Politecnico di Torino

Master's Degree in Mathematical Engineering

Material for Thesis

Thesis Structure

Elisabetta Roviera s328422

Contents

Note The following outline represents the current planned structure of the thesis. It is intended as a working version and may be refined during the development of the analysis, for example depending on data availability, the stability of the proposed CpG signature, and the outcome of validation experiments on external cohorts. The logical flow (from biological motivation, to computational framework, to application, validation, and quantitative interpretation of epigenetic instability) is expected to remain consistent, even if specific subsections are adjusted or merged in the final document.

1 Introduction

1.1 Motivation

Breast cancer initiation is thought to proceed through intermediate states in which histologically normal tissue located near the primary lesion ("normal-adjacent" tissue) already displays epigenetic disruption. Detecting and quantifying these early alterations is relevant for local risk assessment and for understanding tumor initiation.

1.2 Background and open problem

Robust DNA methylation signatures distinguishing Tumor vs. histologically normal tissue exist. In contrast, it is still unclear whether it is possible to reproducibly distinguish truly healthy normal tissue from normal-adjacent tissue using DNA methylation alone. This is an important question, since normal-adjacent tissue is often considered a "field at risk".

1.3 Aim of the thesis

The thesis aims to:

- build a quantitative pipeline to identify CpG sites that discriminate normal vs. normal-adjacent breast tissue;
- define a continuous epigenetic score that places each sample along a Normal → Normal-adjacent → Tumor trajectory;
- characterize normal-adjacent tissue as a potentially unstable, pre-neoplastic state using quantitative instability measures (entropy and Dynamical Network Biomarker concepts).

1.4 Outline of the work

The thesis proceeds from biological motivation to computational formulation, application on real breast tissue methylation data, validation in external cohorts, and quantitative interpretation of normal-adjacent tissue as a critical pre-neoplastic state.

2 Biological and Clinical Background

2.1 Breast tissue organization

Overview of the cellular compartments of breast tissue (luminal and basal epithelial cells, stromal compartment, adipocytes, immune infiltration) and their relevance in bulk methylation data.

2.2 DNA methylation and cancer development

DNA methylation is measured as a β-value in $[0, 1]$, reflecting the fraction of methylated copies at a CpG site. Aberrant hypermethylation of regulatory regions is linked to the silencing of differentiation, cell cycle control, and tumor suppressor programs. Such alterations often appear before large-scale genetic changes.

2.3 Normal-adjacent tissue and field cancerization

Definition of normal-adjacent tissue: histologically normal tissue sampled near the tumor. Literature reports indicate that this tissue frequently exhibits focal hypermethylation, increased variability, and repression of developmental regulators, suggesting the presence of a pre-neoplastic “field defect”.

2.4 Why normal vs. normal-adjacent matters

Distinguishing normal from normal-adjacent is clinically relevant (margin assessment, local risk stratification) and mechanistically relevant (capturing the earliest epigenetic instability). The central question is whether a reproducible CpG-based signature can be derived specifically for this distinction, beyond standard Tumor vs. Normal comparisons.

3 Mathematical and Computational Framework

3.1 Data model

Let $X \in [0,1]^{p\times n}$ be the DNA methylation matrix, with p CpG sites and n samples labelled as Normal, Normal-adjacent, or Tumor. Public datasets provide methylation levels as β-values, already preprocessed at source.

3.2 Probe-level quality filtering

Sondes/CpG probes flagged as technically unreliable are removed, including:

- cross-reactive probes (off-target hybridization),
- probes overlapping polymorphisms that alter the measured signal,
- optionally, probes on sex chromosomes when they may introduce sex-driven confounding.

The goal is to retain CpG sites with interpretable and reproducible biological signal.

3.3 β-to-M transformation and statistical harmonization

The raw methylation proportion β is transformed into an M-value:

$$M = \log_2\left(\frac{\beta}{1-\beta}\right).$$

This transformation stabilizes variance and improves the applicability of linear modeling and t-type tests, since β-values are bounded and highly skewed near 0 and 1.

3.4 CpG-level differential analysis

Within-group comparisons are performed on M-values to detect CpG sites that:

- change in mean methylation between Normal and Normal-adjacent;
- and/or show increased variability (differential variance) in Normal-adjacent.

The second criterion is motivated by the idea that early pre-neoplastic change is often heterogeneous and focal rather than uniform across all cells.

3.5 Outlier modelling and variance-based instability

Difference between mean shifts (classical differential methylation) and variance-based changes is formalized. For each CpG and each sample, an “outlier burden” is defined as the presence of extreme methylation values in normal-adjacent tissue compared to healthy normal tissue. Both simple statistical rules (e.g. IQR/quantiles, skewness) and machine-learning anomaly scores (e.g. isolation forest, autoencoder) are considered as reference approaches.

3.6 Feature selection and classifier construction

Candidate CpGs are further reduced using multivariate feature selection methods (e.g. L1-penalized logistic regression, SVM recursive feature elimination, tree-based importance ranking). The aim is to obtain a compact CpG panel that best separates Normal and Normal-adjacent samples. Performance is quantified using metrics such as ROC/AUC, sensitivity, and specificity.

3.7 Definition of an epigenetic score

An epigenetic score is defined for each sample as a weighted combination of the selected CpG sites. This score is intended to order all samples along a molecular axis Normal → Normal-adjacent → Tumor, and to provide a single interpretable readout of "epigenetic deviation from healthy normal".

3.8 Instability metrics (link to Chapter 6)

Two quantitative indicators of epigenetic instability are introduced for later analysis:

- an entropy-like measure of global disorder in the methylation profile,
- a Dynamical Network Biomarker (DNB)-like instability index based on covariance structure within a subset of CpG sites.

These measures will be formally defined and applied in Chapter 6.

4 Application to the Primary Dataset

4.1 Dataset description

Detailed description of the main breast cancer DNA methylation dataset: number of Normal, Normal-adjacent, and Tumor samples; platform (e.g. Illumina 450K); availability of paired samples from the same patients.

4.2 Preprocessing and harmonization

Concrete application of the Chapter 3 pipeline:

- probe filtering for technical reliability,
- $\beta \to M$ transformation and its practical benefit for statistical testing,
- optional centering/scaling or batch-effect adjustment at the M-value level,
- genomic/functional annotation of the retained CpGs.

4.3 Exploratory data analysis

Exploration of global structure using PCA, hierarchical clustering, and heatmaps of informative CpGs. Visual positioning of Normal, Normal-adjacent, and Tumor samples in the reduced-dimensional space to assess whether Normal-adjacent occupies an intermediate or unstable state.

4.4 Baseline classification: Normal vs. Tumor

Supervised models (e.g. XGBoost) are trained on healthy normal vs. tumor samples to confirm that standard tumor/normal separation can be reproduced on the primary dataset. Performance is reported with ROC–AUC, MCC, and confusion matrices. The most informative CpG sites for this task are extracted as a positive control, before addressing Normal vs. Normal-adjacent.

4.5 Identification of a Normal vs. Normal-adjacent CpG panel

Selection of CpG sites with the strongest discriminatory power between Normal and Normal-adjacent using the criteria and feature selection of Chapter 3. Comparison with classical Tumor vs. Normal signatures from the literature, to highlight that early field effects are not necessarily captured by late tumor signatures.

4.6 Epigenetic score on the primary dataset

Calculation of the epigenetic score for each sample in the primary cohort. Analysis of score distributions across Normal, Normal-adjacent, and Tumor, and interpretation of Normal-adjacent as an intermediate or "destabilized" state.

4.7 Genomic and functional interpretation

Annotation of the selected CpGs in terms of genomic location and associated pathways (e.g. cell cycle control, developmental regulators, chromatin modifiers). Biological interpretation in the context of early breast carcinogenesis and field cancerization.

5 Cross-cohort Validation

5.1 Independent datasets

Selection of one or more external cohorts containing Normal, Normal-adjacent (or equivalent "at-risk" tissue), and/or Tumor samples profiled on compatible DNA methylation arrays.

5.2 External projection of the CpG panel and the epigenetic score

Application of the CpG panel derived in Chapter 4 to the independent datasets, without retraining. Computation of the epigenetic score in these external cohorts and evaluation of its ability to distinguish Normal and Normal-adjacent.

5.3 Robustness and reproducibility

Discussion of cross-cohort stability, including:

- differences in sample handling and clinical composition,
- residual batch effects,
- residual differences in cell-type composition.

Assessment of how transferable the proposed panel/score is as an early epigenetic biomarker.

6 Quantitative Indicators of Epigenetic Instability

6.1 Entropy-based instability

Definition of an entropy metric Definition of an entropy-like measure for each sample, computed from the methylation profile (for example, restricted to the CpG panel identified in Chapter 4). Interpretation of higher entropy as higher global epigenetic disorder and loss of a stable tissue identity.

Application to Normal / Normal-adjacent / Tumor Comparison of entropy values across Normal, Normal-adjacent, and Tumor groups to test whether normal-adjacent tissue exhibits increased global disorder relative to healthy normal.

Relation to the epigenetic score Analysis of the relationship between entropy and the epigenetic score introduced in Chapter 4, to assess whether both measures consistently identify normal-adjacent tissue as epigenetically altered with respect to normal.

6.2 Dynamical Network Biomarker (DNB)-like instability index

Theoretical basis Before a critical transition in a complex biological system, a specific subnetwork often shows:

- increased internal variability,
- increased internal correlation,
- decreased coupling to the rest of the system.

This phenomenon is known as a Dynamical Network Biomarker (DNB) and signals entry into an unstable, critical state.

Operational definition for CpG methylation data A subset of CpG sites (for example, those most informative for Normal vs. Normal-adjacent in Chapter 4) is treated as a candidate "critical subnetwork". For each group (Normal, Normal-adjacent, Tumor), an instability index is computed by combining:

- the average standard deviation within the subset,
- the average pairwise correlation within the subset,
- the (inverse of the) average correlation between the subset and the rest of the CpGs.

The resulting index is expected to peak in a critical, unstable state.

Application and interpretation Evaluation of the instability index across Normal, Normal-adjacent, and Tumor. Hypothesis: the index is maximal in Normal-adjacent, indicating that this tissue is in a high-instability, pre-transition regime.

Joint interpretation Integration of:

- the epigenetic score (Chapter 4),
- the entropy measure (Section 6.1),
- the DNB-like instability index (Section 6.2).

Goal: interpret Normal-adjacent tissue as an epigenetically unstable, critical state located between stable healthy tissue and fully transformed tumor tissue.

6.3 Link between entropy, outlier burden, and the epigenetic score

The epigenetic score (Chapter 4) is compared with entropy-based instability (Section 6.1) and with outlier burden (Section 3.5). The aim is to test whether samples with higher score also show higher entropy and a heavier accumulation of outlier CpG values, supporting the interpretation of normal-adjacent tissue as an epigenetically destabilized state.

7 Discussion and Conclusions

7.1 Summary of findings

Summary of the main results:

- identification of CpG sites that discriminate Normal and Normal-adjacent tissue,
- definition of an epigenetic score that orders samples along a Normal → Normal-adjacent → Tumor axis,
- evidence that normal-adjacent tissue displays quantitative signs of epigenetic instability.

7.2 Biological and clinical implications

Potential use of early DNA methylation alterations as local risk indicators, e.g. to assess surgical margins or local preneoplastic fields. Relevance for understanding how epigenetic disruption precedes (and possibly facilitates) malignant transformation.

7.3 Limitations and Future work

Main limitations The main limitations are the lack of raw IDAT files, which prevents full low-level preprocessing and platform-specific normalization, and the possibility of residual technical or compositional effects in the data.

Future extensions Future extensions include applying the same framework to additional breast cancer subtypes and to other solid tumors, in order to assess subtype specificity and generalizability.