



UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Farmacia

Corso di Laurea in Farmacia

Tesi sperimentale in

Chimica generale ed inorganica

Identificazione di un nuovo modello di attività di peptidi antimicrobici attraverso machine learning

Relatore:

Ch.mo Prof. Stefano Piotto

Candidato:

Mariaclaudia Nicolai

matr. 0760100911

Anno Accademico 2018/2019

Indice

Anno Accademico 2018/2019	1
Abstract	3
1. Introduzione	4
1.1. Obiettivo.....	5
1.1.1. <i>Peptidi antimicrobici (storia, struttura e attività)</i>	7
1.2. Stato dell'arte.....	10
1.3. Metodi computazionali	11
2. Metodi	12
2.1. Database	12
2.2. <i>Elicità</i>	12
2.3. Algoritmi GMDH.....	13
2.3.1. <i>Modello di regressione</i>	14
2.4. Elementi di statistica	16
2.5.1. <i>Modello di validazione</i>	16
2.5.2. <i>Confusion Matrix</i>	18
3. Risultati	21
3.1. Selezione del modello.....	21
3.1.1. <i>Analisi Subset</i>	24
3.1.2. <i>Analisi subset Split</i>	33
3.1.3. <i>Analisi subset Overlap</i>	37
4. Conclusioni	43
Bibliografia.....	45

Abstract

Il crescente uso degli antibiotici ha portato all'incremento delle specie batteriche resistenti ai trattamenti farmacologici. Le terapie multi-farmaco per il trattamento di specie multi-farmaco resistenti risultano spesso rischiose e costose. La ridotta disponibilità di nuovi antibiotici richiama l'attenzione sulla necessità di trovare delle nuove cure. Recenti studi hanno proposto l'uso dei peptidi antimicrobici (AMP) la cui progettazione sfrutta l'introduzione di tecniche computazionali nella ricerca farmacologica.

L'obiettivo di questo studio è l'identificazione dei modelli di predizione di attività degli AMP attraverso l'uso di tecniche di machine learning. Questa ricerca si avvale degli algoritmi di apprendimento forniti dal software commerciale GMDH (Group Method of Data Handling) scegliendo di applicare l'analisi di regressione ai tre dataset contenenti descrittori molecolari delle specie batteriche: *E. coli*, *P. aeruginosa* e *S. aureus*. I modelli risultanti sono stati valutati sulla base di due osservazioni per accertarne l'accuratezza. Sono stati considerati tre elementi fondamentali per la verifica dell'adeguatezza dei dataset quali l'analisi dei residui, della Confusion Matrix e del coefficiente di determinazione (R^2). È stato eseguito un confronto con modelli già esistenti in letteratura ottenuti con diverse tecniche di predizione.

Dai risultati dello studio si deduce che, con opportuni accorgimenti, gli algoritmi di machine learning permettono di ottenere modelli accettabili per la predizione dell'attività dei peptidi antimicrobici.

1. Introduzione

La scoperta degli antibiotici ha avuto effetti sul trattamento delle malattie infettive e sulla società cambiando l'atteggiamento verso il processo patologico e la morte [1]. Introdotti nel 1911, i primi rimedi antimicrobici hanno dato avvio alla scoperta di farmaci sempre più nuovi fornendo ai medici un'ampia scelta di trattamenti per le cure cliniche. In questo modo alcune delle malattie precedentemente pericolose hanno trovato cura. Durante il XIX secolo, la principale causa di morte era costituita dalla polmonite, dalla diarrea e dalla difterite. Al tempo della rivoluzione industriale e dell'urbanizzazione tale elenco si espande aggiungendo malattie come tubercolosi e sifilide. In quel periodo viene registrato un aumento dell'incidenza delle malattie come conseguenza della migrazione della popolazione verso le città. Nel 1867, l'introduzione e l'uso degli antisettici in ambito ospedaliero e post-operatorio, per ridurre infezioni causate da batteri Gram-negativi, non è bastato a limitare i decessi. Bisogna aspettare l'arrivo della chemioterapia antimicrobica per ridurre il tasso delle infezioni chirurgiche dal 40% al 2%. Il progresso degli antibiotici è un percorso affascinante che parte dalla necessità di trattare le infezioni più comuni causa di morte nel passato per arrivare al continuo accrescersi delle indagini scientifiche.

La recente ricerca scientifica ha scoperto che la resistenza antimicrobica sta diventando sempre più pericolosa [2]. Gli organismi resistenti sono in aumento e le aree geografiche colpite sono in espansione. Gli agenti patogeni, che erano sotto il controllo farmacologico, sono diventati resistenti ai trattamenti e questo è iniziato a verificarsi soprattutto negli ospedali dove inizialmente si era diffuso l'uso degli antibiotici. La frequenza della resistenza aumenta in molti differenti batteri a causa del crescente uso di antibiotici, specialmente nei

paesi in via di sviluppo dove gli antibiotici sono facilmente disponibili senza prescrizioni. Le scarse condizioni igieniche ne aiutano la diffusione e il budget limitato per la cura della propria salute impedisce l'accesso a nuovi efficaci ma molto più costosi antibiotici. La gravità e la difficoltà nel trattare specie multi-farmaco resistenti (*multidrug resistance*, MDR) richiede l'uso contemporaneo di più farmaci, in certe circostanze anche sei o sette. Le terapie multi-farmaco risultano rischiose e costose e possono fallire soprattutto nei paesi in via di sviluppo. Come si potrebbe gestire la resistenza ai farmaci e come si potrebbe prevenire la sua diffusione? Una prima ipotesi è quella di segnalare la frequenza dei casi di resistenza a livello locale, nazionale e globale. Viene istituito un sistema di vigilanza della sensibilità e reattività ai farmaci che si attiva in caso la resistenza al farmaco si manifesta in un luogo. Il sistema permette di agevolare la scelta del trattamento più appropriato e di tenere sotto controllo i nuovi agenti patogeni. Una seconda teoria è di isolare in ospedale i pazienti colpiti da infezioni batteriche resistenti ai farmaci e potenzialmente pericolose perché causate da agenti patogeni difficili da trattare. La probabilità di diffusione dell'infezione risulta ridotta secondo alcuni studi. Una terza idea è di avvalersi di nuovi approcci terapeutici. I farmaci finora disponibili sono quelli ancora da preferire rispetto ai pochi nuovi sviluppati e perciò devono essere usati con prudenza, limitandone l'uso nelle terapie per avere ancora effetto. Lo sviluppo di nuovi antibiotici è, tuttavia, fondamentale. È necessario che i nuovi farmaci non agiscano con l'attuale meccanismo d'azione che ne impedirebbe il successo ma che attacchino nuovi target. Negli ultimi anni la produzione di nuovi farmaci antibiotici è calata sia a causa della ridotta ricerca nel campo da parte delle grandi aziende farmaceutiche che dalla insufficiente disponibilità finanziaria delle start-up che rivolgono la loro ricerca al problema della resistenza antibiotica.

1.1. Obiettivo

La resistenza antibatterica e la disponibilità di nuovi farmaci è un problema che deve essere affrontato. Malgrado i tentativi per la scoperta di nuovi farmaci, nessuna nuova classe di antibiotici è stata sviluppata negli ultimi decenni, per lo più a causa di severi requisiti chimici, biologici e farmacologici per ottenere farmaci antibiotici efficaci. Tra le possibili soluzioni al problema, Tyers e Wright hanno proposto l'uso di una combinazione di antibiotici distinguendo in combinazione congruente, combinazione sincretica e combinazione sinergica [3]. La combinazione congruente è attuata da molte terapie e prevede l'uso di antibiotici che inibiscono la crescita cellulare nei confronti di un target. La combinazione sincretica, al contrario, include che venga usato almeno un composto senza attività antibiotica. Infine, la combinazione sinergica comporta l'uso di composti che usati da soli non avrebbero attività antibiotica perciò vengono usati insieme in terapia. Secondo gli studiosi l'uso contemporaneo delle tre combinazioni rappresenta un grande punto di partenza per la scoperta e lo sviluppo di cure contro le infezioni nel ventunesimo secolo. Le difficoltà di produzione di un singolo antibiotico riguardano il raggiungimento del livello terapeutico e la sua durata. La complessità di sviluppo aumenta quando si devono produrre farmaci che devono avere azione sinergica in quanto si deve tener conto della farmacocinetica e della farmacodinamica di entrambi.

Brown e Wright sostengono che le scoperte storiche sul funzionamento degli antibiotici e delle cellule dei microrganismi sono la base degli studi necessari per affrontare il periodo della resistenza antibiotica [4]. I primi agenti antimicrobici sono stati scoperti analizzando una raccolta di sostanze chimiche trattate con coloranti da cui sono ottenute delle molecole sintetiche. Successivamente, si è concretizzata l'ipotesi che i batteri in uno specifico ambiente producono metaboliti capaci di agire contro le infezioni batteriche umane con buoni effetti e con minimi effetti collaterali. Tuttavia, quei metaboliti non sono da considerarsi strutture di nuovi farmaci disponibili perché sono il risultato di una evoluzione microbica.

L'epoca in cui si è avuto lo sviluppo di composti capaci di inibire la crescita microbica è seguita da anni di sospensione dell'identificazione di nuove molecole. La ripresa della ricerca di nuove molecole riparte quando la resistenza batterica colpisce anche quelle poche molecole che erano in uso. In questo periodo si fanno strada delle innovazioni in campo terapeutico e la tecnologia è la chiave del rinnovamento. I nuovi metodi permettono, ad esempio, di manipolare il DNA per produrre proteine desiderate, creando raccolte sempre più ricche di informazioni chimiche, e di individuare strutture proteiche che possono portare alla progettazione di nuovi farmaci. L'arrivo dell'informatica ha fornito nuovi metodi per la ricerca. È stato possibile così ampliare le raccolte dati a disposizione per lo studio di nuovi farmaci antibiotici. Czuplewski *et al.* suggeriscono di ampliare i metodi per il trattamento delle infezioni batteriche [5]. Le terapie in grado di sostituire gli antibiotici possono prevedere una combinazione di antibiotici e metodi profilattici.

Per Lazar *et al.* i peptidi antimicrobici (*Antimicrobial Peptides*, AMP) rappresentano una valida alternativa per fronteggiare la crescente resistenza agli antibiotici [6]. Gli AMP mostrano un ampio spettro d'azione che include batteri Gram-positivi e Gram-negativi, rapida insorgenza d'azione battericida, basso tasso di resistenza per il target a cui sono destinati, lenta comparsa di immunogenicità. Negli ultimi tempi hanno, infatti, suscitato molto interesse e sono diventati un vero e proprio settore di ricerca. Gli esperti hanno scoperto che batteri resistenti agli antibiotici mostrano, invece, sensibilità ai peptidi antimicrobici. Tale scoperta permette di identificare combinazioni di farmaci, antibiotico - peptide antimicrobico, capaci di migliorare l'attività antimicrobica contro batteri multi-resistenti rallentando il processo di nuova resistenza.

1.1.1. Peptidi antimicrobici (storia, struttura e attività)

I peptidi antimicrobici sono oligopeptidi con un numero di amminoacidi variabile (da 5 a 100) i cui target sono microrganismi, dai virus ai parassiti [7]. La loro scoperta, nel 1939, si deve a Dubos che estrasse un agente antimicrobico da una specie di *Bacillus* del terreno. In seguito, è stato dimostrato che tale agente proteggesse i topi dall'infezione da pneumococchi [8]. Hotchkiss e Dubos, analizzando ancora l'agente microbico, hanno identificato un AMP chiamato gramicidina, risultata efficace per il trattamento topico di ferite e ulcere [9]. L'AMP tirocidina, scoperto nel 1941, è risultato efficace sia contro batteri Gram-negativi che Gram-positivi ma tossico per le cellule del sangue umano [10]. Il primo AMP di origine animale, chiamato defensina, è stato isolato dai leucociti di coniglio nel 1956 [11]. Inoltre, è stato dimostrato che i leucociti umani contengono AMP nei loro lisosomi [12]. Ad oggi sono stati scoperti e isolati ben oltre 5.000 AMP di cui circa il 70% sono naturali e il 30% sono peptidi sintetici [13]. Si ritiene che i peptidi antimicrobici siano un antibiotico endogeno con l'obiettivo di uccidere i microbi. Si ritiene che i peptidi antimicrobici siano un antibiotico endogeno con l'obiettivo di uccidere i microbi.

La maggior parte dei peptidi antimicrobici conosciuti finora può essere descritta sulla base della struttura secondaria ovvero β -foglietto e α -elica [14]. A differenza degli antibiotici che hanno un target specifico (es. sintesi di DNA, proteine o parete cellulare), gli AMP mirano allo strato lipopolisaccaridico della membrana cellulare. Tuttavia, le cellule eucariotiche sono fuori dalla portata della loro attività a causa degli alti livelli di colesterolo e la bassa carica anionica. Una caratteristica da menzionare riguardo gli AMP è la loro rapida capacità di uccidere [15]. Alcuni uccidono pochi secondi dopo il contatto con le membrane, altri aumentano l'attività degli antibiotici mediante un effetto sinergico [16]. Nonostante i vantaggi degli AMP, ci sono ancora incertezze riguardo l'applicazione a causa della loro potenziale tossicità sugli umani, la mancanza di selettività contro specifiche specie, ed elevati costi di produzione [17]. La classificazione degli AMP si basava in passato sulla loro origine [18].

Tale classificazione permette di comprendere la correlazione che c'è tra la funzione di peptidi antimicrobici provenienti da un gruppo simile di animali e le condizioni di vita degli animali. Questa classificazione è superata dalla nuova che si basa su caratteristiche chimiche e biochimiche dei peptidi, sulla similitudine strutturale e funzionale. In questo modo si creano cinque classi:

1. Peptidi lineari, prevalentemente α -elica senza residui di cisteina, con o senza regione di cerniera
2. Peptidi antimicrobici con un legame disolfuro che formano struttura a loop con una coda
3. Peptidi antimicrobici con due o più legami disolfuro che hanno struttura β -foglietto
4. Peptidi lineari senza residui di cisteina e con una composizione insolita di amminoacidi regolari
5. Peptidi antimicrobici derivati da peptidi o proteine più grandi con altre funzioni note

Il meccanismo d'azione preciso dei peptidi antimicrobici, tuttavia, non è del tutto spiegato e si hanno ancora limitate informazioni sugli elementi che causano attività e selettività [18]. Per ricostruire il meccanismo sono state avanzate delle ipotesi. I peptidi antimicrobici uccidono le cellule distruggendo l'integrità della membrana (interagendo con membrane cariche negativamente) inibendo proteine, interferendo con sintesi di DNA e RNA, o interagendo con target intracellulari [19]. Gli AMP sono per la maggior parte delle sostanze cationiche (per la presenza dei residui di lisina e arginina) e interagiscono con membrane biologiche cariche negativamente, in particolare con il lipopolisaccaride ovvero con la membrana esterna dei batteri Gram-negativi. Per le caratteristiche intrinseche dei peptidi antimicrobici, i procarioti sono il target d'azione perché hanno membrane cellulari anioniche senza colesterolo [20]. Alcuni studi hanno evidenziato che gli AMP sfruttano la

permeabilizzazione della membrana cellulare batterica come meccanismo primario di uccisione. Infatti, AMP a concentrazioni elevate possono uccidere i microrganismi distruggendo la membrana per mezzo della formazione di canali e pori [21]. L'incorporazione dei peptidi nella membrana forma dei pori e destabilizza la sua struttura portando alla lisi della cellula batterica. Un AMP ha, generalmente, attività solo contro una classe di microrganismi (o batteri o funghi). Tuttavia, ci sono eccezioni e alcuni AMP hanno modi differenti di agire contro differenti tipi di microrganismi [22].

1.2. Stato dell'arte

Il riconoscimento di peptidi antimicrobici esistenti in natura e la progettazione di nuovi peptidi sintetici richiede l'investimento di tempo e denaro per lo studio di potenziali peptidi idonei [23]. L'innovazione in campo informatico fornisce nuove tecniche computazionali di predizione di attività antimicrobica e altre significative attività biologiche delle sequenze peptidiche. Per Gull *et al.* un metodo computazionale ideale si concentra sulla predizione della possibile attività biologica (antimicrobica, antibatterica, antivirale, ecc.) di nuove sequenze peptidiche. La tecnologia ha, inoltre, reso disponibili database di peptidi antimicrobici ricchi di informazioni riguardo differenti attività biologiche, sperimentalmente verificate, di peptidi naturali e sintetici. In contemporanea, si sviluppano differenti modelli predittivi di attività degli AMP.

Molti dei tool di predizione disponibili si occupano della predizione dell'attività antimicrobica dei peptidi. L'uso di tecniche computazionali è uno strumento utile per continuare la ricerca. I modelli creati con le tecniche di machine learning sono un'ipotesi per lo sviluppo di futuri nuovi metodi per contrastare la resistenza batterica. La ricerca porterebbe ad ottenere delle nuove sequenze peptidiche con potenziale attività

antimicrobica e farebbe ripartire le indagini sulle possibili alternative al superamento della crisi della resistenza agli antibiotici.

1.3. Metodi computazionali

I metodi tradizionali usati per analizzare l'attività dei peptidi sono costosi, richiedono tempo, impiegano manodopera e le industrie farmaceutiche, che investono in ricerca e sviluppo, riescono ad immettere sul mercato solo pochi farmaci ogni anno [24]. Rispetto alle molecole chimiche, sembra che i peptidi abbiano dei vantaggi nelle terapie come una minore tossicità ed effetti collaterali e più alta selettività. I peptidi sono impiegati nel drug discovery e per Wu *et al.* si deve far ricorso ad approcci *in silico* per lo sviluppo di strutture con potenziale attività terapeutica. Tra i sistemi computazionali, l'uso del machine learning (ML) ha facilitato la scoperta di peptidi attivi. Nel drug discovery le tecniche più utilizzate sono il support vector machine (SVM), random forest (RF), artificial neural network (ANN) e di recente è stato introdotto il deep learning (DL). La procedura di impiego del machine learning prevede quattro fondamentali step: la raccolta dei dati, la descrizione dei dati (*data labeling*), la costruzione del modello e la messa a punto di un modello di valutazione (Figura 1). Il ML è stato applicato in campo bioinformatico per la predizione di funzione e struttura dei peptidi, oltre che per la predizione della potenziale attività.

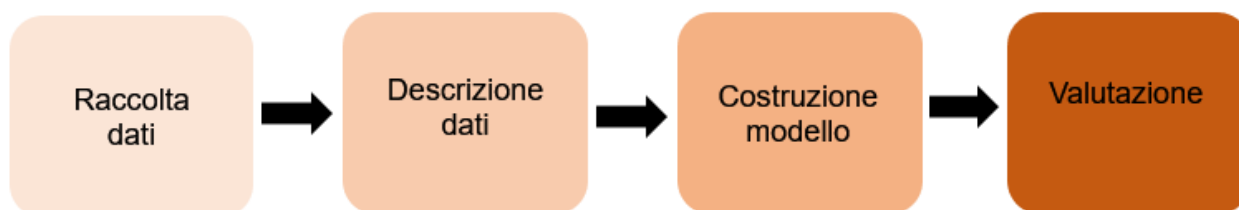


Figura 1 Approccio computazionale per la previsione dell'attività dei peptidi.

2. Metodi

2.1. Database

Per la mia indagine ho avuto bisogno di collezionare dati per costruire il dataset. Mi sono servita del database YADAMP (Yet another database of antimicrobial peptides) [25] (Figura 2), realizzato dal gruppo di ricerca del Prof. Piotto, presso il quale questo lavoro di tesi è stato svolto. YADAMP è stato costruito raggruppando e ripulendo dati presenti in database preesistenti di AMP e dati presenti in letteratura per ottenere informazioni utili su caratteristiche strutturali. Alcune proprietà chimico-fisiche sono state, invece, calcolate come la carica, l'elicità, la flessibilità, il punto isoelettrico, l'indice di Boman e l'indice di instabilità. Fondamentali, ai fini del progetto, sono i dati contenuti sull'attività antimicrobica dei peptidi. Per condurre un'analisi statistica sui peptidi si fa riferimento alla MIC (*Minimal Inhibitory Concentration*). In microbiologia il termine MIC indica la più bassa concentrazione di una sostanza antimicrobica capace di inibire la crescita di un batterio. Tra i principali organismi causa di infezione dell'uomo ritroviamo *Pseudomonas aeruginosa*, *Staphylococcus aureus* ed *Escherichia coli* dei quali in YADAMP sono riportate le MIC.



Figura 2 YADAMP logo

2.2. Elicità

Il termine elicITÀ fa riferimento alla tendenza di una proteina a formare l' α -elica. L' α -elica è parte della struttura secondaria delle proteine ed è dovuta alla formazione dei legami

idrogeno tra il gruppo N-H e il gruppo C=O della sequenza proteica. L'elicità ha un ruolo cruciale nella specificità e nella tossicità di peptidi antimicrobici [26]. Huang *et al.* hanno considerato proprio l'elicità per studiare le relazioni della struttura secondaria e della selettività di peptidi antimicrobici. Hanno scoperto che i peptidi con un alto valore di elicità manifestano una forte attività antimicrobica a dimostrazione dell'importanza che questa copre nel loro meccanismo d'azione contro le specie microbiche.

Nel mio studio ho considerato proprio questa proprietà per riorganizzare e rendere omogenei i dataset al fine di poter istruire la rete neurale nel modo migliore.

2.3. Algoritmi GMDH

Per creare i modelli di attività di peptidi antimicrobici ho utilizzato algoritmi di apprendimento (*learning algorithms*) GMDH (Group Method of Data Handling) ovvero una tecnica sviluppata dal Prof. A. G. Ivakhnenko nel 1968 [27]. Nello specifico, ho utilizzato l'applicazione GMDH Shell DS 3.8.9 (Figura 3) che permettere di ottenere un'analisi predittiva avanzata fornendo strumenti per la previsione di serie temporali, regressione, classificazione, clustering e fitting di curve. L'algoritmo di apprendimento da me utilizzato è il *GMDH-type neural network*. Le reti neurali di questo tipo, conosciute anche come reti neurali polinomiali, utilizzano l'algoritmo combinatorio per migliorare la connessione neuronale.



Figura 3 GMDH Shell logo

Secondo Muller *et al.* sia i metodi statistici che le reti neurali sono metodi deduttivi che richiedono una grande quantità di informazioni e non sono in grado di riconoscere oggetti complessi [28]. Invece, gli algoritmi GMDH possono essere considerati come un metodo di regressione che combina sia reti statistiche che neurali, risultando un metodo induttivo. Tuttavia, non esiste un metodo accettato ed utilizzato comunemente in quanto la scelta dipende dalla finalità della sua applicazione e bisogna considerare la correttezza dei dati.

2.3.1. Modello di regressione

GMDH Shell permette di scegliere il modello più appropriato per il proprio lavoro: nel mio caso ho scelto il *Regression template*. È una tecnica utilizzata per l'analisi di un dataset che ha una variabile dipendente e una o più variabili indipendenti e per valutare la relazioni tra essi. Il metodo di regressione permette di trovare in output la previsione di dati in input.

I passaggi che vengono eseguiti dal modello sono i seguenti (Figura 4):

1. Fornire un numero di osservazioni per valutare l'accuratezza della previsione;
2. Riordinare le righe del dataset selezionando una delle due scelte (*odd/even reordering*);
3. Usare la strategia di valutazione per selezionare i migliori modelli (*k-fold validation*);

4. Usare la RMSE (*Root Mean Square Error*) per la misura dell'errore durante la validazione;
5. Impiegare un generatore *Neural-type* per i modelli;
6. Applicare una funzione lineare, polinomiale o polinomiale quadratica
7. Modulare la selezione del numero dei modelli *top-ranked*

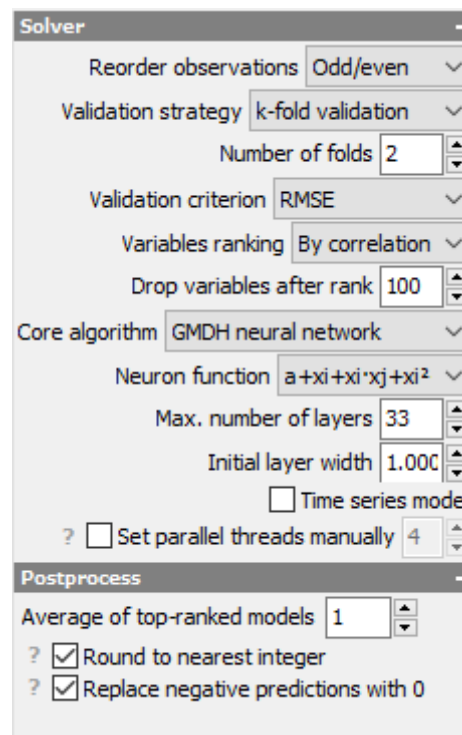


Figura 4 Pannello di controllo di GMDH

Per il mio studio ho selezionato la *2-fold validation* e l'RMSE per applicare l'algoritmo di regressione. Tuttavia, è bene ricordare che il modello è ottimizzato per la velocità piuttosto che per l'accuratezza della previsione.

L'analisi di regressione è uno tra gli strumenti di statistica più importanti ed è ampiamente utilizzato in vari campi scientifici [29]. La relazione tra i dati in analisi permette di formulare un'ipotesi utilizzata per sviluppare un'equazione di regressione. Seguono vari test per valutare se il modello è soddisfacente. La convalida del modello (*model validation*) è un

passaggio importante nella creazione del modello e aiuta a valutare l'affidabilità dei modelli prima che possano essere utilizzati nel processo decisionale.

2.4. Elementi di statistica

Durante il mio lavoro mi sono servita di alcune nozioni di statistica per validare i risultati ottenuti. Li tratto nel dettaglio nei paragrafi seguenti in cui li distinguo in modello di validazione e *Confusion Matrix*. Il modello di valutazione presenta più metodi usati per la verifica dell'accuratezza del risultato dell'analisi.

2.5.1. Modello di validazione

Quando applichiamo una tecnica di predizione per la nostra ricerca, ci chiediamo quanto sia buono il modello risultante [30]. Quando si parla della bontà del modello si fa riferimento a quanto il modello sia coerente con i dati. Il metodo *hold out* è la tecnica più semplice di validazione del modello. Il dataset viene diviso in due parti di cui una rappresenta il *training set* e l'altra il *testing set*.

GMDH Shell dispone della caratteristica di *preprocessing* quando si utilizza il modello di regressione applicato ad un subset. Il pannello di controllo permette di settare il numero delle osservazioni volute per quel subset. Il *sample* che si ottiene può essere utilizzato per verificare l'accuratezza della previsione per valori sconosciuti. Nel mio studio ho preferito proseguire con l'applicazione uniforme sui dati della fase *preprocessing* impegnando il 20% del dataset per le osservazioni.

La maggior parte dei criteri di valutazione dei valori predetti, quando si applicano gli algoritmi di GMDH, non sono precisi soprattutto in presenza di rumore nel dataset. Il criterio MSE (*Mean Squared Error*) risulta molto utile quando si presentano anomalie nei dati, inoltre

tiene in considerazione la loro complessità [28]. Nel mio studio ho scelto di visualizzare i risultati con il parametro RMSPE (*Root mean square percentage error*) e MAPE (*Mean absolute percentage error*) (Figura 5). Il MAPE è usato in statistica per misurare l'accuratezza della previsione di un metodo. Il coefficiente di determinazione o R^2 è la proporzione della varianza nella variabile dipendente che è prevedibile dalle variabili indipendenti ed è una misura di quanto i dati si avvicinino alla linea della regressione. Secondo la definizione, R^2 può assumere il valore massimo di 1 [31]. Un modello sviluppato da un dataset di interesse deve essere sottoposto ad una serie di test. Tra questi è da considerare l'analisi dei residui che vengono riportati nella sezione *processing results*. Aprendo la scheda dedicata ai residui viene mostrato un grafico in cui i residui sono disposti sull'asse verticale e ai valori predetti dal modello sull'asse orizzontale. I residui sono importanti per dichiarare l'appropriatezza dei presupposti fatti. Nelle analisi di regressione c'è da considerare una sottile differenza tra errore e residuo [32]. Secondo la statistica, gli errori e i residui hanno definizioni simili e spesso vengono confuse. Sono misure della deviazione dei valori di un campione in analisi. L'errore di un valore è la deviazione (differenza tra valori) del valore osservato dal valore vero non osservabile di interesse, mentre il residuo di un valore osservato è la differenza tra il valore osservato e il valore stimato del campione.

Error measure		
	Target percentage	Target: "E.coli_MIC"
Postprocessed results	Model fit	Predictions
Number of observations	1166	292
Max. negative error	-100 %	-100 %
Max. positive error	117400 %	31438,5 %
Mean absolute percentage error (MAPE)	1190,14 %	999,45 %
Root mean square percentage error (RMSPE)	6155,33 %	3050,47 %
Residual sum	0,793633%	-0,221201%
Standard deviation of residuals	6046,58 %	2893,66 %
Coefficient of determination (R ²)	0,579619	0,523693
Correlation	0,7614	0,724599

Figura 5 Etichetta dell'accuracy in GMDH

Durante l'analisi di regressione ci si può domandare se il dataset che si sta usando sia adeguato. È importante tenere presente che i dataset possono essere sbilanciati ovvero presentare una minoranza o maggioranza di dati di una particolare categoria [33]. Il modo per migliorare il problema è quello di ricampionare il dataset attraverso il *resample method* o aggiungendo nuovi elementi o eliminando quelli esistenti. Gli step del metodo consistono in:

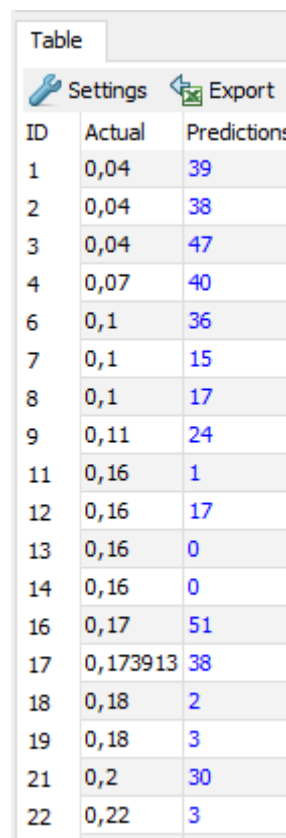
1. Prendere il dataset d'origine e stimare la quantità dei dati da ricampionare;
2. Modificare il dataset d'inizio aggiungendo oggetti per le categorie con minore rappresentanza (eseguendo un *oversampling*), o diminuendo gli oggetti maggiormente rappresentati (*undersampling*);
3. Si ricava come risultato un *resampled dataset*.

Il dataset di cui dispongo presenta uno sbilanciamento: maggiore presenza di dati per le sequenze con inattività. Ho dovuto applicare un *undersampling* su tutte e tre le specie e il loro dataset di inizio.

2.5.2. Confusion Matrix

Una volta ottenuti i modelli di regressione ho costruito le *Confusion Matrix*. Queste sono tabelle che permettono di visualizzare l'accuratezza dell'algorithmo usato.

La sezione *processing results* del software fornisce i risultati dell'analisi sotto forma di plot e sotto forma di tabella (Figura 6) e ho utilizzato quest'ultima per costruire le *Confusion Matrix*. I valori di questa tabella fanno riferimento ai dati reali e ai dati predetti dal programma che sono, quindi, riportati nelle *Confusion Matrix* in modo che le righe rappresentino i valori reali (*Actual*) mentre le colonne i valori predetti (*Predicted*).



ID	Actual	Predictions
1	0,04	39
2	0,04	38
3	0,04	47
4	0,07	40
6	0,1	36
7	0,1	15
8	0,1	17
9	0,11	24
11	0,16	1
12	0,16	17
13	0,16	0
14	0,16	0
16	0,17	51
17	0,173913	38
18	0,18	2
19	0,18	3
21	0,2	30
22	0,22	3

Figura 6 Esempio di tabella nella sezione *processing results*

Nella *Confusion Matrix*, *True Negatives* (TN) sono i risultati negativi di una corretta classificazione, *False Positives* (FP) sono i risultati negativi impropriamente classificati come positivi, *False Negatives* (FN) sono i risultati positivi impropriamente classificati come negativi e *True Positives* (TP) sono i risultati positivi correttamente classificati [34] (Tabella

1). La predizione accurata di un algoritmo di machine learning è la misura della sua performance ed è definita come $Accuracy = (TP + TN)/(TP + FP + TN + FN)$. Nel caso di un'analisi di un dataset bilanciato e tenendo conto del peso dei residui, il valore di *accuracy* ha un massimo di 1.

Tabella 1 Distribuzione dei dati nella Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	True Negatives	False Positives
Actual Positive	False Negatives	True Positives

Infine, le Confusion Matrix contengono i valori di *Precision* e *Recall* che sono due comuni classificazioni statistiche usate in più campi [35]. *Precision* rappresenta la quantità dei dati rilevanti recuperati da un'analisi diviso il numero totale dei dati recuperati dalla stessa analisi, *Recall* è la quantità dei dati rilevanti recuperati da un'analisi diviso la quantità totale dei dati rilevanti noti.

3. Risultati

In questo capitolo discuto dei risultati ottenuti dal lavoro descrivendo i risultati ottenuti applicando il metodo di regressione come spiegato nel capitolo precedente.

3.1. Selezione del modello

Gli algoritmi di machine learning si applicano ad una raccolta di dati. Nel mio studio mi sono concentrata sulla creazione di modelli di predizione riguardo a tre batteri: *Escherichia coli*, *Pseudomonas aeruginosa* e *Staphylococcus aureus*.

Utilizzando il software GMDH ho applicato il metodo della regressione su tre dataset diversi. Nel plot la linea grigia rappresenta i dati di input reali, la linea blu è la parte che il modello apprende dai dati, la linea rossa è la predizione; inoltre, l'area rossa intorno alla predizione è la zona di confidenza calcolata per le previsioni. In ordinata al grafico è riportata la MIC intesa come variabile target mentre in ascissa si considerano le variabili di input. I primi risultati hanno mostrato che c'era un errore nel dataset di partenza (Figura 7, Figura 8, Figura 9). Infatti, i plot evidenziano delle zone di discordanza tra valori reali, valori appresi e valori predetti. Osserviamo in particolare la parte di grafico riferita all'apprendimento (linea blu) riguardo i valori arbitrari di MIC delle sequenze peptidiche. Esaminando l'analisi di regressione di *E. coli* si nota che manca una sovrapposizione di dati dall'istanza 1136 all'istanza 1459. Per l'analisi eseguita sul dataset di *P. aeruginosa* la disparità cade dall'istanza 609 alla 893 e, infine, per la regressione del dataset di *S. aureus* dall'istanza 1070 alla 1364.

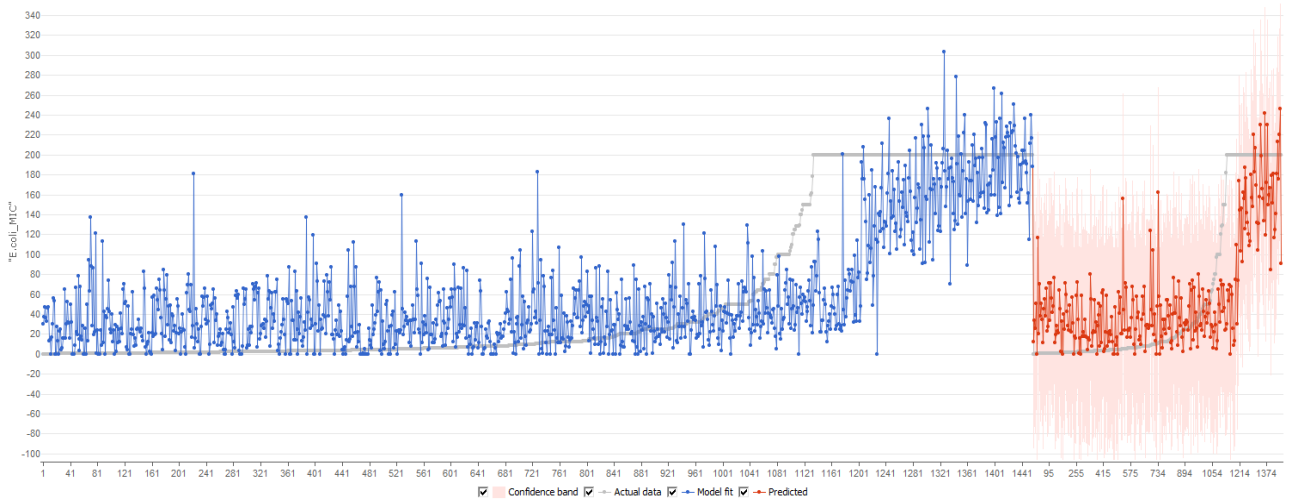


Figura 7 Plot di *E. coli* con valori arbitrari di MIC

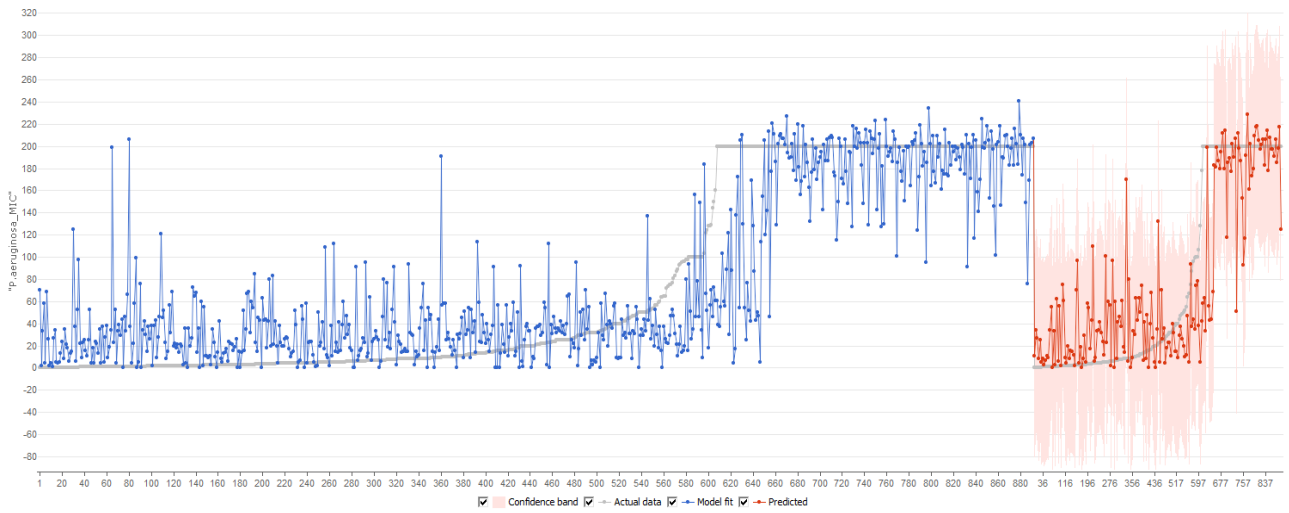


Figura 8 Plot di *P. aeruginosa* con valori arbitrari di MIC

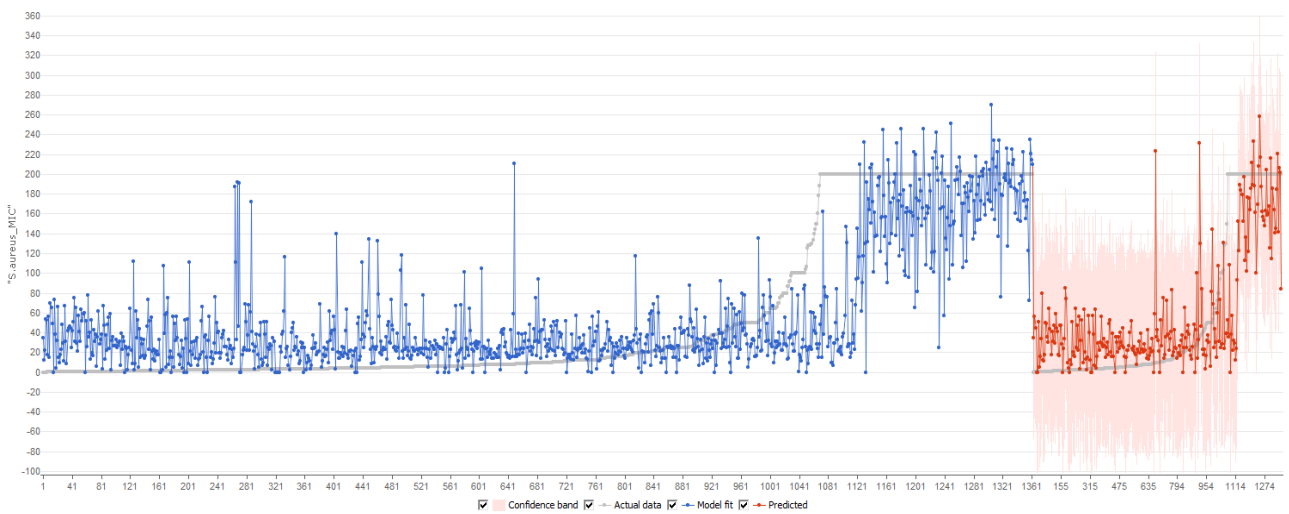


Figura 9 Plot di *S. aureus* con valori arbitrari di MIC

I dataset presentavano valori di MIC fissi al valore di 200 μ M, valore soglia scelto per indicarne l'inattività verso i batteri. Questo impediva una corretta predizione ed è stato necessario ricalcolare i nuovi valori di MIC. I valori di MIC ricalcolati comportano una migliore distribuzione dei dati, al fine di ridurre l'errore nella predizione. Ho fatto ricorso alla seguente formula:

$$MIC_{new} = MIC_{200} + (1 - distance) \times \left(\frac{MIC_{sperim.} - \theta}{\theta} \right) \times 100$$

in cui MIC_{new} rappresenta la MIC che deve essere calcolata, MIC_{200} corrisponde al valore di origine quindi 200, $distance$ rappresenta la distanza calcolata dalla matrice in MATLAB, $MIC_{sperim.}$ è il valore sperimentale di MIC presente nel dataset che corrisponde appunto alla distanza trovata, θ rappresenta il valore soglia pari a 55.

Con questi nuovi valori di MIC, ho creato dei nuovi dataset. Ho dapprima creato dataset chiamati *subset*, da utilizzare nella prima fase dello studio e, successivamente, ho riorganizzato questi dataset creandone altri nuovi chiamati *split* e *overlap*, caratterizzati rispettivamente dalla suddivisione e dalla sovrapposizione di sequenze.

Una volta avviato lo studio mi sono chiesta se il dataset fosse accurato per il tipo di analisi che dovevo eseguire, ovvero se fosse opportunamente bilanciato e presentasse una omogenea distribuzione dei dati. Analizzando i dataset di tutte e tre le specie, ho notato la presenza di un numero maggiore di sequenze peptidiche con inattività. Per rendere omogenei i dati all'interno dei dataset è stata necessaria l'applicazione della tecnica dell'*undersampling*: sono stati eliminati oggetti. In questo modo ho ottenuto dei nuovi dataset che ho definito *light* e che ho sottoposto all'algoritmo di regressione ottenendo nuovi modelli.

Infine, per poter sostenere che i modelli risultanti fossero accurati, ho applicato i metodi sopra descritti. La procedura utilizzata è la medesima per ogni specie in analisi.

3.1.1. Analisi Subset

Inizio la mia analisi studiando la specie *E. coli*. Il primo modello di regressione che ottengo mostra il seguente plot (Figura 10):

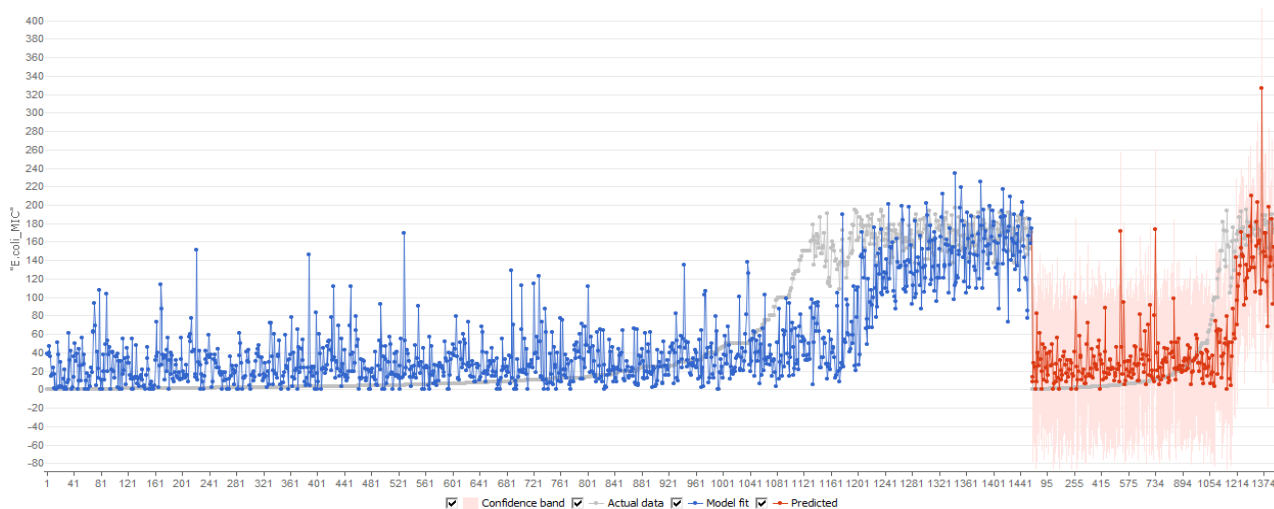


Figura 10 Plot Ecoli_subset

La linea blu di apprendimento segue, in questo caso, con maggiore precisione la linea grigia che rappresenta i dati di riferimento e questo si nota soprattutto se ci soffermiamo nella sezione riguardante i peptidi con inattività (istanza 1136-1459).

Tabella 2 *E. coli* R^2

Ecoli		Peptides
Model fit	Prediction	1458
0.58	0.52	

In Tabella 2 sono riportati valori del coefficiente di determinazione risultanti dall'analisi di regressione. La colonna *Model fit* contiene la misura dell'accuratezza calcolata per le

osservazioni usate per creare il modello, mentre la colonna *Prediction* le misure calcolate per le osservazioni trattenute per la validazione del modello ottenuto. Il valore massimo che il coefficiente di determinazione può assumere è pari a 1 e in questo caso il valore ottenuto indica che il modello può essere migliorato.

Il dataset che ho analizzato è adatto per il tipo di analisi che ho seguito? Il dataset si compone di:

- 526 inattivi
- 442 parzialmente attivi
- 491 attivi

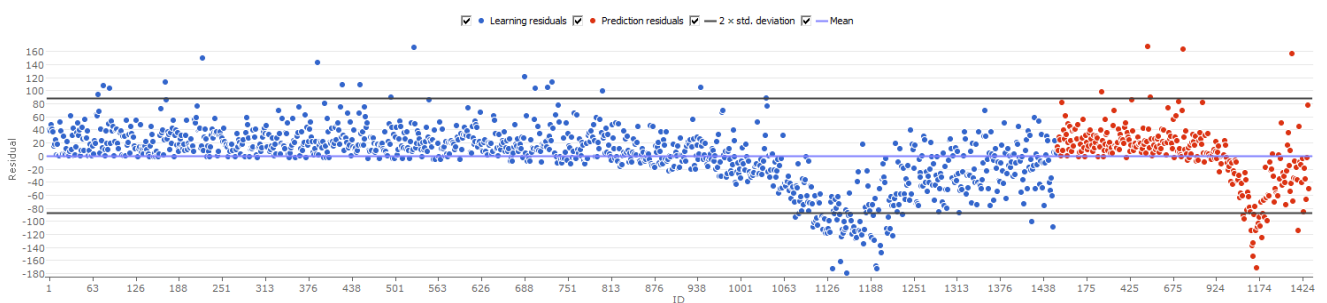


Figura 11 Residui di *E. coli* prima dell'*undersampling*.

Su questo modello (Figura 11) applico il metodo *undersampling* per rendere più omogeneo il dataset in analisi e ricreo dei nuovi dataset su cui applico nuovamente l'algoritmo di regressione. Il ricampionamento parte dal dataset iniziale da cui vengono eliminate 69 istanze creando un nuovo dataset chiamato *Ecoli_light1*. A sua volta questo diventa il dataset di inizio su cui applicare l'*undersampling* e questo vale per i successivi dataset creati. Quindi il dataset *Ecoli_light2* deriva dall'eliminazione di 5 istanze dal precedente, *Ecoli_light3* dall'eliminazione di 11 istanze, *Ecoli_light4* dall'eliminazione di 49 istanze, *Ecoli_light5* dall'eliminazione di 13 istanze e *Ecoli_light6* dall'eliminazione di 15 istanze. L'analisi eseguita sui nuovi dataset ha prodotto i seguenti risultati (Tabella 3):

Tabella 3 R^2 di *Ecoli_subset_light*

Ecoli_light1		Peptides
Model fit	Prediction	1389
0.76	0.70	
Ecoli_light2		Peptides
Model fit	Prediction	1384
0.71	0.72	
Ecoli_light3		Peptides
Model fit	Prediction	1373
0.78	0.77	
Ecoli_light4		Peptides
Model fit	Prediction	1324
0.78	0.76	
Ecoli_light5		Peptides
Model fit	Prediction	1311
0.87	0.82	
Ecoli_light6		Peptides
Model fit	Prediction	1296
0.84	0.82	

Passo all'analisi della seconda specie in esame ovvero *P. aeruginosa*.

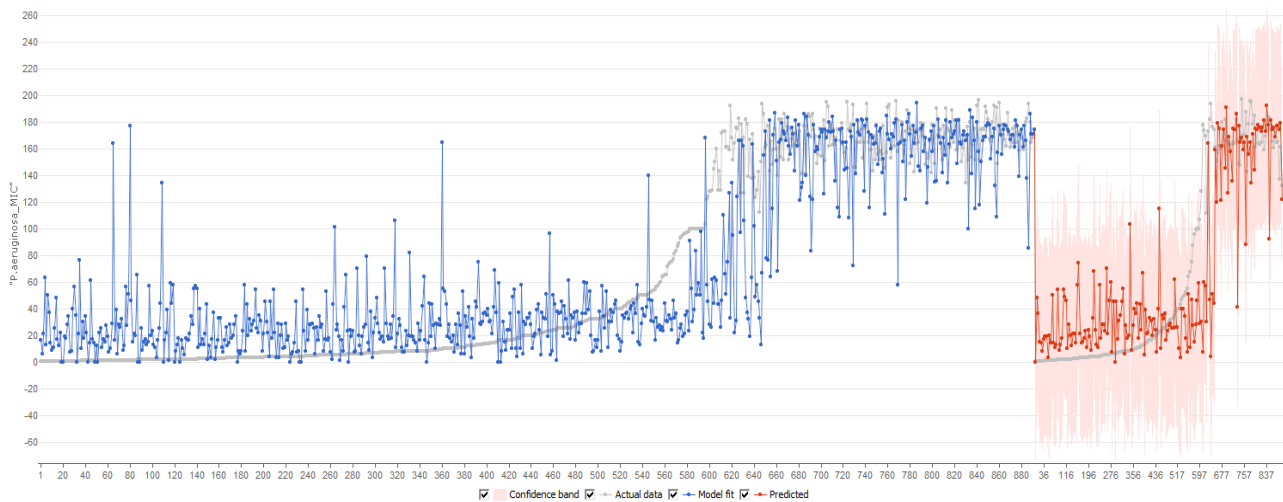


Figura 12 Plot di *P. aeruginosa*

La linea blu di apprendimento (Figura 12) anche in questo caso si avvicina maggiormente alla linea grigia dei dati di riferimento. Se osserviamo la sezione riguardante i peptidi con inattività, tuttavia, abbiamo che le due linee continuano a non sovrapporsi, sia nella zona di apprendimento che di predizione. L'algoritmo di regressione usato per questo studio ha mostrato i seguenti risultati del coefficiente di determinazione (Tabella 4):

Tabella 4 *P. aeruginosa* R^2

Paeruginosa		Peptides
Model fit	Prediction	892
0.74	0.65	

Ho analizzato il numero di oggetti presenti nel dataset:

- 411 inattivi
- 253 parzialmente attivi
- 228 attivi

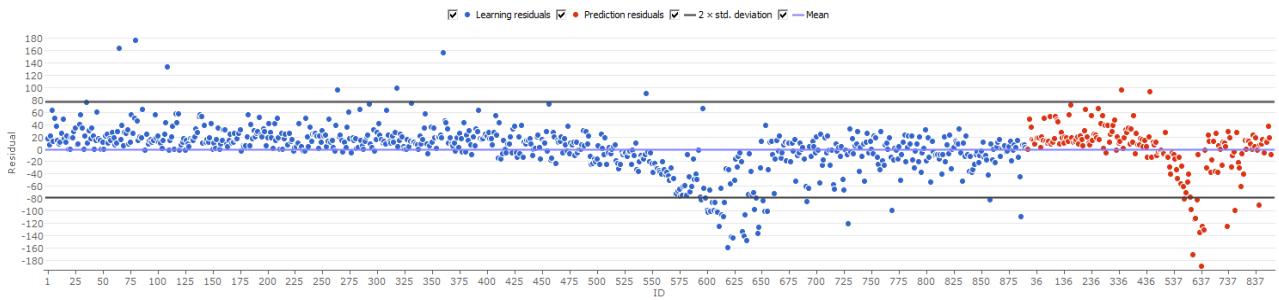


Figura 13 Residui di *P. aeruginosa* prima dell'*undersampling*

Per rendere più omogeneo il dataset (Figura 13) ho, quindi, eseguito il metodo *undersampling* creando nuovi dataset_light che ho sottoposto nuovamente all'algoritmo di regressione. In particolare dal dataset di partenza sono state eliminate 25 istanze. L'analisi di regressione ha poi prodotto i seguenti risultati (Tabella 5):

Tabella 5 R^2 di *P. aeruginosa subset_light*

Paeruginosa_light1		Peptides
Model fit	Prediction	867
0.80	0.78	

In questo caso l'applicazione del metodo *undersampling* è stata più veloce ed è bastato applicarlo una sola una volta in quanto i risultati sono subito migliorati.

Infine, ho eseguito l'analisi sulla specie *S. aureus* (Figura 14).

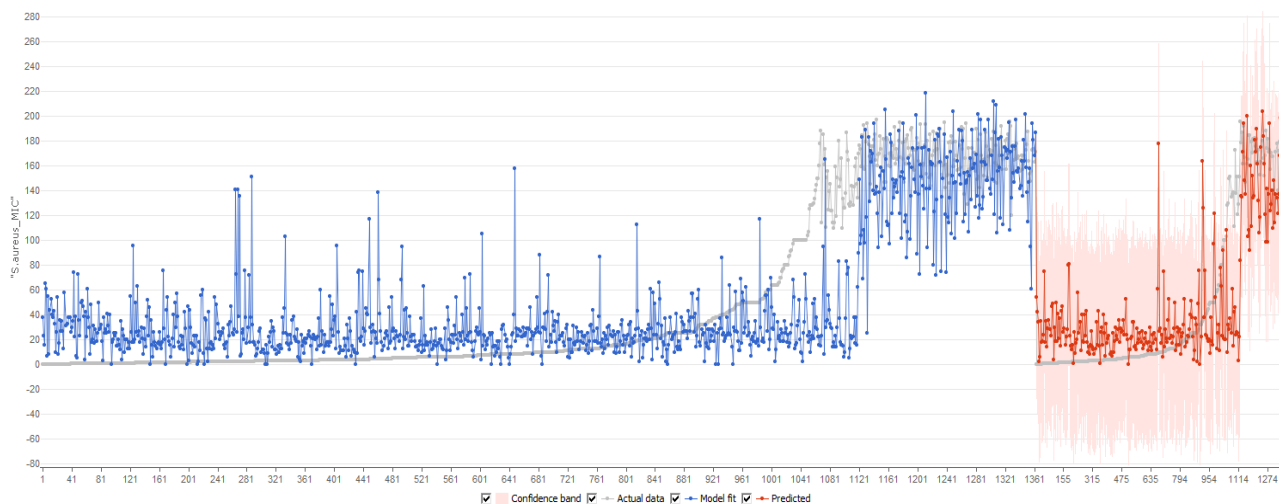


Figura 14 Plot di *S. aureus*

Per quanto riguarda le linee di apprendimento e di predizione, si nota una imprecisione nella sovrapposizione delle linee blue-grigia e rossa-grigia e dando uno sguardo ai risultati del coefficiente di determinazione ottengo (Tabella 6):

Tabella 6 R^2 di *S. aureus*

Saureus		Peptides
Model fit	Prediction	1363
0.62	0.60	

Gli oggetti presenti nel dataset sono così distribuiti:

- 467 inattivi
- 459 parzialmente attivi
- 437 attivi

E la distribuzione dei residui è mostrata nel seguente plot (Figura 15):

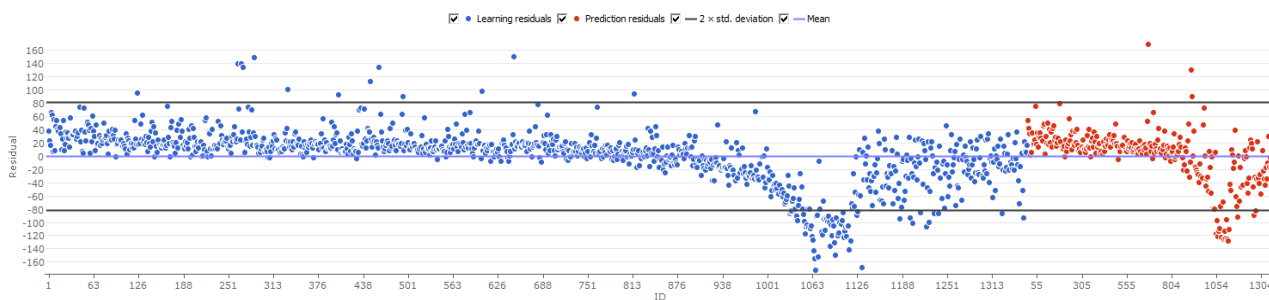


Figura 15 Residui di *S. aureus* prima dell'undersampling

Il metodo del ricampionamento viene applicato sul dataset di partenza e vengono eliminate 54 istanze ottenendo il nuovo dataset Saureus_light1. Questo che diventa il nuovo dataset di inizio è sottoposto a un nuovo ricampionamento ed eliminando 8 istanze si ottiene il nuovo dataset Saureus_light2. Questo metodo si ripete eliminando 15 istanze e creando il dataset Saureus_light3.

L'applicazione del metodo *undersampling* mi ha portato ai seguenti nuovi risultati (Tabella 7):

Tabella 7 R^2 di *S. aureus* subset_light

Saureus_light1		Peptides
Model fit	Prediction	1309
0.75	0.70	
Saureus_light2		Peptides
Model fit	Prediction	1301
0.76	0.71	
Saureus_light3		Peptides
Model fit	Prediction	1286
0.77	0.76	

Il ricampionamento è stato utile per ottenere un miglioramento dei modelli in quanto i valori di R^2 sono più alti.

Dall'analisi dei risultati dei modelli ottenuto finora con questo primo approccio, i modelli accettabili sono i seguenti (Tabella 8):

- Ecoli_light6
- Paeruginosa_light1
- Saureus_light3

Tabella 8 Modelli accettabili

Ecoli_light6		Peptides
Model fit	Prediction	1296
0.84	0.82	

Paeruginosa_light1		Peptides
Model fit	Prediction	867
0.80	0.78	

Saureus_light3		Peptides
Model fit	Prediction	1286
0.77	0.76	

Per dimostrare l'affidabilità di questi risultati analizzo i dati delle *Confusion Matrix* (Tabella 9, Tabella 10, Tabella 11).

Tabella 9 Confusion Matrix E. coli

Confusion Matrix Ecoli_light6						
<i>Model fit</i>	Actual	Predicted				Recall
		active	inactive	mild	total	
	active	516	19	132	667	0.77
	inactive	25	203	17	245	0.83
	mild	86	5	34	125	0.27

		total	627	227	183	1037	
		Precision	0.82	0.89	0.19		0.73

<i>Prediction</i>	Actual	Predicted					
			active	inactive	mild	total	Recall
		active	132	6	28	166	0.80
		inactive	7	51	4	62	0.82
		mild	23	2	6	31	0.19
		total	162	59	38	259	
		Precision	0.81	0.86	0.16		0.73

Tabella 10 Confusion Matrix *P. aeruginosa*

Confusion Matrix *Paeruginosa_light1*

<i>Model fit</i>	Actual	Predicted					
			active	inactive	mild	total	Recall
		active	317	8	74	399	0.79
		inactive	18	204	29	251	0.81
		mild	20	4	20	44	0.45
		total	355	216	123	694	
		Precision	0.89	0.94	0.16		0.78

<i>Prediction</i>	Actual	Predicted					
			active	inactive	mild	total	Recall
		active	75	3	21	99	0.76
		inactive	6	54	3	63	0.86
		mild	7	0	4	11	0.36
		total	88	57	28	173	
		Precision	0.85	0.95	0.14		0.77

Tabella 11 Confusion Matrix di *S. aureus*

Confusion Matrix *Saureus_light3*

<i>Model fit</i>	Actual	Predicted					
			active	inactive	mild	total	Recall
		active	577	12	120	709	0.81

	inactive	44	189	21	254	0.74
	mild	56	6	24	86	0.28
	total	677	207	165	1049	
	Precision	0.85	0.91	0.15		0.75

		Predicted					
		active	inactive	mild	total	Recall	
Prediction	Actual	active	121	2	29	152	0.80
		inactive	10	50	4	64	0.78
		mild	16	0	0	16	0.00
		total	147	52	33	232	
		Precision	0.82	0.96	0.00		0.74

L'efficacia del modello la possiamo leggere sulla base del valore di *accuracy* (in grassetto) riportato nelle *Confusion Matrix*. Questo valore si può leggere come la misura delle previsioni corrette in relazione a tutte le istanze prese in considerazione. Il valore dei tre modelli finora considerati va da un minimo di 0.73 a un massimo di 0.78 che è un valore accettabile considerando che il valore massimo che l'*accuracy* può assumere è pari a 1.

3.1.2. Analisi subset Split

Il lavoro fatto finora lo possiamo immaginare nel modo seguente: ho un grande insieme di partenza che è il dataset di origine che modifico riducendo il numero delle sequenze peptidiche. Tuttavia, in questo modo ho formato sottoinsiemi dello stesso insieme. Ho pensato di creare dall'insieme di origine due nuovi sottoinsiemi dati dalla suddivisione (*split*) dei dati. Ho scelto di costruire dei nuovi subset sulla base dei loro diversi valori di elicità. Per rendere omogeneamente distribuiti i dati all'interno dei dataset ho scelto opportunamente dei valori soglia (*threshold*). Per la divisione in due parti ho scelto un valore

soglia di 1.07 e di conseguenza ho ottenuto una tabella contenente i valori inferiori (chiamati min) e una con valori superiori al valore fissato (max).

L'idea dello *split* è quella di andare a migliorare i modelli creati con la precedente tecnica. Segue un'analisi dei valori risultanti. Vediamo adesso una comparazione dei valori di R^2 per ognuna delle specie in analisi (Tabella 12, Tabella 13, Tabella 14):

Tabella 12 Confronto R^2 per i dataset split di *E. coli*

E. coli	R^2		Peptides	Threshold
	Model fit	Prediction		
<i>Ecoli</i>	0.58	0.52	1457	
Ecoli_split_min	0.55	0.48	775	1.07
Ecoli_split_max	0.66	0.62	683	1.07
<i>Ecoli_light6</i>	0.84	0.82	1296	
Ecoli_light6_split_min	0.89	0.81	668	1.07
Ecoli_light6_split_max	0.84	0.78	628	1.07

Da *Ecoli* originale ai successivi R^2 si ha un peggioramento nel caso di valori minori e un miglioramento per i valori maggiori del valore soglia (0.58->0.66 e 0.52->0.62).

Per il modello ricampionato ho valori nell'intorno di quelli del primo modello e leggero miglioramento solo per il model fit dei valori minori del valore soglia (0.84->0.89).

Tabella 13 Confronto R^2 per i dataset split di *P. aeruginosa*

P. aeruginosa	R^2		Peptides	Threshold
	Model fit	Prediction		
<i>Paeruginosa</i>	0.74	0.65	892	
Paeruginosa_split_min	0.76	0.61	441	1.07
Paeruginosa_split_max	0.78	0.73	451	1.07
<i>Paeruginosa_light1</i>	0.80	0.78	867	
Paeruginosa_light1_split_min	0.78	0.76	422	1.07

Paeruginosa_light1_split_max 0.83 0.78 445 1.07

Il confronto con il dataset di origine mostra un miglioramento di R^2 per l'analisi del dataset con valori maggiori del valore soglia.

Nel caso del modello ricampionato non c'è evidente miglioramento.

Tabella 14 Confronto R^2 per i dataset split di *S. aureus*

S. aureus	R^2		Peptides	Threshold
	Model fit	Prediction		
<i>Saureus</i>	0.62	0.60	1363	
Saureus_split_min	0.72	0.62	732	1.07
Saureus_split_max	0.75	0.66	631	1.07
<i>Saureus_light3</i>	0.77	0.76	1286	
Saureus_light3_split_min	0.80	0.74	685	1.07
Saureus_light3_split_max	0.75	0.74	601	1.07

Per il dataset di origine si ha un miglioramento soltanto nel nuovo dataset con valori maggiori del valore soglia.

Per il modello ricampionato non si rileva un evidente miglioramento se non per il solo valore di *model fit* per il modello con valori inferiori al valore soglia.

L'analisi dell'accuratezza del modello creato è eseguita valutando la *Confusion Matrix* (Tabella 15, Tabella 16, Tabella 17).

Tabella 15 Confusion Matrix del dataset split di *E. coli*

Confusion Matrix Ecoli_split_max						
<i>Model fit</i>	Actual	Predicted				Recall
		active	inactive	mild	total	
		active	254	17	90	
inactive	16	90	22	128	0.70	

		mild	29	5	23	57	0.40
		total	299	112	135	546	
		Precision	0.85	0.80	0.17		0.67

			Predicted				
			active	inactive	mild	total	Recall
<i>Prediction</i>	Actual	active	62	6	22	90	0.69
		inactive	4	23	6	33	0.70
		mild	9	1	4	14	0.29
		total	75	30	32	137	
		Precision	0.83	0.77	0.13		0.65

Tabella 16 Confusion Matrix del dataset split di *P. aeruginosa*

Confusion Matrix Paeruginosa_split_max							
			Predicted				
			active	inactive	mild	total	Recall
<i>Model fit</i>	Actual	active	113	10	73	196	0.58
		inactive	9	95	19	123	0.77
		mild	18	1	23	42	0.55
		total	140	106	115	361	
		Precision	0.81	0.90	0.20		0.64

			Predicted				
			active	inactive	mild	total	Recall
<i>Prediction</i>	Actual	active	22	2	24	48	0.46
		inactive	0	24	7	31	0.77
		mild	3	0	8	11	0.73
		total	25	26	39	90	
		Precision	0.88	0.92	0.21		0.60

Tabella 17 Confusion Matrix del dataset split di *S. aureus*

Confusion Matrix Saureus_split_max							
			Predicted				
			active	inactive	mild	total	Recall
<i>Model fit</i>							

	Actual	active	231	4	97	332	0.70
		inactive	17	81	30	128	0.63
		mild	27	1	17	45	0.38
		total	275	86	144	505	
		Precision	0.84	0.94	0.12		0.65

			Predicted				
			active	inactive	mild	total	Recall
Prediction	Actual	active	50	3	29	82	0.61
		inactive	7	21	4	32	0.66
		mild	7	1	4	12	0.33
		total	64	25	37	126	
		Precision	0.78	0.84	0.11		0.60

Con la tecnica dello *split* si ha una leggera diminuzione dell'efficacia del modello in quanto i valori hanno un range di 0.60 e 0.67 di *accuracy* (in grassetto). Questo significa che si sono ottenuti ancora modelli accettabili.

3.1.3. Analisi subset Overlap

Costruisco, infine, dei subset che abbiamo delle sequenze in comune (*overlap*) scegliendo un valore soglia. Il valore threshold è nell'intorno di quello scelto precedentemente (1.07) creando stavolta disomogeneità tra i valori minimi e massimi riguardo l'elicità.

Ho voluto fare un ulteriore prova per migliorare i risultati dei modelli ottenuti finora. Stavolta ho pensato di sovrapporre una parte dei valori presenti nel dataset e di lanciare una nuova regressione. I due valori soglia usati sono 1.03 e 1.09 (Tabella 18, Tabella 19, Tabella 20).

Tabella 18 Confronto R^2 per i dataset overlap di *E. coli*

E. coli	R^2		Peptides	Threshold
	Model fit	Prediction		
<i>Ecoli</i>	0.58	0.52	1458	
Ecoli_overlap_min	0.60	0.49	939	1.09
Ecoli_overlap_max	0.59	0.39	435	1.03
<i>Ecoli_light6</i>	0.84	0.82	1296	
Ecoli_light6_overlap_min	0.90	0.86	811	1.09
Ecoli_light6_overlap_max	0.83	0.82	375	1.03

Il modello ricampionato con valori <1.09 mostra valori di R^2 maggiori del modello di riferimento.

Tabella 19 Confronto R^2 per i dataset overlap di *P. aeruginosa*

P. aeruginosa	R^2		Peptides	Threshold
	Model fit	Prediction		
<i>Paeruginosa</i>	0.74	0.65	892	
Paeruginosa_overlap_min	0.71	0.56	226	1.09
Paeruginosa_overlap_max	0.70	0.60	544	1.03
<i>Paeruginosa_light1</i>	0.80	0.78	867	
Paeruginosa_light1_overlap_min	0.72	0.66	215	1.09
Paeruginosa_light1_overlap_max	0.75	0.64	523	1.03

Non ci sono miglioramenti rispetto a quelli ottenuti dai dataset di riferimento.

Tabella 20 Confronto R^2 per i dataset overlap di *S. aureus*

S. aureus	R^2		Peptides	Threshold
	Model fit	Prediction		
<i>Saureus</i>	0.62	0.60	1363	
Saureus_overlap_min	0.71	0.72	402	1.09
Saureus_overlap_max	0.64	0.65	895	1.03

<i>Saureus_light3</i>	0.77	0.76	1286	
Saureus_light3_overlap_min	0.70	0.68	376	1.09
Saureus_light3_overlap_max	0.81	0.79	835	1.03

Per il modello *overlap* ottenuto a partire dal dataset di origine si ottiene un miglioramento di R^2 del modello con valori <1.09 . Del modello ricampionato il miglioramento si ottiene per valori <1.03

L'analisi della performance della predizione è valutata con l'analisi delle *Confusion Matrix* (Tabella 21, Tabella 22, Tabella 23).

Tabella 21 Confusion Matrix del dataset overlap di *E. coli*

Confusion Matrix Ecoli_light6_overlap_min							
		Predicted				Recall	
		active	inactive	mild	total		
<i>Model fit</i>	Actual	active	347	5	47	399	0.87
		inactive	11	152	10	173	0.88
		mild	57	3	17	77	0.22
		total	415	160	74	649	
		Precision	0.84	0.95	0.23		0.80

		Predicted				Recall	
		active	inactive	mild	total		
<i>Prediction</i>	Actual	active	91	1	9	101	0.90
		inactive	7	36	2	45	0.80
		mild	13	0	3	16	0.19
		total	111	37	14	162	
		Precision	0.82	0.97	0.21		0.80

Tabella 22 Confusion Matrix del dataset overlap di *S. aureus (min)*

Confusion Matrix Saureus_overlap_min						
<i>Model fit</i>		Predicted				Recall
		active	inactive	mild	total	

		active	130	19	46	195	0.67
		inactive	9	75	9	93	0.81
	Actual	mild	14	3	17	34	0.50
		total	153	97	72	322	
		Precision	0.85	0.77	0.24		0.69

			Predicted				
			active	inactive	mild	total	Recall
<i>Prediction</i>	Actual	active	30	4	14	48	0.63
		inactive	1	19	4	24	0.79
		mild	4	0	4	8	0.50
		total	35	23	22	80	
		Precision	0.86	0.83	0.18		0.66

Tabella 23 Confusion Matrix del dataset overlap di *S. aureus* (max)

Confusion Matrix Saureus_light3_overlap_max

			Predicted				
			active	inactive	mild	total	Recall
<i>Model fit</i>	Actual	active	353	17	46	416	0.85
		inactive	27	150	5	182	0.82
		mild	49	8	13	70	0.19
		total	429	175	64	668	
		Precision	0.82	0.86	0.20		0.77

			Predicted				
			active	inactive	mild	total	Recall
<i>Prediction</i>	Actual	active	87	2	15	104	0.84
		inactive	7	36	3	46	0.78
		mild	15	0	2	17	0.12
		total	109	38	20	167	
		Precision	0.80	0.95	0.10		0.75

L'analisi delle *Confusion Matrix* in questo caso riporta un valore di *accuracy* che oscilla tra 0.66 e 0.80. Questo indica che la tecnica dell'*overlap* ossia quella in cui ho sovrapposto una parte delle istanze ha prodotto accettabile efficacia dei modelli.

Con la mia analisi ho ottenuto *nove modelli* validi, riportati in Tabella 24.

Tabella 24 Valori di R^2 dei modelli accettabili

Dataset	R^2		Peptides
	Model fit	Prediction	
Ecoli_light6	0.84	0.82	1296
Ecoli_split_max	0.66	0.62	683
Ecoli_light6_overlap_min	0.90	0.86	811
Paeruginosa_light1	0.80	0.78	867
Paeruginosa_split_max	0.78	0.73	451
Saureus_light3	0.77	0.76	1286
Saureus_split_max	0.75	0.66	631
Saureus_overlap_min	0.71	0.72	402
Saureus_light3_overlap_max	0.81	0.79	835

L'analisi di regressione è usata anche da Shu *et al.* per caratterizzare gli AMP [36]. In particolare, per il loro studio prendono in considerazione AMP con alta attività contro specie batteriche sia Gram-positive che Gram-negative. I modelli risultanti dell'applicazione della tecnica di regressione mostrano con un coefficiente di determinazione (R^2) pari a 0.73. In generale, risulta difficile una corretta analisi statistica in quanto la maggior parte delle volte si usa un dataset in cui il numero dei descrittori è maggiore dei peptidi usati come campioni. Infatti, essi suggeriscono di usare dataset che comprendono descrittori contenenti molte più informazioni chimiche riguardo l'attività biologica dei peptidi.

Il range di valori di R^2 ottenuti dai modelli ottenuti dal mio studio è di 0.66-0.90 per il *Model fit* e 0.62-0.86 per la *Prediction*, con valori di media pari a 0.78 e 0.74 rispettivamente. Sono

dunque valori accettabili se confrontati con il valore di R^2 ottenuto nello studio di Shu *et al.* pari a 0.73.

4. Conclusioni

Negli ultimi anni stiamo vivendo la crisi della resistenza batterica e i ricercatori stanno cercando valide alternative al problema. La ricerca si concentra su diversi aspetti del problema e riguarda più ambiti. Specialisti del settore medico-scientifico per primi cercano di contrastare l'espandersi della difficoltà del trattamento delle infezioni. Si cerca di prescrivere farmaci antibiotici solo quando è necessario ma sembra che questa pratica non basti. I microbiologi studiano i batteri sotto ogni punto di vista per trovare ispirazione e trovare un modo per fermare la loro azione nociva. Contemporaneamente anche gli altri campi di ricerca sono chiamati in causa. Tra questi ci sono gli esperti bioinformatici. La tecnologia ha permesso, negli anni, di poter sfruttare i computer e i software per aiutare la ricerca scientifica. I computer diventano sempre più potenti e questo va di pari passo con la necessità di potenza di calcolo per analizzare un numero di informazioni sempre maggiore. Gli informatici sviluppano software adatti per le architetture hardware di cui anno dopo anno dispongono. Le ricerche scientifiche multidisciplinari stanno prendendo sempre più piede proprio grazie ai recenti sviluppi. D'altro canto, l'unione delle conoscenze di ambiti diversi è aiutata dal fatto di dover far fronte ad un altro problema quello finanziario. La ricerca di laboratorio tradizionale deve far fronte al problema delle tempistiche di analisi abbastanza lunghe. Gli esperimenti di laboratorio infatti richiedono molto tempo e per la maggior parte delle volte sono ricerche che non hanno esiti positivi. La bioinformatica può dare un aiuto contro la perdita di tempo. Con i software si possono simulare le condizioni di analisi e studiare cosa accade se si vogliono apportare modifiche al sistema sotto indagine. Dalla simulazione al computer si può poi proseguire con i metodi tradizionali per testare le predizioni del software.

Il mio studio si concentra su un lavoro computazionale e il metodo che ho utilizzato si basa sull'uso di algoritmi di machine learning applicati a raccolte di dati. I modelli che questa tecnica fornisce devono essere verificati e analizzati. Esistono diversi metodi per verificare l'accuratezza dei risultati e nel mio studio ho scelto quelli che meglio si adattassero all'analisi. Il miglioramento dei dataset risulta necessario quando i dati sono troppo sbilanciati e questo compromette la buona riuscita della predizione. Gli algoritmi di machine learning mimano le reti neurali umane e sono capaci di apprendere ma sono necessari degli accorgimenti affinché l'addestramento delle reti neurali dia risultati soddisfacenti. Nel mio studio ho fatto ricorso a più metodiche per migliorare i dataset che ho sottoposto all'analisi di regressione e rendere affidabili i modelli risultanti.

I modelli non sono fine a sé stessi ma diventano un punto di partenza per nuovi studi. Una prima ipotesi di lavoro futuro è quello di integrare i modelli in tool computazionali per produrre nuove sequenze peptidiche. La conoscenza dell'attività antimicrobica viene applicata a nuove sequenze che a loro volta vengono modificate con lo scopo di ottenere sequenze di cui si ha conoscenza dell'attività. Tuttavia, la tecnologia evolve giorno dopo giorno e con essa la disponibilità di computer sempre più potenti e capaci di elaborare sempre più grosse quantità di dati e lo sviluppo di nuovi software e nuovi approcci di analisi in futuro daranno vita a nuove soluzioni al problema della resistenza batterica. Una seconda ipotesi è di utilizzare una tecnica diversa e quindi fare uso del deep learning, oppure prendere in esame dati che riguardano altre specie batteriche.

Bibliografia

1. Zaffiri L, Gardner J, Toledo-Pereyra LH. History of Antibiotics. From Salvarsan to Cephalosporins. *Journal of Investigative Surgery* 2012; 25: 67-77.
2. Levy SB, Marshall B. Antibacterial resistance worldwide: causes, challenges and responses. *Nature Medicine* 2004; 10: S122.
3. Tyers M, Wright GD. Drug combinations: a strategy to extend the life of antibiotics in the 21st century. *Nature Reviews Microbiology* 2019; 17: 141-155.
4. Brown ED, Wright GD. Antibacterial drug discovery in the resistance era. *Nature* 2016; 529: 336.
5. Czaplewski L, Bax R, Clokie M et al. Alternatives to antibiotics—a pipeline portfolio review. *The Lancet Infectious Diseases* 2016; 16: 239-251.
6. Lázár V, Martins A, Spohn R et al. Antibiotic-resistant bacteria show widespread collateral sensitivity to antimicrobial peptides. *Nature Microbiology* 2018; 3: 718.
7. Dubos RJ, Cattaneo C. Studies on a bactericidal agent extracted from a soil bacillus: III. Preparation and activity of a protein-free fraction. *Journal of Experimental Medicine* 1939; 70: 249-256.
8. Dubos RJ. Studies on a bactericidal agent extracted from a soil bacillus: II. Protective effect of the bactericidal agent against experimental *Pneumococcus* infections in mice. *Journal of Experimental Medicine* 1939; 70: 11-17.
9. Hotchkiss RD, Dubos RJ. Fractionation of the bactericidal agent from cultures of a soil bacillus. *Journal of Biological Chemistry* 1940; 132: 791-792.
10. Rammelkamp CH, Weinstein L. Toxic effects of tyrothricin, gramicidin and tyrocidine. *The Journal of Infectious Diseases* 1942; 71: 166-173.
11. Hirsch JG. Phagocytin: a bactericidal substance from polymorphonuclear leucocytes. *Journal of Experimental Medicine* 1956; 103: 589-611.
12. Zeya HI, Spitznagel JK. Antibacterial and enzymic basic proteins from leukocyte lysosomes: separation and identification. *Science* 1963; 142: 1085-1087.
13. Zhao X, Wu H, Lu H et al. LAMP: a database linking antimicrobial peptides. *PloS one* 2013; 8: e66557.
14. Ganz T. Defensins: antimicrobial peptides of innate immunity. *Nature reviews immunology* 2003; 3: 710.
15. Loeffler JM, Nelson D, Fischetti VA. Rapid killing of *Streptococcus pneumoniae* with a bacteriophage cell wall hydrolase. *Science* 2001; 294: 2170-2172.
16. Naghmouchi K, Le Lay C, Baah J et al. Antibiotic and antimicrobial peptide combinations: synergistic inhibition of *Pseudomonas fluorescens* and antibiotic-resistant variants. *Research in microbiology* 2012; 163: 101-108.
17. Marr AK, Gooderham WJ, Hancock REW. Antibacterial peptides for therapeutic use: obstacles and realistic outlook. *Current opinion in pharmacology* 2006; 6: 468-472.
18. Kamysz W. Are antimicrobial peptides an alternative for conventional antibiotics. *Nucl Med Rev Cent East Eur* 2005; 8: 78-86.
19. Hancock REW, Diamond G. The role of cationic antimicrobial peptides in innate host defences. *Trends in microbiology* 2000; 8: 402-410.

20. Matsuzaki K, Sugishita K, Fujii N et al. Molecular basis for membrane selectivity of an antimicrobial peptide, magainin 2. *Biochemistry* 1995; 34: 3423-3429.
21. Cudic M, Otvos Jr L. Intracellular targets of antibacterial peptides. *Current drug targets* 2002; 3: 101-106.
22. Yeaman MR, Yount NY. Mechanisms of antimicrobial peptide action and resistance. *Pharmacological reviews* 2003; 55: 27-55.
23. Gull S, Shamim N, Minhas F. AMAP: Hierarchical multi-label prediction of biologically active and antimicrobial peptides. *Computers in biology and medicine* 2019; 107: 172-181.
24. Wu Q, Ke H, Li D et al. Recent progress in machine learning-based prediction of peptide activity for drug discovery. *Current topics in medicinal chemistry* 2019; 19: 4-16.
25. Piotto SP, Sessa L, Concilio S et al. YADAMP: yet another database of antimicrobial peptides. *International journal of antimicrobial agents* 2012; 39: 346-351.
26. Huang Y, He L, Li G et al. Role of helicity of α -helical antimicrobial peptides to improve specificity. *Protein & Cell* 2014; 5: 631-642.
27. Anastasakis L, Mort N. The development of self-organization techniques in modelling: a review of the group method of data handling (GMDH). RESEARCH REPORT-UNIVERSITY OF SHEFFIELD DEPARTMENT OF AUTOMATIC CONTROL AND SYSTEMS ENGINEERING 2001.
28. Mueller JA, Ivachnenko AG, Lemke F. GMDH algorithms for complex systems modelling. *Mathematical and Computer Modelling of Dynamical Systems* 1998; 4: 275-316.
29. Ostertagová E. Modelling using polynomial regression. *Procedia Engineering* 2012; 48: 500-506.
30. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:181112808 2018.
31. Dancer D, Tremayne A. R-squared and prediction in regression with ordered quantitative response. *Journal of Applied Statistics* 2005; 32: 483-493.
32. Dekking F, Kraaikamp C, Lopuhaä H et al. A modern introduction to probability and statistics. *Understanding why and how*; 2005
33. Burnaev E, Erofeev P, Papanov A. Influence of resampling on accuracy of imbalanced classification. In: *International Society for Optics and Photonics*:987521
34. Chawla NV, Bowyer KW, Hall LO et al. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 2002; 16: 321-357.
35. Buckland M, Gey F. The relationship between recall and precision. *Journal of the American society for information science* 1994; 45: 12-19.
36. Shu M, Yu R, Zhang Y et al. Predicting the activity of antimicrobial peptides with amino acid topological information. *Medicinal Chemistry* 2013; 9: 32-44.