UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA DELL'INFORMAZIONE

STUDIO DELLA CINETICA DELLE PROTEINE TRAMITE *LABELING* METABOLICO CON $^{2}H_2O$

Relatore: Prof.ssa Chiara Dalla Man

Laureanda: Silvia Salviato

ANNO ACCADEMICO 2021 – 2022

Data di laurea 23 settembre 2022

SOMMARIO **3**

INTRODUZIONE **5**

1. IL *LABELING* METABOLICO CON ^2H_2O **7**

1.1. Composizione, proprietà e utilizzo di ^2H_2O 7

1.1.1. Isotopi dell'idrogeno 7

1.1.2. Caratteristiche chimiche delle molecole di acqua pesante 8

1.1.3. Tossicità 9

1.1.4. Utilizzi comuni delle molecole di ^2H_2O 10

1.2. Aspetti clinici 10

1.3. Metodo di raccolta dei campioni e considerazioni sperimentali 11

1.3.1. Il *labeling* metabolico delle proteine 12

1.3.2. Studio *in vivo* del comportamento delle proteine 13

1.3.3. Raggiungimento del *plateau* 14

1.3.4. Criticità dell'utilizzo di *labeling* con ^2H_2O e dell'analisi *in vivo* 14

1.4. Utilizzo di isotopi radioattivi per i processi di marcatura metabolica 15

2. TECNICHE DI MISURA **17**

2.1. Spettrometria di massa (MS) 17

2.1.1. La strumentazione e il suo funzionamento 18

2.1.2. Procedure di analisi in presenza di composti proteici 21

2.1.3. Identificazione delle distribuzioni relative 24

2.2. *Mass Isotopomer Distribution Analysis* (MIDA) 24

2.2.1. Il parametro *n* 26

2.2.2. Impiego dell'analisi combinatoria nell'ambito della sintesi proteica 26

2.3. Limiti dell'analisi combinatoria per lo studio della dinamica dei composti 27

2.3.1. Cautele da adottare in fase di somministrazione e raccolta dei campioni 28

2.3.2. Raccolta dei campioni in relazione ai tempi di *turn over* 29

2.3.3. Somministrazioni successive 30

2.3.4. Ulteriori cautele da adottare in seguito alla raccolta dei campioni 31

3. STRUMENTI INFORMATICI A SUPPORTO DELLE ANALISI **33**

3.1. Introduzione al concetto di bioinformatica 33

3.2. Il perfezionamento degli strumenti di analisi 35

3.3. Il progresso della bioinformatica 36

3.4. Considerazioni generali sulle applicazioni degli strumenti informatici 39

3.5. Nuove strategie di analisi e sviluppi futuri 41

4. APPLICAZIONE DEL *LABELING* METABOLICO NELLO STUDIO DI PATOLOGIE CLINICHE 49

4.1. Considerazioni iniziali 49

4.2. Malattia di Alzheimer 51

4.3. Morbo di Parkinson 53

4.4. Impiego della marcatura con $^{2}H_2O$ nell'industria farmaceutica 56

CONCLUSIONI 59

Bibliografia 61

SOMMARIO

La comprensione della struttura del proteoma degli esseri viventi si basa sull'osservazione della dinamica delle proteine, per la quale si sono incontrati molti ostacoli in passato soprattutto a causa delle difficoltà nella raccolta e nell'analisi dei campioni. Queste sono state superate grazie all'impiego di $^{2}H_2O$, o acqua pesante, come marcatore dei processi metabolici.

L'acqua pesante è un composto sicuro per gli esseri viventi e ha caratteristiche chimico-fisiche che ne permettono l'utilizzo per diversi scopi. Tuttavia durante le prime fasi di questi studi, l'incompletezza dei dati raccolti dai ricercatori e l'inadeguatezza degli strumenti di analisi è stata di ostacolo alla diffusione di tale tipologia di marcatura metabolica. In questo tipo di studi, le tecniche di misura maggiormente impiegate si basano sulla spettrometria di massa: questa tecnica permette di identificare e distinguere i composti marcati dagli analoghi non marcati misurando il diverso valore di massa su carica che li caratterizza, in questo modo è possibile monitorare le variazioni nella concentrazione delle proteine marcate all'interno dei fluidi biologici in esame. L'avvento del *World Wide Web* negli anni Novanta del secolo scorso ha permesso ai ricercatori di condividere i risultati delle analisi, nonché sviluppare software per l'elaborazione dei dati adattabili alle esigenze di ciascun esperimento.

Nonostante si siano raggiunte alte prestazioni computazionali, la bioinformatica è un settore in rapido sviluppo che mira a migliorare e ampliare l'offerta di strumenti di analisi coinvolgendo un sempre più vasto gruppo di ricercatori.

Il perfezionamento della tecnica di marcatura metabolica con $^{2}H_2O$ e dei relativi strumenti di analisi ne permette il diffuso utilizzo in ambito medico-diagnostico soprattutto in caso di patologie difficilmente osservabili con le tradizionali tecniche come, per esempio, il morbo di Parkinson e la malattia di Alzheimer, per le quali non esistono ancora tecniche di diagnosi precoce affidabili. Grazie all'impiego del *labeling* metabolico negli studi sulle malattie neurodegenerative, è stato possibile mettere in relazione la comparsa di queste patologie e il comportamento anomalo delle proteine osservate nel fluido cerebrospinale.

In questa tesi si analizzano i metodi di *labeling* metabolico utilizzati per caratterizzare la cinetica delle proteine ed il loro utilizzo in ambito medico-diagnostico.

INTRODUZIONE

L'osservazione della dinamica delle proteine ha da sempre suscitato molto interesse nel campo biologico e medico allo scopo di comprendere la struttura e il comportamento dei tessuti biologici più approfonditamente e con il fine di scoprire l'esistenza di eventuali correlazioni tra l'insorgenza di patologie e le particolari modalità di aggregazione dei composti biologici in modo tale da poter sviluppare nuovi metodi diagnostici.

L'analisi della cinetica dei composti proteici è stata limitata a causa delle difficoltà incontrate dai ricercatori in fase di raccolta dati. Infatti, i prelievi tissutali necessari all'osservazione del proteoma di un certo organo, sono invasivi e costosi, oltre a non fornire alcuna indicazione utile a comprendere le variazioni dinamiche del sistema.

Introducendo marcatori metabolici nel soggetto vivente si possono osservare i cambiamenti che incorrono nel proteoma rendendo possibile ricavare informazioni utili riguardo la sua evoluzione nel tempo con la possibilità di osservare eventuali anomalie o disfunzioni.

Nelle prime fasi di questi studi, erano comunemente impiegati isotopi radioattivi, i quali nonostante assolvessero in pieno alle loro funzioni di marcatura, sono rischiosi per la salute umana e animale cosicché si è reso necessario lo sviluppo di nuove metodologie che potessero supportare queste analisi in modo più sicuro. L'impiego di molecole deuterate di acqua ha permesso di ampliare la gamma di marcatori metabolici costituendo una valida e più sicura alternativa ai composti utilizzati tradizionalmente.

Le analisi sul proteoma attraverso marcatori adatti sono supportate dal metodo conosciuto come biopsia virtuale, il quale consiste nel prelievo non invasivo di campioni di alcuni fluidi corporei, dal quale, a partire dalle diverse concentrazioni di proteine presenti, si possono dedurre le variazioni coinvolte nel metabolismo di una certa proteina.

A causa dell'inadeguatezza o della mancanza di strumenti di misura, il *labeling* metabolico per questa tipologia di studi non è stato sfruttato appieno se non in tempi recenti, quando la tecnologia ha raggiunto un livello di efficienza tale da poter condurre analisi più precise e accurate.

Grazie alla crescente disponibilità di *software* dedicati alle analisi biomediche, oggigiorno è possibile disporre di una vasta gamma di dati ottenuti dai ricercatori in tutto il mondo. Inizialmente, le limitazioni che gli scienziati incontravano nelle procedure di analisi dati erano da imputare, oltre alla mancanza di strumenti adatti, alla scarsità di informazioni disponibili per

condurre efficientemente le operazioni di confronto e validazione. La nascita nei primi anni Novanta del Novecento del *World Wide Web* ha supportato la diffusione di dati e di strumenti di analisi facilitando queste procedure. Tuttavia, spesso i dati raccolti non sono organizzati e catalogati, comportando notevoli ritardi nella conduzione di tali ricerche a cui si aggiungono le difficoltà legate alla scelta dello strumento di analisi adeguato in termini di efficienza computazionale e temporali. Nonostante siano stati compiuti notevoli progressi in questo ambito, il settore della bioinformatica è in continua evoluzione con il fine di sviluppare strumenti sempre più performanti.

Nell'ambito medico, la marcatura con acqua pesante ha aperto la strada alla scoperta di nuovi metodi diagnostici i quali permettono di eseguire le analisi in modo sempre più affidabile, permettendo di comprendere la progressione di patologie come le malattie neurodegenerative, per le quali non esistono ancora metodi ottimali e cure definitive.

1. IL *LABELING* METABOLICO CON ^2H_2O

Lo studio della cinetica delle proteine permesso dalla marcatura tramite acqua pesante venne descritto per la prima volta nel 1941 [1]. Da allora, questo metodo ha incontrato un notevole sviluppo per i suoi molteplici vantaggi e per la semplicità di utilizzo. Attraverso l'impiego di molecole di ^2H_2O utilizzate per effettuare marcature di polimeri biologici si sono condotti studi riguardanti la cinetica delle proteine nei tessuti di vari organi quali fegato, cuore, reni e muscoli, nei quali l'utilizzo di altre tipologie di marcature è ostacolato principalmente dalle notevoli difficoltà in fase di raccolta dei campioni, ma anche nel plasma, decisamente più accessibile per le analisi rispetto agli organi precedentemente citati.

Le motivazioni che hanno portato i ricercatori verso la scelta frequente di utilizzare ^2H_2O nei processi di *labeling* metabolico sono da ricondurre primariamente alle proprietà chimiche possedute, nonché alla versatilità della molecola.

1.1. Composizione, proprietà e utilizzo di ^2H_2O

Con il termine acqua pesante ci si riferisce alle molecole d'acqua in cui uno o entrambi gli atomi di idrogeno presenti nel composto sono stati sostituiti dall'isotopo dell'elemento secondo per abbondanza in natura: il deuterio ^2H.

Le molecole di ^2H_2O presentano due composizioni chimiche diverse: possono essere del tipo ossido di deuterio, rappresentate con il simbolo D_2O (dove D indica l'isotopo deuterio), oppure ossido di deuterio e prozio, rappresentate con il simbolo DHO. In entrambi i casi, sono caratterizzate dalla presenza di uno specifico isotopo dell'atomo di idrogeno, il deuterio ^2H.

L'acqua pesante, a cui talvolta ci si riferisce con il nome di acqua arricchita, termine che sarà chiarito in seguito, si trova normalmente in natura, tuttavia la sua diffusione è di gran lunga inferiore a quella dell'acqua comune. Si stima che in natura su 20 milioni di molecole d'acqua solamente 1 sia del tipo arricchito. [2]

1.1.1. Isotopi dell'idrogeno

L'idrogeno è un atomo che possiede un protone nel nucleo e un elettrone nel suo guscio elettronico: il suo numero atomico è pari a 1. In natura si incontrano tre isotopi dell'idrogeno:

- prozio ➔ è l'isotopo con maggiore diffusione e nel quale non sono contenuti neutroni;

- deuterio ➔ è l'isotopo secondo per diffusione e nel quale nucleo atomico è contenuto un neutrone;
- trizio ➔ è l'isotopo con diffusione in natura minore e nel suo nucleo risiedono due neutroni, è l'unico isotopo dell'idrogeno a presentare attività radioattiva.

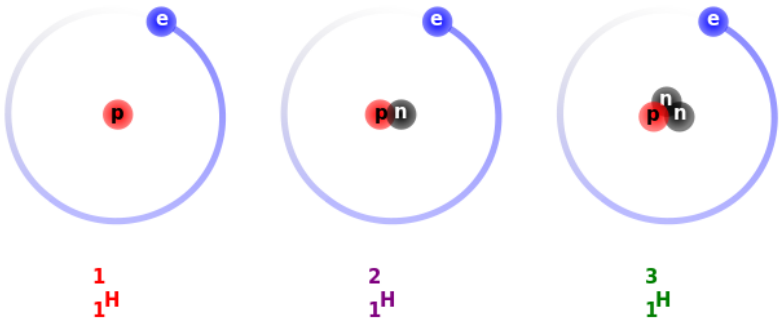

Figura 1. I tre isotopi dell'idrogeno, tratto da it.wikipedia.org.

Nelle molecole di acqua comune si trova il prozio. Il prozio e il deuterio sono isotopi stabili dell'idrogeno, ovvero non si osserva alcuna radioattività prodotta da questi atomi.

Il deuterio è stato scoperto nel 1931 dal chimico statunitense Harold Urey, il quale attribuì a questo atomo il nome correntemente usato. [3] Il termine deuterio deriva dal greco antico *deuteros,* ovvero secondo, a sottolineare il fatto che presenta una seconda particella subatomica nel suo nucleo, la quale presenza essenzialmente rende la massa del deuterio doppia in valore rispetto al prozio.

Dalla parola deuterio deriva il termine deuterazione, il quale indica il processo che si verifica quando l'atomo di idrogeno prozio presente in una molecola viene sostituito da un atomo di deuterio. Gli effetti sulla cinetica di questi composti sono definiti come il rapporto tra il tasso di reazione in cui i reagenti coinvolti differiscono unicamente per la loro composizione isotopica.

1.1.2. Caratteristiche chimiche delle molecole di acqua pesante

I composti di $^{2}H_2O$ hanno la medesima capacità di formare legami chimici con altri atomi, rispetto all'acqua comune (H_2O), tuttavia presentano chiare differenze nella stechiometria e nella massa molecolare: data la struttura atomica del deuterio, le molecole di $^{2}H_2O$ hanno una massa maggiore dovuta alla presenza di un neutrone nel nucleo del deuterio, grazie a tale caratteristica, si ricava il nome comunemente usato di acqua pesante o acqua arricchita.

Le molecole di $^{2}H_2O$, sebbene posseggano l'isotopo di deuterio, non presentano radioattività in quanto il deuterio è un isotopo stabile dell'idrogeno, questo permette un utilizzo sicuro, almeno in quantità contenute, dell'acqua pesante per lo studio dei processi metabolici negli animali così come negli umani.

Basandosi sulla sicurezza del composto, è possibile effettuare somministrazioni di acqua pesante ai soggetti in esame a una distanza temporale minore rispetto a quella valutata per le molecole radioattive. Non di meno, l'acqua arricchita permette il legame molecolare con tutti gli amminoacidi non essenziali, a differenza di altri composti chimici, i quali possono interagire esclusivamente con uno specifico amminoacido.

Di seguito sono riassunte alcune delle principali proprietà chimico fisiche dell'acqua pesante in relazione all'acqua comune:

	Acqua pesante	**Acqua comune**
Formula chimica	$^{2}H_2O$	H_2O
Massa molare	20,0276 g/mol	18,0153 g/mol
Punto di fusione	3,8°C	0°C
Punto di ebollizione	101,4°C	100°C
Indice di rifrazione	1,328	1,333

Tabella 1. tratto da [4].

1.1.3. Tossicità

Nonostante l'acqua arricchita sia sicura per gli animali e, nello specifico, per gli esseri umani, in grandi quantità presenta caratteristiche di nocività. La pericolosità dell'acqua pesante deriva dal diverso comportamento chimico del deuterio rispetto al prozio. Si stima che quando la presenza di acqua arricchita nel corpo umano si aggira intorno al 25-50% della massa totale si può incorrere in avvelenamento da acqua pesante. [2]

Nei mammiferi, il raggiungimento del livello di 25% di acqua pesante sostituita sul totale presente nel corpo causa sterilità, mentre il 50% è mortale. È necessario inoltre considerare che generalmente una certa quantità, sebbene ampiamente trascurabile, di acqua pesante è comunque presente nel corpo umano, la gran parte di questa viene assunta anche attraverso gli

alimenti presenti nella dieta dell'individuo e quindi non unicamente dall'acqua assunta direttamente, la quale di per sé ne contiene una certa parte. [5]

1.1.4. Utilizzi comuni delle molecole di ^2H_2O

L'acqua pesante trova il suo utilizzo più comune nei reattori nucleari, nei processi di moderazione della velocità dei neutroni generati nei processi di fissione nucleare. Tuttavia, il composto è largamente usato anche per lo studio dei processi biologici e nelle reazioni chimiche. L'ossido di deuterio è infatti utilizzato nella spettroscopia a risonanza magnetica nucleare in soluzione acquosa, procedura impiegata negli studi riguardo il nuclide dell'idrogeno. Inoltre, nella chimica organica si utilizza il deuterio principalmente come marcatore degli atomi di idrogeno, soprattutto per il tracciamento di reazioni chimiche che coinvolgono l'acqua, oppure al posto dell'acqua comune negli studi sulle proteine effettuati attraverso la spettroscopia a infrarossi. Le molecole di acqua pesante sono impiegate nelle reazioni nucleari indotte per la produzione di trizio e in aggiunta è impiegato nei processi di riconoscimento dei neutrini. Infine, trova largo impiego come marcatore negli studi metabolici negli umani e negli animali, aspetto su cui si ci si concentrerà in seguito.

1.2. Aspetti clinici

Date le caratteristiche e la composizione chimica di ^2H_2O la somministrazione nei soggetti è semplificata rispetto ad altre molecole. In particolare, è possibile somministrare come bolo una dose di acqua arricchita in soluzione acquosa e una volta raggiunto l'equilibrio procedere con la somministrazione di una certa quantità di acqua pesante tale da mantenere costante l'equilibrio per tutta la durata dell'esperimento. La procedura di marcatura è in questo modo resa più agile da effettuare, poiché non prevede una somministrazione per via endovenosa. Se così non fosse, sarebbe necessario il controllo costante di un medico che valuti le condizioni di salute del paziente sottoposto alla somministrazione e, al caso peggiore, si risolverebbe con un'ospedalizzazione, aspetto che è necessario ricordare se si considera il caso di altri marcatori metabolici.

Si noti che, sebbene la somministrazione di acqua arricchita non presenti alcun rischio per la salute umana o animale, è necessario limitarne l'assunzione a piccole quantità in quanto ^2H_2O è pur sempre un composto chimico che provoca un'alterazione nel sistema biologico. In ambito

medico, le soluzioni di acqua pesante sono caratterizzate da concentrazioni uguali o inferiori al 5%. Concentrazioni superiori presentano un'elevata tossicità per gli animali, come considerato in precedenza (si veda par. 1.1.3). [6] [7]

Limitatamente ai livelli di assunzione di acqua pesante ritenuti sicuri per la salute umana si è osservato che un trentesimo dei soggetti soffre come effetto collaterale della comparsa di vertigini transitorie durante la fase iniziale della somministrazione. Probabilmente, questo fenomeno si deve ricondurre al fatto che, durante la somministrazione di acqua arricchita, si genera una variazione nel flusso della massa d'acqua che si verifica all'interno delle cavità ossee situate nell'orecchio interno, il quale è interpretato a livello neurologico allo stesso modo del caso in cui si effettui una qualsiasi attività motoria, generando la sensazione di perdita di equilibrio e capogiro. Questo effetto collaterale può essere evitato limitando la quantità di acqua pesante somministrata per ogni singola dose, in modo tale da non generare una rapida variazione nei livelli di $^{2}H_2O$. [8]

1.3. Metodo di raccolta dei campioni e considerazioni sperimentali

Il *labeling* metabolico basa il suo funzionamento sulla proteostasi, fusione delle parole proteina e omeostasi, ossia al concetto di equilibrio dinamico delle proteine: il processo di generazione e degradazione delle proteine in un soggetto può subire delle alterazioni durante la sua vita che portano all'instaurazione di un nuovo equilibrio. Questo processo avviene in caso di crescita fisica o differenziazione cellulare ma è osservabile anche in relazione all'insorgenza di disfunzioni.

La marcatura delle proteine permette l'osservazione del processo di generazione e degradazione di questi composti in modo da poter analizzare eventuali anomalie riconducibili a patologie cliniche dell'individuo oggetto di studio. [9]

Per stimare il livello di arricchimento raggiunto in un soggetto si analizzano campioni di fluidi corporei, come sangue, saliva e fluido cerebrospinale. Tra tutti, la saliva risulta essere il fluido corporeo ottimale per il prelievo in fase di ricerca di quelle proteine che forniscono informazioni riguardo la presenza e l'eventuale decorso di patologie cliniche, anche se talvolta si preferiscono altri fluidi in cui la concentrazione di proteine è più alta. Infatti, la saliva è il fluido biologico che permette la raccolta dei campioni in modo meno invasivo rispetto, per esempio, al prelievo di sangue o di fluido cerebrospinale. Questo criterio ha appunto messo in luce proteine caratterizzate da un'espressione genica insolita e di potenziali indicatori biologici in pazienti

che presentavano patologie come schizofrenia, disordine bipolare, sclerosi multipla e stadi iniziali di Alzheimer. [10]

La pratica che prevede la raccolta e l'analisi dei fluidi corporei in ambito di *labeling* metabolico è denominata biopsia virtuale e viene condotta allo scopo di minimizzare l'invasività che caratterizza questa procedura. In alternativa alla biopsia virtuale è possibile osservare le variazioni delle concentrazioni di proteine marcate solamente effettuando biopsie dei tessuti, le quali non possono essere condotte nel caso di attività di *screening* oppure in ambito di studi neurologici. Questo metodo consiste nel misurare la quantità di proteine presenti in un certo fluido le quali, dopo essere state prodotte all'interno del tessuto oggetto di studio, vengono rilasciate nel fluido in questione, si consideri come esempio il sangue.

1.3.1. Il *labeling* metabolico delle proteine

L'interesse degli scienziati si rivolge principalmente verso il comportamento delle proteine, sebbene sia possibile procede con la marcatura di altri composti biologici, in quanto sono responsabili della maggior parte dei processi catabolici e strutturali e più in generale presentano un'elevata versatilità nella maggioranza dei processi biologici.

Le precedenti tecniche utilizzate per lo studio del comportamento delle proteine non fornivano alcuna spiegazione per quanto concerne la dinamica dei polimeri, ovvero la nascita e la rigenerazione di questi. Infatti, le tecniche di *labeling* metabolico precedenti consideravano unicamente la distribuzione e le caratteristiche delle macromolecole al momento della raccolta dei campioni, fotografando un solo istante temporale.

Questo approccio sperimentale, sebbene spieghi ampiamente le caratteristiche strutturali dei tessuti e in quale modo i macropolimeri ne siano coinvolti, non fornisce alcuna interpretazione sulla loro evoluzione e sul loro sviluppo, poiché queste sono ricavabili solo osservando l'evoluzione dinamica delle concentrazioni e delle distribuzioni.

Per lo studio della dinamica dei flussi delle proteine è necessario porre l'attenzione sulle molecole marcate con isotopi in quanto l'utilizzo di questa strategia permette di osservare le variazioni delle loro concentrazioni relative a livello temporale, permettendo di esaminare l'evoluzione dei fenomeni in cui sono coinvolti i polimeri.

Nessuna caratteristica osservabile in modo statico a nessun livello può dare una spiegazione per quanto riguarda l'evoluzione di una malattia la quale è possibile solo attraverso l'osservazione della dinamica del sistema e delle variazioni che incorre.

L'importanza dell'utilizzo del labeling metabolico si basa per l'appunto su questo principio: marcando le molecole di interesse con isotopi specifici si genera un'asimmetria nell'organismo, il quale presenta allo stesso tempo e nello stesso spazio dei polimeri caratterizzati dal possedere il marcatore e altri che non presentano questa peculiarità, dando origine a un sistema che evolve in funzione del tempo. Infatti, all'aumentare del tempo trascorso dalla somministrazione del marcatore, un numero sempre maggiore di composti presenterà il marcatore, fino a raggiungere il momento in cui circa la totalità delle proteine sarà marcata.

1.3.2. Studio *in vivo* del comportamento delle proteine

Per analizzare il comportamento dei composti proteici è necessaria una biopsia dei tessuti oggetto dell'analisi. Nel caso di cavie, sebbene sia possibile un prelievo di piccoli campioni di sangue, per l'analisi di più tessuti è necessario sopprimere l'esemplare per poter procedere allo studio. Nell'uomo, l'analisi prevede una biopsia, ovvero un prelievo chirurgico dei tessuti, pratica oltremodo invasiva.

Seguendo questo approccio sperimentale l'analisi presenta caratteristiche di irriproducibilità: molto semplicemente, nel caso di una cavia soppressa non è possibile effettuare prelievi successivi sullo stesso soggetto, d'altra parte, nell'uomo, è sconsigliabile effettuare molteplici biopsie sullo stesso tessuto, specialmente per quanto riguarda il tessuto muscolare, in quanto la procedura risulta essere invasiva. [6]

D'altra parte, non bisogna dimenticare l'elevato costo caratterizzante una simile procedura di analisi, la quale prevede l'utilizzo di strumenti chirurgici e cavie, nonché di effettuare vere e proprie operazioni chirurgiche, in quanto una biopsia si effettua con le stesse procedure cliniche di qualsiasi altra operazione.

Grazie allo sviluppo di strumenti più precisi e affidabili, è possibile e consigliabile optare per un'analisi *in vivo* del comportamento delle proteine con vantaggi evidenti, tuttavia bisogna considerare l'esistenza di errori nei risultati degli studi riconducibili alle fluttuazioni statistiche derivati dall'imprecisione e dall'inaccuratezza degli strumenti di misura.

Grazie all'utilizzo dei processi di marcatura metabolica, è possibile osservare la dinamica delle proteine attraverso la loro presenza nei fluidi, osservando le variazioni nella concentrazione di un certo tipo di polimero si può comprendere l'evoluzione che ha incontrato all'interno di un certo tessuto.

1.3.3. Raggiungimento del *plateau*

Per alcune molecole utilizzate nel *labeling* il raggiungimento dell'equilibrio, o *plateau*, richiede un tempo decisamente più lungo rispetto ai valori osservati nel caso ivi considerato e questo comporta inevitabilmente un dispendio in termini temporali di risorse per attuare l'indagine. Nel caso dell'acqua pesante, poiché le sue caratteristiche non sono dissimili a quelle della comune acqua, essa raggiunge velocemente ogni parte del corpo del soggetto in esame accelerando la procedura. [9]

1.3.4. Criticità dell'utilizzo di *labeling* con ^2H_2O e dell'analisi *in vivo*

Sebbene a livello globale, la marcatura ottenuta grazie alla somministrazione di molecole di acqua pesante abbia molteplici vantaggi, le maggiori difficoltà riscontrate dagli scienziati sono principalmente imputabili a due aspetti fondamentali legati alla fase di analisi dei dati raccolti.

In primo luogo, a causa della bassa concentrazione di deuterio somministrabile per ragioni sanitarie, si possono raggiungere livelli di marcatura delle proteine solo parziali. Questo limite è legato al fatto che solo una piccola parte degli atomi di prozio presenti nelle macromolecole vengono sostituiti dagli atomi di deuterio assunti. Ciò implica una scarsa presenza di ^2H, rendendo la stima della quantità di molecole marcate osservate nei fluidi corporei più complessa. [7]

In aggiunta all'incompleta marcatura, un ulteriore aspetto che rende complesso l'utilizzo del *labeling* metabolico con acqua pesante è da imputare alla mancanza o all'inadeguatezza di strumenti informatici idonei alle caratteristiche del metodo in analisi, le quali necessitano di alta affidabilità e accuratezza, proprietà che non tutti i *software* utilizzati nell'ambito proteomico possiedono. [11]

1.4. Utilizzo di isotopi radioattivi per i processi di marcatura metabolica

La necessità e l'importanza del *labeling* impiegante isotopi stabili è emersa gradualmente dopo un lungo periodo di utilizzo esclusivo di marcatori isotopici radioattivi. La maggioranza degli studi che impiegavano composti isotopicamente stabili veniva effettuata principalmente *in vitro* impiegando reagenti complessi e talvolta costosi. Inoltre, l'analisi *in vitro* necessita di frammentare le proteine derivate dal campione esaminato e così facendo non è possibile intraprendere qualsiasi tipo di analisi sul metabolismo e sulla cinetica delle proteine, il quale prevede necessariamente che i composti siano intatti e che l'incorporazione del marcatore avvenga in seguito alla somministrazione *in vivo*.

Considerando gli isotopi radioattivi utilizzati per il *labeling* è importante notare come il grado di radioattività che li caratterizza possa raggiungere valori molti alti, questo aspetto è sicuramente vantaggioso perlomeno per quanto riguarda la quantità di sostanza somministrabile: infatti, date queste caratteristiche, anche livelli molto bassi di incorporazione dell'isotopo sono facilmente rintracciabili, si prenda come esempio l'isotopo ^{35}S per cui si osserva questo comportamento, il quale permette l'assunzione di basse quantità di marcatore e breve tempo di esposizione e presenta valori di radioattività pari a circa 40 TBq/mmol. Tuttavia, con gli strumenti utilizzati comunemente per effettuare analisi spettrometriche, bassi livelli di marcatura non sono rintracciabili a causa dell'alto valore di segnale-rumore che presentano i dati, dovuto alle interazioni chimiche che si verificano spontaneamente nel soggetto e agli strumenti di misura stessi. Infatti, affinché il marcatore sia rilevato dalla strumentazione, la sua concentrazione deve essere necessariamente pari ad almeno il 5-10% del totale per poter ricavare valori significativi. [12]

2. TECNICHE DI MISURA

Il *labeling* metabolico attraverso la somministrazione di molecole di ^2H_2O è di fondamentale importanza nelle analisi riguardanti la cinetica delle proteine, in quanto è un approccio estremamente potente che permette di condurre questo tipo di studi in modo più semplice, tuttavia ciò su cui si basa e che rende possibile applicare un simile metodo è sostanzialmente la strumentazione a cui seguono dei chiari protocolli impiegati per condurre le successive analisi.

Grazie al metodo spettrometrico, necessario per stimare l'abbondanza di massa di isotopomeri marcati, si possono analizzare allo stesso tempo i flussi e l'andamento temporale della concentrazione di tali polimeri e la variazione dei macro polimeri analoghi, i quali non presentano marcatura, rendendo in questo modo possibile l'approccio *in vivo* per lo studio dei macropolimeri. [8]

2.1. Spettrometria di massa (MS)

La spettrometria di massa, denotata per semplicità con la sigla MS, dall'Inglese *Mass Spectrometry*, è una tecnica molto diffusa in ambito sperimentale, in particolare nell'ambito delle analisi chimiche. Consente l'identificazione della tipologia e della quantità di composti presenti in un campione portato in fase gassosa, in modo tale da poter misurare il rapporto massa su carica m/z degli ioni. Lo strumento utilizzato per questa selezione è conosciuto come spettrometro di massa, da cui deriva il nome della tecnica di analisi.

Figura 2. Spettrometro di massa Agilent 8890/5977B con sistema GC/MSD e vassoio aggiuntivo, tratto da agilent.com.

Si definisce come spettro di massa il grafico in cui viene rappresentato il segnale rilevato dallo spettrometro in funzione del rapporto m/z. Dall'analisi degli spettri è possibile identificare la composizione e la distribuzione isotopica di un certo campione, cosicché si possa definire in modo non ambiguo la struttura delle molecole presenti in quest'ultimo.

Nei primi stadi di sviluppo e impiego di questa tecnologia, sebbene fosse chiaro fin dal principio la sua rilevanza e i suoi possibili utilizzi, si sono riscontrate alcune limitazioni, dovute principalmente alla mancanza di strumenti tecnologicamente adeguati alle esigenze dei ricercatori, perciò, agli esordi, tale procedimento veniva applicato in modo efficace esclusivamente a campioni contenenti una quantità limitata di proteine isolate in specifici ambiti funzionali. Non di meno, la scarsità o addirittura la totale mancanza di dati riguardanti il proteoma non permettevano l'identificazione di tutti quei composti i quali fino a quel momento non erano mai stati considerati, comportando notevoli limitazioni nell'applicazione di questa procedura di analisi. [13]

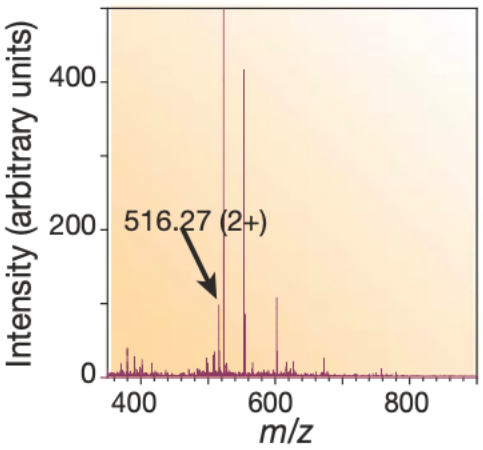

Figura 3. Esempio dello spettro osservato in un generico esperimento di proteomica basato sulla spettrometria di massa, riadattato da [13].

Per quanto riguarda i composti marcati, ciò che rende possibile la separazione dai loro analoghi non marcati è solamente la differente massa che li caratterizza. Infatti, poiché l'atomo di deuterio possiede un neutrone aggiuntivo nel nucleo, questo aggiunge una certa quantità di massa alla molecola di acqua arricchita, permettendo così l'identificazione univoca dei composti marcati (si veda par. 1.1.1).

2.1.1. La strumentazione e il suo funzionamento

Per definizione, uno spettrometro di massa è uno strumento composto da una sorgente di ioni, un analizzatore di massa che misura il valore del rapporto massa su carica m/z degli ioni emessi dalla sorgente e un rilevatore che registra il numero delle particelle che possiedono un dato valore. [13]

La procedura che permette la preparazione dei campioni per l'analisi è composta da una sequenza di fasi ben definite.

Per prima cosa, il composto subisce il processo di ionizzazione, purtroppo però in questa fase si può incorrere facilmente nella frammentazione dei costituenti. Tuttavia, il processo di ionizzazione avviene esclusivamente nel caso in cui il campione non si trovi già nello stato gassoso.

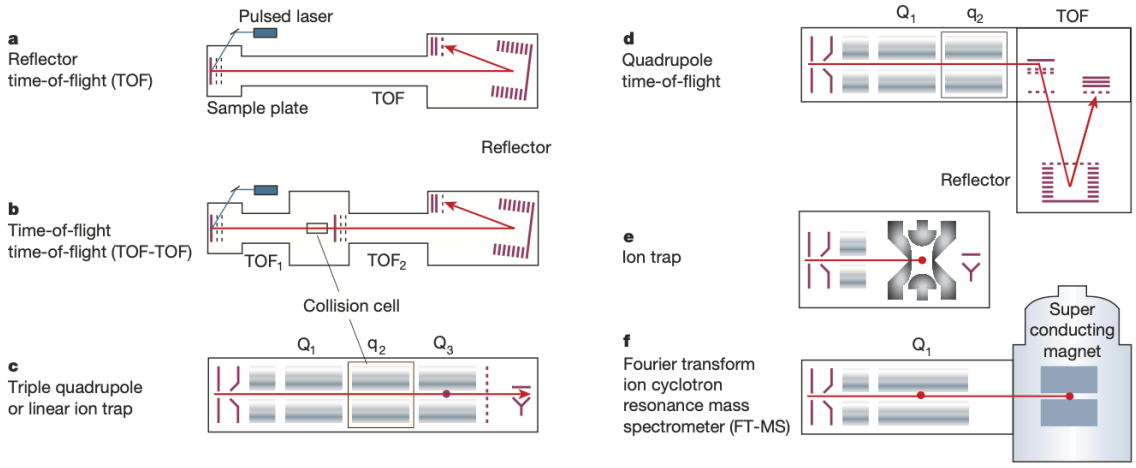

Figura 4. Rappresentazione delle diverse configurazioni degli spettrometri di massa utilizzati negli studi di proteomica, tratto da [13].

In seguito, gli ioni vengono distinti in base al rapporto m/z posseduto. Il dispositivo utilizzato in questo stadio è posto in cascata allo ionizzatore. Questa fase è indispensabile, in quanto successivamente permette di identificare e quantificare gli elementi costituenti del composto analizzato attraverso il processo di *matching*.

Dopo la fase di separazione, gli ioni vengono identificati attraverso un meccanismo adatto a questo scopo, come per esempio, il moltiplicatore di elettroni, dispositivo che ha lo scopo di amplificare il debole fascio di elettroni prodotto dalla sorgente, provocando emissioni multiple di elettroni in cascata ottenuti facendo incidere di volta in volta il fascio principale su una lastra metallica.

A questo punto i risultati ricavati sono graficati come spettro delle abbondanze relative in funzione del rapporto m/z da cui sono caratterizzati.

La fase finale è fondamentale, infatti permette di riconoscere la composizione del campione in analisi attraverso il *matching* dei risultati ricavati dall'analisi dello spettro con i dati presenti nelle banche dati. [14] Considerando lo studio condotto da Holmes e colleghi [8], i dati ricavati dalle analisi spettrometriche sono stati analizzati avvalendosi del *database* UniProt/SwissProt e del *software* Agilent Spectrum Mill Proteomics Workbench adottando i seguenti criteri:

- i parametri per l'estrazione dei dati si ricavano considerando la stessa massa del precursore ottenuta dalla somiglianza degli spettri entro la tolleranza di ±10 s per il tempo di ritenzione, definito come l'istante in cui lo spettrometro registra la fuoriuscita del peptide dalla colonna liquido cromatografica in un sistema LC/MS, e ±1.4 m/z per la massa, minimo valore del rapporto segnale rumore ottenuto nel primo stadio

dell'analisi spettrometrica per il precursore pari a $S/N = 10$ e valore di m/z del precursore di ^{12}C assegnato durante la fase di estrazione;

- per condurre la ricerca dei parametri, il numero massimo di picchi mancanti è fissato a 2, l'intensità minima di somiglianza del picco pari al 30%, la tolleranza sulla massa dei precursori pari a 10 ppm, tolleranza sulla massa dei prodotti pari a 30 ppm, il numero minimo di picchi identificati pari a 4 e la carica massima del precursore pari a 3;
- i risultati sono considerati validi a livello di proteine e peptidi con un tasso globale di identificazione dei valori errati pari all'1%;
- infine, il *software* fornisce una lista di peptidi con punteggio maggiore di 6 e intensità dei picchi maggiore al 50% per le proteine che presentano un punteggio superiore a 11.

Lo sviluppo tecnologico che si è verificato in questo ambito ha oltretutto reso possibile l'introduzione di nuove tecniche che sono state sviluppate in tempi relativamente recenti, come, per esempio, l'impiego della tecnica di ionizzazione leggera, la quale non causa la frammentazione del campione, che altrimenti sarebbe reso inutilizzabile per l'applicazione dell'approccio di analisi MIDA, descritto in seguito.

Se le analisi effettuate sul campione hanno lo scopo di esaminare la distribuzione di proteine e peptidi, le tradizionali tecniche di ionizzazione risultano inadatte, in quanto danneggiano il campione, impedendo le analisi ulteriori. Peptidi e proteine sono composti polari, non volatili e termicamente instabili perciò si richiede l'utilizzo di tecniche di ionizzazione adatte a non compromettere questi composti, ovvero tali per cui questi siano portati in fase gassosa senza subire una degradazione eccessiva. Le tecniche di nuovo impiego che assicurano queste necessità sono denominate MALDI (acronimo di *Mass-Assisted Laser Desorption Ionization*) ed ESI (acronimo di *Electrospray Ionization*):

MALDI: la matrice utilizzata in questa tecnica assorbe l'energia rilasciata dai laser, la quale viene trasferita al campione acidificato, in questo modo il rapido riscaldamento causa la degradazione della matrice e vengono rilasciati in fase gassosa ioni provenienti dal campione. Affinché questo avvenga, è necessario l'impiego di qualche centinaio di impulsi laser per ottenere un rapporto segnale-rumore accettabile. [15]

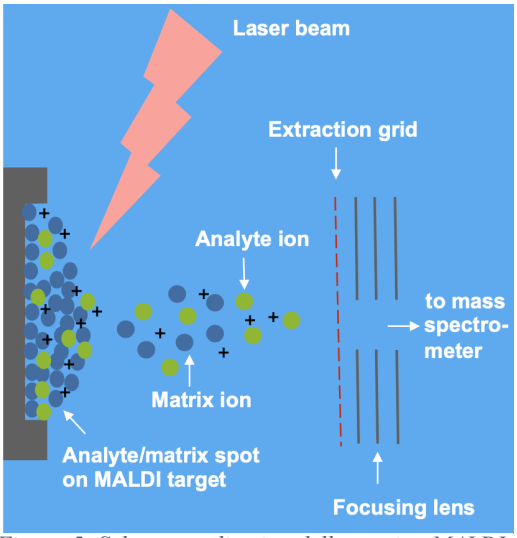

Figura 5. Schema applicativo della tecnica MALDI: un impulso laser irradia il campione, il quale rilascia ioni in fase gassosa che in seguito vengono convogliati nello spettrometro di massa, tratto da [14].

ESI: la produzione di ioni attraverso questa tecnica avviene a partire dal campione in soluzione. La ionizzazione attraverso vaporizzazione elettrica è veicolata da un alto voltaggio (dai 2 ai 6 kV) applicato tra l'emettitore nella parte terminale della catena e l'ingresso dello spettrometro di massa. Questo processo implica la creazione di un composto nebulizzato elettricamente carico, al quale segue la formazione e la successiva dissoluzione in goccioline, restituendo in questo modo una miscela formata dal campione in soluzione. [15]

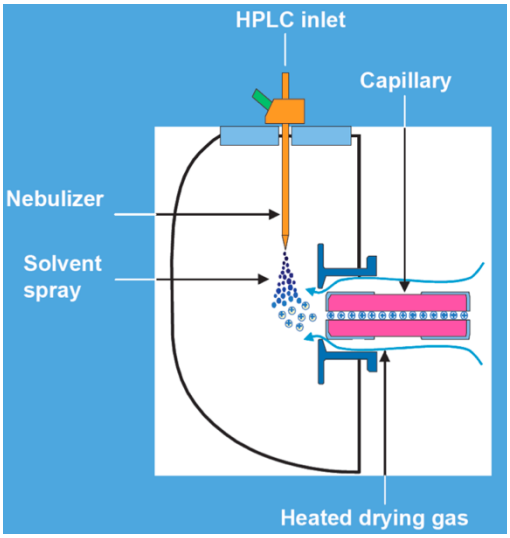

Figura 6. Schema applicativo della tecnica ESI: il campione nebulizzato viene introdotto in un contenitore che si trova alla pressione atmosferica e in cui persiste un forte campo magnetico, in seguito condotto all'interno di un capillare e quindi nello spettrometro di massa, tratto da [14].

Di fondamentale importanza rimane la decisione riguardante quali siano le proteine di maggiore interesse presenti nel campione considerato, poiché un numero eccessivamente elevato di elementi da analizzare aumenta la probabilità di essere in presenza di dati contaminati. [16]

2.1.2. Procedure di analisi in presenza di composti proteici

Per le analisi spettrometriche in cui sono coinvolte le proteine, generalmente si adottano due diverse strategie di analisi: strategia *bottom up* e strategia *top down*. Nel primo caso, l'analisi proteomica comincia con la preparazione del campione attraverso la digestione enzimatica delle proteine ivi presenti, nel secondo caso, i composti vengono analizzati intatti.

L'approccio *bottom up* è il procedimento tradizionalmente più usato e diffuso soprattutto nel caso in cui i campioni oggetto di studio presentino un'elevata complessità oppure per tutte quelle applicazioni le quali coinvolgono l'analisi proteomica su larga scala. Come già menzionato, il primo passo per l'applicazione di questo processo prevede la digestione attraverso l'uso di enzimi specifici delle proteine oggetto di interesse, in modo tale da ridurre il composto in polipeptidi che saranno analizzati allo scopo di identificare i composti da cui derivano. [15]

Per poter attuare questa strategia, la lunghezza ottima dei peptidi da analizzare è da considerarsi nell'intervallo che va tra i 6 e i 50 amminoacidi costituenti, perciò è spesso scelto, come digerente enzimatico delle proteine, l'enzima tripsina, il quale genera peptidi che presentano

una lunghezza media di 14 amminoacidi. La procedura di riduzione in frammenti costituiti da pochi peptidi offre molteplici vantaggi, che includono la separazione efficiente dei composti in modo tale da osservare un numero limitato di cariche per ogni peptide oltre a un aumento nell'omogeneità del campione analizzato: tutti questi aspetti costituiscono un beneficio per le fasi di distinzione e identificazione successive. [10]

La strategia *top down* impiega le masse dei composti intatti allo scopo di identificarli successivamente. Alcuni aspetti rilevanti che derivano dall'impiego di questo criterio includono la vasta copertura di proteine rilevate e la più efficace caratterizzazione delle alterazioni posttraslazionali. Utilizzando questo metodo, è possibile ridurre l'ambiguità inerente all'identificazione dei composti proteici la quale, nel caso *bottom up*, avviene a partire dai singoli peptidi, permettendo in aggiunta il riconoscimento di isomeri. Non di meno, esso presenta affidabilità superiore poiché la quantità delle proteine presenti è misurata direttamente, mentre nel caso *bottom up* avviene impiegando le relative abbondanze peptidiche, le quali introducono inevitabilmente ulteriori fonti di errore. [15]

Questa metodologia ha incontrato una crescente popolarità grazie principalmente al perfezionamento tecnologico degli strumenti impiegati nella spettrometria ad alta risoluzione, all'aumentata efficienza del metodo di separazione con cromatografia liquida e ai progressi riguardanti i *software* di analisi.

Tra i vantaggi del metodo *top down* si rileva la possibilità di studiare il composto proteico nella sua totalità. Questo aspetto permette di analizzare la struttura del proteoma, la sua cinetica e il ruolo che occupa nei processi biologici. In questo modo, è possibile identificare alcune patologie, attraverso l'osservazione delle variazioni avvenute nella struttura del proteoma: infatti, questa proprietà rende possibile l'identificazione di alcune particolari proteine, le quali permettono di comprendere se ci si trovi in presenza di patologie per mezzo dell'osservazione delle alterazioni nella loro cinetica.

Per condurre analisi spettrometriche attraverso la strategia *top down* esistono fondamentalmente tre procedure: quantificazioni senza marcatura, *labeling* metabolico o chimico.

Quantificazione senza marcatura: questa procedura valuta il proteoma effettuando un confronto diretto di analisi liquido cromatografiche in combinazione con la spettrometria di massa eseguite in successione.

L'applicazione di questo metodo, come suggerisce il nome, non impiega alcun marcatore isotopico o di massa. Tali caratteristiche lo rendono più facilmente impiegabile ed economicamente più accessibile da implementare in laboratorio. Tuttavia la risposta degli strumenti di misura impiegati per la quantificazione delle proteine intatte può essere complicata dalla presenza di ioni nel campione nebulizzato i quali possiedono diversa carica e diversi valori di m/z , ma coesistono in un singolo proteoma. Le masse ricavate attraverso l'applicazione di questo metodo sono in seguito comparate con le masse delle proteine raccolte nei *database* considerando una certa tolleranza nella stima. Non bisogna dimenticare il fatto che in ogni singola analisi, l'errore dovuto allo strumento di misura ha un notevole impatto sul processo comparativo effettuato a partire dall'osservazione dei picchi, aspetto che grava fortemente sul successo dell'analisi dei dati raccolti e sulla loro validità, nonché sulla riproducibilità dell'esperimento, comportando una limitazione sul numero di osservazioni eseguibili.

Labeling metabolico: il *labeling* metabolico può essere effettuato applicando tre possibili strategie: SILAC (dall'Inglese, *Stable Isotope Labeling by Amino acids in Cell culture*), *Neutron Encoding* SILAC e infine TIPMI (acronimo di *Tunable Intact Protein Mass Increases*).

Il criterio SILAC prevede che si effettui una coltura cellulare in un mezzo contenente amminoacidi isotopicamente marcati in modo tale da rendere possibile l'espressione di proteine con una massa superiore rispetto alle proteine che si originerebbero in assenza di marcatori, chiamate appunto "pesanti". Il campione marcato viene in seguito combinato con un campione che è cresciuto in un mezzo ordinario, non marcato, tale così da generare proteine "leggere" e rendendo possibile l'applicazione successiva di analisi comparative. Uno svantaggio nell'applicazione del metodo SILAC risiede nella mancata possibilità di effettuare analisi multiple, in quanto in genere è possibile procedere al confronto di solamente due popolazioni. Nel caso della procedura modificata *Neutron Encoding* SILAC si introduce la teoria su cui si fonda il metodo di marcatura isobarica, il quale permette l'analisi di sistemi con elevata complessità: le proteine vengono generate in una coltura in cui sono presenti isotopologhi differenti dello stesso amminoacido tutti dotati di una certa massa nota. L'ultima strategia, TIPMI, differisce da entrambe le versioni di SILAC in quanto il marcatore viene somministrato per via alimentare al soggetto di cui si intende osservare il proteoma, come per esempio attraverso acqua deuterata oppure zucchero contenente isotopi di ^{13}C, il principale ostacolo nell'applicazione di questa procedura è la scarsa marcatura metabolica che ne risulta.

Come già menzionato, l'applicazione della strategia *top down* ha trovato recentemente larga applicazione soprattutto nell'ambito della quantificazione dell'espressione del proteoma in presenza di patologie mediche, come risposta biologica a stimoli esterni, per gli studi

riguardanti l'invecchiamento e, in generale, nell'ambito microbiologico. Grazie all'applicazione di questo metodo, si possono identificare le proteine specifiche che forniscono informazioni riguardo l'avanzamento di una malattia, in questo ambito sono stati condotti studi sul Parkinson e sulle malattie autoimmuni oltre ad altre applicazioni, come verrà in seguito chiarito. Inoltre, l'applicazione di questo metodo negli studi microbiologici, ha reso possibile ricavare informazioni riguardo lo sviluppo e il meccanismo di infezione di patogeni, le quali forniscono delle soluzioni applicabili in sede di trattamento terapeutico. [10]

2.1.3. Identificazione delle distribuzioni relative

Alla base dello studio del comportamento dinamico delle proteine si trova la necessità di quantificare la distribuzione delle masse degli isotopomeri che permette di valutare la frazione relativa di sintesi di ciascuna proteina singolarmente in modo tale da poter seguire le variazioni dell'abbondanza relativa dei peptidi.

Di seguito si elencano le grandezze fondamentali per la descrizione dei processi combinatori:

$p = valore\ dei\ precursori\ arricchiti\ (generalmente\ espresso\ in\ forma\ percentuale)$

$n = numero\ dei\ siti\ attivi\ di\ legame\ carbonio\ idrogeno\ in\ ciascun\ peptide$

$\%M_0 = \frac{M_0}{M_0 + M_1 + M_2 + M_3} * 100$ dove il valore $\%M_{0\ (iniziale)}$ è calcolato utilizzando il metodo MIDA (si veda par. 2.2)

$EM_{0t} = \%M_{0t\ (misurata)} - \%M_{0\ (iniziale)}\ calcolata\ al\ tempo\ t$

$EM_0^* = valore\ asintotico\ di\ EM_0 =$
$massimo\ valore\ assunto\ da\ EM_0\ in\ relazione\ a\ una\ certa\ p$

$f = quantità\ sintetizzata\ al\ tempo\ t = \frac{EM_{0t}}{EM_0^*}$

Inoltre, è possibile ricavare il valore di FSR, ovvero il tasso di sintesi frazionale delle singole proteine durante il periodo di marcatura, che è pari al valore di k ottenuto dalla relazione $f = 1 - e^{-kt}$ dove t si riferisce ai giorni di somministrazione.

2.2. *Mass Isotopomer Distribution Analysis* (MIDA)

L'acronimo MIDA identifica la tecnica di analisi basata sulla distribuzione delle masse degli isotopomeri presenti in un certo campione. Questa tecnica si basa sullo studio delle probabilità combinatorie e sulla distribuzione dei marcatori nei polimeri in modo tale da poter ricavare un'equazione di base utile alla comprensione e alla descrizione della biosintesi dei polimeri.

La sintesi di nuovi polimeri può essere descritta attraverso un processo combinatorio, alla base del quale si considerano due gruppi distinti di subunità monomeriche caratterizzate dal possedere una composizione chimica differente: i monomeri che contengono l'isotopo marcatore e i monomeri non marcati. In queste condizioni ciò che si osserva è una popolazione di polimeri che non presenta una composizione isotopicamente uniforme e ogni polimero generato è caratterizzato dal possedere un certo numero, variabile in ciascuna macromolecola, di subunità marcate.

La probabilità posseduta da ciascun polimero di possedere un certo numero di isotopi è descritta dalla distribuzione che si ricava a partire da un'espansione binomiale la quale necessita di due variabili descrittive: il numero n di subunità presenti e la probabilità p di ciascuna subunità di presentare certe caratteristiche, ovvero il possedere o meno un certo numero di monomeri marcati.

L'unico parametro che deve essere stimato dal campione raccolto è la probabilità p, la quale viene ricavata a partire dalla distribuzione di massa degli isotopomeri, indipendentemente dalla quantità di molecole di nuova sintesi presenti nel campione e per un certo valore fissato di n. Il dato n possiede un valore costante e noto a priori per ciascun polimero, perciò, una volta stimato il valore per il composto esaminato, è possibile utilizzarlo anche nelle analisi successive. Si può osservare in Figura 7 la relazione che esiste tra i valori di p, f e $|EM_0|$ per il peptide VLEDLRSGLF, con $m = 1147$ e $n = 17$.

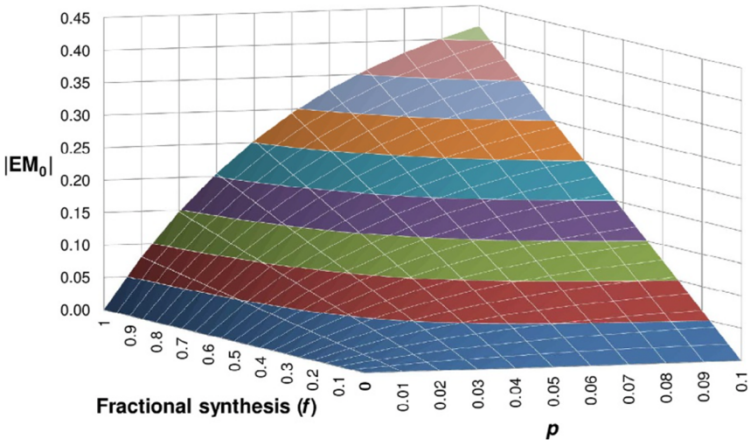

Figura 7. Relazione tra i valori di p, f e EM_0 assunti dal peptide VLEDLRSGLF, tratto da [8].

Con queste premesse, è possibile ricavare informazioni rilevanti sulle caratteristiche dei precursori delle molecole frutto di nuova sintesi, le quali sarebbero altrimenti impossibili da

ricavare in quanto è impossibile distinguere un polimero frutto di una sintesi recente da un polimero preesistente. [8]

2.2.1. Il parametro *n*

Con il simbolo *n* si denota il numero di siti di scambio degli isotopi nelle macromolecole, ovvero la somma dei siti di scambio in ciascun amminoacido che costituisce un peptide.

Questo valore è stato stimato da Commerford effettuando uno studio sui ratti e confermato in seguito da ricerche successive. [8] La stima del valore di *n* può essere effettuata seguendo due approcci sperimentali diversi:

- la misurazione diretta della presenza di deuterio negli amminoacidi liberi servendosi del metodo GC/MS (*Gas Chromatography – Mass Spectrometry)*;
- l'utilizzo dell'approccio MIDA che fornisce il valore finale di *n* osservato nei peptidi a seguito della marcatura con acqua pesante.

Tuttavia, è possibile impiegare il valore di *p* calcolato sperimentalmente da cui si può stimare il valore di *n* quando la presenza di isotopi marcati ha raggiunto la quantità sufficiente tale da poter soddisfare i criteri di analisi in un numero di isotopomeri superiore all'unità. Riassumendo è possibile stimare il valore di *n* adottando l'approccio sperimentale che si basa sull'analisi combinatoria. In modo analogo, è possibile ricavare il valore di *f* per ogni isotopomero come il rapporto, costante, dato da $\frac{EM_3^*}{EM_2^*} = \frac{EM_3}{EM_2} = costante$, da cui si ricava facilmente $\frac{EM_2}{EM_2^0} = \frac{EM_3}{EM_3^0}$.

Nel momento in cui si procede con questa strategia di analisi, lo scienziato può decidere se procedere impiegando i valori di *n* precedentemente raccolti oppure se condurre la stima *ex novo*.

Nel caso delle proteine, se si osservano valori non coerenti di *n* misurati su tessuti diversi, si può dedurre che le macro molecole, indipendentemente dal sito di produzione, il quale può essere il tessuto in analisi o un altro qualsiasi, vengono trasportate da e verso altri siti.

2.2.2. Impiego dell'analisi combinatoria nell'ambito della sintesi proteica

L'approccio che prevede l'impiego dell'analisi combinatoria nello studio sulla dinamica delle proteine si fonda sul fatto che è possibile ricavare tutte le informazioni riguardo le

caratteristiche isotopiche dei precursori biologici nella sintesi proteica per ogni tipo di polimero senza richiedere misurazioni ulteriori a quelle effettuate sulle molecole contenute nel campione raccolto. Grazie a questo aspetto, si supera il problema dell'esistenza di altre possibili fonti di precursori che potrebbero non essere considerate, in quanto le informazioni che si ricavano dalle molecole sono sufficienti a descrivere completamente le condizioni iniziali del sistema. Tutto ciò che è necessario per procedere con le analisi successive sono le abbondanze relative delle molecole oggetto di analisi, indipendentemente dai possibili cambiamenti nelle condizioni del sistema biologico o dalla diversa sensibilità degli strumenti di misura.

Inoltre, le misurazioni da effettuare necessariamente in due momenti distinti, qualora non si potessero ricavare informazioni utili dal solo campione raccolto, sarebbero inevitabilmente affette da errori dovuti a cambiamenti spontanei avvenuti nel sistema biologico durante il periodo che intercorre tra le due misurazioni oltre alla diversa sensibilità degli strumenti utilizzati.

Non per ultimo, poiché il metodo di MIDA prevede l'utilizzo di marcatura con $^{2}H_2O$, che tra le sue vantaggiose caratteristiche possiede la proprietà di non essere radioattiva, esso è totalmente sicuro in quanto l'analisi combinatoria abbinata alla marcatura con acqua pesante è applicabile alla maggior parte dei composti biologici.

Nel caso in cui fosse necessario ripetere la misurazione, il metodo che coinvolge l'analisi combinatoria riesce a modellare perfettamente la presenza di eventuali residui di marcatori che non sono stati eliminati durante il periodo che intercorre tra la prima somministrazione e la successiva. Infatti, è possibile stimare in anticipo la quantità di isotopi residui contenuti all'inizio della seconda somministrazione, che saranno successivamente sottratti nel momento delle osservazioni condotte per la seconda procedura.

Si osserva inoltre che le relazioni precedenti sono alla base dei criteri di filtraggio, i quali assicurano che la distribuzione osservata per un certo isotopomero marcato sia esattamente quella che ci si aspettava di ricavare.

2.3. Limiti dell'analisi combinatoria per lo studio della dinamica dei composti

Le osservazioni realizzabili a partire dall'analisi combinatoria implicano notevoli vantaggi, tuttavia affinché la procedura condotta fornisca dati rilevanti da poter esaustivamente spiegare il comportamento dei composti proteici, bisogna adottare alcuni accorgimenti tali da non inficiare tutto il processo.

Lo studio dei campioni attraverso MIDA possiede dei limiti strutturali, in quanto dipende inevitabilmente dall'accuratezza dei valori ricavati utilizzando strumenti che per loro natura sono imprecisi, per questo motivo i dati raccolti soffrono di una certa indeterminazione che può alterare i risultati: qualora i valori ricavati con l'analisi combinatoria fossero affetti da errori troppo gravi, tutte le successive considerazioni non rifletterebbero la reale distribuzione degli isotopi nel sistema esaminato.

2.3.1. Cautele da adottare in fase di somministrazione e raccolta dei campioni

A partire dalla pianificazione delle tempistiche in cui si svolge l'esperimento, è essenziale notare come anche l'esposizione del soggetto alle molecole di acqua arricchita può risultare un limite alla buona riuscita dell'analisi. Infatti, valori di arricchimento troppo vicini al *plateau* comportano variazioni minime nell'incorporazione del deuterio, provocando al contrario grandi variazioni nei valori riferiti alla cinetica dei polimeri.

Si può infatti osservare che in corrispondenza di una variazione minima della quantità di marcatori nelle vicinanze del *plateau*, ovvero da 87,5% a 93,75%, si ottengono gli stessi risultati che si possono ricavare passando dallo 0% dei polimeri marcati al 50%. Da questo si deduce che valori di *f* maggiori di 0,75-0,80 non sono adatti a condurre osservazioni sulla cinetica dei polimeri, fatto che impone di prestare una certa attenzione nel momento della progettazione del protocollo sperimentale da adottare in modo tale da evitare il raggiungimento di questi valori, modificando la metodologia di somministrazione del marcatore e il processo di raccolta dei campioni. [17]

Nel grafico riportato in Figura 8, si osservano le variazioni nei valori delle abbondanze relative degli isotopomeri del peptide FEDGDLTLYQSNAILR, con n=29. Nel corso dei 32 giorni di somministrazione del marcatore, le variazioni delle abbondanze approssimano una costante avvicinandosi alla condizione di completa marcatura. A causa di questo comportamento, risulta necessario calcolare *f* secondo i criteri precedentemente descritti. Per

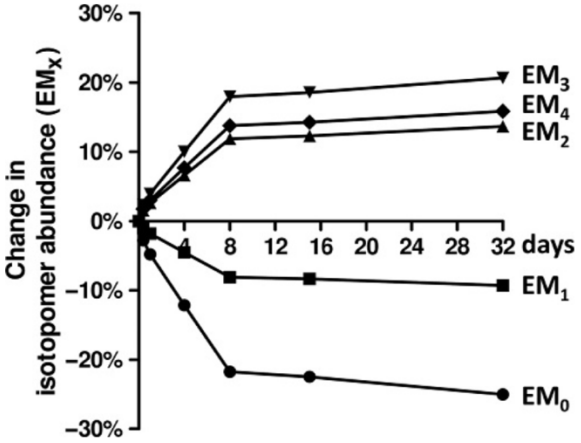

Figura 8. Il grafico riporta le variazioni delle abbondanze relative degli isotopomeri in relazione ai giorni trascorsi dalla prima somministrazione del marcatore, tratto da [8].

questo specifico caso, il giorno 8 risulta essere ideale per la raccolta dei campioni da analizzare, poiché si è raggiunto il livello di marcatura adeguato. [8]

Di fondamentale importanza è il momento in cui vengono raccolti i campioni da analizzare. Ogni amminoacido presenta un periodo di *turn over* diverso, questa caratteristica comporta alcune accortezze: in particolare, è possibile che, nel momento in cui si procede con la raccolta dei campioni, per alcune delle proteine oggetto di studio i valori di marcatura raggiunti non siano sufficienti a fornire alcuna spiegazione riguardo l'andamento della cinetica oppure che il livello di marcatura ottimale sia già stato superato fornendo dati anche in questo caso inutilizzabili.

A seconda della tipologia del polimero in esame lo sperimentatore dovrà ricavare le tempistiche adatte alla raccolta dati in modo tale che i valori ottenuti siano significativi. Generalmente, si sceglie come tempo di riferimento il massimo valore assunto per raggiungere il livello di marcatura ottimale nel campione di proteine che si intende studiare, tuttavia, questo approccio può invalidare i dati raccolti per altri polipeptidi per i quali si è già superato il valore ottimale nel momento in cui avviene la raccolta dei campioni.

La quantità di molecole marcate ideale a condurre delle valutazioni significative si aggira intorno all'intervallo che va dal 25% al 50% in modo tale che l'incremento o il decremento di *f* sia stimabile con una certa affidabilità e assuma valori compresi nell'intervallo tra il 10% e il 75% del valore di *f* stessa. [8]

Per analisi effettuate su un proteoma di nuovo studio è consigliato eseguire uno studio pilota, somministrando acqua pesante per un periodo di tempo variabile nel quale si procede alla raccolta dei campioni in tre o più momenti, i quali permettono in seguito di valutare le tempistiche adeguate a fornire dati significativi per le analisi successive. Nel caso in cui si raggiunga il *plateau* a valori inferiori del valore teorico, ovvero il 100% delle molecole marcate, risulta indispensabile considerare l'eventualità dell'esistenza di più di una fonte di sintesi proteica, sebbene questo fenomeno sia possibile anche nel caso contrario, e che non tutti i tessuti di origine siano considerati nello studio.

2.3.2. Raccolta dei campioni in relazione ai tempi di *turn over*

A livello pratico, non è generalmente possibile somministrare il marcatore sotto forma di bolo, secondo le considerazioni precedenti (si veda par. 1.2). Più comunemente, si è in presenza di un'esposizione non costante del marcatore. In svariati scenari clinici non è ammissibile

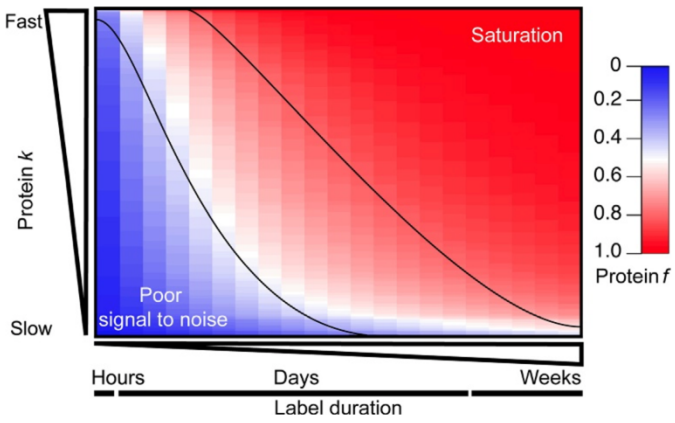

Figura 9. Si osserva l'intervallo ideale in cui effettuare la raccolta dei campioni in relazione alle caratteristiche di turn over *della proteina in esame e della durata della somministrazione del marcatore: l'area ideale è riportata tra le due curve nere, tratto da [8].*

somministrare la quantità di $^{2}H_2O$ tutta in un unico momento, appunto come un bolo, questa eventualità comporta il fatto che nella maggior parte dei soggetti l'andamento dei dati raccolti non è tale da descrivere il fenomeno per mezzo di una rampa.

In questi casi, la somministrazione di acqua arricchita, di conseguenza anche il contributo esatto degli isotopi nella sintesi delle nuove proteine in relazione alla totalità dei polimeri di nuova sintesi, non è costante nel tempo. Se la proteina in analisi presenta un periodo di *turn over* lungo, in corrispondenza di ogni campionamento saranno presenti polimeri che erano presenti anche nella precedente osservazione. D'altra parte, considerando proteine con *turn over* più veloce, i polimeri di nuova sintesi domineranno sul computo totale in corrispondenza a ogni misurazione effettuata.

Dalle precedenti considerazioni si ricava che, nel caso di proteine dal *turn over* lento, è consigliabile applicare un approccio tempo mediato, il quale è considerato più conveniente per approssimare l'andamento medio delle proteine di nuova sintesi per la stima del valore di *f*.

Al contrario, per proteine con un tempo di *turn over* breve, le proteine sintetizzate in precedenza hanno un peso poco rilevante sul computo totale, perciò l'unico valore di polimeri precursori da considerare sarà il valore ricavato in prossimità al giorno di raccolta dei campioni, rendendo possibile ignorare i valori precedenti.

2.3.3. Somministrazioni successive

Nel caso in cui si presentasse la necessità di eseguire delle analisi successive alla prima, affinché i dati raccolti siano significativi, bisogna procedere con la stima della quantità residua di marcatori presenti nel sistema biologico in cui si effettuano le analisi, con lo scopo di tenere traccia degli isotopi preesistenti, cosicché i risultati ricavati dai dati della seconda osservazione non siano falsati dalla presenza di residui di marcatori della prima analisi.

La procedura ottimale prevede di effettuare una biopsia preliminare tale da stimare l'effettiva quantità di marcatori residui, tuttavia questa è una pratica sconsigliata. Il percorso che si preferisce intraprendere prevede l'impiego di modelli matematici e di simulazioni numeriche. Inoltre, si suggerisce di tenere traccia delle condizioni precedenti alla seconda somministrazione di ^2H_2O, definendo in questo modo le sue condizioni iniziali, le quali saranno usate per stimare i valori di marcatura residua alla seconda somministrazione, in modo tale da rendere più agile un'ulteriore analisi qualora questa risultasse necessaria.

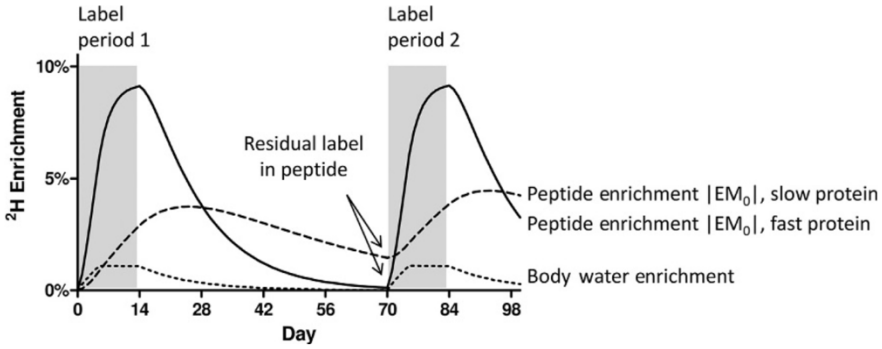

Figura 10. Nel grafico è riportato l'andamento delle curve di arricchimento nel caso di proteine con turn over *veloce (curva continua) e di proteine con* turn over *lento (curva tratteggiata) nelle condizioni di una seconda somministrazione di acqua pesante (curva con tratteggio più fitto), dove si pone l'attenzione sulla quantità di marcatori presenti nel soggetto residui dalla prima fase di somministrazione, tratto da [8].*

Per studi longitudinali che prevedono somministrazioni in successione, si adotta il protocollo di analisi tempo mediato, lo stesso in uso nel caso di proteine con lento *turn over*, per simulare la risposta a questo tipo di studi, adottando tutte le accortezze già evidenziate in precedenza.

2.3.4. Ulteriori cautele da adottare in seguito alla raccolta dei campioni

Alla base del successo del metodo di studio fondato sull'analisi combinatoria vi è la necessità di mantenere intatta l'integrità dei polimeri che si intende analizzare, il che implica necessariamente che non deve essere impiegato alcun tipo di processo chimico che comporti la riduzione in monomeri dei polimeri presenti nel campione. La disgregazione del composto proteico comporta la perdita di tutte le informazioni ricavabili dalla struttura dei polimeri, soprattutto riguardo le caratteristiche che possiedono in seguito alla sintesi avvenuta in presenza del marcatore. A ciò si aggiunge la perdita di identificabilità e quantificabilità delle sottopopolazioni di molecole, la quale non permette di ricavare alcuna informazione utile applicando l'analisi combinatoria.

A causa di quanto sopra considerato, risulta indispensabile l'utilizzo di un marcatore che distingua efficacemente i composti marcati da quelli non marcati il quale risulti allo stesso tempo isotopicamente stabile, ragion per cui, non possono essere impiegati radioisotopi per questo tipo di analisi.

Considerata la natura chimica del composto utilizzato come marcatore, non vi è alcuna particolare procedura da seguire necessaria per la raccolta e la conservazione dei campioni. Questo aspetto è dovuto al fatto che non possono intercorrere scambi di idrogeno tra gli atomi di deuterio incorporati nei peptidi del campione e il solvente (acqua comune) in cui si conserva data la stabilità dei legami H-C, contrariamente a quanto accade in presenza dei legami H-N oppure H-O, impiegati in altre tipologie di marcature metaboliche. [8]

3. STRUMENTI INFORMATICI A SUPPORTO DELLE ANALISI

Il settore informatico che si occupa di analizzare e organizzare i dati biologici derivanti dagli esperimenti discussi nei paragrafi precedenti, così come in tutte le altre aree della biologia, prende il nome di bioinformatica. Gli strumenti bioinformatici sono stati sviluppati per ben oltre trent'anni e la vasta varietà esistente non consente di determinare in modo univoco quale sia lo strumento migliore in termini di computabilità e affidabilità, soprattutto perché la scelta deve essere effettuata basandosi primariamente sullo scopo dell'analisi e sui metodi in cui si sviluppa l'attività di ricerca.

3.1. Introduzione al concetto di bioinformatica

Fin dal principio, negli studi biologici che coinvolgevano il sequenziamento di geni risultò evidente la necessità di creare un sistema di analisi efficiente che permettesse di confrontare ed elaborare i dati in tempi ragionevolmente accettabili; la difficoltà che gli scienziati riscontravano nel momento in cui dovevano trattare una vasta mole di dati era di ostacolo per la buona riuscita di questi studi.

Il momento in cui la bioinformatica nacque è databile intorno agli inizi degli anni Cinquanta del secolo scorso: sebbene non fosse stato ancora sequenziato il DNA e gli apparati computazionali fossero ancora in fase di sviluppo e scarsamente diffusi, date le considerevoli dimensioni oltre al costo spesso proibitivo, fu chiaro che rendere automatiche molte procedure di confronto sarebbe diventato l'unica via per effettuare analisi che altrimenti sarebbero state insostenibili.

Grazie al procedimento sviluppato da Edman, biochimico svedese il quale sviluppò il metodo di sequenziamento polipeptidico noto con il nome di *Degradazione di Edman*, divenne possibile sequenziare un numero di amminoacidi pari a circa 50 o 60, il problema vero e proprio risiedeva però nella ricostituzione delle informazioni derivanti da centinaia di tali sequenziamenti distinti, impresa impossibile in presenza di proteine generate da grandi sequenze amminoacidiche. Agli inizi degli anni Sessanta Dayhoff, scienziata statunitense nota per le sue ricerche nei campi della fisica e della biochimica e considerata la fondatrice della bioinformatica, in collaborazione con Ledley, sviluppò uno dei primi software che permetteva questa tipologia di analisi. Il risultato del loro lavoro è conosciuto con il nome di COMPROTEIN, ideato con lo scopo di identificare la struttura primaria delle proteine utilizzando i dati ricavati dalle singole sequenze

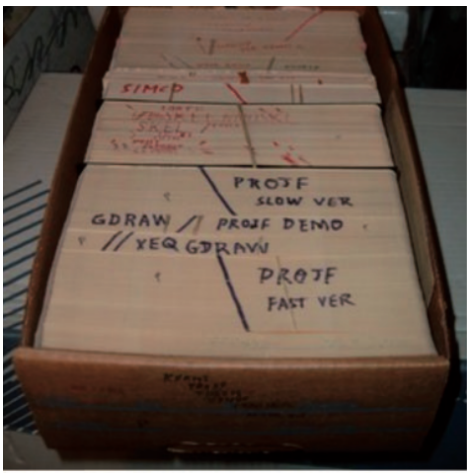

Figura 11. Codice sorgente completo di un programma per il software *COMPROTEIN su schede perforate, tratto da [18].*

di Edman. COMPROTEIN venne interamente scritto in linguaggio FORTRAN ed essenzialmente può essere definito come il primo software del tipo identificato come assemblatori *de novo*. A livello pratico, venne scelto di indicare gli amminoacidi per mezzo di sequenze di tre lettere, più tardi, risultò più agevole definire gli amminoacidi utilizzando esclusivamente una lettera. Quest'ultima codifica venne in seguito utilizzata nel 1965, quando la stessa Dayhoff in collaborazione con Eck crearono il primo database biologico, conosciuto con il nome di *Atlas of Protein Sequence and Structure*.

Nell'anno 1970, gli scienziati Needleman e Wunsch svilupparono il primo algoritmo dinamico per l'allineamento di coppie di sequenze proteiche. Tuttavia, solo negli anni ottanta emerse un algoritmo che consentiva l'allineamento di sequenze multiple, chiamato MSA (da *Multiple Sequence Alignment*). MSA propose una soluzione al problema che sorgeva nel momento in cui le proteine confrontate possedevano un discendente comune troppo distante per essere identificato come tale oppure nel momento in cui le sequenze in esame presentavano lunghezze diverse. Nel 1978, Dayhoff in collaborazione con altri colleghi sviluppò il primo modello probabilistico che descriveva la sostituzione degli amminoacidi. Esso si basava sull'osservazione dei punti di mutazione nell'albero filogenetico di varie famiglie di proteine le quali possedevano oltre l'85% di aspetti in comune. Poco più tardi, vide la luce il linguaggio di programmazione PERL (acronimo di *Practical Extraction and Reporting Language*). PERL risolveva l'esigenza di manipolare sequenze biologiche, le quali, adattandosi a essere rappresentate perfettamente da sequenze testuali, vennero rappresentate per mezzo di stringhe. Con caratteristiche analoghe a PERL, nacque Python, il quale era caratterizzato da un vocabolario contenuto e una sintassi semplice, tali da rendere il codice leggibile e di più facile interpretazione, sebbene, sostanzialmente, i due linguaggi potessero risolvere problemi analoghi. [18]

Negli anni successivi, oltre al miglioramento e alla nascita di altri software di analisi, si assistette all'evento che determinò la svolta nel campo delle ricerche scientifiche. Agli inizi degli anni Novanta venne creato il *World Wide Web*, una via che permetteva di scambiare informazioni e risorse in modo veloce e con diffusione globale. Questo aspetto è di fondamentale importanza perché, date le sue caratteristiche, forniva la possibilità di condividere

in tempo reale un'enorme quantità di dati e di strumenti bioinformatici che fino a quel punto era sempre stata ostacolata dalla mancanza di mezzi.

Se, all'inizio, la nascita del *World Wide Web* sostenne lo sviluppo di questi studi, analogamente a tutti i settori scientifici, oggi si sta assistendo a uno scenario diverso. In origine, la quantità di informazioni che venivano condivise era limitata in quanto tali strumenti di ricerca a loro volta non erano molto diffusi e l'utilizzo della rete era comunque una novità e non tutti gli scienziati aderirono nell'immediato, al contrario oggi, grazie ai progressi tecnologici e alla capillare diffusione della rete, i dati prodotti, raccolti e condivisi hanno raggiunto livelli considerevoli. Si stima che tutti i dati derivati dalla ricerca scientifica a livello globale raggiungano l'ordine degli Exabyte e questi spesso non sono usufruibili direttamente, poiché non vengono catalogati e ordinati risultando sostanzialmente inutilizzabili, talvolta complicando le operazioni di confronto e rendendole inconcludenti.

Da ciò deriva inevitabilmente la necessità di disporre di maggiori capacità computazionali che consentano di trattare un numero di informazioni così alto, non solo per quanto riguarda la raccolta, ma anche per la loro catalogazione e per renderle accessibili e fruibili. Il futuro della bioinformatica si fonda sulla creazione di strumenti intuitivi e facilmente adoperabili anche e soprattutto per incontrare le esigenze di tutti coloro i quali non hanno grande esperienza dell'ambito informatico. A questo scopo, sono nati negli ultimi decenni dei gruppi di collaborazione tra ricercatori i quali operano per raggiungere questi obiettivi.

Oltre alle difficoltà nel trattare una così vasta mole di dati, bisogna ricordare che, in qualche caso, si assiste alla totale o parziale mancanza di informazioni riguardanti intere molecole. Questo aspetto rende inevitabilmente ancora molto complesso generare dei modelli capaci di descrivere la dinamica di un sistema attraverso simulazioni ottenibili in tempi ragionevolmente contenuti quando si trattano composti ancora sconosciuti o per i quali le informazioni raccolte richiedono particolari attenzioni. [18]

3.2. Il perfezionamento degli strumenti di analisi

Negli ultimi decenni per gli studi riguardanti la proteomica si è assistito a un notevole sviluppo in merito alle metodologie di ricerca e questo aspetto ha comportato necessariamente la nascita di nuovi approcci informatici i quali consentono di elaborare velocemente, e con una certa affidabilità, i dati provenienti dagli esperimenti. Da queste considerazioni nasce l'esigenza di sviluppare strumenti informatici sempre più potenti e performanti che possano adempiere a

queste richieste oltre alle proprietà precedentemente citate in merito alla capacità di gestire molti dati allo stesso tempo. [19]

Nell'ambito della proteomica computazionale e della spettrometria di massa si assiste a una tendenza crescente nell'utilizzo di architetture ad alte prestazioni definite HPC (dall'Inglese, *High Performance Computing*) e, allo stesso tempo, ci si rivolge preferibilmente ad applicazioni in *Cloud* piuttosto che a desktop, in quanto presentano migliori caratteristiche computazionali in presenza di vaste quantità di dati e di analisi molto complesse. Sebbene la maggior parte degli strumenti commerciali rimangano esclusivi di Windows, come per esempio ProteomeDiscover e Spectronaut, esistono altre applicazioni *open-source* e strumenti non commerciali come MaxQuant, OpensMS e Skyline, i quali sono stati resi utilizzabili da piattaforme differenti. [20]

3.3. Il progresso della bioinformatica

Gli ostacoli che ponevano la raccolta e la condivisione dei dati biologici all'inizio di questi studi sono stati decisamente superati con la nascita del *World Wide Web*, tuttavia permangono delle difficoltà nel momento in cui si decide di riprodurre un esperimento e le relative analisi condotte in precedenza da altri ricercatori.

Per l'appunto, l'ampia varietà di strumenti informatici sviluppati negli anni ha permesso agli scienziati di adottare il sistema che ritenevano di volta in volta migliore, aspetto che, se da una parte facilita il lavoro del primo ricercatore in quanto utilizza uno strumento familiare, dall'altra rende difficile per altri scienziati le operazioni di validazione delle informazioni derivanti dallo studio servendosi della stessa procedura, poiché lo strumento ritenuto adatto dal primo ricercatore, potrebbe essere complicato da utilizzare per i ricercatori successi. Per ovviare a questo problema, negli anni sono nate numerose iniziative da parte di informatici, per raccogliere e categorizzare gli strumenti impiegati comunemente in bioinformatica. Così facendo, si semplifica la fase di ricerca dello strumento adatto e inoltre si coinvolgono tutti i fruitori dei *software* a fornire informazioni riguardanti il corretto utilizzo, così come i punti di forza e gli ambiti in cui mostra le prestazioni migliori.

A tal proposito si cita BioContainers, una comunità di bioinformatici e sviluppatori di *software*, la quale, tra i molti progetti, ha come fine quello di sviluppare e accrescere nuovi contenitori di *software* per condurre analisi bioinformatiche, nonché lo sviluppo e l'implementazione di ricerche facilitate che permettano di velocizzare la fase iniziale oltre a rendere l'utilizzo di tali

strumenti più intuitivo e immediato. In questo modo sarà più semplice riprodurre le procedure sperimentali e saranno velocizzate le operazioni di validazione dei protocolli, cosicché anche i ricercatori meno esperti nell'ambito informatico non siano ostacolati da queste pratiche, altrimenti complesse e dispendiose in termini temporali.

BioContainers è stato sviluppato seguendo i principi di condotta della politica FAIR (*Findable, Accesible, Interoperable and Reusable*) proposte nel 2017 da Jiménez secondo le quali un *software* deve essere facile da trovare, accessibile, interoperabile e riutilizzabile. [21] A tal proposito, per facile da trovare si intende che gli scienziati devono fornire ai colleghi informazioni tra le quali il titolo, una descrizione generale, la pubblicazione, la licenza e non da ultimo la versione utilizzata per condurre le analisi. In merito all'accessibilità dello strumento ci si riferisce al fatto che il codice del *software* deve essere reso disponibile in collegamento. Per interoperabilità si intende di fatto che deve essere interscambiabile tra i maggiori registri di *software*. Infine esso deve essere riutilizzabile, cioè deve essere adottata una licenza, la quale definisce i limiti imposti all'utilizzo dello strumento.

Infatti, uno degli aspetti che sono considerati di maggiore ostacolo riguarda l'installazione e la configurazione del *software*. Inoltre, comprendere quali strumenti tra quelli disponibili e, non solo, quali strategie operative siano legate alla pubblicazione in questione, risultano essere, in genere, operazioni molto complesse. Benché vengano pubblicati il codice sorgente del *software* e i dati di riferimento, potrebbe esistere una qualche forma di dipendenza con altri *software*, altri sistemi operativi o, addirittura, opzioni di configurazione le quali rendono impossibile replicare fedelmente l'esperimento. Esattamente per queste ragioni è risultato necessario costruire dei pacchetti *software* per superare queste limitazioni e difficoltà.

Nel caso di BioContainers, viene fornito un *Restful* API (*Application Programming Interface*) che introduce due principali funzionalità, la ricerca di strumenti bioinformatici e dei loro corrispondenti contenitori, e fornisce informazioni inerenti a quello specifico strumento e al contenitore di riferimento. Il *Restful* API è sostanzialmente l'implementazione dello standard GA4GH (*Global Alliance for Genomic and Health*) il quale consente la compatibilità con altre risorse. [20]

Il GA4GH è un'associazione internazionale e non-profit nata nel 2013 con l'intento di massimizzare il potenziale nell'ambito della ricerca e della medicina per garantire lo sviluppo della salute umana, mira ad accelerare in particolare il progresso nel campo genomico garantendo l'adesione a un protocollo comune di procedure standard e protocolli per la condivisione consapevole di tutti i dati relativi al genoma e alla salute umana. [22]

Con caratteristiche analoghe a BioContainers, si cita anche MPI *Bioinformatic Toolkit*, introdotto nel 2005, esso fornisce ai ricercatori coinvolti nell'ambito biologico uno strumento facile da utilizzare, il quale permette di accedere via web e fornisce i migliori strumenti e database utili per condurre le analisi bioinformatiche. [23]

Data l'estensione degli ambiti di ricerca in cui questi strumenti sono utilizzati e di conseguenza il gran numero di differenti procedure sperimentali impiegate, si assiste allo sviluppo di molti *software* con caratteristiche specifiche, i quali sono caratterizzati da certe peculiarità e non sono sempre adatti a condurre ogni tipologia di analisi. Nel momento in cui si progetta un esperimento, la scelta dello strumento con cui condurre le analisi relative può costituire un ostacolo alla sua buona riuscita: lo strumento che presenta caratteristiche migliori a livello computazionale e di precisione nel caso di un certo esperimento, può non essere adatto per un altro, ciò può causare una certa lentezza nelle procedure computazionali e, nei casi peggiori, compromettere i risultati.

Per facilitare lo sviluppo della proteomica computazionale è necessario l'impegno di tutti i ricercatori e gli accademici coinvolti direttamente nelle procedure di analisi. Infatti, affinché si possa progredire in questo ambito è essenziale che ogni studioso condivida informazioni riguardo l'utilizzabilità e le caratteristiche di tali *software*, in modo tale da permettere un'estensione della platea di ricercatori che può usufruire di questi sistemi. Per ottenere programmi sempre più efficienti e orientati a un utilizzo specifico è necessario suddividere in categorie i vari strumenti attraverso la descrizione quanto più possibilmente completa e accurata delle loro peculiarità e funzioni. Dato il gran numero di strumenti esistenti, la diffusione di ciascun *software* e la corretta definizione delle sue funzionalità dipende primariamente dagli sforzi collettivi degli scienziati e dei ricercatori che hanno potuto conoscere e lavorare con un certo strumento in modo tale da agevolare la scelta in merito al programma più adatto ai loro colleghi. Poiché i programmi utilizzati in questo ambito di ricerca sono sviluppati da un insieme eterogeneo di interpreti a partire dai singoli scienziati o piccoli gruppi di ricerca, fino a *team* di sviluppo internazionali e aziende private, è necessario che gli sviluppatori stessi rendano note le caratteristiche principali dei loro codici, rendendo accessibile e identificabile il *software* immediatamente dopo il suo rilascio. Iniziative come *bio.tools* e *ms-utils.org* raccolgono informazioni in merito alle risorse utilizzate in ambito di proteomica computazionale per poter facilitare l'identificazione di strumenti efficienti in grado di soddisfare le richieste di una specifica attività. [24]

3.4. Considerazioni generali sulle applicazioni degli strumenti informatici

Un prerequisito fondamentale nel caso in cui si voglia comprendere le funzioni e il comportamento di una proteina si basa sull'identificazione corretta e caratterizzata da un certo livello di affidabilità in modo tale da evitare di incappare in deduzioni erronee.

L'identificazione di un composto proteico non è un'impresa facile, infatti, la struttura interna del composto, ossia il fatto che è formato da una lunga sequenza di peptidi, insieme alla capacità posseduta di assumere funzioni diverse, le quali sono influenzate da più di 28 possibili mutazioni posttraslazionali che comportano oltre 10 mila potenziali variazioni, rendono questo passaggio molto articolato.

Un *data system* dedicato in modo esclusivo alle analisi spettrometriche è descrivibile sostanzialmente come il potere computazionale, ossia *hardware* e *software*, dedicato esclusivamente all'acquisizione dei dati derivati dalla spettrometria di massa, il loro processamento e le successive analisi dei dati eseguite in modo efficiente e automatico. Lo spettro di massa identifica le relazioni esistenti tra le abbondanze di massa relative e l'intensità di ciascun ione proveniente da una certa proteina o peptide e il suo valore di m/z.

Un aspetto limitante nelle analisi bioinformatiche e nel caso di identificazione computazionale dei composti è dovuto dalla natura degli strumenti, i loro parametri e la loro configurazione. Per poter utilizzare gli strumenti informatici adatti a questo scopo è necessario innanzitutto convertire i dati raccolti in un formato openXML, il quale può essere letto e interpretato da diversi *software* di analisi, in seguito, si effettua l'identificazione per mezzo del *database*, quindi si procede con la validazione dei valori tipicamente mostrati dalla proteina e, se necessario, la si quantifica. [25]

In riferimento alle metodologie di analisi spettrometriche descritte in precedenza (si veda par. 2.1) sono necessarie strategie computazionali e analitiche diverse a seconda del caso che si sta trattando.

Per quanto riguarda l'approccio di analisi spettrometrica *bottom up*, dopo aver effettuato la raccolta dati, il primo passo da compiere è l'identificazione delle sequenze peptidiche grazie alla consultazione delle raccolte di spettri di frammentazione presenti nei database oppure è possibile effettuare il sequenziamento *de novo* di volta in volta. Considerando il metodo di confronto con gli spettri già catalogati, si assegna un punteggio in relazione alla qualità della corrispondenza risultante dal confronto tra lo spettro ottenuto sperimentalmente e tutti i possibili spettri presenti nella raccolta. La sequenza peptidica per cui si ottiene il punteggio tra

tutti più alto rappresenta il peptide candidato. Per poter effettuare queste analisi ci si serve di alcuni *database*, tra i quali si citano SEQUEST e MASCOT, che restituiscono il peptide più simile a quello in esame servendosi di analisi statistiche. Affinché i risultati ottenuti siano considerati affidabili, è necessario prestare molta attenzione alla scelta dei valori di tolleranza entro i quali si può operare la ricerca: se l'intervallo tra il valore massimo e il valore minimo è troppo stretto, c'è la possibilità di esclusione di alcuni spettri che in realtà dovrebbero essere considerati perché potenzialmente validi, al contrario, in caso di intervallo troppo ampio, il rischio in cui si incorre è quello di ottenere un risultato che in effetti è ricavato da valori che non sono significativi. Nel caso del sequenziamento *de novo*, la sequenza peptidica si ricava unicamente dalle informazioni ottenute analizzando lo spettro di frammentazione e dalle stesse proprietà di frammentazione, tuttavia, in presenza di dati incompleti, non è possibile in ogni caso ottenere una stima valida.

Con il fine di migliorare le capacità di identificazione, è pratica attuale utilizzare una combinazione tra le tecniche descritte in precedenza, in modo tale da sfruttare le peculiarità e i vantaggi derivanti da ciascuna.

In seguito all'identificazione del peptide, ci si occupa della ricostruzione della sequenza peptidica che costituisce una certa proteina. Di norma, non è possibile costruire una lista affidabile di proteine a partire da peptidi corti in quanto alcuni di questi potrebbero essere condivisi da più proteine e, perciò, per questo tipo di peptidi sono accettabili più soluzioni.

Per stimare l'abbondanza relativa delle proteine all'interno del proteoma considerato in un certo campione, è possibile avvalersi dell'analisi proteomica quantitativa. Per il momento, non esiste alcuna procedura di analisi standard per ricostruire totalmente il profilo genetico di una proteina a partire da un campione, poiché esistono vari algoritmi che fanno riferimento a tecniche sperimentali specifiche e non è possibile determinare un criterio univoco.

Per quanto riguarda l'analisi in assenza di marcatura, gli spettri dei singoli saggi sono acquisiti servendosi di distinte analisi LC-MS/MS, questo aspetto comporta inevitabilmente alcune variazioni indesiderate come, per esempio, considerare distinte misurazioni liquido cromatografiche le quali sono tra loro inconsistenti. La maggior parte dei *software* utilizzati in supporto a questa strategia si basano sull'analisi dell'intensità del segnale ricavato dagli ioni precursori dei peptidi costituenti le proteine precedentemente frammentate. Avvalendosi di questo approccio, risulta possibile effettuare correzioni in merito al numero di peptidi esaminati a partire dalla probabilità di osservazione di quel peptide attraverso procedure di *machine learning* impiegando le proprietà fisico chimiche che possiede.

Nelle procedure di quantificazione dei peptidi basate sulla spettrometria di massa, l'abbondanza di proteine è ricavata a partire da un numero limitato di peptidi, a volte addirittura solamente uno. Per aumentare l'affidabilità delle analisi condotte, risulta necessario apportare un notevole dimensionamento in merito alla quantità minima di peptidi considerati; se così non fosse, si osserverebbe una maggiore interferenza causata dalla presenza di un numero elevato di elementi. Inoltre, per generare dei risultati confrontabili e affidabili per condurre le osservazioni successive, è necessario normalizzare e rielaborare i dati grezzi ottenuti in fase di raccolta. Anche in questo caso, la scelta delle procedure da effettuare è estremamente delicata in quanto decidere di effettuare un'operazione di pulizia troppo incisiva comporta la perdita o il danneggiamento delle informazioni. In ogni caso, la scelta dipende esclusivamente dalla tipologia di studio che si intende condurre e quindi non esiste alcuna linea di condotta che indichi come effettuare queste decisioni, ma deve essere una scelta valutata in relazione al caso singolo. Nel caso di dati incompleti, l'efficacia di queste analisi diminuisce in modo considerevole. [19]

Nello specifico, per quanto riguarda la tecnica di analisi *in vivo* effettuata utilizzando il *labeling* metabolico attraverso molecole di $^{2}H_2O$, si osservano alcune limitazioni dovute agli strumenti informatici. Infatti, la presenza di marcatura introduce nel sistema una dimensione aggiuntiva, ossia quella temporale, la quale non si considera per gli studi condotti *in vitro* o di proteomica statica, nei quali non esiste la variazione nel tempo e si osservano solamente le condizioni al tempo presente. In questo caso è necessario sviluppare un *software* che tenga in considerazione anche questa peculiarità. Non solo, data la scarsa marcatura che si osserva, dovuta alle caratteristiche proprie del composto, il problema generato dall'incompletezza dei dati deve essere rimediato attraverso procedure di rielaborazione con strumenti informatici. A questo scopo, è necessario sviluppare un *software* ad accesso libero che raccolga tutte le caratteristiche utili per effettuare delle analisi coerenti con i dati raccolti. [11]

3.5. Nuove strategie di analisi e sviluppi futuri

Nonostante i numerosi sforzi condotti dalla comunità scientifica per elaborare manualmente gli spettri, solo una minima parte dei composti biologici è stata totalmente sequenziata. Si stima che solamente il 2% degli spettri acquisiti con il metodo di spettrometria di massa siano stati estratti attraverso il *matching* diretto, anche se per composti biologici noti come l'*Escherichia coli* e plasma, urina e linea cellulare umani si sono raggiunti negli ultimi tempi livelli pari al 10%. [19]

Al fine di colmare queste mancanze, si stanno studiando degli algoritmi che sappiano trattare con la mancanza di spettri di riferimento in modo particolare per quanto riguarda il caso di piccole molecole. In questa prospettiva, si stanno sviluppando algoritmi di *machine learning* e, più recentemente, di *Deep Learning*, i quali, rispetto ai metodi bioinformatici tradizionali, semplificano le procedure di identificazione dei polimeri. Inoltre, questi strumenti, grazie alle loro caratteristiche, potrebbero essere in grado di accertare regole di frammentazione rare e di difficile riconoscimento così come di configurazioni inaspettate, caratteristiche che non possono essere ricavate per mezzo di approcci basati unicamente sull'osservazione dei dati raccolti dalle procedure sperimentali poiché questi comportamenti inusuali non sottostanno a regole predefinite.

Le strategie di *machine learning* mirano a ricavare l'andamento degli spettri teorici dei composti non presenti nelle raccolte, avvalendosi delle regole di frammentazione note e dei calcoli sugli standard energetici. Ogni modello generato con tecniche di *machine learning* può essere pensato come un'approssimazione di una funzione matematica, che analizza le relazioni che esistono tra i dati in *input* e i risultati in *output* nei casi in cui non è nota fin dal principio la relazione che li lega. Servendosi di questi concetti, si possono ricavare informazioni a partire da un'enorme quantità di dati proteomici ed è possibile costruire modelli adattati in modo specifico al caso in esame, i quali possono essere validati utilizzando *dataset* diversi oltre alla possibilità di impiegare unicamente informazioni descrittive. In particolare, per lo sviluppo di algoritmi basati sul *machine learning* si citano due approcci: il *supervised learning* e l'*unsupervised learning*.

Nel caso *supervised*, si utilizzano ampi *dataset* contenenti i dati sperimentali senza essere a conoscenza di aspetti riguardanti le caratteristiche fisiche e chimiche che caratterizzano il processo di frammentazione, con il fine di ricostruire modelli capaci di predire risultati clinici a partire dai dati ottenuti da un paziente. Questa strategia si basa sull'ausilio di algoritmi che predicono la sottostruttura dei composti, a questo scopo, ogni piccola molecola è trasformata nel suo descrittore molecolare seguendo alcune regole predefinite. In riferimento a questa tipologia di algoritmi, un descrittore molecolare è sostanzialmente un vettore binario, il quale indica la presenza o meno di certe proprietà chimiche e di sottostrutture oppure l'appartenenza a una certa classe di composti chimici. In sostanza, il modello è costruito a partire dagli spettri ricavati dalla spettrometria di massa di molecole note, le quali vengono talvolta trasformate, per esempio, in alberi di frammentazione o combinati con i dati ottenuti da altre procedure di analisi. A tal proposito, è stato possibile predire i risultati clinici, come il tempo di

sopravvivenza, a partire dall'analisi dei dati proteomici per mezzo di modelli elaborati servendosi di campioni raccolti da soggetti le cui condizioni cliniche erano note a priori. [19]

Nel caso *unsupervised*, lo scopo è quello di ricavare la struttura biologica originaria e le dipendenze che esistono all'interno di un definito insieme di dati fornendo unicamente i valori in ingresso. In merito a questo aspetto, si sottolinea l'importanza dell'approccio *Deep Learning*, il quale si sta recentemente diffondendo nell'ambito della proteomica e permette di estrarre autonomamente strutture ricorrenti nei dati a partire da alti livelli di astrazione senza avvalersi dell'ausilio di procedure di ingegnerizzazione dei dati grezzi e introducendo potenziali interazioni tra migliaia di dati. Questo metodo basa il suo funzionamento sulla creazione di reti neurali (*Neural Networks*) che emulano il comportamento e la struttura del cervello umano. Le reti neurali sono costituite da semplici unità chiamate neuroni, in analogia con i neuroni del sistema nervoso, connesse le une alle altre in cui ciascuna connessione è caratterizzata da un peso differente. I pesi assegnati a ciascun collegamento hanno lo stesso ruolo delle interconnessioni sinaptiche esistenti negli organismi biologici. In base alla tipologia dei collegamenti esistenti tra i neuroni si possono identificare vari modelli: *Convolutional Neural Network* (CNN) e *Reccurrent Neural Network* (RNN), che a sua volta si differenzia in *Gated Reccurent Unit* (GRU) e in *Long Short-Term Memory* (LSTM). [26]

Data la grande disponibilità di dati raccolti negli anni di ricerca, in particolare ci si riferisce agli spettri ricavati dagli esperimenti, il metodo *Deep Neural Networks* (DNN) si sta imponendo come valida alternativa rispetto ai modelli più comuni di *machine learning* basati sullo studio della regressione lineare, sulle considerazioni ricavabili a partire dai discriminanti e dai *Support Vector Machines*, fatto che è stato possibile anche grazie all'incremento delle capacità computazionali

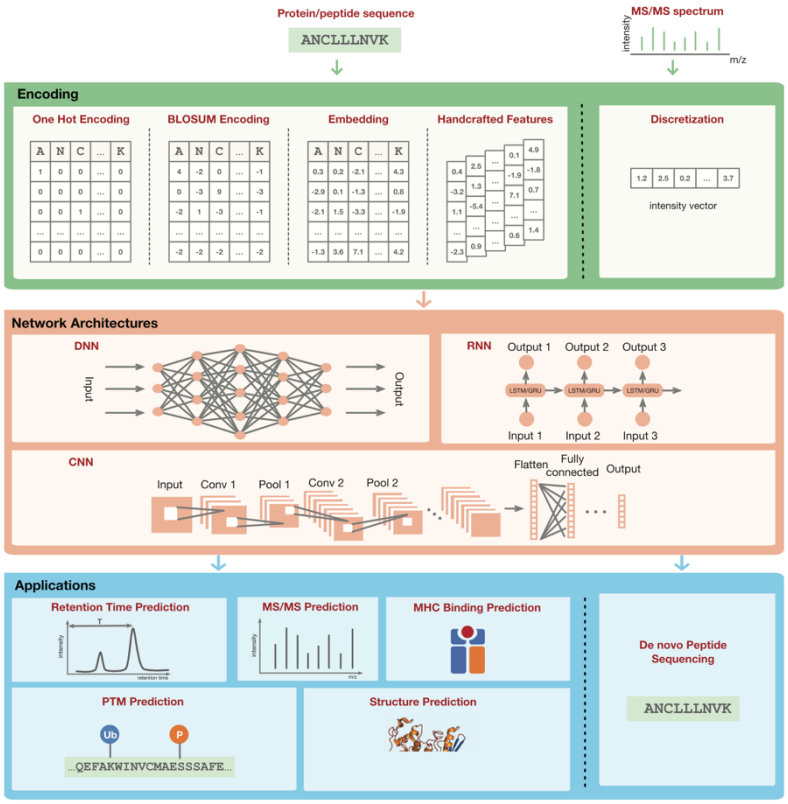

Figura 12. Riassunto degli elementi di base del Deep Learning *e delle sue applicazioni in proteomica, tratto da [27].*

disponibili. Per elaborare una rete neurale è infatti necessario avvalersi di un ampio campione su cui far esercitare la rete (denominato *training data*), alla quale vengono forniti successivamente dei dati in ingresso, per esempio sequenze peptidiche, delle caratteristiche associate ai dati, come i tempi di ritenzione dei peptidi, di un modello di rete, di una funzione che stima la bontà della previsione chiamata *loss function* e, per ultimo, di un metodo di ottimizzazione della rete. I modelli così costruiti hanno dimostrato di superare di gran lunga le prestazioni dei metodi tradizionali di *machine learning*, inoltre, tra tutte le tipologie di modelli *unsupervised* possibili, quelli ottenuti a partire dal *Deep Learning*, risultano essere caratterizzati dalla più alta capacità di generalizzazione.

Se si considera, per esempio, il caso in cui si utilizza il *Deep Learning* per prevedere lo spettro di un certo peptide, i risultati che si ottengono sono molto simili a quelli ottenuti a livello sperimentale. In più, è necessario sottolineare che spesso le similitudini tra gli spettri ottenuti sperimentalmente e quelli predetti con il *Deep Learning* sono maggiori rispetto al caso in cui si considerano gli spettri di uno stesso peptide ottenuti da analisi spettrometriche successive. [27]

In aggiunta, le *Deep Neural Networks* sono risultate molto efficienti nel predire il sequenziamento *de novo* dei peptidi. A tal proposito, le sequenze peptidiche sono generate direttamente a partire dagli spettri ottenuti dalle analisi spettrometriche senza l'ausilio di alcun *database*. Il sequenziamento *de novo* condivide delle somiglianze con i metodi di descrizione delle immagini ottenuti per mezzo del *Deep Learning*, in merito si cita il funzionamento del *software* DeepNovo, il quale tratta lo spettro in *input* come una generica immagine e la sequenza peptidica in *output* che ne deriva, come una frase scritta nel "linguaggio delle proteine". [27] Dal lato pratico, DeepNovo ha dimostrato prestazioni nettamente superiori rispetto ad altri algoritmi creati allo stesso scopo sotto molteplici aspetti: è noto che DeepNovo è stato in grado di ricostruire oltre il 97% delle sequenze di anticorpi con precisione superiore al 97% senza il supporto di *database*. [26]

Tuttavia, per poter trattare gli spettri ricavati sperimentalmente, è necessario operare un processo di discretizzazione iniziale con lo scopo di creare un vettore di intensità che li descriva. Per i dati riconducibili a proteine e peptidi, la sequenza primaria viene segmentata ulteriormente in sotto sequenze, che corrispondono sostanzialmente agli amminoacidi. A questo punto, a ciascuna sotto sequenza viene assegnato un vettore numerico che assume caratteristiche diverse a seconda del tipo di associazione che si considera. La tipologia di associazione più semplice e diffusa è conosciuta con il nome *One-Hot Encoding*, secondo la quale ciascun amminoacido viene rappresentato da un vettore binario di lunghezza pari a n contente un solo elemento di valore 1 e i restanti $n-1$ elementi di valore nullo, così facendo ciascuna sotto sequenza è

trattata in modo uguale alle altre senza possedere alcuna conoscenza a priori che le possa distinguere. Un secondo approccio chiamato BLOSUM (da *Blocks Substitution Matrix*) prevede l'utilizzo di matrici di sostituzione a blocco, in cui ogni amminoacido è rappresentato da una specifica riga nella matrice. In questo caso, viene mantenuta l'informazione sulla struttura evolutiva della proteina, ovvero si sottolinea quale coppia di amminoacidi può commutare durante l'evoluzione, in questo modo gli amminoacidi non possono più essere trattati in modo indipendente l'uno dall'altro. Oltre alle tipologie precedenti, si ricorda il metodo *Word Embedding*, il quale utilizza vettori numerici densi, ampiamente utilizzato nei casi in cui si conducono analisi in merito al linguaggio naturale, oppure è possibile costruire delle strutture su misura da utilizzare in ingresso al modello che si intende elaborare. [27]

Un limite esistente all'utilizzo e all'effettiva ampia diffusione di questi modelli si fonda sul fatto che sono necessarie decine di migliaia di esempi di singole distribuzioni di dati per creare una rete neurale efficiente. Tuttavia, una volta che il modello è stato creato, è possibile predire i risultati di analisi future fornendo unicamente i dati in ingresso e, non di meno, avvalendosi del processo di *transfer learning*, è possibile applicare reti generate a uno scopo iniziale in tutti quei casi in cui i dati in ingresso presentano una struttura simile a quelli con cui si è creato il modello, anche in ambiti molto diversi da quello originale, fornendo delle architetture molto più agili e altamente personalizzabili, come se si considerano i modelli di classificazione di oggetti generici e di classificazione di farmaci.

L'approccio *Deep Learning* è stato utilizzato, per esempio, nel caso della ricostruzione di un modello per verificare l'efficacia di alcuni farmaci utilizzati in presenza di cellule tumorali ed è stato creato a partire da informazioni riguardanti le caratteristiche del genoma e del proteoma del caso esaminato. [19] Nello specifico, lo studio condotto da Ding e colleghi [28] ha dimostrato l'efficacia dei metodi di *machine learning*, e principalmente, di *Deep Learning*, nel perfezionamento dell'oncologia di precisione. Infatti, l'assenza di *biomarkers* capaci di stabilire l'efficacia dei farmaci impiegati in ambito chemioterapico rende impossibile intraprendere fin dal principio, una terapia personalizzata in base al paziente e alla tipologia di cellule cancerogene, perciò è possibile applicare esclusivamente terapie non specifiche, le quali, spesso si rivelano inefficaci e causano come unico risultato l'intossicazione dell'organismo già compromesso dalla malattia. Ding e colleghi hanno sviluppato dei modelli di classificazione basati sulla regressione e su *Support Vector Machines* (SVM) a partire da dati genomici per 140 farmaci comunemente utilizzati nelle terapie osservando risultati ancora lontani dall'essere affidabili. Le maggiori limitazioni subentrano principalmente a causa delle notevoli dimensioni dei dati da elaborare e dal corrispondente ristretto campione di *training data* disponibile,

conducendo a condizioni di *overfitting*. Il modello basato sulla regressione utilizzato è noto con il nome di *elastic net*, ossia un tipo di regressione logistica con un termine di regolarizzazione ibrido che combina i valori di *lasso* e *ridge regularization*. Per ciascun farmaco analizzato sono stati costruiti 6 modelli, ciascuno caratterizzato da vettori in ingresso di dimensioni diversa e il *target vector* costituito dai dati discretizzati riguardanti la sensibilità delle cellule per quel particolare farmaco. Nel secondo caso, il modello basato su *Support Vector Machines* è caratterizzato da un *kernel* gaussiano e per ciascun farmaco sono stati creati due modelli: il primo servendosi dei dati sperimentali completi, il secondo, invece, basato solamente su una parte delle caratteristiche ricavabili dai dati. I valori utilizzati per questo studio contengono informazioni su 624 linee cellulari distinte e ogni composto è stato testato in media solamente per 586 di queste, con i seguenti risultati: 11 linee cellulari non hanno interagito con nessun farmaco, per le restanti 613, in media 14,5 composti hanno dato dei risultati. Valutando le *performance* dei modelli, i quali determinano se il farmaco con cui una certa cellula interagisce ha una risposta efficace, è risultato che nella maggioranza dei casi i modelli elaborati falliscono nell'identificazione delle cellule cancerogene sensibili ai farmaci specifici. Il modello *elastic net* ha ottenuto risultati migliori rispetto al modello SVM: nel primo caso, il valore di AUROC, acronimo di *Area Under the Receiver Operating Characteristic* termine utilizzato come statistica riassuntiva della bontà del modello, è pari a 0,81 in relazione a un valore medio di sensibilità di 0,75 e di specificità pari a 0,78, mentre nel secondo caso il valore medio di sensibilità è pari a 0,59 e il valore medio di specificità è 0,56 con AUROC di 0,55. In base a questi risultati, è chiaro che l'algoritmo basato sulla regressione ha superato le prestazioni del modello SVM. Tuttavia, è emerso che per alcuni farmaci, anche nel caso del modello più performante, esistono problemi legati all'*overfitting* principalmente dove le dimensioni dell'*input* sono considerevoli e, in certi casi, neppure il modello *elastic net* ha fornito alcun risultato significativo, portando i ricercatori a considerare l'esistenza di alcuni legami e *pattern* tra i dati che sfuggono a questa tipologia di modelli. Perciò si è optato per l'utilizzo di un *autoencoder*, ossia un algoritmo generato attraverso una *Deep Neural Network*, allo scopo di identificare quei legami non ancora considerati dai metodi precedenti. In ciascun livello della rete neurale i dati in *input*

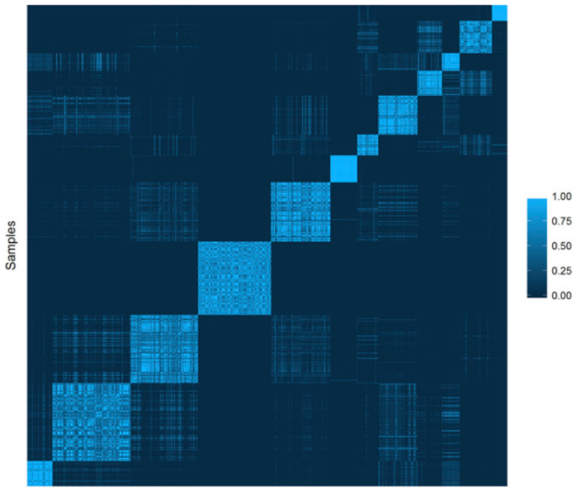

Figura 13. Frequenza relativa del clustering *delle linee cellulari basato sull'*autoencoder *con caratteristiche* Hidden 1. *L'intensità del* plot *indica la frequenza relativa, con la quale una coppia di campioni si associa nel caso di* clustering *gerarchico successivo o di sotto insiemi dello stesso* dataset*, tratto da [28].*

sono rielaborati assegnando dei pesi i quali sono conseguentemente propagati al successivo, in modo tale da permettere in ogni livello la decodifica della distribuzione statistica che soggiace ai dati genomici. Grazie a questo programma, sebbene le prestazioni globali non migliorino, per alcuni farmaci scarsamente descritti dai modelli precedenti si sono ricavate informazioni più complete e precise. Da tali considerazioni è stato possibile osservare che l'utilizzo di un algoritmo basato esclusivamente sui dati sperimentali ha prestazioni nettamente superiori ai metodi sviluppati in precedenza, consentendo di determinare a priori l'efficacia di un particolare farmaco in base alla tipologia di cellula, considerando le caratteristiche genomiche delle cellule obiettivo di un farmaco come indicatori dell'efficacia della terapia che ne deriva e aprendo nuove strade all'applicazione dell'oncologia di precisione.

Nonostante si dispongano di molti dati nonché di *software* potenti, l'identificazione di composti ignoti richiede tuttora una laboriosa attività di interpretazione manuale e l'intervento diretto dei ricercatori. Molti spettri non possono essere analizzati utilizzando gli strumenti bioinformatici usuali a causa della mancanza di campioni oppure, in alcuni casi, di algoritmi appropriati per condurre uno specifico studio. Anche nell'applicazione di strumenti computazionalmente potenti come quelli ivi citati, affinché rivelino le loro migliori prestazioni è di fondamentale importanza ottimizzare i dati di partenza, il che implica aumentare la quantità, la qualità e la varietà di questi, ovvero eliminare gli spettri di scarsa qualità, riscontrare e correggere gli errori compiuti nei processi di identificazione precedenti e integrare in questi studi anche i dati provenienti da altre tipologie di analisi. A tutto questo si aggiunge la necessità di condurre studi ulteriori che prendono in considerazione anche strutture che sono al di fuori dei database utilizzati durante i primi stadi di analisi.

Come già sottolineato, è basilare rendere questi software più accessibili, creando un tramite tra sviluppatori e utilizzatori, soprattutto nel caso di algoritmi così sofisticati e adattabili, per il quale utilizzo è necessario fare riferimento a un protocollo, il quale deve includere il metodo necessario per trattare i dati grezzi, le varie possibilità di personalizzazione, l'interpretazione dei modelli e il *batch processing*, in modo tale da rendere *user-friendly* questi strumenti altamente potenti ma, finora, complessi. [29]

4. APPLICAZIONE DEL *LABELING* METABOLICO NELLO STUDIO DI PATOLOGIE CLINICHE

Gli studi condotti sul proteoma umano e animale hanno trovato un grande uso pratico nella medicina, non solo a fini di ricerca, bensì trovando impiego nello sviluppo di nuove terapie. A livello biologico, la conoscenza delle interazioni tra i composti proteici in relazione alla loro genesi ed evoluzione aiuta a comprendere il funzionamento dei normali processi cellulari, tuttavia, osservando delle variazioni inaspettate di questi, è possibile ricavare informazioni riguardanti la presenza di malattie e disfunzioni.

Nei paragrafi successivi ci si concentrerà nell'analisi di alcune patologie le quali non sono state ancora del tutto chiarite, sebbene ci siano stati numerosi sforzi da parte della comunità scientifica, poiché attaccano organi difficilmente raggiungibili dalle tradizionali tecniche diagnostiche.

4.1. Considerazioni iniziali

Le malattie neurodegenerative sono un disturbo in rapida diffusione negli ultimi decenni, aspetto legato inevitabilmente anche all'invecchiamento progressivo della popolazione e ciò che ne ostacola maggiormente la cura e la diagnosi è da imputare alla mancanza di efficaci strumenti diagnostici. Questi disturbi possono essere normalmente accertati solo con una biopsia del cervello e quindi non è possibile conoscere con sicurezza se si è tratti effettivamente di tale patologia fino al decesso del paziente. Le tecniche tradizionali si basano sull'osservazione di anomalie in corrispondenza dell'organo, tuttavia non sono ritenute affidabili, in quanto i sintomi iniziali sono comuni alla maggior parte di questi disturbi e non è possibile affermare con certezza di quale patologia si tratti. Inoltre, gli esami clinici sono effettuati solo in seguito alla comparsa dei primi sintomi, rallentando in questo modo il processo di diagnosi e, non solo, compromettendo l'efficacia delle terapie rallentanti il processo di degenerazione della patologia.

Per poter efficacemente investigare i sintomi legati ai disturbi neurodegenerativi è necessario utilizzare degli indicatori biologici (*biomarker*), ossia dei composti biologici facilmente e chiaramente identificabili grazie ai quali, osservandone la variazione nelle concentrazioni a livello temporale, è possibile stabilire univocamente la presenza di una certa patologia. L'indicatore biologico può essere il DNA, l'RNA oppure una specifica proteina. Negli studi umani, l'indicatore biologico deve possedere, come caratteristiche principali, alta

identificabilità e manipolabilità. L'utilizzo di queste strategie richiede un'attenta analisi preliminare delle caratteristiche proprie dei composti candidati, infatti *biomarker* non ottimali possono potenzialmente condurre a errori nella diagnosi. Gli indicatori biologici sono utilizzati non unicamente per scopi diagnostici, ma anche per esaminare la risposta biologica in risposta a una terapia farmacologica. Alcuni di questi sono tuttavia visti in modo più coerente come segnali di rischio piuttosto che come veri indicatori di patologie, dato che, per le considerazioni precedenti, ciò che può indubitabilmente accertarne la presenza è una biopsia dei tessuti.

In assenza di *biomarker* adatti, per poter ottenere risultati attendibili, l'unica strategia su cui fondare le osservazioni si basa su studi effettuati su soggetti guariti e su dati clinici altamente generici ottenuti da un ampio campione di pazienti.

A partire da studi effettuati su culture cellulari, è stato possibile riscontrare che il comportamento dinamico anomalo dei microtubuli insieme al trasporto neuronale basato su questi ultimi può avere un ruolo fondamentale nella comparsa di disturbi neurologici. [30] Sebbene si sia riscontrato questo legame tra la disfunzione dei microtubuli e la comparsa di tali patologie, non è possibile determinare la validità di queste affermazioni anche nell'uomo, poiché le analisi sui microtubuli necessitano inevitabilmente di effettuare un prelievo tissutale a livello cerebrale. Tuttavia, grazie all'impiego del *labeling* metabolico per mezzo di acqua deuterata, è stato possibile quantificare l'efficienza del trasporto attraverso gli assoni presenti nel sistema nervoso centrale nel caso di animali vivi e di esseri umani. [30] Il campione analizzato è stato ottenuto attraverso un prelievo di fluido cerebrospinale dopo aver effettuato la marcatura metabolica delle proteine coinvolte in questi processi.

Dalle precedenti considerazioni, risulta evidente il motivo per cui si rivolge maggiormente l'attenzione alle anomalie nel trasporto neuronale in relazione alla comparsa di patologie neurodegenerative come potenziale causa nell'Alzheimer, nel morbo di Parkinson, nella malattia di Huntington e nella sclerosi amiotrofica laterale. [30]

Per poter convalidare le considerazioni tratte dall'osservazione dei *biomarker* è necessario effettuare degli studi che coinvolgono un campione molto ampio di soggetti i quali devono possedere delle caratteristiche tra loro simili come l'età, il sesso, lo stile di vita e anche il livello di istruzione, legato principalmente alla dinamicità del cervello correlata con l'attività di studio, in modo tale da poter condurre analisi allo scopo di verificare se effettivamente le molecole candidate come indicatori biologici presentino comportamenti differenti nei vari soggetti in relazioni alle condizioni cliniche o meno, ma questo processo di validazione richiede un lungo lavoro poiché i dati raccolti sono molti e non è possibile trattarli velocemente. [31]

4.2. Malattia di Alzheimer

L'Alzheimer è una patologia neurologica degenerativa la cui sintomatologia è caratterizzata da perdita di memoria e declino cognitivo. Esso rappresenta il disturbo neurodegenerativo più comune al mondo nonché la forma di demenza che si osserva con la frequenza più alta, assumendo valori che oscillano tra il 60% e 80% dei casi. [32] La maggior parte dei casi diagnosticati si presenta dopo i sessant'anni e in meno del 2,5% delle diagnosi si possono riscontrare disposizioni genetiche per la comparsa di questo disturbo. Si è stimato che nell'anno 2050, circa ottanta milioni di persone in tutto il mondo soffriranno di questa malattia, è perciò di fondamentale importanza effettuare diagnosi tempestive e sviluppare terapie adatte a rallentarne la progressione. [31] L'unica via attualmente percorribile per poter raggiungere questi obiettivi si basa sulla ricerca di *biomarker* affidabili e stabili, i quali permettano di condurre diagnosi precise e veloci.

La strategia ritenuta più valida prevede l'utilizzo di più proteine la cui espressione combinata funge da indicatore della patologia. Infatti, affinché l'impiego di indicatori biologici risulti efficace nel processo di diagnosi, le proteine devono presentare caratteristiche di elevata stabilità e, purtroppo, è raro incontrare composti singoli che possiedano queste caratteristiche, affidandosi a un insieme di proteine in relazione tra loro, i risultati ricavabili assumono maggiore attendibilità.

L'aspetto che differenzia l'Alzheimer da altre patologie con analoga sintomatologia è la presenza di placche costituite dalla proteina amiloide e di grovigli neurofibrillari e da questo

Biomarker		**Valori di controllo (pg/ml)**	**Diagnosi di Alzheimer (pg/ml)**
Aβ(1-42)	Si deposita in placche a livello extracellulare.	749 ± 20	< 500
Total Tau	Si osserva l'inclusione intraneuronale dei microtuboli ad essa associati.	$136 \pm 89\ (21 - 50\ anni)$ $243 \pm 127\ (51 - 70\ anni)$ $341 \pm 171\ (> 71\ anni)$	> 450 > 600
Phospho-tau-181	Provoca la degenerazione delle normali funzioni e il malfunzionamento del trasporto assonale.	23 ± 2	> 60

Tabella 2. Biomarker del fluido cerebrospinale riconosciuti internazionalmente nella diagnosi di Alzheimer, (i dati sono stati ottenuti avvalendosi dei kit Innogenetics single 96-well ELISA), tratto da [31].

	Aβ(1-42)	Total tau	Phospho-tau-181
Morbo di Alzheimer	↓	↑	↑
Malattia di Parkinson	↔	↔	↔
Invecchiamento	↔	↔	↔

Tabella 3. Cambiamenti nei livelli dei marcatori nel fluido cerebrospinale in presenza di morbo di Alzheimer, malattia di Parkinson e normale invecchiamento (↔ nessuna variazione, ↑ aumento, ↓ diminuzione), tratto da [31].

deriva il fatto che la diagnosi certa basata sull'osservazione di queste conformazioni può essere condotta solo *post mortem*. Nonostante la presenza di placche sia un tratto distintivo legato alla manifestazione del disturbo, in alcuni studi in cui venivano coinvolti individui in età avanzata, si poteva osservare l'accrescimento di tali strutture, sebbene non si assistesse alla perdita delle funzioni cognitive, fatto che ha indotto gli scienziati a ipotizzare che la diagnosi di questa malattia fosse in realtà molto più complessa.

Grazie all'impiego della marcatura metabolica è inoltre possibile monitorare vari stadi della malattia, in modo tale da chiarire quali siano le tappe della degenerazione progressiva. È ormai diffusamente noto che i sintomi di questa malattia si manifestino solo in seguito ad anni di accumulo della proteina amiloide e l'unica strategia per prevenire e diagnosticare tempestivamente la presenza dell'Alzheimer è basata sull'analisi del comportamento mostrato dagli indicatori biologici specifici effettuata prima di assistere alla comparsa dei sintomi collegati.

La ricerca di indicatori biologici ottimi si effettua più facilmente a partire dall'analisi delle proteine contenute nel fluido cerebrospinale, poiché rispetto al plasma sanguigno, questo fluido comunica direttamente con l'encefalo e, di conseguenza, con i composti secreti in questa sede. Per prelevare tale fluido è necessario effettuare una puntura lombare, procedura altamente invasiva e la quale comporta dei rischi per la salute del paziente, inoltre, date le sue caratteristiche, non è possibile adottare questa strategia per condurre attività di *screening* oppure in caso di studi longitudinali, i quali necessitano di prelievi frequenti, non indicati a causa dei rischi insiti nella procedura e per i quali si preferisce l'utilizzo di prelievi di sangue, più facilmente prelevabile.

La ricerca di *biomarker* nel plasma è un'attività ancora limitata, poiché in esso sono contenute solamente piccole quantità di molecole individuate come possibili indicatori delle patologie neurodegenerative. La matrice sanguigna è una struttura molto articolata e isolare una certa tipologia di proteine richiede un approfondito lavoro di analisi e ricerche con strumenti altamente sensibili capaci di identificare anche la limitata presenza di questi composti in un

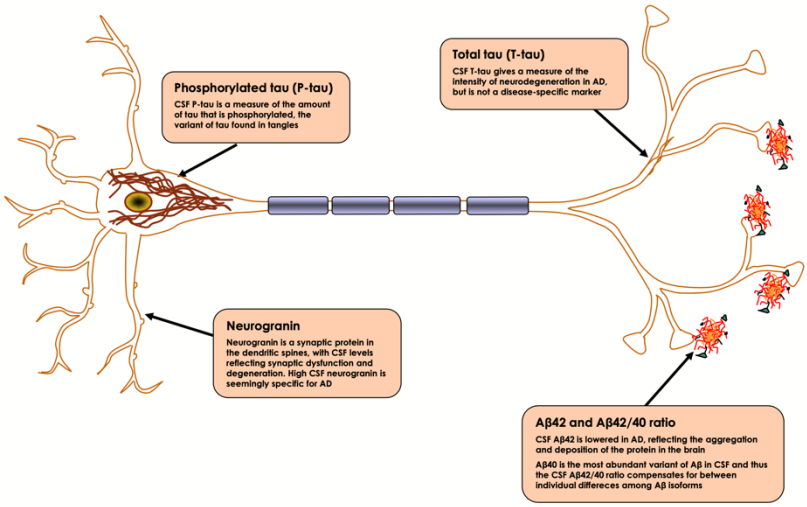

Figura 14. Diagramma schematico di un neurone con i grovigli neurofibrillari intracellulari e con le placche amioloidi neuritiche extracellulari. Sono indicati i biomarcatori specifici per l'Alzheimer presenti nel fluido cerebrospinale e la neuragenina, un nuovo candidato come biomarcatore sinaptico, tratto da [33].

saggio. Infatti, nel plasma è presente solo una quantità minima di proteine riconducibili al sistema nervoso, mentre allo stesso tempo si osservano alte concentrazioni di altre tipologie di proteine le quali introducono inevitabilmente una fonte di interferenza in sede di analisi. Non bisogna dimenticare che tali proteine, prima di raggiungere il flusso sanguigno, hanno incontrato dei processi di degradazione dovuti all'intervento della proteasi, metabolizzata nel fegato oppure ripulita dai reni, aspetto che introduce ulteriori variazioni nella quantità di proteine osservabili e per cui non si ha nessun mezzo di controllo e gestione a causa delle quali non è possibile in alcun modo ricavare le concentrazioni e le caratteristiche effettive possedute originariamente. [33]

4.3. Morbo di Parkinson

Gli studi clinici riguardanti il morbo di Parkinson hanno avuto uno sviluppo successivo a quelli condotti sull'Alzheimer e, in parte, hanno seguito un percorso affine a questi. La necessità di trovare nuove metodologie di diagnosi nel caso del morbo di Parkinson è legata all'impossibilità di fornire diagnosi prima della comparsa dei sintomi. A differenza dell'Alzheimer, il morbo di Parkinson colpisce soggetti di qualsiasi età e le terapie farmacologiche utilizzabili per rallentare il decorrere della malattia hanno funzionalità migliori nelle prime fasi della comparsa, quando purtroppo è tuttora difficile stabilirne la presenza.

Il morbo di Parkinson interessa i gangli basali costituenti del sistema extrapiramidale, ossia l'insieme dei centri nervosi che controllano e regolano la postura e i movimenti volontari e involontari. Questa malattia è contraddistinta da rigidità muscolare manifestata come resistenza ai movimenti passivi, insorgenza di tremore durante lo stato di riposo, i quali degenerano in seguito come disturbi dell'equilibrio, andatura impacciata e postura curva. Oltre ai sintomi precedentemente citati si può osservare anche depressione e lentezza nel parlare. [34]

A partire da alcuni studi effettuati su popolazioni di ratti nell'ambito della ricerca riguardante le malattie neurodegenerative è stato osservato che il tasso di comparsa di alcune proteine deuterate era ritardato proporzionalmente allo stato di degradazione dei microtubuli. Questa relazione ha condotto i ricercatori a ipotizzare l'esistenza di un legame tra la comparsa del disturbo e il danneggiamento dei microtubuli, anche nel caso umano, in quanto comportamenti affini sono stati osservati nel caso di pazienti affetti da Parkinson, per i quali si assisteva a un ritardo nella comparsa di proteine, analogamente allo studio sui ratti, dopo aver subito marcatura metabolica con acqua pesante. Da queste considerazioni è evidente l'importanza di concentrare questi studi nel caso umano poiché è ragionevole pensare che sussistano gli stessi legami riscontrabili in ambito animale. Prima di queste ricerche, le quali sono state condotte da Fanara, direttrice del dipartimento che si occupa dello studio sulle malattie neurodegenerative presso l'istituto privato KineMed, Inc in California, non vi era alcuna possibilità di collegare ragionevolmente le alterazioni subite dai composti prelevati dal fluido cerebrospinale o da altri tessuti con le disfunzioni dei microtubuli.

Per mezzo della marcatura metabolica effettuata con molecole di acqua arricchita è possibile identificare univocamente la presenza e la quantità di un dato composto presente nel fluido cerebrospinale, il quale è riconducibile a uno specifico processo nei modelli di analisi preclinica. Risulta perciò necessario che questi studi ottengano l'approvazione per l'applicazione in ambito umano cosicché sia possibile condurre l'identificazione di indicatori biologici, i quali permetterebbero di riconoscere i legami esistenti tra la comparsa del disturbo e la disfunzione dei microtubuli nel cervello. Tuttavia, i dati ricavabili attraverso questa strategia non godono di una certezza assoluta: se il paziente presenta danni a livello neuronale, il tasso di variazione nella concentrazione dei composti analizzati a partire del fluido cerebrospinale può essere alterata conseguentemente alla lesione, non permettendo di collegare tali alterazioni a una disgregazione dei microtubuli. Non di meno, se non si conoscono con precisione i valori dei tassi di formazione e di rimozione dei composti presenti

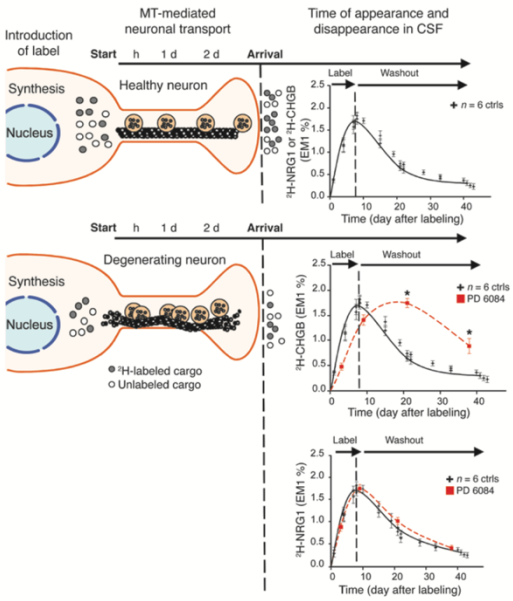

Figura 15. Modello e interpretazione cinetica della dinamica del trasporto delle proteine cargo rilasciate nel fluido cerebrospinale. La cinetica del trasporto veicolato dai microtubuli si basa sull'intervallo tra la comparsa e la scomparsa delle proteine cargo di nuova sintesi marcate con ^{2}H nel fluido cerebrospinale. Si osservano i diversi andamenti nella concentrazione di ^{2}H-CHGB e ^{2}H-NRG1: la curva continua si riferisce ai soggetti di controllo e la curva tratteggiata ai pazienti con Parkinson. In entrambi i casi, la concentrazione delle proteine è marcatamente inferiore nei pazienti affetti dal morbo rispetto ai pazienti di controllo, tratto da [30].

nel fluido cerebrospinale così come l'eventuale presenza di precursori in tale fluido, non è possibile ricavare dati certi.

Anche in questo caso, così come per l'Alzheimer, per poter comprendere l'evoluzione della patologia è indispensabile ricorrere all'uso di indicatori biologici, ricavabili grazie allo studio della presenza di tali composti in seguito al *labeling* metabolico, altrimenti impossibile da effettuare se si considerano unicamente le analisi statiche del proteoma. Questa caratteristica è importante per stabilire la presenza del morbo di Parkinson nelle primissime fasi e intervenire farmacologicamente per arrestarne la progressione e, allo stesso tempo, attraverso l'applicazione di queste metodologie, è possibile cercare nuovi strumenti per testare e sviluppare terapie sempre più efficienti. [35]

In una delle ricerche condotte da Fanara, è stato osservato che i pazienti sani hanno mostrato un ritardo minimale tra il momento in cui era stato completato l'arricchimento e quello in cui si registrava la presenza di proteine marcate nel fluido cerebrospinale, fatto che indica la presenza di un rapido transiente delle proteine di nuova sintesi dal sito di generazione al sito di raccolta dei campioni. D'altra parte, nel caso dei pazienti malati si è osservato un ritardo notevole in relazione al trasporto delle proteine marcate, il quale presentava uno schema di rilascio all'interno del fluido cerebrospinale relativamente prolungato, risultato imputabile a una disfunzione nel trasporto assonico. [30]

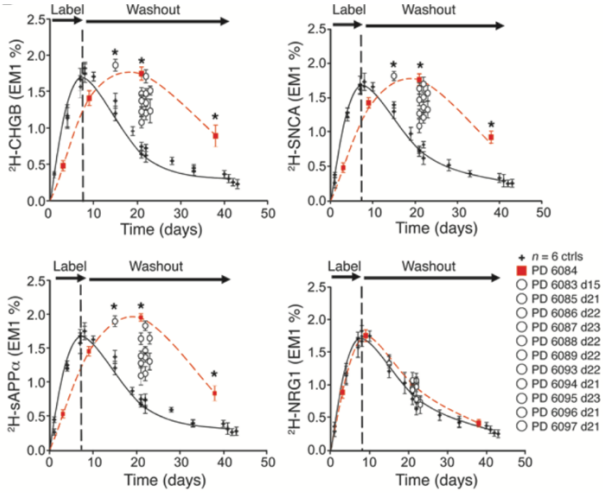

Figura 16. Confronto dei ritardi nei tassi di trasporto delle proteine cargo dei neuroni nel fluido cerebrospinale in soggetti affetti da Parkinson. Si osservano ritardi nel tempo di comparsa delle proteine ^{2}H-CHGB, ^{2}H-SNCA e ^{2}H-sAPPα ma non per ^{2}H-NRG1 in soggetti malati rispetto ai pazienti di controllo. La curva continua si riferisce ai dati di controllo, mentre la curva tratteggiata ai valori ricavati dai pazienti affetti da Parkinson, tratto da [30].

Sebbene tali risultati dimostrino la validità del processo di ricerca di *biomarker* e il loro potenziale impiego in ambito diagnostico, questi forniscono utili informazioni riguardanti unicamente la presenza o meno di una lesione neuronale ma, tuttavia, non è possibile ricavare dettagli riguardanti l'area dell'encefalo interessata, poiché tutto ciò che è desumibile riguarda unicamente il ritardo nel trasporto delle proteine lungo le terminazioni nervose. Per poter ricavare queste informazioni è necessario avvalersi di tecniche di *labeling* cinetico, in modo tale da poter ricavare dei valori integrati di trasporto del marcatore che differiscono non solo in termini temporali, ma anche spaziali.

Così come nel caso dell'Alzheimer, il fluido cerebrospinale rimane la fonte di indicatori biologici più ricca, tuttavia, per le stesse motivazioni citate in precedenza, non lo si può considerare come fluido di riferimento in ambito clinico e diagnostico. In assenza di *biomarker*, i metodi diagnostici utilizzati tradizionalmente per il morbo di Parkinson sono validi e conducibili solamente in seguito alla comparsa dei primi sintomi, quando purtroppo il processo degenerativo ha già raggiunto livelli avanzati. Inoltre, i metodi diagnostici effettuabili per mezzo di strumentazione medica come nel caso di PET e SPECT, rispettivamente Tomografia a Emissione di Positroni e Tomografia Computerizzata a Emissione Singola di Fotoni, sebbene siano utili per determinare la presenza di riduzione nella densità delle terminazioni nervose dopaminergiche nei gangli basali, non forniscono alcuna indicazione specifica riguardo la presenza del morbo di Parkinson, oltre a essere procedure molto costose e implicando la necessità di esporre il paziente a radiazioni.

Ancora una volta, le diagnosi più precise sono state condotte per mezzo della combinazione di indicatori biologici diversi. In questa fase di ricerca, risulta essere di fondamentale rilevanza la procedura di convalida dei dati ottenuti dai vari studi riguardanti il morbo di Parkinson cosicché si possano validare in modo definitivo le metodologie di ricerca derivate da queste ricerche, piuttosto che impiegare ulteriori sforzi nell'indagine di nuovi indicatori biologici. [36]

4.4. Impiego della marcatura con $^{2}H_2O$ nell'industria farmaceutica

Nei precedenti paragrafi è risultata chiara l'importanza nell'utilizzo delle procedure di *labeling* metabolico per la diagnosi di malattie e per gli studi clinici. Un altro campo in cui la marcatura metabolica per mezzo di acqua arricchita può essere impiegata con successo si riferisce allo sviluppo di nuovi farmaci più efficaci e con minori effetti collaterali.

A tal proposito, è possibile sviluppare dei medicinali che possiedono $^{2}H_2O$ in sostituzione alle molecole di acqua comune. Inizialmente, la deuterazione dei composti destinati alle terapie farmacologiche veniva impiegata al solo scopo di ricerca e studio riguardanti la farmacocinetica e gli effetti indotti nel paziente in seguito all'assunzione di tali farmaci. Più tardi, si osservò come questa pratica potesse trovare anche un impiego medicale e non rimanere esclusiva dell'ambito di sviluppo e ricerca, permettendo di produrre farmaci caratterizzati da particolari proprietà dovute alla presenza di atomi di deuterio ma con funzioni analoghe ai farmaci non deuterati.

Alcuni studi hanno permesso di avanzare ipotesi riguardanti la possibilità di trovarsi in presenza di farmaci con caratteristiche di degradazione e di efficacia migliori rispetto agli analoghi non deuterati. Infatti, la diversa cinetica di questi composti può avere effetti sulle interazioni tra le molecole deuterate e gli enzimi coinvolti nel metabolismo del farmaco.

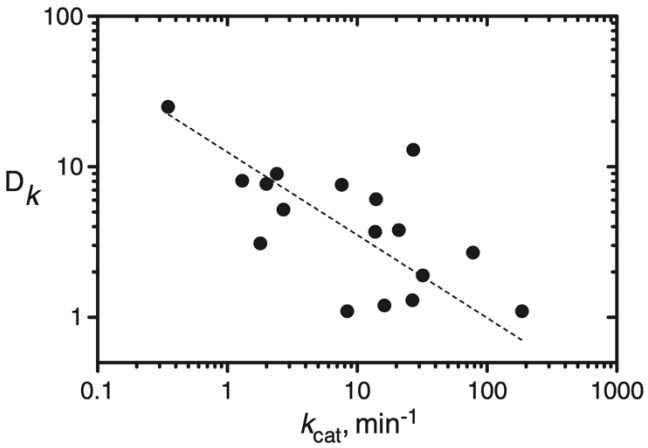

Figura 17. Relazione tra gli effetti della cinetica dell'isotopo (intramolecolare non competitiva) e della costante catalitica k_{cat} per un certo numero di reazioni del citocromo P450, enzima coinvolto in circa il 75% del metabolismo dei farmaci. La curva presenta un coefficiente di correlazione di 0,62, indicando che la maggioranza delle variazioni degli effetti cinetici dell'isotopo è inversamente legata alla costante catalitica, ossia in reazioni veloci, la rottura dei legami C-H è più facile, conseguentemente gli ulteriori processi nella catalisi sono limitati, tratto da [37].

Il primo farmaco di questo tipo che ha ottenuto l'approvazione da parte dell'organismo di vigilanza statunitense, ovvero *Food and Drug Administration* (FDA), è la deutetrabenazina, la versione deuterata della tetrabenazina, per la quale è previsto l'utilizzo nel trattamento della corea associata alla malattia di Huntington e per il trattamento della discinesia tardiva negli adulti, inoltre alcuni studi stanno cercando di provare la sua efficacia anche per il trattamento della sindrome di Tourette. Tuttavia, non è stata ancora condotta alcuna analisi comparativa che dimostri un aumento dell'efficacia del farmaco deuterato rispetto al tradizionale, le uniche analisi mettono in risalto temi quali la sicurezza per la salute del nuovo composto.

In generale, gli studi condotti sulla versione deuterata di alcuni farmaci rilevano come il metabolismo dei questi composti sia più lento, anche se, tuttavia, questa affermazione non è valida in assoluto, come nel caso della Paroxetina, farmaco utilizzato per il trattamento di stati depressivi e ansiosi, per il quale si assiste a una velocizzazione del metabolismo.

Per ora non è stata ancora dimostrata chiaramente la migliore efficacia dei farmaci deuterati, tuttavia, molte aziende farmaceutiche hanno iniziato a sviluppare questa nuova tipologia di medicinali. Se, per esempio, si considerano alcuni composti chirali, i quali a seconda della tipologia di isomero considerato possono mostrare diverse caratteristiche farmacologiche, di farmacocinetica e minori effetti indesiderati, la presenza di deuterio ne aumenta la stabilità, in quanto questa molecola riduce oppure, talvolta, elimina l'instabilità causata dall'attività ottica della molecola.

Per quanto riguarda il tema della tossicità dei composti deuterati ci si riferisce alle caratteristiche proprie dell'atomo di deuterio, il quale presenta, limitatamente a piccole dosi, una certa affidabilità. Tuttavia, sebbene limitati, gli effetti indesiderati conseguenti alla presenza del deuterio possono interagire e alterare i risultati dell'assunzione del farmaco, come per esempio, amplificare gli effetti collaterali dovuti al farmaco stesso oppure alterarne l'efficacia, determinando in questo modo una scarsa capacità di valutare i cambiamenti che derivano dall'assunzione del medicinale deuterato in relazione alla tipologia tradizionale. In questa fase bisogna sottolineare che la deuterazione non porta sempre e necessariamente migliorie in merito alle prestazioni del farmaco.

Limitatamente al fattore economico, sono state avanzate preoccupazioni riguardanti il possibile aumento dei prezzi dei farmaci deuterati, tuttavia, nonostante siano riscontrabili dei costi aggiuntivi dovuti alla produzione dell'isotopo, la relativamente piccola quantità di deuterio utilizzata nella sintesi del farmaco, non introduce aumenti significativi nel costo finale. Inoltre, è plausibile che, date le caratteristiche di maggiore durata nell'efficacia di questi farmaci, sia possibile ridurre la frequenza di assunzione, anche se, tuttavia, questa possibilità probabilmente è limitata a un numero ristretto di farmaci deuterati. [3]

CONCLUSIONI

In molteplici casi riscontrabili nella storia dell'umanità, le idee e le tesi degli scienziati non sono state applicate o verificate se non molti anni dopo la loro formulazione a causa del ritardo tecnologico che limitava lo sviluppo e la creazione di strumenti adeguati a condurre le ricerche necessarie.

Anche nel caso ivi discusso, è possibile concludere che le difficoltà nella ricerca di biomarcatori sono derivate principalmente dalle limitazioni dovute a strumenti di analisi inefficienti. Si consideri, come esempio, il caso sopra citato dei marcatori radioattivi, sebbene questi assolvessero alle loro funzioni e fossero rilevati dalla strumentazione, il loro utilizzo era limitato, data la loro natura chimica che non consente un'esposizione prolungata. Solo con la comparsa di strumenti più precisi e accurati, è stato possibile considerare l'uso di molecole sicure per la salute permettendo in questo modo di ampliare l'utilizzo di tale metodologia di analisi. Infatti, nonostante i ricercatori intuissero l'importanza e l'urgenza di estendere la conoscenza del proteoma umano a fini medici, solo in seguito all'avvento di tecnologie più precise e performanti è stato possibile condurre i primi passi di una ricerca che aspira a migliorare sensibilmente le condizioni di vita dei malati.

Benché questa metodologia sia ancora in fase iniziale, è possibile intuirne la rilevanza in ambito diagnostico e nelle attività di *screening*, in particolare per le patologie neurologiche, che affliggono un sempre maggior numero di individui nel mondo e per le quali non esistono mezzi affidabili per l'individuazione precoce, impedendone il trattamento nei primi stadi di comparsa della malattia.

Oggi più che mai, è necessario concentrare gli sforzi della comunità scientifica in tutti i campi al fine di sviluppare e ampliare le tecnologie esistenti affinché in futuro sia possibile affrontare le esigenze e le richieste di una società in continua evoluzione. Questo è appunto il caso osservato per il morbo di Parkinson e per la malattia di Alzheimer trattati in precedenza, per i quali si sta notando un rilevante incremento nelle diagnosi conseguente all'innalzamento dell'età media della popolazione, specialmente nei Paesi Occidentali. In relazione all'evoluzione della struttura sociale, variano le richieste e i quesiti a cui la scienza è chiamata a rispondere e la tecnologia deve fornire i mezzi adatti a questo scopo.

Gli obiettivi di sviluppo tecnico e scientifico possono essere raggiunti solamente grazie alla collaborazione e alla cooperazione dei ricercatori a livello globale, in modo tale che le informazioni raccolte e le conoscenze acquisite possano essere la base per ulteriori progressi e

non il punto di arrivo. A tal proposito, si sottolinea la rilevanza dell'avvento del *World Wide Web*: la vastissima diffusione di questo strumento ha permesso la condivisione dei dati e degli strumenti informatici a livello globale, permettendo lo scambio in tempo reale di informazioni e la collaborazione tra gli scienziati.

Oggi è indispensabile sfruttare appieno le potenzialità di questi strumenti, le quali sono tuttavia limitate a causa dell'enorme quantità di dati non catalogati, che compromettono la qualità delle ricerche, come pure di *software* di recente sviluppo e scarsa diffusione, che rendono difficile ripetere e convalidare le analisi precedenti.

Grazie all'adozione di protocolli internazionali per le analisi bioinformatiche sarà possibile migliorare le ricerche e le strategie di analisi permettendo allo studio di biomarcatori e ai relativi ambiti di applicazione di diffondersi su larga scala.

Bibliografia

[1] R. G. Sadygov, «High-Resolution Mass Spectrometry for In Vivo Proteome Dynamics using Heavy Water Metaboli Labeling,» *International Journal of Molecular Sciences,* vol. 21, n. 21, October 2020.

[2] A. M. Helmenstine, «What Is Heavy Water?,» 16 February 2021. [Online]. Available: http://www.thoughtco.com/what-is-heavy-water-609412. [Consultato il giorno 30 ottobre 2021].

[3] E. M. Russak e E. M. Bednarczyk, «Impact of Deuterium Substitution on the Pharmacokinetics of Pharmaceuticals,» *Annals of Pharmacotherapy,* vol. 53, n. 2, pp. 211-216, February 2019.

[4] A. M. Helmenstine, «Water Properties and Facts You Should Know,» 28 August 2020. [Online]. Available: http://www.thoughtco.com/water-chemistry-facts-and-properties-609401. [Consultato il giorno 30 ottobre 2021].

[5] A. M. Helmestine, «Heavy Water Facts,» 27 August 2020. [Online]. Available: http://thoughtco.com/properties-of-heavy-water-609397. [Consultato il giorno 30 ottobre 2021].

[6] S. Ilchenko, A. Haddad, P. Sadana, F. A. Recchia, R. G. Sadygov e T. Kasumov, «Calculation of the Protein Turnover Rate Using the Number of Incorporated 2H Atoms and Proteomics Analysis of a singlle Labeled Sample,» *Analytical Chemistry,* vol. 91 , n. 22, pp. 14340-14351, 2019.

[7] R. G. Sadygov, «Using Heavy Mass Isotopomer for Protein Turnover in Heavy Water Metabolic Labeling,» *Journal of Proteome Research,* vol. 20 , n. 4, pp. 2035-2041, 2021.

[8] W. E. Holmes, T. E. Angel, K. W. Li e M. K. Hellerstein, «Dynamic Proteomics: In Vivo Proteome-Wide Measurement of Protein Kinetics Using Metabolic Labeling,» *Methods in Enzymology,* vol. 561, pp. 219-276, 2015.

[9] A. Borzou, V. R. Sadygov, W. Zhang e R. G. Sadygov, «Proteome Dynamics from Heavy Water Metabolic Labeling and Peptide Tandem Mass Spectrometry,» *International Journal of Mass Spectrometry,* vol. 445, 2019.

[10] K. A. Cupp-Sutton e S. Wu, «High-throughput Quantitative Top-Down Proteomics,» *Molecular Omics,* vol. 16, pp. 91-99, 2020.

[11] R. G. Sadygov, J. Ayya, M. Rahman, K. Lee, S. Ilchenko, T. Kasumov e A. Borzou, «d2ome, Software for in Vivo Protein Turnove Analysis Using Heavy Water Labeling and LC-MS, Reveals Alteration of Hepatic Proteome Dynamics in a Mouse Model of NAFLD,» *Journal of Proteome Reasearch,* vol. 17 , n. 11, pp. 3740-3748, 2018.

[12] R. J. Beynon e J. M. Pratt, «Metabolic Labeling of Proteins for Proteomics,» *Molecular & Cellular Proteomics,* vol. 4 , n. 7, pp. 857-872, 2005.

[13] R. Aebersold e M. Mann, «Mass Spectrometry-Based Proteomics,» *Nature,* vol. 422, 2003.

[14] Agilent Technologies, *Mass Spectrometry Fundamentals - Theory,* 2016.

[15] J. R. Yates, C. I. Ruse e A. Nakorchevsky, «Proteomics by Mass Spectrometry: Approches, Advances and Applications,» *Annual Review of Biomedical Engineering,* 2009.

[16] R. G. Sadygov, «Partial Isotope Profiles are Sufficient for Protein Turnover Analysis Using Closed-Form Equations of Mass Isotopomer Dynamics,» *Analytical Chemistry,* vol. 92 , n. 21, pp. 14747-14753, 2020.

[17] J. C. Price, W. E. Holmes, K. W. Li, N. A. Floreani, R. A. Neese, S. M. Turner e M. K. Hellerstein, «Measurement of Human Plasma Proteome Dynamics with 2H2O and Liquid Chromatography Tandem Mass Spectrometry,» *Analytical Biochemistry,* vol. 420, pp. 73-83, 2012.

[18] J. Gauthier, A. T. Vincent, S. J. Charette e N. Derome, «A Brief History of Bioinformatics,» *Briefings in Bioinformatics,* vol. 20, n. 6, pp. 1981-1996, 2019.

[19] C. Chen, J. Hou, J. J. Tanner e J. Cheng, «Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis,» *International Journal of Molecular Sciences,* vol. 21, n. 8, 2020.

[20] J. Bai, C. Bandla, J. Guo, R. V. Alvarez, M. Bai, J. A. Vizcaìno, P. Moreno, GrU, B. Gruning, O. Sallou e Y. Perez-Riverol, «BioContainers Registry: Searching Bioinformatics and Proteomics Tools, Packages and Containers,» *Journal of Proteome Research,* vol. 20, n. 4, pp. 2056-2061, 2021.

[21] R. C. Jiménez, M. Kuzak, M. Alhamdoosh e altri, «Four Simple Recommendations to Encurage Best Practices in Research Software,» *F100 Reasearch,* 2017.

[22] «Global Alliance for Genomic & Health,» 2021. [Online]. Available: https://www.ga4gh.org/about-us/. [Consultato il giorno 6 dicembre 2021].

[23] J. Pereira e V. Alva, «How Do I get the Most Out of my Protein Sequence Using Bioinformatics Tools?,» *Acta Crystallographica,* vol. 77, pp. 1116-1126, 2021.

[24] V. Schwammle, J. Harrow e H. Ienasescu, «Proteomics Software in bio.tools: Coverage and Annotations,» *Journal of Proteome Research,* vol. 20, n. 4, pp. 1821-1825, 2021.

[25] Z. Noor, S. Beom Ahn, M. S. Baker, S. Ranganathan e A. Mohamedali, «Mass Spectrometry-BAsed Protein Identification in Proteomics-A Review,» *Briefings in Bioinformatics,* vol. 22, n. 2, pp. 1620-1638, 2021.

[26] J. G. Meyer, «Deep Learning neural network tools for proteomics,» *Cell Reports Methods,* vol. 1, 2021.

[27] B. Wen, W.-F. Zeng, Y. Liao, Z. Shi, S. R. Savage, W. Jiang e B. Zhang, «Deep Learning in Proteomics,» *Proteomics,* n. 20, 2020.

[28] M. Q. Ding, L. Chen, G. F. Cooper, J. D. Young e X. Lu, «Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics,» *Molecular Cancer Research,* pp. 269-278, 2018.

[29] Y. Liu, T. Vijlder, W. Bittermieux, K. Laukens e W. Heyndrickx, «Current and Future Deep Learning Algorithms for Tandem Mass Spectrometry (MS/MS)-Base Small Molecules Stucture Elucidation,» *Rapid Communnications in Mass Spectrometry,* 2021.

[30] P. Fanara, P.-Y. A. Wong, K. H. Husted, S. Liu, V. M. Liu, L. A. Kohlstaedt, T. Riiff, J. C. Protasio, D. Boban, S. Killion, M. Killian, L. Epling, E. Sinclair, J. Peterson, R. W. Price, D. E. Cabin, R. L. Nussbaum, J. Bruhmann, R. Brandt, C. W. Christine, M. J. Aminoff e M. K. Hellerstein, «Cerebrospinal Fluid-Based Kinetic Biomarkers of Axonal Transport in Monitoring Neurodegeneration,» *Journal of Clinical Investigation,* vol. 122, n. 9, pp. 3159-3169, 2012.

[31] C. Humpel, «Identifying and Validating Biomarkers for Alzheimer's Disease,» *Trend in Biotechnology,* vol. 29, n. 1, pp. 26-32, 2011.

[32] T. J. Hark e J. N. Savas, «Using Stable Isotope Labeling to Advance our Understanding of Alzheimer's Diases Etiology and Pathology,» *Journal of Neurochemistry,* vol. 159, n. 2, pp. 318-329, 2021.

[33] K. Blennow e H. Zetterberg, «Biomarkers for Alzheimer's Disease: Current Status and Prospects for the Future,» *Journal of Internal Medicine,* vol. 284, n. 6, pp. 643-663, 2018.

[34] Istituto Superiore di Sanità, «Malattia di Parkinson-Informazioni Generali,» 19 September 2013. [Online]. Available: https://www.epicentro.iss.it/parkinson/. [Consultato il giorno 7 dicembre 2021].

[35] W. Z. Potter, «Mining the Secrets of the CSF: Developing Biomarkers of Neurodegeneration,» *Journal of Clinical Investigation,* vol. 122, n. 9, pp. 3051-3053, 2012.

[36] L. Parnetti, L. Gaetani, P. Eusebi, S. Paciotti, O. Hansson, O. El-Agnaf, B. Mollenhauer, K. Blennow e P. Calabresi, «CSF and Blood Biomarkers for Parkinson's Disease,» *The Lancet Neurology,* vol. 18, n. 6, pp. 573-586, 2019.

[37] F. P. Guengerich, «Kinetic deuterium isotope effects in cytochrome P450 oxidation reactions,» vol. 56, pp. 428-431, 2013.