Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Informatica

ANALISI E MIGLIORAMENTO PRESTAZIONALE DELL'IMPUTAZIONE DI DATI DI METILAZIONE CON *methyLImp2*

Relatore:
Chiar.mo Prof.
Pietro Di Lena

Presentata da:
Federico Piozzi

III Sessione
Anno Accademico 2022/2023

A Mamma e Papà che ammiro.

A Margherita e Rolly che amo.

A Elisabetta per il supporto.

Agli amici per il sostegno.

Abstract

Background:
La metilazione del DNA è una modificazione epigenetica che interessa i siti CpG (citosina-fosfato-guanina) del DNA. Essa ha numerose implicazioni nello sviluppo e nell'invecchiamento fisico umano nonché nello sviluppo di patologie come il cancro. Per questo numerose ricerche si interessano a questo processo biologico stimolate anche dall'introduzione di tecnologie avanzate che permettono un'elevata caratterizzazione dei livelli di metilazione del genoma umano. Le sperimentazioni svolte su profili di metilazione contengono spesso dati mancanti che possono peggiorare o compromettere l'analisi sui dati di metilazione ottenuti. Per questo motivo si utilizzano tecniche di imputazione per stimare i dati mancanti. Da un confronto tra le tecniche disponibili si è rilevato come alcune siano risultate più adatte nel trattare dati di metilazione. In particolare si è distinta una tecnica di imputazione chiamata *methyLImp* in quanto pensata appositamente per trattare dati di metilazione. Essa ha dimostrato, nel confronto con tecniche già presenti ed utilizzate, di performare meglio sia in termini di accuratezza che in termini di efficienza.
Risultati:
Si sono svolti test utilizzando la versione più recente *methyLImp2* di *methyLImp* lavorando su set di dati di metilazione con rappresentazione in β-values, sui quali sono stati simulati valori mancanti rifacendosi alle distribuzioni di valori mancanti estrapolate da set di dati di metilazione reali. I test si sono svolti in due fasi: la prima imputando i valori mancanti senza sfruttare le funzionalità introdotte con *methyLImp2*, la seconda sfruttando la funzionalità di parallelizzazione su cromosomi. Si è poi svolto un confronto tra le imputazioni delle due fasi in termini di accuratezza ed efficienza osservando le differenze ed i miglioramenti ottenuti.
Conclusioni:
Le analisi svolte sui risultati ottenuti nei test hanno permesso di capire come sfruttare alcune delle nuove funzionalità fornite da *methyLImp2* ottenendo tempi di esecuzione nettamente minori pur mantenendo l'accuratezza invariata.

Indice

Elenco delle tabelle

Capitolo 1

Introduzione

1.1 BioInformatica

La BioInformatica è una disciplina scientifica che, come suggerisce il nome, unisce le discipline di biologia ed informatica. In particolare, è una disciplina che si occupa di sviluppare algoritmi, metodologie e strumenti software per l'analisi di dati biologici. L'applicazione dell'informatica al campo della biologia ha consentito una maggior comprensione dei sistemi biologici permettendo, ad esempio, di sequenziare i genomi di molti organismi, di conservare i dati raccolti in database biologici e di analizzare questi dati mediante algoritmi specializzati. Oltre all'analisi di sequenze, è stato possibile analizzare l'espressione genica ed attuare analisi predittive relativamente a mutazioni cancerogene, a strutture e funzioni proteiche ecc. È chiaro, dunque, come la bioinformatica sia divenuta, con lo sviluppo dell'informatica stessa, di vitale importanza nella ricerca in campo biologico.

Una delle aree di ricerca più attive nel campo della biologia è la cosiddetta *Epigenomica*: un settore specializzato nello studio di alterazioni fenotipiche che il genoma subisce senza che venga modificata la sequenza del DNA. La rigogliosa ricerca sull'argomento deriva dalla suddetta caratteristica di ereditabilità a livello cellulare di queste alterazioni, la quale permette di studiare e comprendere meglio meccanismi biologici causali alla base della nascita di patologie come il cancro, l'Alzheimer o il Parkinson.

Come l'epigenomica studia globalmente i cambiamenti all'interno dell'intero genoma, similmente l'epigenetica si riferisce allo studio di singoli o gruppi di geni, ed in particolare si concentra su 3 tipi di alterazioni: metilazione del DNA, modificazioni istoniche e cambiamenti successivi alla trascrizione del DNA.

Di seguito verrà introdotta brevemente la metilazione del DNA, una delle 3 alterazioni studiate dall'epigenetica, che sarà il processo biologico alla base di tutta la successiva trattazione.

1.2 Metilazione del DNA

La metilazione del DNA è una modificazione epigenetica che coinvolge il DNA stesso. Descrivendola più nello specifico, per i mammiferi, essa consiste nell'addizione covalente di un gruppo metile (-CH3) al livello del carbonio-5 della citosina andando a formare la 5-metil-citosina (5mC). Questa modificazione avviene quasi esclusivamente nel contesto del dinucleotide CpG (citosina seguita da una guanina) nella regione codificante dei geni.

La metilazione del DNA viene studiata in diversi ambiti della biologia e interessa tutto l'arco della vita umana, dallo sviluppo embrionale all'invecchiamento [4]. Per lo sviluppo embrionale, definisce l'identità di cellule di tessuti specifici [5] mentre, per l'invecchiamento, comporta il silenziamento selettivo, stabile e reversibile, di porzioni del genoma. Inoltre la metilazione del DNA interviene nella modificazione delle cellule immunitarie ed è fondamentale nella costituzione della cosiddetta memoria cellulare, la quale è un fattore cruciale, ad esempio, nella vaccinazione [6].
Per questi motivi, una comprensione profonda delle metriche e delle statistiche più adatte a modellare correttamente l'attività del processo di metilazione permetterebbe di utilizzare i dati di metilazione, da essa estrapolati, come potenziale biomarcatore, ovvero come indicatore biologico correlato alla predisposizione di uno specifico individuo allo sviluppo di diverse patologie e malattie o utile nel determinare risposte biologiche a specifici trattamenti.

1.3 Dati di Metilazione

I dati di metilazione sono intesi, nel loro senso più ampio, come tutte le informazioni estrapolate dal processo di metilazione del DNA.

Questi dati permettono di stimare la cosiddetta età di metilazione (indicata come *mAge*). Quest'ultima non è altro che una misura dell'età biologica di un individuo basata su modelli di metilazione del DNA estrapolati dai dati di metilazione stessi. Negli ultimi anni, molte ricerche si sono concentrate nello stabilire una correlazione tra età cronologica (=tempo trascorso dalla data di nascita) ed mAge mediante

l'utilizzo di orologi epigenetici [13]. La logica dietro questi ultimi si fonda sull'aver osservato come, la divergenza, rispetto alla correlazione stimata, tra età cronologica ed mAge in uno specifico CpG, sia indicativa, e spesso predittiva, di diversi fattori. Tra questi generalmente si distinguono indicazioni relativamente allo sviluppo di gravi patologie, se l'età di metilazione è maggiore di quella cronologica [1] o, al contrario, un invecchiamento biologico sano nel caso in qui si osservi un rallentamento nella crescita dei livelli di metilazione con l'aumentare dell'età biologica.

Oltre ad essere fondamentali nel campo della ricerca, i dati di metilazione saranno il soggetto principale di questa tesi poiché protagonisti del processo di imputazione e delle successive analisi. Per questo verranno ulteriormente approfonditi nel seguito della trattazione (Vedi capitolo 2.4).

1.4 *methyLImp* e *methyLImp2*

I tool *methyLImp* [7] e *methyLImp2* [17] sono due versioni, introdotte recentemente, di un metodo definito appositamente per stimare valori mancanti di dati di metilazione del DNA. Sviluppati in ambiente R, nella pratica sono pacchetti di quest'ultimo. Essi si sono distinti, tra le varie tecniche di imputazione, per applicare un metodo pensato, come detto, per applicazioni riguardanti il processo di metilazione del DNA, riscuotendo particolare successo in termini di prestazioni, se confrontati con altri metodi, con scopi similari, più usati ad oggi [7] [8].

Questa tesi si pone l'obiettivo di analizzare e descrivere il funzionamento delle due versioni di *methyLImp*, ponendo maggior attenzione sulla seconda versione *methyLImp2*, allo scopo di migliorarne due aspetti cruciali in termini computazionali, quali efficienza e costo. Verrà inizialmente affrontato il problema dell'imputazione in termini generali e relativamente ai dati di metilazione per poi passare ad una descrizione di *methyLImp* e *methyLImp2*, sia relativamente alle modalità in cui operano, sia in termini prestazionali mediante l'analisi computazionale degli stessi. Quest'ultima sarà fondamentale per individuare possibili modifiche e sfruttare al meglio determinate funzionalità fornite dagli algoritmi stessi al fine di migliorarli negli aspetti suddetti. Verrà inoltre descritto, in tutte le sue fasi, il processo propedeutico ai test, che riguarda la preparazione dei dati di metilazione da utilizzare.

Il miglioramento di *methyLImp2* permetterà di attuare imputazioni di dati di metilazione ancora più precise, efficienti ed efficaci, con l'intento di dare stimoli ad ulteriori ottimizzazioni e alla ricerca relativa alla metilazione del DNA.

Capitolo 2

Imputazione dei dati di metilazione

In questo capitolo si vuole chiarire cosa significhi imputare una qualsiasi variabile e si vuole spiegare in cosa consiste il processo di imputazione. Si inizierà presentando le tecnologie e gli strumenti che permettono di raccogliere i dati di metilazione; si cercherà poi di dare una definizione formale al problema dell'imputazione non strettamente relativa ai dati di metilazione. Su questi ultimi si tornerà, al termine del capitolo, riprendendo il concetto di imputazione nel contesto della metilazione del DNA, fondamentale per il seguito della trattazione.

2.1 Infinium Methylation Assay

L'*Infinium Methylation Assay* [15], sviluppato dall'azienda Illumina, è al momento la miglior tecnologia in termini di costo ed efficienza che permetta di stimare i livelli di metilazione del DNA. Questa tecnologia fa utilizzo di una coppia di sonde progettate per misurare l'intensità di metilazione negli alleli metilati e non metilati all'interno di ogni sito CpG. Svolte queste misurazioni, i livelli di metilazione sono calcolati come il rapporto tra i segnali degli alleli metilati e i segnali degli alleli non metilati raccolti in precedenza su tutte le cellule del tessuto considerato.

Questa tecnologia ha permesso di reperire una grande quantità di dati, pubblici ed accessibili, consentendo di poter attuare meta-analisi approfondite su di essi e conseguenti analisi statistiche migliori che hanno, a loro volta, stimolato ricerca e scoperte. Contemporaneamente, sono stati sviluppati approcci di pre-elaborazione [18] pen-

sati appositamente per i dati di metilazione, che hanno ulteriormente migliorato il processo di analisi.

2.2 β-Value e M-Value

I dati di metilazione sono comunemente misurati utilizzando due metriche: i β-Values e gli M-Values.

I β-Values hanno valori compresi tra 0 e 1. Rispettivamente, il valore zero indica che tutte le coppie dei siti CpG (vedi 2.4.1) analizzati sono completamente non metilate, mentre il valore uno indica che tutte le coppie dei siti CpG sono completamente metilate. La definizione di β-Value, considerato l'i-esimo sito CpG analizzato, è la seguente:

$$\beta_i = \frac{max\left(x_i^{meth}, 0\right)}{max\left(x_i^{meth}, 0\right) + max\left(x_i^{unmeth}, 0\right) + \alpha} \tag{2.1}$$

dove x_i^{meth} e x_i^{unmeth} sono rispettivamente le intensità metilate e non metilate misurate dalle due sonde.

Gli M-Values hanno valori compresi tra $+\infty$ e $-\infty$ e sono definiti come segue:

$$M_i = log_2\left(\frac{max\left(x_i^{meth}, 0\right) + \alpha}{max\left(x_i^{unmeth}, 0\right) + \alpha}\right) \tag{2.2}$$

Tipicamente i β-Values forniscono una misurazione più intuitiva dei livelli di metilazione rispetto agli M-Values rivelandosi più adatti alla produzione di array [2]. Al contrario, per condurre analisi differenziali di metilazione, si è mostrato come i β-Values siano meno adatti rispetto agli M-Values, a causa della loro eteroschedasticità (che si verifica quando all'interno di un campione esistono sotto-popolazioni con varianze diverse) per siti CpG altamente metilati o non metilati [8]. A riprova di ciò si può notare come la deviazione standard nei β-Values è bassa su intervalli di valori all'estremità (più vicini allo zero o all'uno), mentre aumenta per gli intervalli di valori centrali [7].

Inoltre è importante notare come questi due sistemi di misurazione siano in realtà correlati tra loro. Se infatti non consideriamo la variabile di offset α (presente in enrambe le formule), che, per valori tipici di x_i^{meth} e x_i^{unmeth} ha effetti trascurabili

[9], la relazione tra β-Value e M-Value è una trasformazione *logit*:

$$\beta_i = \frac{2^{M_i}}{2^{M_i} + 1} \tag{2.3}$$

$$M_i = log_2 \left(\frac{\beta_i}{1 - \beta_i} \right) \tag{2.4}$$

In questa tesi tratteremo dati di metilazione in notazione β-Value poiché, come già detto, più intuitiva da interpretare ed analizzare e dunque maggiormente funzionale all'obiettivo della trattazione.

2.3 Il Problema dell'Imputazione

In ogni campo in cui si attuano sperimentazioni sono presenti dati mancanti. Questo è determinato da diversi fattori, molti dei quali inevitabili, ad esempio: errori di registrazione durante la raccolta dei dati, perdita di follow-up nel tempo di alcuni soggetti presenti nello studio, problemi tecnici come il malfunzionamento della strumentazione utilizzata o valori mancanti risultato di errori casuali nel processo di raccolta.

La presenza di dati mancanti può compromettere la validità di intere sperimentazioni per diversi motivi, tra i quali, un peggioramento della potenza statistica e della rappresentatività dei risultati delle analisi i quali, a loro volta, comportano nei casi peggiori l'eliminazione di interi set di dati obbligando i ricercatori a ripetere interamente parte dell'analisi con conseguente aumento dei costi, oppure, nel migliore dei casi, obbligando a ridurre il volume di dati analizzati compromettendo, di conseguenza, la precisione della sperimentazione. Si capisce dunque quanto sia importante trattare i set di dati che presentano dati mancanti, andando a stimare questi ultimi piuttosto che eliminandoli. Questo processo è chiamato "imputazione" e il problema, ad esso associato, è detto "problema dell'imputazione di dati mancanti".

2.3.1 Il processo di imputazione

Con il termine *imputazione* si fa riferimento alla sostituzione dei dati mancanti, presenti in uno specifico set di dati, con stime ottenute mediante processi statistici di vario tipo basati su informazioni ricavate dal set di dati stesso. Questi processi, applicati a diversi set di dati, seguono un approccio metodologico sistematico

che si compone di diverse fasi (al momento descritte sinteticamente e approfondite successivamente):

1. Fase di Analisi: analisi ed estrazione dei metadati, standardizzazione del set di dati e simulazione dei dati mancanti.
2. Fase di Imputazione: scelta, implementazione ed applicazione della tecnica di imputazione.
3. Fase di Valutazione: raccolta ed analisi di dati statistici relativi all'imputazione e valutazione dell'efficienza e della precisione dell'imputazione.

2.3.2 Principali tecniche di imputazione

Una tecnica di imputazione è un modello, avente determinate regole, che analizzando le informazioni presenti in un set di dati, tra cui i dati stessi, attua stime dei dati mancanti sostituendole poi a quest'ultimi. Come già detto in precedenza, queste tecniche sono applicate massicciamente su set di dati provenienti da sperimentazioni nei campi di ricerca più disparati (ad esempio in campo biologico, economico ecc.). Il principale vantaggio derivante dal loro utilizzo è il mantenimento dell'ampiezza campionaria con un conseguente miglioramento di tutto il successivo processo di analisi e studio dei risultati. Applicare queste tecniche consente, inoltre, di evitare l'applicazione di altre tecniche molto più impattanti sulla qualità dei risultati e quindi sula validità degli stessi, come una tecnica che consiste nell'eliminazione dei set di dati contenenti unità mancanti. Nonostante ciò ci sono casi in cui, se la presenza di dati mancanti è eccessiva, l'applicazione di tecniche di imputazione è sconsigliata poiché si andrebbero a sostituire i dati mancanti con stime completamente fuori scala. In questi casi diventa più conveniente, ad esempio, eliminare direttamente i set di dati piuttosto che imputarne i dati mancanti.

In generale, le tecniche di imputazione possono essere classificate secondo diverse metodologie. Una si basa sulle variabili considerate in queste tecniche. Queste potranno essere: variabili continue, variabili categoriali, o entrambe, oppure variabili a range limitato. Un'altra distinzione può essere fatta tra metodi di imputazione singola (SI) e metodi di imputazione multipla (MI). I primi consistono nel sostituire al valore mancante un unico valore stimato, mentre i secondi svolgono imputazioni singole multiple, andando a sostituire al valore mancante la media dei valori stimati ottenuti da ogni imputazione singola effettuata.
La più semplice tecnica di imputazione singola è la media[20] (*mean*) che tratta variabili continue in un range limitato sostituendo, al valore mancante di una certa variabile, la media di tutti i valori conosciuti per quella variabile.
Avente la stessa logica è il cosiddetto *impute.knn*[12], un'altra tecnica che, a diffe-

renza della precedente, sostituisce al valore mancante di una variabile la media dei valori (non mancanti) delle variabili più correlate a quella considerata.
Esempi, invece, di metodi multipli iterativi sono: *SVDmiss*[11], *SoftImpute*[16], *imputePCA*[14] e *missForest*[19]. Tutte queste tecniche aggiornano iterativamente i valori mancanti, seguendo determinati metodi, fino a convergenza.

2.3.3 Tipologie di dati mancanti

I dati mancanti sono classificati in 3 classi distinte in relazione alla motivazione più probabile per la quale i dati risultino effettivamente mancanti:

- *Missing completely at random*(MCAR): il motivo non dipende dai dati osservati o non osservati.
- *Missing at random*(MAR): il motivo può essere spiegato dai dati osservati.
- *Missing not at random*(MNAR): il motivo è da attribuire a dati non osservati.

Generalmente le prime due classi (MCAR,MAR) sono considerate "ignorabili" poiché legate ad eventi casuali che sono per definizione non controllabili. Al contrario la terza classe (MNAR) è considerata "non ignorabile" poiché definisce esplicitamente il modello di distribuzione che seguono i dati mancanti, permettendo di modellare su quest'ultimo il processo di imputazione al fine di evitare stime distorte.
Tuttavia, non esiste un metodo statistico generalmente valido per rilevare, in ogni contesto, lo specifico meccanismo di mancanza dei dati [21]; diviene quindi necessario formulare ipotesi basate sulla conoscenza del dato considerato nello specifico e sulle fonti di acquisizione di quest'ultimo.

2.4 Imputare dati di metilazione

Il processo di imputazione dei dati di metilazione presenta diverse problematicità. Queste derivano principalmente da caratteristiche specifiche legate alla struttura e alla natura biologica dei dati di metilazione. Di seguito verrà approfondito come queste caratteristiche influenzano il processo di imputazione rendendolo complesso da organizzare e gestire.

2.4.1 Caratteristiche dei dati di metilazione

I dati di metilazione sono estrapolati dal processo di metilazione del DNA che coinvolge milioni di sequenze di due nucleotidi (citosina e guanina, da cui la sigla CpG)

presenti nel DNA stesso. Per molti di questi siti, se coinvolti nel processo di metilazione, possono essere raccolti questi tipi di dato. L'ottenimento dei dati di metilazione può essere ostacolato da numerosi fattori, ad esempio legati a difficoltà di misurazione relativamente ad alcune posizioni dei siti CpG nel genoma, i quali possono essere soggetti ad errori maggiori.
Tutto questo impatta sull'organizzazione e la composizione dei set di dati di metilazione che, ad esempio, si differenziano a seconda del formato dei dati di metilazione, che a loro volta dipendono dalle piattaforme di sequenziamento utilizzate per la raccolta dei dati stessi.

I dati di metilazione sono inoltre influenzati da diversi fattori legati al soggetto analizzato. Tra questi ci sono il sesso e l'età cronologica del paziente, ma ancora più importanti, il tessuto analizzato e lo stato del paziente (se in salute o malato). Sarà dunque fondamentale che i set di dati di metilazione siano accompagnati da queste informazioni, definite come metadati. Ulteriore caratteristica dei set di dati di metilazione sono le grandi dimensioni e la complessità, entrambi influenzati dalla copertura del genoma e dal numero di campioni utilizzati. Le notevoli variazioni in termini di dimensioni complicano l'analisi e la conseguente progettazione di metodi di imputazione genericamente efficaci poiché si necessita di ricoprire casistiche strettamente dipendenti dalle singole sperimentazioni. Non a caso, non si trovano espliciti riferimenti in letteratura che riescano ad individuare e classificare i dati mancanti relativi a set di dati di metilazione del DNA [7].

2.4.2 Tecniche di imputazione su dati di metilazione

Oltre al già citato *methyLImp*, non sono ancora presenti in letteratura altri metodi di imputazione specificatamente progettati per i dati di metilazione del DNA.

In generale, si nota come tutti i metodi di imputazione che lavorano con variabili continue siano potenzialmente adatti ad essere applicati ai dati di metilazione, poiché questi vengono rappresentati proprio come variabili continue (vedi 2.2). A tal riguardo sono state analizzate e confrontate, oltre a *methyLImp*, alcune tecniche di imputazione (vedi 2.3.2), sotto diverse assunzioni relativamente alla classificazione sul tipo di dato di metilazione mancante (MCAR,MAR,MNAR). In particolare si conclude come, in termini di prestazioni e accuratezza dell'imputazione, *methyLImp* sia il metodo con prestazioni medie migliori [8].

Nel prossimo capitolo approfondiremo proprio *methyLImp*, in particolare nella sua seconda versione: *methyLImp2*

Capitolo 3

Software di Imputazione: methyLImp

In questo capitolo si descriveranno approfonditamente le due versioni del metodo di imputazione di dati di metilazione: *methyLImp* e *methyLImp2*. Entrambe le versioni sono implementate, come pacchetti, in ambiente R. Si inizierà riportando lo pseudocodice della prima versione di *methyLImp*, seguito da una descrizione dell'algoritmo necessaria per comprendere al meglio il metodo ed il funzionamento adottati. Verrà poi approfondita la seconda versione *methyLImp2*, ponendo particolare attenzione sulle differenze rispetto alla sua versione precedente. Questo permetterà nel prosieguo della trattazione di comprendere meglio le motivazioni che hanno portato alla scelta di utilizzare l'ultima versione, piuttosto che la precedente. Successiva alla descrizione dell'algoritmo, è inoltre presente l'analisi computazionale dello stesso, fondamentale per valutarne l'efficienza ed il costo, al fine di migliorarne le prestazioni relativamente a questi aspetti.

3.1 Pseudocodice e Descrizione dell'Algoritmo

Il metodo di *methyLImp* sfrutta le correlazioni esistenti tra i livelli di metilazione di CpG differenti, che possono essere catturati con una semplice regressione lineare sui dati osservabili.

Algorithm 1 Pseudocodice methyLImp

1: **function** METHYLIMP($M \in [0,1]^{n \times m}$)
2: $M' \leftarrow M$
3: $R \leftarrow \{1, \ldots, n\}$ ▷ Indici di tutte le righe
4: $C \leftarrow \{1, \ldots, m\}$ ▷ Indici di tutte le colonne
5: NA $\leftarrow \{c \mid \exists r \in R, M[r,c]$ è mancante$\}$ ▷ Colonne con valori mancanti
6: $L \leftarrow$ NA
7: **while** $L \neq \emptyset$ **do**
8: Seleziona delle $col \in L$
9: $R_{NA} \leftarrow \{r \mid M[r, col]$ è mancante$\}$ ▷ Righe con valori mancanti nella colonna selezionata
10: $C_{NA} \leftarrow \{c \mid M[r,c]$ è mancante $\Leftrightarrow r \in R_{NA}\}$
11: **if** $R \setminus R_{NA} \neq \emptyset$ and $C \setminus$ NA $\neq \emptyset$ **then**
12: $A \leftarrow M[R \setminus R_{NA}, C \setminus \text{NA}]$
13: $B \leftarrow M[R \setminus R_{NA}, C \setminus C_{NA}]$
14: $X \leftarrow M[R_{NA}, C \setminus \text{NA}]$
15: $M'[R_{NA}, C_{NA}] \leftarrow \text{logit}^{-1}(X[A^{-1}\text{logit}(B)])$
16: **end if**
17: $L \leftarrow L \setminus C_{NA}$
18: **end while**
19: **return** M'
20: **end function**

Assumiamo di avere in input una matrice contenente livelli di metilazione espressi in notazione β-value, di dimensione n×m dove "n" è il numero di campioni (sulle righe R) e "m" è il numero di siti CpG (sulle colonne C).

Inizialmente si identificano tutte le colonne contenenti valori mancanti (riga 5) e si assegnano ad L (riga 6). Poi, per ogni iterazione, si selezionano delle colonne (riga 8) e si raggruppano tutte le variabili mancanti che seguono uno stesso modello di distribuzione. In particolare si vanno a costruire due set:

- R_{NA} che raggruppa tutte le righe contenenti valori mancanti in un determinato gruppo delle colonne L identificate inizialmente (riga 9).
- C_{NA} che raggruppa tutte le colonne che contengono valori mancanti date le righe appartenenti ad R_{NA} (riga 10).

L'obiettivo di *methyLImp* è di stimare tutti i valori mancanti simultaneamente, per ognuno dei gruppi di variabili appena definiti, costruendo come risultato una sotto-matrice di valori imputati.

Per ottenerla, l'algoritmo applica una regressione lineare simultanea su tutte le colonne appartenenti a C_{NA}, risolvendo la seguente formula (riga 15):

$$valori\_imputati \leftarrow \text{logit}^{-1}(X[A^{-1}\text{logit}(B)]) \tag{3.1}$$

dove:

- A è una sotto-matrice della matrice di partenza in input (M). Essa contiene tutte le variabili non mancanti indicizzate sulle righe appartenenti ad R, a meno delle righe contenute in R_{NA}, e sulle colonne C, a meno di quelle contenute in C_{NA} (riga 12).
- X è una sotto-matrice della matrice di partenza. Essa contiene tutte le variabili non mancanti indicizzate sulle righe appartenenti ad R_{NA} (riga 14)
- B è una sotto-matrice della matrice di partenza. Essa contiene tutte le variabili indicizzate sulle colonne appartenenti a C_{NA} e sulle righe non in R_{NA} (riga 13).
- logit(p) = log(p/(1-p)), p $\in$ [0,1] e la sua inversa:
  logit^{-1}(q)= 1/(1-exp(-q)), q $\in$ (-∞,+∞).
  Queste formule sono utilizzate per assicurare che i valori imputati siano nell'intervallo [0,1], rispettando la notazione β-value della matrice di input M.

L'esecuzione si ripete per un certo numero di iterazioni finché ci sono colonne contenenti valori mancanti in L(riga 7).

methyLImp ritorna infine la matrice M', inizialmente uguale alla matrice in input contenente valori mancanti M (riga 2) e, durante la computazione, modificata andando a sostituire ogni valore mancante con il corrispondente valore imputato.

3.2 Analisi Costo Computazionale

Il costo computazionale di *methyLImp* dipende dal numero di gruppi, appartenenti alla matrice di input, che presentano lo stesso modello di distribuzione dei valori mancanti, il quale determina direttamente il numero di problemi di regressione lineare da risolvere.

Il costo è inoltre dipendente dalle dimensioni n e m della matrice di input (M). Esse, come già accennato in precedenza (vedi 2.4.1), possono essere molto grandi se si ha a che fare con dati di metilazione di un genoma ampio.

Detto K il numero di gruppi con stesso modello di distribuzione ed n e m le dimensioni della matrice di input, sapendo che queste ultime sono generalmente grandi (vedi tabella 4.2) e comprendendo come questo influenzi anche il valore K in maniera direttamente proporzionale (si suppone che più grande sia la matrice in input e più gruppi K si andranno a determinare), si capisce come il problema sia computazionalmente complesso.

3.3 *methyLImp2*

La seconda versione di *methyLImp* è implementata, proprio come la prima, come un pacchetto R chiamato *methyLImp2*, testato su più piattaforme (Windows, Mac, Linux) e disponibile su GitHub al seguente link:
`https://github.com/annaplaksienko/methyLImp2`

Algorithm 2 Prototipo *methyLImp2*

function *methyLImp2*(
input,
type = c("450K", "EPIC", "user"),
annotation = NULL,
range = NULL,
groups = NULL,
skip_imputation_ids = NULL,
ncores = NULL,
minibatch_frac = 1,
minibatch_reps = 1
)

Nelle sezioni riportate sopra, si è osservato come, in particolare su matrici di grandi dimensioni, *methyLImp* possa avere problemi relativamente al tempo di esecuzione, pur mantenendo buone prestazioni in termini di precisione dei valori stimati. Per questo motivo, si è deciso di crearne una nuova versione allo scopo di ridurne proprio il tempo di esecuzione e possibilmente migliorarne la precisione. *methyLImp2* presenta, a tale scopo, nuove funzionalità rispetto alla versione precedente che verranno di seguito approfondite permettendo di capirne al meglio il suo funzionamento. Queste funzionalità verranno poi sfruttate per ottenere i risultati in successive fasi di test (vedi 5.3).

3.3.1 Parallelizzazione sui cromosomi

In informatica, la parallelizzazione è un metodo che permette l'esecuzione simultanea del codice sorgente di un dato programma su più istanze di calcolo (*es.* microprocessori, più core di uno stesso processore ecc.) allo scopo di aumentarne le prestazioni.

La parallelizzazione del calcolo dei valori stimati è una funzionalità introdotta in *methyLImp2* che permette una gestione molto più efficiente dei set di dati di metilazione notoriamente caratterizzati da un alto numero di variabili.

La parallelizzazione in *methyLImp2* si basa sulla presenza di correlazioni tra livelli di metilazione dei siti CpG all'interno di un cromosoma. Queste correlazioni, ben note in letteratura, sono dovute principalmente alla distribuzione spaziale della cromatina e del DNA [22].
Su queste premesse, dato un set di dati di metilazione, si è deciso di suddividere i siti CpG in base ai rispettivi cromosomi di appartenenza, reperendo queste informazioni dal sito dell'azienda Illumina.

Formati i sotto-gruppi di CpG corrispondenti ai cromosomi, si è andato ad applicare su ognuno di essi l'algoritmo *methyLImp*, già descritto in precedenza (vedi 3.1). Come è intuibile, questa suddivisione permette di ottenere matrici con un numero di colonne minore se comparato a quelle del set di dati di partenza. Su di esse l'applicazione di *methyLImp* risulterà in tempi di calcolo nettamente inferiori [17].

Le computazioni svolte su ognuno di questi sotto-gruppi avvengono in parallelo mediante l'utilizzo di un pacchetto R chiamato *BiocParallel* [3]. In particolare, poiché i cromosomi differiscono notevolmente in termini di dimensioni, si è scelto di distribuire il carico computazionale su tutte le risorse di calcolo (core); di conseguenza questa distribuzione non è predeterminata e definita staticamente ma, al contrario, è dinamica ovvero ogni compito viene assegnato, durante l'esecuzione, al primo core disponibile.

Dal lato utente, questa funzionalità di parallelizzazione su cromosomi è disponibile impostando l'argomento *type* di *methyLImp2* con valori che specificano il tipo di dato considerato (450K, EPIC) oppure, come si farà in fase di test (vedi 5.3), impostando l'argomento a *user*. Esso permette all'utente di definire, in un ulteriore argomento detto *annotation*, i sotto-gruppi di CpG, associando questi ultimi ai vari cromosomi sui quali si baserà il calcolo parallelo.

Infine, relativamente alla parallelizzazione, attraverso l'argomento *ncores* si permette all'utente di specificare il numero di core da poter utilizzare in parallelo. Se non specificato in modo esplicito, questo numero verrà impostato automaticamente come il numero di core disponibili meno uno.

3.3.2 Mini-batch

Il metodo Mini-batch è un'altra funzionalità introdotta con *methyLImp2*. Anch'essa, come la parallelizzazione sui cromosomi, è stata implementata allo scopo di ridurre i tempi di calcolo del processo di imputazione dei dati di metilazione mancanti senza intaccarne l'accuratezza.

Questo approccio risulta particolarmente utile quando si considerano set di dati di metilazione con un gran numero di campioni [17] (nell'ordine delle centinaia).
Lo pseudocodice di questo algoritmo è presentato di seguito.

Algorithm 3 Pseudocodice algoritmo Mini-batch

1: **function** MINI-BATCH(data X, sample size n)
2: $P \leftarrow 10/20/30$ ▷ Percentuale di campioni scelti
3: $R \leftarrow 1/2/3$ ▷ Numero di ripetizioni
4: **for** r=1 to R **do**
5: $X_P \leftarrow$ Campione casuale di X di dimensione n/P*100
6: $Imp_r \leftarrow methyLImp(X_P)$
7: **end for**
8: Imp $\leftarrow \sum_r Imp_r/R$ ▷ Calcolo la media dei risultati su R
9: **end function**

Mini-batch si applica allo scopo di ridurre le dimensioni delle matrici A e B (vedi 3.1). Si noti come, nello pseudocodice 3 sopra riportato, non è esplicitata la fase preliminare (vedi 3.1), nella quale il set di dati di metilazione in input X viene suddiviso nelle sotto-matrici A,B e C ,così da permettere di concentrarsi sul funzionamento dell'algoritmo Mini-batch senza mostrare concetti già visti in precedenza nel paragrafo 3.1.

L'algoritmo mini-batch viene applicato per ogni regressione lineare successivamente alla suddivisione della matrice in input nelle rispettive sotto-matrici. In particolare, avendo lo scopo di ridurre la dimensione delle sotto-matrici A e B, l'algoritmo selezionerà solo una percentuale P dei campioni in C (riga 2). Successivamente selezionerà casualmente dei campioni, basandosi sulla percentuale scelta P, e su di

essi applicherà una regressione lineare. Questo passaggio verrà ripetuto un numero R di volte (riga 3). Si calcolerà, infine, la media dei risultati ottenuti dalle varie iterazioni per ricavare il valore imputato (riga 8). Si noti inoltre come, nel caso in cui le righe della sotto-matrice A siano minori della quantità dei campioni iniziali n diviso il prodotto tra la percentuale di campioni scelti P e 100, ovviamente non verranno attuate alcune sotto-selezioni di campioni su A (riga 5).

Dal lato utente, questa funzionalità mini-batch è utilizzabile mediante gli argomenti *minibatch_frac* e *minibatch_reps*. Sono entrambi valori numerici e corrispondono rispettivamente alla percentuale di campioni che verranno selezionati P (di default impostato a 1) ed al numero di ripetizioni dell'imputazione R sui sotto-gruppi di campioni selezionati.

3.3.3 Confronto con *methyLImp*

Si intuisce, già dalle precedenti sezioni, come *methyLImp2*, confrontato con la sua versione precedente, sia andato a migliorare in particolare l'aspetto del tempo di calcolo dell'imputazione. Quest'ultimo risultava essere un fattore problematico per *methyLImp* in particolar modo quando si aveva a che fare con matrici in input di grandi dimensioni.

Queste funzionalità hanno inoltre permesso un maggior controllo da parte dell'utente utilizzatore della funzione *methyLImp2* che può potenzialmente definire i gruppi (o cromosomi) assegnandogli diversi siti CpG o, ad esempio, specificare il numero di core da utilizzare durante la computazione.
Tutto ciò ha permesso di migliorare le problematicità che *methyLImp* presentava, consentendo inoltre di introdurre un nuovo approccio al processo di imputazione di dati di metilazione.

Capitolo 4

Dataset di Riferimento e Setup

I dati estrapolati dal processo di metilazione del DNA vengono memorizzati all'interno di set di dati che successivamente vengono pubblicati e resi disponibili ad essere analizzati e trattati. Per questa trattazione è necessario avere a disposizione alcuni di questi set di dati di metilazione sui quali poter testare *methyLImp2*.

In questo capitolo si tratteranno le diverse fasi che hanno portato ad avere un set di dati pronto per essere utilizzato nei test, con all'inizio la presentazione delle piattaforme online dove questi set di dati sono scaricabili, per passare poi a descriverne la loro struttura, che, vedremo essere non sempre coerente, e per finire col descrivere le varie fasi del processo di selezione dei set di dati più adatti ad essere utilizzati in fase di test. Inoltre si approfondiranno le varie statistiche scelte e successivamente utilizzate nella fase di analisi e confronto dei risultati ottenuti dai test.

4.1 Banche Dati Biologiche: Biostudies, GEO

Una banca dati biologica è genericamente definita come un'infrastruttura informatica che raccoglie, archivia e distribuisce dati biologici geneticamente correlati e, insieme, fornisce strumenti per analizzare tali dati.

Esistono diverse tipologie di banche dati che, relativamente alla qualità dei dati in esse contenuti, si dividono in due classi principali:

- Banche dati primarie: raccolgono quotidianamente informazioni che riguardano biomolecole prodotte in tutti i laboratori del mondo e rendendole disponibili.

- Banche dati secondarie: esaminano i dati contenuti nelle banche dati primarie e ne correggono eventuali errori includendo informazioni aggiuntive. Rendono poi disponibili i risultati ottenuti.

In questa tesi è stata consultata principalmente una banca dati biologica di tipo primario detta GEO (Gene Expression Omnibus) [10].

I registri di dati in GEO sono organizzati come: registri di piattaforma, registri di campioni e registri di serie. Questi ultimi sono quelli che verranno estrapolati e poi opportunamente selezionati per le fasi di test. Un registro di serie è composto in GEO da un gruppo di campioni correlati e, ad ognuno di essi, è assegnato un identificativo univoco detto GEO Accession Number, espresso come segue: *GSExxx*.

I registri di serie possono inoltre contenere: tabelle che descrivono sommariamente i dati in essi contenuti, informazioni sul tipo di sperimentazione dalla quale questi dati sono stati estratti, informazioni sulla piattaforma utilizzata (ad esempio EPIC, 450K ecc.) ed informazioni sui campioni presenti.
A sua volta, ogni campione presente nella serie sarà associato ad un identificativo univoco nella forma *GSMxxx*.

4.2 Struttura Dataset

Di seguito si descriverà la struttura di partenza dei set di dati che verranno utilizzati per i test di *methyLImp2*. Come già accennato, questi set di dati non hanno una struttura standard ben definita, di conseguenza si cercherà di descrivere le caratteristiche strutturali più comuni ai vari set considerati.

Associato ad ogni registro di serie (vedi 4.1), si hanno diversi dati a disposizione. Quello di maggior interesse, in questo contesto, è il set di dati contenente i dati di metilazione. Questo, però, nella maggior parte dei casi non contiene esclusivamente i dati di metilazione propriamente detti ma anche altre informazioni relative ai campioni esaminati. Queste informazioni, insieme ai dati di metilazione, saranno importanti nel processo di analisi dei dati ma, al momento, ci si concentrerà solo sulla struttura dei set di dati di metilazione, lasciando il resto delle informazioni ad un approfondimento successivo (vedi 4.3.1).

Generalmente il set di dati contenente i dati di metilazione è una matrice di grandi dimensioni. In particolare i dati di metilazione sono indicizzati su un certo numero di campioni (variabili a seconda dell'esperimento effettuato) e un certo numero di

CpG, che per i set di dati considerati in questo contesto, varieranno dalle 700.000 alle 800.000 unità circa. Gli identificativi dei campioni sono solitamente riportati sulla prima riga della matrice mentre gli identificativi delle CpG sono riportati sulla prima colonna. È da notare come questa disposizione sia richiesta anche per la matrice in input su entrambe le versioni di *methyLImp*. Quando questa struttura non viene rispettata sarà dunque necessario operare su una matrice di dati di metilazione trasposta rispetto a quella di partenza.

Ora che si è dettagliatamente spiegato come sia strutturato un set di dati di metilazione, si descriverà in seguito come questo venga modificato al fine di arrivare ad ottenere una matrice di dati di metilazione adatta ad essere utilizzata in fase di test.

4.3 Processi di Analisi Dati

Una volta scaricati i set di dati di metilazione, si inizia il processo di analisi e selezione di questi ultimi allo scopo di capire quali siano i più adatti ad essere utilizzati in fase di test. Questo processo si divide in due fasi fondamentali: l'estrazione dei metadati e la standardizzazione dei set di dati.

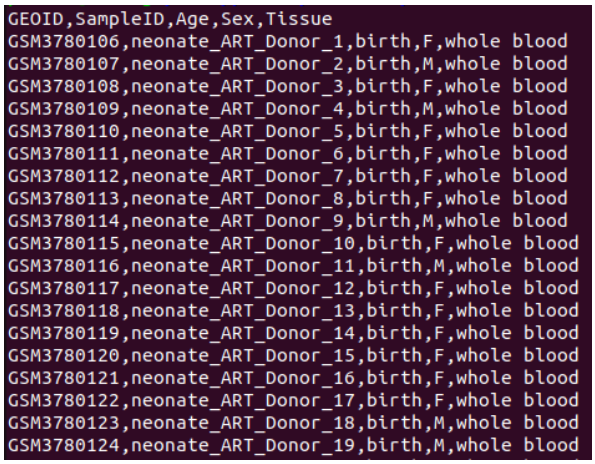
Per comprendere al meglio questo processo si porterà, come esempio, il set di dati GSE131433, poi utilizzato anche in fase di test.

4.3.1 Estrazione Metadati

Con il termine metadati si intende indicare tutte le informazioni, non strettamente riferite ai dati di metilazione, che permettono di effettuare un'analisi degli stessi a partire dal campione dai quali sono stati estrapolati. Tra queste informazioni sono presenti:

- GEO Accession Number: utilizzato per l'identificazione del campione al quale si fa riferimento. Esso permette inoltre di notificare la presenza di eventuali campioni duplicati.
- Sesso: utilizzato come informazione per meri scopi statistici.
- Età: espressa come età cronologica o come status (ad esempio adulto, neonato ecc.). Essa permette di individuare la presenza di eventuali differenze considerevoli, in termini di età appunto, tra campioni considerati.
- Tessuto: specifica il tessuto dal quale i dati di metilazione sono stati estrapolati. Esso permette di capire se la serie considerata include tessuti diversi e quindi dati di metilazione non correlati tra loro.

Il primo passo, una volta estratte queste informazioni, è quello di mantenerle in matrici indicizzate seguendo una struttura predefinita.
Si riporta di seguito un esempio della struttura di una matrice contenente metadati.



```
GEOID,SampleID,Age,Sex,Tissue
GSM3780106,neonate_ART_Donor_1,birth,F,whole blood
GSM3780107,neonate_ART_Donor_2,birth,M,whole blood
GSM3780108,neonate_ART_Donor_3,birth,F,whole blood
GSM3780109,neonate_ART_Donor_4,birth,M,whole blood
GSM3780110,neonate_ART_Donor_5,birth,F,whole blood
GSM3780111,neonate_ART_Donor_6,birth,F,whole blood
GSM3780112,neonate_ART_Donor_7,birth,F,whole blood
GSM3780113,neonate_ART_Donor_8,birth,F,whole blood
GSM3780114,neonate_ART_Donor_9,birth,M,whole blood
GSM3780115,neonate_ART_Donor_10,birth,F,whole blood
GSM3780116,neonate_ART_Donor_11,birth,M,whole blood
GSM3780117,neonate_ART_Donor_12,birth,F,whole blood
GSM3780118,neonate_ART_Donor_13,birth,F,whole blood
GSM3780119,neonate_ART_Donor_14,birth,F,whole blood
GSM3780120,neonate_ART_Donor_15,birth,F,whole blood
GSM3780121,neonate_ART_Donor_16,birth,F,whole blood
GSM3780122,neonate_ART_Donor_17,birth,F,whole blood
GSM3780123,neonate_ART_Donor_18,birth,M,whole blood
GSM3780124,neonate_ART_Donor_19,birth,M,whole blood
```

Figura 4.1: GSE131433.metadata.csv

Una volta organizzati correttamente i metadati, questi verranno analizzati, categoria per categoria, al fine di individuare se presentano differenze o caratteristiche che possono avere un impatto sui rispettivi dati di metilazione. In particolare, differenze impattanti saranno, ad esempio, la presenza di tessuti differenti nello stesso set di dati, o la presenza di fasce di età variabili, o ancora, la presenza di duplicati negli identificativi dei campioni.
Ognuno di questi fattori comporta la suddivisione del set di dati in sottogruppi dello stesso che, dal momento della suddivisione, vengono considerati come dataset separati e distinti. In questa trattazione questi sottogruppi saranno distinguibili dall'aggiunta, all'identificativo del set di dati, di lettere alfabetiche (in particolare A e B) ognuna relativa ad uno specifico sottogruppo.
Caso distinto è quello della presenza di duplicati che approfondiremo in fase di standardizzazione (vedi 4.3.2).

Ad esempio, il dataset GSE131433 presentava, nei metadati, campioni categorizzati come adulti e campioni categorizzati come neonati. Questa è un evidente differenza di età che può avere un impatto sulle prestazioni in fase di imputazione dei dati di metilazione. Si è quindi suddiviso il set di dati in due sottogruppi contenenti, rispettivamente, gli adulti e i neonati, come si vede nella tabella 4.1 riportata di seguito.

Tabella 4.1: Informazioni relative ai sottogruppi del set di dati di partenza GSE131433

GEOID	CAMPIONI	CpGs	ETÀ	TESSUTO
GSE131433_A	233	722,292	Adulti	Sangue
GSE131433_B	207	722,292	Neonati	Sangue

L'estrazione dei metadati è una fase propedeutica alla successiva fase di standardizzazione del dataset che verrà presentata di seguito.

4.3.2 Standardizzazione Dataset

La fase di standardizzazione è in realtà composta di più momenti distribuiti in tutto il processo di estrazione ed analisi dei set di dati. Come già accennato, infatti, al momento della raccolta dei dati, questi vengono inizialmente trattati al fine di individuare ed estrapolare la matrice contenente i dati di metilazione e, a loro volta, estrapolare i metadati che vengono mantenuti in una matrice con struttura standard prefissata.

Successivamente si sfruttano le osservazioni fatte sui metadati ed in particolare sulla presenza di duplicati. Questi, per le differenti caratteristiche dei set di dati grezzi reperiti dalle banche dati biologiche, non sono facili da individuare. Oltre ai metadati, vengono infatti studiate le informazioni presenti sulle banche dati stesse (vedi 4.1), in particolare relativamente alla sperimentazione effettuata, al fine di individuare la presenza di duplicati e le modalità con le quali si è deciso di esprimere gli identificativi degli stessi.
Per questa trattazione si è previsto, una volta individuati con certezza i duplicati, di eliminarli poiché avrebbero influenzato negativamente la fase di imputazione di *methyLImp2* che avrebbe dovuto potenzialmente imputare dati di metilazione dello stesso soggetto più volte, aumentando inevitabilmente il tempo di calcolo e potenzialmente peggiorando la qualità dei valori stimati.

4.4 Criteri di Selezione

Una volta completata l'analisi dei dati e quindi ottenuto un certo insieme di set di dati, eventualmente suddivisi in sottogruppi e con struttura comune, si sono utilizzati dei criteri di selezione su di essi per capire quali effettivamente fossero utilizzabili per i test.

In primo luogo, si sono distinti i set di dati contenenti già dati mancanti e i set di dati che non ne contenevano (poiché si erano già adoperati a priori metodi di imputazione in fase di sperimentazione). Si approfondiranno le motivazioni rispetto a questa scelta nel seguito della trattazione (vedi 5.1).

Successivamente, sono stati selezionati i set di dati che contenevano un sufficiente numero di campioni e CpG. Rispettivamente, i primi dovevano essere almeno superiori a 100, mentre i secondi quantomeno superiori a 700'000. Questo è fondamentale per garantire la validità delle statistiche (vedi 4.5) che, rispettando i criteri appena descritti, saranno calcolate su set di dati tra loro coerenti permettendo un confronto dei risultati adeguato. Si noti che molti sottogruppi di dataset, andatisi a creare in fase di standardizzazione, sono risultati troppo piccoli relativamente alle dimensioni sopra specificate e, di conseguenza, sono stati scartati riducendo la dimensione del gruppo di set di dati utilizzabili nei test.

4.5 Metriche di Valutazione

Al fine di valutare e confrontare le prestazioni del processo di imputazione svolto da *methyLImp2*, sono state utilizzate 5 metriche di valutazione: RMSE, MAE, PCC, MAPE e tempo di esecuzione. Esse verrano presentate di seguito.

Lo scarto quadratico medio (RMSE) è una metrica utilizzata per misurare la differenza, e quindi l'errore, che intercorre tra il valore stimato (P) e il valore reale (T). RMSE è definito dalla seguente formula:

$$RMSE(P,T) = \sqrt{\frac{\sum_{i=1}^{n}(P_i - T_i)^2}{n}} \tag{4.1}$$

Similmente al RMSE, si è scelto di utilizzare anche l'errore assoluto medio (MAE). Quest'ultimo è una metrica che indica la differenza, e quindi l'errore, in valore assoluto tra il valore stimato (P) e il valore reale (T).
MAE è definita dalla seguente formula:

$$MAE(P,T) = \frac{\sum_{i=1}^{n} \|P_i - T_i\|}{T} \tag{4.2}$$

Le due metriche, appena descritte, sono sostanzialmente equivalenti, essendo complementari. Si sceglie di utilizzarle entrambe per delle loro caratteristiche specifiche,

infatti, RMSE dà maggior peso a errori di grandi dimensioni rispetto al MSE, risultando più adatto nel classificare la qualità di una prestazione e permettendo di dare maggior evidenza a stime pessime che, nel contesto dell'imputazione di dati di metilazione, sono particolarmente problematiche e indesiderate. Si sceglie però di utilizzare anche MSE poiché fornisce un' interpretazione più immediata dei risultati rispetto a RMSE ed in particolare indica l'errore medio che, prevedibilmente, sarà presente sui valori imputati.

Un'altra metrica utilizzata è il cosiddetto coefficiente di correlazione di Pearson (PCC). Esso misura la correlazione lineare tra due variabili continue che, come detto, sono le variabili con cui si ha a che fare nel processo di imputazione, ed è definito dalla seguente formula:

$$PCC(P,T) = \frac{COV(P,T)}{DevSt(P) * DevSt(T)} \tag{4.3}$$

Il PCC ha valori compresi tra -1 e 1 che rispettivamente indicano nel primo caso una correlazione prefetta negativa, ovvero, a valori elevati per una variabile corrispondono valori bassi per l'altra e viceversa, mentre nel secondo caso indica una correlazione perfetta positiva, ovvero a valori elevati per una variabile corrispondo valori elevati per l'altra, e viceversa. Inoltre una correlazione uguale a zero indica che tra le due variabili non vi è alcuna correlazione. Nel contesto di questa trattazione, si cercherà di ottenere valori di PCC il più possibili vicino al valore 1 poiché indicheranno una buona qualità di imputazione dei valori imputati rispetto a quelli reali.

Nell'analisi delle prestazioni, si utilizzerà anche l'errore medio assoluto percentuale (MAPE) che indica, come nel caso del MAE, l'errore assoluto che intercorre tra i valori stimati (P) e quelli reali (T). Esso è utile poiché esprime questo errore non in unità della variabile ma in unità percentuali permettendo un'analisi ed un confronto delle prestazioni anche sotto questa forma.
Il MAPE è definito dalla seguente formula:

$$MAPE(P,T) = \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{T_i - P_i}{T_i}\right\| \tag{4.4}$$

Si terrà infine conto del tempo di esecuzione dell'algoritmo. Quest'ultimo è un parametro fondamentale per comprendere se i test e le modifiche effettuate por-

tino un miglioramento effettivo nella prestazioni, non in termini di accuratezza dell'imputazione, quanto in termini di efficienza.

4.6 Set di dati di riferimento

Svolte le fasi di estrazione, standardizzazione e selezione appena descritte, si arriva a definire l'insieme dei set di dati che verranno utilizzati in fase di test (vedi 5). Questi sono riportati nella tabella 4.2 che mostra, per ogni dataset, le seguenti informazioni: identificativo assegnato automaticamente (ID), identificativo assegnato dalla banca dati biologica dalla quale il set di dati è stato estratto (GEOID), numero di campioni (#Samples), numero di Cpg (#CpGS), media dell'età dei campioni appartenenti al set di dati (μAge) e tessuto analizzato.

Tabella 4.2: Set di dati di riferimento

ID	GEOID	#Samples	#CpGS	μ Age	Tissue
D1	GSE128821	542	706,323	NA	Buccal swab
D2	GSE131433_A	233	722,292	Adult	Blood
D3	GSE131433_B	207	722,292	Neonate	Blood
D4	GSE144129_A	227	706,435	31	Placenta
D5	GSE144129_B	227	706,435	31	Placenta
D6	GSE150643_A	118	797,603	11	Saliva
D7	GSE150643_B	118	797,603	13	Saliva
D8	GSE151732_A	128	754,231	57	Colon
D9	GSE151732_B	128	754,231	57	Colon

Capitolo 5

Test e Risultati

In questo capitolo verranno descritti i test effettuati prima sulla versione base di *methyLImp2* e in seguito sulla versione modificata dello stesso. Successivamente, una volta ottenuti e analizzati i risultati, si commenterà e si valuterà l'accuratezza e l'efficienza computazionale rilevata.

I test sono stati effettuati su un totale di 9 set di dati di metilazione (vedi tabella 4.2) sui quali sono stati introdotti dati mancanti. Quest'ultima è una fase propedeutica ai test e permette di facilitare la successiva analisi dei risultati (vedi 5.1).

5.1 Procedura di Simulazione dei Dati mancanti

La procedura di simulazione dei dati mancanti consiste, essenzialmente, nell'inserire, all'interno dei set di dati considerati, una certa percentuale di dati mancanti.
In questa trattazione le percentuali da inserire sono state estrapolate da un sottogruppo di set di dati venutosi a creare durante la fase di estrazione degli stessi (vedi 4.4). In particolare questo sottogruppo contiene tutti i set di dati che, pur avendo caratteristiche adatte per essere utilizzati in fase di test, presentavano già dei dati mancanti e dunque sono stati scartati (non fanno parte dei dataset di riferimento che verranno utilizzati nei test).
I set di dati appartenenti a questo sottogruppo sono 6 e in tabella 5.1 sono descritte le loro principali caratteristiche.

Tabella 5.1: Set di dati con valori mancanti

ID	GEOID	#Samples	#CpGS	μ Age	Tissue	%NA	%NACpG	NASamples
D1	GSE116339	679	816,999	54	Blood	530,063 (0.096)	53,253 (6.52)	679 (100)
D2	GSE122408	180	866,895	NA	PBMC	153,064 (0.098)	113,723 (13.12)	180 (100)
D3	GSE137682	152	265,087	NA	Buccal cells	47,981 (0.119)	33,121 (12.49)	152(100)
D4	GSE147040	221	866,091	46	NAc	9,903 (0.005)	6,309 (0.73)	221 (100)
D5	GSE149318	180	709,521	35	Blood/Sperm	1,524,304 (1.194)	17,244 (2.43)	180 (100)
D6	GSE147740	1134	761,652	41	PBMC	9,276,799 (1.074)	86,325 (11.33)	1134 (100)
Media						0.43%	7.77%	100%

Tramite questi set di dati si è potuto calcolare una distribuzione di valori mancanti da simulare sui set di dati di riferimento. Così facendo, si è ottenuta una distribuzione che rispecchia quella realmente presente su un set di dati di metilazione generico.

Dalla media delle percentuali di valori mancanti presenti rispettivamente su tutto il dataset, tutti i CpG e tutti i campioni, si è calcolato che su ogni set di dati di riferimento verranno inseriti lo 0.43% di valori mancanti ristretti al 7.77% delle CpG e al 100% dei campioni (vedi tabella 5.1).

Nella pratica di questo processo, si sono andati a selezionare casualmente i valori reali da sostituire con dati mancanti, nelle quantità specificate dalle percentuali sopra definite, avendo cura di salvare i valori reali per il successivo confronto.

Infine, per rendere l'analisi ancor più solida, poiché i dati mancanti sono inseriti casualmente all'interno dei set di dati, si è scelto di ripetere questa procedura di simulazione di valori mancanti per 5 volte, ottenendo così, per ogni set di riferimento, 5 set di dati con percentuale fissa di valori mancanti ma distribuita in modo differente su ognuno di questi set.
Dunque, seppur come già detto in precedenza, i set di dati di riferimento sono 9, per ogni test effettuato si sono utilizzati nella pratica 45 set di dati differenti. I risultati derivanti da questi test saranno quindi, per ogni dataset, la media dei risultati ottenuti dai 5 set estrapolati da quello di partenza ognuno dei quali avrà una disposizione differente dei valori mancanti al loro interno.

5.2 Prima Fase di Test: *methyLImp2*

La prima fase di test consiste nell'imputare i dati mancanti dei set di dati di metilazione di riferimento presentati in tabella 4.2. Su di essi sono stati opportunamente simulati dati mancanti nelle percentuali specificate nel paragrafo precedente.

Come già anticipato, i test si sono svolti su 5 repliche di ogni dataset di riferimento, ognuna distinta per la distribuzione dei valori mancanti contenuti, i quali sono stati simulati, su ognuna, rispettando le percentuali prefissate.

Su ogni replica dei vari dataset è stato eseguito *methyLImp2* nelle modalità più basilari utilizzando tutte le CpG per l'imputazione di ogni dato di metilazione mancante. In particolare si è specificato esclusivamente il tipo di dato dei set di dati considerati, ovvero *EPIC*, che fa riferimento ad una tecnica di analisi della metilazione del DNA basata su tecnologia Illumina (vedi 2.1). Inoltre il numero di core utilizzati per l'imputazione è stato impostato a 9 (come nell'articolo di riferimento [17]). Allo stesso modo è stato impostato automaticamente il range dei valori imputati, definito tra 0 e 1, coerentemente con la notazione in β-Value dei dati di metilazione considerati (vedi 2.2).
Per la maggior parte dei set di dati è stato inoltre necessario calcolarne la trasposta per poter applicare *methyLImp2*. Quest'ultimo passaggio è dovuto al fatto che l'algoritmo lavora esclusivamente su set di dati in input con i campioni indicizzati sulle righe e i siti CpG indicizzati sulle colonne.

Durante i test sono stati salvati, per ogni set di dati, la media dei dati statistici e del tempo di esecuzione (in secondi) associati ad ogni replica imputata. Questi dati sono rispettivamente presentati nelle tabelle 5.2 e 5.3. Sono stati inoltre memorizzati i dati imputati stessi per permettere un confronto più immediato tra questi ultimi e i valori reali e per assicurare la mancanza di errori nel calcolo automatico dei dati statistici durante l'esecuzione. Quest'ultimi, infatti, sono stati opportunamente ricalcolati e confrontati direttamente sui valori reali e i valori imputati.

Tabella 5.2: Statistiche imputazione dei valori mancanti con la versione base di *methyLImp2*

ID	GEOID	#Samples	μRMSE	μMAE	μPCC	μMAPE
D1	GSE128821	542	0.043	0.025	0.993	10.270
D2	GSE131433_A	233	0.025	0.015	0.997	5.258
D3	GSE131433_B	207	0.041	0.027	0.992	7.931
D4	GSE144129_A	227	0.038	0.022	0.993	7.668
D5	GSE144129_B	227	0.028	0.017	0.826	6.797
D6	GSE150643_A	118	0.029	0.016	0.997	6.911
D7	GSE150643_B	118	0.029	0.016	0.997	6.862
D8	GSE151732_A	128	0.033	0.018	0.996	7.144
D9	GSE151732_B	128	0.031	0.018	0.996	6.868
Media		214	0.033±0.006	0.019±0.004	0.976±0.053	7.301±1.261

Tabella 5.3: Tempi di imputazione dei valori mancanti con la versione base di *methyLImp2*

ID	GEOID	#Samples	μTime (sec.)
D1	GSE128821	542	33,152
D2	GSE131433_A	233	14,040
D3	GSE131433_B	207	12,723
D4	GSE144129_A	227	13,065
D5	GSE144129_B	227	13,238
D6	GSE150643_A	118	9,344
D7	GSE150643_B	118	9,471
D8	GSE151732_A	128	9,045
D9	GSE151732_B	128	9,028
Media		214	13,678±7,152

In generale si può subito notare come, utilizzando *methyLImp2* nelle sue funzioni più basilari, i tempi di imputazione siano direttamente proporzionali alla dimensione dei campioni considerati, similmente a quanto avviene nella sua versione precedente methyLImp.

5.3 Seconda Fase di Test: *methyLImp2* modificato

La seconda fase di test consiste nell'imputare dati di metilazione mancanti sui set di dati di riferimento sfruttando alcune delle funzionalità fornite da *methyLImp2*.

In particolare, per i test svolti in questa fase, si è sfruttata la funzionalità di parallelizzazione su cromosomi (vedi 3.3.1). Essa, come già spiegato in precedenza, consiste nell'associare gruppi di CpG a singoli cromosomi.
Questa associazione può avvenire automaticamente, ovvero *methyLImp2* si rifarà ad informazioni prese da terzi (sito di Illumina) per individuare quali raggruppamenti di CpG corrispondano effettivamente ad un certo cromosoma, oppure l'associazione tra CpG e cromosomi può esser definita dall'utente. In quest'ultimo caso è impreciso parlare di parallelizzazione sui cromosomi. Questo perché i gruppi di CpG che vengono associati ad un cromosoma non vanno realmente a costituire quest'ultimo ma sono appunto definiti, a priori, dall'utente senza rispettare le effettive composizioni dei cromosomi.
Si può dunque dedurre come i cromosomi, che si andranno a definire in questa fase di test, vadano ad identificare un sottogruppo di CpG specifico che non ha alcuna corrispondenza nella realtà ma che è stato definito dall'utente, come nel caso di

questa trattazione, per fini di analisi e valutazione delle prestazioni di *methyLImp2*.

Prima di presentare nel merito le scelte operate in questa fase di test relativamente ai set di dati è opportuno chiarire in che modo questi sottogruppi di CpG vengano definiti ed associati ai vari cromosomi.
La procedura consiste nell'estrarre tutti i CpG contenuti nel set di dati considerato e si inseriscono in una tabella composta da due colonne denominate rispettivamente *cpg* e *chr*. Come è intuibile nella prima colonna si andranno ad inserire i CpG estrapolati dal set di dati mentre, nella seconda, si andranno a definire i cromosomi, i quali sono identificati singolarmente con il nome "chr*n*", dove *n* è il numero del gruppo andatosi a definire. Dunque ad ogni CpG in tabella che si trova sulla colonna cpg, corrisponderà il cromosoma (identificativo del gruppo) di appartenenza che si trova sulla colonna chr.

Si sono svolti due test separati, distinti dalle modalità con le quali si sono andati a suddividere la totalità dei CpG relativi ad ogni singolo set di dati.
In un primo momento si sono andati a suddividere i CpG relativamente ai processori a disposizione sulla macchina utilizzata per svolgere i test. Quest'ultima ha a disposizione 9 processori, di conseguenza sono stati definiti 9 gruppi distinti formati da CpG assegnatigli casualmente. Una volta eseguito *methyLImp2* sui set di dati di metilazione specificando la suddivisione in questi 9 sottogruppi di CpG, l'algoritmo ha eseguito in parallelo su ognuno di essi la funzione *methyLImp* (vedi 3.1). I risultati non vengono direttamente riportati in questa trattazione. Ci limitiamo infatti a citare questo primo test poiché non ha ottenuto risultati soddisfacenti relativamente alle prestazioni e all'accuratezza dell'imputazione. Infatti i dati statistici relativi all'accuratezza dell'imputazione e i tempi di calcolo rimangono sostanzialmente invariati se confrontati con i risultati della prima fase di test (vedi 5.2).

Nel secondo test, i CpG sono stati suddivisi in gruppi (cromosomi) in relazione alla dimensione di quest'ultimi. In particolare, ogni gruppo conterrà circa 10.000 CpG selezionati casualmente. Dunque a differenza del test descritto appena sopra, il numero di gruppi definiti sarà nettamente superiore al numero di processori utilizzati per imputarne i dati mancanti, ma i gruppi presi singolarmente avranno dimensioni ristrette poiché ognuno sarà composto da una percentuale di CpG molto bassa, sul totale considerato per un certo set di dati. Essa si attesta attorno al 1,43% per i set di dati contenenti circa 700.000 CpG mentre attorno al 1,45% per i set di dati contenenti circa 800.000 CpG.

I risultati relativi a questo secondo test sono presentati nelle tabelle 5.4 e 5.5, riportate di seguito.

Tabella 5.4: Statistiche imputazione valori mancanti su gruppi di 10.000 CpG con *methyLImp2*

ID	GEOID	#Samples	μRMSE	μMAE	μPCC	μMAPE
D1	GSE128821	542	0.046	0.026	0.992	10.949
D2	GSE131433_A	233	0.026	0.016	0.997	5.440
D3	GSE131433_B	207	0.041	0.027	0.991	8.093
D4	GSE144129_A	227	0.040	0.023	0.992	8.182
D5	GSE144129_B	227	0.031	0.017	0.797	7.289
D6	GSE150643_A	118	0.031	0.016	0.996	7.202
D7	GSE150643_B	118	0.031	0.016	0.996	7.150
D8	GSE151732_A	128	0.034	0.019	0.996	7.348
D9	GSE151732_B	128	0.032	0.018	0.996	7.046
Media		214	0.034±0.006	0.019±0.004	0.873±0.281	7.633±1.385

Tabella 5.5: Tempi di imputazione dei valori mancanti su gruppi di 10.000 CpG con *methyLImp2*

ID	GEOID	#Samples	μTime (sec.)
D1	GSE128821	542	17,512
D2	GSE131433_A	233	6,217
D3	GSE131433_B	207	5,425
D4	GSE144129_A	227	6,038
D5	GSE144129_B	227	6,052
D6	GSE150643_A	118	3,674
D7	GSE150643_B	118	3,696
D8	GSE151732_A	128	3,746
D9	GSE151732_B	128	3,699
Media		214	6,228±4,130

5.4 Valutazione Accuratezza ed Efficenza computazionale

In questa sezione si vogliono confrontare i risultati ottenuti nelle due fasi di test per valutare quali effetti abbia avuto sull'accuratezza e sull'efficienza computazionale del processo di imputazione l'applicazione della funzionalità di parallelizzazione su cromosomi di *methyLImp2* applicata nella seconda fase di test.
Essendo questo confronto fondamentalmente basato sui dati statistici e i tempi di imputazione raccolti, si riportano di seguito delle tabelle che affiancano i risultati rispettivamente della prima e della seconda fase di test. Così facendo si potranno cogliere con maggior facilità le eventuali differenze riscontrate sui dati ottenuti.

Tabella 5.6: Confronto media statistiche test 1 e test 2

MEDIA	TEST1	TEST2
RMSE	0.033±0.006	0.034±0.006
MAE	0.019±0.004	0.019±0.004
PCC	0.976±0.053	0.873±0.081
MAPE	7.301±1.261	7.633±1.385

Relativamente alle medie dei dati statistici, presentati in tabella 5.6, si può osservare come i valori ottenuti abbiano subito piccole variazioni generalmente tendenti ad indicare un peggioramento dell'accuratezza dell'imputazione. Essendo però, come detto, variazioni minime, eventuali perdite di accuratezza possono essere considerate trascurabili. La perdita minima di accuratezza è in ogni caso dovuta alla riduzione della dimensione delle matrici sulle quali *methyLImp2* ha svolto il processo di imputazione. Quest'ultima ha infatti conseguentemente diminuito il numero di dati di metilazione sui quali *methyLImp2* si è basato per stimare i vari dati di metilazione mancanti.

Tabella 5.7: Confronto tempi di imputazione test 1 e test 2

ID	GEOID	TEST1 μTime (sec.)	TEST2 μTime (sec.)
D1	GSE128821	33,152	17,512
D2	GSE131433_A	14,040	6,217
D3	GSE131433_B	12,723	5,425
D4	GSE144129_A	13,065	6,038
D5	GSE144129_B	13,238	6,052
D6	GSE150643_A	9,344	3,674
D7	GSE150643_B	9,471	3,696
D8	GSE151732_A	9,045	3,746
D9	GSE151732_B	9,028	3,699
Media		13,678±7,152	6,228±4,130

Come si può vedere nella tabella 5.7 riportata appena sopra, con la suddivisione in gruppi di 10.000 Cpg si è ottenuto un netto miglioramento in termini di velocità di imputazione. Infatti, con la strategia testata, i tempi medi di calcolo di *methyLImp2* sono stati ridotti da circa 4 ore a poco più 1 ora e 30 minuti.

Tra i principali motivi dietro questo risultato vi sono le correlazioni esistenti tra livelli di metilazione di CpG differenti distribuite casualmente in tutto il genoma.

Suddividendo infatti l'imputazione su gruppi di 10,000 CpG si è riusciti a cogliere un alto numero di queste correlazioni. Esse sono state sufficienti ad ottenere le stesse performance in termini di accuratezza dell'imputazione se confrontate con il test precedente.

Ulteriori motivazioni relativamente ai risultati ottenuti sono riconducibili al già citato impatto negativo che le matrici di grandi dimensioni hanno sui tempi di imputazione di *methyLImp* (vedi 3.2). Dunque suddividere i CpG appartenenti ad ogni set di dati in gruppi di dimensioni ridotte ha permesso di eseguire *methyLImp* effettuando inversioni su matrici con dimensioni molto piccole se comparate a quelle che tipicamente contengono dati di metilazione e che sono utilizzate sia nel caso del test iniziale sia nella prima versione di *methyLImp*, la quale considera per ogni imputazione tutte le CpG. Ciò non deve però portare a pensare, erroneamente, che basti ridurre il più possibile la dimensione dei gruppi di CpG per diminuire i tempi di imputazione. Infatti, se da un lato questo può sembrare logico, dall'altro non si deve sottovalutare il peggioramento in termini di accuratezza dei valori imputati che questo approccio può comportare. Basti ricordare che *methyLImp* (vedi 3.1) basa l'imputazione dei dati di metilazione su regressioni lineari effettuate utilizzando sotto-matrici della matrice di partenza (ad ogni gruppo di CpG corrisponderà una matrice di partenza distinta), le quali sono ottenute a partire da un sotto-insieme di colonne della matrice di partenza stessa. Se quest'ultima ha dimensioni molto ridotte, saranno a loro volta numericamente ridotte le sue colonne perciò la regressione lineare, che andrà ad imputare il valore mancante, baserà questo calcolo su pochissimi dati, catturando poche correlazioni tra i vari CpG e, di conseguenza, peggiorando inevitabilmente l'accuratezza dell'imputazione.

Si può dunque dedurre come basare la suddivisione dei CpG a partire dalla dimensione dei sottogruppi che si vanno a creare, sia un approccio potenzialmente migliorativo del processo di imputazione, con un impatto particolarmente positivo relativamente ai tempi di calcolo. Nel definire la dimensione di questi sottogruppi vi è però un trade-off tra velocità di imputazione e accuratezza dei dati di metilazione stimati. Se, infatti, questa dimensione viene portata a valori molto piccoli, si rischia di ottenere un accuratezza dei dati imputati insufficiente e, viceversa, se si costituiscono questi sottogruppi basandosi sul numero di processori della macchina utilizzata o, più generalmente, se si costituiscono gruppi di CpG di dimensioni molto grandi, si rischia di non ottenere alcun vantaggio dall'utilizzo della funzionalità di parallelizzazione di *methyLImp2* poiché i tempi di imputazione risulterebbero molto elevati.

Capitolo 6

Conclusioni

In questa trattazione si è ottenuto un miglioramento prestazionale del processo di imputazione di dati di metilazione svolto da *methyLImp2*, in particolar modo relativamente alla riduzione del tempo di calcolo.
Ciò è stato stato possibile applicando e testando la funzionalità di parallelizzazione sui cromosomi fornita da *methyLImp2* su diversi set di dati dai quali sono state generate repliche contenenti dati di metilazione mancanti distribuiti casualmente in determinate percentuali. La possibilità di organizzare i dati in sottogruppi predefiniti ha permesso di ridurre il fattore della dimensionalità delle matrici di dati di metilazione soggette al processo di imputazione. Si può osservare come *methyLImp2* sia adatto anche ad imputare matrici di grandi dimensioni in modo efficiente a differenza del predecessore *methyLImp*.
Il numero di 10,000 CpG, sul quale si è basata la costituzione dei sotto-gruppi, è stato scelto in maniera empirica. Sarebbe opportuno quindi testare sperimentalmente quale dimensione sia più adatta nell'ottenere il miglior trade-off tra velocità e accuratezza di imputazione.
Inoltre rimane aperta la possibilità di approfondire ancor di più questa tecnica di imputazione sfruttando, ad esempio, le altre funzionalità fornite dalla stessa, come la tecnica mini-batch che in questa trattazione ci si è limitati a presentare e descrivere. Inoltre, potrebbe essere interessante ampliare il set di dati di riferimenti per avere un attendibilità delle analisi sui risultati ancora maggiore.
Si rileva, in ogni caso, quanto potenziale dimostri *methyLImp2* nell'ambito dell'imputazione dei dati di metilazione e quanto margine di miglioramento si possa ottenere dai test della stessa per poter arrivare a svolgere stime sui dati di metilazione mancanti sempre più accurate e veloci così da arricchire ulteriormente la ricerca e la sperimentazione relativa alla metilazione del DNA.

Riferimenti bibliografici

[1] L M R B Arantes, A C de Carvalho, M E Melendez, A L Carvalho, and E M Goloni-Bertollo. Methylation as a biomarker for head and neck cancer. *Oral Oncol.*, 50(6):587–592, June 2014.

[2] Marina Bibikova, Zhenwu Lin, Lixin Zhou, Eugene Chudin, Eliza Wickham Garcia, Bonnie Wu, Dennis Doucet, Neal J. Thomas, Yunhua Wang, Ekkehard Vollmer, Torsten Goldmann, Carola Seifart, Wei Jiang, David L. Barker, Mark S. Chee, Joanna Floros, and Jian Bing Fan. High-throughput dna methylation profiling using universal bead arrays. *Genome research*, 16(3):383–393, March 2006.

[3] Bioconductor Package Maintainer [cre], Martin Morgan [aut], Valerie Obenchain [aut], Michel Lang [aut], Ryan Thompson [aut]. BiocParallel, 2017.

[4] Ozren Bogdanović and Ryan Lister. DNA methylation and the preservation of cell identity. *Curr. Opin. Genet. Dev.*, 46:9–14, October 2017.

[5] Ozren Bogdanović and Ryan Lister. DNA methylation and the preservation of cell identity. *Curr. Opin. Genet. Dev.*, 46:9–14, October 2017.

[6] Annalisa Ciabattini, Christine Nardini, Francesco Santoro, Paolo Garagnani, Claudio Franceschi, and Donata Medaglini. Vaccination in the elderly: The challenge of immune changes with aging. *Semin. Immunol.*, 40:83–94, December 2018.

[7] Pietro Di Lena, Claudia Sala, Andrea Prodi, and Christine Nardini. Missing value estimation methods for DNA methylation data. *Bioinformatics*, 35(19):3786–3793, 02 2019.

[8] Pietro Di Lena, Claudia Sala, Andrea Prodi, and Christine Nardini. Methylation data imputation performances under different representations and missingness patterns. *BMC Bioinformatics*, 21, 06 2020.

[9] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11(1):587, November 2010.

[10] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–210, January 2002.

[11] Montserrat Fuentes, Peter Guttorp, and Paul Sampson. Using transforms to analyze space-time processes. In *C&H/CRC Monographs on Statistics & Applied Probability*, pages 77–149. Chapman and Hall/CRC, October 2006.

[12] Steve Horvath. DNA methylation age of human tissues and cell types. *Genome Biol.*, 14(10):R115, 2013.

[13] Steve Horvath and Kenneth Raj. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.*, 19(6):371–384, June 2018.

[14] Julie Josse and François Husson. Handling missing values in exploratory multivariate data analysis methods. *Journal de la société française de statistique*, 153(2):79–99, 2012.

[15] Diljeet Kaur, Sol Moe Lee, David Goldberg, Nathan J Spix, Toshinori Hinoue, Hong-Tao Li, Varun B Dwaraka, Ryan Smith, Hui Shen, Gangning Liang, Nicole Renke, Peter W Laird, and Wanding Zhou. Comprehensive evaluation of the infinium human MethylationEPIC v2 BeadChip. *Epigenetics Commun.*, 3(1), September 2023.

[16] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, 11:2287–2322, March 2010.

[17] Anna Plaksienko, Pietro Di Lena, Christine Nardini, and Claudia Angelini. methyLImp2: faster missing value estimation for DNA methylation data. *Bioinformatics*, 40(1), January 2024.

[18] Claudia Sala, Pietro Di Lena, Danielle Fernandes Durso, Andrea Prodi, Gastone Castellani, and Christine Nardini. Evaluation of pre-processing on the meta-analysis of DNA methylation data from the illumina HumanMethylation450 BeadChip platform. *PLoS One*, 15(3):e0229763, March 2020.

[19] Daniel J Stekhoven and Peter Bühlmann. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, January 2012.

[20] O Troyanskaya, M Cantor, G Sherlock, P Brown, T Hastie, R Tibshirani, D Botstein, and R B Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, June 2001.

[21] Stef van Buuren. *Flexible imputation of missing data, second edition.* Taylor & Francis, London, England, 2 edition, September 2021.

[22] Yingying Zhang, Christian Rohde, Sascha Tierling, Tomasz P Jurkowski, Christoph Bock, Diana Santacruz, Sergey Ragozin, Richard Reinhardt, Marco Groth, Jörn Walter, and Albert Jeltsch. DNA methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution. *PLoS Genet.*, 5(3):e1000438, March 2009.

Ringraziamenti

Ringrazio il Professor Di Lena per la professionalità e la pazienza dimostratami durante tutta la stesura della tesi. Ringrazio inoltre i miei amici Luca, Stefano, Filippo, Francesca, Federico, Enrico, Simone, Giulio, Cosimo, Marco, Leonardo, Alessio, Lorenzo, Daniele, Edoardo, Simone e Francesco per avermi, ognuno a suo modo, accompagnato ed aiutato in tutto il percorso universitario.