Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

염기분포와 대치 비교를 이용한 염기서열 집단의 고유 시그너쳐 추출

Characteristic Signature Extraction using the Base Distribution and Substitution Comparison

황경순[1], 이혜리[1], 이건명[1], 김성수[1], 이찬희[2], 이성덕[3], 윤형우[4]

[1]충북대학교 전기전자컴퓨터공학부
E-mail: kmlee@cbnu.ac.kr
[2]충북대학교 생명과학부
[3]충북대학교 정보통계학과
[4]주성대학 임상병리학과

요 약1)

유전자 변이가 쉽게 일어나는 바이러스 등은 변이 계통에 따라 집단을 형성하게 된다. 이러한 집단들에 대한 분석은 해당 바이러스 집단에 대한 추적, 백신 및 치료약 개발에서 필수적이다. 어떤 집단의 염기 서열의 특성을 효과적으로 표현하는 패턴을 시그너쳐라 하며, 이러한 시그너쳐는 특정 염기서열 집단의 고유한 특성을 나타내면서 다른 집단과 구별되는 정보를 포함하는 것이 바람직하다. 이 논문에서는 가능한 후보 시그너쳐들을 염기분포를 이용하여 생성해가면서, 시그너쳐 해당부위의 염기를 상대 서열집단의 공통 서열의 염기로 변환하여 집단간의 상대거리를 측정함으로써, 후보 시그너쳐에 의한 집단의 고유성질 표현능력과 집단간 차별화 능력을 고려하여 시그너쳐를 추출하는 방법을 제안한다.

Key Words : 바이오인포매틱스, 염기서열 분류, 시그너쳐, 기계학습

1. 서 론

HIV, 감기 바이러스 등과 같은 바이러스에서는 변이가 심하게 나타나며, 지역적 차이도 보일 수 있다. 이러한 변이 때문에 바이러스 백신이나 치료약은 바이러스 변이 형태를 고려하여 개발되어야 효과적일 수 있다. 바이러스의 염기 서열 분석에서는, 특정 바이러스 집단에 대한 기원 예측, 바이러스 종의 분화 정보 등 백신이나 치료약 개발에 유용한 기반 정보를 확보하게 된다. 서열 수준에서의 바이러스 분석의 하나로 바이러스 집단을 다른 집단과 구별하여 특성짓는 특징을 찾는 시그너쳐(signature) 추출이 있다. 이전 연구[1]에서는 하나의 집단을 다른 집단과 차별화하는 특징을 추출하여 이를 집단 분류하는 패턴인 시그너쳐를 찾는데 관심을 가졌었다. 관련 분야 연구자들에게는 해당 집단을 다른 집단으로부터 구별할 수 있는 특징뿐만 아니라, 해당 집단 고유의 특징을 함께 나타낼 수 있는 특징도 확인할 수 있는 시그너쳐가 유용하다. 집단을 차별화하는 특징은 해당 집단을 다른 집단과 구별될 수 있도록 하는 것이고, 집단 고유의 특징은 해당 성질이 집단에서 없어지면 다른 집단과 차별화가 되지 않는 특징이다. 집단을 차별화하는 특징이 집단 고유특성이기는 하지만, 어떤 차별화 특징은 다른 차별화 특징에 비하여 중요하지 않거나, 수반되는 것이 때문에 집단 자체를 특성짓는 데는 큰 의미가 없을 수 있다.

하나의 집단을 다른 집단으로 차별화하는 데 사용될 수 있는 특징은 많을 수 있기 때문에, 이 중의 일부 만으로 집단이 구별될 수 있다. 한편, 집단의 고유한 특성을 나타내는 데 필요한 특징은 집단의 정체성을 손상시키지 않는데 필요한 최소한의 특징들이면 충분하다. 집단의 정체성을 나타내는데 필요한 최소한의 특징을 초과하는 특징은 집단을 기술하는 부수적인 특징이므로 볼 수 있다. 최소한의 특징만을 사용하여 집단을 서로 분류할 수 있는 경우도 있지만, 이 때 사용되는 되는 특징들이 집단의 정체성을 충분히 대표하지 못할 수 있다. 따라서, 집단에 대한 가장 바람직한 시그너쳐는 집단의 정체성을 유지하면서 다른 집단과 차별화시키

이 논문는 2007년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (지방연구중심대학육성사업/충북BIT연구중심대학육성사업단)

는 최소한의 특징 집합이다. 이 논문에서 이러한 성질을 갖는 시그너쳐를 추출하는 방법을 제안한다.

2. 관련 연구

생명과학 연구자들은 주어진 염기서열을 집단별로 나눈 다음, 집단별로 다중 서열정렬을 하여, 정렬된 다중서열을 집단간에 비교하여, 특정 집단에 대한 시그너쳐를 수작업으로 추출하는 방법을 사용해왔다. 집단의 특성을 나타내기 위해 HMM등과 같은 확률모델을 사용하여 집단에 대한 확률적인 모델을 구성하는 경우도 있다.[3] 이러한 확률모델에서는 연구자들이 서열의 위치별 정보를 확인하는 것이 쉽지 않는 경우가 많다.

논문 [1]에서는 시그너쳐 추출 문제를 기계학습 관점에서 분류기(classifier)를 학습하는 것으로 간주하고, 시그너쳐 추출을 위해서 이미 집단이 구별된 서열(학습 데이터)을 사용하여 시그너쳐를 학습하고, 학습에 사용되지 않은 집단이 알려진 서열(검증 데이터)을 사용하여, 시그너쳐의 정확성을 추정하는 학습문제로 보고 시그너쳐를 추출하는 방법을 제안하였다. 이 방법에서는 학습을 위한 특징으로 염기서열의 위치별 염기의 상대적 빈도를 이용하여, 최적 상대빈도 차이값을 갖는 염기를 특징으로 결정하였다. 최종 시그너쳐로는 가장 단순하면서, 민감도(sensitivity)와 특이도(specificity)의 값이 큰 것을 선택하였다. 민감도는 자신의 집단에 속하는 것을 시그너쳐가 자신의 집단으로 분류하는 비율이고, 특이도는 다른 집단에 속하는 것을 시그너쳐가 다른 집단으로 분류하는 비율이다. 이 방법은 하나의 집단을 다른 집단으로부터 차별화하는 가장 간단한 특징으로 시그너쳐를 구성하게 되는데, 집단을 분류하는 시그너쳐로서는 매우 우수한 장점이 있지만, 집단의 정체성을 시그너쳐가 충분히 나타낼 수 없는 단점이 있다. 이 논문에서는 염기서열 분포 정보를 이용하면서, 집단에 대한 차별화 및 정체성 유지 측면에서 최선의 시그너쳐를 추출하는 방법을 제안한다.

3. 집단간 차별화와 집단 정체성을 갖는 염기서열 집단의 시그너쳐 추출

염기서열 집단에 대한 시그너쳐를 결정할 때는, 해당 집단을 다른 집단으로부터 구별할 수 있으면서, 집단 고유의 특성을 나타낼 수 있도록 하는 요소를 추출해야 한다. 이를 위해 제안한 방법에서는 베이스별 염기분포를 이용하여 시그너쳐 후보들을 생성하고, 이들 각 후보 시그너쳐가 해당 집단의 정체성을 얼마나 잘 나타내는지 평가하여, 최적의 시그너쳐를 선정하도록 한다.

3.1 베이스별 염기분포를 이용한 시그너쳐 후보 선정

후보 시그너쳐를 생성하기 위해서는 기존의 제안한 방법[1]에서 이용한 염기분포를 이용하는 기법을 활용한다. 후보 시그너쳐 추출방법을 설명하기 위해 다음과 같은 표기법을 사용한다.

$SG=\{SG_1,SG_2\}$: 서열집단의 모임

$SG_i=\left\{N_1^i,N_2^i,...,N_{E_i}^i\right\}$: 집단 i에 속하는 정렬된 서열의 집합, E_i은 집단 i에 속하는 서열의 개수

$N_i^j=(n_1^{ij},n_2^{ij},...,n_L^{ij})$: SG_i에 속하는 j번째 서열, L은 정렬된 서열의 길이로서 모든 서열이 동일한 길이를 가짐

$m(k,j)$: 집단 k에서 위치 j에서 최대빈도를 갖는 염기

$s(k,j)$: $m(k,j)$의 상대빈도값

$f(k,j,n)$: 집단 k의 위치 j에서 염기 n의 상대빈도수

$S_k^\delta=(s_1^k,s_2^k,...,s_L^k)_\delta$: 집단 k에 대한 식별임계값 δ에서의 후보 시그너쳐

$SG_i=(c_1^i,c_2^i,...,c_L^i)$: 집단 i에 서열에 대한 공통 서열

$CN_i^j=(cn_1^{ij},cn_2^{ij},...,cn_L^{ij})$: SG_i에 속하는 j번째 서열 N_i^j에서 시그너쳐 S_k에 속하는 위치의 베이스를 상대 집단의 공통 서열 CS_j에 해당 위치 염기로 대체한 서열

제안한 방법에서는 두 개의 서열 집단이 있는 것을 가정한다. 또한, 시그너쳐 추출을 위해 사용되는 염기서열들은 ClustalX 등의 도구를 사용하여 이미 다중 서열정렬이 된 상태로서 동일한 길이를 가지고 있다고 전제한다. 바이러스 등과 같이 동일 종에서 변이가 생겨 분화된 염기서열 집단간의 차이는 특정위치의 염기의 변이에 기인하는 경우가 많다. 이러한 생물학적인 관찰에 기반하여, 각 집단에 대해서 위치별 염기의 발생빈도를 계산하여, 최대빈도 염기가 상대 집단에서 얼마나 자주 발생하는지 비교하여, 빈도차이가 어떤 임계값 이상이면, 해당 위치가 시그너쳐를 구성하는데 사용되고, 최대 빈도의 염기가 해당 위치의 특징 염기로 사용된다. 이때 사용되는 임계값을 식별 임계값(discrimination threshold) δ라고 한다. 집단 SG_1에서 위치 i가 시그너쳐를 정의하는데 사용될 수 있으려면 다음과 같은 성질을 만족해야 한다.

$$f(1,i,m(1,i)) \geq f(2,i,m(1,i))+\delta \qquad (1)$$

식 (1)에서 기술한 것은 집단 SG_1의 위치 i

에서 가장 빈발하는 염기 $m(1,i)$의 해당 위치에서의 발생빈도 $f(1,i,m(1,i))$가 집단 SG_2에서의 발생빈도 $f(2,i,m(1,i))$에서 보다 δ이상 클 때 해당위치가 시그너쳐에 포함된다는 것을 나타낸다. 한편 식 (1)의 조건을 만족할 때, 해당위치의 최대빈도 염기 $m(1,i)$가 해당위치의 시그너쳐 문자가 되고, 상대빈도값 $s(1,i)$는 해당 시그너쳐 문자에 대한 가중치가 된다.

집단 k에 대한 식별임계값 δ에서의 시그너쳐 S_k^δ는 다음과 같이 정의된다. 식(2)에서 보는 바와 같이 시그너쳐는 해당 집단을 다른 집단과 차별화시키는 염기의 위치와 해당 위치의 최대빈도 염기, 이에 대응하는 가중치 즉, 상대빈도값으로 표기한다.

$$S_k^\delta = (s_1^k, s_2^k, \ldots, s_L^k)_\delta \tag{2}$$

$$s_i^k = \begin{cases} <m(k,i), s(k,i)> & m(k,i)\text{가 시그너쳐 문자} \\ <'-', 0> & \text{그렇지않을경우} \end{cases}$$

3.2 공통 서열

서열 집단의 전형적인 특징은 공통(consensus) 서열로 나타낼 수 있다. 임의의 집단 SG_i에 대한 공통 서열 $CS_i = (c_1^i, c_2^i, \ldots, c_L^i)$는 SG_i에 속하는 서열들을 다중 정렬한 후 각 베이스별로 가장 빈번하게 출현하는 염기로 서열을 구성한 것이다.

$$CS_i = (c_1^i, c_2^i, \ldots, c_L^i)$$

$$c_k^i = m(k,j) \tag{3}$$

공통 서열은 특정 집단의 전형적인 형태를 보이는 것이기는 하지만, 해당 집단의 정체성을 대표하지는 못한다. 어떤 서열에서 집단의 정체성을 나타내는 것은 해당되는 요소가 다른 것을 변경될 때 집단의 정체성을 잃어버려서 해당 집단에 속한다고 할 수 없게 되는 것이다.

3.3 후보 시그너쳐에 대한 서열 부합도 평가

시그너쳐는 집단을 다른 집단과 구별짓는 특징으로서, 새로운 염기 서열이 주어질 때 이를 해당 집단으로 간주할 것인지 여부를 판정하는데 사용될 수 있다. 이를 위해서는 임의의 염기서열 N이 주어질 때, 시그너쳐 S_k에 대한 부합정도 $G(N,S_k)$를 평가하기 위해 다음과 같은 함수를 사용한다.

$$G(N,S_k) = \frac{\sum_{j=1}^{L} w(k,j) \cdot 1(n_j = m(k,j))}{\sum_{j=1}^{L} w(k,j)} \tag{4}$$

식(4)에서 $1(n_j = m(k,j))$는 $n_j = m(k,j)$ 성질을 만족하면 1이고, 그렇지 않으면 0값을 주는 함수이고, $w(k,j)$는 시그너쳐 S_k의 j번째 위치의 가중치 값을 나타낸다. 평가되는 서열 N은 길이 L이 되도록 집단 1의 서열과 다중 서열정렬되어 있다고 전제한다. $G(N,S_k)$ 함수에 의해서 평가되는 부합정도값은 구간 [0,1]의 값을 가지게 된다.

두 서열 N_i^j, N_i^k간의 부합정도 $Gs(N_i^j, N_i^k)$는 다음과 같이 일치하는 염기 개수의 비로 계산한다.

$$G(N_i^j, N_i^k) = \frac{\sum_{l=1}^{L} 1(n_l^{ij} = n_l^{ij})}{L} \tag{5}$$

3.4 후보 시그너쳐의 집단 정체성 평가

공통 서열은 집단의 전형적인 특성을 보이는 것이기 때문에, 집단의 정체성을 대표하지는 못하더라도, 다른 집단의 서열에 있는 특징을 제거하는데 사용할 수 있다. 즉, 어떤 집단에 속하는 서열에 대해서 특정 부분을 다른 집단의 공통 서열의 대응하는 부분으로 대치하게 되면, 해당 서열이 그 집단에 속하지 않게 될 수 있는, 즉 집단의 정체성을 상실하게 될 수 있다.

집단 SG_1의 식별 임계값 δ에서의 후보 시그너쳐 S_k^δ에 대한 정체성 유지 여부를 측정하기 위해서, SG_1의 각 서열 N_1^j의 S_k^δ에 해당하는 베이스의 염기를 상대 집단 SG_2의 공통 서열 CS_2에서 대응하는 위치의 염기로 대치하여 치환서열 CN_1^j를 만든다.

$$CN_1^j = (cn_1^{1j}, cn_2^{1j}, \ldots, cn_L^{1j})$$

$$cn_k^{1j} = \begin{cases} cn_k^2 & \text{if } s_1^k \neq <'-', 0> \\ n_k^{1j} & otherwise \end{cases} \tag{6}$$

치환서열 CN_1^j가 집단 SG_2보다 SG_1에 더 가깝다고 할 수 없게되면, 서열 N_1^j는 집단 SG_1의 정체성을 상실한다고 간주할 수 있다. 이러한 관점에서 복잡한 형태의 후보 시그너쳐에서 시작하여 단순한 후보 시그너쳐로 점검하면서 집단 SG_1의 각 서열 N_1^j의 상대 집단 SG_2에 대응하는 치환서열 CN_1^j을 구하여, CN_1^j와 SG_2의 각 서열 N_2^j의 평균유사도들의 평균 $BetweenSim(S_1^\delta, SG_2)$과, CN_1^j와 $SC_1 - \{N_1^j\}$에 속하는 각 서열 N_1^k와의 평균유사도의 평균 $WithinSim(S_1^\delta, SG_1)$을 비교하여, 대치서열 CN_1^j이 자신의 집단 SG_1에 더 가깝다고 할 수 없는 상태에서의 시그너쳐를 선택한다. 이것이 집단의 정체성을 유지하면서, 가장 간단한 시그너쳐가 된다.

$$WithinSim(S_1^{\delta}, SG_1) = \frac{\sum_{j=1}^{|SG_1|}\sum_{k=1,k\neq j}^{|SG_1|} Gs(CN_1^j, N_1^k)}{|SG_1| \cdot (|SG_1|-1)} \quad (7)$$

$$BetweenSim(S_1^{\delta}, SG_2) = \frac{\sum_{j=1}^{|SG_1|}\sum_{k=1}^{|SG_2|} Gs(CN_1^j, N_2^k)}{|SG_1| \cdot |SG_2|} \quad (8)$$

상대 집단 SG_2과 차별화면서 집단 SG_1의 정체성을 가장 잘 표현하는 시그너쳐 S_1는 다음과 같이 결정한다.

$$S_1 = \arg\min_{S_1^{\delta}} |S_1^{\delta}| \text{ s.t.}$$

$$WithinSim(S_1^{\delta}, SG_1) \geq BetweenSim(S_1^{\delta}, SG_2) \quad (9)$$

4. 적용 실험

HIV-1 한국형 바이러스 264개 염기서열, HIV-1 외래종 바이러스 71개 서열에 대해서 두 집단을 효과적으로 구별할 수 있는 시그너쳐를 찾기 위해 개발된 방법을 적용하는 실험을 하였다. 실험을 위해서 각 집단의 서열의 반씩 나누어서 반은 시그너쳐를 생성하는데 사용하고, 나머지 반은 시그너쳐에 대한 각 서열의 유사도를 확인함으로써 얼마나 효과적으로 집단을 구별할 수 있는 검증하는데 사용하였다.

(그림 1)은 식별 임계값 δ의 증가에 따른 후보 시그너쳐들에 대한 한국형 바이러스 집단(KBV)와 외래종 바이러스 집단(OBV)의 집단내 또는 집단간의 서열간 평균거리를 나타낸 것이다. 그림은 KBV 집단의 시그너쳐를 결정하기 위한 실험에서의 특성을 보인 것으로, 1번 점선은 KBV 내의 서열간의 평균거리이고, 2번은 KBV 집단 서열과 OBV 집단 서열간의 평균거리를 나타낸다. 3번 점선은 표현된 KBV 집단의 치환서열(KVOV)을 KBV 서열들과 평균 거리를 나타내고, 4번 점선은 KBV 집단의 치환서열(KVOV)들의 KBV 집단 서열들과의 평균 거리를 나타낸다. 가장 바람직한 시그너쳐는 3번 점선과 4번 점선이 만나는 부근에서의 시그너쳐가 된다.

(그림 2-(a))는 KBV에 대해 선정된 시그너쳐(kbS)에 대한 KBV 집단의 검증 서열에 대한 부합정도를 나타낸 것이고, (그림 2-(b))는 kbS에 대한 OBV 집단의 검증 서열에 대한 부합정도를 나타낸 것이다. 보는 바와 같이 서열이 집단을 명확히 구별할 수 있는 것을 확인할 수 있다.

5. 결 론

이 논문에서는 서열집단을 상대집단으로 부터 효과적으로 차별화면서 해당 서열의 정체성을 유지하는 효과적인 시그너쳐를 추출하는 기법을 제안하고, 실제 데이터에 대해 적용하여, 제안된 기법의 유용성을 확인하였다. 향후에는 세 개 이상의 집단이 있는 상황에서 효과적인 시그너쳐를 찾는 방법에 대해서 추가적인 연구를 수행할 예정이다.

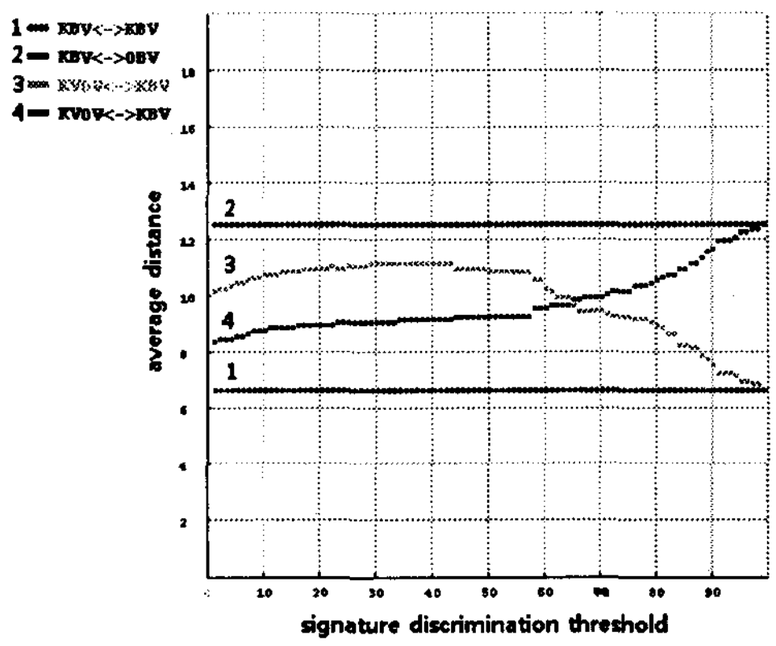

그림 1. 시그너쳐 크기에 따른 집단 정체성 표현 정도를 표현하는 집단내/집단간 평균거리

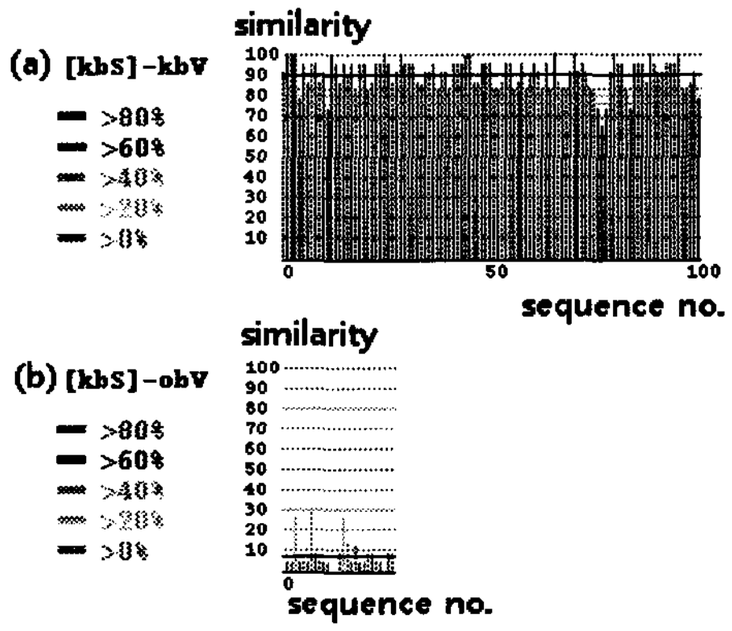

그림 2. 시그너쳐에 대한 집단서열의유사도

참 고 문 헌

[1] 황경순, 이혜리, 이건명, 이찬희, 윤형우, 김성수, "위치기반 상대빈도차 기반의 바이러스 염기서열 시그너쳐 추출 기법," 한국퍼지및지능시스템학회 2007춘계학술대회 논문집, 제17권 제1호, pp.167-170, 2007.04.

[2] M. Kanehisa, Post-Gemome Informatics, Oxford, 199

[3] D. W. Mount, "Bioinformatics: Sequence and Genome Analysis," Cold Spring Habor Lab Press, 2004.

[4] G. B. Forgel, D. W. Corne, "Evolutionary Computation in Bioinformatics," Morgan Kaufmann Publishers, 2003