Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 4, pp. 745~758

Analysis on the Amino Acid Distributions with Position in Transmembrane Proteins

Sang-Mun Chi[1)]

Abstract

This paper presents a statistical analysis on the position-specific distributions of amino acid residues in transmembrane proteins. A hidden Markov model segments membrane proteins to produce segmented regions of homogeneous statistical property from variable-length amino acids sequences. These segmented residues are analyzed by using chi-square statistic and relative-entropy in order to find position-specific amino acids. This analysis showed that isoleucine and valine concentrated on the center of membrane-spanning regions, tryptophan, tyrosine and positive residues were found frequently near both ends of membrane.

Keywords : chi-square statistic, position-specific distribution, relative-entropy, transmembrane protein

1. 서론

막단백질은 펌프나 채널과 같은 운반 시스템, 수용체, 에너지 변환기, 그리고 촉매와 같은 생물학적인 역할을 수행함으로써, 광합성, 시각작용, 뉴런의 흥분작용, 호흡, 면역반응, 한 세포에서 다른 세포로의 신호 전달 등과 관련 된다 (Berg et al., 2002). 약학적으로 GPCR (G-protein coupled receptor) 등의 막단백질은 대부분의 약물치료의 대상이며 (Dahl et al., 2002), 모든 유전체의 20% - 30%를 차지하고 있으므로 (Hessa et al., 2005), 막단백질의 구조와 기능에 대한 연구가 활발하다.

막횡단 단백질의 구조형성에는 위치특이적인 아미노산이 결정적인 역할을 한다. 아미노산의 분포에 대한 연구로 널리 알려진 예로는 양으로 하전된 잔기를 가진 아미노산이 세포질쪽에 많이 존재한다는 법칙이다 (positive-inside rule; Heijne, 1989). 막횡단 단백질의 구조를 형성하는 과정에서 단백질 전도채널 (protein conducting channel)인 트란스로콘 (translocon)과 정전기적인 상호작용을 통해 방향성이 결정되

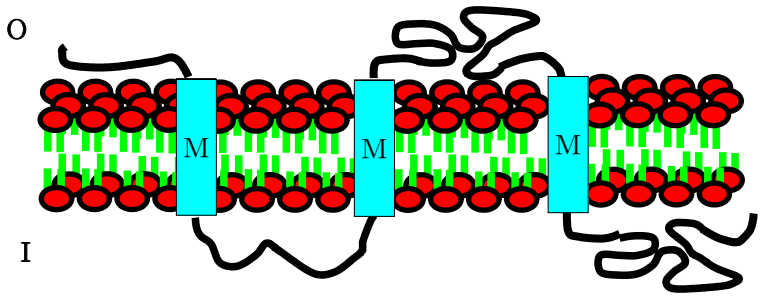
1) Assistant Professor, Department of Computer Science, College of Multimedia, Kyungsung University, 110-1 Daeyeon-dong, Nam-gu, Pusan, 608-736, Korea
E-mail: smchiks@ks.ac.kr

고, 이러한 정전기적인 상호작용에서 트란스로콘에 가까운 아미노산 잔기일수록 영향을 크게 미친다 (Goder and Spiess, 2003; Goder et al., 2004). 전하에 의한 영향은 N-말단부의 아미노산 고리에서 제일 강하고, 세포막에 내재되는 나선형 분절과 분절 사이의 아미노산 고리에도 적용이 된다. (Gafvelin et al., 1997). 처음으로 세포막에 내재되는 나선형 분절인 신호서열의 소수성 또한 막단백질의 방향에 중요한 역할을 하는데, 소수성이 강할수록 처음에 트란스로콘과 막내부 사이에 묶인 상태를 유지하는 특성이 있다 (Goder and Spiess, 2003).

아미노산 분포에 대한 연구는 삼차원 구조가 알려진 막횡단 단백질을 사용하여 위치별로 나타나는 아미노산의 빈도를 분석하거나 생화학적인 모의실험을 통해서 이루어진다 (Granseth et al., 2005; Bowie, 1997; Hessa et al,, 2005). 소수성 환경인 세포막 내부에 막 단백질의 일부가 존재하므로, 구조를 알아내기 위한 준비 과정인 막 단백질을 결정화하는 단계에 난점이 있어서, 많은 수의 막 단백질의 구조는 밝혀져 있지 않다. 따라서 삼차원 구조가 알려진 막단백질을 이용하는 경우에는 자료의 수가 많지 않아 조사 분석의 유의성이 작다. 본 논문에서는 보다 많은 자료의 사용을 가능하게 하는 동시에 위치특이적인 분포에 따라 아미노산 서열을 분할하기 위하여, 아미노산 서열을 카이제곱 검정과 상대 엔트로피를 이용하여 부분영역으로 분할한다. 분할된 영역의 위치특이적인 아미노산 분포를 조사하고, 알려진 실험적인 사실들과 비교하여 생화학적 의미를 살펴본다. 또한, 빈도로부터 유도한 통계량을 사용하여 아미노산의 위치특이성에 대한 판단을 용이하게 한다.

2. 막횡단 단백질의 분할

막횡단 단백질은 그림 1과 같이 세포막을 가로지르는 형태를 갖는다. 막횡단 단백질을 구성하는 아미노산 서열을 세포막 내부의 고리 I, 세포막 외부의 고리 O, 세포막에 내재되는 나선형 분절 M으로 나타낼 때, 이 세 개의 구조적 단위에 속하는 아미노산의 분포는 크게 다르다. 동일한 단위내에서도 위치특이적인 아미노산의 분포가 막단백질의 구조와 기능에 중요한 역할을 한다. 본 논문에서는 구조적 단위내에서의 위치특이적인 분포를 조사하기 위해서 구조적 단위를 부분영역으로 분할하고, 분할된 아미노산 서열의 차이를 알아본다.



<그림 1> 막횡단 단백질 구조의 간략도

세포막에 내재하는 나선형 분절은 여러 개의 분절로 이루어 질수 있다. 각 분절이 세포막과 이루는 각도가 달라 서열내의 아미노산의 위치와 세포막에서의 실제 위치가 선형적으로 비례하지는 않는다. 따라서, 서열을 길이에 따라 등분할하는 대신에 아미노산 서열이 위치특이적인 여러 개의 확률분포로 가진다고 가정하고, 확률분포에 따라 아미노산 서열을 분할한다. 세포막에 내재하는 아미노산 서열의 길이 자체와 각각의 확률분포들에 속하는 아미노산 서열의 길이가 모두 가변적이므로, 가변적인 길이의 아미노산 서열을 확률분포에 따라 분할할 수 있어야 한다. HMM (Hidden Markov Model)은 막횡단 단백질을 모델링 하는데 높은 성능을 보이는 통계적인 모델로서 (Möller et al., 2001; Chen et al., 2002; Arai et al., 2004), HMM의 상태별로 서열을 분할하면 가변 길이의 아미노산 서열을 확률분포에 따라 분할할 수 있다. 본 논문에서는 상태에 종속적인 출력확률과 천이확률을 갖는 표 1의 정의와 같은 HMM을 사용한다.

<표 1> HMM 정의

$O = o_1, o_2, \ldots, o_T$	출력 관측치로서, 본 논문에서는 아미노산 서열
$\Omega = \{1, 2, \ldots, N\}$	상태집합
$A = \{a_{ij}\}$	천이확률 행렬로서, a_{ij}는 상태 i에서 상태 j로 천이할 때의 확률
$B = \{b_j(k)\}$	출력확률 행렬로서, $b_j(k)$는 상태 j에서 아미노산 o_k를 출력할 확률
$\Pi = \{\pi_i\}$	π_i는 초기 상태가 i일 확률

아미노산 서열을 상태별로 분할하기 위해 Viterbi 알고리즘을 사용한다. Viterbi 알고리즘은 모든 가능한 상태의 경로를 고려하지 않고, 최대확률을 가지는 최적 상태열만을 고려하는 근사적인 방법이다. $\delta_t(j)$를 서열위치 t와 상태 j에서 아미노산 서열 o_1부터 o_t까지를 관찰할 최대 확률을 나타내고, $\psi_t(j)$를 서열위치 t에서 현재 상태 j에 도달하는 최적의 이전 상태를 저장하는 변수라면, 다음의 반복식이 성립한다.

$$\delta_t(j) = \max_i [\ \delta_{t-1}(i)\ a_{ij}]\ b_j(\ o_t).$$

$$\psi_t(j) = \arg\max_i [\ \delta_{t-1}(i)\ a_{ij}]$$

초기값은 $\psi_1(j) = 0, \quad \delta_1(j) = \pi_j\, b_j(\ o_1)$ 이다. $s_T^* = \arg\max_{s \in \text{최종상태집합}} [\ \delta_T(s)]$ 을 초기값으로 갖고 t = T-1, ..., 1 에 대해서 다음을 수행하여 최적 상태열을 얻는다.

$$s_t^* = \psi_{t+1}(\ s_{t+1}^*)$$

본 논문에서는 5개의 상태를 갖고 현재 상태자체와 오른쪽의 인접한 하나의 상태로만 천이하는 left-to-right 구조의 HMM을 사용하여 세포막에 내재하는 부분을 모델링 하였다. 세포질에서 세포외부로 향하는 서열을 모델링하는 HMM인 i_M_o과 세포외부에서 세포질로 향하는 HMM인 o_M_i을 사용하였다. 실험대상의 아미노산 서열이 전체적으로 비슷한 개수로 세포질쪽, 세포외부쪽, 그리고 가운데 부분으로 분할되도록 표 2와 같이 각 부분영역에 속하는 상태를 결정하였다.

<표 2> 막에 내재되는 아미노산 서열내에서 부분 영역의 정의

부분영역	부분 영역의 위치	부분 영역에 속하는 아미노산
M_i	세포질에 가까운 부분	i_M_o의 상태 1과 2, o_M_i의 상태 5에 해당하는 아미노산의 집합
M_m	가운데 부분	i_M_o의 상태 3과 4, o_M_i의 상태 3과 4에 해당하는 아미노산의 집합
M_o	세포외부에 가까운 부분	i_M_o의 상태 5, o_M_i의 상태 1과 2에 해당하는 아미노산의 집합

세포질과 세포외부에 존재하는 아미노산 서열은 HMM의 특정상태에 아미노산의 서열이 집중되는 현상이 발생하여 HMM을 사용하여 분할하지 않았다. 이들 아미노산 서열은 기능적인 구조인 영역 (domain)을 이루기 위해서 복잡한 구조를 갖기 때문에 서열상의 위치와 세포막에서의 거리는 근사적으로도 비례하지 않는다. 비록 불규칙한 구조를 이루기는 하지만, 그림 1과 같이 세포막에 연결된 부근의 아미노산은 세포막과의 실제 거리가 인접하다. 표 3과 같은 부분 영역을 정의하고, 이들 부분영역사이의 확률분포 차이를 크게 만드는 N을 사용하여 부분영역을 만든다.

<표 3> 세포질과 세포외부의 아미노산 서열내에서 부분 영역의 정의

부분영역	부분 영역의 위치	부분 영역에 속하는 아미노산
I_near, O_near	세포막에 가까운 부분	세포막에 연결된 아미노산과 서열상의 위치가 N 이하인 아미노산의 집합
I_far, O_far	세포막에서 먼 부분	I_near와 O_near에 속하지 않는 아미노산의 집합

I는 세포질내의 아미노산을 O는 세포외부의 아미노산을 의미한다.

3. 분할된 영역 사이의 차이 측정

분할된 아미노산 서열을 이질적인 확률분포로 만드는 위치특이적인 아미노산으로 찾기 위하여, 분할된 영역사이의 차이를 카이제곱 검정과 상대 엔트로피를 사용하여 조사한다.

3.1 카이제곱 검정을 사용한 동질성 분석

범주형 자료의 분석에 널리 이용되는 카이제곱 통계량을 사용하여, 분할된 각 영역에 존재하는 20가지 아미노산의 분포가 동일한지에 대한 검정을 한다. 영역 i에서 아미노산 j의 확률을 p_{ij}라 할 때

$$H_0: p_{1j} = p_{2j} = \cdots = p_{Nj}, \quad j = 1,2,\cdots,20$$

을 검정한다. 여기서, N은 분할된 영역의 수이다. i번째 영역의 아미노산의 총개수를 n_i, 아미노산 j의 개수를 X_{ij}라 하면, 검정통계량

$$\sum_{j=1}^{20}\left[\frac{(X_{ij}-n_i p_{ij})^2}{n_i p_{ij}}\right]$$

은 근사적으로 자유도가 20인 카이제곱 분포를 따른다. $H_0: p_{1j} = p_{2j} = \cdots = p_{Nj} = p_j$가 사실이라면 전체 모집단에서의 검정통계량

$$\sum_{i=1}^{N}\sum_{j=1}^{20}\left[\frac{(X_{ij}-n_i p_j)^2}{n_i p_j}\right]$$

은 근사적으로 자유도가 N*19인 카이제곱 분포를 따른다. p_j대신 최대가능도 추정량 $\hat{p}_j = \sum_{i=1}^{N} X_{ij} / \sum_{i=1}^{N} n_i$ 을 사용한 통계량

$$X^2 = \sum_{i=1}^{N}\sum_{j=1}^{20}\left[\frac{(X_{ij}-n_i\hat{p}_j)^2}{n_i\hat{p}_j}\right] \tag{1}$$

은 근사적으로 자유도가 (N-1)*19인 카이제곱분포를 따른다.

본 논문에서는 통계량 X^2을 사용하여 분할된 영역사이의 동질성을 검정하는데 사용한다. 또한, 확률분포사이의 이질성에 크게 기여하는 아미노산을 찾기 위해 (1)식에서 아미노산별로 계산되는 다음 값을 사용한다.

$$X_j^2 = \sum_{i=1}^{N}\left[\frac{(X_{ij}-n_i\hat{p}_j)^2}{n_i\hat{p}_j}\right] \tag{2}$$

비록 개념과 샘플링과정이 동질성 검정과는 다르나 통계량과 검정의 수행과정이 수학적으로 동일하므로, 본 논문에서는 각 분할영역과 아미노산의 분포가 독립적인지를

알아보는 독립성 검정에도 통계량 X^2을 사용한다. 즉, 통계량 X^2이 일정값보다 클 때는 분할영역과 아미노산의 분포가 독립적이라는 귀무가설을 기각한다.

3.2 상대 엔트로피를 사용한 확률분포간의 차이

카이제곱 통계량 X^2과 더불어서 상대 엔트로피 (relative entropy 또는 Kullback-Leibler divergence)

$$\sum_{j=1}^{20} p_j \log \frac{p_j}{q_j}$$

를 이용하여 분할된 두 영역의 아미노산 확률분포 사이의 차이를 측정한다. 여기서, p_j와 q_j는 분할된 두 영역에서 아미노산 j의 확률이다. 상대 엔트로피는 음이 아닌 값만을 가지며, $p_j = q_j, j = 1, 2, \ldots 20$일 경우에는 0이므로 여러 분야에서 확률분포간의 거리로 유용하게 쓰이고 있고, 실제 확률분포가 p_j일 때 확률분포를 q_j라고 가정하는 경우의 비효율성을 측정한다 (Cover and Thomas, 1991). 하지만 상대 엔트로피는 p_j와 q_j에 대해서 대칭적이지 않으며, 삼각부등식을 만족하지 않으므로 수학적으로 거리척도는 아니다.

본 논문에서는 두 분할영역을 합병한 영역으로부터 각 영역의 차이를 계산하는 방법을 사용한다 (Hwang, et al., 1996). 이 방법은 두 분포에 속하는 아미노산의 개수를 고려하여 차이를 정의할 수 있다는 장점이 있다. 두개의 분할영역 a, b에서 아미노산 j의 개수를 X_{aj}, X_{bj}로 나타내고, 각 영역의 아미노산 총개수를 n_a, n_b라 하자. 합병된 영역의 엔트로피는

$$H_{a+b} = -\sum_{j=1}^{20} \frac{X_{aj} + X_{bj}}{n_a + n_b} \log \frac{X_{aj} + X_{bj}}{n_a + n_b}$$

이다. 합병함으로서 증가한 엔트로피를 아미노산의 개수로 가중한

$$\begin{aligned} d_{WE} &= (n_a + n_b) H_{a+b} - n_a H_a - n_b H_b \\ &= n_a \sum_{i=1}^{20} \frac{X_{ai}}{n_a} \log \left(\frac{X_{ai}}{n_a} / \frac{X_{ai} + X_{bi}}{n_a + n_b} \right) + n_b \sum_{i=1}^{20} \frac{X_{bi}}{n_b} \log \left(\frac{X_{bi}}{n_b} / \frac{X_{ai} + X_{bi}}{n_a + n_b} \right) \end{aligned} \quad (3)$$

는 상대 엔트로피의 가중합이다.

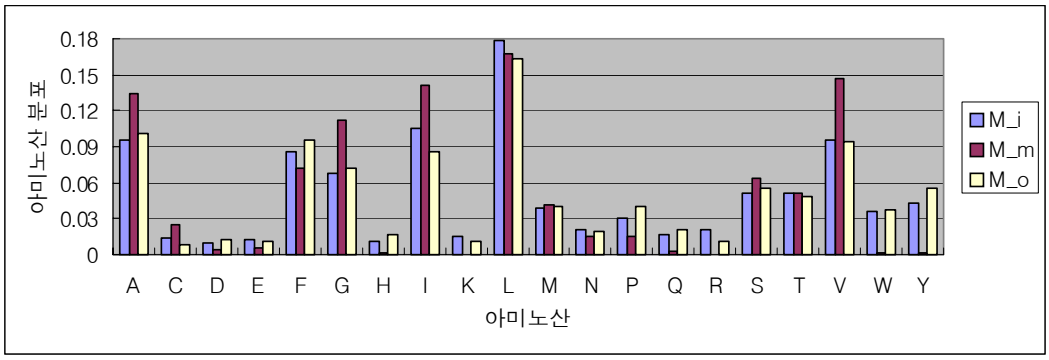
4. 아미노산의 분포 조사

4.1 막횡단 단백질 자료

아민노산의 위치특이성을 조사하기 위해서는 막횡단 부분이 실험적으로 확인된 막횡단 단백질 자료가 필요하다. 본 논문에서는 대용량자료를 확보하기 위해서 두개의 자료를 병합하여 사용한다. 하나의 자료는 ftp://ftp.ebi.ac.uk/databases/testsets/transmembrane에서 다운로드하였다. 이 자료는 883개의 막횡단 부분을 가지고 있는 188개의 단백질이다 (Möller et al., 2000). 또 다른 자료를 ftp://bioinfo.si.hirosaki-u.ac.jp/TMPDB로부터 다운로드 하였다. 이 자료는 231개의 단백질로 구성되어 있다 (Ikeda et al., 2003). 이들 두 자료는 115개의 공통된 단백질 자료를 포함하고 있다. 아미노산 잔기, 방향, 막에 내재되는 부분의 위치가 불일치되는 정보는 Möller et al.,(2000)의 주석을 사용하였다.

4.2 위치특이적인 아미노산의 분포

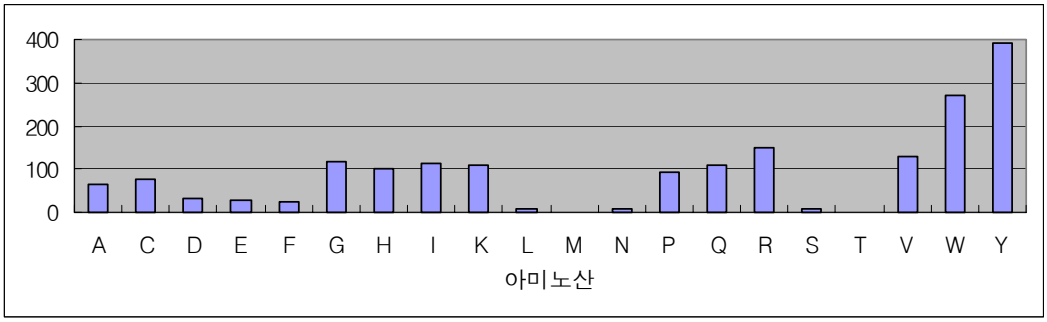
먼저 세포막에 내재하는 아미노산의 분포를 조사한다. 표 2의 정의에 따라 분할한 영역인 M_i에는 8736, M_m에는 7945, M_o에는 8215개의 아미노산이 존재하여 비교적 균등하게 분할되었다. 분할영역별로 각 아미노산의 비율을 계산하여 그림 2에 나타내었다.



<그림 2> 세포막에 내재하는 아미노산의 부분영역별 조성

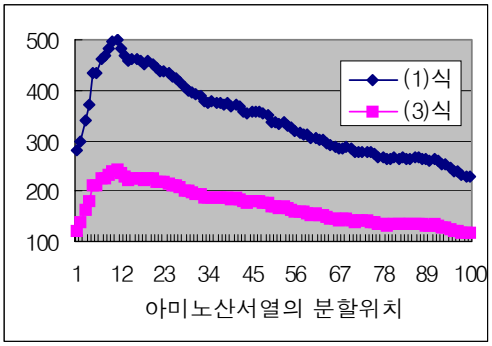
그림 2에서 보듯이 세포막의 중심부와 막의 끝단에 존재하는 아미노산의 분포가 확연히 다름을 볼 수 있다. 이들 세 부분의 아미노산 분포의 동질성을 평가하기 위해, 카이제곱 검정을 수행하였다. 통계량 X^2이 1840.174로서 자유도 38, 유의수준 0.01일 때 61.162, 0.001일때 70.703보다 크므로, 동일한 분포라는 가설을 기각할 수 있다. 분할된 영역의 확률분포사이의 차이를 수치적으로 알아보기 위해 3장에서 정의한 카이제곱 통계량과 상대엔트로피의 가중합으로 정의된 (1)식과 (3)식을 사용한다. 그림 2

에서 보듯이 M_i와 M_o는 서로 유사한 분포이고 M_m과는 구별되는데, 카이제곱 통계량 X^2으로는 M_i와 M_o의 거리는 110인 반면 M_i와 M_m은 1437, M_o와 M_m은 1625로 차이가 크다. 마찬가지로 d_{WE}는 M_i와 M_o의 거리는 55인 반면 M_i와 M_m은 871 M_o와 M_m은 965로 카이제곱 통계량과 일관되는 경향성을 갖는다.

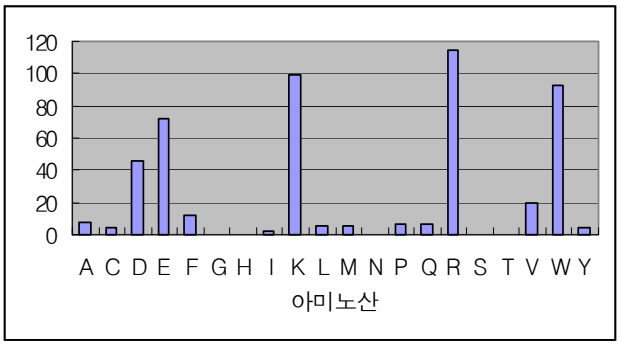


<그림 3> 아미노산별 X_j^2의 값

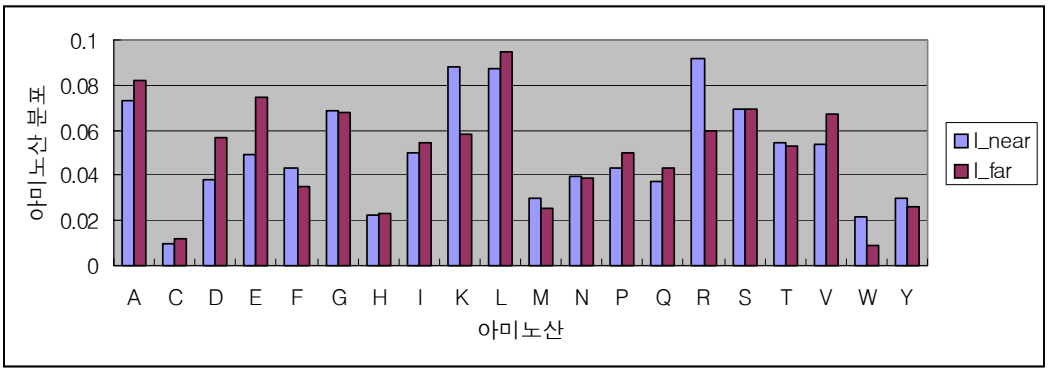
동질성 분석에 사용한 통계량 X^2을 독립성 검정에 사용한다. 같은 계산과정과 통계량을 사용하여 표 2의 정의에 따른 세 개의 부분영역과 아미노산의 분포가 독립이라는 귀무가설을 기각할 수 있다. 본 논문에서는 분할된 영역별로 아미노산의 분포가 관련되어 있다고 가정하고, 분할영역에 특이적으로 나타나는 아미노산을 조사한다. 식(2)의 X_j^2값을 사용하여 아미노산의 위치특이성을 조사한다. 그림 3에 나타나듯이 티로신(Y), 트립토판(W), 아르기닌(R), 발린(V)이 영역별 차이가 크다. 그림 2에서 보듯이 티로신, 트립신, 아르기닌은 세포막의 양쪽 경계부분에서 많이 나타났고, 발린은 막의 중심부에 나타나는 빈도가 높았다.



<그림 4> 세포질내 분할영역간의 차이

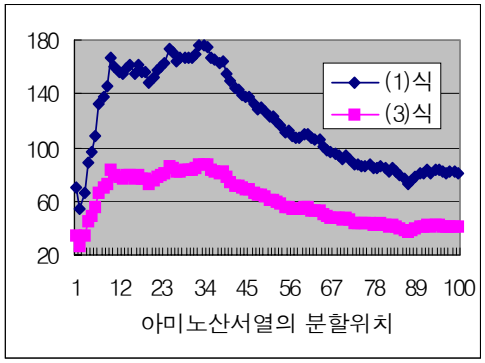


<그림 5> 아미노산별 비동질성의 기여도

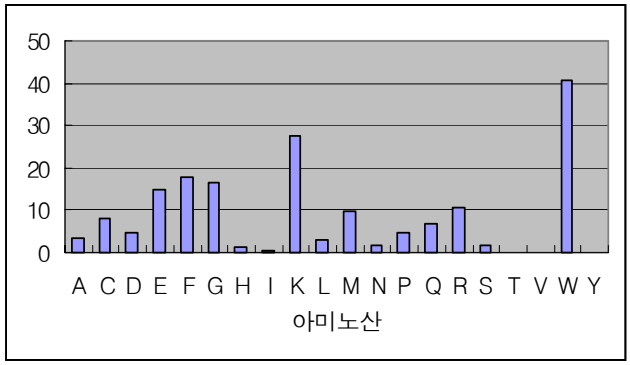


<그림 6> 세포질내의 분할영역에서 아미노산 조성

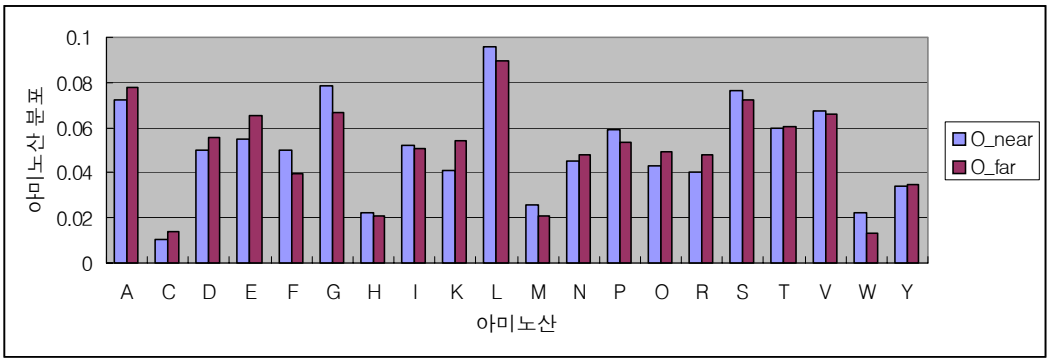
세포질에 존재하는 아미노산에 대한 조사를 위해서, 표 3의 정의에 따라 세포막에 연결된 아미노산과 서열위치로 N번째 이내의 아미노산 집합과 N번째 이후의 아미노산 집합을 나눈다. 그림 4는 (1)식의 X^2과 (3)식의 d_{WE}를 사용하여 두 분할영역의 차이를 측정한 결과이다. N이 9에서 12사이에서 확률분포간의 차이가 크다. 그림 4에 나타난 카이제곱 통계량이 자유도 19, 유의수준 0.01일때 36.191보다 모두 크므로 통계적으로 동일한 분포라는 가설을 기각할 수 있다. 아미노산 분포가 가장 큰 차이를 보이는 서열 위치 11을 기준으로 아미노산 서열을 두개의 집합으로 나누었다. 분할한 두 아미노산 집합의 비동질성에 기여하는 아미노산을 찾기 위해 X_j^2를 그림 5에 나타내었다. 아르기닌(R), 글루탐산(E), 리신(K), 트립토판(W), 아스파르트산(D)의 값이 크다. 그림 6에 보듯이 아르기닌, 리신과 트립토판은 세포막에 가까운 부분영역에 높은 빈도로 나타났고, 글루탐산과 아스파르트산은 그 이외의 위치에서 빈도가 높게 나타났다. 이러한 경향성은 11이외의 다른 위치로 분할하였을 경우에도 비슷한 경향성을 보였다.



<그림 7> 세포외부 분할영역간의 차이



<그림 8> 아미노산별 비동질성의 기여도



<그림 9> 세포외부의 분할영역에서 아미노산 조성

세포외부에 존재하는 아미노산 서열에 대해서 동일한 방법으로 조사한다. 그림 7에 보듯이 세포 외부의 서열내의 확률분포의 차이는 그림 4의 세포질 내부의 서열내의

확률분포의 차이보다 작다. 즉, 세포외부의 분할영역이 세포질내부의 분할영역보다 확연히 구별되지는 않는다는 것을 알 수 있다. 하지만 자유도 19, 유의수준 0.01일때 36.191보다 모든 위치에서 카이제곱 통계량의 값이 크므로 통계적으로 동일한 분포라는 가설을 기각할 수 있다. 아미노산 분포가 가장 큰 차이를 보이는 서열 위치는 10, 25, 33이다. 이 세 개의 위치를 기준으로 아미노산 서열을 두개의 집합으로 나누었을 경우 비슷한 결과를 보였다. 본 논문에서는 서열 위치 25를 기준으로 분할된 서열에서 아미노산별로 비동질성에 기여하는 부분을 찾기 위해 X_j^2를 그림 8에 나타내었다. 트립토판(W), 리신(K), 페닐알라닌(F), 글리신(G), 글루탐산(E)의 기여도가 크게 나타났다. 그림 9에서 보듯이 트립토판, 페닐알라닌, 글리신은 세포막쪽에 상대적으로 많이 존재하고, 리신, 글루탐산은 그 외의 위치에 더 많이 존재한다. 리신과 같이 양으로 하전된 아르기닌(R)도 비교적 세포막에서 먼 위치에 더 많이 존재한다.

4.3 토의

확률분포간의 차이를 알기위해 카이제곱 통계량과 상대 엔트로피를 이용하였는데, 두 값의 경향성이 거의 일치하므로 카이제곱 통계량만으로 구조적 단위간의 차이를 살펴본다. 세포막에 내재하는 아미노산 서열에서 세포질쪽과 세포외부쪽은 카이제곱 통계량은 110인 반면, 막의 중심부와 세포질쪽은 1437, 막의 중심부와 세포외부쪽은 1625로 매우 차이가 크므로 이질적인 아미노산 분포를 가짐을 알 수 있다. 세포질내의 부분영역간의 차이는 최대 499로 세포막에 내재하는 아미노산의 분포만큼 차이가 크지는 않았지만 K와 R등의 명확히 위치특이적인 아미노산이 존재한다. 세포외부에 존재하는 아미노산 서열의 부분영역간의 차이는 최대 176으로 다른 구조적 단위보다 작기는 하였으나 통계적로 같은 분포라는 가설을 기각할 수는 있다. 전체적으로 막횡단 단백질의 구조에 따라 영향을 받는 아미노산의 위치특이적인 분포는 막에 내재하는 아미노산 서열에서 가장 크고, 세포질내의 아미노산 서열과 세포외부의 아미노산 서열 순서로 작아짐을 확인할 수 있었다.

먼저 막횡단 단백질의 구조적인 단위중에 세포막에 내재하는 서열에 존재하는 아미노산의 위치특이성을 알아본다. 이 부분의 아미노산은 소수성이 강하고, 소수성이 강할수록 막내부 사이에 묶인 상태를 유지하는 특성이 있다 (Goder and Spiess, 2003). 아미노산의 상대적인 빈도를 알기위해 그림 2, 그림 6, 그림 9을 보면, 소수성 지방족 곁사슬을 가진 아미노산인 I, L, V가 세포막에 내재하는 아미노산서열에서 상대적으로 빈도가 높다. 또 다른 소수성 아미노산인 A, M, F는 차이는 크지 않으나 세포막에 내재하는 부분에 많이 나타났으나, P는 고리를 이루어 나선구조를 구성하기에는 입체형태에 제약이 있어서 막에 내재하는 부분에 더 적게 존재함을 볼 수 있다.

세포막에 내재하는 아미노산 분포는 중심부와 막의 양끝과 이질성이 명확하다고 알려져 있다. Sankararamakrishnan과 Weinstein (2000)은 방향족 잔기는 수소결합을 형성하고 인지질의 머리들과 상호작용을 한다고 보고하고 있으므로 이들은 막의 경계에 위치하는 빈도가 높다. 이러한 연구는 극성 방향족 곁사슬을 가진 W, Y의 분포에 관한 Granseth et al.(2005)와 Hessa et al.(2005)의 연구에서도 확인할 수 있다. Heijne(1994)의 논문에서는 방향족 곁사슬을 가진 F, W, Y는 양 끝단에 에 대한 집중되어 있고, 지방족 곁사슬을 가진 I, L, V는 중심부에 집중되어 있음을 볼 수 있다. 그림 3에 카이제곱 통계량을 이용한 위치특이적인 아미노산을 표시하였는데, W와 Y

는 가장 크게 위치특이적이고 그림 2에서 보듯이 막의 양끝에 집중되어 있다. 그러나 F는 비교적 집중도의 차이가 크지 않다. I와 V의 경우에는 Heijne(1994)의 결과와 일치하게 중심부에 집중되어 있으나, L의 경우에는 집중도의 차이가 크지 않다.

전하를 가진 아미노산 D, E, H, K, R은 막과의 경계부에 많이 존재한다고 알려졌다. 그림 3에 위치특이적으로 나타난 K와 R은 그림 2에 보듯이 세포막의 중심부보다 양끝단에 집중되어 있으며, 세포질쪽이 세포외부쪽보다 빈도가 높다. 이는 친수성을 가지므로 세포막에 내재되기 보다는 막과의 경계에 위치하는 것을 선호하고, 양전하를 가진 곁사슬이 트란스로콘의 세포질쪽의 음전하와 상호작용하기 때문이라고 여겨진다. 음전하를 가진 D, E의 경우에는 세포막에 내재하는 서열에 나타나는 빈도가 적고, 위치특이적인 특성도 작으므로 세포막에 내재하는 구조의 형성에는 기여도가 작다고 판단된다. Granseth et al.(2005)의 연구에도 D, E의 경우에는 특이한 경향성을 보이지 않았다. 이밖에도 중간의 소수성을 가진 G가 막의 중심부에 집중되어 있고, 세포막에 내재하는 빈도는 적으나 H, P, Q는 막의 양끝단에 집중되어 있다.

세포질에 존재하는 아미노산에는 양으로 하전된 잔기를 가진 K, R이 세포막 외부보다 훨씬 많이 존재함이 알려져 있다 (positive-inside rule; Heijne, 1989). 양으로 하전된 잔기는 트란스로콘의 세포질쪽과 정전기적인 상호작용을 함으로서 막단백질의 방향성에 영향을 미치고 (Goder and Spiess, 2003; Goder et al., 2004), 이러한 전하에 의한 영향은 세포막에 내재되는 나선형 분절사이의 아미노산 고리에도 적용이 된다 (Gafvelin et al., 1997). K와 R의 비율에 대해서 그림 6의 세포질내와 그림 9의 세포외부를 비교하면 세포질내에 더 많이 존재함을 알 수 있다. 하지만, 음으로 하전된 잔기를 갖는 D와 E의 경우에는 별 차이가 없었다. 양전하를 가진 잔기의 빈도차가 큰 것으로 보아, 음전하를 가진 잔기보다 구조 형성에 중요한 역할을 하는 것으로 판단된다. 더구나 전하를 가진 아미노산은 위치특이적인 분포를 갖는데, 세포막에 가까운 전하가 결정적이다 (Whitley et al., 1995). 예로서, 아미노산 고리에서 C-단말방향에 K와 D를 갖고 나선형 머리핀구조를 형성하는 아미노산 서열에서 K와 R이 막에 내재되는 소수성 서열에서 멀어질수록 머리핀 구조가 형성되지 않는다는 보고가 있다 (Hermansson et al., 2001). 이런 현상은 본 논문의 조사에도 그대로 나타났다. 그림 5에서 세포질내의 부분영역에서 위치특이성이 큰 아미노산인 K와 R이 세포막에 근접한 부분(I_near)에서 훨씬 더 빈도가 높게 존재하였다. 그러나, 음전하의 잔기를 갖는 아미노산인 D, E의 경우에는 그림 8에 나타난 위치특이성이 그림 5의 K와 R에 비해 작았다. 이밖에도 그림 5에 보이듯이 W, D, E가 위치특이성이 크다. 트립토판(W)은 인지질의 머리들과 상호작용을 하기 때문에 세포막에 근접한 위치에서 빈도가 높은 것이라 판단되며 (Sankararamakrishnan and Weinstein, 2000), D와 E는 세포막에 근접하지 않은 영역(I_far)에 더 많이 존재하여 트란스로콘의 세포질부분의 음전하와 상호작용이 최소화되도록 놓여 있다고 여겨진다.

세포 외부에 존재하는 아미노산은 전체적으로 다른 부분에 존재하는 아미노산보다 위치특이성이 적은 것으로 조사되었다. 세포 외부의 아미노산에서 위치특이성이 큰 것은 그림 8에 보이듯이 W와 K이다. 그림 9에 보듯이 W는 세포막에 근접한 부분(O_near)에 빈도가 높은데, 이는 역시 인지질의 머리들과 상호작용을 하기 때문이라고 판단된다. K는 양전하를 가졌기 때문에 세포외부에 존재하면 막횡단 단백질의 형성에 불리하지만, 세포막에 근접하지 않은 영역(O_far)에 더 많이 존재하여 트란스로콘에서 세포막 외부쪽의 양전하와 상호작용이 최소화되도록 놓여 있다고 여겨진다.

5. 결론

본 연구에서는 막횡단 단백질을 구성하는 아미노산의 분포특성을 통계적 절차와 정보이론을 이용하여 조사하고, 이를 생화학적 실험에 의해서 알아낸 사실들과 비교하였다. 또한 위치특이적인 아미노산을 쉽게 찾기 위해 통계량을 정의하여 사용하였다.

막횡단 단백질의 구조에 종속적인 위치특이적인 아미노산은 막에 내재하는 아미노산 서열에서 가장 많이 나타나고, 세포질내의 아미노산 서열과 세포외부의 아미노산 서열 순서로 작아졌다. 세포막에 내재하는 아미노산 서열에는 소수성 지방족 곁사슬을 가진 아미노산인 I, V가 나타나는 빈도도 높고, 중심부에 집중되는 위치특이성을 보였고, W, Y는 막의 경계부근에 집중되는 선호도를 보였다. 전하를 가진 K와 R은 막의 중심부보다 양끝단에 집중되어 있으며, 널리 알려진 대로 세포질쪽이 세포외부쪽보다 빈도가 높았다. 세포질내의 부분영역에서 위치특이성이 큰 아미노산인 K와 R은 세포막에 근접한 부분에 빈도가 높았고, 트립토판(W)은 세포막에 근접한 위치에서, D와 E는 세포막에 근접하지 않은 영역에 더 많이 존재하였다. 세포막 외부에 존재하는 아미노산은 위치특이성이 적었지만, W는 세포막에 근접한 부분, K는 세포막에 근접하지 않은 영역에 더 많이 존재하였다.

이러한 아미노산의 위치특이적인 분포에 대한 조사는 막단백질에 대해 보다 깊은 이해를 가능하게 하는 동시에, 위치별로 존재하는 아미노산 분포의 이질성을 이용하여 막단백질을 효과적으로 모델링하는 단위를 선정하는 데에 적용할 수 있으리라 판단된다.

References

1. Arai, M., Mitsuke, H., Ikeda, M., Xia, J., Kikuchi, T., Satake, M., and Shimizu, T. (2004). ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability, *Nucleic Acids Research,* 32, w390-w393.
2. Berg, J. M., Tymoczk, J. L., Stryer, L. (2002). *Biochemistry* 5th Ed., W. H. Freeman and Company.
3. Bowie, J. U. (1997). Helix packing in membrane proteins, *Journal of Molecular Biology,* 272, 780-789.
4. Chen, C. P., Kernytsky, A., and Rost, B. (2002). Transmembrane helix predictions revisited, *Protein Science*, 11, 2774-2791.
5. Cover, T. M. and Thomas, J. A. (1991), *Elements of information theory*, John Wiley & Sons.
6. Dahl, S. G., Kristiansen, K. and Sylte, I. (2002). Bioinformatics: from gene to drug targets, *Ann. Med.*, 34, 306-312
7. Gafvelin, G., Sakaguchi, M., Andersson, H., and von Heijne, G. (1997). Topological rules for membrane protein assembly in eukaryotic cells, *Journal of Biological Chemistry*, 273, 6119-6127.

8. Goder, V. and Spiess, M. (2003). Molecular mechanism of signal sequence orientation in the endoplasmic reticulum, *The EMBO Journal*. 22, 14, 3645-3653.
9. Goder, V., Junne, T., and Spiess, M. (2004). Sec61p contributes to signal sequence orientation according to the positive-inside rule, *Molecular Biology of the Cell,* 15, 1470-1478.
10. Granseth, E., von Heijne, G. and Elofsson, A. (2005). A study of the membrane-water interface region of membrane proteins, *Journal of Molecular Biology,* 346, 377-385.
11. Hermansson, M., Monne, M., and von Heijne, G. (2001). Formation of Helical Hairpin, *Journal of Molecular Biology,* 313, 1171-1179.
12. Hessa, R., Kim, H., Bihlmaler, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S. H., and von Heijne, G. (2005). Recognition of transmembrane helices by the endoplasmic reticulum translocon, *Nature*, 433, 377-381.
13. Hwang, M., Huang, X. and Alleva, F. A. (1996). Predicting unseen triphones with senones, IEEE trans. SAP., 4. 412-419.
14. Ikeda, M., Arai, M., Okuno, T., and Shimizu, T. (2003). TMPDB: a database of exper imentallly-characterized transmembrane topologies. *Nucleic Acids Research,* 31, 406-409.
15. Möller, S., Kriventseva, E. V.. and Apweiler, R. (2000). A collection of well characterised integral membrane proteins. *Bioinformatics*, 16, 12, 1159-1160.
16. Möller, S., Croning, M. D. R.. and Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions, *Bioinformatics*, 17, 7, 646-653.
17. Sankararamakrishnan, R. and Weinstein, H. (2000). Molecular dynamics simulations predict a tilted orientation for the helical region of dynorphin A(1-17) in dimysistoylphosphatidylcholine bilayers. *Biophysical Journal,* 79, 2331-1344.
18. von Heijne, G. (1989). Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues, *Nature*, 341, 456-458.
19. von Heijne, G. (1994). Membrane proteins: from sequence to structure. *Annu. Rev. Biophys. Biomol. Struct.* 23, 167-192.
20. Whitley, P., Gafvelin, G., and von Heijne, G. (1995). SecA-independent translocation of the periplasmic n-terminal tail of an escherichia coli inner membrane protein, *Journal of Biological Chemistry*, 50, 29831-29835.

[2005년 8월 접수, 2005년 9월 채택]