Università degli Studi di Padova
Facoltà di Ingegneria

Corso di Laurea Magistrale in Bioingegneria

Tesi di laurea magistrale

Analisi modellistica di dati high-throughput di spettrometria di massa per la quantificazione del turnover proteico

Candidato:
Gloria Pasqualetto

Relatore:
Prof.ssa Gianna Maria Toffolo

Correlatore:
Prof.ssa Barbara Di Camillo

Anno Accademico 2012-2013

“Più contempli un pericolo, meno ti piacerà.
Affrontalo con decisione e ti accorgerai
che non è poi così brutto come sembra”.

Robert Baden-Powell

INDICE

1 MODELLO DEL TURNOVER PROTEICO 1
- 1.1 Il turnover delle proteine 1
- 1.2 Assunzioni di base 1
- 1.3 Modello di sintesi e degradazione delle proteine 2
- 1.4 Stima dell'emivita 5

2 DIABETE E NEFROPATIA 7
- 2.1 Nefropatia diabetica 7
  - 2.1.1 Ereditarietà della nefropatia 8
  - 2.1.2 L'espressione proteica nella nefropatia diabetica 8

3 ACQUISIZIONE DELLE MISURE 11
- 3.1 Protocollo SILAC DINAMICO 11
  - 3.1.1 Coltura cellulare 11
  - 3.1.2 Estrazione e purificazione delle cellule 12
  - 3.1.3 Lisi cellulare 12
  - 3.1.4 Elettroforesi monodimensionale 12
  - 3.1.5 Taglio del gel 13
  - 3.1.6 Digestione delle proteine 13
  - 3.1.7 Spettrometria di massa 14
  - 3.1.8 Dati forniti dal software 14

4 IDENTIFICAZIONE DEI PARAMETRI DEL MODELLO 17
- 4.1 Stima con i minimi quadrati pesati 17
  - 4.1.1 Stima con 'pesi relativi' 19
  - 4.1.2 Residui 20
  - 4.1.3 Precisione delle stime 21
- 4.2 Implementazione 22

5 FILTRAGGIO 25
- 5.1 Scelta dei tempi di campionamento 25
- 5.2 Variabilità tecnica delle misure 29
- 5.3 Prefiltraggio e proteine considerate 34
  - 5.3.1 Prefiltraggio 34
  - 5.3.2 Soggetti e proteine considerate 35
- 5.4 Analisi peptidi 37
- 5.5 Normalizzazione dei dati 38

6 ANALISI DELL'ESPRESSIONE DIFFERENZIALI: METODI 43
- 6.1 Test d'ipotesi 43

6.1.1 Test di Student su campioni indipendenti 43
6.1.2 Test di normalità di Shapiro-Wilk 46
6.1.3 Correzione per test multipli 47
6.1.4 Implementazione 48
6.2 GSEA 48
6.2.1 Metodo 49
6.2.2 Settaggio dei parametri per l'analisi 50
6.2.3 Risultati forniti 51

7 ANALISI DELL'ESPRESSIONE DIFFERENZIALE: RISULTATI 53
7.1 Identificazione del parametro 53
7.2 Test statistici: risultati 54
7.3 GSEA: risultati 58

8 CONCLUSIONI 67
8.1 Sviluppi futuri 68

BIBLIOGRAFIA 69

SOMMARIO

Un'alterazione a livello cellulare può comportare variazioni nel metabolismo delle proteine al suo interno; questo può determinare sia mutamenti del loro livello di espressione sia del loro turnover. Non necessariamente però andando ad analizzare il solo livello di espressione si riesce a identificare variazioni del turnover: la concentrazione di una proteina, infatti, è determinata dal bilancio tra la sua degradazione e la sua sintesi, ma esso può mantenersi invariato pur variando singolarmente questi due processi.
In questa tesi ci si è concentrati sull'analizzare la velocità di turnover delle proteine in fibroblasti cutanei al fine di individuare se per alcune di esse la degradazione fosse significativamente diversa in 2 classi di soggetti: diabetici di tipo 1 affetti o meno anche da nefropatia diabetica. L'obiettivo è infatti stato quello di evidenziare dei biomarkers per tale complicazione, che permettessero, non solo di diagnosticarla in uno stato avanzato, ma anche di prevederla prima della comparsa.
Per questo studio si sono avute a disposizione misure high-throughput di spettrometria di massa, a partire dalle quali si è formulato, identificato e validato un modello del turnover a livello di singola proteina. In questo modo si è potuto stimare il parametro che quantifica la velocità di degradazione di ogni proteina in ogni soggetto.
È stato poi applicato il test di Student al fine di discriminare le proteine per cui la velocità di degradazione fosse significativamente diversa nelle 2 classi di soggetti. Dovendo considerare molte proteine e avendo a disposizione pochi soggetti, si è preferito concentrarsi non tanto su singole proteine ma su gruppi di esse, che avessero, se pur bassa, una coerente espressione differenziale. Questo è stato fatto attraverso la GSEA (Gene Set Enrichment Analysis) che considera set di proteine accomunate da una stessa caratteristica (es. appartenenza allo stesso pathway biologico, condivisione della stessa funzione cellulare...).
È emerso che la velocità di degradazione è significativamente maggiore per la classe dei soggetti diabetici e nefropatici, rispetto ai diabetici senza tale complicazione, nelle proteine legate ai ribosomi (costituenti dei ribosomi stessi o appartenenti a pathway che li coinvolgano); essa invece è significativamente minore in pathway legati all'attività dei proteosomi. Questi risultati sembrano essere coerenti in quanto entrambi evidenziano una maggior attività di degradazione di tutte le proteine a livello cellulare nei pazienti affetti da nefropatia diabetica rispetto ai soggetti diabetici dove tale complicazione non si è riscontrata.

RINGRAZIAMENTI

Alla fine di questo percorso, desidero ringraziare delle persone che sono state per me molto importanti durante questi anni. Scriverò solo poche righe ma spero che riescano a trasmettere ad ognuno il messaggio che vorrei arrivasse.

La mia famiglia: **i miei genitori** in primis che mi hanno permesso di raggiungere questo traguardo, sostenendomi moralmente e materialmente e non avendomi mai fatto mancare consigli, aiuti e rassicurazioni. **Aligi**, che durante questi mesi mi ha fatto divertire quando, con le sue domande e osservazioni, ha cercato di capire cosa stessi facendo. La **zia Elda** che è sempre stata interessata e partecipe a ogni esame e ogni novità della mia vita universitaria.

Alvise: che mi ha supportato e sopportato in ogni momento, trasmettendomi calma e stemperando i momenti di nervosismo. Sul quale ho sempre potuto contare e senza del quale sarebbe stato tutto molto più difficile. Un grazie veramente grandissimo.

Gli amici dell'università: con i quali ho condiviso buona parte di questa avventura universitaria e in cui ho sempre trovato aiuto, scambio e fiducia. Grazie alle brioches di **Ilaria** che hanno rallegrato molte mattine e alle giornate passate a fare homeworks e progetti (che insieme sono stati decisamente più leggeri da affrontare). Grazie ad **Angela** per aver condiviso, tra gli altri, questo ultimo periodo di tesi: senza le 'ricapitolazioni burocratiche' probabilmente sarei ancora a compilare scartoffie. Grazie a **Marco**, che è stato compagno di molti esami, e con cui è sempre bello passare qualche ora ad aggiornarsi sulle rispettive novità. Grazie ad **Alessandro** con cui ho passato molte giornate in aula computer e che ha ascoltato pazientemente le gioie e i dolori dei risultati della mia tesi.

Alvi, Matteo e Nicola: che, pur essendo (chi più chi meno) lontani, e pur vedendosi poco, mi hanno sempre dimostrato grande amicizia: sono sempre stati pronti ad ascoltare e consigliare, ma anche a ridere e scherzare. Alvi, che si è pazientemente sorbito lunghe spiegazioni su spettrometria e proteine; Matteo, che si è sempre preoccupato di sentire come procedeva il lavoro, dispensando consigli su come contrastare l'agitazione; Nicola, che quando torna a casa, tra

le mille cose da fare, trova sempre un angolino per fare due chiacchiere assieme.

Alle professoresse **Gianna Toffolo**, **Barbara di Camillo**, a **Lucia Puricelli** e a **Giorgio Arrigoni** del VIMM e al professor **Paolo Tessari** del DMCS che mi hanno permesso di intraprendere questo lavoro, aiutandomi a portarlo a termine, e da cui ho imparato moltissimo.

INTRODUZIONE

La proteomica differenziale è una branca della proteomica che ha come obiettivo la determinazione dell'espressione differenziale delle proteine o in cellule diverse o nella stessa tipologia di cellule ma in differenti condizioni (per esempio prima e dopo l'insorgenza di una malattia o in fenotipi diversi). In questo modo, tra le altre cose, si possono individuare i biomarkers che identificano una determinata patologia o un fenotipo.
In questa tesi si sono analizzati dati di spettrometria di massa della concentrazione relativa (non assoluta) di proteine provenienti da fibroblasti di pazienti diabetici di tipo 1 e di pazienti in cui si sia diagnosticata anche la nefropatia diabetica, al fine di capire se alcune proteine, o gruppi funzionali di esse, potessero essere considerate biomarkers di tale malattia. La nefropatia diabetica è infatti la principale complicazione del diabete, ed è particolarmente grave, in quanto può portare alla morte del paziente. Ad oggi esistono metodi di diagnosi di essa, ma solo in fase avanzata: prevedere tale malattia o rilevarla in fase precoce potrebbe servire ad una cura più efficace.
Sono già stati fatti studi con lo stesso obiettivo, ma si sono concentrati sull'analisi dei livelli cellulari assoluti di espressione proteica. In essa però non sono riflesse tutte le possibili alterazioni che il sistema può subire, in quanto non dà indicazione su come siano variate le velocità di sintesi e degradazione della proteina, ma solo sul risultato del loro bilancio (che è l'espressione stessa).
È inoltre molto difficile avere dati di proteomica quantitativa: infatti gli esperimenti per ottenerli sono molto costosi e complicati da attuare.
L'obiettivo di questa tesi è quindi stato quello di determinare un modello del turnover delle proteine da applicare ai dati ottenuti dalla spettrometria, al fine di capire se la velocità di degradazione di alcune proteine differisse in maniera significativa tra le 2 classi di soggetti.
Per l'acquisizione dei dati ci si è avvalsi della collaborazione del Dipartimento di Medicina Clinica e Sperimentale (Prof. Paolo Tessari) e del VIMM - Venetian Institute of Molecular Medecine (Dott. Giorgio Arrigoni). È stato utilizzato a tal fine il protocollo SILAC: in esso viene misurato il rapporto tra le proteine sintetizzate dopo l'istante d'inizio dell'esperimento (che saranno marcate con un isotopo stabile) e quelle già presenti prima di tale istante (non marcate). Ci si aspetta che queste ultime si degradino nel tempo, mentre le prime siano soggette sia a degradazione che a sintesi.
Si sono avuti a disposizione i dati di 10 soggetti: 5 diabetici e 5 diabetici affetti anche da nefropatia, e per ognuno è stato misurato il rapporto tra le proteine marcate e non in 3 istanti temporali: 4 h, 7.5 h e 24 h. Oltre alla misura globale di ogni proteina lo spettrometro ha fornito anche le misure di tutti i peptidi in essa contenuti (da cui si ricava attraverso una media pesata il dato globale della

proteina). Di conseguenza si ha avuto a che fare con una grande mole di dati: per rendere l'idea, di ogni paziente si sono misurate circa 1000 proteine per ognuna delle quali si hanno le misure di circa 10 peptidi, in 3 istanti temporali. È stato quindi necessario, prima dell'identificazione vera e propria del parametro del modello (velocità di degradazione), assicurarsi dell'affidabilità dei dati, e rendere minima l'influenza dell'errore di misura. A tal fine si sono attuate una serie di operazioni di preprocessing dei dati.
In primo luogo i dati sono stati normalizzati, in modo da eliminare l'errore sistematico introdotto a causa del fatto che gli esperimenti sulle cellule dei differenti soggetti sono avvenuti in momenti diversi. Ci si è quindi assicurati di poter usare direttamente i dati globali di proteina, e non quelli dei peptidi, senza perdita di precisione. Sono state poi filtrate le proteine per cui l'andamento contraddicesse il modello assunto. Infine, si è dovuto formulare un modello per l'errore di misura: come verrà spiegato in seguito i modelli classici (a SD o CV costante) si sono ritenuti non adatti, optando per un compromesso tra i 2. L'identificazione del parametro (implementata con il software R) è stata fatta attraverso la stima con il metodo dei minimi quadrati pesati, usando come peso il modello dell'SD determinato. Ci si è quindi preoccupati di validare il modello andando a considerare la precisione delle stime.
Per l'analisi successiva, delle proteine per cui si sono stimati i parametri, si sono considerate solo quelle per cui si avessero i parametri stimati per almeno 3 soggetti diabetici e 3 soggetti diabetici nefropatici: questo è infatti il numero minimo necessario per poter attuare il test d'ipotesi di Student. Considerando però un numero elevato di proteine contro un numero molto più basso di soggetti, si è preferito concentrarsi non tanto sulle singole proteine, ma su gruppi di esse, in cui si riscontrasse una seppur bassa, coerente espressione differenziale. Questo è stato fatto attraverso la GSEA (Gene Set Enrichment Analysis), che analizza set di proteine che condividano una determinata funzione cellulare o appartengano alla stesso pathway biologico, selezionando quelli per cui si riscontri una significativa differenza nella velocità di degradazione nelle 2 classi di soggetti.

Nello specifico, il primo capitolo di questa tesi è dedicato al modello utilizzato per la descrizione del turnover proteico, evidenziando le ipotesi di base che l'esperimento permette di fissare e la formalizzazione matematica dei processi coinvolti.
I capitoli 2 e 3 illustrano il background su cui questa tesi si inserisce; in primo luogo viene fatta una breve panoramica sul diabete e sulla sua degenerazione in nefropatia diabetica, citando anche precedenti studi che si sono interessati all'analisi dell'espressione proteica nei 2 casi; si procede poi con la descrizione del protocollo utilizzato per l'acquisizione dei dati, necessario per comprendere le scelte fatte nella loro successiva elaborazione.
Nel capitolo 4 viene invece descritto sia a livello teorico che implementativo il metodo dei minimi quadrati pesati applicato per identificare i parametri del

modello, su cui si baserà la successiva analisi.
Nel capitolo 5 vengono descritte dettagliatamente tutte le operazioni di filtraggio: la scelta dei tempi di campionamento e della variabilità tecnica delle misure (modello dell'errore di misura), il filtraggio delle proteine con misure non conformi al modello assunto e la normalizzazione dei dati.
Il capitolo 6 è dedicato ai metodi usati nel processing dei dati: vengono quindi illustrati i test statistici utilizzati e la Gene Set Enrichment Analysis che permette di evidenziare gruppi funzionali proteici associati ai 2 fenotipi considerati. Essi vengono descritti sia dal punto di vista teorico che a livello implementativo.
Il capitolo 7 illustra i risultati dell'analisi differenziale delle proteine considerate e, infine, nel capitolo 8 vengono riportate le conclusioni di tale studio e le prospettive di sviluppo.

1 MODELLO DEL TURNOVER PROTEICO

1.1 IL TURNOVER DELLE PROTEINE

All'interno di ogni cellula si ha un continuo ricambio delle proteine, grazie a processi di degradazione e nuova sintesi; essi fan sì di mantenere la concentrazione proteica a valori costanti, o di rispondere a bisogni momentanei. Il livello di espressione di ogni proteina può essere quindi determinato dal bilancio tra la produzione della proteina (a seguito di trascrizione e traduzione) e la sua distruzione da parte di altre proteine specializzate [1].
Dall'analisi del turnover (degradazione) si riesce a determinare un parametro molto importante: l'*emivita delle proteine*. Esso indica il tempo necessario affinché venga degradata la metà della concentrazione iniziale di una certa proteina (per questo viene anche chiamato tempo di dimezzamento). Proteine con emivita maggiore saranno quindi soggette ad una degradazione più lenta; tanto più alta sarà la velocità di degradazione, tanto più piccolo sarà il valore dell'emivita.
A partire dai dati sulla concentrazione delle proteine in diversi istanti temporali il primo obiettivo è stato quello di trovare un modello che riuscisse a spiegarne il turnover.
L'esperimento attuato per l'acquisizione dei dati verrà illustrato nei dettagli nel capitolo 3, ma è opportuno citarne i passi fondamentali in questa sede per comprendere le scelte fatte nella determinazione del modello.
Le cellule inizialmente poste in un terreno di amminoacidi non marcati (*light*) verranno poi trasferite in un terreno di amminoacidi marcati (*heavy*). Di conseguenza, dall'istante in cui avviene lo scambio del terreno, si potranno distinguere le proteine di nuova sintesi da quelle sintetizzate in precedenza la cui quantità verrà misurata attraverso spettrometria di massa. Quest'ultima fornisce i valori del $\frac{P_H}{P_L}$ (dove P_H è la concentrazione di proteine *heavy* e P_L quella delle proteine light) delle proteine in vari istanti temporali.

1.2 ASSUNZIONI DI BASE

Le ipotesi su cui ci si è basati sono state:

- il turnover proteico può essere descritto attraverso un modello monocompartimentale, Figura 1;

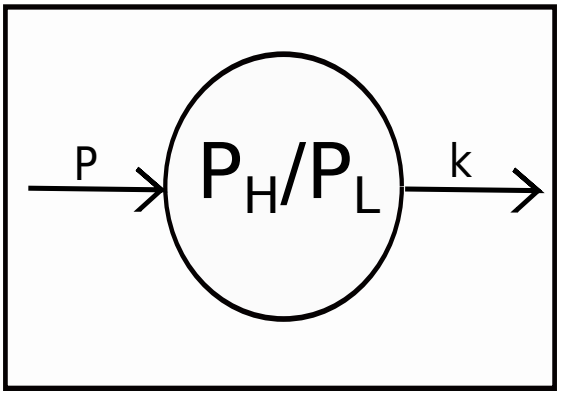


Figura 1: Modello compartimentale del turnover delle proteine: k e P sono rispettivamente la velocità di degradazione (turnover) e la velocità di sintesi.

- le cellule sono in stato stazionario; questo significa che le proteine intracellulari sono caratterizzate da valori costanti di produzione (P), degradazione (k) e quindi di concentrazione. Quest'ultima, nello specifico, in ogni istante sarà $P_{tot} = P_L(t) + P_H(t)$;
- il ricircolo di amminoacidi light (cioè il riuso da parte della cellula degli amminoacidi provenienti dalla degradazione delle proteine sintetizzate prima del cambio di terreno per sintetizzare nuove proteine) viene considerato non significativamente influente e quindi può essere trascurato nell'analisi.

1.3 MODELLO DI SINTESI E DEGRADAZIONE DELLE PROTEINE

Innanzitutto bisogna distinguere i 2 intervalli temporali: prima e dopo l'istante t_0 in cui avviene il cambiamento del terreno di coltura. Quello che succede è descritto in Figura 2: prima di t_0 la coltura è composta solamente da amminoacidi non marcati, dopo essi vengono sostituiti totalmente da quelli marcati.

A partire da tale istante si ha che:

- **Proteine light**: sono interessate solo da degradazione; infatti, non essendo più presenti amminoacidi light, e potendone trascurare il ricircolo, non si avrà nuova sintesi di proteine non marcate. Esse, partendo dal valore iniziale P_{tot} (essendoci all'inizio all'interno della cellula solo proteine light), diminuiscono nel tempo.
- **Proteine heavy**: sono interessate sia da degradazione che da nuova sintesi. Esse inizialmente partiranno da una concentrazione nulla per giungere, all'infinito, ad un valore pari a P_{tot}.

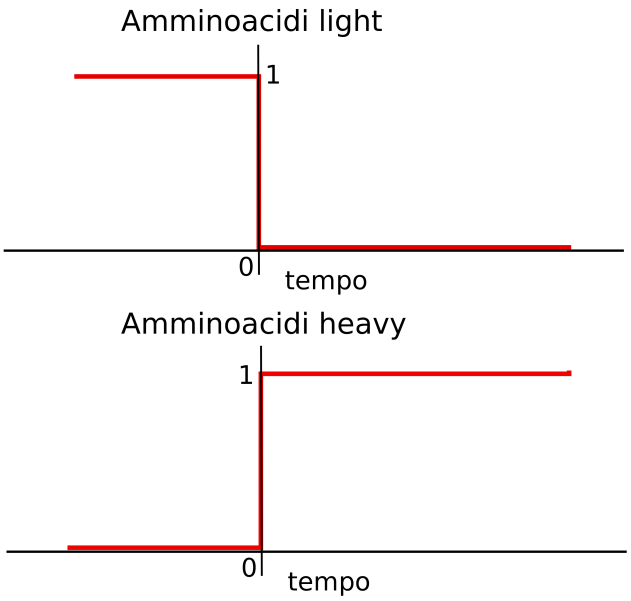


Figura 2: Cambiamento di coltura: all'istante t=0 gli amminoacidi light vengono totalmente eliminati (in alto); essi vengono rimpiazzati da amminoacidi heavy.

Si possono descrivere entrambi i comportamenti con delle equazioni differenziali:

Proteine light:

$$P_L(t)' = -k * P_L(t) \quad \text{con} \quad P_L(0) = \frac{P}{k} = P_{tot} \tag{1}$$

Proteine heavy:

$$P_H(t)' = P - k * P_H(t) \quad \text{con} \quad P_H(0) = 0 \tag{2}$$

dove k e P sono rispettivamente la velocità di degradazione e di sintesi delle proteine. Ovviamente ogni proteina avrà i suoi specifici k e P.
Le equazioni descrivono bene il comportamento delle proteine dall'istante t=0:

- la marcatura degli amminoacidi non altera i processi cellulari, quindi si può assumere che la velocità di degradazione e sintesi nei 2 casi sia la stessa;
- $\frac{P}{k}$ è il bilancio tra la velocità di sintesi e quella di degradazione ed è quindi uguale a P_{tot}, livello a cui si trovano le proteine light subito dopo il cambio di terreno;
- per quanto riguarda quelle non marcate, non sono interessate da nuova sintesi;

- le proteine heavy sono soggette sia a sintesi che a degradazione con le rispettive velocità. All'istante iniziale la loro concentrazione è nulla.

Andando a risolvere tali equazioni differenziali, risulta che:

Proteine light:

$$P_L(t) = \frac{P}{k} * e^{-kt} = P_{tot} * e^{-kt} \tag{3}$$

Proteine heavy:

$$P_H(t) = \frac{P}{k} * (1 - e^{-kt}) = P_{tot} * (1 - e^{-kt}) \tag{4}$$

Il comportamento nel tempo così trovato è in accordo con le ipotesi sopra elencate; infatti P_L si degrada nel tempo (con legge esponenziale), mentre $P_H = P_{tot} - P_L$ cresce esponenzialmente. Nella Figura 3 viene riportato l'andamento di entrambe.

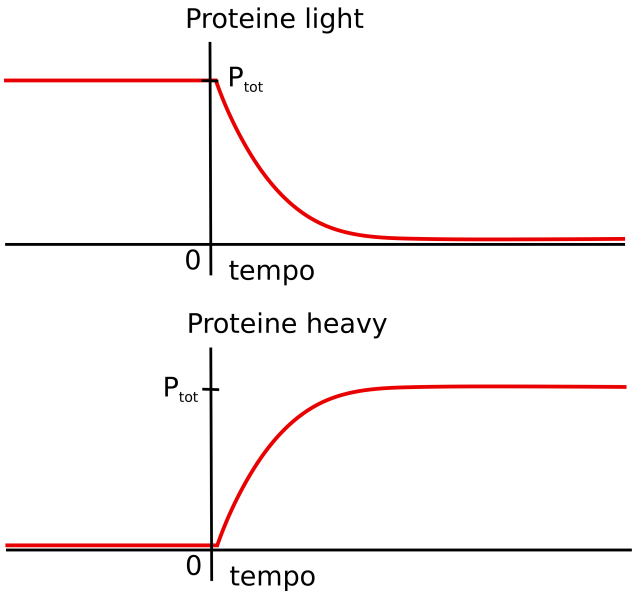


Figura 3: Andamento nel tempo della concentrazione delle proteine light (sopra) e di quelle heavy (sotto).

Avendo ora a disposizione sia il modello delle proteine *heavy* sia di quelle *light* si può ricavare quello del loro rapporto:

$$\frac{P_H(t)}{P_L(t)} = \frac{P_{tot} * (1 - e^{-kt})}{P_{tot} * e^{-kt}} = \frac{(1 - e^{-kt})}{e^{-kt}} \tag{5}$$

che ha un andamento crescente, come riportato in Figura 4.

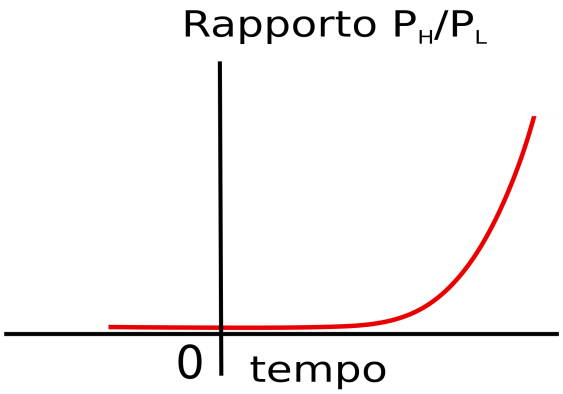


Figura 4: Andamento nel tempo del rapporto tra la concentrazione delle proteine *heavy* e quella delle *light*.

1.4 STIMA DELL'EMIVITA

Grazie a questo modello, e ai campioni ottenuti dall'esperimento, è possibile stimare il parametro k (si veda capitolo seguente per i dettagli) per ogni proteina. Poiché la degradazione proteica ha un andamento esponenziale decrescente con costante di tempo pari a $\frac{1}{k}$, è immediato calcolarne il tempo di dimezzamento:

$$T_{1/2} = \frac{\ln(2)}{k} \tag{6}$$

che coincide con l'emivita della proteina. Emivita e valore del parametro k sono quindi inversamente proporzionali: tanto più piccolo è il k tanto maggiore sarà l'emivita e viceversa.
Nell'analisi, per comodità, ci si è sempre riferiti al k. Di conseguenza l'obiettivo è stato quello di capire se per alcune proteine il k fosse significativamente diverso in soggetti diabetici affetti anche da nefropatia rispetto a quelli solo diabetici.

2 DIABETE E NEFROPATIA

Con il termine *diabete mellito* si intende un disturbo metabolico che si manifesta come un'iperglicemia cronica che comporta delle alterazioni al metabolismo delle macromolecole (carboidrati, grassi, proteine...). Essa è causata dalla mancanza di secrezione dell'insulina o dalla sua inattività, o da entrambe [2]. Tale malattia può portare a danni a lungo termine e disfunzioni, e può compromettere l'attività di vari organi.
Bisogna distinguere due tipi di diabete mellito:

- *Tipo* 1 *(T1DM)*: detto anche insulino-dipendente, è una malattia che colpisce prevalentemente durante l'infanzia e l'adolescenza e dura tutta la vita. È caratterizzato dall'assoluta mancanza di insulina e comporta quindi un alto livello di glucosio nel sangue. I pazienti per poter vivere devono assumere insulina dall'esterno.

- *Tipo* 2 *(T2DM)*: detto anche insulino-indipendente, è caratterizzato da una scarsa produzione di insulina e dalla resistenza alla sua azione da parte dei tessuti periferici. Esso si riscontra principalmente in persone di età adulta, ed è spesso associato ad altri problemi quali obesità, ipertensione, dislipidemia, malattie cardiovascolari. La terapia nella maggior parte dei casi consiste nell'aumento dell'attività fisica e in una dieta equilibrata.

Il diabete è una malattia diffusa: nel 2012 è stato stimato che più di 350 milioni di persone il tutto il mondo ne sono affette. Di esse, il 90% è colpito dal tipo 2, mentre, per quanto riguarda il diabete insulino-dipendente, esso ha un'incidenza minore ma riguarda soprattutto persone giovani. Tale malattia, a causa delle complicanze a livello sistemico che comporta, nel 2012 ha portato alla morte 4.8 milioni di persone [3][4].

2.1 NEFROPATIA DIABETICA

Come è stato detto precedentemente, una delle possibili conseguenze del diabete è il danneggiamento di alcuni organi. Tra questi sono inclusi anche i reni, che, a causa di esso, possono essere afflitti da una malattia cronica chiamata *nefropatia diabetica* (*DN*). Questa complicanza si riscontra sia in persone affette da T1DM che da T2DM.
Per quanto riguarda il T1DM (in questa tesi), essa insorge in circa il 25% dei soggetti con età maggiore di 30 anni e sono svariate le modificazioni che può

causare sia a livello strutturale che a livello funzionale; innanzitutto, l'iperglicemia propria del diabete comporta un'iperfiltrazione renale che a lungo andare può provocare danni a livello dei glomeruli (parti dell'unità funzionale renale). L'immediata conseguenza è l'aumento di proteine nelle urine (proteinuria), ma anche (cosa che può avvenire in parallelo o sostituirsi a quest'ultima) la diminuzione della filtrazione da parte dei glomeruli [5][6]. Altre modificazioni a livello strutturale che spesso si verificano sono l'ispessimento della membrana basale dei glomeruli, l'accumulazione di cellule mesangiali, l'aumento della grandezza dei tubuli prossimali e il mutamento dei podociti [7].
La presenza di microalbuminuria è ad oggi è il miglior predittore della nefropatia diabetica, ma non è un biomarker della nefropatia diabetica in fase iniziale: infatti in molti casi essa si presenta solo nella fase avanzata della malattia, quando si hanno già gravi danni a livello renale [8]. Sono quindi necessari degli altri predittori che consentano non solo di diagnosticare tale complicazione, ma anche di determinare se ne esista, e di quale entità sia, il rischio.
Il fatto che la nefropatia induca l'alterazione della normale struttura e attività dei reni può far ipotizzare che implichi un'alterazione anche dell'espressione e/o del turnover delle proteine [5][9]; infatti è al loro studio che ci si sta muovendo al fine di trovare nuovi markers per tale malattia.

2.1.1 Ereditarietà della nefropatia

Diversi studi [10][11] hanno evidenziato il fatto che almeno il 40% di pazienti con diabete mellito insulino-indipendente sviluppano anche nefropatia diabetica. Non è ancora perfettamente chiaro quali siano i fattori che comportano tale esito della malattia: sicuramente un largo contributo è dato da fattori ambientali, ma essi non possono essere i soli. Infatti, pur avendo simili caratteristiche (lunghezza della malattia, controllo metabolico, esposizione a medesimi fattori ambientali...), ci sono pazienti in cui tale complicazione insorge e altri per cui questo non avviene [9].
Tutto ciò fa pensare che i pazienti diabetici abbiano una predisposizione genetica a tale malattia e ciò è supportato dal fatto che l'insorgere della DN si verifica spesso in cluster famigliari [11][12].

2.1.2 L'espressione proteica nella nefropatia diabetica

Per studiare l'implicazione genetica nella nefropatia diabetica si possono andare ad analizzare i prodotti genici, cioè l'mRNA (che viene sintetizzato dal DNA a seguito della trascrizione) e le proteine (che vengono tradotte a partire dall'mRNA). Per quanto riguarda queste ultime (a cui ci si è interessati in questo studio) un parametro fondamentale che le caratterizza è il loro livello di espressione all'interno della cellula: esso rappresenta la concentrazione di tale proteina e può essere misurato attraverso tecniche di spettrometria di mas-

sa. Bisogna però tener presente da che cos'è determinata la concentrazione di una proteina: essa è il bilancio tra la sua velocità di sintesi e la sua velocità di degradazione. Questi 2 processi possono subire singolarmente delle variazioni (a seguito di perturbazioni dovute ad esempio a malattie, modifiche strutturali delle cellule...), ma nel complesso rimanere prossimi all'equilibrio, e quindi far rimanere inalterata l'espressione proteica. Quindi, la sola analisi di essa non sempre rispecchia totalmente eventuali perturbazioni del sistema, perché queste ultime potrebbero provocare una variazione nella velocità di sintesi o di degradazione della proteina ma non nel loro bilancio.
In letteratura sono già presenti studi che, confrontando il livello di espressione in pazienti diabetici e pazienti anche affetti da nefropatia, hanno individuato alcune funzioni biologiche per cui c'è una differente espressione delle proteine nelle 2 classi di soggetti [13][9].
Essi però possono non rispecchiare tutte le alterazioni indotte dal sistema che potrebbero invece determinare un'alterazione del turnover proteico.
Da queste considerazioni nasce l'interesse per il comportamento delle proteine nelle due classi di soggetti precedentemente citate. Lo studio si propone di analizzare il turnover delle proteine in fibroblasti di pazienti T1DM e pazienti anche affetti da DN, al fine di determinare se alcune di esse abbiano una velocità di degradazione significativamente diversa in questi ultimi rispetto ai primi.

3 ACQUISIZIONE DELLE MISURE

L'esperimento utilizzato per l'acquisizione dei dati, che verrà descritto dettagliatamente in seguito, utilizza fibroblasti cutanei ottenuti tramite biopsia dall'avambraccio dei pazienti e compie un'analisi *in vitro* del turnover delle proteine presenti in essi. Sono state scelte queste cellule in quanto, pur non essendo direttamente collegate ai reni, hanno evidenziato significative differenze fenotipiche nelle due classi di soggetti di interesse [14].
Per fare qualche esempio, in studi precedenti[15][16] in cui sono stati usati fibroblasti cutanei, si è riscontrato un aumento dell'antiporto (trasporto contemporaneo di due soluti attraverso la membrana cellulare) in pazienti diabetici in cui è stata diagnosticata anche nefropatia rispetto a quelli solo diabetici. È anche emerso che nei primi c'è una maggior sintesi di DNA rispetto ai secondi.

3.1 PROTOCOLLO SILAC DINAMICO

Il protocollo SILAC (stable-isotope labelling by amino acids in cell culture) è una tecnologia usata in proteomica quantitativa. Essa ha molte applicazioni [17][18][19], tra cui quella della determinazione contemporanea del turnover di tutte le proteine presenti in una popolazione cellulare (definita *dinamica*).
In sintesi (i dettagli verranno forniti più sotto) le cellule inizialmente poste in un terreno di amminoacidi non marcati (*light*) verranno poi trasferite in un terreno di amminoacidi marcati (*heavy*). Di conseguenza, dall'istante in cui avviene lo scambio del terreno, si potranno distinguere le proteine di nuova sintesi da quelle sintetizzate in precedenza la cui quantità verrà misurata attraverso spettrometria di massa.

3.1.1 Coltura cellulare

- *Coltura non marcata*: dopo la biopsia di alcune cellule epiteliali dell'avambraccio, esse vengono portate a confluenza. Subito dopo vengono raccolte e conservate in azoto liquido.

- *Coltura marcata*: all'istante di inizio dell'esperimento (t=0, t_0) viene aspirato il terreno freddo e le cellule vengono lavate 2 volte con PBS (soluzione salina tampone). Il medium appena tolto viene sostituito con:

  – terreno DMEM (privo di Arginina e Lisina) in cui vengono aggiunte Arginina e Lisina-$^{13}C_6 - 2HCl$

- siero dializzato (10%, contenente fattori nutritivi)
- glutammina (amminoacido essenziale)
- penicillina e streptomocina (antibiotici)

Subito dopo aver aggiunto gli amminoacidi marcati le cellule vengono trasferite in un incubatore a 37°.

3.1.2 Estrazione e purificazione delle cellule

Ad ogni istante temporale in cui si vuole effettuare la misura (compreso il t_0), viene prelevato il terreno di coltura e le cellule sono trattate come segue:

- lavate 3 volte con PBS;
- viene aggiunta la tripsina e vengono trasferite in incubatore;
- al fine di bloccare l'azione della tripsina dopo 3-4 minuti vengono aggiunti 10 ml di terreno privo di Lisina e Arginina e con siero dializzato (10%);
- vengono poi centrifugate a bassa velocità per eliminare il terreno e lavate più volte per eliminare il surnatante;
- infine vengono congelate a $-80°$.

3.1.3 Lisi cellulare

Le cellule devono essere lisate per poter estrarre le proteine in esse contenute. Viene quindi preparato il tampone di lisi che viene poi messo nella provetta contenente le cellule precedentemente estratte. Le cellule in questa soluzione vengono congelate in N_2 liquido, sonicate e ricongelate più volte, per poi essere centrifugate. Dopo questa serie di passaggi vengono demolite tutte le strutture cellulari (membrana, organelli...) e si ottengono proteine purificate.

3.1.4 Elettroforesi monodimensionale

Le proteine estratte vengono poste in un gel di poliacrilammide (al 12%) a cui viene applicato un campo elettrico. In questo modo esse sono separate in base alla loro dimensione e carica (in realtà essa è molto simile per tutte). Le proteine più piccole sono più veloci, mentre le più grandi sono più lente.
Nello specifico:

- viene aggiunto alle proteine il DTT che rompe i legami disolfuro delle proteine;

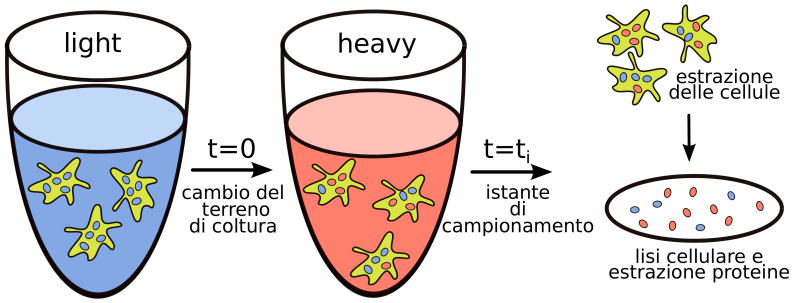


Figura 5: Primi passaggi del protocollo SILAC.

- viene caricato per ogni pozzetto (fessura nel gel) lo stesso numero di proteine;
- lo strumento viene impostato con una ddp di 80 V e una corrente di 25 mA;
- dopo la corsa il gel viene tolto e lasciato 3 ore a bagno nel colorante Blu Comassie colloidale;
- infine il gel viene lavato più volte per far si che rimangano colorate solo le proteine.

3.1.5 Taglio del gel

La corsa elettroforetica viene divisa in più parti, in maniera tale di avere meno proteine per volta nell'analisi successiva. Si è scelto di dividerla in 5 bandine.

3.1.6 Digestione delle proteine

Questa fase della procedura serve a digerire le proteine in peptidi che verranno poi analizzati con la spettrometria di massa.

- Le bandine vengono lavate con H_2O, e poi vortexate e centrifugate per eliminare l'acqua.
- Le fasi successive consistono in una serie di passaggi che servono a disidratare e decolorare le bandine ed eliminare i sali.
- Digestione con Endoproteinase Lys-C: questo enzima serve a tagliare ogni proteina in peptidi a livello della Lisina marcata (in questo modo tutti i peptidi avranno un amminoacido marcato). Esso viene combinato

con una soluzione tampone e le bandine vengono lasciate in digestione overnight a 37° all'interno di provette.

- Le provette vengono centrifugate e infine lavate con l'acetonitrile e vortexate ripetutamente al fine di far uscire i peptidi dal gel.

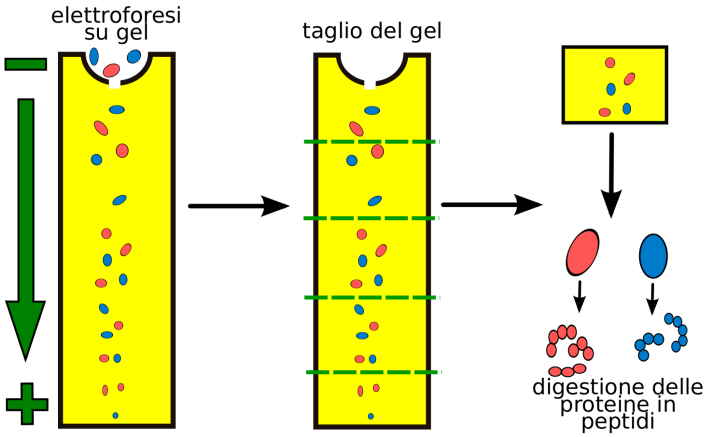


Figura 6: Ultimi passaggi del SILAC per preparare i campioni alla spettrometria di massa.

3.1.7 Spettrometria di massa

I peptidi isolati possono essere analizzati attraverso spettrometria di massa. Lo strumento utilizzato è l'*Orbitrap*.

- Ognuna delle 5 bandine viene analizzata singolarmente: l'*Orbitrap* è un sistema combinato di cromatografia liquida e spettrometria di massa (LC/MS). La prima serve a separare cromatograficamente i peptidi e permette di far entrare i peptidi nello spettrometro non tutti contemporaneamente. Con la seconda si effettua l'analisi vera e propria.
- I files provenienti dallo spettrometro vengono elaborati con il software Discoverer Daemon 1.2 che ha la funzione di riconoscere i peptidi in base agli spettri misurati. Infine i dati vengono analizzati e quantificati.
- Al fine di essere sicuri che non rimangano residui, dopo l'analisi di ogni campione vengono fatte 3 corse 'in bianco'.

3.1.8 Dati forniti dal software

Il software, per ottenere il dato globale di proteina, analizza e rielabora i dati dei peptidi ad essa associati. Per ogni peptide, riconosciuto attraverso

l'analisi dei picchi generati dallo spettrometro, viene quindi calcolato il valore del *picco light* e quello del *picco heavy*, che rappresentano la loro abbondanza, e successivamente il loro rapporto, come mostrato in Figura 7.

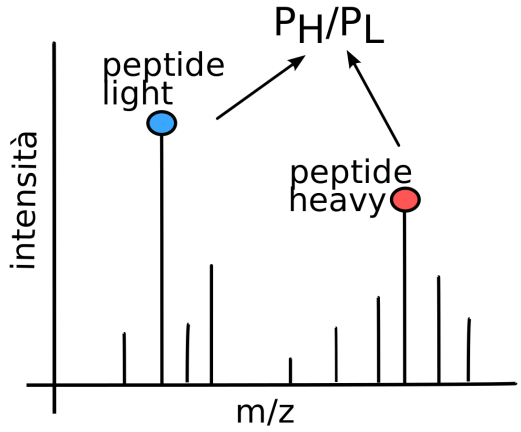


Figura 7: Lo spettrometro identifica i picchi relativi al peptide heavy e a quello light per poi calcolarne il rapporto.

Vengono considerate, per ogni istante di campionamento, solo le proteine a cui il software riesca ad associare almeno 2 peptidi.
Il software è segretato quindi non si sa con precisione in che modo, dai peptidi, ricava il rapporto $\frac{P_H}{P_L}$ della proteina; probabilmente attua una sorta di media pesata dei dati dei peptidi in cui come peso considera l'affidabilità della misura da cui essi sono stati ottenuti (quindi l'affidabilità di individuazione del picco dallo spettro). In corrispondenza di ogni istante di campionamento vengono quindi forniti dal software:

- il $\frac{P_H}{P_L}$ di ogni proteina;
- un valore del CV (coefficiente di variazione) relativo alla misura (non si sa come è stato ricavato);
- il $\frac{P_H}{P_L}$ di ogni peptide;
- il grado di accuratezza della misura del peptide;
- un riferimento a se è stata usata oppure no questa misura nel calcolo del valore della proteina.

In realtà in aggiunta a questi sono molti altri i dati riportati, ma non sono stati citati in quanto non utili all'analisi successiva.

4 IDENTIFICAZIONE DEI PARAMETRI DEL MODELLO

4.1 STIMA CON I MINIMI QUADRATI PESATI

Considerando un modello univocamente identificabile a priori a tempo continuo, l'uscita osservabile $Y(t)$ può essere predetta come

$$Y(t) = f(t, p) \text{ con p = parametri del modello} \tag{7}$$

dove f è la *funzione di predizione del modello* che dipende dai suoi parametri e dal tempo. Generalmente si hanno a disposizione un certo numero di misure m_i agli istanti t_i, con $i = 1, 2..., N$ (dove N è il numero di campioni); esse sono però affette da un errore e_i, in genere considerato additivo, e quindi possono essere espresse come:

$$m_i = y_i + e_i = f(t_i, p) + e_i \text{ con i=1,2...,N} \tag{8}$$

Nella maggior parte dei casi e_i è incognito e può essere descritto come una variabile aleatoria con:

$$E[e_i] = 0 \text{ con i=1,2...,N} \tag{9}$$

$$var[e_i] = SD_i^2 \text{ con i=1,2...,N} \tag{10}$$

SD_i è la *deviazione standard* dell'errore di misura che può essere costante (e quindi $SD_i^2 = SD^2$) oppure dipendere dall'istante di campionamento.

È utile anche riferirsi all'e_i in termini di coefficiente di variazione (CV):

$$CV_i = \frac{SD_i}{y_i} \text{ con i=1,2...,N} \tag{11}$$

Anch'esso, a seconda di come variano la SD_i e l'uscita y_i può essere costante (e quindi $CV_i = CV$) oppure dipendere dall'istante di campionamento. Il modello generale che descrive la SD_i^2 è:

$$SD_i^2 = a + b * (y_i)^c \text{ con i=1,2...,N} \tag{12}$$

Due casi particolari sono:

$$b = 0 \implies SD_i^2 = a \implies SD_i = \sqrt{a} \implies \text{SD COSTANTE} \tag{13}$$

$$\begin{aligned} a = 0 \text{ e } c = 2 \implies SD_i^2 = b(y_i)^2 \implies SD_i = \sqrt{b}(y_i) \implies CV_i = \sqrt{b} \\ \implies \text{CV COSTANTE} \end{aligned} \tag{14}$$

La scelta del modello dell'errore di misura influisce sull'affidabilità che si si vuole dare ai dati:

- nel caso a SD COSTANTE viene dato lo stesso peso a tutti i dati (a prescindere che essi abbiano valore maggiore o minore);
- nel caso a CV COSTANTE si ha invece che l'SD è maggiore per i dati più grandi in modulo. Di conseguenza si assume che l'errore di misura sia proporzionale al modulo dei dati; nella pratica, nella predizione si darà più peso alle misure minori reputandole *più credibili* e invece si ipotizzerà che ci possa essere un errore più grande nei dati con valore assoluto più alto.

Si può riscrivere la 8 in forma vettoriale:

$$m = F(t, p) + e \tag{15}$$

dove:

$$m = [m_1 \ \ m_1 \ \ ... \ \ m_N]^T \tag{16}$$

$$F(t, p) = [f(t_1, p) \ \ f(t_2, p) \ \ ... \ \ f(t_N, p)]^T \tag{17}$$

$$e = [e_1 \ \ e_1 \ \ ... \ \ e_N]^T \tag{18}$$

$$E[e] = 0 \text{ vettore media dell'errore di misura} \tag{19}$$

$$E[ee^T] = \Sigma_e \text{ matrice di covarianza dell'errore di misura} \tag{20}$$

La 20, qualsiasi sia il modello dell'errore di misura, può essere scritta come:

$$\Sigma_e = \sigma^2 * B \tag{21}$$

dove:

- B è sempre noto
- σ^2 può essere noto o incognito

A seconda del modello dell'errore di misura:

- SD COSTANTE: $B = I_N$ e σ^2 = varianza costante
- CV COSTANTE: $B = diag(m_1^2, m_2^2, ..., m_N^2)$ e σ = CV costante

Per attuare la stima ai minimi quadrati pesati, l'obiettivo è quello di trovare il vettore dei parametri p che renda minima la distanza pesata rispetto all'errore di misura tra il modello e i dati. Tale distanza è così definita:

$$\| (m - F(t,p))^2 \|_{\Sigma_e^{-1}} = [m - F(t,p)]^T \Sigma_e^{-1} [m - F(t,p)] \tag{22}$$

Il parametro stimato sarà quindi:

$$\hat{p} = \text{argmin}_p [m - F(t,p)]^T \Sigma_e^{-1} [m - F(t,p)] \tag{23}$$

Se il modello è lineare nei parametri la 23 ha la seguente soluzione in forma chiusa:

$$\hat{p} = (F^T \Sigma_e^{-1} F)^{-1} F^T \Sigma_e^{-1} m \tag{24}$$

altrimenti bisogna usare un metodo iterativo di ottimizzazione.

4.1.1 Stima con 'pesi relativi'

Qualora si ipotizzi che l'errore di misura sia a SD costante o a CV costante, ma non si conosca il valore di tali parametri (cioè si conosca il valore di B ma non quello di σ^2) si può lo stesso procedere con la stima. L'idea è quella di impostare dei *pesi relativi* (per differenziarli da quelli che conosciamo totalmente che chiamiamo *pesi assoluti*). La 22 può essere riscritta come:

$$\| (m - F(t,p))^2 \|_{\Sigma_e^{-1}} = \frac{1}{\sigma^2}[m - F(t,p)]^T B^{-1}[m - F(t,p)] \tag{25}$$

e posso quindi stimare $\hat{p}$ come:

$$\hat{p} = \text{argmin}_p [m - F(t,p)]^T B^{-1}[m - F(t,p)] \tag{26}$$

Infatti 23 e 26 differiscono solo per un fattore costante ($\frac{1}{\sigma^2}$) che non influenza il calcolo del minimo.
Il σ^2 può essere poi stimato a posteriori come:

$$\hat{\sigma^2} = \frac{WRSS(\hat{p})}{gradi\ di\ liberta} = \frac{[m - F(t,p)]^T B^{-1}[m - F(t,p)]}{N - M} \tag{27}$$

con N=numero di campioni e M=numero di parametri.
Pur ottenendo lo stesso valore dei parametri stimati, si avrà però una diversa precisione delle stime, essendo diversa la σ^2 (vedi 4.1.3).

4.1.2 Residui

Il vettore dei residui rappresenta la distanza tra le misure e il valore ottenuto sostituendo al modello il vettore dei parametri $\hat{p}$ ottenuto dalla stima; esso quindi sarà:

$$res = m - F(t,\hat{p}) \tag{28}$$

Esso, confrontandolo col l'eq. 8, oltre a costituire l'errore di predizione, può anche essere considerato come una stima dell'errore di misura.

Si può definire il vettore dei *residui pesati* come:

$$wres = \frac{res}{SD} \tag{29}$$

Sei il modello scelto è buono, ci si deve attendere che i residui rispecchino le proprietà statistiche dell'errore di misura. Quindi, se come assunto l'errore è a campioni scorrelati e varianza nota (o stimata a posteriori se incognita), e poiché:

$$var(\frac{e_i}{SD_i}) = 1 \tag{30}$$

i residui pesati dovrebbero essere scorrelati e in modulo <1.

4.1.3 Precisione delle stime

L'errore commesso nella stima del parametro è definito come:

$$\tilde{p} = p - \hat{p} \tag{31}$$

Nel caso di modello lineare nei parametri, sfruttando 24, tale valore è determinabile come:

$$\tilde{p} = [I_M - F^T\Sigma_e^{-1}F)^{-1}F^T\Sigma_e^{-1}F]p - F^T\Sigma_e^{-1}F)^{-1}F^T\Sigma_e^{-1}e \tag{32}$$

in cui la prima parte è deterministica, mentre la seconda è random.

Da qui si può determinare la matrice di covarianza dell'errore di stima, che fornisce un'informazione relativa al range di valori che esso può assumere e quindi una stima della precisione della stima:

$$\Sigma_{\tilde{p}} = (F^T\Sigma_e^{-1}F)^{-1} \tag{33}$$

Nel caso di modello non lineare nei parametri tale valore non è determinabile in forma chiusa, ma può essere approssimato a:

$$\Sigma_{\tilde{p}} = (S^T\Sigma_e^{-1}S)^{-1} \tag{34}$$

dove:

$$S = \begin{pmatrix} \frac{\partial f(t_1,p)}{\partial p_1}|_{p=\hat{p}} & \frac{\partial f(t_1,p)}{\partial p_2}|_{p=\hat{p}} & \cdots & \frac{\partial f(t_1,p)}{\partial p_M}|_{p=\hat{p}} \\ \frac{\partial f(t_2,p)}{\partial p_1}|_{p=\hat{p}} & \frac{\partial f(t_2,p)}{\partial p_2}|_{p=\hat{p}} & \cdots & \frac{\partial f(t_2,p)}{\partial p_M}|_{p=\hat{p}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f(t_N,p)}{\partial p_1}|_{p=\hat{p}} & \frac{\partial f(t_N,p)}{\partial p_2}|_{p=\hat{p}} & \cdots & \frac{\partial f(t_N,p)}{\partial p_M}|_{p=\hat{p}} \end{pmatrix}$$

4.2 IMPLEMENTAZIONE

Il modello considerato, che rappresenta l'andamento del rapporto tra la quantità di proteine *heavy* e di proteine *light*, è:

$$\frac{P_H(t)}{P_L(t)} = \frac{P_{tot} * (1 - e^{-kt})}{P_{tot} * e^{-kt}} = \frac{(1 - e^{-kt})}{e^{-kt}} \tag{35}$$

L'identificazione del parametro k è stata implementata nel linguaggio R (versione 2.15.3) seguendo il procedimento descritto in 4.1.
Il modello è non lineare nei parametri, quindi si è dovuto usare un metodo iterativo per l'identificazione del parametro. A tale scopo è stata scelta la function *optim* di R in cui è stato impostato *method='Brent'*. Esso infatti è adatto a problemi uno-dimensionali (con un solo parametro da stimare) e permette di vincolare la stima ad un certo intervallo di valori. Nello specifico si è cercato il valore del parametro tra i soli valori positivi compresi tra 0 e 1 (impostando *lower=c(0), upper=c(1)*). Valori negativi infatti non avrebbero avuto senso dal punto di vista sperimentale.
Si sono quindi definiti il modello e la funzione costo da minimizzare in *optim*:

```
# mod1 = modello dei dati
# parametri d'ingresso: temp = istanti di campionamento,
#                       k = parametro del modello
# parametri d'uscita: y = uscita calcolata applicando al modello
#                         i parametri d'ingresso
mod1 = function(temp,k)
{
        y = (1-exp(-k*temp))/exp(-k*temp)
        return(y)
}
```

```
# fcosto = funzione costo
# parametri d'ingresso: k = parametro del modello,
#                       modello = modello applicato,
#                       temp = istanti di campionamento,
#                       dati = misure,
#                       w = pesi
# paramentri d'uscita: COST = valore della funzione costo
fcosto = function(k,modello,temp,dati,w)
{
        y = mod1(temp,k)
        COST = sum(w*(y-dati)^2)
        return(COST)
}
```

Nell'**impostazione dei pesi** bisogna tener conto del tipo di stima che si sta facendo:

1. *stima con pesi assoluti*: in questo caso l'errore di misura è totalmente conosciuto (sia il modello che i valori). Il vettore dei pesi sarà quindi ottenuto a partire dal vettore delle misure (*dati*) come:
   - a *SD costante*

     ```
     sd=rep(SD,length(dati))
     w=1/((sd)^2)
     ```

   - a *CV costante*

     ```
     w=1/((dati*CV)^2)
     ```

2. *stima con pesi relativi*: in questo caso dell'errore di misura si conosce solo il modello, ma non il valore. Quindi si impostano i pesi, a meno di una costante come:
   - a *SD costante*

     ```
     w=rep(1,length(dati))
     ```

   - a *CV costante*

     ```
     w=1/(dati)^2))
     ```

In uscita dalla function *optim* si ottengono il vettore dei parametri (K), che nel caso specifico è uno solo e quindi un float, e la WRSS:

```
K=res$par #parametro stimato
WRSS=res$value #somma dei residui al quadrato pesati
```

Nel caso di stima con *pesi relativi* si può quindi calcolare la $\hat{\sigma}^2$ e la Σ_e, cioè la matrice contenente sulla diagonale i valori della varianza stimata a posteriori:

- a *SD costante*

  ```
  sigma2=WRSS/(N-1)
  B=diag(rep(1,N))
  sigma_e=sigma2*B
  ```

- a *CV costante*

  ```
  sigma2=WRSS/(N-1)
  B=diag((dati)^2)
  sigma_e=sigma2*B
  ```

Per il calcolo della **precisione della stima**, cioè della $\Sigma_{\tilde{p}}$, si fa la derivata parziale, rispetto al parametro, della funzione del modello e sostituendo in essa i tempi di campionamento e il valore del parametro stimato si ottiene un vettore di lunghezza pari al numero di misure. Con esso, sfruttando la 34, si può quindi calcolare la varianza della stima e il suo CV:

```
S=temp*exp(K*temp) #derivata del modello rispetto a k
var_stim=solve(t(S)%*%solve(sigma_e)%*%S)
cv_stim=sqrt(var_stim)/K #cv del parametro stimato
```

5 FILTRAGGIO

5.1 SCELTA DEI TEMPI DI CAMPIONAMENTO

In primo luogo è stato necessario capire quali fossero gli istanti di campionamento che permettessero di avere una stima accurata del parametro k. Per ogni esperimento è stato individuato 3 come numero di campioni per ogni soggetto e si è poi dovuto scegliere quale fosse la loro miglior collocazione nel tempo.
Sono stati fatti 2 esperimenti pilota su 2 soggetti con i seguenti istanti di campionamento:

- 1 soggetto sano: 1h, 2h, 4h, 7.5h e 24h;
- 1 soggetto T1DM+DN: 4h, 7.5h, 24h, 48h e 72h.

Per completezza, il soggetto sano è stato considerato solo nell'analisi preliminare in quanto ci si è concentranti sulle 2 classi di soggetti T1DM e T1DM+DN per l'analisi vera e propria. La strada che si è seguita è stata quella di confrontare le stime dei k fatte con tutti i campioni con quelle eseguite eliminando alcuni istanti di campionamento: se la correlazione tra la stima con o senza un determinato campione fosse risultata alta allora ciò avrebbe significato che tale campione non sarebbe stato essenziale per la stima corretta; viceversa, una bassa correlazione avrebbe implicato la necessità dell'utilizzo di esso per non inficiare la stima.
Dopo aver eliminato le proteine per cui i dati contraddicevano il modello assunto (vedi 5.3), è stata fatta la stima dei k. In essa si è ipotizzato prima un errore a deviazione standard (*SD*) costante e poi a coefficiente di variazione *(CV)* costante, non avendo ancora chiarito quale fosse l'errore più coerente (vedi 5.2). Non conoscendo il valore né dell'SD né del CV è stata fatta la stima con i *pesi relativi*. In realtà si sono ottenuti risultati concordi, e molto simili, in entrambi i casi; quindi nella trattazione vengono esposti solo quelli relativi alla stima a *CV* costante.
È risultato che:

- *Confronto tra la stima con tutti i campioni (full) e senza il campione delle 72h e delle 48h*: togliendo prima solo il campione alle 72h e poi anche quello alle 48h, come si vede dai grafici in Figura 8, c'è una correlazione piuttosto bassa tra le stime relative ai 2 esperimenti e quelle ottenute mediante il fit *full*. Questo in un primo momento ha fatto pensare che tali campioni fossero necessari. Analizzando però l'andamento di tutti i campioni si è visto che per moltissime proteine il valore $\frac{P_H}{P_L}$, dalle 48 ore in poi, cresce

molto più lentamente, fino talvolta a decrescere (2 esempi sono riportati in Figura 9), cosa che trova riscontro nel confronto tra le stime in cui si vede che i k sono generalmente più bassi nella stima *full* rispetto a quella senza i 2 campioni. Questo comportamento contraddice l'assunzione iniziale che il rapporto $\frac{P_H}{P_L}$ debba sempre crescere e significherebbe che dalle 48 ore in poi il contributo *ligth* aumenterebbe. Come spiegazione del fenomeno si è ipotizzato che il terreno di coltura non fosse sufficiente a soddisfare il fabbisogno di amminoacidi della cellula e quindi che essa rimettesse a disposizione anche gli amminoacidi *light* derivanti dalla degradazione delle proteine non marcate. In questo modo il ricircolo, che era stato assunto non significativo, lo sarebbe diventato dalle 48 ore in poi.
Per questo motivo si è scelto di escludere come istanti di campionamento sia le 48h che le 72h.

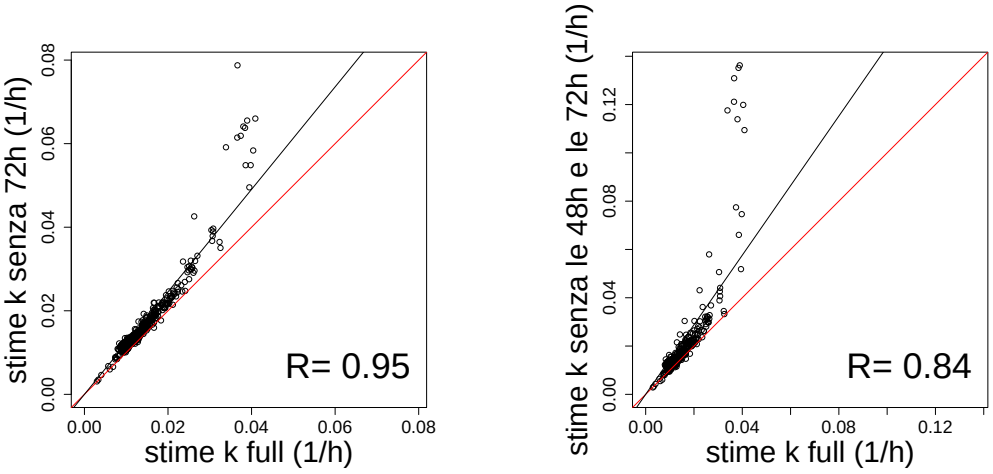


Figura 8: Grafici che rappresentano il confronto tra le stime full vs stime senza il campione delle 72 h (sopra) e full vs stime senza i campioni delle 48h e 72h (sotto). Le stime sono state fatte ipotizzando un errore a CV costante e sono relative al soggetto diabetico e nefropatico. Viene anche riportato il valore del coefficiente di correlazione dei 2 casi.

- *Confronto tra la stima full e senza il campione delle 1h e delle 2h*: da questo confronto è emersa un'altissima correlazione (Figura 10).
  Quindi i 2 istanti temporali sono stati ritenuti non necessari alla stima e si sono potuti escludere.

- *Confronto tra la stima con i 3 istanti temporali 4h, 7.5h e 24h e quelle ottenute escludendone uno alla volta*: escludendo il campione della 4ª ora oppure quello delle 7.5h nel fare la stima, la correlazione tra esse e la con tutti e 3 istanti temporali è risultata molto alta (Figura 11). Escludendo invece il campione delle ore 24 la correlazione è scesa di molto. Di conseguenza il campione delle 24h è stato ritenuto necessario, mentre, in linea teorica, si sarebbe potuto eliminarne uno degli altri due. In realtà è stato però

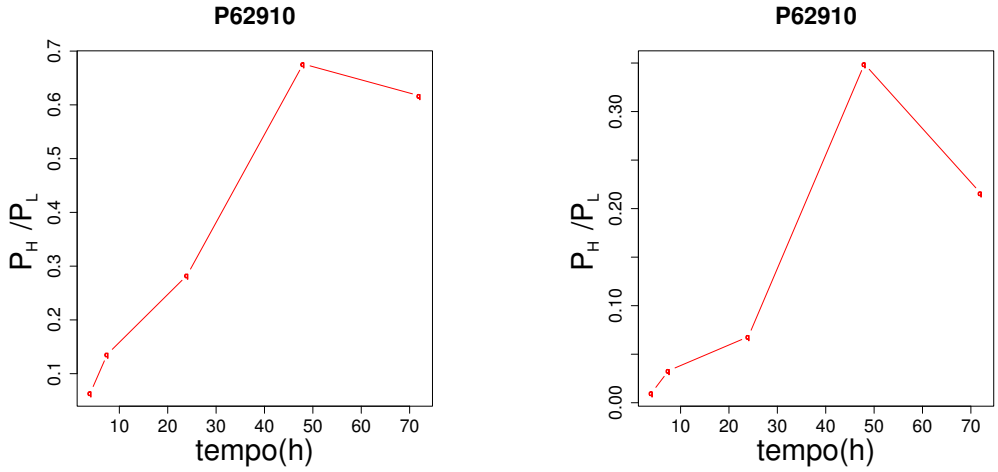


Figura 9: Grafici che rappresentano l'andamento del rapporto $\frac{P_H}{P_L}$ di alcune proteine. Si vede che dal campione delle 48h si ha una decrescenza.

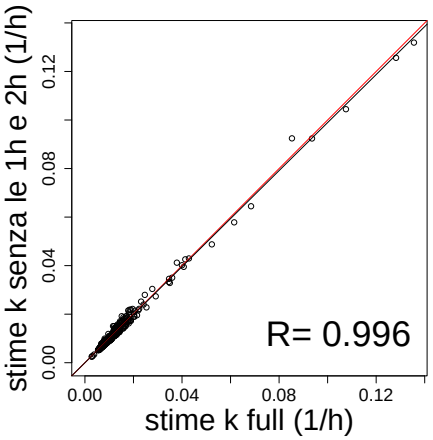


Figura 10: Grafico che rappresenta il confronto tra le stime full vs senza il campione delle 1h e 2h. Le stime sono state fatte ipotizzando un errore a CV costante e sono relative al soggetto di controllo. Viene anche riportato il valore del coefficiente di correlazione.

scelto di tenerli entrambi per per poter considerare più proteine possibili nell'analisi; infatti, talvolta succede che lo spettrometro non riesca ad identificare i picchi di tutti gli istanti temporali per tutte le proteine e quindi, qualora per una proteina questo si verificasse al tempo 4h o 7.5h, la stima si sarebbe potuta comunque fare con i soli 2 istanti acquisiti.

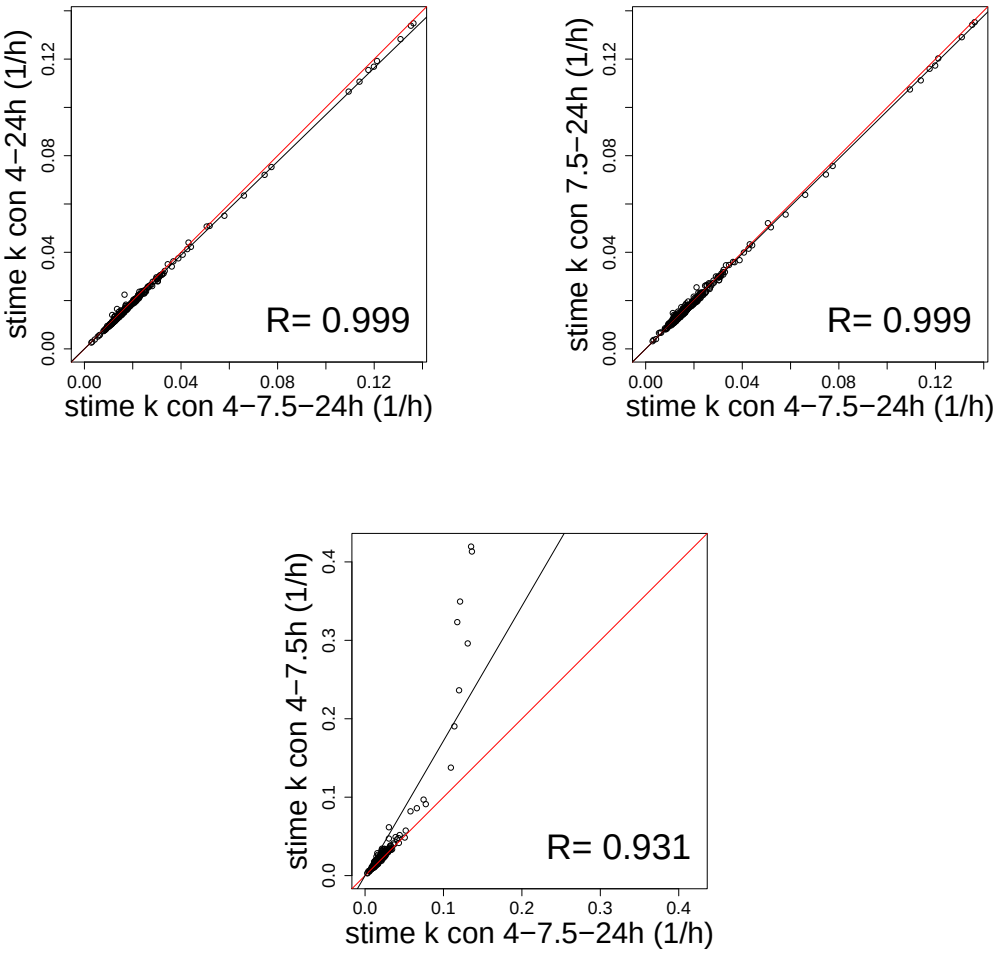


Figura 11: Grafici che rappresentano il confronto tra le stime con i campioni 4h, 7.5h e 24h vs le stime con i campioni 4h e 24h (sopra), tra le stime con i campioni 4h, 7.5h e 24h vs le stime con i campioni 7.5h e 24h (in mezzo) e tra le stime con i campioni 4h, 7.5h e 24h vs le stime con i campioni 4h e 7.5h (sotto). Le stime sono state fatte ipotizzando un errore a *CV* costante e sono relative al soggetto diabetico e nefropatico. Viene anche riportato il valore del coefficiente di correlazione dei 2 casi.

5.2 VARIABILITÀ TECNICA DELLE MISURE

Il software di elaborazione dei dati dello spettrometro di massa fornisce un valore per il CV delle misure per ogni campione dei dati globali di proteina. Esso è stato però ritenuto inutilizzabile in quanto, non solo non si hanno informazioni su come è stato ottenuto, ma anche perché i valori spaziavano in un range molto ampio arrivando ad essere molto alti e poco verosimili (2%-700%). Per poter procedere con la stima è stato quindi necessario capire quale fosse il modello più indicato per l'errore di misura.
Questa scelta è stata fatta considerando 8 soggetti: 4 diabetici e 4 diabetici e nefropatici. Tale numero infatti è stato ritenuto una percentuale sufficientemente alta (80%) dei soggetti totali per poter evincere delle informazioni che valessero per tutti i soggetti dell'analisi. Sono state utilizzate solo le proteine per cui si avessero le misure per tutti e 3 gli istanti temporali o quella delle 24h e una tra le 4h e le 7.5h; sono state poi eliminate quelle per cui i dati fossero in contraddizione con il modello adottato (vedi 5.3).
Inizialmente ci si è concentrati su due modelli: a SD costante e a CV costante. Delle proteine considerate è stato quindi stimato il k impostando i *pesi relativi* e ipotizzando:

- $SD = \alpha$ (con α costante) per il modello a SD costante;
- $SD = \beta * x$ (con β costante) per il modello a CV costante.

Le costanti α e β sono ignote.
I risultati ottenuti non hanno però permesso di decidere per una delle due ipotesi. Infatti:

1. Dal *confronto diretto dei parametri* è emerso che i risultati ottenuti dalle 2 stime differiscono soprattutto nei valori bassi. In generale infatti, la stima a CV costante, dando maggior affidabilità al primo campione, tende ad abbassare le stime, facendo abbassare anche la correlazione (Figura 12). Questo confronto non ha permesso però di decidere per una delle 2 stime.
2. Si è quindi proceduto con l'*analisi dei residui pesati a posteriori*. È stata calcolata, in entrambe le stime, la loro mediana per ogni istante temporale. Come si vede in Figura 13 entrambe le pesature sono in modulo quasi sempre minori di 1. Inoltre, pur essendo minori i residui a SD costante, visto il basso numero di campioni, non si può dire nulla sulla loro bianchezza. Non si può quindi propendere per uno o l'altro.

Decisiva si è rivelata l'*analisi dei peptidi*. Come spiegato precedentemente (vedi 3.1.8) la misura del rapporto $\frac{P_H}{P_L}$ viene fornita sia come dato globale di proteina, sia per ogni peptide associato alla proteina stessa. Quindi le misure dei peptidi possono essere considerate come replica della misura della proteina a cui appartengono.

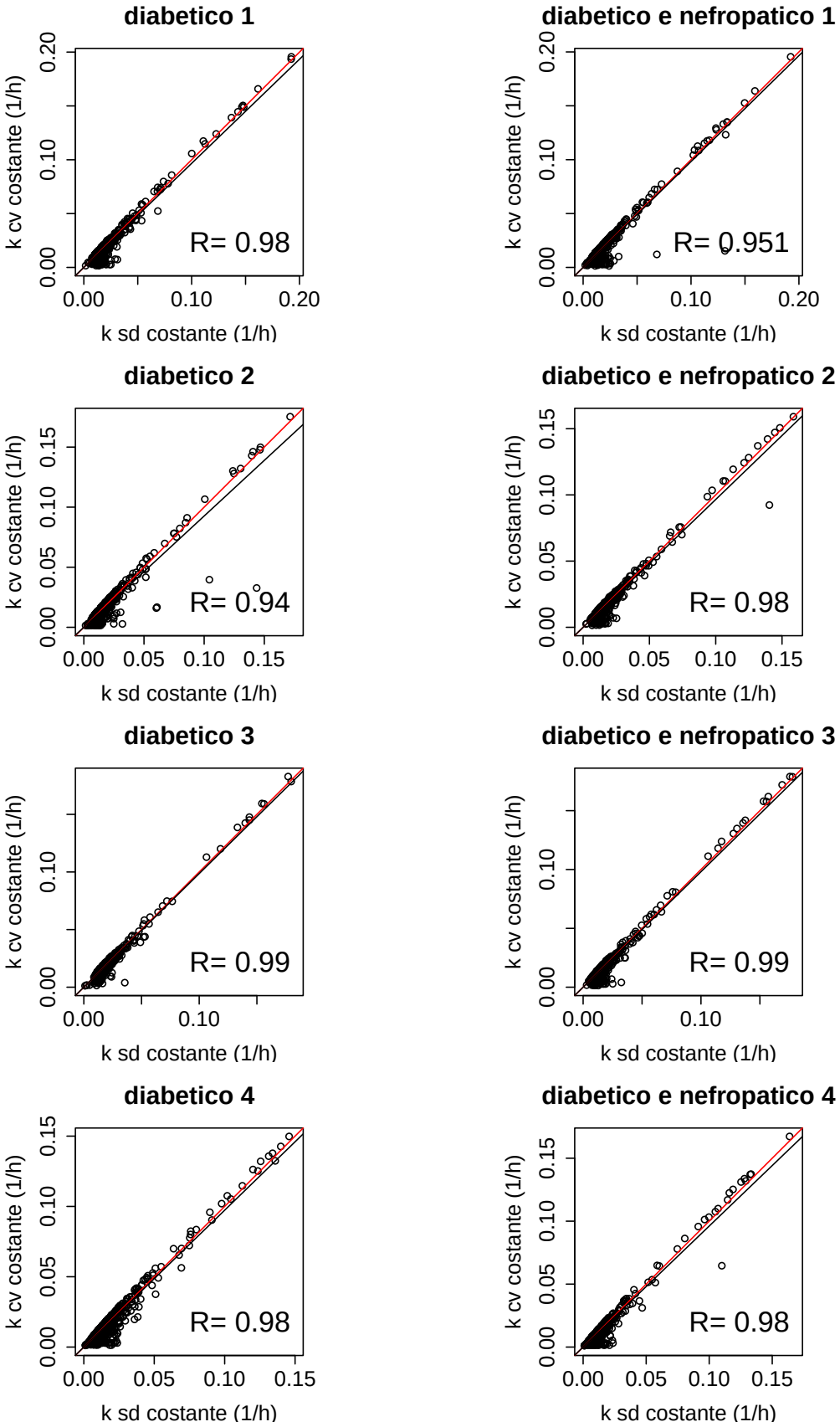


Figura 12: Grafici che rappresentano il confronto tra le stime a SD costante e CV costante negli 8 soggetti considerati. La retta nera rappresenta la retta di regressione dei punti, quella rossa la bisettrice del quadrante (di riferimento). Viene riportato anche il valore della correlazione.

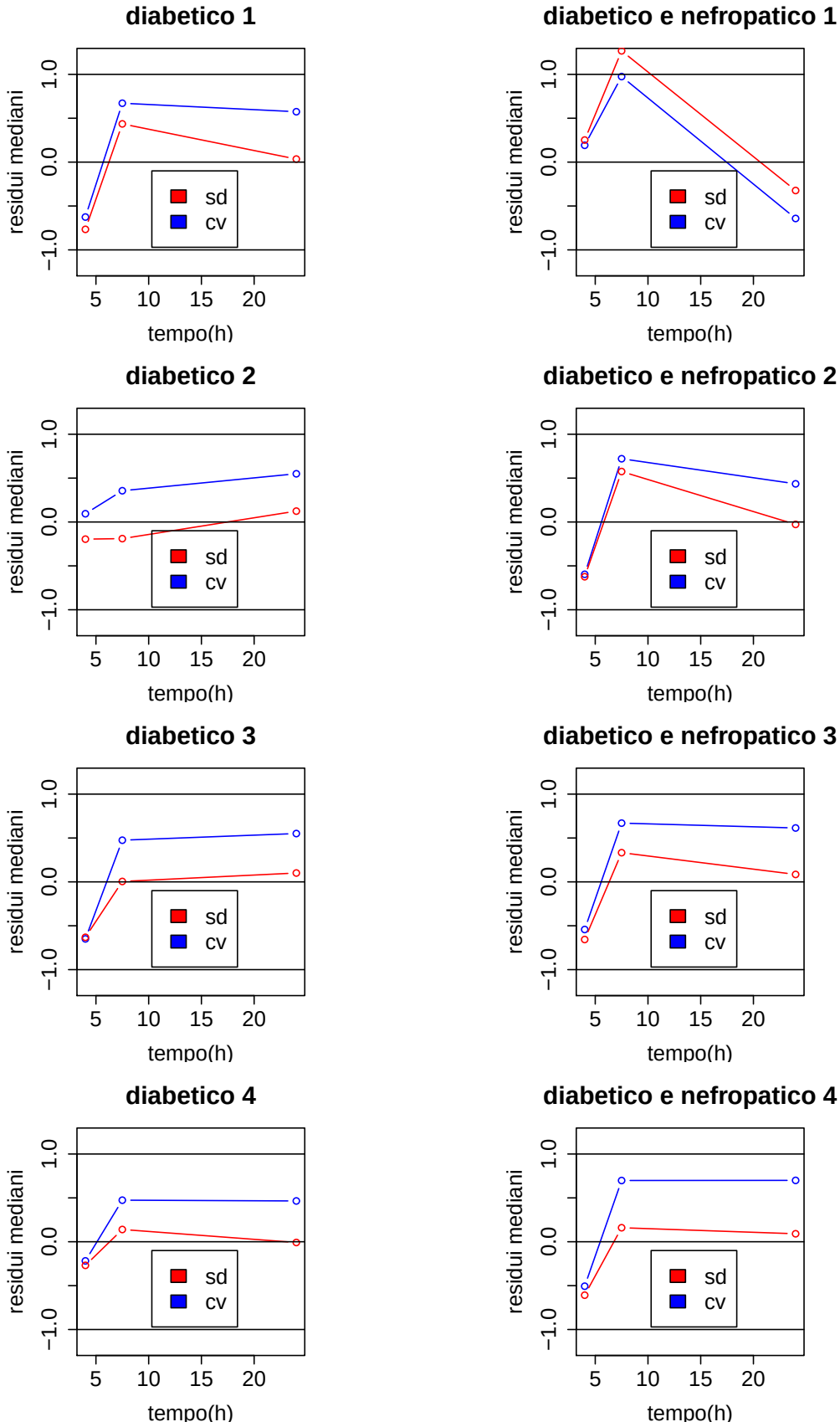


Figura 13: Grafici delle mediane dei residui pesati a posteriori delle stime degli 8 soggetti a SD costante (in rosso) e a CV costante (in blu).

Per ognuno degli 8 soggetti sono stati stimati la deviazione standard e il coefficiente di variazione dell'errore di misura del rapporto $\frac{P_H}{P_L}$ al variare del valore del rapporto stesso.
Nel dettaglio: per ogni proteina, per ogni istante temporale, sono state calcolate la media, la SD e il CV del valore di $\frac{P_H}{P_L}$ dei peptidi ad essa associati; considerando tutte le medie così calcolate, esse sono state divise in intervalli (di ampiezza 0.05, fino al valore 0.4, e da lì in poi in intervalli contenenti ognuno lo stesso numero di elementi). Per ogni intervallo si è poi calcolato il valore mediano della SD e del CV delle proteine in esso contenute.

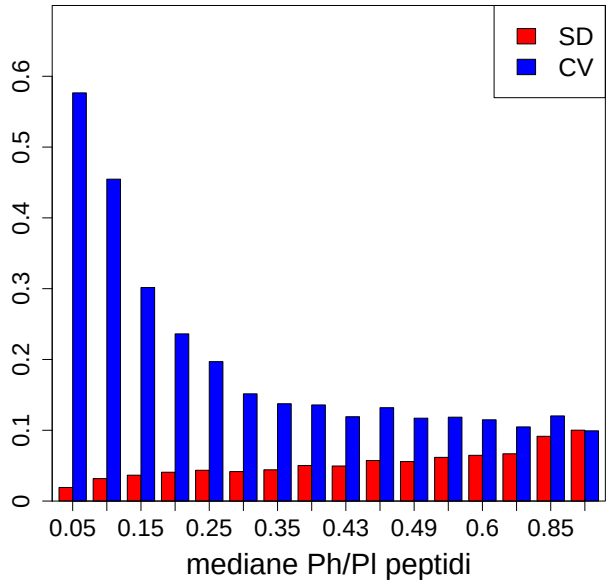


Figura 14: Barplot relativo all'andamento della SD (in rosso) e del CV (in blu) dei peptidi di ogni proteina considerando la mediana di tutte le proteine degli 8 soggetti.

Come si vede in Figura 14 né la SD né il CV hanno un andamento costante, ma la prima tende a crescere all'aumentare della media, mentre il secondo tende a diminuire. Si è quindi escluso che il modello dell'errore di misura potesse essere a SD o a CV costante, ma si è pensato ad un modello che rispecchiasse meglio tale andamento.
La scelta è caduta su:

$$SD = \sqrt{\alpha^2 + \beta^2 * x^2} \quad \text{con } \alpha \text{ e } \beta \text{ costanti e x valore di } \frac{P_H}{P_L} \tag{36}$$

Per verificare se tale modello potesse essere adatto, si è plottato l'andamento della SD e del CV così calcolato sopra il barplot di entrambe. Sono stati impostati alle 2 costanti i seguenti valori:

- α=0.02
- β=0.1

Come si vede in Figura 15 l'andamento di quest'ultimo modello (in verde) a differenza degli altri 2 segue bene l'andamento sia del CV sia della SD.

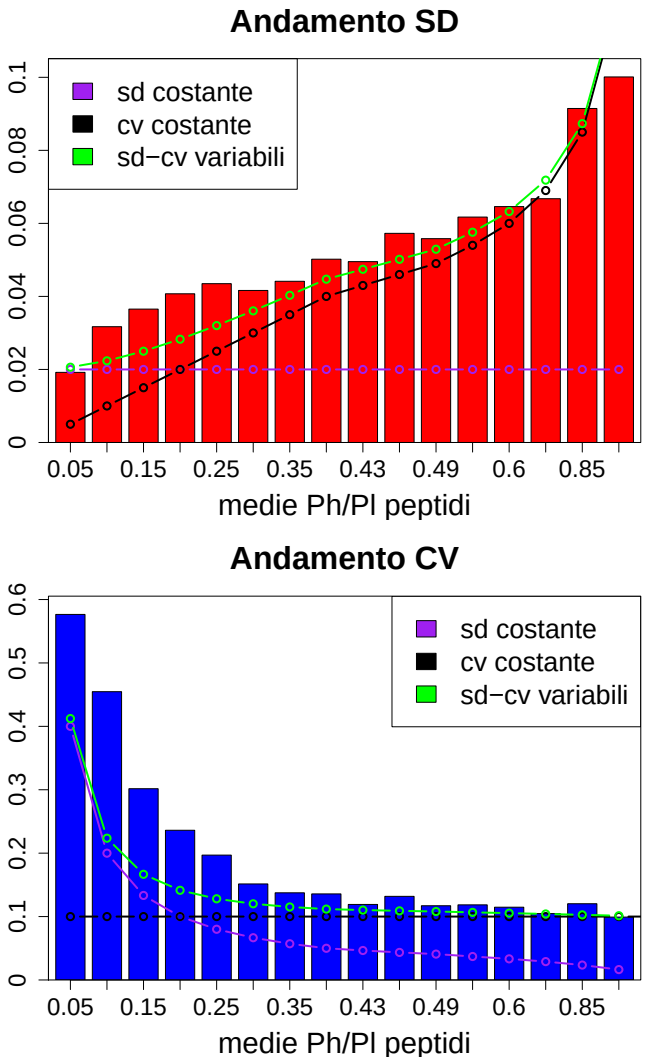


Figura 15: Barplot relativi all'andamento della SD (in alto) e del CV (in basso) dei peptidi. Sono stati anche riportati i plot delle predizioni della SD e del CV calcolati applicando i 3 modelli considerati di SD. Per le costanti sono stati impostati i valori di: α=0.02 e β=0.1.

Da tutte le precedenti considerazioni, la scelta è caduta sul modello dell'errore di misura variabile, che è stato usato per le successive stime dei k.

5.3 PREFILTRAGGIO E PROTEINE CONSIDERATE

5.3.1 Prefiltraggio

Non tutte le proteine di cui il software ha fornito le misure si sono potute considerare. Si è dovuto attuare un filtraggio preventivo, secondo i criteri illustrati qui sotto.

1. *Proteine senza il numero minimo di campioni per poter procedere con la stima*
   Sono state eliminate tutte le proteine per cui:
   - si avessero meno di due campioni temporali;
   - si avessero solo il campione delle 4h e delle 7.5h.

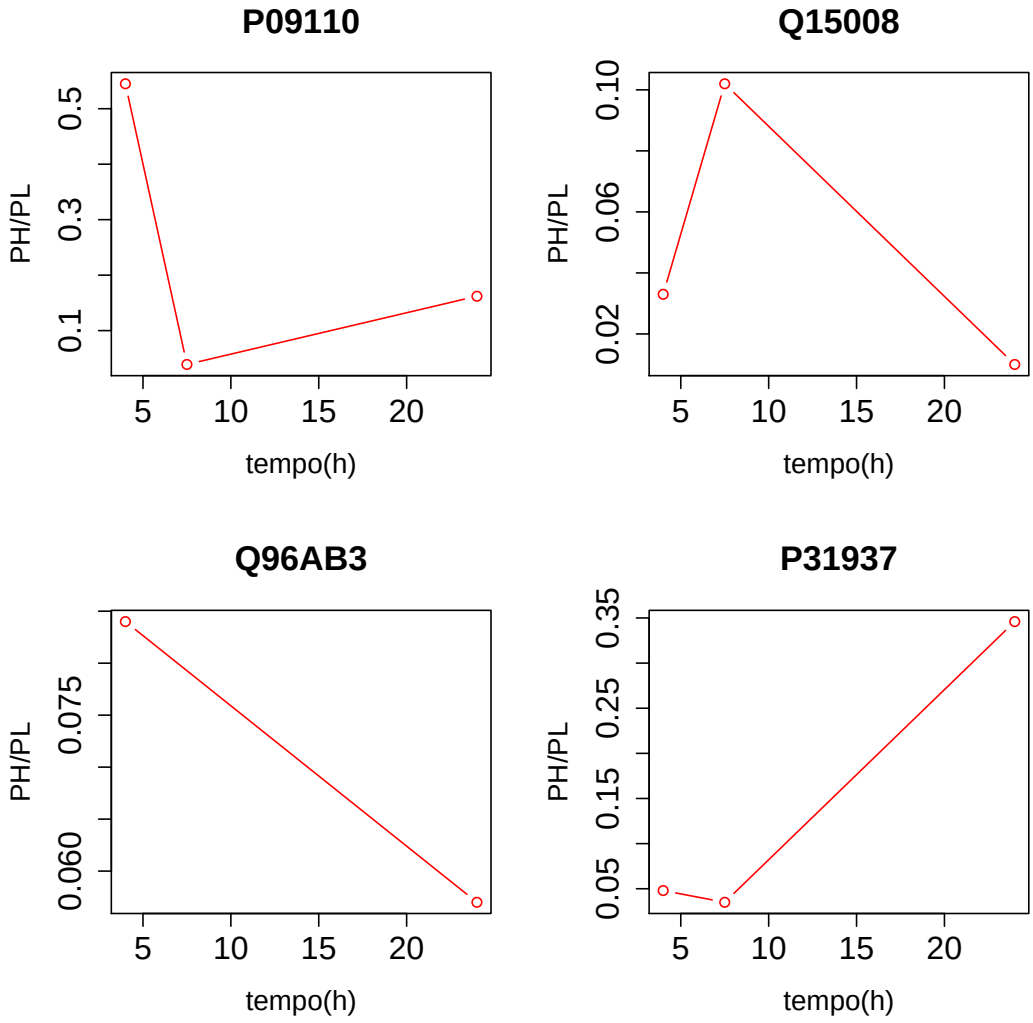


Figura 16: Grafici rappresentanti l'andamento del rapporto $\frac{P_H}{P_L}$ di alcune proteine decrescenti. Le prime 3 sono state scartate, l'ultima in basso a destra rappresenta il caso in cui la decrescenza si riscontra tra i primi 2 campioni e quindi è stata considerata.

2. *Proteine le cui misure sono decrescenti*
   Sono state considerate come decrescenti solo le proteine in cui almeno uno tra i campioni delle 4h e delle 7.5h fosse maggiore del campione delle 24h. Queste proteine sono state eliminate in quanto palesemente in disaccordo con il modello assunto. Infatti per ipotesi il rapporto $\frac{P_H}{P_L}$ cresce esponenzialmente (vedi Figura 16).
   Per alcune proteine la decrescenza si riscontra tra il primo e il secondo campione, e nella maggior parte di questi casi non è molto accentuata: in tale situazione probabilmente la decrescenza è dovuta all'errore di misura. Per questo motivo, si è deciso di considerarle e non eliminarle.

3. *Proteine le cui misure sono prossime a 0*
   Si sono riscontrati alcuni casi in cui il rapporto $\frac{P_H}{P_L}$ mantiene sempre valori prossimi a 0 (Figura 17). Anche in questo caso l'andamento contraddice il modello assunto, poiché significa che non c'è mai nuova sintesi dal cambio di terreno in poi. Per questo motivo sono state eliminate anche le proteine che presentassero tale andamento.

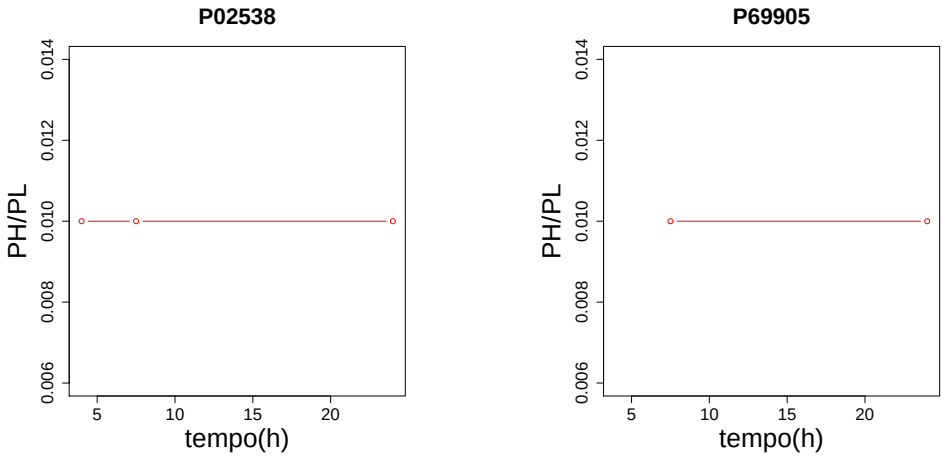


Figura 17: Grafici che rappresentano l'andamento del rapporto $\frac{P_H}{P_L}$ di alcune proteine per cui esso rimane prossimo a 0.

5.3.2 Soggetti e proteine considerate

I dati a disposizione sono stati ottenuti dall'analisi di 10 soggetti, 5 diabetici e 5 diabetici nefropatici. Nelle tabelle 1 e 2 sono riportati alcuni dati relativi ad essi.

Nelle seguenti tabelle 3 e 4 è invece riportato il numero di proteine considerate per ogni soggetto.

	ID	Sesso	Età alla biopsia (anni)	Durata malattia (anni)
Diabetico1	D1	M	33	11
Diabetico2	D2	M	57	27
Diabetico3	D3	F	41	25
Diabetico4	D4	F	36	29
Diabetico5	D5	F	27	22

Tabella 1: Dati relativi ai soggetti diabetici

	ID	Sesso	Età alla biopsia (anni)	Durata malattia (anni)
Diab.-nefropatico1	DN1	F	25	10
Diab.-nefropatico2	DN2	M	41	25
Diab.-nefropatico3	DN3	F	48	37
Diab.-nefropatico4	DN4	F	32	21
Diab.-nefropatico5	DN5	M	30	13

Tabella 2: Dati relativi ai soggetti diabetici-nefropatici

	D1	D2	D3	D4	D5
Proteine considerate	1005	955	705	1059	904

Tabella 3: Numero di proteine considerate per ognuno dei soggetti diabetici

	DN1	DN2	DN3	DN4	DN5
Proteine considerate	977	889	915	1026	894

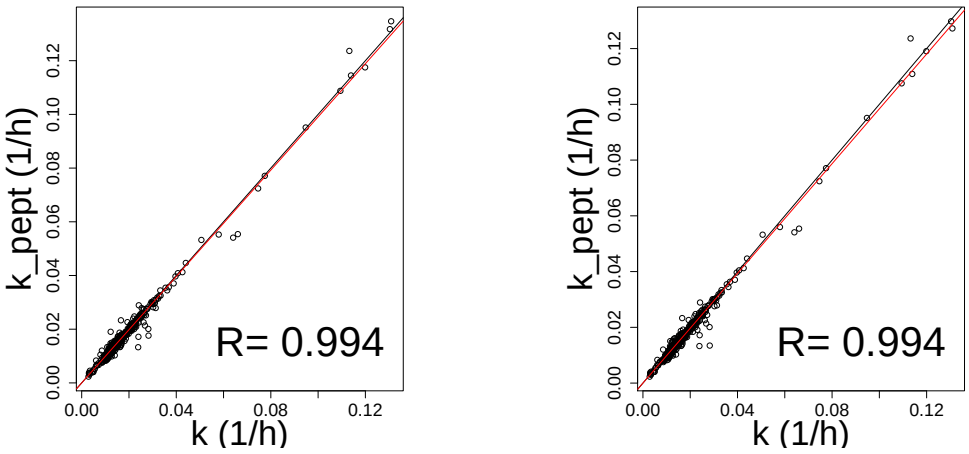
Tabella 4: Numero di proteine considerate per ognuno dei soggetti diabetici-nefropatici

5.4 ANALISI PEPTIDI

Prima di procedere con l'analisi usando direttamente i dati globali di proteina si è voluto accertare che, stimando il k a partire dai dati dei peptidi, non ci fossero significative differenze tra le 2 stime. Per ogni proteina si è quindi:

1. isolato ogni peptide per cui si avesse la misura $\frac{P_H}{P_L}$ per almeno 2 istanti temporali tra cui le 24h;
2. stimato il k per ognuno dei peptidi isolati usando i *pesi relativi* a CV costante e SD costante. Infatti questa parte dell'analisi è stata fatta prima della scelta del modello dell'errore di misura e per questo sono stati usati i pesi relativi. Il fatto di aver fatto questa scelta non inficia il risultato in quanto, come detto in 4.1.1, il k stimato é simile;
3. calcolato la media e la mediana dei k stimati dai peptidi per ottenere il k totale della proteina (k_{pept});
4. confrontato il k_{pept} con quello ottenuto dalla stima con pesi relativi a CV costante usando le misure globali della proteina.

Plottando i k e i k_{pept} gli uni contro gli altri (Figura 18) si vede che la correlazione è altissima, sia nel caso che questi ultimi siano stati calcolati come media dei k stimati dai peptidi, sia come mediana (che peraltro hanno valori molto simili).
Per questo motivo si è concluso che per l'analisi futura si possono usare direttamente i dati globali di proteina.



(a) k_{pept} calcolati come media **(b)** k_{pept} calcolati come mediana

Figura 18: Grafici che rappresentano il confronto tra le stime calcolate dai fit dei peptidi: a sinistra ottenendo il k_{pept} come media dei k dei peptidi, a destra come mediana.

5.5 NORMALIZZAZIONE DEI DATI

Confrontando le misure acquisite dai vari soggetti, si è riscontrata una grande variabilità inter-soggetto. Vengono riportati in Figura 19 gli MvA plot tra il $\log_2$ dei dati delle proteine comuni di alcune coppie di soggetti (riportare tutte le possibili coppie sarebbe troppo oneroso). Essi sono stati fatti distinguendo i 3 istanti temporali.

Come si vede dai grafici la nuvola dei punti è shiftata rispetto all'asse y=0 (negli esempi riportati verso il basso, ma spesso anche verso l'alto). Avendo a disposizione dati high-throghput, si ipotizza che la maggior parte delle proteine abbia un andamento simile, mentre solo per poche le misure siano molto diverse. Quindi ci si dovrebbe aspettare che la nuvola di punti si distribuisse lungo l'asse delle ascisse, dove la differenza è minima. Questo, come detto sopra, non succede, e la causa è da imputare ad un errore sistematico dovuto al fatto che gli esperimenti sono stati eseguiti in momenti diversi e da persone diverse. Per questo motivo si è attuato lo scaling dei dati, attraverso i seguenti passi:

1. considerando il $\log_2$ delle misure, per ogni paziente è stata calcolata la mediana di tutte le misure di ogni istante di campionamento separatamente. Quindi si avranno le mediane:

$$M_{ij} \text{ con i=4,7.5,24 h (istante temporale) e j=1,...,10 (soggetto)} \tag{37}$$

2. è stata poi calcolata la mediana delle mediane (separatamente per ogni istante di campionamento). Quindi otterrò:

$$Mtot_i \text{ con i=4,7.5,24 h (istante temporale)} \tag{38}$$

3. per ogni soggetto e ogni istante temporale è stato calcolato il fattore di scala come:

$$FS_{ij} = Mtot_i - M_{ij} \text{ con i=4,7.5,24 (istante temporale) e j=1,...,10 (soggetto)} \tag{39}$$

4. ogni fattore di scala è stato sommato alle rispettive misure.

In Figura 20 sono rappresentati gli MvA plot dei confronti tra gli stessi soggetti di quelli in Figura 19, ma ottenuti dai dati ricalcolati. Come si vede, le nuvole di punti sono ora centrate rispetto all'asse y=0.
In tabella 5 sono riportati gli FS, in scala logaritmica. In generale si ha che per ogni soggetto gli FS sono simili nei 3 istanti temporali; questo però non è sempre vero. Questo fenomeno si spiega considerando il fatto che lo spettrometro analizza indipendentemente i vari istanti di campionamento. Quindi, anche all'interno di ogni soggetto si possono avere errori sistematici diversi per i 3 campioni a disposizione.

	D1	**D2**	**D3**	**D4**	**D5**	**DN1**	**DN2**	**DN3**	**DN4**	**DN5**
4 h	-0.234	-0.073	-0.332	0.140	0.107	-0.488	-0.045	0.092	0.792	0.045
7.5 h	-0.268	0.028	-0.302	0.132	0.042	-0.543	-0.082	-0.028	0.536	0.147
24 h	-0.320	0.006	-0.378	0.149	0.065	-0.211	-0.139	-0.006	0.560	0.121

Tabella 5: Tabella contenente per ogni soggetto gli FS, in scala logaritmica, da applicare ai dati per attuare lo scaling.

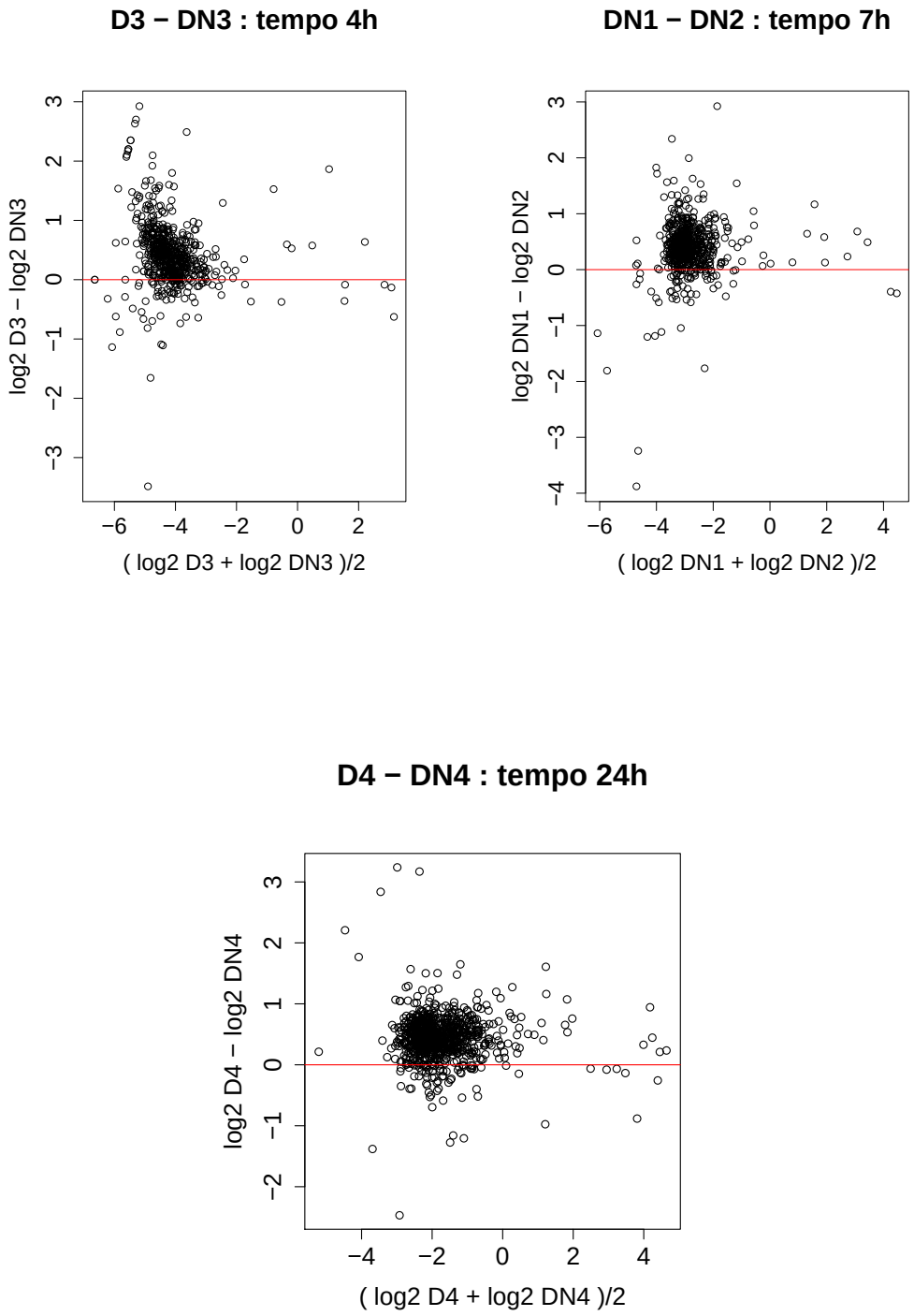


Figura 19: Esempi di MvA plot tra alcune coppie di soggetti; in alto a sinistra il confronto è tra le misure al tempo 4h del paziente diabetico 3 e del diabetico e nefropatico 3; in alto a destra il confronto è tra le misure al tempo 7h dei pazienti diabetici e nefropatici 1 e 2; in basso il confronto è tra le misure al tempo 24h del paziente diabetico 4 e del diabetico e nefropatico 4.

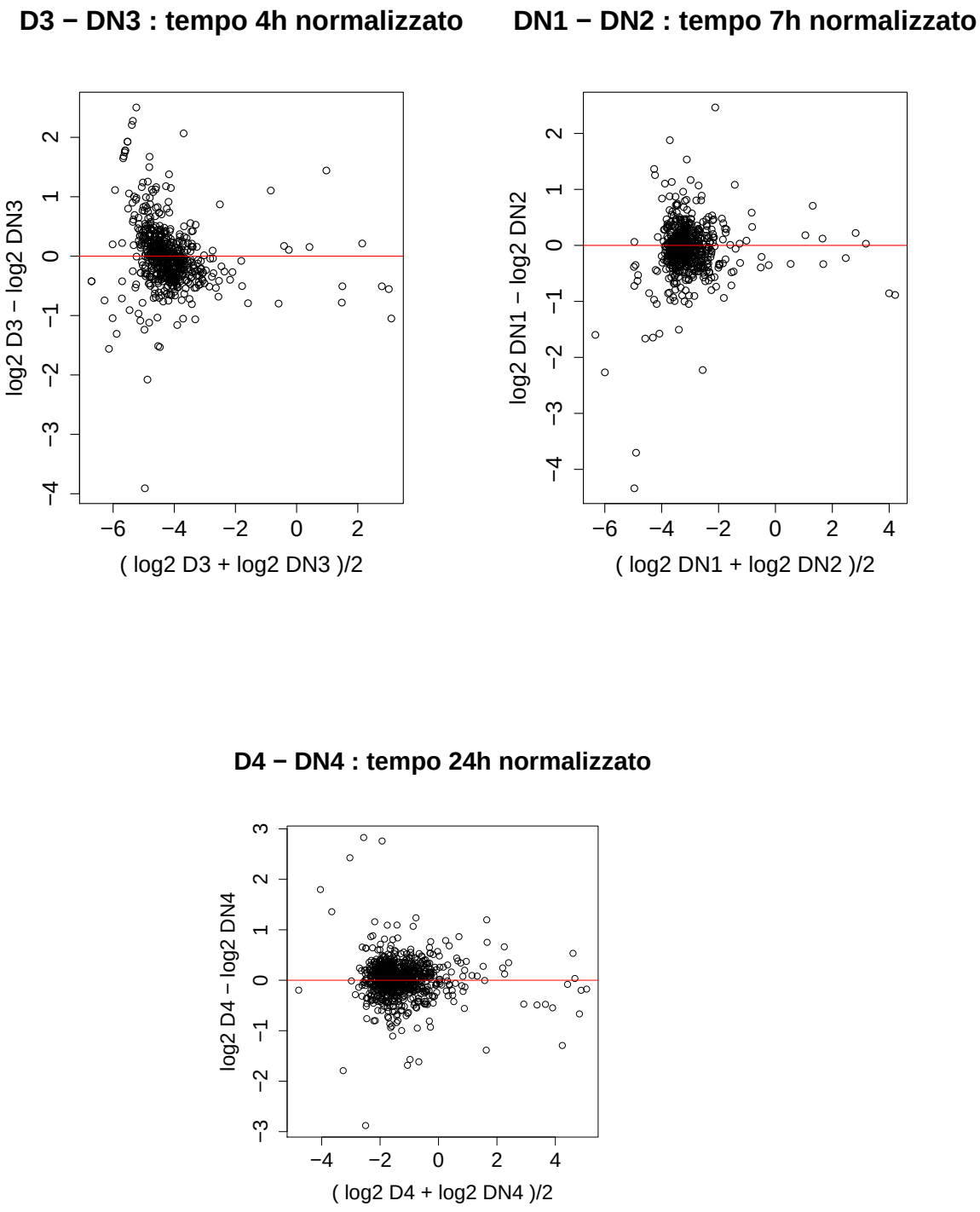


Figura 20: Esempi di MvA plot tra alcune coppie di soggetti ottenuti dai dati normalizzati con lo scaling; in alto a sinistra il confronto è tra le misure al tempo 4h del paziente diabetico 3 e del diabetico e nefropatico 3; in alto a destra il confronto è tra le misure al tempo 7h dei pazienti diabetici e nefropatici 1 e 2; in basso il confronto è tra le misure al tempo 24h del paziente diabetico 4 e del diabetico e nefropatico 4.

6 ANALISI DELL'ESPRESSIONE DIFFERENZIALI: METODI

6.1 TEST D'IPOTESI

I test d'ipotesi sono un potente strumento della statistica inferenziale. Essi, prendendo in considerazione una o più variabili tratte da una o più popolazioni, hanno l'obiettivo di formulare un'ipotesi relativa alla distribuzione delle variabili, e, in base ai dati che si hanno a disposizione, capire se essa può essere accettata o meno. L'ipotesi che viene formulata viene chiamata *ipotesi nulla* (H_0), mentre l'opposta viene denominata *alternativa* (H_1).
A seconda della decisione che viene presa si possono verificare 4 situazioni:

1. VERO NEGATIVO: l'H_0 viene assunta vera e H_0 è vera;
2. FALSO POSITIVO (errore di tipo 1): l'H_0 viene assunta falsa e H_0 è vera;
3. FALSO NEGATIVO (errore di tipo 2): l'H_0 viene assunta vera e H_0 è falsa;
4. VERO POSITIVO: l'H_0 viene assunta falsa e H_0 è falsa.

Quando viene presa una decisione non si sa con certezza in quale dei 4 casi si sia, ma si può stimare la probabilità con cui si incorre in uno dei 2 tipi di errore.
I test d'ipotesi possono essere applicati in molti casi: in particolare, quando si hanno campioni relativi ad una classe e si vuole determinare se essa ha una distribuzione significativamente diversa da una data o quando si hanno 2 o più classi e si vuole stabilire se hanno distribuzioni simili o meno. In seguito saranno presi in considerazione i test su due classi in quanto saranno usati nell'analisi successiva.

6.1.1 Test di Student su campioni indipendenti

Il test di Student, chiamato anche test *t di Student*, su 2 classi è un test parametrico; questo significa che si suppone come nota la distribuzione di probabilità dell'ipotesi nulla e ci si basa su di essa per capire se nel caso in esame essa può essere accettata o rifiutata. Per ognuno dei due gruppi (che saranno chiamati 1 e 2) si hanno a disposizione un serie di osservazioni.
Affinché il test sia valido, devono verificarsi alcune assunzioni di base:

- i campioni dei due gruppi devono essere indipendenti;
- le osservazioni sui due gruppi sono realizzazioni di variabili gaussiane con media μ_1 e μ_2 e deviazione standard σ_1 e σ_2;

- le deviazioni standard devono essere uguali ($\sigma_1 = \sigma_2$).

Qualora valgano gli assunti sopra elencati si possono definire le ipotesi:

- H_0: $\mu_1 = \mu_2$
- H_1: $\mu_1 \neq \mu_2$

Quello appena descritto è il caso più generale, chiamato *two-sided*: in esso l'ipotesi alternativa indica solo che esiste una diversità tra le medie, ma non fa distinzione sulla direzione di tale disuguaglianza. Questo viene invece specificato nei test di Student *one-sided*, in cui si ha che:

- H_0: $\mu_1 \leqslant \mu_2$
- H_1: $\mu_1 > \mu_2$

oppure

- H_0: $\mu_1 \geqslant \mu_2$
- H_1: $\mu_1 < \mu_2$

La distribuzione in ipotesi nulla si può ottenere supponendo di campionare N volte n_1 e n_2 osservazioni dalla stessa distribuzione gaussiana, simulando che esse siano le osservazioni dei 2 gruppi che si andranno a testare. Appartenendo alla medesima distribuzione, per essi varrà sicuramente l' H_0. Viene ora definita la variabile

$$t = \frac{m_1 - m_2}{s_p * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{40}$$

dove

$$s_p{}^2 = \frac{(n_1 - 1) * s_1{}^2 + (n_2 - 1) * s_2{}^2}{(n_1 + n_2 - 2)} \tag{41}$$

con m_1 e m_2 le medie e s_1 e s_2 le deviazioni standard delle n_1 e n_2 osservazioni, rispettivamente. Per ognuno degli N campionamenti viene quindi calcolato il valore di t, ottenendo il vettore $T=t_1,t_2,...,t_N$. La distribuzione che si ottiene è del tipo in Figura 21.

Il test vero e proprio consiste, avendo i dati reali (osservazioni del primo e del secondo gruppo), nel calcolare il t_{obs} relativo ad essi e nel verificare in che punto della distribuzione t_{obs} si collochi.
Fissata poi una certa soglia θ, l'ipotesi H_0 viene rifiutata se $|t_{obs}| > \theta$ (nel caso del test *two-sided*) oppure se $t_{obs} > \theta$ (nel caso del test *one-sided*). Invece di considerare la soglia prefissata, si può considerare l'area verde sottesa dalla curva che tale soglia identifica (livello di significatività α); l'area compresa tra $|t_{obs}|$ e $|Inf|$ (omettendo il modulo se si tratta di un test one-sided) viene chiamata

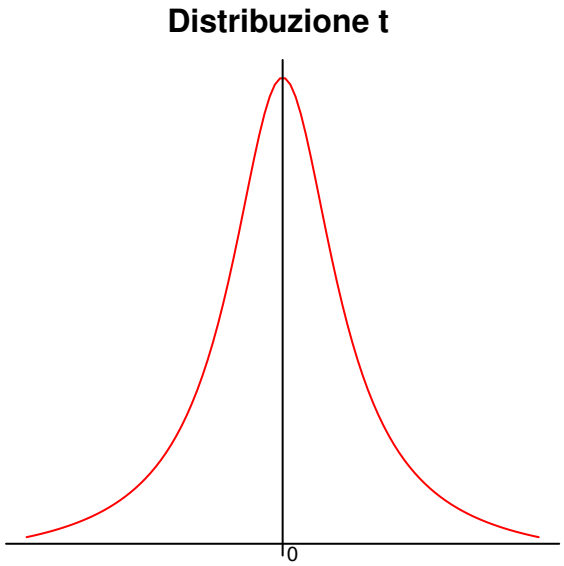


Figura 21: Tipico andamento della distribuzione t in ipotesi nulla.

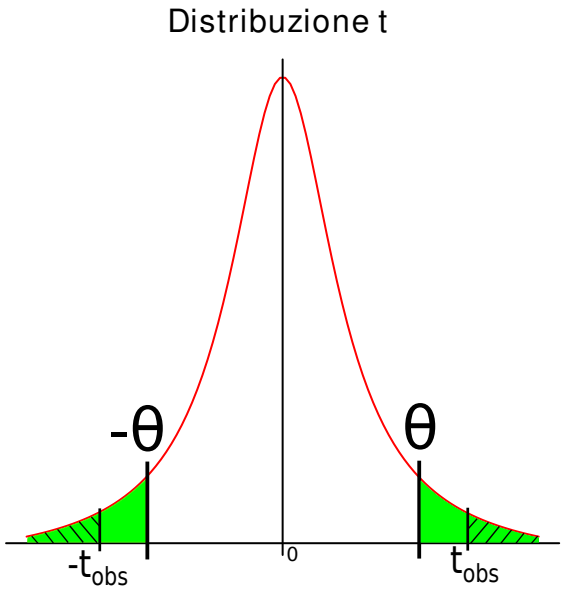


Figura 22: Distribuzione t con l'esempio di un t_{obs}. Vengono evidenzianti la soglia θ e il livello di significatività α (area in verde) per un test two-sided. In questo caso l'ipotesi nulla si sarebbe dovuta rifiutare in quanto il p-value (area tratteggiata) è minore di α.

p-value: congruentemente a quanto detto prima, rifiuto H_0 se *p-value*<α (vedi Figura 22).

Per fissare θ bisogna considerare qual è il rischio di commettere un errore di tipo 1 (falso positivo) che si può accettare di correre: esso è proprio il livello di significatività α, che di norma viene fissato al 5%.

Variante di Welch

Non sempre la terza ipotesi del test di Student ($\sigma_1 = \sigma_2$) è verificata. Qualora $\sigma_1 \neq \sigma_2$ e si applicasse la stessa statistica descritta sopra, si rischierebbe di sovrastimare la varianza, commettendo quindi più errori di tipo 2. Il questo caso è meglio applicare la variante di Welch del test di Student in cui:

$$t = \frac{m_1 - m_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}} \tag{42}$$

e attuare la seguente correzione dei gradi di libertà:

$$df = \frac{\left(\frac{\sigma_1^2}{n_1}\right)^2 + \left(\frac{\sigma_2^2}{n_2}\right)^2}{\frac{1}{(n_1-1)} * \left(\frac{\sigma_1^2}{n_1}\right)^2 + \frac{1}{(n_2-1)} * \left(\frac{\sigma_2^2}{n_2}\right)^2} \tag{43}$$

6.1.2 Test di normalità di Shapiro-Wilk

Molti test statistici, come anche quello di Student appena descritto, impongono come assunzione che i dati analizzati seguano una determinata distribuzione (quella gaussiana nello specifico). Se vengono applicati senza che essa sia verificata, possono far incorrere in errore e quindi è necessario scegliere un metodo alternativo.
Esistono dei test statistici che permettono di capire se i campioni che si hanno a disposizione appartengano o meno ad una distribuzione gaussiana, in modo tale di avere la certezza di poter applicare un test dove essa è assunta senza commettere errore. Uno di essi è il test di normalità di Shapiro-Wilk, che si ritiene essere molto potente anche qualora si abbiano poche osservazioni.
Essendo $Y = y_1, y_2, ..., y_N$ il vettore contenente gli N campioni che si hanno a disposizione, il test procede come segue:

1. i campioni vengono ordinati in senso crescente;

2. se i dati sono dei campioni casuali tratti da una distribuzione normale, di cui non si conoscono la media μ e al varianza σ^2, possono essere rappresentati come un'equazione lineare del tipo:

$$y_i = \mu + \sigma * x_i \text{ con i=1,...,N} \tag{44}$$

dove le x_i sono in insieme ordinato di campioni casuali estratti dalla distribuzione N(0,1);

3. si calcola quindi il vettore B

$$B = \frac{m^T * V^{-1}}{\sqrt{m^T * V^{-1} * V^1 * m}} \tag{45}$$

dove V è la matrice di covarianza degli x_i e m è il vettore contenente i valori attesi degli x_i;

4. viene calcolato il parametro

$$W = \frac{\sum_{i=1}^{N} b_i * y_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{46}$$

dove $\bar{y}$ è la media dei campioni e quindi al denominatore è presente la somma degli scarti quadratici;

5. un alto valore di W indica che è molto probabile che i campioni siano tratti da una distribuzione normale;

6. dalla statistica W può essere calcolato il p-value e, fissato un livello di significatività α (di solito al 0.05%) se:
   - p-value<soglia la distribuzione da cui sono tratti i campioni non può essere considerata normale;
   - p-value>soglia la distribuzione da cui sono tratti i campioni può essere considerate normale;

6.1.3 Correzione per test multipli

Come detto sopra, di norma una buona scelta del livello di significatività per un singolo test statistico è del 5%. Questo significa che la probabilità di rifiutare l'ipotesi nulla quando invece essa è vera è del 5%. Se però si deve compiere un numero elevato di test statistici (supponendo che ognuno sia indipendente dagli altri) con ogni volta α=5%, la probabilità di selezionare almeno 1 falso positivo diventerà $1 - (1 - \alpha)^N$, con N numero dei test svolti. Di conseguenza il livello di significatività non è più quello voluto ma è molto più alto.
Per attuare la correzione del livello di significatività per test multipli si sono scelte 2 strade:

1. **Correzione di Bonferroni**: viene impostato $\alpha = \frac{\text{liv.di significatività voluto}}{N}$. Questo metodo è molto conservativo, in quanto α diventa molto piccolo. C'è quindi il rischio di commettere errori di tipo 2 e cioè di accettare l'ipotesi nulla quando invece non è vera;

2. **False Discovery Rate** (FDR): essa è così definita

$$\begin{aligned} FDR &= E\left[\frac{\#FP}{\#Selezionati}\right] && \text{se } \#Selezionati > 0 \\ FDR &= 0 && \text{se } \#Selezionati = 0 \end{aligned} \tag{47}$$

6.1.4 Implementazione

Il linguaggio R (versione 2.15.3) ha a disposizione delle funzioni che permettono di eseguire entrambi i test; esse sono:

- **Test di Student**: funzione *t.test*. Essa prende in ingresso 2 vettori contenenti i campioni rispettivamente della classe 1 e 2. Permette di impostare il tipo di test che si vuole svolgere con il parametro *alternative* che può assumere i valori di *two-sided* per l'omonima tipologia di test o *less* o *greater* per i test *one-sided*. In uscita si ottengono i p-value attraverso *$p.value*. ;

- **Test di Shapiro-Wilk**: funzione *shapiro.test*. Essa prende in ingresso un vettore contenente i campioni della distribuzione che si vuole testare. Anche in questo caso in uscita si ottengono i p-value attraverso *$p.value*

6.2 GSEA

La *Gene Set Enrichment Analysis* (GSEA) [20][21] è un metodo computazionale che serve a determinare se un *gene set* mostra delle significative differenze tra due stati biologici (come possono essere due fenotipi). Un *gene set* è un raggruppamento di geni accomunati secondo un determinato criterio; esistono molti tipi di gene sets, di cui qui sotto vengono riportati alcuni esempi:

- GO gene sets: sono composti da geni annotati nel medesimo termine GO;

- gene sets che derivano da database di pathway che si trovano online o pubblicati in importanti giornali scientifici.

Questo metodo è molto utile quando si ha un alto numero di geni/proteine e un numero molto più basso di campioni; infatti, in questi casi, a causa della correzione per test multipli, è difficile identificare dei gruppi funzionali di geni caratterizzati da una bassa, se pur presente, espressione differenziale e questo è proprio l'obiettivo che si pone la GSEA. Generalmente, tale analisi prende

in ingresso dei campioni dei profili di espressione dei geni (o prodotti genici) appartenenti alle 2 classi, e fornisce in uscita, per ogni gene set considerato, dei parametri che indicano il grado di *arricchimento* di una delle 2 classi in tale set. È disponibile online un tool che permette di compiere le GSEA, sviluppato dal Broad Institute of MIT and Harvard [22].

6.2.1 Metodo

La GSEA considera i valori di espressione dei geni, di cui si hanno più campioni appartenenti alle 2 classi di interesse. Viene poi calcolato per ogni gene uno *score* che sta ad indicare quanto alta è la correlazione tra le 2 classi; i geni vengono quindi ordinati secondo il valore dello *score* (ottenendo una *ranked list* L).
Considerando poi un gene set S definito a priori, l'obiettivo della GSEA è determinare se i membri di tale insieme siano distribuiti in modo casuale in L oppure siano maggiormente concentrati all'inizio o alla fine. Se si verifica l'ultima condizione significa che c'è un arricchimento in una delle 2 classi.
Per determinare se c'è o meno un arricchimento vengono calcolati dei parametri:

- **ES (Enrichment Score)**: esso viene calcolato percorrendo la lista ordinata e andando ad incrementare una variabile Σ ogniqualvolta il gene della lista appartiene al set considerato, decrementandola se il gene non è in S. L'incremento può essere costante o dipendere dallo score del gene. L'ES è la massima deviazione da 0, positiva o negativa, di Σ, ed è tanto più alto in modulo quanto più i geni appartenenti al set sono associati al fenotipo.

- **Leading edge subset**: è il sottoinsieme dei geni appartenenti al gene set che contribuiscono maggiormente all'ES, quelli quindi che stanno tra l'inizio della lista e l'ES nel caso di ES positivo, oppure tra l'ES e la fine della lista in caso di ES negativo.

- **P-value**: esso indica il livello di significatività statistica dell'ES stimato. Viene calcolato attraverso la permutazione delle etichette delle classi di appartenenza dei campioni oppure creando dei gene sets casuali e ricalcolando per ognuno di essi l'ES. In questo modo viene calcolata la distribuzione nulla e il p-value viene calcolato in relazione ad essa.

- **NES (Normalized Enrichment Score)**: per capire quali gene sets sono maggiormente arricchiti è necessario avere un valore di riferimento che tenga conto delle differenze nelle loro dimensioni. L'ES viene quindi normalizzato:

$$\mathrm{ES}(S_i) = \text{ES dell'i-esimo gene set} \tag{48}$$

$$ES(S_i, p_j) = \text{ES della j-esima permutazione dell'i-esimo gene set} \quad (49)$$

$$NES(S_i) = \frac{ES(S_i)}{media(ES(S_i, p_j))} = \text{NES dell'i-esimo gene set} \quad (50)$$

Tanto più grande in modulo è il NES, tanto più significativo è l'arricchimento. Se esso ha un valore positivo allora l'arricchimento si avrà in cima alla *ranked list*, viceversa se esso è negativo. Considerando il caso in cui si stiano confrontando 2 fenotipi, nel primo caso si ha una correlazione con il primo fenotipo, nell'altro con il secondo.

- **FDR (False Discovery Rate)**: è la stima della probabilità che un gene set con un dato NES rappresenti un falso positivo.

6.2.2 Settaggio dei parametri per l'analisi

Il tool richiede di fornire:

- il **database dei gene sets**: esso contiene tutti i gene sets su cui si vuole effettuare l'analisi. Il sito mette a disposizione molte collezioni di gene sets [23];
- i **dati di espressione** contenenti i campioni che si hanno a disposizione per entrambi i fenotipi;
- le **etichette dei fenotipi** in cui viene specificato quali sono i fenotipi considerati e quali sono i campioni associati ad ognuno di essi;
- il **numero e tipo di permutazioni** da usare per la stima dei parametri;
- la **pesatura** con cui incrementare/decrementare la somma percorrendo la ranked list;
- la **statistica** con cui generare la ranked list (test di Student, foldchange e foldchange logartimica, differenza tra le medie e rapporto segnale/rumore - inteso come differenza tra le medie diviso per la somma delle SD):
- il numero massimo e minimo di geni appartenenti alla ranked list e al gene set; se tali limiti vengono superati il gene set viene filtrato (e quindi non considerato) dall'analisi.

6.2.3 Risultati forniti

Come valori di ritorno il tool fornisce:

- il numero di gene sets arricchiti per ciascun fenotipo, indicando per quanti tale arricchimento è significativo (FDR<25%) e per quanti il p-value è minore del 5% e dell'1%;
- per i gene sets con maggior NES viene fornito l'**enrichment plot**, in cui viene riportato l'andamento della Σ man mano che si percorre la ranked list e la posizione dei geni presenti in essa e appartenenti al gene set; viene anche fornito un report in cui si specifica, per ogni set, quali geni delle lista sono presenti e a che punto di essa, e quali fanno parte del *leading edge subset*;
- la **heat map** dei geni appartenenti ai data set arricchiti: essa rappresenta i valori di espressione con dei colori (rosso, rosa, azzurro e blu). Ad ogni campione viene attribuito un colore tanto più caldo quanto più alto è il suo livello di espressione (alto, moderato, basso, bassissimo);
- informazioni generali relative all'analisi svolta (per ogni set quanti geni si trovano anche nella ranked list, numero di set filtrati, ranked list con relativi scores, i geni che fanno parte del leading edge subset).

7 ANALISI DELL'ESPRESSIONE DIFFERENZIALE: RISULTATI

7.1 IDENTIFICAZIONE DEL PARAMETRO

Nella Figura 23 sono riportati alcuni esempi di fit del $\frac{P_H}{P_L}$ di alcune proteine, ottenuti mediante la stima con i pesi assoluti, con il modello dell'errore di misura precedentemente scelto.
Come si vede, l'andamento delle proteine viene seguito molto bene sia nei casi in cui si abbiano a disposizione tutti e 3 i campioni, sia in quelli in cui se ne abbiano solo 2. Il grafico in alto a destra rappresenta il caso in cui si riscontra una decrescenza tra il primo e il secondo campione. Come detto in 5.3 sono state considerate anche le proteine con tale discrepanza rispetto alle assunzioni iniziali; l'errore di misura così impostato permette nel fit di scegliere un compromesso tra i 2 istanti di campionamento iniziali.
La bontà del modello viene validata con l'analisi della precisione delle stime: come si vede in Figura 24 il CV del parametro stimato è sempre $< 50\%$.

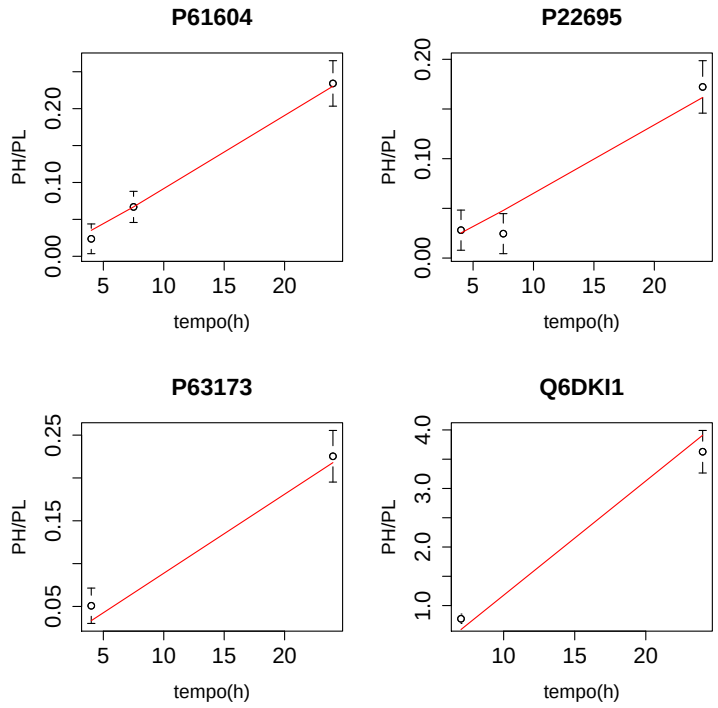


Figura 23: Esempi di fit di proteine ottenuti attraverso la stima con i pesi assoluti, con il modello dell'errore di misura precedentemente scelto.

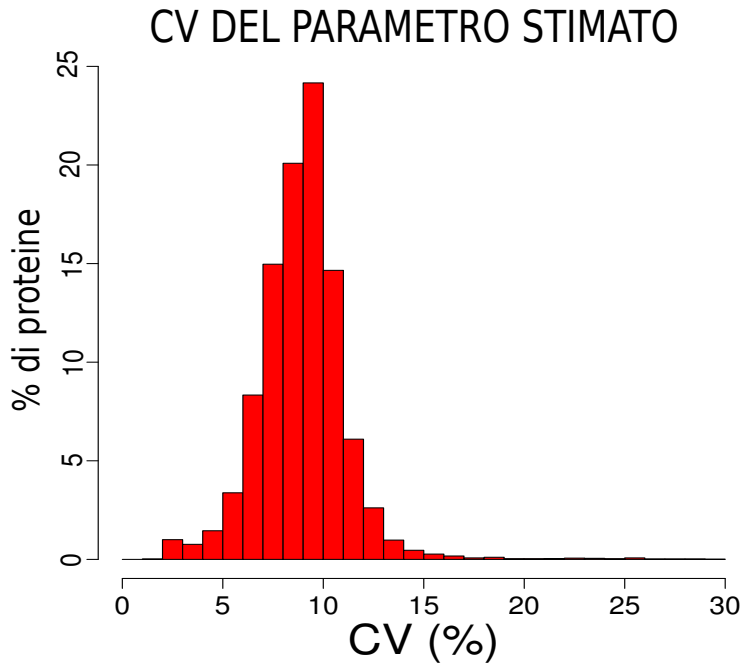


Figura 24: Istogramma del CV del parametro stimato in tutte le proteine di tutti i soggetti. Come si vede esso è sempre $< 50\%$.

Vengono inoltre riportati in 25 e 26 gli istogrammi dei k ottenuti per i 10 soggetti.

7.2 TEST STATISTICI: RISULTATI

Il fine dell'analisi svolta da qui in poi è stato quello di determinare se per alcune proteine si riscontrasse una significativa differenza nei valori dei k (e quindi nelle emivite) di alcune di esse tra le 2 classi di soggetti. È opportuno chiarire cosa questo comporti a livello intracellulare:

- proteine per cui il k risulti maggiore nei soggetti diabetici e nefropatici: esso indica una loro degradazione più veloce, e quindi, supponendo che rimanga inalterata la loro produzione, si dovrebbero riscontrare livelli di espressione più bassi;
- proteine per cui il k risulti minore nei soggetti diabetici e nefropatici: esso indica una loro degradazione più lenta, e quindi, determinerebbe livelli di espressione più alti.

Per far questo si è andati innanzitutto a selezionare tutte le proteine per cui ci fossero almeno 3 campioni per entrambe le classi di soggetti (D e DN), numero minimo necessario per poter attuare il test di Student. Esse sono risultate essere 776.

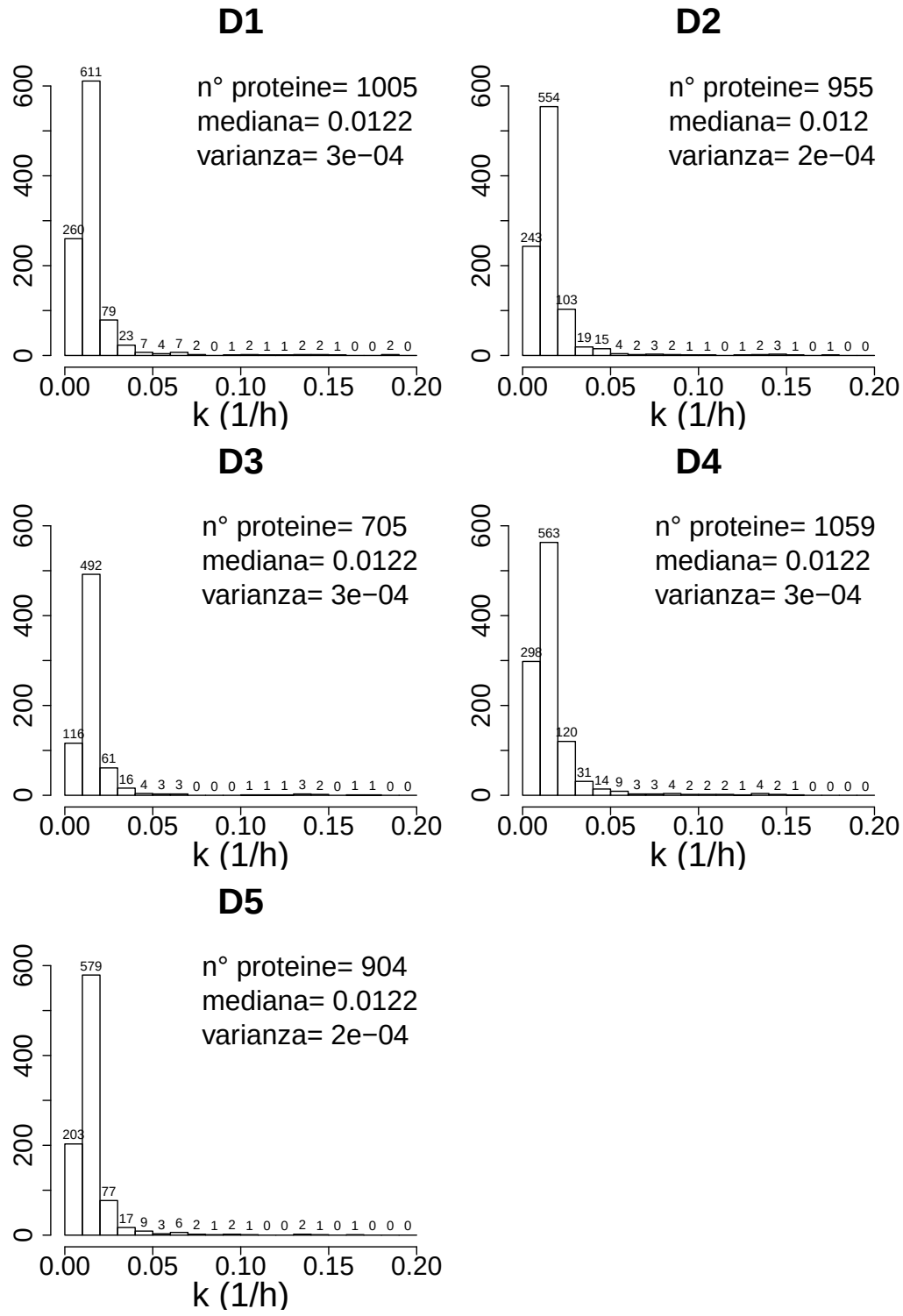


Figura 25: Istogrammi del valore del parametro k stimato dalle proteine dei soggetti diabetici. Sono riportati anche il numero di proteine considerate e mediana e varianza dei k.

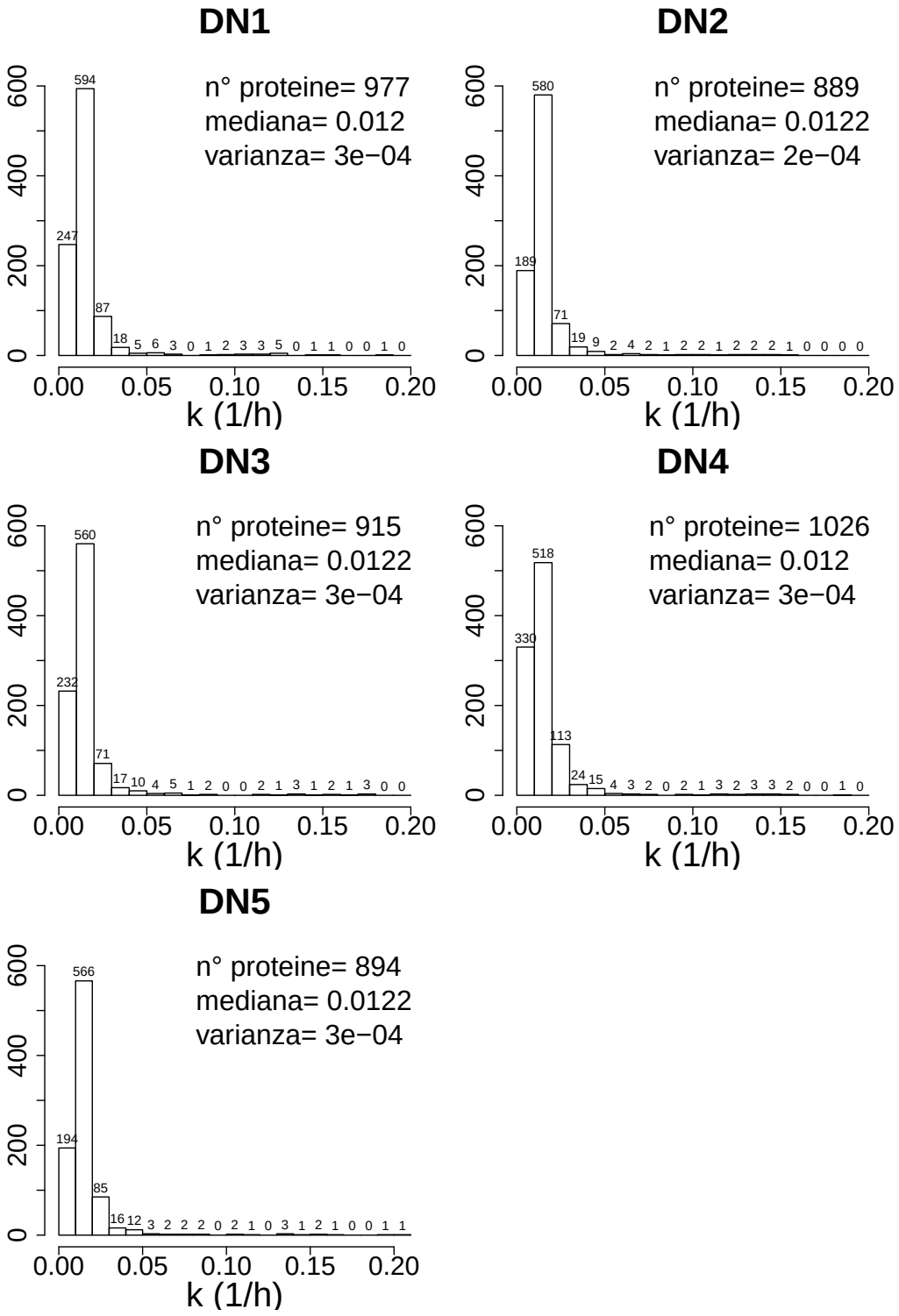


Figura 26: Istogrammi del valore del parametro k stimato dalle proteine dei soggetti diabetici e nefropatici. Sono riportati anche il numero di proteine considerate e mediana e varianza dei k.

Il t test richiede che le distribuzioni da cui sono tratti i dati siano gaussiane, e si è quindi applicato il test di Shapiro-Wilk per verificare se tale ipotesi fosse soddisfatta. Avendo a disposizione pochi soggetti, esso è stato fatto considerando tutte le proteine; prendendo come livello di significatività $\alpha = 0.05$, è risultato che per il 96% il p-value>α (il 100% applicando la correzione di Bonferroni). Si è quindi ritenuto corretto ipotizzare la gaussianità delle distribuzioni del k.
Dal test di Student *two-sided* è risultato che:

- per 13 proteine il p-value $\leqslant$ 0.05 (vedi tabella 6). Di queste per 8/13 il k è maggiore (e quindi si ha una degradazione più veloce) nei pazienti diabetici; per 5/13 invece il k è maggiore nei pazienti diabetici e nefropatici;

GENE NAME	P-VALUE	mediana D	mediana DN
TPP2	0.011	0.0117	0.0094
PLOD1	0.012	0.0234	0.0325
PDCD6IP	0.015	0.016	0.0149
TWF1	0,016	0,0175	0,0163
CCT7	0.023	0.0128	0.0111
NCKAP1	0.027	0.0136	0.0147
ACTB	0.033	0.0103	0.0085
TBCA	0.037	0.0167	0.0155
AKR1C1	0.04	0.0111	0.0134
HGS	0.049	0.016	0.0178
CAP1	0.05	0.0104	0.0099
RAB5C	0.05	0.0175	0.0167
COL6A3	005	0.0781	0.0973

Tabella 6: Proteine per cui il p-value ottenuto mediante il test di Student risulta <0.05. In rosa quelle per cui il k è maggiore per i soggetti DN, in azzurro quelle per cui il k è significativamente maggiore per i pazienti D.

- andando ad attuare la correzione per test multipli con Bonferroni, non ne risulta nessuna con k significativamente diverso nelle 2 classi;
- anche ponendo come soglia la FDR<0.05 non si hanno proteine con significative differenze nelle 2 classi.

Il fatto di non aver selezionato nessuna proteina a seguito dei test multipli è dovuto all'esiguo numero di campioni rispetto all'invece alto numero di proteine considerate. Provenendo inoltre i campioni da soggetti diversi, entrano in gioco molti fattori che contribuiscono alla variabilità (e quindi aumentando la varianza diminuiscono in modulo il valore della statistica t).

7.3 GSEA: RISULTATI

I test statistici sulle proteine non individuano delle chiare correlazioni tra proteine e classi di soggetti. Si è quindi è deciso di concentrarsi non tanto sulle singole proteine, quanto su gruppi di esse legate da determinate caratteristiche (funzione biologica, appartenenza ad un pathway...). Per far questo si è applicata ai dati la GSEA, seguendo le linee guida sotto esposte:

- sono stati usati 4 differenti database di gene sets:
  - GO gene sets: che includono i set derivanti dai termini GO legati a processi biologici, funzioni molecolari e componenti cellulari;
  - BIOCARTA, KEGG e REACTOME gene sets: che includono i set derivanti dai pathways inclusi negli omonimi database.
- il software dello spettrometro di massa fornisce gli identificatori delle proteine in formato *Uniprot*. Esse sono state mappate attraverso il portale *Biomart* [24] nei *Gene Symbols* (ID) che permettono di interfacciarsi ai database. In alcuni casi (rarissimi) si avevano a disposizione i k stimati da 2 isoforme della stessa proteina, che vengono mappate nello stesso ID; non potendo comparire nella ranked list due volte la stessa proteina, si è deciso di mediare, per ogni paziente di cui si ha il dato, il valore dei k stimati e usare tali valori per la GSEA;
- la statistica impostata, coerentemente alle considerazioni sulla normalità dei dati precedentemente illustrate, è quella di Student; per un'ulteriore verifica si è anche applicata la statistica segnale/rumore;
- è stato impostato a 10000 il numero di permutazioni, che sono state fatte sui geni in quanto si hanno a disposizione un basso numero di campioni per ogni fenotipo;
- si è scelta una pesatura basata sullo score di ogni proteina per l'incremento della Σ nella determinazione dell'ES;
- si sono filtrati i gene sets a cui appartenessero meno di 5 proteine;

I risultati ottenuti con entrambe le statistiche sono concordi e pressoché identici; a seguire sono quindi riportati solo quelli in cui si è impostata la statistica t.
Per identificare i gene sets significativamente arricchiti in una delle classi di soggetti, oltre a considerare quelli con un alto valore del modulo del NES, di norma sono state scelte le soglie di FDR<5% (infatti, avendo usato la tipologia di permutazioni dei gene sets e avendo pochi campioni, il valore della significatività è poco stringente, per cui è necessario considerare una stringente soglia per l'FDR) e p-value<1%. Si è imposto inoltre che il numero di geni appartenenti alla ranked list e presente nel gene set non fosse troppo esiguo rispetto al numero totale di geni da cui è composto il gene set.

Arricchimenti significativi per la classe D

I gene sets riportati in Tabell 7 sono arricchiti nei diabetici: questo sta a significare che i k delle proteine in essi contenute sono generalmente più alti nei pazienti diabetici e quindi più bassi nei pazienti affetti anche da nefropatia.

- Sia nel database KEGG (Fig:27), sia in BIOCARTA (Fig:28) si è riscontrato un arricchimento nei gene set che identificano il **Proteaosome Pathway**. Il proteosoma è un complesso multiproteico che ha la funzione di degradare polipeptidi.

- In REACTOME (Fig:29) risultano significativamente arricchiti i pathway che coinvolgono il **TRiC** (CCT for chaperonin containing TCP-1). I chaperoni sono proteine che creano le condizioni favorevoli affinché avvenga il corretto ripiegamento di altre proteine. Più precisamente è stato identificato il gruppo di geni coinvolti nel folding e prefolding della tubulina attraverso l'intermediazione dei CCT/Tric (vengono riportati solo i grafici del secondo gene set in quanto le proteine coinvolte sono le stesse in entrambi).
  Grazie a questa famiglia di proteine sia in REACTOME che nel database dei GO gene sets (anche se con un FDR=15%) c'è un arricchimento per il gene set del **Protein Folding**, cioè il processo con cui una proteina si ripiega per assumere la sua conformazione tridimensionale. Questo arricchimento non è però da considerarsi significativo in quanto le proteine maggiormente influenti sono quelle già presenti nei 2 precedenti pathway e il numero di proteine nella ranked list è molto minore rispetto a tutte le proteine presenti nel gene set testato.

DATABASE	GENE SET	NUM.PROT.	NES	P-VALUE	FDR
GO	Protein Folding	13/55	1.89	0.002	0.15
KEGG	Proteaosome	23/44	1.93	$<10^{-4}$	0.033
REACTOME	Protein Folding	13/42	2.31	$<10^{-4}$	$<10^{-4}$
REACTOME	Prefoldin mediated transfer of substrate to CCT/TRIC	11/21	2.25	$<10^{-4}$	$<10^{-4}$
REACTOME	Formation of tubulin folding intermediate by CCT/TRIC	10/14	2.17	$<10^{-4}$	<0.001
BIOCARTA	Proteaosome pathway	17/27	1.85	0.003	0.026

Tabella 7: Gene set significativamente arricchiti nella classe D. Per ognuno è riportato: database da cui sono tratti, nome del gene set, numero di proteine appartenenti alla ranked list rispetto al dimensione del gene set, NES, p-value e FDR

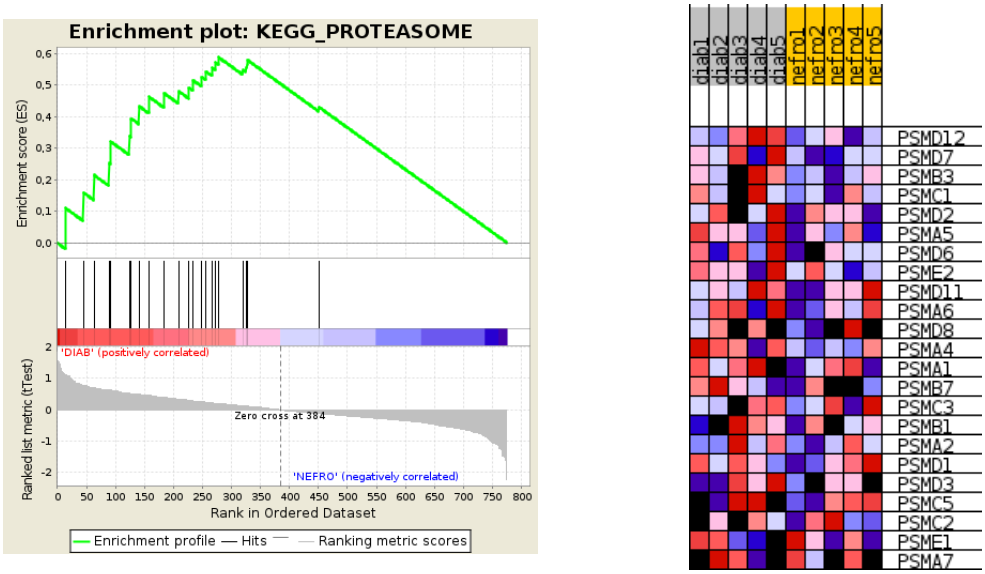


Figura 27: KEGG PROTEAOSOME PATHWAY: enrichment plot (a destra) e heat map (a sinistra)

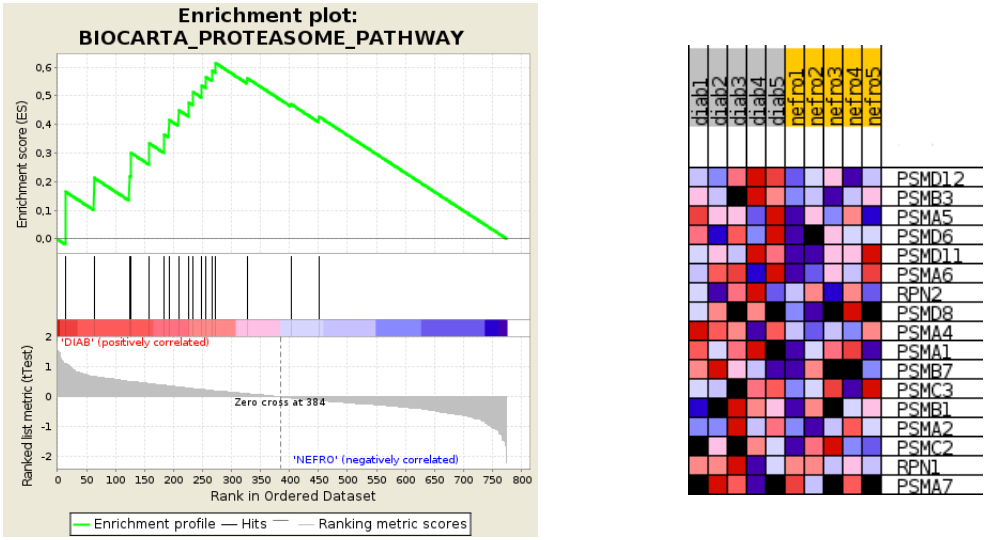


Figura 28: BIOCARTA PROTEAOSOME PATHWAY: enrichment plot (a destra) e heat map (a sinistra)

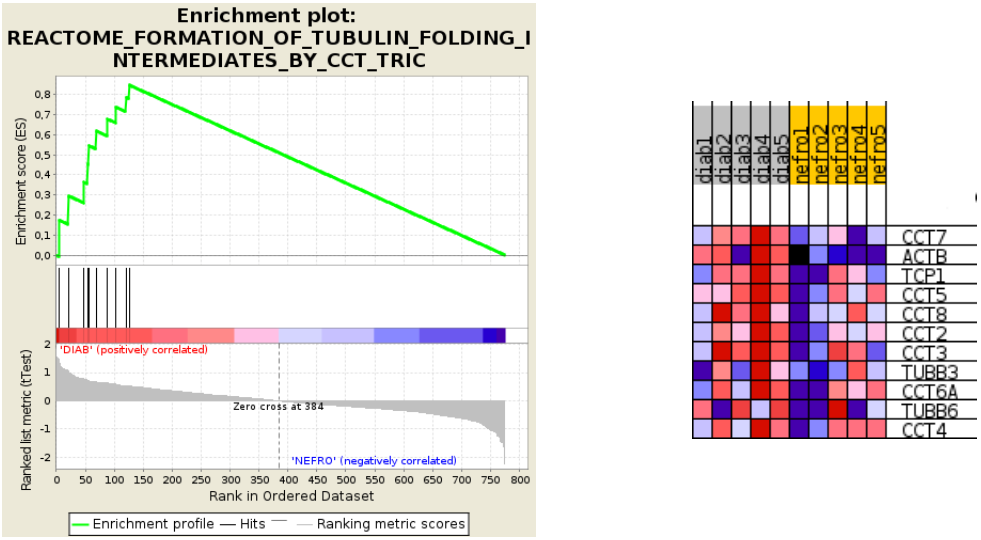


Figura 29: REACTOME FORMATION OF TUBULIN FOLDING INTERMEDIATE BY CCT: enrichment plot (a destra) e heat map (a sinistra)

Arricchimenti significativi per la classe DN

I gene sets riportati in Tabella 8 sono arricchiti nella classe dei diabetici e nefropatici: questo sta a significare che i k delle proteine in essi contenute sono generalmente più alti nei pazienti diabetici e nefropatici rispetto a quelli solo diabetici.

- Sia nel database KEGG (Fig:30) che nella GO (Fig:31) risultano arricchiti significativamente i gene sets che interessano l'**attività ribosomiale**. I ribosomi sono gruppi di molecole e proteine responsabili della sintesi proteica; più precisamente sintetizzano le proteine leggendo le informazioni contenute nell'mRNA. In KEGG si ha un arricchimento nel pathway dei ribosomi, mentre in GO nel gene sets in cui sono annotati i termini GO legati ai costituenti strutturali dei ribosomi.

- anche in REACTOME 32 risultano arricchiti molti pathway che coinvolgono i ribosomi. Quello con NES maggiore è il **SRP-dependent cotranslational protein targeting to membrane**: SRP è la particella di riconoscimento del segnale che è una ribonucleoproteina che riconosce e trasporta le proteine verso la parete citosolica del reticolo endoplasmatico rugoso. Altri pathway, in cui sono presenti molte proteine ribosomiali, sono:
  - **peptide chain elongation pathway**: è l'insieme di processi che determina l'allungamento della catena polipeptidica, grazie all'aggiunta di un amminoacido, che si svolge all'interno del ribosoma;
  - **3-UTR mediated translation regulation pathway**: meccanismi di controllo della traduzione che avvengono, o sono mediati, all'estremità 3-UTR a seguito del legame ad essa di proteine specializzate;

DATABASE	GENE SET	NUM.PROT.	NES	P-VALUE	FDR
GO	Structural constituent of ribosome	53/80	-2.30	$<10^{-4}$	$<10^{-4}$
KEGG	Ribosome	65/86	-2.33	$<10^{-4}$	$<10^{-4}$
REACTOME	SRP-dependent cotranslational protein targeting to membrane	74/127	-2.22	$<10^{-4}$	0.001
REACTOME	Peptide chain elongation	67/104	-2.18	$<10^{-4}$	$<10^{-4}$
REACTOME	Nonsense mediated decay enanched by the exon junction complex	70/124	-2.07	$<10^{-4}$	0.002
REACTOME	3-URT mediated translation regulation	71/114	-1.94	$<10^{-4}$	0.009
REACTOME	Collagen formation	11/58	-1.73	0.012	0.05

Tabella 8: Gene set significativamente arricchiti nella classe DN. Per ognuno è riportato: database da cui sono tratti, nome del gene set, numero di proteine appartenenti alla ranked list rispetto al dimensione del gene set, NES, p-value e FDR

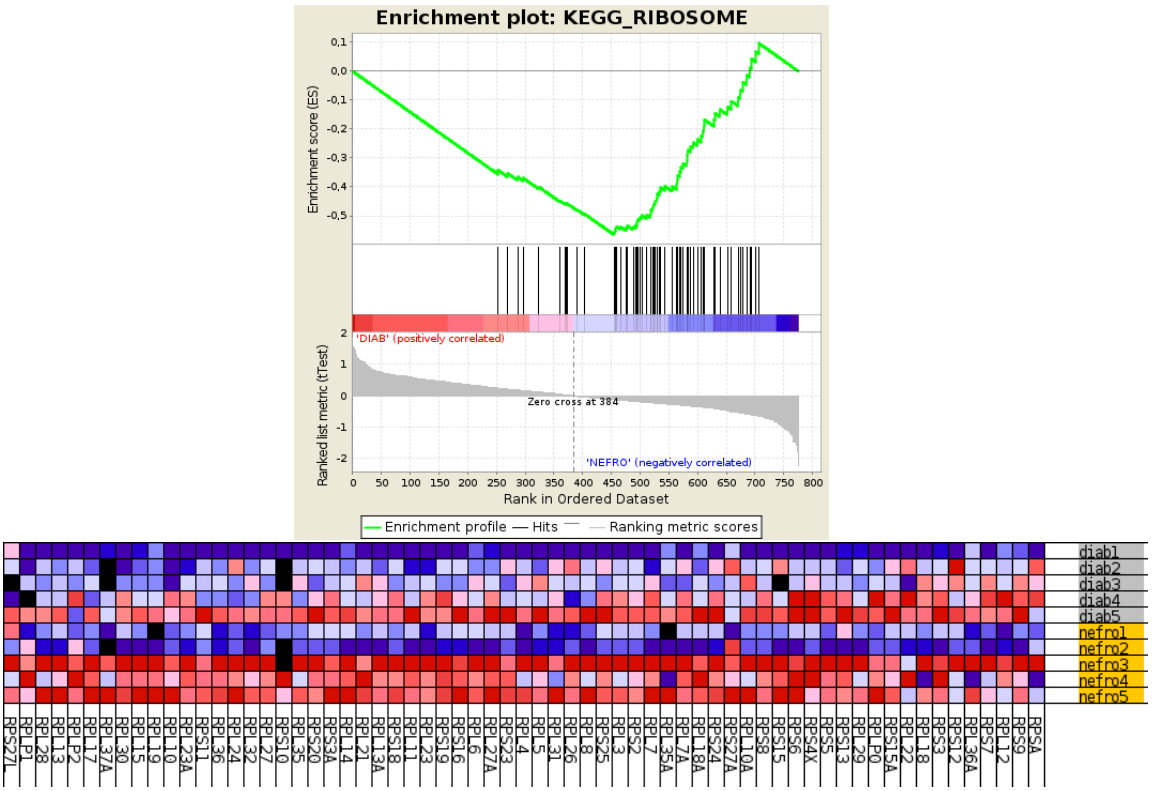


Figura 30: KEGG RIBOSOME PATHWAY: enrichment plot (in alto) e heat map (in basso)

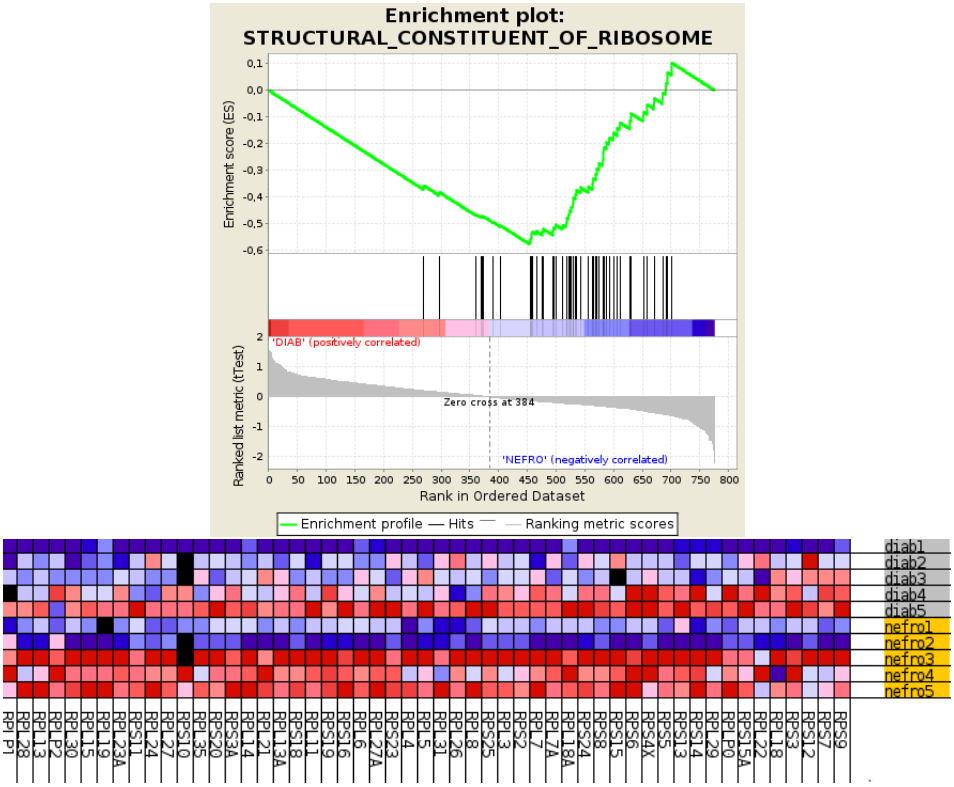


Figura 31: GO STRUCTURAL CONSTITUENT OF RIBOSOME: enrichment plot (in alto) e heat map (in basso)

  - **nonsense mediated decay enhanced by the exon junction complex**: questo pathway ha una funzione di controllo, in quanto provvede a eliminare i trascritti di mRNA che contengono codoni di stop prematuri, che durante la traduzione potrebbero portare a proteine degeneri. Si serve di diversi meccanismi tra cui quello che coinvolge il complesso di giunzione esonica.

  Di questi ultimi 3 non vengono riportati i plot in quanto le proteine coinvolte sono nella quasi totalità le stesse dei gene sets REACTOME precedenti.

- pur essendo i parametri un po' al di sopra della soglia (p-value=0.012 e FDR=5%), è di interesse segnalare l'arricchimento nel pathway realativo alla **formazione del collagene** in REACTOME 33. Infatti studi precedenti [25] hanno già rilevato un legame tra la nefropatia e la variazione nell'espressione del collagene.

Gli heat plot riguardanti i gene sets arricchiti nella classe DN mostrano come si abbia una maggior concentrazione di colori caldi nei k stimati dai dati dei pazienti nefropatici, rispetto a quelli dei pazienti solo diabetici. Si ha un ulteriore riscontro negli enrichment plot in cui si vede che le proteine sono concentrate alla fine della ranked list. Le stesse osservazioni, valgono, anche se in senso inverso, per i gene sets arricchiti nella classe D.

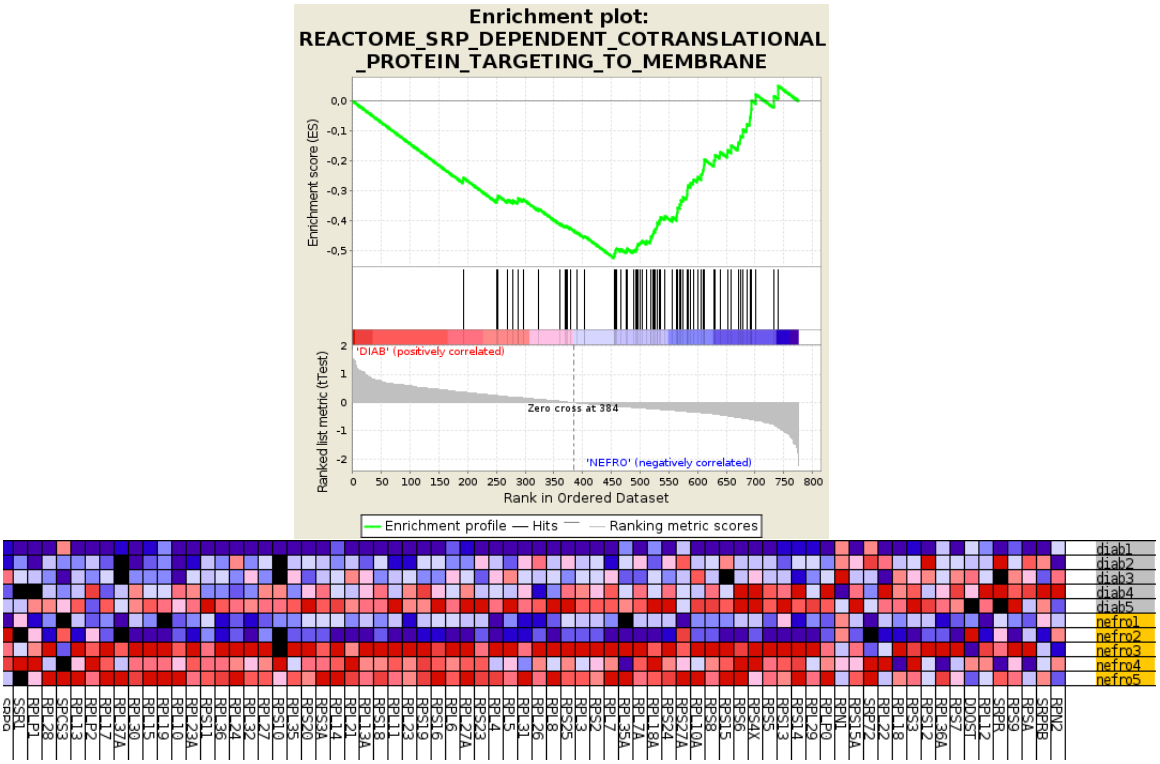


Figura 32: **SRP-dependent cotranslational protein targeting to membrane**: enrichment plot (in alto) e heat map (in basso)

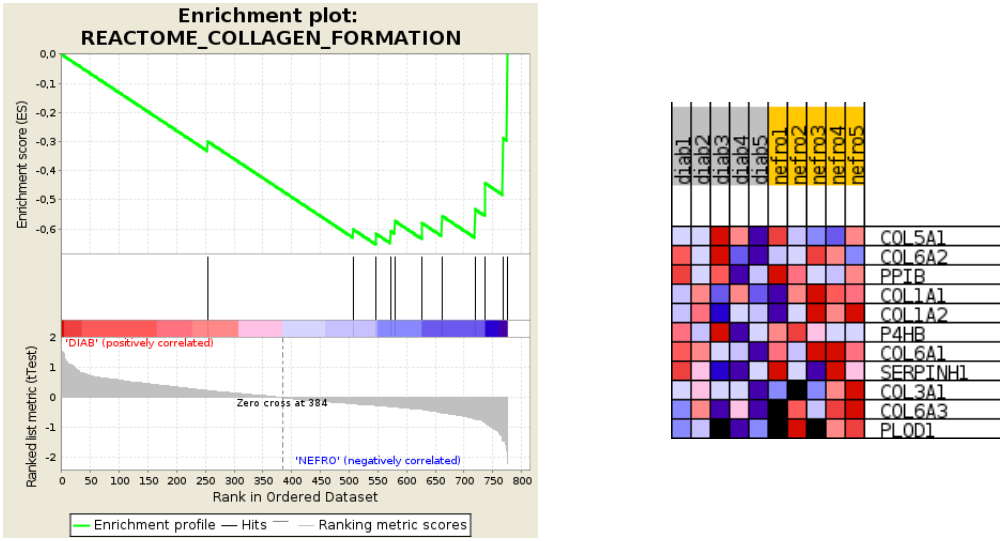


Figura 33: REACTOME FORMATION OF COLLAGENE: enrichment plot (a destra) e heat map (a sinistra)

Dal punto di vista biologico avere una degradazione più veloce per le proteine legate ai ribosomi, essendo essi responsabili della sintesi proteica, potrebbe far pensare che nei pazienti diabetici e nefropatici ci possa essere una globale diminuzione del livello di espressione di tutte le proteine; infatti una minor quantità di ribosomi determinerebbe una minor sintesi di nuove proteine. Questa tesi può trovare ulteriore conferma nel fatto che i proteosomi sembrano invece avere un turnover più lento: essi, essendo responsabili della degradazione delle proteine, se presenti in maggior quantità, implicherebbero una maggior degradazione generale delle proteine.

8 CONCLUSIONI

L'obiettivo della presente tesi, è stato quello di formulare, identificare e validare un modello per la stima del turnover a livello di proteine singole all'interno di fibroblasti cutanei; esso è stato poi utilizzato per confrontare la velocità di degradazione in pazienti diabetici di tipo 1 affetti e non da nefropatia diabetica al fine di rilevare se esistano dei biomarkers per tale malattia. Dai risultati ottenuti innanzitutto si è potuto stabilire che il modello applicato per rappresentare il turnover proteico, pur essendo semplice, descrive bene tale processo. La scelta di non renderlo troppo complicato è stata determinata dal fatto di avere a disposizione un numero limitato di istanti di campionamento: avere troppi parametri da stimare, avrebbe sicuramente compromesso l'affidabilità della stima.
Queste circostanze hanno anche implicato un'attenta analisi dell'errore che la metodica con cui sono state fatte le misure introduceva: pur usando lo stesso strumento, gli esperimenti sono stati fatti in tempi diversi, e questo comporta una variabilità tecnica, seppur piccola, delle misure. Essendo i dati high-troughput l'errore viene ulteriormente amplificato. Si è quindi posta molta cura nella definizione del corretto modello dell'errore di misura e nella normalizzazione dei dati.
Si è anche dovuto porre attenzione a che fossero soddisfatte le ipotesi di valenza del modello: il $\frac{P_H}{P_L}$ doveva rispettare un andamento sempre crescente. Nella scelta dei tempi di campionamento si è tenuto conto di questo, limitando la durata dell'esperimento ai tempi che non contraddicessero tale assunzione.
È stata anche attentamente vagliata la strada di affidarsi ai dati globali delle proteine, assicurandosi che i risultati con essi ottenuti non fossero in disaccordo con quelli ottenuti usando le misure sui singoli peptidi.

Pur non rilevando significative differenze sulle singole proteine attraverso il test statistico di Student, la successiva analisi di arricchimento, ha evidenziato delle correlazioni tra processi biologici e fenotipi. Esse sembrano avere una plausibilità biologica: un'emivita maggiore (e quindi una degradazione più lenta) delle proteine legate ai proteosomi e un'emivita minore (e quindi una degradazione più veloce) di quelle legate a pathway che coinvolgono i ribosomi, farebbe pensare che la nefropatia implichi una generale alterazione del livello di tutte le proteine. Dai dati a disposizione, non avendo informazioni sulla concentrazione totale di ogni proteina, non se ne può però ancora trovare conferma.

8.1 SVILUPPI FUTURI

Nel futuro l'obiettivo sarà quello di dare maggior valenza ai risultati esposti nel capitolo precedente.
Innanzitutto si punterà ad aumentare il numero di soggetti, al fine di accrescere la potenza statistica. Infatti, i risultati che sono stati ottenuti fino ad ora possono essere inficiati da caratteristiche proprie dei vari soggetti (diverso metabolismo, caratteristiche fisiologiche...); con l'aumentare del loro numero, acquisterebbero affidabilità.
Aumentando la cardinalità di ogni classe si potrà anche pensare ad un modello alternativo e più complesso con cui spiegare più nel dettaglio il turnover proteico.
Si potrà inoltre cercare di aumentare il numero di proteine considerate: lo spettrometro di massa infatti spesso non fornisce il dato globale di una proteina in un determinato istante di campionamento perché non ha disposizione tale misura per almeno 2 peptidi appartenenti ad essa. Per questo motivo non si è potuto fare il fit di molte proteine (tutte quelle per cui ciò si è verificato al campione delle 24 ore, $\simeq 15\%$). L'idea è quindi in quei casi di provare a considerare come dato globale di proteina quello ottenuto dall'unico peptide, verificandone però prima l'attendibilità.
Infine, l'obiettivo sarà quello di integrare i risultati ottenuti dall'analisi del turnover proteico con delle misure del livello di espressione delle proteine e dell'mRNA nelle stesse cellule. In questo modo si riuscirebbe ad avere una visione globale dell'alterazione che la nefropatia diabetica comporta sia a livello di trascrizione che di traduzione all'interno delle cellule considerate.

BIBLIOGRAFIA

[1] Cobelli C Toffolo G Di Camillo B. *Modelli del turnover e della regolazione proteica*. 2007.

[2] Zimmet PZ Alberti KGMM. "Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications - Part 1". In: *Diabetic Medicine* 15 (1998), pp. 539–553.

[3] *International Diabetes Federation - diabetes atlas*. 2012. URL: http://www.idf.org/diabetesatlas/.

[4] *World Health Organzation - Diabetes*. 2012. URL: http://www.who.int/mediacentre/factsheets/fs312/en/.

[5] Wahab NA Mason RM. "Extracellular Matrix Metabolism in Diabetic Nephropathy". In: *Journal of the American Society of Nephrology* 14 (2003), 1358–1373.

[6] Balthazar APS Thomazelli FCS Matos JD Canani LH Zelmanovitz T Gerchman F. "Diabetic nephropathy". In: *Diabetology and Metabolic* (2009), 1–10.

[7] Bortoloso E Mauer M Fioretto P Dalla Vestra M Saller A. "Structural involvement in type 1 and type 2 diabetic nephropathy". In: *Diabetes Metab* 26 (2000), Suppl 4 :1954–1960.

[8] Mauer M Caramori ML Fioretto P. "The need for early predictors of diabetic nephropathy risk: is albumin excretion rate sufficient?" In: *Diabetes* 49(9) (2000), pp. 1399–1408.

[9] Iori E Arrigoni G Vedovato M James P Coracina A Millioni R Tessari P Puricelli L. "Altered Chaperone and Protein Turnover Regulators Expression in Cultured Skin Fibroblasts from Type 1 Diabetes Mellitus with Nephropathy". In: *Journal of Proteome Research* 6 (2007), pp. 976–986.

[10] Newman JM Katz PP Sepe S Showstack J Selby JV FitzSimmons SC. "The natural history and epidemiology of diabetic nephropathy. Implications for prevention and control". In: *JAMA* 263 (1990), pp. 1954–1960.

[11] Hommel E Mathiesen ER Jensen JS Deckert T Parving HH Borch-Johnsen K Nørgaard K. "Is diabetic nephropathy an inherited complication?" In: *Kidney International* 41 (1992), 719–722.

[12] Rich S Barbosa J Seaquist ER Goetz FC. "Familial Clustering of Diabetic Kidney Disease". In: *The new England Journal of Medecine* 320 (1989), pp. 1161–1165.

[13] Iori E Trevisan R Tessari P Millioni R Puricelli L. "Skin fibroblasts as a tool for identifying the risk of nephropathy in the type 1 diabetic population". In: *Diabetes Metab Res Rev* 28 (2012), 62–70.

[14] Batlle D LaPointe MS. "Cultured skin fibroblasts as an in vitro model to assess phenotypic features in subjects with diabetic nephropathy". In: *American Journal of Kidney Diseases* 38 (2001), pp. 1239–1246.

[15] Messent J Tariq T Earle K Walker JD Viberti G Trevisan R Li LK. "Na+/H+ Antiport Activity and Cell Growth in Cultured Skin Fibroblasts of IDDM Patients With Nephropathy". In: *Diabetes* 41 (1992), pp. 1239–1246.

[16] Kofoed-Enevoldsen A Li LK Earle KA Trevisan R Viberti G Davies JE Ng LL. "Intracellular pH and Na+/H+ antiport activity of cultured skin fibroblasts from diabetics". In: *Kidney* 42 (1992), pp. 1184–1190.

[17] Mann M. "Functional and quantitative proteomics using SILAC". In: *Nature Reviews Molecular Cell Biology* 7 (2006), pp. 952–958.

[18] Mann M Ong SE. "A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC)". In: *Nature Protocols* 6 (2006), pp. 2650–2660.

[19] Kratchmarova I Kristensen DB Steen H Pandey A Mann M Ong SE Blagoev B. "Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics". In: *Molecular and Cellular Proteomics* 1 (2002), 376–386.

[20] Mootha VK Mukherjee S Ebert BL Gillette MA Paulovich A Pomeroy SL Golub TR Lander ES Mesirov JP Subramanian A Tamayo P. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proc. Natl. Acad. Sci. USA* 102 (2005), pp. 15545–15550.

[21] Eriksson KF Subramanian A Sihag S Lehar J Puigserver P Carlsson E Ridderstrale M Laurila E et al Mootha VK Lindgren CM. "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes". In: *Nature Genetics* 34 (2003), pp. 267–273.

[22] *GSEA, Broad Insitute*. URL: http://www.broadinstitute.org/gsea/.

[23] Pinchback R Thorvaldsdóttir H Tamayo P Mesirov JP Liberzon A Subramanian A. "Molecular signatures database (MSigDB) 3.0". In: *Bioinformatics* 27(12) (2011), pp. 1739–1740.

[24] *BioMart Central Portal*. URL: http://central.biomart.org/.

[25] In: ().

[26] "American Diabetes Association: Nephropathy in Diabetes". In: *Diabetes Care* 27 (2004), pp. 79–83.

[27] Li N Dittmar G Schuchhardt J Wolf J Chen W Selbach M Schwanhausser B Busse D. "Global quantification of mammalian gene expression control". In: *Nature* 473 (2011), 337–342.

[28] Puricelli L Arrigoni G Vedovato M Trevisan R James P Tiengo A Tessari P Millioni R Iori E. “Abnormal cytoskeletal protein expression in cultured skin fibroblasts from type 1 diabetes mellitus patients with nephropathy: A proteomic approach”. In: *Proteomics Clin. Appl.* 2(4) (2008), pp. 492–503.