PPI 네트워크를 이용한 SNP 군집화 및 질병 연관성 분석

이규범*, 이선원**, 강재우**※
*고려대학교 바이오정보학 협동과정
**고려대학교 정보통신대학, ※교신저자
e-mail : {kyubumlee, sunwonl, kangj}@korea.ac.kr

SNP Grouping Method Based on PPI Network Information

Kyubum Lee*, Sunwon Lee**, Jaewoo Kang**※
*Program in Bioinformatics, Korea University
**College of Information and Communications, Korea University
※Corresponding Author

요 약

대용량 고차원의 생물학 데이터가 매우 빠른 속도로 생산되는 현재, 단순히 고전적인 알고리즘들로는 풀 수 없는 문제들을 맞이하게 되었다. 이러한 문제들의 경우 시스템 생물학의 관점으로 다양한 생물 데이터의 융합을 통하여 접근할 경우 효율적으로 Computational Infeasibility(계산 불가능)를 해결함은 물론 그 해석 및 새로운 정보 획득에 매우 유리하다. 인간 DNA 의 고차원 SNP 정보들의 군집화 및 질병 발현 패턴 분석은 그 조합의 수가 입력 데이터의 차원수에 따라 지수적(Exponentially)으로 증가하지만 PPI(단백질 상호작용) 네트워크 정보에 결합하여 필요한 중요부위를 선택적으로 이용할 경우 효율적으로 필요 SNP 들의 선택 및 이로 인한 공간 축소가 가능하다.

1. 서론

Human Genome Project 의 완료와 함께 쏟아지기 시작한 유전체 정보들의 양적인 증가는 최근 들어 Next Generation Sequencing 기술의 발달과 함께 더욱 가속화 되고 있다[1]. 이러한 데이터량의 증가는 생물정보학(Bioinformatics)의 탄생과 비약적 발전을 이루었다. 또한 다양한 형태의 생물학 데이터들의 축적은 이들의 통합을 통하여 하나의 시스템 차원에서 원리를 이해하고자 하는 시스템 생물학(Systems biology)적 접근을 가능하게 하였다[2].

현재 모든 DNA 상의 염기서열을 보지 않고 사람들이 차이를 보이는 약 1%의 염기서열들, 즉 SNP(Single Nucleotide Polymorphism, 단일 염기서열 다형성)만을 이용한 유전체 데이터 분석 방법이 많이 사용되고 있다. 이렇게 SNP 만을 이용할 경우 모든 유전체를 시퀀싱하지 않아도 되므로 시간적으로나 경제적으로 매우 유리하다. 이렇게 1%도 안되는 유전체의 정보라고 할지라도, SNP 데이터 역시 각 개인당 50~100 만개의 염기서열 정보로 이루어져 있다. 이러한 SNP 데이터와 특정 질병들과의 연관관계를 알아보는 GWAS(Genome-wide Association Study)가 많이 이루어졌으나, 각각 한개의 SNP 과 질병의 연관관계는 그리 크지 않음이 많은 연구를 통해 밝혀졌다[3]. 이는 질병이나 유전적 특성 발현의 복잡성 때문으로 이해되고 있다. 이러한 문제의 해결을 위해서 다양한 SNP 들의 조합들과 질병 발현의 관계를 보려는 시도가 이루어지고 있다[4]. 이는 하나의 질병 발현에 여러가지 유전자의 상호작용이 관련되어 있음을 기반으로 한다. 이러한 연구를 위해서는 SNP 의 다양한 조합과 질병간의 연관관계를 확인해야 하지만, 약 50 만개의 SNP 의 조합을 단순히 생성할 경우 그 조합의 수가 지수적(exponential)으로 증가한다. 그러한 수천~수만명의 데이터 분석에 이러한 조합들을 수십개의 SNP 들의 조합까지 생각할 경우 이는 실질적으로 계산 불가능하다. 이러한 문제를 풀기 위해 효율적으로 SNP 들을 조합하고자 하는 연구들이 수행되어 왔다. 우리는 그러한 효율적 조합의 방법으로 PPI Network Information 을 이용한 시스템생물학적 접근 방법을 이용하고자 한다.

2. 배경

(1) PPI Network

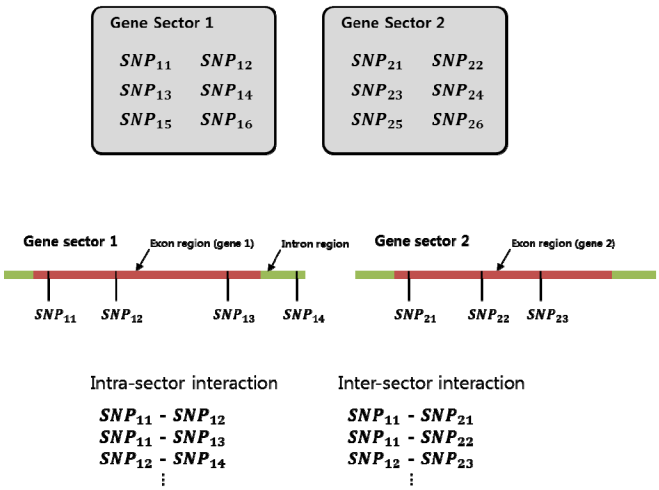
생물체 내에서 이루어지는 많은 기능들은 수많은 단백질들에 의해서 이루어진다. 이러한 단백질들은 단독으로 작용하기 보다는 단백질 복합체(Protein Complex)를 구성하여 작용한다. 또 단백질들은 생물학적 과정(Biological Pathway)에 속하여서 다른 단백질 또는 유전자들의 기능들을 활성, 비활성 하기도 한다. 이러한 단백질들간의 동시발현, 활성, 비활성 등의 작용을 주고받는 것을 PPI(Protein-Protein Interaction, 단

백질 상호작용)이라고 한다. 이러한 PPI 정보는 Yeast two-hybrid, Immunoprecipitation 등과 같은 생물학적 실험, DNA 내부의 공간적 위치의 근접성, 다양한 의·생물학 논문 초록 정보의 Text-mining 등을 이용하여 계산되고 있다.[5] 이러한 PPI 정보는 서비스를 제공하는 DB 에 따라 다양한 방법으로 수치화 되어서 나타나고 있다.

이러한 PPI 정보들을 한데 모으면, 각 단백질을 노드(Node), 각 단백질간의 PPI 정보를 간선(Edge)의 가중치(Weight)로 하는 네트워크 그래프를 구성할 수 있다. 이를 PPI Network 라고 한다. 이러한 PPI Network 분석을 통하여 아직 알려지지 않은 단백질의 기능들을 예측하거나, 유전적 특성의 발현 과정에 대한 이해, 질병과의 연관성 등을 밝혀내는데 도움을 줄 수 있다[5].

(2) Protein – Gene – DNA – SNP 매핑

각 Protein 은 DNA 상에 매핑되는 유전자를 가진다. DNA 상에 존재하는 SNP 들은 유전자 부분에 속하거나, 유전자 부분에 속하지 않을경우 인접한 유전자에 매핑되는 방법으로 SNP 과 Gene, Protein 간의 관계를 유추할 수 있다. (그림 1) SNP 이 반드시 유전자부분에 속하지 않더라도 유전자의 인접부분은 유전자의 발현 및 억제에 영향을 줄 수 있다는 많은 보고가 있으므로 이러한 SNP-단백질간의 매핑이 가능하다.



(그림 1) SNP 과 Gene 의 관계 및 Gene Sector

3. SNP Pairing 과 Computational Feasibility

일반적인 SNP 데이터셋이 50 만~100 만개의 SNP 정보를 포함한다. 이는 유전정보를 얻어내는 SNP Microarray 칩의 발달로 한개의 칩을 이용하여 50 만~100 만개의 SNP 정보를 한번에 알아낼 수 있기 때문이다. 이러한 50 만개의 SNP 두개의 조합을 구해볼 경우 (50 만)^2/2=1.25*10^{11} 의 속성(feature)이 생겨난다. 이러한 SNP 의 조합이 3 개 이상으로 증가할수록 이러한 속성(Feature)의 개수는 지수적(exponential)하게 증가하게 된다. 최근 SNP 데이터셋이 수천~수만명의 정보를 포함하는 상황에서 이러한 SNP 의 무분별한 조합은 결국 Computational Infeasibility(계산 불가능)를 야기하게 된다.

만약 이를 각 유전자에 매핑 시킬 경우 현재 알려진 유전자의 개수가 약 23,000 개이므로 약 50 만개의 SNP 은 평균 약 22 개정도가 한개의 유전자에 매핑되게 된다. 이러한 유전자를 다시 단백질과 매핑하면 두개의 단백질 상호작용 사이에 매핑되는 SNP 의 개수는 평균 44 개 정도로 예측할 수 있으며, 이러한 SNP 들의 조합들의 경우 약 470 개의 조합에 이름을 알 수 있다. PPI Network 의 간선(Edge)정보의 개수가 HPRD 기준으로 약 4 만개임을 가정해보면 총 feature 의 수는 2 천만개에 미치지 못함을 알 수 있다. 이는 이전의 무작위 페어링과 비교하여 1 만분의 1 에 해당하는 수치이다.

또한 수많은 유전자와 단백질의 질병과의 연관성이 이미 많은 실험과 검증을 통하여 널리 알려져 있으므로 이러한 정보를 이용하면 모든 단백질 및 PPI 정보를 사용하지 않고 이미 알려진 단백질이나 유전자의 선택을 통하여 효율적인 계산량 축소가 가능하다.

4. 실험

(1) 데이터 셋

우리는 WTCCC(Wellcome Trust Case Control Consortium)의 코호트 SNP 데이터를 사용하였다[6]. 이중 Inflammatory Bowel Disease(i.e. Crohn's disease, 이하 IBD)에 걸린 2005 명의 환자와 대조군 1504 명 500,568 개의 SNP 데이터를 이용하여 실험하였다.

OMIM 데이터베이스[7]에서 IBD 와 관련된 유전자와 관련된 단백질 121 개를 이용하였다.

PPI Interaction 데이터로는 HPRD(Human Protein Reference Database) [8]의 현재(2012 년 2 월) 최신 업데이트인 HPRD release 9 에 속한 39,194 개의 상호작용 정보 중 OMIM 에서 IBD 와 관련된 것으로 알려진 상호작용만을 이용하였다.

(2) 실험 방법

먼저 OMIM(Online Mendelian Inheritance in Man) DB 를 이용하여 IBD 와 관련된 단백질 121 개의 목록을 얻었다. OMIM 은 다양한 질병 및 유전적 특성에 관련된 유전자 및 단백질의 정보들을 제공하는 데이터베이스다. 이렇게 얻은 단백질에 속하는 SNP 들을 OMIM 에서 제공하는 Protein-SNP Mapping Table 을 이용하여 각 단백질의 그룹에 속하게 하였다. 또한 HPRD 에서 얻은 39,194 개의 PPI 정보 중 IBD 와 관련된 121 개 중 하나라도 포함하는 PPI 는 모두 실험에 포함시켰다. 이렇게 얻어진 SNP 쌍들의 염기서열 정보를 이용하여 질병의 유무를 판단하고 이의 정확도 및 통계적 유의성을 판단하였다.

(3) 실험 결과

이렇게 IBD 와 관련된 PPI 가 하나라도 속하는 단백질쌍으로 인해 구해진 단백질의 개수는 총 262 개

였고 이들에 속한 SNP 들의 개수는 22,229 개였다. 이러한 염기서열들의 종류의 조합 별 환자 및 대조군의 수를 비교하여 카이제곱 검정(Chi-square Test) 및 Bonferroni 보정을 통하여 p-value 를 계산하였으며 상위결과는 아래 <표 1>과 같다.

표에서 볼 수 있듯이 가장 낮은 p-value 를 가지는 최상위 세개의 SNP 조합들중 두가지는 PPI 를 이용하지 않는 하나의 단백질(Protein) 내부 SNP 들의 그룹이었다. 두번째 SNP 조합그룹의 경우 PPI Interaction 이 밝혀져 있는 CCR3 단백질과 FGR 단백질내부 SNP 들의 조합이었다. CCR3 는 이미 IBD 와의 연관성이 알려져 있지만[9] FGR 의 경우 그렇지 않다. 하지만 이러한 PPI 를 통하여 FGR 단백질과 IBD 와의 연관성을 유추해 볼 수 있는 근거를 제공할 수도 있을것이다.

SNPs	Protein	Protein	p-Value
rs17759529 rs16822665 rs12617656 rs10490422 rs7608798	DPP4	-	$4.9 * 10^{-5}$
rs13096142 rs4987053 rs1491962 rs17217831 rs1292089	CCR3	FGR	$4.5 * 10^{-4}$
rs946486 rs1800609 rs2253070 rs2253084 rs2583845	ABL1	-	$1.1 * 10^{-3}$

<표 1> 연관된 SNP 들의 그룹과 그에 따른 p-value

5. 결론

다양한 종류의 대용량 생물정보 데이터들이 생산되는 현재 데이터들의 적절한 융합을 통한 시스템 생물학 및 생물정보학적 접근은 고차원 대용량 데이터 마이닝의 Computational Feasibility(계산 불가능) 문제의 해결 뿐만 아니라 결과 해석 및 새로운 생물학적 정보 획득에도 많은 도움을 줄 수 있다. PPI Network 뿐만 아니라 Gene Regulatory Network 및 다양한 Genomics, Proteomics 관련 정보들이 공개되어 있으며 이를 어떻게 융합 및 활용하느냐가 효율적인 생물학 데이터 분석의 방향을 제시할 것이다.

참고문헌

[1] Green, E. D. and M. S. Guyer (2011). "Charting a course for genomic medicine from base pairs to bedside." Nature 470(7333): 204-213.

[2] Dollery, C., Kitney, R. (2007). "Systems Biology: a vision for engineering and medicine." The Academy of Medical Sciences

[3] Listgarten, J., S. Damaraju, et al. (2004). "Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms." Clin Cancer Res 10(8): 2725-2737.

[4] Cordell, H. J. (2009). "Detecting gene-gene interactions that underlie human diseases." Nat Rev Genet 10(6): 392-404.

[5] Droit, A., G. G. Poirier, et al. (2005). "Experimental and bioinformatic approaches for interrogating protein-protein interactions to determine protein function." J Mol Endocrinol 34(2): 263-280.

[6] The Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature 447(7145): 661-678.

[7] Online Mendelian Inheritance in Man (http://www.omim.org/)

[8] Human Protein Reference Database (http://www.hprd.org/)

[9] El-Shazly, A., N. Yamaguchi, et al. (1999). "Novel association of the src family kinases, hck and c-fgr, with CCR3 receptor stimulation: A possible mechanism for eotaxin-induced human eosinophil chemotaxis." Biochem Biophys Res Commun 264(1): 163-170.