MOD^f: 대규모 단백질 DB에서 효과적이고 빠르게 PTM을 동정하는 알고리즘

신성호*, 박희진**, 백은옥***
*한양대학교 전자컴퓨터통신공학과
**한양대학교 컴퓨터공학부
***서울시립대학교 기계정보공학과
e-mail: landdog@hanyang.ac.kr hjpark@hanyang.ac.kr paek@uos.ac.kr

MOD^f: An Effective and Fast Algorithm for Identification of PTM in Large Protein Sequence Database

Seong-Ho Shin*, Heejin Park**, Eunok Paek***
*Dept of Electronics and Computer Engineering, Hanyang University
**Dept of Computer Science and Engineering, Hanyang University
***Dept of Mechanical and Information Engineering, University of Seoul

요 약

프로테오믹스는 세포 안 또는 개체 안의 모든 단백질을 총체적으로 연구하는 분야이다. 단백질 동정은 단백질이 어떤 아미노산의 서열로 구성되었는지를 확인하는 것이다. 하지만 Post-translational modification과 같은 단백질 변형을 고려하게 되면 단백질 동정은 매우 어렵게 된다. MOD^i 알고리즘은 단백질 동정을 할 때 Post-translational modification의 종류나 개수에 제한 없이 단백질 동정을 정확하게 수행한다. 하지만, 대용량 단백질 서열 데이터베이스를 사용하면 수행시간이 많이 걸리는 단점이 있다. 본 논문에서는 MOD^{i}를 보완하기 위해 대용량 데이터베이스에서 후보 단백질을 선정하는 알고리즘을 통해서 개선된 MOD^f 알고리즘을 제안하고 Target-decoy search strategy를 적용하여 정확성을 분석한다. 후보 단백질 선정 알고리즘과 Target-decoy search strategy 적용 결과 MOD^{f}는 MOD^{i}에 비해 정확도를 희생하지 않으면서 수행속도는 약 2배 향상되었다.

1. 서론

프로테오믹스(Proteomics)[1]는 단백질(Protein)의 기능과 성질을 연구하는 학문이다. 단백질은 하나 이상의 아미노산(Amino acid) 서열로 이루어지며 단백질 동정(Protein identification)은 단백질이 어떤 아미노산 서열로 구성되는지 확인하는 과정이다. 단백질은 아미노산 서열에 따라 기능과 성질에 차이가 있기 때문에 단백질 동정은 프로테오믹스에서 매우 중요하다.

단백질 동정을 위해서는 주어진 탠덤 질량 스펙트럼(Tandem mass spectrum)을 정확히 분석해야 올바른 결과를 얻을 수 있다. 정확한 분석을 위해서는 주어진 스펙트럼에 대한 정보를 통해 이 스펙트럼과 비슷한 스펙트럼 형태를 갖는 펩타이드를 찾아야 한다. 이러한 문제를 해결하기 위해 SEQUEST[2], Mascot[3], InsPecT[4] 같이 많은 알고리즘들이 사용되고 있다. 하지만 스펙트럼을 구성하는 정보가 분명하게 나타나지 않거나 실험 데이터를 이온화하는 과정에서 생기는 변이로 인해서 잘못된 결과를 출력 할 수 있다. 이러한 문제 때문에 정말 올바르게 동정되었는지 확인할 필요가 있다.

Post-translational modification(PTM)이란 단백질 합성 후 나타나는 변이를 말한다. 단백질 변이는 단백질 합성이 일어난 후 하나이상 발생하며 단백질의 구조를 복잡하게 만들어 단백질 동정을 어렵게 한다. PTM을 고려하여 단백질 동정을 수행하는 것은 PTM을 고려하지 않은 것에 비해 어렵다. 왜냐하면, PTM의 가능한 모든 조합은 그 수와 종류가 증가할수록 더 많은 PTM을 고려해야 하기 때문이다. 그래서 기존 단백질 동정 알고리즘들은 PTM의 종류와 개수에 제한한다. 하지만 PTM의 종류와 개수를 제한하게 되면 올바른 단백질 동정 결과를 얻을 수 없다.

현재 단백질 동정 알고리즘에서 단백질 서열 데이터베이스의 크기와 PTM의 종류와 개수에 제한을 두지 않고 단백질을 동정하는 알고리즘은 존재하지 않는다. MOD^i[5]는 드 노보 시퀀싱 방법[6]과 데이터베이스 검색방법을 모두 사용하여 단백질을 동정하는 알고리즘으로 PTM의 종류와 개수에 제한 없이 단백질 동정을 수행한다. 하지만, 대규모 단백질 서열 데이터베이스를 사용할 경우 단백질을 동정하는데 매우 많은 수행시간이 필요하다.

단백질 동정의 결과에 있어서 올바르지 않은 동정과 올바른 동정 사이에 겹치는 부분이 존재하는데, 이 문제는 true positives를 포기하여 false positives의 수를 최소로 만들거나 false positives를 허용하여 true positives의 수

를 최대로 만들도록 제한하여 false positives에 대한 비율을 조절 할 수 있다.

Target-decoy search strategy[7][8]는 동정된 결과가 올바른지 올바르지 않은지 정확하게 구별하지는 않지만, Target-decoy의 통합된 데이터베이스를 통해 동정 결과에 대한 false positives의 비율을 측정할 수 있다. Target-decoy search strategy에 필요한 통합 데이터베이스는 쉽게 만들 수 있는데, Target 데이터베이스에 포함된 단백질 서열과 반대되는 서열을 갖는 단백질을 생성한 후 Target 데이터베이스 덧붙여서 사용한다.

후보 단백질 선정 알고리즘(Candidate protein filtering algorithm)[9]은 대용량 단백질 서열 데이터베이스에서 후보 단백질을 선정하는 알고리즘이다. 실험스펙트럼의 정보를 추출하여 이와 유사한 모양을 갖는 후보 단백질들을 선정하여 새로운 데이터베이스를 생성한다. 이 알고리즘은 MOD^i에서 대규모 단백질 서열 데이터베이스를 사용하였을 때 수행시간 많이 걸리는 문제를 해소하기 위해 제안되었다.

본 논문에서는 Target-decoy search strategy와 후보 단백질 선정 알고리즘을 이용하여 새로운 MOD^f 알고리즘을 제안하였다. FDR 1%를 적용하여 MOD^f와 MOD^i의 동정 결과를 비교한 결과 99.1%의 동일한 결과를 나타냈는데 이것은 MOD^f 알고리즘이 기존 MOD^i에 비해 정확도는 유지하면서 수행시간을 단축한 것을 보여준다.

2. 후보 단백질 선정 알고리즘

후보 단백질 선정 알고리즘은 실험스펙트럼과 단백질 서열 데이터베이스를 입력으로 받아 실험스펙트럼으로부터 추출된 정보(sequence tag, prefix mass, suffix mass, parent mass)를 이용하여 후보 펩타이드(Candidate peptide)를 선정하고, 선정된 후보 펩타이드를 많이 포함하는 단백질을 후보 단백질(Candidate protein)로 선정하는 알고리즘이다.

후보 단백질 선정 알고리즘은 다음의 네 단계 과정으로 진행된다.

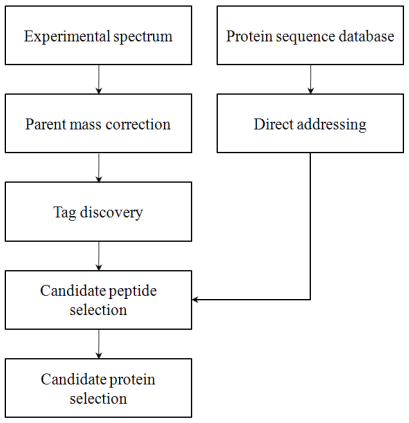

그림 1 후보 단백질 선정 알고리즘

먼저, parent mass correction 단계에서 펩타이드의 질량인 parent mass를 보정한다. 두 번째, tag discovery 단계에서 sequence tag를 추출하고, 각 sequence tag에 대한 prefix mass, suffix mass 정보를 얻는다. 세 번째, candidate peptide selection 단계에서 추출된 sequence tag를 포함하면서 펩타이드의 두 flanking mass(prefix mass, suffix mass)가 sequence tag의 두 flanking mass와 동일한 값을 갖는 펩타이드를 후보 펩타이드로 선정한다. 마지막으로, candidate protein selection 단계에서 후보 펩타이드를 많이 갖는 단백질을 후보 단백질로 선정한다.

3. Target-decoy Search Strategy

대부분의 단백질 동정 알고리즘들은 올바르지 않은 동정을 상당부분 포함하고 있다. 올바르지 않은 동정이 일어나는 이유로는 스펙트럼의 특성을 나타내기 위해 필요한 정보가 빈약하거나 확인된 결과가 단백질 서열 데이터베이스에 존재하지 않기 때문이다. 그렇기 때문에 올바른 동정과 올바르지 않은 동정을 구별하는 것이 필요하다.

Target-decoy search strategy에서는 동정이 올바른지 올바르지 않은지 정확하게 결정하지는 않지만, 통합된 Target-decoy 데이터베이스를 통해서 많은 동정 결과에 대한 false positives의 비율을 측정 할 수 있다. 이렇게 측정된 false positives의 비율을 통해서 올바르다고 생각되어지는 동정을 구별하는데, 이렇게 선별을 한다고 해도 그 중에 false positives가 완전히 제거되는 것은 아니다. 하지만 이 방법을 통해서 올바른 동정에 대한 특징(예를 들어, 펩타이드의 길이, 전하, 알고리즘에 할당된 점수)을 제시할 수 있다.

Target-decoy search strategy를 위한 데이터베이스는 쉽게 생성할 수 있다. 입력받은 단백질 서열 데이터베이스를 Target 데이터베이스로 하고 이 Target 데이터베이스에 포함된 단백질 서열과 반대되는 순서를 갖는 단백질을 생성한다. Target 데이터베이스에 대한 모든 단백질에 대해 역방향 단백질을 생성한 것을 Decoy 데이터베이스라고 한다. 이 Decoy 데이터베이스를 Target 데이터베이스에 덧붙여 하나의 통합된 데이터베이스를 생성한다. 이때 Decoy 데이터베이스에는 검색결과에서 확실히 구분이 될 수 있도록 명확하게 표시되어야 한다.

$$FDR = \frac{2 \times FP}{(TP + FP)}$$

FP: false positive의 수

TP: true positive의 수

FDR은 위 식으로 측정할 수 있다. FDR이 주어지면 해당 FDR을 만족하는 score threshold 값을 찾는데 이 sc

ore가 올바른 것과 올바르지 않은 것에 대한 기준이 된다.

4. 정확도 분석

정확도를 분석하기 위해 human plasma의 142,494개 실험 스펙트럼과 Swissprot_Human_57.12의 20,328개 단백질 서열을 포함한 데이터베이스를 사용하였다. 그리고 실험 데이터를 후보 단백질 선정 알고리즘을 사용하여 후보 펩타이드를 16개 이상 갖는 4,805개의 후보 단백질을 갖는 데이터베이스를 만들었다.

Target-decoy search strategy를 적용하기 위해서 Swissprot_Human_57.12와 4.805개의 후보 단백질 데이터베이스에 대한 decoy 데이터베이스를 만들어 Target과 Decoy가 더해진 데이터베이스를 생성하여 실험을 진행하였다.

구 분	선정된 DB 사용	전체 DB 사용
후보 단백질 선정	2시간 40분	-
MOD^{i}	209시간 11분	394시간 34분
총 시간	211시 51분	394시간 34분

그림 2 수행시간 비교

먼저, 후보 단백질 선정 알고리즘을 통해 선정된 4,805개의 단백질을 갖는 데이터베이스를 만드는 시간은 2시간 40분이었다. 그리고 이 선정된 데이터베이스에 Decoy를 추가하여 MOD^{f}를 수행한 시간은 209시간 11분이었다. 반면 전체 데이터베이스를 사용하여 MOD^{i}를 수행한 시간은 394시간 34분으로 후보 단백질 선정 데이터베이스를 사용한 MOD^{f}가 MOD^{i}에 비해 필요한 수행시간이 약 2배 가량 감소된 것을 확인 할 수 있다.

Decoy 데이터베이스가 첨가된 전체 Swissprot_Human_57.12 데이터베이스를 사용한 결과와 후보 단백질 선정 알고리즘으로 선정된 데이터베이스를 통해 나온 동정 결과를 비교하였다.

구 분	MODi 동정	MODi-MODf 동일	불일치
1% FDR	18,382	18,233(99.1%)	149

그림 3 동정 결과 비교

FDR 1%를 적용한 결과, MOD^{i}는 18,382개의 스펙트럼을 동정하였고, MOD^{f}는 19,807개의 스펙트럼을 동정하였다. 이 중 두 결과 모두에서 동일하게 동정한 수는 18,233개로 99.1%의 정확성을 보였다. 이 실험을 통해서 MOD^{f}와 MOD^{i}의 정확도가 크게 차이나지 않는다는 것을 확인하였다.

5. 결론

본 논문에서는 Decoy가 추가된 Swissprot_Human 데이터베이스를 통한 실험 MOD^{i}와 후보 단백질 선정 알고리즘을 사용한 실험 MOD^{f}를 Target-decoy search strategy를 사용하여 비교 분석하였다. FDR 1%를 적용한 결과 99.1%가 동일하게 동정되었다는 것을 확인하였다. 이러한 결과를 통해 MOD^{f} 알고리즘이 기존 MOD^{i}에 비해 정확도를 희생하지 않으며 수행시간이 약 2배 빨라졌다는 것을 확인하였다.

참고문헌

[1] Marc R. Wilkins, Keith L. Williams, Ron D. Appel, Denis F. Hochstrasser, Proteome Research: New Frontiers in Functional Genomics, 1 edition, Springer, 1997

[2] J. Eng, A. L. McCormack, J. R. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, J. Am.Soc.MassSpectrom, 5, 976-989, 1994

[3] D. N. Perkins, D. J. Pappin, D. M. Creasy and J. S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data, Electrophoresis, 20, 3551-3567, 1999

[4] S. Tanner, H. Shu, A. Frank, L. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna, InsPecT: Identification of Posttranslationally Modified Peptides from Tandem Mass Spectra, Analytical Chemistry, Vol. 77, No. 14, 4626-4639, July 15, 2005

[5] S. Kim, S. Na, J. Sim, H. Park, J. Jeong, H. Kim, Y. Seo, J. Seo, K. Lee and E. Paek, MODi: a powerful and convenient web server for identifying multiple post-translational peptide modifications from tandem mass spectra, Nucleic Acids Research, 34, W258-W263, 2006

[6] V. Dancik, T. S. Addona and K. R. Clauser, *De Novo* Peptide Sequencing via Tandem Mass Spectrometry, Journal of computatoinal bology, Volume 6, Number 3, 327-342, 1999

[7] Elias, J.; Gygi, S. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 2007, *4*, 207 - 214.

[8] Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, S. W. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **2008**, *7*, 29 - -34.

[9] J. Lee, H. Park, E. Paek, An Effective Candidate Protein Filtering Algorithm in Large Protein Sequence Database, KOREA INFORMATION SCIENCE SOCIETY, Vol. 36, No. 2B 2009. 11, PP. 479-484