Knock-out Data 를 이용한 S. Cerevisiae 유전자 조절망의 재구성

홍성룡*, 손기락*
*한국외국어대학교 컴퓨터 및 정보통신공학과
e-mail : HongVsPark@paran.com

Reverse-engineering of Gene Regulatory Network of S. cerevisiae using Knock-out Data

Seong-yong Hong*, Kirack Sohn*
*Dept. of Computer and Information Communications Engineering
HanKuk University of Foreign Studies

요 약

하나의 유전자는 또 다른 유전자의 단백질과 프로모터 영역에서 Binding 함으로써 그 유전자의 발현에 영향을 미칠 수 있다. 이러한 두 유전자간의 조절 상호 작용을 유전자 조절망이라 하며 유전체의 핵심적인 기능을 보다 간결하게 표현하는 조절망을 설계할 수 있다.
대표적인 설계 방법으로는 Time-Series Data 를 이용한 방법과 Steady-State Data 를 이용하는 방법이 있으며 이 논문에서는 Steady-State Data 즉, Knock-out Data 를 이용하여 유전자 조절망을 재구성함으로써 기존의 방법을 개선하여 보다 정확한 결과 예측을 목표로 한다.

1. 서론

생물 정보학(Bioinformatics)이란 매우 폭넓은 분야를 담고 있는 "생명 현상 연구에 필요한 전산학 / 통계학 / 수학적인 것들"이라 표현 할 수 있다. 생물 정보학의 대상이 되는 중요한 데이터는 DNA 와 이에 코딩된 정보로부터 생성되는 단백질의 서열 정보, 단백질의 3 차원 구조에 대한 정보라 할 수 있다. 그 중 단백질들과 염색체 상의 DNA 형태로 존재하는 유전자, 그리고 이들 사이의 중간 단계인 RNA 가 어떻게 상호작용을 하며 어디에 얼마나 존재하고 어떠한 환경이나 조건에서 양이나 구조들이 변하는지에 대한 데이터가 있는데 이를 밝혀내는 도구가 DNA Microarray 또는 Proteomics 이다. 기존의 생명체 내부의 분자적인 메커니즘을 규명하기 위해 사용된 방법들은 적은 양의 RNA 나 단백질을 추적하는 식에 반해서 이 도구들의 특징은 대상이 되는 세포나 조직 속에 들어 있는 모든 RNA, 단백질을 추적해 볼 수 있다는 것이다. Microarray chip 에는 수천 개에서 많게는 수만 개의 단편들에 대응하는 양이 모두 측정치로 얻어지게 되며 이들은 개별적으로 고려해야만 하는 데이터 포인트들이 된다. 그리고 Microarray 실험에서는 단 한 장의 실험으로는 의미 있는 결과는 얻을 수 없고 가능한 많은 chip 을 사용해야만 한다는 점으로 인해 더욱 많은 양의 데이터가 얻어지게 된다.
이러한 Microarray chip 기술로 유전자의 발현 데이터를 대량으로 얻고 다양한 실험 조건하에 유전자의 발현 양상을 관찰하고 유전자간의 조절 관계를 분석 및예측 할 수 있게 되었다. 본 논문에서는 유전자간 조절 관계를 분석하는데 이용되는 방법 중에서 Time -Series Approach 보다는 더 확실한 결과 예측이 가능한 Knock-out Data 를 이용한 Steady-State Approach[1]를 개선하여 보다 향상된 결과를 예측할 수 있는 알고리즘을 제시한다.

2. 관련 연구

2. 1 DNA 와 RNA

세포는 모든 생명체의 가장 기본적인 활동단위로써 모든 세포의 핵은 생명체를 만들기 위한 유전정보를 가지고 있으며 단백질은 아미노산의 조합으로 만들어지는데 단백질의 아미노산서열은 핵산(necleic acid)의

염기 서열에 의해 결정된다.

핵산에는 DNA(Deoxyribonucleic acid)와 RNA(ribonucleic acid) 두 종류가 있으며 뉴클레오티드(nucleotide)의 중합체이다. 뉴클레오티드는 염기, 오탄당과 인산으로 구성된 DNA 와 RNA 를 구성하는 단위체이다. 뉴클레오티드는 오탄당(DNA 의 경우는 데옥시리보오스, 그리고 RNA 의 경우는 리보오스), 인산염기(PO_4)와 염기로 구성된다. 염기에는 아데닌(A), 구아닌(G), 시토신(C), 티민(T)과 우라실(U)이 있다.

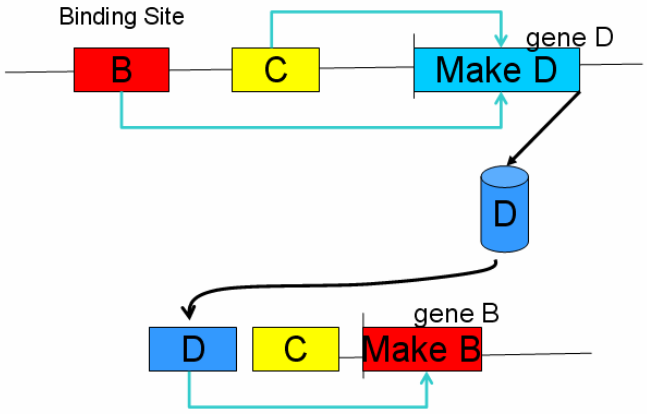
DNA 는 A, C, G 와 T 를 갖고 있고 RNA 는 A, C, G 와 U 를 갖는다. DNA 중합체는 이중나선이고 오탄당과 인산이 서로 교대로 연결된 두 개의 축을 형성하고 그 축을 연결하는 나선형 층계 모양을 하고 있다.

2.2 유전 정보의 흐름과 단백질 생성

DNA 는 단백질의 조절과 합성에 대한 정보를 가지고 있는 하나의 설계도이다. 특정 단백질의 합성 명령과 조합의 정보는 Gene 이라 불리는 DNA 의 단편(Segment)에 기록되어 있다. 분자 생물학의 시발점이라고 할 수 있는 Central Dogma 에 의하면 유전 정보는 DNA 에서 mRNA 로 전사(Transcription)되고 mRNA 에서 단백질로 번역(Translation) 되어 흐른다

- DNA → RNA : RNA Polymerase 가 DNA Sequence 에서 유전자의 mRNA 를 생산하는 Transcription 단계이며 전사된 mRNA 가 세포 내에 존재하고 활동할 때 Gene Expression 이라고 한다.
- RNA → Protein : 리보솜이라는 Micro-molecule 에서 mRNA 에 있는 유전 정보 코드에 따라 단백질을 합성하여 만들어내는 Translation 단계이다.

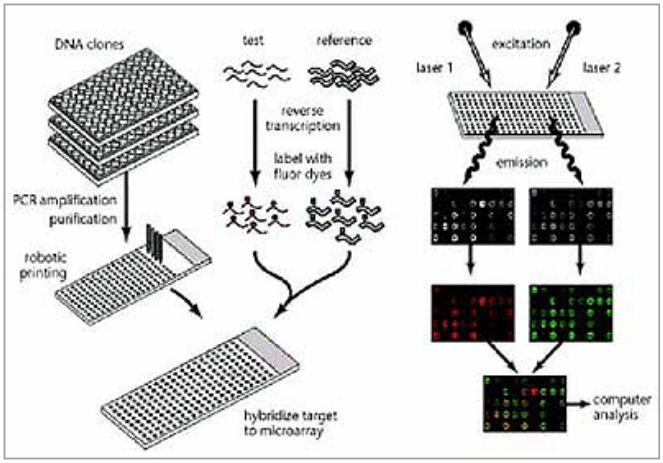
위의 두 단계를 거쳐서 생산된 단백질이 또 다른 유전자의 발현에 영향을 줄 수 있는데 Cis-Regulatory logic 에 따라서 도식화하면 아래의 그림 1 과 같다.



(그림 1) Cis-Regulatory Logic

유전자 B 가 발현하여 생성된 단백질이 유전자 D 의 발현에 영향을 주게 된다. 이 때 유전자 D 가 발현하여 다시 유전자 B 의 발현에 영향을 주는 단백질을 생산하는 관계를 나타내고 있다. 즉, 단백질은 다른 세포로 이동하여 다른 유전자의 발현을 억제 또는 촉진하고 다른 단백질과 결합하여 조직을 만들어 모든 생명활동을 조절하는 기능을 수행하게 된다.

2.3 Microarray Chip 과 실험 과정



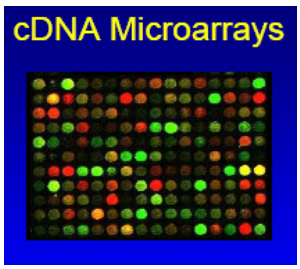
(그림 2) Microarray Chip 및 실험 단계[1]

위의 그림 2 와 같이 Microarray chip 또는 DNA chip 은 방대한 양의 유전 정보를 이용하여 유전자 발현, 변이나 다형성 등을 대량으로 고속 처리 검색함으로써 유전자의 기능을 밝히는 데 매우 유용하 도구이다. DNA 는 A-T, C-G 간의 강하고 선택적인 결합으로 이중나선을 이루는 고유한 특징을 가지고 있으며 DNA chip 은 그 상보적인 서열을 인지하고 선택적인 결합성을 활용한 chip 이다. 즉 chip 위에 한 가닥의 DNA 분자가 탐침으로 붙어있고 검색하고자 하는 DNA 시료를 침 위에 뿌려주면 상보적인 DNA 염기서열을 가진 DNA 들만이 탐침 서열에 가서 달라붙게 되는 것이다. 탐침 및 시료 DNA 간의 결합을 정량적으로 검출하기 위해서는 시료 DNA 를 형광 표시해서 처리하고 스캐너로 DNA 칩의 영상을 입력 받아 격자의 색깔과 농도를 판별하여 Microarray data 로 유전자들의 일치 정도, 발현 정도를 수치화하여 얻을 수 있다.

Microarray chip 의 일반적인 실험 단계는 다음과 같다.

- 서로 다른 두 환경의 세포들로부터 mRNA 추출
- mRNA 를 역전사(Reverse Transcription)시킬 때 각각 다른 색의 형광 불질을 띤 염기를 삽입하고 빨간색과 녹색을 띤 cDNA 를 합성.
- 합성된 두 개의 cDNA 를 같은 양으로 섞어서 하나의 cDNA Microarray chip 과 결합.
- 결합이 안된 유전자들을 씻어낸 칩은 스캐너에 의해 읽혀진다.
- 각 유전자의 형광 정도는 그 유전자의 발현 정도를 알려주는 것이며 컴퓨터로 분석.

2.4 Gene Expression Level 측정



(그림 3)

위의 그림 3 과 같이 Microarray chip 의 색깔과 형광 정도에 따라 유전자 발현 레벨은 기준(Control, Reference)에 대한 테스트(Test, Sample)의 발현양에 로그를 취한 값이다.
M = log Red / Green = log Red - log Green

- M < 0 : Gene 이 Under-Expressed, 녹색
- M > 0 : Gene 이 Over-Expressed, 빨간색
- M = 0 : Test 시료와 Reference 시료의 발현 양이 거의 동일

3. The Steady-State Approach

본 논문에서 사용된 S. Cerevisiae 데이터는 Hughes et al[2].의 Rosetta Inpharmatics[3]에서 구하였으며 효모의 일종인 S. Cerevisiea 의 내부적으로는 global mRNA 발현 하위 집합들을 포함하고 있다. 우선 한 쪽 집합은 Control Set 이라 부르는 정상적인 야생 S. Cerevisiea 를 63 번 샘플링한 knockout mRNA 발현 데이터를 제공하고 다른 집합은 Perturbation Set 인 정상적인 효모와 동일한 조건에서 배양한 287 개의 유전자에 대해서 각각 knockout 된 287 개의 S. Cerevisiea 의 발현양을 측정한 수치를 제공하고 있다[4]. Gene 에 대한 다량의 발현 level 데이터를 가지고 있다면 조절망을 설계할 수 있다. Steady-State Approach 는 다른 Gene 들이 발현할 때 임의의 한 Gene 을 삭제시킴으로써 삭제하기 전과 후의 다른 Gene 들의 발현 level 결과를 보다 명확하게 예측할 수 있는 접근 방법이다. 아래의 표 1 은 S. Cerevisiea 데이터에서 5 개의 유전자간의 상호 반응 매트릭스를 구성하기 위한 전단계로 log10(intensity)값을 추출한 표이다[5][6].

	ade1	aep2	ald5	anp1	ard1
ade1		**-0.06**	**0.07**	**-0.12**	**0.13**
aep2	**-0.22**		**-0.28**	**-0.42**	**-0.22**
ald5	**0.58**	**0.41**		**-0.19**	**0.61**
anp1	**0.2**	**-0.06**	**0.07**		**-0.1**
ard1	**-0.15**	**-0.12**	**0.19**	**-0.13**	

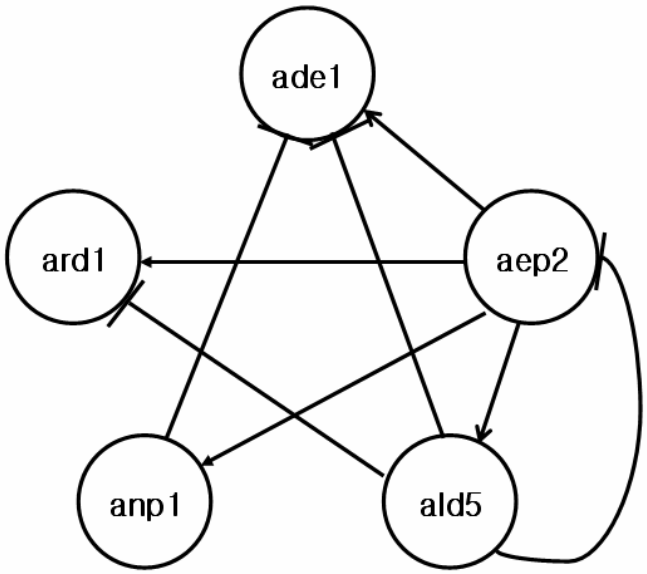
(표 1) log10(intensity) between five genes

아래의 표 2 는 표 1 의 intensity 값에 대해 임계값을 -0.2~+0.2 로 하여 실제 상호 반응 매트릭스를 구성한 표이다. 임계값을 설정하지 않았을 경우 모든 유전자간의 매우 작은 영향까지도 모두 고려해야 하기 때문에 조절망이 더 복잡해진다.

	ade1	aep2	ald5	anp1	ard1
ade1					
aep2	+		+	+	+
ald5	-	-			-
anp1	-				
ard1					

(표 2) Interaction Matrix between five genes

표 1 과 표 2 에서 1 열의 유전자들은 knockout 된 유전자를 의미하고 1 행의 유전자들은 각각의 Knockout 된 유전자에 대해 변화를 보이는 유전자들을 의미한다.또한 '-'는 negative 또는 repressor 역할을, '+'는 positive 또는 enhancer 의 역할을 한다.
표 2 의 상호 반응 매트릭스를 구하여 Redundant Network 을 설계하면 다음과 같다.



(그림 3) Redundant Network 구성도

그림 3 에서 정상적인 화살표는 negative 또는 repressor 의 역할을 하고 막힌 화살표는 positive 또는 enhancer 의 역할을 한다.

4. Redundant Network 제거 알고리즘

상호 반응 매트릭스에서 각 유전자간의 모든 관계를 도식화하면 매우 복잡해진다. 따라서 Redundant 조절망에서 중복되는 관계를 제거하여 간소화된 조절망을 구성할 수 있는데 여기에 적용할 수 있는 알고리즘은 다음과 같다.

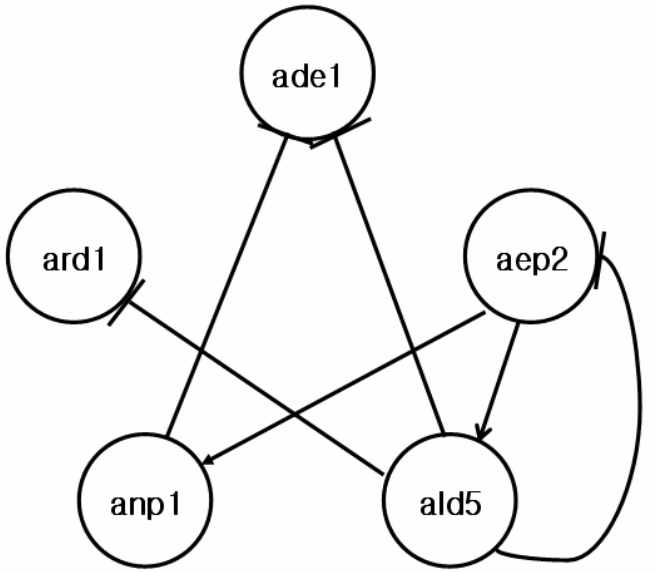
1. For each arc ***A*** from ***U*** to ***V*** on G
   if arc A is negative
   and if there exists an alternative path from ***U*** to ***V***
   where number of negative arcs is odd
   remove arc A from G

2. For each arc ***A*** from ***U*** to ***V*** on G
   if arc ***A*** is positive
   and if there exists an alternative path from ***U*** to ***V***
   where number of negative arcs is even
   remove arc ***A*** from G

(표 3) Redundant Network 의 제거 알고리즘

위 알고리즘에서 U 와 V 는 유전자를, A 는 arc 를 의미한다. 유전자의 발현을 억제(repressor)하는 역할을 한다면 negative 영향을 주는 것이며 촉진(enhancer)하는 역할을 한다면 positive 영향을 주게 된다. 또한 표 3 의 1 에서 유전자 U 에서 V 로 가는 arc A 에 대해 억

제를 하며 negative arc 의 개수가 홀수인 다른 선택경로가 존재하면 arc A 삭제 가능하다. 2 에서도 역시 arc A 가 positive 하고 negative arc 의 개수가 짝수인 U 에서 V 로 가는 선택적인 경로가 존재한다면 arc A 는 삭제 가능하다. 따라서 그림 3 과 같이 유전자 aep2 는 ade1 와 ard1 에 repressor 의 역할을 하고 있다. 여기서 aep2 는 ald5 를 거쳐 ade1 으로의 또다른 경로가 존재하며 negative arc 가 홀수 이므로 aep2 에서 ade1 로 가는 직접적인 arc 는 삭제할 수 있다. 또한 ae2p2 에서 ard1 으로 가는 직접적인 negative arc 가 존재하지만 aep2 에서 ald5 를 거쳐 ard1 으로 가는 선택적 경로가 존재하고 negative arc 의 개수 역시 홀수 이므로 삭제할 수 있다. 위의 알고리즘을 적용하여 간소화된 조절망을 구성한 것을 도식화 하면 그림 4 와 같다.



(그림 4) 간소화된 Redundant Network

5. 결론 및 향후 연구 과제

본 논문에서는 Knockout 데이터를 이용하여 각 각의 knockout 된 유전자의 log10(intencity) 값을 추출하여 상호 반응 매트릭스를 구성하면 각 유전자의 발현 결과물인 단백질이 다른 유전자의 발현을 억제 또는 촉진하는 역할을 예측할 수 있는 조절망을 설계하였다. 이 조절망을 Redundant Network 라고 하는데 이렇게 설계된 조절망은 knockout 데이터가 많아지면 굉장히 복잡해진다. 따라서 필요 없는 관계를 삭제하여 기존의 조절망을 개선하여 간략히 설계할 수가 있다.
하지만 knockout 된 각 유전자의 log10(intensity)값을 추출하여 상호 반응 매트릭스를 구성할 때 유의미한 임계값의 결정 방법에 대한 연구가 필요하다. 본 논문에서는 여러 번 바꾸어 설계해 봤으나 -0.2 ~ +0.2 로 설정했을 경우에 가장 적절한 조절망이 구성되었다. 이는 임계값의 범위에 따라서 조절망의 복잡도가 바뀌기 때문에 앞으로 가장 적절한 임계값 계산 알고리즘에 대한 연구가 필요하다.

참고문헌

[1] http:// http://www.genocheck.com.

[2] Isaac S. Kohne, Alvin T. Kho, and Atul J. Butte, *Microarrays for an Integrative Genomics*, MIT Press, 2003.

[3] http://www.rii.com Rosetta Inpharmatics.

[4] Farkas et al. "The topology of the transcription regulatory network in the yeast, Saccharomyces cerevisiea," Physica A 318 (2003) 601-612.

[5] Steen Knudsen, *A Biologist's Guide to ANALYSIS OF DNA MICROARRAY DATA,* Wiley & Sons, 2001.

[6] Andreas D. Baxevanis, B. F. Francis Ouellette, *Bioinfomatics : A Practical Guide to the Analysis of Genns and Proteins Second Edition*, ISBN 0-471-38391-0284-299, 393-410, Wilye & Sons, 2001.