를 구동하여 검색하는 알고리즘으로 이루어지며, 이들을 각각 'autotf' 프로그램과 TFSCAN 프로그램으로 구현하였다. 또, TFSCAN에서 문헌 검색을 하기 위한 자료구조를 만들기 위해, 문헌 데이타베이스를 변화하는 프로그램으로 'conref'를 구현하였다.

TFD의 seq	automata 구조
aacgt accgg acctg	a → c → g → t (1) a → c → c → g → g (2) a → c → c → g → t → g (3)
cgat cgatgtaa	c → g → a → t (4) c → g → a → t → g → t → a → a (5)
ggatc ggta	g → g → a → t → c (6) g → g → a → t → a (7)
tacacg tagaaaaa	t → a → c → a → c → g (8) t → a → g → a → a → a → a → a → a (9)

Fig. 1. Automata to recognize TFD signal patterns. Each letter in automata means a state to which the previous state should transfer with the letter, and each number in parenthesis means the index of coresponding pattern which each final state accepts.

Automata 구성 알고리즘

'autotf'는 TFD에서 sequence만을 인지하여 이들로부터 sequence 패턴을 나타내는 automata를 구성한 다음, sequence와 automata 전체를 화일로 저장하며, TFSCAN은 이 automata 화일을 읽어들인 후, 입력 DNA sequence에 따라 automata를 구동해 가면서 signal이 있으면 저장하여 최종단계에서 출력한다. Fig. 1의 automata는 염기로 표시된 state와 화살표로 표시된 state 이동으로 이루어지며 다음과 같이 구성되었다.

i) TFD에서 sequence를 하나씩 읽는다.

ii) 각 sequence를 한 base씩 읽어, 그 염기에 따라 이동할 state를 찾는다. 이동할 state가 없으면 그 염기를 위한 새로운 state를 만든다. 즉 'accgg'에서는 'c-c-g-g'로 4개의 state를 만들고, 'acctg'에서는 't-g'로 2개의 state를 만든다.

iii) 각 패턴의 마지막 염기에 해당하는 state는 그 패턴의 accepting state가 되며, 여기에는 그 패턴의 인덱스(index)를 기록한다.

iv) 모든 패턴에 대해 i)에서 iii)을 수행한다.

v) TFD의 sequence와 i)에서 iv)로 만들어진 각 state들을 화일에 저장한다. 각 state는 state의 인덱스와 accept되는 패턴에 대한 인덱스, 그리고 각 염기에 의해 이동되는 다음 state의 인덱스들을 포함하고 있다. ii)에서 패턴에 나타난 불확실한 염기의 경우, TFD가 그 염기를 나타내기 위해서 사용하는 IUB-IUPAC 심볼에서(9) N을 제외한 나머지 염기를 모두 A, C, G, T로 변환하여 복수 개의 패턴으로 처리하였다. N은 transcription factor의 작용 부위가 아닌 중간 부위로 몇 개씩 연속적으로 나타나는 경우가 많으므로, N을 4개의 염기로 전환한 다음 automata를 구성할 경우, automata의 state 수가 급격히 증가한다. 따라서 N은 복수 개로 변환하지 않고, automata 상에서 특별한 염기로 처리하였다. Automata 구동에서도 N은 별도로 처리하였다.

Automata 구동 알고리즘

입력 DNA sequence를 읽어들여 상보적인 가닥을 만든 후, 이 두 가닥(strand)에 대해 아래와 같이 automata를 따라 가면서 accepting state가 있으면, 그 인덱스와 sequence상의 위치를 저장한다.

i) 입력 sequence의 각 위치를 시작 위치로 하여 automata에서 각 염기에 따라 이동할 수 있는 state로 이동한다. accepting state에 도달하지 않고 이동이 중단되면 중지한다.

ii) accepting state에 도달하면, 그 state에 대한 패턴 인덱스와 입력 sequence상의 시작 위치를 저장한다.

iii) accepting state 이후에도 이동할 state가 있으면 계속 이동한다.

iv) 이동 state가 없을 때까지 ii)에서 iii)을 계속한다.

v) 입력 sequence에서 3′ 방향으로 시작 위치를 한

염기 옮겨 i)에서 iv)를 수행한다.

vi) 상보 가닥 sequence에 대해서도 i)에서 v)를 수행한다.

vii) 입력 sequence와 상보 sequence에 대해 signal이 시작되는 위치에 각 signal의 이름을 출력한다. 발견된 signal에 관한 sequence 정보들을 보여주고, 관련 문헌도 검색하여 출력한다. State 이동에서 N의 경우는 A, C, G, T와 별도로 처리한다. 즉, 현재 state에서 N에 의한 이동이 있으면 특정 염기에 대한 state 이름과 N에 따른 state 이동을 동시에 수행한다. 따라서, 동시에 2개 이상의 이동 state를 유지해야 하는 경우가 발생하는데, 이는 리스트로 관리한다.

관련 문헌 검색

SIGNAL SCAN에서 수행되는 문헌 검색 질의에서는, 문헌 데이타베이스 화일을 읽어가면서 문자열 비교에 의해 화일의 처음부터 직렬 검색(serial search)하고, 한 번의 질의에 한 개의 질의 문헌 번호만 입력 가능하도록 되어 있다. 본 프로그램 중 'conref'는 TFD의 문헌 데이타베이스 화일을 읽어 문헌 번호의 일련 번호에 따라 1번부터 고정크기(2000)로 각 문헌 정보를 저장하여 화일을 구성하고, TFSCAN은 문헌 번호만으로 해당 문헌이 있는 위치를 계산해서 직접 접근할 수 있다. 따라서, TFSCAN의 문헌 검색은 확인된 signal sequence에 해당하는 문헌 번호를 모두 저장한 후, 각 문헌을 직접 추출함으로써 검색 시간을 대폭 개선하였다.

Web 서버에의 설치

GINet Web 서버인 grcsys1의 응용 프로그램 란에 TFSCAN을 연결하고, Fig. 2와 같은 TFSCAN 제공용 form을 작성하였다. 사용자가 입력한 sequence를 입력 화일로 하고 signal class를 표준 입력으로 하여 TFSCAN을 실행한 뒤, 실행 결과를 표준 출력으로 Web 서버에 보내주도록, CGI 프로그램인 'tfscan-srch'를 구현하였다.

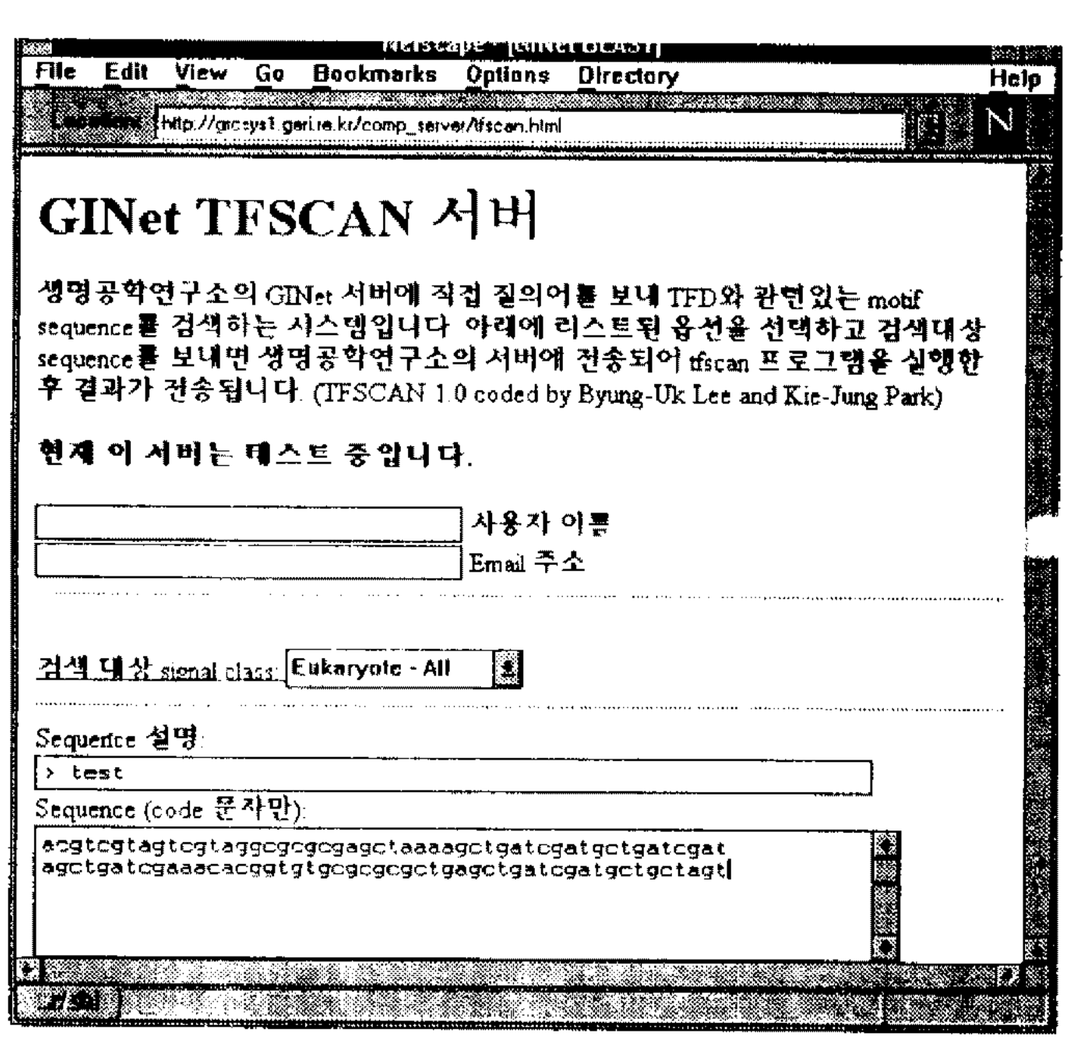

Fig. 2. TFSCAN in GLNet Web server.

결과 및 고찰

입력 sequence를 SIGNAL SCAN과 TFSCAN 프로그램으로 실행시킨 검색 결과는 서로 일치하지만, 두 프로그램의 검색 알고리즘 차이로 인해 계산 상의 속도는 현격한 차이를 나타내었다. 크기가 2000인 전체 TFD를 대상으로, TFSCAN은 한 염기당 약 30번의 문자 비교를 하는 반면, SIGNAL SCAN은 약 1484번의 비교를 하는 것으로 나타났다(Table 1). TFSCAN에서 automata를 읽으면서 이동하는 state 하나에 대한 부담은 SIGNAL SCAN의 문자 비교의 2번 내지 3번에 해당하므로, TFSCAN과 SIGNAL SCAN의 한 염기당 비교 회수에서 20배 정도 차이가 나타났다. 이러한 결과는, 패턴의 중복성으로 인한 오토메타의 비교 횟수 감소와, 일치하지 않는 문자에 대해 문자 비교 자체를 피하는 (일치하는 문자에 대한 state로만 이동함으로) automata의 특성에 기인한 것이다. 따라서, 검색 데이타베이스의 크기가 커질수록 중복성이 커지므로 그 차이는 더욱 증가하게 된다. 실제로 패턴 데이터베이스의 크기가 증가할수록 TFSCAN의 효율이 증가하는 것을

Table 1. The number of character comparisons for sequences of several size and a TFD of fixed size (2000): the numbers in parenthesis are average comparison numbers for a base in TFSCAN and in SIGNAL SCAN.

sequence size	TFSCAN	SIGNAL SCAN
50	1414 (28.28)	72366 (1447.32)
100	3184 (31.84)	152524 (1525.24)
200	5868 (29.34)	293245 (1466.26)
2500	79810 (31.92)	3724183 (1489.67)
10000	319856 (31.98)	14884574 (1488.46)

Table 2. The number of character comparisons for a sequence of fixed size (2500) and TFDs of several size in TFSCAN and in SIGNAL SCAN.

database size	TFSCAN	SIGNAL SCAN	(SIGNAL SCAN/ TFSCAN)
500	41458	904734	21.83
1000	49498	1900158	38.39
2000	79810	3721346	46.63

```
1  AGCGCTTGTGACTGCACGGCTAGAGTACGGTGACTGCACGGCTAGAGTCA
          *                    *          *
          HIS4 US              HIS4 US    GCN4-ILV1.1
                                          GCN4-HIS3.2

     >>>>>>>>>>>>>> Complementary strand <<<<<<<<<<<<<<

1  TGACTCTAGCCGTGCAGTCACCGTACTCTAGCCGTGCAGTCACAAGCGCT
       *        *          *        *
       LBP-1 RS  GCN4-HIS4.3 LBP-1 RS  GCN4-HIS4.3


     ************************************************************
                          SEQUENCE REFERENCE
     ************************************************************

  factor       site                    signal sequence              journal
-------------------------------------------------------------------------------
LBP-1          [LBP-1 RS               ] : WCTRG                    (S00487 )
GCN4           [GCN4-HIS3.2            ] : GAGTCA                   (S00724 )
GCN4           [GCN4-HIS4.3            ] : CAGTCA                   (S00732 )
GCN4           [GCN4-ILV1.1            ] : GAGTCA                   (S00806 )
unknown        [HIS4 US                ] : TGACT                    (S01125 )


     ************************************************************
                          JOURNAL REFERENCE
     ************************************************************

ID #    FACTOR          SITE              JOURNAL REFERENCE
-------------------------------------------------------------------------------
S00487  LBP-1           LBP-1 RS          Genes Dev 2: 1101-14 (1988)
S00724  GCN4            GCN4-HIS3.2       Proc Natl Acad Sci U S A 83: 8516-20 (1986)
S00732  GCN4            GCN4-HIS4.3       Proc Natl Acad Sci U S A 83: 8516-20 (1986)
S00806  GCN4            GCN4-ILV1.1       Proc Natl Acad Sci U S A 83: 8516-20 (1986)
S01125  unknown         HIS4 US           Mol Cell Biol 4: 1326-33 (1984)
```

Fig. 3. An output of TFSCAN program.

다양한 크기의 TFD에 대한 실행에서 볼 수 있었다(Table 2). 패턴을 구성하는 각 염기의 자리당 출현 빈도를 균일하게 보고 패턴 길이에 대한 분포를 조사하면, SIGNAL SCAN과 TFSCAN 각각에 대해 문자 비교 횟수의 평균을 확률적으로 구할 수 있을 것이며, 데이터베이스 크기와 sequence 크기에 대한 두 알고리즘의 계산 효율성(computational complexity)을 구할 수 있을 것이다.

TFSCAN의 직접 사용 명령어는 'tfscan 〈input_file〉 [output_file]'로 간단하게 구성하였고, 결과는 화면으로 출력하거나 출력화일로 저장할 수 있으며, 내용은 크게 세 부분으로 나누었다(Fig. 3). 입력 sequence와 상보 sequence를 나열하여 그 밑 부분에 해당 signal을 나타내고, 그 signal에 대한 sequence를 참조하도록 sequence reference 부분을 넣었으며, 그 signal이 발표된 문헌에 대한 정보도 마지막 부분에 출력하여 사용자가 원하는 signal에 대한 정보를 모두 보여주도록 하였다.

TFSCAN은 프로그램 수행의 모든 명령을 표준 입력으로부터 받고, 결과는 표출 출력으로 나타내므로, World Wide Web에 연결하기 위한 CGI 프로그램을 용이하게 구현할 수 있었다. 이를 통해 TFSCAN은 생명공학연구소 GINet Web 서버의 생물학 관련 응용 소프트웨어 부분(http://grcsys1.geri.re.kr/comp_server/TFSCAN.html)에서 수행된다.

Automata 구성에서 signal sequence에 불확실한 염기를 나타내는 IUB-IUPAC 심볼이 연이어 많이 나오면, automata state의 수가 프로그램의 허용 범위를 넘어설 수 있다. TFD의 경우, mammal.dat 화일에 'LOB RS'가 'GCWWWWWWWWWWWWWWWWWG'의 signal sequence를 가져(W가 연속하여 17개) 2^{17}개의 state를 가지므로, automata 형성에 장애가 된다. N을 제외한 IUB-IUPAC 심볼이 연속되는 경우는 거의 발생하지 않지만, 이러한 경우가 다수 발생하면 예외적으로 검색하도록 처리할 수 있을 것이다. 실제로 TFD에서는 앞의 한 경우만 이에 해당하여 이를 automata 구성에서 제외하였다.

TFD와 같은 생물학 패턴 데이타베이스를 검색하는 방법으로 automata를 이용하면, 본 연구에서와 같이 수행 시간에서 매우 효과적인 결과를 얻을 수 있을 것이며, 특히 그 크기가 증가하는 경우 활용도가 더욱 커질 것이다. Table 2의 결과는 데이타베이스 크기 증가에 따른 automata 응용의 효과를 보여주는 것으로, 크기가 급격히 증가하는 생물학 데이타베이스에서도 이러한 검색 방법이 매우 유용하게 사용될 것으로 볼 수 있다. 이 알고리즘을 개선하기 위하여, 패턴의 시작 부위가 다른 패턴의 일부분이 되는 경우를 automata에 반영하기 위한 방법을 고안하여 시험하고 있다. 이 알고리즘은 입력 sequence에 대해 하나의 시작 부위만 설정하는 변칙적인 automata를 사용하는데, 이 알고리즘이 성공적으로 수행되면 계산 시간에서 보다 효율적인 프로그램을 개발할 수 있을 것이며, 이 또한 여러 패턴 데이터베이스의 검색에 활용될 수 있을 것이다.

TFSCAN 1.0은 signal sequence와 완전히 일치하는 입력 sequence만을 검색한다. Signal sequence와 특정 값 이내의 상동성을 보이는 부분도 검색하면서, automata를 응용하여 검색 시간을 줄일 수 있는 알고리즘에 대한 연구를 현재 계속하고 있는데, 이 알고리즘을 이용할 수 있게 되면 보다 민감한(sensitive) signal 검색을 할 수 있고, 기존 signal의 변경과 개선을 위해서도 활용할 수 있을 것이다.

현재 TFSCAN 1.0이 실행되는 곳은 생명공학연구소 grcsys1(IP address: 134.75.131.101)과 GINet Web 서버이다. 여기서는 SIGNAL SCAN도 동시에 지원하고 있으나, TFD의 변경에 따른 update 시의 문제점 등으로, 앞으로 새로운 버전의 TFD는 TFSCAN만으로 지원할 것이다. TFSCAN과 관련 프로그램의 소스 코드는 bulee@grcsys1.geri.re.kr로 요청 메세지를 메일로 보내면 보내 줄 예정이다.

요 약

TFD는 기능이 알려진 짧은 DNA sequence(signal)들과 그와 연관된 저널 자료로 구성된 데이타베이스이다. 임의의 DNA에서 이 데이타베이스의 sequence들을 검색하여 signal을 찾는 프로그램으로 Dan S. Prestridge가 개발한 SIGNAL SCAN이라는 프로그램이 사용

되고 있는데, 이는 간단한 문자열 비교 알고리즘을 사용한다. 본 연구에서는 계산상 보다 효율적인 검색을 위해서, TFD의 sequence를 검색하기 위한 automata를 구성하는 프로그램과, 이 automata에 따라 signal을 검색하도록 하는 TFSCAN이라는 프로그램을 개발하였다. 검색된 signal에 대한 관련 문헌의 검색에서도 인덱싱 방법을 이용하여 계산 속도를 향상시켰다. 프로그램의 사용을 단순화시켰고, 결과 내용을 signal과 관련된 모든 정보를 일목요연하게 보여줄 수 있도록 구성하였다. 이 프로그램을 Web을 통해서도 사용할 수 있도록, GINet Web 서버에 TFSCAN 입력용 form과 CGI 프로그램을 개발·설치하였다.

본 연구의 특정 Motif 패턴으로 구성된 데이타베이스 검색에서, automata를 응용한 알고리즘을 이용하여 계산상 급격히 향상된 결과를 얻을 수 있음을 알 수 있었는데, 이는 생물학의 여러 패턴 검색에 응용될 수 있을 것이다. 더욱 민감한(sensitive) signal 검색을 위해서, 이와 같이 automata를 활용하고, 이 automata를 최적화하는 알고리즘 연구를 계속하고 있다.

참고문헌

1. Ghosh, D. 1990. A relational database of transcription factors. *Nucleic Acids Research* **16**: 1749-1756.
2. Ghosh, D. 1992. TFD: The transcription factor database. *Nucleic Acids Research* **20**: 2091-2093.
3. Ghosh, D. 1993. Status of the transcription factors database (TFD). *Nucleic Acids Research* **21**: 3117-3118.
4. Prestridge, D.S. 1991. SIGNAL SCAN: a computer program that scans DNA sequences for eukaryotic transcriptional elements. *Computer Applications in the Biosciences* 7: 203-206.
5. Denning, P.J., Jack B. Dennis, and Joseph E. Qualitz. 1978. Machines, Languages, and Computation (1-136). Prentice-Hall Inc.
6. Park, K.J., C.K. Park, and Y.H. Park. Application of automata to recognition of biological sequence patterns. in preparation.
7. 이병욱, 박기정, 김기봉, 김혁, 김록희, 김대겸, 박완, 박용하. 1995. GINet Web 서버의 설치 및 운영. *The Microorganisms & Industry* **21**: 383-387.
8. 박기정, 이병욱, 박용하. 1996. 제놈 분석용 계산 Web 서버의 구성. *Kor. J. Appl. Microbiol. Biotechnol.* **24**(1): in press.
9. Cornish-Bowden, A. 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendation 1984. *Nucleic Acids Reasearch* **16**: 3021-3030.

(Received 1 December 1995)