단백질 삼차 구조 예측: 비교 모형(comparative modeling) 방법

김 승 연

1. 비교 모형 방법

단백질 삼차 구조 예측(protein tertiary structure prediction)에서 가장 많이 사용되는 방법은 비교 모형(comparative modeling 또는 homology modeling) 방법이다. 단백질 서열들의 유사도(similarity)가 높은 경우에는 단백질 서열들의 친족 관계(homology)를 쉽게 찾을 수 있다. 따라서, 주어진 문제의 단백질 아미노산 서열(보통 query 단백질이라고 부른다)이 삼차 구조가 알려진 다른 단백질의 서열(이것을 template 단백질이라고 부른다)과 많이 유사한 경우에는 comparative modeling 방법을 이용하여 주어진 단백질의 삼차 구조를 쉽게 예측할 수 있으며, 이 경우 예측 정확도가 아주 높다.

comparative modeling 방법에는 크게 두 가지가 있다. 첫 번째는 조각에 기초한 방법(fragment-based method)[1]이고 두 번째 방법은 구속조건에 기초한 방법(restraint-based method)[2]이다. fragment-based method 에서는 서열 중에서 잘 보존되는 부분(conserved region)과 잘 변하는 부분(variable region)을 먼저 구분한다. 여기에서 잘 보존되는 부분은 query 단백질과 template 단백질의 서열이 잘 일치하는 부분을 뜻하며, 많은 경우에 helix 나 beta 구조와 같은 이차 구조가 잘 보존된다. 반면에 잘 변하는 부분은 query 단백질과 template 단백질의 서열이 잘 일치하지 않는 부분을 지칭하며, 많은 경우에 helix 와 helix(또는 helix 와 beta 구조, beta 구조와 beta 구조)를 연결해주는 loop 부분이 보존되지 않고 잘 변한다. fragment-based method 에서는 먼저 conserved region 의 구조를 만들어내고, 그 이후에 variable region 의 구조를 만들어 낸다. 최종적으로 옆 가지(side chain)를 붙여주어서 예측된 삼차 구조를 만들어 준다. 옆 가지에 관련된 프로그램으로는 SCWRL(http://dunbrack.fccc.edu/SCWRL3.php)이 가장 많이 사용되고 있다. fragment-based method 의 대표적인 예는 COMPOSER 와 SWISS-MODEL 이다. COMPOSER 는 상용이어서 쉽게 이용할 수 없다. 반면에

SWISS-MODEL 은 웹(http://www.expasy.org/swissmod/SWISS-MODEL.html)에서 query 단백질의 서열을 입력해주면, 예측된 삼차 구조를 쉽게 얻을 수 있다.

restraint-based method 에서는 먼저 template 단백질의 아미노산들 또는 원자들 사이의 거리를 측정한 후, 이 거리를 구속조건(restraint)으로 사용하여 query 단백질의 삼차 구조를 만들어낸다. restraint-based method 의 대표적인 방법으로는 MODELLER 가 있다. MODELLER 의 실행파일(executables)을 “http://salilab.org/modeller/”에서 학문적 연구가 목적인 경우 공짜로 얻을 수 있기 때문에, MODELLER 는 comparative modeling 방법들 중에서 가장 많이 사용되고 있다.

comparative modeling 방법들 중에서 프로그램까지 제공하는 경우로는 Jackal(C++ code, 대략 90000 줄)이 있다. comparative modeling 방법을 자체적으로 직접 개발하고 싶으신 분은 Jackal 을 참고하기 바란다. Jackal 은 “http://trantor.bioc.columbia.edu/programs/jackal”에서 얻을 수 있다. 웹(http://www.bmm.icnet.uk/~3djigsaw/)에서 사용하기 편리한 comparative modeling 방법으로는 3D-JIGSAW 가 있다. 가장 최근에 발표된 comparative modeling 방법으로는 MolIDE(http://dunbrack.fccc.edu/molide/molide.php)가 있다[3].

2. MODELLER 사용법

지금부터는 comparative modeling 방법들 중에서 가장 많이 사용되고 있는 MODELLER 의 사용법을 간단하게 소개하고자 한다. MODELLER 는 Windows, Mac, Linux, Unix 와 같은 operating system 에서 사용할 수 있다. 따라서 자신에게 알맞은 파일을 “http://salilab.org/modeller/”에서 얻어야 한다. 이 웹 페이지에서 매뉴얼도 얻을 수 있으며, 매뉴얼을 따라 MODELLER 를 쉽게 설치할 수 있다. 여기에서는 MODELLER 버전 7v7(operating system: Linux)을 기준으로 하여 사용법을 소개하려고 한다. MODELLER 를 이용한 삼차 구조 예측 과정은 다음과 같이 크게 세 단계로 나누어진다.

(1) template 단백질 찾기
(2) query 단백질 서열과 template 단백질 서열의 정렬(alignment)
(3) query 단백질의 삼차 구조 만들기

첫 번째 단계에서는 서열 데이터베이스 검색 프로그램들을 이용하여 template 단백질을 찾는다. 검색 프로그램으로는 BLAST 나 PSI-BLAST 가 많이 이용된다(두 번째 글인 “단백질 구조 예측의 기초: BLAST 와 프로파일(profile)” 참고). BLAST 로 template 단백질을 찾을 수 있으면 쉬운 CM(Comparative Modeling) 문제라고 부른다. 반면에 PSI-BLAST 를 사용해야만 template 단백질을 찾을 수 있는 경우에는 어려운 CM 문제라고 부른다. 여기에서는 제 5 회 세계 단백질 구조 예측 대회(CASP5)에 출제되었던 T0137 문제(query 단백질 서열)를 예로 들어 설명하겠다. 이 단백질은 133 개의 아미노산으로 이루어져 있으며, 지방산에 결합하는 단백질이다. CASP5 당시에는 T0137 문제의 삼차 구조가 알려지지 않았지만, 지금은 삼차 구조가 알려져서 PDB(Protein Data Bank)에 이미 포함되어 있다(PDB ID: 1O8V). 따라서 PSI-BLAST 검색을 실행하면 1O8V 가 가장 좋은 점수를 보여주지만, query 단백질 서열과 동일하므로 버려야 한다. 자기자신을 버리고 나면 query 단백질 서열에 가장 가까운 단백질 중의 하나는 1BWY 이다. 1BWY(132 개의 아미노산으로 이루어짐)는 template 단백질이 되며, 삼차 구조의 좌표를 담고 있는 1BWY.pdb 파일을 PDB(http://www.rcsb.org)에서 얻어온다.

두 번째 단계에서는 T0137 문제의 서열(query)과 1BWY.pdb(template)를 이용하여 query 와 template 의 서열을 정렬한다. 이때 align2d.top 이라는 파일을 만들어주어야 한다. align2d.top 파일의 내용은 다음과 같다.

```
INCLUDE
READ_MODEL FILE = '1BWY.pdb'
SEQUENCE_TO_ALI ALIGN_CODES = '1BWY'
READ_ALIGNMENT FILE = 'T137.seq', ALIGN_CODES = ALIGN_CODES ‘T137', ADD_SEQUENCE = on
ALIGN2D
WRITE_ALIGNMENT FILE='T137.ali', ALIGNMENT_FORMAT = 'PIR'
```

여기에서 T137.seq 는 query 의 서열을 담고 있는 파일의 이름을 뜻하며, PIR format 으로 기록되어야 한다. T137.seq 파일의 내용은 다음과 같다.

```
>P1;T137
sequence:T137:::::::0.00: 0.00
MEAFLGTWKMEKSEGFDKIMERLGVDFVTRKMGNLVKPNLIVTDLGGGKYKMRSESTFKTTECSFKLGEKFKEVTPDSREVA
SLITVENGVMKHEQDDKTKVTYIERVVEGNELKATVKVDEVVCVRTYSKVA*
```

이 파일의 끝은 항상 “*”로 끝난다는 점에 주의해야 한다. 한편 정렬된 결과는 T137.ali 파일에 기록된다. 모든 준비가 끝났으면 다음과 같이 MODELLER 명령어를 실행해 본다.

```
mod7v7   align2d.top
```

위 명령어의 실행 결과로서 T137.ali 파일을 얻을 수 있으며, 파일의 내용은 다음과 같다.

```
>P1;1BWY

structureX:1BWY:    1 :A: 132 :A:undefined:undefined:-1.00:-1.00

VDAFVGTWKLVDSKNFDDYMKSLGVGFATRQVGNMTKPTTIIEVNG-

DTVIIKTQSTFKNTEISFKLGVEFDETTADDRKVKSIVTLDGGKLVHVQKWNGQETSLVREMVDGKLILTLTHGTAVCTRTYEKQ

A*

>P1;T137

sequence:T137:       : :       : ::: 0.00: 0.00

MEAFLGTWKMEKSEGFDKIMERLGVDFVTRKMGNLVKPNLIVTDLGGGKYKMRSESTFKTTECSFKLGEKFKEVTPDSREVA

SLITVENGVMKHEQDDKTKVTYIERVVEGNELKATVKVDEVVCVRTYSKVA*
```

세 번째 단계에서는 model.top 이라는 파일이 있어야 한다. model.top 파일의 내용은 다음과 같다.

```
INCLUDE

SET ALNFILE = 'T137.ali'

SET KNOWNS = '1BWY'

SET SEQUENCE = 'T137'

SET STARTING_MODEL = 1

SET ENDING_MODEL = 5

CALL ROUTINE = 'model'
```

여기에서는 서로 다른 다섯 개의 삼차 구조가 만들어지도록 기록하였다. 만약 서로 다른 세 개의 삼차 구조를 만들고 싶다면 “SET ENDING_MODEL = 3”이라고 적으면 된다. 모든 준비가 끝났으면 다음과 같이 MODELLER 명령어를 실행해 본다.

```
mod7v7   model.top
```

위 명령어의 실행 결과로서 T137.B99990001,…,T137.B99990005 의 다섯 개의 파일들을 얻을 수 있는데, 이들이 예측된 삼차 구조의 좌표를 담고 있다.

[참고문헌]

[1] T. L. Blundell, M. E. J. Sternberg, B. L. Sibanda, and J. M. Thornton, Knowledge-based prediction of protein structures and the design of novel molecules, Nature **326** (1987) 347-352.

[2] A. Sali and T. L. Blundell, Comparative protein modeling by satisfaction of spatial restraints, J. Mol. Biol. **234** (1993) 779-815.

[3] A. A. Canutescu and R. L. Dunbrack, MolIDE: a homology modeling framework you can click with, Bioinformatics **21** (2005) 2914-2916.