Preprocessing Model for Operon Prediction Using Relative Distance of Genes and COG Distance

COG 거리와 유전자 간의 상대 위치정보를 이용한 오페론 예측 전처리 모델

Bong-Kyung Chun[1*], Chul-Jin Jang[1*], Eun-Mi Kang[2], Hwan-Gue Cho[1]

[1] Department of Computer Engineering, Pusan National University, Pusan, Korea
[2] Bioinformatics Research Team, Computer & Software Research Laboratory, ETRI

[*] E-mail: {bkchun, cjjang}@pearl.cs.pusan.ac.kr

요약

오페론(operon)은 보통 미생물에서 다수의 인접한 유전자들로 구성된 그룹으로 하나의 유전자처럼 공통된 프로모터에 의해 전사되는 단위이다. 오페론을 구성하는 유전자들은 기능적으로 서로 유사하거나 같은 물질대사경로(metabolic pathway) 상에 존재하는 특징을 지니기 때문에 이들은 중요한 의미를 가지며, 미생물 유전체 분석에서 오페론을 구성하는 유전자들을 예측하는 것은 상당히 중요하다. 오페론을 예측하는 이전 연구들로는 이미 알려진 오페론의 특징인 유전자간 거리나 오페론을 구성하는 평균 유전자 개수 등을 이용하는 방법, 마이크로어레이 발현 실험을 이용한 방법, 전유전체(whole genome)들 간의 보존된 유전자 집합(conserved gene cluster)을 이용한 방법 그리고 물질대사경로를 이용한 방법 등이 있다. 본 논문에서는 COG 기능(function) 거리, 유전자 간의 거리, 코돈 사용빈도(codon usage) 그리고 COG 기능 거리와 유전자간 거리를 같이 적용한 방법을 이용하여 오페론 예측을 위한 전처리 모델을 생성하였다. 전처리 모델을 *E. coli* 전유전체에 적용해본 결과, 알려진 오페론들의 약 90%가 이를 포함하였다. 따라서 본 논문에서 제시한 전처리 모델은, 추후 오페론 예측을 위한 좋은 도구로 활용할 수 있을 것이다.

서론

생명체는 생명 활동을 영위하기 위해 유전자의 정보에 따라 단백질을 만든다. 만들어진 단백질은 필요한 영양소를 세포 속에 공급하고, 그 영양을 바탕으로 생명체의 몸을 구성하며, 생명현상을 영위하기 위해 필요한 효소로써 작용한다. 유전자의 정보에 따라서 단백질이 만들어 질 때에는, 우선 DNA에 쓰여진 유전자의 유전정보가 mRNA에 복사된다. 이 과정을 전사(transcription)

과정이라 하며, 보통 미생물에서는 다수의 인접한 유전자들이 하나의 유전자처럼 공통된 프로모터에 의해 전사된다. 이렇게 함께 전사되는 유전자 집합을 오페론이라 한다. 오페론을 구성하는 유전자들은 비슷한 기능을 하거나 같은 물질대사경로 상에 존재하기 때문에 중요한 의미를 가지며, 미생물 유전체 분석에서 오페론을 구성하는 유전자들을 예측하는 것은 상당히 중요하다.

오페론을 실험을 통해서 예측하거나 확인하는 것은 매우 복잡한 과정이며, 전과정을 실험실에서 구현하기가 어렵다. 그리고 실험적 방법을 통한 오페론 예측은 상당한 시간과 비용이 든다. 따라서 최근에는 많은 전유전체의 서열 데이터, 알려진 오페론 데이터 및 기타 생물학적 데이터 등을 이용하여 오페론을 예측하는 방법들이 많이 연구되고 있다.

이들 방법은 오페론을 예측할 때 사용되는 특징에 따라 다음과 같이 분류할 수 있다.

오페론의 여러 특징을 이용한 방법

a) 전사신호 특징 (Transcription signals feature)

오페론에 속하는 유전자들은 함께 전사되기 때문에 같은 전사신호(transcription signals)에 의해 조절된다. 프로모터(promoter)와 터미네이터(terminator)는 오페론 전사과정을 조절하는 전사신호이며, 유전체 서열에서 프로모터와 터미네이터를 예측함으로 오페론의 시작과 끝을 예측할 수 있다. 기존의 알려진 오페론의 프로모터와 터미네이터 서열을 이용하여 마코브 모델(markov model)을 생성한 뒤, 이를 유전체 서열에서 프로모터와 터미네이터를 예측하여 오페론을 찾는 방법이다[12].

b) 기능클래스 특징 (Functional class feature)

오페론은 구성하는 유전자들은 비슷한 기능을 수행한다. 따라서 기능적으로 유사하고, 서열에서 인접한 유전자 그룹이 같은 오페론을 형성할 가능성이 있다. 유전자들의 기능주석(functional annotations)을 이용하여, 비슷한 기능을 하는 유전자 그룹을 조사하여 오페론을 예측할 수 있다. 그러나 모든 유전자들에 대해 기능주석이 되어 있지 않기 때문에, 예측률이 낮으며 다른 유전자예측 방법들과 같이 사용되고 있다[1].

c) 크기 및 공간 특징 (Length and spacing feature)

미생물 유전체에 있어 오페론에 속하는 유전자들은 함께 전사되기 때문에 유전자간에 전사되지 않는 부분이 있으면 비효율적이다. 따라서 오페론에 속하는 유전자들은 최대한 가깝게 위치하려는 특성을 가진다. *E. coli*를 포함하여 여러 미생물의 데이터를 가지고 조사한 결과, 오페론의 약 80%가 5개의 이하의 유전자들로 구성되며, 오페론에 속하는 유전자 간의 거리는 -1 bp ~ -4 bp 사이인 것이 많이 나타났다[1].

이러한 특징을 이용하여 오페론을 예측하는 방법이 발표되어 있으며, 특히 유전자 간의 거리를 이용하는 경우 높은 예측률을 보였다.

d) 유전자 발현 특징 (Gene expression feature)

마이크로어레이(microarray) 칩은 수천개 이상의 유전자 발현 변이를 한번의 실험으로 확인 할 수 있고, 실험 기술의 발전으로 인해 최근 마이크로어레이 실험을 이용한 유

전자 발현 실험이 증가하고 있다. 같은 오페론에 속하는 유전자들은 함께 발현되기 때문에 마이크로어레이 실험을 통하여, 같이 발현되는 유전자들의 상관 계수 값들을 구해봄으로써 유전자 간의 관계를 파악할 수 있다. 상관관계가 높은 유전자들을 후보 오페론으로 예측할 수 있으며, 이것이 유전자 발현 특징을 이용한 오페론 예측 방법이다[3, 11].

보존 유전자 클러스터(Conserved gene cluster)를 이용한 방법

비슷한 종에서 같은 기능을 하는 오페론들은 세대를 거듭하더라도 보존되는 특성이 있다. 이 특징을 이용하여 다수의 전유전체에서 유전자들과 유전자들 순서가 보존된 유전자 그룹이 같은 것을 찾아 이를 오페론으로 예측하는 방법이 있다. 이 방법은 다수의 전유전체들이 서열화되어야 하고, 다수의 전유전체에서 보존된 오페론들만을 찾을 수 있다는 제약이 있다[9].

물질대사경로(Metabolic pathway)를 이용한 방법

오페론은 기능적으로 관련 있는 유전자들로 구성된다는 특징이 있으며, 따라서 이 유전자들은 같은 물질대사경로에 나타날 가능성이 높다. 기존의 물질대사경로에 대한 지식을 이용하여 같은 물질대사경로에 있는 유전자 그룹들을 오페론이라 예측할 수 있다. 이 방법은 물질대사경로의 서브그래프(subgraph)를 찾는 문제로 볼 수 있으며 NP-Complete 문제이다. 또한 물질대사경로가 잘 밝혀져 있는 경우에만 효과적인 예측을 기대할 수 있다[2].

앞에서 설명한 방법인 전사신호, 기능 클래스, 크기 및 공간 특징, 유전자 발현, 그리고 물질대사경로를 이용하여 오페론을 예측하는 방법은 실험 데이터나 알려진 오페론에 대한 지식이 필요하다는 제약이 있으며, 보존 유전자 클러스터를 이용하는 방법은 다수의 전유전체에서 보존된 오페론들만 찾을 수 있다는 제약이 있다.

본 논문에서는 오페론 예측 시 도움을 줄 수 있는 기본이 되는 두 개의 유전자 쌍인 전처리 모델을 생성하는 방법을 제안한다. 이전 연구에서와 달리 본 논문에서는 오페론 예측을 위한 전처리 모델을 생성할 때 이미 알려진 오페론 실험 데이터나 학습 데이터를 사용하지 않는다.

본 논문의 구성은 다음과 같다. 우선 실험적으로 밝혀진 오페론 데이터들의 특징을 분석하고, 본 논문에서 제시하는 전처리 모델을 생성하는 방법과 실험 및 결론 그리고 향후 연구과제를 알아 보도록 하겠다.

*E. coli*의 오페론 데이터 분석

본 장에서는 오페론들의 특징을 알아보기 위해 실제 오페론 데이터를 이용하여, 오페론을 구성하는 유전자들의 COG 기능과 인접한 유전자간 거리 분포에 관한 실험을 하였다. 본 실험에서는 RegulonDB[7]에서 얻은 *E. coli* 오페론 데이터를 사용하여 오페론 에 속하는 유전자들의 COG 기능 값과 오페론에 속하는 유전자 간의 거리를 알아보았다. 사용한 오페론 데이터는 RegulonDB version 3.1 에서 정의한 tu_list(transcription unit list)의 487개의 데이터 중 2개 이상의 유전자로 구성된 293개이다.

COG 기능 분석

일반적으로 같은 오페론에 속하는 유전자들은 공통된 기능을 수행하기 때문에, 기능이 유사한 유전자들이 같은 오페론을 형성할 가능성이 높다. RegulonDB에서 얻은 *E. coli*의 오페론 데이터들의 각 유전자 서열을 블라스트(BLAST)[5] 에 입력하여, NCBI[6]의 COG 데이터베이스[4]와 가장 유사한 e-value값을 갖는 COG 기능 값을 계산하였다. 표 1은 COG 기능 점수(오페론 내에서 가장 빈도수가 높은 COG값을 가지는 유전자수 / 오페론 내 유전자수 × 100)에 따른 오페론의 비율을 보여 준다.

COG 기능 점수	전체 오페론 중 차지하는 비율 (%)
50 이상	90.102
60 이상	66.552
70 이상	54.266
80 이상	48.122
90 이상	43.344
100	43.003

표 1: *E. coli*에서 일정 이상의 COG 기능 점수를 가지는 오페론 비율

표 1에서 보듯이 오페론을 구성하는 유전자는 대체로 같은 COG 기능 값을 가지는 것을 알 수 있다.

유전자간 거리 분석

유전자 간의 거리는 서열에서 연속하여 위치하는 유전자 간의 염기 쌍의 개수를 나타낸다. 이는 g_i를 i번째 유전자라고 했을 때, g_i[start], g_i[end]를 서열에서 g_i의 시작 염기와 마지막 염기의 위치라고 정의하면, 유전자간 거리는 g_{i+1}[start] - (g_i[end] + 1) 식으로 계산된다. 그림 1은 *E. coli*에서 오페론을 구성하는 유전자들 간의 거리와 오페론 경계 간의 거리를 나타낸다. 그림 1에서, 오페론을 구성하는 유전자들은 오페론 경계의 유전자들보다 가까이 위치하는 경우가 많다는 것을 알 수 있다.

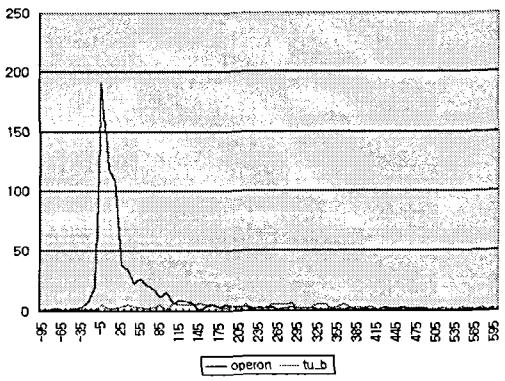

그림 1: *E. coli*에서 오페론을 구성하는 유전자들 간의 상대적 위치와 오페론 경계 간의 상대적 위치. x축은 base pair, y축은 해당 항목 수를 나타냄. Operon(파란선)은 오페론 내부에 속하는 유전자들 사이의 거리(base pair), tu_b(tu border, 적색선)는 오페론 경계(오페론 양 끝 유전자와 이에 인접한 오페론에 속하지 않는 유전자 사이) 부분의 거리를 나타낸다.

코돈 사용빈도(Codon Usage) 분석

코돈이 특정 아미노산으로 전사되는 정도를 코돈 사용빈도라고 한다. 코돈 사용빈도 특성은 유전자의 기능이나 발현 정도와 진화의 흔적등과 같은 요소에 의해서 영향을 받는 것으로 알려져 있다(Karlin, 1998)[11]. 이는 또한 유전자들이 오페론을 구성하는데 영향을 미치는 것으로 알려져 있다. 이를 근거로 같은 오페론에 속하는 유전자들에서 나타나는 코돈 사용빈도를 분석하였다.

코돈 사용빈도값을 계산하기 위해서 각 유전자 g_k에서 코돈 bias vector(Bockhorst, 2002)[3]를 구한다. 이는 $\{ \vec{b}_a^k \}$로 표시하며 a는 아미노산을 나타낸다. n_{uvw}가 코돈 uvw가 유전자 g_k에 나타나는 횟수라고 했을 때, bias vector 의 요소는 다음과 같이 나타낼

수 있다.

$$b^{k}_{a,uvw} = \hat{f}_{(uvw|a)} - \bar{f}_{(uvw|a)}$$

이때 *uvw*는 아미노산 *a*를 인코딩하는 뉴클레오티드이며, $\bar{f}_{(uvw|a)}$는 전유전체 상에서 *a*가 *uvw*에 의해서 인코딩되는 빈도수를 나타낸다.

$$\hat{f}_{(uvw|a)} = \frac{n_{uvw} + \bar{f}_{(uvw|a)}}{\sum_{xyz \in codons(a)} n_{xyz} + 1}$$

두 유전자 g_k, g_l간의 코돈 사용빈도 유사도는 다음과 같이 계산한다.[3,11].

$$Similarity(g_k, g_l) = \sum_{a} \vec{b}^{k}_{a} \cdot \vec{b}^{l}_{a}$$

위와 같이 코돈 사용빈도를 통해서 두 유전자간의 유사도를 비교할 수 있었으며, 이웃한 유전자들의 코돈 사용빈도 유사도와 오페론에 속한 유전자들의 코돈 사용빈도 유사도를 측정하여 분석해보았다.

전처리 모델 생성 방법

본 논문에서 제안하는 오페론 예측을 위한 전처리 모델 생성하는 과정은 디렉톤(directon) 생성, reciprocal pair 생성, 그리고 reciprocal pair cut을 거친다.

실험에 사용된 입력 데이터들은 *E. coli* 전유전체에서 각 유전자들의 유전자 위치와 서열, 전사 방향 그리고 블라스트를 통해 얻어진 COG 값이다.

Directon 생성

오페론을 구성하는 유전자들은 같은 방향으로 전사되는 특징이 있다. 따라서, 오페론을 분석하기 위해 먼저 입력 데이터를 전사 방향이 같은 연속된 유전자 집합 단위인 디렉톤[1]으로 분류한다. 오페론은 디렉톤 안에 존재하며, 서로 이웃하더라도 전사 방향이 다른 디렉톤에 걸쳐서 존재하지 않는다. *E. coli*에서 나타나는 오페론들은 2개 이상의 유전자들로 이루어진 813개의 디렉톤으로 구성되며, 이중 약 80%정도가 10개 이하의 유전자들로 구성되어있다.

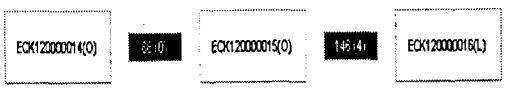

그림 2: 디렉톤(directon), 전사 방향이 같은 연속된 유전자 집합. 큰 사각형이 유전자를 나타내며 ECK0000(L)과 같이 [유전자ID(COG기능)]을 표시하였다. 검은색 작은 사각형은 유전자간 거리를 나타내며 [유전자간실제거리(COG 기능거리)]로 표시.

Reciprocal Pair 생성

디렉톤을 생성한 후, 각 디렉톤 내의 유전자들에 대해서 reciprocal pair를 생성한다. Reciprocal pair는 특정 기준을 적용하여 이웃한 유전자 간의 수치적인 거리를 구하고, 현재 위치의 유전자에서 좌우 유전자 중 가까운 한 곳으로만 연결했을 때, 서로 연결되는 쌍을 일컫는다. 여기서는 COG 기능 거리, 유전자 간 거리, 코돈 사용빈도 그리고 COG 기능 거리와 유전자간 거리를 함께 적용한 방법을 이용하여 reciprocal pair를 생성한다.

? NCBI에서 정의하고 있는 COG 기능은 총 18가지이다. 이들은 아직 제대로 파악되지 않은 기능을 나타내는 2가지(R과 S)를 제외하면, 총 3개의 그룹으로 나뉘어진다 (Information storage and processing, Cellular processes, Metabolism). COG 기능 거리는 해당 유전자가 얼마나 유사한 기능을 하는지 살펴보기 위한 것으로, 동일한 COG

기능 < 같은 그룹에 속하는 COG 기능 < 다른 그룹에 속하는 COG 기능 순으로 거리값에 차등을 주어 계산된다.

- 유전자 간의 거리는 유전자 간의 실제 거리와 상대적인 거리로 나눌 수 있다. 유전자 간의 실제 거리는 유전자의 크기를 고려하지 않은 거리로 g_{i+1}[start] - (g_i[end] + 1) 식으로 계산된다. 상대적인 거리는 유전자 간의 크기를 고려한 거리로 (g_{i+1}[start] - (g_i[end] + 1)) / (g_i[length] + g_{i+1}[length]) 식으로 계산된다. 여기서 g_i는 i번째 유전자, g_i[start], g_i[end]는 서열에서 i 번째 유전자의 시작 염기와 마지막 염기의 위치, g_i[length]는 유전자의 크기를 나타낸다.
- COG 기능 거리와 유전자간 거리를 함께 적용한 방법은 COG[g_i, g_{i+1}] × Max{Dis[g_*, g_{*+1}]} + Dis[g_i, g_{i+1}]로 거리를 구할 수 있다. COG[g_i, g_{i+1}]은 i, i+1번째 유전자 간의 COG 기능 거리를, Max{Dis[g_*, g_{*+1}]}는 모든 유전자 간의 거리 중 가장 큰 거리를 그리고 Dis[g_i, g_{i+1}]은 i, i+1번째 유전자 간의 거리를 나타낸다.

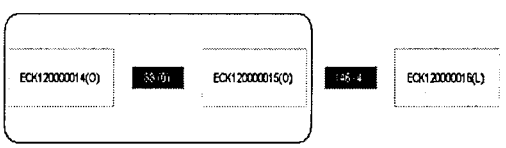

그림 3: Reciprocal pair. 디렉톤 내에서 서로 거리가 가까운 것끼리 연결된 pair로 그림에서는 COG 기능 거리에 따른 pair 구성을 보여준다.

주어진 기준에 의해 작성된 reciprocal pair를 실제 알려진 오페론 데이터와 비교해보면 그림 4와 같이 나타난다. Pair는 두 개의 유전자로 이루어지기 때문에 크기가 2인 오페론을 의미하거나, 또는 오페론의 부분집합을 의미한다.

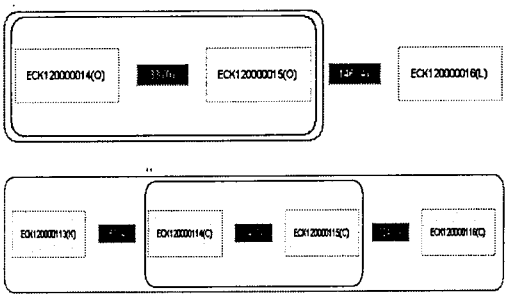

그림 4: Reciprocal pair와 실제 알려진 오페론(적색)을 표시한 그림.

Reciprocal Pair Cut

이 과정에서는 전체 reciprocal pair 중 유전자간 상호 관계 거리가 가까운 상위 pair들을 추출한다. 이것은 reciprocal pair 생성시 사용된 기준을 통해 생성된 reciprocal pair를 정렬하여, 원하는 상위 그룹을 추출 과정을 통해 이루어 진다. 추출된 상위 그룹들은 다른 pair들에 비해서 오페론이 될 확률이 높다고 볼 수 있다.

Reciprocal Pair 확장

생성된 reciprocal pair를 확장하기 위하여, 각 pair를 하나의 유전자로 가정하고, pair 생성과정을 반복한다. *E. coli* 유전체에서 COG 기능 거리를 이용한 두 번의 pair 구성 과정을 거쳐 나온 reciprocal pair와 3개의 유전자로 구성된 오페론 67개와 비교하면, 그 중 29개(43.3%)는 reciprocal pair의 양 끝 유전자와 오페론의 양 끝 유전자가 일치하였고, 적어도 한쪽 끝 유전자가 일치한 경우는 55개(82.1%)가 있었다.

실험 결과

본 논문에서는 *E. coli* 전유전체에 COG, 유전자간 거리, 코돈 사용빈도 그리고 COG

와 유전자간 거리를 같이 적용한 각각의 전처리 모델을 분석해 보았다. 사용된 입력데이터는 RegulonDB 3.1에서 얻은 유전자 위치와 서열 그리고 각 유전자 서열을 BLAST를 통해 얻은 COG 기능 값을 사용하였다. 그리고 생성된 전처리 모델을 평가하기 위해 RegulonDB의 tu_list 데이터 중 2개 이상의 유전자로 구성된 데이터를 사용하였다.

각 기준에 따른 전처리 모델의 분석 결과는 다음과 같다.

그림 5는 RegulonDB에 밝혀진 *E. coli*의 오페론 데이터와 COG 기능 거리를 이용하여 생성된 reciprocal pair를 비교한 결과이다. TP(true / predicted_reciprocal_num)는 *x*축을 따라 약 0.3 값을 가지며, TK(true / known_tu_num)은 *x*축이 1일 때 최고 0.83 값을 가진다(생성된 pair의 수는 1292개, 알려진 오페론 수 293개, 오페론에 속하는 pair 수 243개). 그림에서 *x*축은 상위 pair의 비율을 나타내며, TK는 생성된 pair 중 알려진 오페론에 속하는 pair 수 / 알려진 오페론 수의 비율을, TP는 생성된 pair 중 알려진 오페론에 속하는 pair 수/ 생성된 pair 수의 비율을 나타낸다.

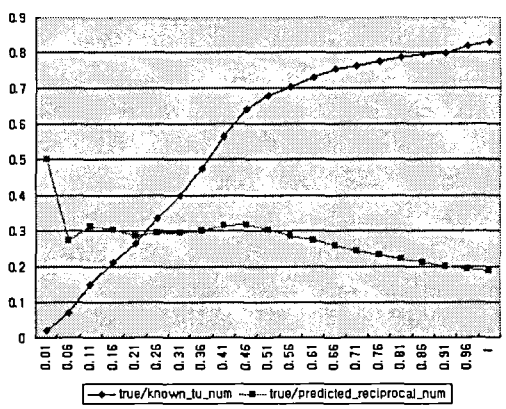

그림 5: COG 기능 거리를 이용한 reciprocal pair를 *E. coli*의 알려진 오페론과 비교한 결과. x축은 전체 reciprocal pair 중 점수가 높은 상위 비율로 1이면 예측한 모든 pair를 의미한다. y축은 오페론의 비율. 붉은 선은 해당하는 상위의 pair들 중 오페론에 해당되는 비율을 나타내며, 파란선은 전체 알려진 오페론 중 pair가 포함된 비율을 나타낸다.

그림 6, 7은 유전자 간 실제거리와 유전자 간 상대거리를 적용하여 생성된 reciprocal pair의 분석 결과이다. 두 경우가 서로 비슷한 결과를 보여주고 있으며, 그림 5와 비교해보면 TK의 최고값은 0.89로 유전자간 거리를 이용한 것이(그림 6, 7) 높으며, TP는 COG 기능 거리를 이용한 것이(그림 5) 전체적으로 높은 것을 알 수 있다.

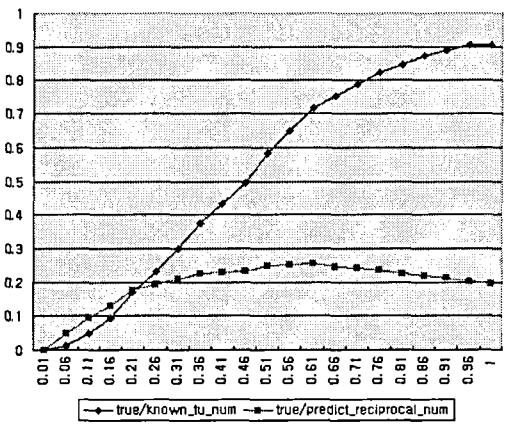

그림 6: 유전자간 실제거리를 이용한 reciprocal pair를 *E. coli*의 알려진 오페론과 비교한 결과.

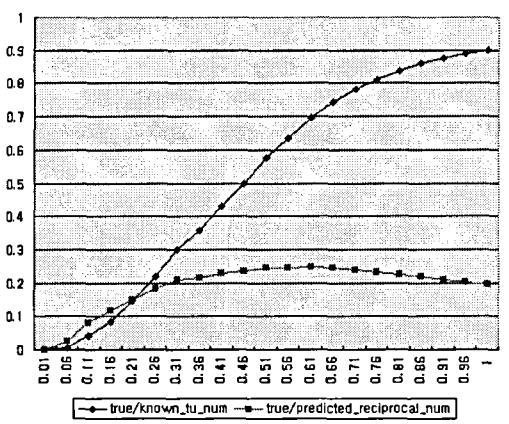

그림 7: 유전자간 상대거리를 이용한 reciprocal pair를 *E. coli*의 알려진 오페론과 비교한 결과.

코돈 사용빈도를 이용하여 reciprocal pair

를 생성한 경우, 그림 8에서 보듯이 앞에서 적용한 방법보다 TK와 TP 값 모두 낮게 나왔다.

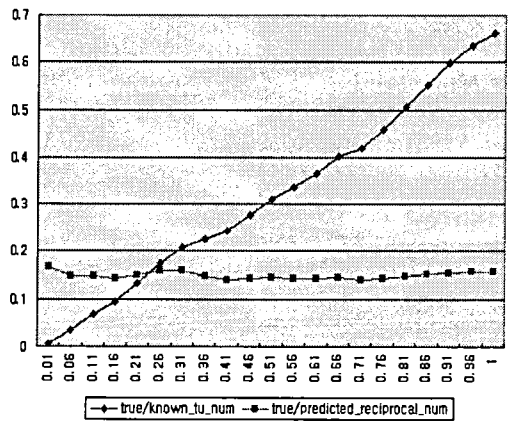

그림 8: 코돈 사용빈도를 적용한 reciprocal pair를 *E. coli*의 알려진 오페론과 비교한 분석결과.

TK와 TP 값 모두 좀 더 높이기 위해서, reciprocal pair 생성시 COG 기능 거리와 유전자간 거리를 함께 적용하여 보았다. 그림 9는 이 결과를 나타내며, 이 경우 TP 값은 최고 0.91로 다른 경우보다 높게 나왔으며, TK 값 또한 전체적으로 다른 경우보다 높게 나타났다.

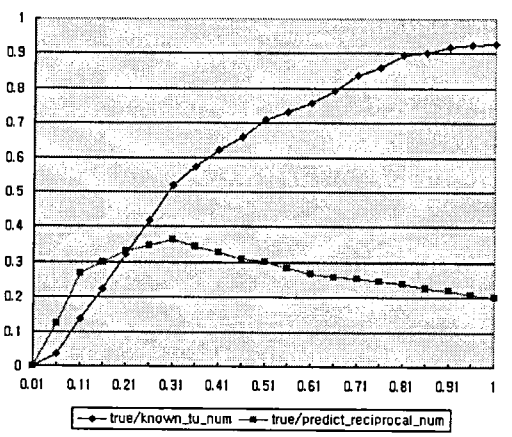

그림 9: COG 기능 값과 실제거리를 같이 적용한 reciprocal pair를 *E. coli*의 알려진 오페론에 적용시킨 결과.

표 2는 전체 알려진 오페론을 유전자 개수에 따라 분류하여, 각 분류된 오페론이 reciprocal pair를 포함하고 있는 비율을 나타낸다. 표에서 보듯이 오페론의 크기에 상관없이, 각 경우는 reciprocal pair를 80%이상 포함하였다.

오페론 크기	COG 기능거리	유전자간 상대거리	COG + 유전자간 거리	임의추출
2	93/133	107/133	109/133	29/133
3	59/67	63/67	66/67	42/67
4	34/36	36/36	36/36	29/36
5	25/25	25/25	25/25	22/25
전체	243/293	263/293	268/293	153/293

표 2: 오페론 크기에 따라 COG 기능거리, 유전자간 상대거리(relative distance), 임의 추출(random)을 통해 얻어진 reciprocal pair을 포함하고 있는 오페론의 비율

RegulonDB를 구성한 Salgado[1]는 현재 오페론이라 명확히 밝혀진 것은 일부이며, 전체 오페론은 적어도 700개 이상이라 예상했다. 2003년 9월 현재 RegulonDB 웹사이트[10]에서 발표한 바에 따르면 *E. coli*에서 오페론이 2325개가 존재할 것으로 예측하고 있다. 따라서 아직 밝혀지지 않은 오페론이 많다는 점을 고려한다면 앞 그림의 TK, TP 값은 향상될 것으로 예상된다.

본 연구에서 실험한 결과를 정리해보면 다음과 같다.

- Reciprocal pair를 적용해 인접한 유전자간의 상대적인 유사도를 고려할 수 있었다.
? Reciprocal Pair를 만들 때 적용하는 여러 가지 방법을 비교해보면 오페론을 예측할 때, 이미 많은 연구에서 쓰이고 있는 유전자간의 거리뿐만 아니라 COG 기능도 영향을 상당히 미친다는 것을 알 수 있다.

- COG 기능 거리와 실제 유전자간 거리를 같이 적용하여 전처리 모델을 생성한 경우, 가장 많은 오페론들이 이들을 포함하였다.
- 현재까지 밝혀진 전체 오페론 중 약 90%의 오페론이 reciprocal pair를 가지며, reciprocal pair를 포함하지 않는 유전자들은 상대적으로 오페론일 가능성이 적다.
- 본 연구에서 생성한 reciprocal pair는 오페론 예측 시 전처리 모델로써 좋은 도구가 될 것이다.

결론 및 향후 과제

본 논문에서는 오페론 예측을 위한 전처리 모델을 생성하는 방법을 제안하였다. 제안된 방법은 이전의 연구와 달리 실험적 데이터 및 알려진 오페론 학습 데이터가 필요 없으며, 현재 알려진 오페론의 약 90%가 전처리 모델을 통해 생성된 reciprocal pair를 포함하였다. 앞으로의 연구 과제는 다음과 같이 생각해 볼 수 있다.

- 최근 계속해서 추가되고 있는 오페론 데이터를 실험에 적용 시켜볼 수 있다. RegulonDB 사이트에서는 2003년 9월 현재 밝혀진 TU (transcription unit)는 705개라고 밝히고 있다. 데이터베이스가 아직 공개 되지는 않았지만 이를 적용한다면 우리가 사용한 487개의 TU를 통한 실험보다 정확도 높은 결과를 얻을 수 있을 것이다.
- 본 논문에서 제안된 전처리 모델인 reciprocal pair 생성을 위한 새로운 방법을 제안해 보아야 할 것이다. COG 기능 거리, 유전자간 거리, 코돈 사용빈도 그리고 COG 기능 거리와 유전자간 거리를 같이 적용한 방법 외 좀 더 높은 결과를 얻을 수 있는 새로운 방법을 찾아 보아야 할 것이다.
- 본 논문에서 제안된 전처리 모델은 완전한 오페론 예측이 아니다. 그러므로 앞으로 전처리 모델에서 완전한 오페론 예측 모델로 확장하는 방법에 대해서 연구해 보아야 할 것이다.

참고문헌

[1] Heladia Salgado, Gavriel Moreno-Hagelsieb, Temple F. Simith, and Julio Collado-Vides. Operons in Escherichia coli: Genomic analyses and predictions. *PNAS*, 97(12), pp. 6652-6657, 2000.

[2] Yu Zheng, Joseph D.Szustakowski, Lance Fortnow, Richard J.Roverts, and Simon Kasif. Computaional Identification of Operons in Microbial Genomes. *Genome Research*, pp.1221~1230, 2002.

[3] Joseph Bockhorst, Mark Craven, David Page, Jude Shavlik and Jeremy Glasner. A Bayesian network approach to operon prediction. *bioinformatics*, 19(10), pp. 1227-1235, 2003.

[4] COG, http://www.ncbi.nlm.nih.gov/COG/

[5] BLAST, http://www.ncbi.nlm.nih.gov/BLAST

[6] NCBI, http://www.ncbi.nlm.nih.gov/

[7] Heladia Salgado et al. RegulonDB(version 3.2) : a database on transcriptional regulation and operon organization in Escherichia coli. *NAR*, 29(1), pp. 72-74, 2001.

[8] Mark Caraven, David Page, Jude Shavlik, Joseph Bockhorst and Jeremy Glasner. A probabilistic Learning Approach to Whole-Genome Operon Prediction. *ISMB*, 2000.

[9] Maria D.Ermolaeva, Owen White and Steven L.Salzgerg. Prediction of operons in microbial genomes. *NAR*, 29(5), pp.1216-1221, 2001.

[10] RegulonDB, http://www.cifn.unam.mx/Computational_Ge

nomics/regulondb

[11] Joseph Bockhorst et al, Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, 19, pp. i34-i43, 2003.

[12] Tetsushi Yada, Mitsuteru Nakao, Yasushi Totoki and Kenta Nakai, Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models. *Bioinformatics*, 15(12), pp. 987-993, 1999.

[13] 강은미, Operon prediction을 위한 E. coli operon data의 Intergenic distance분석. *Technical Report*, Graphics Application Lab, Pusan National University, 2003.

[14] 천봉경, Operon Prediction In Microbial Genomes Using Some Characteristic Features. *Technical Report*, Graphics Application Lab, Pusan National University, 2003.

[15] 장철진, 오페론 예측 시스템의 개발. *Technical Report*, Graphics Application Lab, Pusan National University, 2003.