약물 부분 구조 검색을 위한 RS3 시스템의 개선

이현구* 자제석
한양대학교 정보통신대학원
prodog@hanyang.ac.kr, chajh@hanyang.ac.kr

The Improvement of RS3 System for Drug Substructure Searching

Hwang Lee* Jaehyuk Cha
The Graduate School of Information & Communications, Hanyang University

요 약
약물의 화학구조와 약리작용관계의 관계는 'Medicinal Chemistry'에서 활발히 연구된다. 이에 도움이 되는 분야로 수많은 약물들에서 사용자가 지정한 구조를 부분구조로 가진 약물들을 자동으로 찾아내는 부분구조검색(Substructure Searching)이 있다. 1950년대부터 연구된 양의 문제는 NP-Complete이나 미리
인덱스를 두어 성능을 높린 RS3 시스템(http://www.acetys.com/rs3)이 미국 보험국에 사용되고 있다. 이 시스템은
화학구조에 대한 설명을 이용하여 약을 RDBMS에 저장하고 검색하는 시스템이다. 하지만 이 시스템은 재현율(Recall)과 정도(Precision)가 매우 낮으므로, 본 논문에서는 새로운 인덱스를 개발하여 재현률
고 정도를 향상시킨 기법을 제시한다.

1. 서 론
약물의 화학구조와 약리작용관계의 관계는 'Medicinal Chemistry'에서 활발히 연구된다[1]. 약학자는 신약개발시 안
았기 때문에 약물과 비슷한 화학구조를 가지고 있는 기존 약물들
에는 어떠한 것들이 있는지 알고 싶어한다. 다시 말하면, 연구자
나 개발자들은 특정 화학구조가 어떤 약물들에 나타나는지 신
속히 검색하기를 원하고 있다. 따라서 수많은 약물 파일들에서
사용자가 지정한 구조를 부분구조로 가진 약물 파일들을 자동
으로 찾아내는 것이 필요하다.
1980년대에 구조를 온차일로 검색하여 이를 RDBMS에 저장한
뒤 사용시 문제를 해결하는 RS3 시스템이 미국 보험국에 사용
되었다. 현재 사용되어 사용되고 있다. 본 논문은 RS3 시
스템을 개선시킨 기법을 제시한다.

2. 관련 연구
2.1 그래프 이론
일반적인 화학구조를 부분적으로 가지고 있는 파일들이 검색하
는 것을 'Substructure Searching'라 하며, 이것은 그래프 이론
에서 'Subgraph Isomorphism' 문제로 해결된다. Subgraph
Isomorphism 문제는 NP-Complete 문제이다[2].

2.2 다른 접근방법들
NP-Complete이란 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 후
보기 없는 편집자들을 정의하는 Algorithm이나 Heuristics,
Mapping이 가능한 동작과 결합관계를 줄이는 다양한 방법
이 연구되었다. 현재 단계에서 비교 불가능한 정보의 단계
로 가러나 올라가 비교를 시도하는 Backtracking Algorithm의
일종인 Ullmann Algorithm[6]이 대표적인, 이는 target atom
T에 mapping되는 query node Q가 아닌 노드 Q를 가지고 있다
면 T도 Q에 mapping되는 T를 가지고 있어야 한다는 조건을 반
복적으로 다른 원자들에 적용함으로써 비교대상 원자수를 줄
인다. 하지만 이러한 방법들은 어려운 일부 시험을 요구하기 때문에,
또 다른 방법으로서 시간이 많이 소요되는 영산을 여러 개정하는
스크린링(Screening) 기법이 개발되었다. 구조를 대표하는 키(Key)
을 개발하여 이를 미리 계산하여 저장한 후, 정확성을 높
도로 검색시킨다. 예로서 BASIS fragment dictionary[7], STN
International for its on-line substructure search system[3] 등
이 있다. 1990년대에 들어 하드디스크 가격이 낮아지고 빠르고, 각 원자들
을 중심으로 구조를 상세히 기록하는 데이터의 키로 개발하는 방법
이 등장했다. 이들 Atom-centered Indexes라 하며
HTSS(Hierarchical Tree Substructure Search)[8, 9], S4
system[10] 등이 있다. HTSS는 각 원자들 정점까지 규칙에 따라
문서화하여 비교하는 방법이고, S4 system은 각 원자들 중심으로
인접 원자들의 관계들 를 bitstring으로 나타내어 비교하는 방법
이다. 1990년대에는 구조를 온차일로 검색하여 이를 RDBMS에 저장

751
2003년도 한국정보과학회 기술 학술발표논문집 Vol. 30. No. 2

한 뒤, 검색된 문자열을 사용하여 RS3 시스템이 더욱 특화된 모델로, 본 논문은 RS3 시스템의 문제점을 논의하고 개선방안을 모색하였다.

2.3 RS3 시스템

RS3 시스템의 정보는 원자 중심의 인덱스화를 하며, 다른 원자들의 결합관계를 문자열로 표현하고, 어휘과 와일드카드(%)를 결합
히 포함한 문자열로 표현하여 부분문자열과 부분문자열의 결합으로 변환시킨 것이다. 이렇게 함으로써 NP-Complete인 부분구조
결합 문제를 O(n)의 시간복잡도로 줄일 수 있다.

아래 [표1]와 같이 이론물자와의 결합관계를 하나의 문자로 정의
한다. 예를 들어 Br의 경우에는 'Br' 문자로 표시한다. 그 후, 각 원자별로 'Br'로 단계를 구성하여 결합구조를 문자열로 한다. 이때 문자간의 순위는 앞부터 순서대로, 순서대로 가
서서가 닫혀 있는 부분을 '로 나누어 기술한다. (그림1) 참조

![표1] 결합관계의 측약

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Br</td>
<td>b</td>
</tr>
<tr>
<td>C</td>
<td>c</td>
</tr>
<tr>
<td>= C</td>
<td>d</td>
</tr>
<tr>
<td>= O</td>
<td>e</td>
</tr>
<tr>
<td>wildcard</td>
<td>%</td>
</tr>
</tbody>
</table>

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

[그림1] DB자장구와 문자열

부분구조를 추리하는 경우에는 아래 [그림2]와 같이, 여기서 결합될 수 있는 부분(추리하는 부분구조가 전체 화학구조에 복을
수 있는 곳)을 "*"로 지정해 주어야 한다.

![그림2] 추리부분구조와 문자열

위의 추리 문자열을 가지고 DB에 저장된 문자열들과 문자열 결
합을 수행하게 되면, 후보가 되는 화학구조를 만들 수 있
다. 그 후, ABAM(Atom-by-Atom Matching)을 수행하여 정확
한 결합구조를 출력한다. RS3 시스템은 현재 Accelys사에서
상용화되었다[12].

2.4 RS3 시스템의 문제점

RS3의 가장 큰 문제점은 '로 지정한 구(부분구조)가 연결될 수
있는 곳이 없어지면 따라 정보가 급격히 넘어간다는 점이다.

[그림2]의 3번째 문자열과 C 원자에 '야 지정되면
1%~5%이지만, 30%의 경우로 보니 이자이지고, 다음과
'으로만 쓰여 지게 되며 이상의 구조기술이 어렵게 된다. 즉, '야'를 만나게
되면 그 이후의 구조들은 기술이 불가능하게 되는 것이다.

또한 추가의 문제점은 동일원자들의 동일결합에 대해 우선순위를
정할 수 없다. RS3 시스템이 중요한 아이디어는 구조
기술 우선순위를 알파벳 순서로 하는 것이다. 예를 들어
[그림1]의 2번째 문자열의 'C b, b, c, c, b, c, c'
은 DB에서 B와의 1차결합, C와의 1차결합이 있다. 이를 나타내는 문자 'b',
'c'을 알파벳 순서대로 나열한 것이다. 하지만 3번째 문자열처럼
같은 'b', 'c'에 대해서는 우선순위를 정할 수 없다. 이후의 구조
물 어느 것보다 기술하려는 것이라면 정확하지 않아도 RS3 시
스템의 전반을 기술하는 데에는 이유가 없다.

3. 개선된 RS3 시스템

3.1 해결 방안

이러한 두 가지 문제점은 각 원자별로 결합구조를 포함하는, 각
Level(경로의 길이, 길이별로 그 Level에 해당하는 원자들의 경
로를 기술하여 이를 정렬하여 저장하는 방법으로 해결할 수 있다. 즉, 경로의 모든 원자에 기호를 갖게 하고, 각 원자별로 최소
비용강화를 구성한 다음 모든 원자까지의 경로를 알파벳
나머지 기술하여 저장하는 것이다. [그림1]의 DB자장구를 분
기법으로 정렬하면 아래 [표2]와 같다.

![표2] 해결방안의 DB자장구

<table>
<thead>
<tr>
<th>원자</th>
<th>Level1</th>
<th>Level2</th>
<th>Level3</th>
<th>Level4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Br</td>
<td>c</td>
<td>c, c</td>
<td>c, c, c</td>
<td>c, c, c</td>
</tr>
<tr>
<td>C</td>
<td>b</td>
<td>c, c, c</td>
<td>c, c</td>
<td>c, c</td>
</tr>
<tr>
<td>C</td>
<td>c, c</td>
<td>c, c</td>
<td>c, c</td>
<td>c, c</td>
</tr>
<tr>
<td>O</td>
<td>d</td>
<td>d, c, c</td>
<td>d, c, c</td>
<td>d, c, c</td>
</tr>
<tr>
<td>C</td>
<td>c</td>
<td>c, c</td>
<td>c, c, c</td>
<td>c, c, c</td>
</tr>
<tr>
<td>C</td>
<td>c</td>
<td>c, c</td>
<td>c, c, c</td>
<td>c, c, c</td>
</tr>
</tbody>
</table>

아마가지로 [그림2]의 추리구조의 분기법 기술은 다음 [표3]과
같다.

![표3] 해결방안의 추리구조 기술

<table>
<thead>
<tr>
<th>원자</th>
<th>Level1</th>
<th>Level2</th>
<th>Level3</th>
<th>Level4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Br</td>
<td>%c%</td>
<td>%c, %c</td>
<td>%c, %c</td>
<td>%c, %c</td>
</tr>
<tr>
<td>C</td>
<td>%b%</td>
<td>%c, %c</td>
<td>%c, %c</td>
<td>%c, %c</td>
</tr>
<tr>
<td>C</td>
<td>%c%</td>
<td>%c, %c</td>
<td>%c, %c</td>
<td>%c, %c</td>
</tr>
<tr>
<td>O</td>
<td>%d%</td>
<td>%d, %c</td>
<td>%d, %c</td>
<td>%d, %c</td>
</tr>
</tbody>
</table>

이러한 기술하면, 연결부분(RS3의 '아')을 지정하지 않아
도 되고 따라서 모든 구조를 기술할 수 있게 되면, 정보가 줄어
들어 전체 경로를 기술하고 이를 알파벳 순서로 정렬하기
때문에 동일원자들의 동일결합에 대한 우선순위가 명확하게 제한

752
물이 흘러가게 된다.

3.2 해결 방안의 알고리즘 순서도

본 기법의 특성에 대한 알고리즘 순서도는 아래 [그림3]과 같다.

유의 결과에서 알 수 있듯이 본 기법은 재현률과 정도를 크게

4.2 본 기법의 재현율 검토

순수 그래프에서 어떤 여부로 본 기법은 사이클의 일부분을 읽어내려고 할 때에는 재현률이 100%가 되지 않는다. 하지만

5. 결론

RS3 시스템은 화학구조를 문자열로 기술하는 방법에서 몇 가지

REFERENCES