

ترجمه فصل اول: چکیده + مقدمه

(From *More Is Better: Recent Progress in Multi-Omics Data Integration Methods*)

چکیده (Abstract)

یکپارچه‌سازی داده‌های چند‌آمیک (Multi-omics) یکی از مهم‌ترین چالش‌ها در دوران پزشکی دقیق (Precision Medicine) است.

با رشد سریع فناوری‌های *High-throughput*, حجم بسیار بزرگی از داده‌ها در دسترس پژوهشگران قرار گرفته و این امکان فراهم شده است که تحلیل‌های عمیق‌تر و هدفمندتری انجام شود.

برای بهبود پیش‌بینی پیامدهای بالینی (مانند تشخیص، پیش‌آگهی و بقا)، مجموعه‌ای بزرگ از ابزارها و نرم‌افزارهای جدید توسعه داده شده‌اند.

این مقاله مروری، پیشرفت‌های انجام‌شده در زمینه یکپارچه‌سازی Multi-omics و ابزارهای جامع موجود را بررسی می‌کند. در پایان مقاله نیز روش‌های یکپارچه‌سازی داده‌ها برای پیش‌بینی بقا بیمار بررسی می‌شود.

کلمات کلیدی

multi-omics, یکپارچه‌سازی, پیش‌آگهی, پیش‌بینی, پژوهشگری دقیق، یادگیری بدون نظارت

مقدمه (Introduction)

دوران جدید «پزشکی شخصی» آغاز شده است؛ دورانی که در آن با توجه به ویژگی‌های فردی هر بیمار، درمان و مدیریت پزشکی مخصوص همان فرد طراحی می‌شود.

در این نوع درمان، نه تنها اطلاعات بالینی، بلکه اطلاعات مولکولی بدن فرد نیز بررسی و برای تصمیم‌گیری استفاده می‌شود.

تحقیقات سرطان که فقط بر یک نوع داده آمیک (مثلاً فقط بیان ژن) تکیه دارند، اطلاعات بسیار محدودی درباره علت ایجاد سرطان و پیشرفت تومور ارائه می‌دهند.

به همین دلیل، تلاش‌های زیادی برای تولید مجموعه داده‌های گستردۀ و چندلایه انجام شده است.

بزرگ‌ترین پروژه در این زمینه TCGA است که از بیش از ۱۰ هزار بیمار اطلاعات جمع کرده و داده‌هایی شامل:

توالی‌بایی اگزوم •

DNA (CNV) تغییرات تعداد نسخه •

DNA متیلاسیون •

بیان ژن •

• miRNA بیان

اطلاعات بالینی مثل نژاد، مرحله تومور، عود بیماری، پاسخ به درمان را فراهم کرده است.

نسخه جهانی تر این پروژه کنسرسیوم جهانی ژنوم سرطان (ICGC) است که برای بیش از ۵۰ نوع سرطان اطلاعات ژنومی، ترانسکریپتومی، اپیژنتیکی و جهش‌های جسمی را جمع‌آوری کرده است.

این مجموعه‌های عظیم داده، فرصت بی‌نظیری برای کشف امضاهای مولکولی پنهان مرتبط با سرطان فراهم کرده‌اند.

اما چالش بزرگ اینجاست:

🔍 ژنوم انسان یک سیستم فوق العاده پیچیده است

و لایه‌های مختلف داده‌های امیک با یکدیگر در تعامل ووابستگی هستند.

هر نوع داده تنها یک «زاویه دید» به سیستم می‌دهد، اما هیچ کدام به تنها یکی کافی نیستند.

به همین دلیل، وقتی چند نوع داده را با هم ترکیب می‌کنیم، پیدا کردن الگوهای منسجم و معنادار کار بسیار دشواری می‌شود.
به خصوص برای پیش‌بینی ویژگی‌های مهم بالینی مثل:

• سالم/سرطانی

• مرحله تومور

• بقای بیمار

بنابراین، هدف این مقاله این است که:

✓ چالش‌های مهم در توسعه روش‌های جدید یکپارچه‌سازی Multi-omics را بررسی کند

✓ ابزارها و الگوریتم‌های مطرح در این حوزه را معرفی کند

✓ و در نهایت روش‌های یکپارچه‌سازی برای پیش‌بینی بقا را توضیح دهد.

📘 فصل دوم: روش‌های بدون نظارت (Unsupervised Data Integration)

روش‌های بدون نظارت به مجموعه‌ای از روش‌ها گفته می‌شود که بدون داشتن برچسب یا پاسخ بالینی (مثل سالم/سرطانی، زنده‌مانی، مرحله بیماری) فقط بر اساس شباهت‌ها و الگوهای موجود در داده‌ها عمل می‌کنند.

هدف اصلی این روش‌ها:

🔍 پیدا کردن الگوهای پنهان، خوش‌ها(Cluster) ها و ساختارهای مشترک بین چند نوع داده امیک

در مقاله، روش‌های بدون نظارت به چند گروه تقسیم شده‌اند:

۱) (روش‌های ماتریس‌فکتورگیری) Matrix Factorization Methods

Joint Non-Negative Matrix Factorization (NMF) ◆

ساده‌ترین و مشهورترین روش در این دسته است.

ایده کلی NMF این است که:

- یک ماتریس بزرگ داده مانند ماتریس بیان ژن یا متیلاسیون DNA
- به دو ماتریس کوچک‌تر تجزیه می‌شود:

$$X \approx W \times H$$

که:

- X: ماتریس داده (مثلاً بیان ژن)
- W: الگوهای مشترک (فاکتورهای زیستی مشترک بین داده‌ها)
- H: میزان مشارکت هر نمونه در آن الگوها

: NMF ویژگی مهم

☞ همه عناصر ماتریس‌ها غیرمنفی هستند، پس تفسیر زیستی راحت‌تر است (مثلاً فعالیت مسیرها همیشه ≤ 0).

مزایا:

- کشف الگوهای مشترک بین چند نوع داده
- مناسب برای خوشبندی بیماران

معایب:

- محاسبات سنگین
- نیاز به نرمال‌سازی دقیق داده‌ها
- فقط برای داده‌های غیرمنفی قابل استفاده است

iCluster ◆

NMF شبیه Cluster است اما محدودیت غیرمنفی بودن را ندارد و انعطاف‌پذیرتر است.

ایده اصلی:

- یک «متغیر پنهان مشترک» بین همه داده‌ها وجود دارد
- این متغیر باعث شکل‌گیری بیماری یا زیرگروه‌های سرطان می‌شود

: iCluster فرمول کلی

$$X = W \times H + E$$

که:

- W : عوامل پنهان مشترک
- H : وزن هر داده
- E : نویز

ویژگی مهم:

از مجازات (L1) LASSO برای ایجاد سختگیری و انتخاب ویژگی استفاده می‌کند.

کاربرد اصلی:

- کشف زیرگروه‌های جدید سرطان
- ادغام چند لایه آمیک

iCluster+ 

.iCluster نسخه پیشرفته‌تر

: iCluster+ مزیت بزرگ

✓پشتیبانی از انواع مختلف داده‌ها:

- پیوسته
- دودوبی
- شمارشی
- طبقه‌ای

هر نوع داده با مدل مناسب خودش (لجستیک، خطی، پواسون...) تحلیل می‌شود .

به دلیل پیچیدگی زیاد:

- نیاز به پیش انتخاب ویژگی دارد
- محاسبات سنگین است

JIVE – Joint and Individual Variation Explained ◆

JIVE داده‌ها را به دو بخش تقسیم می‌کند:

1. بخش مشترک بین همه لایه‌های داده
2. بخش ویژه هر لایه (مثلًا متیلاسیون تنها در برخی ژن‌ها اثر دارد)

این روش:

- قسمت مشترک بین omics‌ها را استخراج می‌کند
- و قسمت منحصر به فرد هر omics را جدا نگه می‌دارد

کاربرد:

- فهمیدن اینکه کدام مسیرها در همه لایه‌ها هماهنگ تغییر کرده‌اند
- و کدام فقط در یک لایه خاص فعال‌اند

Joint Bayes Factor ◆

یک روش مبتنی بر بیز که هدفش:

کشف فاکتورهای مشترک و اختصاصی بین سطوح مختلف داده

ویژگی مهم:

- استفاده از توزیع Student-t برای تشویق به تنکشدن (Sparsity)
- مناسب برای داده‌های نویزی یا با حجم خیلی بالا

۲) روش‌های همبستگی محور و CCA

این دسته روش‌ها مثل CCA یا Sparse CCA به دنبال:

یافتن ترکیب‌هایی از ژن‌ها در هر لایه هستند که بیشترین همبستگی را با لایه دیگر دارند.

مثال‌ها:

ssCCA •

Sparse-group CCA •

این‌ها برای پیدا کردن لینک‌های میان بیان ژن و متیلاسیون یا بیان ژن و پاسخ دارویی بسیار مفیدند.

۳ (روش‌های مبتنی بر شبکه) (Network-Based Methods)

داده‌ها را به شبکه تبدیل می‌کنند

(گره‌ها = ژن یا بیمار، یال‌ها = شباهت یا تعامل)

سه روش مهم:

PARADIGM ◆

ترکیب چند نوع داده برای محاسبه فعالیت مسیرهای بیولوژیک

→ خروجی: فعالیت هر مسیر در هر بیمار

SNF – Similarity Network Fusion ◆

یکی از بهترین روش‌ها برای کشف زیرگروه‌های سرطان.

مراحل:

1. ساخت شبکه شباهت بیماران برای هر لایهٔ داده

2. ترکیب شبکه‌ها

3. تولید یک شبکه نهایی قوی‌تر و کم‌نوفر

نتیجه:

• خوشه‌بندی بیماران بسیار دقیق‌تر می‌شود.

• زیرگروه‌های جدید سرطان کشف می‌شوند.

Lemon-Tree ◆

برای کشف شبکه‌های ماژولی (module networks) استفاده می‌شود.

لایه‌ها را یکی‌یکی وارد مدل می‌کند و سپس شبکه نهایی را می‌سازد.

۴ (روش‌های چند-کرنلی) (Multiple Kernel Learning)

این روش‌ها:

- هر نوع داده را به صورت یک کرنل (Kernel) نمایش می‌دهند
- سپس کرنل‌ها را وزن‌دهی و ادغام می‌کنند

مثال:

rMKL-LPP

خروجی:

- خوشبندی
- کاهش ابعاد
- نمایش ساختار چند لایه‌ای داده

(۵) روش‌های چند مرحله‌ای (Multi-Step Analysis)

این روش‌ها لایه‌ها را جداگانه تحلیل کرده و در آخر نتایج را ترکیب می‌کنند.

مثال‌های مهم:

- CNAmet
- iPAC

این‌ها معمولاً برای پیدا کردن ژن‌هایی استفاده می‌شوند که:

- در CNV تغییر دارند
- در متیلاسیون تغییر دارند
- و در بیان ژن نیز تغییری مرتبط با آن دو مشاهده می‌شود

■ فصل سوم: روش‌های با نظارت (Supervised Data Integration)

این فصل درباره روش‌هایی است که داده‌های چندآمیک را همراه با برچسب یا اطلاعات بالینی تحلیل می‌کنند؛ یعنی مدل می‌داند چه چیزی باید پیش‌بینی کند، مثل:

- نوع سرطان
- پاسخ بیمار به درمان
- میزان بقا
- عود بیماری
- متاستاز

برخلاف روش‌های بدون نظارت که فقط الگوها را کشف می‌کردند، در روش‌های با نظارت هدف این است که:

از چند نوع داده مولکولی → برای ساخت یک مدل پیش‌بینی دقیق استفاده شود. 🔥

این روش‌ها در پژوهشی دقیق بسیار مهم‌اند.

۳.۱. روش‌های مبتنی بر یادگیری ماشین کلاسیک

در این بخش، روش‌هایی معرفی می‌شوند که از مدل‌های آماری یا یادگیری ماشین معمولی برای پیش‌بینی استفاده می‌کنند.

◆ (1) مدل‌های رگرسیونی (Regression Models)

LASSO و Elastic Net

برای انتخاب تعداد کمی از ژن‌های مهم، از جریمه‌های L1 و L2 استفاده می‌کنند.

این مدل‌ها روی داده‌های زیاد مؤثرند.

کاربرد:

- انتخاب ویژگی از چند نوع داده
- پیش‌بینی مرحله سرطان یا بقا

مزایای این مدل‌ها:

- قابل تفسیر

مناسب برای داده‌های با ویژگی‌های زیاد •

جلوگیری از بیش‌بازش(Overfitting) •

(یادگیری چندکرنلی) **Multiple Kernel Learning (۲◆)**

در این روش:

هر نوع omics → تبدیل به یک Kernel •

سپس کرنل‌ها ترکیب و وزن‌دهی می‌شوند •

کاربرد:

پیش‌بینی طبقه سرطان •

پیش‌بینی پاسخ به دارو •

مزیت:

هر لایه داده می‌تواند به شکل ریاضی متفاوتی مدل‌سازی شود •

نیاز به فروض ساده‌کننده ندارد •

۳.۲. روش‌های مبتنی بر شبکه (Network-Based Supervised Methods)

در این مدل‌ها، دانش زیستی مثل شبکه‌های پروتئین–پروتئین یا مسیرهای KEGG وارد تحلیل می‌شود.

NetGAIN و NetICS ◆

این ابزارها از شبکه‌های زیستی برای:

رتیبه‌بندی ژن‌های مهم •

پیدا کردن مسیرهای فعال •

اولویت‌بندی عوامل بیماری •

استفاده می‌کنند.

مزیت مهم آن‌ها:

ترکیب تعاملات ژنی → با داده‌های چندآمیک → برای پیش‌بینی بهتر بیماری

۳.۳. یادگیری عمیق (Deep Learning) در چندامیک

در سال‌های اخیر، یادگیری عمیق نتایج خیلی قوی تولید کرده است.

مقاله چند مدل مهم را معرفی می‌کند:

۱) مدل‌های مبتنی بر Autoencoder ◆

:Autoencoder

- داده‌ها را برای هر omics جداگانه فشرده می‌کنند
- سپس نمایش فشرده را ادغام می‌کنند
- و مدل نهایی برای طبقه‌بندی یا پیش‌بینی بقا استفاده می‌شود

ویژگی‌ها:

- کشف ویژگی‌های پنهان
- استخراج الگوهای مشترک بین omics ها

۲) روش‌های چندشاخه‌ای (Multi-branch Neural Networks) ◆

در این روش:

- هر نوع omics یک «شاخه جدا» از شبکه عصبی دارد
- در انتهای خروجی شاخه‌ها ترکیب و وارد لایه تصمیم‌گیری می‌شوند

کاربرد:

- تشخیص نوع تومور
- پیش‌بینی بقا
- طبقه‌بندی بیماری‌ها

مزیت:

- هر omics ویژگی‌های مخصوص خودش را یاد می‌گیرد
- در مرحله نهایی این ویژگی‌ها با هم ادغام می‌شوند

۳) شبکه‌های توجه (Attention Networks)

این مدل‌ها وزن بیشتری روی لایه‌هایی می‌گذارند که اطلاعات مهم‌تری برای پیش‌بینی دارند.

مثال:

- بیان زن اهمیت بیشتری دارد
- یا
- متیلاسیون برای یک نوع سرطان خاص نقش پررنگ دارد

شبکه‌های Attention کمک می‌کنند مدل هوشمندانه تشخیص دهد کدام omics مهم‌تر است.

۴) مدل‌های ترکیبی Deep + Network Biology

مثال:

- گراف کانولوشنال نتورک (GCN)
- مدل‌های یادگیری عمیق مبتنی بر شبکه‌های پروتئینی

این‌ها ساختار روابط زن‌ها را نیز در مدل لحاظ می‌کنند.

کاربرد:

- پیش‌بینی مسیرهای فعال
- پیش‌بینی زیرگروه‌های سرطان
- کشف بیومارکر

۳.۴. روش‌های دانش‌محور (Knowledge-Guided)

در این روش‌ها از اطلاعات بیولوژیکی قبلی استفاده می‌شود:

- مسیرهای KEGG
- شبکه تعامل پروتئینی
- دیتابیس‌های ژنتیکی
- Signature‌های شناخته شده

این دانش باعث می‌شود:

مدل دقیق‌تر آموزش ببیند و تفسیرپذیری آن بالا بود.

مثال‌ها:

- DIGEST •
- PARADIGM Supervised Version •
- (Evaluation) ۳.۵ ارزیابی عملکرد

برای ارزیابی مدل‌های با نظارت از معیارهای زیر استفاده می‌شود:

- AUC (Area Under Curve) •
- Accuracy •
- F1-score •
- C-index برای مدل‌های بقا •

مقاله تأکید می‌کند که:

ادغام چندآمیک تقریباً همیشه دقت بیشتری نسبت به استفاده تکآمیک دارد.

به خصوص در:

- تشخیص سرطان •
- پیش‌بینی زنده‌مانی •
- پاسخ به دارو •
- کشف زیرگروه‌های تومور •

جمع‌بندی فصل سوم

روش‌های با نظارت:

- از برچسب‌های بیماری یا بقا استفاده می‌کنند •
- می‌توانند از شبکه، یادگیری ماشین کلاسیک، یا یادگیری عمیق استفاده کنند •
- دقت بسیار بیشتری در پیش‌بینی‌های بالینی دارند •

■ فصل چهارم: روش‌های نیمه‌ناظارت‌شده (Semi-Supervised Data Integration)

این روش‌ها ترکیبی از دو حالت قبلی هستند:

- بخشی از داده‌ها برچسب دارند (مثلاً برای برخی بیماران نوع سرطان یا میزان بقا مشخص است)
- بخشی از داده‌ها برچسب ندارند (معمولًاً داده‌های زیاد omics بدون اطلاعات بالینی دقیق)

روش‌های نیمه‌ناظارت‌شده برای زمانی مناسب‌اند که:

برچسب‌ها کمیاب هستند، اما داده‌های خام زیادند. 

این وضعیت در پزشکی و بهخصوص در پژوهش‌های بزرگ مانند TCGA بسیار رایج است.

چرا روش‌های نیمه‌ناظارت‌شده مهم هستند؟ 

زیرا:

- گرفتن اطلاعات بالینی دقیق برای هر بیمار سخت، گران و زمان‌بر است
- اما گرفتن داده‌های مولکولی (omics) برای تعداد زیاد بیمار ساده‌تر است

پس بهتر است بتوانیم:

از داده‌های بدون برچسب هم برای بهبود مدل استفاده کنیم 

این کار باعث می‌شود:

- پیش‌بینی‌های دقیق‌تر شود
- مدل از ساختار طبیعی داده‌ها بهتر استفاده کند
- تعمیم‌پذیری (Generalization) بالاتر باشد

۴.۱. روش‌های مبتنی بر گراف (Graph-Based Semi-Supervised Learning)

متداول‌ترین دسته روش‌ها در این فصل.

ایده اصلی این مدل‌ها:

۱. ابتدا یک گراف شباهت بین بیماران ساخته می‌شود

(مثل روش SNF اما نیمه‌ناظارت شده)

۲. بیماران با برچسب مشخص مثل نقاط راهنمای (Seed) عمل می‌کنند

۳. برچسب‌ها از بیماران شناخته شده → به بیماران ناشناخته

روی گراف پخش می‌شود

(Label Propagation (به این روش می‌گویند

این کار باعث می‌شود:

• مدل از ساختار محلی داده‌ها استفاده کند

• نمونه‌های نزدیک به هم برچسب مشابه بگیرند

◆ (Label Propagation) انتشار برچسب

یکی از ساده‌ترین و قوی‌ترین روش‌ها.

مراحل:

۱. ساخت گراف شباهت بین بیماران

۲. نگه داشتن برچسب‌های بیماران شناخته شده

۳. گسترش برچسب‌ها در داخل گراف

ویژگی مهم:

• فقط از روابط طبیعی بین داده‌ها استفاده می‌کند

• مناسب برای پژوهه‌هایی که تعداد بیماران برچسب‌دار کم است

کاربرد:

• تشخیص یا طبقه‌بندی نوع سرطان

• گروه‌بندی بیماران

• تکمیل داده‌های گمشده

٤.٢ روش‌های مبتنی بر یادگیری چندمنظوره(Multi-Task Learning)

در این روش‌ها:

- چند نوع omics → به صورت «چند وظیفه مرتبط» مدل‌سازی می‌شود
مثالاً:
- پیش‌بینی نوع تومور
- پیش‌بینی بقا
- شناسایی زیرگروه‌های مولکولی

و همه وظایف از اطلاعات هم‌دیگر یاد می‌گیرند.

مزیت:

🔥 هنگامی که یک وظیفه برچسب کمی دارد، از وظایف دیگر کمک می‌گیرد.

۴.۳. مدل‌های مبتنی بر ماتریس و فاکتورگیری نیمه‌نظرارت شده

این مدل‌ها مثل NMF و iCluster هستند اما:

- بخشی از فاکتورهای پنهان را با برچسب‌ها هماهنگ می‌کنند
- بخشی دیگر بدون نظرارت یاد گرفته می‌شود

نتیجه:

- استخراج الگوهای بین‌omics ها
- استفاده از برچسب‌های بالینی برای هدایت یادگیری

مثال‌ها (در مقاله اشاره عمومی شده):

- Semi-Supervised NMF
- Joint Latent Variable Models

۴.۴. روش‌های Deep Semi-Supervised

ترکیب یادگیری عمیق + انتشار برچسب + دوره پیش‌آموزش بدون نظرارت.

این مدل‌ها معمولاً شامل:

۱) مرحله بدون ناظارت

Autoencoder ها ویژگی‌های چندامیک را یاد می‌گیرند.

۲) مرحله ناظارت شده

لایه‌های بالا برای طبقه‌بندی یا پیش‌بینی بقا آموزش می‌بینند.

۳) مرحله نیمه ناظارت شده

برچسب‌های ناقص با روش‌هایی مثل زیر تقویت می‌شوند:

- pseudo-labeling
- consistency training
- graph regularization

🔥 مزیت کلیدی مدل‌های نیمه ناظارت شده

⭐ دقیقی نزدیک به مدل‌های با ناظارت

📌 اما با نیاز بسیار کمتر به داده‌های برچسب‌دار

به همین دلیل برای پژوهه‌های بزرگ مقیاس پزشکی مناسب‌اند.

✳️ ۴) جمع‌بندی فصل چهارم

روش‌های نیمه ناظارت شده:

- ترکیبی از یادگیری با ناظارت و بدون ناظارت
- از گراف‌ها و شباهت بین بیماران استفاده می‌کنند
- برای زمانی مناسب‌اند که برچسب‌گذاری محدود است
- دقت مدل را با هزینه کم بالا می‌برند

📘 فصل پنجم: یکپارچه‌سازی داده‌های چندامیک برای پیش‌بینی بقا (Survival Prediction Integration)

(ترجمه کامل متن موجود از بخش MKGI و CoxPath در PDF)

بخش ۵ – روش‌های پیشرفته برای پیش‌بینی بقا

هدف این فصل، معرفی دو مدل مهم برای پیش‌بینی بقا با استفاده از داده‌های چند‌امیک است:

.CoxPath و MKGI

هر دو روش تلاش می‌کنند اطلاعات ژنومی چندلایه را با اطلاعات بالینی ترکیب کنند تا پیش‌بینی بقا بیماران دقیق‌تر شود.

MKGI (Metadimensional Knowledge-Guided Genomic Interactions) (مدل)

هدف MKGI چیست؟

این مدل برای کشف ژن‌های بحرانی و ارتباط بین مسیرهای بیولوژیک طراحی شده است؛ به‌طوری‌که:

- نقش لایه‌های مختلف omics
- همراه با دانش زیستی موجود
- برای پیش‌بینی بقا

به شکل یک «مدل تکاملی» ادغام می‌شود.

چگونه MKGI کار می‌کند؟

این روش سه مرحله دارد:

مرحله ۱: انتخاب ژن‌های مرتبط با بقا

در این مرحله، ژن‌هایی انتخاب می‌شوند که:

- در مسیرهای KEGG حضور دارند
- با بقا ارتباط آماری دارند
- تغییرات omics (بیان، متیلاسیون، CNV) در آن‌ها دیده می‌شود

مرحله ۲: محاسبه ارتباط بین مسیرها

مسیرهای بیولوژیک به کمک سه نوع شباهت ارزیابی می‌شوند:

1. شباهت در توالی ژن‌ها
2. شباهت عملکردی ژن‌ها
3. شباهت ساختاری در سطح شبکه

سپس مسیرهای مشابه در «گام‌های تکاملی» با هم ادغام می‌شوند.

مرحله ۳: الگوریتم تکاملی برای انتخاب بهترین ترکیب

مدل با استفاده از یک الگوریتم تکاملی تلاش می‌کند:

- بهترین ترکیب ژن‌ها
 - بهترین وزن برای لایه‌های مختلف omics
 - بهترین ساختار مسیرهای بیولوژیک
- را پیدا کند تا بقا را با بیشترین دقت پیش‌بینی کند.

MKGI نتیجه چیست؟

- انتخاب دقیق ژن‌های کلیدی
- درک بهتر نقش تعاملات بین مسیرهای بیولوژیکی
- افزایش چشمگیر دقت پیش‌بینی بقای بیماران

بر اساس متن CoxPath، MKGI دقت بهتری نسبت به CoxPath نشان داده است.

CoxPath (۲) مدل

CoxPath هدف چیست؟

Cox Proportional Hazards مدل کلاسیک Regularized CoxPath نسخه پیشرفته و است.

ایده اصلی:

- ابتدا داده‌های Multi-omics را ادغام می‌کند
- سپس با جریمه L1 و L2 (LASSO) یک مدل Cox می‌سازد
- و بهترین مسیر را برای انتخاب ژن‌ها پیدا می‌کند

CoxPath چگونه کار می‌کند؟

۱. شناسایی روابط بین omics ها

ابتدا ارتباط بین انواع داده‌ها را پیدا می‌کند:

• بیان ژن

• متیلاسیون

• CNV

• جهش‌ها

این مرحله منجر به استخراج ویژگی‌های مشترک بین لایه‌ها می‌شود.

۲. مدل‌سازی با جریمه Cox با L1

سپس مدل Cox با LASSO ساخته می‌شود:

• ژن‌های غیرمهم حذف می‌شوند

• ژن‌های تأثیرگذار روی بقا باقی می‌مانند

• از بیش‌برازش جلوگیری می‌شود

۳. مسیر Cox (Cox Path)

این بخش شبیه مسیر Regularization است:

• مقدار جریمه L را تغییر می‌دهند

• مسیر تغییر وزن ژن‌ها بررسی می‌شود

• بهترین نقطه (Optimal Path) انتخاب می‌شود

CoxPath نتیجه ★

• انتخاب سریع ژن‌های مهم

• مناسب برای داده‌های بسیار پر حجم

• اما نسبت به MKGI کمتر «دانش محور» است

• دقیق‌تر از MKGI طبق (PDF)

CoxPath و MKGI مقایسه

ویژگی	MKGI	CoxPath
استفاده از دانش زیستی	دارد	کم
درک تعاملات بین مسیرها	دارد	محدود
مبتنی بر شبکه + تکاملی	انتخاب ژن	LASSO
دقت پیش‌بینی بقا	بالاتر	متوسط
هزینه محاسباتی	زیاد	کم

جمع‌بندی فصل پنجم 🔥

فصل مربوط به پیش‌بینی بقا نشان می‌دهد که:

- ترکیب چندآمیک + دانش زیستی → بهترین عملکرد
- MKGI یک روش «تکاملی» و «دانش‌محور» است که تعامل بین مسیرهای زیستی را در نظر می‌گیرد
- CoxPath روش کلاسیک‌تر و سریع‌تر است، ولی دقیق‌تر از MKGI است

این فصل نتیجه می‌گیرد که در پزشکی دقیق، برای پیش‌بینی بقا بیماران، بهتر است از:

مدل‌های ادغامی + دانش‌محور + چندلایه آمیک

استفاده شود.

فصل ششم: جمع‌بندی و چشم‌انداز آینده (Conclusion & Future Directions)

جمع‌بندی (Conclusion) ⭐

مقاله نتیجه می‌گیرد که:

درک کامل بیماری‌ها — مخصوصاً سرطان — تنها با یک نوع داده آمیک ممکن نیست. 🔥

هر لایه داده (بیان ژن، متیلاسیون، miRNA، CNV، ...) فقط بخشی از واقعیت زیستی را نشان می‌دهد.

به همین دلیل، یکپارچه‌سازی چندآمیک (Multi-omics Integration) برای پزشکی دقیق ضروری است.

نویسنده‌گان تأکید می‌کنند که:

- روش‌های بدون نظارت می‌توانند زیرگروه‌های پنهان سرطان را کشف کنند
- روش‌های با نظارت اطلاعات بالینی را وارد مدل کرده و پیش‌بینی‌های دقیقی ایجاد می‌کنند
- روش‌های نیمه‌نظرارت شده زمانی عالی هستند که برچسب‌ها کم ولی داده‌ها زیاد باشد
- مدل‌های شبکه‌ای و دانش‌محور نقش مهمی در تفسیرپذیری دارند

به طور کلی:

هرچه داده‌های بیشتری ادغام شود → دقت و قدرت تحلیل بیشتر می‌شود.

چالش‌های موجود (Current Challenges)

مقاله چند چالش بزرگ در ادغام داده‌های چندآمیک را مطرح می‌کند:

۱. تنوع و ناهمگنی داده‌ها

لایه‌های omics دارای:

- توزیع‌های آماری متفاوت
- مقیاس‌های مختلف
- نویزهای خاص خودشان هستند.
- یکپارچه‌سازی آن‌ها کار پیچیده‌ای است.

۲. ابعاد بسیار بزرگ

هر لایه هزاران ویژگی دارد.
ترکیب چند لایه → میلیون‌ها ویژگی بالقوه.
مدل‌های ساده نمی‌توانند از پس این حجم بر بیایند.

۳. کمبود برچسب بالینی

در اکثر پژوهه‌ها:

- داده‌های ژئومی زیادند
- اما داده‌های بالینی دقیق کم هستند

همین موضوع یادگیری نظارت شده را سخت می کند.

◆ ٤. چالش در تفسیرپذیری

بسیاری از مدل‌های عمیق خوب کار می کنند اما:

- چرا نتیجه می دهند؟
- کدام مسیرهای زیستی مهم بودند؟
- این‌ها مشخص نیست.

◆ ٥. هزینه محاسباتی بسیار بالا

به خصوص برای روش‌هایی مثل SNF، MKGI و مدل‌های عمیق.

چشم‌انداز آینده (Future Directions) ★

مقاله پیش‌بینی می کند که جهت پژوهش‌ها در آینده به سمت موارد زیر خواهد بود:

١. مدل‌های ادغامی عمیق‌تر و هوشمندتر

مثل:

- Graph Neural Networks
- Attention-based Multi-omics Models
- Transformers مخصوص داده‌های زیستی

این مدل‌ها قادرند ساختار پیچیده تعاملات ژنی را بهتر یاد بگیرند.

٢. ترکیب داده‌های امیک + داده‌های تک‌سلولی (Single-cell Omics)

امیک‌های تک‌سلولی در حال رشدند و با ادغام چندامیک، قدرت تشخیص زیرگروه‌ها شدیداً افزایش می یابد.

٣. استفاده از دانش زیستی (Knowledge Integration)

ترکیب omics با:

- شبکه‌های پروتئینی
- پایگاه‌های Pathway مثل KEGG

• causal مدل‌های

باعث افزایش دقت و تفسیرپذیری می‌شود.

٤. مدل‌های Multi-modal با داده‌های پزشکی تصویری

در آینده احتمالاً داده‌های:

CT •

MRI •

عکسبرداری میکروسکوپی •

omics همراه با •

در یک مدل ادغام می‌شوند.

٥. ادغام داده‌های زمانی (Longitudinal Multi-omics)

يعنى بررسى اين که بيمار در طول زمان چگونه تغيير مى‌کند.

جمع‌بندی نهایی ★

مقاله نتیجه می‌گيرد که:

★ پزشکی آينده بر پايه ادغام داده‌هاست هرچه بيشتر، بهتر.

★ Multi-omics کليد فهم دقيق بيماري‌هاست.

★ تركيب يادگيري ماشين + دانش زيسنی بهترین راه حل است.