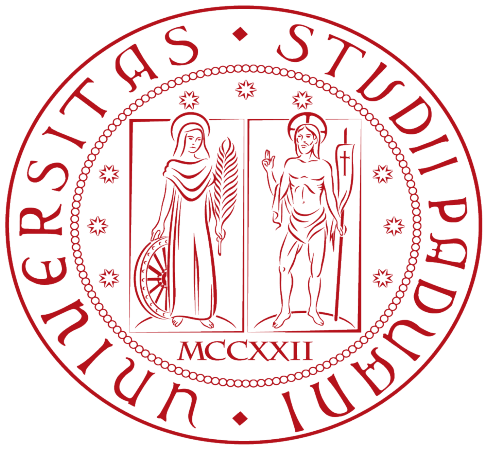
UNIVERSITA' DEGLI STUDI DI PADOVA



FACOLTA' DI INGEGNERIA
CORSO DI LAUREA SPECIALISTICA IN BIOINGEGNERIA

ALGORITMI PER L'ANALISI ED IL TROUBLESHOOTING DI SEGNALI DI SEQUENZIAMENTO SANGER DEL DNA

ALGORITHMS FOR ANALYSIS AND TROUBLESHOOTING OF SANGER DNA SEQUENCING DATA

Relatore: Prof. Barbara Di Camillo

Correlatore: Prof. Giovanni Sparacino
Correlatore: Dott. Barbara Simionati
Correlatore: Ing. Fabrizio Levorin

Laureanda: Jole Costanza

ANNO ACCADEMICO 2009/2010

Indice

Sommario

Una volta effettuato il sequenziamento Sanger di un campione di DNA, è necessario valutare la correttezza del sequenziamento analizzando i segnali Raw Data ed Analyzed Data forniti dal sequenziatore per poter risalire ai problemi, legati alle tecniche di preparazione del campione o alla procedura di sequenziamento, che possono generare errori nella determinazione della sequenza. Tali analisi, detta in gergo troubleshooting, attualmente viene effettuata da biologi esperti, in quanto non esistono software che realizzano automaticamente questa analisi. In questa tesi è stato realizzato un algoritmo che compie l'analisi automatica dei segnali e classifica le problematiche in opportune categorie al fine di fornire un supporto efficiente e veloce al troubleshooting. Per realizzare l'algoritmo è stato utilizzato un training set di 167 sequenze con problematiche note. L'algoritmo, implementato in Matlab, utilizza tecniche di filtro a media mobile, di peak detection e altre metodologie tipiche dell'analisi del segnale. Per validare l'algoritmo sono state utilizzate 1200 sequenze con problematiche note. Per ogni problematica è stata testata la performance dell'algoritmo, valutando quante classificazioni corrette vengono da esso compiute. I risultati ottenuti sono buoni, superando per ogni problematica mediamente il 93% dell'assegnazione corretta, definita come il rapporto tra il numero dato dalla somma dei veri positivi e negativi e il numero di sequenze del validation test.

Introduzione

Il sequenziamento del patrimonio genetico, principalmente di quello umano, ha aperto nuovi scenari di ricerca grazie anche allo sviluppo e ai progressi nei metodi e nelle tecnologie di analisi.

Molte malattie dipendono da mutazioni localizzate su geni specifici. Nel 1986, Renato Dulbecco suggerì che determinando la sequenza normale del DNA umano si sarebbero potuti ricavare alcuni vantaggi anche per la ricerca sul cancro. Prese così corpo il Progetto Genoma Umano (HGP), inaugurato ufficialmente nel 1990, con l'obiettivo di mappare il patrimonio genetico umano (genoma), ovvero di descrivere la struttura, la posizione e la funzione dei geni che caratterizzano la specie umana. In seguito al sequenziamento del genoma umano sono emersi fatti interessanti ed inattesi. Dei 3,2 miliardi di paia di basi, meno del 2% fa parte di regioni codificanti e il numero totale di geni ammonta a circa 24000. Prima delle tecniche del sequenziamento, si stimava che il numero dei geni del genoma umano variasse da un minimo di 80000 a un massimo di 100000. Un numero di geni tanto inferiore alle aspettative sta a significare che la diversità osservata nelle molecole proteiche deve esser dovuto a modifiche post-traduzionali. In altre parole,un gene eucariotico di medie dimensioni codifica in realtà più di una molecola proteica.

Un'altra scoperta interessante è che quasi tutto il genoma (circa il 99,9%) è identico in tutti gli individui di una specie. Nonostante questa apparente omogeneità, esistono comunque molte differenze tra gli individui. I ricercatori hanno mappato oltre 2 milioni di polimorfismi di singoli nucleotidi (SNPs), ossia di basi che differiscono almeno nell'1% delle persone.

Per conseguire tutte queste importanti scoperte, la genomica si è avvalsa delle diverse tecniche di sequenziamento del DNA. Il termine sequenziamento, in biologia molecolare indica il processo per la determinazione dell'esatta struttura primaria di un biopolimero e cioè dell'ordine delle basi nel caso di un acido

nucleico o degli amminoacidi nel caso di proteine. La metodica per il sequenziamento di DNA principalmente utilizzata finora si basa sul metodo della terminazione della catena sviluppato da Frederick Sanger. Questa tecnica si basa sull'utilizzo di nucleotidi modificati (dideossitrifosfato, ddNTPs) per interrompere la reazione di sintesi in posizioni specifiche lungo la sequenza. Il sequenziamento tramite il cosiddetto metodo Sanger risulta oggigiorno nella capacità di sequenziare frammenti fino a 1000 basi e l'automazione ha reso possibile la corsa di 384 reazioni contemporaneamente.

Il risultato del sequenziamento è rappresentato nell'*elettroferogramma* che mostra una successione di picchi che corrisponde alla sequenza dei nucleotidi: il colore del picco corrisponde al tipo di base azotata (un colore per ogni base). Normalmente il sistema del sequenziatore interpreta automaticamente l'elettroferogramma, e quando l'interpretazione non è ovvia, il sistema inserisce o una N al posto della lettera (A, T, C, G) che identifica la base azotata mancante (questa, quando è possibile, può essere corretta manualmente dopo aver analizzato visivamente l'elettroferogramma), oppure inserisce comunque una lettera associata ad un quality score, un punteggio legato alla probabilità di errore nella determinazione dell'identità della base.

Il *troubleshooting* è la procedura di controllo di qualità dell'elettroferogramma necessaria per l'identificazione, l'analisi e la risoluzione di problemi inerenti alla preparazione del campione di DNA e/o alle reazioni che stanno alla base del sequenziamento, e/o a malfunzionamenti legati alla strumentazione. La lettura dell'elettroferogramma è un passo necessario per valutare la correttezza del sequenziamento, ma è anche un lavoro che richiede importanti risorse umane (personale esperto per una giusta interpretazione dei risultati), economiche, nonché dispendio di tempo. Lo scopo della tesi è quello di automatizzare questo processo attraverso la realizzazione di un algoritmo che sia un supporto d'analisi efficiente e veloce per il personale dell'azienda addetto all'analisi visiva dell'elettroferogramma.

Il **Capitolo 1** della tesi presenta una panoramica sulla biologia del DNA,

in cui vengono descritti i suoi costituenti principali, i nucleotidi, e come sono organizzati (struttura, composizione e disposizione) lungo la doppia elica. Viene spiegato il processo di replicazione del DNA, e l'importanza funzionale che questa macromolecola assume nelle cellule degli organismi viventi in cui è contenuta.

Nel **Capitolo 2** vengono illustrate le varie tecniche di sequenziamento del DNA, in particolar modo viene descritto nel dettaglio il metodo Sanger, con le varie tecniche di preparazione del campione biologico oggetto di studio. Sono riportate anche le tecniche di sequenziamento di nuova generazione che si basano sul principio del pirosequenziamento.

Come accennato precedentemente, nel sequenziamento automatico si possono presentare dei problemi che possono originare degli errori nella determinazione della sequenza. Scoprire dall'analisi dell'elettroferogramma (**Capitolo 3**) la causa degli errori permette di risalire all'origine del problema, risolverlo se possibile, e di ottenere quindi un risultato migliore. La manifestazione di questi errori nell'elettroferogramma e le loro cause sono argomenti trattati nel **Capitolo 4**, mentre nel **Capitolo 5** della tesi è descritto l'algoritmo proposto per l'analisi ed il riconoscimento automatico degli errori direttamente sull'elettroferogramma.

L'algoritmo, realizzato in Matlab, è organizzato in vari step, sequenziali in alcune fasi dell'analisi e paralleli in altre. Verte in un'analisi del segnale di sequenziamento fornito dal sequenziatore Applied Biosystems 3730xl, con lo scopo di compiere una decisione e di attribuire al segnale le classi che ne rappresentano i problemi che generano errori nella determinazione della sequenza. L'algoritmo elabora i dati forniti dal sequenziatore per approssimare l'andamento del segnale, il suo inviluppo, valutare la sua intensità e per analizzare la forma e la regolarità del singolo picco dell'elettroferogramma. Per far ciò ci si avvale di strategie di peak detection, strategie di sogliatura per la valutazione del superamento di un certo valore limite, e di tecniche per l'approssimazione di dati. L'algoritmo è stato realizzato grazie ad un training set composto da 167 sequenze di cui erano note le problematiche. Utilizzando un altro set di dati di 1200

sequenze, con problematiche note, è stata valutata la performance dell'algoritmo, ovvero quante classificazioni corrette vengono da esso compiute per ogni problematica (risultati riportati nel **Capitolo 6**). La valutazione della performance è stata fatta ricavando i Veri Positivi, Veri Negativi, Falsi Positivi, Falsi Negativi per ogni classificazione del troubleshooting automatico. Una ulteriore verifica è stata fatta valutando come funziona l'algoritmo di fronte ad un sequenziamento andato a buon fine (test con 150 sequenze che non presentano problematiche). In fine, il **Capitolo 7** espone le conclusioni che si possono trarre da questo lavoro di tesi.

Questa tesi è stata realizzata durante uno stage universitario iniziato nel mese di Ottobre 2009 presso l'azienda BMR Genomics di Padova. Quest'azienda offre servizi di analisi del DNA, tra cui quello del sequenziamento attraverso il metodo Sanger, per mezzo del sequenziatore Applied Biosystems 3730xl. I dati, utilizzati per implementare e testare l'algoritmo, derivano proprio da questa strumentazione. È da notare anche l'originalità e l'unicità di questo progetto, il primo in assoluto ad affrontare il problema del troubleshooting automatico di sequenze di DNA.

Capitolo 1

La biologia del DNA

1.1 La struttura chimica del DNA

Il DNA, acido desossiribonucleico, è una molecola presente sia nelle cellule procariotiche, che eucariotiche, che codifica l'informazione ereditaria e la trasferisce da una generazione all'altra. Gli acidi nucleici, tra cui il DNA, sono costituiti da monomeri chiamati nucleotidi, ognuno dei quali consiste di uno zucchero pentoso, un gruppo fosfato e una base azotata.

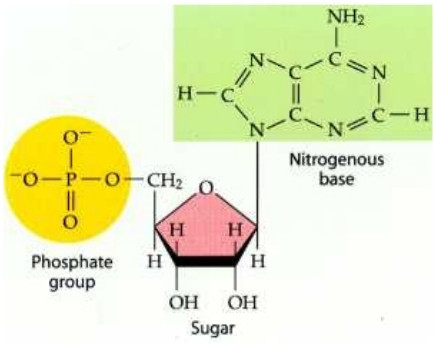


Figura 1.1: *rappresentazione della molecola nucleotidica di DNA costituita da un gruppo fosfato, un gruppo zuccherino e una base azotata (Adenina).*

Nel DNA lo zucchero pentoso è il desossiribosio. Lo scheletro della macromolecola consiste di una catena di zuccheri pentosi alternati a gruppi fosfato

(zucchero-fosfato-zucchero-fosfato). Le basi sono legate agli zuccheri e sporgono dalla catena polinucleotidica. I nucleotidi sono uniti da legami fosfodiestere presenti tra lo zucchero di un nucleotide e il fosfato del successivo (diestere si riferisce ai legami covalenti formati dai gruppi –OH che reagiscono con i gruppi fosfato acidi). I gruppi fosfato uniscono il carbonio in 3' di uno zucchero pentoso al carbonio in 5' dello zucchero adiacente.

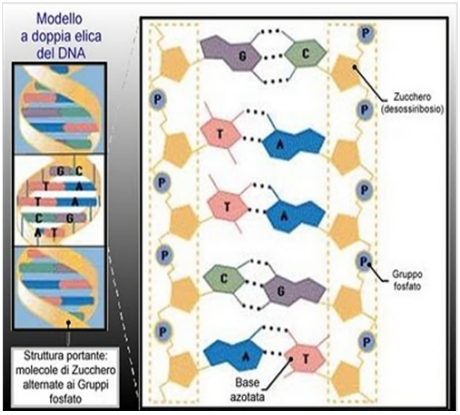


Figura 1.2: *modello a doppia elica del DNA. In evidenza i legami a ponte idrogeno tra le basi azotate dei due filamenti complementari, e i legami fosfodiestere presenti tra lo zucchero di un nucleotide e il fosfato del successivo del singolo filamento.*

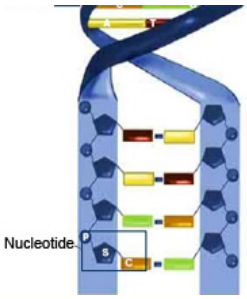


Figura 1.3: *disposizione dei nucleotidi nella doppia elica del DNA.*

Il DNA è a doppio filamento e le due catene polinucleotidiche sono tenute assieme da legami a ponte idrogeno tra le basi azotate. I due filamenti di DNA hanno direzione opposte, e tale orientamento antiparallelo permette ai due filamenti di adattarsi l'uno all'altro nello spazio tridimensionale.

Nel DNA si trovano quattro basi azotate, e quindi quattro nucleotidi; queste basi e le loro abbreviazioni sono adenina (A), citosina (C), guanina (G), timina (T). Adenina e timina si appaiono sempre tra loro, così come citosina e guanina.

1.2 La replicazione del DNA

La replicazione del DNA ha lo scopo di trasmettere l'informazione, e quindi il suo patrimonio genetico, ad una cellula figlia quando la cellula si riproduce. La replicazione si svolge in due tappe: la doppia elica viene svolta (denaturata) ad opera dell'enzima DNA elicasi; nuovi nucleotidi vengono aggiunti mediante legami fosfodiesterico ad ogni nuovo filamento in via di accrescimento, in base ad una sequenza determinata dall'appaiamento complementare con le basi presenti sul filamento stampo, processo catalizzato dall'enzima DNA polimerasi.

Le DNA polimerasi possono allungare un filamento polinucleotidico legando in modo covalente nuovi nucleotidi ad un filamento preesistente, ma non sono in grado di iniziare un filamento dal nulla. Per catalizzare la replicazione l'enzima necessita di un innesco definito primer. Il primer corrisponde ad una breve sequenza di RNA a singolo filamento.

Successivamente, la DNA polimerasi aggiunge nucleotidi in corrispondenza dell'estremità 3' del primer fino a quando non è completata la replicazione di quella sezione di DNA.

Un'osservazione fondamentale è che i nucleotidi sono aggiunti in corrispondenza dell'estremità 3' del filamento in accrescimento, ovvero l'estremità in cui il filamento di DNA possiede un gruppo ossidrilico (-OH) libero legato al carbonio 3' del desossiribosio terminale.

In questo modo il DNA si replica esattamente uguale a se stesso e la nuova cellula avrà un patrimonio genetico identico alla cellula di partenza.

1.3 L'informazione contenuta nel DNA

Il DNA è una molecola informazionale. L'informazione contenuta nella macromolecola è codificata nella sequenza delle basi che ne costituiscono i filamenti. Tale informazione è poi usata dall'RNA per specificare la sequenza aminoacida che definisce la struttura primaria di una proteina.

Le conoscenze relative alla struttura e ai meccanismi di replicazione del DNA hanno permesso lo sviluppo di tecniche capaci di fornire copie multiple di sequenze di DNA e di determinare la sequenza nucleotidica di molecole di DNA: la tecnica del sequenziamento.

Capitolo 2

Il sequenziamento

2.1 Il sequenziamento Sanger

Il sequenziamento è il metodo fondamentale per caratterizzare una macromolecola, sia che si tratti di determinare l'ordine degli aminoacidi di una proteina o la sequenza di basi di un acido nucleico. Per capire l'importanza di questa tecnica e la conseguente informazione che fornisce, basti pensare che il sequenziamento di un intero genoma può permettere di predire la sequenza di tutte le proteine che potenzialmente questo può produrre.
Il metodo di sequenziamento del DNA con dideossinucleoterminatori o *metodo di Sanger* consiste di 3 fasi: la preparazione del campione, la reazione di sequenziamento, l'elettroforesi.

2.1.1 La preparazione del campione

Nella prima fase di preparazione del campione, il filamento di DNA che si vuole sequenziare viene copiato artificialmente, in modo da ottenere diverse copie identiche dello stesso. Sono due le tecniche che permettono questo processo di duplicazione: il *DNA ricombinante*, e la *PCR*, dall'inglese *Polymerase Chain Reaction*.

Il DNA ricombinante è una sequenza di DNA ottenuta artificialmente dalla

combinazione di materiale genetico di origini differenti. Per ottenerlo ci si serve di sistemi biologici come i plasmidi. Quest'ultimi sono piccoli filamenti circolari di DNA presenti nel citoplasma batterico e distinguibili dal cromosoma batterico per le loro dimensioni ridotte. I batteri usano plasmidi come veicolo per trasportare DNA ad un altro batterio. È possibile usare la capacità dei plasmidi di integrarsi in un batterio ospite (ad esempio *Escherichia Coli*) per integrare un gene o una sequenza di DNA di interesse. Perché la sequenza di DNA o il gene di interesse possa essere correttamente trasportato, occorre anzitutto che sia tagliato e ridotto nei minimi termini possibili. In seguito al taglio, il gene potrà essere integrato (cioè inserito) all'interno del vettore. L'ultima fase consiste nell'inserimento di un plasmide all'interno del batterio, che viene fatto replicare. Dopo la replicazione, si ottengono così molti batteri che contengono lo stesso materiale genetico e i plasmidi con la sequenza di DNA di interesse. Attraverso opportune tecniche di laboratorio, le sequenze di DNA oggetto allo studio vengono estratte dai plasmidi, ricavando più copie del campione biologico d'interesse.

Col processo di reazione a catena della polimerasi (PCR), è possibile produrre milioni di copie di una quantità molto ristretta di DNA iniziale. È dunque una tecnica più rapida rispetto al DNA ricombinante, sebbene necessiti delle sequenze nucleotidiche (primer) che possano appaiarsi correttamente con le estremità del segmento da amplificare. La PCR è un processo ciclico in cui una sequenza di processi vengono ripetuti molte volte. Frammenti di DNA a doppio filamento vengono denaturati attraverso la somministrazione di calore (circa 90°C). Alla miscela viene aggiunto un primer sintetizzato chimicamente (alla temperatura di 50-60°C per favorire il legame del primer col filamento). Vengono aggiunti chimicamente i quattro deossinucleosidi trifosfato (dNTPs) e la DNA polimerasi, che catalizza la produzione di nuovi filamenti complementari a quelli da copiare. Un singolo ciclo di reazioni impiega pochi minuti per raddoppiare la quantità di DNA iniziale, e il DNA neosintetizzato viene a trovarsi nello stato a doppia elica. Ripetendo il ciclo di reazioni (denaturazione, primer annealing,

copia del filamento) molte volte si ottiene un aumento esponenziale del numero di copie di DNA iniziale. La DNA polimerasi è però un enzima che viene in gran parte distrutto dalle alte temperature necessarie durante la fase di denaturazione. Per ovviare a tal problema viene utilizzata la DNA polimerasi del batterio *Thermus Aquaticus* che vivendo in acque caldissime dispone di un intero meccanismo termoresistente, compresa la DNA polimerasi.

2.1.2 La reazione di sequenziamento

Durante la fase di reazione di sequenziamento, il campione biologico viene sottoposto a quattro processi: denaturazione, primer annealing, copia del filamento, terminazione. Con la denaturazione i singoli filamenti di DNA così separati sono posti in provetta. Il primer annealing è la fase in cui viene aggiunto un primer all'estremità 3' di uno dei due filamenti. Il primer è sintetizzato artificialmente e appositamente per la sequenza di DNA da sequenziare. Vengono allestite quattro miscele in quattro provette, una per ogni base. In ogni provetta viene aggiunta la DNA polimerasi (fase della copia del filamento), i quattro nucleotidi (dATP, dCTP, dGTP, dTTP), e una piccola quantità di un dideossinucleoside trifosfato (ad esempio ddATP).

Un dideossinucleoside trifosfato (ddNTP), è un nucleotide che non ha il gruppo –OH in posizione 3' dello zucchero.

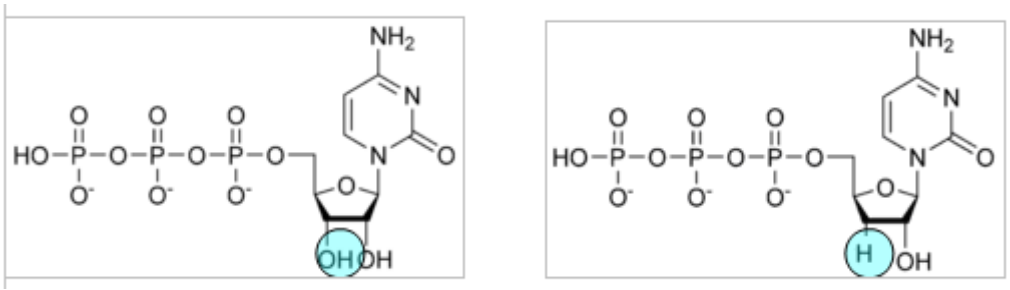


Figura 2.1: *Rappresentazione di un deossinucleoside trifosfato, a sinistra, e di un dideossinucleoside trifosfato, a destra. Il dideossinucleoside trifosfato è un nucleotide che non ha il gruppo -OH in posizione 3' dello zucchero.*

Questi comunque potranno esser aggiunti dalla DNA polimerasi a un filamento di DNA in corso di sintesi mediante la formazione di un legame fosfodiesterico tra il suo 5'-fosfato e il 3'-OH del residuo precedente. Tuttavia, poiché i ddNTPs mancano del gruppo -OH in posizione 3', il nucleotide successivo non potrà esser legato come avviene nella replicazione naturale del DNA. Per questo motivo la sintesi si arresta alla posizione in cui un ddNTP è stato incorporato all'estremità in accrescimento di un filamento di DNA (fase di terminazione). A differenza di quanto accade nella PCR questa volta viene copiato solo il filamento specifico per il primer utilizzato, in direzione 5'-3'. La replicazione del DNA procede e nella provetta si viene a trovare una miscela di filamenti stampo del DNA insieme ad una varietà di filamenti neosintetizzati più brevi. I filamenti nuovi, ognuno dei quali termina con un ddATP, per l'esempio riportato, avranno lunghezze differenti.

Una volta che la DNA polimerasi incontra una base T sul filamento stampo, essa potrà aggiungere o un dATP o un ddATP. Se viene aggiunto un dATP la crescita del filamento continua, mentre se viene aggiunto un ddATP, la crescita del filamento si arresta. Questo processo viene ripetuto in altre provette rispettivamente per il ddGTP, ddTTP, ddCTP. Dopo aver fatto proseguire per un po' la replicazione del DNA, i filamenti neosintetizzati vengono denaturati e tramite opportune tecniche di laboratorio separati dai filamenti stampo. A questo punto il preparato è pronto per la successiva fase e viene sottoposto ad elettroforesi.

2.1.3 L'elettroforesi

L'elettroforesi è un processo elettrocinetico nel quale molecole e particelle cariche, sotto l'influenza di un campo elettrico, migrano in direzione del polo che ha carica opposta. Grazie alla presenza dei gruppi fosfato, le molecole di DNA sono cariche negativamente e quindi migreranno verso il polo positivo (anodo) se sottoposte a un campo elettrico, con velocità che dipende anche dalla loro

lunghezza, oltre che dall'intensità della corrente.

Nel sequenziamento manuale le quattro miscele di frammenti di terminazione della catena, una per ogni analogo ddNTP marcato radioattivamente, vengono sottoposte ad elettroforesi su gel di poliacrilamide o gel di agarosio in differenti corsie. Durante la corsa i frammenti più corti si muoveranno più agevolmente attraverso il gel rispetto i frammenti più lunghi. Se il campo elettrico viene tolto prima che le molecole abbiano raggiunto l'elettrodo, si ha una separazione dei singoli componenti in base alla loro mobilità elettroforetica. La sequenza delle basi del DNA complementare a quella cercata viene letta dall'autoradiogramma delle quattro linee.

Nel sequenziamento automatico, invece non è necessario separare le quattro reazioni di terminazione in quattro provette differenti, ma si può allestire una singola reazione inserendo i quattro ddNTPs marcati mediante l'incorporazione di un composto fluorescente, diverso per ogni base, e l'elettroforesi avviene lungo dei capillari, all'interno dei quali è sempre ricostruita la rete di gel o in polimero. Durante la corsa elettroforetica, i frammenti vengono letti in ordine di lunghezza crescente da un fascio laser che eccita i marcatori fluorescenti. Quindi l'intensità della luce emessa viene misurata e tale informazione, cioè quale colore di fluorescenza e dunque il tipo di ddNTP è presente all'estremità di ogni filamento di differente lunghezza, viene inviata ad un computer.

L'introduzione dell'elettroforesi capillare per la separazione dei frammenti marcati ha consentito un notevole aumento della processività. Sono stati inoltre sviluppati modelli di sequenziatori automatici che sono in grado di eseguire corse elettroforetiche multiple su apparecchi multicapillari.

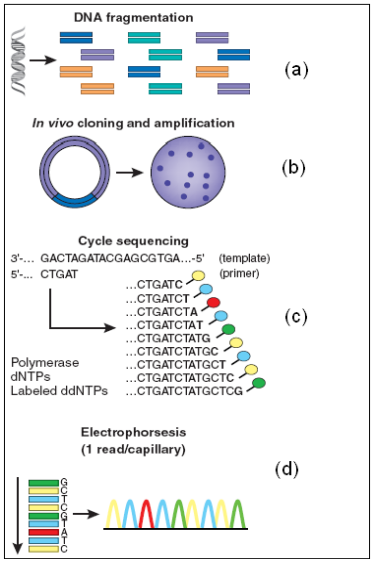


Figura 2.2: *(a) estrazione del filamento di DNA da sequenziare. (b) preparazione del campione. (c) reazione di sequenziamento. (d) elettroforesi capillare*

2.2 Altri tipi di sequenziamento

2.2.1 Sequenziatori di nuova generazione

I sequenziatori di nuova generazione presentano il pregio, rispetto al metodo Sanger, di servirsi di tecniche di amplificazione in vitro piuttosto che della complicata procedura dei plasmidi, ma soprattutto di utilizzare array (di diversa natura) per sequenziare contemporaneamente milioni di frammenti di DNA. Queste migliorie hanno permesso alle nuove piattaforme di ridurre drasticamente i tempi e i costi richiesti. Nonostante questi incoraggianti passi in avanti, i sequenziatori di nuova generazione, a eccezione del Genome Sequencer FLX 454 (GS FLX), non sono ancora le tecnologie di riferimento per il sequenziamento genomico, a causa soprattutto dei pesanti limiti in termini di lunghezza delle *read* e di accuratezza nella determinazione delle basi.

Il GS FLX 454 utilizza una tecnica detta pirosequenziamento. Questa

tecnica si articola in più fasi. È necessaria una prima fase di amplificazione durante la quale viene utilizzata una variante della PCR, detta PCR ad emulsione o *em*PCR, tramite la quale si ottengono milioni di copie identiche di ogni frammento di DNA. La lettura delle sequenze avviene invece per mezzo della tecnica del pirosequenziamento, che utilizza l'enzima DNA-polimerasi.

Uno dei limiti di questa tecnica è il fatto che essa garantisce la linearità del segnale solo fino a otto incorporazioni contemporanee; la lettura di sequenze contenenti omopolimeri più lunghi di otto basi può portare a risultati non corretti.

Questa tecnica è sviluppata dal Genome Sequencer FLX della 454 Life Sciences Corporation (centro di eccellenza della Roche Applied Science) ed è stato il primo sequenziatore di nuova generazione disponibile sul mercato come prodotto commerciale. Le *read* ottenute con il GS FLX sono lunghe 200-300 basi (400 con la nuova versione *Titanium*), molto meno rispetto a quelle ottenuti con il metodo Sanger, e sono inoltre caratterizzate da un'accuratezza inferiore. Tuttavia, dato che il volume dei reagenti può essere ammortizzato sull'intero set di sequenze presenti sull'array, i costi si riducono molto: 60 dollari per Mb (costi approssimati che comprendono i soli reagenti). Inoltre, questa nuova tecnica è caratterizzata da un maggiore grado di "parallelismo", ovvero può trattare milioni di sequenze diverse contemporaneamente, sia in fase di amplificazione che in fase di sequenziamento. L'enorme mole di dati in output permette di aumentare di molto il coverage del genoma originale.

Oltre al GS FLX sono state proposte altre tecnologie di sequenziamento *high throughput*. Tra queste, quelle che godono di maggiore popolarità sono l'AB SOLiD, della Applied Biosystems, e il Genome Analizer (o Solexa), della Illumina. La tecnologia AB SOLiD utilizza l'emPCR per la prima fase di amplificazione, mentre per il sequenziamento delle basi si serve di una tecnica che utilizza l'enzima ligasi al posto della DNA-polimerasi. La piattaforma Solexa, invece, amplifica i frammenti servendosi di un'altra versione della PCR, detta bridge PCR. Per la fase di sequenziamento viene usata ancora la DNA-polimerasi, ma in una procedura diversa, che prevede l'utilizzo anche di particolari

deossinucleotidi modificati. Come anticipato in precedenza, queste tecnologie di nuova generazione sono caratterizzate da un livello di accuratezza non confrontabile con quello ottenuto con il metodo di Sanger. Inoltre, le *read* ottenute risultano lunghe circa settanta basi, quindi molto più corte anche di quelle fornite in output dal GS FLX. Nuovamente, gli aspetti positivi sono dati dai costi e dalla quantità di dati in output: ad ogni corsa dello strumento vengono sequenziate più di 3000 Mb, ad un costo di circa 2 dollari per Mb (costi approssimati che comprendono i soli reagenti).

La metodologia Sanger è tuttora la più utilizzata perché permette di ottenere read lunghe fino a 1000 basi, con un'accuratezza nella determinazione delle basi pari al 99.999%. Gli aspetti negativi più gravosi sono rappresentati dalla lentezza di questa tecnologia, che permette di sequenziare solamente 1 Mb alla volta, e dai costi elevati, che superano i 500 dollari per Mb (costi approssimati che comprendono i soli reagenti).

2.3 Scopo della tesi

L'Applied Biosystems 3730xl (mostrato in figura 2.3) è uno strumento per il sequenziamento del DNA che sfrutta il metodo Sanger, ed usa un sistema di elettroforesi capillare, ognuno dei quali riempito col polimero POP-7, ed uno di rilevamento delle molecole di fluorescente contenute nei frammenti.

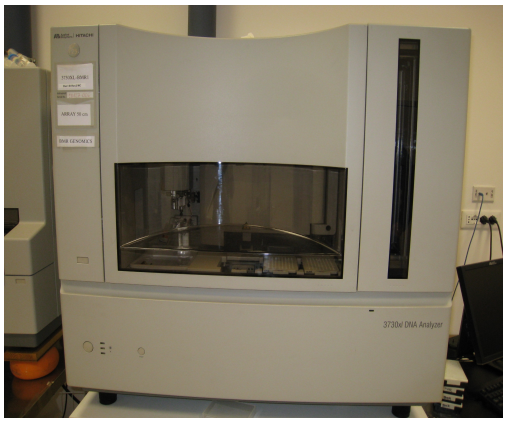


Figura 2.3: *sequenziatore Applied Biosystems 3730xl per il sequenziamento di DNA con il metodo Sanger utilizzato nell'azieda BMR Genomics di Padova.*

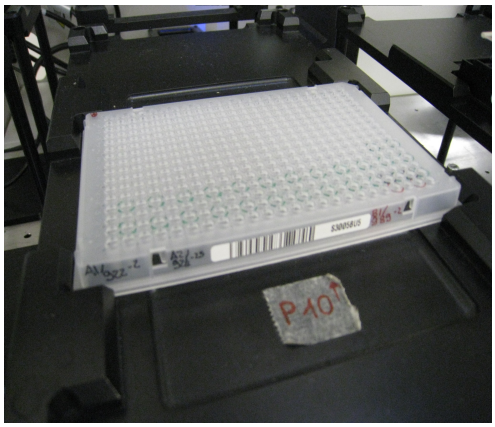


Figura 2.4: *piastra da 384 pozzetti contenenti la soluzione col DNA da sequenziare.*

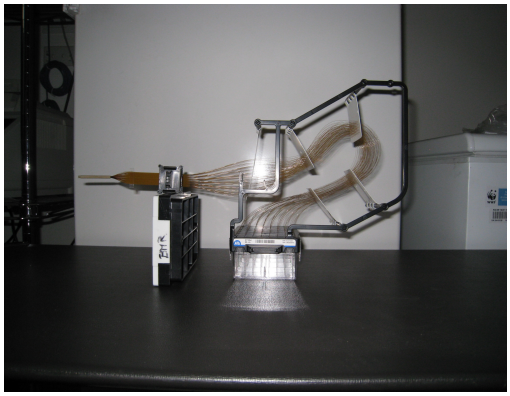


Figura 2.5: *capillari per elettroforesi*

Questo sequenziatore riesce a sequenziare più di 1000 basi, utilizzando piastre con 96 o 384 pozzetti, permettendo quindi l'analisi di 96 o 384 campioni di DNA (figura 2.4).

La qualità e la quantità del DNA stampo, del primer, dei reagenti usati per la reazione di sequenziamento Sanger e la procedura di elettroforesi capillare sono fattori importanti che influiscono sulla qualità del segnale rilevato ed elaborato, e quindi sull'accuratezza nella determinazione delle basi.

Figura 2.6: *programma Sequencing Analysis 5.2 per l'analisi visiva dei segnali di sequenziamento*

Il *troubleshooting* è la procedura di analisi del risultato del sequenziamento necessaria per capire se la qualità dello stesso è buona. In caso contrario, bisogna riconoscerne le cause e ripetere o meno la procedura di sequenziamento modificando qualità o quantità del template. L'osservazione dei dati viene compiuta da personale esperto che analizza visivamente (attraverso opportuni programmi d'analisi come il *Sequencing Analysis 5.2* mostrato in figura 2.6) il risultato e l'andamento di ogni segnale che viene fornito dal software dall'Applied Biosystems 3730xl, decidendo se l'elettroferogramma presenta o no problematiche, e se il campione deve essere risequenziato.

Il lavoro di questa tesi ha l'obiettivo di automatizzare l'analisi e la ricerca di tali problematiche nell'elettroferogramma, fornendo un software d'ausilio alla decisione. I dati, utilizzati per implementare e testare l'algoritmo di ricerca delle problematiche, sono stati forniti dalla BMR Genomics, un'azienda di Padova che offre servizi di analisi del DNA, tra cui anche il sequenziamento. A seconda dei casi, al segnale viene attribuita un classe. Le cause degli errori possono esser di più tipi: imputabili allo strumento (ad esempio la corsa nei capillari), o legati alla reazione di sequenziamento (ad esempio la mancata reazione), al templato o al primer. Gli effetti invece sono legati alla risoluzione, all'andamento o alla qualità del segnale.

Capitolo 3

I dati

3.1 Il sequenziatore Applied Biosystems 3730xl

I dati utilizzati per questo lavoro di tesi sono stati forniti dall'azienda padovana BMR Genomics e sono relativi a campioni di DNA sequenziati con il metodo Sanger attraverso la tecnologia del sequenziatore Applied Biosystems 3730xl. I capillari per elettroforesi sono realizzati in silice fusa o teflon, avente diametro interno nel range di 25-75 μm e diametro esterno di 300-400 μm, rivestito da uno strato protettivo di poliammide che lo rende resistente e maneggevole. La lunghezza del capillare (25-75 cm) non influisce sull'efficienza del processo, ma gioca un ruolo importante sul tempo di migrazione e quindi sulla durata dell'analisi. All'interno di ogni capillare è presente il polimero POP-7 che costruisce la rete attraverso cui si muovono i frammenti.

Una volta applicata la differenza di potenziale (che varia tra 8.5 kV e 13.2 kV a seconda delle impostazioni del sequenziatore) inizia la corsa elettroforetica (alla temperatura di 60°C). Passando dalla finestra di rilevamento le molecole di fluorescente vengono colpite da un fascio laser ed emettono luce. Un rilevatore registra lo spettro delle onde elettromagnetiche e il 3730xl Data Collection software legge e interpreta i dati di fluorescenza, mostrandoli in un *elettroferogramma* che riporta lo spettro di emissione dei vari fluorescenti in

funzione del tempo. Dato che vengono identificati prima i nucleotidi terminali dei segmenti più corti, l'elettroferogramma rappresenta proprio la sequenza ordinata dei nucleotidi letti, detta *read*. Durante questa procedura chiamata *basecalling* viene anche assegnato un punteggio ad ogni base identificata. Questi punteggi, detti *quality score*, vengono determinati grazie ad un algoritmo simile a Phred e sono legati alla probabilità di errore nella determinazione dell'identità delle basi a partire dal tracciato elettroforetico.

Generalmente durante la fase di risospensione del campione di DNA da caricare sul sequenziatore viene inserito anche un primer di controllo, detto *marcatore*, contenente la prima molecola marcata con lo stesso fluorescente usato per marcare i ddGTPs. L'uso del marcatore, lungo circa 20 basi, è utile per capire se un possibile fallimento del sequenziamento è stato causato da una mancata corsa nei capillari o perché la reazione di sequenziamento non è stata eseguita.

3.1.1 Raw Data ed Analyzed Data: il file ABIF

L'Applied Biosystems 3730xl salva i dati in un file binario con un formato definito ABIF (Applied Biosystems Inc. Format). Questo file contiene il *Raw Data, l'Analyzed Data*, e altre informazioni relative al sequenziamento. Il Raw Data è la sequenza di dati grezzi raccolti durante l'emissione dei fluorescenti: è la combinazione dei segnali di fluorescenza associati ad ogni base azotata. Il segnale associato ai nucleotidi contenenti Guanina (G) è di colore nero, il segnale associato ai nucleotidi contenenti Timina (T) è di colore rosso, per la Citosina (C) è blu, per l'Adenina (A) è verde, come mostrato in figura 3.1. Durante la corsa elettroforetica, i frammenti vengono letti in ordine di lunghezza crescente, per cui l'asse x del Raw Data rappresenta la lunghezza del filamento di DNA, mentre l'asse y rappresenta la scala d'intensità luminosa rilevata dalla fluorescenza dei marcatori.

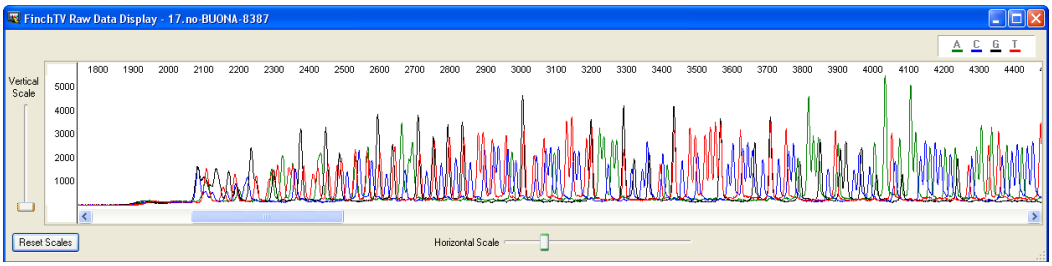


***Figura 3.1**: RawData della sequenza "17.no-BUONA-B387.ab1".*

Il software del sequenziatore, il cui algoritmo non è noto, rielabora il Raw Data e normalizza i dati in ampiezza e nel tempo, fornendo l'A*nalyzed Data,* l'elettroferogramma vero e proprio. È da questo segnale che vengono identificate le basi, e dalla qualità del singolo picco viene assegnato il quality score.

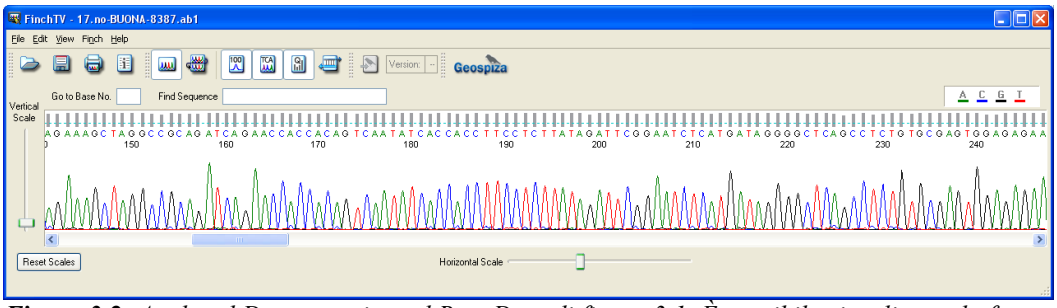


***Figura 3.2**: Analyzed Data associato al Raw Data di figura 3.1. È possibile visualizzare la forma dei picchi relativi ad ogni base azotata del DNA. Questa assume un colore differenze, come mostrato nella legenda. Ad ogni lettera della sequenza è associata il quality score, rappresentato dalla barra grigia in verticale. Più alta è la barra migliore è la qualità del risultato.*

Il file ABIF fornisce inoltre altre informazioni relative al sequenziamento: la *sequenza* delle basi nucleotidiche, il *signal strength* associato ad ogni segnale, il *base spacing* ovvero l'ampiezza media di ogni singolo picco, e le informazioni sulle impostazioni del macchinario, come il numero del capillare in cui è stata fatta la corsa elettroforetica, l'identificativo del campione, la data e il tempo di inizio e di fine della corsa.

Capitolo 4

Controllo di qualità del segnale

4.1 Caratteristiche dei dati, controllo di qualità e troubleshooting

Sia il sequenziamento manuale che quello automatico [Cap. 2, paragrafo 2.1.3] presentano dei problemi, che possono originare errori nella determinazione della sequenza. Questi tipi di problemi, come già detto precedentemente [Cap. 2, paragrafo 2.3], possono essere imputabili allo strumento o alla reazione di sequenziamento.

Analizzare il Raw Data è utile per valutare diverse caratteristiche del processo e/o del segnale: il rapporto segnale-rumore, l'ampiezza e l'andamento del segnale, la giusta riuscita della reazione di sequenziamento e della corsa nei capillari, o anomalie nel segnale (spikes).

Le figure 3.1 e 3.2 mostrate nel Capitolo 3 rappresentano i segnali di un buon sequenziamento. Visualizzando l'intero asse dei tempi del Raw Data (figura 4.1), è possibile osservare che l'andamento e l'ampiezza del segnale rimangono costanti lungo l'asse temporale, l'ampiezza media ha valori che non superano i 5000 nella scala di intensità luminosa, il sequenziamento parte intorno ai 2000 nell'asse temporale, e non ci sono picchi anomali o *spikes*.

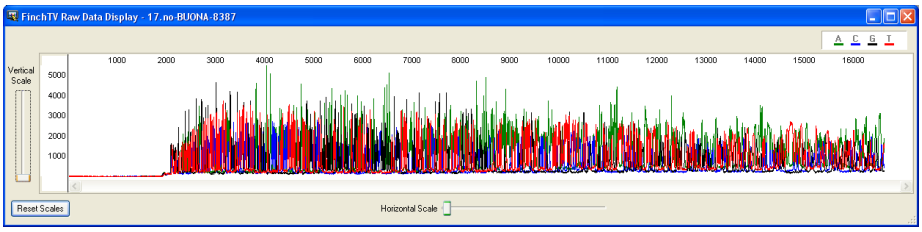


***Figura 4.1**: intero asse temporale del RawData di figura 3.1*

L'A*nalyzed Data* (Figura 3.2) ha anch'esso una buona qualità. Ogni picco ha una forma che somiglia ad una curva gaussiana, e ogni base assegnata è associata ad un solo picco. I *signals strength* [Cap. 3 paragrafo 3.1.1], rientrano nel range dei valori accettabili (deve essere superiore a 100), e il numero di basi sequenziate è 1001. L'esempio mostrato rappresenta il segnale di sequenziamento di un campione di DNA nella cui fase di *preparazione del campione* è stata utilizzata la tecnica del DNA ricombinante.

Quando viene utilizzata la tecnica della PCR, se la dimensione del filamento è inferiore al limite di lettura del sequenziatore, si ottengono invece segnali come quelli riportati in figura 4.2. Si può osservare che il Raw Data non occupa l'intero asse dei tempi, come per i campioni amplificati attraverso plasmidi, e l'ultimo picco in verde rappresenta l'adenina. Questo tipo di preparazione del campione infatti può esser utilizzata per campioni di DNA "corti", con lunghezza inferiore alle mille basi, come invece avviene attraverso le tecniche del DNA ricombinante [Cap. 2 paragrafo 2.1.1].

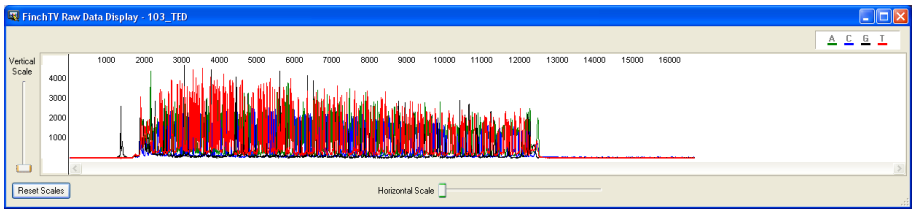


***Figura 4.2**: RawData di un campione di DNA amplificato attraverso la tecnica della PCR. La lunghezza del filamento è minore alla lunghezza massima leggibile dal sequenziatore 3730xl. In genere i campioni preparati con PCR presentano per costruzione un ultimo picco verde rappresentante l'adenina.*

Il processo di preparazione del campione, il funzionamento dello

strumento, la reazione di sequenziamento e la successiva run elettroforetica sono step del processo di sequenziamento che possono interferire nella riuscita del processo stesso. Questi step spesso causano degli errori che impediscono una corretta individuazione delle basi nucleotidiche di cui è composto il filamento di DNA sequenziato. Nei paragrafi successivi verranno descritti i diversi problemi che interessano i segnali di sequenziamento, con le relative cause. Alcuni problemi hanno una gravità differente rispetto ad altri, ed assumono una certa priorità durante l'analisi.

Il flow-chart di figura 4.3 rappresenta le diverse classificazioni del troubleshooting per segnali di sequenziamento Sanger di DNA attraverso il sequenziatore 3730 xl.

Il troubleshooting [Cap. 2 paragrafo 2.3] consiste nell'assegnare ad ogni sequenza una o più classi che identificano la problematica o le problematiche che la caratterizzano. Il primo step fondamentale è quello di valutare la presenza del segnale di sequenziamento nei dati, raccolti durante il passaggio dei frammenti marcati dalla CCD camera del sequenziatore [Cap. 3 par. 3.1]. Quest'analisi riguarda esclusivamente il Raw Data: l'assenza di segnale può esser dovuta all'insuccesso della corsa elettroforetica [Cap. 2 paragrafo 2.1.3], o ad un fallimento della reazione di sequenziamento [Cap.2 paragrafo 2.1.2]. In questi casi, l'esito del sequenziamento viene classificato come “no signal” (*NS)* o “no reaction” (*NR)*, classificazioni che verranno descritte nel dettaglio nel paragrafo 4.2 di questo capitolo. Altre problematiche sono legate all'ampiezza e all'andamento del Raw Data (paragrafo 4.3): a seconda dei casi, il segnale potrebbe esser classificato come “segnale basso” *(SB)*, “segnale alto” *(SA)*, “fuori scala” *(FS)*, segnale che “muore” *(M)*, segnale con “struttura” *(ST)*. La classe “segnale inarcato” *(SI)*, a se stante, verrà descritta nel paragrafo 4.4. La presenza di picchi multipli nell'Analyzed Data (paragrafo 4.5) rappresenta una terza tipologia di problematiche che classifica il segnale come “doppio” *(D)*, “fondo doppio” *(FD)*, “diventa doppio” *(DD-)*, “tratto doppio” *(TD-)*. Una quarta tipologia, relativa alla presenza di picchi anomali sporadici lungo l'Analyzed Data

(paragrafo 4.6), raggruppa le problematiche “dimeri di primer” *(DP)*, “blob” *(BL)*, “problemi di risoluzione” *(PR)*, e “agglomerati non incorporati” *(DG)*.

Queste classificazioni non sono presenti in letteratura, ma sono state introdotte e adottate dai biologi della BMR Genomics per riconoscere l'insieme delle problematiche che caratterizzano un sequenziamento.

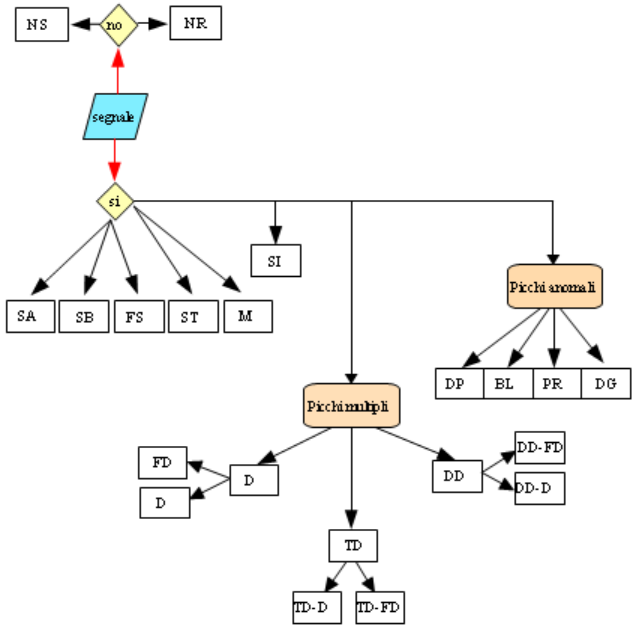


***Figura 4.3**: Rappresentazione degli aspetti relativi ai dati di sequenziamento Raw Data, Analyzed Data del sequenziatore 3730xl che possono compromettere l'identificazione della sequenza, suddivise nelle tre tipologie principali: aspetti legati all'ampiezza e all'andamento del RawData, aspetti legati alla presenza di picchi multipli e anomali nell'Analyzed Data. Solo la classe SI non rientra in nessun gruppo specifico.*

4.2 Problematiche

4.2.1 No signal e no reaction

4.2.1.1 No signal

Nel caso in cui i frammenti di DNA non vengano caricati nei capillari il Raw Data si presenta come in figura 4.3.

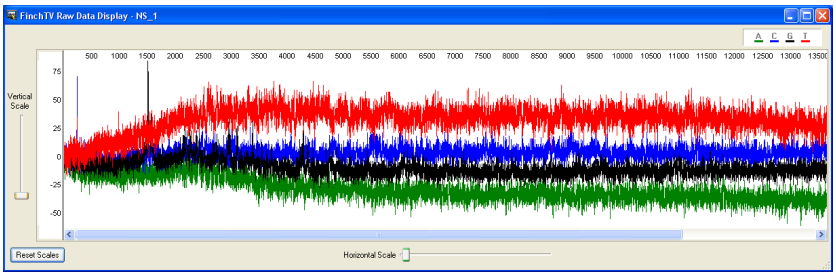


***Figura 4.4**: RawData della sequenza contenuta nel file "NS_1.ab1". Il sequenziamento non ha prodotto nessun segnale utile per l'identificazione delle basi. I dati costituiscono solo rumore di strumentazione. Sequenza NS*

Ciò che viene registrato è solo rumore di fondo, in quanto il rilevatore non ha registrato nessuna emissione di luce. L'Analyzed Data si presenta molto rumoroso, e il software non riesce a riconoscere le basi azotate. Il procedimento di basecalling quindi fallisce.

Questo aspetto dipende dalla strumentazione e non dal processo di reazione di sequenziamento. Nel caso in cui venisse usato un marcatore [Cap 2, paragrafo 3.1], nemmeno questo sarebbe visibile nel Raw Data. Questo tipo di sequenze vengono classificate come *NS*, *no signal*.

4.2.1.2 No reaction

I casi come quello mostrato in figura 4.5 invece vengono classificati come *NR, no reaction.*

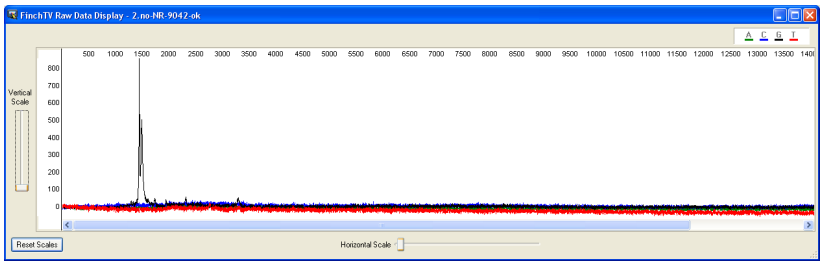


***Figura 4.5**: RawData della sequenza contenuta nel file "2.no-NR-9042-ok.ab1". I dati registrano il picco del marcatore intorno ai 1400-1500 valori dell'asse x. Il resto dell'acquisizione non contiene segnale, ma solo rumore di fondo. La reazione di sequenziamento non ha prodotto frammenti marcati, per cui non sono stati caricati nei capillari del sequenziatore. Sequenza NR.*

Il marcatore viene rilevato. Ciò sta ad indicare che la mixture viene

caricata nei capillari, ma la reazione di sequenziamento non ha prodotto i frammenti marcati con le molecole di fluorescente che non vengono quindi rilevate e il segnale presenta solo rumore di fondo.

Le cause di una mancata reazione di sequenziamento possono essere di diverso tipo :

- fallimento della fase di primer annealing durante la reazione di sequenziamento;
- bassa concentrazione di DNA;
- presenza di contaminanti che hanno ostacolato l'innesco della reazione;
- bassa concentrazione di primer o una sua errata progettazione;
- troppi o pochi reagenti;
- fallimento del ciclo termico della reazione di sequenziamento(denaturazione-primer annealing);
- fallimento dell'estensione dei frammenti;
- prodotti non risospesi.

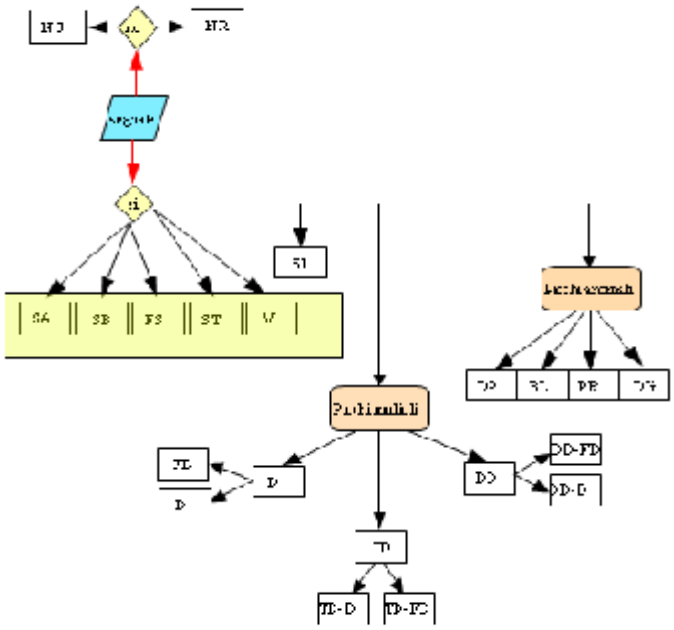


Figura 4.6: *Rappresentazione degli aspetti relativi ai dati di sequenziamento Raw Data, Analyzed Data del sequenziatore 3730xl che possono compromettere l'identificazione della sequenza, suddivise nelle tre tipologie principali. L'area evidenzia le classi di problematiche che interessano l'ampiezza e l'andamento del RawData*

4.2.2 Problematiche legate all'ampiezza e all'andamento del Raw Data

In questo paragrafo vengono presentate le problematiche legate all'ampiezza e all'andamento del Raw Data con i relativi effetti nell'Analyzed Data, tali da rendere nei casi peggiori, impossibile il processo di basecalling. Un sequenziamento andato a buon fine presenta un Raw Data che ha un andamento regolare, costante, e che non supera determinati valori nella scala di intensità luminosa. Questo tipo di problematiche sono raggruppate nel riquadro evidenziato nel flow chart di figura 4.6, mentre nei successivi paragrafi verranno descritte nel dettaglio le caratteristiche principali di ogni classe, con le relative cause.

4.2.2.1 Segnale basso

Dopo la fase di preparazione del campione, è importante eliminare sali residui, proteine, detergenti e residui di RNA. La presenza di queste molecole inibisce la reazione di sequenziamento o interferisce sul processo dell'elettroforesi.

La presenza di proteine interferisce durante il processo dell'elettroforesi perché queste tendono ad attaccarsi alla parete dei capillari, ostacolando e rallentando la corsa dei frammenti di DNA marcati. I contaminanti invece possono degradare il campione di DNA: i frammenti prodotti quindi sono in numero scarso per generare un segnale significativo. Anche una bassa concentrazione di DNA o primer nella reazione di sequenziamento riduce la potenza del segnale. Tutto ciò si manifesta come una registrazione di un segnale grezzo "debole", a volte anche molto rumoroso da rendere impossibile nell'Analyzed Data la distinzione dei picchi e di conseguenza la lettura della sequenza nucleotidica. È questo il caso di sequenze classificate come *segnale basso (SB)*.

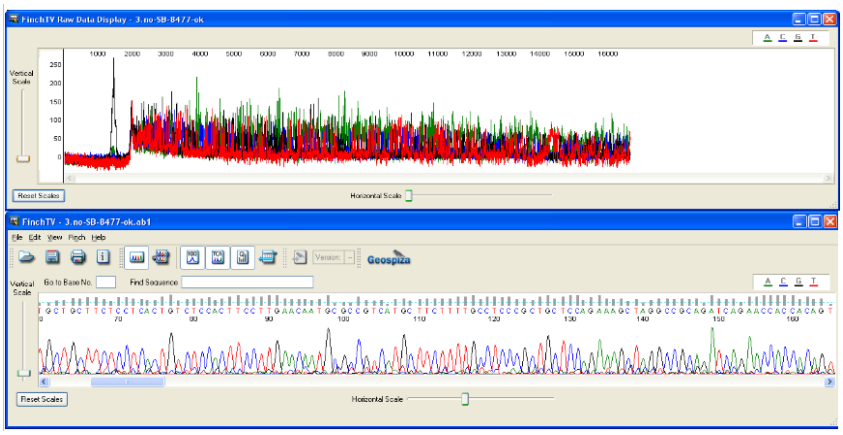


Figura 4.7: *RawData ed Analyzed Data di una sequenza SB. Il segnale grezzo è debole e nella trasformazione in Analyzed Data è presente rumore di fondo che compromette la "pulizia" del singolo picco. Il software riesce ad identificare le basi, ma la qualità dell'identificazione non è ottima, come mostrano le barre grigie dei quality scores associate alla sequenza.*

4.2.2.2 Segnale alto e fuori scala

Una quantità elevata di DNA e primer nella reazione di sequenziamento genera un grande numero di frammenti corti, per cui il segnale relativo alle prime basi risulta molto alto (è il caso di sequenze classificate come *segnali alti SA* illustrato in figura 4.8) e in certi casi è così elevato che supera la scala di intensità massima che lo strumento può rilevare, che nel caso del sequenziatore 3730xl è pari a 32000. Quest'ultimo è il caso delle sequenze classificate come *fuori scala, FS,* figura 4.9 .

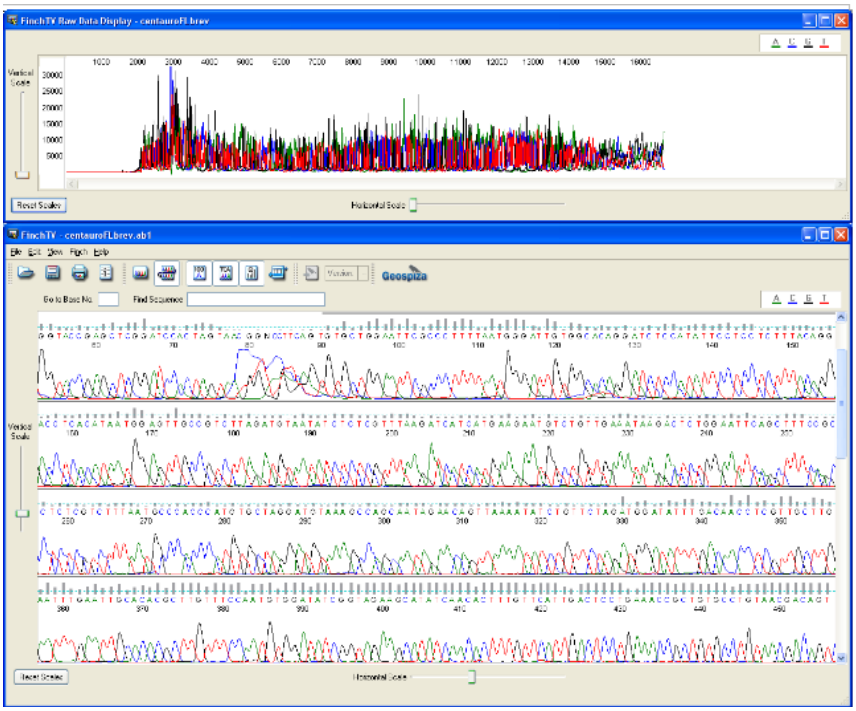


Figura 4.8: *Raw Data ed Analyzed Data di una sequenza alta (SA). Il Raw Data ha un ampiezza che supera i 5000 valori della scala di intensità luminosa, ma non satura. Questo causa delle anomalie nell'Analyzed Data. I singoli picchi non hanno una forma delineata, soprattutto nelle prime e ultime basi.*

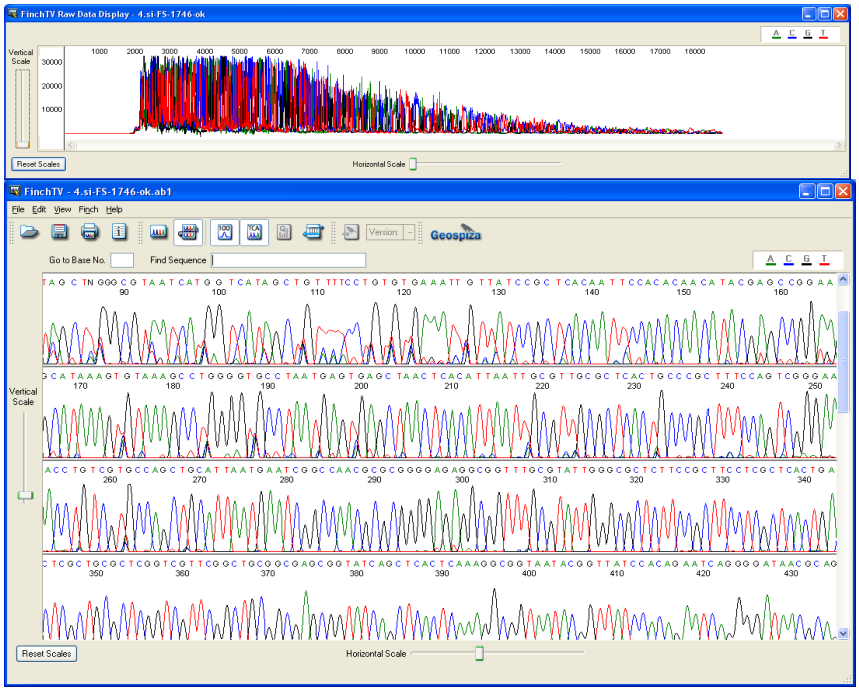


Figura 4.9: *Raw Data ed Analyzed Data di una sequenza fuori scala (FS). Questo è un effetto dovuto all'incapacità dello strumento di registrare una luminescenza delle molecole così intensa da superare la scala di intensità luminosa massima del sequenziatore 3730xl, per cui il segnale satura. Genera picchi multipli nelle singole posizioni dell'Analyzed Data. In genere il FS può interessare solo le prime 200-300 basi del campione di DNA.*

4.2.2.3 Segnale che cala gradualmente

Nel caso in cui vi sia un'alta concentrazione di DNA, i frammenti lunghi prodotti dalla reazione di sequenziamento potrebbero accumularsi all'interno dei capillari ostacolando l'elettroforesi. Durante questo processo anche gli ioni negativi presenti nei sali possono interferire con l'aspirazione dei frammenti più lunghi, che quindi giungono in numero minore alla finestra di rilevazione. Così la potenza del segnale diventa sempre più debole all'aumentare della lunghezza del frammento, l'andamento del segnale cala gradualmente, e certe volte giunge a "morire". Per cui diventa difficile interpretare l'elettroferogramma e identificare correttamente la base azotata. La presenza di RNA causa gli stessi effetti in

quanto entra in competizione con la corsa dei frammenti di DNA durante l'aspirazione.

Lo stesso fenomeno si manifesta nel caso di campioni di DNA che hanno un'alta concentrazione di nucleotidi contenenti G o T. In questi casi il DNA è difficile da sequenziare usando le condizioni standard di reazione. Questo è probabilmente dovuto alle alte temperature di denaturazione necessarie per consentire lo svolgimento della doppia elica di DNA durante la reazione di sequenziamento [Cap. 2 paragrafo 2.1.2]. Per questi campioni utilizzando temperature di denaturazione inferiori ai 95°C, possono verificarsi due casi: è possibile che il DNA venga denaturato, ma non completamente, oppure che il DNA mantenga la struttura completa a doppia elica. Nel primo caso il primer riesce ad agganciarsi al filamento campione, la DNA polimerasi sintetizza i nuovi nucleotidi, ma non riesce a proseguire nel momento in cui incontra il filamento non svolto. Nel secondo caso invece il primer non riesce completamente ad agganciarsi, la fase di primer annealing fallisce, e la DNA polimerasi non può sintetizzare un nuovo filamento marcato (fallimento della fase di copia del filamento e di terminazione). Per entrambi i casi, alla fine della reazione si hanno pochi frammenti marcati, e il segnale rilevato dallo strumento è debole. Queste situazioni influiscono sull'ampiezza e anche sull'andamento del dato grezzo, causando per l'appunto un calo graduale del segnale nel caso in cui la doppia elica del DNA venga denaturata solo parzialmente (sequenza che *muore M*, figura 4.10), oppure ad un segnale basso, caso già presentato precedentemente, nel caso in cui molte copie dello stampo di DNA non vengano completamente denaturate (figura 4.7).

Aspetti relativi alla reazione di sequenziamento, per l'appunto la difficoltà di svolgere particolari sezioni del campione, e aspetti che interessano lo strumento, e in particolar modo la corsa elettroforetica, possono compromettere la corretta acquisizione dei dati, generando quindi i segnali appena descritti.

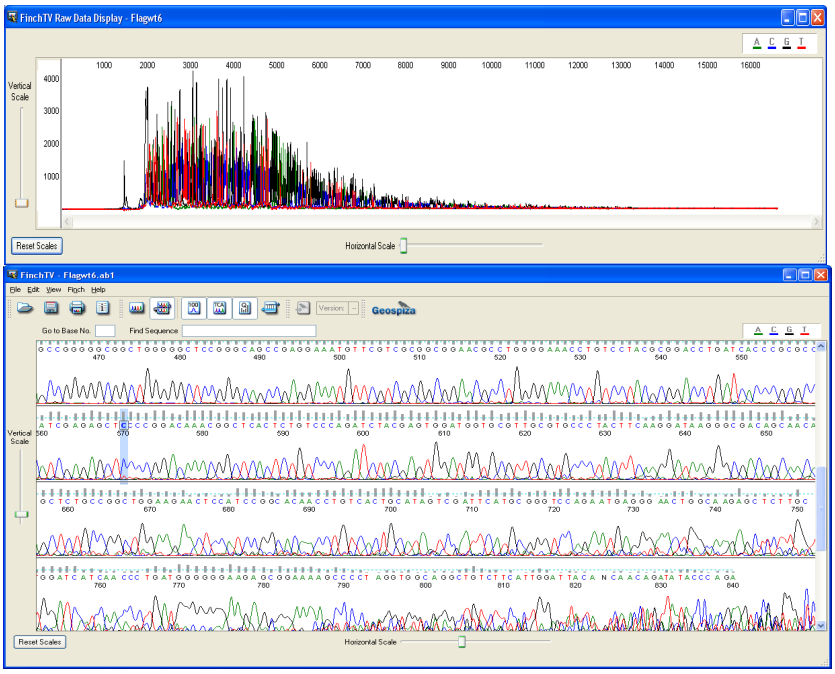


Figura 4.10: *Raw Data ed Analyzed Data di una sequenza che cala gradualmente. L'elettroferogramma risulta più corto, e non si riesce a sequenziare l'intero filamento di DNA. L'intensità dei dati grezzi relativi alle prime basi è nella norma (fino ai 5000 campioni dell'asse delle ascisse), mentre successivamente il segnale cala. In corrispondenza a questo calo aumenta il rapporto segnale-disturbo. Dalla 560° base in poi, il rumore diventa significativo: l'interpretazione dei dati diventa difficile, ed il basecalling ha un quality scores basso.*

4.2.2.4 Segnale che cala repentinamente: strutture secondarie

Nel paragrafo 3.2.2.3 si è visto che oltre a valutare l'ampiezza del Raw Data, è necessario analizzare anche il suo andamento: se si mantiene costante (condizione ottimale), o se cala gradualmente come nel caso delle sequenze classificate come *M*. Un altro problema consiste nel calo repentino del RawData.

Il calo repentino del segnale è dovuto generalmente alla presenza di strutture secondarie nel campione. La sequenza di nucleotidi complementari lungo lo stesso filamento del DNA può far si che questo si ripieghi su se stesso.

Figura 4.11*: Esempio di un filamento di DNA che può generare una struttura secondaria.*

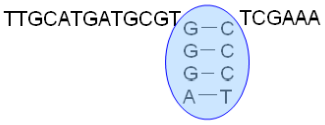


Figura 4.12: *Appaiamento di 4 nucleotidi lungo il singolo filamento di DNA. Formazione di una struttura secondaria.*

Nell'esempio riportato in figura 4.11 il filamento presenta la sequenza "GGGATCCC": le prime quattro basi azotate sono complementari alle successive quattro.

Il filamento ripiegandosi su se stesso, impedisce alla DNA polimerasi di continuare la sintesi di nuovi nucleotidi durante la fase di copia della reazione di sequenziamento [Cap. 2 paragrafo 2.1.2], per cui si stacca, dopo aver sintetizzato un nuovo frammento che può non contenere il dideossinucleotide terminale marcato. Alla fine della reazione si avrà un certo numero di frammenti marcati corti (di lunghezza inferiore al numero di basi che precedono la zona di appaiamento secondario, inferiore a 14 basi nell'esempio illustrato), ed un numero ridotto di frammenti lunghi marcati (sintetizzati da quei filamenti che non si ripiegano).

In figura 4.13, la sequenza presenta infatti un andamento regolare per le prime basi, fino a circa 5500 campioni dell'asse dei tempi, successivamente si assiste ad un calo repentino del segnale, che assume valori molto ridotti rispetto la precedente porzione del RawData.

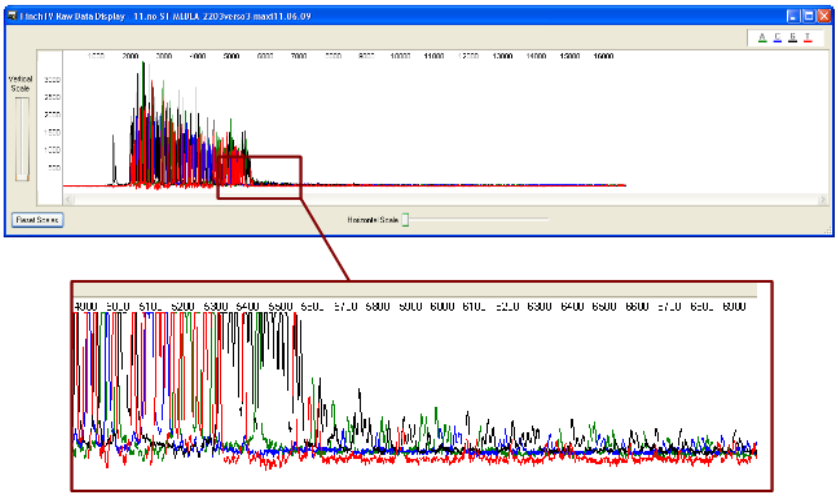


Figura 4.13: *RawData di una sequenza ST. Il segnale subisce un calo repentino, mostrato nel riquadro in rosso.*

L'Analyzed Data si presenta come in figura 4.14. Il filamento contiene una struttura secondaria, e la sequenza viene riconosciuta come *ST (struttura secondaria).*

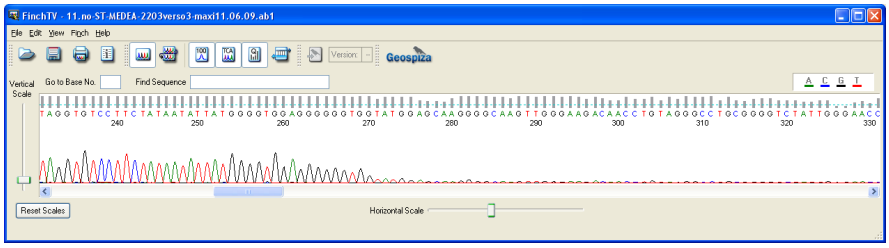


Figura 4.14: *Analyzed Data della sequenza di figura 4.13. In corrispondenza della struttura secondaria i picchi dell'elettroferogramma decrescono fino a scomparire e rendendo difficile il basecalling.*

4.2.3 Segnale inarcato

L'effetto di un segnale inarcato è dovuto alla presenza di contaminanti all'interno della soluzione contenente il campione di DNA. Ciò che avviene è un innalzamento della linea di base del segnale registrato durante il passaggio dei frammenti marcati dalla finestra di rilevamento del sequenziatore. La presenza di queste molecole incrementa il rumore di fondo durante l'acquisizione dei dati, per cui viene sempre registrato un segnale, creando l'effetto cerchiato in figura 4.15.

L'effetto è visibile nel Raw Data di una sequenza, e può interessare i primi dati di fluorescenza (come per l'esempio mostrato), ma anche una finestra più ampia del Raw Data, se non tutto l'asse.

Figura 4.15: *RawData ed Analyzed Data di una sequenza SI. La presenza di contaminanti nella reazione di sequenziamento causa l'effetto visibile nel primo riquadro e interessante i primi picchi acquisiti. Nell'Analyzed Data le prime 30-40 basi non sono caratterizzate da un segnale pulito, è difficile riconoscere i picchi. Per queste basi il quality score assume valori scarsi.*

4.2.4 Picchi multipli nell'Analyzed Data

Si parla di picchi multipli quando ci sono due o più picchi nella stessa posizione dell'Analyzed Data. Le cause possono essere molteplici e interessano sia i processi caratterizzanti la reazione di sequenziamento, sia la composizione e la struttura del campione di DNA d'interesse. L'area evidenziata nel flow chart di figura 4.16 contiene le diverse classificazioni dei picchi multipli, descritte nei punti successivi di questo paragrafo.

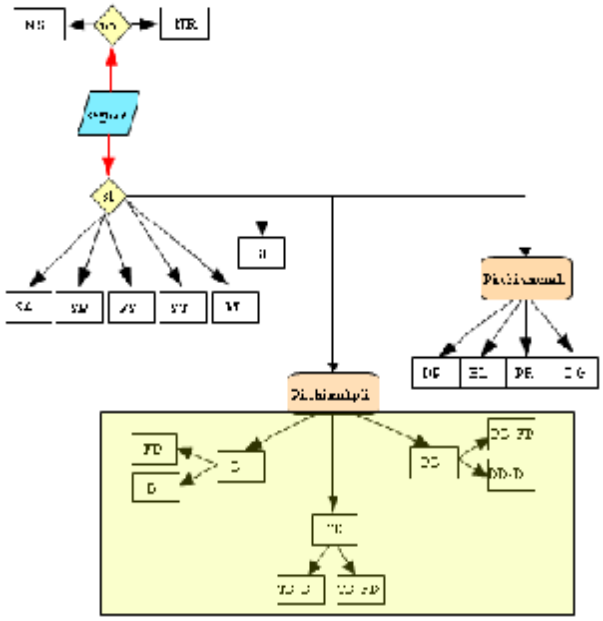


Figura 4.16: *Rappresentazione degli aspetti relativi ai dati di sequenziamento Raw Data, Analyzed Data del sequenziatore 3730xl che possono compromettere l'identificazione della sequenza, suddivise nelle tre tipologie principali. Nel riquadro evidenziato sono messe in risalto le classi relative all'analisi dei picchi multipli nell'Analyzed Data.*

- Una scarsa qualità del campione interferisce durante la reazione di sequenziamento (come si è visto per sequenze *SB, FS, SA o M*). Ciò genera un Analyzed Data che presenta un alto rumore di fondo dovuto all'accavallamento di più segnali nella stessa posizione. Quando questo accavallamento di segnali è talmente importante da impedire l'interpretazione del dato, la procedura di assegnazione della base fallisce e il software del sequenziatore assegna una N (*noise*) in quella posizione del cromatogramma.

  In genere le sequenze SB presentano nell'Analyzed Data un alto rumore di fondo lungo tutto la sequenza e ciò classifica il segnale anche come *FD.* I picchi nell'Analyzed Data sono multipli, e i picchi secondari hanno un'intensità minore rispetto quelli principali.

Figura 4.17: Analyzed Data e Raw Data di una sequenza SB. Si osservano nel primo riquadro picchi multipli, anche se è evidente la presenza di una sequenza dominante. I picchi sono distribuiti lungo l'intero Analyzed Data. La sequenza è SB ed FD.

- La presenza di più di un template nella reazione di sequenziamento causa la lettura contemporanea di più sequenze, che quindi risultano sovrapposte nell'Analyzed Data.

  Il sequenziatore legge contemporaneamente due o più sequenze e tale effetto, che non è visibile nel Raw Data, si manifesta in maniera evidente nell'Analyzed Data.

  I picchi risultano doppi nel caso in cui i campioni di DNA sono due, multipli nel caso di più campioni. I picchi possono avere stessa intensità, ed è il caso di sequenze classificate come *D*, o possono avere intensità differente, con la presenza di una sequenza dominante, ed è il caso delle sequenze riconosciute come *FD*.

  Nelle figure 4.18 e 4.19 sono riportati due casi di picchi multipli: per entrambi i casi i picchi multipli interessano l'intero RawData, ma mentre la prima figura mostra una sequenza D, in quanto i picchi hanno uguale intensità, la seconda figura riporta una sequenza FD, con picchi di intensità differente.

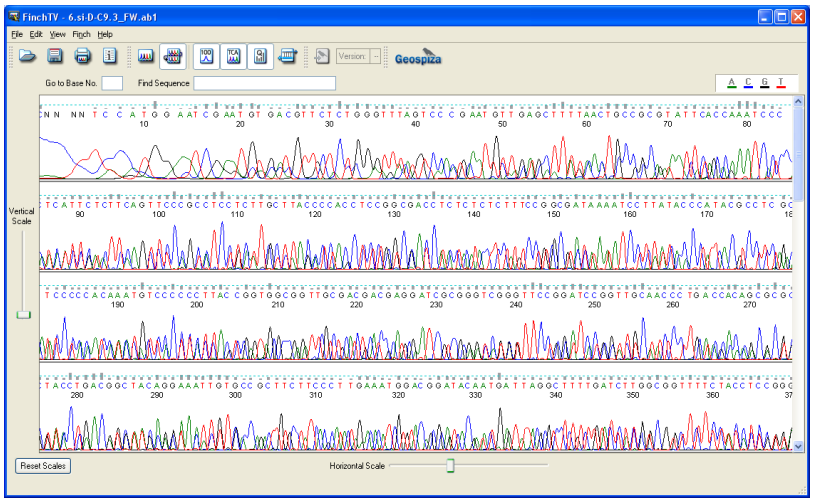


***Figura 4.18**: picchi doppi di uguale intensità lungo l'intero Analyzed Data. La sequenza è D*

***Figura 4.19:** picchi doppi di diversa intensità lungo l'intero Analyzed Data. La sequenza è FD*

Questi problemi ricorrono sia per filamenti di DNA amplificati con la PCR, sia per filamenti clonati attraverso plasmidi. Durante la reazione a catena della Polimerasi infatti è possibile che vengano amplificati filamenti di DNA differenti a causa di un appaiamento aspecifico del primer progettato. Generalmente le procedure di cleanup che seguono la PCR rimuovono i nucleotidi non incorporati e primers residui, ma non i prodotti secondari della PCR.

Nel caso di clonaggio attraverso plasmidi invece, può accadere inavvertitamente, che il DNA venga estratto da più colonie con contenuto plasmidico diverso, anziché da una sola. Il DNA non è unico, ci sono quindi diversi campioni. Per entrambe le situazioni il macchinario legge più di una sequenza contemporaneamente.

- Una ottimale progettazione del primer e il successo della fase di primer annealing della reazione di sequenziamento sono dei processi fondamentali per un giusto esito del sequenziamento. Precedentemente si è visto come il fallimento del primer annealing può causare la formazione di pochi frammenti marcati, ma anche nel peggiore dei casi il totale fallimento della reazione di sequenziamento. Una scorretta progettazione del primer invece porta alla lettura contemporanea di due o più sequenze. Ciò accade quando:
  1. Nella reazione è presente oltre al primer ideale anche un primer che è di una base più corto rispetto l'altro. In questo caso il sequenziamento mostra la sovrapposizione di due sequenze identiche, una shiftata rispetto l'altra.
  2. Più di un primer è presente nella reazione di sequenziamento. Ciò accade quando durante la fase di cleanup successiva alla PCR non vengono rimossi i primer non utilizzati. Se nel template vi è un sito di ancoraggio per i primer spuri, verranno sintetizzati frammenti marcati costituenti sequenze nucleotidiche differenti.
  3. c'è una seconda zona di appaiamento col primer nel template.

A causa delle motivazioni esposte in questi tre punti, è possibile che da un certo punto in poi dell'Analyzed Data si possano presentare dei picchi multipli nella sequenza: i picchi possono avere stessa intensità (la sequenza viene classificata come *DD-D*), o i picchi possono avere intensità differente, con la presenza di una sequenza dominante (*DD-FD*).

Figura 4.20: *Picchi doppi di uguale intensità a partire dalla 221 base dell'Analyzed Data (DD-D)*

Viceversa è possibile riscontrare sequenze che hanno picchi multipli solo nelle prime basi. In questi casi la sequenza viene classificata utilizzando il prefisso *TD-*, e in base all'intensità dei picchi, viene aggiunta la sigla *-D* (uguale intensità), *-FD* (diversa intensità). É il caso mostrato in figura 4.21.

Figura 4.21: *picchi doppi nelle prime 150 basi dell'Analyzed Data. Il picco secondario ha intensità inferiore rispetto a quello principale. Si osserva come i quality scores delle prime basi abbiano valori più bassi rispetto quelli delle basi successive alla 150. La sequenza è TD-FD.*

- Anche la struttura e la composizione del campione di DNA costituiscono un potenziale problema nella procedura di sequenziamento. Si è visto come campioni con un alto contenuto in GC siano difficili da sequenziare a causa delle alte temperature di fusione richieste per la denaturazione della doppia elica del DNA. Campioni di questo tipo devono esser perciò trattati con temperature adeguate, altrimenti il segnale rilevato dallo strumento risulta debole in ampiezza [paragrafo 3.6.1]. Problematico è anche il sequenziamento di campioni che contengono regioni con lo stesso omopolimero (figura 4.22): sequenze classificate come *PA* nel caso di ripetizione di basi A, *PT* nel caso di ripetizione di basi T, *PG* nel caso di ripetizione di basi G, PC nel caso di ripetizione di basi C. La DNA polimerasi slitta di posizione durante la reazione di sintesi di nuovi nucleotidi. Nell'Analyzed Data, la sequenza successiva all'omopolimero presenta picchi multipli. Come nel punto 1., il sequenziamento mostra la

sovrapposizione di due sequenze identiche, una shiftata rispetto l'altra. Il riconoscimento di picchi multipli o doppi a partire da un punto intermedio alla sequenza viene classificato con il suffisso *DD-,* e a seconda dell'entità dei picchi, di uguale intensità, o diversa e con la presenza di un picco dominante, viene aggiunta al suffisso DD- rispettivamente la voce *-D* o *-FD*.

Figura 4.22: *Analyzed Data di un filamento di DNA che presenta una regione con omopolimero. Dopo la 77 base è presente una successione di nucleotidi contenente la base azotata T. Successivamente a tal regione si osservano dei picchi doppi (uno dominante, l'altro secondario) che interessano la restante parte della sequenza (DD-FD).*

4.2.5 Picchi anomali

Questo tipo di problematiche interessano solo alcune regioni dei segnali di sequenziamento; sono raggruppate nell'area evidenziata del flow chart di figura 4.23.

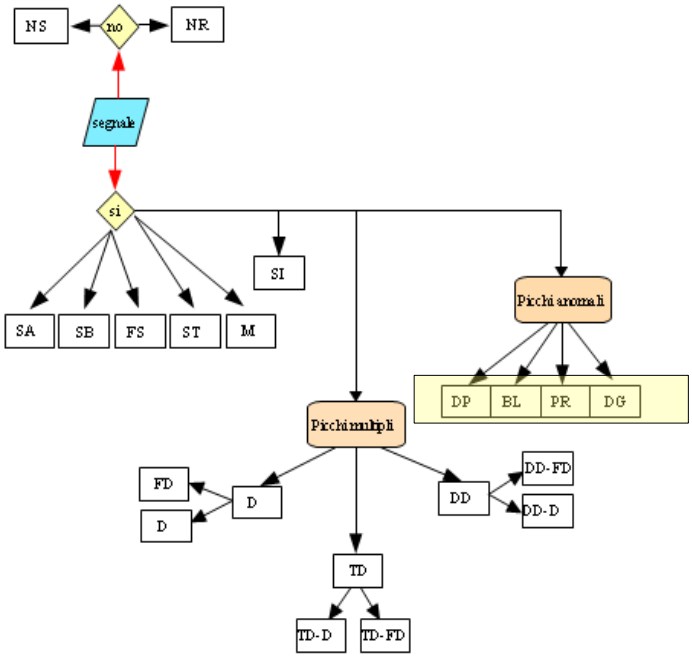


Figura 4.23: *Rappresentazione degli aspetti relativi ai dati di sequenziamento Raw Data, Analyzed Data del sequenziatore 3730xl che possono compromettere l'identificazione della sequenza, suddivise nelle tre tipologie principali. Nel riquadro evidenziato sono messe in risalto le classi relative all'analisi dei picchi anomali nell'Analyzed Data.*

Durante la fase del primer-annealing della reazione di sequenziamento [Cap. 2 paragrafo2.1.2], i primer si agganciano al filamento di DNA da sequenziare. Può accadere che invece dell'appaiamento del primer col filamento campione di DNA, due primer si agganciano tra loro. Quindi la DNA polimerasi sintetizza un filamento che sarà la copia complementare del primer e non del campione.

```
5'  ACTATAGGGCACGCGTGGT
                |||||||||
3'              TGGTGCGCACGGGATATCA
```

Figura 4.24: *esempio dell'appaiamento di due primer identici*

In figura 4.24 è riportato l'appaiamento di due primer. Alla fine della reazione di sequenziamento, si avranno dei frammenti di DNA marcati copia sia del primer che del filamento di DNA da sequenziare. I frammenti copia del primer sono tutti di lunghezza molto piccola, essendo il primer costituito da circa 20 nucleotidi. Ciò si manifesta con la registrazione di un segnale intenso durante il passaggio dei frammenti più corti dalla CCD camera, e quindi i primi picchi visibili nel Raw Data hanno un'intensità maggiore rispetto l'intero segnale. La figura 4.25 rappresenta il Raw Data di una sequenziamento in cui sono visibili i picchi dei dimeri di primer. Questo tipo di problema viene classificato come *DP*, ed è dovuto a fenomeni che coinvolgono la sola reazione di sequenziamento. Anche nell'Analyzed Data è possibile riscontrare la presenza di questi picchi.

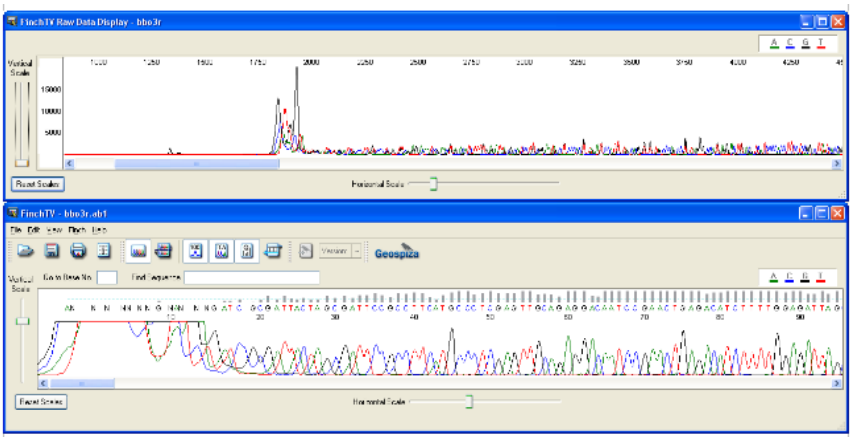


Figura 4.25: *RawData e Analyzed Data di una sequenza di DNA in cui è visibile la presenza di dimeri di primer (DP) visibili nei primi picchi. Si nota come questi abbiano intensità maggiore rispetto al resto del RawData.*

L'effetto della presenza di bolle d'aria, o di cristalli di polimero all'interno della soluzione contenente il campione di DNA, viene registrato durante l'acquisizione dei dati nel processo di elettroforesi capillare. Ciò che si osserva è la presenza di un picco stretto e molto alto, spike, sia nel RawData, che nell'Analyzed Data. Sotto, nelle figure 4.26 e 4.27, sono riportati il RawData e l'Analyzed Data di una sequenza che presenta questo artefatto e che viene classificata nella categoria *BL (blob)*.

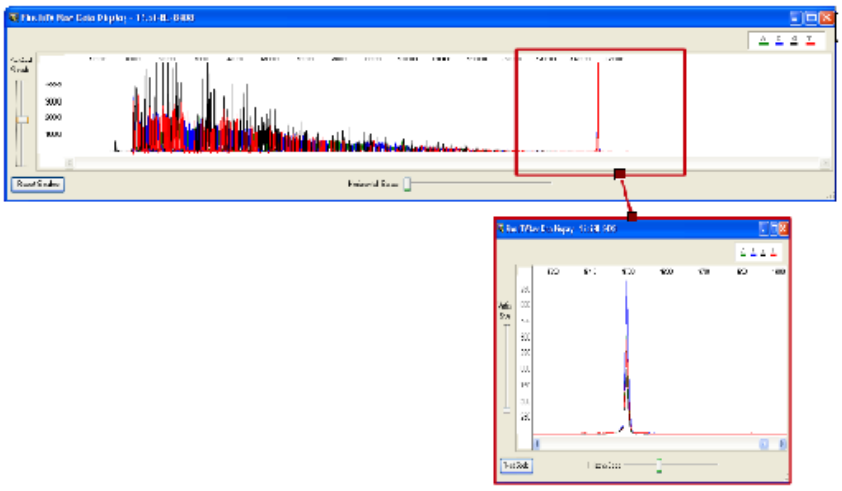


Figura 4.26: *RawData di una sequenza di DNA che contiene uno spike anomalo dovuto alla presenza di bolle d'aria o di cristalli di polimero (BL).*

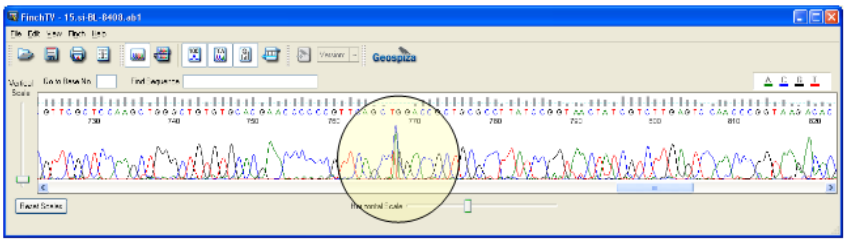


Figura 4.27*: Analyzed Data della stessa sequenza di DNA di figura 4.26, in cui è evidente un picco anomalo.*

Durante la reazione di sequenziamento, i nucleotidi marcati non incorporati possono formare agglomerati che migrano durante il processo di elettroforesi e vengono quindi rilevati dallo strumento. Questi si presentano come dei picchi spuri che compaiono all'interno del segnale, visibili sia nel RawData che nell'Analyzed Data e che occupano delle posizioni ben precise in quest'ultimo: tra la 70-esima e l' 80-esima base, come per il caso riportato in figura 4.28, o tra la 230-esima e la 350-esima base. Ciò è dovuto al fatto che evidentemente questi agglomerati migrano lungo i capillari in funzione del loro volume e giungono al rilevatore in istanti della corsa determinanti quelle posizioni nell'Analyzed Data. In questi casi le sequenze vengono assegnate alla classe *DG*.

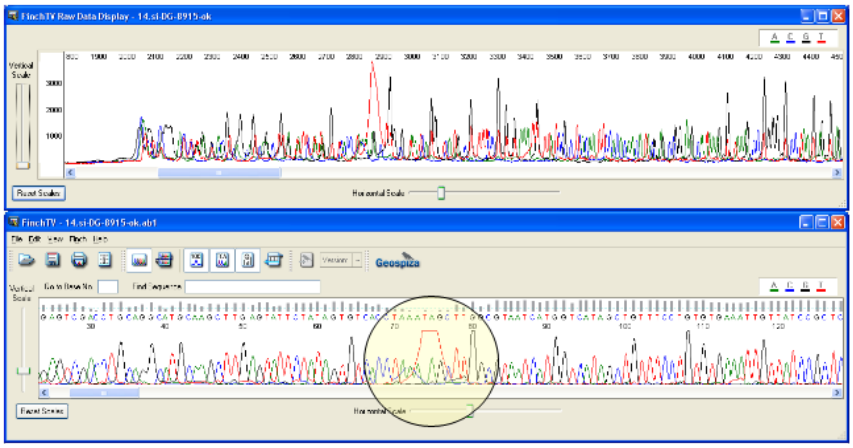


Figura 4.28: *RawData ed Analyzed Data di una sequenza di DNA che contiene uno spike anomalo dovuto al rilevamento di agglomerati dei nucleotidi marcati non incorporati (DG).*

La presenza di picchi spuri rende difficile l'identificazione della base azotata, e sia nella figura 4.27, che nel secondo riquadro della figura 4.28, si può osservare un basso quality score delle basi che contengono le anomalie appena descritte.

In gran parte dei segnali di sequenziamento, è possibile osservare che sia i primi picchi contenenti le 30-40 basi del cromatogramma, che gli ultimi picchi nell'Analyzed Data non hanno una forma ben delineata (soprattutto per campioni di DNA amplificati tramite DNA ricombinante): questi non somigliano più ad una curva gaussiana, e diventano molto irregolari nell'ampiezza e nella forma. Quando l'irregolarità dei picchi interessa gran parte dell'ultima porzione del segnale, la sequenza viene riconosciuta come *PR,* problemi di risoluzione.

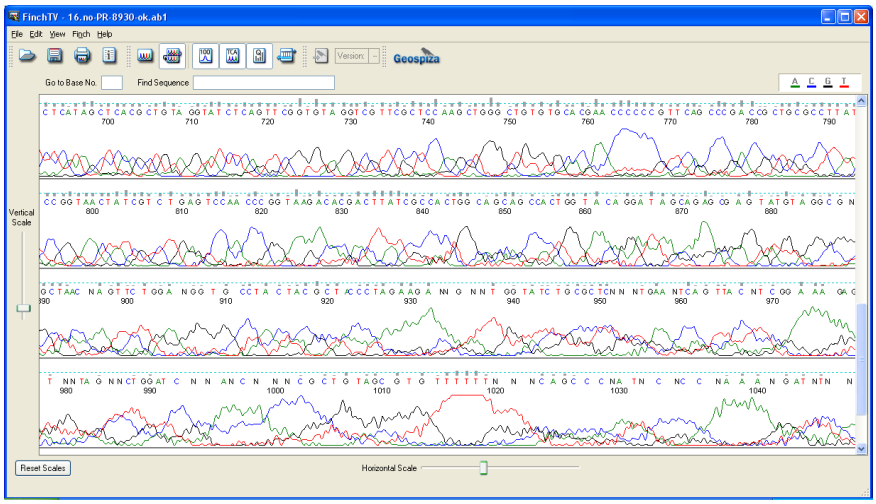


Figura 4.29: *Analyzed Data di una sequenza di DNA con problemi di risoluzione (PR).*

Viene mostrata di seguito la tabella 4.1, riportata nella guida Applied Biosystems 3730/3730xl, "Sequencing Chemistry Guide" che descrive le diverse problematiche relative al sequenziamento del DNA: nella prima colonna vengono riportate le problematiche, mentre nella seconda e terza colonna della tabella vengono descritte rispettivamente le possibili cause della problematica e le tecniche per correggere il problema. In tabella 4.2 vengono riportati i simboli associati ad ogni problematica.

Observation	Possible Cause	Recommended Action
Poor data resolution	Clogged capillary array caused by an excess of protein, template, other sample impurities, or dried polymer	Replace the array.
	• Degradation of samples in formamide • Degradation due to formamide exposed to air	Re-prepare the samples.
	Overloading of the sample	Dilute the sample and adjust the injection parameter. Refer to "Optimizing Electrokinetic Injection" on page 4-5.
Weak signal	Quantity of template or primers in the sequencing reaction or the quantity of sample injected too low	Refer to "Template Quantity" on page 2-5 for a description of recommended template quantities.
		If possible, resuspend the template in a smaller volume.
		Increase injection time. Refer to "Optimizing Electrokinetic Injection" on page 4-5.
	Excess salt present in the sample	Clean up the sample using a spin column or a 70% ethanol wash.
	Bad post-reaction clean-up	Repeat sample preparation.
High background	Dirty template, bad primers, bad post-reaction clean-up	Refer to the documents listed on page vii of the Preface section for a description of how to clean up dirty templates.
Top-heavy data	Amount of template in the sequencing reaction too high, creating an excess of short fragments that are preferentially injected into the capillary array	Refer to "Template Quantity" on page 2-5 for a description of recommended template quantities.
	Concentration of extension products too high	Dilute the sample or decrease the injection time.
	Diluted reactions	Use more BigDye reagent.

Blank lanes or no signal	Cycle sequencing reaction failed	Repeat the cycle sequencing reaction, adjust primer and template concentration.
	Bad post-reaction clean-up	Repeat sample preparation.
	Blocked capillary array caused by an excess of protein, template, or other impurities, or by dried polymer	Replace the capillary array.
Failed injection	Reaction plate not centrifuged prior to injection, air bubbles in the sample wells	Centrifuge the reaction plate.
Breakdown of BigDye® G nucleotide	Formamide degradation caused by exposure to the air	Refer to "Problems with Commercial Formamide" on page A-2. Use formamide as recommended in Appendix A.
		Cover reaction plates with septa or film.
Abrupt signal loss	Reactions too dilute	Repeat run using more BigDye reagent.
	Poor quantitation of primer and/or template, leading to top-heavy data	Adjust concentrations and repeat reactions.
Poor template quality	Residual salts or organic chemicals carried over from template preparation	Precipitate the template with ethanol and resequence. Refer to Chapter 3, "Purifying the Extension Products."
	Incomplete removal of cellular components such as RNA, proteins, polysaccharides, and contaminating chromosomal DNA	
	Degradation of DNA in storage	
	More than one template DNA in the sequencing reaction	
Inhibition of the sequencing reaction	Various types of contaminates present during template preparation	Precipitate the template with ethanol and resequence. Refer to Chapter 3, "Purifying the Extension Products."
Multiple, overlapping sequences in the data (PCR templates)	More than one template present in the reaction (*i.e.*, secondary PCR products) due to lack of specificity	The majority of cleanup procedures for PCR products are designed to remove unincorporated nucleotides and residual PCR primers, not secondary PCR products. Use agarose gel electrophoresis to detect the presence of secondary PCR products.
		Optimize the PCR conditions and/or use a Hot Start method.
		Purify the PCR products using a gel before sequencing.

Multiple, overlapping sequences in the data (cloned DNA templates)	More than one sequence present in the reaction due to mixed plaques or colonies	Re-isolate the DNA from a pure colony and re-sequence.
		When picking bacterial colonies for growth and DNA isolation, choose a colony that is well isolated.
		With M13 plaques, use fresh plates for plaque picking. Check the DNA purity by running it on an agarose gel.
Multiple peaks in the same position at some points (*pull up peaks* or *bleed-through*)	Very strong signals saturating the instrument's detector, causing the signals to be truncated The Sequencing Analysis software underestimates the amount of signal at these positions, therefore underestimating the amount of spectral overlap to correct.	Very strong signals are common when sequencing short PCR fragments, because the sequencing reaction is often very efficient. You may need to load less of this type of sample to compensate for the increased signal.
Excess dye peaks	Incomplete removal of unincorporated, fluorescently labeled dcNTPs during alcohol precipitation	Use only room-temperature alcohol. Cold alcohol will also precipitate unincorporated dye terminators. Do not use denatured alcohol. Denatured alcohol has inconsistent quality. The concentration of the alcohol and purity of the additives can vary.
		Use the concentration of alcohol recommended in the precipitation procedures.
		Use a precipitation method appropriate for your sequencing chemistry.

Observation	Possible Cause	Recommended Action
Difficulty sequencing GC-rich templates, resulting in weak signa	The DNA is melting at a higher temperature due to the high proportion of GC base pairs Note: Even a template that has a fairly average base composition overall can have a very GC-rich region that effects its ability to be sequenced.	Increase the denaturation temperature.
		Add DMSO to a final concentration (v/v) of 5%.* Note: Adding a mixture of 5% DMSO and 5% glycerol has also been used successfully for some templates.
		Incubate the reaction at 95 °C for 10 mn before cycling.
		Add betaine to a final concentration of 1 M.†
		Double all reaction components and incubate at 98 °C for 10 min before cycling.
		Add 5 to 10% formamide or 5 to 10% glycerol to the reactions.
		Linearize the plasmids with a restriction enzyme.
		Shear the insert into smaller fragments (<200 bp) and subclone.
Secondary structure in the template making it difficult to obtain good sequencing data beyond the region of secondary structure	Self-annealing DNA	Increase the denaturation temperature.
		Add DMSO to a final concentration (v/v) of 5%.‡ Note: Adding a mixture of 5% DMSO and 5% glycerol has also been used successfully for some templates.
		Incubate the reaction at 98 °C for 10 min before cycling.
		Add betaine to a final concentration of 1 M.§
		Double all reaction components and incubate at 98 °C for 10 min before cycling.
		Add 5 to 10% formamide or 5 to 10% glycerol to the reactions.
		Linearize the plasmids with a restriction enzyme.
		Shear the insert into smaller fragments (< 200 bp) and subclone.
Slippage in the region of the homopolymer (DNA sequencing reactions)	Long homopolymer T (or A) regions	Use an anchored primer (i.e., a sequencing primer that is polyT containing an A, C, or G base at the 3′ end of a polyA region). The 3′ base will anchor the primer into place at the end of the homopolymer region.
	Use of dUTP in the deoxynucleotide mixture	
	Unknown	

Tabella 4.1: *troubleshooting del sequenziamento di DNA, tabella riportata nella guida "Sequencing Cheministry guide" dell'Applied Biosystems 3730/3730xl*

Problematiche	***Simbolo***
Assenza di segnale	*NS*
	NR
Problematiche legate all'ampiezza del Raw Data	*SB*
	SA
	FS
Problematiche legate all'andamento del Raw Data	*M*
	ST
Linea di base inarcata	*SI*
Picchi multipli	*D*
	FD
	DD-
	TD-
Picchi anomali	*DP*
	BL
	PR
	DG

Tabella 4.2: *simbologia associata alle problematiche*

Capitolo 5

Troubleshooting: algoritmi e soluzioni proposte

Nel capitolo 4 sono state descritte tutte le problematiche che possono generare errori nella determinazione della sequenza delle basi azotate di un campione di DNA sottoposto a sequenziamento Sanger.

Il troubleshooting è la procedura di analisi dei dati di sequenziamento necessaria per capire se bisogna o meno ripetere il procedimento cambiando eventualmente qualità o quantità del template, cioè del campione biologico da processare. L'osservazione dei dati viene compiuta da biologi esperti che analizzano visivamente il risultato e l'andamento dei segnali forniti dal sequenziatore. L'analisi consiste nel riconoscere particolari problematiche nei dati, che possono compromettere l'identificazione delle basi azotate del campione. I segnali Analyzed Data e Raw Data (che in seguito verranno chiamati *AD* e *RD* rispettivamente) [Cap. 3 paragrafo 3.1.1] devono per ciò rispettare delle caratteristiche ben precise.

Lo scopo di questa tesi è stato quello di automatizzare l'analisi dei dati in modo di fornire un algoritmo che sia un supporto efficiente e veloce al troubleshooting.

Nel capitolo 4 è stato presentato in figura 4.3 il flow chart con le principali

classificazioni del troubleshooting. Viene riportato lo stesso diagramma in figura 5.1, con in evidenza (in verde) le analisi realizzate dall'algoritmo proposto e descritto in questo Capitolo.

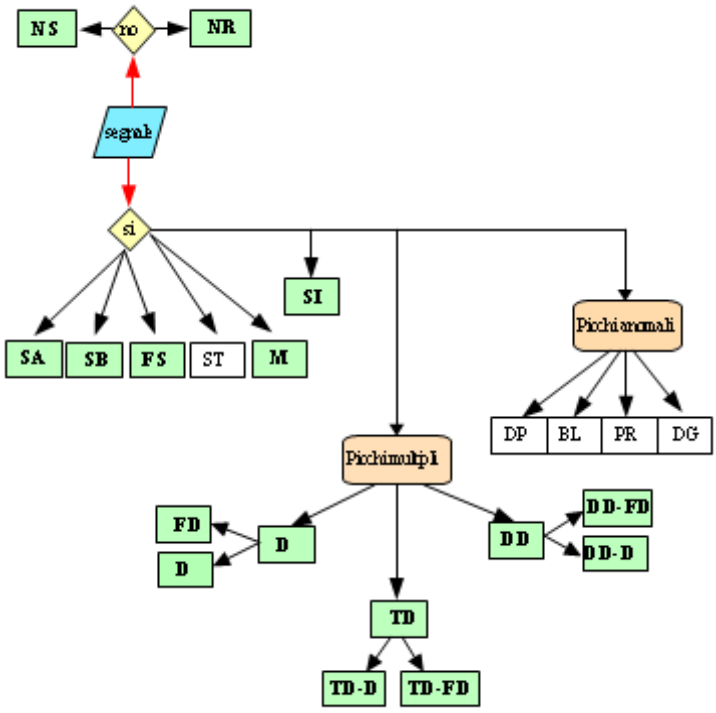


Figura 5.1: *Classificazione degli aspetti relativi ai dati di sequenziamento Raw Data, Analyzed Data del sequenziatore 3730xl che possono compromettere l'identificazione della sequenza, suddivise nelle tre tipologie principali. L'algoritmo è stato realizzato per riconoscere le classi evidenziate in verde.*

Le figura 5.2 e 5.3 riportano i segnali ricavati dal sequenziamento di un campione di DNA contenuti nel file ABIF "s_5f.ab1". La prima mostra il RD, mentre la seconda l'AD della sequenza. Dal RD si osserva come il campione di DNA sia stato preparato attraverso la tecnica della PCR [Cap. 2 paragrafo 2.1.1], in quanto il segnale non occupa l'intero asse dei tempi e presenta un ultimo picco finale di colore verde (Adenina) [Cap. 3 paragrafo 3.2]. L'ampiezza del segnale è alta, quindi è classificato come *SA* [Cap. 3 paragrafo 3.2.2.2], ed è inarcato, *SI* [Cap. 3 paragrafo 3.2.3]. L'AD è invece caratterizzato da picchi multipli lungo tutto il segnale, la sequenza risulta anche *FD* [Cap.3 paragrafo 3.2.3].

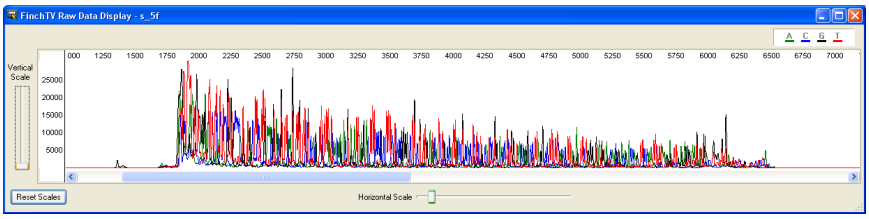


***Figura 5.2**: RD ottenuto dal sequenziamento di un filamento di DNA Il segnale è alto, SA, ed inarcato, SI. Il campione è stato amplificato attraverso la tecnica della PCR, e lo si coglie dall'ultimo picco verde che rappresenta l'adenina, e dal fatto che il segnale non occupa l'intero asse temporale.*

***Figura 5.3**: AD ottenuto dal sequenziamento dello stesso campione di DNA della figura 5.2. Si osservano picchi multipli lungo tutta la sequenza.*

L'algoritmo riesce ad individuare tutti e tre gli aspetti che riguardano la sequenza e in figura 5.4 viene mostrato l'output dell'analisi automatica. L'algoritmo:

- riconosce la presenza del marcatore, *m(t)*, nel RD;
- individua l'istante in cui la CCD camera del sequenziatore inizia a rilevare i picchi del segnale di sequenziamento[Cap. 2, paragrafo 2.1.3], *y(t)*;
- riconosce che il campione è stato preparato attraverso la tecnica della PCR;
- riconosce la presenza di picchi multipli nell'AD;
- individua un RD inarcato;

- riconosce un calo del RD, ma non così drastico da portarlo a morire.

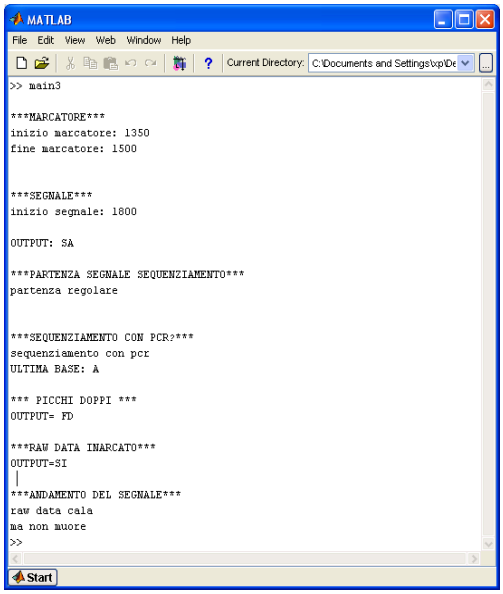


Figura 5.4: *risultato del troubleshooting automatico per la sequenza "s_sf.ab1"*

Nel flow chart (figura 5.5) sono rappresentati gli step su cui è articolato l'algoritmo e nei paragrafi successivi verranno descritte nel dettaglio le specifiche di ogni step.

L'algoritmo è stato implementato in Matlab (Matrix Laboratory), un software che utilizza un linguaggio ad alto livello, molto utilizzato nel settore del calcolo ingegneristico e della simulazione. La logica con cui opera è una logica prettamente matriciale; inoltre, essendo un metalinguaggio, Matlab consente l'utilizzo di molte funzioni predefinite senza che l'utente debba preoccuparsi di programmare a basso livello. Matlab viene spesso utilizzato per l'analisi numerica, per la modellazione di sistemi dinamici, per lo sviluppo di algoritmi, l'elaborazione grafica e l'analisi del segnale. Tutte queste caratteristiche lo rendono facile ed efficiente ed è per questo che è stato utilizzato per implementare l'algoritmo di questa tesi. Per questo lavoro sono state realizzate 11 funzioni. In un unico script invece sono stati memorizzati i parametri utilizzati nell'algoritmo (come le soglie decisionali), e che possono esser modificati da chi utilizza il

software. Il software riporta il risultato dell'analisi per ogni singola sequenza in una interfaccia (figura 5.4), e in diversi plot viene mostrato come il segnale viene elaborato, fornendo all'operatore uno strumento in più per la valutazione del risultato. Nel cd allegato alla tesi, è contenuto il codice dell'algoritmo, e tre file ABIF che rappresentano tre sequenze esempio (dati forniti dall'azienda BMR Genomics): due caratterizzate da problematiche, una invece che non ne contiene.

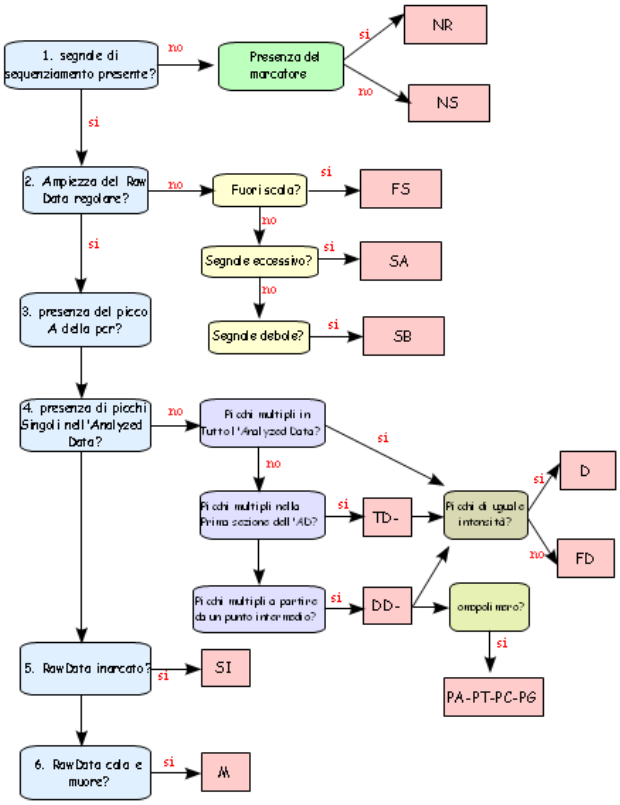


Figura 5.5: *flow chart rappresentante gli step dell'algoritmo che effettua il troubleshooting automatico.*

5.1 No signal e no reaction

Ogni qual volta si debba analizzare l'esito di un sequenziamento, è necessario innanzitutto capire se i frammenti marcati vengano caricati nello strumento [Cap.2 paragrafo 2.1.3]. Il primo passo da compiere è quello di verificare la presenza del marcatore [Cap.3 paragrafo 3.1] e la successione dei picchi di sequenziamento nel

dato grezzo.

Per quest'analisi sono state realizzate due funzioni chiamate `presenza_marcatore` e `presenza_segnale`. La prima serve, come dice il nome stesso, a verificare la presenza del marcatore nel RD.

Nel caso in cui si usasse il marcatore, come avviene per la maggior parte delle sequenze analizzate dalla BMR Genomics, questo si presenta come una forma d'onda caratterizzata dalla successione di due picchi che ricoprono una finestra di circa 150 campioni in corrispondenza dei 1200-1500 valori dell'asse temporale, e assume valori d'ampiezza variabili dai 300 ai 3500 nella scala d'intensità luminosa. Come già spiegato, questo non è altro che un breve filamento di DNA avente il primo nucleotide marcato con lo stesso fluorocromo utilizzato per marcare i nucleotidi che contengono la G, per cui assume la stessa colorazione di questa base (nera), come è mostrato in figura 5.6.

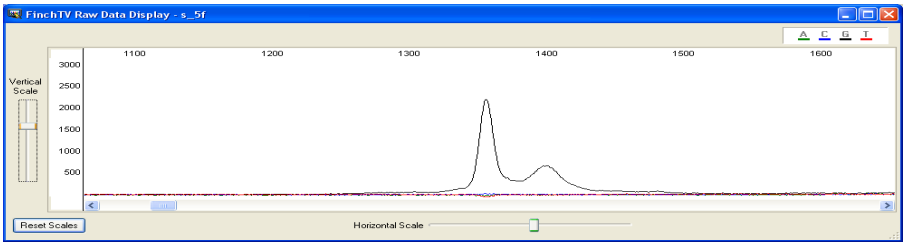


Figura 5.6: *Picco caratteristico del marcatore. Questo occupa una finestra temporale di 150-200 campioni circa, nell'intervallo 1300-1400 del RawData.*

Il RD di ogni nucleotide è costituito da due componenti: il rumore di fondo *e(t)*, e la successione dei picchi di sequenziamento *y(t)*:

$$RD_A(t) = y_A(t) + e_A(t)$$
$$RD_T(t) = y_T(t) + e_T(t)$$
$$RD_C(t) = y_C(t) + e_C(t)$$

Quello associato al nucleotide G, $RD_G(t)$, può contenere anche il marcatore *m(t)*:

$$RD_G(t) = y_G(t) + e_G(t) + m(t)$$

Il segnale $RD(t)=y(t)+e(t)$ non è altro che la sovrapposizione dei quattro segnali $RD_A(t)$, $RD_T(t)$, $RD_G(t)$, $RD_C(t)$. Il RD_G della sequenza "s_5f.ab1" è mostrato in figura 5.7, e si può osservare dallo zoom come il marcatore occupi, per questo esempio, proprio una finestra di circa 100 campioni, intorno ai 1320-1420 valori dell'asse temporale, con un'ampiezza pari a circa 2000 nella scala di intensità luminosa.

L'algoritmo esamina solo il dato grezzo della G. Per il riconoscimento del marcatore $m(t)$ calcola, all'interno di una finestra-mobile di 150 campioni che scansiona il RD_G a partire dall'inizio, la varianza s^2 dei dati. La varianza viene calcolata utilizzando la formula:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(RD_{Gi} - RD_{Gm}\right)^2$$

dove N è il numero dei campioni contenuti nella finestra corrente, mentre

$$RD_{Gm} = \frac{1}{N}\sum_{i=1}^{N} RD_{Gi}$$

è la media dei campioni di RD_G contenuti nella stessa finestra. La scelta di esaminare finestra per finestra la varianza del segnale,e di non concentrare la ricerca del marcatore esclusivamente nell'intervallo 1200-1500, è dovuta al fatto che per alcuni sequenziamenti la migrazione dei frammenti di DNA parte in ritardo, per cui sia il marcatore m(t), che il segnale y(t) di sequenziamento, sono traslati in avanti lungo l'asse delle ascisse.

Nelle finestre che non contengono né marcatore m(t), né il segnale $y_G(t)$, la varianza è quella dei dati che rappresentano solo rumore di fondo $e_G(t)$. Una "cambiamento significativo" della varianza dei dati della finestra corrente rispetto a quelli della precedente e della seguente, certifica la presenza del marcatore.

L'algoritmo riconosce il "cambiamento significativo" quando la varianza della finestra corrente è 8 volte più grande della varianza di quella precedente e successiva.

La funzione fornisce in output le variabili:

- *out:* pari a 1 quando viene riconosciuto il marcatore, pari a 0 nel caso contrario;
- *ampiezza:* che rappresenta l'ampiezza del marcatore;
- *inizio_marc* e *fine_marc:* rappresentano gli estremi della finestra contenente il marcatore.

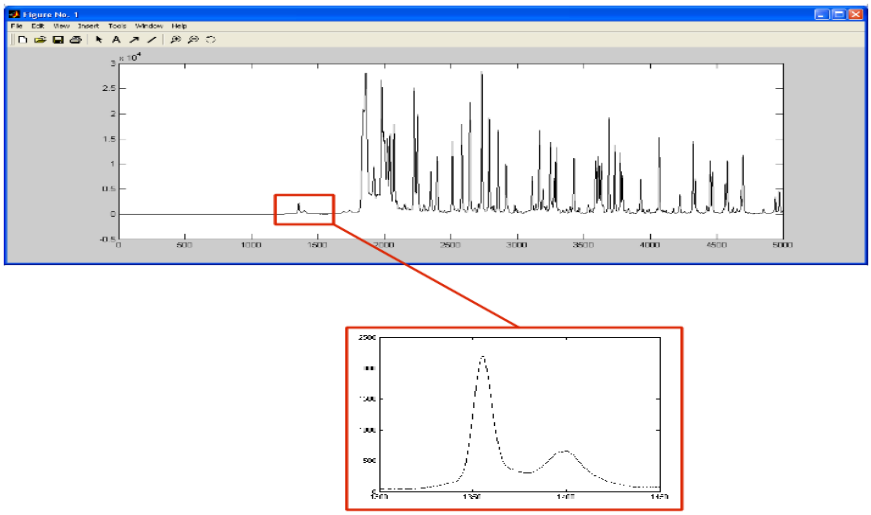


Figura 5.7: *RD_G con il picco del marcatore.*

La seconda funzione `presenza_segnale` è stata realizzata per riconoscere la successione dei picchi di sequenziamento *y(t)* nel RD. In un sequenziamento corretto, il RD è rappresentato da un segnale i cui primi 1800-2000 campioni contengono solo rumore di fondo *e(t)* (a parte il marcatore per RD_G). Successivamente, il segnale cresce e si osserva la registrazione simultanea dei segnali di sequenziamento (i picchi) per ogni base azotata [Cap.2 paragrafo 2.1.3].

L'algoritmo analizza solo i primi 4000 campioni del RD. Utilizzando, come in `presenza_marcatore`, una finestra mobile, ma stavolta composta da

100 campioni, cerca il punto di massimo e minimo di tutti e quattro i RD_G, RD_T, RD_A, RD_C in ogni finestra e ne calcola la media.

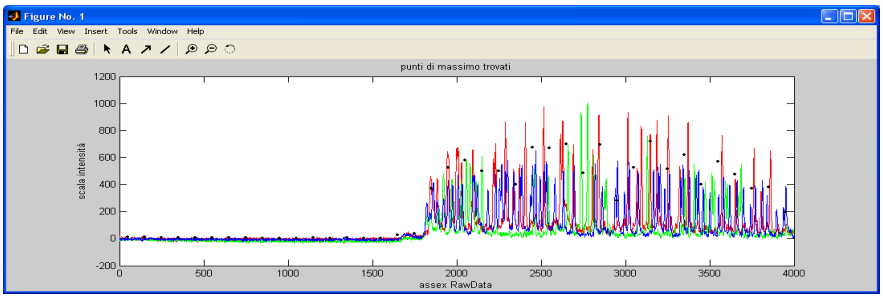


***Figura 5.8**: In nero sono rappresentati i punti "massimi" calcolati dalla funzione* `presenza_segnale` *per il riconoscimento dei picchi di sequenziamento nel RD.*

Come è evidente nel plot di figura 5.8, i punti di massimo,definiti x_i, nella prima porzione del segnale sono bassi, mentre intorno ai valori 1800-1900 dell'asse temporale, in corrispondenza dell'acquisizione del segnale di sequenziamento vero e proprio, i punti iniziano ad assumere valori elevati. Questa netta variazione dei valori viene riconosciuta dall'algoritmo attraverso la valutazione della derivata dei punti di massimo trovati. In prossimità della variazione la derivata cresce a differenza delle altre porzioni del segnale.

La derivata viene calcolata tramite approssimazione, come *differenza finita all'indietro:*

$$x'(t_i) = \frac{x(t_i) - x(t_{i-1})}{t_i - t_{i-1}}$$

dove i=2,..., M, ed M è il numero dei punti di massimo *x* calcolati.

La variazione significativa del segnale è riconosciuta nell'istante *t** quando la derivata in *t** supera di 3/2 la media delle derivate:

$$x'(t^*) \geq \frac{3}{2} \frac{1}{M-1} \sum_{i=2}^{M} x'(t_i)$$

Una volta trovato l'istante *t**, salvato nella variabile *inizio_segnale* della

funzione, l'algoritmo si accerta che l'ampiezza del segnale sia tale da rappresentare dei picchi di sequenziamento reali. Questo ulteriore controllo scaturisce dall'esigenza di riconoscere un segnale di sequenziamento che sia corretto e che non rappresenti un segnale che in effetti non è il risultato di una corretta reazione.

Il RD viene suddiviso in due regioni: la prima, che precede l'istante *t** e che contiene solo rumore di fondo *e(t)*, la seconda, successiva all'istante *t**, che contiene anche il segnale di sequenziamento *y(t)*. Il criterio è quello di valutare l'intensità del segnale nella seconda regione e l'algoritmo lo fa confrontando le due bande Δ contenute dalle linee continue nere mostrate in figura 5.9. Gli estremi di ciascuna banda, y_1 e y_2 (non visibili in figura, perché molto vicini) per la prima Δ_1, e y_3 e y_4 per la seconda Δ_2, vengono trovati attraverso il calcolo della media dei punti di massimo e di minimo trovati precedentemente e rispettivamente nelle due regioni. In particolare y_1 e y_2 vengono calcolati considerando solo i primi 5 punti di massimo e minimo, in quanto il segnale in prossimità di t* non è più stazionario.

Se Δ_2 è 3 volte più ampia di Δ_1, allora viene riconosciuto il segnale di sequenziamento, ed il risultato viene fornito in output dalla funzione: la variabile *esito* è uguale ad 1 qualora y(t) viene riconosciuto, viceversa è uguale a 0. La funzione `presenza_segnale` fornisce in output anche la variabile *banda1max* ovvero il livello superiore di Δ_1, quello indicato come y_1. Questo valore, come sarà spiegato nel paragrafo 5.4, verrà utilizzato come argomento di input della funzione `segnale_inarcato`.

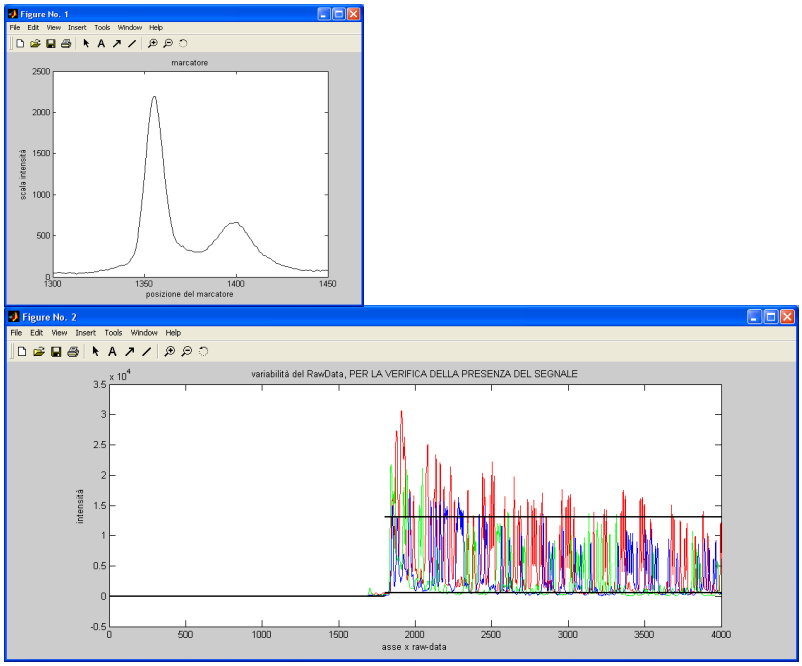


Figura 5.9: *plot forniti dalle funzione presenza_marcatore e presenza_segnale per la sequenza di figura 5.2. L'algoritmo riconosce correttamente la presenza e la posizione del marcatore nel RD, e riconosce, secondo plot, la presenza dei picchi del segnale di sequenziamento.*

Se l'algoritmo non trova né il marcatore, né il segnale di sequenziamento, attribuisce alla sequenza la classe *NS*: *RD=e(t)*.

Se invece trova il marcatore, ma non il segnale di sequenziamento, l'algoritmo classifica la sequenza come NR: RD=m(t)+e(t). Per entrambi i casi, l'algoritmo non prosegue con i successivi step, in quanto i dati costituirebbero solo disturbo, e non segnale utile y(t) da cui poter leggere la sequenza di basi

.

5.2 Problematiche legate all'ampiezza e all'andamento del Raw Data

5.2.1 Problematiche legate all'ampiezza del Raw Data

Quando l'algoritmo riconosce il segnale y(t), prosegue con il secondo step

dell'analisi che verte nel valutare l'ampiezza di *y(t)*, argomento trattato nel paragrafo 4.3 del Cap.4. La funzione Matlab che realizza questa analisi si chiama `ampiezza_segnale`, e determina se il segnale è basso (SB), se è alto (SA) o se è fuori scala (FS). Quando l'analisi dà esito negativo, ed il segnale quindi non rientra in nessuna delle tre classi, l'algoritmo ne riconosce un'ampiezza regolare.

Per valutare solo i picchi del sequenziamento, e quindi escludere il rumore di fondo e(t), l'intero RD viene sottoposto ad un processo di sogliatura (realizzato dalla funzione `sogliatura`). Per la maggior parte dei segnali, e(t) resta contenuto all'interno di una banda di valori -50,50 nella scala delle ordinate, prima dell'istante t* [paragrafo 5.1]. La scelta è stata quella di scegliere una soglia costante, *soglia_rumore,* pari a 75, e di considerare solo i dati che superano tale valore, e di porre uguali a zero quelli che non lo superano.

Il passo successivo consiste nel ricercare i punti più alti del RD, come rappresentato dai pallini in blu della figura 5.10 per la sequenza "s_5f.ab1". Il segnale y(t) viene visitato attraverso finestre di 200 campioni, dentro ognuna delle quali viene trovato il *picco* più alto per ogni RD_A, RD_T, RD_C, RD_G, attraverso una strategia di peak detection. Il picco finale, quello in blu, viene scelto tra i quattro picchi trovati in ogni finestra, e corrisponde al picco più alto fra i quattro.

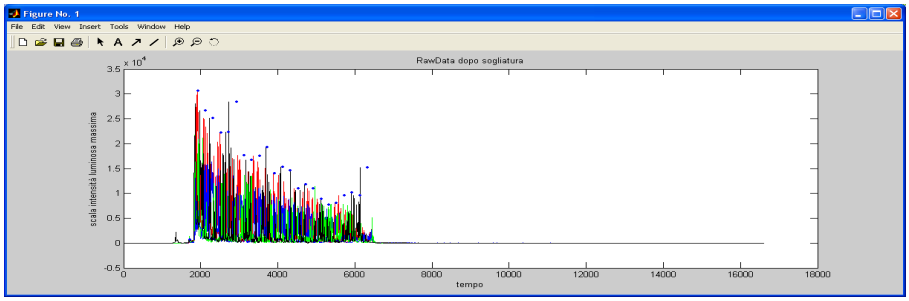


Figura 5.10: *I punti in blu rappresentano l'inviluppo del RawData elaborato dalla funzione* `ampiezza_segnale`

A questo punto la serie di valori viene confrontata con tre soglie costanti i cui valori sono stati suggeriti dai biologi della BMR Genomics che si occupano di

troubleshooting:

- la prima soglia, scelta pari a 500, serve per verificare se il segnale è basso. Se l'80% dei valori sta al di sotto di questa soglia, il segnale viene classificato *SB;*
- la seconda soglia, pari a 5000, serve per riconoscere un segnale alto. Se il 20% dei valori supera questa soglia, il segnale viene classificato *SA;*
- la terza soglia, pari a 27000 (la scala di intensità del RD arriva fino a 32000) serve per riconoscere un segnale fuori scala. Se il 15% dei valori supera questa soglia, il segnale viene classificato *FS.*

In particolare, la scelta delle diverse percentuali per FS, SA e SB ha un motivo: come già spiegato nel capitolo 4, nel paragrafo 4.3.2, segnali alti e fuori scala si presentano a volte con un RD che assume valori elevati soprattutto nei primi istanti di acquisizione, e pian piano cala raggiungendo valori regolari.

Come detto precedentemente, la *soglia_rumore* che viene utilizzata per escludere il disturbo dall'analisi è inizializzata a 75. Il riconoscimento di un segnale basso, piuttosto che di un segnale alto, o fuori scala, fa sì che il valore della soglia cambi. Per i successivi step dell'algoritmo, nel caso di segnali bassi, la soglia da 75 è corretta a 50, per segnali alti è corretta a 300, per segnali fuori scala è corretta a 500. Il motivo di questa correzione verrà spiegato nel paragrafo 5.2.2.

La figura 5.11 mostra l'output della funzione `ampiezza_segnale` per la sequenza “s_5f.ab1”.

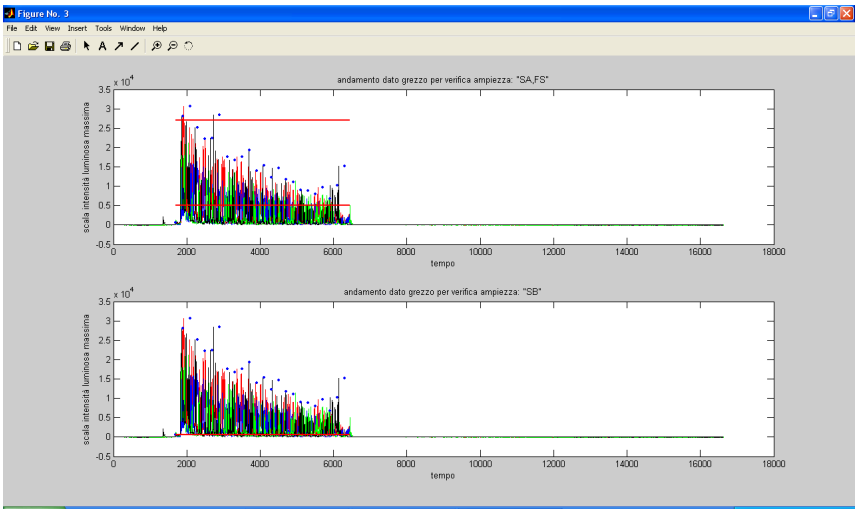


***Figura 5.11**: plot output della funzione* `ampiezza_segnale` *per il troubleshooting automatico della sequenza "s_5f.ab1". Il primo riquadro mostra il RD e in rosso sono tracciate le due soglie per il confronto dell'ampiezza del segnale per il riconoscimento di segnali SA e FS. I punti in blu tracciano l'inviluppo del segnale. Il RD supera nettamente la soglia pari a 5000, e l'algoritmo classifica il segnale come SA. Il secondo riquadro invece rappresenta il confronto del RD con la soglia per il riconoscimento di segnali SB; è evidente come il RD sia nettamente superiore a tal livello.*

5.2.2 Problematiche legate all'andamento del Raw Data

La funzione `segnale_calante` di Matlab ha lo scopo di analizzare l'andamento del RD [Cap.3 paragrafo 3.2.2]. Prima di procede con quest'analisi, l'algoritmo testa la lunghezza di y(t), per riconoscere un campione preparato o attraverso PCR o attraverso DNA ricombinante. `Presenza_pcr` è la funzione realizzata per la ricerca del picco di PCR. L'algoritmo trova i picchi del RD, dopo un processo di sogliatura per escludere rumore di fondo, identico a quello descritto nel paragrafo 5.2.1. Se l'ultimo picco è verde (A), e se occupa una posizione inferiore ai 16000 valori delle ascisse (il RD ha un'asse x che raggiunge al massimo 16200 valori) viene riconosciuto un picco di PCR. Questa funzione riceve come argomento il parametro *soglia_rumore* descritto nel paragrafo precedente, che assume differenti valori a seconda dell'ampiezza del RD. In figura 5.12 è rappresentata una finestra temporale del RD di un segnale fuori scala,

contenuta nel file ABIF "Per1Rhod1-16S.ab1". Si nota l'ultimo picco verde rappresentativo della PCR. Il RD successivo al picco non contiene il segnale di sequenziamento y(t), ma rumore di fondo e(t). Se venisse utilizzata una *soglia_rumore* pari a 75, verrebbero trovati altri picchi nel RD riconosciuti come appartenenti a y(t). In questo esempio è necessario usare una soglia maggiore almeno a 200 per riconoscere la fine di y(t). Vi sono casi più gravi in cui la soglia deve anche superare i 300, o 400 valori. Dopo diverse prove empiriche, la *soglia_rumore* è stata fissata a 500 per segnali FS. Le stesse osservazioni valgono per segnali alti (la soglia è corretta a 300), e per segnali bassi (la soglia è corretta a 50). Per quest'ultimo caso, essendo il segnale debole, la *soglia_rumore* è stata abbassata in modo da non perdere i picchi di sequenziamento.

La funzione `presenza_pcr`, qualora venisse trovato il picco di PCR, fornisce in output la posizione dell'ultimo picco T_{fin} che rappresenta quindi la lunghezza del filamento sequenziato.

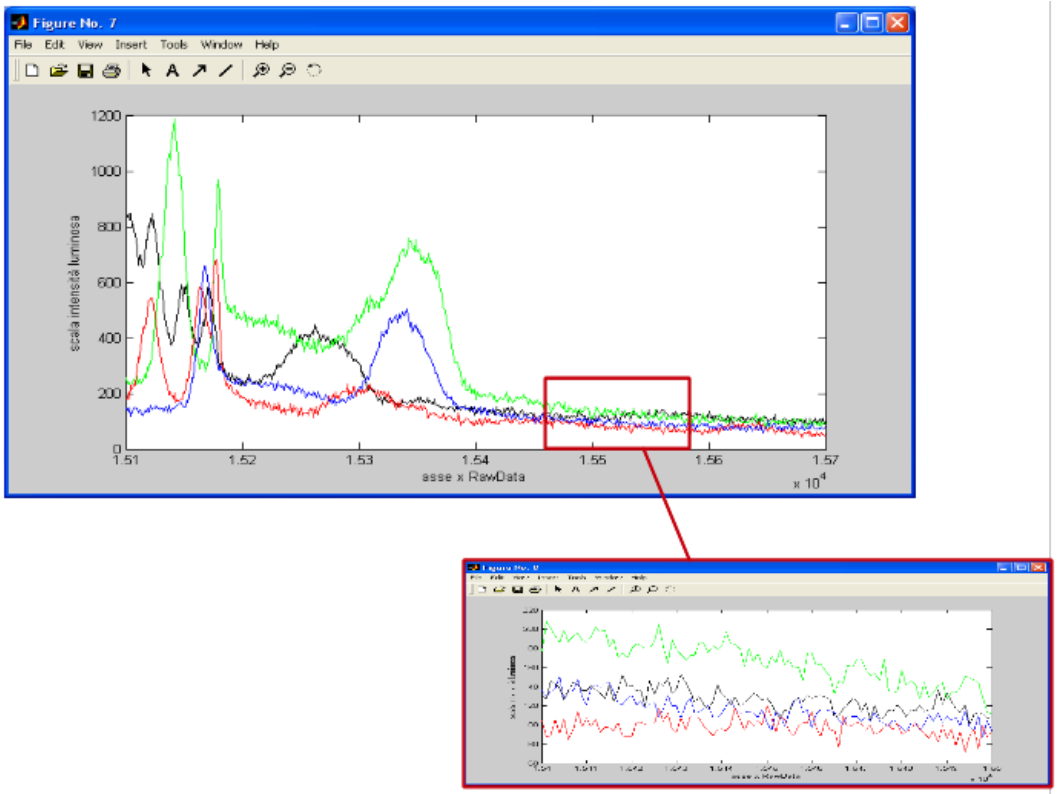


Figura 5.12: *finestra temporale del RD di una segnale FS, contenuta nel file ABIF PerlRhod1-16S.ab1. L'ultimo picco A rappresenta la fine di y(t). Il resto del segnale costituisce rumore di fondo.*

In `segnale_calante` l'informazione sulla presenza del picco di PCR è necessaria per riconoscere la fine del segnale y(t) registrato, e limitare l'analisi solo nella regione che lo contiene. L'obiettivo è quello di tracciare la curva *F(t)* per approssimare l'andamento di y(t). Ciò lo si fa attraverso una ricerca del picco "massimo" del RD, stessa strategia di peak detection utilizzata nella funzione `ampiezza_segnale` nel paragrafo 5.2.1, all'interno di una finestra di 200 campioni che scansiona y(t) dall'istante *[t*+200]*, fino alla fine di y(t). Questo passo fornisce la serie di picchi p_i.

La scelta di non considerare i primi 200 campioni di y(t) dipende dal fatto che questi primi dati possono registrare anche la fluorescenza dei dimeri di primer *DP* [Cap. 4 paragrafo 4.6], che incrementerebbero di molto il valore del RD proprio in questa regione del segnale. Prima di quest'operazione, come per tutte le

analisi fatte precedentemente, il RD viene sogliato, con soglie pari a 50 nel caso di segnali bassi, 200 per segnali alti e fuori scala. Nel caso in cui il segnale non rientrasse in nessuna di queste categorie la *soglia_rumore* resta pari a 75.

L'istante T che indica la fine di y(t) invece corrisponde alla posizione dell'ultimo picco del RD, altrimenti alla posizione in cui è stato riconosciuto il picco A di PCR, T_{fin}.

Una volta trovata la serie di picchi p_i, con $i=1,...P$, dove P è il numero dei picchi, viene effettuata un'operazione di *media mobile* per approssimare questa serie:

$$f_i = \frac{1}{2n+1} \sum_{i=1}^{2n+1} p_i$$

Il risultato di questa operazione è una serie di valori, f_i: per ogni p_i, si identifica l'insieme N^{2n} dei 2n picchi più vicini a p_i. N^{2n} contiene 2n+1 punti perché contiene pure p_i. Il parametro n viene posto uguale a 3. Dopo aver reiterato l'operazione di media mobile su f_i per 3 volte si ottiene la curva dell'andamento F(t).

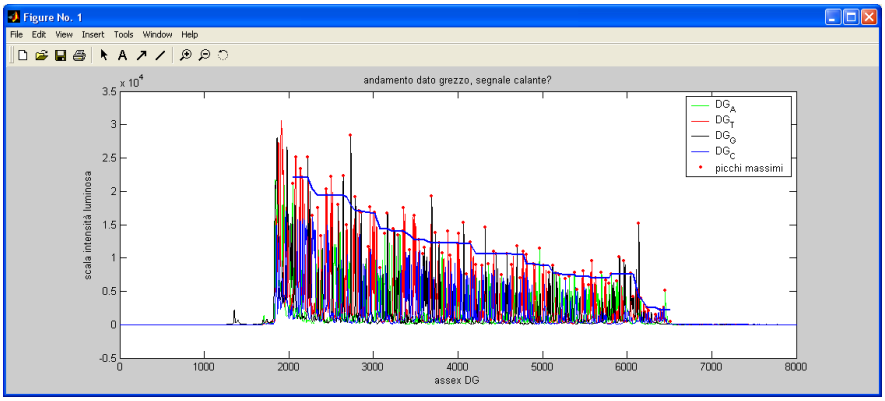


Figura 5.13: *RD della sequenza "s_5f.ab1". La linea blu rappresenta la curva dell'andamento F(t) di y(t) calcolata dall'algoritmo per l'analisi del decremento del segnale*

Il passo successivo consiste nel valutare l'istogramma dell'andamento. Il criterio è il seguente: se la "variazione" di F(t) è significativa, l'algoritmo

riconosce un andamento calante del segnale, anziché un andamento costante. Inoltre, se l'ultimo valore della curva è "basso", l'algoritmo riconosce una *morte M* del segnale. La "variazione significativa" e il "valore basso" vengono affrontati dall'algoritmo in modi differenti a seconda di un segnale SB, SA, FS o di uno che non rientri in nessuna delle tre classi. Per quest'ultimi la *variazione significativa* corrisponde ad un Δ di 1000 valori, mentre il *valore basso* è un parametro fissato a 200. Quindi quando F(t) subisce una variazione di 1000 valori nell'asse y, e quando decresce raggiungendo il valore 200, l'algoritmo classifica il segnale come *M*. Nel caso di segnali SB, SA, FS, i parametri cambiano come mostra la tabella 5.1. La scelta di questi valori è stata fatta in base a test su segnali che hanno composto il training set per la realizzazione dell'algoritmo, come sarà mostrato nel Capitolo 6.

segnale	*Variazione significativa*	*Valore basso*
SB	50	100
SA	4000	350
FS	15000	500

Tabella 5.1: *tabella con i valori scelti per i parametri "variazione significativa" e "valore_basso" usati nella funzione* `segnale_calante` *per sequenze SB, SA o FS.*

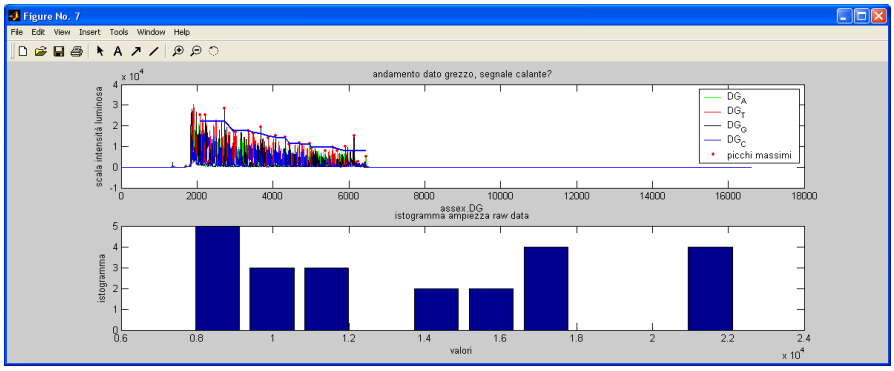


Figura 5.14: *plot output della funzione* `segnale_calante` *per il troubleshooting automatico della sequenza "s_5f.ab1". Nel primo riquadro è tracciata in blu la curva che rappresenta l'andamento di y(t): F(t) Il secondo riquadro rappresenta il suo istogramma. Il segnale inizialmente assume valori piuttosto alti, sopra i 20000, e man mano cala fino a raggiungere valori più bassi, 3000. La variazione è pari in valore assoluto a 17000, e l'algoritmo riconosce un calo del segnale. Il segnale però non viene classificato come M, in quanto il calo non è così drastico da portare il segnale a morire. Nella parte finale del segnale, i dati assumono valori alti (3000 nella scala di intensità luminosa) e costituiscono picchi di y(t). Nel caso in cui l'andamento avesse raggiunto valori inferiori a 350, parametro dell'algoritmo, la sequenza sarebbe stata riconosciuta come M.*

La figura 5.14 mostra il plot di output della funzione `segnale_calante` descritta.

5.3 Segnale inarcato

La funzione Matlab *segnale_inarcato* è stata realizzata per riconoscere nel RD l'innalzamento della linea di base del segnale [Cap.3 paragrafo 3.2.3]. Come linea di base viene intesa quel livello medio di valori registrati prima dell'acquisizione dei picchi di sequenziamento. Questa funzione riceve come argomento di input una variabile fornita dalla funzione `presenza_segnale` (paragrafo 5.1), chiamata *banda1max,* utilizzata per rappresentare la linea di base del segnale.

L'algoritmo proposto confronta la larghezza di banda contenuta tra le due linee rosse mostrate in figura 5.15, chiamata *banda_segnale*, e la larghezza tra la linea rossa inferiore e la linea di base del segnale, inizializzata come *banda_segnale_lineaBase* nella funzione e che qua verrà indicata come Δ.

Il livello di ciascuna linea viene calcolato attraverso il seguente procedimento: a partire dall'istante di inizio di y(t), t* (paragrafo 5.1), vengono

calcolati, in ogni finestra-mobile di 50 campioni che scandisce il segnale, il picco e il nadir del RD (fino a 4000 campioni), rappresentati nel plot con i pallini neri. La media dei picchi dà il valore della soglia superiore, chiamata *soglia_alta*, mentre quella inferiore, *soglia_bassa*, è uguagliata al valore del nadir più alto. Una sequenza viene riconosciuta come *SI* quando Δ è "grande" in confronto alla banda del segnale. L'algoritmo compie quest'operazione testando se il Δ moltiplicato per 8 è maggiore di *banda_segnale*. L'esito positivo di questo controllo classifica il segnale come inarcato, *SI*.

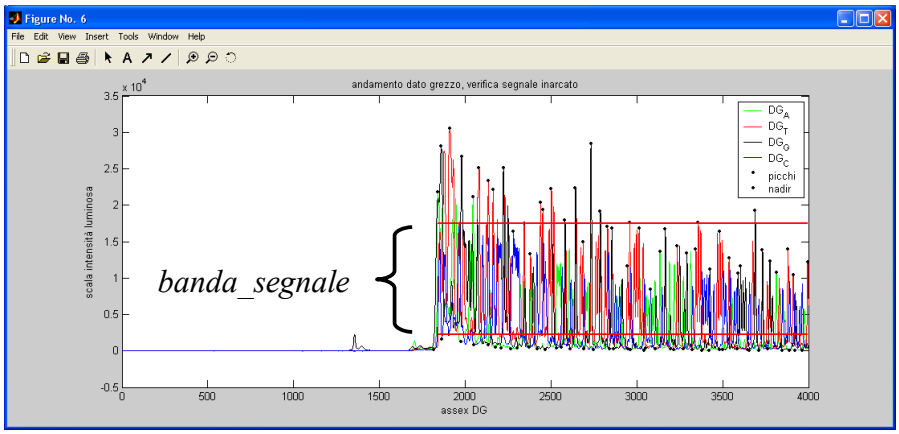


***Figura 5.15**: plot output della funzione "segnale_inarcato" per il troubleshooting automatico della sequenza "s_5f.ab1". La figura rappresenta i primi 4000 campioni del RD. Le linee in rosso vengono tracciate dall'algoritmo, e per questa sequenza la linea di base del segnale è inarcata, e lo si coglie dalla distanza della linea rossa inferiore rispetto la linea di base (vicina allo zero).*

5.4 Picchi multipli nell'Analyzed Data

Il riconoscimento di picchi multipli nell'AD [Cap.3 paragrafo 3.2.4] è svolto dalla funzione `presenza_picchiDoppi`. A differenza dei casi precedenti, in questo step il segnale sottoposto ad analisi è l'AD, e non il RD. Questo passo dell'algoritmo ha lo scopo di verificare, picco per picco del segnale, la presenza di picchi multipli lungo la sequenza. Come per RD, AD è la sovrapposizione dei quattro segnali associati ad ogni base:

$$AD_A(t) = Y_A(t) + E_A(t)$$
$$AD_T(t) = Y_T(t) + E_T(t)$$
$$AD_C(t) = Y_C(t) + E_C(t)$$
$$AD_G(t) = Y_G(t) + E_G(t)$$

dove Y(t), e E(t) sono il risultato della trasformazione di y(t) ed e(t) da RD ad AD.

Anche in quest'analisi, il segnale subisce una sogliatura per escludere il rumore di fondo. Dopo diverse prove empiriche, si è giunti alla conclusione che la soglia pari a 80 offre le migliori prestazioni per l'analisi. Il file ABIF fornisce, nell'array *base_location,* le posizioni che occupano le singole basi nell'AD. Utilizzando quindi questa informazione, per ogni segnale, l'algoritmo procede nella seguente maniera: considera un intorno (costituito da 11 campioni) di ogni punto del *base_location,* e all'interno di questo controlla la presenza di picchi di AD_A, AD_T, AD_C, AD_G. Se vi sono almeno due picchi, l'algoritmo riconosce la presenza di picchi multipli in quella posizione della sequenza. I picchi vengono ordinati in base alla loro intensità e vengono presi in considerazione solo i due picchi più alti. Se il picco più basso è inferiore alla metà dell'intensità del picco più alto, viene riconosciuto un picco multiplo. Viceversa viene riconosciuto un picco multiplo i cui picchi coinvolti hanno stessa intensità. L'analisi viene ripetuta per tutti i punti del *base_location*.

Alla fine di questo passo, si ottiene un array chiamato *presenza* che ha le stesse dimensioni di *base_location,* e che per ogni posizione contiene 1 se in quella posizione della sequenza è stato trovato un picco multiplo, altrimenti contiene 0. Per ogni picco multiplo è associato anche un valore di "intensità", anch'esso contenuto in un array chiamato *intensità.* Ogni elemento di *intensità* è pari ad 1 se il picco multiplo è caratterizzato da due picchi di intensità molto simile, altrimenti è uguale a 0 se il picco multiplo è caratterizzato dalla presenza di due picchi di intensità differente. In figura 4.17 è riportato una sezione dell'AD della sequenza contenuta nel file ABIF "s_5f.ab1". La sezione contiene i picchi associati alla sequenza di basi che vanno dalla 64 alla 99. L'AD presenta dei

picchi multipli, ed in particolar modo i sotto-picchi hanno un'intensità bassa rispetto la sequenza dominante. Il plot di figura 5.17 mostra la stessa porzione dell'AD di figura 5.16 elaborata dall'algoritmo. La linea in nero rappresenta la soglia utilizzata per escludere il rumore di fondo dall'analisi, mentre gli asterischi rossi rappresentano le posizioni del *base_location*.

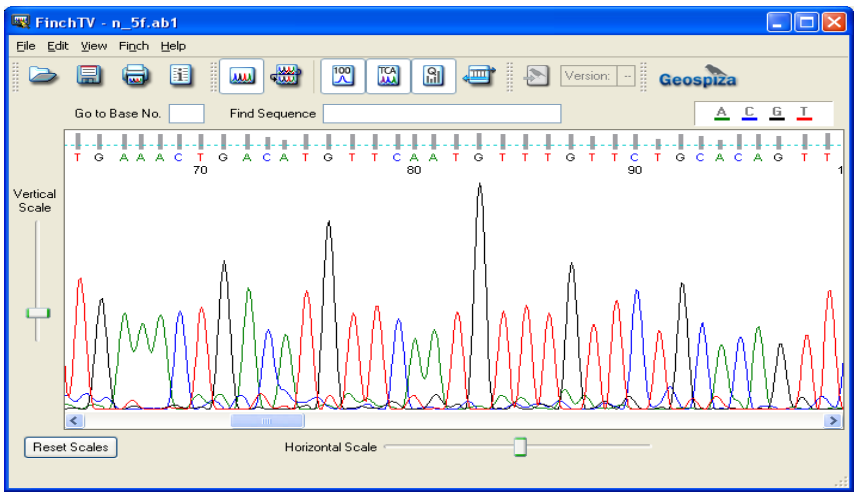


Figura 5.16: *una finestra dell'Analyzed Data della sequenza s_5f.ab1. Si osserva la presenza di un alto rumore di fondo nel segnale.*

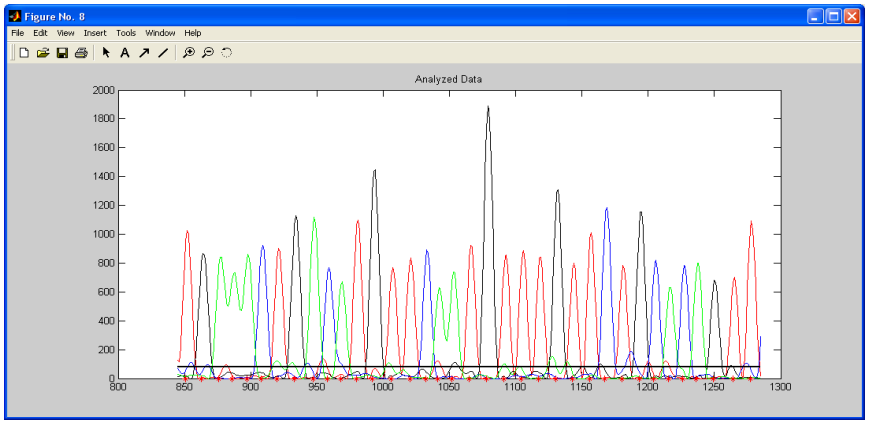


Figura 5.17: *una finestra dell'Analyzed Data della sequenza s_5f.ab1, identica a quella di figura 5.16. La figura mostra il primo passo dell'algoritmo per il riconoscimento dei picchi multipli. In nero è rappresentata la soglia per l'eliminazione del rumore di fondo presente in ogni Analyzed Data, mentre gli asterischi in rosso rappresentano le posizioni delle basi.*

In figura 5.18 viene mostrata invece l'analisi compiuta dall'algoritmo per la

sequenza d'esempio:

- il primo vettore costituisce la successione di basi azotate identificate dal sequenziatore per la sezione del segnale considerata;
- il secondo vettore rappresenta il *base_location:* la posizione nell'asse x dell'AD occupata da ogni base azotata;
- il terzo vettore rappresenta la sezione corrispondente dell'array *presenza*;
- il quarto vettore rappresenta invece la sezione corrispondente dell'array *intensità:* l'array contiene in tutte le posizioni solo lo 0 in quanto tutti i picchi multipli trovati sono relativi a picchi di diversa intensità.

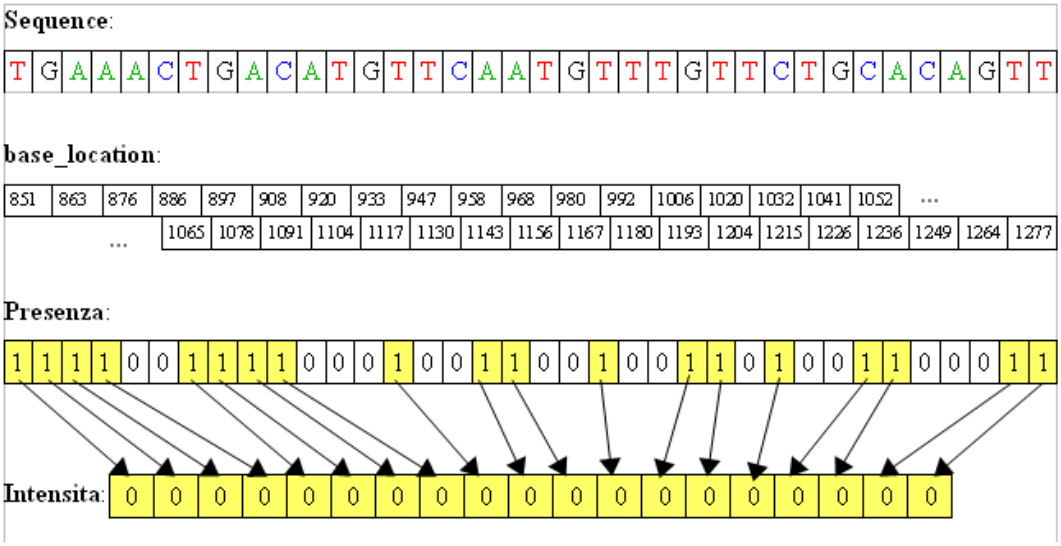


Figura 5.18: *la figura mostra 4 array: sequence rappresenta la sequenza di basi identificata dal sequenziatore relativa alla porzione del segnale di figura 5.16. Il secondo array rappresenta il base_location, contenente la posizione nell'asse x dell'AD delle basi (gli asterischi in rosso di figura 5.17). Il terzo array, presenza, è calcolato dall'algoritmo, e contiene 1 nelle posizioni dove viene riconosciuto un picco multiplo, 0 viceversa. Sono colorate in giallo le posizioni in cui è stato trovato il picco multiplo, e per ognuno di esso corrisponde un valore che viene memorizzato nel vettore intensità. Viene analizzata la natura del picco multiplo: l'array intensità contiene 1 se il picco è doppio (D), 0 nel caso contrario (FD).*

L'algoritmo prosegue calcolando la mediana dei valori contenuti nel vettore *intensità,* e memorizza il risultato nella variabile *intensitàTOT.* Attraverso quest'operazione, viene riconosciuto se la sequenza è nel suo complesso *FD* o *D*

[Cap. 4 paragrafo 4.5]. Inoltre l'algoritmo ignora le prime 30 basi della sequenza: questa scelta si basa sul fatto che i primi dati acquisiti hanno sempre una risoluzione scarsa. Quasi tutte le sequenze presentano infatti picchi irregolari e senza forma proprio per queste basi.

Dopo aver analizzato la natura dei picchi (*FD, D*), viene analizzata la "quantità" di picchi multipli nella sequenza. Quando i picchi multipli trovati sono pochi rispetto il numero di basi dell'AD, non viene riconosciuta una sequenza multipla. Per far ciò, l'algoritmo scandisce il vettore *presenza* in blocchi di 10 posizioni. Quando su 10 basi, più di tre sono costituite da picchi multipli, in quel blocco viene riconosciuta l'esistenza di picchi multipli. Il risultato viene salvato in un altro vettore, chiamato *presenza10:* questo contiene 1 se in 10 basi dell'AD almeno 3 contengono picchi multipli, 0 viceversa. La funzione fornisce come output il vettore *presenza10,* e la variabile *intensitàTOT.*

Per la sequenza "s_5f.ab1" presa come esempio in questo capitolo, *base_location* è un vettore riga che ha dimensioni [1, 328]. Per quello che è stato detto in questo paragrafo, *presenza* sarà anch'esso un vettore che avrà le stesse dimensioni. Ignorando le prime 30 basi, la dimensione principale del vettore passerà da 328 a 298. Raggruppando le basi in gruppi di 10, *presenza10* invece avrà dimensioni [1,29] (figura 5.21).

Il passo successivo ha la finalità di trovare la ragione dell'AD che presenta picchi multipli:

- i picchi multipli interessano tutto l'asse x dell'AD *(FD, D)*;
- i picchi multipli interessano solo la parte iniziale dell'AD (*TD-*);
- i picchi multipli iniziano da un punto intermedio dell'AD (*DD-*).

Se la sequenza è costituita da almeno 200 basi, l'algoritmo funziona come segue: il vettore *presenza10* viene anch'esso raggruppato in blocchi da dieci, e per ogni blocco viene calcolata la mediana dei valori che contiene; il dato calcolato viene memorizzato nel vettore chiamato *Doppio.* Così ogni elemento di quest'ultimo riassume la natura di 100 basi della sequenza. La logica è sempre la

stessa: se la mediana calcolata è pari a 1 le 100 basi sotto analisi sono rappresentate da picchi multipli, se la mediana calcolata è pari a 0, le 100 basi invece sono rappresentate da una sequenza che nel suo complesso non presenta picchi multipli. Quando il risultato dell'operazione di mediana è uguale a 0.5, questa viene corretta a 1 (per tutte le volte che viene calcolata nell'algoritmo). Una volta calcolato il vettore *Doppio,* lo si sottopone a tre test:

1. se il vettore *Doppio* è costituito da elementi pari solo a 0, nella sequenza non viene riconosciuta la presenza di picchi multipli;
2. se il vettore *Doppio* contiene l'elemento 1 in tutte le sue posizioni, allora nella sequenza viene riconosciuta la presenza di picchi multipli in tutta la sua lunghezza;
3. se il vettore *Doppio* contiene anche degli elementi pari a 0, vuol dire che la distribuzione dei picchi multipli interessa solo alcune porzioni dell'AD. L'analisi viene suddivisa in altri due sotto-problemi:

   3.1 il vettore *Doppio* contiene l'elemento 1 nella sua prima posizione: la sequenza viene classificata nella classe *TD*-, perché l'algoritmo riconosce picchi multipli solo nelle prime basi. In più, per i motivi spiegati nel paragrafo 4.5 del Capitolo 4, l'algoritmo può non riconoscere la presenza di picchi multipli alla fine di una sequenza, a causa della scarsa risoluzione che possono avere quest'ultimi. Quindi, in funzione ai *quality scores* associati ad ogni base identificata, l'algoritmo rianalizza l'AD solo fino al picco P della sequenza che ha un *quality_score** superiore al valor medio di tutti i quality scores (N è il numero dei quality scores):

   $$quality\_score^{*} \geq \frac{1}{N}\sum_{i=1}^{N} quality\_scores_i$$

   Questo passo viene compiuto solo per sequenze che risultano inizialmente *TD*-, attraverso la funzione `analisi_qualità`. Per riportare un esempio di come funziona l'algoritmo di fronte a queste

situazioni, in figura 5.19 viene riportata la sequenza "6.no-FD-8624.ab1" già vista in figura 4.19, del Capitolo 4, i cui ultimi 200 picchi dell'AD hanno, come si osserva, una scarsa risoluzione. La sequenza è *FD*, ma l'algoritmo nella prima analisi la classifica *TD-FD*. Successivamente, l'analisi di qualità ignora appunto gli ultimi 200 picchi, e la sequenza viene correttamente classificata come *FD*.

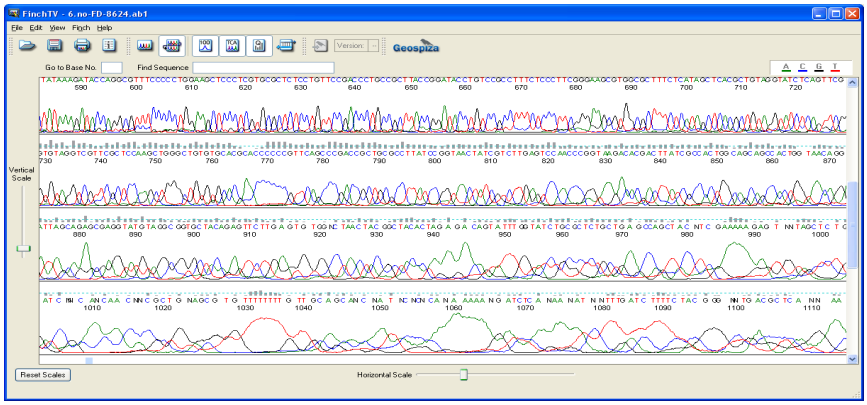


Figura 5.19: *AD della sequenza "6.no-FD-8624.ab1". La sequenza, che presenza picchi multipli in tutta la sua lunghezza, ha dei problemi di risoluzione relativi alle ultime 200 basi circa. I picchi non hanno una forma regolare, si allargano, e non è possibile riconoscere il picco principale.*

3.2 il vettore *Doppio* contiene l'elemento 1 dopo la prima posizione: la sequenza viene classificata nella classe *DD-*, perché l'algoritmo riconosce picchi multipli solo a partire da un *punto intermedio* della sequenza. Nel caso in cui il filamento di DNA sia stato preparato attraverso la tecnica di DNA ricombinante [Cap.2 paragrafo 2.1.1], e se i picchi doppi interessano solo le ultime 250 basi, queste vengono ignorate dall'algoritmo in quanto i picchi multipli riconosciuti potrebbero esser dovuti alla bassa risoluzione del segnale.

Una sequenza DD- può nascere dalla presenza di un omopolimero nella regione che precede il "doppiaggio" (paragrafo 4.5 del Cap. 4). Vengono prese in considerazioni le 30 basi precedenti e seguenti il *punto intermedio*. Viene analizzata la stringa di caratteri rappresentante le basi, e se viene individuata una ripetizione continua di una singola

base azotata, l'algoritmo riconosce la presenza di un omopolimero, e classifica la sequenza nella classe *PA,* se l'omopolimero è l'adenina, *PC* se l'omopolimero è la citosina, *PG* se l'omopolimero è la guanina, *PT* se l'omopolimero è la timina.

Se la sequenza è costituita da meno di 200 basi, invece di compiere l'analisi appena vista sul vettore *Doppio*, viene fatta direttamente sul vettore *presenza10,* output della funzione `presenza_picchiDoppi`.

In figura 5.20 è mostrato il plot fornito in seguito alle elaborazioni appena descritte. I pallini in blu rappresentano i valori contenuti nel vettore *presenza10.* Ogni punto riassume la natura di 10 basi della sequenza sotto analisi, in questo caso per la sequenza "s_5f.ab1" riportata in molte figure di questo capitolo. Si nota come le prime 30 basi vengano escluse dall'analisi. La sequenza è lunga circa 330 basi, come mostrato dall'asse x del grafico (con precisione 328); i punti possono assumere due soli valori: 1 quando le 10 basi contenute in quella finestra sono rappresentate da un AD che contiene picchi multipli, 0 viceversa. Ad esempio, il punto indicato dalla freccia rappresenta le 10 basi contenute tra la 190-esima e la 200-esima base. Si osserva che la sequenza nel suo complesso presenta picchi multipli. L'algoritmo classifica la sequenza come *FD* (figura 5.4).

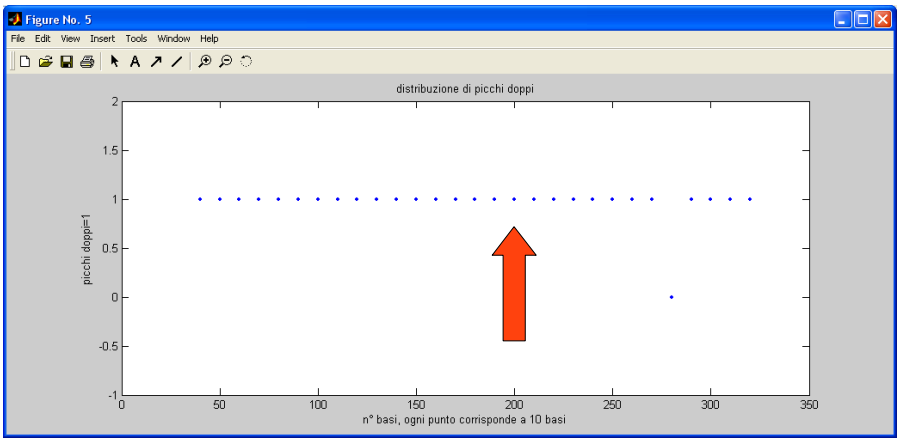


Figura 5.20: *plot fornito dalla funzione* `presenza_picchiDoppi`*. Il grafico riporta in ascissa le basi della sequenza raggruppate in gruppi da 10, mentre in ordinata i valori rappresentati possono occupare solo due livelli: 1 se in quella posizione dell'asse x dell'AD ci sono picchi multipli, 0 viceversa. Il plot è il risultato dell'analisi della sequenza "s_5f.ab1" e si evince dal grafico come l'intera sequenza presenta picchi multipli. È possibile verificare la riuscita del troubleshooting automatico controllando l'AD di figura 5.3, relativo alla sequenza in questione.*

Capitolo 6

Risultati

L'algoritmo descritto nel Capitolo 5 è stato realizzato utilizzando un insieme di 167 sequenze di DNA, con problematiche note, fornite dalla BMR Genomics. Per ogni problematica, in tabella 6.1, è indicato il numero di sequenze utilizzate per impostare in maniera empirica i diversi parametri dell'algoritmo; si osserva che la somma delle sequenze utilizzate per la ricerca di ogni problematica non corrisponde a 167, in quanto una sequenza può presentare una o più problematiche.

Classificazione	***N° sequenze per implementazione***
NR	25
NS	10
SB	29
SA	38
FS	7
M	42
SI	44
Picchi Multipli	80
Totale	**167**

Tabella 6.1: *sequenze utilizzate per la realizzazione dell'algoritmo*

L'algoritmo è stata testato con altre 1200 sequenze, anche queste con

problematiche note. Per la classificazione di ogni problematica sono stati ricavati i Veri Positivi (VP), i Veri Negativi (VN), i Falsi Positivi (FP) e i Falsi Negativi (FN), i cui risultati sono riportati nelle successive tabelle di questo capitolo. Per ogni classificazione vengono riportate due tabelle: la prima che riporta in valore assoluto il numero dei FP, FN, VP e VN, mentre la seconda li riporta in percentuale. I FP e i FN danno una misura degli errori compiuti dall'algoritmo: i Falsi Positivi rappresentano il numero di sequenze riconosciute come appartenenti alla classe oggetto di studio, ma che in realtà non vi appartengono, mentre i Falsi Negativi rappresentano il numero di sequenze realmente appartenenti alla classe, ma non riconosciute dall'algoritmo. In particolare i VP e VN in percentuale corrispondono rispettivamente alla sensibilità (*Se*) e alla specificità (*Sp*).

$$Se = \frac{TP}{TP + FN} \qquad Sp = \frac{TN}{TN + FP}$$

Essendo il numero dei Veri Negativi molto alto, non si riescono sempre a discriminare efficacemente gli errori commessi dall'algoritmo per diverse problematiche, per cui viene calcolato anche il valore predittivo positivo (PPV), definito come:

$$PPV = \frac{TP}{TP + FP}$$

Per ogni classificazione viene calcolata anche la probabilità di assegnazione corretta (*Pc*):

$$Pc = \frac{TN + TP}{N}$$

dove N è pari a 1200. Pc, che varia tra 0 e 1, permette di quantizzare la correttezza dell'algoritmo nel classificare una sequenza. Per cui, maggiore è Pc, migliore è la performance dell'algoritmo. È possibile calcolare anche la probabilità d'errore, pari a 1-*Pc (Pe)*.

Più problematiche posso interessare una sequenze. Per cui è possibile

valutare la bontà della classificazione considerando globalmente il numero di sequenze per l'algoritmo ha riconosciuto tutte le problematiche presenti in essa. E in particolare, su 1200 sequenze problematiche, 716 è il numero di sequenze correttamente classificate (59,67%). L'algoritmo è stato testato anche su 150 sequenze che invece non presentano nessuna problematica. Solo 5 di queste vengono classificate erroneamente, in quanto viene loro attribuita almeno una classe di problematiche: tre di queste sequenze vengono classificate come DD-, un'altra M, e un'altra ancora come DD- ed M.

6.1 No signal e No reaction

L'algoritmo offre buone prestazioni per la ricerca di queste problematiche. La probabilità di assegnazione corretta è molto alta, raggiungendo quasi il 98% delle assegnazioni corrette per il riconoscimento di sequenze con mancata reazione (NR), e addirittura il 99% delle assegnazioni corrette per il riconoscimento di sequenze con assenza di segnale (NS). L'analisi automatica riesce quasi sempre a distinguere nel RD il rumore di fondo e(t) dall'acquisizione dei picchi di sequenziamento y(t) [Cap. 5 paragrafo 5.1]. Dal valore predittivo positivo si evince che delle 115 sequenze classificate come NR dall'algoritmo, l'83% di queste sono correttamente classificate. Invece, per le 59 sequenze classificate come NS, l'88% di queste sono classificate correttamente. Come è stato detto precedentemente la probabilità di assegnazione corretta non è in grado di dare una giusta discriminazione per quantizzare l'entità dell'errore commesso, e per le sequenze NS si osserva che anche se probabilità di assegnazione corretta è alta, il valore predittivo positivo è più basso (figura 6.3). La sensitività è del 91% per sequenze NS, del 90% per sequenze NR.

Gli errori più frequenti compiuti da questo step dell'algoritmo, e che costituiscono quella percentuale di FP (Tabella 6.3), interessano il riconoscimento dell'istante t* di inizio del segnale di sequenziamento y(t) [paragrafo 5.1 Cap.5]. In figura 6.1 è mostrato il plot fornito dalla funzione `presenza_segnale`

descritta nel paragrafo 5.1: si osserva che l'istante t* non è stato individuato correttamente e ciò causa un'errata costruzione delle bande Δ_1 e Δ_2 [Cap.5 paragrafo 5.1] che distinguono il rumore dai picchi di sequenziamento.

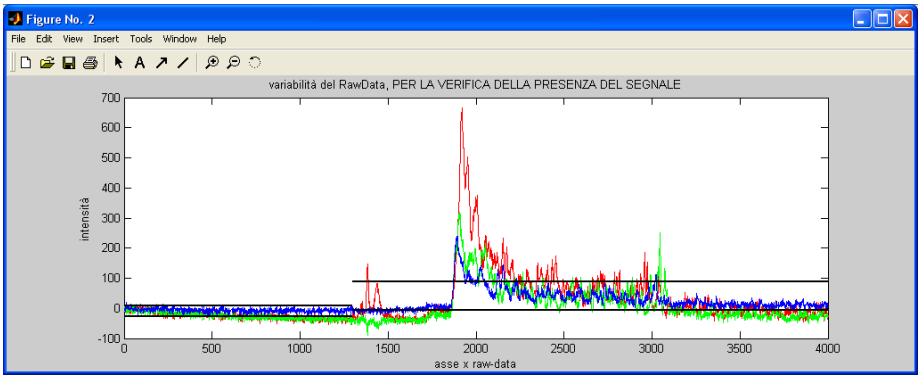


Figura 6.1: *esempio di sequenza riconosciuta NR dall'algoritmo. In realtà il Raw Data in questione presenta i picchi del segnale y(t), e non è NR. L'algoritmo fallisce (FalsiPositivi).*

Viceversa, è possibile che alcune sequenze NR non vengano riconosciute correttamente, e questo è il caso dei Falsi Negativi (Tabelle 6.2 e 6.3). In figura 6.2 si può osservare un esempio di sequenza NR, che viene riconosciuta dell'algoritmo come un segnale basso (SB). La causa di questo errore è attribuito anche in questo caso ad un'errata costruzione delle bande Δ_1 e Δ_2: la presenza di spikes sporadici (DP, DG, BL, Capitolo 4 paragrafo 4.6), rende difficoltosa la procedura di riconoscimento del segnale y(t).

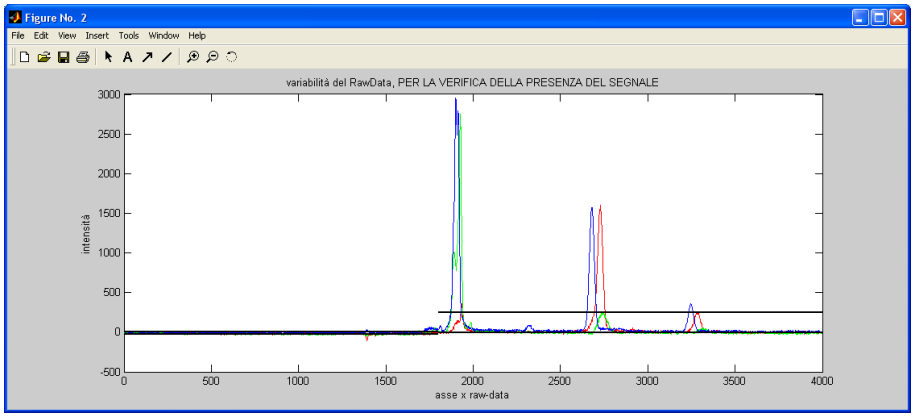


Figura 6.2: *sequenza riconosciuta dall'algoritmo come SB. In realtà si tratta di una sequenza NR, che contiene picchi anomali. L'algoritmo compie un Falso Negativo*

		risultato		
		VERO	***FALSO***	***TOTALE***
realtà	***VERO***	*VP=52*	*FN=5*	*57*
	FALSO	*FP=7*	*VN=1136*	*1143*

Tabella 6.2: *Numero di VP, VN, FP, FN per le sequenze NS*

		risultato	
		VERO	***FALSO***
realtà	***VERO***	*VP=91,23%*	*FN=8,77%*
	FALSO	*FP=0,61%*	*VN=99,39%*

Tabella 6.3: *VP, VN, FP, FN in percentuale per le sequenze NS*

		risultato		
		VERO	***FALSO***	***TOTALE***
realtà	***VERO***	*VP=96*	*FN=10*	*106*
	FALSO	*FP=19*	*VN=1075*	*1094*

Tabella 6.4: *Numero di VP, VN, FP, FN per le sequenze NR*

		risultato	
		VERO	***FALSO***
realtà	***VERO***	VP=90,57%	FN=9,43%
	FALSO	FP=1,74%	VN=98,26%

Tabella 6.5: *VP, VN, FP, FN in percentuale per le sequenze NR*

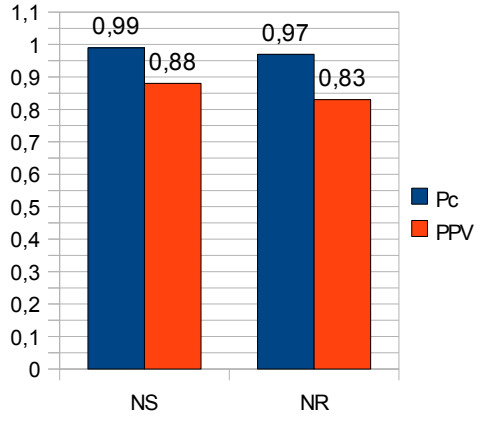


Figura 6.3: *grafico che rappresenta la probabilità di assegnazione corretta (in blu) e il valore predittivo positivo (in rosso) per sequenze NS e NR*

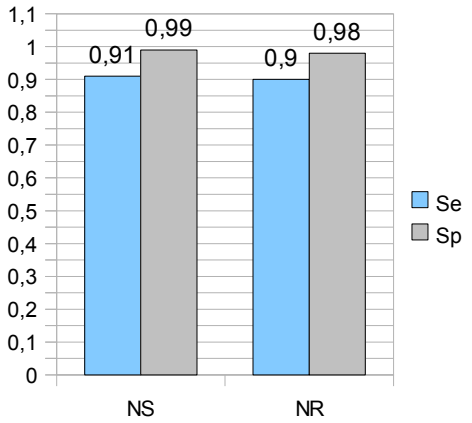


Figura 6.4: *grafico che rappresenta la sensitività (in azzurro) e la specificità (in grigio) per sequenze NS e NR*

6.2 Problematiche legate all'ampiezza e all'andamento del Raw Data

6.2.1 Problematiche legate all'ampiezza del Raw Data

La funzione `ampiezza_segnale` [Cap. 5 paragrafo 5.2.1] è stata realizzata per riconoscere segnali alti (SA), fuori scala (FS), e bassi (SB). Nelle tabelle 6.6-6.11, è possibile osservare i risultati della riuscita dell'algoritmo per quest'analisi. La probabilità di assegnazione corretta (Pc) è di 96.25% per SA, 99% per FS, 95.92% per SB (figura 6.6). L'algoritmo offre ottime prestazioni per il riconoscimento di sequenze FS, mentre per sequenze SA, probabilità di assegnazione corretta cala al 96.25%. Ciò può dipendere dalla percentuale (20%, parametro memorizzato nel file di configurazione) della serie dei "punti di inviluppo" presi in considerazione nella fase di sogliatura [Cap. 5 paragrafo 5.2.1, figura 5.10]. La maggior parte degli errori è dovuta ad un alto numero di FN (circa 9% rispetto l'1.9% del FP), perché non il 20% della serie dei "punti di inviluppo", ma meno, superano la soglia considerata. Si osserva dal primo riquadro di figura 6.5, come il segnale RD per la sequenza "B4.71.ab1" supera la soglia costante pari a 5000, ma la percentuale dei punti che la superano è inferiore al 20% di tutta le serie. Per questo il controllo fallisce, e la sequenza non viene classificata come SA.

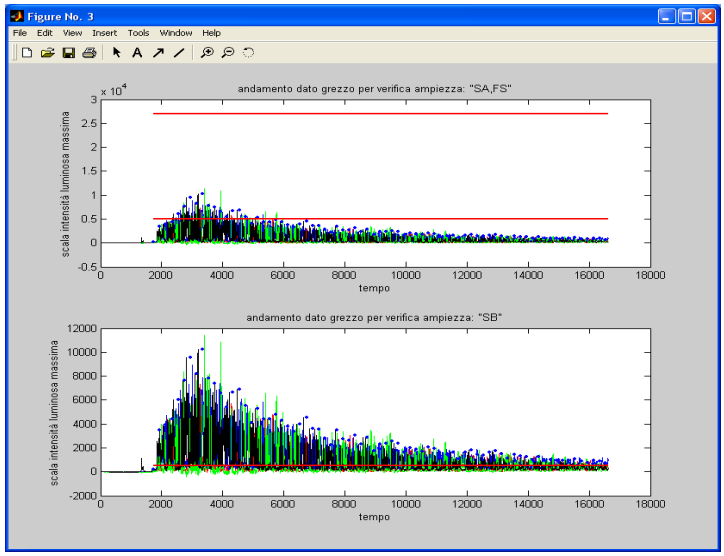


Figura 6.5: *plot di output della funzione* `ampiezza_segnale` *per la sequenza "B4.71.ab1"*

Un possibile miglioramento dell'algoritmo si potrebbe avere quindi riducendo questa percentuale. Bisogna far attenzione però a non considerare così solo i picchi dei dimeri di primer [Cap. 4 paragrafo 4.6] (ciò farebbe aumentare il numero dei FP), che normalmente hanno intensità elevata rispetto l'intero segnale. Scende invece a 95.92% la probabilità di assegnazione corretta per SB. Anche per questo caso, la maggior parte dell'errore dipende dai FN: per questo tipo di sequenze l'algoritmo ha difficoltà a distinguere una sequenza SB da una sequenza NR (figura 6.1). Per migliorare questo step si potrebbe pensare di analizzare non solo il RD, ma anche la sequenza di basi [Cap.5 figura 5.16] fornita dal file ABIF. La maggior parte delle sequenze NR sono caratterizzate, ma non sempre, da una sequenza costituita da sole e poche *N (noice)* al posto delle lettere (A, C, T, G) che rappresentano le basi azotate; ciò potrebbe esser una ulteriore strumento per operare la distinzione tra sequenze SB e NR. Di seguito sono riportate le tabelle che contengono il numero dei VP, FN, FP, VN (in valore assoluto e in percentuale) rispettivamente per SA, FS, SB. Le figura 6.6 riporta in valori percentuali la probabilità di assegnazione corretta e il valore predittivo positivo, mentre la figura 6.7 riporta la sensitività e la sensibilità per le tre problematiche trattate.

		risultato		
		VERO	***FALSO***	***TOTALE***
realtà	***VERO***	*VP=279*	*FN=28*	*307*
	FALSO	*FP=17*	*VN=876*	*893*

Tabella 6.6: *Numero di VP, VN, FP, FN per le sequenze SA*

		risultato	
		VERO	***FALSO***
realtà	***VERO***	*VP=90,88%*	*FN=9,12%*
	FALSO	*FP=1,9%*	*VN=98,1%*

Tabella 6.7: *VP, VN, FP, FN in percentuale per le sequenze SA*

		risultato		
		VERO	**FALSO**	**TOTALE**
realtà	**VERO**	*VP=56*	*FN=11*	*67*
	FALSO	*FP=1*	*VN=1132*	*1133*

Tabella 6.8: *Numero di VP, VN, FP, FN per le sequenze FS*

		risultato	
		VERO	**FALSO**
realtà	**VERO**	*VP=83,58%*	*FN=16,42%*
	FALSO	*FP=0.08%*	*VN=99.92%*

Tabella 6.9: *VP, VN, FP, FN in percentuale per le sequenze FS*

		risultato		
		VERO	**FALSO**	**TOTALE**
realtà	**VERO**	*VP=128*	*FN=36*	*164*
	FALSO	*FP=13*	*VN=1023*	*1036*

Tabella 6.10: *Numero di VP, VN, FP, FN per le sequenze SB*

		risultato	
		VERO	**FALSO**
realtà	**VERO**	*VP=78,05%*	*FN=21,95%*
	FALSO	*FP=1,25%*	*VN=98,75%*

Tabella 6.11: *VP, VN, FP, FN in percentuale per le sequenze SB*

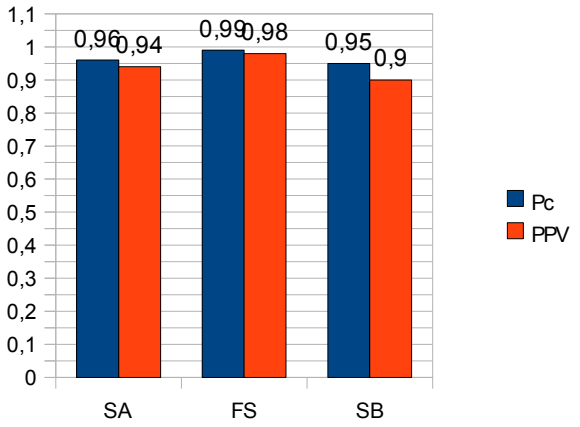


Figura 6.6: *grafico che rappresenta la probabilità di assegnazione corretta (in blu) e il valore predittivo positivo (in rosso) per sequenze SA, FS, SB*

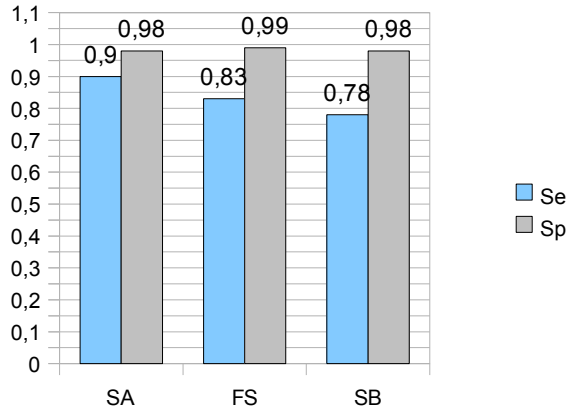


Figura 6.7: *grafico che rappresenta la sensibilità (in azzurro) e la specificità (in grigio) per sequenze SA, FS, SB*

6.2.2 Problematiche legate all'andamento del Raw Data

Per questo tipo di analisi, la probabilità di assegnazione corretta è del 91,33%. Su 331 sequenze classificate come M, l'84% di esse viene classificato correttamente (figura 6.10). Per poter migliorare il metodo si potrebbe pensare di usare tecniche di filtraggio del segnale ad esempio un filtro passa-basso, in modo da eliminare la componente di rumore e(t) che è presente nel RD. L'uso di questi filtri elimina le componenti ad alta frequenza, che sono tipiche del rumore, ma in parte anche del segnale utile y(t). Quindi è opportuno scegliere una frequenza di taglio adeguata. Inoltre c'è da considerare che l'uscita di un filtro passa-basso reale non è uguale a quella di un filtro passa-basso ideale, in quanto il segnale viene modificato anche in modulo. Per questo, l'elaborazione del RD attraverso filtraggio non è adatta invece per l'analisi della sua ampiezza.

Gli errori di tipo FP (Tabelle 6.12, 6.13), sono dovuti nella maggior parte dei casi al mancato riconoscimento del picco di PCR [Cap. 4 paragrafo 4.1], che rappresenta la fine del segnale y(t) di sequenziamento. Quando il picco della A non viene riconosciuto, i picchi di e(t) vengono attribuiti al segnale, per cui, assumendo valori bassi, la sequenza viene classificata come M (figura 6.8).

Gli errori di tipo FN invece interessano per lo più segnali in cui non viene

riconosciuta una variazione significativa dell'andamento, come il caso mostrato in figura 6.9.

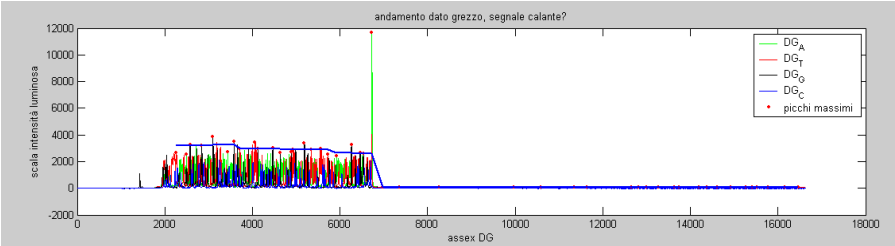


Figura 6.8: *plot fornito dalla funzione segnale_calante per la sequenza "577-09_146_F.ab1" . Il picco finale della PCR non viene riconosciuto e il segnale viene classificato come M, falso positivo.*

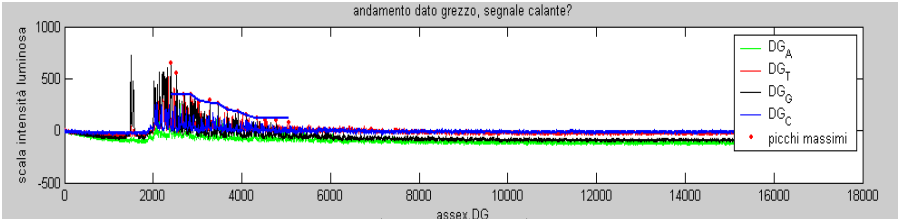


Figura 6.9: *plot fornito dalla funzione* `segnale_calante` *per la sequenza "NEOR2.ab1". La variazione dell'andamento del segnale non viene riconosciuta come significativa e di conseguenza il segnale non viene classificato come M, falso negativo.*

		risultato		
		VERO	***FALSO***	***TOTALE***
realtà	***VERO***	*VP=279*	*FN=88*	*367*
	FALSO	*FP=52*	*VN=817*	*869*

Tabella 6.12: *Numero di VP, VN, FP, FN per le sequenze M*

		risultato	
		VERO	***FALSO***
realtà	***VERO***	*VP=84,29%*	*FN=26,59%*
	FALSO	*FP=5,98%*	*VN=94,02%*

Tabella 6.13: *VP, VN, FP, FN in percentuale per le sequenze M*

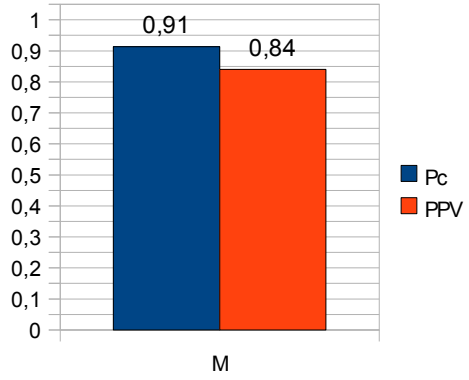


Figura 6.10: *grafico che rappresenta la probabilità di assegnazione corretta (in blu) e il valore predittivo positivo (in rosso) per sequenze M*

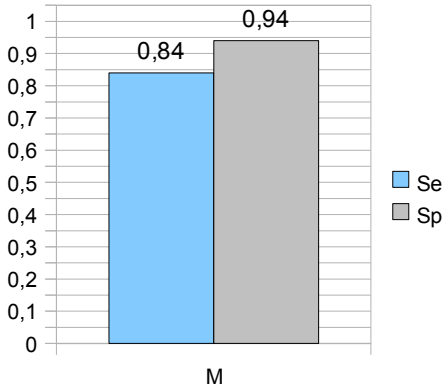


Figura 6.11: *grafico che rappresenta la sensibilità (in azzurro) e la specificità (in grigio) per sequenze M*

6.3 Segnale inarcato

Il numero dei FP (Tabella 6.14 e 6.15) compiuti in questo step dell'algoritmo sono dovuti agli errori compiuti dagli step precedenti. É il caso per esempio di sequenze in cui non viene riconosciuto correttamente l'istante t* [Cap. 5 paragrafo 5.1]. Per diminuire invece il numero dei FN si potrebbe modificare il parametro *rapporto_bande_SI* che è stato scelto pari a 8 [paragrafo 5.3 Cap.5]. Bisogna porre attenzione però nel caso di segnali SB, che hanno ampiezza bassa, e si rischia di selezionare erroneamente molte sequenze come SI (aumentando i FP). Sono comunque buone le percentuali trovate sia per la probabilità di assegnazione corretta che per il valore predittivo positivo, entrambe del 90% (figura 6.12). Minore è invece la sensitività dell'algoritmo, pari al 75% (figura 6.13).

		risultato		
		VERO	***FALSO***	***TOTALE***
realtà	***VERO***	*VP=273*	*FN=88*	*381*
	FALSO	*FP=28*	*VN=811*	*839*

Tabella 6.14: *Numero di VP, VN, FP, FN per le sequenze SI*

		risultato	
		VERO	***FALSO***
realtà	***VERO***	*VP=75,62%*	*FN=24,38%*
	FALSO	*FP=3,33%*	*VN=96,67%*

Tabella 6.15: *VP, VN, FP, FN in percentuale per le sequenze SI*

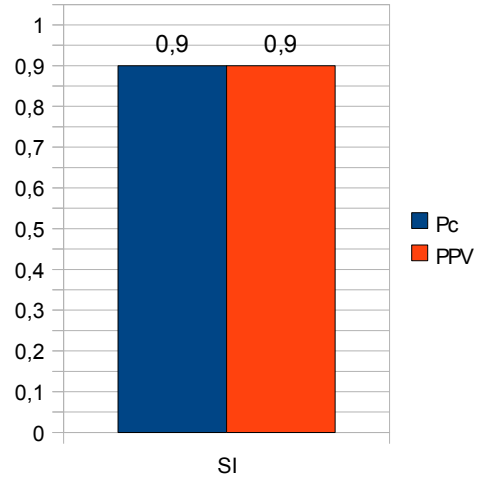


Figura 6.12: *grafico che rappresenta la probabilità di assegnazione corretta (in blu) e il valore predittivo positivo (in rosso) per sequenze SI*

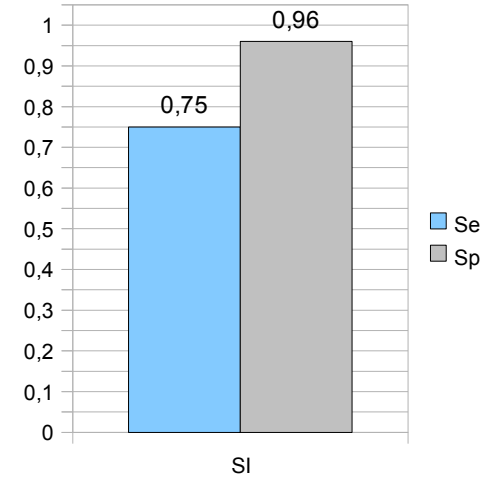


Figura 6.13: *grafico che rappresenta la sensibilità (in azzurro) e la specificità (in grigio) per sequenze SI*

6.4 Picchi multipli nell'Analyzed Data

Per quanto riguarda l'analisi dell'AD, ed in particolare il riconoscimento dei picchi multipli, l'algoritmo ha nel complesso buone prestazioni. Ci sono casi in cui però un picco viene riconosciuto singolo e non multiplo, ed è per questo che vengono compiuti i FN per le sequenze FD (Tabelle 6.22 e 6.23) e la sensibilità scende al 73% (figura 6.19). In questo casi, l'algoritmo non trova i picchi multipli lungo tutta la sequenza, ma solo in porzioni di essa, per questo viene classificata a volte come TD- oppure come DD-. In figura 6.14, viene mostrato l'output della funzione `presenza_picchiDoppi` [Cap.5 paragrafo 5.4] un esempio di sequenza FD, contenuta nel file "pGem-0000009409.ab1" (figura 6.15), i cui

picchi multipli non vengono correttamente individuati. L'algoritmo classifica la sequenza come TD-FD.

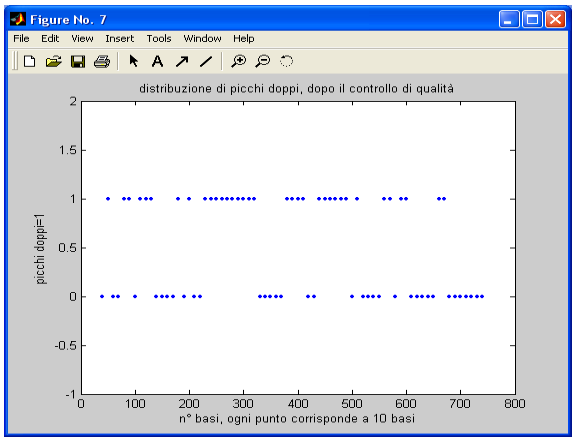


Figura 6.14: *plot di output della funzione presenza_picchiDoppi per individuare i picchi multipli lungo la sequenza dell'AD.*

Figura 6.15: *Analyzed Data della sequenza "pGem-0000009409.ab1" analizzato in figura 6.14.*

Per migliorare quest'analisi si potrebbe pensare di considerare un intorno del picco [Cap.5 paragrafo 5.4] più ampio. Questo può esser utile per sequenze FD, i cui picchi multipli sono dovuti soprattutto al rumore di strumentazione, risultando quindi meno delineati e sfasati. Nel complesso l'algoritmo riesce a riconoscere sequenze D, con una probabilità di assegnazione corretta del 96,33%, e sequenze FD con una probabilità di assegnazione corretta di circa 89%. L'algoritmo classifica 312 sequenze come FD e 104 sequenze come D, e rispettivamente il

90,06% ed il 93,27% sono correttamente classificate (figura 6.18).

L'algoritmo offre prestazioni meno buone per le classificazioni DD- e TD-. Di 178 sequenze classificate DD-, il 74,72% sono correttamente classificate, e di 155 sequenze classificate TD-, solo il 56,13% sono correttamente classificate (figura 6.18). La sensitività cala al 76% (figura 6.19) sia per TD- che per DD-. Per ottenere risultati migliori una possibile soluzione potrebbe esser quella di considerare blocchi con meno basi (ad esempio 50) e non di 100 per individuare i picchi multipli lungo la sequenza e per evitare di approssimare in un unico risultato la natura di 100 basi [Cap. 5 paragrafo 5.4]. Questa approssimazione tra l'altro è la fonte di errori per il mancato riconoscimento di sequenze DD-. Quando sono solo le ultime 50-20 basi a presentare picchi multipli, l'uso della mediana fa si che queste vengono ignorate, per cui nel risultato finale, che riporta la natura di tutte le 100 basi, se non di più, non compaiono. Inoltre quando non viene riconosciuto il picco della PCR e i picchi multipli interessano solo le ultime 250 basi, il software ignora queste basi [Cap.5 paragrafo 5.4, punto 3.2] e l'algoritmo compie un FN. È il caso riportato in figura 6.16, relativo alla sequenza "SABCASTELITS1R.ab1", che a partire dalla 219-esima base, presenta picchi multipli. Il software ignora le basi, commettendo un Falso Negativo, perché non viene classificata come DD-D.

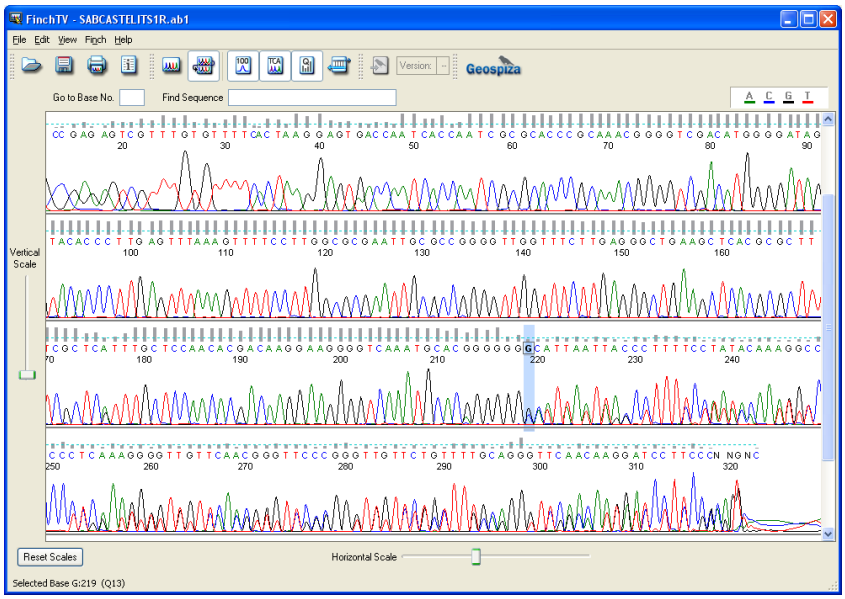


Figura 6.16: *Analyzed Data della sequenza "SABCASTELITS1R.ab1", DD-D.*

Quando l'algoritmo riconosce una sequenza DD-, riesce approssimativamente a dare una posizione della base in cui si sdoppiano i picchi [Cap. 5 paragrafo 5.4], e nel caso in cui questo si trovasse lontano della zona poli-A/T/C/G, l'omopolimero non viene individuato (FN di 46,67%, Tabella 6.25). Inoltre una parte dei FN per i poli-A/T/C/G viene compiuta qualora non venisse riconosciuta una sequenza DD-, in quanto solo questo riconoscimento implica la ricerca dell'omopolimero [Cap. 5 paragrafo 5.4]. La specificità e il valore predittivo positivo per questa problematica assumono valore massimo, avvero pari ad 1 (figura 6.19). Questo è dovuto al fatto che l'algoritmo è stato testato solo per il riconoscimento di questa problematica. Specificità e valore predittivo positivo non sono quindi, in questo caso, utili per valutare la performance.

La funzione `presenza_picchiDoppi` proposta fornisce un unico risultato (*intensitàTOT* [Cap. 5 paragrafo 5.4]) per riassumere la natura dei picchi dell'intera sequenza (D o FD). Nel caso in cui la sequenza presentasse sin dall'inizio un alto rumore di fondo, tale da riconoscere un FD, ma anche un punto a partire dal quale la sequenza si sdoppia divenendo ad esempio DD-D, è

necessario fare una distinzione delle due porzioni dell'AD: riconoscere quindi un FD, ma anche un DD-D fornendo due risultati distinti per quanto riguarda l'intensità dei picchi. Vale anche per le sequenze TD-. É il caso mostrato in figura 6.17, in cui la sequenza presenta picchi multipli D nelle prime 220 basi, dopo di queste la sequenza è FD. L'algoritmo classifica la sequenza come FD. L'unica operazione di mediana [Cap.5 paragrafo 5.4] quindi non è più adeguata. Di seguito sono riportate le tabelle col numero dei VP, VN, FP, FN per sequenze TD-, D, DD-, FD, poli-A/T/G/C, mentre in figura 6.18 è riportato il grafico con la probabilità di assegnazione corretta e il valore predittivo positivo, in figura 6.19 il grafico con la sensitività e la specificità per le sequenze appena elencate.

Figura 6.17: *sequenza "AP5THRREV.ab1". I primi 220 picchi sono multipli (TD-D), successivamente i picchi multipli calano di intensità. L'algoritmo classifica la sequenza come FD.*

		risultato		
		VERO	***FALSO***	***TOTALE***
realtà	***VERO***	*VP=87*	*FN=23*	*110*
	FALSO	*FP=68*	*VN=1022*	*1090*

Tabella 6.16: *Numero di VP, VN, FP, FN per le sequenze TD-*

		risultato	
		VERO	***FALSO***
realtà	***VERO***	*VP=79,09%*	*FN=20,91%*
	FALSO	*FP=6,24%*	*VN=93,76%*

Tabella 6.17: VP, VN, FP, FN in percentuale per le sequenze TD-

		risultato		
		VERO	***FALSO***	***TOTALE***
realtà	***VERO***	*VP=97*	*FN=37*	*130*
	FALSO	*FP=7*	*VN=1059*	*1066*

Tabella 6.18: *Numero di VP, VN, FP, FN per le sequenze D*

		risultato	
		VERO	***FALSO***
realtà	***VERO***	*VP=72,39%*	*FN=27,61%*
	FALSO	*FP=0,66%*	*VN=99,34%*

Tabella 6.19: *VP, VN, FP, FN in percentuale per le sequenze D*

		risultato		
		VERO	***FALSO***	***TOTALE***
realtà	***VERO***	*VP=133*	*FN=41*	*174*
	FALSO	*FP=45*	*VN=981*	*1026*

Tabella 6.20: *Numero di VP, VN, FP, FN per le sequenze DD-*

		risultato	
		VERO	***FALSO***
realtà	***VERO***	*VP=76,44%*	*FN=23,56%*
	FALSO	*FP=4,39%*	*VN=95,61%*

Tabella 6.21: *VP, VN, FP, FN in percentuale per le sequenze DD-*

		risultato		
		VERO	***FALSO***	***TOTALE***
realtà	***VERO***	*VP=281*	*FN=102*	*383*
	FALSO	*FP=31*	*VN=786*	*817*

Tabella 6.22: *Numero di VP, VN, FP, FN per le sequenze FD*

		risultato	
		VERO	***FALSO***
realtà	***VERO***	*VP=73,37%*	*FN=26,63%*
	FALSO	*FP=3,79%*	*VN=96,21%*

Tabella 6.23: *VP, VN, FP, FN in percentuale per le sequenze FD*

		risultato		
		VERO	***FALSO***	***TOTALE***
realtà	***VERO***	*VP=32*	*FN=28*	*60*
	FALSO	*FP=0*	*VN=1140*	*1140*

Tabella 6.24: *Numero di VP, VN, FP, FN per le sequenze P-A/C/G/T*

		risultato	
		VERO	***FALSO***
realtà	***VERO***	*VP=53,33%*	*FN=46,67%*
	FALSO	*FP=0%*	*VN= 100%*

Tabella 6.25: *VP, VN, FP, FN in percentuale per le sequenze Poli-A/C/G/T*

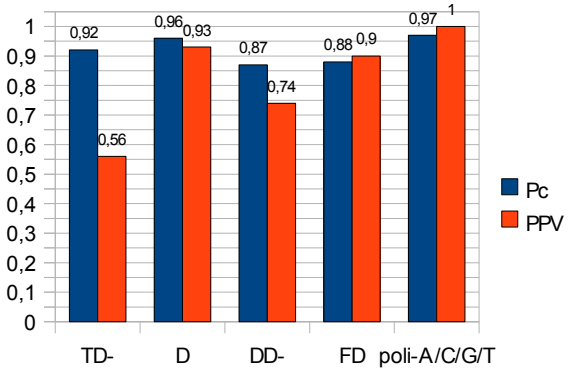


Figura 6.18: *grafico che rappresenta la probabilità di assegnazione corretta (in blu) e il valore predittivo positivo (in rosso) per sequenze TD, D, DD-, FD, poli-A/C/G/T*

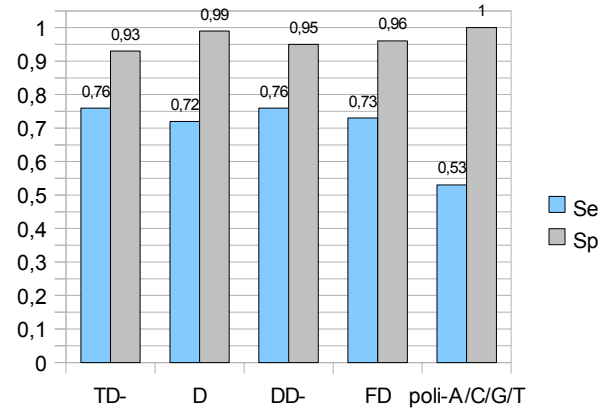


Figura 6.19: *grafico che rappresenta la sensibilità (in azzurro) e la specificità (in grigio) per sequenze TD, D, DD-, FD, poli-A/C/G/T*

Per concludere in figura 6.20 viene riportato il grafico con la probabilità di assegnazione compiuta dall'algoritmo per tutte le problematiche. Mediamente la probabilità supera il 93%.

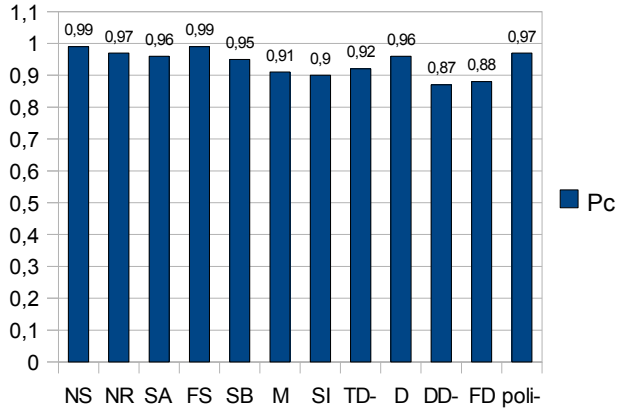


Figura 6.20: *probabilità di assegnazione corretta dell'algoritmo per ogni problematica.*

Capitolo 7

Conclusioni

L'obiettivo di questa tesi è stato quello di realizzare un algoritmo di analisi dei segnali di sequenziamento di DNA per valutare la presenza di problematiche che possono generare errori nella determinazione della sequenza nucleotidica. Queste problematiche possono essere causate sia dalla qualità e/o dalla quantità del campione da sequenziare, che dal processo di sequenziamento. Le cause appena elencate vanno ad inficiare la regolarità, l'ampiezza e l'andamento del segnale.

L'analisi per il controllo delle eventuali problematiche viene in genere compiuta da biologi esperti che valutano visivamente l'esito di ogni sequenziamento. Successivamente a questa analisi, il segnale viene assegnato ad opportune classi che ne rappresentano il problema. A seconda della gravità della problematica, si decide se è opportuno o meno sequenziare il campione. Il software per l'analisi delle problematiche sviluppato nell'ambito di questa tesi fornisce quindi un utile ausilio per la procedura di troubleshooting.

I segnali analizzati sono relativi alla metodologia di sequenziamento Sanger, e ottenuti dal sequenziatore Applied Biosystems 3730xl.

L'algoritmo è stato implementato in Matlab, e per compiere l'analisi dei segnali sono state utilizzate tecniche di approssimazione dei dati (filtro a media

mobile), di calcolo numerico della derivate (per valutare le variazioni del segnale), di peak detection. Il software ha lo scopo di riconoscere dai dati la presenza del segnale di sequenziamento, e una volta riconosciuto, il segnale viene sottoposto a diversi controlli che riguardano la sua ampiezza, il suo andamento e la sua regolarità.

L'esito del sequenziamento viene classificato come “no signal” (*NS)* o “no reaction” (*NR)* quando non è presente il segnale di sequenziamento (la differenza sta nella causa che ha provocato l'assenza del segnale: un *NS* dipende dal fallimento da parte della strumentazione ad esempio la corsa nei capillari, un *NR* dal fallimento della reazione chimica, e quindi la sintesi dei frammenti marcati). Altre problematiche sono legate all'ampiezza e all'andamento del Raw Data: a seconda dei casi, il segnale potrebbe esser classificato come “segnale basso” *(SB)*, “segnale alto” *(SA)*, “fuori scala” *(FS)*, segnale che “muore” *(M)*, segnale con “struttura” (*ST,* problematica che non è stata affrontata in questa tesi). La classe “segnale inarcato” *(SI)* è a se stante, e viene assegnata quando l'esito del sequenziamento fornisce un Raw Data con una deriva della linea di base. La presenza di picchi multipli nell'Analyzed Data rappresenta una terza tipologia di problematiche che portano a classificare il segnale come “doppio” *(D)*, “fondo doppio” *(FD)*, “diventa doppio” *(DD-)*, “tratto doppio” *(TD-)*.

Successivamente all'analisi, l'algoritmo fornisce in output la classificazione delle problematiche attribuite alla sequenza sotto analisi. L'algoritmo è stato addestrato su 167 sequenze e le sue performance sono state testate su 1200 sequenze.

Per quasi tutte le problematiche ricercate dall'algoritmo, la probabilità di assegnazione corretta supera il 90%, ed il positive predictive value supera l'80%, tranne per le sequenze TD- e DD-, per cui la precisione scende a 56% e 74%, rispettivamente. Questi risultati costituiscono un buon punto di partenza.

Non esistono in letteratura lavori che hanno affrontato il problema del troubleshooting automatico per segnali di sequenziamento Sanger, e pertando questa tesi è il primo lavoro che affronta l'analisi automatica del segnale per il

troubleshooting. Possibili sviluppi futuri comprendono un raffinamento dell'algoritmo per migliorarne la performance e l'utilizzo di tecniche di elaborazione del segnale per risolvere le problematiche legate ad artefatti dello stesso.

Elenco dei simboli

Simbolo	***Significato***
A	*adenina*
AD	*Analyzed Data*
BL	*"blob"*
C	*citosina*
D	*"doppio"*
DD-	*"diventa doppio"*
dNTP	*deossinucleoside trifosfato*
ddNTP	*dideossinucleoside trifosfato*
DG	*"agglomerati non incorporati"*
DP	*"dimeri di primer"*
emPCR	*PCR ad emulsione*
FD	*"fondo doppio"*
FN	*falso negativo*
FP	*falso positivo*
FS	*"fuori scala"*
G	*guanina*
M	*"segnale che muore"*
NR	*"no reaction"*
NS	*"no signal"*
Pc	*Probabilità di assegnazione corretta*
PCR	*Polimerase chain reaction*
PPV	*Valore predittivo positivo*
poli-A/T/C/G	*Omopolimero- A/T/C/G*
PR	*"problemi di risoluzione"*
RD	*Raw Data*
SA	*"segnale alto"*
SB	*"segnale basso"*
Se	*sensitività*

Simbolo	***Significato***
SI	*“segnale inarcato”*
Sp	*specificità*
ST	*“segnale con struttura”*
T	*timina*
TD	*“tratto doppio”*
VN	*vero negativo*
VP	*vero positivo*

Ringraziamenti

Ringrazio vivamente tutto il personale dell'azienda BMR Genomics, dai biologi agli informatici per avermi fornito tutti i dati indispensabili per la realizzazione della tesi. In particolare ringrazio Barbara e Fabrizio che mi hanno seguito costantemente durante i sei mesi di lavoro, e risposto alle mie domande in maniera esaustiva e chiara.

Inoltre ringrazio sentitamente la prof. Barbara Di Camillo che è stata sempre disponibile a dirimere i miei dubbi durante la stesura di questo lavoro.

Ringrazio i professori, che con la loro precisione e passione hanno contribuito alla mia formazione professionale, tra cui il prof. Sparacino. Ringrazio tutti i miei colleghi studenti, in particolare Francesca e Marco.

Grazie Padova per avermi accolta, ospitata e accompagnata fino a questo traguardo, che oggi festeggerò con te.

Inoltre desidero ringraziare Anna e Roberto per il supporto e il “tifo” che mi hanno fatto.

Infine ho desiderio di ringraziare i miei genitori e le mie sorelle per il sostegno e il grande aiuto che mi hanno dato.

Bibliografia

1. Dale, Jeremy, W. (c2008) *Dai geni ai genomi, applicazioni del DNA ricombinante,* seconda edizione, Napoli, EdiSES.
2. Ewing B, Green P. (1998) “Base-calling of automated sequencer traces using phred. II. Error Probabilities ”, *Genome Research,* Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA, Vol 8, 3, pp. 186-194.
3. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) “DNA sequencing with chain-terminating inhibitors”, *Biochemistry,* Proc. Natl. Acad. Sci. USA, Vol 74, 12, pp. 5463-5467.
4. Savada, D., et al. (2009) B*iologia la cellula,* terza edizione, Bologna, Zanichelli.
5. Wu SM, Blomberg LA, Chan WY (1996) ”Recovery of unlabeled PCR product from polyacrylamide gel for sequencing”, *Biotechniques,* Dept. of Pediatrics, Georgetown University Children's Medical Center, Washington, DC 20007-2196, USA, Vol 21, 3 , (Sep), pp. 358-362.
6. *Automated DNA sequencing, Chemistry Guide,* (1998), Applied Biosystems, Capitolo 7, pp. 1-38.
7. *Applied Biosystems 3730/3730xl, DNA Analyzers, Sequencing Chemistry guide*, (2002), cap. 1, 2, 4, Appendix A, B.
8. *The qiagen guide to template Purification and DNA Sequencing,* (1998), 2a edizione, QIAGEN.