Journal of Knowledge Information Technology and Systems

ISSN 1975-7700

http://www.kkits.or.kr

Structural Similarity Evaluation between Proteins Based on 3D Model Shape Analysis

Sung-Hwan Chun[1], **Yoo-Joo Choi**[1,2], **Jung-Keun Suh**[*1,2]

[1]*Division of Newmedia Content, Seoul Media Institute of Technology*
[2]*Immersive Media Lab, Seoul Media Institute of Technology*

ABSTRACT

Over 130,000 protein structures are solved and deposited in the PDB with molecular structural models of high resolution determined by X-ray diffraction or NMR studies. Those conventional methods for structural determination and similarity comparison are not applicable for therapeutic proteins due to their own technical limitations. However, assessing structural comparability for therapeutic proteins is critical during biopharmaceutical development or manufacturing processes. Currently, those assessments were done by various spectroscopic methods but those methods are not giving specific structural information. Moreover, recent developments such as single angle X-ray scattering (SAXS) techniques can provide low-resolution structural models for proteins more easily than those for X-ray diffraction or NMR studies. These enable us to develop fast and reliable method for the similarity assessment of low-resolution surface models of therapeutic proteins using geometric shape descriptor. Our descriptor consists of two features, local and global. The local feature calculates the distance distribution for each vertex from the center in 510 bins. For global feature, the ratio, x to y axis and x to z axis from the surface model of proteins were determined. A geometric shape descriptor is then constructed by combining local and global features with weights. We applied this geometric shape descriptor to assess structural similarity for the therapeutic protein, insulin models. Our geometric shape descriptor can clearly classify human insulin models and insulin analog models which having locally different structures. The performance of the geometric shape descriptor was evaluated by comparing to the conventional method of the root mean square deviation measure (RMSD) which computes the minimum average distance between the backbones of superimposed. The result shows that the performance of the global shape descriptor is comparable to that of RMSD. This provides potential applications to classify protein structure and compare low-resolution protein structures for EM and SAXS.

KEYWORDS: Protein similarity evaluation, Protein surface model, Shape descriptor, root mean square deviation measure (RMSD), Small-angle X-ray scattering (SAXS)

ARTICLE INFO: Received 2 January 2018, Revised 28 January 2018, Accepted 8 February 2018.

[*]Corresponding author is with the Division of Newmedia Content, Seoul Media Institute of Technology, 99 Hwagok-ro 61-gil Gangseo-gu Seoul, 07590, Korea.
E-mail address: jksuh@smit.ac.kr

1. 서 론

단백질은 두 개 이상의 단위구조가 연결되면서 형성된 이중체의 구조를 가지며 세포 내에서 구별된 기능을 수행하게 된다. 현재 43,000 여개의 단백질 서열에 대한 3차구조가 규명되어 PDB(Protein Data Bank) 웹사이트에 공개되어 있지만,[24] 아직 많은 단백질 구조가 밝혀지지 않은 상태이다. 단백질 3차구조 분석은 X-ray 결정 방식이나 NMR 분광기 분석과 같은 실험적인 방법으로 수세기 동안 진행되어 원자 수준에서의 고해상도 단백질 3차 구조를 파악하고 있다. X-ray 결정 방식은 결정을 만들 수 없는 단백질의 경우 3차구조 분석이 불가능하며 NMR 분광기 분석은 단백질 결합 등의 이유로 단백질의 크기가 커지면 분석을 못하는 한계가 있다. 또한 분석이 가능하더라도 단백질 3차구조 분석을 위해 많은 시간 및 비용이 든다. 이러한 어려움 때문에 수치적 계산을 통한 단백질 구조를 예측하는 방법들이 오랫동안 연구되어 왔다. 단백질 구조 예측은 1994년부터 CASP (The Critical Assessment of protein Structure Prediction)라고 하는 community-wide experiment를 통해서 이루어지고 있다. 2016년 개최된 CASP2016에서 목표 단백질 중 하나인 Cysteine synthase A에 대한 단백질 구조 예측 모델이 (CASP2016: T0861-D1) <그림 1> (a) 에 나타나 있으며 목표 단백질에 대한 X-ray 결정 분석을 통해 얻어진 단백질 3차 구조 모델이 (PDB:5J5V) <그림 1> (b)에 나타나 있다[1].

예측 모델과 실험 모델 사이의 유사성은 두 단백질 모델을 Fig 2.와 같이 맞춘 후 각 원자 사이의 거리를 통해 두 단백질 사이에서의 차이를 계산하는 RMSD 값을 통해 평가할 수 있다. RMSD 값 0.5를 기준으로 0.5 이하로 나타나는 경우, 예측 모델과 실험 모델 사이에 구조적 유사성이 높은 것으로 확인된다. <그림 2>의 경우, RMSD 값은 0.28로 확인되었다.

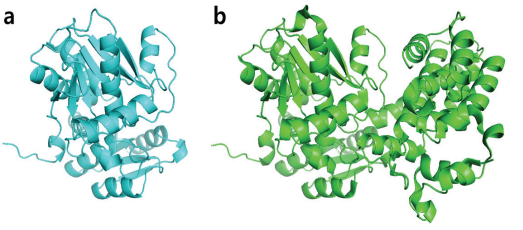

그림 1. CASP2016의 단백질 모델
Fig 1. Protein models from CASP2016

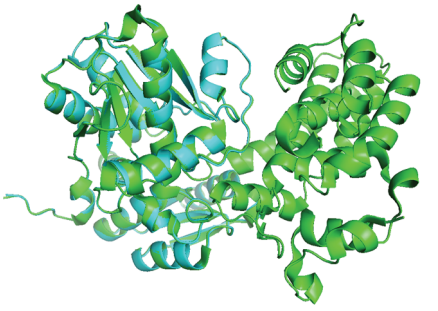

그림 2. 정렬된 단백질 모델 구조
Fig 2. Protein model structure with alignment

단백질 구조 예측 방법은 단백질 구조 자체의 모델 시각화를 통해 서로 다른 단백질 사이의 상보성을 규정하기 위해 사용되고 있으며 CASP와 유사한 community-wide experiment 활동인 CAPRI (Critical Assessment of Prediction of Interactions)를 통해 이루어지고 있다[2].

최근에 와서 SAXS 분석을 통해 저해상도 단백질 구조 모델을 실험적으로 확보할 수 있다[3]. SAXS 분석을 통해 확보한 X-ray 산란 데이터의 경우 Ab initio Modeling 및 Rigid Body Modeling이 가능하며 이러한 모델링을 통해 어떠한 사전 정보 없이 단백질에 대한 저해상도를 가지는 표면 모델을 생성할 수 있다. <그림 3>에 저해상도 표면모델이 나타나 있으며 이러한 표면모델은 단백질의 3차 구조에 대한 형태적 특성을 가지고 있다.

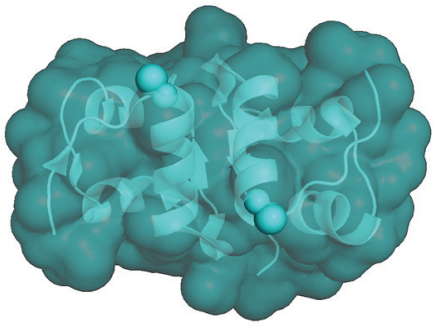

그림 3. 단백질의 저해상도 구조 모델
Fig 3. Low-resolution structural model for a protein

현재까지 X-ray 결정 방식이나 NMR 분광기 분석과 같은 고해상도 구조 분석 데이터를 활용한 단백질 유사도 분석은 많은 연구가 이루어지고 있지만 이에 반해 형태적 특성 기반의 형태 기술자를 활용한 유사성 분석 연구는 제한적으로 이루어지고 있다.

단백질 의약품의 경우 생산된 단백질이 서로 다른 생산 배치 사이에서 구조적으로 동일한 지를 규명해야 한다. 하지만 시간과 비용 및 실험적인 한계로 인해 X-ray 결정 방식이나 NMR 분광기 분석과 같은 고해상도 구조 분석을 통한 3차 구조 동등성 평가가 현실적으로 불가능하다. 이러한 한계를 극복하기 위해 저비용/단시간에 확보할 수 있는 저해상도 3차 구조 (표면 구조) 정보를 활용하여 구조적 유사성을 평가할 수 있는 평가 기술 연구가 필요하다.

이에 본 논문에서는 단백질에 대한 저해상도 모델의 구조적 유사도 평가를 위한 방법론을 제안하였다. 저해상도 구조 모델의 구조적 유사성 평가를 위해 형태적 특성 기반의 형태 기술자(shape descriptor)를 구성하였으며 형태 기술자는 3D 모델에 대한 지역적 또는 전역적 특성 데이터를 기반으로 한다.

저해상도 구조 모델의 형태 데이터를 활용하여 단백질 구조를 비교하기 위해서는 3D 모델로 시각화된 단백질 구조 형태를 규정짓는 특징을 추출하는 것이 중요하다. 본 논문에서는 모델의 입력 상태에 관계없이 구분되는 모델의 특성값을 추출하여 형태 기술자를 새롭게 제안하였다. 제안된 형태 기술자의 효용성을 검증하기 위해 X-ray 결정 방법으로 확인된 human insulin 단백질의 3차구조 모델을 사용하였으며 고해상도 단백질 3차 구조 정보를 활용하여 유사성을 판별한 RMSD 값과 비교하여 제안된 방법의 실효성을 검증하였다.

본 논문은 2절에서 3차원 형상 분석에 대한 관련 연구를 기술하고 3절에서는 본 연구진이 제안하는 형태기술자를 기술하였다. 4절에서는 인슐린 단백질 구조에 본 연구진이 제안하는 형태기술자를 적용하여 그 성능을 RMSD 분석 결과와 비교하여 평가하였다.

2. 관련 연구

2.1 3차원 형상 데이터 비교 관련 연구

3D 모델 기반 검색 및 3D 프린터 등 활용 요구가 많아짐에 따라 3D 모델 형상 비교에 대한 효과적인 방법의 필요성이 증가 하고 있다.

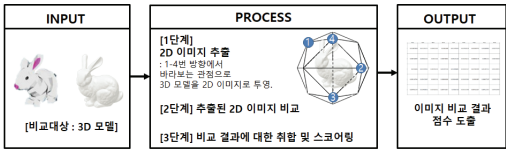

그림 4. 시각 기반 비교방법 순서도
Fig 4. Flow of View-based retrieval

3D 모델 형상 비교는 접근 방법에 따라 크게 시각 기반 비교(view based retrieval)와 모델 정보 기반 비교(geometric based retrieval)로 나뉜다. 시각

기반 비교 접근 방법은 <그림 4>와 같이 3D 모델을 바라보는 방향에 따라 다르게 나오는 수많은 2D 이미지를 추출하여 서로 비교하는 방법이다[4-7].

모델 정보 기반 비교 방법은 형태 기반 접근법(Shape-based approach)과 토폴로지 기반 접근(Topology-based approach)으로 구분할 수 있다. 형태 기반 접근법은 정점(vertex), 변(edge), 면(facet), 곡률(Curvature) 과 같은 모델 데이터를 추출하여 형태 특성이 반영된 형태 기술자를 만들어 비교하는 방법이다[8,-12]. 토폴로지 기반 접근법 은 입력 받은 모델 내의 정점 간의 거리 및 좌표로 그래프 알고리즘을 활용하여 두 모델 간 비교를 한다[13]. 형태 기반 접근법과 달리 입력 받은 모델들의 변환 상태에 영향을 받지 않기 때문에 더 빠른 장점이 있다. 하지만, 토폴로지 기반 비교 방법은 형태적 특성에 대한 데이터를 기반으로 하지 않기 때문에 모델 데이터가 나뉘게 되는 경우 서로 다른 모델로 인식하는 경우가 생긴다. A'가 A 모델의 부분이 되는 모델이고, A 모델과 B 모델이 동일한 모델일 때 A'가 B의 부분이라는 것을 토폴로지 기반 비교 방법으로 구분하는 것에 어려움이 있다.

2.1.1. 시각 기반 비교 접근 (view based retrieval)

3D 모델을 바라보는 방향에 따라 투영된 2D 이미지들을 활용하여 서로 다른 3D 모델들을 비교하는 방법이므로 2D 이미지를 생성하는 단계와 생성된 2D 이미지를 비교하는 단계로 나뉜다. 2D 이미지를 생성하는 단계에서는 고려되어야 할 부분은 생성되는 투영된 이미지의 개수와 품질이다. 바라보는 방향의 수가 많을수록 3D 모델 비교 결과의 정확도가 높아지지만 계산 소요 시간 또한 비례하여 증가한다.

초기에는 주축 3개와 주축을 보조하는 4개의 보조축을 정하여 총 7개의 방향에서 투영된 이미지를 바탕으로 비교하였다[4]. 비교 대상이 되는 2D 이미지에 3D 모델의 Z-축 깊이를 반영하고자 3D 모델의 외곽선 상자(Bounding box)를 만들어 8개의 면에서 바라보는 깊이 이미지를 추가하여 비교하였다[5]. 주축을 기준으로 모델들의 회전 상태를 맞추게 되면, 주축이 정확하게 회전 정도를 표현하지 못하는 경우 모델들의 유사성 판단이 어렵다. 서로 동일한 3D 모델이라 하더라도 모델의 회전 정도 및 크기, 이동 위치에 따라 투영된 2D 이미지는 다르게 나온다. 회전 정도 및 크기, 이동 위치와 같은 변환 상태를 동일하게 맞추는 것이 매우 중요하다.

시각 기반 비교 접근 방법의 기존 연구들에서 모델들의 변환 상태;회전 정도, 크기, 이동 위치를 동일하게 맞추는 과정이 3D 모델 비교에 있어서 공통적인 단계임을 알게 되었다. 모델들의 변환 상태를 동일하게 하는 과정은 주축을 찾아 서로의 상태를 맞추는 방법[6]과 같이 서로 다른 많은 회전 각도 위치에서 바라본 3D 모델의 투영된 2D 이미지를 가지고 서로의 상태를 맞추는 방법이 있다. 본 연구에서 비교할 단백질 3D 모델 데이터의 경우, 모델의 시각적 유사도가 매우 높고 전체적인 모델이 타원형으로 구성되어 있어 주축을 구하는 방법으로도 서로 다른 단백질 모델 간의 회전 변화 정도를 동일하게 맞출 수 있었다.

본 연구에서는 시각적 비교 접근 방법은 단백질 모델 데이터 비교에 적용할 경우 불필요한 계산이 많을 것으로 판단하여 형태 기반 비교 접근 방법을 사용하였다.

2.1.2. 모델 기반 비교 접근 (geometric based retrieval)

모델 기반 비교 접근 방법은 형태 기반 접근법과 토폴로지 기반 접근법으로 나뉜다. 본 연구에서 중점을 두고 있는 단백질 3D 모델 데이터 비교의 경우 단백질의 기능의 변화가 구조의 변화; 통합과 해제에 의해 일어나기 때문에 단백질 구조의 같은 부분이 연결되면, 같은 모델로 인식해야 한다. 토폴로지 기반 접근법은 구조의 변화에 따라 동일 모델로 평가하지 못하는 한계가 있어 본 연구에서는 모델 기반 비교 접근 방법 중, 형태 기반 접근법에 범위를 한정한다.

형태 기반 접근법은 형태 기술자를 어떻게 정의하느냐에 따라 유사도 측정 결과가 좌우된다. 형태 기술자는 3D 모델의 특정 부분에 있는 지역적 특성을 포함하고 있는 데이터 세트이다. Osada 및 Funkhouser는 랜덤한 정점들 간의 거리, 각도, 면적 및 볼륨을 구하여 형태 기술자를 구성하여 형태의 분산 정도에 따라 모델 간의 유사도를 비교하였다[9]. Liu 등은 일정 거리에 위치한 특정 모양;볼록(convex), 오목(concave), 평평(flat)의 개수에 근거한 3D 히스토그램으로 형태 기술자를 구성하여 제안하였다[11]. 형태 기반 접근법은 비교할 3D 모델에서 비교 대상이 되는 모델과 구분 지을 수 있는 특성을 규정짓고 모델 자체에서 특성 데이터를 반영하는 범위를 어떻게 규정하느냐에 따라 성능 차이가 있음이 나타났다.

2.2 형태기술자를 적용한 단백질 구조 유사성 분석 기법

형태기술자를 활용한 단백질 구조 유사성 분석은 고해상도 구조 분석 데이터를 활용한 단백질 유사도 분석 연구에 비해 제한적으로 이루어지고 있다. 대표적인 형태 기술자는 단백질 모델을 일정한 구역별로 나누어 각 구역 내의 지역적 특성값으로 표현된다.

PatchDock[14,15]에서는 단백질의 형상을 볼록, 오목, 평평한 부분으로 분류한 후, 한 단백질의 표면 삼각형(face triangle)과 비교 대상의 단백질 모델의 한 정점간의 유클리디안(Euclidean) 거리를 바탕으로 한 형태 기술자가 제안 되었다. 이외에도 3D 저니크 형태 기술자(3D Zernike shape descriptor)[16,17], 컨텍스트 형상(Context Shape)을 기반으로 한 형태 기술자[18] 등이 있다.

2.3 기존 단백질 구조 유사성 분석 기법의 한계점

현재까지 X-ray 결정 방식이나 NMR 분광기 분석과 같은 고해상도 구조 분석 데이터를 활용한 단백질 유사도 분석은 많은 연구가 이루어지고 있지만 X-ray 결정 방식이나 NMR 분광기 분석과 같은 실험적인 방법으로 단백질 구조 유사성을 분석하고 비교하는 것은 매우 어렵고 시간과 비용이 많이 소요되는 작업이다. 단백질 의약품의 경우 다른 단백질 보다 더 큰 한계에 직면하게 된다. 일반적으로 단백질 구조 분석을 위해서 적용할 수 있는 X-ray 결정 방식이나 NMR 분광기 분석 등의 방법을 단백질 의약품 생산 과정에서 일상적으로 사용할 수 없다는 한계가 있다. 단백질 의약품에 대한 생산 과정에서 활용되는 분석법은 UV 분석법, 형광 분석법, CD 및 FT-IR과 같은 분광학적 분석법이다. 이러한 분석법은 특이성 문제로 인해 서로 구조가 같은지 다른지에 대한 구조적인 정보를 주지 못하며 특히 3차 구조에 대한 유사성 비교 정보를 제공하지 못하는 한계가 있다.

저해상도 구조 모델 기반 형태 기술자를 활용한 유사성 분석 연구는 고해상도 구조 모델 기반 연구에 비해 제한적으로 이루어지고 있지만 현재까지 알려진 형태기술자들은 매우 비슷한 구조 유사성을 판별하기에는 아직 한계가 있다.

이러한 한계를 극복하기 위해 본 연구에서는 저해상도 3차 구조 정보를 제공하는 형태 기반 모델인 표면 모델을 활용하여 단백질 의약품 3차 구조의 유사성을 평가할 수 있는 평가 기술을 제안하였다.

3. 3차원 형태 분석을 통한 단백질 유사성 분석

3.1 3차원 형태 분석 기법 정의

3차원 형태 분석은 3D 모델링에서 3D 모델 검색 등 많은 분야에서 많이 사용되는 중요한 연구 과제이다. 3차원 형태 분석을 통해 모델 비교 및 검색을 하기 위해서 입력받은 모델들 간의 위치, 회전, 크기 변환 상태를 동일하게 하는 공간 정규화 과정이 있다. 공간 정규화 과정을 거친 후, 3차원 형태에 대한 분석은 지역적 특성 및 전역적 특성을 통해 할 수 있다. 지역적 특성은 곡률, 빛 반사율과 같이 3D 모델의 특정 부분에 집중된 특성을 말한다. 전역적 특성은 정점 간 거리, 모델의 방향성과 같이 전체적인 모델의 윤곽을 나타내는 특성을 말한다. 단백질 데이터의 3차원 형태 분석 비교의 경우, 지역적 특성 보다는 전체적인 구조의 윤곽을 다른 단백질 데이터와 구분할 수 있는 전역적 특성이 더 중요하다. 본 연구에서는 공간 정규화가 필요 없는 특성 데이터에서 전역적 지표와 정점 간의 거리를 이용한 지역적 지표로 형태 기술자를 구성하여 단백질의 유사성 판단을 하였다.

3.2 알고리즘 개요

본 연구에서는 단백질 3D 모델 데이터 비교를 위해 정점 간 유클리디안 거리를 활용한 히스토그램과 입력 받은 3D 모델의 전역적 특성을 나타내는 X-Y축 및 X-Z축 비율을 형태 기술자로 사용하고 있다. 전체 알고리즘의 순서는 <그림 5> 와 같다.

그림 5. 알고리즘 프로세스 개요도
Fig 5. Algorithm Process Outline

1단계 : 특성 데이터 추출(Feature extraction)

1단계에서 입력받은 모델 데이터의 변환 상태를 모두 맞춘 다음, 형태 기술자에 사용될 특성 데이터를 추출한다. 본 연구에서는 단백질의 3D 모델 데이터를 구성하고 있는 모든 정점들과 모델 데이터의 중심점 사이의 유클리디안 거리를 구하여 510개의 방(bin)으로 구성된 히스토그램을 만들었다. 또한, 모델의 x축, y축, z축 최대 길이를 구하여 X-Y축의 비율과 X-Z축의 비율을 각각 구하여 모델의 방향성에 대한 추가 지표로 추출하였다.

2단계 : 유사성 평가 함수

X-ray 결정 방법으로 나온 데이터의 결과를 단백질 비교에 주로 사용되는 평균 제곱근 편차(RMSD)를 사용하여 분석한 결과와 제안된 형태 기술자를 사용하여 나온 결과를 박스 플롯 그래프로 도식화 하여 비교하였다. RMSD 및 형태 기술자 값의 분포에 대한 통계적 유의성을 분석하기 위해 t-검정을 수행하였다.

3.3 정점 기반 형태 특성 묘사 (Shape Descriptor)

단백질 3D 모델 데이터를 묘사하기 위해 본 연

구에서는 두 가지 특성 데이터를 사용하였다. 단백질 모델의 외곽 형태 특성을 알기 위해 모델의 X-Y축 및 X-Z축 비율을 전역 지표로 사용하였고, 모델의 중심점과 모든 정점 간의 유클리디안 거리들의 히스토그램을 지역 지표로 사용하였다.

3.3.1. 전역 지표 (Global Feature)

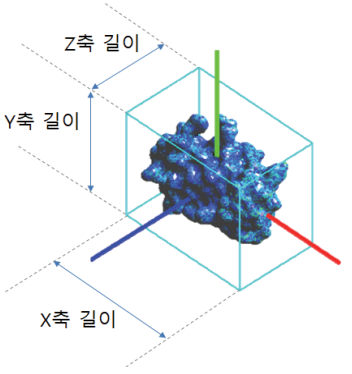

그림 6. 단백질 모델의 X, Y, Z축 길이
Fig 6. X, Y, Z length of Protein Model

입력 받은 단백질 데이터의 X축에 대한 Y축의 방향성과 X축에 대한 Z축의 방향성을 식 (1), 식 (2)와 같이 구한다 (그림 6)

$$Ratio\ Y\ by\ X = (Max_y - Min_y) / (Max_x - Min_x) \quad (1)$$

$$Ratio\ Z\ by\ X = (Max_z - Min_z) / (Max_x - Min_x) \quad (2)$$

3.3.2. 지역 지표 (Local Feature)

지역 지표로 사용된 모델의 중심과 정점 간의 유클리디안 거리는 식 (3) 와 같이 구할 수 있다. $C=(C_x, C_y, C_z)$ 는 모델 중심 좌표이고 $V=(V_x, V_y, V_z)$ 는 단백질 모델의 구성하고 있는 정점 좌표이다. $D_{V with C}$는 유클리디안 거리이다.

$$D_{V with C} = \sqrt{(C_x - V_x)^2 + (C_y - V_y)^2 + (C_z - V_z)^2} \quad (3)$$

실험 대상의 단백질 데이터들의 최대 거리 값을 구한 결과 최대 거리로 정하였다. 히스토그램의 방(bin) 개수는 510개로 하였으며, 히스토그램은 아래와 같은 순서로 구하였다. 히스토그램을 만들기 전 모델 데이터의 좌표들은 모두 정규화 과정을 통해 값을 축소하였다.

Algorithm 1. 유클리디안 거리 히스토그램 구하기
Input :
nVert : 모델 데이터의 총 개수
Range : 히스토그램 한 방(bin)의 값 범위
C : 모델 데이터의 중심 3차원 벡터
Vi : 모델 데이터를 구성하는 3차원 벡터
D : C와 Vi 사이의 유클리디안 거리
γ : 모델 내 버텍스 간의 최대 거리
Output:
DistHistogram : 유클리디안 거리 히스토그램

```
1 Range = γ / 510
2 DistHistogram[510] ← 0
3 for( i = 0, i < nVert, i++){
4   D ← √((C_x − V_x)^2 + (C_y − V_y)^2 + (C_z − V_z)^2)
5   idx ← D / Range
6   DistHistogram[idx]++
7 }
```

3.4 형태 특성 기반 유사성 평가 함수

유사성 평가 함수는 아래와 같이 512차원으로 구성된다.

(1) X축 대한 Y축 길이 비율 (Ratio of Y to X)
: 1차원

(2) X축 대한 Z축 길이 비율 (Ratio of Z to X)
: 1차원

(3) 모델 중점과 각 정점 간 거리에 대한 히스토그램 : 510차원(Histogram of distance between vertex and center point)

입력된 단백질 모델 M, N에 대해 각 단백질 모델의 X-Y축 길이 비율을 $R(M_{X-Ylength})$, $R(N_{X-Ylength})$, X-Z축에 대한 비율을 $R(M_{X-Zlength})$, $R(N_{X-Zlength})$, 모델 중점과 각 정점 간의 거리에 대한 히스토그램을 $Histogram(M(Distance_{C-V}), Histogram(N(Distance_{C-V})$, 서로 차원을 맞추기 위한 임의의 값을 λ이라 할 때, 두 단백질 M, N에 대한 유사성 평가 함수 (S)는 식 (4)와 같이 정의된다.

$$R_{X\text{ and }Y}=((R(M_{X-Ylength})-R(M_{X-Ylength}))^2$$
$$R_{X\text{ and }Z}=(R(M_{X-Zlength})-R(M_{X-Zlength}))^2$$
$$Histogram_{M\text{ and }N}=(Histogram(M(Distance_{C-V})-Histogram(N(Distance_{C-V}))^2$$
$$S_{M\text{ and }N}=\sum(R_{X\text{ and }Y}+R_{X\text{ and }Z}+\lambda Histogram_{M\text{ and }N}) \quad (4)$$

동일 그룹의 단백질 모델들의 유사도 평가 값들의 평균 및 표준편차, 최대값을 구해 그룹 간 차이를 비교하고 박스 플롯(Box plot)으로 유사도 평가 값들의 분포도를 평가하였다. 동일 그룹 내에서 단백질 모델 간의 형태 특성을 비교하여 동일 그룹 안의 단백질 유사도를 평가하였다.

4. 실험 결과

4.1 실험 방향

본 연구에서는 실험 대상 단백질로 대표적인 단백질의약품의 하나인 인슐린을 선정하였으며 X-ray 결정 방법으로 고해상도 구조 모델이 확보된 휴먼 인슐린(Human insulin) 모델과 인슐린 아날로그인 Aspart 인슐린의 모델을 사용하였다[19]. 인슐린은 21개의 아미노산으로 구성된 a-chain과 30개의 아미노산으로 구성된 b-chain이 이황화결합에 의해 연결된 구조를 가진다 <그림 7>.

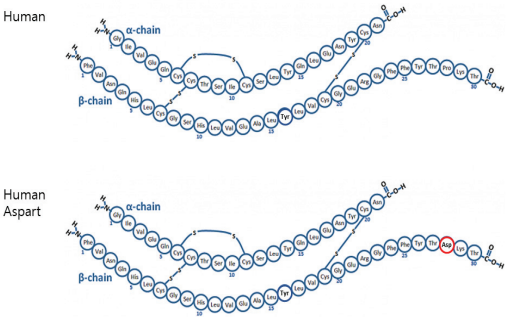

그림 7. 휴먼 인슐린과 Aspart의 구조
Fig 7. Structure of human insulin and aspart

실험 대상이 되는 휴먼 인슐린은 28개 및 Aspart 인슐린은 5개 모델로 구성되어 있으며 이 중, 휴먼 인슐린은 인슐린 단백질의약품 제조사에 따라 4개의 다른 그룹 (Batalin, Humulin, Insunorm, Novolin)으로 구분된다.

Dynamic light scattering 분석, SAXS 분석 및 NMR 분석을 통해 휴먼 인슐린 사이에서는 구조가 유사함이 확인되었으며 휴먼 인슐린과 Aspart 인슐린 사이에서는 구조적 차이가 나타남이 확인되었다[19]. 서로 다른 그룹에 속해 있는 휴먼 인슐린 사이에서의 구조적 유사도를 고해상도 3차구조 모델 분석에 주로 사용되는 RMSD (식(5)) 방법으로 비교하였다. PyMOL 소프트웨어를 활용하여 휴먼 인슐린 28개 및 Aspart 인슐린 5개 모델들 중 2개씩 alignment tool을 활용하여 align시킨 후 쌍별 비교를 통해 RMSD 값을 계산하였다.

$$RMSD=\sqrt{\frac{1}{n}\sum_{i=1}^{n}d_i^2} \quad (5)$$

d_i: i번째 쌍의 두 원자 사이의 거리 (단, n개의

동등한 한 쌍의 원자 비교 그룹이 있는 경우에 한함)

본 연구에서 제안하는 형태기술자를 적용하여 휴먼 인슐린 28개 및 Aspart 인슐린 5개 모델 사이에서의 쌍별 비교를 통해 유사성 평가함수 값을 계산하였다.

4.2 인슐린의 유사성 비교 결과

공간 정규화가 된 인슐린 데이터를 제안된 형태기술자를 사용하여 확보한 유사도 측정값과 RMSD 값의 분포를 비교하여 본 연구에서 제안하는 형태기술자의 유효성을 검증하였다.

4.2.1 휴먼 인슐린 사이의 유사성 비교

휴먼 인슐린의 경우 Betaline 제품 중 두 개 고해상도 단백질 3차구조 모델을 (PDB ID: 4EWZ, 4EX0) 대표로 선정하여 비교한 결과가 Fig 8에 나타나 있다.

고해상도 모델 상에서 두 개의 구조를 나란히 맞추어 비교한 결과 구조가 매우 잘 겹치게 나타나는 것을 확인하였다 <그림 8>.

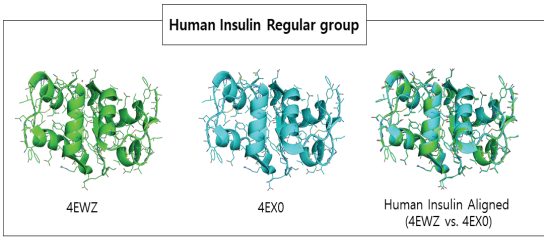

그림 8. 휴먼 인슐린의 분자모델 정렬 결과

Fig 8 Alignment of 3D structure between Human Insulins

휴먼 인슐린 Betaline 제품의 단백질 고해상도 구조 모델에서 (PDB ID: 4EWZ, 4EX0) 계산된 형태 모델이 <그림 9>에 나타나 있으며 형태 측면에서도 두 개의 구조가 겹치게 나타나는 것을 확인하였다.

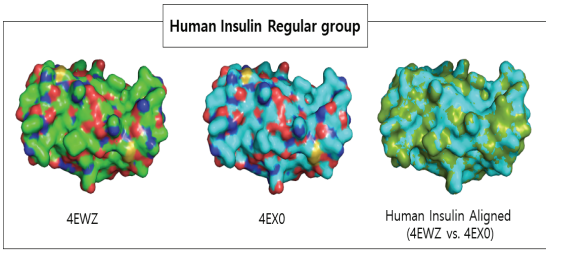

그림 9. 휴먼 인슐린의 표면모델 정렬 결과

Fig 9. Alignment of surface structure between Human Insulins

휴먼 인슐린 사이에서의 구조적 유사성을 분석하기 위해 휴먼 인슐린 단백질의 고해상도 모델 사이에서의 RMSD 값을 확인하였으며 휴먼 인슐린 네 가지 제품인 Betalin, Humulin, Insunorm, Novolin 사이에서 계산된 RMSD 값의 평균과 표준편차가 <표 1>에 나타나 있다. RMSD 값은 일반적으로 0.5이하로 계산되면 구조적 유사성이 높은 것으로 판단하는데 휴먼 인슐린 단백질 사이에서의 RMSD 분석 결과 모두 0.5이하로 확인되어 휴먼 인슐린 단백질 사이에서 3차구조가 유사함을 확인하였으며 이는 보고된 결과와 동일하였다 [19].

본 연구에서 제안한 형태 기술자를 사용하여 휴먼 인슐린 표면 구조 사이에서의 형태 기술자 값을 확인하였으며 휴먼 인슐린 네 가지 제품인 Betalin, Humulin, Insunorm, Novolin 사이에서 계산된 형태 기술자 값의 평균과 표준편차가 <표 2>에 나타나 있다. 휴먼 인슐린 사이에서 형태 기술자 평균값은 0.000673~0.001106 사이에서 확인되었다.

휴먼 인슐린 단백질 사이에서 계산된 RMSD 및 형태 기술자 값의 분포를 박스 플롯 그래프로 시각화하였다 <그림 10>. RMSD 및 형태 기술자 값의 경우 모두 휴먼 인슐린 단백질 제품 사이에서 중

앙값과 분포 정도가 차이가 있음을 확인하였다. 또한 동일한 제품 사이에서도 분포의 형태가 고르지 않으며 이는 n 수가 적기 때문으로 예측된다.

표 1. 휴먼 인슐린 사이에서의 RMSD 평균 및 표준편차
Table 1. Avg. and std. deviation of RMSD between human insulins

		Betalin	Humulin	Insunorm	Novolin
Betalin	n	10			
	평균	0.2045			
	표준편차	0.0698			
Humulin	n	40	28		
	평균	0.2399	0.1661		
	표준편차	0.0735	0.0982		
Insunorm	n	30	48	15	
	평균	0.1875	0.2464	0.1277	
	표준편차	0.0690	0.0838	0.0711	
Novolin	n	45	72	54	36
	평균	0.2690	0.2272	0.2770	0.2617
	표준편차	0.0949	0.1331	0.0991	0.1394

표 2. 휴먼 인슐린 사이의 형태 기술자 평균 및 표준편차
Table 2. Avg. and std. deviation of shape descriptor between human insulins

		Betalin	Humulin	Insunorm	Novolin
Betalin	n	10			
	평균	0.001106			
	표준편차	0.000346			
Humulin	n	40	28		
	평균	0.000959	0.000997		
	표준편차	0.000707	0.000914		
Insunorm	n	30	48	15	
	평균	0.000748	0.000765	0.000435	
	표준편차	0.000359	0.000745	0.000142	
Novolin	n	45	72	54	36
	평균	0.000867	0.000848	0.000673	0.000825
	표준편차	0.000387	0.000629	0.000220	0.000352

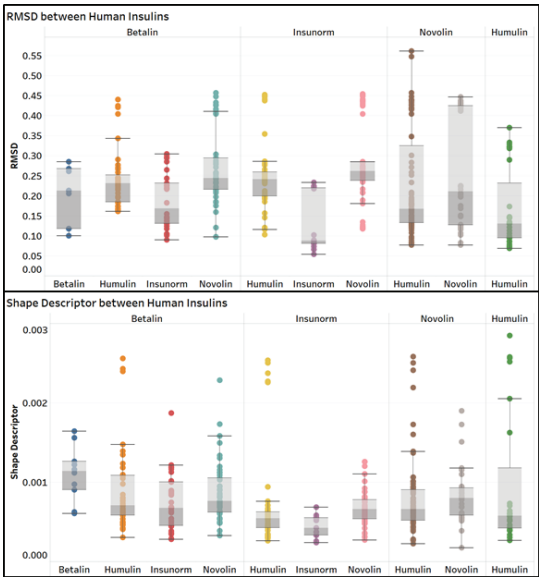

그림 10. 휴먼 인슐린에 대한 RMSD와 형태 기술자 값의 박스플롯 그래프
Fig 10. Box-Plot analysis of human insulin based on RMSD and Shape descriptor

표 3. 휴먼 인슐린 사이에서의 RMSD 값의 평균과 분산값
Table 3. Avg. and deviation of RMSD between human insulins

인자의 수준	관측수	합	평균	분산
Betalin to Humulin	40	9.595	0.239875	0.005409
Betalin to Insunorm	30	5.624	0.187467	0.004764
Betalin to Novolin	45	12.106	0.269022	0.009004
Humulin to Novolin	72	16.357	0.227181	0.017706
Insunorm to Humulin	48	11.826	0.246375	0.007016
Insunorm to Novolin	54	14.958	0.277	0.009812

동일한 제품 사이에서의 결과를 제외한 후 휴먼 인슐린 단백질 사이에서 계산된 RMSD 및 형태 기술자 값의 분포에 대한 통계적 유의성을 분석하기 위해 분산 분석을 수행하였다. 휴먼 인슐린 제품 사이에서 확인된 RMSD 값에 대한 분산분석 결과, 평균과 분산값이 <표 3>에 나타나 있다.

RMSD 값에 대한 분산 분석 결과가 <표 4>에 나

타나 있다. P-값이 0.05 보다 작게 나와 휴먼 인슐린 단백질 제품 사이에서 확인되는 RMSD 분포의 차이는 통계적으로 유의함을 확인하였다.

표 4. 휴먼 인슐린 사이에서 RMSD 값의 분산 분석
Table 4. Variance analysis of RMSD between human insulins

변동의 요인	제곱합	자유도	제곱평균	F 비	P-값	F 기각치
처리	0.2041	5	0.04084	4.0517	0.0014	2.245902
잔차	2.8521	283	0.01008			
계	3.0563	288				

휴먼 인슐린 제품 사이에서 확인된 형태 기술자 값에 대한 평균과 분산값이 <표 5>에 나타나 있으며 분산 분석 결과가 <표 6>에 나타나 있다.

표 5. 휴먼 인슐린 사이에서 형태 기술자 값의 평균과 분산값
Table 5. Avg. and deviation of shape descriptor between human insulins

인자의 수준	관측수	합	평균	분산
Betalin to Humulin	40	0.038371	0.000959	4.99E-07
Betalin to Insunorm	30	0.02245	0.000748	1.29E-07
Betalin to Novolin	45	0.039019	0.000867	1.5E-07
Humulin to Novolin	72	0.061035	0.000848	3.96E-07
Insunorm to Humulin	48	0.036737	0.000765	5.55E-07
Insunorm to Novolin	54	0.036356	0.000673	4.83E-08

표 6. 휴먼 인슐린 사이에서 형태 기술자 값의 분산 분석
Table 6. Variance analysis of Shape descriptor between human insulins

변동의 요인	제곱합	자유도	제곱평균	F 비	P-값	F 기각치
처리	2.36E-06	5	4.72E-07	1.5427	0.1764	2.24590
잔차	8.66E-05	283	3.06E-07			
계	8.89E-05	288				

P-값이 0.05 보다 크게 확인되어 서로 다른 휴먼 인슐린 단백질 제품 사이에서 확인되는 형태 기술자 값의 분포는 차이가 없음을 통계적으로 확인하였다. 이러한 형태 기술자 값의 분산 분석 결과는 통계적으로 유의하게 확인된 RMSD 값의 분산 분석 결과와는 다르게 나타났다.

4.2.2 휴먼 인슐린과 인슐린 아나로그 Aspart 인슐린 사이의 비교

실험적 분석을 통해 구조적으로 차이가 나는 것으로 확인된 휴먼 인슐린과 인슐린 아나로그 Aspart 인슐린 사이에서의 구조를 RMSD 값과 본 연구에서 제안한 형태기술자 값으로 유사성을 확인하였다.

휴먼 인슐린의 경우 Betaline 제품과 (4EWZ) 인슐린 아나로그인 Asprt 제품 (4GBC)에 대한 고해상도 단백질 3차구조 모델을 (PDB ID: 4EWZ, 4EX0) 대표로 선정하여 비교한 결과가 <그림 11>에 나타나 있다. 고해상도 모델 상에서 두 개의 구조를 나란히 맞추어 비교한 결과 구조가 잘 겹치지 않는 지역이 있음을 확인하였다 <그림 11>.

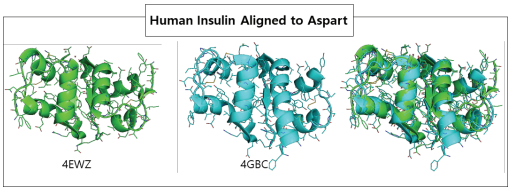

그림 11. 휴먼 인슐린과 Aspart 인슐린의 분자모델 정렬 결과
Fig 11. Alignment of 3D structures between Human Insulin and Aspart insulin

휴먼 인슐린 Betaline 제품과 인슐린 아날로그 Aspart 인슐린 단백질 고해상도 구조 모델에서 (PDB ID: 4EWZ, 4GBC) 계산된 형태 모델이 <그림 12>에 나타나 있으며 형태 측면에서도 두 개의 구

조가 잘 겹치지 않는 부분이 있음을 확인하였다.

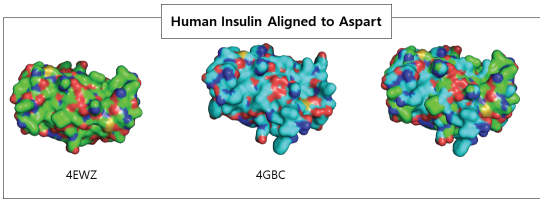

그림 12. 휴먼 인슐린과 Aspart 인슐린의 표면모델 정렬 결과
Fig 12. Alignment of surface structures between Human Insulin and Aspart insulin

휴먼 인슐린과 Aspart 인슐린 사이에서의 구조적 유사성을 분석하기 위해 휴먼 인슐린과 Asprt 인슐린의 고해상도 구조 모델 사이에서 RMSD 값을 확인하였으며 계산된 RMSD 값의 평균과 표준편차가 <표 7>에 나타나 있다. RMSD 값은 일반적으로 0.5 이하로 계산되면 구조적 유사성이 높은 것으로 판단하는데[19], 휴먼 인슐린과 Aspart 인슐린 사이에서의 RMSD 분석 결과 모두 0.5 보다 큰 것으로 확인되어 휴먼 인슐린과 Asparat 인슐린 사이에서 3차 구조가 차이가 있음을 확인하였다.

표 7. 휴먼 인슐린과 Aspart 인슐린 사이의 RMSD 평균과 표준편차 값
Table 7. Avg. and std. deviation of RMSD between Human Insulin and Aspart insulin

		Betalin	Humulin	Insunorm	Novolin
Aspart	n	25	40	30	45
	평균	1.3264	1.2970	1.3397	1.2720
	표준편차	0.0259	0.0438	0.0282	0.0458

본 연구에서 제안한 형태 기술자를 사용하여 휴먼 인슐린과 Aspart 인슐린에 대한 표면 구조 사이에서의 형태 기술자 값을 확인하였으며 계산된 형태 기술자 값의 평균과 표준편차가 <표 8>에 나타나 있다.

표 8. 휴먼 인슐린과 Aspart 인슐린 사이의 형태 기술자 평균과 표준편차 값
Table 8. Avg. and std. deviation of Shape descriptor between Human Insulin and Aspart insulin

		Betalin	Humulin	Insunorm	Novolin
Aspart	n	25	40	30	45
	평균	0.010825	0.011406	0.010103	0.011061
	표준편차	0.002546	0.003992	0.001919	0.002457

표 9. 휴먼 인슐린 및 Aspart 인슐린 사이의 RMSD 평균과 표준편차 값
Table 9. Avg. and st. deviation of RMSD between human insulin and Aspart insulin

		Betalin	Humulin	Insunorm	Novolin
Betalin	n	10			
	평균	0.2045			
	표준편차	0.0698			
Humulin	n	40	28		
	평균	0.2399	0.1661		
	표준편차	0.0735	0.0982		
Insunorm	n	30	48	15	
	평균	0.1875	0.2464	0.1277	
	표준편차	0.0690	0.0838	0.0711	
Novolin	n	45	72	54	36
	평균	0.2690	0.2272	0.2770	0.2617
	표준편차	0.0949	0.1331	0.0991	0.1394
Aspart	n	45	72	54	36
	평균	1.3264	1.2970	1.3397	1.2720
	표준편차	0.0259	0.0438	0.0282	0.0458

휴먼 인슐린과 Aspart 인슐린 사이에서 확인된 RMSD 값을 휴먼 인슐린 사이에서 확인된 RMSD 값과 비교하여 구조적 유사성과의 연관성을 확인하였다 <표 9>. 휴먼 인슐린 사이의 RMSD 평균값은 0.1277~0.269로 확인되었으며 휴먼 인슐린과

Aspart 인슐린 사이의 RMSD 평균값은 1.27~1.3397로 확인되어 휴먼 인슐린과 Aspart 인슐린 사이에서 확인되는 RMSD 값이 훨씬 더 크게 나타나 RMSD와 구조적 유사성과의 연관성을 확인할 수 있었다.

표 10. 휴먼 인슐린 및 Aspart 인슐린 사이의 형태 기술자 평균과 분산값

Table 10. Avg. and std. deviation of Shape descriptor uman insulin and Aspart insulin

		Betalin	Humulin	Insunorm	Novolin
Betalin	n	10			
	평균	0.001106			
	표준편차	0.000346			
Humulin	n	40	28		
	평균	0.000959	0.000997		
	표준편차	0.000707	0.000914		
Insunorm	n	30	48	15	
	평균	0.000748	0.000765	0.000435	
	표준편차	0.000359	0.000745	0.000142	
Novolin	n	45	72	54	36
	평균	0.000867	0.000848	0.000673	0.000825
	표준편차	0.000387	0.000629	0.000220	0.000352
Aspart	n	25	40	30	45
	평균	0.010825	0.011406	0.010103	0.011061
	표준편차	0.002546	0.003992	0.001919	0.002457

휴먼 인슐린 과 Aspart 인슐린 사이에서 확인된 형태 기술자 값을 휴먼 인슐린 사이에서 확인된 형태 기술자 값과 비교하여 형태 기술자와 구조적 유사성과의 연관성을 확인하였다 <표 10>. 휴먼 인슐린의 형태 기술자 평균값은 0.000435~0.001106로 확인되었으며 휴먼 인슐린과 Aspart 인슐린 사이의 형태 기술자 평균값은 0.010103~0.011406으로 확인되어 휴먼 인슐린과 Aspart 인슐린 사이에서 확인되는 형태 기술자 값이 훨씬 더 크게 나타나 형태 기술자와 구조적 유사성과의 연관성을 확인할 수 있었다.

휴먼 인슐린 사이 및 휴먼 인슐린과 Aspart 인슐린 사이에서 계산된 RMSD 및 형태 기술자 값의 분포를 박스 플롯 그래프로 시각화하였다 <그림 13>. RMSD 및 형태 기술자 값의 분포의 경우 모두 휴먼 인슐린과 Aspart 인슐린 사이에서의 값이 휴먼 인슐린 사이에서의 값보다 크게 나타남을 확인하였다 <그림 13>.

밧스 플롯 그래프에서 확인된 휴먼 인슐린과 Aspart 인슐린 사이에서의 분포 차이를 통계적으로 분석하였다. 휴먼 인슐린 사이 및 휴먼 인슐린과 Aspart 단백질 사이에서 계산된 RMSD 및 형태 기술자 값의 분포에 대한 통계적 유의성을 분석하기 위해 t-검정을 수행하였다. 휴먼 인슐린 사이 및 휴먼 인슐린과 Aspart 인슐린 사이에서 계산된 RMSD 분포와 형태 기술자 분포에 대한 t-검증 결과가 <표 11> 및 <표 12>에 나타나 있다.

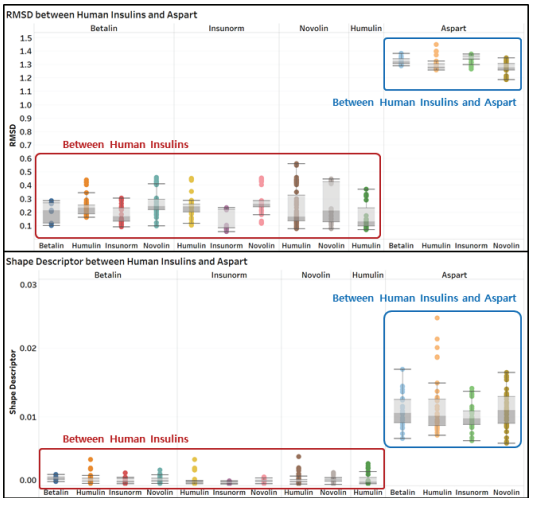

그림 13. 형태 기술자 특성값과 RMSD의 박스플롯 그래프

Fig 13. Box-Plot of RMSD and Data from Shape descriptor

표 11. RMSD 분포에 대한 t-검증 결과
Table 11. t-test result about RMSD distribution

	Human Insulin to Human Insulin	Human Insulin to Aspart
평균	0.234121693	1.303357143
분산	0.011869078	0.002196735
관측수	378	140
가설 평균차	0	
자유도	506	
t 통계량	-155.8138472	
P(T<=t) 단측 검정	0.E+00	
t 기각치 단측 검정	1.647870583	
P(T<=t) 양측 검정	0.E+00	
t 기각치 양측 검정	1.964663311	

RSMD 분포 및 형태 기술자 분포의 P-값이 모두 0.05 보다 작아 휴먼 인슐린 제품 사이에서 확인되는 RMSD 및 형태 기술자 분포와 휴먼 인슐린과 Aspart 인슐린 제품 사이에서 확인되는 RMSD 및 형태 기술자 분포는 확실히 다르다는 것을 통계적으로 확인하였다.

표 12 형태 기술자 특성값에 대한 t-검증 결과
Table 12. t-test result about shape descriptor's data

	Human Insulin to Human Insulin	Human Insulin to Aspart
평균	0.000817868	0.01091215
분산	3.2124E-07	8.48908E-06
관측수	378	140
가설 평균차	0	
자유도	143	
t 통계량	-40.70866841	
P(T<=t) 단측 검정	7.7869E-81	
t 기각치 단측 검정	1.655579143	
P(T<=t) 양측 검정	1.55738E-80	
t 기각치 양측 검정	1.976692198	

통계적 분석을 통해 휴먼 인슐린 사이에서 계산된 RMSD 및 형태 기술자 값의 분포와 휴먼 인슐린과 Aspart 인슐린 사이에서 계산된 RMSD 및 형태 기술자 값의 분포가 차이가 남을 확인하였기에 형태적 기술자를 활용하여 단백질 표면 구조의 유사성을 판별할 수 있는 가능성을 확보하였다.

휴먼 인슐린 사이에서 계산된 RMSD 및 형태 기술자 값의 분포와 휴먼 인슐린과 Aspart 인슐린 사이에서 계산된 RMSD 및 형태 기술자 값의 분포를 정규 분포화 하여 확률밀도함수를 확보하고 이로부터 유사도 판별을 위한 결정 경계를 정하였다.

휴먼 인슐린 사이에서 계산된 RMSD 값의 분포와 휴먼 인슐린과 Aspart 인슐린 사이에서 계산된 RMSD 값의 분포로부터 추정한 정규 분포 곡선이 <그림 14>에 나타나 있다. 휴먼 인슐린 사이에서 계산된 RMSD 값의 분포와 휴먼 인슐린과 Aspart 인슐린 사이에서 계산된 RMSD 값의 분포가 확연하게 구분됨을 확인하였다.

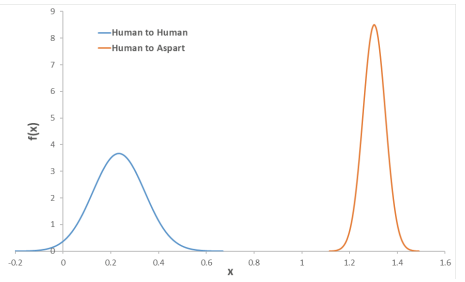

그림 14. RMSD의 분포도
Fig 14. Normal distribution of RMSD

휴먼 인슐린 사이에서 계산된 형태 기술자 값의 분포와 휴먼 인슐린과 Aspart 인슐린 사이에서 계산된 형태 기술자 값의 분포로부터 추정한 정규 분포 곡선이 <그림 15>에 나타나 있다. 휴먼 인슐린 사이에서 계산된 형태 기술자 값의 분포와 휴먼 인슐린과 Aspart 인슐린 사이에서 계산된 형태 기술자 값의 분포 역시 확연하게 구분됨을 확인하

였다. 휴먼 인슐린 사이에서 계산된 형태 기술자 값의 분포와 휴먼 인슐린과 Aspart 인슐린 사이에서 계산된 형태 기술자 값의 분포가 조금 겹치는 부분이 있으며 이에 대한 결정 경계를 LRT (likelihood ratio test) 분석을 통해 결정하였다< 그림 15>. x=0.002714 형태 기술자 값을 기준으로 단백질 표면 구조의 유사성을 판별할 수 있는 결정 규칙을 확보하였다.

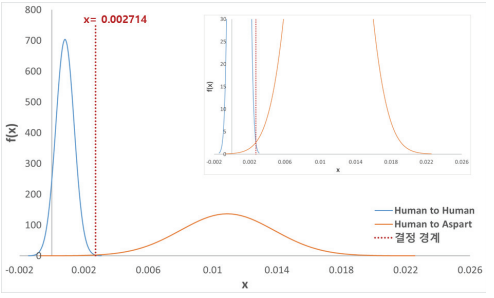

그림 15. 형태 기술자 특성값의 분포도
Fig 15. Normal distribution of Shape descriptor

5. 결 론

본 연구에서는 저해상도 구조 모델에 적용하여 단백질의 구조적 유사성을 평가할 수 있는 형태기술자를 설계하는 방법론을 제시하였다. 본 연구에서 제시된 형태기술자는 고해상도 구조 모델에 일반적으로 적용하는 RMSD 방법과 동일하게 서로 다른 표면 구조를 가지는 단백질을 구별할 수 있음을 확인하였다.

특히 본 연구에서는 아미노산 서열에서 1개만의 차이를 보이는 구조적으로 매우 유사한 휴먼 인슐린과 Aspart 인슐린을 비교하였다. 그 결과 본 연구에서 제시하는 형태기술자는 특정 부위에서 지역적 차이를 보이는 아주 유사한 구조를 가지는 단백질에 대해서도 RMSD 방법과 동일하게 그 유사성을 구별할 수 있음을 확인하였다.

이러한 결과로부터 본 연구에서 제시하는 형태기술자는 SAXS 분석과 같이 저해상도 단백질 구조 모델을 확보할 수 있는 분석에 적용하여 단백질의 구조적 유사성을 판별할 수 있는 기반을 제공할 수 있으며 이를 통해 단백질의약품에 대한 구조적 유사성을 판별할 수 있는 방법으로 활용될 수 있을 것으로 판단된다. 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.

References

[1] A. Kryshtafovych, J. Moult, A. Basle, A. Burgin, T. K. Craig, R. A. Edwards, R, D. Fass, M. D. Hartmann, M. Korycinski, R. J. Lewis, D. Lorimer, A. N. Lupas, J. Newman, T. S. Peat, T. K. H. Piepenbrink, J. Prahlad, M. J. van Raaij, F. Rohwer, A. M. Segall, V. Seguritan, E. J. Sundberg, A. K. Singh, M. A. Wilson, and T. Schwede, *Some of the most interesting CASP11 targets through the eyes of their authors.* Proteins, Vol. 84, pp. 34-50, 2016.

[2] M. F. Lensink, S. Velankar, and S. J. Wodak, *Modeling protein‐protein and protein‐peptide complexe*, CAPRI 6th edition. Proteins, Vol. 85, pp. 359-377, 2017.

[3] A. G. Kikhney, and D. I. Svergun, *A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins*, FEBS Letters, Vol. 589, Issue 19, Part A, pp. 2570-2577, 2015.

[4] S. Mahmoudi, and M. Daoudi, *3D models retrieval by using characteristic views*, CPR'02, pp. 11-15. Quebec, Canada, Aug. 2002.

[5] J-L. Shih, C-H. Lee, and J. T. A. Wang, *A new*

3D model retrieval approach based on the elevation descriptor, Pattern Recognition, Vol. 40, No. 1, pp. 283-295, 2007.

[6] D-Y. Chen, M. Ouhyoung, X-P. Tian, Y-T. Shen, and M. Ouhyoung, *On visual similarity based 3d model retrieval*, Proceedings of Eurographics, pp. 223-232. Granada, Spain, 2003.

[7] R. Ohbuchi, K. Osada, T. Furuya, and T. Banno, *Salient local visual features for shape-based 3D model retrieval*, Proceedings of the IEEE international conference on shape modeling and applications (SMI 2008), pp. 93-102, 2008.

[8] D. Vranic, D. Saupe, and J. Richter, *Tools for 3d-object retrieval: Karhunen-loeve transform and spherical harmonics*, Proceedings of the IEEE fourth workshop on multimedia signal processing, pp. 293-29, 2001.

[9] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, *Matching 3D models with shape distributions*, Proceedings of the IEEE international conference on shape modeling and applications (SMI2001), pp. 154-166, 2001.

[10] R. Ohbuchi, T. Minamitani, and T. Takei, *Shape similarity search of 3D models by using enhanced shape functions*, International Journal of Computer Applications in Technology, Vol. 23, No. 2, pp. 97-104, 2005.

[11] Y. Liu, H. Zha, and H. Qin, *The generalized shape distributions for shape matching and analysis*, Proceedings of the IEEE international conference on shape modeling and applications (SMI2006). Matsushima, Japan, 2006.

[12] M. Ankerst, G. Kastenmuller, H. P. Kriegel, and T. Seidl, *3D shape histograms for similarity search and classification in spatial databases*, Proceedings of the 6th international symposium of spatial databases (SSD1999). Hong Kong, 1999.

[13] M. Hilaga, Y. Shinagawa, T. Kohmura, and T. L. Kunii, *Topology matching for fully automatic similarity estimation of 3D shapes*, Proceedings of ACM SIGGRAPH 2001, pp. 203-212, 2001.

[14] D. Duhovny, R. Nussinov, and H. J. Wolfson, *Efficient unbound docking of rigid molecules*, Proceedings of the Second International Workshop on Algorithms in Bioinformatics, SpringerVerlag: London, UK. pp. 185-200, 2002.

[15] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H. J. Wolfson, *Geometry-based flexible and symmetric protein docking*, Proteins, Vol. 60, No. 2, pp. 224-231, 2005.

[16] V. Venkatraman, Y. Yang, L. Sael, and D. Kihara, *Protein-protein docking using region-based 3D Zernike descriptors*, BMC Bioinformatics, Vol. 10:407, pp. 1-21, 2009.

[17] D. Kihara, L. Seal, and J. Esquivel-Rodriguez, *Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking*, Current Protein & Peptide Science, Vol. 12, No. 6, pp. 520-530, 2011.

[18] Z. Shentu, M. Hasan, C. Bystroff, and M. Zaki, *Context shapes: efficient complementary shape matching for protein–protein docking*, Proteins Vol. 70, pp. 1056-1073, 2008.

[19] M. P. Favero-Retto, L. C. Palmieri, T. A. Souza, F. C. Almeida, L. M. Lima, *Structural meta-analysis of regular human insulin in pharmaceutical formulations*, European Journal of Pharmaceutics and Biopharmaceutics, Vol. 85, pp. 1112-1121, 2013.

3차원 형상 분석 기반 단백질 유사성 분석

전성환[1], 최유주[1,2], 서정근[1,2]

[1]*서울미디어대학원대학교(SMIT) 뉴미디어학부*

[2]*서울미디어대학원대학교(SMIT) 실감미디어연구소*

요 약

단백질 구조 분석 및 비교는 X-ray 결정 방식이나 NMR 분광기 분석과 같은 실험적인 방법으로 수세기 동안 진행되어 십삼만개 이상의 원자 수준에서의 고해상도 단백질 3차 구조를 파악하고 있다. X-ray 결정 방식은 결정(Crystal structure)을 만들 수 없는 단백질은 구조 분석이 안 되는 단점이 있고, NMR 분광기 분석은 단백질 결합 등의 이유로 실험대상의 크기가 커지면 분석을 못하는 한계가 있다. 하지만 단백질의약품의 구조적 유사성을 평가하는 것은 단백질의 약품 개발과 생산과정에서 아주 중요한 부분이다. 현재 이러한 유사성 평가는 다양한 분광학적 방법을 활용하지만 기술적 제한으로 인해 구조적인 정보를 확인하기에는 한계가 있다. 더구나 최근 Single Angle X-ray Scattering(SAXS) 분석법을 통해 해상도가 낮은 표면 구조 정보를 보다 쉽게 확보할 수 있게 되었다. 이에 본 연구에서는 저비용/단시간에 확보할 수 있는 저해상도 3차 구조 (표면 구조) 정보를 활용하여 동등성을 평가할 수 있는 기술을 개발하고자 하였다. 본 연구에서는 단백질 저해상도 모델로 표면구조 데이터의 특성을 이용하여 3D 모델 형태 특성 데이터를 기반으로 한 새로운 형태 기술자(Shape Descriptor)를 정의하고 이를 활용하여 단백질의 구조적 유사성을 평가할 수 있는 방법을 제안한다. 제안한 방법은 기존 방법인 고 해상도 단백질 3차 구조 정보를 활용하여 유사성을 판별한 RMSD 값과 비교하여 제안된 형태 기술자를 적용한 방법의 실효성을 검증하였다.

Sung Hwan Chun received the bachelor's degree in the Department of Mass-communication from the HanDong University in 2002. He is in undergraduate course in the Department of New Media Engineering from *Seoul Media Institute of Technology from 2015*. From 2015 to 2016, he was a researcher at *Immersive Media Lab*. His current research interests include computer graphics, 3D model retrieval, 3D model recognition, pattern recognition and artificial intelligence.

E-mail address: alexchun78@gmail.com

Yoo-Joo Choi is a associate professor in Department of Newmedia at Seoul Media Institute of Technology(SMIT), Korea. She received her M.S. and Ph.D. degrees in Computer Science from Ewha Womans University in 1991 and 2005, respectively. She was a researcher at R&D Department of KCI Co. and POSDATA Co. in Korea between 1991 and 1999. Her research interests include image processing, computer vision, computer graphics, augmented reality and human-computer interaction.

E-mail address: yjchoi@smit.ac.kr

Jung-Keun Suh received the bachelor's degree in the Department of Botany from Seoul National University in 1987. He received the M.S. degree in the Department of Botany from Seoul National University in 1989 and the Ph.D. degree in the Department of Chemistry & Biochemistry from the University of Texas at Austin in 1996. From 1996 to 2000, he was a post-doctoral fellow in the Department of Chemistry & Biochemistry from the University of Texas at Austin. He was a professor in Department of Newmedia at Seoul Media Institute of Technology(SMIT), from 2009 to present.

E-mail address: jksuh@smit.ac.kr