Project Title: CHO세포의 UPR균형지수와 생산성 간의 상관성 분석

바이오의공학부 2023250054 윤현주

1. Problem Statement

CHO 세포는 항체와 같은 단백질 의약품 생산에 가장 널리 사용되는 세포주이다. 그러나 같은 배양 조건에서도 세포마다 단위 세포당 생산성, 즉 Qp가 크게 달라지고 이를 예측하기가 쉽지 않다. 단백질 생산 과정에서 소포체에 접히지 않은 단백질이 쌓이면 ER 스트레스가 유발되고 이를 완화하기 위해 UPR이라는 반응이 활성화된다. UPR에는 스트레스를 완화해 세포를 살리려는 방향과 스트레스가 심할 때 세포사멸을 유도하는 방향이 함께 존재한다. 이 두 방향의 균형이 세포의 운명과 생산성을 결정할 가능성이 크다.

이 프로젝트의 목표는 두 가지이다. 첫째, UPR과 단백질 분비와 관련된 유전자 발현으로부터 UPR 균형지수와 secretory 점수를 정의하고, 이 점수들이 CHO 세포의 Qp와 어떤 관계를 가지는지 정량적으로 확인한다. 둘째, 유전자 발현 데이터를 이용해 고생산 세포와 저생산 세포를 구분하는 분류 모델을 만들고, 단순한 경로 점수만으로도 어느 정도까지 생산성을 예측할 수 있는지, 그리고 Autoencoder로 학습한 저차원 표현이 이 예측에 얼마나 유용한지 평가한다.

2. Introduction to the Dataset

2.1 데이터셋 개요

본 연구에서는 공개 데이터베이스 GEO에 등록된 [GS E30321 Gene expression profiling of CHO production cell lines]를 사용하였다. 이 데이터는 여러 가지 배양 조건에서 얻은 CHO 세포의 마이크로어레이 전사체를 포함한다. 사용된 플랫폼은 GPL13791이라는 Affymetrix CHO 전용 어레이로, 원본 파일에서는 각 행이 probe ID 형태로 주어진다. 플랫폼 정보 파일에는 각 probe ID와 gene symbol의 대응 관계가 정리되어 있다.

각 샘플에는 유전자 발현 정보 외에도 다양한 공정 관련 메타데이터가 함께 제공된다. 예를 들어 샘플별 성장 속도, 배양 온도, lactate와 ammonia 농도, 세포 생존율과 같은 지표와 더불어 cell-specific productivity 즉 Qproductivity 값이 포함된다. 본 분석에서는 이 중 Qp를 생산성의 대표 지표로 사용하고 나머지 공정 변수는 해석을 위한 참고 정보로만 활용하였다.

2.2 전처리 및 사용 샘플 정의

전처리 과정에서 먼저 series matrix 파일에서 샘플 ID와 Qp 값을 추출하였다. `!Sample_geo_accession` 줄에서 각 샘플의 ID를 읽어 들였고, `!Sample_characteristics_ch1` 중 qproductivity 항목이 포함된 줄에서 정규표현식을 사용해 실제 숫자 값을 뽑아냈다. Qp 값이 존재하지 않는 샘플은 이후 분석에서 제외하였다. 남은 샘플에 대해 Qp 분포를 보고 하위 30퍼센트 구간을 저생산 그룹, 상위 30퍼센트 구간을 고생산 그룹으로 정의하였다. 이렇게 정의된 두 구간에 속하지 않는 중간 40퍼센트 샘플은 분류 문제를 명확히 하기 위해 사용하지 않았다.

유전자 발현 데이터는 series matrix에서 probe × sample 형태로 읽어 왔다. 이후 플랫폼 파일에서 ID와 gene symbol 정보를 불러와 probe ID를 gene symbol로 매핑하였다. 하나의 유전자에 여러 probe가 매핑되는 경우 probe 값의 평균을 사용하였다. 최종적으로는 각 행이 샘플, 각 열이 gene symbol인 발현 행렬을 구성하였다. 이 행렬에서 Qp 정보가 있는 샘플만 남겨 표현 분석과 모델 학습에 사용하였다.

3. Proposed Method

3.1 전처리

전처리 단계에서 먼저 발현 행렬에 로그 변환을 적용하였다. 구체적으로는 각 값에 log1p를 적용하여 저발현 유전자의 분산을 완화하였다. 그 다음 각 유전자에 대해 평균이 0, 분산이 1이 되도록 z 점수 정규화를 수행하였다.

이렇게 얻은 z 점수 기반 행렬을 이후 분석과 모델의 기본 입력으로 사용하였다.

고차원 유전자 전체를 그대로 사용하는 대신, 생산성과 관련된 주요 변동을 잘 포착하기 위해 분산이 큰 유전자만 고르는 단계를 거쳤다. 전체 유전자 중 분산이 가장 큰 상위 5,000개 유전자를 선택하여 이를 top variable gene feature로 정의하였다. 이 데이터는 PCA와 Autoencoder의 입력으로 사용하였다.

3.2 UPR 및 secretory gene set 기반 feature

UPR과 secretory pathway의 영향력을 보기 위해 문헌을 참고하여 세 가지 gene set을 구성하였다. Pro survival UPR 그룹에는 HSPA5, HSP90B1, PDIA4, PDIA3, PDIA6, XBP1, ATF6와 같은 샤페론 및 UPR 관련 유전자를 포함하였다. Pro apoptotic UPR 그룹에는 DDIT3, BBC3, BAX, PMAIP1과 같이 세포사멸을 유도하는 유전자를 포함하였다. Secretory 그룹에는 SEC61A1, SEC24D, SAR1A, EDEM1, HERPUD1, HYOU1과 같이 단백질 분비와 품질관리 과정에 관여하는 유전자를 포함하였다. 실제 데이터에서 해당 gene symbol이 존재하는 유전자만 선별하여 사용하였다.

각 gene set에 대해 해당 유전자들의 z 점수 발현값 평균을 계산하여 점수로 사용하였다. Pro survival UPR 점수는 upr_survival, pro apoptotic UPR 점수는 upr_apoptotic, 둘의 차이를 upr_balance로 정의하였다. Secretory gene set 평균은 secretory 점수로 사용하였다. 이렇게 해서 각 샘플마다 네 개의 연속값을 갖는 pathway feature 벡터를 얻었다. 일부 샘플에서 upr_survival이 양수이고 upr_apoptotic이 음수인 경우 upr_balance가 큰 양수 값을 보였으며, 동시에 secretory 점수도 높은 경향을 보였다. 샘플 몇 개에 대한 예시는 표 1에 정리할 수 있다.

	upr_survival	upr_apoptotic	upr_balance	secretory
GSM751635	0.180922	-0.096607	0.277529	0.524675
GSM751636	0.455325	-0.139309	0.594634	0.766122
GSM751637	0.071089	-0.191192	0.262281	0.412323
GSM751638	-0.098628	-0.848739	0.750111	0.120581
GSM751639	-0.587966	-0.131143	-0.456823	-0.065400

표 1. 선택된 CHO 세포 샘플에서 계산된 upr_survival, upr_apoptotic, upr_balance, secretory 점수 예시

실제 데이터에서는 제안한 모든 유전자가 어레이에 포함되어 있지는 않았다. pro-survival UPR 그룹 중에서는 HSP90B1, PDIA4, PDIA6 세 유전자의 발현이 검출되었고, pro-apoptotic 그룹에서는 DDIT3 하나만 데이터에 존재하였다. secretory 그룹에서는 SEC24D, HERPUD1, HYOU1 세 유전자가 사용 가능했다. 따라서 최종적인 점수 계산은 이 일곱 개의 유전자에 기반하여 이루어졌으며, 표 1에는 이 점수들이 고생산 샘플 몇 개에서 어떤 값 범위를 가지는지 예시를 정리하였다.

발현 데이터와 pathway feature, 그리고 이후에 설명할 Autoencoder latent 벡터에 대해 NaN 여부를 모두 확인한 결과 결측값은 존재하지 않았다. 따라서 별도의 결측값 처리나 대체 과정은 필요하지 않았다.

3.3 차원 축소와 Autoencoder

차원 축소와 비지도 표현 학습을 위해 먼저 PCA를 적용하였다. 첫 번째 경우에는 상위 5,000개 변동성 유전자 전체를 입력으로 하여 두 개의 주성분으로 데이터를 줄였다. 이 결과를 고생산과 저생산 샘플을 다른 색으로 표시한 산점도로 나타내었다. 이 그림은 보고서의 그림 1로 제시할 수 있다. 두 번째 경우에는 UPR과 secretory 관련 유전자만 골라 PCA를 수행했다. 이 결과를 역시 고생산과 저생산을 색으로 구분하여 산점도로 표현하였고, 그림 2에 담을 수 있다.

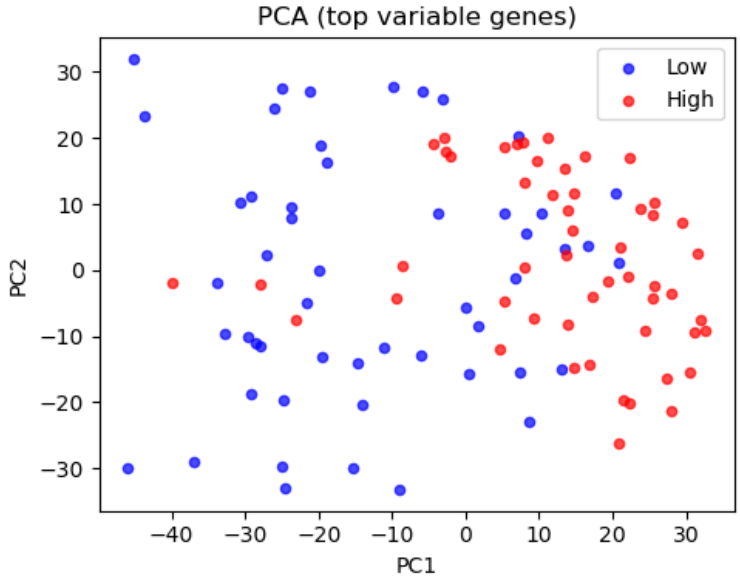

그림 1. 상위 변동성 유전자 5,000개를 사용한 PCA 결과 (PC1–PC2 공간에서 Low/High producer 분포)

Autoencoder를 이용해 비지도 학습 기반 저차원 표현을 얻었다. 입력은 상위 변동성 유전자 5,000개였고, 인코더는 5,000차원에서 256차원을 거쳐 2차원 latent로 압축하는 구조로 설계하였다. 디코더는 다시 2차원에서 256차원을 거쳐 원래 5,000차원으로 복원하도록 구성하였다. 손실 함수는 평균제곱오차를 사용하였고, 최적화는 Adam을 사용하였다. 학습은 50 epoch 동안 진행하였다. 학습 과정에서 reconstruction loss는 처음 약 1.05 수준에서 점차 감소하여 마지막에는 약 0.48까지 떨어졌다. 인코더의 2차원 출력을 latent 벡터로 두고, 이를 고생산과 저생산 샘플을 다른 색으로 표시한 산점도로 시각화했다. 이 결과는 그림 3에 제시할 수 있다.

3.4 분류 모델과 검증

분류 모델은 크게 두 종류의 입력을 사용하여 구성하였다. 첫 번째는 UPR 및 secretory 경로에서 계산한 네 개의 pathway feature를 사용하는 방법이다. 두 번째는 Autoencoder로 얻은 2차원 latent 표현을 사용하는 방법이다. 우선 pathway feature를 입력으로 Logistic Regression 모델을 학습하였다. 데이터는 고생산과 저생산이 유지되도록 층화하여 훈련 80퍼센트, 테스트 20퍼센트로 나누었다. Logistic Regression의 반복 횟수는 1,000으로 설정하였다. 같은 입력에 대해 단순 Perceptron 모델도 학습하여 비교하였다.

Autoencoder latent를 이용한 분류에서는 2차원 latent를 입력으로 Logistic Regression을 적용하였다. 이 경우에도 동일하게 80 대 20 비율로 훈련과 테스트를 나누었다. 또한 같은 latent 입력에 대해 간단한 다층 퍼셉트론을 구성하였다. 이 MLP는 입력 2차원, 은닉층 32차원, 출력 2차원 구조로 만들었다. 훈련과 검증 데이터는 8 대 2 비율로 나누고, 교차 엔트로피 손실과 Adam 최적화를 사용하여 50 epoch 동안 학습하였다. 각 epoch마다 훈련 손실과 검증 손실, 정확도, AUC를 저장하여 수렴 상태와 과적합 여부를 확인하였다.

마지막으로 단일 데이터 분할에서 나온 성능이 우연한 결과인지 확인하기 위해 pathway feature를 이용한 Logistic Regression에 대해 더 엄격한 검증을 수행하였다. 먼저 데이터 전체를 대상으로 층화 5-fold 교차 검증을 수행하여 각 fold에서의 정확도와 ROC AUC를 계산하였다. 또 다른 검증으로는 train test split의 random_state를 0부터 9까지 바꿔가며 10회 반복 분할을 수행하고, 매번 정확도와 AUC를 계산한 뒤 평균과 표준편차를 확인하였다. 이 결과는 표 2로 요약할 수 있다. 또한 pathway 기반 모델, latent 기반 모델, 그리고 비교용 Perceptron의 최종 테스트 성능은 표 3에 함께 정리하여 서로 비교할 수 있도록 하였다.

검증 방식	Accuracy (mean ± std)	ROC AUC (mean ± std)

5-fold 교차 검증	0.805 ± 0.057	0.854 ± 0.062
10 회 random train/test	0.810 ± 0.048	0.881 ± 0.071

표 2. Pathway 기반 Logistic Regression의 5-fold 교차 검증 및 10회 random split에서의 정확도와 ROC AUC 평균 및 표준편차

4. Results

이 장에서는 실제로 얻은 결과를 정리하고, 초기 가설과 어떤 점에서 맞았는지 살펴본다.

4.1 UPR 및 secretory 점수

먼저 UPR 및 secretory gene set에서 계산한 점수를 보면 네 개의 점수가 모두 0 근처에서 적당한 범위의 양수와 음수 값으로 분포하였다. upr_survival과 secretory 점수가 높은 샘플은 대체로 고생산 그룹에 많이 포함되었다. upr_apoptotic 점수가 낮고 upr_balance 값이 큰 양수인 샘플에서 생산성이 높은 경우가 자주 관찰되었다. 이러한 경향은 UPR에서 프로 생존 신호가 강하고 분비 경로가 활성화된 상태가 Qp 증가와 연관될 수 있다는 가설을 뒷받침한다. 이 점수들의 구체적인 예시는 표 1에 제시하였다. 표 1에서는 일부 샘플의 upr_survival, upr_apoptotic, upr_balance, secretory 값을 함께 보여 주며, upr_balance가 양수이면서 secretory 점수가 높은 샘플이 실제로 고생산 그룹에 속하는 모습을 확인할 수 있다.

4.2 PCA 결과

상위 변동성 유전자 전체를 이용한 PCA 결과를 보면, 그림 1에서 PC1과 PC2로 줄인 공간에서 고생산과 저생산 샘플이 완전히 섞여 있지는 않고 어느 정도 방향성을 가지고 분리되는 것을 볼 수 있다. 특히 첫 번째 주성분을 따라 한쪽에는 고생산 샘플이 상대적으로 많이 모이고 반대쪽에는 저생산 샘플이 많다. 이는 전사체 전체의 주요 변동 축이 생산성과 관련된 정보를 상당 부분 담고 있다는 뜻이다.

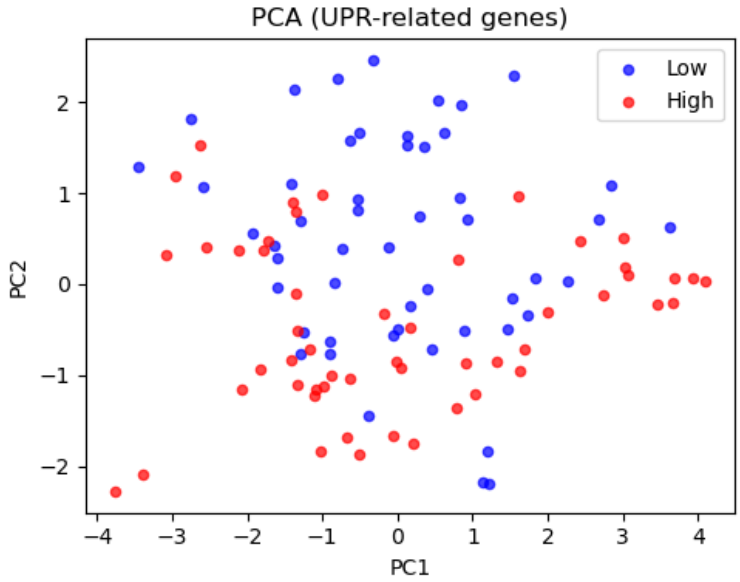

그림 2. UPR 및 secretory 관련 유전자만을 사용한 PCA 결과 (PC1–PC2 공간에서 Low/High producer 분포)

UPR과 secretory 관련 유전자만을 사용한 PCA에서는 그림 2에서 보는 것처럼 두 그룹의 분리가 상대적으로 약하다. 고생산과 저생산 샘플이 같은 영역에 섞여 있는 경우가 많다. 이는 UPR 관련 유전자만의 분산만으로는 전체 Qp 차이를 설명하기에는 정보가 부족할 수 있음을 보여준다. 하지만 이러한 정보를 단순 평균 형태의 경로 점수로 요약하여 사용했을 때는 분류 성능이 크게 개선되었다는 점에서, raw expression보다 pathway-level 요약이 더 유용할 수 있음을 시사한다.

4.3 Autoencoder latent 결과

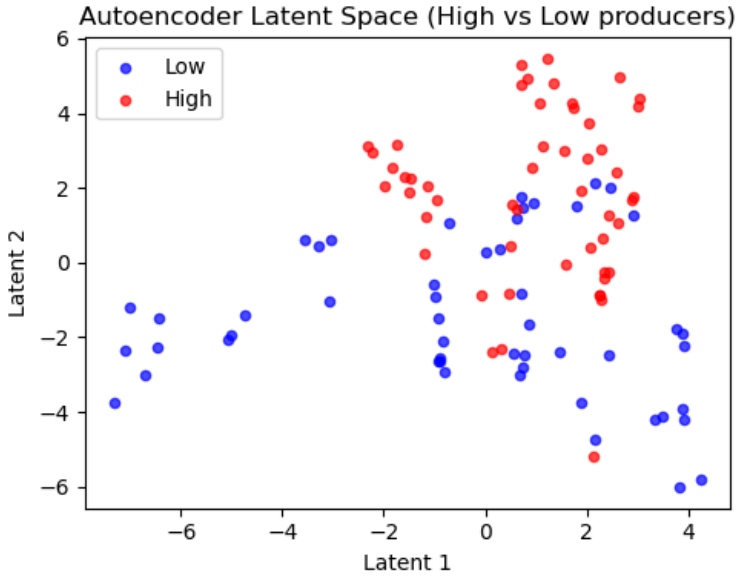

그림 3. Autoencoder로 얻은 2차원 latent 공간에서의 Low/High producer 분포

Autoencoder latent 공간에서는 그림 3에서 보는 것처럼 두 차원으로 줄인 표현 상에서 고생산과 저생산 샘플이 비교적 뚜렷하게 다른 영역에 모여 있다. 일부 영역은 고생산 샘플이 집중되어 있고, 다른 영역은 저생산 샘플이 주로 분포한다. 학습 과정에서 reconstruction loss가 epoch이 진행될수록 꾸준히 감소하여 50 epoch 시점에는 0.5 이하 수준에서 수렴하였다. 이는 Autoencoder가 전사체 전체 정보 속에서 생산성과 관련된 구조를 잘 포착하고 있음을 의미한다.

4.4 분류 성능 요약

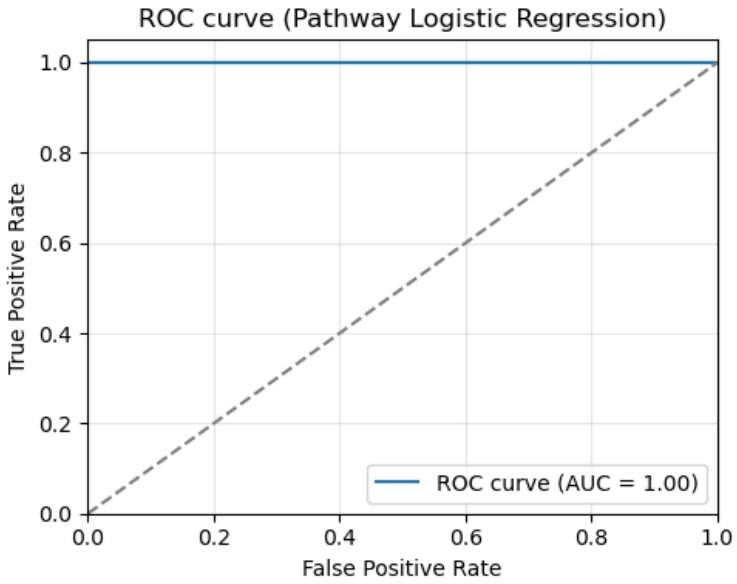

그림 4. Pathway 기반 Logistic Regression의 ROC 곡선 (테스트 세트, AUC = 1.0)

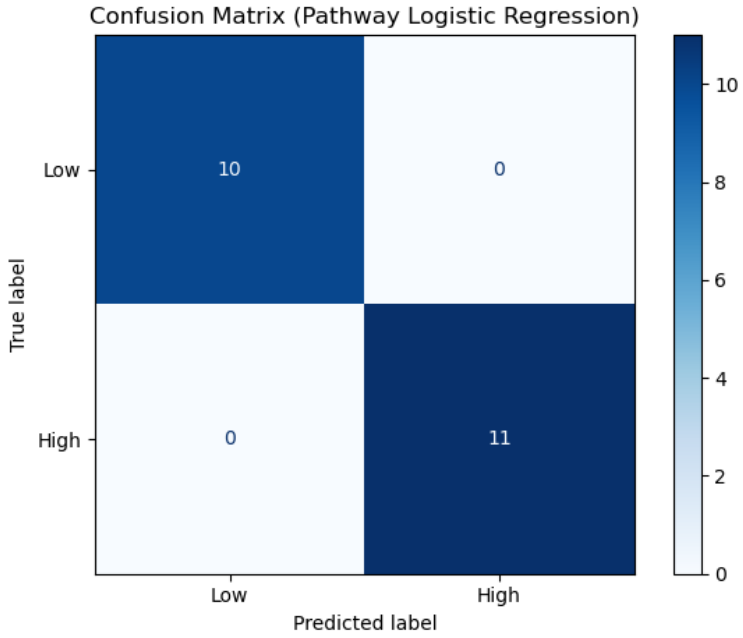

그림 5. Pathway 기반 Logistic Regression의 혼동행렬 (테스트 세트)

분류 성능을 보면, 먼저 pathway feature를 입력으로 한 Logistic Regression에서 단일 80 대 20 분할에서는 테스트 세트에 대해 정확도 1.0, ROC AUC 1.0이라는 매우 높은 값이 나왔다. 이 분할에서는 네 개의 경로 점수만으로 고생산과 저생산 샘플을 완벽히 맞출 수 있었다. 그러나 샘플 수가 많지 않고 분할이 한 번뿐이기 때문에 이 결과만으로 일반화 성능을 판단하기는 어렵다. 이 단일 분할에서의 분류 과정을 ROC 곡선과 혼동행렬로 나타내면 그림 4와 그림 5와 같다. 그림 4에서 ROC 곡선은 거의 좌측 상단 모서리를 따라가며 AUC가 1.0에 가까운 이상적인 형태를 보인다. 그림 5의 혼동행렬에서도 저생산과 고생산 샘플이 각각 하나도 틀리지 않고 올바른 클래스에 배정되어 있어, 이 분할에서는 네 개의 pathway 점수만으로 두 그룹을 완전히 구분할 수 있었음을 확인할 수 있다.

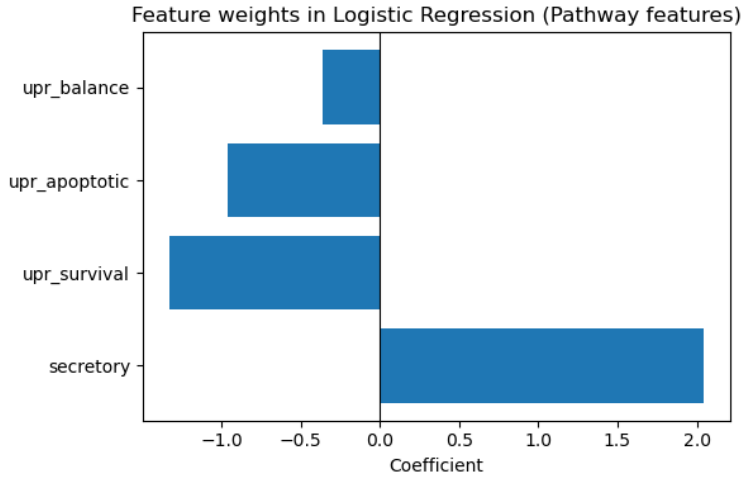

그림 6. Pathway feature 네 개에 대해 학습된 Logistic Regression 계수 (특징별 가중치 비교)

Logistic Regression에서 학습된 계수는 그림 6에 정리하였다. secretory 점수의 계수가 가장 큰 양수 값을 가져 고생산 방향으로 강하게 기여하는 특징으로 나타났고, upr_survival의 계수도 양수여서 프로 생존 UPR 신호가 강할수록 고생산 쪽으로 분류될 가능성이 커졌다. 반대로 upr_apoptotic의 계수는 음수 값으로 학습되어, 세포사멸 관련 UPR 신호가 강한 샘플일수록 저생산 그룹으로 분류되는 경향을 보였다. upr_balance 계수는 상대적으로 크기가 작았지만 양수 방향이어서, 전체적으로 "생존 UPR + 분비 경로 활성" 조합이 생산성 증가와 함께 나타난다는 해석과 잘 맞는다.

이를 보완하기 위해 수행한 5-fold 교차 검증에서는 fold 별 정확도가 약 0.75에서 0.90 사이에 분포했고, 평균 정확도는 약 0.81, 표준편차는 약 0.06 수준이었다. ROC AUC는 fold 별로 0.77에서 0.92 사이였고, 평균 AUC는 약 0.85, 표준편차는 약 0.06이었다. train test split을 random_state 10개에 대해 반복했을 때도 평균 정확도는 약 0.81, 표준편차는 약 0.05였으며, 평균 AUC는 약 0.88, 표준편차는 약 0.07이었다. 이러한 교차 검증 결과는 표 2에 정리하였다. 또 pathway 기반 Logistic Regression, latent 기반 Logistic Regression, latent 기반 MLP, 그리고 Perceptron을 한눈에 비교하면 표 3과 같으며, 대부분의 설정에서 정확도는 약 0.8, AUC는 0.85 이상 수준을 유지한다. 이를 통해 이번 분석이 특정 분할에만 우연히 맞아떨어진 것이 아니라, 실제로 생산성과 관련된 안정적인 패턴을 포착하고 있음을 다시 확인할 수 있다.

모델	입력 feature	평가 데이터	Accuracy	ROC AUC
Pathway Logistic Regression	UPR/secretory pathway 4 개 점수	테스트셋	1.00	1.00
Latent Logistic Regression	Autoencoder latent 2 차원	테스트셋	0.81	0.92
Latent MLP	Autoencoder latent 2 차원	검증셋	0.81	0.95
Perceptron	UPR/secretory pathway 4 개 점수	테스트셋	0.71	– (미측정)

표 3. Pathway feature와 Autoencoder latent를 사용한 주요 분류 모델들의 테스트 성능 비교 (Accuracy 및 ROC AUC)

Autoencoder latent를 입력으로 한 Logistic Regression에서도 테스트 세트에서 정확도는 약 0.81 정도였고 ROC AUC는 약 0.92 수준이었다. 이는 pathway feature를 사용한 모델과 비슷하거나 AUC 측면에서는 조금 더 좋은 성능이다. latent 공간을 입력으로 한 MLP에서는 검증 셋 기준으로 정확도 약 0.81, AUC 약 0.95 정도에서 수렴하였다. 초기 epoch에서 검증 정확도와 AUC가 빠르게 상승하고 이후에는 큰 변화 없이 안정된 곡선을 보였으며, train loss와 val loss가 함께 감소해서 심한 과적합은 관찰되지 않았다.

이러한 결과를 종합하면, UPR과 secretory 경로 점수만으로도 생산성 고저를 상당히 잘 예측할 수 있고, Autoencoder로 얻은 2차원 표현 역시 생산성과 관련성이 높은 정보를 잘 담고 있음을 알 수 있다. 여러 가지 분할과 검증 방법에서도 성능이 크게 떨어지지 않았다는 점에서, 이번 분석이 특정 분할에만 우연히 맞아떨어진 것이 아니라 어느 정도 일반화 가능한 패턴을 포착했다고 볼 수 있다.

5. Significance

이번 분석은 공공 CHO 전사체 데이터에서 나온 정보를 바탕으로 UPR과 secretory pathway의 상태를 간단한 점수 몇 개로 정리하고, 이 점수들이 CHO 세포의 단위 생산성과 실제로 어떤 관계를 가지는지 확인했다는 점에서 의미가 있다. upr_balance와 secretory 점수만으로도 고생산과 저생산 샘플을 약 80퍼센트 수준의 정확도로 구분할 수 있었고, 여러 분할과 교차 검증에서도 이 수준의 성능이 유지되었다. 이는 소포체 스트레스 반응의 방향성과 분비 경로의 활성 정도가 실제 생산성과 상당히 밀접하게 연관되어 있을 수 있음을 뒷받침한다.

pathway feature는 단 네 개의 숫자만으로 구성되어 있으면서 각 점수의 의미가 명확하기 때문에 공정 개발이나 세포주 스크리닝 과정에서 설명 가능한 지표로 활용하기 좋다. 예를 들어 특정 세포주나 배양 조건에서 upr_balance와 secretory 점수가 높게 나타난다면, 향후 고생산 세포주로 발전할 가능성이 높다고 판단할 수 있다. 또한 Autoencoder로 얻은 2차원 latent 표현이 별도 지도 정보 없이도 생산성 차이를 잘 반영했다는 점은, 향후 더 큰 RNA 시퀀싱 데이터나 다른 세포주 데이터에 representation learning을 적용해 품질이나 생산성을 예측하는 모델을 개발할 수 있는 가능성을 보여준다.

물론 몇 가지 한계도 존재한다. Qp가 측정되지 않은 샘플이 많아 실제로 학습에 사용한 샘플 수는 제한적이었고, 데이터가 마이크로어레이 기반이라는 점에서 노이즈와 probe와 gene symbol 매핑의 불확실성이 남아 있다. UPR과 secretory gene set 정의도 특정 문헌을 바탕으로 하였기 때문에 추후 다른 정의를 사용하면 결과가 다소 달

라질 수 있다. 그럼에도 불구하고 여러 분할과 검증 방법에서 성능이 안정적으로 유지되었다는 점에서, 이번 분석은 실제 생물학적 신호를 어느 정도 신뢰할 수 있는 수준으로 포착했다고 볼 수 있다. 특히 pathway 기반 Logistic Regression에서 secretory 점수와 upr_survival 계수가 양수, upr_apoptotic 계수가 음수로 나타난 점은, 단순한 성능 수치뿐 아니라 모델 해석 측면에서도 "분비 경로와 프로 생존 UPR의 강화, 프로 사멸 UPR의 억제"라는 직관적인 그림을 제공해 준다.

6. References

1. Clarke, C. et al. (2011) Predicting cell-specific productivity from CHO gene expression. *Journal of Biotechnology*, 151(2), 159–165. (GSE30321 기반 CHO 전사체–생산성 분석 논문)
2. Li, Z.-M. et al. (2022) Factors Affecting the Expression of Recombinant Protein and Improvement Strategies in Chinese Hamster Ovary Cells. *Frontiers in Bioengineering and Biotechnology*, 10, 880155. (CHO 세포에서 재조합 단백질 생산성에 영향을 주는 요인 및 개선 전략 리뷰)
3. Chakrabarti, A., Chen, A. W., Varner, J. D. (2011) A review of the mammalian unfolded protein response. *Biotechnology and Bioengineering*, 108(12), 2777–2793. (포유류 세포에서 UPR 메커니즘을 정리한 리뷰)
4. Prashad, K., Mehra, S. (2015) Dynamics of unfolded protein response in recombinant CHO cells. *Cytotechnology*, 67(2), 237–254. (재조합 CHO 세포에서 UPR 활성과 생산성/세포 운명 간 관계를 다룬 연구)