韓國 컴퓨터情報學會 論文誌
第14卷 第1號, 2009. 1.

2009-14-1-2-1

단백질 2차 구조를 이용한 유사 GPCR 검출에 관한연구

구 자 효*, 한 찬 명*, 윤 영 우*

A Study on the Detection of Similarity GPCRs by using protein Secondary structure

Ja-Hyo Ku*, Chan-Myung Han*, Young-Woo Yoon*

요 약

GPCR(G protein-coupled receptors) 패밀리(family)는 세포막 단백질로서, 외부 신호를 세포막을 경유하여 세포 내로 전달하는 신호전달 기전에서 중요한 역할을 담당한다. 그러나 GPCR마다 다양하고 복잡한 조절기전을 보이며 매우 특이적인 신호전달 기전을 가지는 것으로 알려져 있다. GPCR의 구조적인 특징과 패밀리 및 서브패밀리 등은 기능별로 잘 알려져 있는데 과거 GPCR을 찾아내는 연구 중에 가장 기본이 되는 일이 주어진 단백질 서열로부터 GPCR을 분류하는 일이다. 이미 발견된 GPCR들을 가지고 수학적인 모델을 이용하여 보다 정확하게 분류하는 연구가 주로 진행되어 왔다. 본 논문에서는 단백질의 기능이 입체적 구조에 의해 결정되는 점에 착안하여 두 GPCR의 아미노산 서열의 유사도가 낮은 경우에 그 2차 구조의 서열을 비교함으로써 기존의 발견된 GPCR의 데이터베이스에서 동일한 기능을 가졌을 것으로 추정되는 미지의 GPCR을 검출하는 방법을 제안한다.

Abstract

G protein-coupled receptors(GPCRs) family is a cell membrane protein, and plays an important role in a signaling mechanism which transmits external signals through cell membranes into cells. But, GPCRs each are known to have various complex control mechanisms and very unique signaling mechanisms. Structural features, and family and subfamily of GPCRs are well known by function, and accordingly, the most fundamental work in studies identifying the previous GPCRs is to classify the GPCRs with given protein sequences. Studies for classifying previously identified GPCRs more easily with mathematical models have been mainly going on. In this paper Considering that functions of proteins are determined by their stereoscopic structures, the present paper proposes a method to compare secondary structures of two GPCRs having different amino acid sequences, and then detect an unknown GPCRs assumed to have a same function in databases of previously identified GPCRs.

▸ Keyword : 서열 정렬(sequence alignment), GPCR(G protein-coupled receptor), 단백질 2차 구조(Protein secondary structure prediction)

• 제1저자 : 구자효 교신저자 : 윤영우
• 투고일 : 2008. 11. 12, 심사일 : 2008. 11. 18, 게재확정일 : 2008. 12. 30.
* 영남대학교 컴퓨터공학과
※ 본 논문은 2006학년도 영남대학교 연구년제 수행에 의한 것임

Ⅰ. 서 론

21세기에 들면서 Human Genome Project 등 대형 유전자 관련 프로젝트가 확대되는 추세에 있고 고 처리율 서열 기술 발전에 의해 대량의 서열 정보가 급격히 생산되고 있다. 1차적으로 유전자의 서열검출은 완성되었다고 할 수 있으나 이는 DNA 지도를 그렸을 뿐 어느 부분이 유전자인지, 또한 유전자라면 유전자의 기능 및 작용 기작이 무엇인지는 아직도 해결해야할 숙제이다. 이에 신약 개발 등 여러 바이오산업의 연구기간 단축을 위한 대량 서열 정보의 효율적인 검출의 필요성이 대두되고 서열정보의 데이터베이스화가 진행되어 기존에 알려져 있는 서열정보를 더욱 빠르고 정확하게 탐색하는 방법들이 개선, 발전되어 왔다[1].

GPCR은 세포막 단백질로서, 외부 신호를 세포막을 경유하여 세포내로 전달하는 신호전달 기전에서 중요한 역할을 담당한다[2]. 현재 약 800~1000개 정도의 많은 GPCR이 사람의 유전체에 존재할 것으로 예측하고 있다[3]. 그러나 GPCR 마다 다양하고 복잡한 조절 기전을 보이며 매우 특이적인 신호전달 기전을 가지는 것으로 알려져 있다. GPCR은 내분비계나 신경계를 비롯한 다양한 인체 조직의 항상성 유지에 관여하며, GPCR의 활성 조절에 이상이 생기면 심혈관계 질환, 대사성 질환, 퇴행성 질환, 발암 등이 중요한 원인이 될 수 있다. 실제로 제약회사들은 60%이상 신약개발을 위해 GPCR 연구에 집중하고 있다[4].

여러 가지 GPCR들을 기능별로 서브패밀리 단위로 분류한 데이터베이스가 공개되어 있는데, 서브패밀리 내의 GPCR들의 기능은 서로 유사하다. GPCR 연구 중에 가장 기본이 되는 일은 주어진 아미노산 서열로써 그 데이터베이스를 검색하는 것이다. 검출 목적은 주어진 아미노산 서열이 'GPCR인가? GPCR이라면 이미 알려진 GPCR과 같은 것인가? 아니면 아직까지 알려지지 않은 새로운 종류의 GPCR인가?' 등이다. 두 아미노산 서열의 유사도가 높으면 GPCR이라고 추정하고 해당 서브패밀리에 포함시킬 수 있다.

현재까지 개발된 서열정보 탐색도구로는 BLAST가 가장 대중적으로 사용되고 있지만 기존의 방법에는 유사한 기능의 GPCR이라도 서열의 유사도가 낮은 경우에는 정밀검출 후보군 대상에서 배제되어 찾지 못하는 오류가 보고되고 있다. 특히 GPCR 패밀리는 아미노산의 서열이 유사성이 낮음에도 불구하고 비슷한 기능을 하는 경우가 있는가 하면, 유사성이 높음에도 전혀 다른 역할을 수행하는 등 기존의 정보가 데이터베이스로 저장되어 있어도 활용하기가 까다로운 편이다. 단백질의 기능이 입체적 구조에 의해 결정되는 점을 감안하면, 동일한 서브패밀리에 속해 있으면서도 아미노산 서열이 서로 다를 수 있는 까닭은 서로 다른 종의 GPCR 단백질이 진화과정에서 입체적 구조가 유사하게 변화된 것으로 추정해 볼 수 있다.

단백질은 단계적으로 구성된다. 단백질의 아미노산 서열을 1차 구조라 하고, 1차 구조의 아미노산들이 서로 결합하여 α-helix와 β-sheet라는 두 가지의 형태를 구성한다. α-helix와 β-sheet와 그들을 연결하는 아미노산 서열들을 단백질의 2차 구조라 한다. α-helix와 β-sheet는 단백질을 구성하는 기본 요소가 되는데, 이들 사슬이 만들어 내는 입체적 형태를 단백질의 3차 구조라 한다. 한 개 이상의 단백질 3차 구조가 복합체를 이루는 형태를 단백질의 4차 구조라 한다. 현재 아미노산 서열로부터 단백질 2차 구조를 예측하는 알고리즘이 개발되어 있으며, 그 정확도는 약 70% 정도인 것으로 알려져 있다. 몇 개의 단백질의 3차 구조 및 4차 구조가 정밀검출에 의해 밝혀져 있으며, 많은 연구가 3차 구조의 예측방법에 집중되고 있으나 아직까지 목적을 달성하지 못하고 있다.

본 논문에서는 단백질의 기능이 입체적 구조에 의해 결정되는 점에 착안하여 두 GPCR의 아미노산 서열이 서로 다른 경우에 그 2차 구조를 비교함으로써 기존의 발견된 GPCR의 데이터베이스에서 동일한 기능을 가진 것으로 추정되는 미지의 GPCR을 검출하는 방법을 제안한다. 따라서 1차 구조를 비교하는 기존의 방법에서는 검출하지 못한 GPCR의 발견 가능성을 높인다.

Ⅱ. 관련 연구

2.1 GPCR

대부분의 GPCR은 400-500 잔기의 단일 폴리펩타이드로 구성되어 있으며 칠중나선(heprahelical)구조를 가지고 있다[5]. 인간 유전체 중 1%이상이 이러한 칠중나선 구조로 발현하고 있으므로 모두 GPCR의 후보물질이 되고 있다[6]. 약 800~1000개의 유전체중 650여개의 GPCR 유전자가 밝혀져 있는 상태이다. 이렇게 많은 GPCR은 비록 그 구조적인 면에서 매우 유사할지라도 내인 리간드의 종류와 특이성은 매우 다양하고 결합방식 또한 복잡하다. 펩티드 혹은 비펩티드성 신경전달물질, 이온, 아미노산, 호르몬, 빛, 그리고 방향물질 등이 GPCR을 활성화시키며, 조직 특이적으로 분포된 GPCR은 수많은 리간드들과 결합을 함으로써 특이적 생리현상을 일으키며, 이미 전 세계적으로 제약회사에서 개발되고 있는 약물의 약 60%정도가 GPCR이 대상이 된다.

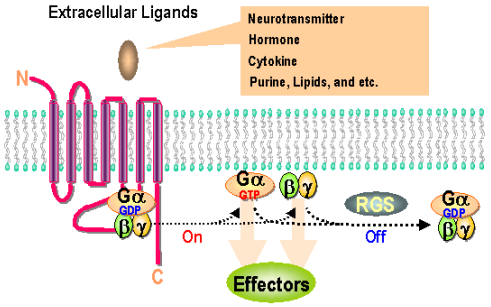

그림 1. GPCR의 신호전달
Fig 1. Signal transduction of GPCR

GPCR은 7개의 막관통 도메인(transmembrane domains)을 가진 막단백질이다. 이 중 3번째 세포질 내 고리는 다른 고리보다 크고 G-단백질과 상호작용한다. G-단백질은 G_{α}와 $G_{\beta\gamma}$이합체로 된 이종삼합체로서 GDP가 G_{α}에서 유리되고 작용제가 GPCR에 결합하게 되고 곧 GTP가 결합하게 된다. 이로써 G_{α}와 $G_{\beta\gamma}$는 서로분리가 되고 각각은 효과기에 영향을 주게 된다. 이때 하나의 인산을 잃게 되어 GTP가 GDP로 변환되고 이것이 다시 G_{α}와 $G_{\beta\gamma}$의 결합을 유발하여 계속적인 순환 고리를 돌게 된다[4].

2.2 단백질 2차 구조 예측

단백질의 1차 구조(DNA 정보로부터 얻어지는 특정 단백질의 아미노산 서열)만으로부터 단백질의 3 차 구조(특정 단백질이 고유의 생물학적 기능을 할 수 있기 위해서 가져야 되는 삼차원적 고유 구조) 및 단백질의 다른 여러 가지 성질들 2차 구조, 도메인 경계(domain boundary), 용매 접근성(solvent accessibility) 등을 예측하는 것을 단백질 구조예측(protein structure prediction)이라고 한다[6].

그림 2. 단백질 2차 구조예측
Fig 2. Protein secondary structure prediction

단백질 2차 구조 예측은 서열의 각 잔기 α-helix, β-sheet, turn 구조 중 어떤 구조를 좋아하는지에 대한 선호도를 표시한 것이라 말할 수 있는데, 이런 선호도를 결정짓기 위해 영향을 미치는 몇 가지 요소들이 있다.

첫째 특정 아미노산들은 특정구조를 선호한다. 예를 들어 알라닌(Ala), 글루탐산(Glu), 루신(Leu), 메티오닌(Met) 등의 아미노산은 α-helix를 선호하는 아미노산으로 아미노산 서열 중 앞서 말한 아미노산들이 있으면 α-helix를 형성할 가능성이 많게 된다.

둘째, 인접 아미노산 잔기의 영향을 들 수 있다. 예를 들어 α-helix에 인접한 사이에 수소결합이 일어나면 구조가 불안정해지는 경우가 있는데, 이런 경우에는 2차 구조에 포함되지 않는 특이한 구조가 생성된다.

셋째, 서열상으로 멀리 위치하지만 공간적으로 인접한 원자들의 영향이 있다. 이것은 3차 구조에 대한 정보가 없는 상황에서는 반영이 불가능한 요인이다.

현재 서열 정보를 기반으로 단백질 2차 구조를 예측하는 방법 중 가장 널리 알려진 두 가지 방법으로 단일 서열을 기반으로 하는 예측 방법과 신경망을 이용한 예측 방법이 있다. 2차 구조를 예측하는 소프트웨어는 PSIPRED, GOR, NNPREDICT등이 있다.

2.3 서열 정렬

두 유전체를 비교하기 위해서는 서열 정렬(sequence alignment) 알고리즘을 사용해서 두 서열간의 유사성을 측정해야 하는데 서열 정렬은 서열 간의 상관관계를 보여주기 위해, 특히 유사성을 나타내기 위해 핵산이나 단백질의 서열을 정렬하는 것을 말한다. 일반적으로 비교할 서열의 길이 범위에 따라서 전역 정렬(global alignment)과 지역 정렬(local alignment)로 분류하는데 유전체 정렬에서는 다양한 재조합과 변이에 의한 유전체 서열의 재정렬(rearrangement)을 전역 정렬로 표현하기 어렵기 때문에, 지역적으로 유사성을 가지는 지역을 검출하고 이들의 연력관계를 검출하여 전체적인 유전체 상에서의 정렬을 재구성하는 방법을 사용해야 한다[7][8]. 지역 서열 정렬 알고리즘으로 가장 광범위하게 사용되는 알고리즘은 Smith-Waterman 알고리즘, BLAST및 FASTA등이 있다. Smith-Waterman 알고리즘은 다이나믹 프로그램을 이용하여 전체 서열에서 유사성 검색을 수행하므로 BLAST와 FASTA에 비해 계산상 좀 더 정확한 검색결과를 얻을 수 있으나, 생물학적인 패턴의 지역 정렬에서는 오히려 BLAST나 FASTA보다 비효율적이고, 검색시간이 엄청나게 오래 걸린다는 단점이 있다. BLAST와 FASTA는 임의의 서열과 유사성을 가진 지역 서열을 서열 데이터베이스로부터 찾는 프로그램으로 일반적으로 유사성 검색을 위해 사용하는 프로그램이다. BLAST는 FASTA와 달리 미리 전 처리된 검색 데이터

베이스 파일을 필요로 한다. 두 가지 프로그램에서 사용하는 알고리즘은 각각 옵션으로 입력한 통계치나 유사성 값을 필터로 하여 지역 유사성을 보이는 서열 부분을 계산 및 검색해준다. 한편, 두 프로그램은 모두 갭(gap)을 포함한 정렬을 검색해 줌으로 유전체 정렬 계산을 위해 사용하는 부분 서열 정렬의 모듈 프로그램으로 적합하게 활용할 수 있다[8].

III. 유사 GPCR 검출

현재 GPCR 검출은 1차 구조 서열비교를 통하여 기능이 알려진 GPCR로부터 기능을 결정하려는 GPCR사이의 유사 아미노산 비교를 통해 이루어진다. 이는 두 가지 조건이 필요하다. 첫 번째는 "기능이 알려진 GPCR과 기능을 결정하려고 하는 GPCR 사이에는 반드시 아미노산의 유사도가 존재하여야한다" 이며 두 번째는 기능이 알려진 GPCR과 기능을 결정하려는 GPCR의 아미노산 서열정보는 정확이 계산되어야 한다."이다. 단백질 구조를 결정하는 방법은 여러 가지가 있지만 현재까지는 아미노산 서열비교 방법이 가장 정확하게 검출하는 것으로 알려져 있다. 현재 까지 개발된 유사성 비교 검색도구로 BLAST가 가장 많이 사용되고 있다. 이는 단백질의 1차 구조 유사성 비교를 통해 1차 구조의 유사성이 높은 집단을 정밀분석을 위한 후보군으로 선정을 하기 위함이다. 하지만 정밀분석 후보군에서 배제된 집단, 즉 1차 구조의 유사성이 낮은 집단에도 정밀분석의 후보 대상이 존재할 가능성 있다. 따라서 본 장에서는 단백질의 2차 구조를 이용하여 정밀분석 후보대상에서 배제된 집단에서 정밀분석을 위한 후보 대상에 추가로 포함하여 기존의 방법에서는 배제된 미지의 GPCR을 검출하는 방법을 제안한다.

3.1 제안 방법

본 논문에서 제안하는 방법은 단백질 1차 구조 서열의 비교에서 유사도가 낮은 것으로 판정되는 경우 2차 구조 서열도 비교하여 유사도가 높으면 동종의 GPCR 후보로 분류하고자 한다. 조건은 i차 구조 서열의 유사도가 높으면, i+1차 구조 서열의 유사성도 높을 가능성이 커질 것이다. 그림 3에 본 논문에서 제안하는 동종의GPCR 후보로 분류하는 방법을 표현하였다.

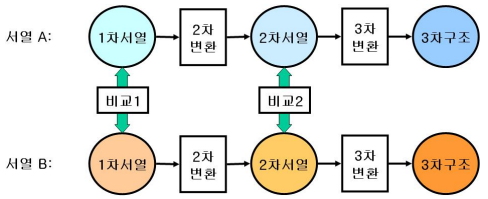

그림 3. GPCR 후보 분류 방법
Fig 3. GPCR candidate classification methods

3.2 제안 알고리즘

네덜란드의 CMBI 연구소를 중심으로 운영되는 GPCRDB에는 현재까지 밝혀진 많은 다양한 생물개체들의 GPCR 데이터베이스가 등재되어 있어 특정 GPCR의 단백질 서열에 대한 정보들을 구할 수가 있다. 또한 동일한 기능을 하는 유사 GPCR에 대하여 패밀리 형식으로 정보를 제공하고 있다. 동일한 패밀리에 속하는 GPCR의 경우 단백질의 1차 구조가 유사한 경우도 있지만 단백질의 1차 구조가 아주 다른데도 동일한 기능을 하는 것으로 판명된 많은 GPCR들이 있다. 단순한 아미노산의 조합이 다를지라도 이들이 형성하는 2차 공간적 구조가 유사할 때 유사 기능을 가지는 것을 살펴볼 필요성이 있다. 유사 기능을 한다는 사실은 이미 밝혀져 있으므로 이들의 2차 구조를 살펴보면 되는데 2차 구조를 실제로 관찰하는 것은 많은 시간과 높은 비용이 발생한다.

본 논문에서는 단백질 1차 구조 서열을 2차 구조 서열로 예측해 주는 소프트웨어 패키지들을 사용하여 방대한 데이터를 손쉽게 비교하도록 한다. 단백질 1차 구조 서열과 2차 구조 서열의 유사성의 비교는 BLAST를 이용하고 2차 구조 예측은 nnPredict를 이용하여 그림 4와 같은 알고리즘을 제안한다.

```
모든 GPCR(G)에 관해서
 {
step1. 미지의 아미노산 서열(X)를 기능이 밝혀
       진 GPCR서열(G)와 비교
step2. 유사도가 높으면 G와 동일 기능의 GPCR
       후보로 인정하고 종료
step3. 유사도가 낮으면 X의 1차 구조 서열로써
       2차 구조 서열을 예측
step4. 예측된 2차 구조 서열로써 G의 GPCR의
       2차 구조와 비교
step5. 유사도가 높으면 G와 동일 기능의 GPCR
       후보로 인정하고 종료
step6. 유사도가 보다 낮으면 step1 반복
 }
 step7. 유사 GPCR 없음
```

그림 4. 제안 알고리즘
Fig 4. Proposal algorithm

3.3 GPCR DB 추출

GPCRDB는 이미 구조가 밝혀진 모든 GPCR의 모든 정보에 대한 데이터베이스를 구축하고 있어 새로운 GPCR에 대한 구조를 밝히거나 그 기능을 예측하기 위해 활용되고 있다. 네덜란드의 CMBI사, 미국 예일대학교의 ORDB와 Frank Kolackowski의 Swiss-Prot DB를 모두 포함하고 있는 통합된 DB이다. Swiss-Prot DB 중에서 그림 5와 같이 GPCR을 "family"로 정렬하고 특정한 ID를 선택한 후 GPCRDB-Family에서 HSSP format으로 데이터들을 조회하면 동일 패밀리에 속하는 GPCR들을 볼 수 있고 각각의 GPCR들을 클릭하면 아미노산의 서열정보를 얻을 수 있다.

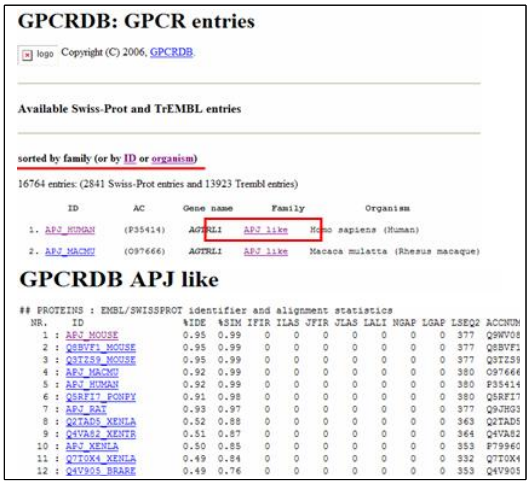

그림 5. GPCRDB의 GPCR entries
Fig 5. GPCR entries of GPCRDB

3.4 GPCR 1차 구조 비교를 통한 유사도 평가

그림 6은 GPCRDB의 APJ_HUMAN의 동일 패밀리에 있는 APJ_MOUSE GPCR 개체의 아미노산 서열을 검색하여 FASTA 포맷의 텍스트 파일에 저장하는 예이다.

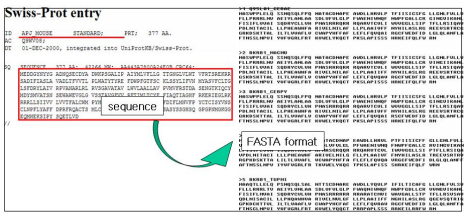

그림 6. FASTA 포맷 저장
Fig 6. FASTA format stores

GPCR 1차 구조를 비교하기 위해서는 BLASTP를 이용한다. 첫 번째로 GPCRDB로부터 생성된 FASTA포맷 데이터를 BLAST의 FormatDB를 이용하여 데이터베이스를 구축하고. 검출을 위한 미지의 GPCR 서열도 FASTA포맷의 데이터로 저장한다. 파일로 저장된 미지의 GPCR 서열을 BLAST의 아미노산 유사성 검색을 위한 알고리즘의 BLSATP를 이용하여 기 구축된 BLAST DB를 대상으로 유사성 비교가 이루어진다. GPCR 1차 구조를 비교하기 위해서는 BLASTP를 이용한다. 그림 7과 같이 1차 구조 비교는 패밀리 내의 GPCR 서열을 선택하여 해당 GPCR이 속한 패밀리의 1차 BLAST 데이터베이스와 비교한다. BLAST 데이터베이스는 FormatDB 프로그램을 실행하여 생성된 FASTA 포맷의 1차 구조 서열 정보인 텍스트 파일을 입력받아 ".phr", ".pin", ".psd", ".psi", ".psq"의 확장자를 가지는 다섯 가지 파일 형태의 1차 구조 데이터베이스를 생성한다. 그 다음, 생성된 1차 구조 데이터베이스를 이용하며 GPCR 서열을 BLASTP 프로그램을 이용하여 서열을 비교하여 유사성 점수를 산출한다.

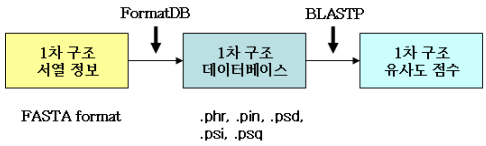

그림 7. 1차 구조 유사도 비교 프로세스
Fig 7. First structure similar compare process

그림 8에서는 BLASTP를 실행 결과를 나타낸 것으로, GPCR의 데이터베이스 개수는 4,289개이며 총 1,358,990개 아미노산 코드로 구성되어 있다. 유사도 비교를 위한 입력 시퀀스는 820개의 아미노산 코드로 구성되며 비교한 결과를 스코어가 높은 순으로 나타낸다. 스코어는 유사도를 점수로 나타낸 것이다.

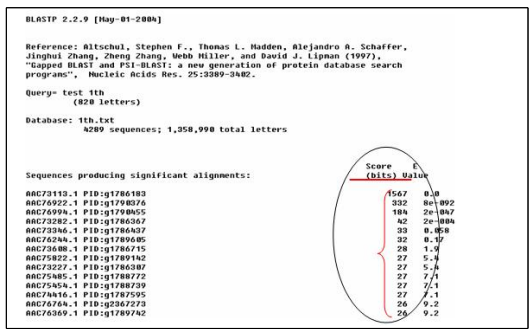

그림 8. BLASTP 실행 후 Score 계산
Fig. 8. Score calculations after BLASTP executing

3.5 GPCR 2차 구조 비교를 통한 유사도 평가

GPCR 2차 구조 비교를 위해 우선적으로 2차 구조의 서열정보를 수집하여야 한다. 본 논문에서는 GPCR 1차 구조 서열정보를 2차 구조로 예측하는 방법으로 nnPredict 서버를 사용하였다. 본 논문에서 사용된 nnPredict 서버는 웹 브라우저를 통해 시퀀스를 입력하여 2차 구조를 예측할 수 있다. GPCRDB는 현재의 웹사이트로부터 추출한 32개의 GPCR 패밀리를 통해 590개의 GPCR개체 정보를 수집하여 2차 구조 데이터베이스를 생성하여 그림 9와 같은 방법으로 분석하였다.

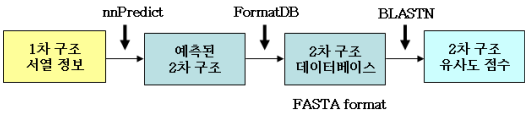

그림 9. 2차 구조 유사도 비교 프로세스
Fig 9. Secondary structure similar compare process

수집된 1차 구조 서열을 2차 구조로 예측하기 위해서는 먼저 nnPredict 서버에 접속한다. 별도의 설정이나 명령을 실행할 필요없이 사용자가 예측하고자 하는 시퀀스를 입력하는 단순한 질의 화면으로 구성되어 있다. 그림 10은 nnPredict 접속 화면에서 APJ_MOUSE의 1차 구조 서열에 대한 2차 구조 예측을 질의하는 것을 나타내는 것으로, 시퀀스 입력 창에 예측하고자 하는 서열을 입력 후 "submit" 버튼을 클릭함으로써 2차 구조로 예측된 결과를 화면에서 즉시 확인할 수 있다.

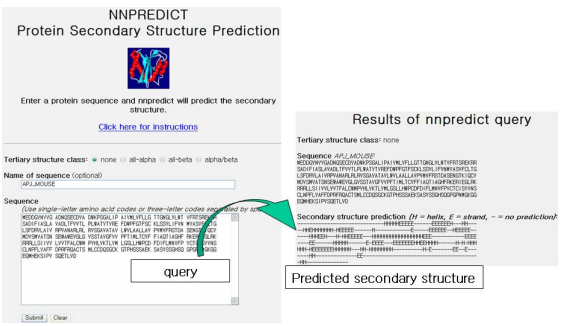

그림 10. nnPredict를 이용한 2차 구조 예측
Fig 10. The secondary structure prediction using nnPredict

IV. 실험 및 결과분석

4.1 알고리즘 검증방법

본 장에서는 본 논문에서 제안한 알고리즘을 실험을 통하여 검증한다. 다음의 2조건의 가정을 두고 검증한다. 첫 번째 "1차 구조 서열은 유사성이 낮지만 2차 구조 서열의 유사성이 높은 GPCR 쌍들이 실제로 존재함을 검증", 두 번째 "1차 구조 서열은 유사도가 낮지만 2차 구조 서열의 유사도가 높은 GPCR 쌍들은 3차 구조도 유사할 가능성이 높음을 검증"한다. 검증방법은 역으로 진행한다. 우선 3차 구조가 유사한 GPCR쌍들 중 1차 구조 서열이 유사하지 않는 것들이 존재하고 그들 중 2차 구조 서열의 유사도가 높은 것들이 존재하며 그들의 방법은 기능이 밝혀져 있는 GPCR 패밀리를 이용하여 패밀리 내의 1차 구조 서열의 유사성이 낮은 GPCR의 2차 구조 서열의 유사성이 높은 GPCR이 존재하는지를 확인한다. 우선 GPCR 패밀리 내에서 1차 구조의 유사성이 낮은 것이 있는지를 확인하고 유사성이 낮은 것 중에서 2차 구조의 유사성이 높은 것이 있는지를 확인한다. 이는 "1차 구조의 유사성이 높은 것"과 1차 구조의 유사성이 낮고 2차 구조의 유사성이 높은 것"이 동일 기능을 발현하는 하나의 패밀리에 존재한다는 것을 의미한다. 따라서, 본 논문에서 제안하는 알고리즘이 타당하다는 것을 검증한다.

4.2 실험 대상

실험에 사용된 GPCR 데이터베이스는 표 1와 같이 32개의 패밀리에 총 590개의 개체로 구성되어 있다. 이러한 GPCR 패밀리 모두 HUMAN이 포함되어 있는 것을 대상으로 무작위로 추출 하였다. 패밀리는 기능이 유사한 GPCR들의 집합이며, 해당 패밀리의 GPCR의 개체 수도 각각 다르다. 이는 미지의 GPCR이 정밀분석을 통해 유사 기능을 하는 GPCR들로 밝혀져서 해당 패밀리에 포함되었기 때문이다. 지속적으로 업데이트 되며 2008년 10월까지 등록된 패밀리를 기준하였으며 향후 개체 수는 변경될 수 있다. 32개의 패밀리 내의 모든 GPCR를 비교하고 1차 구조 서열 유사도는 낮지만 2차 구조 서열의 유사도가 높은 GPCR은 다른 모든 패밀리들의 2차 구조 서열과 비교한다.

표 1. 실험에 사용된 패밀리의 목록
Table 1. List of the family using in experiment

NO.	GPCR 패밀리명	개체 수	NO.	GPCR 패밀리명	개체 수
1	APJ_like	12	17	Somatostatin_and angiogenin_like	15
2	Orexin	14	18	Vasopressin_type_1	15
3	Thrombin	10	19	Somatostatin_type_ 2	12
4	Proteinase_activated	26	20	Neuropeptide_Y other	19
5	Fungal_pheromone_ B_like	40	21	RDC1	9
6	Serotonin_type_2	36	22	Secretin	6
7	Prostacyclin	7	23	Somatostatin_type 5	14
8	Prolactin_releasing peptide	9	24	Neuropeptide_Y type 6/7	5
9	Platelet_activating_ factor	9	25	Orexigenic neuropeptide QRFP	5
10	Thyrotropin_releasi ng_hormone	18	26	Dopamin_Insect	7
11	Rhodopsin Vertebrate_type_	82	27	Smoothened	32
12	Somatostatin_type_1	11	28	Very large G-protein coupled receptor	4
13	Somatostatin_type 1	5	29	Purinoceptor P2RY1-4,6,11	40
14	Thyrotropin	18	30	Serotonin_type_1	41
15	Thyrotropin_1	6	31	Prokineticin receptors	13
16	Purinoceptor P2RY5,8,9,10	37	32	Opioid_type_K	8

4.3 결과 분석 및 고찰

그림 11은 GPCR1을 미지의 GPCR로 선정하여 GPCR1과 GPCR2~GPCR20 간의 1차/2차 구조의 유사도를 구한 후 그래프로 나타낸 것이다. 그림 12에서와 같이 GPCR1과 GPCR7과 GPCR1과 GPCR8간의 데이터에서 1차 구조의 유사도는 낮지만 2차 구조 유사도 비교에서 높은 값을 보임을 확인할 수 있다. 실험결과 32개의 패밀리 중에서 1차 구조의 유사도는 낮지만 2차 구조의 유사도가 높은 개체가 포함된 패밀리는 4개의 패밀리에 161쌍의 GPCR 개체가 존재함을 확인하였다.

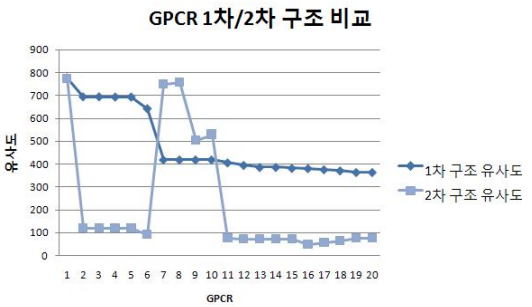

그림 11. GPCR1을 기준으로 한 1차/2차 구조 비교
Fig 11. GPCR1 in standard grudge first/secondary structure comparison

V. 결론

GPCR의 방대한 데이터가 축적된 현시대에서 미지의 GPCR에 대하여 아미노산의 서열정보만을 이용하여 유사한 기능을 가진 GPCR을 분류하는 것은 GPCR 연구에서 아주 중요한 일이 되었다. 현재까지는 GPCR의 1차 구조 서열만을 이용하여 GPCR을 분류하는 연구를 주로 해 왔다. 하지만 1차 구조 서열이 서로 달라도 동일한 기능을 하는 GPCR들이 존재할 가능성이 있다. 이들은 현재 수행되고 있는 1차 구조 서열만을 이용한 방법으로는 분류할 수 없으며 정밀분석 대상에서 사전에 배제될 가능성이 존재한다.

본 논문에서는 단백질의 2차 구조의 유사성이 높으면 동일한 기능을 가진다는 일반적인 사실에 근거하여 동일한 GPCR을 분류해 내는 새로운 방법을 제안하였다. 실험에서는 두 가정을 제시하여 역의 방법으로 검증하였다. HUMAN이 포함되어 있는 GPCR 패밀리 중에서 무작위로 추출한 32개의 패밀리 총 590개의 GPCR 개체에 대하여 실험을 행하였다. 실험 결과 총 590개의 GPCR 개체 상호비교에서 동일 패밀리 내의 1차 구조 서열의 유사성이 낮고 그 2차 구조 서열의 유사성이 높은 GPCR 쌍이 161쌍, 전체적으로 약 0.05%의 비율을 나타내었다. 반면에 다른 패밀리에 속한 GPCR과의 비교에서는 1차 구조 서열의 유사성이 낮고 그 2차 구조서열의 유사성이 높은 GPCR 쌍은 2쌍만 발견되었다. 따라서 GPCR들의 1차 구조 서열의 유사성이 낮은 경우에도 2차 구

조 서열의 유사성이 높은 경우, 그 3차 구조도 유사할 가능성이 있음이 확인되어 제안 알고리즘의 타당성이 검증되었다.

실험 결과가 나타내는 또 다른 특기 사항은 1차 구조 서열의 유사성이 낮고 2차 구조 서열의 유사성도 낮은 경우에도, 그 3차 구조가 유사한 경우도 총 1,570쌍, 전체적으로 약 0.5%나 되었다는 점이다. 그러나 GPCR 2차 구조 서열을 예측하는 소프트웨어의 정확도가 약 70%정도 밖에 되지 않음을 고려할 때 실제 GPCR의 1차 구조 서열은 다르지만 2차 구조가 유사한 더 많은 GPCR이 존재할 것으로 추정되며, 또한 이는 GPCR의 2차 구조에 주목해야하는 충분한 이유를 제시해 준다. GPCR을 분류할 때 1차 구조뿐만이 아니라 2차 구조도 고려의 대상이 되어야 함을 확인 하였으며, 정밀분석을 위한 후보 GPCR에서 배제되는 확률을 낮춤으로써 서열 분석의 정밀도를 향상시킬 수 있다.

참고문헌

〔1〕 이상주, *"IT기반 바이오인포매틱스 인프라 구축 및 응용 연구,"* 한국과학기술정보연구원, 연구보고서, 2007. 2.

〔2〕 Clare Ellis et al, *"THE STATE OF GPCR RESEARCH IN 2004"* Nature Reviews Drug Discovery, Vol. 3, pp. 557-626, 2004.

〔3〕 The International Human Genome Mapping Consortium, *"A physical map of the human genome,"* Nature, Vol. 409, pp. 934-941, 2001. 〔

〔4〕 전주홍, *"GPCR 분석 기반기술"*, KOSEN Expert Review, 2006. 2.

〔5〕 Galperin, M.Y. *"The Molecular Biology Database Collection: 2007 update"* Nucleic Acid Res., 35, D3-D4, 2007.

〔6〕 김진홍, *"XML 기반의 단백질 구조 기술 언어를 이용한 단백질 구조 비교 시스템,"* 울산대학교 박사학위논문, 2004. 12.

〔7〕 장종원, *"상동성 기반 마르코프 모델을 이용한 미생물 유전자 예측 기법에 관한연구,"* 영남대학교 박사학위논문, 2004.

〔8〕 장종원, 류윤규, 구자효, 윤영우, *"통합형 미생물 유전자 예측 시스템의 구축에 관한 연구,"* 신호처리 시스템학회 논문지, 제6권 1호, 27-32쪽, 2005년 1월

저 자 소 개

구 자 효(Ja-hyo Ku)

2000년2월 계명대 컴퓨터공학과(공학사)
2002년8월 영남대 컴퓨터공학(공학석사)
2005년2월 영남대 컴퓨터공학(박사과정)

한 찬 명(Chan-Myung Han)

2002년2월 영남대 컴퓨터공학과(공학사)
2007년8월 영남대 컴퓨터공학(공학석사)
2007년9월 영남대 컴퓨터공학(박사과정)

윤 영 우(Young-woo Yoon)

1972년2월 영남대 전자공학과(공학사)
1983년2월 영남대 전자공학과(공학석사)
1988년2월 영남대 전자공학과(공학박사)
1988년9월 현재 영남대 컴퓨터공학과 교수