Bioinformatics and Biosystems 2007, Vol. 2, No. 1, pp. 7-11

Kogs데이타베이스로부터 얻은 계통학적인 아미노산 치환행렬

(A phylogenetic amino acid substitution matrix from Kogs database)

안 희 성 , 김 상 수

숭실대학교 생명정보학과

초 록

하나의 아미노산이 다른 아미노산으로 바뀌는 가능성을 계통학적인 나무를 이용해서 치환행렬로 만들었다. PFMT(Phylogenetic Focused Mutaion Tendency)행렬은 기존의 PAM160이나 BLOSUM62와 다르게 공통조상으로부터 상위 종으로 치환되는 가능성을 점수화 하였다. COGs의 데이터베이스에 있는 152KOGs를 뽑아서 아미노산의 치환횟수를 점수화 하였다. PFMT 행렬은 어떤 서열보다 더 상위 종의 서열을 비교할 때 유용하게 쓰일 수 있으며 20개의 아미노산간의 치환 관계를 더 자세하게 볼 수 있게 한다.

키워드: 계통학적인 나무, 치환 행렬

Abstract

Methods for making scoring matrix based on phylogenetic tree for all possible exchanges of one amino acid with another. PFMT(Phylogenetic focused Mutation Tendency) matrix is different BLOSUM62 and PAM160 which are the most used scoring matrixes. This matrix calculates possibility of substitution from common ancestor to high spices. PFMT matrix scores substitution frequency in COGs databases which contain 152 KOGs's dataset. PFMT matrix usefully is able to compare between query sequence and sequences of more higher species and show detailed substitution relation of 20 amino acids.

Keywords: phylogenetic tree, substitution matrix

서 론

현재 생물학에는 컴퓨터를 기반으로 하는 서열정렬을 하는 유용한 도구들이 많이 있다. 이 도구들은 유전자와 단백질 기능을 살펴보는데 중요한 인식을 준다. 서열을 정렬(alignment)하는 서로 다른 종류의 방법들이 있는데 공통조상으로 연관된 길이가 비슷한 두 서열을 비교하는 경우 전체정렬(global alignment), 단백질의 연관된 부분을 알아보는 지역정렬(local alignment), 단백질 family의 구성요소를 알아보는 다중 정렬(multiple alignment), 정렬은 데이터베이스를 검색해서 상동성을 갖는 서열(homologous sequence)을 찾는 일을 하기 위해서 만들어 진다. 각각의 경우마다 정렬은 점수배합(scoring scheme)을 이용해서 측정되어진다 (Steven, H. et al, 1992). 최근에는 서열을 정렬하는 경우 BLOSUM(BLOcks SUbstitution Matrix)62과 PAM(Pointed Accepted Mutation) 160이 기본값으로 쓰이고 있다. Dayhoff의 PAM은 적어도 85%동일한 서열들을 이용하여 각각의 아미노산이 가장 최소한 돌연변이되는 서로 치환된 아미노산 쌍의 횟수를 이용해서 점수행렬(scoring matrix)을 만들었다. 1% 치환된 행렬을 PAM1, 이를 이용해서 먼 상동하는 것을 Marcov chain rule을 이용해서 PAM250까지 만들었다.[2-5] 한편, BLOSUM62의 경우는 서열간의 유사성(similarity)이 62%이상 되는 단백질 과(protein family)를 BLOCKs데이터베이스를 이용해서 찾아서 점수행렬을 만들었다 (Steven, H. et al, 1992). 이 두 점수행렬이 BLAST에서 서로 상동한 서열을 찾는데 도움을 주고 있다. 하지만 현재 많은 종(species)의 모든 서열이 완전하게 해독되었고 종들 간의 관계를 파악하는 분야가 연구되고 있다. 이러한 정보를 이용하면 서열 간을 비교할 때 계통학적인 관점에서 유사한 서열을 찾을 수 있을 것이다. 이는 BLOSUM62, PAM160과는 다른 점수행렬이 필요할 것이고 이를 극복하기 위해서 공통조상(common ancestor)에서 상위생물로의 방향성을 가진 새로운 점수행렬을 만들었다. 여기에 추가적으로 20개의 아미노산이 치환되는 경향성도 확인해 보았다.

방 법

1. Kogs데이터베이스로부터 횟수 테이블 유도하기

COGs(Cluster of Orthologous Groups of proteins)의 데이터베이스는 단백질 과(protein family)별로 서로 다른 종의 아미노산 서열을 모아놓은 곳이다.(Roman L. T. et al, 2001) (http://www.ncbi.nlm.nih.gov/COG/) 여기에 진핵의 완전한 지놈(Eukaryotic complete gonomes)의 특정 7종(Arabidopsis thaliana(thale cress), Caenorhabditis elegans(worm), Drosophila melanogaster(fruit fly), Homo sapiens(human), Saccharomyces cerevisiae(baker yeast), Schizosaccharomyces pombe(fission yeast), Encephalitozoon cuniculi (Microsporidia))을 기준으로 해서 모아 둔 KOGs가 있다. 각 KOG는 하나의 단백질 과를 지정한다. 여기서 5종(Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens, Saccharomyces cerevisiae, Schizosaccharomyces pombe)만이 가지고 있는 공통된 152KOGs를 뽑는다. 이 표본데이터를 clustalW를 이용해서 다중서열을 한 후 이 5종의 계통학적인 나무(Phylogenetic tree)을 이용하여 그림1에서와 같은 tree를 보이는, fission yeast와 baker yeast가 조상(ancestor)라고 나온, 132KOGs를 뽑았다. 여기서 2종(fission yeast, baker yeast)은 외집단(outgroup)으로 선택하고 나머지 3종(worm, human, fruit fly)을 비교했는데 2종(fission yeast, baker yeast)을 선택한 이유는 계통학적으로 봤을 때 나머지 3종(worm, human, fruit fly)의 공통조상(common ancestor)에 가까운 기준점을 찾기 위해서였다. 그림1처럼 종간의 계통학적인 나무를 만들고 3종(worm, human, fruit fly)과 진화적인 거리가 가까운 종을 나머지 2종(fission yeast, baker yeast)에서 선택해서 그 종의 단백질 과와 아미노산이 같은 종을 3종(worm, human, fruit fly)에서 선택해서 그 종을 공통조상이라고 가정한다. 예를 들어 그림2에서 fission yeast와 baker yeast(2종의 순서는 바뀌어도 무관하다.)중에 3종(3종의 순서 또한 바뀌어도 무관하다.)과 가까운 fission yeast를 외집단으로 잡고 fission yeast의 각각의 아미노산을 인간, 초파리, 벌레와 비교하여 같은 아미노산을 가진 종을 공통조상으로 잡고 그림2와 같이 나머지 2종의 아미노산과 비교한다. 만약 같은 아미노산이 외집단과 공통조상으로 잡은 종에 없을 경우는 제외한다. 또한 각 아미노산끼리 비교할 때 gap이 있는 경우 또한 제외한다. 20x20 횟수행렬을 구한 것을 토대로 발견된 횟수와 기대되는 횟수간의 odds ratio를 계산해서 이를 테이블 행렬로 사용한다.

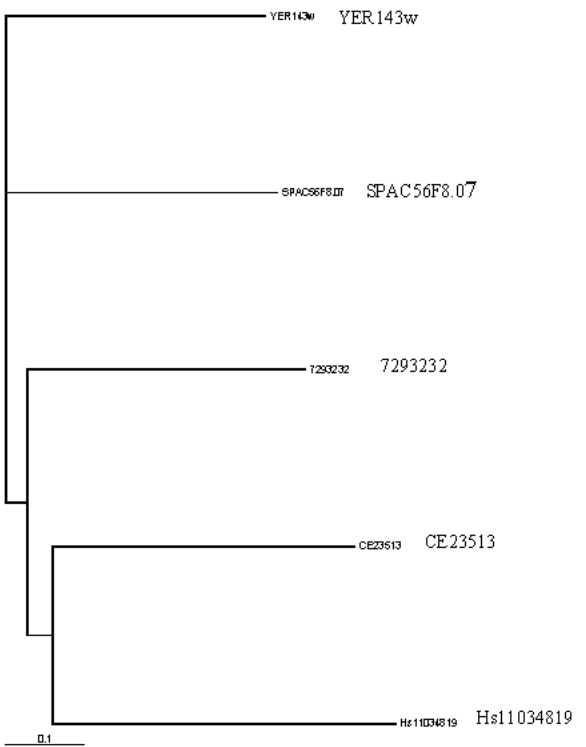

그림 1. Pyhlogenetic tree in KOG0012.
YER143w는 Saccharomyces cerevisiae의 서열, SPAC56F8.07는 Schizosaccharomyces pombe의 서열, 7293232는 Drosophila melanogaster의 서열, CE23513은 Caenorhabditis elegans의 서열, Hs11034819는 Homo sapiens의 서열이다.

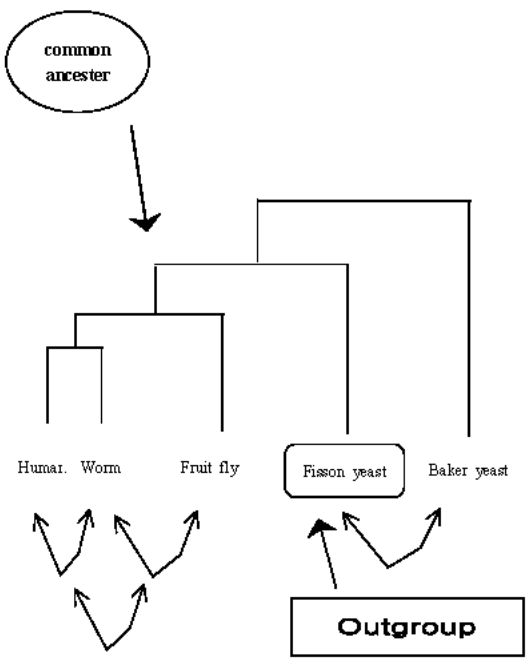

그림 2. KOG0012에서 전형적인 계통학적인 나무. 공통조상을 찾기 위해서 fission yeast를 외집단으로 선택함. Human, worm, fruit fly는 하나의 그룹으로 각 종의 계통학적 위치는 상관이 없지만 이 그룹안에 존재해야하고 이와 마찬가지로 fisson yeast, baker yeast도 이 원칙이 적용된다.

2. Logarithm odds(Lod) Matrix 계산하기 (Steven, H. et al, 1992)

비대칭적인 20 X 20 횟수 행렬의 아미노산 i ,j쌍들의 총 숫자가 f_{ij}에 써있다. 그 다음 발견된 확율을 각 I ,j쌍의 발견을 통해서 구하면

$$q_{ij} = f_{ij} / \sum_{20}^{i=1} \sum_{i}^{j=1} f_{ij}$$

예를 들어 한 열(column)에 4A와 1D가 있다고 가정하자. 그러면 이 열에서 아미노산을 비교하는 횟수는 4+3+2+1 = 10이 된다. 여기서 f_{AA} = 6이고 f_{AD} = 4이 된다. $q_{AA} = 6/10 = 0.6$, $q_{AD} = 4/10 = 0.4$이다. 그 다음에 나는 각 (i, j)쌍이 일어날 기대확률을 측정한다. A는 6번 양쪽에 나타났고 4번 한 쌍에서 한곳에만 나타났다. 따라서 A가 이 열에 나타날 기대확률은 [6+[4/2]]/10 = 0.8가 된다. D는 [4/2]/10 = 0.2이다. 일반적으로 (i, j)쌍에서 i번째 아미노산이 나타날 확률은

$$p_i = q_{ij} + \sum_{j \neq i} q_{ij}/2$$

각 (i, j)쌍에 나타날 기대확률은 $i = j$일 때는 $p_i p_j$가 되고 $i \neq j$일 때는 $p_i p_j + p_j p_i = 2 p_i p_j$가 된다. 예를 들어 AA의 기대확률은 0.8 * 0.8 = 0.64 이고 AD와 DA는 2 * (0.8 * 0.2) = 0.32이 되고 DD는 0.2 * 0.2 = 0.04이 된다. 이 odd ratio 행렬은 각 요소의 q_{ij}/e_{ij}이고 log ratio 행렬을 구하면 $s_{ij} = \log_2 (q_{ij}/e_{ij})$로 나타낸다. $s_{ij} < 0$이면 기대확률보다 실제 일어날 확률이 낮은 것이고 $s_{ij} = 0$이면 기대치와 실제 일어날 확률이 같은 것이고 $s_{ij} < 0$이면 기대확률보다 실제 일어날 확률이 더 높다고 할 수 있다. Log ratio는 규모 요소(scaling factor)를 2로 PAM이나 BLOSUM과 같이 곱해주었다.(Altschul, S. F et al, 1990) 이 행렬 안에서 각 아미노산의 평균 상호적인 정보, H(상대적인 엔트로피(relative entropy))를 계산할 수 있고 각 비트 유닛마다 기대되는 점수, E(expected score)를 구할 수 있다.

$$H = \sum_{i=1}^{20} \sum_{j=1}^{i} q_{ij} \times s_{ij}, \qquad E = \sum_{i=1}^{20} \sum_{j=1}^{i} p_i \times p_j \times s_{ij}.$$

3. KOGs 데이타베이스 구축하기

여기서 이 일을 하면서 NCBI(National Center for Biotechnology Information)에서 만들어 놓은 COGs의 안에 KOGs를 이용하였다. 총 4852KOGs 중에서 5종에만 있는 152KOGs를 이용하였고 여기에는 1481개의 단백질이 있었다. 152KOGs는 작용기작(mechanism), 정보저장(infomation storage)과 전달(information processing), 세포신호(cullular signal)과 세포처리(cellular processes) ,등 여러 가지의 기능을 가지고 있었다. 각 KOG는 단백질 family별로 서로 다른 종의 서열정보를 알고 있고 GENBANK와 연결되어 있다. 이 때 한 종이 한 개 혹은 여러 개의 서열정보를 가지고 있다. 한 단백질 가족에 한 종이 여러 개의 서열이 있다면 서로 paralog일 가능성이 있기 때문에 한 종에서 임의적으로 서열을 하나씩만 뽑았다. 이 서열정보를 이용해서 5종간의 clustalW를 이용하여 다중서열정렬결과(multiple sequence alignment result)를 얻었다. 그리드 컴퓨팅에서 효율적인 결과를 보여주고 있다. 따라서 다양한 길이의 서열을 대상으로 하게 되는 실제 분석에서도 유용할 것으로 예측된다.

결 과

1. 비대칭적인 행렬(asymmetric matrix)에서 얻은 정보

PFMT(Phylogenetic Focused Mutation Tendency)행렬(그림 3)은 비대칭적인 값을 가지고 있다. PAM과 BLOSUM과 다른 가정에서 출발하였는데 아미노산 쌍에서 서로 돌연변이(mutation)되는 확률이 다르다는 것이다. 예를 들어 알라닌(Alanine)이 히스티딘(Histidine)으로 바뀌는 확률과 히스티딘이 알라닌으로 바뀌는 확률은 서로 다르다고 본다. 이 PFMT 행렬의 가정은 공통조상에서 변화된 아미노산의 치환을 점수행렬로 나타낸 것이기 때문에 각각의 아미노산의 쌍에는 다른 점이 있다고 본다. 예를 들어 A아미노산에서 B아미노산으로 치환되는 경우와 B아미노산에서 A아미노산으로 치환되는 경우가 서로 같은 확률이 아닌 두 아미노산은 다른 경향성(tendency)를 가지고 진화해 온 것이다. 부분적으로 차이가 큰 아미노산을 보면 아르기닌(Arginine)에서 세린(Serine)으로 갈 때는 값이 '0'인데 반대방향으로 갈 때는 값이 '-4'로 차이가 있었다. 다시 말해서 아르기닌에서 세린으로 치환될 가능성이 더 크다는 뜻이다. 20개의 아미노산은 특성에 따라서 극성그룹(polar group), 소수성 그룹(Hydrophobic group), 전하를 띈 그룹(charged group)으로 분류할 수 있다. 아르기닌은 전하를 띈 그룹에 속하고 세린은 극성그룹에 속한다. 이 예로보아 그룹과 그룹 간에 서로 다른 두 아미노산이 치환이 이루어 질 때 아미노산 서열을 전체적으로 봤을 때 경향성이 있다는 것을 확인할 수 있었다. 반대로 생각해 보면 위의 경향성과 반대되는 경향성으로 아미노산이 돌연변이 된다면 그 단백질의 아미노산의 어느 부분인가도 중요하겠지만 확률적으로 그 단백질은 기능을 잃어버릴 가능성이 크다. 다른 예로는 히스티딘에서 트립토판(Tryptophan)으로 변하기는 쉬운데 반대로 변하는 것은 어려운 경우와 메티오닌(Methionine)에서 트립토판으로 치환되기는 쉬운데 반대로 변하는 것은 어려운 경우가 있었다. 이 2가지 경우는 생물학적인 관점에서 봤을 때 의미가 있어 보인다.

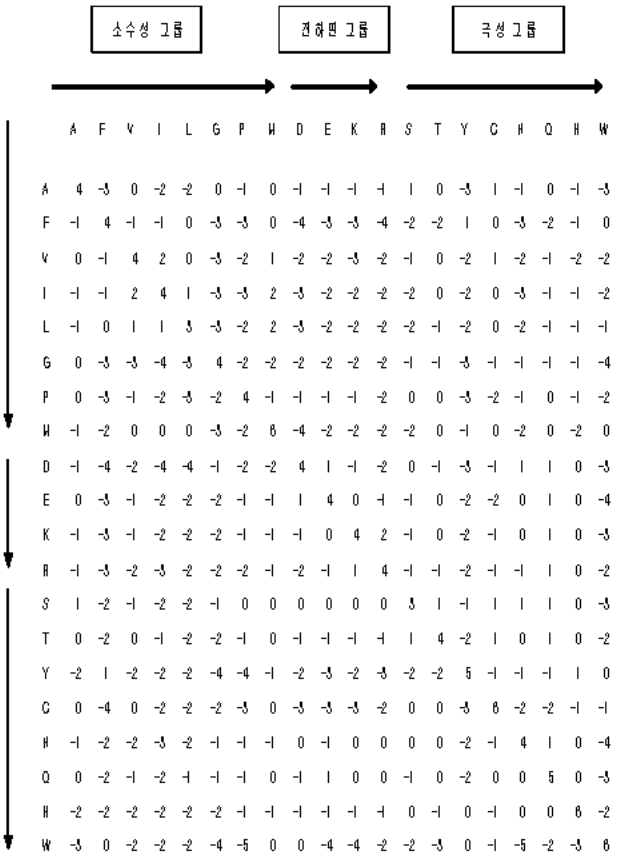

그림 3. PFMT Matrix.

2. PAM70과 비교

여기서 구한 PFMT행렬과 기존의 점수행렬을 비교하기 위해서 PFMT행렬의 상대적인 엔트로피를 구했다. 이 행렬의 상대적인 엔트로피 값은 1.66이었고 이 수치는 PAM70과 유사하며 BLOSUM의 어떠한 행렬보다 더 높게 나왔다. 특히 BLOSUM90보다 높게 나온 이유는 내가 만든 행렬의 표본의 성질 때문이다. 표본데이터가 진핵세포에서 계통학적으로 가까운 3종(fruit fly, human, worm에 속하는 단백질 family의 서열들을 뽑았기 때문일 것이다. 그래서 BLOSUM행렬과는 비교하기 어렵고 PAM행렬 중에 PAM70과 비교하였다. PAM70과 상대적인 엔트로피는 비슷하지만 기대값(Expected value)는 PFMT 행렬이 더 높았다. (e.g PAM70: -2.770, PFMT: -0.53)PAM행렬에서 PAM의 수가 높아질수록 기댓값은 낮아진다 (Steven, H. et al, 1992). 두 행렬 상에서 각 요소들을 비교하여 서로 부호가 반대로 나타나는 아미노산 쌍을 찾아보았다. 그 중에 특별히 글루타민(Glutamine)과 시스테인(cysteine)으로 치환되는 성분에 값이 서로 반대로 나왔다. {알라닌, 세린, 트레오닌(Threonine)} -> 시스테인으로 {아스파라긴(Asparagine), 리신(Lysine), 세린, 트레오닌} -> 글루타민으로 치환되는 확률이 PAM70행렬에서보다 PFMT 행렬에서 더 경향성이 컸다. 시스테인은 20개의 아미노산에서 극성 그룹안에 속하고 세린과 트레오닌 역시 이 그룹에 속해서 별 특이한 점은 없지만 소수성그룹에 속하는 알라닌이 시스테인으로 변하는 경향이 PAM70에서 보다 PFMT에서 더 크다는 뜻이 특별하다. 한편 글루타민은 아스파라긴, 세린, 트레오닌과 함께 극성그룹에 속하는데 양전하를 띄는 리신에서 글루타민으로 치환되는 경향이 더 크다는 점을 알 수 있었다. PAM70은 어떤 아미노산이 특정 아미노산 그룹 속할 경우 다른 그룹의 아미노산으로 치환될 가능성이 매우 희박한데 비해서 PFMT행렬은 이러한 변화를 민감하게 측정하였다.

3. 허브 아미노산 찾기

PFMT행렬과 PAM70행렬의 각 구성요소의 값을 비교해 보았다. PAM70에서는 음수값을 갖지만 PFMT행렬에서는 양수값을 갖는 구성요소를 찾아보았더니 특이한 점을 발견하였다. 극성그룹에 속하는 시스테인으로 소수성 그룹의 알라닌과 발린이 치환되는 확률이 PAM70보다 높았고 전하를 띤 그룹의 리신은 극성그룹의 글루탐산으로 치환되는 점수가 PAM70보다 높았다.(그림4) 보여주고 있다.

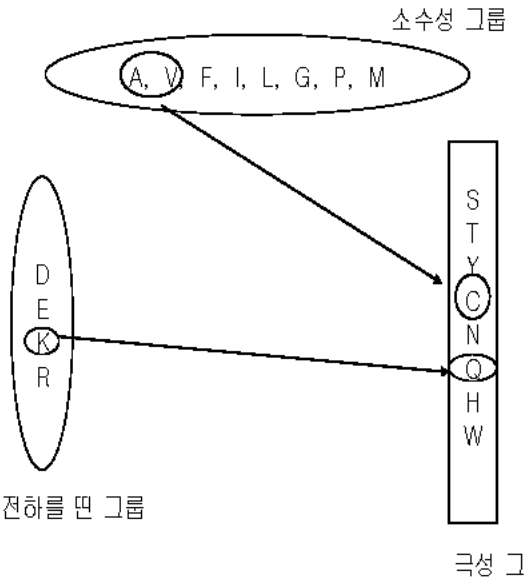

그림 4. 허브 아미노산. 소수성 그룹의 알라닌과 발린이 극성 그룹의 시스테인으로 전하를 띤 그룹의 리신이 극성 그룹의 글루탐산으로 치환되는 확률이 PFMT 행렬에서 PAM70보다 더 크게 나타남.

토 의

여기서 계통학적인 나무에 근거한 점수행렬을 찾았다. 기존의 점수행렬에서 다루기 힘들었던 아미노산의 특성그룹간의 이동 정보를 포함한 이 행렬은 공통조상을 중심으로 상위의 존재하는 종의 서열을 찾는데 유리할 것이다. PFMT행렬에서는 하나의 아미노산이 다른 아미노산으로 치환될 때 방향성이 존재하며 이것을 비대칭적인 행렬로 표현하는 것이 더 의미가 있어보였다. PAM과 BLOSUM과 다르게 말이다. PFMT행렬을 이용해서 BLAST를 통한 서열 정렬을 하고 싶지만 이 점수함수의 가정 상 BLAST를 통한 서열검색은 의미가 없다고 판단하여서 검색은 하지 않았다. 이유는 공통조상을 중심으로 그 상위 종의 서열을 찾는데 유리한 이 점수행렬을 사용하기 위해서는 query 서열을 공통조상으로 넣어주어야 하는데 현재 공통조상은 예측만 가능하지 실제로 존재하지 않기 때문에 이 점수행렬에 맞는 query 서열을 찾기가 어렵기 때문이다. 만약 BLAST로 상동성의 갖는 서열을 검색을 할 때 이 query 서열에서 공통조상을 예측한 후 BLAST를 돌리거나 아니면 이 점수행렬에 맞는 검색 도구를 만든다면 우리는 서열간의 종의 정보와 함께 그 서열간의 진화적인 시간을 측정이 가능할 수 있을 것이다.

기존의 PAM과 BLOSUM에 비해 표본의 크기가 매우 적다. PFMT행렬을 구성하기 위한 표본의 크기는 진핵세포 중에서 3종(Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens)으로 한정되어 있고 단백질의 개수도 약 660개 정도로 작았다. 하지만 명확하게 그리고 연구가 가장 많이 된 종으로 이 실험을 수행하여서 새로운 결과들을 알아내었다. 아미노산들이 치환되는 방향성이라든지 서로간의 진화적으로 상위의 종의 서열을 측정할 때는 비대칭적인 행렬이 더 의미 있다는 사실을 알아내었다. 또한 각 아미노산의 특성에 따른 분류에서 특정한 아미노산이 다른 그룹으로 옮겨가는데 허브역할을 하는 것도 알아 내었다. 시스테인과 트립토판은 자연계에 존재하는 비율이 다른 아미노산에 비해 적지만 PFMT행렬에서 극성 그룹의 시스테인이 소수성 그룹의 아미노산과 치환되는 확률이 높고 두 그룹을 이어지는 다리역할을 한다는 사실은 흥미롭다. 시스테인의 진화적인 방향성에 대해서 더 연구해 볼 필요성이 있겠다. 표본 데이터를 여러 종으로 늘려서 좀 더 먼 종간의 상동성을 가진 서열을 찾는 점수행렬을 만들고 싶다.

참 고 문 헌

[1] Steven, H. & Jorja G. H (1992) Amino acid substitution matrices from protein blocks (Proc. Natl. Acad. Sci. USA) Vol. 89, pp. 10915-10919, November 1992.

[2] George, D. G., Barker, W. C. & Hunt, L. T. (1990) Methods Enzymol. 183, 333-351.

[3] Dayhoff, M. (1978) Atlas of Protein Sequence and Structure (Natl. Biomed. Res. Found., Washington), Vol. 5, Suppl. 3, pp. 345-358.

[4] Altschul, S. F. (1991) J. Mol. Biol. 219, 555-565.

[5] Dayhoff, M, O. & Eck, R. V., eds. (1968) Atlas of Protein Sequence and Structure (Natl. Biomed. Res. Found., Silver Spring, MD), Vol. 3, p. 33

[6] Roman L. T & Darren A. N, (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes Nucleic Acids Research, 2001, Vol. 29, No. 1 22-28

[7] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) J. Mol. Biol. 215, 403-410.