약물 부분 구조 검색을 위한 RS3 시스템의 개선

이환구[O] 차재혁
한양대학교 정보통신대학원
prodog@ihanyang.ac.kr[O], chajh@hanyang.ac.kr

The Improvement of RS3 System for Drug Substructure Searching

Hwangu Lee[O] Jaehyuk Cha
The Graduate School of Information & Communications, Hanyang University

요 약

약물의 화학구조와 약리작용간의 관계는 'Medicinal Chemistry'에서 활발히 연구된다. 이에 도움이 되는 분야로 수많은 약물들에서 사용자가 지정한 구조를 부분구조로 가지는 약물들을 자동으로 빠르게 찾아내는 부분구조검색(Substructure Searching)이 있다. 1950년대부터 연구된 앞의 문제는 NP-Complete이나 미리 인덱스를 두어 성능을 높인 RS3 시스템(http://www.acelrys.com/rs3)이 미국 특허를 받았다. 이 시스템은 화학구조에 대한 설명을 대용량으로 기술하여 이를 RDBMS에 저장하고 검색하는 시스템이다. 하지만 이 시스템은 재현율(Recall)과 정도(Precision)가 매우 낮으므로, 본 논문에서는 새로운 인덱스를 개발하여 재현율과 정도를 향상시킨 기법을 제시한다.

1. 서 론

약물의 화학구조와 약리작용간의 관계는 'Medicinal Chemistry'에서 활발히 연구된다[1]. 약학자는 신약개발시 만들고자 하는 약물과 비슷한 화학구조를 가지고 있는 기존 약물들에는 어떠한 것들이 있는지 알고 싶어한다. 다시 말하면, 연구자나 개발자들은 특정 화학구조가 어떤 약물들에서 나타나는지 신속히 검색하기를 원하고 있다. 따라서 수많은 약물 파일들에서 사용자가 지정한 구조를 부분구조로 가지는 약물 파일들을 자동으로 빠르게 찾아내는 일이 필요하다.

1990년대에 구조를 문자열로 기술하여 이를 RDBMS에 저장한 뒤, 검색시 문자열 검색을 사용하는 RS3 시스템이 미국 특허를 받았고, 현재 상용화되어 사용되어지고 있다. 본 논문은 RS3 시스템을 개선시킨 기법을 제시한다.

2. 관련 연구

2.1 그래프 이론

임의의 화학구조를 부분적으로 가지고 있는 화합물들을 검색하는 것을 'Substructure Searching'이라 하며, 이는 그래프 이론에서 'Subgraph Isomorphism' 문제로 귀결된다. Subgraph Isomorphism 문제는 NP-Complete 문제이다[2].

2.2 다른 접근방법들

NP-Complete인 이 문제에 대해 여러 가지 다른 접근방법들이 있었다. 예를 들면 좀 더 빠른 컴퓨터를 사용하거나, 하드웨어 병렬화 기법을 사용하는 방법으로 CAS[3], Daylight Chemical Information System[4], Synopsys[5] 등이 있다. 또한 검색 후보가 되지 않는 원자들을 제거하는 Algorithm이나 Heuristics, Mapping이 가능한 원자와 결합관계를 줄이거나 정렬시키는 법 등이 연구되었다. 현재 단계에서 비교 불가능하면 바로 전 단계로 거슬러 올라가 비교를 시도하는 Backtracking Algorithm의 일종인 Ullmann Algorithm[6]이 대표적으로, 이는 target atom T_i에 mapping되는 query node Q_i가 이웃노드 Q_j를 가지고 있다면 T_i도 Q_j에 mapping되는 T_j를 가지고 있어야 한다는 조건을 반복적으로 다른 원자들에게 적용함으로써 비교대상 원자수를 줄인다.

하지만 이러한 방법들은 여전히 많은 시간을 요구했기 때문에, 또 다른 방법으로서 시간이 많이 소모되는 연산은 미리 계산하는 스크리닝(Screening) 기법이 개발되었다. 구조를 대표하는 키(Key)를 개발하여 이를 미리 계산하여 저장함으로써 검색시 속도를 개선시킨다. 예로서 BASIS fragment dictionary[7], STN International for its on-line substructure search system[3] 등이 있다.

1980년대 들어 하드디스크 가격이 낮아짐과 맞물려, 각 원자를 중심으로 구조를 상세히 기술하는 대용량의 키를 개발하는 방법이 등장했다. 이를 Atom-centered Indexes라 하며 HTSS(Hierarchical Tree Substructure Search)[8,9], S4 system[10] 등이 있다. HTSS는 각 원자를 정해진 규칙에 따라 분류하여 비교하는 방법이고, S4 system은 각 원자를 중심으로 인접 원자와의 관계들을 bitstring으로 나타내어 비교하는 방법이다.

1990년대에는 구조를 문자열로 기술하여 이를 RDBMS에 저장

한 뒤, 검색시 문자열 검색을 사용하는 RS3 시스템이 미국 특허를 받았다[11]. 이 RS3 시스템이 본 논문의 모태로서, 본 논문은 RS3 시스템의 문제점을 규명하고 개선방안을 모색하였다.

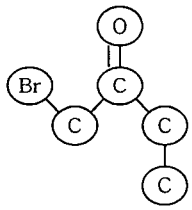
2.3 RS3 시스템

RS3 시스템의 장점은 원자 중심의 인덱스를 하되, 다른 원자와의 결합관계를 문자열로 표현하고 쿼리도 와일드카드(%)를 적절히 포함한 문자열로 표현하여 부분구조검색을 부분문자열검색으로 변환시킨 점이다. 이렇게 함으로써 NP-Complete인 부분구조검색 문제를 O(n)의 시간복잡도로 줄일 수 있다.

아래 [표1]처럼 이웃원자와의 결합관계를 하나의 문자로 정의한다. 예를 들어 Br원자와의 일차결합은 'b'문자로 표시한다. 그 후, 각 원자별로 '.'로 단계를 구분하여 결합구조를 문자열화한다. 이때 문자간의 순위는 알파벳 순서를 따르며, 순서대로 기술되지 않은 부분을 '.'으로 나누며 기술한다.([그림1]참조)

[표1] 결합관계의 축약

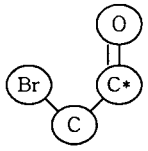
- Br	b
- C	c
= C	d
= O	e
wildcard	%



1. Br c . c . c e . c . . .
2. C b c . . c e . c . . .
3. C c c e . b . c . . . .
4. O d . c c . b . c . . .
5. C c c . . c e . b . . .
6. C c . c . c e . b . . .

[그림1] DB저장구조와 문자열

부분구조를 쿼리하는 경우는 아래 [그림2]와 같은데, 여기서 결합될 수 있는 부분(쿼리하는 부분구조가 전체 화학구조에 붙을 수 있는 곳)을 '*'로 지정해 주어야 한다.



1. Br c . c . % e % . %
2. C b c . . % e % . %
3. C % c % e % . %
4. O d . % c % . %

[그림2] 쿼리부분구조와 문자열

위의 쿼리 문자열을 가지고 DB에 저장된 문자열들과 문자열 검색을 수행하게 되면, 후보가 되는 화학구조파일들을 추릴 수 있다. 그 후, ABAM(Atom-by-Atom Matching)을 수행하여 정확한 결과구조를 출력한다. RS3 시스템은 현재 Accelrys사에서 상용화되었다[12].

2.4 RS3 시스템의 문제점

RS3의 가장 큰 문제점은 '*'로 지정한 곳(부분구조가 연결될 수 있는 곳)이 많아짐에 따라 정도가 급격히 낮아진다는 점이다. [그림2]의 3번째 문자열처럼 C원자에 '*'이 지정되면 'C%c%e%'처럼 곧바로 %가 이어지고, 다음엔 '.%'으로만 쓰여지게 되서 더 이상의 구조기술이 어렵게 된다. 즉, '*'를 만나게 되면 그 이후의 구조들은 기술이 불가능하게 되는 것이다.

또 한가지의 문제점은 동일원자의 동일결합에 대해 우선순위를 정할 수 없다는 것이다. RS3 시스템의 중요한 아이디어는 구조기술시 우선순위를 알파벳 순서로 한다는 점이었다. 예를 들어 [그림1]의 2번째 문자열인 'Cbc..ce.c...'은 C원자에 Br원자와의 1차결합, C원자와의 1차결합이 있어, 이를 나타내는 문자 'b', 'c'를 알파벳 순서대로 나열한 것이다. 하지만 3번째 문자열처럼 같은 'b', 'b'에 대해서는 우선순위를 정할 수 없다. 이후의 구조를 어느 것부터 기술하느냐 하는 것이 명확하지 않음이 RS3 시스템의 재현율을 낮게 하는 이유가 된다.

3. 개선된 RS3 시스템

3.1 해결 방안

이러한 두가지 문제점은 각 원자별로 결합구조를 표현하되, 각 Level(경로의 길이, 깊이)별로 그 Level에 해당하는 원자들의 경로를 기술하여 이를 정렬시켜 저장하는 방법으로 해결할 수 있다. 즉, 그래프의 모든 에지에 가중치를 같게 주고, 각 원자별로 최소비용신장트리를 구성한 다음 모든 원자까지의 경로를 깊이별로 나누어 기술하여 저장하는 것이다. [그림1]의 DB저장구조를 본 기법으로 저장하면 아래 [표2]와 같다.

[표2] 해결방안의 DB저장형태

원자	Level1	Level2	Level3	Level4
Br	c	c.c	c.c.c c.c.e	c.c.c.c
C	bc	c.c c.e	c.c.c	
C	cce	c.b c.c		
O	d	d.c d.c	d.c.b d.c.c	
C	cc	c.c c.e	c.c.b	
C	c	c.c	c.c.c c.c.e	c.c.c.b

마찬가지로 [그림2]의 쿼리구조의 본 기법 기술은 다음 [표3]과 같다.

[표3] 해결방안의 쿼리기술형태

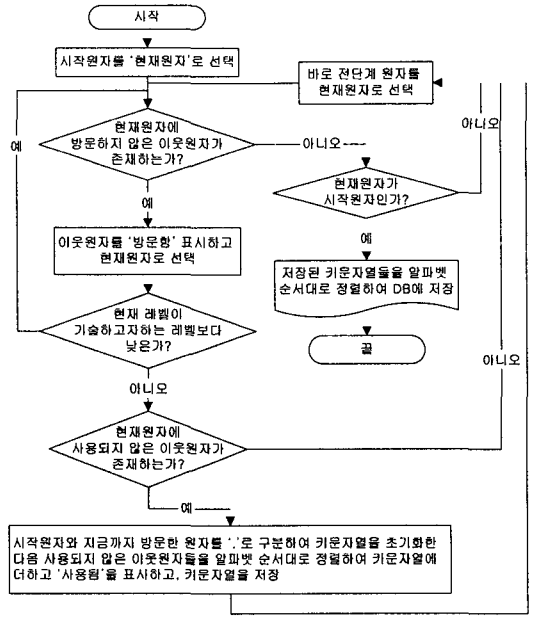
원자	Level1	Level2	Level3	Level4
Br	%c%	%c.%c%	%c.c.%e%	
C	%b%c%	%c.%e%		
C	%c%e%	%c.%b%		
O	%d%	%d.%c%	%d.c.%b%	

이와 같이 기술하면, 연결부분(RS3에서의 '*')을 지정하지 않아도 되고 따라서 모든 구조를 기술할 수 있기 때문에, 정도가 좋아진다. 또한 전체 경로를 기술하고 이를 알파벳 순서로 정렬하기 때문에 동일원자의 동일결합에 대한 우선순위가 명확하여 재현

율이 향상되게 된다.

3.2 해결 방안의 알고리즘 순서도

본 기법의 핵심에 대한 알고리즘 순서도는 아래 [그림3]과 같다.

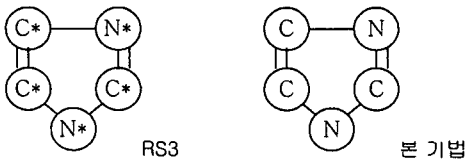


[그림3] 알고리즘 순서도

4. 실험 및 성능 평가

4.1 RS3 시스템과의 성능 비교

258개의 약물파일을 가지고 본 기법과 RS3 시스템의 성능을 비교하였다. 둘 다 동일하게 핵심 엔진에서 나온 결과의 원자종류별 갯수를 쿼리의 원자종류별 갯수와 비교하여 적은 것은 제외시키는 과정을 거쳤다. 즉 다차시간 알고리즘까지는 수행시켰다. 쿼리 구조는 아래 [그림4]와 같고 그 결과는 아래 [표4]와 같다.



[그림4] 쿼리 구조

[표4] 실험 결과

	RS3 (%)	본 기법 (%)
재현율	62.5	100
정도	22.2	94.1

위의 결과에서 알 수 있듯이 본 기법은 재현율과 정도를 크게 향상시킨다.

4.2 본 기법의 재현율 검토

순수 그래프적인 측면에서 보면 본 기법은 사이클의 일부분을 검색하려고 할 때에는 재현율이 100%가 되지 않는다. 하지만 화학구조에서 사이클은 사이클 자체로 의미가 있다. 즉, 사이클과 그 사이클의 일부분과는 화학적 특성이 확연이 달라지게 되므로 사이클의 일부분을 검색하고자 하는 시도는 무의미하다. 이에, 본 기법은 화학 부분 구조 검색에서 100%의 재현율을 달성할 수 있을 것으로 보여진다.

5. 결 론

RS3 시스템은 화학구조를 문자열로 기술하는 방법에서 몇가지 치명적인 단점이 있다. 이를 해결한 본 기법은 좀 더 많은 문자열을 생성하여 저장공간이 증가하는 단점이 있으나, 재현율과 정도는 급격히 향상된다.

REFERENCES

[1] Alfred Burger, A Guide to the Chemical Basis of Drug Design, 1983

[2] R. C. Read and D. G. Corneil, J. Graph Theory, 1, 339-363, 1977

[3] N. Farmer, J. Amoss, W. Farel, J. Fehribach, and C. R. Zeidner, in 'Chemical Structures: The International Language of Chemistry', ed. W. A. Warr, Springer-Verlag, Heidelberg, pp. 283-295, 1988

[4] Daylight Chemical Information Systems, Inc., 27401 Los Altos, Suite 370, Mission Viejo, CA 92691, USA

[5] G. A. Hopkinson, J. Chem. Inf. Comput. Sci., 37, 143-145, 1997

[6] J. R. Ullmann, J. Assoc. Comput. Mach., 23, 31-42, 1976

[7] W. Graf, H. K. Kaindl, H. Kniess, and R. Warszawski, J. Chem. Inf. Comput. Sci., 22, 177-181, 1982

[8] Z. M. Nagy, S. Kozics, T. Veszpremi, and P. Bruck, in 'Chemical Structures: The International Language of Chemistry', ed. W. A. Warr, Springer-Verlag, Heidelberg, pp. 127-130, 1988

[9] Z. M. Nagy, J. Chem. Inf. Comput. Sci., 33, 542-544, 1993

[10] A. Bartmann, H. Maier, D. Walkowiak, B. Roth, and M. G. Hicks, J. Chem. Inf. Comput. Sci., 33, 539-541, 1993

[11] J. Moore and J. R. Hoover, US Patent 5 577 239, 1996

[12] RS3 system, http://www.acelrys.com/rs3