Gene expression

Friday, December 5, 2025 2:00 PM

Τι είναι τα χαρακτηριστικά του κάθε κυτάρου (περιλαμβάνονται > 17700 χαρ/κα για κάθε Cell line);

Τα χαρακτηριστικά είναι γονιδιακή έκφραση (gene expression) του κάθε κυττάρου,που μετράται με RNA-seq τεχνολογία και εκφράζεται σε TPM

RNA-seq(RNA Sequencing):Αναλυτικη μέθοδος που μετρά την ποσότητα mRNA σε ένα δείγμα.

TPM (Transcripts per million): Κανονικοποιημένη μονάδα που δείχνει πόσα αντιγράμμα mRNA ενός γονιδίου υπάρχουν ανά εκατομμύριο μεταγράφων

GDSC1

Dataset Description: Genomics in Drug Sensitivity in Cancer (GDSC) is a resource for therapeutic biomarker discovery in cancer cells. It contains wet lab IC50 for 100s of drugs in 1000 cancer cell lines. In this dataset, we use RMD normalized gene expression for cancer lines and SMILES for drugs. Y is the log normalized IC50. This is the version 1 of GDSC.

Task Description: Regression. Given the gene expression of cell lines and the SMILES of drug, predict the drug sensitivity level.

Dataset Statistics: 177,310 pairs, 958 cancer cells and 208 drugs

RNA: Ζωτικής σημασίας μόριο στα κύτταρα, που μεταφέρει γενετικές οδηγίες από το DNA για τη σύνθεση πρωτεϊνών, αλλά και παίζει ρόλο στη ρύθμιση των γονιδίων και αποτελεί γενετικό υλικό σε ορισμένους ιούς.
mRNA: Μεταφέρει τον κώδικα από το DNA (πυρήνα) στα ριβοσώματα (κυτταρόπλασμα) για την παραγωγή πρωτεϊνών.

Πχ

Drug_ ID		Cell Line_ ID
Erlotinib	COCCOC1=C(C=C2C(=C1)C(=NC=N2)NC3=CC=CC(=C3)C#C...	MC-CAR

	1ο γονιδιο	2ο γονιδιο	3ο γονιδιο	...17.000	Y
Διάνυσμα 17.742 τιμών γονιδιακής έκφρασης σε μορφή $\log_2$(TPM+1).	3.23827250519154	2.98225419469807	10.235490		2.395685

Σε αυτο το παραδειγμα σημαινει οτι αμα παρουμε το ζευγαρι drug:Erlotinib με το κυττραρο:MC-CAR
Τοτε ,το πρωτο γονιδιο του κυτταρου MC-CAR εχει την τιμη :

TPM value = 3.23827250519154 (log2(TPM+1))

Αντίστροφος υπολογισμός:
TPM actual = 2^3.23827250519154 - 1
= 9.43 - 1 = 8.43 transcripts per million
Αρα το πρωτο γονιδιο του κυτταρου MC-CAR συνεισφέρει ~ 8.43 μεταγράφων ανά εκατομμύριο συνολικού mRNA

Ερμηνεία Τιμών TPM
Η τιμή 8.43 TPM για το γονίδιο A1BG στο κύτταρο MC-CAR σημαίνει:

Από κάθε 1,000,000 mRNA μόρια στο κύτταρο MC-CAR,
περίπου 8.43 αντιστοιχούν στο γονίδιο A1BG.
Βιολογικά:
Η χαμηλή τιμή 8.43 TPM στο MC-CAR υποδεικνύει ότι αυτό το γονίδιο
δεν παίζει σημαντικό ρόλο στον φαινότυπο αυτής της καρκινικής σειράς.

Σημασία για το μοντέλο:
Αυτή η αναπαράσταση επιτρέπει στο μοντέλο να:
1. Κατανοήσει τον κυτταρικό φαινότυπο από τα μοτίβα έκφρασης
2. Αναζητήσει συσχετίσεις μεταξύ γονιδίων-στόχων και φαρμάκων
3. Προβλέψει την απόκριση βάσει του συνδυασμού φαρμάκου-κυττάρου

Προκλήσεις

1) Πολύ μεγάλη διάσταση
   Curse of Dimensionality (17,700 features!)
   Λυσεις
   PCA , Autoencoder

2) Τα 17,700 γονίδια δεν είναι ανεξάρτητα.
   Πιθανη Λυση
   GNN οπως για τα drugs (γονιδια Nodes edges:protein-protein interactions

3) Πολλά γονίδια έχουν 0 έκφραση
   Πιθανη Λυση
   Αφαίρεση γονιδίων με πολύ χαμηλή διακύμανση