Novel Clustering Algorithm Exploiting Bottleneck Nodes

- TEAM : O O
- MEMBER : 이O O O
- PROFESSOR : 박O 교수

Introduction

- 최근의 연구에 따르면 Overlapping을 허용하는 것이 더 높은 정확도를 가지고 있지만, 기존의 알고리즘들은 대개 Overlapping을 허용하지 않고 있음
- Overlapping을 허용하는 알고리즘들 또한 너무 많은 Overlapping으로 각 결과의 Uniqueness가 떨어짐

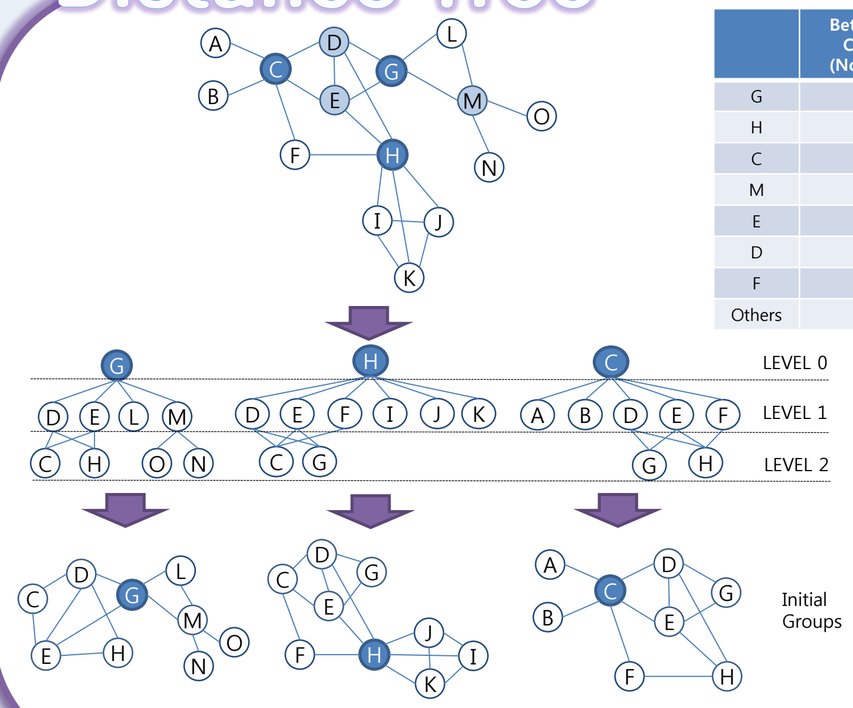
Goal

- Bottleneck node를 이용하여 Overlapping을 허용하는 알고리즘 개발
- Clustering Coefficient, Betweenness Centrality를 이용하여 성능을 개선
- DNA Microarray를 이용한 Differentially Expressed Gene Method의 문제점 개선

Idea

- **STEP1**
PPI Network에 있는 모든 노드의 Betweenness Centrality(이하 BC)를 계산
- **STEP2**
BC가 Threshold이상인 노드를 Bottleneck이라 가정, 각각에 대한 Distance Tree를 구축
- **STEP3**
만들어진 Distance Tree를 Initial Group으로 이용하여 PPI Network 분리
- **STEP4**
결과 그룹에서 Clustering Coefficient 들을 Threshold로 사용하여 일정 수준이상 Dense 한 subgroup을 최종결과로 출력

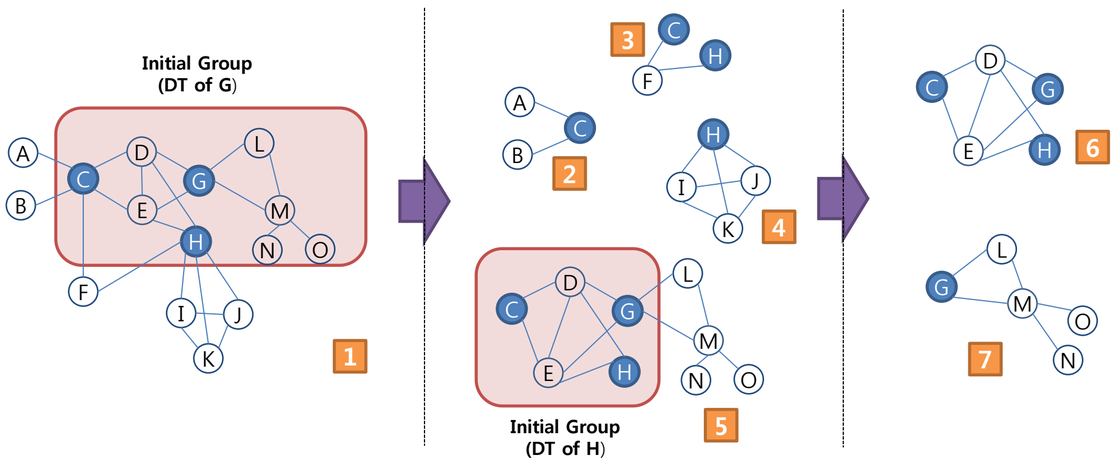
Distance Tree



	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

- Bottleneck node를 root로 시작
- Root node에서부터 특정 노드까지 2개 이상의 Shortest Path가 존재하거나, 더 이상 진행할 노드가 없다면 Building Process 종료

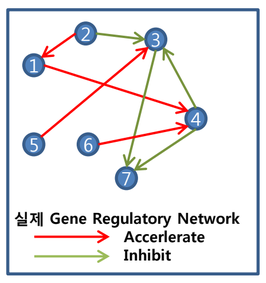
Divide PPI Network

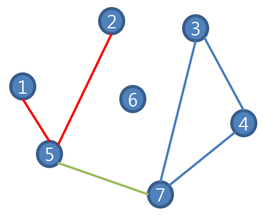


- PPI를 위에서 구해진 Distance Tree Element로 구해진 Set으로 나눔
- Distance Tree 구성과정에서 생긴Bottleneck Point는 Overlap이 가능하도록 함

Further Work

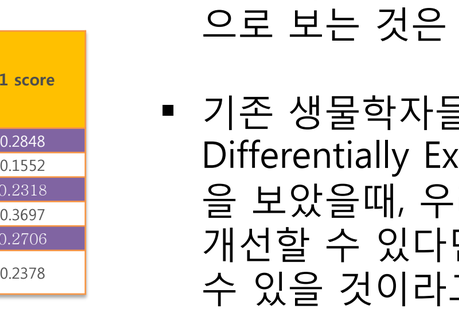
Improvement of Differentially Expressed Gene Method





	Cond1	Cond2	Cond3	Cond4
1	1	4	1	1
2	1	1	1	4
3	1	1	4	1
4	1	1	2	1
5	3	2	1	2
6	1	1	1	1
7	0.2	1	0.1	1

Differentially Expressed Gene method로 구해질 수 있는 Regulatory Information



Bottleneck Based Approach를 통해 얻어진 Subgroup

- 찾아낸 Subgroup에만 국한시켜 Microarray Data의 Expression을 분석한다면 기존 Differentially Expressed Gene에 비해 의미있는 Regulatory Information을 얻을 수 있을 것임

Result & Validation

Database (version)	Species	Number Of proteins	Number of PPIs
BioGRID (3.1.69)	Saccharomyces cerevisiae	5,920	162,378
I2D (1.95)	Homo Sapiens	13,665	109,086

PPI Network Datasets

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

Reference Datasets

Protein Interaction Network Dataset	Reference Dataset	Algorithm	Optimal parameters	Number of protein complexes	Recall	Precision	F1 score
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
	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
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

Result of comparison test

- 다른 알고리즘들에 비해 대체로 훨씬 높은 F1 Score값을 가가진다는 것 확인할 수 있음
- 기존에 주로 사용되지 않았던 Approach임에도 불구하고 기존보다 나은 성능을 보여주는 것으로 보아 Bottleneck을 Gene의 기능적 분기점으로 보는 것은 합당하다고 보임
- 기존 생물학자들이 보이는 모순에도 불구하고 Differentially Expressed Gene을 선호하는 것을 보았을때, 우리의 방법을 통하여 위 기술을 개선할 수 있다면 생물학적으로 많이 사용될 수 있을 것이라고 기대함