[소프트웨어종합설계(2)]

Community Clustering

최종보고서

Proffesor
Team
Member

1. 연구목표

지난 학기에 개발한 PPI Network(Protein-Protein Interaction Network)에 관한 클러스터링 알고리즘을 그래프 특성과 생물학적 실험결과 등을 이용하여 개선, 발전시킨다.

2. 연구의 필요성

스마트 디바이스가 대중화 되면서, 폭발적으로 많은 데이터가 생성되고 있는 Big Data의 시대에 들어섰다. 따라서 이러한 대규모 데이터를 처리하는 기술은 최근 들어 점점 더 중요한 이슈로 부각되고 있다.

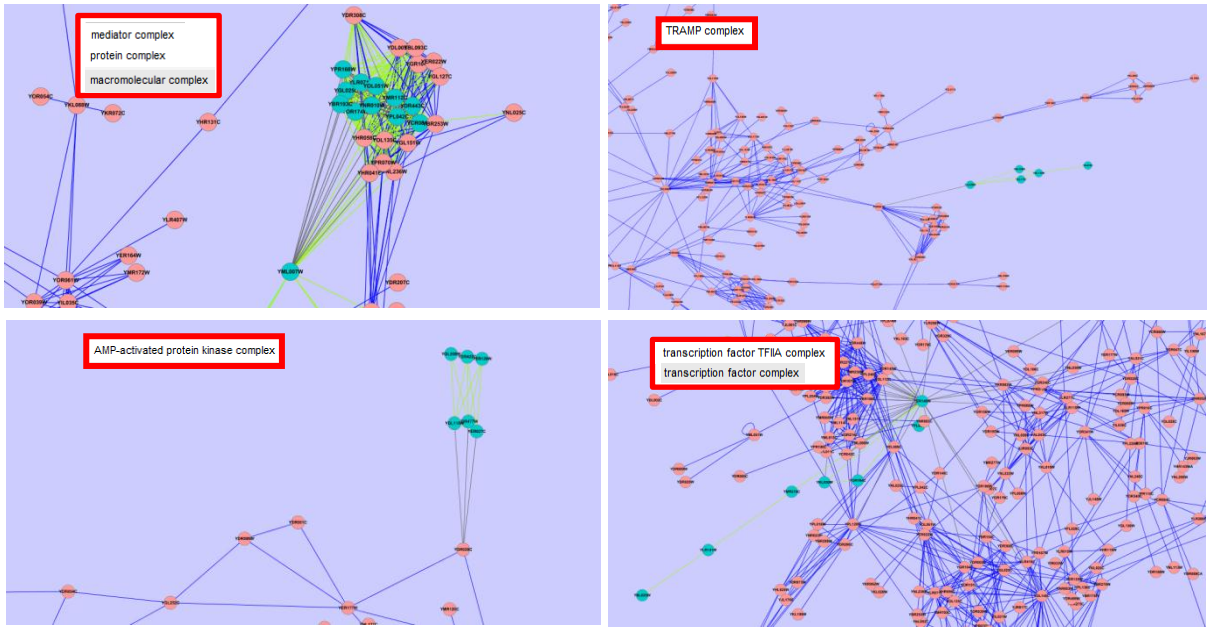
생물학 분야에서는 유전자 정보나 단백질의 역할 등을 파악하기 위해 이러한 데이터 처리 기술이 이용되어 왔다. 대부분의 단백질들은 다른 단백질과 Complex를 형성한다거나. 하나의 세포 안에서 어떤 Function 또는 Complex와 관련되어 있다. 따라서 Protein Complex를 찾는 것은 생물학적 기능이나 프로세스를 알기 위해서는 필수적이라고 할 수 있으며, 이 중 Clustering 기법은 효율적인 방법 중 하나로 널리 쓰이고 있다.

따라서 기존의 알고리즘들이 갖고 있는 문제점들을 해결할 수 있는 새로운 접근법의 Clustering알고리즘을 설계하는 것이 상당히 의미 있는 작업이 될 수 있을 것이다.

3. 기존 연구의 문제점

현재 잘 알려진 대표적인 Clustering 알고리즘들은 대부분 Overlapping을 허용하지 않고 있다. 하지만, 최근의 연구에 따르면 Overlapping을 허용하는 것이 더 정확도가 높다는 점을 확인 할 수 있다. 그러나 Overlapping을 허용하고 있는 알고리즘들의 경우, Overlapping을 지나치게 많이 허용하는 경향이 있어서 결과가 Agile하지 못하다는 단점이 있다.

지난 학기에 우리는 이러한 문제인식을 바탕으로 새로운 Clustering 알고리즘을 설계하였다. 아래 그림과 같이 우리의 알고리즘이 Complex를 검출 할 수 있었고, 각각의 Cluster들이 Protein Complex들로 어느 정도 의미가 있는 그룹으로 Clustering된 것을 확인해 볼 수 있었다.



APMS_Collins　　LC_Multiple

(괄호안은 총 Cluster개수)

	APMS_Collins (167)	LC_Multiple (126)
Complex 검출됨	151	98
1개의 Complex만 검출됨	25	19
Complex 검출되지 않음	16	28
평균 Complex 검출개수	3.63	3.23

하지만 우선 수행시간이 너무 오래 걸렸고, 정확한 Precision이나 Recall값을 모르기 때문에 알고리즘이 어느 정도의 정확성을 가지고 있는지 확실하게 알 수 없었다는 문제가 있었다.

따라서 지난 학기에 개발한 알고리즘이 갖고 있는 장점과 새로운 접근방식을 유지하면서 성능을 개선시키는 방향으로 이번 학기의 연구를 진행하였다.

4. 연구동향 조사

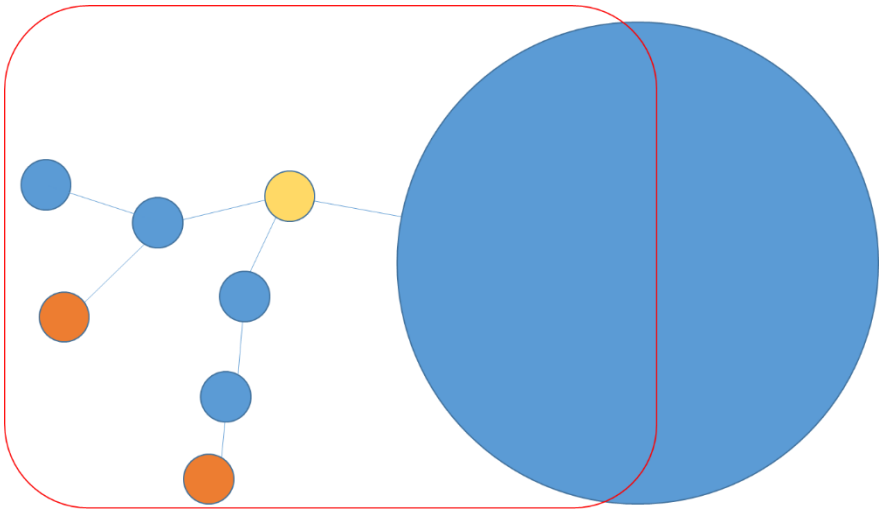
현 알고리즘의 문제점에 대한 인식이 정립된 상태에서 Bioinformatics에서의 연구동향을 살펴보았다. 기존 Wet Lab에서만 이루어지던 연구들은 동형의 데이터들에 치중해서 이루어졌다. 각 실험 기구에서 만들어내는 데이터들 역시 적절한 표준화가 이루어지지 않았고, 근래와 같이 다량의 실험결과를 양산하는 방향보다는 정확도가 높은 소량의 실험결과를 만들어내는 방향으로 진행되었었다. 또한, Dry Lab의 경우, Wet Lab에서 만들어내는 결과가 통계적으로 접근할 정도로 많지 않았고, Data Transformation에 대한 이해가 부족했기 때문에 이형의 데이터에 대한 동형화는 많이 이루어지지 않았다.

그 후, Human Genome Project의 영향으로 실험결과들의 스케일이 커지면서 이에 대한 처리와 분석의 필요성이 대두되었고, 이를 통해 Bioinformatics가 학문적으로 입지를 잡게 되었다. 초기 Bioinformatics의 연구에서는 대부분의 연구자들이 생물학적으로 치우치거나 컴퓨터 과학 쪽에 치우친 연구결과들을 보여주었다. 이러한 접근방법은 점차 서로의 정보를 통합하는 방향으로 발전되어 왔으며, 여러 개의 Heterogeneous한 데이터들을 기반으로 하여 의미 있는 결과를 도출해내는 방향으로 진행되고 있다. 이렇게 함으로서 Computer Scientific Approach는 기존과 달리 생물학적인 신뢰성을 가지게 되는 것이다. 따라서 이미 실험적으로 얻어진 binary protein interaction에 기반한 Network Analysis를 이용한 우리의 방법은 실제 생물학적 발현정보를 갖고 있는 MicroArray의 결과를 사용함으로써 생물학적인 신뢰성이 더욱 높아질 수 있을 것으로 판단된다.

5. 연구방향 설정

A. Betweenness Centrality

기존의 우리 연구는 모든 Node를 Starting Node로 삼고, 각 Node에 대한 Bottleneck Boundary를 구한 후, 이를 이용하여 개별적인 그룹을 잘라냈다. 그러나 아래서 보는 것과 같이 Spoke들에 속하는 모든 Node들은 동일한 그룹으로 Grouping 되는 것을 볼 수 있다(by Orange node, with Red boundary). 또한, 이러한 Spoke들은 인접하고 있는 Hub Node를 포함하는 Bottleneck boundary를 포함한다(by Yellow node, with Red boundary). 따라서, 이러한 Hub Node를 기준으로 Bottleneck boundary를 구한다면 Starting Node를 줄여서 Run-Time을 줄이는 동시에 기존 결과와 크게 다르지 않은 결과를 구할 수 있을 것이라고 생각한다.



B. Clustering Coefficient

기존의 연구는 Bottleneck Boundary만 이용하여 주어진 Graph를 Clustering하였다. 이를 통하여 구해진 결과들은 대부분 Protein Complex를 이룬다고 나왔으나 그렇지 않은 것들도 있었다. 결과를 검토해본 결과 Complex가 아니라고 판독 된 것들은 여러 패턴들이 존재하였나 Hollow한 결과들이 가장 많이 등장하는 것을 보았다. 따라서 Clustering Coefficient를 이용하여 기존 방법을 이용하여 구해진 결과들에 대해 Clustering Coefficient Threshold를 줌으로서 일정 수준 이상의 Cluster들만 결과로 출력한다면 Specificity와 Sensitivity가 더 올라갈 것이라고 예상하고 있다. 또한, PPI Network들은 서로 다른 실험적인 방법으로 Protein간의 Interaction을 구하기 때문에 이러한 Clustering Coefficient는 PPI Network자체의 경향성도 결과에 반영할 수 있을 것으로 기대된다.

C. Microarray

위의 Graph Centrality와 Clustering Coefficient를 이용하여 도출된 결과들을 최종적으로 Micro Array를 이용하여 각 Gene들의 유전자 발현 경향이 동일한 방향성을 갖고 있는지 검사하게 된다. 기존에 Micro Array에서 SOM이나 K-Nearest Neighbor를 이용한 방법들이 존재하였으나, 이는 Micro Array만을 이용하여 접근하는 방법이기 때문에, 우리의 방법을 통해 얻어진 결과에 대한 Micro Array를 이용한 사전결과 검증은 실험으로 구해진 이형의 Protein-Protein Interaction 데이터와 Microarray 데이터를 통합하는 것으로 생각할 수 있으며 이를 통하여 생물학적인 정보를 더욱 포함할 수 있을 것이라고 예상한다.

6. 연구내용

A. Idea

우리의 알고리즘에서 제안하는 Protein Complex 검출 방법은 크게 네 부분으로 구성되어 있다.

STEP1, PPI Network에 있는 모든 노드의 Betweenness Centrality를 계산한다.

STEP2, Betweenness Centrality가 Threshold보다 큰 각 노드에 대해 Distance Tree를 구축한다.

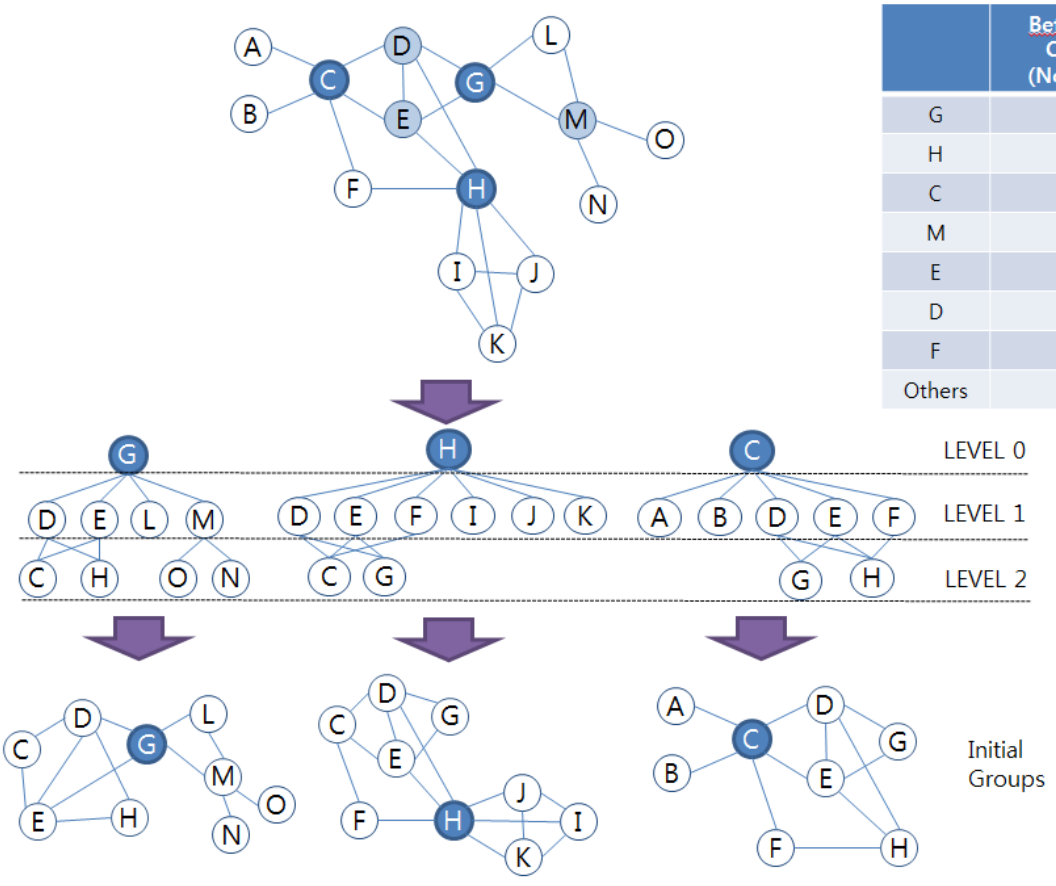
STEP3, 만들어진 Distance Tree를 Initial Group으로 이용하여 Protein Complex를 검색한다.

STEP4, 결과 그룹에서 Clustering Coefficient 들을 Threshold로 사용하여 결과를 재정리한다.

B. Distance Tree

구해진 Betweenness Centrality를 바탕으로 이 값이 Threshold보다 큰 노드를 Bottleneck노드로 간주한다. Bottleneck 노드를 Distance Tree에서 Root 노드로 삼아 Distance Tree를 구성한다. Root 노드에 직접 연결된 노드는 자식 노드가 되며, 같은 방식으로 하위 노드를 구성하여 Distance Tree를 구성한다. 확장 할 더 많은 노드가 없으면 프로세스가 종료 된다. 만일 프로세스가 두 개 이상의 Parent 노드를 만난다면 이것은 Bottleneck 노드일 가능성이 매우 높으므로 Tree를 확장하지 않고 프로세스를 종료한다. 또한 Root 노드도 Bottleneck 노드이기 때문에 Boundary가 Bottleneck 노드인 Sub-network로 볼 수 있다.

예를 들어 아래 그림에서 Bottleneck 노드인 G, H, C에 대한 Distance Tree를 구축해보자. G에 대한 Distance Tree의 경우, Root 노드는 G가 되고, G에 직접 연결된 D, E, L, M이 자식 노드가 된다. 또한 같은 방식으로 이들의 자식노드를 구하면 D, E는 공동으로 C와 H를 자식노드를 가지고, M은 O, N을 자식노드로 가진다. C와 H는 두 개 이상의 Parent노드를 가지기 때문에 더 이상 확장되지 않는다. 이를 통해 구축된 각 Distance Tree는 Initial Group으로 삼는다.



	Betweenness Centrality (Normalized)
G	1
H	0.9125
C	0.7125
M	0.625
E	0.6
D	0.6
F	0.1
Others	0

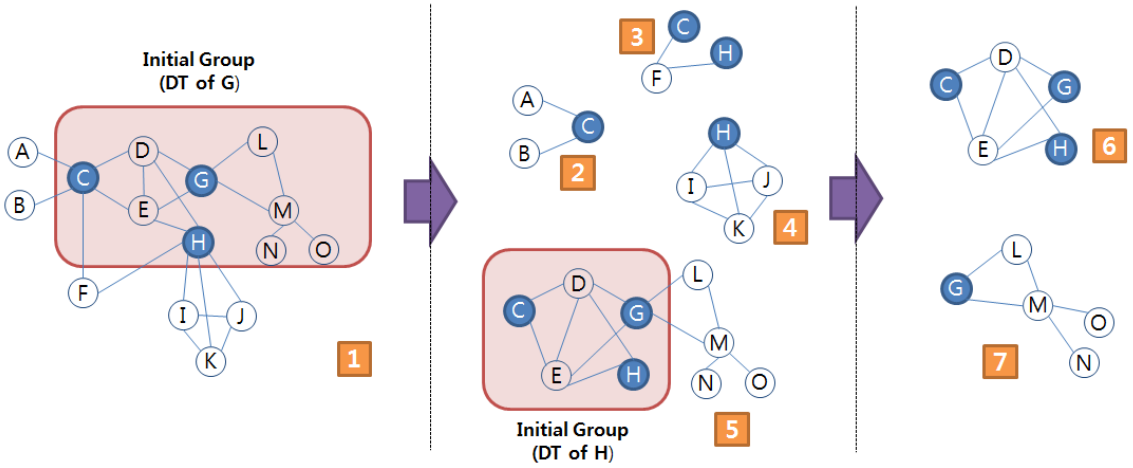
C. Divide PPI Network

위의 과정에서 생성된 Initial Group은 전체 PPI network를 두 개 혹은 그 이상의 부분으로 나눈다. 만일, 전체 Network를 두 개의 Initial Group A, B로 나눈 경우 이 두 개의 그룹은

Bottleneck 노드를 공유하게 된다.

아래의 그림을 보면 위의 과정에서 생성된 Initial Group(G에 대한 Distance Tree) Bottleneck 노드에 의해 2, 3, 4, 5로 나뉘어질 수 있다. 따라서 2와 5는 노드 C를 공유하고, 3과 5는 노드C와 H, 4와 5는 노드H를 공유한다. 이러한 공유노드를 이용하여 Protein Complex를 감지할 수 있게 된다.

다음 과정에서 앞서 나눈 2, 3, 4, 5를 다른 Initial Group에 의해 더 작게 나눌 수 있는지를 판단한다. 따라서 H에 대한 Initial Group으로 5를 6과 7의 더 작은 그룹으로 쪼갤 수 있다. 마찬가지로 6과 7은 노드 G를 공유하며, 다른 Initial Group에 의해 더 작은 그룹으로 쪼갤 수 있는지 판단한다. 하지만 C에 대한 Initial Group으로 더 작게 쪼개는 것은 불가능하기 때문에 여기에서 프로세스를 멈추고 각각 쪼개진 그룹들을 Protein Complex로 판단한다.



이때, 어떤 Initial Group을 먼저 사용하는가는 결과에 영향을 미치지 않는다. 어떠한 Initial Group으로 시작하더라도 최종 결과는 항상 같다.

위의 결과에서 얻어낸 Subgroup들을 Clustering Coefficient에 의해 Subgroup내의 노드들이 얼마나 밀집되어있는지(서로 연관되어있는지) 판별한다. 이 각 Subgroup에서 Clustering Coefficient를 계산하고 Threshold와 비교하여 의미가 없다고 판단되는 부분은 제외한다. 이 과정을 통해 Spoke형태처럼 서로간의 밀도가 낮은 그룹과 충분히 밀도가 높은 일반적인 형태의 Protein Complex를 구분해 낼 수 있고, 이것으로 알고리즘의 정확도를 향상시킬 수 있다.

7. 결과 및 검증

A. 실험 환경

우리는 결과 분석을 위해 Saccharomyces cervisiae에 관한 DIP와 BioGIRID, 2가지 PPI Network

Database와 Human PPI인 I2D Database를 이용하여 결과를 실험하였다. 각 PPI Network 에 관한 정보는 아래 표와 같다.

Database (version)	Species	Number of proteins	Number of PPIs
DIP (20071007)	Saccharomyces cerevisiae	4,823	16,914
BioGRID (3.1.69)	Saccharomyces cerevisiae	5,920	162,378
I2D (1.95)	Homo Sapiens	13,665	109,086

PPI Network Datasets

또한 우리가 설계한 알고리즘의 결과 검증을 위해 현재 알려져 있는 각각의 Protein Complex에 관한 Dataset인 MIPS, CYC2008 그리고 CORUM Database를 이용하였다. Reference Database에 관한 정보는 아래 표와 같다.

Database (version)	Number of protein complexes	Number of proteins	Average number of proteins in protein complexes
MIPS	81	885	12.358
CYC2008 (2.0)	236	1,627	6.678
CORUM (17.02.2012)	1,942	4,394	5.789

Refrence Datasets

B. 실험 결과

우선 설계한 알고리즘에 의해 식별된 Complex가 Reference Dataset의 Protein Complex와 일치하는지 여부를 판단하기 위해 우리는 Affinity Score를 사용하였다.

Reference Dataset의 Protein Complex와 우리가 구한 Protein Complex를 각각 A, B라고 하면 Affinity Score는 다음과 같이 계산 될 수 있다.

$$aff(A, B) = n\ (A \cap B)^2 / (n(A) \times n(B))$$

일반적으로 Affinity Score ≥ 0.2 인 경우, Protein Complex를 성공적으로 찾았다고 볼 수 있다.

또한 알고리즘에서 찾아낸 Protein Complex의 집합을 C, Reference Dataset의 Protein Complex 집합을 R이라 하면, 알고리즘의 성능은 Recall, Precision, F1 Score에 의해서 측정할 수 있다. Recall은 찾아낸 Protein Complex가 Reference dataset에 얼마나 존재하는가에 대한 비율이며, Precision은 찾아낸 단백질이 Reference Dataset의 Protein Complex와 얼마나 일치하는가에 대한 비율이다. 그리고 F1 Score는 테스트의 전반적인 정확도를 의미한다. 각각의 값은 다음과 같이 계산된다.

$$Recall = |R_{hit}| / |R|$$

Precision = $|C_{hit}| / |C|$

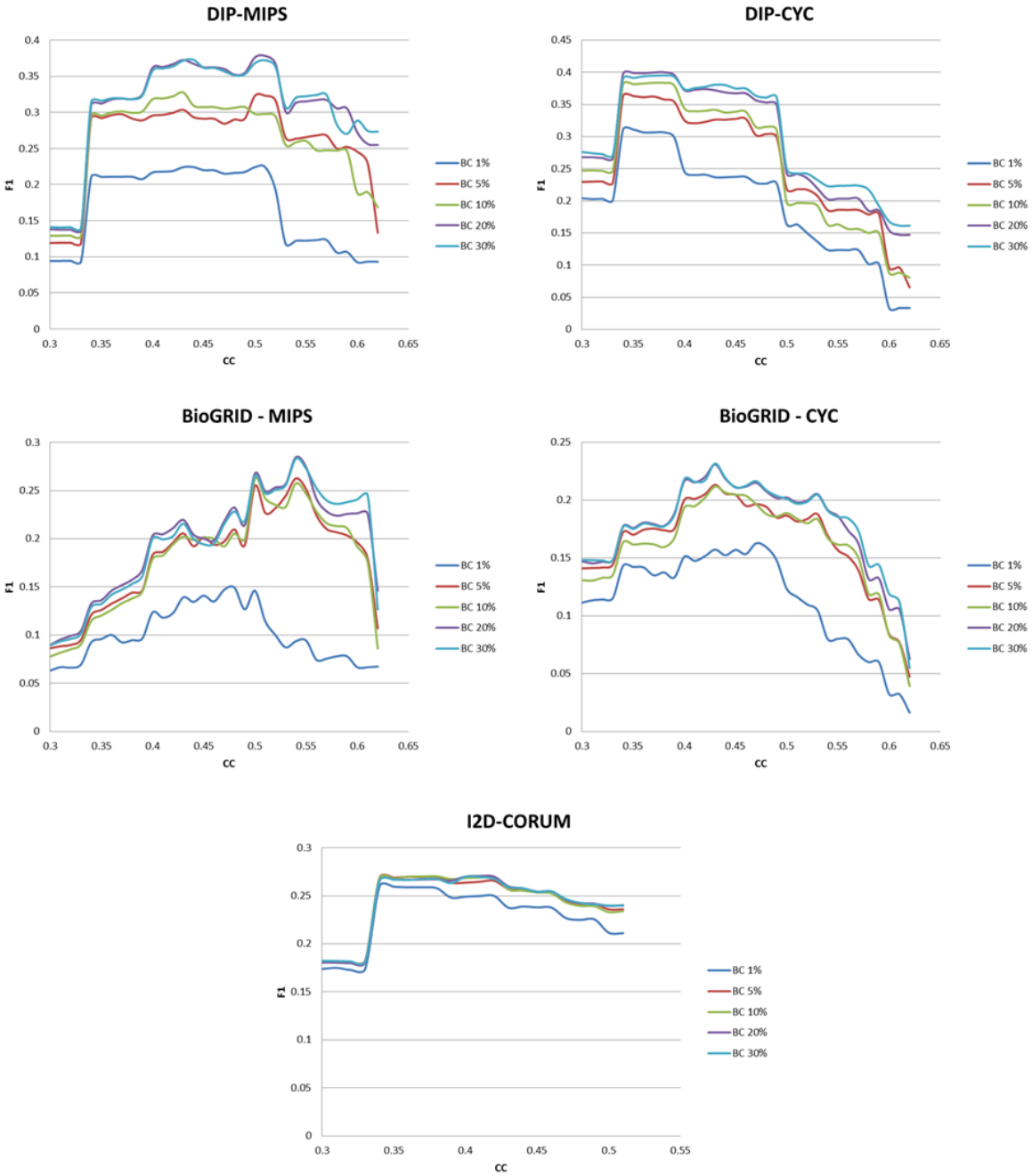
F1 score = Recall과 Precision의 Harmonic Mean

$R_{hit} = \{ R_i \in R \mid aff(R_i, C_j) \geq 0.2, C_j \in C \}$

$C_{hit} = \{ C_i \in C \mid aff(C_i, R_j) \geq 0.2, R_j \in R \}$

우리는 이를 바탕으로 다양한 Betweenness Centrality와 Clustering Coefficieint의 변화에 따라 결과가 어떻게 달라지는지 실험을 진행하였고 다음과 같은 결과를 얻을 수 있었다.

.



위의 그림을 보면 BC(Betweenness Centrality)와 CC(Clustering Coefficient)의 변화에 따라 F1 Score의 변화를 볼 수 있다. BC를 높이면 높일수록 정확도(F1 Score)는 증가하는 경향을 띄는데 너무 높인다면 수행시간이 길어지고, BC 상위 20%이상에서는 F1 Score에 큰 변화가 없기 때문에 우리는 모든 노드에서 클러스터링을 할 필요 없이 BC 상위 20~30%정도에서 수행하는 것이 효율적이라고 판단하였다.

또한 기존의 알고리즘들과의 성능 비교도 진행하였다. 아래의 표는 각각의 알고리즘들에 대한 Recall, Precision, F1 Score를 보여준다.

PPI Network Dataset	Reference Dataset	Algorithm	Optimal parameters	Number of protein complexes	Recall	Precision	F1 score
DIP	MIPS	Ours	CC = 0.51, BC = 20%	76	0.3210	0.4605	0.3783
		Ahn *et al.*	Partition_density = 0.30	1,177	0.7037	0.1427	0.2373
		MCL	Granularity = 2.00	614	0.5679	0.0739	0.1298
		MCODE	Node_score = 0.10	83	0.2930	0.2530	0.2729
	CYC2008	Ours	CC = 0.38, BC = 20%	333	0.3898	0.4114	0.4003
		Ahn *et al.*	Partition_density = 0.29	1,179	0.5932	0.2858	0.3857
		MCL	Granularity = 2.40	639	0.4746	0.1690	0.2493
		MCODE	Node_score = 0.10	83	0.2119	0.5542	0.3065
BioGRID	MIPS	Ours.	CC = 0.54, BC = 20%	69	0.2346	0.3623	0.2848
		Ahn *et al.*	Partition_density = 0.30	10,463	0.5926	0.0893	0.1552
		MCL	Granularity = 3.60	216	0.2099	0.0556	0.0879
		MCODE	Node_score = 0.10	120	0.086	0.0500	0.0633
	CYC2008	Ours	CC = 0.43, BC = 30%	324	0.2500	0.2160	0.2318
		Ahn *et al.*	Partition_density = 0.28	10,915	0.5297	0.2802	0.3697
		MCL	Granularity = 3.00	225	0.1144	0.1111	0.1127
		MCODE	Node_score = 0.10	120	0.0593	0.1167	0.0787
I2D	CORUM	Ours	CC = 0.41, BC = 20%	1132	0.2961	0.2491	0.2706
		Ahn *et al.*	Partition_density = 0.21	8,033	0.4576	0.1595	0.2378
		MCL	Granularity = 1.60	750	0.0623	0.0587	0.0604
		MCODE	Node_score = 0.10	251	0.0469	0.1076	0.0652

위의 표에서 보듯 우리의 알고리즘은 다른 알고리즘들에 비해 대체로 훨씬 높은 F1 Score값을 가진다는 것을 확인할 수 있다. 이로서 우리가 설계한 알고리즘이 상당히 높은 정확성을 가지고 있다는 것을 확인해 볼 수 있다.

8. MicroArray를 이용한 결과의 정확도 향상

Microarray는 특정 시점, 혹은 상태에서 Capture한 각 유전자의 발현상태를 Normal 샘플과 비교했을 때의 발현량 차이를 수치로서 나타낸 테이블이다. 우리는 PPI Network를 분할하면서 얻은 결과를 Microarray와 접목하여 정확도를 향상하고자 하였다.

Microarray는 위에서 설명한 것과 같이 각 Condition에서의 유전자 발현량과 정상 Sample의 Ratio에 로그 값을 취한 데이터를 나열한다. 따라서, 우리는 기능적으로 연관된 Gene Complex는 Microarray에서 특정 Condition에서 Inhibit이든 Activate이든 정상 발현량과는 다른 성향을 보일 것이라고 예상하였다.

Table A. Absolute Measurement

	C1	C2	C3	C4
Gene A	10	80	40	20
Gene B	100	200	400	200
Gene C	30	240	60	60
Gene D	20	160	80	80

Table B. Relative Measurement

	C1/C4	C2/C4	C3/C4
Gene A	0.50	4.00	2.00
Gene B	0.50	1.00	2.00
Gene C	0.50	4.00	1.00
Gene D	0.25	2.00	1.00

Table C. Log_2(Relative Measurement)

	Log_2 (C1/C4)	Log_2 (C2/C4)	Log_2 (C3/C4)
Gene A	-1	2	1
Gene B	-1	0	1
Gene C	-1	2	0
Gene D	-2	1	0

Table D. Discrete Values

	D[Log_2 (C1/C4)]	D[Log_2 (C2/C4)]	D[Log_2 (C3/C4)]
Gene A	-1	2	1
Gene B	-1	0	1
Gene C	-1	2	0
Gene D	-2	1	0

위의 Table C에서 보이는 것과 같이, Microarray에서 Condition C2에서는 정상 sample에 비해 Gene A와 Gene C가 activate 되었다고 나타난다. 따라서, 우리가 우리의 Method로 해당 Organism에 대해 구한 Subgroup에 대해 A, C와 같은 결과가 얼마나 포함되었는지 Count하고, 특정 Threshold 이상의 결과만 출력하고자 하였다.

하지만, 테스트해본 결과 모든 실험에서 오히려 결과가 나빠지는 것을 볼 수 있었다. 여러 논문들과 문서들을 참조해본 결과, 아래와 같은 결론을 얻었다.

우선, Microarray는 실험으로 구해지는 값으로, 우리가 생각하는 것 보다 Noise가 많았다. 또한, Microarray는 Target으로 한 Gene에 대해서만 결과를 얻기 때문에 전체적인 결과를 얻기 위해서는 여러 개의 Microarray를 중첩해서 사용해야 한다. 대부분의 Microarray는 비용문제 때문에 해당 Organism의 모든 Gene을 한 Plate에 포함하지는 않는다.

마지막으로 가장 중요한 점은, Microarray는 해당 Condition에서의 발현량의 여부만 기록한다. 즉, Gene 간의 거리는 아무런 상관이 없으며, 그저 Target의 발현량에 대한 기록만 있기 때문에, Microarray에서의 발현은 Inhibit관계와 Activate관계를 직접적으로 나타낸다고 보기 힘들다.

따라서 처음에 Microarray를 이용하여 알고리즘의 정확도를 높이고자 했던 방향대로 Microarray를 이용하기는 어려울 것으로 결론을 내렸고, 우리의 알고리즘을 Microarray에 적용한다면 현재 Microarray기술에서 가지고 있는 한계를 일정부분 개선할 수 있을 것으로 판단하였다.

9. Further Work

→ Differentially Expressed Gene Method를 개선한 Microarray Gene Regulatory Network 구축

한 유기체 내에 존재하는 Gene들의 개수는 전체를 Microscopic Level으로 연구하기에는 너무 많다. 따라서 전통적으로 생물학자들은 Microarray Data의 분석을 이용하여 다른 유전자와 다르게 발현된 유전자, 즉 Differentially Expressed Gene을 Target으로 삼아 연구를 진행하였다. 물론 이전보다 더욱 다양하고 많은 양의 의미있는 유전자에 대한 연구가 진행되었으나, 이렇게 드러나 있는 Gene에 대한 연구도 어느정도 한계에 봉착하였고, 이에 따라 연구자들은 Serendipity라고 칭할 정도로 운에 맡긴 채 연구의 Target을 정하고 있다.

따라서, 우리는 이러한 Differentially Expressed Gene에 가려져 발현량이 드러나지는 않지만 이러한 발현을 조절하는 핵심적인 유전자를 Targeting 하고자 한다. 이를 위해, 우리는 현재의 Clustering의 성능을 개선하고 각 Subgroup간의 Hierarchical한 구조 역시 밝혀내어 더욱 많은 정보를 얻고자 한다. 또한, Microarray 데이터를 이용한 Supervised Learning Model을 만들어 분명하게 드러나는 관계를 분석하여 발현을 조절하는 핵심 유전자를 밝혀내고자 한다.

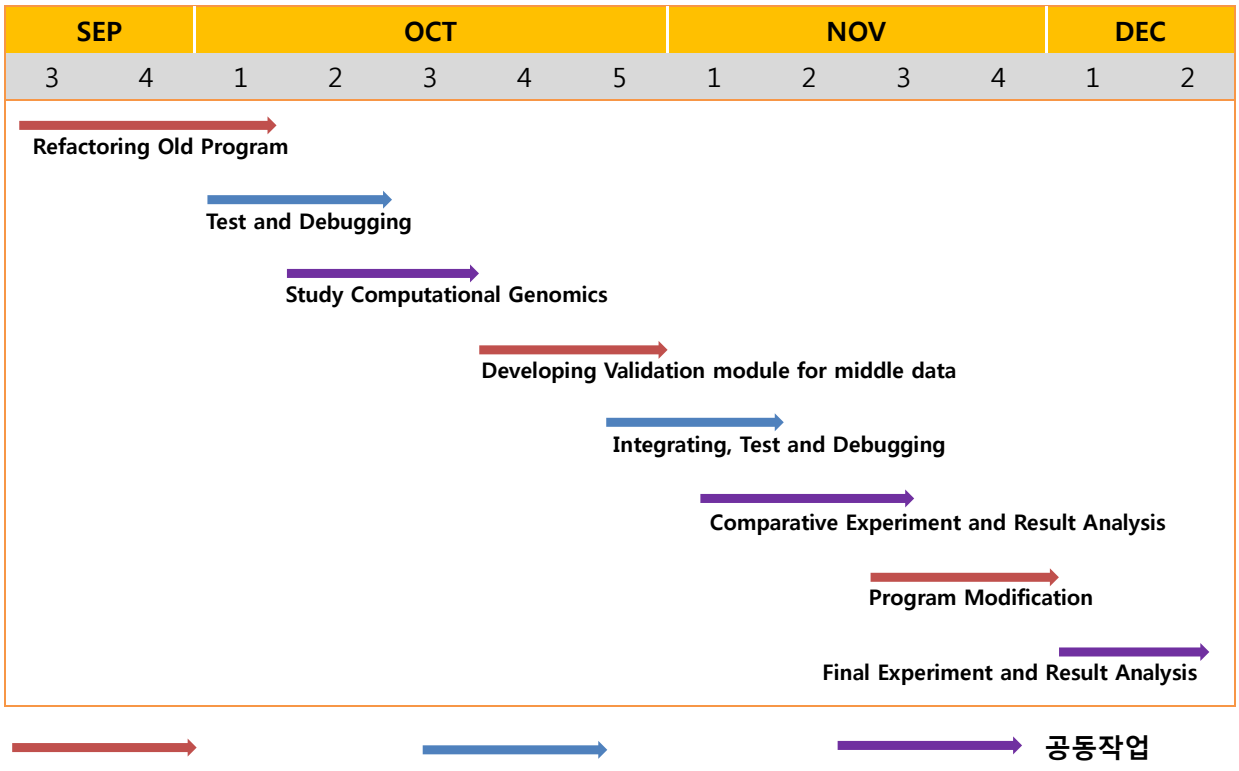
10. 결론 및 의의

우리의 알고리즘은 한 단백질이 많은 생물학적 기능이나 프로세스에 관여 할 수 있다는 사실을 바탕으로 서로 Overlapping을 허용하는 단백질 Complex를 발견 할 수 있다. 또한 Betweenness Centrality와 Clustering Coefficient를 이용하여 결과의 신뢰성을 유지하면서도 수행시간을 상당히 줄일 수 있었다.

그 결과, 전반적으로 우리의 알고리즘이 정확성 면에서 이전 알고리즘에 비해 우수하다는 것을 확인할 수 있었다. 기존에 주로 사용되지 않았던 Approach임에도 불구하고 기존보다 나은 성능을 보여주는 것으로 보아 Bottleneck을 Gene의 기능적 분기점으로 보는 것은 합당하다고 보인다.

우리의 알고리즘은 세포에서 단백질 Complex, 또는 기능 모듈의 계층 구조를 밝혀내는데 도움이 될 것이다. 또한 유전자 Ontology 데이터베이스와 같은 다양한 Biological 데이터베이스에서 그들의 기능을 추론하는데 도움이 될 수 있을 것이다. 또한 기존 생물학자들이 Microarray에서 보이는 모순에도 불구하고 Differentially Expressed Gene을 선호하는 것을 보았을 때, 우리의 방법을 통하여 위 기술을 개선할 수 있다면 많은 가치를 지닐 수 있을 것으로 기대한다.

11. 역할분담



12.참고문헌

- Link communities reveal multiscale complexity in networks, Ahn et al., 2010, *Nature*
- Multifunctional proteins revealed by overlapping clustering in protein interaction network, Emmanuelle Becker et al., 2012, *Bioinformatics*
- Uncovering the overlapping community structure of complex networks in nature and society, 2005, Nature
- On the evolution of random graphs, Erdős, Paul; A. Rényi, 1960, *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*
- Fast algorithm for detecting community structure in networks, Newman,M.E., 2004, *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.*
- Bioinformatics - Trends and Methodologies, Mahmood A. Mahdav, 2011, *InTech*
- Data Integration in Bioinformatics: Current Efforts and Challenges, Zhang Zhang, Vladimir B. Bajic, Jun Yu, Kei-Hoi Cheung and Jeffrey P. Townsend
- Neural networks - A comprehensive foundation, Simon Hykin, 1999, "9. Self-organizing maps", *Prentice-Hall*
- Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data, Lisa M. McShane1, Michael D. Radmacher, Boris Freidlin, Ren Yu, Ming-Chung Li and Richard Simon, 2002, *Oxford Univ Press*
- Gene Expression Profi ling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis, Sotiriou C, Wirapati P, Loi S, Harris A, Fox S and meds J, 2006, *J Natl Cancer Inst.*
- Meta-Analysis of Microarrays: Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Cancer, Rhodes DR, Barrette TR, Rubin MA, Ghosh D and Chinnaiyan AM, 2002, *Cancer Res.*
- Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, Aravind Subramanian, Pablo Tamayo, Vamsi K. Mooth, Sayan Mukherjee, Benjamin L. Ebert, 2005
- Mining Graph Data, Diane J. Cook, Lawrence B. Holder, 2006, *Wiley*