군집화를 통한 메타유전체 종 분류 방법

박치현° 박상현
연세대학교 컴퓨터과학과
tianell@cs.yonsei.ac.kr, sanghyun@cs.yonsei.ac.kr

A Species Classification Method by Clustering in Metagenomes

Chihyun Park°, Sanghyun Park
Department of Computer Science, Yonsei University

요 약

전체 생물 중 가장 많은 비율을 차지하는 미생물을 분리 배양 시켜 연구하는 방법은 과거부터 이루어져 왔다. 하지만 다양한 유전체가 섞여 있기 때문에 실험실 환경에서 우리가 원하는 모든 유전체를 분리 배양하기란 어렵다. 최근 들어 다양한 미생물 유전체 전체를 하나의 집합으로 보고 그 자체를 분석하는 연구가 수행되고 있다. 미생물의 서식환경을 그대로 분석하는 연구는 최근 연구는 메타지놈이라고 불리고 있다. 본 논문에서는 메타지놈 환경에서 추출한 짧은 길이의 염기서열 조각을 통해 메타지놈 환경의 유전체의 종을 분석할 수 있는 방법을 제시한다. 반복적인 군집화 방법과 염기서열에서 특징을 추출하는 방법을 통해 가장 근사한 종을 수를 통계적으로 밝히며 최종적으로 하나의 메타지놈 환경에 존재하는 미생물 종의 수를 밝히는 것을 본 논문의 목표로 한다. 본 논문에서는 이 방법에 대한 전반적인 아이디어를 제시한다.

1. 서 론

미생물은 전체 생물 종에서 가장 많은 비율을 차지하고 있으며, 아직 전체 미생물들 중에서 대부분이 기능적으로, 유전적으로 밝혀진 바가 없다. 현재 실험실 환경에서 밝혀낸 미생물은 전체 미생물 종에서 약 1% 밖에 안 되는 것으로 추정되고 있다. 현대 생물학이 배양 기술과 실험 환경이 계속적으로 발전하고 있다고 하지만, 미생물의 종이 방대하기 때문에 이를 분리 배양하기란 쉽지 않다. 따라서 아직까지 미생물의 분리 배양이 쉽게 되지 않기 때문에 미생물에 대한 유전적 연구가 활발히 이루어지지 않고 있다. 미생물이 존재하는 환경은 예를 들어 하수구, 동물의 소화기관내, 지구, 심해저 해수 등으로 사실상 대부분의 자연 환경이라고 간주해도 무방하다. 예를 들어 160/ml 토양의 경우 약 6,400~38,000/g 정도의 미생물이 존재한다고 예상한다 [1]. 이는 지금까지 유전체 연구의 환경과는 완전히 다른 분포이며, 기존 대부분의 유전체 분석 방법이 통하지 않는 이유이기도 하다.

이러한 대규모 미생물들이 함께 섞여서 존재하는 유전체의 집합을 메타지놈[2] 이라고 정의 한다. 이러한 메타지놈을 연구하는 방법은 크게 두 가지로 나눌 수 있다. 실험실 환경에서 BAC-클로닝(BAC-cloning) 시스템을 통하여 미생물을 배양한 후 유전자 분석을 통해 미생물 라이브러리를 구축하는 것이 첫 번째 방법이다 [3]. 두 번째 방법은 최근 시도되고 있는 방법인데, 클로닝 시스템을 이용하지 않고 바로 염기서열을 샷건 방식으로 밝힌 후 라이브러리를 구축하는 것이다. 이른바 WGSS(Whole Genome Shotgun Sequencing)이라고 불리는 방법으로 Celera사의 Craig Venter 박사에 의해서 고안되었다. 최근 염기서열을 결정하는 이른바 시퀀싱 기술의 급속한 발전으로 실험실 환경에서의 유전체 분석을 양과 시간적으로 압도할 수 있는 기술들이 나오고 있기 때문에 최근 메타지놈 연구 연구에 도입되고 있다.

본 논문에서는 메타지놈 연구에 있어 WGSS 방식에 적용 가능한 종에 따른 유전체 분류법을 제안한다. 메타지놈에 대한 연구가 아직 초기 단계이고, 실험실 차원에서의 배양도 한계가 있기 때문에 메타지놈에 존재하는 종의 수를 확인하는 연구 또한 초기 단계이다.

현재까지 이루어진 대부분의 연구는 염기서열의 분석을 통해 메타지놈에 존재하는 미생물을 분석하기 위해서 WGS을 통해 얻어진 짧은 길이의 염기서열 조각의 특징을 추출한 후 그 특징에 따라서 메타지놈 환경에 존재하는 유전체를 추정해가는 리버스 엔지니어링 방법을 주로 사용하였다. 그렇지만 짧은 길이의 유전체 조각에서 종별 특징을 추출하는 것을 오차가 생길 수 도 있기 때문에 대부분의 연구에서는 짧은 길이의 유전체 조각을 가능한 긴 조각으로 조립한 후 그 조각들의 특징을 통해 밝혀내는 방법이 주를 이루어왔다. [4]에서는 SOM(Self Organizing Map)과 신경망(Neural Network)을 통해 2~4Bp 길이의 뉴클레오타이드에서 나타나는 유전체의 특징을

연구했다. [5]은 GC염기 비율과 테트라뉴클레오타이드(Tetra-nucleotide)의 빈도수를 통해 유전체 조각들 사이의 상관관계를 밝혔다. 하지만 메타지놈 환경에서 짧은 길이의 유전체 조각을 긴 길이로 조립하는 방법은 매우 어렵고 오차가 많다. 짧은 길이의 유전체 조각을 긴 조각으로 조립하기 위해서는 보통 전체 지놈 크기보다 몇 배수(Coverage)의 크리고 짧은 조각들을 생성해 내고 서로 오버랩 되는 짧은 조각들을 큰 조각으로 조립하는 방식을 택하는데, 메타지놈 환경에서는 이런 배수의 짧은 유전체조각을 생성할 때 다양한 종에서부터 유전체 조각을 얻지 못하고 한 종에 편중되어서 얻어 질 수 도 있고, 단일 유전체에서 배수 조각을 얻었다고 하여도 너무 많은 반복이 일정 부분에 존재한다면 제대로 조립되지 않을 수 있다.

따라서 메타지놈 분석에서는 다음과 문제점이 존재한다고 볼 수 있다. 짧은 길이의 조각으로 유전체 분석을 한다면, 각 종별 미생물의 특징을 제대로 추출할 수 없을 가능성이 크기 때문에 오차율이 높아 질 수 있다. 반대로 짧은 길이의 염기서열 조각을 긴 조각으로 조립한 후 유전체 분석을 수행한다면, 제대로 염기서열 조각이 조립되지 않았다면 오차율이 높아지며, 제대로 조립되었다면 오차율이 떨어지게 된다. 따라서 가장 좋은 해결책은 적당한 길이로 짧은 염기서열 조각들을 조립한 후 그 조각들로부터 유전체 특징을 추출한 후 종을 밝히는 것이다. 하지만 이 어느 정도의 cut-off가 적당한지 찾는 연구는 쉽지 않다. 또한 짧은 길이의 염기서열 조각을 긴 조각으로 조립하는 방법은 사실상 메타지놈 환경에 대한 아무런 정보가 없는 상태이기 때문에 상당히 어렵다고 볼 수 있다. 본 논문에서는 가능한 한 짧은 길이의 염기서열 조각을 사용하기 위해서, 짧은 길이의 유전체 조각들로부터 각 종별 특징을 추출할 수 있는 효과적인 방법을 제시하고, 이 특징을 통해서 염기서열 조각들의 군집화를 통해서 종의 수를 밝히는 방법을 제시한다. 기본적으로 본 논문에서는 K-means 군집화 방법 사용하는데 K를 변경함으로써 메타유전체에 존재하는 미생물의 수를 최적으로 결정하는 알고리즘을 제시한다. 기존의 논문들에서 짧은 길이의 유전체조각들과 연관된 단백질 정보를 이용해서 메타지놈을 군집화하는 시도는 있었고 [7], 본 논문에서 제안하는 방법과 비슷하게 일정 길이의 뉴클레오타이드에 대한 빈도수를 얻은 후 분류자를 만든 연구가 있었지만, 이는 이미 종의 수를 알고 있어야 가능하다고 볼 수 있다. [8] 본 논문에서 제안하는 바와 같이 종의 수를 가정하지 않고 군집화하는 연구는 없었다.

2. 본 론

2.1에서는 메타유전체의 짧은 염기서열 조각으로부터 특징을 추출하고 사용하는 방법에 대한 서술을 할 것이고, 2.2에서는 군집화 방법을 통해 메타지놈에 존재하는 미생물의 수를 밝힐 수 있는 알고리즘을 제안한다.

2.1 유전체의 염기서열 조각으로부터 특징 추출 방법

본 논문에서 사용하는 짧은 길이의 염기서열 조각은 대략 700Bp 길이이다. 하나의 Bp는 A,C,G,T 중에 하나가 되기 때문에 700Bp길이의 염기서열 이론적으로 조각은 4700(21400) 개의 서로 다른 조각으로 나타날 수 있다. 메타지놈 연구에서는 대략 700만개 정도의 짧은 길이의 염기서열 조각을 추출하는데, 이론적으로 700만개의 랜덤한 리드는 서로 다른 시퀀스를 갖는다고 볼 수 있다. 이론적으로 이렇게 서로 다른 염기서열의 시퀀스들에서 종에 따른 특징을 추출하는 방법은 대부분 일정 길이의 뉴클레오타이드의 빈도를 통해 얻을 수 있다. 유전체에 존재하는 뉴클레오타이드의 빈도를 통한 종 분류 연구는 일반 유전체 연구에서는 많이 시도되고 있지 않은 연구이다. [6]은 생물학 전체의 종 분류연구에 이 방법을 사용하여, 기존 종 분류와는 다른 결과를 도출하고 있다. 이 연구에서는 영문책 분류에서 사용한 자연어비교 기법을 사용하여 생물의 종 분류를 하였다. 결과적으로 겉으로 드러나는 표현형만으로 종 분류를 했던 전통적인 분류법과 조금 다른 결과가 도출되었다. 하지만, 이는 신선한 시도로, 어떤 종의 유전체의 특징을 추출하는 연구에 특정 유전자나 부분을 지표로 하지 않고 전체 염기서열에 존재하는 뉴클레오타이드의 빈도수가 중요한 요소로 작용할 수 있다는 것을 보여주었다. 이러한 연구가 메타지놈 연구에 있어서도 적용될 가능성이 충분히 있으며 최근 연구들은 이러한 방법으로 연구를 진행하고 있다.

본 논문에서도 이러한 방법을 통해 평균 700bp 길이의 염기서열에서 종의 특징을 추출한다. [9]에서는 2~6 뉴클레오타이드의 빈도를 사용했을 때 비교를 수행했었는데 본 논문에서는 4-윈도우 크기의 뉴클레오타이드의 빈도수를 통해서 종의 특징을 추출한다. 다음은 식 (1)은 4-윈도우 크기의 빈도벡터를 나타낸다. 빈도는 슬라이딩 윈도우 방식을 통해서 계산된다.

$$Freq_{VEC} = (f_1, f_2, f_3, ..., f_{256})$$
$$, f = Number of\ frequency \quad .... (1)$$

2.2 메타지놈 유전체의 군집화를 통한 종 분류 모델

본 논문에서 제안하는 방법의 핵심은 군집화를 통한 종 분류 모델에 있다. 이 방식은 기존 연구에서는 다루어지지 않았던 새로운 접근법으로 기본적으로 K-means 군집화 방법을 사용한다. 본 연구의 주제인 메타지놈 연구와 K-means 군집화 방식을 다음과 같은 측면에서 본다면 공통된 점이 있다. 본 논문의 목표는 메타지놈에 존재하는 종의 수를 밝히는 것인데, 이는 K-means 군집화 방법의 K로 치환될 수 있다. K-means 방법은 가장 대표적인 군집화 방식이고, 빠르고, 간단한 알고리즘 때문에 데이터마이닝 분야 전반에 걸쳐서 많이 사용되고 있는 방법이다. 하지만 초기 클러스터의 센터를 결정하는 방식에 따라 결과가 달라지는 비결정적인 방식(Non-deterministic)이라는 점과 근본적으로 K를 결정하기가 쉽지 않다는 것이 대표적인 단점이다. 분류하고자 하는 데이터셋이 몇 개의 분류가 되야 하는지 미리 알 수 있다면 K는 쉽게 결정되지만 이를 모른 상태에서 발견적으로(heuristic) K를 찾아야 한다면 K-mean 방식은 좋은 결과를 도출할 수 없을 것이다.

본 논문에서도 메타지놈 환경에 K-means를 도입하고 K를 메타지놈 환경에 존재하는 미생물의 종의 수라고 한다면, 우리는 쉽게 K를 결정할 수 없다. 즉 메타지놈에 존재하는 문제가 K-means 군집화 방식이 갖고 있는 문제점으로 변환이 될 수 있다. K를 효과적으로 결정하기 위해서 기존 연구들은 진화연산(Genetic Algorithm)과 같은 지역탐색(local search) 방식을 활용하기도 한다. 이러한 방식을 사용한다면 전체적인 최적해(global optimum)를 구할 수는 없더라도 이에 근접한 해를 허용 가능한 시간복잡성 안에서 찾아낼 수 있다. 본 논문에서 제안하는 방법은 메타지놈 연구에 있어서 완벽한 종의 수를 제시하는 것이 아니라, 메타지놈 조립알고리즘에 대한 개발과 생물학 연구자들에 있어서 유의한 미생물 종의 추정치를 제공하는 데 있다. 따라서 이를 전산학적 알고리즘을 통해서 제공하고자 한다.

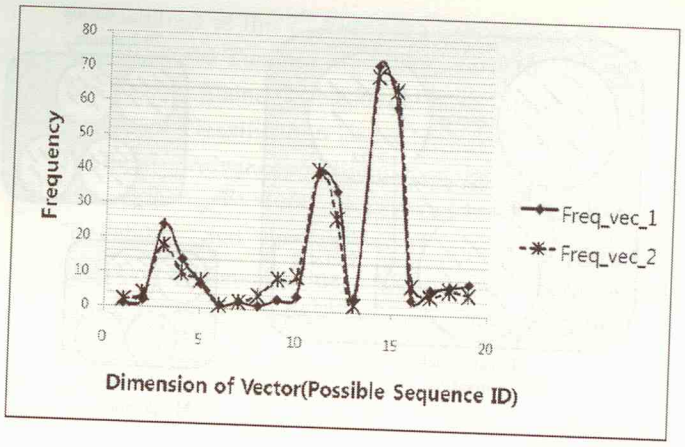

그림 1. 두 염기서열조각의 빈도수 벡터의 그래프 표현

본 논문에서 제안하는 K-means 군집화 방식을 제시하기 전 군집화 방식에 사용되는 거리측정치를 제시한다. 보통 K-means 군집화 방식은 수치데이터셋을 사용하며 유클리디안 거리를 군집화의 측정기준으로 사용한다. 하지만 본 논문에서 사용하는 K-mean 군집화의 데이터셋은 벡터셋이다. 2.1에서 언급한 FreqVEC 가 한 오브젝트가 되어서 군집화가 이루어 진다. 따라서 벡터사이의 거리를 측정하기 위해서 본 논문에서는 MSR(Mean Squared Residue)을 사용한다. MSR은 어떤 matrix의 sub-matirx를 결정하는 문제에 있어서, 해당 sub-matrix를 구성하는 벡터들이 얼마나 같은 패턴을 가지면서 적합도를 나타내느냐를 측정하는 측정치이다. 본 논문에서 사용하는 서로 다른 벡터들도 여러 개가 모이면 사실상 전체 벡터셋의 sub-matrix형태가 되기 때문에 MSR을 통해서 벡터들 사이의 거리를 측정할 수 있다. MSR은 다음과 같다. 서로 다른 두 벡터를 다음과 같은 matrix 형태로 변환한다면 MSR은 (2)와 같이 계산된다.

$$\begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,256} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,256} \end{bmatrix} = \begin{bmatrix} Freq_{VEC\_1} \\ Freq_{VEC\_2} \end{bmatrix}$$

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \quad ... (2)$$

여기서, I는 matrix의 행을 J는 열을 나타낸다. MSR은 0에서 1사이의 값으로 나타내지며, 0에 가까울수록 두 벡터의 pattern이 적합한 것을 나타낸다. 그림 1은 두 벡터를 그래프 형태로 나타낸 그림이다. 두 벡터의 차이는 (2)를 통해 계산된다.

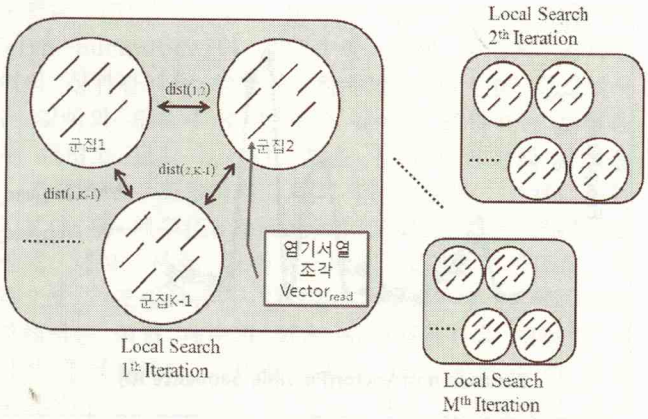

그림 2. 각 반복자 별 군집화 후 군집들 사이의 거리측정을 통한 스코어 구하기

제안하는 모델에서 또 한가지 중요한 부분은 K를 결정하기 위한 방법이다. 본 논문에서는 K를 결정하기 위해 지역 탐색 방식을 적용하며, K=k 일 때의 군집화 모델들의 스코어를 계산하여, 이를 k개의 군집들로 대표되는 하나의 후보 모델을 대표 값으로 사용한다. 그런 후 이 스코어 값들을 통해서 가장 최적의 K를 찾아낸다. 하나의 k를 대표하는 스코어는 다음과 같은 원리로 계산된다. k개의 군집들로 이루어진 결과셋이 k개의 종을 분류해낸 것이라면, 하나의 군집 안에 포함된 짧은 길이의 원본 염기서열 시퀀스들은 하나의 종에서 나타나는 특징들을 가지고 있어야 한다. 이런 특징을 수치화해서 스코어 계산에 사용한다. 스코어는 다음과 같이 계산된다.

$$S_{iteration} = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} Dist\,(FreqVec\_all)_{ij} \quad \ldots (3)$$

여기서 FreqVec_all은 하나의 군집에 존재하는 모든 짧은 길이의 염기서열 조각들 중 서로 80% 이상 중복되는 조각들은 제외한 조각들을 한 군집을 대표하는 조각들로 보고, 그 조각들에서 모두 4-윈도우를 통해 빈도수를 계산한 벡터가 된다. 한 군집을 하나의 미생물로 가정하기 때문에 서로 80% 이상 중복되는 조각들은 사실상 하나의 미생물로부터 추출되었으리라 판단할 수 있다. 따라서 이런 요소를 제거한 후 한 미생물의 전체 뉴클레오타이드 빈도수를 계산하면 그 벡터가 한 생물을 대표하는 특징이 된다. 이런 벡터들을 각 군집 즉 k개에 대해서 모두 구한 후 그 벡터들의 모든 가능한 쌍에 대해서 벡터 사이의 거리를 MSR을 사용하여 구한다. 그림 2는 식 (3)을 통해 구하려는 군집들 사이의 관계를 그림으로 표현한 것이다. 군집화 알고리즘은 지역탐색 방법을 통해서 스코어 값의 최적해를 찾게 되고 더 이상 최적해를 구할 수 없다면 반복을 중단한다. 전체 알고리즘의 시간복잡도는 원래 K-means 군집화 방법의 복잡도인 O(NKt)에서 지역 탐색 방식에 대한 복잡도를 고려한 O(NKtL)이 된다. 여기서 N은 군집화할 데이터의 수, K는 K-means 군집화의 K, t는 군집화 알고리즘이 수행될 때의 반복횟수이다. 여기서 N을 제외한 나머지는 상수로 봐도 무방할 만큼 그 값이 N에 비하여 작다. 따라서 실제 복잡도는 데이터의 수인 N와 로컬 탐색 횟수인 L만 고려하면 된다. L은 보통 N보다는 작으며 따라서 시간 복잡도는 최대 O(N2)가 된다. 전체 알고리즘은 아래 알고리즘 1과 같다.

Input : Short sequence read set, $R = \{r_1, r_2, \ldots, r_N\}$
Output : Inferred number of species

```
1:  Compute nucleotide frequency from R
2:  Make Frequency vector V for each element of R
3:  New dataset is obtained (R, V)
4:  while(Until get a local optimal solution)
5:    Choose initial K center
6:    Compute distance between V & each centers
7:    Assign each vector to closet centers
8:    Recompute centets using current cluster member
9:    if(Center ≠ newCenter) then
10:     goto step 6 with newCenter
11:   else
12:     Complete Clustering
13:   for(i=1; i≤K; i++)
14:     Remove more than 80% repeating reads
15:     Compute frequency from all remaining reads
16:     Make Frequency vector Cv
17:   end for
18:   Compute S_iteration from all possible pair-wise Cv
19:   Score ← S_iteration
20:   if(Score = local optimal) then
        break;
21:   else
22:     change the K
23: return optimal K
```

알고리즘 1. 전체 군집화 및 종 분류 알고리즘

3. 결 론

본 논문에서는 메타지놈에 존재하는 종의 수를 밝히기 위해서 반복적인 군집화 알고리즘을 제시하였다. 군집화 알고리즘을 수행하기 위해서 종 사이의 차이점을 구하기 위한 테트라 뉴클레오타이드의 빈도수를 특징으로 결정하였다. 본 논문을 제안하는

가장 큰 동기는 메타지놈 내에 존재하는 미지의 값인 종의 수를 풀기 위해서는 전산학에서 자주 사용되는 발견적이고 부분적인 최적해를 도출하는 알고리즘이 적당한 해결책이 되리라고 생각했기 때문이다. 아직 본 논문에서는 구체적인 실험결과와 방법이 제시되고 있지 않다. 향후 구체적인 실험이 진행될 예정이며, 예상하는 실험데이터셋은 실제메타유전체 프로젝트에서 도출된 염기서열을 사용할 예정이며, 광산수(acid mine) 환경에서 추출한 염기서열 시퀀스들을 사용할 예정이다. 이 실험데이터는 약 166만개의 짧은 길이의 염기서열 시퀀스가 있으며 크기는 92MB정도이다. 실제데이터를 통한 실험이 완료되면 도출되는 종의 수는 메타지놈 어셈블러를 개발하기 위해 꼭 필요한 중요한 팩터로 사용될 것이다.

참고문헌

[1] Curtis, T. P., Sloan, W. T. & Scannell, J. W. (2002). From the Cover: Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci U S A 99, 10494-10499.

[2] Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5, R245-249.

[3] Lorenz, P., Liebeton, K., Niehaus, F. & Eck, J. (2002). Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space. Curr Opin Biotechnol 13, 572-577.

[4] Takashi Abe, Shigehiko Kanaya, Makoto Kinouchi, Yuta Ichiba, Tokio Kozuki, Toshimichi Ikemura, "Informatics for Unveiling Hidden Genome Signatures", Genome Research, 13, 693-702, 2003.

[5] Hanno Teeling, Anke Meyerdierks, Margarete Bauer, Rudolf Amann, Frank Oliver Glöckner, "Application of tetranucleotide frequencies for the assignment of genomic fragments", Environmental Microbiology, 6, 938-947, 2004

[6] Gregory E. Sims, Se-Ran Jun, Guohong A. Wu, Sung-Hou Kim, "Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions", PNAS, 106, 8, 2677-2682, 2009.

[7] Gianluigi Folino, Fabio Gori, et al., "Clustering Metagenome Short Reads using Weighted Proteins", Proceedings of the 7th European Conference on, EvoBIO 2009 Tubigen, Germany, April 15-17, 2009.

[8] Gail Rosen, Elaine Garbarine, et al., "Metagenome Fragment Classification Using N-Mer Frequency Profiles", Advances in Bioinformatics, 2008.

[9] Jon Bohlin, Eystein Skjerve, David W. Ussery, "Reliability and application of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes", BMC Genomics, 9, 104, 2008.