유전자 중요도 계산 관련 고찰

조건: 특정 cancer(예를 들어, 췌장암)와 관련해서 중요한 유전자를 찾는 것이라고 가정.

처음에 논문에서 텍스트 기반으로 유전자 이름을 마이닝한 후 → 유전자 중요도를 계산

기준 1: 유전자 발현 기반 지표

- 발현량이 암 조직과 정상 조직 사이에서 크면 → 암의 발현과 관련해서 중요한 유전자라고 추론 가능.
- 이때 유전자 발현은 RNA의 양으로 정량화. RNA-seq 등의 연구 방법이 관련된 논문을 데이터로 사용해야 함.
- 그 중에서 특히 발현의 변동성이 크면 암 관련 GRN에서 중요한 역할을 한다고 생각할 수 있음. (변동성을 점수에서 가산점 느낌으로 부여)
- 장점: 유전자 자체의 발현 변화를 보기 때문에 확실한 중요도 계산 지표로 활용 가능
- 단점: 정상세포, 암세포 각각에서 특정 유전자의 발현량에 대한 raw data를 얻어야 함

기준 2: 네트워크 중심성 기반 지표

- GRN, PPI 네트워크 내에서 중심에 위치할수록 중요한 유전자라고 추론 가능.
- 네트워크는 Cytoscape, StringApp 등에 접근해서 가져와야 함. → 오픈소스 데이터 접근해서 확인하기.
- 노드의 개수, 상호작용하는 유전자의 개수를 중심성 점수 평가에서 사용하는 척도로 활용.
- 장점:
- 단점:

현재까지 찾은 논문들에서 동일한 척도를 사용해 유전자 중요도를 측정하는 논문은 몇 개 있는 걸 확인했음(DiCE, NGP 등). 그러나 수식화한 계산 알고리즘이 동일한 논문은 아직까지 확인하지 못함.

아이디어가 겹친다는 점은 좀 아쉽지만, 동일 아이디어로 논문이 여러 편 나오는 걸 봤을 땐 계산 방식이 다르다는 점 하나만으로 독창성을 가진다고 할 수 있는 것으로 보임.

이외에도 진화적 보존성(conservative), knockout 실험 결과, GWAS 결과 등을 가지고도 중요성 확인 가능하나, 모델 제작에 부적합하다고 판단함.

유전자 발현 기반의 중요도 계산

- 유전자 이름은 g로 함
- 실험에서 얻은 샘플의 데이터는 s로 함

1. LFC(log fold change) 계산

$$LFC_g = \log_2(\frac{\bar{x}_{g,cancer} + \epsilon}{\bar{x}_{g,normal} + \epsilon})$$

정상세포, 암세포 각각에서 g의 평균 발현량 평균의 비율에 로그를 취함 → 값의 범위를 줄일 수 있고, 배수 단위로 발현 변화 해석 가능

ϵ는 log의 분모가 0이 되는 것을 방지하는 용도

$$Z_g^{LFC} = \frac{|LFC_g| - \mu_{|LFC|}}{\sigma_{|LFC|}}$$

이후 LFC의 평균, 표준편차를 통해 표준화 → 전체 genome(또는 연구에 활용하는 유전자 pool) 내에서 g의 LFC가 어느 정도로 치우친 값인지 계산 가능

이후 t test 결과와 함께 발현 차이 정량화에 활용

2. Two-sample t test를 활용한 통계적 유의미성 계산

R 등을 활용해 raw data가 있으면 손쉽게 two-sample t test를 진행 가능 (이미 툴이 존재!)

이를 통해 p value를 구할 수 있음 → p_g

Quantile function(표준정규분포 역누적분포함수)을 이용해 LFC의 표준화와 동일하게 Z-score로 변환

$$Z_g^p = -\Phi^{-1}(\frac{p_g}{2})$$

3. 1번과 2번 값을 이용한 발현 차이 점수 부여 (점수 1)

$$Diff_g^* = \alpha_{Diff} Z_g^{LFC} + (1 - \alpha_{Diff}) Z_g^p$$

α_{Diff}는 0과 1 사이의 가중치 → 두 값을 적절히 섞어서 점수로 도출하는 것. LFC와 t test 결과 사이의 비중 차이를 고려해야 하나, 일반적으로 0.5로 둠

$$Diff_g = \frac{Diff_g^* - \min(Diff^*)}{\max(Diff^*) - \min(Diff^*)}$$

전체 연구하는 유전자 pool의 표준화 전 Diff 값을 이용 → 0과 1 사이의 숫자로 발현 차이 점수를 표준화

4. 유전자 발현의 변동성 (점수 2)

$$MAD_{g,cancer} = median(|x_{g,s} - median(x_{g,cancer})|)$$

우선 중앙값 절대 편차를 구함 → 이상치에 영향을 많이 받지 않고 샘플 데이터의 이산도 확인 가능

S에는 정상세포와 암세포 모두, cancer에는 암세포의 데이터만 넣어서 구하면 됨

이후 중앙값 절대 편차를 활용해 표준화된 변동성을 구함

$$Rob_g = \frac{MAD_{g,cancer} - \min(MAD)}{\max(MAD) - \min(MAD)}$$

마찬가지로 0과 1 사이의 값을 얻을 수 있고, 유전자 발현의 변동이 얼마나 강한지 확인 가능

5. **분류력(ROC AUC, 구분의 정확도) (점수 3)**

$$AUC_g = P(x_{g,cancer} > x_{g,normal})$$

AUC(Area Under the Curve)의 값이 0.5면 무작위로 분류된 수준, 0이나 1에 가까울수록 분류력이 더 강하다고 판단할 수 있음

이후 마찬가지로 표준화된 분류력을 구함

$$AUC'_g = 2(AUC_g - 0.5)$$

표준화 값이 0에 가까울수록 분류력이 약한 것, 1에 가까울수록 분류력이 강한 것으로 해석

6. **최종 발현 점수**

$$SCORE1_g = \beta_1 Diff_g + \beta_2 Rob_g + \beta_3 AUC_g$$

이때 $\beta_1 + \beta_2 + \beta_3 = 1$이 되어야 함. 각각은 점수의 가중치.

Diff가 50%, Rob이 20%, AUC가 30%의 중요도를 가진다고 가정하면 β_1=0.5, β_2=0.2, β_3=0.3의 값을 가지게 됨.

SCORE1의 값이 높을수록(즉, 1에 가까울수록) 암과 관련해서 발현의 변화가 크게 나타나고, 따라서 암 관련 기능에서 중요한 역할을 함을 시사.

네트워크 중심성 기반의 중요도 계산

1. **연결 노드 중심성**

$$C_g^{node} = \frac{node(g)}{N-1}$$

N은 전체 노드(유전자) 수, node(g)는 g와 연결된 노드 수. N-1을 한 건 correction을 위한 것.

N을 얻기 위해선 암과 관련된 전체 GRN 데이터가 필요함 → 관련 데이터베이스의 데이터 활용해야 함

상호작용하는 노드의 수가 많을수록 네트워크의 중심일 가능성이 높다고 해석

2. **근접 중심성**

$$C_g^{close} = \frac{N-1}{\sum_{g \neq h} D(g,h)}$$

D(g, h)는 g와 h 사이 최단 경로 거리로 정의. 가중치를 따로 가정하지 않으면 h와 g 사이 연결하는 화살표의 개수로 생각하면 됨. (다만, 가중치를 부여하려면 플로이드-워셜 알고리즘 등 복잡한 영역으로 감)

근접 중심성이 클수록 신호 전달의 거리가 짧기 때문에, 변화를 주면 주변 유전자에게 빠르게 변화를 미칠 것이라고 해석

3. **매개 중심성**

$$C_g^{med} = \sum_{g \neq h \neq i} \frac{short_{hi}(g)}{short_{hi}}$$

$short_{hi}$는 h부터 i까지 가는 최단거리의 개수, $short_{hi}(g)$는 그 중에서 g를 지나는 최단거리의 개수를 의미. 근접 중심성 구할 때 최단거리 계산을 해야 하므로 같이 해결 가능.

값이 높으면 유전자 사이를 연결하는 매개자 역할로써의 중요성이 높음을 의미 → 변화를 주면 네트워크 흐름에 큰 영향을 줄 것이라고 해석

4. **고유벡터 중심성**

먼저 N×N 크기의 행렬 A를 만들고, 이를 인접행렬(Adjacency Matrix)로 정의

인접행렬 A는 노드의 연결 여부에 따라 0 또는 w_{ij}의 값을 가지게 함

$$A_{ij} = \begin{cases} w_{ij}\ if\ there\ are\ some\ connections\ between\ two\ nodes \\ 0\ if\ there\ is\ no\ connection\ between\ two\ nodes \end{cases}$$

상호작용의 strength에 대한 정보를 얻을 수 있으면 w_{ij}는 가중치 점수를 반영한 수를 가지게 할 수 있음(연속적인 수 분포를 가지는 행렬을 얻을 수 있음). 그렇지 않을 경우 무가중 네트워크를 가정하고, w_{ij}의 값을 1로 고정(0 또는 1의 2진법 형태의 행렬을 얻게 됨).

$$C_g^{eigen} = \frac{1}{\lambda}\sum_{i=1}^{N} A_{gi} C_i^{eigen}$$

$Ac = \lambda c,\ c = (C_1^{eigen}, C_2^{eigen}, \dots, C_N^{eigen})^T$를 만족하는 스칼라 λ을 고유값으로 사용.

제일 큰 λ 값을 가지는 고유벡터를 정규화 → 이걸 토대로 정규화된 값으로 점수 부여

구글 PageRnak 알고리즘 활용. 중요한 노드와 연결되어 있으면 상대적으로 중요도 점수가 올라가게 됨.

- 선형대수 배워서 아마 너가 고유값이나 고유벡터 관련해선 더 잘 알지 않을까 싶다. 읽어보고 허점 있나 확인 한 번 해줘.

5. **최종 중심성 점수**

$$SCORE2_g = \gamma_1 C_g^{node} + \gamma_2 C_g^{close} + \gamma_3 C_g^{med} + \gamma_4 C_g^{eigen}$$

이때 $\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 = 1$이 되어야 함. 각각은 점수의 가중치. (앞의 SCORE1에서 가중치와 동일한 역할을 수행.

SCORE2가 높을수록(즉, 1에 가까울수록) 암 관련 네트워크의 중심에 있고, 그만큼 중요한 유전자일 가능성이 높음을 시사.

- Random Walk with Restart(RWR)도 구현 가능하면 중심성 지표에서 더 나아가 맥락적 중요성까지 반영 가능. 그러나 restart 확률 설정과 완벽한 무작위성 구현에서 한계가 있을 것으로 생각됨. (만약 추가하면, 최종 점수에서 RWR 요소를 추가하면 됨.)

최종 중요도 점수 부여

SCORE1이 0~1점 사이, SCORE2가 0~1점 사이 값이 나옴

따라서 두 점수를 단순히 합치면 0~2점 사이의 값이 나오게 됨

이를 0~1점 사이의 값으로 나오게 하려면 마찬가지로 표준화를 진행하면 됨.

$$SCORE = \omega_1 SCORE1 + \omega_2 SCORE2$$

가중치인 ω_1, ω_2는 합하면 1이 나오는 두 수. 작성자 본인은 발현량이 결국 기능과 직결되기 때문에 SCORE1의 비중이 더 높다고 생각, $\omega_1 = 0.7, \omega_2 = 0.3$을 제안함.

이후 최종 SCORE를 가지고 분포도를 제시하는 것도 좋다고 생각. 유전자 수가 많을수록 중심극한정리에 따라 정규분포를 따라갈 것이고, 통계적 유의미성을 검정하기도 좋을 것으로 기대.