

지역적/전역적 형태기술자 최적화를 통한 단백질 구조 동등성 평가

서정근^{*,*}, 전성환^{*}, 최유주^{*}

^{*}서울미디어대학원대학교 뉴미디어학부, ^{*}교신저자

e-mail: jksuh@smit.ac.kr, alexchun78@gmail.com, yjchoi@smit.ac.kr

Optimization of Shape Descriptor for Comparability Assessment of Protein Structure

Jung-Keun Suh^{*,*}, Sung-Hwan Chun^{*}, Yoo-Joo Choi^{*},

^{*}Dept of NewMedia, Seoul Media Institute of Technology

^{*}Corresponding author

요 약

단백질의 구조적 동등성을 평가를 위한 형태 기반의 기술자에 대한 연구는 제한적으로 이루어지고 있으며 대부분 지역적 특성 값으로 표현된 지역적 접근 방법이 다수를 이루고 있다. 지역적 특성과 전역적 특성을 포함하는 형태기술자의 경우 각 특성들이 동등한 중요도로 결합되어 있다. 본 연구에서는 선형 회귀분석을 적용하여 각 특성에 대한 중요도를 최적화하여 형태기술자를 재정의 하였다. 최적화된 형태기술자를 단백질의약품인 인슐린 모델에 적용하여 구조적 동등성을 평가할 수 있는 방법론을 제시하였다. 최적화된 형태기술자는 동일한 그룹에 속한 인간 인슐린 단백질 모델과 지역적으로 다른 구조를 가지는 인슐린 아날로그 그룹을 명확히 구분할 수 있음을 확인하였고 이러한 성능은 이전 연구의 형태기술자와 3D 저니크 기술자보다 더 좋은 성능을 보였다. 또한 제안한 방법은 고해상도 단백질 3차 구조 정보를 활용하여 유사성을 판별한 RMSD 방법과 유사하게 서로 다른 표면 구조를 가지는 단백질을 구별할 수 있음을 확인하였다. 이러한 결과로부터 본 연구에서 제시하는 형태기술자 및 최적화된 동등성 평가 함수는 SAXS 분석과 같이 저해상도 단백질 표면 모델을 확보할 수 있는 분석에 적용하여 단백질의 구조적 동등성을 판별할 수 있는 기반을 제공할 수 있을 것으로 판단된다.

1. 서론

단백질은 모든 생물체의 몸을 구성하는 대표적인 고분자 유기물로서 생체 내에서 중요한 기능을 담당한다. 각 단백질은 분자구조가 서로 다르며 이에 대응하여 고유한 기능을 갖기 때문에 단백질 구조분석은 매우 중요하다. 특히 단백질의약품의 경우 복잡한 생산 및 정제 과정을 거쳐 제조되기 때문에 주성분 단백질에 대해 빠르고 간단하게 구조를 확인할 수 있는 분석법 개발이 요구되고 있다 [1,2]. 단백질 구조 분석은 X-ray 회절 분석법 및 NMR 분광 분석법과 같은 고전적인 방법을 적용하여 분자 수준에서의 고해상도 구조를 확인할 수 있으나 X-ray 회절 분석법은 결정(Crystal)을 만들 수 없는 단백질의 경우 구조 확인이 불가능하며 NMR 분광 분석법은 크기가 큰 단백질의 경우 구조 확인이 어려운 한계가 있다[3]. 고전적인 방법을 이용하여 구조 확인이 가능하더라도 구조 분석을 위해 많은 시간과 비용이 소요되기 때문에 빠른 개발이 요구되는 단백질의약품에 적용하기에는 한계가 있다. 단백질 구조분석의 또 다른 방법으로 푸리에변환 적외선법(Fourier Transform Infrared, FTIR) 및 UV 원편광 이색성 분광법(Ultraviolet Circular Dichroism, UV CD)과 같은 분광 분석법이 사용되고 있다[4]. 분광 분석법은 액상에서 분석이 가능하다는 장점이 있으나 해상도 및 특이

도의 문제로 고해상도 구조 확인이 불가능하다. X-ray 회절 분석법과 분광 분석법의 대안으로 단백질 구조 분석을 위해 Small Angle X-ray Scattering (SAXS) 분석법이 제시되었다[5]. SAXS 분석법은 빠르고 간편하게 액상에서 구조 분석이 가능하며 저해상도 단백질 구조 모델을 확보할 수 있다는 장점이 있다. SAXS 분석을 통해 확보한 X-ray 산란 데이터의 경우 Ab initio Modeling 및 Rigid Body Modeling이 가능하며 단백질의 3차 구조에 대한 형태적 특성을 가지는 저해상도 표면 모델을 생성할 수 있다[5]. 이러한 표면 모델의 형태적 특성을 이용하여 단백질의약품의 생산 배치 사이에서 주성분 단백질에 대한 구조적 동등성 평가가 가능하다. 이전 연구에서 단백질의약품의 단백질에 대한 저해상도 모델의 구조적 동등성 평가를 위한 방법론을 제시하였다[5]. 저해상도 구조 모델의 구조적 유사성 평가를 위해 지역적 또는 전역적 형태적 특성 기반의 형태기술자(shape descriptor)를 구성하였으며 이를 적용하여 단백질의 구조적 동등성 평가 방법론의 실효성을 검증하였다. 본 연구에서는 지역적 또는 전역적 형태기술자를 최적화하여 재정의 하고 이를 이전 결과와 비교하여 우수성을 검증하였으며 다른 방법론인 3D 저니크 형태기술자 (3D Zernike shape descriptor)[6,7] 방법론과 비교하여 우수성을 검증하였다.

2. 관련 연구

저해상도 표면모델 기반 형태기술자를 활용한 구조적 동등성 평가는 단백질의 고해상도 3차구조 데이터를 이용한 동등성 평가에 비해 제한적으로 이루어지고 있으며 대표적인 형태기술자는 단백질 모델을 일정한 구역별로 나누어 각 구역 내의 지역적 특성 값 (geometric feature)으로 표현된 지역적 접근 방법이 다수를 이루고 있다[5].

PatchDock 알고리즘에서는 단백질의 지역적 형상 특성을 볼록(convex), 오목(concave), 평평한 부분(flat)으로 분류한 후, 특정 단백질의 표면 삼각형(face triangle)과 비교 대상의 단백질 모델의 한 정점(vertex)간의 유클리디안(Euclidean) 거리를 바탕으로 한 형태 기술자를 제안하였다[8,9].

저니크 형태기술자는 단백질의 표면 모델을 확보한 후, 표면 모델에 대한 3D 그리드 지도에 각 그리드 셀(voxel)을 할당하여 저니크 모멘트를 계산한다. 특정 단백질이 가지는 저니크 형태기술자의 계산 값과 비교 대상의 단백질이 가지는 저니크 형태기술자와 비교함으로써 두 단백질 사이의 유사성을 평가한다[6,7].

본 연구진의 기존 형태기술자의 설계를 위해 단백질의 표면 모델을 확보한 후, 단백질 모델의 외곽 형태 특성을 모델의 X-Y축 및 X-Z축 비율을 전역 지표로 사용하였고, 모델의 중심점과 모든 정점 간의 유클리디안 거리들의 히스토그램을 지역 지표로 사용하였다. 1차원으로 이루어진 전역 지표와 510차원으로 이루어진 지역 지표를 동등하게 합한 형태기술자를 제안하였다.

PatchDock 및 3D 저니크 형태기술자의 경우 단백질 구조 비교를 위해서 활용되고는 있지만 매우 비슷한 구조를 가진 단백질 사이에서의 구조적 동등성을 평가하기에는 한계가 있다[3]. 특히 단백질의약품의 경우 동일한 아미노산 서열을 가지고 있는 단백질에 대한 생산 배치 사이에서의 구조적 동등성을 평가해야하기 때문에 작은 구조적 차이를 판별할 수 있는 동등성 평가 방법이 필요하다. 또한 본 연구진의 이전 연구의 경우 전역 지표와 지역 지표를 동등하게 처리한 형태기술자이기에 이를 개선할 필요가 있다.

이러한 한계를 극복하기 위해 본 연구에서는 이기존 지역적 그리고 전역적 지표에 대한 형태 형태기술자를 재정의 하여 새로운 형태기술자를 제안하고자 하며 새로운 형태기술자의 우수성을 기존 및 다른 형태기술자와 비교하여 우수성을 검증하였다.

3. 형태기술자 최적화

이전 연구에서 단백질에 대한 구조적 동등성 비교를

위해 정점 간 유클리디안 거리(Euclidean distance)를 활용한 지역적 지표를 나타내는 히스토그램과 입력 받은 표면 모델의 전역적 지표를 나타내는 X-Y축 비율과 X-Z축 비율을 형태기술자로 정의하였으며 단백질 구조 정보를 활용하여 그 성능을 평가하였다[3]. 본 연구에서는 이전 연구의 기존 형태기술자에 대한 최적화를 진행하여 최적화된 형태기술자를 재정의 하였으며 이에 대한 성능 평가를 수행하였다. 전체 연구 과정은 표 1과 같다.

표 1. 연구 과정 개요

1	단백질 표면 모델
	분자구조 3D 구조로부터 표면 모델 변환
2	주성분 분석과 공간 정규화
	회전 /이동 정보 도출
	모든 모델을 동일한 조건으로 변환
3	특성 데이터 추출
	모델 XYZ 축의 비율 정보
	정점 간 거리에 대한 히스토그램
	동등성 평가 함수 정의 및 성능 평가
4	동등성 평가 함수 최적화
	X-ray 회절 방식 분자 구조 데이터와 비교
	회귀 분석을 통한 최적 계수 결정
5	최적화 동등성 평가 함수의 성능 평가
	박스 플롯으로 동등성 정보 시각화
	기존 연구 방법론 결과와 비교

입력된 단백질 모델 M, N에 대해 각 단백질 모델의 X-Y축 길이 비율을 $R(N_{X-Ylength})$, X-Z축에 대한 비율을, $R(N_{X-Zlength})$, 모델 중심점과 각 버텍스 간의 거리에 대한 히스토그램을, $Histogram(M(Distance_{C-V}))$, 서로 차원을 맞추기 위한 임의의 값을 λ 이라 할 때, 이전 연구에서 두 단백질 M, N에 대한 기존 형태기술자 (S) 는 식(1)과 같다.

$$R_{X \text{ and } Y} = ((R(M_{X-Ylength}) - R(N_{X-Ylength}))^2) \tag{1}$$

$$R_{X \text{ and } Z} = (R(M_{X-Zlength}) - R(N_{X-Zlength}))^2$$

$$Histogram_{M \text{ and } N} = (Histogram(M(Distance_{C-V})) - Histogram(N(Distance_{C-V})))^2$$

$$S_{M \text{ and } N} = \sum (R_{X \text{ and } Y} + R_{X \text{ and } Z} + \lambda Histogram_{M \text{ and } N})$$

본 연구에서는 식(1)에서 표현된 지역적 지표와 전역적 지표에 미지의 계수를 할당하여 이를 새로운 형태기술자 (C)로 식(2)와 같이 정의하였다.

$$C_{M \text{ and } N} = \sum (\alpha R_{X \text{ and } Y} + \beta R_{X \text{ and } Z} + \lambda Histogram_{M \text{ and } N}) \tag{2}$$

각 특성에 할당된 계수를 선형 회귀분석을 통해 식(3)과 같이 최적화하였으며 이를 통해 최종적으로 최적화 형태기술자 (C)를 확정하였다.

$$X = [R_{X \text{ and } Y}, R_{X \text{ and } Z}, \text{Histogram}_{M \text{ and } N}^{(3)}]$$

$$Y = [RMSD]$$

$$Y = X w$$

$$w = (X^T X)^{-1} X^T Y$$

단백질 구조 동등성 평가를 위한 형태기술자를 최적화를 위해 OLS 모델을 적용하였으며 의존형 변수 (Y) 값으로 단백질 분자 모델로부터 계산된 RMSD 값을 사용하였다. 표 2에 선형 회귀분석 결과가 나타나 있다.

표 2. 선형 회귀분석 결과

	Coef.	std err	t	P> t
$R_{X \text{ and } Y}$	7.76	0.301	25.754	0.0000
$R_{X \text{ and } Z}$	9.19	0.412	22.285	0.0000
$\text{Histogram}_{M \text{ and } N}$	5.59	0.974	5.736	0.0000

$R_{X \text{ and } Y}$, $R_{X \text{ and } Z}$, $\text{Histogram}_{M \text{ and } N}$ 변수의 계수 값이 선형 회귀분석 결과 확인되었으며 p value의 경우 모든 변수에 대해 통계적으로 유의한 값으로 확인되었다. 선형 회귀분석을 통해서 최적화된 계수들을 바탕으로 기존 형태기술자 기반 단백질 구조 동등성 평가 함수 (C)를 식 (4)와 같이 확정하였다.

$$C_{M \text{ and } N} = \sum(7.76R_{X \text{ and } Y} + 9.19R_{X \text{ and } Z} + 5.59\text{Histogram}_{M \text{ and } N}) \quad (4)$$

4. 최적화 형태기술자 기반 구조 동등성 평가

동일 그룹에 속하는 단백질 표면 모델 사이에서 계산된 형태기술자 값들에 대한 평균 및 표준편차, 최대값을 확보한 후 서로 다른 그룹 간 차이를 비교하고 박스 플롯(Box plot)으로 형태기술자 값들의 분포도를 평가하였다.

최적화 형태기술자에 대한 분포 값을 이전 형태기술자와 비교하여 성능 개선 여부를 확인하였으며 다른 형태기술자인 저니크 형태기술자 기반 동등성 평가 방법과 비교하여 성능을 검증하였다.

이전 형태기술자, 최적화 형태기술자, 저니크 형태기술자로부터 계산된 동등성 평가 결과 값의 분포가 그림 1에 RMSD로부터 계산된 결과 값과 비교하여 나타나 있다.

저니크 형태기술자는 동일 그룹 내에서의 분포와 서로 다른 그룹 간의 분포를 구분할 수 없음을 확인하였고 최적화 형태기술자의 경우 기존 형태기술자 보다 동일 그룹 내에서의 분포와 서로 다른 그룹 간의 분포를 조금 더 잘 구분하는 것으로 나타났다.

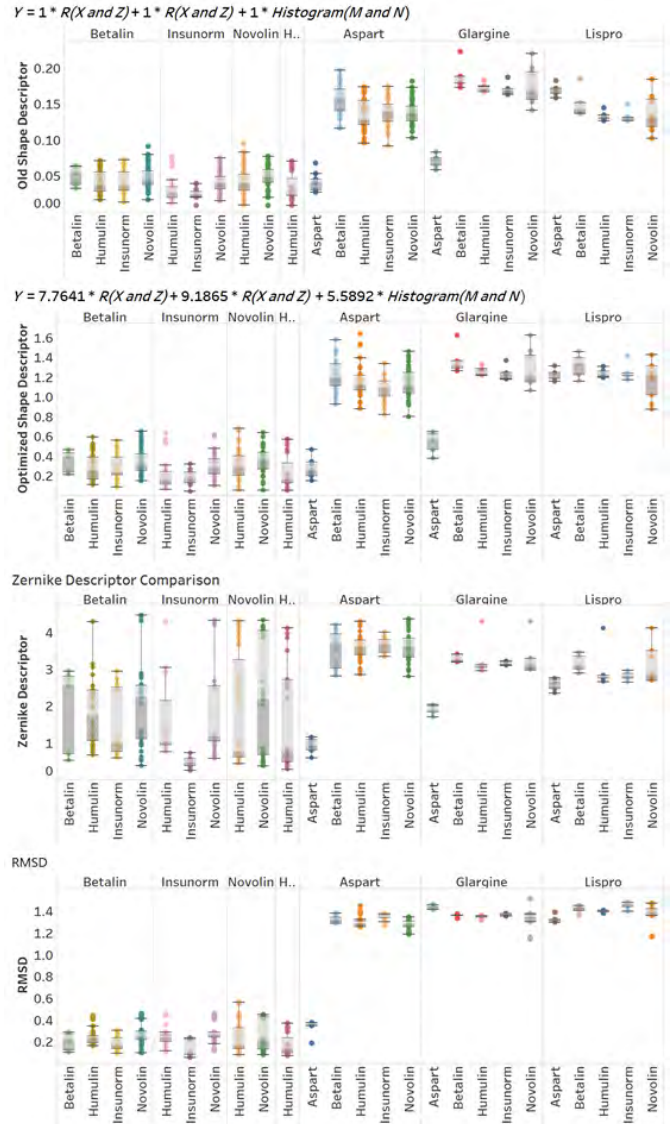


그림 1. 형태기술자 별 인슐린 모델 사이에서의 단백질 구조 동등성 평가 결과

좀 더 정확하게 비교하기 위해 동일 그룹 내에서의 결과 (Same), 유사한 그룹 사이에서의 결과 (Homology), 그리고 다른 그룹 사이에서의 결과 (Different) 값들의 분포를 비교하였으며 그 결과가 그림 2에 나타나 있다.

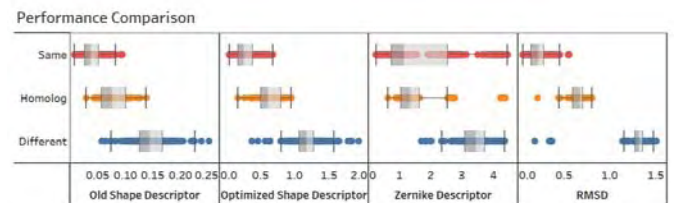


그림 2. 형태기술자 별 단백질 그룹별 구조 동등성 평가 결과 비교

분자 구조 모델로부터 RMSD의 계산을 통해 동등성 평가 값의 분포를 비교한 결과 확연하게 세 개의 그룹에서 차이가 남을 확인하였다. 하지만 저니크 형태기술자의 경우 세 그룹을 전혀 구분할 수 없었다. 최적화 형태기술자의 경우 Same 그룹과 Different 그룹을 확실히 구분할 수 있었으며 Homolog 그룹과 Different 그룹도 어느 정도 구분할 수 있는 것으로 확인되어 기존 형태기술자보다 성능이 개선되었음을 확인하였다.

5. 결론

본 연구에서는 표면 구조 모델에 적용하여 단백질의 구조적 동등성을 평가할 수 있는 최적화 형태기술자를 설계하는 제시하였다. 이전 연구에서 제시된 기존 형태기술자에 대한 최적화를 통해 성능이 개선된 형태기술자를 확보하였고 이를 통해 고해상도 구조 모델에 일반적으로 적용하는 RMSD 방법과 유사하게 서로 다른 표면 구조를 가지는 단백질을 구별할 수 있음을 확인하였다.

이러한 결과로부터 본 연구에서 제시하는 형태기술자는 SAXS 분석과 같이 저해상도 단백질 표면 모델을 확보할 수 있는 분석에 적용하여 단백질의 구조적 동등성을 판별할 수 있는 기반을 제공할 수 있을 것으로 예상된다. 특히 단백질의약품 생산 배치 사이에서의 구조적 동등성 평가 및 생산 공정 변경에 따른 변경 전, 변경 후 생산된 단백질의약품에 대한 구조적 동등성 평가에 활용될 수 있을 것이다.

참고문헌

- [1] K.-E. Song, J. Byeon, D.-B. Moon, H.-H. Kim, Y.-J. Choi, and J.-K. Suh. (2014) Structural Identification of Modified Amino Acids on the Interface between EPO and Its Receptor from EPO BRP, Human Recombinant Erythropoietin by LC/MS Analysis. *Mol. Cells*, 37(11), 819-826.
- [2] G. Walsh. (2010) Post-translational modifications of protein biopharmaceuticals. *Drug Discovery Today*. 15, 773-780.
- [3] S.-H. Chun, Y.-J. Choi, and J.-K. Suh. (2018) Structural Similarity Evaluation between Proteins Based on 3D Model Shape Analysis. *Journal of Knowledge Information Technology and Systems*, 13(1), 87-103.
- [4] P.C. Gross and M. Zeppezauer. (2010) Infrared spectroscopy for biopharmaceutical protein analysis, *J. Pharm. Biomed. Anal.* 53, 29-36.
- [5] A.G.. Kikhney and D.I. Svergun. (2015) A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Letters*, 589, 2570-2577.
- [6] V. Venkatraman, Y. Yang, L. Sael, and D. Kihara. (2009) Protein-protein docking using region-based 3D Zernike descriptors. *BMC Bioinform.* 10, 407.
- [7] D. Kihara, L. Seal and J. Esquivel-Rodriguez. (2011) Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking. *Curr. Protein. Pept. Sci.* 12(6), 520-30.
- [8] D. Duhovny, R. Nussinov, and H.J. Wolfson. (2002) Efficient unbound docking of rigid molecules. *Proceedings of the Second International Workshop on Algorithms in Bioinformatics*. SpringerVerlag: London, UK. pp.185-200.
- [9] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H.J. Wolfson.. (2005) Geometry-based flexible and symmetric protein docking. *Proteins:Struct. Funct. Bioinform.* 60(2), 224-231.