# 단백질 구조 비교에서 유사성 그래프의 효율적인 생성

최경호[0*], 김진홍**, 이명준**, 이수현*

*창원대학교 컴퓨터·정보통신공학부, **울산대학교 컴퓨터·정보통신공학부
*fates_forever@hotmail.com

# Efficient Generation of Docking Graph in Protein Structure Comparison

Kyung-Ho Choi[0*] Jin-Hong Kim** Myung-Joon Lee** Su-Hyun Lee*

*School of Computer & Information Technology, Changwon National University
**School of Computer Engineering & Information Technology, University of Ulsan

**요 약**

단백질간 구조 비교는 기능적 또는 구조적으로 연관된 단백질을 분류하거나 모티프(motif)를 찾는데 유용하게 사용되고 있다. 여러 가지 단백질간 구조 비교 방법 중에서 단백질 2차구조를 이용하는 방법은 실행속도의 측면에서 장점이 있다. 본 논문에서는 단백질 2차 구조와 그들 사이의 관계를 기반으로 한 단백질 구조 비교에서 사용될 유사성 그래프를 생성하는 방법을 기술하였다. 유사성 그래프는 단백질의 2차구조 사이의 관계를 노드로 하여 생성되는데, 그 시간복잡도가 $O(n^4)$이다. 이에 본 논문에서는 유사성 그래프의 생성을 효율적으로 할 수 있는 알고리즘을 개발하였다.

## 1. 서론

현재 새롭게 밝혀지는 단백질 3차구조 정보의 증가량이 날이 갈수록 높아짐에 따라 단백질 구조 데이터베이스를 대상으로 입력된 단백질 구조 비교 방법은 보다 효과적으로 빠르게 결과를 산출할 수 있어야 한다. 이는 단백질 구조 비교 시 요구되는 많은 데이터를 효과적으로 처리할 수 있는 방법과 고성능의 계산 능력을 가진 고급 장비가 요구된다. 또한, 단백질 구조를 비교하는데 있어 효과적으로 하기 위한 단백질 구조 표현 방법의 개선이 필요하다.

PSAML(PSA Markup Language)[1,2]은 단백질의 2차구조와 2차구조 사이에서 발견되는 상호 관계를 이용하여 단백질 구조를 표현하는 방법을 제공하는 PSA (Protein Structure Abstraction) [1,2]를 기준으로 단백질 구조를 표준화된 문서 표현 양식인 XML로 기술할 수 있는 언어이다. PSA 및 PSAML 기반으로 표현된 단백질 구조를 이용하여 2차구조의 특징과 2차구조 사이의 관계를 비교하여 효과적으로 두 단백질 사이의 유사한 부분 구조를 찾을 수 있다.

단백질 2차구조 기반의 단백질간 구조 비교 방법은 주어진 PSAML 정보를 바탕으로 유사한 부분구조를 내포하는 유사성 그래프를 생성한 후, 모든 노드사이에 간선이 존재하는 서브 그래프인 clique를 찾는 알고리즘을 이용하여 최대로 유사한 부분 구조를 파악할 수 있다.

유사성 그래프는 단백질의 2차구조 사이의 관계를 노드로 하여 생성되는데, 대상 단백질의 2차구조의 수가 $n$이라고 할 때 노드의 수는 $O(n^2)$이 된다. 따라서 유사성 그래프의 간선의 수는 최악의 경우 $O(n^4)$만큼 있게 되며, 따라서 유사성 그래프를 구성하는데 $O(n^4)$의 시간이 걸리게 된다. 본 논문에서는 유사성 그래프의 생성을 효율적으로 할 수 있는 알고리즘을 개발하였으며, 제안한 알고리즘의 최악시간복잡도는 $O(n^2 \log n + |E|)$이다. 여기에서 $|E|$는 생성되는 간선의 수이다.

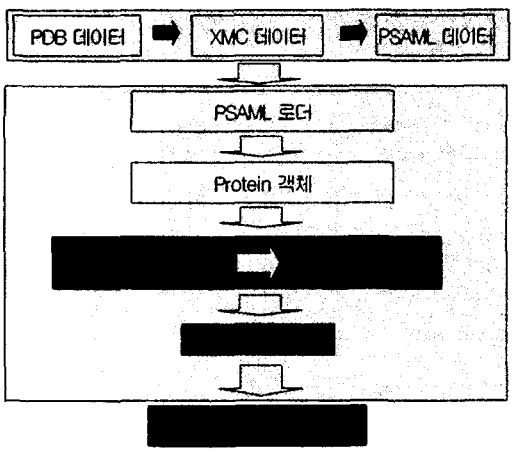
본 논문의 구성은 다음과 같다. 2장에서는 단백질 구조 비교 방법에 대하여 살펴보고, 3장에서는 유사성 그래프의 생성에 관하여 설명하고 효율적인 알고리즘을 제시한다. 끝으로 4장에서는 결론 및 향후 연구방향에 대하여 기술한다.

## 2. 단백질 구조 비교 방법

대표적인 단백질 구조 알고리즘은 단백질 구조의 내부 분자들 사이의 거리 정보를 동적 프로그래밍 기법을 이용한 DALI[3], Cα 원자들 사이에 RMSD가 최소가 되는 부분을 찾는 LOCK[4], 단백질 이차 구조의 3차원 위치정보를 유사 부분을 찾기 위하여 기하학적 해싱 기법 사용하는 3dSEARCH[5], 그리고 단백질 이차구조 사이

의 거리 및 각도 관계를 이용한 SARF2[6] 등이 있다.

PSAML을 이용한 단백질간 구조 비교 방법[7]은 주어진 PSAML 정보를 바탕으로 유사한 부분구조를 내포하는 유사성 그래프를 생성한 후, 모든 노드사이에 간선이 존재하는 서브 그래프인 clique를 찾는 알고리즘[8]을 이용한다. (그림 1)은 PDB[9] 형태로 입력되는 두 단백질간의 구조 비교 방법의 과정을 보여주고 있다.




(그림 1) PSAML 기반의 단백질 구조 비교

입력되는 PDB 형태의 단백질 구조 데이터는 변환 도구를 통하여 PSAML 문서로 변환된다. 단백질 구조 비교를 수행하는 방법은 변환된 PSAML을 읽어 Protein 객체를 생성한다. 생성된 Protein 객체는 PSAML에서 정의하는 2차구조의 특징(아이디, 타입, 3차원 좌표)과 2차구조 사이의 각도와 같은 관계에 대한 정보를 가진다. 생성된 두 Protein 객체로부터 단백질간 유사성을 내포하는 유사성 그래프를 생성하여 모든 노드들 사이의 간선이 존재하는 서브 그래프를 찾는 알고리즘을 이용하여 최대로 유사한 부분 구조를 파악할 수 있다.

## 3. 유사성 그래프의 생성

PSAML을 이용한 단백질간 구조 비교 방법에서 가장 중요한 부분 중의 하나는 유사한 부분구조를 내포하는 유사성 그래프를 생성하는 것이다.

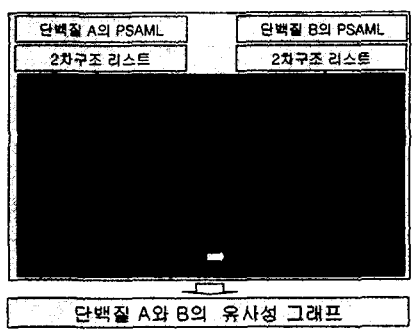
### 3.1 유사성 그래프

PSAML 데이터를 기반으로 단백질 구조간의 유사성을 내포하는 그래프 G는 다음과 같이 정의된다.

<표 1> 유사성을 내포한 그래프 표현

| |
|---|
| $G(A, B) = \{V, E\}$, $A$와 $B$는 단백질 |
| $V = \{(ai, bj) \mid ai \in A, bj \in B, T(ai)=T(bj)\}$ |
| $E = \{[(ai, bk), (aj, bl)] \mid \lvert\Theta(ai, aj)-\Theta(bk, bl)\rvert<\Theta d\}$ |

<표 1>에서 $V$와 $E$는 그래프 $G$의 노드 및 간선의 집합을 나타내고 있다. $V$에 속한 각 노드는 단백질 $A$의 한 이차 구조와 단백질 $B$의 한 이차 구조의 쌍으로 이루어져 있다. $E$에 속한 각 노드사이의 간선은 노드에 포함된 단백질 이차 구조간의 관계가 유사할 경우에 존재한다. 즉, 노드 $(ai, bk)$와 노드 $(aj, bl)$ 사이의 간선은 이차 구조 $ai$와 $aj$에 존재하는 관계와 $bk$와 $bl$에 존재하는 관계가 유사할 경우 생성된다. 이 경우, 단백질 $A$의 $ai$와 단백질 $B$의 $bk$ 및 단백질 $A$의 $aj$와 단백질 $B$의 $bl$가 유사하다는 것을 의미한다. 두 단백질간 유사성 그래프의 생성과정은 (그림 2)와 같다.




(그림 2) 유사성 그래프 생성 방법

(그림 2)에서, 단백질 $A$와 $B$ 각각에서 두 2차구조 요소를 선택한다. 이때, 2차구조 사이의 거리가 일정 거리(threshold distance) 이내에 있는 2차구조를 선택한다. 선택되는 2차구조의 쌍은 $[ai, aj]$와 $[bk, bl]$과 같은 형태이며, $(ai, bk)$, $(aj, bl)$, $(aj, bk)$, $(ai, bl)$와 같은 유사성 그래프 노드가 생성된다. 유사성 그래프 노드 사이의 간선은 각도 관계의 인자 값($\theta d$)보다 작으며 같은 타입일 때 생성된다.

### 3.2 유사성 그래프 생성 알고리즘

제안하는 알고리즘은 단백질의 각도 정보를 정렬함으로서 계산속도를 빠르게 한다. 이를 위하여 다음의 자료구조를 사용한다. angleA[]와 angleB[]는 대상 단백질의 2차구조 사이의 각도정보를 가진 배열로서 정렬되어 있다.

```
Algorithm
  Diff = 10; low = 1; high = 1
  for i=1 to N {
    a = angleA[i]
    for j=low to N {                          - ①
      if (angleB[j] >= a - Diff)
        break
    }
    low = j
    for j=high to N {                         - ②
      if (angleB[j] > a + Diff)
          break
    }
    high = j

    for k=low to high                         - ③
      make an edge between i and k
  }
```

알고리즘에서 N은 두 단백질의 2차구조사이의 관계의 개수로서 2차구조의 수가 $n$개라고 했을 때, $n(n-1)/2$개의 상관관계가 존재한다. Diff는 각도의 차이를 의미하는데 이 차이보다 작은 각도를 가지는 2차구조들은 상관이 있다는 의미이다.

제안한 알고리즘은 중첩된 for-문을 사용하고 있으나, ①, ②의 for-문은 전체적으로 N번의 반복을 수행하며, ③의 for-문은 생성되는 간선의 수만큼 실행된다. 또한 angleA[], angleB[] 배열을 정렬해야하는데, 정렬의 복잡도는 $O(N\log N)$으로서, 이는 $O(n^2\log n)$에 해당한다. 따라서 전체 알고리즘의 시간복잡도는 $O(n^2\log n + |E|)$이다.

### 3.3 실험 결과

다양한 크기의 단백질에 대하여 제안한 알고리즘의 성능을 실험하였다. 실험 환경은 SunOS 5.6에서 자바를 이용하였으며, time 유틸리티를 이용하여 시간을 측정하였다. 시간 측정에 대한 오차를 줄이기 위하여 각 경우마다 5번씩 실행하여 평균을 구하였다.

실험 결과를 <표 2>에 나타내었다. 대상단백질에서 괄호속의 수는 해당 단백질의 2차구조의 수를 나타낸다. 표에서 보듯이 2차구조의 수가 커질수록 유사성 그래프 생성의 효율이 좋아짐을 알 수 있다.

<표 3> 실험결과

| | 대상단백질 | | 생성되는 간선의 수 | Naive 알고리즘 | 제안 알고리즘 |
|---|---|---|---|---|---|
| | 이름 (수) | 이름 (수) | | | |
| 1 | 107L (10) | 108L (10) | 1984 | 1.4초 | 1.4초 |
| 2 | 174L (24) | 175L (24) | 58676 | 1.62초 | 1.42초 |
| 3 | 1ahe (52) | 1ahf (52) | 1232680 | 7.22초 | 1.6초 |
| 4 | 1a0s (98) | 1a0t (102) | 9576768 | 56.14초 | 2.12초 |

## 4. 결론

본 논문에서는 단백질 이차 구조와 그들 사이의 관계를 이용하여 단백질 구조를 비교하는 방법에서 유사성 그래프를 생성하는 방법에 대하여 기술하였다. 그리고 제안한 알고리즘이 좋은 성능을 가지는 것을 실험을 통하여 확인하였다.

두 단백질 구조에서 유사한 부분 구조는 단백질 이차 구조의 정보(형태, 길이)와 그들 사이의 관계(각도, 거리, 길이)를 바탕으로 그래프 형태로 표현하여 최대 유사한 하위 그래프를 찾는 알고리즘을 이용하여 찾아가는데, 본 논문에서는 각도만을 고려하여 그래프를 생성하였다.

추후연구 과제로는 여러 가지의 관계를 모두 고려하는 효율적인 알고리즘을 개발할 예정이며, 제한 프로그래밍 기법을 이용하여 보다 빠른 다중 단백질 구조 비교 방법을 개발할 예정이다.

### [참고문헌]


[1] Su-Hyun Lee, Jin-Hong Kim, Geon-Tae Ahn, and Myung-Joon Lee, "An XML Representation of Protein Datafor Efficient Structure Comparison," Proc. of Second *ICIS*, 2002.

[2] 김진홍, 안건태, 변경익, 윤형석, 이수현, 이명준, "단백질 3차 구조의 추상적인 표현기법," 한국정보과학회, '2001 가을 학술발표논문집(B) 제 28권 2호, 595-597, 2001.

[3] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," Journal of Molecular Biology, Vol. 233, pp. 123-138, 1993.

[4] A. P. Singh and D. L. Brutlag, "Hierarchical Protein Structure Superposition using both Secondary Structure and Atomic Representations," Proc. Intelligent Systems for Molecular Biology, 1997.

[5] A. P. Singh and D. L. Brutlag, "Protein Structure Alignment: A Comparison of Methods", 1999.

[6] N.Alexandrov and D Fischer, "Analysis of topological and nontopological structural similarities in the PDB: New examples with old structures," Proteins, Structure, Function, and Genetics, Vol 25, No. 3, pp.354-365, 1996.

[7] Jin-Hong Kim, Geon-Tae Ahn, Min-Su Cho, Su-Hyun Lee, and Myung-Joon Lee, "An XML Representation of Protein Data for Structure Comparison and Its Application," Proc. of 2002 KSBI annual meeting, 2002.

[8] C. Bron and J. Kerbosch, "Algorithm 457: Finding All Cliques of an Uundirected Graph," CACM, 16(9):575-577, 1973.

[9] H. M. Berman, J. D. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," Nucleic Acid Research, Vol. 28, No. 1, pp. 235-242, 2000.