UNIVERSITA' DEGLI STUDI DI PADOVA
Facoltà di Scienze MM. FF. NN.

Sede Amministrativa: Dipartimento di Biologia

CRIBI, UNIVERSITÀ DEGLI STUDI DI PADOVA

DOTTORATO DI RICERCA IN BIOTECNOLOGIE
CICLO XIX

Regolatori di RNA e loro sequenze bersaglio

Coordinatore : Ch.mo Prof. Giuseppe Zanotti
(Dipartimento di Chimica)

Supervisore : Ch.mo Prof. Giorgio Valle
(Dipartimento di Biologia)

Dottoranda : Dr. Alessandra Bilardi

DATA CONSEGNA TESI
31 Gennaio 2008

Ai miei due uomini che mi rendono
ogni giorno molto più che dottoressa

Sommario

La bioinformatica è una materia interdisciplinare che integra diversi campi di biologia, chimica, biofisica, biostatistica e informatica al fine di risolvere problemi biologici. Una delle questioni più importanti è l'annotazione strutturale e funzionale dei geni nascosti nella sequenza genomica. Di notevole interesse sono anche la predizione delle strutture proteiche, la scoperta di farmaci e la manipolazione di vie metaboliche. Questi problemi si possono affrontare in tempi più ragionevoli utilizzando metodiche computazionali, attraverso l'uso della bioinformatica come catalizzatore.

In questi anni di dottorato mi sono accupata di annotazione genica sia strutturale che funzionale. Nei procarioti, ho studiato i terminatori Rho dipendenti (Rho Dependent Terminator (RDT)) come attenuatori e/o terminatori delle unità trascrizionali (Transcriptional Unit (TU)). Nei batteri gram negativi, è importante studiare i RDT per conoscere le TU alternative degli operoni ma anche perché è singolare che un sistema così complesso serva per la regolazione di pochi geni. Conoscere la posizione dei RDT e dei terminatori Rho indipendenti servirebbe a dividere i geni in operoni e così contribuire alla loro annotazione strutturale e funzionale. In letteratura i RDT documentati sono meno di 40 e ce ne sono solo 18 in *Escherichia coli*. Non esistono programmi di predizione dei RDT e nemmeno sequenze consenso ma solo alcune descrizioni delle caratteristiche dei RDT. Ho disegnato il primo algoritmo per la predizione dei RDT e ho proposto il primo consenso della sequenza dei RDT con una matrice pesata. Per confermare le predizioni ottenute con l'implementazione dell'algoritmo di predizione e l'allineamento della matrice e comprendere quale ruolo funzionale può avere la proteina Rho, ho disegnato degli oligo e progettato delle condizioni per degli esperimenti di microarray. Con

gli esperimenti di microarray è stato confermato dai controlli che la metodica di ricercare i RDT predetti nel leader con gli oligo che si sono accesi o spenti funziona. Ma dovranno essere disegnati ulteriori esperimenti per poter validare i RDT predetti.
Negli eucarioti ho studiato i microRNA delle piante e i loro bersagli. In letteratura e online sono descritti e scaricabili molti programmi di predizione di geni codificanti proteine, geni di RNA ribosomale, transfer e microRNA ma sono pochi quelli che predicono geni e siti bersaglio di microRNA in modo specifico per le piante. Il gruppo del prof. Valle partecipa al progetto di sequenziamento e annotazione del genoma di *Vitis vinifera*. Per questo motivo ho implementato e testato numerosi programmi di predizione di geni e bersagli di microRNA provenienti da studi sia su piante che su animali. Ho proposto un metodo integrato che prevede l'uso in contemporanea di risultati di più programmi di predizione ed ho ottenuto dei risultati positivi. I risultati di predizione saranno validati sfruttando una nuova metodica di sequenziamento, il Sequencing by Synthesis (SBS) presentato da Solexa.
Inoltre, ho contribuito alla visualizzazione di dati e all'interrogazione avanzata di database per i progetti dei genomi a cui partecipa il laboratorio. Ho acquisito esperienza nel gestire dei browser genomici e dei tool per interrogazioni avanzate che sono forniti dal progetto GMOD. GMOD sta per Generic Model Organism Database ed è una collezione di software per la creazione e gestione di database biologici su scala genomica. Ho implementato nuovi programmi e rivisto degli script che fanno parte di questo tool che verranno proposti nella prossima distribuzione del software. Il tool che ho implementato presenta delle correzioni di errori che ci sono negli script della versione distribuita e ha una velocità comparabile alle versioni più veloci fornite dalla distribuzione stessa. I risultati di configurazione e miglioramento degli script riguardo la visualizzazione e l'interrogazione dei database sono importanti in quanto il laboratorio del prof. Valle sarà responsabile dello sviluppo della piattaforma di annotazione del progetto di sequenziamento del genoma di *Vitis vinifera*.

Abstract

Bioinformatics is an interdisciplinary subject which integrates fields like biology, chemistry, biophysics, biostatist ics and computer science to solve biological problems. Its goal is to enable the discovery of new biological insights as well as to create global perspective from which unifying principles in biology can be discerned. The first and foremost biological problem is the annotation of functional genes hidden in sequenced genomes which can be effectively sca nned by computational approaches. The other problems include protein structure prediction, drug discovery, metabolic pathway manipulation and much more. These problems can be consolved in a faster way by using bioinformatics as a catalyst.

In these four years I have been involved in the annotation of structural and functional genes. In prokariotes I study Rho dependent terminator (RDT) as attenuator and/or terminator of transcriptional unit (TU). In gram negative bacteria, RDT study is important to understand alternative TUs inside each operon so as it is weighty to identify Rho independent terminator (RIT). If we know where terminators (RDT and RIT) are located then we could cluster genes in operons and so we could contribute in annotating structural and functional genes. In the literature, documented RDT are less than 40 and there are 18 only in *Escherichia coli*. There is not a RDT prediction program and not even a RDT consensus sequence but only some features that they describe RDT shares. I designed first algorithm about RDT prediction and I suggested first RDT consensus sequence with a weighed matrix. I designed oligos and conditions about microarray experiments in order to confirm our RDT predictions and to understand about Rho functional correlation with genes that they have got a RDT. Microarray results confirmed that I can use intensity of oligos to calculate if one putative RDT exits. But

to confirm putative RDTs we do other microarray experiments.

In eukariotes I study plants microRNA and their targets as genes annotated hardly and their function. Standard automatic gene annotation identifies and/or predicts genes that they will be translate. There are programs that they identify and/or predict genes about transfer, ribosomal and microRNA but there is little about microRNA and their targets in plants. I am involved in the annotation of microRNA genes and their targets in *Vitis vinifira* sequencing genome project. I implemented and tested more finder and predictor programs about microRNA genes and their targets for plants and animals. I proposted an integrated method with which I obtain positive results. Results about microRNA gene prediction could be validate with highthroughput sequencing approach. Results about microRNA target prediction protocol could be the starting point for lab experiments necessary for the validation of microRNA targets.

Beyond, I contributed to visualize data and to do advanced queries with databases about lab projects. I gained through experience and manage a generic genome browser and a advanced query tool about GMOD project. GMOD is the Generic Model Organism Database Toolkit, a collection of software tools for creating and managing genome-scale biological databases. I realized new versions of some scripts about GMOD tool that they will be proposed in the next GBrowse release. I implemented and tested scripts more fast than real GBrowse scripts and without some bugs. Results of configuration and design about visualization and querying of databases are important since our genomics research group will be the responsible for the development of the *Vitis vinifera* annotation platform.

Ringraziamenti

Volevo ringraziare tutte le persone del laboratorio del prof. Valle.
In particolare Davide Campagna, Laura Colluto, Chiara Rigobello, Andrea Telatin, Alessandro Vezzi, Nicola Vitulo per avermi aiutato nella stesura della tesi ma soprattutto per aver contribuito con suggerimenti e tempo.
Inoltre volevo ringraziare anche Elisa Caniato e Chiara Romualdi per i suggerimenti relativi alle metodi statistici che ho utilizzato.
E per finire volevo ringraziare Alessandro Albiero, Stefano Campanaro, Micky Del Favero, Erika Feltrin, Claudio Forcato, Fabrizio Levorin per avermi aiutato con suggerimenti di carattere tecnico-computazionale.

Indice

Capitolo 1

Introduzione

La bioinformatica è una materia interdisciplinare che integra diversi campi di biologia, chimica, biofisica, biostatistica e informatica al fine di risolvere problemi biologici. Una delle questioni più importanti è l'annotazione strutturale e funzionale dei geni nascosti nella sequenza genomica. Di notevole interesse sono anche la predizione delle strutture proteiche, la scoperta di farmaci e la manipolazione di vie metaboliche. Questi problemi si possono affrontare in tempi più ragionevoli utilizzando metodiche computazionali, attraverso l'uso della bioinformatica come catalizzatore.

In questi anni di dottorato hostudiato in modo approfondito i terminatori Rho dipendenti (Rho Dependent Terminator (RDT)). Conoscere quali geni sono regolati dai RDT è importante sia per annotare la loro funzione che per comprendere la struttura degli operoni e perciò per contribuire all'annotazione strutturale genica di batteri appena sequenziati.

Sono stati affrontati inoltre degli studi computazionali per la predizione di geni di microRNA, in quanto il gruppo del prof. Valle partecipa al progetto di senquenziamento del genoma di *Vitis vinifera* (1) e le metodiche computazionali di base utilizzate per la ricerca dei RDT sono le stesse.

Sono stati fatti infine degli studi di carattere tecnico per il perfezionamento di tool forniti dal progetto Generic Model Organism Database (GMOD). Di seguito sono riportati i dati presenti in letteratura riguardo i RDT, i microRNA e il progetto GMOD.

1.1 RDT

Il sequenziamento di nuovi genomi produce ogni giorno nuovo materiale da annotare ed analizzare. L'annotazione dei genomi batterici può risultare utile, ad esempio, per progettare antibiotici, sfruttare determinate proteine nell'ingegneria genetica o conoscere il ruolo di un microrganismo all'interno del suo ambiente.
Uno dei maggiori problemi dell'annotazione strutturale dei geni nei batteri è quella di predire correttamente gli operoni e le unità trascrizionali (TU) alternative che possono derivare dalla loro trascrizione. Nei procarioti le TU possono essere mono o policistroniche e possono comprendere tutti i geni dell'operone o solo una parte (ad esempio, solo i primi due o solo gli ultimi due).

1.1.1 Operoni

Comparando le sequenze dei genomi di batteri ed archea, è stato notato che la maggior parte degli operoni si sono largamente riarrangiati durante l'evoluzione; ciò può avvenire sia all'interno dello stesso operone sia tra operoni diversi. Solamente una minoranza, tipicamente quelli codificanti per proteine che compongono uno stesso complesso, o interagiscono nella stessa via metabolica, è conserva nella maggior parte dei genomi. Gli operoni trasferiti, riarrangiati, rimossi e probabilmente anche la loro formazione sono i maggiori fattori coinvolti nell'evoluzione di batteri ed archea (2). Questo comporta una notevole difficoltà nel predire l'esatta struttura degli operoni.
Dall'ampia collezione di sequenze genomiche analizzate, è emerso che i geni contigui, coespressi e collegati per la funzionalità delle proteine codificate molto spesso formano un operone. Se un insieme di geni sintenici conservati appare in due o, meglio ancora, in tre o più genomi batterici tassonomicamente vicini, ci sono pochi dubbi sul fatto che questi geni formino un operone e siano funzionalmente correlati. Se uno (o più) di questi geni non ha una funzione conosciuta o ha una funzione definita in termini generali, è possibile predirne la funzione (3).
Questi concetti sono alla base della maggior parte dei programmi disponibili per la predizione

concetto	**e programmi che lo sfruttano**
uguale direzione dei geni	tutti
distanza intergenica	(4),(5),(6),(7),(8),(9),(10),(11),(12)
predizione TSS, TFBS, SD o RIT	(4),(5),(13),(14)
classificazione funzionale dei geni	(5),(6),(7),(9),(10),(11),(12),(13)
comparazione sintenica dei geni	(4),(5),(8)
comparazione sintenica di TSS, TFBS e SD	(5)
valutazione di geni contigui per (*)	(5),(8),(13)
valutazione di geni contigui per (+)	(9),(15),(16)
espressione genica per (x)	(15),(16),(17)

Tabella 1.1: Concetti considerati nei programmi esistenti. TSS, sito di inizio della trascrizione; TFBS, sito di legame di fattore di trascrizione; SD, Shine-Dalgarno; (*), conservazione dell'adiacenza; (+), distanza intergenica, codon usage, classe funzionale, via metabolica, monomeri della stessa proteina, trasporto dello stesso substrato di una via metabolica, gene a monte o a valle della coppia correlato; (x), identificazione di regioni codificanti e intergeniche

degli operoni. In Tabella 1.1 sono riportati i principali programmi di predizione e le strategie adottate.

L'accuratezza maggiore, ancora oggi, è stata ottenuta dall'algoritmo di Wang *et al.* (4) in cui si arriva al 91% di predizioni corrette degli operoni sfruttando pochi concetti fondamentali: comparazione sintenica per allineamento di sequenze e predizione dei terminatori Rho indipendenti (RIT). Il metodo sembra aver avuto buoni risultati perchè il batterio preso in esame (*Staphylococcus aureus*) cresce ed è virulento anche con la delezione del gene *rho*, perciò la maggior parte degli operoni probabilmente termina con un RIT.
Nei batteri gram negativi, invece, la delezione del gene *rho* è letale. Nel gruppo del prof. Valle è stato sequenziato *Photobacterim profundum* SS9 (18; 19; 20), un gram negativo che presenta la proteina Rho che potrebbe essere essenziale come per *Escherichia coli*. Implementare un algoritmo che integri la maggior parte dei parametri elencati in Tabella 1.1 e che tenga conto della sintenia, della predizione dei RIT e RDT, potrebbe portare a una buona predizione degli operoni e delle TU (anche alternative) per nuovi genomi batterici gram negativi ed essere un punto di partenza per sviluppare un'annotazione strutturale e funzionale più fine.

1.1.2 Caratteristiche dei RDT

Nei procarioti sono conosciuti due tipi di terminatori della trascrizione che si distinguono per il loro meccanismo d'azione e la sequenza di RNA coivolta. Quando la RNA polimerasi incontra un terminatore intrinseco (precedentemente definito con RIT), rilascia spontaneamente il trascritto nascente. Quando incontra un RDT, il rilascio della molecola di RNA dipende dall'azione di una proteina chiamata Rho (21). Gli RDT sono coinvolti nell'espressione genica come attenuatori nel leader (5' non tradotto) e come terminatori all'interno o alla fine degli operoni. Un RDT è formato da tre parti distinte che possono superare a volte le 200 basi. La prima parte del RDT è il sito Rho UTilization (rut), destinazione della proteina Rho, che fa parte di un segmento del trascritto nascente al quale Rho si può legare ed è essenziale per l'inizio della terminazione (23; 24; 25). La seconda parte del RDT è il secondo sito di attacco di Rho ed è essenziale per l'attività di traslocazione dell'elicasi (26; 27). L'ultima parte del RDT è il Transcription Stop Point (tsp), regione in cui la RNA polimerasi si ferma alcuni momenti durante l'allungamento del trascritto in assenza di Rho (28; 29) e sito di terminazione della trascrizione. Rho si lega a rut instaurando un legame col trascritto nascente; ciò porta ad una modificazione conformazionale della proteina che premette il legame del dominio con attività di traslocazione al trascritto. Il dominio si sposta lungo il trascritto spostando la proteina Rho verso la RNA polimerasi; l'attività di traslocazione provoca il rilascio del trascritto che si distacco dal DNA nella regione del tsp.
Ciascun monomero di Rho interagisce con nove residui nucleotidici di RNA ricchi di citosina (30; 31; 32). Il sito rut è una regione ricca di citosina e povera di guanina ed è lungo almeno 40 basi (32; 33) e, alcune volte, presenta una debole struttura secondaria (33). Il sito in cui si lega il dominio con attività ATPasica è un'altra regione ricca di citosina, povera di guanina, spesso ricca di uracile con una composizione CU>>AG (30; 34) e lunga almeno 40 basi (27). La regione tsp può essere distante fino a 150 basi a valle del sito di rut, è lunga al più 100 basi e consiste di gruppi di punti di stop (Stop Point (sp)) con intervalli di 20-30 basi (32; 35). Un sp è il sito dove la RNA polimerasi, in assenza del legame con Rho, rallenta durante l'elongazione del trascritto. I punti di stop si estendono da 5 a 40 basi (32) e sono preceduti da regioni ricche

di citosina e guanina (Strong Island (si)) e/o da almeno 3 basi di timina (36).
In letteratura, i RDT sono descritti nel leader degli operoni (37; 38; 39; 40; 41; 42; 43; 44), all'interno degli operoni in siti intragenici (45; 46; 47; 48; 49; 50), o intergenici (51; 52; 53), e alla fine degli operoni (23; 54; 55; 56; 57).
La maggior parte dei batteri gram negativi presenta la proteina Rho, ma solo in *Escherichia coli* sono documentati 18 RDT. Per la maggior parte, sono geni coinvolti nel ripiegamento strutturale delle proteine, nella traduzione e biogenesi di RNA ribosomale e nella motolità batterica (vedi Tabella 2.3). I RDT documentati sono meno di una quarantina e non sono ancora chiari i motivi per cui un sistema così complesso sembra regolare pochi geni rispetto ai RIT. Inoltre non è ancora noto se esistono delle vie metaboliche regolate dalla proteina Rho o se la sua regolazione è limitata solo a singoli geni.
Per cercare di rispondere a questi quesiti, sarebbe importante predire le regioni dei RDT per eseguire degli esperimenti mirati all'identificazione di nuovi geni regolati da Rho. Ad oggi, non esistono programmi di predizione dei RDT e nemmeno un consenso della sequenza riconosciuta dalla proteina Rho però si sa che tutti i RDT documentati presentano una struttura simile (RDT organizzato in tre parti) e che le loro specifiche possono essere sfruttate per implementare un algoritmo per la loro ricerca e/o predizione. Si è pensato perciò di disegnare un algoritmo per la ricerca dei RDT in modo da poter preparare, in un secondo momento, un protocollo più completo per l'identificazione degli operoni.

1.2 microRNA

Per molti anni, i geni non codificanti (noncoding RNA (ncRNA)) sono stati considerati *relitti* di una origine della vita basata sull'RNA (58; 59) e molti di loro erano sconosciuti. Grazie ai nuovi approcci sperimentali e computazionali (60; 61; 62; 63; 64), si è cominciato a comprendere che molti dei ncRNA hanno un ruolo biologico e esso è, in alcuni casi, altamente specializzato.
Negli anni '90, cominciarono ad essere descritti i primi microRNA in *Caenorhabditis elegans* (65; 66) e nel 2002, alcuni gruppi trovarono i primi microRNA nelle piante (67; 68; 69). Diversi

laboratori hanno dimostrato che i microRNA hanno un ruolo importante nella crescita e nello sviluppo delle piante (70).

La biogenesi ed i meccanismi d'azione dei microRNA sono illustrati in Figura 1.1.

I microRNA sono una classe endogena di small RNA non codificanti che, normalmente, sono formati da 20-22 basi negli animali e 20-24 basi nelle piante (68; 70). I loro precursori (pre-miRNA) hanno una struttura a loop caratterizzata da bassa energia libera, come dimostrato da Bonnet *et al* (81) Molti microRNA sono evolutivamente conservati non solo tra le varie specie, ma anche tra generi diversi, alcuni dai nematodi all'uomo (82; 83; 84) o dalle felci ai mono/dicotiledoni (67; 68; 85; 86; 87).

Nelle piante i precursori dei microRNA sono meno conservati rispetto a quelli degli animali: generalmente solo i microRNA maturi sono conservati mentre negli animali sono conservati i microRNA precursori (70). L'appaiamento che avviene tra microRNA maturo e microRNA star (vedi Figura 1.1) o tra microRNA maturo e sito bersaglio ha delle peculiarità che sono state analizzate e implementate in quasi tutti i programmi di predizione. L'allineamento prevede che non vi siano più di sette basi non appaiate fra i primi 25 nucleotidi centrati sul microRNA maturo. Di questi, più di tre non possono essere consecutivi e non più di due sono senza base corrispondente nell'altro filamento (88). Inoltre, in questi allineamenti, avvengono appaiamenti tra le basi guanina e uracile. Esistono molti programmi atti all'identificazione e/o predizione di geni codificanti, molti meno sono invece quelli per RNA ribosomale, trasfer e geni e bersagli dei microRNA (89).

Il gruppo del prof. Valle partecipa al progetto del sequenziamento genomico di *Vitis vinifera* (1), con contributi di sequenziamento genomico e di annotazione genica. Per partecipare all'annotazione genica e funzionale del progetto sono stati condotti degli studi sui microRNA.

Ci sono quattro approcci per identificare i microRNA: screening genetico (65; 66), clonaggio diretto dopo isolamento di small RNA (90), strategie computazionali (91) ed analisi di Expressed Sequence Tag (EST) (87).

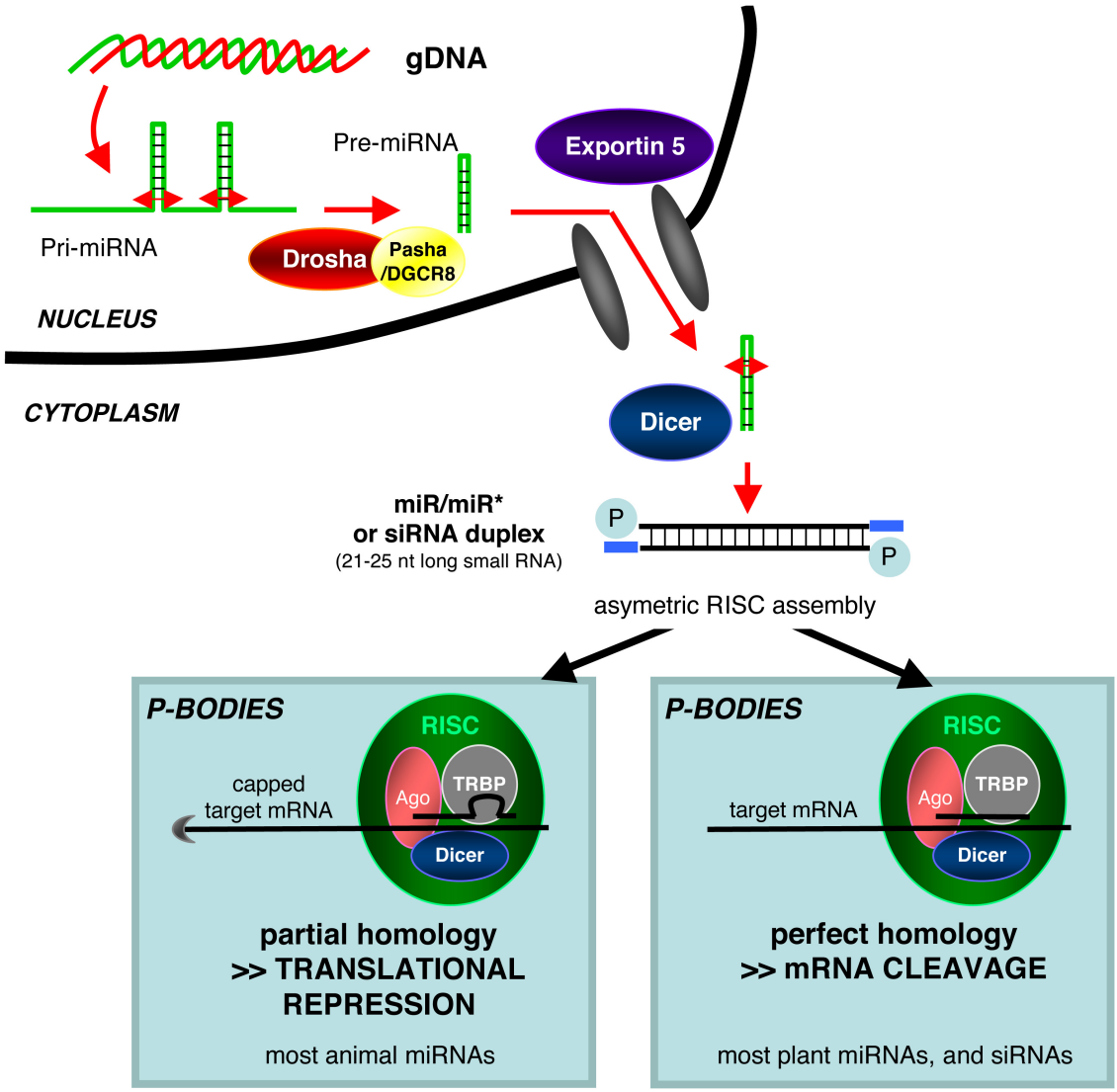


Figura 1.1: Biogenesi e azione dei microRNA (figura tratta da (71)). Il trascritto primario (Primary transcript of microRNA gene (pri-miRNA)) contiene uno o più microRNA. È trascritto dalla RNA polimerasi e processato da un complesso proteico che contiene una endonucleasi RNAasi III (Drosha, negli animali (70; 72), Dicer-like 1 enzyme (DCL1), nelle piante (73; 74)). Questo complesso riconosce la struttura a doppia elica del pri-miRNA e taglia esattamente alla base della forcina (loop), creando il precursore del microRNA (Precursor of microRNA (pre-miRNA)) lungo da 60-70 a 350 basi (75). Il pre-miRNA è trasportato nel citoplasma (dall'esportina 5, negli animali (76; 77; 78), da HASTY, nelle piante (69)) dove viene processato da una seconda edonucleasi RNAsi III (Dicer, negli animali (79), DCL1, nelle piante (80)) nel microRNA a doppio filamento (miR/miR* duplex, dove miR* è il filamento complementare del microRNA chiamato microRNA star). Il miR/miR* duplex è caricato nel complesso proteico RNA-induced Silencing Complex (RISC). Il miR serve come guida per riconoscere mRNA bersaglio mentre il miR* viene tagliato da una proteina del complesso (Argonaute (Ago)) e liberato nel citoplasma. Negli animali, il metodo di inibizione dell'espressione genica avviene principalmente per repressione della traduzione atta dall'interferenza del complesso con il riconoscimento del 5' del mRNA (cap). Nelle piante, generalmente, il complesso porta alla degradazione del mRNA bersaglio. Questa fase avviene nella regione chiamata Processing Bodies (P-bodies) nella quale ci sono gli mRNA non tradotti e in cui può avvenire la loro degradazione.

1.2.1 Approcci di sequenziamento

Il primo approccio per scoprire nuovi microRNA è stato attraverso clonaggio e sequenziamento dei singoli small RNA usando metodi tradizionali (vedi Figura 1.2). La maggioranza dei microRNA correntemente conosciuti sono stati identificati con questo approccio. Recentemente è stato dimostrato che il metodo di Massively Parallel Signature Sequencing (MPSS) può essere applicato al sequenziamento di small RNA (90). I dati di MPSS su small RNA hanno documentato che alcuni microRNA sono tra i più abbondanti degli small RNA.

1.2.2 Strategie computazionali

I metodi per l'identificazione dei microRNA sfruttano le caratteristiche strutturali e di allineamento della molecola. Questi approcci sono integrati con i risultati di analisi della struttura secondaria dell'RNA e conservazione di sequenze di microRNA considerate in genomi correlati. Una delle più grandi sfide nel cercare i geni dei microRNA è che i genomi eucariotici contengono un numero molto alto di sequenze ripetute che quando vengono trascritte possono formare dei loop. Il problema di base della predizione dei geni di microRNA è perciò selezionare i loop *giusti* per ridurre il numero di falsi positivi.
I metodi disponibili per la ricerca di geni e bersagli di microRNA si basano su tre principali strategie: l'identificazione e classificazione dei loop, l'omologia con i microRNA documentati e/o l'omologia con i ncRNA identificati su due o più genomi (89). I passaggi principali per la ricerca dei microRNA sono riassunti nella Tabella 1.2.

In ogni predizione, ciascuno dei punti può essere considerato come un filtro che la sequenza candidata potrà passare o no. Questo approccio ha il vantaggio che ciascun punto può essere migliorato o riutilizzato all'interno della stessa predizione. Comunque, queste metodologie hanno degli svantaggi poiché un candidato che è buono secondo un criterio di predizione può essere eliminato da un altro filtro che utilizza diversi valori di soglia. In generale, i criteri di predizione avrebbero dei risultati migliori se si potessero usare tutte le informazioni in modo simultaneo per determinare una predizione ottimale con un modello statistico integrato

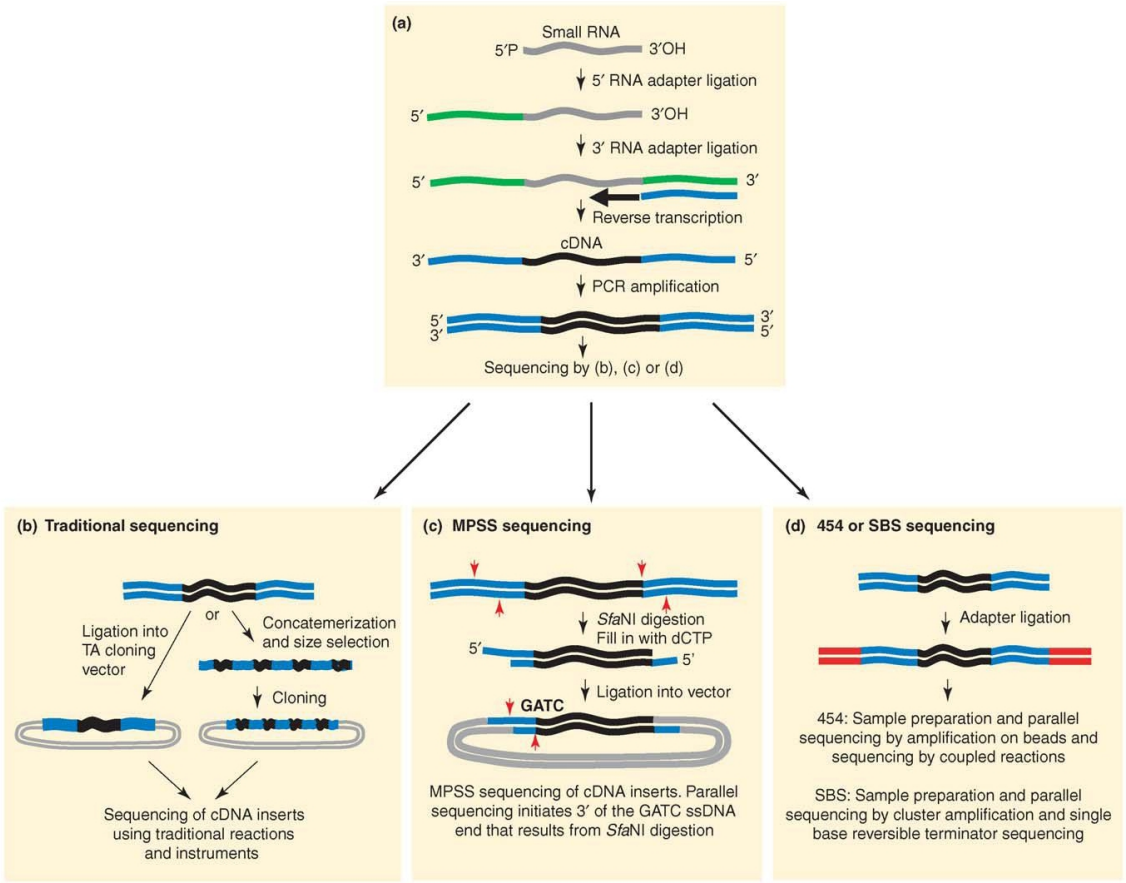


Figura 1.2: Metodi di clonaggio e sequenziamento di small RNA (figura tratta da (92)). **(a)** I metodi per clonare small RNA cominciano con la ligazione di adattatori di RNA agli small RNA inizialmente isolati per frazionamento su un gel di poliacrilamide (67; 68; 93). Per ottenere una quantità sufficiente di templato sono usati un numero basso di cicli di PCR. Gli adattatori di RNA sono indicati in verde, gli small RNA in grigio, il cDNA in blu (adattatori) e in nero (small RNA). **(b)** Per il sequenziamento tradizionale, il prodotto di PCR è clonato in un vettore standard (TA cloning vector) per poter usare primer standard per il sequenziamento dell'inserto. Il cDNA può essere concatemerizzato e selezionato per dimensione per catturare e sequenziare small RNA multipli. **(c)** Per MPSS, Solexa (Hayward, California) crea una libreria di prodotti di PCR che poi sono trasferiti alle gocce e sequenziate come descritto in Brenner S, *et al* (94). Le frecce rosse indicano il sito di restrizione *Sfa*NI. **(d)** La tecnologia 454 (o Sequencing by Synthesis (SBS)) sequenzia direttamente le molecole di templato amplificate; i templati del 454 sono amplificati su gocce mentre i templati del metodo SBS sono amplificati a gruppi su di una superficie solida. Il metodo di sequenziamento 454 è stato descritto in (95) mentre il metodo di SBS è stato sviluppato quest'anno dal Solexa (96). Le linee rosse indicano gli adattatori di DNA standard usati per entrambe le tecnologie per iniziare il sequenziamento; la sequenza dell'adattatore rosso può essere incorporato negli adattatori di RNA nel punto (a) per evitare una seconda ligazione.

geni di microRNA
→ ricerca dei loop (pri-miRNA e pre-miRNA)
→ riduzione delle regioni genomiche da analizzare (come mascheramento delle regioni ripetute o focalizzazione dell'analisi nelle regioni conservate) in modo da poter ridurre anche il numero di falsi positivi
→ classificazione dei loop (distinti per struttura secondaria e comparazione dei microRNA maturi)
→ identificazione dei microRNA maturi

bersagli di microRNA
→ ricerca della sequenza che si allinea al microRNA maturo
→ riduzione delle regioni genomiche da analizzare (focalizzazione dell'analisi nelle regioni conservate o nelle TU o parte di esse)
→ identificazione delle correlazioni tra regolatore e geni bersaglio

Tabella 1.2: Passaggi più comuni per la ricerca dei microRNA.

(89). Quindi si è pensato di testare un numero cospicuo di programmi di predizione per poter integrare in un unico consenso i risultati migliori, secondo la sensibilità e precisione di ciascun programma.

1.3 GMOD

La mole di dati prodotta da metodiche sfruttate per l'annotazione genica, quali esperimenti di microarray, sequenziamento e predizione genica, sono difficili da visualizzare nell'insieme e a tal proposito è nato il Generic Model Organism Database (GMOD). Tale progetto prevede un programma di lavoro *open source* per lo sviluppo di un pacchetto completo di software atto a creare e amministrare un database di un organismo modello (99). Componenti di questo progetto includono: lo sviluppo di tool per la visualizzazione e la manipolazione dei dati genomici, organizzazione e stoccaggio della letteratura, organizzazione e manipolazione di ontologie biologiche, un robusto schema di database, e un set di procedure operative standard (vedi Tabella 1.3).

Il GMOD è stato fondato dal National Institutes of Health (NIH) e dall'USDA Agricultural Research Service (ARS) (100; 101), con la partecipazione di altri membri che lavorano su progetti riguardanti database di differenti organismi modello quali: WormBase (102; 103), FlyBase

tool	**breve descrizione**
GBrowse	browser per visualizzare dati di annotazione genica
Apollo	browser per manipolare dati di annotazione genica
Sybil	browser per visualizzare dati di comparazione genica
CMap	browser per visualizzare dati di comparazione di mappe
Pathway Tools	predizione di vie metaboliche e loro visualizzazione
Chado	schema del database per salvare tutti i dati
BioMart	interrogazioni avanzate a uno o più database

Tabella 1.3: Tool in sviluppo nel progetto GMOD. Per i tool sfruttati in questa tesi, vedi Paragrafo 2.3.2

(104; 105), Mouse Genome Informatics (106), Gramene (107; 108), Rat Genome Database (109; 110), The Arabidopsis Information Resource (TAIR) (111; 112), EcoCyc (113; 114), e Saccharomyces Genome Database (115; 116).

L'uso di tool in sviluppo presentano complicazioni quali errori negli script e poca documentazione. Tuttavia, la possibilità di contribuire attivamente allo sviluppo dei tool e di interagire dinamicamente con gli sviluppatori, rende il progetto GMOD interessante per molte applicazioni bioinformatiche. Oltretutto i tool forniti dal progetto GMOD stanno diventanto uno standard con il quale salvare i propri dati rendendoli anche scaricabili senza dover aggiungere altra documentazione sul come sono organizzati.

Capitolo 2

Materiali e Metodi

Tutti i programmi son stati eseguiti nell'ambiente di lavoro linux 2.6 (Debian etch) su server Dual-Core AMD64 Opteron(tm) con quadriprocessore a 2GHz e 16GB di RAM. In Tabella 2.1 sono riportate le versioni delle applicazioni principali sfruttate dai vari programmi scaricati e/o implementati.

2.1 Database

I database utilizzati sono stati scaricati e usati in locale per poterli interrogare in tempi ragionevoli. Il materiale scaricato è relativo alle sequenze in formato FASTA di cromosomi, regioni ripetute, microRNA maturo e precursore, ed ai tabulati in formato GFF relativi alle informazioni estrapolate dai tabulati in formato GenBank corrispondenti ad annotazione di

applicazione	**breve descrizione**	**versione**
g++	compilatore C/C++	4.1.2
perl	interprete Perl	5.8.8
java	interprete java	1.5.0_10
mysql	sistema di gestione di basi di dati relazionale (MySQL)	14.12
psql	sistema di gestione di basi di dati relazionale (PostgreSQL)	8.1.9
Bioperl	collezione di moduli Perl	1.5.2_102

Tabella 2.1: Versione delle applicazioni e moduli di base.

geni (vedi Tabella 2.2). La lista completa dei cromosomi batterici analizzati è nell'Appendice B. Per l'interpretazione dei dati ottenuti sono stati sfruttati i database UniprotKB versione 12.3 (123) e Gene Ontology Annotation (GOA) versione del 16 gennaio 2008 (124).
I database che invece sono stati creati per salvare, visualizzare e manipolare i database scaricati saranno descritti nei paragrafi successivi.

2.2 Microarray

Le posizioni dei RDT documentati che sono stati usati per lo studio dei terminatori sono elencati nella Tabella 2.3. Il deposito internazionale Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSMZ) (Braunschweig, Germania) ha fornito *Escherichia coli* K12 MG1655 (DSM498) (125). È stata eseguita la trasformazione del batterio con il vettore pUC19 (126) per conferirgli resistenza all'ampicillina. Gli esperimenti sono stati condotti in terreno minimo a 37°C in presenza di ampicillina (50µg/ml). Un litro di terreno minimo contiene 5g di glucosio, 6 di Na_2HPO_4, 3 di KH_2PO_4, 1 di NH_4Cl, 0.5 di NaCl, 0.12 di $MgSO_4$ e 0.01 di $CaCl_3$. L'antibiotico biciclomicina (Bicyclomycin (BCM)) è stato fornito dalla ditta farmaceutica Astellas Pharmaceutical Co., Ltd. (Osaka, Giappone) ed è stato usato ad una concentrazione di 50µg/ml. RNA batterico è stato estratto con il kit della ditta Qiagen e marcato con fluoroforo Cy5 con il kit della ditta Kreatech.
Per condurre gli esperimenti di microarray è stata utilizzata la nuova piattaforma Combimatrix con CustumArray 12K (127; 128; 129). Le scansioni dei vetrini sono state eseguite con ScanArray Lite (Packard).

2.3 Programmi

2.3.1 Indispensabili

Sono stati usati due programmi di allineamento: BLAST (130), per allineamenti locali, e sim4 (131), per allineamenti più precisi. Sono stati usati i parametri di default ad esclusione del parametro W che, per il programma di sim4, è stato settato a 4 per ottenere degli allineamenti

gruppo	**sito**	**organismo**	**database**	**versione**
CRIBI (*)	(117)	*Photobacterium profundum* SS9	cromosomi, gff	20/06/2005
		Vitis vinifera	cromosomi, gff	03/09/2007
EMBL-EBI	(118)	*Escherichia coli*	GOA	08/06/2007
JGI	(119)	*Populus trichocarpa*	cromosomi, gff	1.1
NCBI	(120)	*Arabidopsis thaliana*	cromosomi, gff	24/04/2007
		Bacillus subtilis	cromosomi, gff	04/12/2007
		Caenorhabditis elegans	cromosomi, gff	16/02/2006
		Enterobacteria lambda tR1	cromosomi, gff	18/10/2007
		Enterobacteria Phage f1	cromosomi, gff	27/04/1993
		Enterobacteria Phage P4	cromosomi, gff	14/11/2006
		Enterobacteria Phage T5	cromosomi, gff	28/11/2007
		Escherichia coli	cromosomi, gff	26/12/2007
		Homo sapiens	cromosomi	6.32
		Mycobacterium tuberculosis H37Rv	cromosomi, gff	01/12/2007
		Neisseria meningitidis MC58	cromosomi, gff	01/12/2007
		Oryza sativa	cromosomi, gff	23/10/2007
		Salmonella typhimurium LT2	cromosomi, gff	01/12/2007
		Thermotoga maritima	cromosomi, gff	04/12/2007
Sanger	(121)	*Arabidopsis thaliana*	microRNA, microgff	10.0
		Caenorhabditis elegans	microRNA, microgff	10.0
		Caenorhabditis briggsae	microRNA, microgff	10.0
		Oryza sativa	microRNA, microgff	10.0
		Populus trichocarpa	microRNA, microgff	10.0
TIGR	(122)	*Arabidopsis thaliana*	repeat	2
		Oryza sativa	repeat	2

Tabella 2.2: Versione dei database scaricati. (*): CRIBI genomics, laboratorio del Prof. Valle; GOA: tabulati formato GOA; cromosomi: sequenze genomiche formato FASTA; gff: tabulati formato GFF dell'annotazione dei geni; microRNA: sequenze dei microRNA formato FASTA; microgff: tabulati formato GFF dell'annotazione dei microRNA; repeat: sequenze delle regioni ripetute formato FASTA

organismo	**operone**	**riferimento**
Bacillus subtilis	*lysC-*	(40)
Enterobacteria lambda tR1	*cro*	(23)
	nin	(46)
Enterobacteria phage F1	*IV*	(55)
Enterobacteria phage P4	*CI-*	(47)
Enterobacteria phage T5	*T5.134*	(51)
Escherichia coli K12	*clpB-*	(42)
	dnaK	(42)
	fim	(52)
	groE	(42)
	lacZ-	(48; 50)
	livJ-	(44)
	livK-	(44)
	ptsI	(53)
	rrnA, B, C, D-, E, G, H	(37; 43; 54)
	tna	(38)
	trp-	(23)
Mycobacterium tuberculosis H37Rv	*rrn*	(39)
Neisseria meningitidis MC58	*sia-*	(49)
Salmonella typhimutium LT2	*his*	(45)
Thermotoga maritima	*his-*	(56)

Tabella 2.3: Regioni in cui è documentato almeno un RDT. Quando il nome dell'operone è seguito da un "-" la regione di interesse risiede nell'elica complementare.

più precisi.

I programmi per la ricerca e l'allineamento di matrici pesate sono, rispettivamente, wconsensus-v5d e patser-v3e.2 (132). I parametri di wconsensus sono -s 1 -c 1 -pf 1000 per considerare nell'analisi entrambi i filamenti e ottere le prime 1000 matrici significative. Le matrici sono state salvate se avevano il logaritmo del p-value minore o uguale a -8 e la lunghezza maggiore o uguale a 10 basi in modo da ottenere molti risultati da filtrare in seguito. I parametri di patser sono -c -ls 3 -ds per ricercare le matrici in entrambi i filamenti con parametri restrittivi (-ls 3).

Con patternScan, un tool implementato nel laboratorio del Prof. Valle dal Dr. Campagna, sono stati ricercati tutti i pattern con struttura NNNNN(-)NNNNN e N-N-N(-)N-N-N dove '-' corrisponde a una base qualsiasi, (-) un numero di basi qualsiasi variabile da 0 a 18, e N le basi

del pattern. I pattern che differenziano solo per il numero di basi centrali (-) costituiscono una famiglia. Sono stati allineati al genoma tutti quei pattern che erano presenti in un numero significativamente diverso rispetto alla propria famiglia.
Sono stati usati due programmi di predizione della struttura secondaria: erpin-4.2.5 (133), per allineamenti globali e veloci e Vienna RNA Package 1.6 (134), per allineamenti locali e precisi. Sono stati usati i parametri di default a parte per erpin a cui sono state fornite le opzioni -2,+2 -nomask per ottenere risultati più permissivi.

2.3.2 GMOD

Sono stati usati i tool GBrowse (Generic genome browser) versione 1.68 (135), BioMart versione 0.5 (138) e lo schema Chado versione 1.6 (99).

GBrowse

Il GBrowse è una combinazione di database e pagine Web interattive per la manipolazione e visualizzazione dell'annotazione su genoma (135). Il tool è in grado di visualizzare dei dati leggendo direttamente il file tabulato in formato GFF oppure interrogando il database MySQL corrispondente. Il file formato GFF presenta nove campi separati da una tabulazione (vedi Tabella 2.4). I database sono stati creati con la terza versione del formato GFF (GFF3 (137)). Il database MySQL corrispondente è costituito da sette tabelle relazionate tra loro come descritto in Figura 2.1.
Esiste un file di configurazione da compilare per definire nome del browser, formato del database da caricare (tabulato o MySQL), nomi, spazi, forme, colori e applicazioni di ogni evidenza e pagina Web del GBrowse. La documentazione di tale file risale alla seconda versione del formato GFF perciò sono di facile configurazione solo le impostazioni di base. In modo analogo, anche la documentazione di ogni script del tool GBrowse è riferita alle impostazioni di base. Le opzioni avanzate hanno una documentazione pressoché inesistente e spesso non sono interamente funzionanti.

Gli script del GMODTools migliorati fanno parte della versione Bioperl riportata in Tabella 2.1, e sono gli stessi per il confronto dei tempi di esecuzione e dei risultati ottenuti (vedi Paragrafo 5.1).

Biomart

BioMart è un sistema integrato orientato all'interrogazione di dati salvati in database (138). Il sistema supporta la possibilità di interrogare in contemporanea uno o più database. I dati provenienti dai database sono adattati al modello di dati di Biomart: un semplice schema di database ottimizzato per le interrogazioni. Il sistema consiste di uno schema di database specifico, un tool di gestione per creare e configurare i database 'mart' specifici e un software per l'accesso ai dati che include un'interfaccia Web.

2.3.3 microRNA

In Tabella 2.5 sono riportati i programmi usati per la ricerca e predizione dei microRNA maturo, precursore e target. La maggior parte dei programmi necessitavano di librerie, moduli o altri programmi che sono stati scaricati a parte. Tra questi, i più conosciuti sono Vienna RNA Package (vedi Paragrago 2.3.1), T-Coffee versione 4.67 (146), RepeatMasker 3.0 (147) e Threaded Blockset Aligner (TBA) (148).

2.3.4 Dati di espressione

L'acquisizione delle immagini dei vetrini di microarray è stata gestita con ScanArray 2.1 (Packard BioChip Technologies). Per la quantificazione delle immagini scansionate, è stato usato Microarray Imager (MI) versione 5.8.1, della Combimatrix (149). Con un tool implementato nel laboratorio del Prof. Valle sono state integrate le intensità derivate dalle singole scansioni in un unico file in formato TIGR Array Viewer (tav) richiesto dal programma Microarray Data Analysis System (MIDAS) versione 2.19 (150) che ha effettuato la normalizzazione dei dati con il metodo Locfit (che utilizza la regressione Localised Weighted Smother Estimator (LOWESS)) (151; 152; 153). Per individuare i geni differenzialmente espressi è stato

##gff-version 3
##sequence-region chr1 1 1497228

seqid	source	type	start	end	score	strand	phase	attributes
chr1	jigsaw	gene	1000	9000	.	+	.	ID=gene00001;Name=EDEN
chr1	patscan	TF_binding_site	1000	1012	10.4	+	.	ID=tfbs00001;Parent=gene00001
chr1	cribi	mRNA	1050	9000	.	+	.	ID=mRNA00001;Parent=gene00001;Name=EDEN.1
chr1	cribi	mRNA	1050	9000	.	+	.	ID=mRNA00002;Parent=gene00001;Name=EDEN.2
chr1	cribi	mRNA	1300	9000	.	+	.	ID=mRNA00003;Parent=gene00001;Name=EDEN.3
chr1	cribi	exon	1300	1500	.	+	.	ID=exon00001;Parent=mRNA00003
chr1	cribi	exon	1050	1500	.	+	.	ID=exon00002;Parent=mRNA00001,mRNA00002
chr1	cribi	exon	3000	3902	.	+	.	ID=exon00003;Parent=mRNA00001,mRNA00003
chr1	cribi	exon	5000	5500	.	+	.	ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
chr1	cribi	exon	7000	9000	.	+	.	ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
chr1	cribi	CDS	3301	3902	.	+	0	ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
chr1	cribi	CDS	5000	5500	.	+	2	ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
chr1	cribi	CDS	7000	7600	.	+	2	ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
chr1	cribi	CDS	3391	3902	.	+	0	ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
chr1	cribi	CDS	5000	5500	.	+	2	ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
chr1	cribi	CDS	7000	7600	.	+	2	ID=cds00004;Parent=mRNA00003;Name=edenprotein.4

Tabella 2.4: Nome dei campi ed esempio di sintassi del formato GFF3. La tabulazione deve essere rispettata in modo che gli script che manipolano questi file li leggano correttamente. Questa è la sintassi corretta per l'annotazione di un gene. Ogni riga è una singola evidenza che potrà essere visulizzata nel browser secondo la configurazione scelta (vedi testo). *seqid* è il nome o codice identificativo della sequenza. *source* è il nome del programma o della società che ha rilasciato quell'evidenza. Per creare un GFF3 standard, il campo *type* deve presentare un termine descritto nell'ontologia delle sequenze (136). Nel campo *attributes* deve essere presente almeno l'attributo ID ma nel caso l'evidenza dipenda da un'altra (come nel caso riportato), deve essere presente anche l'attributo Parent separato dall'ID per mezzo di un punto e virgola.

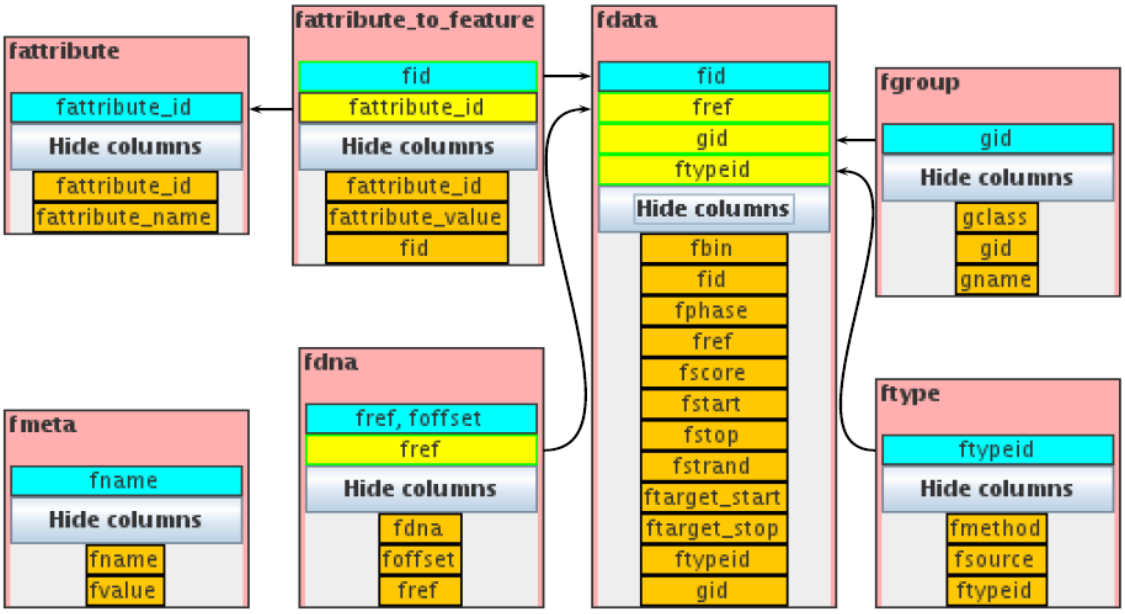


Figura 2.1: Schema del database GBrowse. La tabella principale è *fdata* in cui ogni record corrisponde ad una evidenza caricata dal tabulato formato GFF. La tabella *fdna* presenta nome e stringa di ogni sequenza caricata a blocchi di 2000 basi. La tabella *ftype* riporta ogni tipo (type descritto in Tabella 2.4) di evidenza caricata con la propria sorgente (source descritto in Tabella 2.4) e un codice identificativo del tipo. La tabella *fgroup* è formata da nome e codice identificativo di ogni singolo valore; il nome coincide prevalentemente con il valore dell'attributo ID. La tabella *fattribute* è composta da nome e codice identificativo di ogni attributo presente nel campo attributes del tabulato di partenza (vedi Tabella 2.4). La tabella *fattribute_to_feature* relaziona il codice identificativo di *fdata* con quello dell'attributo e il suo valore; il valore dell'attributo può essere diviso a blocchi definiti dalle virgole che erano presenti nel tabulato di partenza (vedi Tabella 2.4). La tabella *fmeta* presenta dei valori prestabiliti al momento della creazione del database che servono per definire una mappa delle posizioni di ogni record di *fdata* su ogni sequenza caricata in *fdna*.

applicazione		**ricerca**	**concetti sfruttati**
AmiRNA	(139)	m, t	complementarietà e statistica
findMiRNA	(140)	m, p, t	ss, complementarietà, pattern
microHarvester	(141)	m, s, p	ss, complementarietà, regioni ripetute
miRanda	(142)	t	complementarietà
precExtract	(140)	m, s, p	ss, complementarietà
RNAhybrid	(143)	t	ss, complementarietà
srnaloop	(144)	p	ss e complementarietà
TargetScanS	(145)	t	ss, complementarietà, regioni conservate

Tabella 2.5: Applicazioni che predicono i microRNA. I parametri sono stati scelti sulla base dei risultati ottenuti messi a confronto con le posizioni dei microRNA maturi depositati (vedi Paragrafo 1.2). p: pre-miRNA; m: microRNA maturo; s: microRNA star (sequenza complementare al microRNA maturo); t: microRNA bersaglio; ss: struttura secondaria.

usato Significance Analysis of Microarrays (SAM), un programma statistico che implementa un t-test attuato ricorsivamente su tutti i geni (154).

Capitolo 3

RDT

I risultati che seguono si basano sui RDT documentati in letteratura riassunti in Tabella 2.3. Per studiare la regolazione della proteina Rho e identificare nuovi RDT sono state usate due metodiche differenti: una computazionale e una di microarray.

3.1 Caratterizzazione dei RDT

Come descritto nell'introduzione, i procarioti hanno due tipi di terminatori della trascrizione che si distinguono per la sequenza di DNA e per il meccanismo molecolare: i terminatori intriseci e i terminatori Rho dipendenti (RDT). Quando la RNA polimerasi incontra un terminatore intrinseco (precedentemente definito RIT), rilascia spontaneamente il trascritto nascente, ma quando incontra un RDT, il rilascio della molecola di RNA dipende dall'azione di una proteina chiamata Rho (21). I RDT sono quindi coinvolti nel controllo dell'espressione genica agendo come attenuatori, oltre che come terminatori. Sono tipicamente localizzati nel leader (5' non tradotto) dell'mRNA oppure all'interno dell'operone dove possono agire come terminatori tra un gene e l'altro, infine si possono trovare alla fine dell'intero operone. In Figura 3.1 è illustrata l'organizzazione di un RDT.
Ci sono evidenze che indicano che Rho interagirebbe con un'ampia varietà di trascritti nascenti (22), ma la sequenza RDT che dovrebbe determinare l'azione di Rho sull'mRNA non è ben caratterizzata e non esistono programmi informatici per la predizione dei RDT. Un motivo che

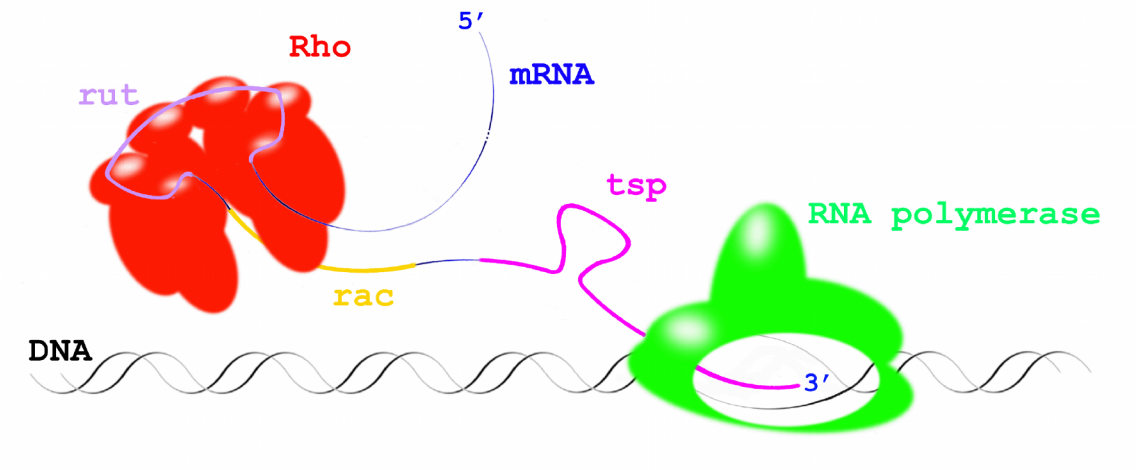


Figura 3.1: Schema di un RDT. Il terminatore Rho dipendente è formato da tre parti distinte: Rho UTilization (rut), Rho ACtivity (rac)[1] e Transcription Stop Point (tsp). rut è il sito di destinazione di Rho, fa parte di un segmento del trascritto nascente al quale Rho si può legare ed è essenziale per l'inizio della terminazione (23; 24; 25). rac è il secondo sito di legame di Rho ed è essenziale per l'attività di traslocazione dell'elicasi (26; 27). tsp è una regione in cui la RNA polimerasi si ferma alcuni momenti durante l'allungamento del trascritto in assenza di Rho (28; 29)), ed è il sito di terminazione della trascrizione.

rende molto difficile la predizione è la mancanza di una esatta sequenza consenso condivisa dai RDT noti elencati nella Tabella 2.3.

Uno degli obiettivi della mia ricerca di dottorato è la caratterizzazione dei RDT per progettare un algoritmo in grado di identificarli. Da quanto è riportato in letteratura e descritto nell'introduzione di questa tesi, un RDT consiste di tre parti distinte, che insieme si possono estendere oltre 150-200 basi di RNA. La prima parte del RDT è il sito Rho UTilization (rut) che corrisponde al segmento del trascritto nascente al quale Rho si lega all'inizio del processo di terminazione (23; 24; 25). La seconda parte del RDT è il secondo sito di attacco di Rho che sembra essere essenziale per l'attività dell'elicasi (26; 27). Dal momento che non aveva ancora un nome, ho deciso di chiamarlo rac, per Rho ACtivity. L'ultima parte del RDT è il Transcription Stop Point (tsp), regione in cui la RNA polimerasi si ferma alcuni momenti durante l'allungamento del trascritto in assenza di Rho (28; 29), e sito di terminazione della trascrizione.

Altre evidenze sperimentali indicano che ciascuno dei sei monomeri di Rho è in grado di inte-

[1] Termine coniato da noi

ragire con sequenze di RNA di 9 basi ricche di citosina (30; 31; 32) che corrispondono al sito rut. Inoltre è riportato che i siti rut sono poveri di guanina, lunghi almeno 40 basi (32; 33) e, alcune volte, presentano una debole struttura secondaria (33). Il sito rac è un altro sito spesso ricco di citosine e povero di guanine, pertanto ho coniato l'acronomo crgpi per indicare Citosine Rich and Guanine Poor Island. Nello specifico, il sito rac è anche ricco di uracile con una composizione CU>>AG (30; 34) e ha una lunghezza di almeno 40 basi (27). La regione tsp può essere distante fino a 150 basi a valle del sito di rut, è lunga al massimo 100 basi e consiste di gruppi di punti di stop (Stop Point (sp)) con intervalli di 20-30 basi (32; 35). Un sp è dove la RNA polimerasi si ferma in assenza di Rho, durante l'elongazione del trascritto. I punti di stop si estendono da 5 a 40 basi (32) e sono preceduti da regioni ricche di citosina e guanina (Strong Island (si)) e/o da almeno 3 basi di timina (36).

In letteratura è inoltre riportato che i RDT sono presenti nel leader (5' non tradotto) degli operoni (37; 38; 39; 40; 41; 42; 43; 44), all'interno degli operoni in siti intragenici (45; 46; 47; 48; 49; 50), o intergenici (51; 52; 53), e alla fine degli operoni (23; 54; 55; 56; 57).

Ad oggi, non esistono programmi di predizione dei RDT e nemmeno un consenso della sequenza riconosciuta dalla proteina Rho, però tutti i RDT documentati presentano una struttura simile (composizione in rut, rac e tsp) che può essere sfruttata per implementare un algoritmo per la loro ricerca e/o predizione. Uno degli obiettivi della mia ricerca è quindi quello di sviluppare un algoritmo per la predizione di RDT e a questo proposito ho condotto delle analisi di composizione nucleotidica e di struttura secondaria sui RDT documentati.

Analizzando la composizione nucleotidica dei RDT documentati ho trovato alcuni ulteriori aspetti che ritengo utili per la caratterizzazione della struttura dei RDT. Un'osservazione riguarda le crgpi che sembrano essere esse stesse gruppi di isole ricche in citosina di lunghezza minima di 6 basi e distanti fra loro al massimo 18 basi. Spesso le posizioni di inizio delle crgpi osservate sono le stesse dei siti di rut (o del box NusA o NusG) documentati, ma alcune volte si estende oltre i limiti descritti in letteratura. Le crgpi sono tipicamente poste tra due regioni ricche di adenina e timina (wi) di circa 19 basi che spesso si trovano anche all'interno delle stesse crgpi.

Dalle osservazioni che ho effettuato su tutte le sequenze rut descritte in letteratura posso con-

cludere che esse sono lunghe almeno 35 basi e che alcune volte esistono delle crgpi a monte di rut anche distanti 50 basi in cui è presente un sito rac (indicando la possibile esistenza di più punti di inizio, vedi Tabella 3.2, operone *lacZ*). Studiando la regione del tsp ho notato che sono presenti forti strutture secondarie a monte delle si (o delle 3 basi di timina) e dei sp. A differenza della letteratura in cui si afferma che i sp di un tsp possono essere intervallati da 20-30 basi, ho osservato che i singoli sp di uno stesso tsp hanno una distanza massima di 7 basi. Inoltre la distanza minima tra la posizione di inizio del RDT e il suo tsp è di 70 basi. Ciò coincide anche con la minima regione coperta da crgpi e rac ed è uguale anche al numero minimo di basi che possono interagire con i sei monomeri di Rho (30; 31; 32). Altro dato osservato è invece che la massima distanza tra la posizione di fine del sito rut e l'inizio del tsp è 218 basi.

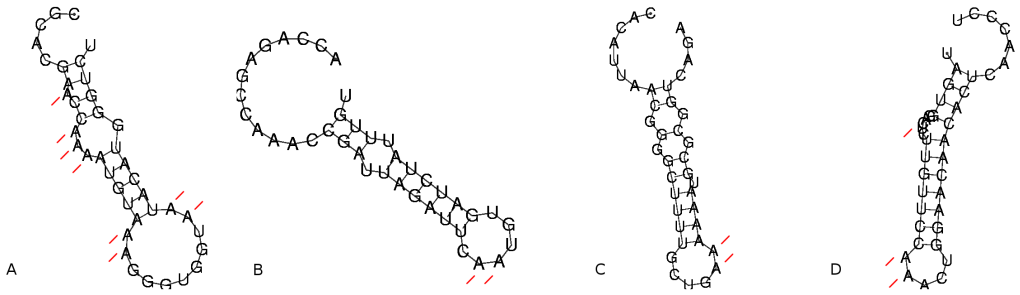


Figura 3.2: Alcune strutture secondarie con almeno due adenine adiacenti nel loop più basso. A: *B. subtilis*, *lysC*-; B: *E. coli*, *ptsI*; C: *E. coli*, *tna*; D: *E. phage* F1, *IV*.

Ho condotto un'analisi approfondita delle strutture secondarie davanti ai tsp di tutti i RDT documentati calcolando tutte le possibili strutture secondarie in un intervallo di 50 basi davanti ad ogni sp predetto. Per ogni intervallo ho eseguito 50 volte RNAfold (vedi Vienna RNA Package, Paragrafo 2.3.1) con una sequenza di 40 basi spostandomi ogni volta di una base a monte dall'inizio di ogni sp. L'analisi ha evidenziato una particolare composizione dei loop a singolo filamento. In particolare i loop sembrano contenere due tipi di composizione (o entrambe): tre adenine divise nello stesso loop o in più loop e almeno due uracili adiacenti in un altro loop, oppure almeno due adenine adiacenti nel loop più basso (vedi Figura 3.2).

3.2 RDTSCAN

Ho disegnato un algoritmo per la ricerca dei RDT a partire dai concetti estrapolati dalla letteratura (vedi Paragrafo 1.1.2) e da nuovi dettagli derivati dallo studio della composizione dei RDT documentati.

L'algoritmo proposto sfrutta una finestra dinamica per ricercare le isole CG (Strong Island (si)) e AT (Weak Island (wi)), il programma erpin (vedi Paragrafo 2.3.1) per identificare le strutture secondarie, l'algoritmo di programmazione dinamica di Mathews *et al.* (157) per generare le strutture secondarie e calcolarne l'energia libera, patser per riconoscere le posizioni della matrice RDT (vedi Paragrafo 3.3), e le soglie elencate in Tabella 3.1.

3.2.1 Metodo della finestra dinamica

Esistono varie tecniche per identificare le si. Il più diffuso è l'uso della finestra mobile che scorre sul genoma contando quante citosine e guanine sono presenti in quella finestra (158; 159; 160). Questa tecnica ha una bassa risoluzione e non riconosce piccoli cambiamenti nel contenuto di CG. Inoltre, la distribuzione del contenuto in CG dipende dalla larghezza della finestra. Recentemente, è stato proposto un nuovo algoritmo per l'identificazione delle si senza finestra: il metodo cumulativo (161). Questo sistema ha un'alta risoluzione ma, come la tecnica precedente, identifica uguali risultati per ciascuna elica.
La metodica che propongo invece prevede una finestra mobile con una larghezza dinamica che permette di avere risultati differenti per le due eliche che ha identificato più inizi di siti di rut. Questa procedura è in accordo con l'orientamento dei rut descritta in Hinde *et al* (162). Con il metodo della finestra dinamica ho identificato le isole crgpi, si, wi, ricche di timina (Timine Island (ti)) e ricche di timina e citosina (Pyrimidine Island (yi) o rac).
A titolo di esempio, per calcolare la larghezza di una crgpi sono state considerate la percentuale minima di citosina e la percentuale massima di guanina (vedi Tabella 3.1, punti 1 e 2) che ho calcolato con l'equazione (3.1):

$$percentuale = \frac{\sum_{n=x}^{y} N_n}{y - x + 1} \tag{3.1}$$

punto	**breve descrizione parametro**	**valore**
1	percentuale minima di C in crgpi	30
2	percentuale massima di G in crgpi	20
3	lunghezza minima di crgpi	6
4	lunghezza massima tra le crgpi	21
5	lunghezza minima dei rut	35
6	lunghezza massima tra crgpi e wi	19
7	percentuale minima di wi per passare da crgpi a rut	30
8	percentuale minima di ci per passare da crgpi a rut	20
9	percentuale minima di wi	80
10	lunghezza minima di wi	2
11	lunghezza minima di wi per passare a sp	3
12	lunghezza massima tra wi	7
13	lunghezza massima tra si (o ti) e wi	25
14	percentuale minima di si (o ti)	70
15	lunghezza minima di si (o ti)	5
16	dalla posizione di inizio di wi, regione dove cercare le ss	-45, +2
17	percentuale minima di sp	60
18	lunghezza minima di tsp	13
19	massimo e-value dei risultati di erpin	2.9
20	percentuale minima di rac	61
21	lunghezza minima di rac	15
22	lunghezza massima tra fine rut e inizio tsp	218
23	lunghezza massima tra inizio rut e inizio dell'ultimo sp	70
24	percentuale minima di crgpi e rac a monte di rut	40
25	oltre le 50 basi a monte di rut il punto 24 vale	80
26	lunghezza massima tra fine crgpi e inizio rut	50
27	numero minimo di basi (crgpi+rac+rut) che legano Rho	70
28	percentuale minima di RDT	39
29	punteggio minimo (ln(p-value)) per passare a RDT	-26.3

Tabella 3.1: Parametri di RDTSCAN. I valori segnati rappresentano le soglie usate per genomi con una percentuale di adenina compresa tra 24% e 30% (vedi testo).

dove x e y sono le posizioni di inizio e fine della regione della sequenza genomica dove calcolare il numero di N (nel caso presentato N è citosina oppure guanina) e la sua percentuale relativa. Nella Figura 3.3 è riportato il listato dello pseudocodice dell'algoritmo in cui viene calcolata lunghezza e percentuale delle isole N.

3.2.2 Strutture secondarie

Per calcolare la struttura secondaria e l'energia libera corrispondente ho usato due metodiche. Il programma proposto richiama prima erpin (vedi Paragrafo 2.3.1) per scremare le regioni che non presentano una struttura secondaria e in un secondo momento, calcola struttura secondaria ed energia libera usando l'algoritmo di Mathews *et al* (157).

3.2.3 Soglie

Per il calcolo delle crgpi, si, rac, ti e wi sono sfruttate delle soglie per analizzare solo determinate isole (vedi Tabella 3.1, punti 1-3, 9-10, 14-15, 20-21). Per il calcolo di rut, tsp e RDT ne vengono applicate altre in modo da ottenere solo RDT putativi con una composizione che rispecchi le informazioni ricavate dalla letteratura (vedi Paragrafo 1.1.2) e quelle ottenute negli studi effettuati (vedi Paragrafo 3.1). Le soglie presentate in Tabella 3.1 sono i valori dei parametri del programma che identificano tutti e 32 i RDT di Tabella 3.2.
L'algoritmo presentato è basato sulla composizione nucleotidica perciò esistono tre gruppi di parametri: uno che presenta meno del 24% di adenina, uno che presenta meno del 30% di adenina e uno che presenta più del 30% di adenina. Usando questo accorgimento si ottengono risultati uniformi per genomi con composizione differente.

3.2.4 Implementazione

Il programma RDTSCAN l'ho implementato in C++. Funziona su linux e si lancia da linea di comando. Ogni parametro della Tabella 3.1 può essere modificato al momento dell'esecuzione del programma. Legge un file formato FASTA, crea un file in formato GFF3 e restituisce nella finestra in cui è stato eseguito alcune informazioni riguardo RDT osservati, aspettati,

```
Set NucleotidePosition to 0
Set NucleotideCount to 0
FOR each nucleotide of sequence
  IF nucleotide is N THEN
    IF Start is null THEN
      Set Start to NucleotidePosition
    END IF
    INCREMENT NucleotideCount
    Set End to NucleotidePosition
    Update Percentage and Length
    IF Percentage is lower Threshold and Length exceed Threshold THEN
      IF N island is crgpi THEN
        IF GuanineCount is lower Threshold or
        GuaninePercentage is lower Threshold and
        CitosinePercentage exceed sum of Threshold and GuaninePercentage THEN
          Save OldStart, OldEnd, OldPercentage and OldLength of N island
      ELSE
        Save OldStart, OldEnd, OldPercentage and OldLength of N island
      END IF
      Set Start and End to NucleotidePosition
      Set NucleotideCount to 1
      Update Percentage and Length
    ELSE
      IF N island is crgpi THEN
        IF GuanineCount exceed Threshold or
        GuaninePercentage exceed Threshold and
        CitosinePercentage is lower sum of Threshold and GuaninePercentage THEN
          IF Percentage and Length exceed Thresholds THEN
            Save OldStart, OldEnd, OldPercentage and OldLength of N island
          END IF
          Set Start and End to NucleotidePosition
          Set NucleotideCount to 1
          Update Percentage and Length
        ELSE
          Set OldEnd, OldPercentage and OldLength correspondingly to End, Percentage and Length
        END IF
      ELSE
        Set OldEnd, OldPercentage and OldLength correspondingly to End, Percentage and Length
      END IF
    END IF
  END IF
END FOR
```

Figura 3.3: Listato dello pseudocodice dell'algoritmo implementato in RDTSCAN per il calcolo del contenuto nucleotidico. *Start*, *End*, *Percentage* e *Length* sono posizioni sulla sequenza analizzata, percentuale e lunghezza dell'isola N.

operon	**type**	**literature**	**RDTSCAN**
lysC-	RDTLE	about into 2909815-2910065	2909719-2909950 *tsp*3(2909752,2909792)
cro	RDTEN	RDT 38249-38368	38046-38344
nin	RDTRA	about 41732-41940	41580-41878 *tsp*2(41849)
nin	RDTRA	about 42020-42231	42046-42282 *tsp*2(42193)
nin	RDTRA	about 42630-42825	42644-42945
IV	RDTEN	about 5500	5544-5907 [5326-5703]
CI-	RDTRA	*rut* 8208-8296	8095-8386
ltf-	RDTER	*into* 84346-86822	84411-84768 [84263-84870]
clpB-	RDTLE	box 2732227-2732341	2732263-2732388
dnaK	RDTLE	box 12106-12118	12094-12288 *tsp*2(12371)
fim	RDTER	into 4540656-4541686	4540828-4541159 *tsp*3(4541081,4541131)
groE	RDTLE	box 4368622-4368632	4368616-4368772
lacZ-	RDTRA	RDT 365099-365249	365073-365298
lacZ-	RDTRA	RDT 365329-365439	365132-365544
livJ-	RDTLE	about into 3597681-3597786	3597736-3597875
livK-	RDTLE	about into 3595583-3595754	3595557-3595716
ptsI	RDTER	about into 2531786-2532088	2531804-2532077
rrnA	RDTLE	box 4032875-4032891	4032875-4033252 *tsp*3(4033152,4033193)(*)
rrnB	RDTLE	about into 4164063-4164682	4163994-4164267 [4163994-4164694]
rrnC	RDTLE	about into 3939258-3939831	3939287-3939529 *tsp*2(3939470)(+)
rrnE	RDTLE	about into 4205441-4206170	4205604-4205850 [4205866-4206082]
rrnH	RDTLE	about into 223596-223771	223452-223683 *tsp*3(223554,223653)
rrnD-	RDTLE	operon 3421690-3426784	3426622-3426966 [3426622-3427085]
rrnG-	RDTLE	about into 2729179-2729349	2729017-2729358 [2729017-2729497]
rrnG-	RDTEN	about RDT 2723912-2724085	2723914-2724073
tna	RDTLE	about 3886554	3886516-3886659 *tsp*2(388616)
trp-	RDTEN	RDT 1314129-1314441	1314230-1314437 *tsp*2(1314274)(x)
rrn	RDTLE	operon 1471844-1477011	1471486-1471759 *tsp*2(1471649)
sia-	RDTRA	about 77013	76897-77222
sia-	RDTRA	about 77090	77012-77267
hisG	RDTRA	upstream 2150064	2149854-2150077
his-	RDTEN	upstream 1052194	1051886-1052188 *tsp*2(1052071)

Tabella 3.2: Confronto delle posizioni dei RDT documentati e predetti. Quando il nome dell'operone è seguito da un '-' la regione di interesse risiede nell'elica complementare; *tsp*N(M): N, numero di *tsp* nei RDT putativi e M, posizione di fine alternativi del RDT putativo; [inizio-fine]: posizioni alternative dei RDT putativi o di gruppi di RDT putativi. (*): [4032875-4033566], (+): [3939287-3939991], (x): [1314112-1314378].

significatività e tempo di esecuzione. Può essere lanciato anche fornendogli un file in cui ogni linea ha i parametri necessari per l'analisi di una singola sequenza alla volta.
Il file GFF3 che produce può presentare le posizioni corrispondenti a: crgpi, si, ti, wi (che presentano nel campo score (vedi Tabella 2.4) la percentuale della loro composizione), rut (con score uguale alla percentuale dei crgpi che lo compongono), tsp (con score uguale alla percentuale dei sp che lo compongono), strutture secondarie (con score uguale al valore fornito da erpin), matrici (con score uguale al logaritmo del p-value) e RDT (con score uguale al logaritmo del p-value della matrice che si allinea in quella stessa posizione).

3.2.5 Attendibilità del programma

Non sono disponibili un numero sufficiente di RDT documentati per calcolare sensibilità e specificità perciò ho stimato solo la significatività e parte dei falsi positivi. Per comprendere se i RDT putativi possono essere significativi, il programma RDTSCAN calcola il numero di attesi come il numero medio degli osservati in 50 genomi random con lunghezza e composizione identica a quella analizzata e il numero degli osservati è significativo quando:

$$valore\ del\ test = \frac{(e-o)^2}{e} > 5 \tag{3.2}$$

dove e è il numero di attesi e o è il numero di RDT osservati nel genoma analizzato. I RDT presentano un orientamento e perciò tutti quei RDT predetti nella direzione inversa a quella dei geni documentati in quella regione sono stati considerati falsi positivi certi.

3.2.6 Analisi dei risultati

In Tabella C.2 sono riportati i risultati dell'analisi dei RDT su 50 cromosomi tra batteri e fagi. Nell'82% dei genomi analizzati, il valore del test è maggiore di 5 e i RDT putativi osservati superano quelli aspettati (media degli osservati in 50 genomi random con lunghezza e composizione identica a quella analizzata). Nel 10%, i RDT putativi osservati sono minori di quelli aspettati. Si tratta di cromosomi di tre fagi e del genoma di *R. sphaeroidesone* che presenta un contenuto in GC del 69%. Nel 8%, il valore del test è minore di 5. Ancora una

volta si tratta di piccoli cromosomi tra cui tre plasmidi e un cromosoma con un contenuto in GC minore del 30%. Complessivamente, nel 18% dei cromosomi analizzati il valore del test non è significativo. Si tratta sempre di piccoli genomi di plasmidi e fagi. Probabilmente la metodica basata sulla composizione nucleotidica non è ancora affidabile per piccoli genomi ed insieme ai limiti di erpin con piccole sequenze è stato amplificato l'errore.

Sicuramente parte dei falsi positivi sono quei RDT putativi che sono stati predetti nel filamento complementare a quello del gene che risiede in quella regione. In Tabella C.3 e Tabella C.5 sono riportate le percentuali dei RDT putativi senza i falsi positivi certi. Osservando le tabelle, tutti i RDT predetti sembrano avere più probabilità di essere un RDT nel leader o all'inttragenico rispetto ad essere un RDT intergenico (all'interno di un operone) od alla fine di un gene (o un operone). Inoltre sembrano esserci più RDT intragenici che RDT nel leader. I RDT putativi osservati hanno meno del 49% di falsi positivi certi ma nel 8% dei cromosomi analizzati hanno più del 62%. Questi sono due fagi e due plasmidi a corta sequenza che non hanno significatività. Questi dati confermano che per piccoli genomi di plasmidi e fagi il programma proposto non è ancora stato settato in modo corretto. Il programma RDTSCAN riconosce dei RDT putativi in tutte le regioni dei RDT documentati (vedi Tabella 3.2). In 18/32 RDT dimostrati, sono identificati almeno la posizione di inizio. In 15/18 dei corrispondenti RDT putativi la posizione di inizio differisce di meno di 20 basi e in 3/18 sempre la posizione di inizio è a monte di quella documentata in letteratura. Di 4/32 RDT confermati sono conosciute solamente le regioni in cui esiste un RDT ma non le posizioni esatte e il programma predice 4/4 RDT in quelle zone. Infine, in 10/32 RDT documentati sono nella regione del leader di operoni di cui sono conosciuti i cDNA. RDTSCAN predice 10/10 RDT putativi all'interno di questi leader ma in 3/10 la posizione di inizio è a monte del cDNA di 200 basi.

In Tabella C.1 sono riportati i tempi di esecuzione del programma. I tempi reali e dell'utente descrivono una funzione lineare in relazione alla lunghezza del genoma. Quei genomi che hanno bassa significatività hanno un tempo di esecuzione che non rientra in questa relazione perché molte funzioni non vengono eseguite se la prima scansione del genoma ed erpin trovano poche regioni da analizzare. Nel complesso, le prestazioni del programma dipendono dalla lunghezza del genoma.

3.3 Matrici pesate

Inizialmente si pensava di cercare una sequenza consenso presente in tutti i RDT ed avevo preso in considerazione i box di NusA o NusG, ma solo alcuni RDT documentati hanno questo elemento. Usando le posizioni dei RDT conosciuti, ho usato wconsensus per produrre delle matrici pesate. Le matrici trovate avevano posizione casuale all'interno dei RDT. Quando ho implementato RDTSCAN, ho sfruttato le posizioni putative dei RDT predetti per calcolare le matrici. Tra le matrici trovate una copre interamente tutta la regione dei RDT documentati (di seguito chiamata matriceRDT).
Calcolando i RDT putativi solamente con l'allineamento della matriceRDT alla sequenza batterica, trovo un numero minore di RDT rispetto a quelli trovati dal programma implementato ed alcune regioni non presentano la composizione strutturale documentata (vedi Figura 3.1). Perciò ho unito le due metodiche per diminuire i falsi positivi.
In Tabella C.4 e Tabella C.6 sono riportate le analisi che ho effettuato sui 50 genomi analizzati solo con la matriceRDT. Come per le predizioni di RDTSCAN, sembrano più probabili i RDT presenti nel leader e all'interno di gene od operone con prevalenza delle regioni leader. Questo però potrebbe dipendere anche dal fatto che quasi la metà dei RDT documentati (15/32) sono presenti nel leader. Il numero di falsi positivi certi è intorno al 30% e ci sono solo due cromosomi che si distinguono per un numero relativamente basso di RDT putativi: i due ceppi di *Mycobacterium tuberculosis*. I parametri che ho usato per allineare la matriceRDT sono stati permissivi ma in una sequenza nucleotidica in cui la composizione in GC è del 66% (come nei *Mycobacterium tuberculosis*), è difficile avere ampie zone ricche di adenina e timina (come i RDT).

3.4 Analisi dei risultati

Allineando la matriceRDT al genoma di *E. coli* e sfruttando parametri specifici riguardo la sua composizione nucleotidica, ho individuato 60 geni che potrebbero essere regolati da un RDT

Tabella C.7. I risultati dei RDT predetti li ho relazionati ai geni e ai codici corrispondenti alle annotazioni di Clusters of Orthologous Groups of proteins (COGs).Il 50% dei 60 geni individuati è coinvolto in trascrizione (Clusters of Orthologous Groups of proteins (COGs) K), traduzione e biogenesi di RNA ribosomale (J), trasporto e metabolismo degli aminoacidi (E), oppure sono chaperonine (O) o geni di RNA ribosomale (vedi Tabella C.8). Il rimanente dei geni non presenta un codice COG e/o sono poco caratterizzati. Ho ottentuto dei risultati differenti dalla stessa analisi ma con i RDT predetti da RDTSCAN. Infatti il 50% dei geni è coinvolto in parte ancora in traduazione e biogenesi di RNA ribosomale (J) e trasporto e metabolismo degli aminoacidi (E) ma anche in produzione e conversione di energia (C), trasporto e metabolismo di carboidrati (G), coenzimi (H) e ioni inorganici (P). Una buona percentuale è poco caratterizzata e il rimanente fa parte di altre categorie funzionali (vedi Tabella C.8). Per avere un quadro generale ho anche calcolato le frequenze di ogni categoria funzionale presente in ognuna delle due predizioni sul totale dei geni di ogni categoria funzionale (vedi Tabella C.10). Ma non ho ottenuto risultati rilevanti.

3.5 Microarray

Per comprendere se le predizioni dei RDT rispecchiano in qualche modo la vera regolazione della proteina Rho, ho progettato degli esperimenti di microarray. Ho disegnato due oligo per ogni regione intorno alle posizioni di inzio e fine di tutti i geni documentati nel file GenBank del NCBI di *Escherichia coli* K12 e di tutti i RDT predetti dal programma RDTSCAN.
I geni regolati dai RDT documentati in *E. coli* sono coinvolti nel ripiegamento strutturale delle proteine (O), nella traduzione (J), nel trasporto e metabolismo degli aminoacidi (E) e motilità batterica (N, vedi Tabella 2.3). Questi geni vengono tutti sottoregolati a parte quelli coinvolti nella sintesi degli aminoacidi che presentano un terminatore alla fine dell'operone.
Nelle condizioni di coltura ho scelto il terreno minimo per avere dei controlli di regolazione negativa e positiva per operoni come tnaCAB e trpLEDCBA coinvolti nel metabolismo degli aminoacidi. Per comprendere, invece, se la maggior parte dei geni coinvolti nel ripiegamento strutturale delle proteine sono sottoregolati da Rho, ho condotto degli esperimenti in cui ho

aggiunto un antibiotico che inibisce la proteina Rho (Biciclomicina (BCM)).
In letteratura sono stati eseguiti esperimenti su ceppi di *E. coli* sensibili alla BCM e per questi sono documentati tempi e concentrazioni per ottenere cambiamenti morfologici e nel pool proteico (163; 164; 165; 166). Non esistono dati in proposito sul ceppo usato. Sono stati perciò eseguiti dei controlli sulla crescita di *E. coli* K12 MG1655 in presenza di BCM a 50µg/ml per conoscere i tempi di risposta del batterio con tale concentrazione (vedi Figura 3.4). In

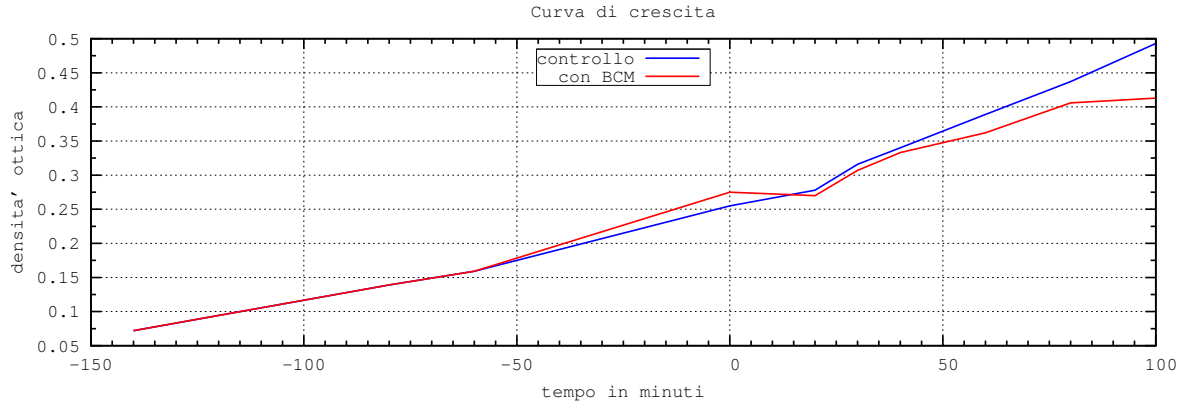


Figura 3.4: Curva di crescita di *E. coli* K12 in presenza di BCM. La BCM è stata aggiunta al tempo zero.

letteratura dati sperimentali eseguiti su ceppi batterici sensibili alla BCM dimostrano che l'antibiotico è nel citoplasma dopo 30 minuti dall'aggiunta in coltura liquida del farmaco. In base a questi risultati e alle analisi di crescita eseguite ho eseguito i primi esperimenti di microarray. Si tratta di un esperimento di *time course* in cui ho prelevato tre campioni: uno al tempo zero, uno al tempo 25 minuti e uno al tempo 35 minuti dall'aggiunta della BCM nel terreno di coltura. L'esperimento è iniziato nel momento in cui la coltura presentava un OD di 0.25 in 550nm di assorbanza (vedi Figura 3.4). Dopo estrazione e marcatura del RNA batterico (vedi Paragrafo 2.2), ho eseguito gli esperimenti di ibridazione in tre repliche per ogni campione ed ogni replica l'ho scansionata quattro volte per integrare in modo più affidabile i dati di intensità.

3.5.1 Analisi dei risultati

Ho considerato un gene sovra o sotto espresso se gli oligo disegnati a valle del gene, a monte o a valle della posizione di un possibile RDT intragenico o a monte della posizione di un RDT predetto alla fine del gene erano rispettivamente sovra o sotto espressi. Con questo metodo ho ricavato quali geni hanno cambiato il livello di espressione in seguito al trattamento con BCM. Ho osservato che nei dati relativi agli esperimenti al tempo 25 minuti, il 50% dei geni differenzialmente espressi sono coivolti in trasduzione e biogenesi di RNA ribosomale (J) e trasporto e metabolismo degli aminoacidi (E), trasporto e metabolismo di carboidrati (G) e meccanismi di trasduzione del segnale (T). Mentre a 35 minuti gli oligo differenzialmente espressi sono molto pochi rispetto sia al tempo zero che a 25 minuti. Questo sembra ipotizzare che i tempi utilizzati sono tardivi rispetto al momento che si voleva osservare (vedi Tabella C.8, righe T25 e T35).

Nel caso in cui un RDT predetto nel leader di un gene A aveva gli oligo espressi a monte, in assenza di sovrapposizione di un gene B nella stessa direzione e in assenza di espressione degli oligo del gene A a valle del RDT, ho considerato che il RDT predetto è effettivamente funzionale ed attivo nelle condizioni testate. Con questo metodo ho trovato che, al tempo zero, i RDT nel leader dei controlli negativi (i geni *clpB*, *dnaK*, *groS*, *tnaC*, *lacZ*) sono funzionali e i geni non sono trascritti, mentre i RDT nel leader dei controlli positivi (i geni *rrsG*, *livJ*, *rrsA*, *rrsB*, *rrsE*, *rrsD*, *trp*) non sono funzionanti e i geni sono trascritti. Per i tempi 25 e 35 minuti le intensità degli oligo erano statisticamente significative solo per alcune regioni che coprivano i RDT nel leader dei geni di controllo positivo e quattro dei controlli negativi (i geni *clpB*, *dnaK*, *groS*, *tnaC*) diventati positivi. Inoltre, al tempo zero, gli oligo relativi al 3' del RDT dell'operone del triptofano (posizionato alla fine dell'operone) risultano spenti mentre, dopo l'aggiunta della BCM, la trascrizione è andata oltre la fine del RDT documentato. Questa è un'ulteriore conferma del metodo per individuare se un oligo è acceso o spento e di conseguenza se un gene è stato trascritto oppure no.

Capitolo 4

microRNA

La ricerca dei geni di microRNA è stata affrontata su due fronti computazionali: predizione dei geni di microRNA e identificazione dei siti bersaglio dei microRNA.

4.1 Approccio computazionale

Come ho detto nell'introduzione, ciò che manca è un protocollo che tenga conto di più metodiche contemporaneamente attraverso un modello statistico integrato. Ho scelto allora di sfruttare più programmi di identificazione e/o predizione di microRNA. Il protocollo che presento prevede l'uso di tutti i risultati derivati dai diversi sistemi per ottenere, da una parte un consenso di predizioni di precursori e dall'altrai siti bersaglio dei microRNA (vedi Figura 4.1). I risultati di ogni programma sono valutati in base alla loro sensibilità e precisione e sono integrati in un'unica predizione.
La scelta dei programmi da utilizzare è dipesa dalla disponibilità del programma *stand alone*, dal suo effettivo funzionamento e dai tempi di esecuzione ragionevoli. Inoltre, i programmi usati (vedi Tabella 2.5) sfruttano metodiche di predizione diverse o hanno input diversi: genoma, regioni intergeniche, regioni trascritte, Untranslated Region (UTR), o altro materiale estrapolato dalla sequenza genomica. Buona parte di questi software sono stati implementati e/o testati su organismi animali, perciò ho dovuto adattarli alle piante settando i parametri diversamente.

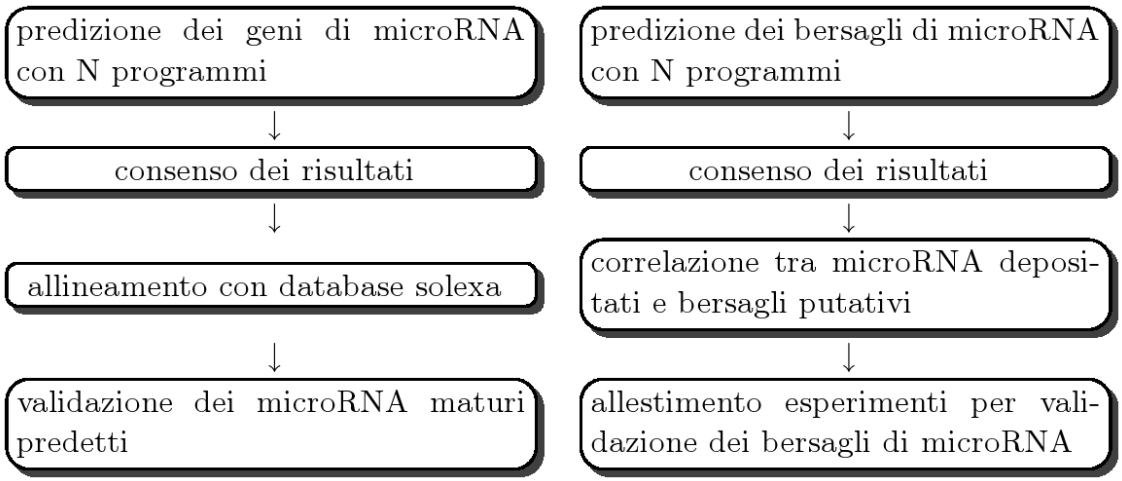


Figura 4.1: Protocollo integrato proposto. L'approccio proposto prevede l'uso di più programmi di predizione da cui ottenere un consenso di pregioni predittive. Entrambe le metodiche sono progettate per fare da supporto ad approcci sperimentali quali il sequenzaimento con metodo SBS (96) di un pool di small RNA e l'allestimento di esperimenti mirati alla validazione dei siti bersaglio dei microRNA.

4.1.1 Calibrazione e F-measure

Ho calibrato ogni programma in modo da poter riconoscere la maggior parte dei microRNA depositati (vedi Tabella 2.2). La calibrazione è avvenuta testando i programmi con tre organismi che presentano più di un centinaio di microRNA depositati ciascuno: *Arabidopsis thaliana*, *Oryza sativa* e *Populus trichocarpa*. Ho eseguito ogni programma più volte per individuare i parametri più convenienti a calcolare i valori migliori di significatività. Quando il programma richiedeva il database dei microRNA depositati, ho utilizzato un pool di microRNA privato di quelli relativi all'organismo analizzato, che ho successivamente sfruttato per il calcolo di sensibilità e specificità usate poi per il calcolo di F-measure (167).

Per calcolare la significatività dei programmi ho usato i valori di sensibilità e del tasso di precisione (Positive Predictive Value (ppv)). Ho scelto di usare il ppv al posto della specificità perchè il calcolo dei veri negativi è pressoché empirico, basato sull'attendibilità dei microRNA maturi depositati e sul calcolo dei loro pre-miRNA, che nelle piante può variare da una lunghezza di 60 a 350 basi (vedi Figura 4.2).

Il punteggio che ho scelto come misura di confronto tra i risultati dei vari programmi è la media geometrica di sensibilità e tasso di precisione (o media armonica pesata) ovvero F-measure.

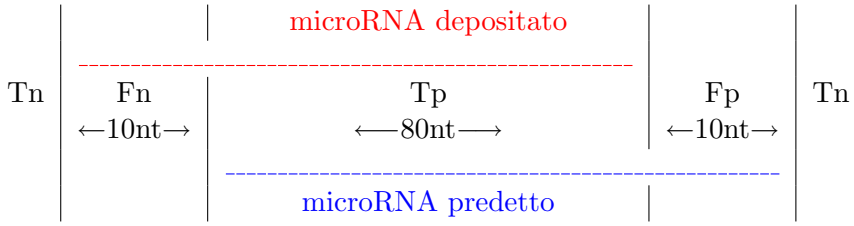


Figura 4.2: Calcolo di veri positivi, falsi positivi e falsi negativi. Tn, veri negativi; Fn, falsi negativi; Tp, veri positivi; Fp, falsi positivi. Tn, Fn, Tp, Fp sono stati calcolati sommando il numero di basi del microRNA predetto secondo la situazione. Definito che i Tn sono calcolati nelle regioni a monte e a valle del microRNA depositato in un intervallo di 100 basi, nell'esempio il calcolo delle basi è il seguente: Tn=180, Fn=10, Tp=80, Fp=10. Il valore della sensibilità è stato mediato con la sensibilità calcolata dal rapporto del numero di microRNA depositati trovati sul numero di microRNA depositati totale.

L'ho calcolata con l'equazione 4.1:

$$F_\alpha = \frac{(1+\alpha)\cdot \mathrm{ppv}\cdot \mathrm{sb}}{(\alpha\cdot \mathrm{ppv})\cdot \mathrm{sb}} \tag{4.1}$$

dove *ppv* sta per tasso di precisione, *sb* per sensibilità e α è un coefficiente uguale a 0.5, cioè il valore di *sb* ha un peso maggiore rispetto a *ppv* (vedi Figura 4.2). Ogni programma, alla fine della calibrazione, ha ottenuto un punteggio (un F-measure) calcolato dalla media ottenuta tra i tre punteggi delle tre prove relative ad *Arabidopsis thaliana*, *Oryza sativa* e *Populus trichocarpa*.

4.1.2 wEvidence

wEvidence è uno script che ho implementato in perl. È pensato per estrapolare da tabulati GFF3 (pesati secondo un F-measure) delle zone in cui sembra più probabile che ci sia un evidenza. Nel caso in cui i tabulati GFF3 siano dei risultati di più programmi che forniscono le posizioni di microRNA putativi su di un determinato cromosoma, wEvidence non fa altro che scorrere tale cromosoma e dare un punteggio ad ogni nucleotide uguale alla somma del F-measure relativo ad ogni programma che predice un ipotetico microRNA in quella determinata base. In questo modo si ottengono delle regioni più probabili che poi vengono filtrate dal rumore di fondo secondo due soglie. Per determinare se una regione possa essere un consenso deve avere un punteggio nucleotidico sempre maggiore di una determinata soglia. Ad esempio,

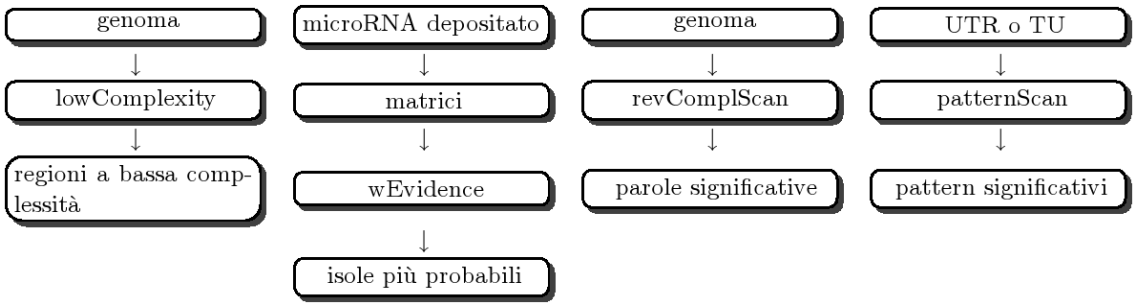


Figura 4.3: Metodiche proposte per la predizione di microRNA. I protocolli sviluppati sono stati usati per la predizione di geni e bersagli di microRNA. I risultati di ogni filtro sono stati 'sommati' agli altri per mezzo di wEvidence.

se il valore della soglia è settato con il più alto tra gli F-measure dei programmi utilizzati, si ha la garanzia che i risultati del programma considerato *migliore* ci sono tutti mentre degli altri programmi si hanno solo quelle evidenze ritrovate da più metodi. Inoltre, la regione da considerare consenso deve presentare l'area creata dai punteggi nucleotidi per la lunghezza della regione maggiore di un determinato valore. Quest'ultima soglia garantisce che regioni minori di una certa lunghezza non rientrino tra i risultati finali o risultino una minima parte. Con questo script ho calcolato le *isole* più probabili dei precursori (Precursor Island (PI)) e dei bersagli (Target Island (TI)) sulla base delle posizioni di matrici pesate (vedi Paragrafo 4.2) e i consensi finali dei risultati dei programmi di predizione dei microRNA.

4.2 Metodiche di predizione

Gli approcci usati dai programmi disponibili non sfruttano matrici pesate calcolate sui microRNA maturi depositati. In letteratura esiste un solo protocollo che si basa su pattern determinati a partire dai microRNA maturi depositati (168) ma non è disponibile *stand alone*. La metodica proposta in questa sede prevede una serie di filtri indipendenti che hanno come input genoma, UTR e/o TU e l'uso di matrici pesate, pattern e conteggio di parole (vedi Figura 4.3). Di seguito sono presentati i programmi che ho implementato e usato nella metodica proposta.

4.2.1 lowComplexity

lowComplexity è uno script che ho implementato in perl. Esso crea una tabulato GFF3 con tutte le posizioni delle regioni a bassa complessità o tutte le posizioni delle regioni inverse. Al momento dell'esecuzione possono essere modificati i parametri di lunghezza, percentuale e tipo di nucleotidi che formano le regioni a bassa complessità. Secondo le analisi che ho eseguito sul database di microRNA maturo e di precursori depositati al Sanger Center, le regioni dei geni di microRNA sono ricche di adenina ed uracile con un contenuto massimo del 65%. Questi dati sono stati utilizzati come filtro per selezionare i risultati dei programmi descritti di seguito.

4.2.2 Matrici pesate

A partire dal database di microRNA depositati ho creato delle matrici di peso con wconsensus (vedi Paragrafo 2.3.1). Una volta allineate al genoma con patser (vedi Paragrafo 2.3.1), ho scartato le zone a bassa complesità calcolate da lowComplexity. Con wEvidence ho calcolato le isole più probabili di precursori (Precursor Island (PI)) e bersagli (Target Island (TI)) dei microRNA. Per il calcolo delle TI, ho calcolato gli inversi complementari delle matrici pesate prima di allinearle al genoma.

4.2.3 revComplScan

revComplScan è un tool implementato in C++ che conta le parole e poi le relaziona al proprio inverso complementare. L'algoritmo di revComplScanP l'ho progettato per ricavare i precursori di microRNA mentre quello di revComplScanT l'ho disegnato per identificare i siti bersaglio dei microRNA.

revComplScanP conta le m coppie di parola e proprio inverso complementare di lunghezza l presenti in una finestra di w basi. La finestra viene spostata sul genoma di $w/3$ basi alla volta. Il programma è eseguibile da linea di comando e i parametri m, l e w sono modificabili. L'output è direttamente un tabulato GFF3 con le posizioni delle finestre e delle coppie di parola e proprio inverso complementare.

È in sviluppo la parte che terrà conto della struttura secondaria delle regioni che presentano più coppie di parola e proprio inverso complementare.

revComplScanT conta le parole di lunghezza l, determina la complessità linguistica di ogni tipo di parola e le divide in famiglie, secondo la complessità linguistica. Dopodiché calcola le parole presenti un numero di volte significativamente differente dalla famiglia di appartenenza e solo per le parole che presentano anche il proprio inverso complementare con la stessa significatività, determina se sono presenti in modo significativamente differente rispetto al proprio inverso complementare. I risultati sono nel formato riconosciuto da patternScan in modo che possano essere sfruttati per allineare le parole di interesse al genoma analizzato.
I modelli statistici che ho utilizzato in questo programma sono di due tipi. Una parola si distingue all'interno della famiglia di appartenenza, se il numero della sua conta ha un valore al di fuori dell'intervallo calcolato con la distribuzione di Poisson. Una parola è significativa rispetto al proprio inverso complementare, se il numero della sua conta ha un valore al di fuori dell'intervallo di 2σ calcolato dalla media dei rapporti di tutte le coppie parola e proprio inverso complementare.

4.2.4 patternScan

patternScan conta i pattern (vedi Paragrafo 2.3.1). Ho analizzato dei file multifasta con le regioni del 3' UTR, 5' UTR e dei CDS. Fra i pattern presenti in modo significativo nella regione analizzata, ho scelto quelli presenti in modo significativo tra le varie regioni. Anche in questo caso la conta dei pattern l'ho considerata significativa quando presentava un valore al di fuori dell'intervallo di 2σ dalla media dei pattern della stessa famiglia di appartenenza (vedi Paragrafo 2.3.1). Ho allineato i pattern così scelti al genoma mentre ho scartato i tipi di pattern che non erano presenti in nessuna delle zone dei TI.
È in sviluppo la parte di metodica che porta alla predizione di zone di microRNA precursore.

applicazione	**sensibilità**	**precisione**	**F-measure**	**user**	**real**
findMiRNA	0.57	0.93	0.71	>864000	>864000
microHarvester	0.50	0.96	0.60	61412	77857
PrecExtract	0.54	0.93	0.63	345403	472474
srnaloop	0.50	0.80	0.57	101826	101856
revComplScanP	0.99	0.35	0.61	1697	1698
MatriciPesateP	0.58	0.92	0.66	6371	6432
MetodoIntP	0.99	0.50	0.75	603	698
ConsensoP	0.99	0.50	0.75	641	655
MatriciPesateT	0.33	0.58	0.39	6571	6632
MetodoIntT	0.48	0.42	0.46	14821	15423

Tabella 4.1: Significatività e tempi di esecuzione delle applicazioni. I valori di sensibilità e precisione sono espressi in percentuale. user e real sono relativi ai tempi di esecuzione dei programmi (vedi Tabella 5.1) e sono espressi in secondi. I tempi di esecuzione sono tutti relativi all'analisi del genoma di vite. MatriciPesate P e T sono i tempi di esecuzione per la ricerca e allineamento delle matrici identificate a partire dal database di microRNA maturo depositati. I tempi di esecuzione relativi ai sistemi integrati sono il tempo di esecuzione di wEvidence. MetodoInt P e T sono relativi ai metodi proposti integrati con i tool di revComplScan, wconsensus e wEvidence. ConsensoP è relativo all'approccio proposto di usare più risultati di programmi di predizione per identificare delle regioni probabili per un microRNA precursore (vedi testo). Nel caso proposto l'approccio tiene conto dei risultati di findMiRNA, microHarvester, PrecExtract, srnaloop, revComplScanP e MatriciPesateP.

4.3 Attendibilità dei programmi

Il gruppo di Miranda *et al* (168), ha analizzato *Caenorhabditis elegans* con una metodica basata sui pattern, ed è riuscito a predire più del 70% dei precursori depositati e a validare sperimentalmente 168/226 siti bersaglio predetti. Il tool revComplScan si basa sulla conta delle parole e la ricerca delle matrici pesate dipende solo dai microRNA depositati perciò la metodica proposta non è specifica per microRNA di animali o di piante. Questo rende possibile mettere a confronto i risultati ottenuti con quelli presentati dalla metodica documentata in letteratura che sfrutta i pattern (168).
Analizzando *C. elegans* con revComplScanP si ottengono dei buoni risultati in quanto si sono identificati 133/136 microRNA maturi depositati del nematode. Con l'allineamento delle matrici PI si sono identificati 41/136 microRNA mentre l'integrazione delle due metodiche, creando i consensi con wEvidence, ne trova 135/136. Ciò conferma che la metodica integrata da risultati migliori.

La Tabella 4.1 riporta la significativià dei programmi che ho utilizzato e dei metodi inte-

grati. È possibile notare che la somma dei risultati per la predizione delle PI ha una sensibilità maggiore dei singoli programmi. Questo vuol dire che l'approccio usato aumenta il numero di veri positivi. Il metodo è pensato per identificare delle isole di precursori e perciò le posizioni di inizio e fine dei pre-miRNA predetti non sono ben definite come quelle calcolate dai singoli programmi per questo motivo il tasso di precisione è visibilmente diminuito. In ogni caso, il numero di risultati totali è minore di quelli ottenuti dai singoli programmi perciò questo metodo diminuisce in parte il numero di falsi positivi. Questo tipo di approccio è utile per filtrare dei risultati provenienti da programmi con alta sensibilità, ma con un numero considerevole di falsi positivi.

Con lo stesso approccio ho ricercato i siti bersaglio dei microRNA. I risultati di revComplScanT sono relativi a parole significativamente rappresentate lunghe 5 basi perciò non è possibile dare sensibilità e precisione rispetto alle posizioni dei microRNA maturi predetti perchè non coprono la lunghezza di un microRNA bersaglio. Per il momento è possibile utilizzarlo per identificare delle posizioni in cui intorno aspettarsi un ipotetico pre-miRNA. Per questo motivo per l'identificazione dei siti bersaglio dei microRNA, ho calcolato solo la significatività del metodo con le sole matrici pesate e il metodo integrato (matrici pesate e revComplScanT). I risultati ottenuti sono molto diversi. Le matrici utilizzate sono le inverse complementari di quelle sfruttate per la ricerca dei microRNA maturi e il loro allineamento al genoma copre pochissimi siti bersaglio documentati e in modo poco preciso. Sono in sviluppo miglioramenti sia dal punto di vista di implementazione dei programmi sia per la scelta dei loro parametri.

La scelta di parametri migliori (sulla base di sensibilità e tasso di precisione che ho calcolato sui microRNA bersaglio documentati) è sicuramente uno dei passaggi della metodica che richiede più tempo. Per la ricerca dei parametri da utilizzare nei programmi di wconsensus (per identificare delle matrici pesate), patser (per l'allineamento delle matrici), patternScan e revComplScanT ho eseguito ancora poche prove. Fattori limitanti, quali tempo di esecuzione e server disponibili, rendono molto lento tutto il processo di identificazione.

Capitolo 5

GMOD

È stato scelto di usare i toolkit del progetto GMOD (99) per creare un database interno di annotazione genomica da visualizzare e interrogare. Come accennato nell'introduzione, usare un tool in sviluppo porta dalle complicazioni quali errori negli script e poca documentazione che costringono l'utente ad interagire con gli sviluppatori per poter implementare un database efficiente. Tuttavia, il progetto GMOD è un progetto *open source* ed è possibile contribuire attivamente allo sviluppo dei tool. Di seguito sono descritti due tool che saranno presentati ai responsabili del progetto GMOD per essere inseriti nella distribuzione ufficiale.

5.1 GBrowse

Per risolvere alcuni problemi di incompatibilità con i file in formato GFF2 e visualizzare i risultati dei programmi usati, ho sviluppato degli script *ad hoc* per creare i file in formato GFF3 (136). I database dei browser sviluppati sono stati implementati a partire dal formato GFF3.

I file tabulati in formato GFF3, scaricabili dal sito del NCBI (120), sono stati creati a partire dai file in formato GenBank con lo script fornito dal tool del GBrowse (bp_genbank2gff3.pl). Questo script crea un tabulato formato GFF3 con degli errori sistematici di tipo sintattico che rendono tale file non immediatamente utilizzabile per la visualizzazione dei dati.

Dal 2007, il Dr. Payan Canaran, uno sviluppatore del progetto WormBase, si sta occupando

dell'implementazione di un tool per il controllo dei file formato GFF3. Ma il tool non è ancora disponibile e per questo motivo è stato scelto di creare un tool *ad hoc* per il controllo dei file in formato GFF3 e uno per il caricamento corretto dei dati nel database MySQL.

5.1.1 Controllo dei dati

In collaborazione con il Dr. Davide Campagna, ho implementato due programmi in linguaggio C per la verifica e la correzione dei file formato GFF3.
Errori sistematici comuni sono ID non unici, mancanza di ID e mancanza dell'evidenza di tipo *gene* per un blocco di evidenze di tipo *mRNA*, *exon* e *CDS* (vedi Tabella 2.4). Altri tipi di errori sono relativi all'uso scorretto dei cinque simboli chiave della sintassi cioé ';', '=', '%', '&', ',' che possono essere usati per altri scopi se inclusi tra virgolette. In più se negli attributi (vedi Tabella 2.4) è necessario usare degli spazi, bisogna farlo all'interno di virgolette oppure è possibile usare il simbolo '+' al posto dello spazio. Errori relativi all'uso della simbologia riconosciuta dagli script sono più difficili da riconoscere e perciò da correggere.
Ma l'errore più comune è sicuramente legato agli ID assenti o non unici. Per controllare l'esistenza e unicità degli ID (indipendentemente dal cromosoma in cui si trova l'evidenza), è stato necessario memorizzare codici identificativi, posizioni e verso per ogni evidenza presente nel tabulato, dividere le evidenze per blocchi (un esempio di blocco sono l'evidenza *gene* ID=1 e tutte le evidenze *mRNA*, *exon* e *CDS* che presentano come Parent=1; vedi Tabella 2.4) e confrontare ogni ID con tutti gli altri presenti nel tabulato (e se non presente l'ID, confrontare altri attributi che possono diventare l'identificativo unico di quell'evidenza).
I programmi di verifica e correzione dei tabulati formato GFF3 controllano gli attributi standard ID, Parent, Note, Target e Gap. In più controllano l'attributo locus_tag che nei tabulati prodotti dallo script bp_genbank2gff3.pl è spesso al posto dell'attributo ID. Il tool di verifica e correzione l'ho testato sul genoma di uomo ed ho creato un database GBrowse interno con tutti i geni visualizzati correttamente per ogni cromosoma (169).

5.1.2 Caricamento dei dati

Il tool scaricabile di caricamento dei tabulati in formato GFF nel database MySQL è costituito da tre script tutti funzionanti: bp_load_gff.pl (presente dalla versione GFF1), bp_bulk_load_gff.pl e bp_fast_load_gff.pl (presenti dalla versione GFF2). Il primo script è il più lento nel caricare i dati perché per ogni record interroga il database MySQL richiedendo se ID e/o valori del record in analisi sono già presenti nel database in modo da inserire corretamente i codici identificativi in ogni tabella (vedi Figura 2.1). Dopo aver analizzato i risultati dell'interrogazione del database, inserisce o meno ogni record singolarmente. Il secondo script prevede invece un'unica interrogazione iniziale e l'inserimento dei dati in un'unica volta alla fine della parserizzazione del tabulato GFF3. Entrambi gli script sono funzionanti con database MySQL sia in locale che in remoto e riconoscono i formati GFF1, GFF2 e GFF3. Il terzo script è la versione del secondo script per l'aggiornamento dei database e funziona solo per database remoti. Quest'ultimo script non è stato sfruttato perché non interamente funzionante e perché ha la stessa velocità del secondo script descritto. Tutti gli script non prevedono l'inserimento di evidenze con errori sintattici qualunque esse siano e non prevedono un semplice controllo di unicità degli ID che comportano il caricamento di codici identificativi errati o mancanti provocando una visualizzazione errata dei dati.

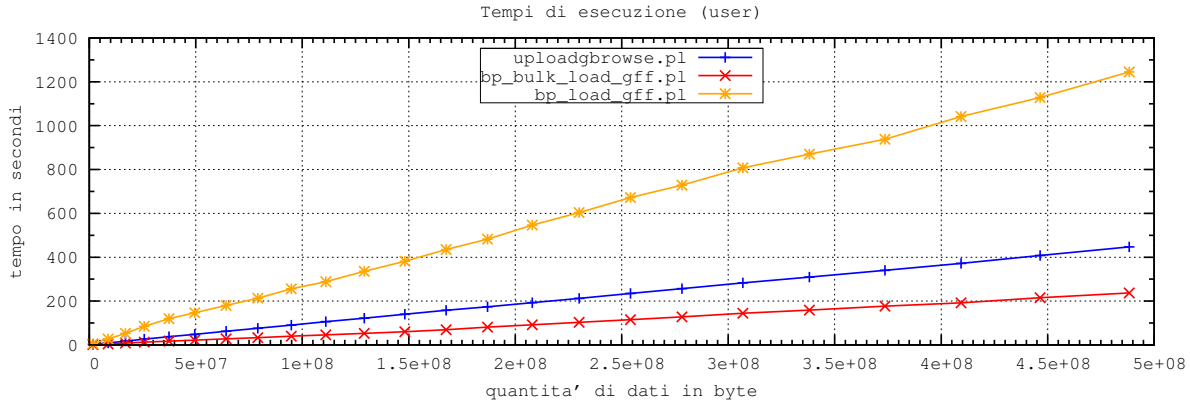
Il tool di caricamento che ho implementato (uploadgbrowse.pl) interroga una volta sola il database MySQL all'inizio, come bp_bulk_load_gff.pl. I dati che vengono richiesti al database MySQL ad ogni record dallo script bp_load_gff.pl, in uploadgbrowse.pl, sono caricati in memoria in una tabella con un'unica interrogazione che viene aggiornata dei nuovi dati da inserire man mano che lo script parserizza il tabulato GFF3, fino a database di dimensioni di 10GB (il database più grande che è stato testato). I record vengono inseriti in una sola volta alla fine dell'intera lettura dei file in formato GFF per mezzo di flat file. Questa tecnica è più veloce rispetto a quella di inserimento dei dati interrogando il database. Lo script riconosce le versioni GFF2 e GFF3. Funziona per database in locale e in remoto sia per la creazione di un database che per il suo aggiornamento. Lo script presentato prevede la stampa di *warning* riguardo ad evidenze con errori sintattici o alla mancanza di codici identificativi di riferimento,

in modo da aiutare l'utente nella correzione del tabulato GFF o per avvertirlo delle modifiche di default che sono state eseguite per un caricamento corretto. Infatti, per ovviare agli errori di visualizzazione, fa dei semplici controlli di unicità dell'ID e, in mancanza, se è un'evidenza che ha parentele (vedi Tabella 2.4 e Figura 2.1), ricerca un possibile codice identificativo unico (ad esempio, il locus_tag). Se viene trovato un possibile codice identificativo unico allora l'evidenza viene inserita e comunque l'utente ne viene informato. In più se sono presenti un numero di dispari di virgolette (e perciò una virgoletta non è stata chiusa), o se sono presenti degli spazi non previsti senza virgolette non inserisce gli attributi e segnala l'errore all'utente.
Il confronto delle prestazioni di bp_load_gff.pl, bp_bulk_load_gff.pl e uploadgbrowse.pl sono illustrate in Figura 5.1 e Figura 5.2.

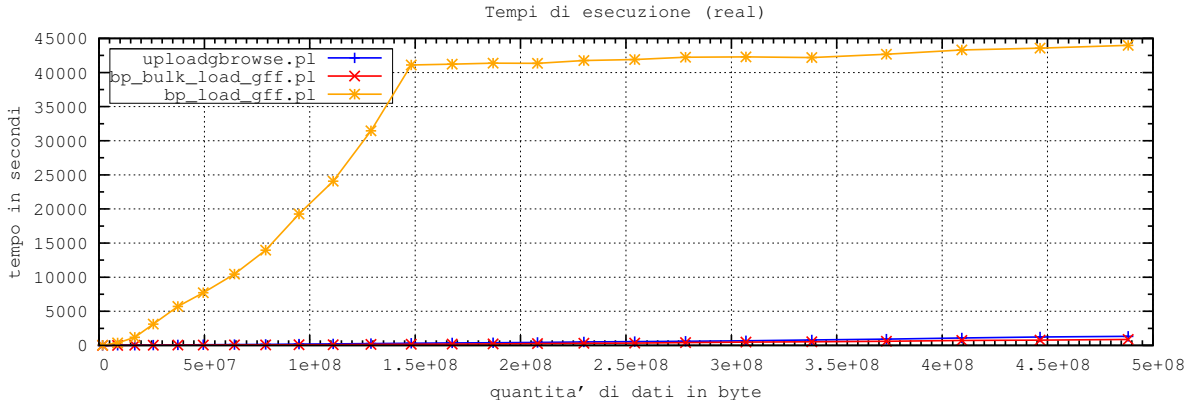
5.2 Prossima distribuzione

Il tool di verifica e correzione dei tabulati formato GFF3 e il tool per il caricamento dei dati nel database MySQL che ho implementato saranno presentati ai responsabili del sottoprogetto GBrowse e ai loro sviluppatori quest'anno per essere valutati prima di essere inseriti nella prossima distribuzione del GBrowse. La distribuzione corrente è la 1.68 perciò i tool prodotti potrebbero essere inseriti nella prossima versione rilasciata.

A



B



C

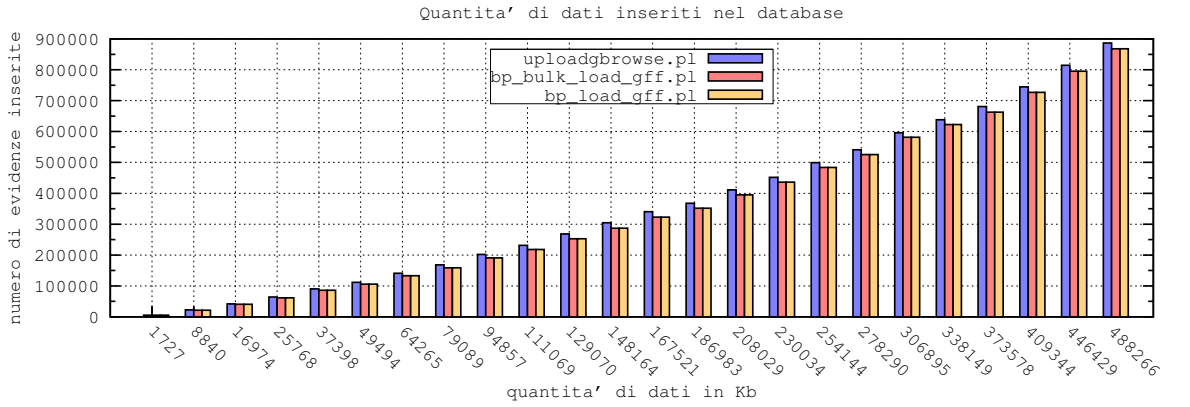
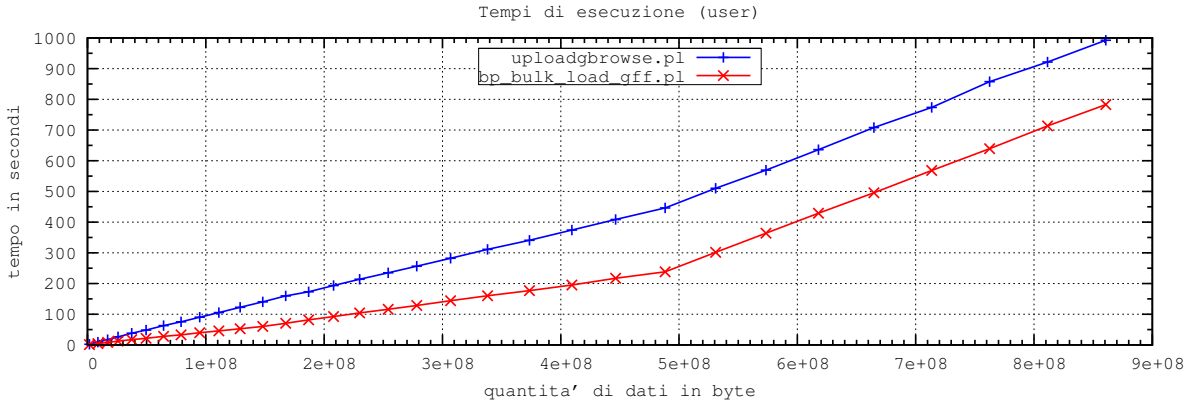
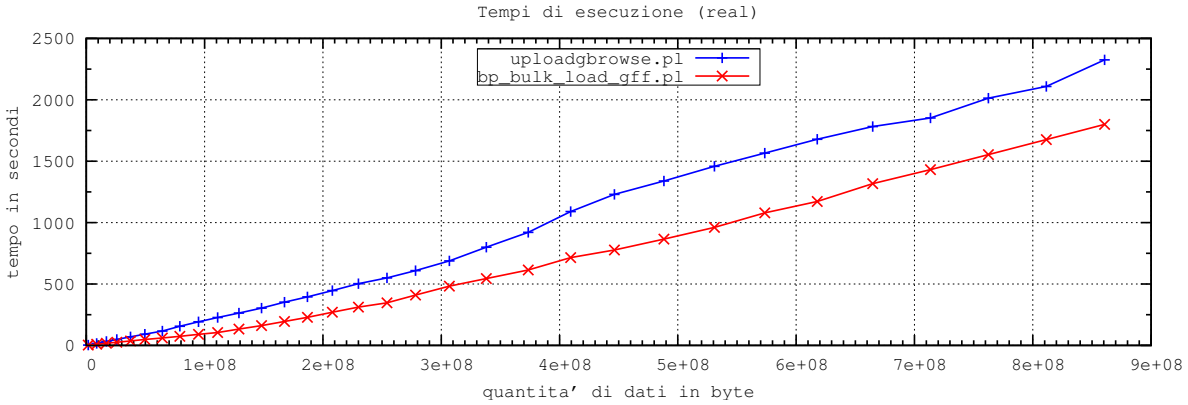


Figura 5.1: Confronto degli script di caricamento. Con user si intende il tempo effettivo in cui il programma ha occupato la CPU mentre con real si intende il tempo reale di esecuzione del programma. Il grafico (A) mostra il tempo occupato dagli script per elaborare i dati del tabulato ed eseguire le chiamate al database MySQL. Il grafico (B) presenta i tempi reali di esecuzione complessivi dell'elaborazione e caricamento dei dati da parte di MySQL. Il tempo di esecuzione dello script bp_load_gff.pl è maggiore a causa delle numerose chiamate al database (vedi testo) mentre sono confrontabili i tempi degli altri due script perché inseriscono entrambi i dati con un'unica chiamata a MySQL. Il diagramma (C) dimostra che il numero di dati inseriti è minore negli script della distribuzione GBrowse. Lo script presentato inserisce in modo esatto tutte le evidenze riportate nel tabulato GFF3 di uomo scaricabili dal sito dell'NCBI.

A



B



C

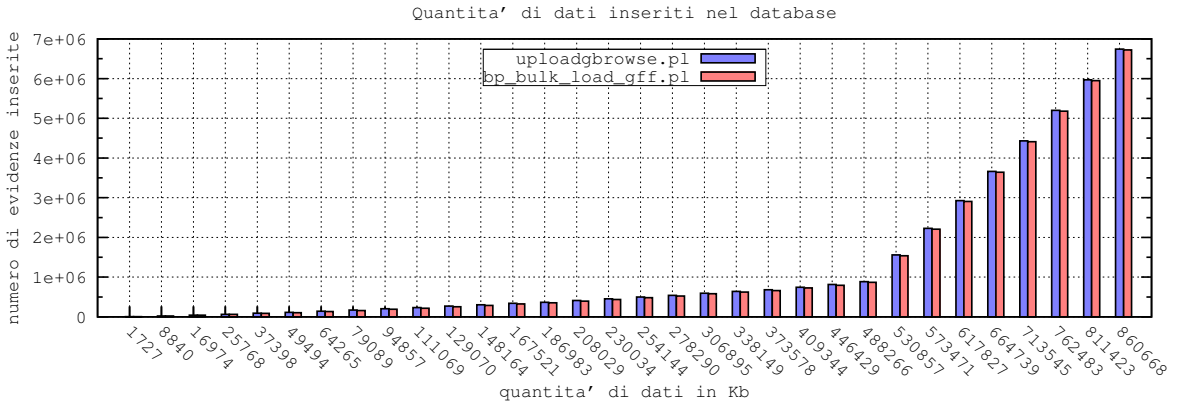


Figura 5.2: Confronto tra gli script di caricamento bp_bulk_load_gff.pl e uploadgbrowse.pl. Per comprendere se anche con altri tipi di dato, diversi da quelli estrapolati dai file del GenBank, ci sono dei problemi di caricamento e per vedere se le prestazioni dei due script rimane simile anche con grandi database, sono stati inseriti gli stessi dati e in più dei dati di tipo quantitativo (sono dati atti alla visualizzazione di un grafico a due dimensioni e presentano un valore per ogni base del genoma visualizzato). Il grafico (A) è paragonabile a quello presentato in Figura 5.1 A. Si può notare che il tempo di esecuzione cambia dopo l'inserimento dei 500Mb, momento in cui si cominciano ad inserire anche i dati di tipo quantitativo. Infatti questo tipo di dati necessita di ulteriori controlli che sembrano essere fatti da entrambi gli script, in tempi simili. Il tempo reale dei due script è simile anche per database grandi (B) e anche l'inserimento dei dati di tipo quantitativo è corretto per entrambi gli script (C).

Capitolo 6

Conclusioni

L'annotazione strutturale e funzionale dei geni è alla base dell'interpretazione della regolazione genica. La conoscenza della regolazione dei geni è fondamentale per manipolare, o semplicemente conoscere meglio, vie metaboliche, differenziamento cellulare ed organogenesi. Il sequenziamento di nuovi genomi, siano essi di procarioti od eucarioti, e la loro prima analisi sono momenti fondamentali per l'uso di una nuova sequenza a scopi comparativi e/o sperimentali.

Ho condotto analisi sui Rho Dependent Terminator (RDT) per comprendere se esistono delle particolari vie metaboliche in cui è coinvolta la proteina Rho e perché un sistema così complesso sembra regolare così pochi geni. Inoltre, conoscere la posizione dei RDT sarebbe utile per l'implementazione di un'applicazione che predica operoni e TU alternative.

Dall'analisi della composizione nucleotidica di RDT sono emerse nuove caratteristiche (vedi Paragrafo 3.1) che ho poi sfruttato per la loro predizione. Ho presentato il primo algoritmo per la predizione *in silico* dei RDT e sulla base dei risultati della sua implementazione, ho prodotto la prima matrice pesata (vedi Paragrafo 3.3) che copre le regioni di tutti i RDT documentati. RDTSCAN predice molti RDT putativi che sono ridotti fino al numero delle posizioni di allineamento della matriceRDT (vedi Paragrafo 3.3) a seconda dei parametri scelti al momento dell'analisi della sequenza genomica. Il programma presenta un alto numero di falsi positivi (vedi Tabelle C.3 e C.5) predetti in regioni in cui è presente un gene nell'elica

opposta. Perciò il programma sarà migliorato tenendo in considerazione le posizioni dei geni. Probabilmente l'implementazione di altre caratteristiche (quali le strutture secondarie descritte nel Paragrafo 3.1) potrebbero ridurre ancora il numero di falsi positivi.
Dall'analisi dei risultati di RDTSCAN è possibile ipotizzare che la proteina Rho potrebbe regolare un alto numero di geni. Inoltre sembrano essereci solo determinate categorie funzionali regolate dalla proteina Rho. Si tratta di geni coinvolti nel trasporto e metabolismo di aminoacidi, zuccheri e piccole molecole, traduzione e biogenesi di RNA ribosomale.
Con gli esperimenti di microarray è stato confermato dai controlli che la metodica di ricercare i RDT nel leader con gli oligo che si sono accesi o spenti funziona. Purtroppo i tempi utilizzati appaiono tardivi perciò confronti di esperimenti condotti in terreno minimo con tempi più piccoli di esposizione alla BCM rispetto a quelli presentati, in terreno ottimale con e senza la presenza di antibiotico e in presenza o meno di Heat Shock potrebbero essere utili a completare il quadro disegnato.
Per contribuire al progetto del sequenziamento di *Vitis vinifera* ho effettuato le prime analisi sui geni di microRNA. L'approccio computazionale presentato ha portato alla messa a punto di un tool di programmi di per la ricerca dei precursori di microRNA. I primi dati evidenziano che la sensibilità delle metodiche proposte sono migliori dei risultati che si possono ottenere con i singoli programmi di predizione. Il metodo è stato pensato per identificare delle isole di precursori perciò le posizioni di inizio e fine non sono ben definite come quelle calcolate dai singoli programmi. Il laboratorio del prof. Valle collabora con Solexa. Sono state eseguite delle estrazioni di small RNA da callo di Vite che sono stati sequenziati con il metodo SBS (96). La possibilità di servirsi di un metodo di sequenziamento su larga scale come il metodo SBS aumenta la probabilità di identificare nuovi microRNA. Ma la mole di dai che ne deriva deve essere filtrata in modo opportuno per ottenere in tempi ragionevoli dei risultati. Una possibilità è quella di allineare il database derivante dal sequenziamento con le isole dei precursori predetti a partire dal metodo integrato che ho presentato. Questo porterebbe alla validazione di nuovi microRNA in poco tempo e ad una miglioria del metodo di predizione sulla base dell'analisi dei nuovi microRNA confermati.
Tutti i dati presentati in questa tesi li ho visualizzati e interrogati con l'uso dei tool del progetto

GMOD. Lo sviluppo e miglioramento dei tool relativi al sottoprogetto GBrowse hanno reso possibile l'utilizzo in tempi ragionevoli di tutta la mole di dati relativa all'annotazione genica di batteri e piante analizzate. Gli strumenti messi a punto saranno proposti quest'anno ai responsabili del sottoprogetto GBrowse per poterli includere nella prossima distribuzione del pacchetto GBrowse. I risultati di configurazione e miglioramento degli script riguardo la visualizzazione e l'interrogazione dei database GBrowse sono importanti in quanto il laboratorio del prof. Valle sarà responsabile dello sviluppo della piattaforma di annotazione del progetto di sequenziamento del genoma di *Vitis vinifera*.

Appendice A

Acronimi usati

Ago Argonaute

ARS Agricultural Research Service

BLAST Basic Local Alignment Search Tool

BCM Bicyclomycin

COGs Clusters of Orthologous Groups of proteins

crgpi Citosine Rich and Guanine Poor Island

DCL1 Dicer-like 1 enzyme

DSMZ Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH

EST Expressed Sequence Tag

GFF General Feature Format

GMOD Generic Model Organism Database

GO Gene Ontology

GOA Gene Ontology Annotation

KEGG Kyoto Encyclopedia of Genes and Genomes

LOWESS Localised Weighted Smother Estimator

MI Microarray Imager

MIDAS Microarray Data Analysis System

MPSS Massively Parallel Signature Sequencing

ncRNA noncoding RNA

NIH National Institutes of Health

OD Optical Density

P-bodies Processing Bodies

PI Precursor Island

ppv Positive Predictive Value

pre-miRNA Precursor of microRNA

pri-miRNA Primary transcript of microRNA gene

SAM Significance Analysis of Microarrays

SBS Sequencing by Synthesis

SD Shine-Dalgarno

si Strong Island

sp Stop Point

rac Rho ACtivity

RISC RNA-induced Silencing Complex

RDT Rho Dependent Terminator

RIT Rho Independent Terminator

rut Rho UTilization

TAIR The Arabidopsis Information Resource

tav TIGR Array Viewer

TBA Threaded Blockset Aligner

ti Timine Island

TI Target Island

TFBS Transcriptional Factor Binding Site

tsp Transcription Stop Point

TSS Transcriptional Start Site

TU Transcriptional Unit

USDA United State Department of Agricolcure

UTR Untranslated Region

wi Weak Island

yi Pyrimidine Island

Appendice B

Cromosomi batterici analizzati

Aquifex aeolicus (AE000657), *Archaeoglobus fulgidus* (AE000782), *Bacillus anthracis* Ames (AE016879), *Bacillus anthracis* Ames 0581 (AE017334), *Bacillus anthracis* str Sterne (AE017225), *Bacillus cereus* ATCC14579 (AE016877), *Bacillus cereus* ATCC 10987 (AE017194), *Bacillus cereus* ZK (CP000001), *Bacillus halodurans* (BA000004), *Bacillus licheniformis* ATCC 14580 (AE017333, CP000002), *Bacillus licheniformis* DSM 13 (AE017333), *Bacillus subtilis* (AL009126), *Bacillus thuringiensis* konkukian (AE017355), *Bifidobacterium longum* (AE014295), *Borrelia burgdorferi* (AE000783), *Caulobacter crescentus* (AE005673), *Clostridium acetobutylicum* (AE001437), *Clostridium perfringens* (BA000016), *Clostridium tetani*E88 (AE015927), *Corynebacterium diphtheriae* (BX248353), *Corynebacterium glutamicum* ATCC 13032 Bielefeld (BX927147), *Corynebacterium glutamicum* ATCC 13032 Kitasato (BA000036), *Dehalococcoides ethenogenes* 195 (CP000027), *Deinococcus radiodurans* (AE000513, AE001825), *Enterobacteria lambda* (J02459), *Enterobacteria phage* F1 (J02448), *Enterobacteria phage* P4 (X51522), *Enterobacteria phage* T5 (NC_005859), *Enterococcus faecalis* V583 (AE016830), *Escherichia coli* K12 (U00096), *Escherichia coli* O157H7 EDL933 (AE005174), *Geobacillus kaustophilus* HTA426 (BA000043), *Helicobacter pylori* 26695 (AE000511), *Helicobacter pylori* J99(AE001439), *Lactobacillus acidophilus* NCFM (CP000033), *Lactobacillus johnsonii* NCC 533 (AE017198), *Lactobacillus plantarum* (AL935263), *Lactococcus lactis* (AE005176), *Listeria innocua* (AL592022), *Listeria monocytogenes* (AL591824), *Listeria monocytogenes* 4b

F2365 (AE017262), *Mesoplasma florum* L1 (AE017263), *Methanobacterium thermoautotrophicum* (AE000666), *Methanococcus jannaschii* (L77117, L77118, L77119), *Mycobacterium avium* paratuberculosis (AE016958), *Mycobacterium bovis* (BX248333), *Mycobacterium leprae* (AL450380), *Mycobacterium tuberculosis* CDC1551 (AE000516), *Mycobacterium tuberculosis* H37Rv (AL123456), *Mycoplasma gallisepticum* (AE015450), *Mycoplasma genitalium* (L43967), *Mycoplasma hyopneumoniae* 232 (AE017332), *Mycoplasma mobile* 163K (AE017308), *Mycoplasma mycoides* (BX293980), *Mycoplasma penetrans* (BA000026), *Mycoplasma pneumoniae* (U00089), *Mycoplasma pulmonis* (AL445566), *Neisseria meningitidis* MC58 (AE002098), *Nocardia farcinica* IFM10152 (AP006618), *Oceanobacillus iheyensis* (BA000028), *Onion yellows* phytoplasma (AP006628), *Photobacterium profundum* SS9 (CR354531, CR354532, CR377818), *Propionibacterium acnes* KPA171202 (AE017283), *Pyrococcus horikoshii* (BA000001), *Rhodobacter sphaeroides* 2 4 1 (CP000143, CP000144, CP000145, CP000146, CP000147), *Salmonella typhimurium* LT2 (AE006468, AE006471), *Staphylococcus aureus* COL (CP000046), *Staphylococcus aureus* MW2 (BA000033), *Staphylococcus aureus* Mu50 (AP003367, BA000017), *Staphylococcus aureus* N315 (BA000018), *Staphylococcus aureus* aureus MRSA252 (BX571856), *Staphylococcus aureus* aureus MSSA476 (BX571857), *Staphylococcus epidermidis* ATCC 12228 (AE015929), *Staphylococcus epidermidis* RP62A (CP000029), *Streptococcus agalactiae* 2603 (AE009948), *Streptococcus agalactiae* NEM316 (AL732656), *Streptococcus mutans* (AE014133), *Streptococcus pneumoniae* R6 (AE007317), *Streptococcus pneumoniae* TIGR4 (AE005672), *Streptococcus pyogenes* M1 GAS (AE004092), *Streptococcus pyogenes* MGAS10394 (CP000003), *Streptococcus pyogenes* MGAS315 (AE014074), *Streptococcus pyogenes* MGAS8232 (AE009949), *Streptococcus thermophilus* CNRZ1066 (CP000024), *Streptococcus thermophilus* LMG 18311 (CP000023), *Streptomyces avermitilis* (BA000030), *Streptomyces coelicolor* (AL645882), *Symbiobacterium thermophilum* IAM14863 (AP006840), *Synechocystis PCC6803* (BA000022), *Thermoanaerobacter tengcongensis* (AE008691), *Thermotoga maritima* (AE000512), *Treponema pallidum* (AE000520), *Tropheryma whipplei* TW08 27 (BX072543), *Tropheryma whipplei* Twist (AE014184), *Ureaplasma urealyticum* (AF222894), *Vibrio cholerae* (AE003852, AE003853), *Vibrio parahaemolyticus* (BA000031, BA000032), *Vibrio vulnificus* CMCP6 (AE016795, AE016796), *Vibrio vulnificus* YJ016 (AP005352, BA000037, BA000038).

Appendice C

Tabelle supplementari

cromosoma	**nt**	**real**	**user**	**system**
Aquifex aeolicus chr1	1551335	1m28.671s	1m28.228s	0m0.410s
Archaeoglobus fulgidus chr1	2178400	2m3.459s	2m2.178s	0m0.934s
Bacillus subtilis chr1	4214630	3m18.922s	3m17.211s	0m1.024s
Bacteriophage P4 chr1	910484	0m23.304s	0m23.068s	0m0.219s
Bacteriophage T5 chr1	4016947	2m20.683s	2m19.940s	0m0.755s
Bacteriophage lambda chr1	48502	0m4.083s	0m4.048s	0m0.025s
Borrelia burgdorferi chr1	6407	0m2.225s	0m2.197s	0m0.016s
Caulobacter crescentus chr1	11624	0m2.321s	0m2.287s	0m0.014s
Enterobacteria phage f1 chr1	121750	0m8.380s	0m8.336s	0m0.039s
Escherichia coli K12 chr1	4639675	3m34.622s	3m33.318s	0m1.213s
Escherichia coli O157H7 EDL933 chr1	5500425	5m10.739s	5m9.236s	0m1.464s
Helicobacter pylori 26695 chr1	1667507	1m32.501s	1m32.048s	0m0.448s
Helicobacter pylori J99 chr1	1643831	1m31.386s	1m30.979s	0m0.405s
Methanobacterium thermoautotrophicum chr1	1751377	1m39.963s	1m39.472s	0m0.490s
Methanococcus jannaschii chr1	1664910	1m9.822s	1m9.400s	0m0.424s
Methanococcus jannaschii chr2	58287	0m3.377s	0m3.333s	0m0.020s
Methanococcus jannaschii chr3	16550	0m3.377s	0m3.333s	0m0.020s
Mycobacterium tuberculosis CDC1551 chr1	4402337	2m36.652s	2m35.769s	0m0.894s
Mycobacterium tuberculosis H37Rv chr1	4411532	2m37.436s	2m36.561s	0m0.885s
Mycoplasma genitalium chr1	580074	0m28.863s	0m21.490s	0m0.177s
Mycoplasma pneumoniae chr1	816394	0m47.090s	0m46.882s	0m0.203s
Neisseria meningitidis MC58 chr1	2272351	2m5.563s	2m4.980s	0m0.585s
Photobacterium profundum SS9 chr1	4085124	3m11.238s	3m10.148s	0m1.093s
Photobacterium profundum SS9 chr2	2237883	2m2.321s	2m1.680s	0m0.611s
Photobacterium profundum SS9 chr3	80033	0m6.387s	0m6.345s	0m0.036s

Tabella C.1 – continua nella prossima pagina

Tabella C.1 – continua dalla pagina precedente

cromosoma	**nt**	**real**	**user**	**system**
Pyrococcus horikoshii chr1	1738505	1m40.307s	1m39.863s	0m0.450s
Rhodobacter sphaeroides 2 4 1 chr1	3188429	1m36.767s	1m36.089s	0m0.688s
Rhodobacter sphaeroides 2 4 1 chr2	943016	0m22.262s	0m22.068s	0m0.192s
Salmonella typhimurium LT2 chr1	4857432	3m49.034s	3m47.871s	0m1.178s
Salmonella typhimurium LT2 chr2	93939	0m6.832s	0m6.782s	0m0.046s
Staphylococcus aureus COL chr1	2809422	2m6.758s	2m6.071s	0m0.683s
Staphylococcus aureus MW2 chr1	2820462	2m6.977s	2m6.329s	0m0.641s
Staphylococcus aureus Mu50 chr1	2878529	2m13.682s	2m9.292s	0m0.707s
Staphylococcus aureus Mu50 chr2	25107	2m13.682s	2m9.292s	0m0.707s
Staphylococcus aureus N315 chr1	2814816	2m6.677s	2m6.024s	0m0.642s
Staphylococcus aureus aureus MRSA252 chr1	2902619	2m9.928s	2m8.138s	0m0.781s
Staphylococcus aureus aureus MSSA476 chr1	2799802	2m6.219s	2m5.584s	0m0.633s
Synechocystis PCC6803 chr1	3573470	3m3.826s	3m2.915s	0m0.921s
Thermotoga maritima chr1	1860725	1m45.747s	1m45.201s	0m0.527s
Treponema pallidum chr1	1137471	1m5.885s	1m5.578s	0m0.292s
Ureaplasma urealyticum chr1	751719	0m20.029s	0m19.845s	0m0.178s
Vibrio cholerae chr1	2960969	2m38.917s	2m38.142s	0m0.786s
Vibrio cholerae chr2	1072255	1m1.758s	1m1.441s	0m0.310s
Vibrio parahaemolyticus chr1	3288558	2m51.644s	2m50.784s	0m0.869s
Vibrio parahaemolyticus chr2	1877212	1m47.320s	1m46.765s	0m0.542s
Vibrio vulnificus CMCP6 chr1	3281945	2m53.466s	2m52.626s	0m0.846s
Vibrio vulnificus CMCP6 chr2	1844853	1m45.703s	1m45.171s	0m0.530s
Vibrio vulnificus YJ016 chr1	3354505	2m55.951s	2m55.077s	0m0.884s
Vibrio vulnificus YJ016 chr2	1857073	1m44.916s	1m44.338s	0m0.575s
Vibrio vulnificus YJ016 chr3	48508	1m44.916s	1m44.338s	0m0.575s

Tabella C.1: Tempi di esecuzione di RDTSCAN. nt: lunghezza del cromosoma in nucleotidi; real: tempo reale di esecuzione del programma; user: tempo effettivo in cui il programma ha occupato la CPU; system: MmSs: M, minuti; S, secondi di tempo dell'esecuzione di un singolo genoma.

info	**cromosoma**	**nt**	**GC**	**RDT**	**exp**	**sig**
ri −	Aae chr1	1551335	0.43	708	268	722.39
i *a*	Afu chr1	2178400	0.49	674	156.69	1707.96
rn +	Bsu chr1	4214630	0.44	1771	733.73	1466.4
r *p*	Bbu chr1	910484	0.29	681	677.49	0.02
r *p*	Ccr chr1	4016947	0.67	186	8.29	3807.44
r *p*	Eph lambda chr1	48502	0.5	15	32.08	9.09
ri −	Eph f1 chr1	6407	0.41	2	12.9	9.21
re −	Eph P4 chr1	11624	0.5	3	67	61.13

Tabella C.2 – continua nella prossima pagina

Tabella C.2 – continua dalla pagina precedente

info	cromosoma	nt	GC	RDT	exp	sig
r *p*	Eph T5 chr1	121750	0.39	85	38.45	56.35
re −	Eco K12 chr1	4639675	0.51	985	211.1	2837.19
re −	Eco O157H7 EDL933 chr1	5500425	0.5	1116	277.14	2539.14
ri −	Hpy 26695 chr1	1667507	0.39	1145	494.39	856.18
ri −	Hpy J99 chr1	1643831	0.39	1120	462.16	936.39
i *a*	Mth chr1	1751377	0.5	347	107.6	532.64
i *a*	Mja chr1	1664910	0.31	1284	1146	16.62
i *a*	Mja chr2	58287	0.28	55	44.94	2.25
i *a*	Mja chr3	16550	0.29	14	8.7	3.23
r +	Mtu CDC1551 chr1	4402337	0.66	328	25.51	3586.87
r +	Mtu H37Rv chr1	4411532	0.66	327	25.86	3506.34
i 0	Mge chr1	580074	0.32	578	388.06	92.97
i 0	Mpn chr1	816394	0.4	472	205.8	344.31
r +	Nme MC58 chr1	2272351	0.52	552	90.47	2354.46
r −	Ppr SS9 chr1	4085124	0.42	1633	907.71	579.54
r −	Ppr SS9 chr2	2237883	0.41	991	545.86	363
r −	Ppr SS9 chr3	80033	0.44	26	12.1	15.97
i *a*	Pho chr1	1738505	0.42	587	382.53	109.29
re −	Rsp 2 4 1 chr1	3188429	0.69	69	632.49	502.02
re −	Rsp 2 4 1 chr2	943016	0.69	31	452.51	392.63
r −	Sty LT2 chr1	4857432	0.52	5241	2088.69	4757.58
r −	Sty LT2 chr2	93939	0.53	76	36.1	44.1
rn +	Sau COL chr1	2809422	0.33	2012	1711.47	52.77
rn +	Sau MW2 chr1	2820462	0.33	2016	1717.59	51.85
rn +	Sau Mu50 chr1	2878529	0.33	2055	1750.51	52.96
rn +	Sau Mu50 chr2	25107	0.29	19	18.3	0.03
rn +	Sau N315 chr1	2814816	0.33	2019	1711.8	55.13
rn +	Sau MRSA252 chr1	2902619	0.33	2092	1774.08	56.97
rn +	Sau MSSA476 chr1	2799802	0.33	1997	1697.47	52.85
ri +	Syn PCC6803 chr1	3573470	0.48	899	301.25	1186.04
r +	Tma chr1	1860725	0.46	635	207.08	884.29
ri −	Tpa chr1	1137471	0.53	795	437.22	292.78
r 0	Uur chr1	751719	0.25	685	611.76	8.77
r −	Vch chr1	2960969	0.48	716	257.8	814.35
r −	Vch chr2	1072255	0.47	296	102.4	366.02
r −	Vpa chr1	3288558	0.45	1048	418.65	946.11
r −	Vpa chr2	1877212	0.45	620	238.9	607.93
r −	Vvu CMCP6 chr1	3281945	0.46	932	347.43	983.56
r −	Vvu CMCP6 chr2	1844853	0.47	521	177.18	667.21
r −	Vvu YJ016 chr1	3354505	0.46	981	366.1	1032.8

Tabella C.2 – continua nella prossima pagina

Tabella C.2 – continua dalla pagina precedente

info	cromosoma	nt	GC	RDT	exp	sig
r −	Vv YJ016 chr2	1857073	0.47	529	171.86	742.14
r −	Vvu YJ016 chr3	48508	0.45	15	6.4	11.56

Tabella C.2: Significatività dei risultati ottenuti con RDTSCAN.

cromosoma	RDT	%	bp/RDT	RDTLE	RDTEN	RDTRA
Aae chr1	708	0.64	3439.77	0.46	0.41	0.56
Afu chr1	674	0.58	5557.14	0.61	0.42	0.51
Bsu chr1	1771	0.58	4087.9	0.52	0.36	0.56
Bbu chr1	3	0.67	5812	1	0	0.5
Ccr chr1	85	0.59	2435	0.86	0.56	0.6
Eph lambda chr1	15	0.47	6928.86	0.86	0.57	0.71
Eph f1 chr1	681	0.54	2480.88	0.53	0.34	0.6
Eph P4 chr1	186	0.56	38256.64	0.68	0.37	0.42
Eph T5 chr1	2	0.5	6407	1	0	1
Eco K12 chr1	985	0.59	7931.07	0.62	0.39	0.47
Eco O157H7 EDL933 chr1	1116	0.56	8857.37	0.54	0.42	0.51
Hpy 26695 chr1	1145	0.6	2437.88	0.56	0.52	0.49
Hpy J99 chr1	1120	0.59	2486.89	0.52	0.54	0.49
Mth chr1	347	0.6	8460.76	0.57	0.53	0.39
Mja chr1	1284	0.6	2176.35	0.49	0.45	0.49
Mja chr2	55	0.38	2775.57	0.43	0.24	0.67
Mja chr3	14	0.64	1838.89	0.44	0.22	0.44
Mtu CDC1551 chr1	328	0.55	24188.66	0.52	0.29	0.62
Mtu H37Rv chr1	327	0.51	26259.12	0.49	0.28	0.64
Mge chr1	578	0.55	1812.73	0.48	0.38	0.63
Mpn chr1	472	0.58	2990.45	0.56	0.56	0.42
Nme MC58 chr1	552	0.56	7330.16	0.55	0.49	0.47
Ppr SS9 chr1	1633	0.56	4454.88	0.54	0.38	0.48
Ppr SS9 chr2	991	0.57	3960.85	0.48	0.38	0.47
Ppr SS9 chr3	26	0.69	4446.28	0.5	0.22	0.44
Pho chr1	587	1	2961.68	0.36	0.29	1
Rsp 2 4 1 chr1	69	0.51	91097.97	0.49	0.34	0.6
Rsp 2 4 1 chr2	31	0.52	58938.5	0.5	0.62	0.5
Sty LT2 chr1	5241	0.58	1600.47	0.52	0.46	0.49
Sty LT2 chr2	76	0.51	2408.69	0.49	0.56	0.49
Sau COL chr1	2012	0.59	2349.02	0.5	0.43	0.53
Sau MW2 chr1	2016	0.59	2388.2	0.49	0.42	0.53
Sau Mu50 chr1	2055	0.58	2398.77	0.51	0.41	0.53

Tabella C.3 – continua nella prossima pagina

Tabella C.3 – continua dalla pagina precedente

cromosoma	**RDT**	**%**	**bp/RDT**	**RDTLE**	**RDTEN**	**RDTRA**
Sau Mu50 chr2	19	0.63	2092.25	0.5	0.58	0.42
Sau N315 chr1	2019	0.58	2395.59	0.49	0.41	0.52
Sau MRSA252 chr1	2092	0.6	2327.68	0.48	0.41	0.55
Sau MSSA476 chr1	1997	0.59	2382.81	0.5	0.42	0.54
Syn PCC6803 chr1	899	0.77	5141.68	0.36	0.28	0.84
Tma chr1	635	0.62	4746.75	0.55	0.4	0.62
Tpa chr1	795	0.57	2527.71	0.49	0.43	0.57
Uur chr1	685	0.54	2020.75	0.48	0.32	0.7
Vch chr1	716	0.57	7257.28	0.51	0.37	0.57
Vch chr2	296	0.6	6057.94	0.54	0.37	0.5
Vpa chr1	1048	0.62	5067.12	0.55	0.39	0.49
Vpa chr2	620	0.63	4801.05	0.51	0.4	0.48
Vvu CMCP6 chr1	932	0.57	6227.6	0.51	0.39	0.51
Vvu CMCP6 chr2	521	0.58	6088.62	0.49	0.35	0.5
Vvu YJ016 chr1	981	0.59	5813.7	0.52	0.39	0.52
Vv YJ016 chr2	529	0.61	5749.45	0.48	0.34	0.53
Vvu YJ016 chr3	15	0.4	8084.67	0.83	0.67	0.33

Tabella C.3: Distribuzione dei RDT putativi come RDTLE, RDTEN e/o RDTRA rispetto ai geni. RDT, numero di RDT putativi osservati; %, percentuale di RDT putativi presenti nelle regioni dei geni. Ciascun RDT predetto può essere contato sia come RDTLE, RDTEN e/o RDTRA simultaneamente.

cromosoma	**RDT**	**%**	**bp/RDT**	**RDTLE**	**RDTEN**	**RDTRA**
Aae chr1	13	0.77	119333,46	0.62	0.15	0.54
Afu chr1	6	1.00	363066,67	0.67	0.33	0.50
Bsu chr1	17	0.94	247919,41	0.82	0.06	0.65
Bbu chr1	124	0.71	7342,61	0.50	0.15	0.42
Ccr chr1	1	1.00	4016947	1.00	0.00	1.00
Eph lambda chr1	6	0.67	8083,67	0.67	0.00	0.50
Eph f1 chr1	2	0.50	3203,5	0.50	0.00	0.00
Eph P4 chr1	2	0.50	5812	0.50	0.00	0.50
Eph T5 chr1	0	0.00	0	0.00	0.00	0.00
Eco K12 chr1	39	0.67	118966,03	0.62	0.03	0.33
Eco O157H7 EDL933 chr1	38	0.74	144748,03	0.45	0.16	0.42
Hpy 26695 chr1	19	0.47	87763,53	0.42	0.05	0.26
Hpy J99 chr1	26	0.81	63224,27	0.58	0.15	0.35
Mth chr1	0	0.00	0	0.00	0.00	0.00
Mja chr1	98	0.82	16988,88	0.68	0.04	0.53
Mja chr2	2	1.00	29143,5	0.50	0.50	1.00

Tabella C.4 – continua nella prossima pagina

Tabella C.4 – continua dalla pagina precedente

cromosoma	**RDT**	**%**	**bp/RDT**	**RDTLE**	**RDTEN**	**RDTRA**
Mja chr3	1	1.00	16550	1.00	0.00	1.00
Mtu CDC1551 chr1	3	0.33	1467445,67	0.00	0.00	0.33
Mtu H37Rv chr1	3	0.33	1470510,67	0.33	0.00	0.33
Mge chr1	37	0.68	15677,68	0.43	0.16	0.49
Mpn chr1	7	0.43	116627,71	0.43	0.00	0.29
Nme MC58 chr1	11	0.82	206577,36	0.73	0.00	0.64
Ppr SS9 chr1	36	0.72	113475,67	0.64	0.08	0.33
Ppr SS9 chr2	6	0.83	372980,5	0.67	0.00	0.67
Ppr SS9 chr3	0	0.00	0	0.00	0.00	0.00
Pho chr1	10	0.80	173850,5	0.70	0.10	0.50
Rsp 2 4 1 chr1	0	0.00	0	0.00	0.00	0.00
Rsp 2 4 1 chr2	1	1.00	943016	1.00	0.00	0.00
Sty LT2 chr1	27	0.67	179904,89	0.67	0.00	0.33
Sty LT2 chr2	0	0.00	0	0.00	0.00	0.00
Sau COL chr1	114	0.75	24644,05	0.62	0.11	0.54
Sau MW2 chr1	114	0.77	24740,89	0.61	0.11	0.49
Sau Mu50 chr1	117	0.73	24602,81	0.56	0.10	0.53
Sau Mu50 chr2	7	0.86	3586,71	0.71	0.00	0.57
Sau N315 chr1	115	0.73	24476,66	0.57	0.10	0.52
Sau MRSA252 chr1	100	0.75	29026,19	0.54	0.14	0.56
Sau MSSA476 chr1	120	0.72	23331,68	0.53	0.11	0.46
Syn PCC6803 chr1	5	0.60	714694	0.60	0.20	0.40
Tma chr1	4	1.00	465181,25	0.00	0.50	1.00
Tpa chr1	2	1.00	568735,5	1.00	0.00	1.00
Uur chr1	271	0.54	2773,87	0.31	0.16	0.32
Vch chr1	11	0.91	269179	0.73	0.00	0.73
Vch chr2	1	1.00	1072255	1.00	0.00	1.00
Vpa chr1	12	1.00	274046,5	1.00	0.00	0.75
Vpa chr2	2	1.00	938606	1.00	0.00	0.50
Vvu CMCP6 chr1	12	0.67	273495,42	0.50	0.00	0.42
Vvu CMCP6 chr2	4	1.00	461213,25	0.75	0.25	0.00
Vvu YJ016 chr1	13	0.69	258038,85	0.69	0.00	0.31
Vv YJ016 chr2	3	1.00	619024,33	0.67	0.33	0.33
Vvu YJ016 chr3	0	0.00	0	0.00	0.00	0.00

Tabella C.4: Distribuzione delle matrici RDT come RDTLE, RDTEN e/o RDTRA rispetto ai geni. RDT, numero di matrici RDT osservate; %, percentuale di matrici RDT presenti nelle regioni dei geni. Ciascuna matrice RDT può essere contata sia come RDTLE, RDTEN e/o RDTRA simultaneamente.

cromosoma	**RDT**	**%**	**bp/RDT**	**RDTLE**	**RDTEN**	**RDTER**	**RDTRA**
Bsu chr1	1771	0.58	4068.18	0.47	0.33	0.03	0.58
Eco K12 chr1	985	0.59	7999.44	0.48	0.31	0.06	0.53

Tabella C.5: Distribuzione dei RDT putativi come RDTLE, RDTEN, RDTER e/o RDTRA rispetto agli operoni. RDT, numero di RDT putativi osservati; %, percentuale di RDT putativi presenti nelle regioni degli operoni. Ciascun RDT predetto può essere contato sia come RDTLE, RDTEN, RDTER e/o RDTRA simultaneamente.

cromosoma	**RDT**	**%**	**bp/RDT**	**RDTLE**	**RDTEN**	**RDTER**	**RDTRA**
Bsu chr1	17	0.94	0	0.82	0.06	0.00	0.65
Eco K12 chr1	39	0.67	0	0.59	0.05	0.08	0.28

Tabella C.6: Distribuzione delle matrici RDT come RDTLE, RDTEN, RDTER e/o RDTRA rispetto agli operoni. RDT, numero delle matrici RDT osservate; %, percentuale delle matrici RDT presenti nelle regioni degli operoni. Ciascuna matrice RDT può essere contata sia come RDTLE, RDTEN, RDTER e/o RDTRA simultaneamente.

elemento	**totale**	**coinvolti**	**RDTLE**	**RDTRA**	**RDTEN**
matrice	50	36	32	2	18
gene	4409	60	42	2	25
matrice	50	36	31	3	16
operone	3454	55	38	3	21

Tabella C.7: Distribuzione delle matrici RDT rispetto a geni e operoni in *Escherichia coli* K12. Ciascuna matrice RDT può essere contata sia come RDTLE, RDTEN e/o RDTRA simultaneamente.

Cog	J	A	K	L	B	D	Y	V	T	M	N	Z	W	U	O	C	G	E	F	H	I	P	Q	R	S	X
Gen	182	2	329	236	-	36	-	50	210	246	115	-	-	136	140	311	426	446	92	152	108	284	81	517	326	-
Mat	3	-	2	-	-	-	-	-	-	-	2	-	-	2	4	1	-	4	1	-	-	-	-	3	2	12
	0.09	-	0.06	-	-	-	-	-	-	-	0.06	-	-	0.06	0.12	0.03	-	0.12	0.03	-	-	-	-	0.09	0.06	0.35
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	1
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.50	-	-	-	-	-	-	-	-	0.50
	2	-	3	1	-	-	-	-	-	-	-	-	-	-	2	-	-	3	-	-	-	-	-	4	2	5
	0.09	-	0.14	0.05	-	-	-	-	-	-	-	-	-	-	0.09	-	-	0.14	-	-	-	-	-	0.18	0.09	0.23
Pro	96	-	56	21	-	5	-	21	83	80	58	-	-	55	31	135	193	230	53	117	50	92	26	80	27	312
	0.06	-	0.03	0.01	-	0.00	-	0.01	0.05	0.05	0.04	-	-	0.03	0.02	0.08	0.12	0.14	0.03	0.07	0.03	0.06	0.02	0.05	0.02	0.19
	10	-	14	7	-	1	-	4	11	16	8	-	-	6	5	10	17	25	3	8	1	12	4	19	7	63
	0.05	-	0.06	0.03	-	0.00	-	0.02	0.05	0.07	0.04	-	-	0.03	0.02	0.05	0.08	0.11	0.01	0.04	0.00	0.05	0.02	0.09	0.03	0.29
	25	-	23	16	-	4	-	4	7	13	9	-	-	10	7	14	31	17	4	10	7	17	5	31	20	87
	0.08	-	0.07	0.05	-	0.01	-	0.01	0.02	0.04	0.03	-	-	0.03	0.02	0.04	0.10	0.05	0.01	0.03	0.02	0.05	0.02	0.10	0.06	0.27
T0s	64	-	116	52	-	14	-	13	40	69	39	-	-	51	50	108	138	157	31	63	40	89	27	180	94	498
	0.04	-	0.07	0.03	-	0.01	-	0.01	0.02	0.04	0.02	-	-	0.03	0.03	0.06	0.08	0.09	0.02	0.04	0.02	0.05	0.02	0.11	0.06	0.29
	48	-	97	53	-	8	-	15	54	57	38	-	-	38	45	83	110	119	25	43	25	83	21	126	100	443
	0.03	-	0.07	0.04	-	0.01	-	0.01	0.04	0.04	0.03	-	-	0.03	0.03	0.06	0.08	0.08	0.02	0.03	0.02	0.06	0.01	0.09	0.07	0.30
T35s	71	-	115	60	-	14	-	11	44	79	39	-	-	51	55	117	140	158	29	66	36	94	31	187	107	527
	0.04	-	0.06	0.03	-	0.01	-	0.01	0.02	0.04	0.02	-	-	0.03	0.03	0.06	0.08	0.09	0.02	0.04	0.02	0.05	0.02	0.10	0.06	0.29
	36	-	87	47	-	9	-	12	44	58	29	-	-	27	32	84	94	102	21	39	20	68	20	117	83	407
	0.03	-	0.07	0.04	-	0.01	-	0.01	0.03	0.04	0.02	-	-	0.02	0.02	0.06	0.07	0.08	0.02	0.03	0.02	0.05	0.02	0.09	0.06	0.31
T25	32	-	11	3	-	-	-	3	22	17	17	-	-	18	4	28	46	49	15	23	10	20	8	8	1	-
	0.11	-	0.04	0.01	-	-	-	0.01	0.08	0.06	0.06	-	-	0.06	0.01	0.10	0.16	0.17	0.05	0.08	0.03	0.07	0.03	0.03	0.00	-
T35	2	-	-	1	-	-	-	-	-	1	-	-	-	1	-	1	2	1	-	2	2	1	1	-	-	-
	0.15	-	-	0.08	-	-	-	-	-	0.08	-	-	-	0.08	-	0.08	0.15	0.08	-	0.15	0.15	0.08	0.08	-	-	-
Tot	156	1	269	145	-	31	-	41	121	193	101	-	-	118	119	242	317	366	72	130	80	227	65	411	275	1262

Tabella C.8: Distribuzione dei geni nelle categorie funzionali in *E. coli*. Cog, nomi delle categorie funzionali. Gen, numero di geni codificanti presenti per ogni categoria. Mat, dati relativi ai risultati ottenuti allineando la matriceRDT al genoma; le tre righe corrispondono ai geni che presentano un RDT nel leader (prima riga), all'interno del gene (seconda riga) o nel terminatore (terza riga). Pro, dati relativi ai risultati ottenuti da RDTSCAN e organizzati come sopra. T0s, dati relativi agli esperimenti di microarray in cui gli oligo sono accesi (prima riga) o spenti (seconda riga) al tempo zero. T35s, come T0s ma per il tempo 35 minuti. T25, dati relativi ai geni differenzialmente espressi rispetto al tempo zero e al tempo 35 minuti. T35, come T25 ma per il tempo 35 minuti. Tot, numero di geni in cui sono stati disegnati degli oligo. I numeri interi sono relativi al numero di geni mentre i numeri decimali sono relativi alle frequenze tra il numero interno presentato e il numero totale di geni di quel esperimento.

INFORMATION STORAGE AND PROCESSING	
J	Translation, ribosomal structure and biogenesis
A	RNA processing and modification
K	Transcription
L	Replication, recombination and repair
B	Chromatin structure and dynamics
CELLULAR PROCESSES AND SIGNALING	
D	Cell cycle control, cell division, chromosome partitioning
Y	Nuclear structure
V	Defense mechanisms
T	Signal transduction mechanisms
M	Cell wall/membrane/envelope biogenesis
N	Cell motility
Z	Cytoskeleton
W	Extracellular structures
U	Intracellular trafficking, secretion, and vesicular transport
O	Posttranslational modification, protein turnover, chaperones
METABOLISM	
C	Energy production and conversion
G	Carbohydrate transport and metabolism
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
H	Coenzyme transport and metabolism
I	Lipid transport and metabolism
P	Inorganic ion transport and metabolism
Q	Secondary metabolites biosynthesis, transport and catabolism
POORLY CHARACTERIZED	
R	General function prediction only
S	Function unknown

Tabella C.9: Categorie funzionali dei Clusters of Orthologous Groups of proteins (COGs).

Cog	J	A	K	L	B	D	Y	V	T	M	N	Z	W	U	O	C	G	E	F	H	I	P	Q	R	S	X
Gen	182	2	329	236	-	36	-	50	210	246	115	-	-	136	140	311	426	446	92	152	108	284	81	517	326	-
Mat	3	-	2	-	-	-	-	-	-	-	2	-	-	2	4	1	-	4	1	-	-	-	-	3	2	12
	0.02	-	0.01	-	-	-	-	-	-	-	0.02	-	-	0.02	0.03	0.00	-	0.01	0.01	-	-	-	-	0.01	0.01	0.01
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	1
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.00	-	-	-	-	-	-	-	-	0.00
	2	-	3	1	-	-	-	-	-	-	-	-	-	-	2	-	-	3	-	-	-	-	-	4	2	5
	0.01	-	0.01	0.01	-	-	-	-	-	-	-	-	-	-	0.02	-	-	0.01	-	-	-	-	-	0.01	0.01	0.00
Pro	96	-	56	21	-	5	-	21	83	80	58	-	-	55	31	135	193	230	53	117	50	92	26	80	27	312
	0.62	-	0.21	0.14	-	0.16	-	0.51	0.69	0.41	0.57	-	-	0.47	0.26	0.56	0.61	0.63	0.74	0.90	0.62	0.41	0.40	0.19	0.10	0.25
	10	-	14	7	-	1	-	4	11	16	8	-	-	6	5	10	17	25	3	8	1	12	4	19	7	63
	0.06	-	0.05	0.05	-	0.03	-	0.10	0.09	0.08	0.08	-	-	0.05	0.04	0.04	0.05	0.07	0.04	0.06	0.01	0.05	0.06	0.05	0.03	0.05
	25	-	23	16	-	4	-	4	7	13	9	-	-	10	7	14	31	17	4	10	7	17	5	31	20	87
	0.16	-	0.09	0.11	-	0.13	-	0.10	0.06	0.07	0.09	-	-	0.08	0.06	0.06	0.10	0.05	0.06	0.08	0.09	0.07	0.08	0.08	0.07	0.07
T0s	64	-	116	52	-	14	-	13	40	69	39	-	-	51	50	108	138	157	31	63	40	89	27	180	94	498
	0.41	-	0.43	0.36	-	0.45	-	0.32	0.33	0.36	0.39	-	-	0.43	0.42	0.45	0.44	0.43	0.43	0.48	0.50	0.39	0.42	0.44	0.34	0.39
	48	-	97	53	-	8	-	15	54	57	38	-	-	38	45	83	110	119	25	43	25	83	21	126	100	443
	0.31	-	0.36	0.37	-	0.26	-	0.37	0.45	0.30	0.38	-	-	0.32	0.38	0.34	0.35	0.33	0.35	0.33	0.31	0.37	0.32	0.31	0.36	0.35
T25s	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	1	-	-	-	1
	-	-	0.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.00	-	-	-	0.00	-	-	-	0.00
T35s	71	-	115	60	-	14	-	11	44	79	39	-	-	51	55	117	140	158	29	66	36	94	31	187	107	527
	0.46	-	0.43	0.41	-	0.45	-	0.27	0.36	0.41	0.39	-	-	0.43	0.46	0.48	0.44	0.43	0.40	0.51	0.45	0.41	0.48	0.45	0.39	0.42
	36	-	87	47	-	9	-	12	44	58	29	-	-	27	32	84	94	102	21	39	20	68	20	117	83	407
	0.23	-	0.32	0.32	-	0.29	-	0.29	0.36	0.30	0.29	-	-	0.23	0.27	0.35	0.30	0.28	0.29	0.30	0.25	0.30	0.31	0.28	0.30	0.32
T25	32	-	11	3	-	-	-	3	22	17	17	-	-	18	4	28	46	49	15	23	10	20	8	8	1	-
	0.21	-	0.04	0.02	-	-	-	0.07	0.18	0.09	0.17	-	-	0.15	0.03	0.12	0.15	0.13	0.21	0.18	0.12	0.09	0.12	0.02	0.00	-
T35	2	-	-	1	-	-	-	-	-	1	-	-	-	1	-	1	2	1	-	2	2	1	1	-	-	-
	0.01	-	-	0.01	-	-	-	-	-	0.01	-	-	-	0.01	-	0.00	0.01	0.00	-	0.02	0.03	0.00	0.02	-	-	-
Tot	156	1	269	145	-	31	-	41	121	193	101	-	-	118	119	242	317	366	72	130	80	227	65	411	275	1262

Tabella C.10: Distribuzione dei geni nelle categorie funzionali in *E. coli*. Cog, nomi delle categorie funzionali. Gen, numero di geni codificanti presenti per ogni categoria. Mat, dati relativi ai risultati ottenuti allineando la matriceRDT al genoma; le tre righe corrispondono ai geni che presentano un RDT nel leader (prima riga), all'interno del gene (seconda riga) o nel terminatore (terza riga). Pro, dati relativi ai risultati ottenuti da RDTSCAN e organizzati come sopra. T0s, dati relativi agli esperimenti di microarray in cui gli oligo sono accesi (prima riga) o spenti (seconda riga) al tempo zero. T35s, come T0s ma per il tempo 35 minuti. T25, dati relativi ai geni differenzialmente espressi rispetto al tempo zero e al tempo 35 minuti. T35, come T25 ma per il tempo 35 minuti. Tot, numero di geni in cui sono stati disegnati degli oligo. I numeri interi sono relativi al numero di geni mentre i numeri decimali sono relativi alle frequenze tra il numero interno presentato e il numero totale di geni in cui sono stati disegnati degli oligo di quella determinata categoria funzionale.

Bibliografia

[1] Jaillon O, *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449: 463-467, 2007

[2] Itoh T, *et al.* Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol*, 16: 332-346, 1999

[3] Wolf I, *et al.* Genome alignment, evolution of prokaryotic genome organization and prediction of gene function using genomic context. *Genome Res*, 11: 356-372, 2001

[4] Wang L, *et al.* Genome-wide operon prediction in *Staphylococcus aureus*. *Nucleic Acids Res*, 32: 3689-3702, 2004

[5] Chen X, *et al.* Operon prediction by comparative genomics: an application to the Synechococcus sp. WH8102 genome. *Nucleic Acids Res*, 32: 2147-2157, 2004

[6] Salgado H, *et al.* Operons in Escherichia coli: genomic analyses and predictions. *Proc Natl Acad Sci USA*, 97: 6652-6657, 2000

[7] Jian-Cheng Lin,Li, *et al.* Prediction of prokaryotic promoters based on prediction of transcriptional units . *Acta Biochimica et Biophysica Sinica*, 35: 317-324, 2003

[8] Ermolaeva D, *et al.* Prediction of operons in microbial genomes. *Nucleic Acids Res*, 29: 1216-1221, 2001

[9] Romero PR and Karp PD, *et al.* Using functional and organizational information to im-

prove genome-wide computational prediction of transcription units on pathway-genome databases. *Bioinformatics*, 20: 709-717, 2004

[10] Paredes J, *et al.* Transcriptional organization of the Clostridium acetobutylicum genome. *Nucleic Acids Res*, 32: 1973-1981, 2004

[11] Moreno-Hagelsieb G and Collado-Vides J, *et al.* A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, 18: S329-S336, 2002

[12] Yellaboina S, *et al.* PredictRegulon: a web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes. *Nucleic Acids Res*, 32: W318-W320, 2004

[13] Alkema WB, *et al.* Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus. *Genome Res*, 14: 1362-1373, 2004

[14] Yada T, *et al.* Modeling and predicting transcriptional units od Escherichia coli genes using hidden Markov models. *Bioinformatics*, 15: 987-993, 1999

[15] Bockhorst J, *et al.* Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, 19: i34-i43, 2003

[16] Bockhorst J, *et al.* A Bayesian network approach to operon prediction. *Bioinformatics*, 19: 1227-1235, 2003

[17] Sabatti C, *et al.* Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res*, 30: 2886-2893, 2002

[18] Lauro FM, *et al.* Large-scale transposon mutagenesis of Photobacterium profundum SS9 reveals new genetic loci important for growth at low temperature and high pressure. *J Bacteriol*, [Epub ahead of print], 2007

[19] Campanaro S, *et al.* Laterally transferred elements and high pressure adaptation in Photobacterium profundum strains. *BMC Genomics*, 6: 122, 2005

[20] Vezzi A, *et al.* Life at depth: Photobacterium profundum genome sequence and expression analysis. *Science*, 307: 1459-1461, 2005

[21] Skordalahes E and Berger JM. Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading. *Cell*, 114: 135-146, 2003

[22] Graham JE *et al.* Sequence-specific Rho-RNA interactions in transcription termination. *Nucleic Acids Res*, 32: 3093-3100, 2004

[23] Chen CY and Richardson JP, *et al.* Sequence elements essential for rho-dependent trascription termination at lambda tR1. *J Biol Chem*, 262: 11292-11299, 1987

[24] Graham JE and Richardson JP. rut Sites in the nascent transcript mediate Rho-dependent transcription termination in vivo. *J Biol Chem*, 273: 20764-20769, 1998

[25] Hart CM and Roberts JW. Rho-dependent transcription termination. Characterization of the requirement for cytidine in the nascent transcript. *J Biol Chem*, 266: 24140-24148, 1991

[26] Walstrom KM, *et al.* Effects of reaction conditions on RNA secondary structure and on the helicase activity of Escherichia coli transcription termination factor Rho. *J Mol Biol*, 279: 713-726, 1998

[27] Wei RR and Richardson JP, *et al.* Identification of an RNA binding site in the ATP binding domain of Escherichia coli Rho by H2O2/Fe-EDTA cleavage protection studies. *J Biol Chem*, 276: 28380-28387, 2001

[28] Lau F, *et al.* RNA polymerase pausing and transcript release at the lambda tR1 terminator in vitro. *J Biol Chem*, 258: 9391-9397, 1983

[29] Morgan D, *et al.* Rho-dependent termination of transcription . *J Biol Chem*, 258: 9565-9574, 1983

[30] Hitchens TK, *et al.* Sequence-specific interactions in the RNA-binding domain of E. coli transcription termination factor Rho. *J Biol Chem*, 44: 33697-33703, 2006

[31] Geiselmann J, *et al.* Functional interactions of ligand cofactors with Escherichia coli transcription termination factor rho. II. Binding of RNA. *Protein Sci*, 1: 861-873, 1992

[32] Richardson JP, *et al.* Rho-dependent termination and ATPases in trascript termination. *BBA*, 1577: 251-260, 2002

[33] Nudler E. Transcription termination and anti-termination in E. coli. *Genes to cells*, 7: 755-768, 2002

[34] Sharp JA and Platt T. Rho-dependent termination and concomitant NTPase activity requires a specific, intact RNA region. *J Biol Chem*, 259: 2268-73, 1984

[35] Richardson LV and Richardson JP. Rho-dependent termination of transcription is governed primarily by the upstream rho utilization (rut) sequences of a terminator. *J Biol Chem*, 271: 21597-21603, 1996

[36] Morgan D, *et al.* Rho-dependent termination of transcription . *J Biol Chem*, 258: 9553-9564, 1983

[37] Arnving B, *et al.* A high-affinity interaction between NusA and the rrn nut site in Mycobacterium tuberculosis. *Proc Natl Acad Sci USA*, 101: 8325-8330, 2004

[38] Gong F and Yanofsky C, *et al.* Analysis of tryptophanase operon expression in vitro. *J Biol Chem*, 277: 17095-17100, 2002

[39] Kempsell KE, *et al.* The nucleotide sequence of the promoter, 16S rRNA and spacer region of the ribosomal RNA operon of Mycobacterium tuberculosis and comparison with Mycobacterium leprae precursor rRNA. *J Gen Microbiol*, 138: 1717-1727, 1992

[40] Kochlar S and Paulus H, *et al.* Lysine-induced premature transcription termination in the lysC operon of Bacillus subtilis . *Microbiology*, 142: 1635-1639, 1996

[41] Vincent KonanYanofsky Charles, *et al.* Rho-dependent transcription termination in the tna operon of Escherichia coli: roles of the boxA sequence and the rut site. *J Bacteriol*, 182: 3981-3988, 2000

[42] Seoh HK, *et al.* rRNA antitermination functions with heat shock promoters. *J Bacteriol*, 185: 6486-6489, 2003

[43] Todo K, *et al.* Comparative analysis of the four rRNA operons in Finegoldia magna ATCC29328. *Syst Appl Microbiol*, 27: 18-26, 2004

[44] Williamson RM and Oxender DL. Premature termination of in vivo transcription of a gene encoding a branched-chain amino acid transport protein in Escherichia coli. *J Bacteriol*, 174: 1777-1782, 1992

[45] Carlomagno S and Nappo A, *et al.* NusA modulates intragenic termination by different pathways. *Gene*, 308: 115-128, 2003

[46] Cheng C, *et al.* Transcription termination signals in the nin region of basteriophage lambda: identification of rho-dependent termination regions. *Genetics*, 140: 875-887, 1995

[47] Briani F, *et al.* A rho-dependent transcription termination site regulated by Bacteriophage P4 RNA imminity factor. *Virology*, 223: 57-67, 1996

[48] Burns CM and Richardson JP. NusG is required to overcome a kinetic limitation to Rho function at an intragenic terminator. *Proc Natl Acad Sci USA*, 92: 4738-4742, 1995

[49] Lavitola A, *et al.* Intracistronic transcription termination in polyltransferase gene (siaD) affects phase variation in Neisseria meningitidis. *Mol Microbiol*, 33: 119-127, 1999

[50] Nowatzke WL and Richardson JP. Characterization of an unusual Rho factor from the high G+C gram-positive bacterium Micrococcus luteus. *J Biol Chem*, 271: 742-747, 1996

[51] Brunel F and Pilaete MF, *et al.* Localisation and characterization of a new rho-dependent transcription terminator from bacteriophage T5. *Nucleic Acids Res*, 13: 7687-7701, 1985

[52] Joyce SA and Dorman CJ. A rho-dependent phase-variable transcription terminator controls expression of the FimE recombinase in Escherichia coli. *Mol Microbiol*, 45: 1107-1117, 2002

[53] Vadeboncoeur C, *et al.* Regulation of the pts operon in low G+C G ram-positive bacteria. *J Mol Microbiol Biotechnol*, 2: 483-490, 2000

[54] Albrechtsen B, *et al.* Transcriptional termination sequence at the end of the Escherichia coli ribosomal RNA G operon: complex terminators and antitermination. *Nucleic Acids Res*, 19: 1845-1852, 1991

[55] La Farina M, *et al.* Readthrough transcription occurs at the rho dependent signal F1 TIV in suppressor cells. *Nucleic Acids Res*, 18: 865-870, 1990

[56] Thoma R, *et al.* A histidine gene cluster of the hyperthermophile Themotoga maritima: sequence analysis and evolutionary significance. *Extremophiles*, 2: 379-389, 1998

[57] Wu M, *et al.* Tandem termination sites in the tryptophan operon of Escherichia coli. *Proc Natl Acad Sci USA*, 78: 2913-2917, 1981

[58] Gilbert W. The RNA world. *Nature*, 319: 618, 1986

[59] Gesteland RF, *et al.* The RNA world. Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press, 1999

[60] Argaman L, *et al.* Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Curr Biol*, 11: 941-950, 2001

[61] Huttenhofer A, *et al.* RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J*, 20: 2943-2953, 2001

[62] Olivas WM, *et al.* Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs. *Nucleic Acids Res*, 25: 4619-4625, 1997

[63] Rivas E, *et al.* Computational identification of noncoding RNAs in E. coli by comparative genomics. *Curr Biol*, 11: 1369-1373, 2001

[64] Wassarman KM, *et al.* Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev*, 15: 1637-1651, 2001

[65] Lee RC, *et al.* The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-4. *Cell*, 75: 843-854, 1993

[66] Wightman B, *et al.* Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75: 855-862, 1993

[67] Llave C, *et al.* Cleavage of scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science*, 297: 2053-2056, 2002

[68] Reinhart BJ, *et al.* MicroRNAs in plants. *Genes Dev*, 16: 1616-1626, 2002. Erratum in: *Genes Dev*, 16: 2313, 2002

[69] Park MY, *et al.* Nuclear processing and export of microRNAs in Arabidopsis. *Proc Natl Acad Sci U S A*, 102: 3691-3696, 2005

[70] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116: 281-287, 2004

[71] Saumet A and Lecellier CH. Anti-viral RNA silencing: do we look like plants? *Retrovirology*, 3: doi:10.1186/1742-4690-3-3, 2006

[72] Lee Y, *et al.* MicroRNA genes are transcribed by RNA polymerase II. *EMBO J*, 23: 4051-4060, 2004

[73] Kurihara Y and Watanabe Y. Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc Natl Acad Sci U S A*, 101: 12753-12758, 2004

[74] Tang G, *et al.* A biochemical framework for RNA silencing in plants. *Genes Dev*, 17: 49-63, 2003

[75] Bonnet E, *et al.* The small RNA world of plants. *New Phytologist*, 171: 451-468, 2006

[76] Yi R, *et al.* Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev*, 17: 3011-3016, 2003

[77] Lund E, *et al.* Nuclear export of microRNA precursors. *Science*, 303: 95-98, 2004

[78] Zeng Y and Cullen BR. Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. *Nucleic Acids Res*, 32: 4776-4785, 2004

[79] Bernstein E, *et al.* Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409: 363-366, 2001

[80] Papp I, *et al.* Evidence for nuclear processing of plant microRNA and short interfering RNA precursor. *Plant Physiol*, 132: 1382-1390, 2003

[81] Bonnet E, *et al.* Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20: 2911-2917, 2004

[82] Weber MJ. New human and mouse microRNA genes found by homology search. *FEBS J*, 272: 59-73, 2005

[83] Berezikov E, *et al.* Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, 120: 21-24, 2005

[84] Altuvia Y, *et al.* Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res*, 33: 2697-2706, 2005

[85] Floyd SK, Bowman JL. Gene regulation: ancient microRNA target sequences in plants. *Nature*, 428: 485-486, 2004

[86] Axtell MJ and Bartel DP. Antiquity of microRNAs and their targets in land plants. *Plant Cell*, 17: 1658-1673, 2005

[87] Zhang BH, *et al.* Identification and characterization of new microRNAs using EST analysis. *Cell Res*, 15: 336-360, 2005

[88] Jones-Rhoades MW, *et al.* MicroRNA and their regulatory roles in plants. *Annu Rev Plant Biol*, 57: 19-53, 2006

[89] Lindow M and Gorodkin J. Principles and limitations of computational microRNA gene and target finding. *DNA Cell Biol*, 26: 339-351, 2007

[90] Lu C, *et al.* Elucidation of the smallRNA component of the trascriptome. *Science*, 309: 1567-1569, 2005

[91] Brown JR and Sanseau P. A computational view of microRNAs and their targets. *Drug Discov Today*, 10: 595-601, 2005

[92] Meyers BC, *et al.* Sweating the small stuff: microRNA discovery in plants. *Curr Opin Biotechnol*, 17: 139-146, 2006

[93] Hamilton AJ and Baulcombe DC. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, 286: 950-952, 1999

[94] Brenner S, *et al.* Gene expression analysis by Massively Parallel Signature Sequencing (MPSS) on microbead arrays. *Nat Biotechnol*, 18: 630-634, 2000

[95] Margulies M, *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 437: 376-380, 2005

[96] Solexa, Inc. Protocol for Whole Genome Sequencing using Solexa Technology. BioTechniques Protocol Guide: 37, 2006

[97] Jacob F and Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3: 318-356, 1961

[98] Britten RJ and Davidson EH. Gene regulation for higher cells: a theory. *Science*, 165: 349-357, 1969

[99] Sito ufficiale del progetto GMOD: *http://www.gmod.org/*

[100] Sito ufficiale del NIH: *http://www.nih.gov/*

[101] Sito ufficiale dell'agenzia ARS: *http://www.ars.usda.gov/*

[102] Stein LD, *et al.* WormBase: network access to the genome and biology of Caenorhabditis elegans. *Nucleic Acids Res*, 1: 82-86, 2001

[103] Rogers A, *et al.* WormBase 2007. *Nucleic Acids Res*, doi:10.1093/nar/gkm975, 2007

[104] Ashburner M and Drysdale R. FlyBase: the Drosophila genetic database. *Development*, 7: 2077-2079, 1994

[105] Wilson RJ, *et al.* FlyBase: integration and improvements to query tools. *Nucleic Acids Res*, doi:10.1093/nar/gkm930, 2007

[106] Zhu Y, *et al.* Integrating computationally assembled mouse transcript sequences with the Mouse Genome Informatics (MGI) database. *Genome Biol*, 2: R16, 2003

[107] Ware D, *et al.* Gramene: a resource for comparative grass genomics. *Nucleic Acids Res*, 1: 103-105, 2002

[108] Liang C, *et al.* Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res*, doi:10.1093/nar/gkm968, 2007

[109] Twigger SN, *et al.* Rat Genome Database (RGD): mapping disease onto the genome. *Nucleic Acids Res*, 1: 125-128, 2002

[110] Twigger SN, *et al.* The Rat Genome Database, update 2007: easing the path from disease to data and back again. *Nucleic Acids Res*, 35: D658-662, 2007

[111] Huala E, *et al.* The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res*, 1: 102-105, 2001

[112] Swarbreck D, *et al.* The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*, doi:10.1093/nar/gkm965, 2007

[113] Karp PD, *et al.* EcoCyc: an encyclopedia of Escherichia coli genes and metabolism. *Nucleic Acids Res*, 1: 32-29, 1996

[114] Keseler IM, *et al.* EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res*, 33: D334-337, 2005

[115] Cherry JM, *et al.* SGD: Saccharomyces Genome Database. *Nucleic Acids Res*, 1: 73-79, 1998

[116] Nash R, *et al.* Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res*, 35: D468-471, 2007

[117] Sito del CRIBI genomics: *http://genomics.cribi.unipd.it/*

[118] Sito ufficiale del EMBL-EBI: *http://www.ebi.ac.uk/*

[119] Sito ufficiale del JGI: *http://www.jgi.doe.gov/*

[120] Sito ufficiale del NCBI: *http://www.ncbi.nlm.nih.gov/*

[121] Sito ufficiale del Sanger Center: *http://www.sanger.ac.uk/*

[122] Sito ufficiale del TIGR: *http://www.tigr.org/*

[123] Gasteiger E, *et al.* SWISS-PROT: connecting biological knowledge via a protein database. *Curr Issues Mol Biol*, 3: 47-55, 2001

[124] Camon E., *et al.* The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 32: D262-D266, 2004

[125] Jensen KF. The Escherichia coli K12 'Wild Types' W3110 and MG1655 have an rph frameshift mutation that leads to pyrimidine starvation due to low pyrE espression levels. *J Bacteriol*, 175: 3401-3407, 1993

[126] Yanisch-Perron C, *et al.* Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene*, 33: 103-119, 1985

[127] Nittler MP, *et al.* Identification of Histoplasma capsulatum transcripts induced in response to reactive nitrogen species. *Mol Biol Cell*, 16: 4792-4813, 2005

[128] Wells CA, *et al.* Alternate transcription of the Toll-like receptor signaling cascade. *Genome Biol*, 7: R10, 2006

[129] Ghindilis AL, *et al.* CombiMatrix oligonucleotide arrays: Genotyping and gene expression assays employing electrochemical detection. *Biosens Bioelectron*, 22: 1853-1860, 2007

[130] Altschul SF, *et al.* Basic local alignment search tool. *J Mol Biol* 215: 403-410, 1990

[131] Florea L, *et al.* A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence. *Genome Res*, 8: 967-974, 1998

[132] Hertz GZ and Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15: 563-577, 1999

[133] Gautheret D and Lambert A, *et al.* Direct RNA definition and identification from multiple sequence alignments using secondary structure profiles. *J Mol Biol*, 313: 1003-1011, 2001

[134] Zuker M and Stiegler P. Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. *Nucl Acid Res*, 9: 133-148, 1981

[135] Stein LD, *et al.* The generic genome browser: a building block for a model organism system database. *Genome Res*, 12: 1599-1610, 2002

[136] Sito ufficiale dell'ontologia delle sequenze: *http://www.sequenceontology.org/*

[137] Sito del tutorial ufficiale della sintassi GFF3: *http://www.sequenceontology.org/gff3.shtml*

[138] Durinck S, *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21: 3439-3440, 2005

[139] Schwab R, *et al.* Highly specific gene silencing by artificial microRNAs in Arabidopsis. *Plant Cell*, 18: 1121-1133, 2006

[140] Adai A, *et al.* Computational prediction of miRNAs in Arabidopsis thaliana. *Genome Res*, 15: 78-91, 2005 identification of C. elegans microRNAs. *Mol Cell*, 11: 1253-1263, 2003

[141] Dezulian T, *et al.* Identification of plant microRNA homologs. *Bioinformatics*, 22: 359-360, 2006

[142] Enright AJ, *et al.* MicroRNA targets in Drosophila. *Genome Biology*, 5: R1, 2003

[143] Rehmsmeier M, *et al.* Fast and effective prediction of microRNA/target duplexes. *RNA*, 10: 1507-1517, 2004

[144] Grad Y, *et al.* Computational and experimental

[145] Lewis BP, *et al.* Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120: 15-20, 2005

[146] Notredame C, *et al.* T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302: 205-217, 2000

[147] Sito ufficiale del programma RepeatMasker: *http://www.repeatmasker.org/*

[148] Blanchette M, *et al.* Aligning multiple genomic sequences with the Threaded Blockset Aligner. *Genome Res*, 14: 708-715, 2004

[149] Sito ufficiale della Combimatrix: *http://www.combimatrix.com/*

[150] Saeed AI, *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, 34: 374-378, 2003

[151] Quackenbush J. Microarray data normalization and transformation. *Nature Genetics*, 32: 496-501, 2002

[152] Yang IV *et al.* Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol*, 3, research0062.1-0062.12, 2002

[153] Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *J Amer Stat Assoc*, 74: 829-836, 1979

[154] Tusher VG, *et al.* Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98: 5116-5121, 2001. Erratum in: *Proc Natl Acad Sci USA* 98: 10515, 2001

[155] Zeeberg BR, *et al.* GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4: R28, doi:10.1186/gb-2003-4-4-r28, 2003

[156] Ashburner M, *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25: 25-29, 2000

[157] Mathews DH, *et al.* Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, 288: 911-40, 1999

[158] Cherry JL. Genome size and operon content. *J theor Biol*, 221: 401-410, 2003

[159] Eom CY, *et al.* Cloning, molecular characterization, and transcriptional analysis of dnaK operon in a methylotrophic bacterium Methylovorus sp. strain SS1 DSM 11726. *Mol Cells*, 14: 245-254, 2002

[160] Robinson K, *et al.* A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. *J Mol Biol.* 284: 241-254, 1998

[161] Zhang R and Zhang CT. A systematic method to identify genomic islands and its applications in analyzing the genomes of Corynebacterium glutamicum and Vibrio vulnificus CMCP6 chromosome I. *Bioinformatics*, 20: 612-622, 2004

[162] Hinde P, *et al.* Characterization of the Detachable Rho-Dependent Transcription Terminator of fimE Gene in Escherichia coli K-12. *J Bacteriol*, 187: 8256-8266, 2005

[163] Someya A, *et al.* Morphological changes of Escherichia coli induced by bicyclomycin. *Antimicrob Agents Chemother*, 16: 87-91, 1978

[164] Tanaka N, *et al.* Mechanism of action of bicyclomycin. *J Antibiot*, 29: 155-168, 1976

[165] Someya A, *et al.* Binding of bicyclomycin to inner membrane proteins of E. coli. *J Antibiot*, 31: 712-718, 1978

[166] Konan KV and Yanofsky C. Regulation of Escherichia coli tna operon: nascent leader peptide control at the tnaC stop codon. *J Bacteriol*, 179: 1774-1779, 1997

[167] Nakaya A, *et al.* Tracing synergetic behavior of the QTLs affecting oral glucose tolerance in the OLETF rat. *Genome Inform Ser Workshop Genome Inform*, 10: 155-165, 1999

[168] Huynh T, *et al.* A pattern-based method for the identification of microRNA-target sites and their corresponding RNA/RNA complexes. *Cell*, 126: 1203-1217, 2006

[169] Sito del CRIBI GBbrowse: *http://gbrowse.cribi.unipd.it/*