
Báo cáo: Pipeline Xử lý Dữ liệu cho Mạng Tương tác miRNA-Gene trong Ung thư Phổi (LUAD)

Ngày: 13 tháng 10, 2025

1. Nguồn Dữ liệu

Pipeline sử dụng các nguồn dữ liệu công khai sau:

- **Dữ liệu Biểu hiện (Expression Data):**
  - **Nguồn:** The Cancer Genome Atlas - Lung Adenocarcinoma (TCGA-LUAD).
  - **Cổng truy cập:**
    - **mRNA & CNV:** cBioPortal (Dữ liệu TCGA PanCancer Atlas).
    - **miRNA:** GDC Data Portal (Dữ liệu miRNA-Seq của LUAD).
- **Dữ liệu Tương tác (Interaction Data):**
  - **Tương tác đã Kiểm chứng: miRTarBase** (Trang tải về).
  - **Tương tác Dự đoán: TargetScan** (Trang tải về).

---

2. Pipeline Xử lý

Pipeline được chia thành các giai đoạn tự động hóa bằng các script Python.

Giai đoạn A: Thu thập & Tiền xử lý Dữ liệu Gốc

1. **Gộp Dữ liệu Biểu hiện miRNA (`scripts/merge_mirna_files.py`):**
   - **Đầu vào:** Khoảng 600 file định lượng miRNA riêng lẻ tải về từ GDC, file `manifest` và `metadata.json`.
   - **Xử lý:** Script tự động đọc tất cả các file, trích xuất giá trị `reads_per_million_miRNA_mapped`, ánh xạ tên file sang mã bệnh nhân (dạng `TCGA-XX-XXXX`), và gộp chúng thành một ma trận duy nhất.
   - **Đầu ra:** File `data/features/mirnas.tsv` (định dạng: hàng là miRNA, cột là bệnh nhân).
2. **Tiền xử lý Dữ liệu Tương tác (`scripts/preprocess_interaction_data.py`):**
   - **Đầu vào:** File `miRTarBase_MTI.csv` và `Predicted_Targets_Info.default_predictions.txt` thô.
   - **Xử lý:**
     - Đối với miRTarBase: Lọc các tương tác của người (`hsa`), chọn cột `miRNA` và `Target Gene`, đổi tên thành `mirna_id`, `gene_id`.
     - Đối với TargetScan: Chọn cột `miR Family` và `Gene Symbol`, đổi tên thành `mirna_id`, `gene_id`.
   - **Đầu ra:** Hai file sạch `data/processed/mirtarbase_processed.csv` và `data/processed/targetscan_processed.csv` (sử dụng Hugo Symbol).

Giai đoạn B: Tính toán Trọng số & Tạo File Cuối cùng (`scripts/build_mirna_gene_edges.py`)

- **Đầu vào:** Các file biểu hiện và tương tác đã được chuẩn hóa Ensembl ID.
- **Xử lý:**
    1. Tải dữ liệu biểu hiện mRNA và miRNA đã được đồng bộ hóa bệnh nhân.
    2. Tải danh sách các tương tác "ứng viên" đã được xử lý.
    3. Lặp qua từng cặp tương tác:
        - Tính **hệ số tương quan Pearson (`r`)** và **`p-value`** giữa vector biểu hiện của miRNA và gene trên tất cả bệnh nhân.
        - **Lọc:** Giữ lại các cặp có tương quan nghịch (`r < 0`) và có ý nghĩa thống kê (`p-value < 0.05`).
        - **Gán Trọng số:** `weight = abs(r)`.
        - **Tăng cường:** Thêm một "điểm thưởng" nhỏ (`+ 0.1`) cho các tương tác có trong miRTarBase.
- **Đầu ra:** File cuối cùng **`data/edges/gene_mirna_ensembl.csv`**.

Giai đoạn C: Chuyển đổi Định danh Gene (`scripts/convert_gene_ids.py`)

- **Mục tiêu:** Chuẩn hóa tất cả các định danh gene về định dạng **Ensembl ID (`ENSG...`)**.

---

3. Cơ sở Khoa học & Tài liệu Tham khảo

Pipeline này được xây dựng dựa trên các phương pháp đã được công bố trong các nghiên cứu về tin sinh học và ung thư. Việc lựa chọn các tài liệu tham khảo dưới đây nhằm mục đích bảo vệ và củng cố tính khoa học cho các quyết định phương pháp luận quan trọng trong pipeline.

- **Về Chiến lược Tích hợp Dữ liệu trong bối cảnh LUAD:**
    - **Trích dẫn:** *Screening and Biological Function Analysis of miRNA and mRNA Related to Lung Adenocarcinoma Based on Bioinformatics Technology*. (PMC9452934, 2022).
    - **Lý do trích dẫn:** Bài báo này cung cấp một **khuôn khổ phương pháp luận tổng thể** tương tự như pipeline của chúng ta. Các tác giả cũng sử dụng dữ liệu TCGA-LUAD, xác định các phân tử quan trọng, và sau đó xây dựng mạng lưới tương tác bằng cách kết hợp các CSDL dự đoán (TargetScan) và kiểm chứng (miRTarBase). Việc trích dẫn bài báo này giúp khẳng định rằng chiến lược chung của chúng ta là hợp lệ và phù hợp với các thực hành tiêu chuẩn trong lĩnh vực.
- **Về Phương pháp Gán Trọng số Dựa trên Dữ liệu (Data-Driven):**
    - **Trích dẫn:** *Integrative analysis of miRNA and mRNA sequencing data reveals potential regulatory mechanisms of ACE2 and TMPRSS2*. (PLOS One, 2020).
    - **Lý do trích dẫn:** Đây là tài liệu tham khảo **cốt lõi** để bảo vệ phương pháp gán trọng số (`weight`). Thay vì chọn các giá trị tùy ý, pipeline của chúng ta áp dụng phương pháp được mô tả trong bài báo này: sử dụng **phân tích tương quan biểu hiện** trên dữ liệu thực tế (TCGA) để xác thực các tương tác và định lượng sức mạnh của chúng. Việc trích dẫn này chứng minh rằng cách chúng ta tính

toán `weight` không phải là tự phát mà dựa trên một phương pháp luận đã được công bố và bình duyệt, liên kết trực tiếp trọng số của cạnh với hoạt tính sinh học trong mô bệnh.

- **Về Việc Lựa chọn và Kết hợp các Cơ sở dữ liệu Tương tác:**
    - **Trích dẫn:** *MiRGraph: A transformer-based feature learning approach to identify microRNA-target interactions*. (bioRxiv, 2023).
    - **Lý do trích dẫn:** Bài báo này, mặc dù có mục tiêu khác, nhưng đã sử dụng chính xác hai nguồn dữ liệu tương tác mà chúng ta đã chọn: **TargetScan** và **miRTarBase**. Nó đại diện cho một nghiên cứu hiện đại trong lĩnh vực học máy trên đồ thị, khẳng định rằng việc kết hợp một CSDL dự đoán có độ bao phủ rộng (TargetScan) với một CSDL kiểm chứng có độ tin cậy cao (miRTarBase) là một **cách tiếp cận mạnh mẽ và hợp lý** để xây dựng một mạng lưới tương tác chất lượng cao làm đầu vào cho mô hình GNN.
    -

2. **Kết hợp CSDL Dự đoán và Kiểm chứng:**
    - Việc kết hợp **TargetScan** (độ bao phủ rộng) và **miRTarBase** (độ tin cậy cao) là một phương pháp phổ biến để tạo ra một mạng lưới vừa toàn diện vừa chính xác.
    - **Tham khảo:** *MiRGraph: A transformer-based feature learning approach...* (bioRxiv, 2023). Bài báo này đã sử dụng cả hai CSDL này làm nền tảng để xây dựng mô hình GNN.

---

4. Kết quả Đầu ra

Sản phẩm cuối cùng của pipeline này là file **data/edges/gene_mirna_ensembl.csv**, một file CSV có 3 cột:

- mirna_id: Định danh của miRNA.
- gene_id: Định danh Ensembl của gene mục tiêu.
- weight: Trọng số của tương tác (từ 0 đến 1), phản ánh độ mạnh của mối quan hệ điều hòa nghịch trong dữ liệu LUAD.

5. Prompt giới thiệu:

Bối cảnh Project:

Đây là một project tin sinh học nhằm phân nhóm bệnh nhân ung thư phổi (LUAD) và dự đoán thời gian sống sót bằng cách sử dụng Mạng Neural Đồ thị Dị thể (Heterogeneous Graph Neural Network - GNN). Mô hình yêu cầu dữ liệu đầu vào được cấu trúc cẩn thận, đặc biệt là các file định nghĩa các cạnh (tương tác) trong đồ thị.

Nhiệm vụ đã thực hiện:

Tôi đã làm việc với một phiên bản trước của Gemini để xây dựng một pipeline xử lý dữ liệu hoàn chỉnh từ đầu đến cuối. Mục tiêu chính là tạo ra file `data/edges/gene_mirna_ensembl.csv`, một thành phần cốt lõi của đồ thị. Pipeline này được thiết kế để đảm bảo tính chính xác, nhất quán và có cơ sở khoa học.

Pipeline đã xây dựng bao gồm các bước sau:

1. **Thu thập dữ liệu thô:**
    - Dữ liệu biểu hiện gen (mRNA) từ **cBioPortal** (bộ TCGA-LUAD, PanCancer Atlas).
    - Dữ liệu biểu hiện miRNA từ **GDC Data Portal**, bao gồm việc gộp ~600 file riêng lẻ thành một ma trận duy nhất bằng script `scripts/merge_mirna_files.py`.
    - Dữ liệu tương tác đã kiểm chứng từ **miRTarBase** và tương tác dự đoán từ **TargetScan**.
2. **Tiền xử lý và Chuẩn hóa (`scripts/preprocess_interaction_data.py`):**
    - Tự động lọc, chọn cột và đổi tên các file tương tác thô thành các file `.csv` sạch sẽ, sẵn sàng cho phân tích.
3. **Thống nhất Định danh Gene (`scripts/convert_gene_ids.py`):**
    - Để đảm bảo tính nhất quán, một script đã được tạo ra để chuyển đổi tất cả các định danh gene (Hugo Symbol) trong tất cả các file liên quan sang định dạng **Ensembl ID (`ENSG...`)** bằng cách sử dụng thư viện `mygene`. Script này cũng xử lý các tên gene cũ hoặc tên đồng nghĩa (alias).
4. **Tính toán Trọng số Cạnh một cách Khoa học (`scripts/build_mirna_gene_edges.py`):**
    - Đây là bước cốt lõi của pipeline. Thay vì gán các giá trị trọng số tùy ý, chúng tôi đã triển khai một phương pháp **data-driven**:
        - Với mỗi cặp tương tác (miRNA, gene) ứng viên, script tính toán **hệ số tương quan Pearson** giữa vector biểu hiện của chúng trên ~500 bệnh nhân LUAD.
        - Các tương tác được lọc dựa trên **tương quan nghịch (`r < 0`)** và **ý nghĩa thống kê (`p-value < 0.05`)**.
        - Trọng số (`weight`) của mỗi cạnh được gán bằng giá trị tuyệt đối của hệ số tương quan (`abs(r)`), phản ánh sức mạnh của mối quan hệ điều hòa trong chính dữ liệu bệnh.
        - Các tương tác đã được kiểm chứng (từ miRTarBase) được cộng một điểm thưởng nhỏ.

Tóm lại: Toàn bộ dữ liệu đầu vào cho project, đặc biệt là mạng lưới tương tác miRNA-gene, đã được xây dựng một cách cẩn thận thông qua một pipeline tự động, dựa trên các phương pháp đã được công bố. Hãy xem xét các script trong thư mục `scripts/` để hiểu rõ hơn về logic xử lý.