Politecnico di Torino

Laurea Magistrale in Ingegneria Matematica

Analisi Comparativa e Strategie Future per lo Studio Epigenetico del Carcinoma Mammario

Revisione dei Lavori di Tesi

Eleonora Guarnaschelli
Elisabetta Roviera

Contents

---

1 Background e Obiettivi Scientifici

L'obiettivo comune dei due progetti di tesi è l'identificazione di **CpG significative** in grado di distinguere il **tessuto normale** da quello **adiacente al tumore** nel carcinoma mammario. Lo studio si basa sul concetto di *field cancerization*. Analizzare le differenze di metilazione permette di comprendere i **meccanismi iniziali della carcinogenesi** e di individuare **biomarcatori di rischio epigenetico precoce**.

2 Struttura dei Progetti di Tesi e Linee Metodologiche

Entrambi i lavori condividono la stessa architettura, ma divergono negli algoritmi implementati e sul metodo di validazione biologica.

Fase	Guarnaschelli	Roviera
Selezione CpG	Metodi statistici di feature selection (ANOVA, Gain Ratio, iEVORA) e modelli ML supervisionati (XGBoost) per la classificazione *Normal vs Adjacent*.	Sviluppo parallelo di metodi statistici, ML e DL (inclusi test di variabilità, iEVORA, reti neurali e LSTM) per individuare CpG discriminanti e pattern non lineari di metilazione.
Validazione biologica	Analisi basata su entropia come misura di instabilità epigenetica intra-gruppo.	Calcolo di un indice Dynamical Network Biomarker-like per misurare l'instabilità epigenetica nei tessuti adiacenti.
Interpretazione finale	Gene & Pathway analysis.	Gene & Pathway analysis.

Le pipeline convergono in una **fase comparativa finale**, che confronterà i loci CpG identificati da ciascun approccio e i pathway associati, con l'obiettivo di evidenziare marcatori comuni e complementari.

3 Pipeline Analitica: Dataset, Pre-processing e Metodi Computazionali

Dataset di Metilazione e Strategia di Utilizzo

Dataset	Piattaforma	Campioni	Link GEO
GSE69914	Illumina 450K	Normal (50) / Tumor (263) / Adjacent (42)	Link
GSE287331	Infinium EPIC (930K CpG)	Normal (250) / Tumor (69) / Adjacent (60)	Link
GSE225845	Infinium EPIC (930K CpG)	Normal (104) / Tumor (185) / Adjacent (113)	Link

Gestione dei dataset I dataset EPIC verranno armonizzati e **troncati alle sole CpG comuni** con la piattaforma 450K, in modo da garantire la comparabilità tra studi. L'idea operativa è di **unire i due dataset EPIC più numerosi** dopo il troncamento, utilizzandoli come **training set** per i modelli di classificazione, e di **testare le pipeline sul dataset meno numeroso** (ad esempio GSE69914), qualora la compatibilità dei dati lo consenta. In alternativa, qualora si ritenga più appropriato mantenere GSE69914 come dataset di addestramento, questa scelta dovrà essere esplicitata.

DOMANDE

- Qual è la strategia più solida: usare i dataset EPIC unificati come *training set* e GSE69914 come *test set*, oppure invertire tale impostazione?
- È necessario che i campioni *Normal* e *Adjacent* siano **appaiati** (provenienti dallo stesso paziente) per garantire maggiore coerenza biologica?
- Come è possibile accedere ai **metadati clinici** (mutazioni, età, sesso, composizione cellulare) per valutarne l'impatto e, se disponibili, includerli come covariate nei modelli ML/DL?
- Devono essere **rimossi i campioni portatori di mutazioni BRCA1/BRCA2**, qualora tali alterazioni possano introdurre bias nei pattern di metilazione?

Strategia di Pre-Processing dei Dati

Pipeline condivisa ispirata a *Newsham et al.* con estensioni adattate ai due progetti.

1. **Importazione e armonizzazione** dei dataset (riduzione dei dataset EPIC alle CpG comuni con la piattaforma 450K).
2. **Rimozione CpG rumorose**, secondo le liste di:
   - *Pidsley et al. (2016) [1]* e *Chen et al. (2013) [2]* – cross-reactive probes;
   - *Naeem et al. (2014) [3]* – sonde con elevata variabilità non biologica.
3. **Gestione dei valori mancanti**: verranno rimossi i CpG o i campioni contenenti NaN, a seconda della loro distribuzione (prevalenza per CpG o per tessuto).
4. **Conversione dei valori β in M-values** per garantire omoschedasticità e validità statistica.
5. **Normalizzazione** dei valori M.
6. **Feature selection preliminare** per ridurre la dimensionalità prima della fase di training.

DOMANDE

- Per quanto riguarda le liste di filtraggio proposte siamo aperte a consigli su eventuali alternative o su criteri per selezionare quella migliore.
- Biologicamente ha senso rimuovere le CpG incluse nelle liste di filtraggio, come quella di Chen [2] (e la sua versione aggiornata e modificata proposta da Pidsley [1])? Infatti, nello studio di Teschendorff [4] 923 CpG considerate significative per distinguere tessuto normale e adiacente risultano incluse nella lista di Chen et al., mentre nello studio di Ding [5] una delle sette CpG ritenute più discriminanti tra tessuto tumorale e normale appartiene alla lista di Pidsley e un'altra a quella di Naeem.
- Inserire subito un livello di feature selection statistica o applicarlo dopo la normalizzazione?

Metodi di Analisi e Selezione delle CpG

Entrambe le tesi mirano a identificare pattern discriminanti e interpretabili, combinando approcci statistici, ML e DL.

Metodi statistici

- *iEVORA* – individuazione di CpG a varianza differenziale.
- *Gain Ratio*, *ANOVA*, test non parametrici – selezione di feature con elevata separabilità tra gruppi.

Machine & Deep Learning

- **XGBoost**, **LightGBM**, **CatBoost** – modelli supervisionati ad alte prestazioni.
- **Reti neurali dense e LSTM** – esplorazione di pattern non lineari nei profili di metilazione.
- **Autoencoder** – riduzione non lineare della dimensionalità e individuazione di feature epigenetiche emergenti.

Criteri di valutazione

- Accuratezza, ROC-AUC, F1 score, MCC e interpretabilità biologica;
- Selezione finale di CpG tramite importance score o p-value thresholding;
- Confronto quantitativo e visivo tra metodi per verificare consistenza e stabilità.

Analisi dei risultati e confronto con precedenti studi L'analisi delle CpG identificate da XGBoost potrà essere approfondita valutando la forma delle loro distribuzioni di metilazione (asimmetria, multimodalità, varianza e presenza di outlier nei tessuti adiacenti rispetto ai normali). Le CpG selezionate verranno annotate per individuare i geni e i pathway maggiormente coinvolti, verificando la coerenza con i risultati riportati da Teschendorff et al. e la possibile espressione differenziale degli stessi geni in dataset indipendenti. Questa fase consentirà di collegare le evidenze statistiche del modello a un contesto biologico verificabile

Prossimi passi

L'approccio congiunto integra la solidità dei metodi statistici e ML con la flessibilità dei modelli DL, offrendo una visione completa del fenomeno epigenetico.

1. Consolidare un pre-processing standardizzato;
2. Integrare e confrontare metodi di feature selection e classificazione;
3. Elaborare un'analisi comparativa finale dei risultati.

4 Proposta di Pubblicazione e Sintesi dei Risultati Attesi

I due progetti, fondati su approcci complementari di **variabilità differenziale** e **outlier detection**, potranno confluire in una pubblicazione congiunta dedicata allo studio dei *field defects* nel carcinoma mammario. L'obiettivo comune è proporre un **modello integrato** per l'identificazione di marcatori epigenetici di rischio precoce, che combini solidità statistica e sensibilità a eventi epigenetici rari.

1. **Ricerca delle CpG e validazione incrociata.** Applicazione dei diversi metodi statistici, di machine e deep learning per individuare CpG discriminanti tra tessuti *Normal* e *Adjacent*. Le pipeline verranno testate mediante **cross-validation** su un dataset indipendente, così da verificare la robustezza e la generalizzabilità dei marcatori individuati.
2. **Analisi di entropia e biomarcatori.** Valutazione dell'instabilità epigenetica attraverso l'entropia e un indice DNB-like ispirato ai Dynamical Network Biomarkers, al fine di caratterizzare lo stato critico dei tessuti adiacenti al tumore.
3. **Confronto dei risultati.** Confronto quantitativo e biologico tra i risultati ottenuti dai due approcci. Saranno analizzati i loci CpG comuni e le differenze di pattern di metilazione.

L'obiettivo finale è delineare un **framework integrato** per l'identificazione di marcatori epigenetici affidabili e biologicamente interpretabili.

References

[1] R. Pidsley, E. Zotenko, T. J. Peters, M. G. Lawrence, G. P. Risbridger, P. Molloy, S. V. Dijk, B. Muhlhausler, C. Stirzaker, and S. J. Clark, "Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome dna methylation profiling," *Genome Biology*, vol. 17, no. 1, p. 208, 2016. DOI: 10.1186/s13059-016-1066-1 [Online]. Available: https://doi.org/10.1186/s13059-016-1066-1

[2] Y. Chen, A. Lemire, S. Choufani, R. Z. Butcher, J. Grafodatskaya, R. Zanke, W. Lou, R. Mill, S. W. Scherer, and R. Weksberg, "Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray," *Epigenetics*, vol. 8, no. 2, pp. 203–209, 2013. DOI: 10.4161/epi.23470 [Online]. Available: https://doi.org/10.4161/epi.23470

[3] H. Naeem, M. Wong, K. Chatterton, M. L. Watson, S. J. Lamont, E. K. McCallum, C. Stirzaker, P. Molloy, S. J. Clark, and T. J. Peters, "Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the illumina humanmethylation450 array," *Genome Biology*, vol. 15, no. 12, R126, 2014. DOI: 10.1186/gb-2014-15-12-r126 [Online]. Available: https://doi.org/10.1186/gb-2014-15-12-r126

[4] A. E. Teschendorff, Y. Gao, A. Jones, M. Ruebner, M. W. Beckmann, D. L. Wachter, P. A. Fasching, and M. Widschwendter, "Dna methylation outliers in normal breast tissue identify field defects that are enriched in cancer," *Nature Communications*, vol. 7, p. 10 478, 2016. DOI: 10.1038/ncomms10478 [Online]. Available: https://doi.org/10.1038/ncomms10478

[5] W. Ding, G. Chen, and T. Shi, "Integrative analysis identifies potential dna methylation biomarkers for pan-cancer diagnosis and prognosis," *Epigenetics*, vol. 14, no. 1, pp. 67–80, 2019. DOI: 10.1080/15592294.2019.1568178