위치기반 상대빈도차 기반의 바이러스 염기서열 시그너쳐 추출 기법

A Nucleotide Sequence Signature Extraction Method based on Position-Specific Relative Base Frequency Differences

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

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

요 약

동일한 집단에 속하는 개체를 다른 집단에 속하는 개체로부터 구별할 수 있는 염기의 특징을 해당 집단의 시그너쳐라고 한다. 학습 데이터는 두 집단에 속하는 염기서열들이고, 염기서열에 대한 시그너쳐는 개체를 다른 집단과 구별할 수 있는 위치의 염기들로 구성된 서열이다. 제안한 방법에서는 각 집단에 대해서 위치별로 염기의 발생빈도를 계산하고, 가장 발생빈도가 높은 염기를 결정한 다음, 다른 집단의 대응 위치에서 해당 염기의 빈도를 계산하여, 빈도차이가 지정한 분류임계값 이상이면, 해당 위치의 염기를 시그너쳐를 구성하는 특징으로 간주한다. 시그너쳐를 대한 임의의 염기서열에 대한 부합정도는 시그너쳐에 속하는 염기의 학습집단에서의 상대빈도값을 가중치로 하여 계산한다. 임의의 염기서열이 특정 집단에 속하는지 판단하기 위해서는 해당 집단의 시그너쳐에 대한 부합정도를 계산하게 되는데, 부합정도가 얼마이상이 되어야 해당 집단에 속하는 것으로 간주할지 기준이 되는 임계값을 엄밀도 임계값이라고 한다. 엄밀도 임계값은 학습 데이터 집합에 대해서 주어진 시그너쳐에 대한 엄밀도 임계값이 민감도와 특이도를 최대로 하는 것을 선택한다. 제안한 방법을 구현한 바이오인포매틱스 도구를 개발하여, 한국형 HIV-1 바이러스 시그너쳐 추출에 적용하여 분류특성이 우수한 시그너쳐를 추출할 수 있음을 확인하였다.

Key Words : bioinformatics, signature, machine learning

1. 서 론

생명체의 유전정보는 염기서열에 코딩되어 있고, 이들 염기서열이 유전을 통해 후손에게 전달된다. 어떤 이유에 의한 염기서열의 변화는 종의 분화 및 변이를 초래해 왔다. 고등 생물인 경우에는 이러한 종의 염기서열 변화가 상대적으로 낮게 관찰되지만 바이러스 등에서는 변이가 쉽게 관찰되고 있다. 바이러스 등과 같이 염기서열의 변이가 빈번한 대상에 대해서는 특정 바이러스의 기원 예측, 바이러스 종의 분류, 백신 개발 등 여러 가지 목적으로 바이러스 서열에 대한 분석을 하고 있다.[1] 바이러스 등에서 같이 변이가 많은 집단을 분류할 때는 집단을 대표하는 특징을 추출하여, 이들 특징과 비교하여 집단을 분류하게 된다. 동일한 집단에 속하는 개체를 다른 집단에 속하는 개체로부터 구별할 수 있는 염기의 특징을 해당 집단의 시그너쳐(signature)라고 한다.[2]

생명과학자들은 주어진 염기서열을 집단별로 나눈 다음, 집단별로 다중 서열정렬을 하여, 정렬된 다중서열을 집단간에 비교하여, 특정 집단에 대한 시그너쳐를 수작업으로 추출하는 방법을 사용해왔다. 또는 HMM등과 같은 확률모델을 사용하여 특정 집단에 대한 확률적인 모델을 구성하는 경우도 있지만, 확률모델이기 때문에 생물학자들이 서열의 위치별 정보를 확인하기에는 다소 복잡한 형태로 정보를 제공한다. 생명과학자들이 이해하기 쉽고, 활용하기 쉬운 형태로 시그너쳐를 추출하는 것이 바람직하다.

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

시그너쳐 추출 문제는 기계학습 관점에서 보면, 분류기(classifier)를 학습하는 것으로 볼 수 있다. 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 간주할 수 있다. 집단의 특징을 특성짓는 정보는 염기서열의 위치별 염기의 상대적인 빈도가 많은 영향을 미친다. 이 논문에서는 염기서열의 위치별 염기의 상대적 빈도를 기반으로 하여, 최적 상대빈도 차이값을 결정하여, 시그너쳐를 추출하는 방법을 제안한다.

2. 위치기반 상대빈도차 기반의 바이러스 염기서열 시그너쳐 추출

시그너쳐를 분석자가 직접 수작업으로 추출할 때는 집단별로 해당 집단의 서열들을 다중정렬한 다음, 각 위치별로 빈도를 비교하여, 빈도차이를 크게하는 염기들에 대해서 주관적으로 해당위치가 시그너쳐에 반영될지 결정하고, 시그너쳐로 반영될 경우에는 해당 위치에 대한 염기를 결정한다. 분석자의 주관적인 판단을 알고리즘에 반영하기 위해서는 객관적으로 이를 모델링하는 것이 필요하다. 이 절에서는 시그너쳐 분석자의 전문적인 분석지식을 반영하여, 시그너쳐를 추출하기 위해 개발한 기법을 소개한다.

3.1 시그너쳐 추출 방법

제안한 시그너쳐 추출방법을 설명하기 위해 다음과 같은 표기법을 사용한다.

$SG = \{SG_1, SG_2\}$: 서열집단의 모임

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

$N_j^i = (n_1, n_2, ..., n_L)$: 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

(그림 1)은 집단 k에 대한 $f(k,j,n)$, $m(k,j)$, $s(k,j)$에 대한 관계를 나타낸 예이다.

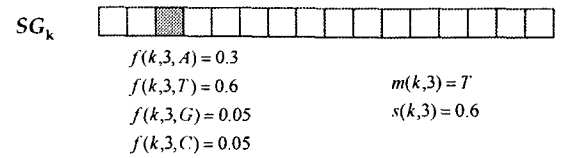

그림 1. $f(k,j,n)$, $m(k,j)$, $s(k,j)$의 관계

제안한 방법에서는 두 개의 서열 집단이 있는 것을 가정하고 있다. 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, 제안한 방법에서는 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 1에서 위치 i가 시그너쳐를 정의하는데 사용될 조건은 다음과 같은 성질을 만족하는 것이다.

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

식 (1)에서 기술한 것은 집단 1의 위치 i에서 가장 빈발하는 염기 $m(1,i)$의 해당 위치에서의 발생빈도 $f(1,i,m(1,i))$가 집단 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 시그너쳐 S_k는 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, 상대 빈도값으로 표기한다.

$$S_k = (s_1^k, s_2^k, ..., s_L^k) \qquad (2)$$

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

3.2 시그너쳐 기반 서열 분류

시그너쳐는 집단을 다른 집단과 구별짓는 특징으로서, 새로운 염기 서열이 주어질 때 이를 해당 집단으로 간주할 것인지 여부를 판정하는

데 사용될 수 있다. 이를 위해서는 임의의 염기서열 N이 주어질 때, 시그너쳐 S_k에 대한 부합정도 $G(N,S_k)$를 평가하는 다음과 같은 함수를 사용한다.

$$G(N,S_k)=\frac{\sum_{j=1}^{L} w(k,j)\cdot 1(n_j=m(k,j))}{\sum_{j=1}^{L} w(k,j)} \quad (3)$$

식(3)에서 $1(n_j=m(k,j))$는 n_j와 $m(k,j)$가 같은 값일 때는 1이고, 그렇지 않을 때는 0값을 주는 함수이고, $w(k,j)$는 시그너쳐 S_k의 j번째 위치의 가중치 값을 나타낸다. 평가되는 서열은 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.

서열 N이 집단 k에 속한다고 판단하기 위해서는 우선 부합정도값 $G(N,S_k)$을 계산한 다음, 이 값이 미리 정해진 임계값이상인지 검사한다. 이 논문에서는 이때 사용되는 임계값 θ를 엄밀도 임계값(stringency threshold)이라고 한다.

$S(N,S_k)>\theta$이면, N는 그룹 k에 속한다.(4)

3.3 최적 시그너쳐 결정

앞 절에서 설명한 것과 같이 시그너쳐를 결정하는 할 때는 식별 임계값 δ과 엄밀도 임계값θ가 결정되어 있어야 한다. 이들 임계값에 따라 추출되는 시그너쳐의 특성이 따라지기 때문에 최적의 임계값들을 선정하는 필수적이다. 최적의 임계값을 선정하기 위해서는 최적의 시그너쳐가 어떤 성질을 만족해야 하는지 결정해야 한다.

제안한 방법에서는 최적의 시그너쳐를 해당 집단을 다른 집단과 구별할 수 있는 가장 간단한 것으로 정의한다. 시그너쳐의 단순도는 다음과 같이 시그너쳐 S_k의 크기 $|S_k|$로 측정한다.

$$|S_k|=|\{s_i=\langle n_i,w_i\rangle | s_i\in S_k \text{ and } n_i\neq '-'\}| \quad (5)$$

시그너쳐에 의한 집단의 구별 특성을 정량화하기 위해 민감도(sensitivity)와 특이도(specificity)를 이용한다. 민감도는 자신의 집단에 속하는 것을 시그너쳐가 해당 집단으로 분류하는 비율로서, 주어진 엄밀도 임계값 θ_i에 대한 시그너쳐 S_k의 민감도 $\sigma_k(\theta_i)$는 다음과 같이 나타낸다.

$$\sigma_k(\theta_i)=\frac{True\ Positive}{True\ Positive+False\ Negative} \quad (6)$$

특이도는 다른 집단에 속하는 것을 시그너쳐가 다른 집단으로 분류하는 비율로서, 주어진 엄밀도 임계값 θ_i에 대한 시그너쳐 S_k의 특이도 $\psi_k(\theta_i)$는 다음과 같이 나타낸다.

$$\psi_k(\theta_i)=\frac{True\ Negative}{True\ Negative+False\ Positive} \quad (7)$$

민감도와 특이성을 모두 1로 하는 시그너쳐가 분류 특성이 가장 좋은 것이다. 이 두가지 측도는 서로 절충(trade-off) 관계에 있기 때문에, 가장 좋은 성질을 가진 것 θ_k^{best}으로 이들 값의 합이 최대인 것을 선택한다.

$$\theta_k^{best}=\arg max_i\{\sigma_k(\theta_i)+\psi_k(\theta_i)\} \quad (8)$$

식(7)의 성질을 만족하는 식별 임계값이 여러 개인 경우에는 ROC(receiver operating characteristic) 곡선을 그려서, 곡선의 아래 면적 $ROC_{area}(\theta_i)$이 가장 큰 것을 선택한다. ROC 곡선은 X축은 $1-\psi_k(\theta_i)$을 나타내고, Y축은 $\sigma_k(\theta_i)$를 나타내는 그래프이다.

최적의 식별 임계값 δ_k^{best}와 엄밀도 임계값 θ_k^{best}를 자동으로 결정하면서, 최적의 시그너쳐를 결정하기 위해 다음과 같은 알고리즘을 사용한다.

procedure find_signature
input : $SG=\{SG_1,SG_2\}$
output : $S_1,\delta_1^{best},\theta_1^{best}$
begin
  각 집단에 대해 위치별 염기의 상대빈도 $f(k,j,n)$를 계산한다.
  각 위치별 최대빈도 염기 $m(k,j)$와 해당 염기의 빈도 $s(k,j)$를 결정한다.
  for $\delta=0.01$ to 1 step 0.01
    집단 1에 대한 식(1)을 사용하여 시그너쳐 S_1^{δ}를 결정한다.
    for $\theta=0.01$ to 1 step 0.01
      식(4)를 사용하여 염기서열의 집단를 판정한다.
      판정결과를 바탕으로 식(6)과 (7)을 사용하여 민감도$\sigma_1(\theta)$와 특이도 $\psi_1(\theta)$를 계산한다.
    endfor
  ROC 곡선 $ROC(\delta)$을 구성하고, 식(8)을

사용하여 δ값에 대한 최적의 $\theta_1^{best}(\delta)$를 결정한다.

endfor

$\Delta_1^{best} = \{\delta \mid \delta = \mathrm{arg}max_\theta\ \theta_1^{best}(\delta)\}$를 구한다.

식(5)를 사용하여 Δ_1^{best}에 속하는 것 중 크기가 가장 작은 것들을 선택한다.

위 단계에서 선택된 δ가 여러 개인 경우에는 해당 δ들에 대한 $ROC_{area}(\delta)$이 최대인 것을 선택한다.

최종의 선택된 δ를 δ_1^{best}로 하고, 이때의 $\theta_1^{best}(\delta)$를 θ_1^{best}를 결과로 반환한다.

end.

3. 구현 및 실험

제안한 시그너쳐 추출방법은 윈도우 환경에서 시각적인 분석이 용이한 형태로 개발하였다. (그림 2)에서 보인 바와 같이 식별 임계값의 변화에 따른 시그너쳐의 크기, 민감도와 특이도 합의 추이, 최적 식별 임계값과 최적 엄밀도 임계값, 최적 시그너쳐가 개발된 도구에서 자동으로 제공된다.

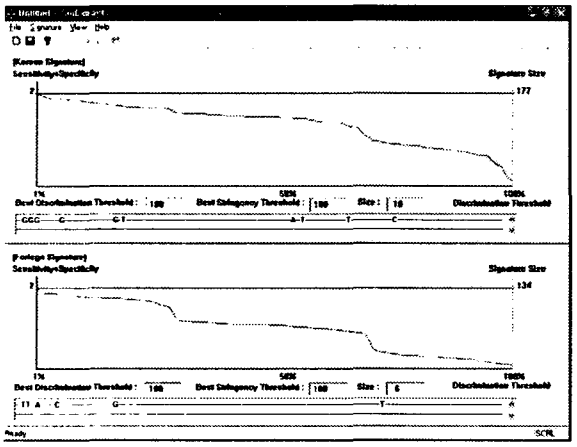

그림 2. 제안한 시그너쳐 추출방법을 구현한 도구

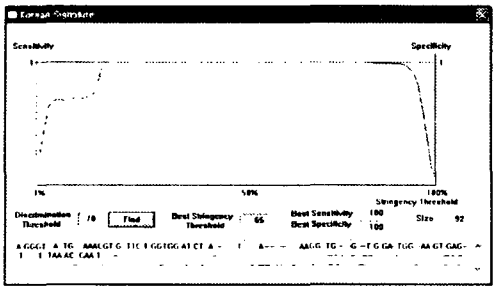

그림 3. 엄밀도 변화에 따른 민감도,특이도 추이 그래프

개발한 도구에서는 (그림 3)과 같은 엄밀도 변화에 따른 민감도, 특이도 추이 그래프, (그림 4)와 같은 ROC 곡선 및 시그너쳐에 대한 부합정도 그래프를 제공함으로써, 도구가 추천하는 시그너쳐 이외에 대한 정보를 분석자가 분석할 수 있도록 하는 시각화 도구를 추가적으로 지원한다.

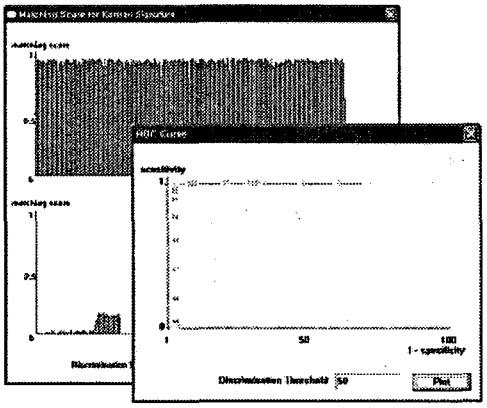

그림 4. ROC 곡선 및 시그너쳐에 대한 부합정도 그래프

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 통해서 구해진 한국형에 대해서 크기 10인 시그너쳐와 외래종에 대해서 크기 6인 시그너쳐는 생물학적으로 의미있는 것으로 분석됨에 따라 제안된 방법이 유용함을 확인할 수 있었다.

4. 결 론

제안한 방법은 두 서열집단을 효과적으로 식별하는 시그너쳐를 추출하는 유용한 기법이다. 향후에는 세 개 이상의 집단이 있는 상황에서 효과적인 시그너쳐를 찾는 방법과, 식별 특성뿐만 아니라 집단의 고유한 성질을 반영하는 시그너쳐를 추출하는 방법에 대해서 추가적인 연구를 수행할 예정이다.

참 고 문 헌

[1] S. Aluru, Handbook of Computational Molecular Biology (eds.), Chapman & Hall/CRC, 2006

[2] M. Kanehisa, Post-Gemome Informatics, Oxford, 1999.