Università degli Studi di Catania
Facoltà di Scienze Matematiche, Fisiche e Naturali
Dottorato di Ricerca in
Informatica
XXII Ciclo – 2006/2009

Aurelio Giudice

Basi di conoscenza e tecniche di Data Mining con applicazioni agli RNA non codificanti

Tesi di Dottorato

Coordinatore:
Chiar.mo Prof. Domenico Cantone _____
Tutor:
Chiar.mo Prof. Alfredo Ferro _____

Sommario

1 INTRODUZIONE

Il lavoro svolto nell'ambito del dottorato di ricerca ha avuto come obiettivo principale quello di sviluppare un sistema web based per la gestione avanzata di dati biologici. Tale sistema, denominato miRò (miR-Ontology), integra tutte quelle informazioni provenienti da varie fonti biologiche riguardanti il processo di silenziamento post-trascrizionale dell'espressione genica (PTGS). In particolare, il sistema si occupa di mettere in evidenza le associazioni tra i microRNA (molecole chiave nel PTGS), le funzioni, i processi e le patologie in cui essi sono potenzialmente coinvolti attraverso i loro geni target. Il sistema nel tempo ha subito varie modifiche al fine di migliorare i criteri e le modalità di ricerca per poter rispondere in maniera efficace ed efficiente alle numerose e varie richieste di natura biologica. In sintesi, è possibile identificare due edizioni del software dove nella prima miRò ha permesso di inferire le associazioni fra i microRNA, i processi e le funzioni biologiche, unite alle patologie che li coinvolgono attraverso le annotazioni funzionali dei geni da essi regolati; nella seconda edizione sono state introdotte informazioni varie su pathway, profili di espressione ed annotazioni genomiche inerenti i microRNA ed i loro target (siti fragili, CpG island, translocation breakpoint), indispensabili per cercare di capire la natura delle associazioni. Contestualmente, per filtrare opportunamente le associazioni trovate dal sistema tra miRNA e fenotipo, è stato realizzato un sistema di annotazione con ontologie mediante text mining dei dati di letteratura biologica pubblicati su Pubmed, in modo da creare una funzione di scoring per le associazioni. In questo modo è stato possibile escludere alcuni dei falsi positivi che il modello trovava, secondo le regole di inferenza definite in esso. Una parte rilevante del lavoro di tesi è consistita nella progettazione e sviluppo di un nuovo tool, *BioXML-Builder*, un'applicazione web che è nata come supporto all'importazione dei dati in miRò, ma che ha le potenzialità per contribuire alla standardizzazione dei dati biologici, avvalendosi di sofisticati parser che producono file in formato XML a partire dai vari formati utilizzati dalle varie fonti biologiche (.gff, .cvs, .xls, .sql, .txt, …). Il tool consente di tradurre i diversi formati di dati biologici presenti su web in un

formato standard comune di tipo XML. Questo lavoro sarà oggetto di un articolo che verrà sottomesso ad una rivista internazionale.

Infine, nella parte conclusiva del dottorato, ci si è occupati dello sviluppo di una base di conoscenza riguardante i piRNA (Piwi-associated RNA), molecole di RNA di recentissima scoperta per le quali è ipotizzata la funzione di regolazione dei trasposoni e la cui localizzazione genomica sarà oggetto di analisi di tipo computazionale e statistico. Questo costituisce un settore altamente avanzato, la cui complessità deriva dall'alto numero di piRNA identificati (oltre 500.000).

1.1 I miRNA

I microRNA (miRNA) sono piccole molecole di RNA regolatore a singolo filamento, di circa 20-22 nucleotidi, in grado di modulare l'espressione genica attraverso la degradazione o la repressione traduzionale di specifiche molecole target. E' stimato che i geni codificanti miRNA siano l'1% dei geni totali, formando la classe più ampia di molecole regolatrici.

1.1.1 La biogenesi dei miRNA

I miRNA sono presenti nelle piante, negli eucarioti superiori ed in alcuni virus, e sono codificati da diversi tipi di geni. La trascrizione dei miRNA è tipicamente eseguita dalla RNA polimerasi II, e i trascritti sono soggetti al *5'-capping* e alla poliadenilazione [24].

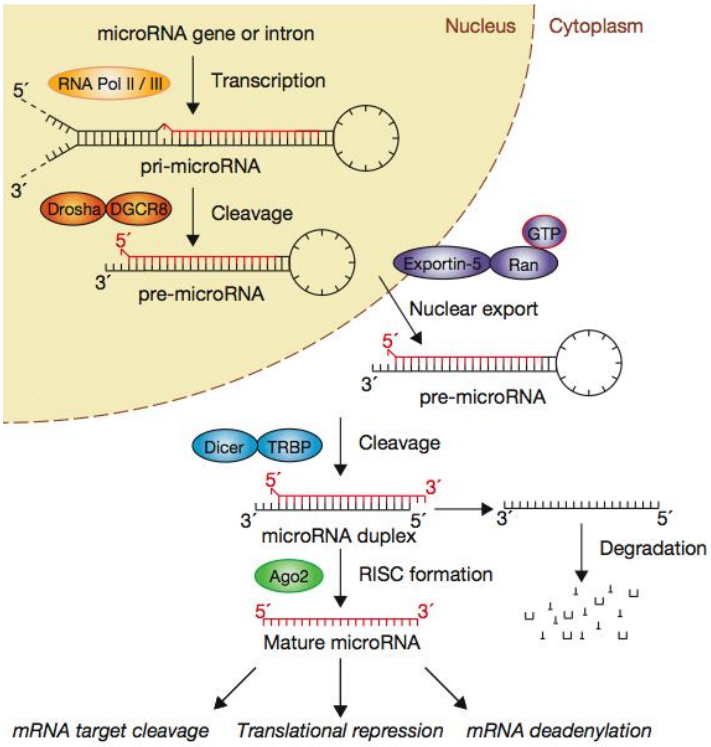


Fig. 1.1– La pathway della biogenesi dei miRNA.

Sebbene alcuni miRNA animali siano individualmente prodotti da unità trascrizionali separate, la maggior parte di essi sono generati da regioni codificanti gruppi di miRNA. Un trascritto può infatti codificare *cluster* di diversi miRNA oppure un miRNA e una proteina. In quest'ultimo caso si tratta di geni che

contengono la sequenza del miRNA all'interno di un introne o, più raramente, di un esone [1].

Il miRNA maturo è ottenuto dal trascritto primario, o pri-miRNA, attraverso due reazioni consecutive. Un tipico pri-miRNA animale consiste di uno stem di circa 33 bp, non perfettamente appaiate, con un loop terminale e dei segmenti fiancheggianti [1]. Il primo step della biogenesi del miRNA avviene nel nucleo e consiste nell'escissione dello stem loop dal resto del trascritto per dar luogo a quello che viene definito pre-miRNA. Per la maggior parte dei pri-miRNA, questa reazione di taglio è effettuata da un membro nucleare della famiglia delle RNasi III (Dcl1 nelle piante, Drosha negli animali) [24]. Sebbene Drosha catalizzi il processamento dei pri-miRNA, esso dipende anche da un cofattore, chiamato DGCR8, che contiene due domini dsRBD e si associa in maniera stabile con la ribonucleasi a formare il complesso detto Microprocessore[25]. DGCR8 interagisce direttamente con lo stem del pri-miRNA e con le sequenze fiancheggianti a singolo filamento. Infatti, questi segmenti fiancheggianti sono determinanti per il processamento, dato che il sito dove avviene il taglio è determinato dalla distanza dalla giunzione stem-sequenza fiancheggiante. Questo tipo di processamento comunque, non è l'unica maniera di produrre pre-miRNA negli animali. Una pathway alternativa utilizza lo splicing di trascritti pri-miRNA intronici, detti Mirtron; queste molecole entrano nella pathway di processamento dei miRNA senza l'intervento del Microprocessore. Si tratta di miRNA noncomuni, presenti però in tutti gli animali [26, 27].

Il secondo step del processamento comprende l'escissione del loop dallo stem del pre-miRNA in modo da creare il cosiddetto duplex miRNA maturo, di circa 22 nucleotidi. Nelle piante, Dcl1 conduce questa reazione nel nucleo. Negli animali, il pre-miRNA viene prima esportato nel citoplasma e successivamente processato dall'enzima Dicer canonico che effettua il taglio. Così come per i siRNA, il dominio PAZ del Dicer interagisce con l'estremità sporgente 3' del dsRNA, mentre il taglio viene effettuato dai siti catalitici RNasi III [1, 24].

La regolazione della biogenesi dei miRNA è chiaramente un meccanismo importante ma non ancora studiato estensivamente. Ad ogni modo, è emersa una tendenza significativa: un sorprendente numero di geni miRNA sono formati sotto il controllo dei molti target che regolano. Ad esempio, la trascrizione del gene miR-

7 in *Drosophila* è repressa da un fattore di trascrizione chiamato Yan, la cui traduzione è a sua volta repressa da miR-7, dando luogo ad un *feedback-loop* negativo [28]. Un altro esempio si ha in *C. elegans*, dove il miRNA let-7 inibisce la traduzione di Lin28 che a sua volta inibisce la trascrizione di let-7 [29].

La logica alla base di queste relazioni regolatorie è definita dalla capacità di regolazione della biogenesi dei miRNA. L'espressione errata dei miRNA mima frequentemente il fenotipo da *perdita di funzione* dei loro target. Questo può essere prevenuto se l'espressione dei miRNA è strettamente controllata dagli stessi target.

I duplex miRNA maturi sono entità che hanno vita breve, in quanto vengono rapidamente svolti quando si associano alla proteina Ago. Così come avviene per i siRNA, lo svolgimento dei miRNA è accompagnato dalla selezione differenziale dei filamenti; un filamento viene ritenuto mentre l'altro viene rilasciato, e la scelta del filamento guida è basata sulla stabilità termodinamica delle estremità del duplex. Il terminale 5' del filamento ritenuto è quello all'estremità meno stabile del duplex [24]. Questa comunque non è una regola assoluta. Il filamento passeggero è difatti ritrovato anch'esso in maniera apprezzabile nei complessi Ago. Sebbene ciascuno dei due filamenti può associarsi stabilmente alle proteine Ago, quello ritrovato più comunemente viene chiamato il filamento miRNA, mentre l'altro viene chiamato il filamento miRNA* [30].

1.1.2 Silenziamento post-trascrizionale operato dai miRNA

I miRNA agiscono come adattatori per i complessi miRISC permettendogli di riconoscere specificamente determinati mRNA target. Con poche eccezioni, i siti di legame dei miRNA negli mRNA animali si trovano nei 3' UTR e solitamente sono presenti in copie multiple. La maggior parte dei miRNA animali si legano ai loro target con complementarietà imperfetta, formando *bulges* e *loop*, sebbene una caratteristica chiave del riconoscimento del target coinvolga l'appaiamento perfetto dei nucleotidi 2-8 del miRNA, che rappresentano la regione *seed*. Al contrario, nella maggior parte delle piante i miRNA si legano con complementarietà quasi perfetta a siti specifici presenti nella regione codificante dei target [1].

Il grado di complementarietà miRNA/target è considerato un fattore chiave del meccanismo regolatorio. La complementarietà perfetta permette il taglio del filamento dell'mRNA catalizzato da Ago, mentre i mismatch centrali del duplex miRNA/mRNA escludono il taglio e promuovono la repressione della traduzione. Tuttavia, esperimenti suggeriscono che il meccanismo predefinito di silenziamento dell'espressione genica operata dai miRNA, sia negli animali che nelle piante, sia la repressione traduzionale, e che la complementarietà perfetta sia un fatto aggiuntivo che può condurre al taglio dell'mRNA, in modo che l'effetto finale sia il risultato di entrambe i meccanismi.

Diversi studi sui miRNA animali indicano che la repressione della traduzione non è accompagnata dalla destabilizzazione dell'mRNA. Tuttavia, per alcune interazioni miRNA-target vi è una riduzione significativa della concentrazione dell'mRNA dovuta ad un incremento della degradazione [31, 32]. Questa degradazione non è causata dall'attività catalitica di Ago ma piuttosto da deadenilazione, *decapping* e digestione esonucleolitica dell'mRNA [32, 33, 34]. Al momento non è chiaro il motivo per cui alcuni target vengono degradati ed altri no. E' stato ipotizzato che il numero, il tipo e la posizione dei mismatch nel duplex miRNA/mRNA giochi un ruolo importante nel determinare la degradazione o l'arresto della traduzione [35].

1.2 Il ruolo dei miRNA nello sviluppo

Le funzioni biologiche dei miRNA sono oggetto di intensi studi, mirati a verificarne il coinvolgimento nei vari processi cellulari. E' ormai stabilito il ruolo cruciale dei miRNA nell'apoptosi, nella proliferazione cellulare, nella resistenza allo stress, nel metabolismo e nella difesa dell'organismo da parte di agenti patogeni. I miRNA giocano inoltre un ruolo essenziale nello sviluppo. Modelli di topo *knock-out* per il Dicer forniscono evidenze significative del ruolo specifico dei miRNA nello sviluppo dei mammiferi. Questi topi, infatti, non sopravvivono oltre il settimo giorno dopo la gastrulazione e mancano di cellule staminali pluripotenti [36]. La rimozione condizionale del Dicer solo su certi tessuti ed organi consente inoltre di valutare il ruolo dei miRNA in contesti specifici. Questo approccio sperimentale ha dimostrato il ruolo fondamentale del Dicer nella morfogenesi di parecchi organi, inclusi i polmoni, gli arti ed i muscoli, e nella differenziazione delle cellule T [37,

38, 39, 40]. Naturalmente, questi esperimenti vanno interpretati sotto l'assunzione che il Dicer non svolga altri ruoli importanti al di fuori del processamento dei miRNA e dei siRNA. Tuttavia l'analisi dell'espressione dei miRNA, supporta le ipotesi formulate.

Attualmente, si sta iniziando a comprendere il ruolo dei singoli miRNA nello sviluppo e nel differenziamento di vertebrati ed invertebrati. Esperimenti mirati hanno permesso, per esempio, di stabilire l'importanza del cluster miR-17-92, la cui rimozione nel topo ne provoca la morte a poche ore della nascita, a causa dello sviluppo incompleto dei polmoni e di un difetto cardiaco. Gli esperimenti mostrano inoltre il ruolo essenziale di tale cluster nella regolazione della proteina pro-apoptotica Bim, correlata allo sviluppo delle cellule B [41].

Altri esperimenti attestano il ruolo essenziale dei miRNA nella proliferazione e nel differenziamento delle cellule staminali. Ad esempio, l'espressione di miR-520h è correlata al differenziamento delle cellule staminali ematopoietiche [42], mentre miR-150 può dirigere il differenziamento dei megacariociti, le cellule del midollo osseo responsabili della produzione delle piastrine [43]. L'espressione specifica di miR-1, miR-133 e miR-206 nei muscoli ne suggerisce il coinvolgimento nella miogenesi [44, 45, 46], processo per il quale è stato dimostrato il ruolo cruciale di miR-26a, in grado di regolare la proteina Ezh2, un soppressore della differenziazione delle cellule del muscolo scheletrico [47].

Sta emergendo anche il ruolo dei miRNA come interruttori di pathway regolatorie, come ad esempio i meccanismi di splicing alternativo, che possono contribuire alla tessuto-specificità. Ad esempio, il miRNA muscolo-specifico miR-133 è in grado di silenziare una proteina regolatrice dello splicing alternativo durante la differenziazione dei mioblasti, per controllare lo splicing di certe combinazioni di esoni [48]. Analogamente, è stato dimostrato il ruolo di miR-124 nello sviluppo del sistema nervoso attraverso la regolazione dello splicing alternativo neurone-specifico [49].

Sebbene i meccanismi regolatori dei miRNA non siano stati ancora completamente elucidati, appare comunque evidente l'importanza di tali molecole nello sviluppo normale di molti organi, incluso il cuore, ed il loro impatto in molte patologie, dalle infezioni al cancro.

1.3 I miRNA e le malattie

Negli ultimi anni si sono moltiplicati gli sforzi orientati allo studio delle alterazioni nell'espressione dei miRNA in molte malattie. Evidenze recenti suggeriscono un potenziale coinvolgimento dei miRNA nella neurodegenerazione. E' stato dimostrata, ad esempio, una sotto-espressione significativa di miR-107 nei pazienti affetti dal morbo di Alzheimer. Tale miRNA potrebbe essere coinvolto nella progressione della malattia, attraverso la regolazione di BACE1, un enzima che taglia la proteina precursore mieloide, generando un peptide amiloide neurotossico. La perdita di miR-107 porta dunque ad un incremento del livello di BACE1, come dimostrano le predizioni bioinformatiche e gli esperimenti di laboratorio. Tale disregolazione potrebbe essere uno dei meccanismi responsabili della patogenesi dell'Alzhaimer [50].

I miRNA potrebbero avere un ruolo rilevante anche nel morbo di Parkinson. Uno studio recente ha indagato il loro ruolo nei neuroni dopaminergici nei mammiferi, identificando miR-133b come miRNA specifico di tali neuroni, ma sottoespresso o addirittura assente in tali cellule dei malati di Parkinson [51]. Questo miRNA regola la maturazione e la funzione dei neuroni dopaminergici attraverso la sotto-regolazione dell'espressione del fattore di trascrizione Pitx3 [51]. Inoltre, diversi studi nei mammiferi e negli invertebrati suggeriscono il coinvolgimento dei miRNA nella neuro-protezione, nella sindrome dell'X fragile e nella schizofrenia [52, 53, 54]. Tutte queste osservazioni indicano che i processi neurodegenerativi potrebbero essere il risultato dell'alterazione di diverse pathway cellulari, nelle quali i miRNA possono giocare un ruolo significativo.

Diversi studi hanno dimostrato un insospettabile ruolo dei miRNA nel controllo dei diversi aspetti della funzione e disfunzione epatica. miR-122, ad esempio, è il miRNA più altamente espresso nel fegato, dove controlla la risposta allo stress regolando il gene CAT-1 [55]. Un recente studio ha consentito la validazione di target di miRNA cellulari nel genoma del virus HCV. Attraverso un'analisi di microarray, effettuata dopo il trattamento con interferone di linee cellulari di epatoma infette da HCV, si è dimostrato che gli interferoni αeβ sono in grado di inibire la replicazione e l'infezione di HCV, sovramodulando l'espressione di numerosi miRNA cellulari. Nello specifico, otto dei miRNA indotti da IFN-β(miR-1, miR-30, miR-128, miR-196, miR-296, miR-351, miR-431 e miR-448) mostrano

complementarietà quasi perfetta nei loro *seed* verso sequenze di RNA virale, e la sovra espressione di questi miRNA riproduce gli effetti antivirali di IFN-β, mentre la loro soppressione ne riduce gli effetti. Esperimenti mostrano l'interazione diretta di miR-196 e miR-448 con l'RNA di HCV. Si deduce quindi che i mammiferi, attraverso gli interferoni, utilizzano i miRNA per la difesa dalle infezioni virali [56].

E' stato dimostrato il coinvolgimento dei miRNA anche nei disordini muscolari primari, che comprendono diverse malattie, incluse la distrofia muscolare e le miopatie infiammatorie e congenite. Uno studio mostra alterazioni nell'espressione di 185 miRNA nella distrofia muscolare di Duchenne, nella miopatia di Miyoshi e nella dermatomiosite. Cinque di essi, miR-146b, miR-221, miR-155, miR-214 e miR-222, sono consistentemente sovra espressi in quasi tutti i campioni analizzati, suggerendo un loro possibile coinvolgimento in una pathway regolatoria comune [57].

I miRNA svolgono un ruolo cruciale anche nelle malattie cardiache. In un lavoro recente è descritta la correlazione tra miR-133, che regola le proteine RhoA e Nelf-A/WHSC2, e l'ipertrofia dei cardiomiociti [58]. miR-1 è sovra espresso negli individui affetti da patologie coronariche, e la sovra espressione di tale miRNA nel cuore dei ratti, esacerba l'aritmia silenziando i geni GJA1 e KCNJ2 [59]. Il knock-out di miR-1 può inibire le aritmie ischemiche, suggerendo una possibile applicazione terapeutica.

Tutti questi studi dimostrano le potenziali funzioni regolatorie dei miRNA nei diversi tipi cellulari e tessuti. I miRNA, attraverso la modulazione di network di centinaia o migliaia di proteine, potrebbero essere coinvolti nella patofisiologia di molte malattie umane, ed un singolo miRNA potrebbe avere effetti su più pathway patologiche, a causa dei diversi target. Come mostrato nel paragrafo precedente, i target dei miRNA comprendono geni coinvolti nel differenziamento e nella trasformazione, quali i fattori di trascrizione e le proteine coinvolte nel controllo del ciclo cellulare. Le malattie potrebbero pertanto essere il risultato della perturbazione di queste pathway a causa di mutazioni nei geni miRNA, dei siti di legame sui loro target o nelle pathway che ne regolano l'espressione.

Nella ricerca di nuove entità molecolari da utilizzare come strumenti terapeutici, sia i miRNA che i loro target sono potenzialmente bersagliabili. Future strategie terapeutiche potrebbero utilizzare i miRNA o gli anti-miRNA come piccole

molecole in grado di mimare o antagonizzare l'azione dei miRNA su target multipli, dando luogo a terapie innovative per malattie attualmente difficili da trattare.

1.4 I miRNA e il cancro

Sebbene gli studi mirati a determinare le correlazioni tra le disfunzioni dei miRNA e le malattie umane siano ancora agli inizi, esiste già una grande quantità di dati che dimostra il ruolo cruciale dei miRNA nella patogenesi del cancro. La prima evidenza del coinvolgimento dei miRNA nel cancro è data dalla sottoespressione o delezione di miR-15 e miR-16nella maggior parte dei pazienti affetti da leucemia linfocitica cronica (CLL) [60]. Questa scoperta ha dato il via a numerosi studi che hanno rivelato l'espressione differenziale dei miRNA non solo tra tessuto normale e tumorale, ma anche tra tumore primario e tessuto metastatico. Queste differenze sono tumore-specifiche ed in alcuni casi associabili alla prognosi. Evidenze attribuiscono ai miRNA tanto la funzione di oncogeni quanto di oncosoppressori.

La famiglia let-7 contiene miRNA in grado di regolare la famiglia degli oncogeni RAS attraverso repressione post-trascrizionale [61]. Uno studio recente mostra che l'espressione di let-7g nelle cellule di tumore polmonare che esprimono il gene K-Ras nel topo, induce l'arresto del ciclo cellulare e la morte delle cellule, rivelando il potenziale terapeutico della famiglia let-7 come oncosoppressori [62].

Un altro studio riporta per la prima volta la capacità di un miRNA di indurre una malattia neoplastica. Infatti, attraverso l'uso di topi transgenici, si è dimostrato che la sovra espressione di miR-155 nelle cellule B induce una proliferazione pre-leucemica, seguita da una malattia maligna [63].

Un'analisi completa dell'espressione dei miRNA nelle diverse fasi della carcinogenesi gastrica, ha rivelato una sovra regolazione del cluster miR-106b-25 con conseguente effetto sulla pathway dell'oncosoppressore TGF-β, interferendo con l'espressione di p21Waf1/Cip1 e Bim. Questi risultati suggeriscono il ruolo chiave di questo cluster nel cancro dello stomaco, causato dall'interferenza con proteine coinvolte nel ciclo cellulare e nell'apoptosi [64].

Studi recenti inoltre, hanno avuto come obiettivo lo studio delle correlazioni tra l'espressione dei miRNA e lo sviluppo di metastasi. miR-10b è altamente espresso nelle cellule metastatiche di cancro della mammella. Tale sovra espressione,

indotta dal fattore di trascrizione Twist, inizia l'invasione e le metastasi attraverso l'inibizione della traduzione del gene homeobox D10, con conseguente sovra espressione del gene pro-metastatico RHOC [65].

Anche i miRNA mir-126 e miR-335, la cui espressione è assente nelle cellule coinvolte nel cancro della mammella, hanno rivelato un potenziale metastatico. Difatti, il ripristino della loro espressione nelle cellule maligne, sopprime le metastasi dei polmoni e delle ossa *in vivo*, principalmente regolando il fattore di trascrizione SOX4 e rivelando il ruolo di tali miRNA come soppressori di metastasi [66]. Ci sono infine delle aree cromosomiche che vanno spesso incontro a rotture. Tale aree cromosomiche vengono chiamate **siti fragili**. Alcuni di loro sono associati a malattie genetiche nell'uomo e molti studi hanno dimostrato la loro importanza nell'instabilità genomica nel cancro.

1.5 I piRNA

Un importante studio che si sta svolgendo è quello sui piRNA (Piwi-associated RNA) ovvero filamenti di RNA non codificante costituiti da 24-30 nucleotidi che, a differenza dei miRNA, interagiscono con una classe diversa di proteine Argonauta chiamate Piwi e, per la loro biogenesi, sono indipendenti dal Dicer. I piRNA derivano principalmente da trasposoni e altri elementi ripetitivi. Studi genetici indicano che i piRNA sono molto importanti nello sviluppo della linea germinale e le proteine coinvolte nella produzione dei piRNA sono coinvolte nella regolazione dell'espressione genica nelle cellule somatiche, nell'apprendimento e nella memoria suggerendo che i piRNA potrebbero avere un ruolo molto importante in diversi processi biologici.

L'espressione alterata dei miRNA è dovuta, a volte, a riarrangiamenti cromosomici o ad eventi epigenetici, quindi è essenziale studiare i miRNA nel contesto della loro localizzazione genomica in modo da poter trovare correlazioni tra la loro espressione aberrante e una determinata malattia.

Al momento sono in corso una serie di attività per mappare tutti i geni miRNA e piRNA nei siti fragili, nei siti dove avvengono rotture dovute alle traslocazioni cromosomiche legate al cancro, nelle aree in cui vi sono elementi ripetitivi e nelle isole CpG e SNP, in modo da poter vedere l'incidenza dei due gruppi di RNA non

codificanti nei siti fragili e le eventuali somiglianze o differenze nella loro distribuzione nei singoli cromosomi umani. Contestualmente, è in corso un'analisi della distanza reciproca dei miRNA e piRNA per analizzare la loro distribuzione nel genoma umano.

1.6 Osservazioni

La ricerca nel campo dell'RNAi è in continuo fermento e sta portando alla luce meccanismi inimmaginabili nell'arco di pochi anni. La scoperta che certe regioni genomiche, precedentemente considerate non trascritte, generino invece una grande quantità di piccoli RNA che partecipano attivamente alla regolazione del genoma, è sorprendente e costituisce una vera e propria rivoluzione nella ricerca biologica di base ed applicata.

Sebbene esperimenti sempre più sofisticati e mirati stiano svelando giorno dopo giorno i meccanismi sottili alla base della biogenesi e delle funzioni dei piccoli ncRNA, sono ancora molte le domande che non trovano una risposta soddisfacente. Ci si chiede il perché di una varietà così ampia di piccole molecole di RNA che condividono molti aspetti ma che si differenziano per altri, sia evolutivamente che a livello di biogenesi e funzione. Ci si chiede il perché di sottoclassi di queste molecole, come ad esempio i Mirtron, e il perché dell'assenza di pathway di RNAi nei procarioti. Ci si aspetta inoltre la scoperta di nuove piccole molecole regolatrici in un futuro immediato.

Lo studio del coinvolgimento di miRNA e siRNA in tutte le pathway regolatorie, rivela la loro importanza fondamentale nella comprensione dei meccanismi alla base dei processi fisio-patologici. L'ipotesi che più di un terzo dei geni umani siano sotto il controllo dei miRNA spiega il loro ampio coinvolgimento in molte malattie, incluso il cancro. I miRNA possono comportarsi tanto da oncogeni che da oncosoppressori. Un miRNA che regola un oncosoppressore o una proteina pro-apoptotica può agire da oncogene, favorendo l'inibizione dell'apoptosi e il potenziamento del ciclo cellulare. Analogamente, un miRNA che regola un oncogene o una proteina anti-apoptotica agisce da oncosoppressore, favorendo l'aumento dell'apoptosi e il blocco del ciclo cellulare.

I numerosi tool computazionali disegnati per lo studio dell'RNAi si rivelano strumenti indispensabili per la comprensione tanto dei meccanismi di base quanto degli effetti sul fenotipo. La predizione computazionale di geni miRNA e siRNA e dei loro target, assieme a sofisticate analisi di Data Mining mirate a svelare correlazioni nascoste tra gli RNA regolatori, i loro target ed i processi fisio-patologici nei quali sono coinvolti, permetterà sempre più di comprenderne i complessi meccanismi molecolari, chiarire le cause di molte malattie, svelare il potenziale diagnostico e prognostico di tali molecole regolatrici e consentire il design di nuove terapie farmacologiche mirate e specifiche.

2 PREDIZIONE DI INTERAZIONI MIRNA/TARGET

La scoperta dei miRNA, abbondantemente presenti nei genomi di numerose specie pluricellulari, ha sollevato diverse questioni, soprattutto in merito alle loro funzioni. Ad oggi sono stati individuati più di 700 miRNA nell'uomo ed ogni giorno esperimenti di laboratorio mirati, permettono di studiare il loro coinvolgimento in pathway fisiologiche e patologiche. La chiave per la determinazione delle funzioni dei miRNA è la scoperta dei loro target. I primi indizi sulle modalità di riconoscimento dei target da parte dei miRNA vennero dall'osservazione che il miRNA lin-4, in C. elegans, aveva complementarietà di sequenza con diversi siti conservati nella regione 3' UTR del messaggero del gene lin-14. Analogamente, si osservò la complementarietà alle regioni 3' UTR dei geni lin-28 e lin-41 da parte dei miRNA lin-4 e let-7, rispettivamente. Numerose altre coppie miRNA/target nel corso degli ultimi anni sono state validate sperimentalmente, e adesso ci si pone l'obiettivo di individuare in maniera sistematica i target delle centinaia di miRNA identificati di recente, con funzione ancora ignota. Nelle piante, molti target possono essere predetti con una buona accuratezza, grazie alla complementarietà totale che i miRNA esibiscono nei loro confronti. Negli animali però, la complementarietà totale occorre solo occasionalmente, rendendo di fatto il problema notevolmente più complesso. La complementarietà perfetta limitata solo alle brevi regioni *seed* dei miRNA (~7-8 nt) verso i loro target, introduce il problema dei falsi positivi. La brevità dei *seed* infatti, rende altamente probabile l'individuazione di sequenze corrispondenti in diversi mRNA, senza che questo implichi una reale interazione con i miRNA. Sono dunque necessarie altre regole che vadano al di là della semplice corrispondenza di sequenza, da utilizzare per filtrare le predizioni e ridurre il numero dei falsi positivi. Tali regole devono necessariamente scaturire dall'osservazione sperimentale e da considerazioni energetiche.

Il lavoro svolto nell'ambito di questa tesi ha portato allo sviluppo di un nuovo tool computazionale per la predizione di interazioni miRNA/target, chiamato miRiam. In particolare ci si è posti l'obiettivo di determinare, dato un potenziale mRNA

target, i suoi miRNA regolatori più probabili, sulla base di considerazioni di carattere empirico e termodinamico.

In questo capitolo verranno discusse le problematiche principali inerenti il problema della predizione di target per i miRNA, verranno introdotti i tool di predizione più usati e verrà descritto l'algoritmo miRiam.

2.1 Il problema del targeting

L'individuazione di siti di legame per miRNA su sequenze di mRNA negli animali è un problema complesso per il quale non si dispone ancora di stumenti realmente efficaci. La maggior parte dei tool sviluppati finora infatti, soffrono di diversi problemi, primo fra tutti quello dell'enorme numero di falsi positivi. Le regolarità dedotte dalle evidenze sperimentali sono una caratteristica comune a tutti i predittori. Ad oggi è disponibile una grande quantità di dati relativi a coppie miRNA/target validate, utilizzabili sia da *training set* dal quale estrarre conoscenza, che da *benchmark*, per valutare le prestazioni dei tool.

La banca dati ufficiale dei miRNA, *miRBase*, conta ad oggi più di 700 miRNA umani, con oltre 800 previsti di esistere, e per ognuno di essi riporta le sequenze dei trascritti maturi e dei precursori [67]. Diverse altre banche dati raccolgono invece le coppie miRNA/target validate, tra queste Tarbase [68] e miRecords [69]. Per ogni coppia miRNA-gene sono date, dove disponibili, informazioni circa l'appaiamento delle basi nel duplex. Questa è un'informazione di importanza fondamentale, in quanto consente di determinare le modalità di appaiamento utilizzabili per addestrare i predittori.

Alle indispensabili regole di interazione si affiancano altre informazioni che possono essere usate per filtrare le coppie miRNA/target predette. Uno dei criteri ampiamente utilizzati dai tool disponibili in rete è la conservazione dei target. Dall'allineamento degli stessi miRNA in specie differenti emerge infatti un'alta conservazione, soprattutto delle sequenze *seed*, che si riflette nella conservazione dei siti di legame sui target. Questa informazione può dunque essere utile per aumentare lo score di una certa predizione, ma non è di nessun aiuto nel caso dei numerosi miRNA specie-specifici [70].

Molti predittori fanno uso di proprietà termodinamiche. La stabilità degli appaiamenti, valutabile attraverso l'energia libera ΔG del duplex miRNA/target predetto è uno dei criteri utilizzati dalla maggior parte dei tool per la valutazione delle coppie predette. Tutti i duplex miRNA/target validati sperimentalmente sono infatti caratterizzati da bassi valori di energia libera (solitamente inferiore a -20 Kcal/mol). Tuttavia si tratta di una condizione necessaria ma non sufficiente. Esperimenti di laboratorio, infatti, mostrano come spesso potenziali interazioni energeticamente favorevoli, non siano riscontrabili all'interno della cellula.

Un altro dei criteri termodinamici utilizzato da alcuni tool è quello dell'accessibilità strutturale della molecola target. Per poter legare il miRNA infatti, il sito target sull'mRNA non deve essere già coinvolto in altri appaiamenti intramolecolari, ed ogni struttura secondaria esistente deve prima essere disfatta. La valutazione dell'accessibilità strutturale è comunque un problema complesso e si basa solitamente sulla predizione della struttura secondaria del target [71].

Altre regole alla base di molti predittori riguardano la composizione nucleotidica e la posizione dei siti di legame sull'UTR del target, così come la presenza di siti di legame multipli su uno stesso UTR. E' infatti noto come uno stesso miRNA possa avere più siti di legame su uno stesso target e che uno stesso target possa avere più siti di legame per diversi miRNA [72].

Nei paragrafi successivi verranno passati in rassegna i principali tool per la predizione di target per miRNA e verrà illustrato il tool miRiam, sviluppato nell'ambito di questa tesi.

2.2 Tool per la predizione di target per miRNA

Ad oggi sono disponibili parecchi tool per la predizione di interazioni miRNA/mRNA. I più popolari sono TargetScan, miRanda, Pictar, Diana-microT, RNA22, RNAHybrid e microInspector.

Uno dei tool più popolari per la ricerca di target per miRNA è TargetScan, un sofisticato algoritmo basato su conservazione e regole di appaiamento [70]. TargetScan ricerca corrispondenze sui target per *seed* di almeno 7 nucleotidi, a partire dal secondo nucleotide all'estremità 5' del miRNA, ed usa il predittore di struttura secondaria RNAFold per calcolare l'energia libera di legame. La presenza di siti multipli su uno stesso target per uno stesso miRNA rafforza lo score della predizione. La versione migliorata, TargetScanS, richiede un *seed* più breve (6 nt) preceduto da un'adenina e situato in una piccola area conservata circondata da regioni meno conservate. Entrambi i tool sfruttano inoltre la conservazione dei siti di legame su diverse specie per l'identificazione dei target più probabili. Sul sito web di TargetScan sono consultabili le predizioni già calcolate su tutti gli UTR di diverse specie, tra cui l'uomo, il topo ed il ratto.

Anche il tool miRanda permette l'identificazione di target per miRNA su uomo, topo e ratto [73]. Il tool usa un algoritmo di allineamento basato su una matrice pesata per enfatizzare il legame della regione 5' del miRNA piuttosto che la regione 3', ed utilizza il tool RNAFold per il calcolo dell'energia libera dei duplex predetti. La conservazione è un criterio utilizzato per la valutazione dei target più probabili. Anche nel caso di miRanda sul sito web sono disponibili le predizioni già calcolate, ma è anche possibile scaricare una versione light del software, privo del modulo relativo all'analisi della conservazione, per effettuare le predizioni su target personalizzati o non presenti sul sito.

PicTar è un altro tool per la predizione di target per miRNA su vertebrati, *C. elegans* e *Drosophila* [74]. L'algoritmo è addestrato per l'identificazione di siti di legame per un singolo miRNA e di siti multipli regolati da diversi miRNA che agiscono in modo cooperativo. Esso utilizza un algoritmo di allineamento *pairwise* per filtrare i siti di legame conservati in molte specie (7 specie di Drosophila e 8 di vertebrati) e considera il clustering e la co-espressione dei miRNA assieme ad

informazioni ontologiche (corrispondenza di miRNA con i potenziali target espressi negli stessi tessuti e nelle stesse fasi dello sviluppo).

L'algoritmo di Diana-MicroT è invece addestrato per l'identificazione di target con un solo sito di legame per miRNA [75]. L'approccio utilizzato si distingue dai precedenti in quanto si focalizza sulla ricerca di duplex miRNA/target dotati di *bulge* centrale ed appaiamento sia del *seed* che della regione 3'.

Ancora diverso è l'approccio seguito dal tool RNA22, il quale calcola le sequenze complementari inverse di pattern statisticamente significativi in target potenziali e li utilizza per identificare possibili miRNA regolatori [76].

RNAHybrid è invece un algoritmo di folding di sequenze di RNA in grado di fornire una stima dell'energia libera di ibridazione di una molecola breve di RNA con una più lunga. E' dunque possibile utilizzarlo per predire target per miRNA, attraverso l'introduzione di regole specifiche, quali ad es. l'appaiamento obbligatorio del seed e la presenza di mismatch in certe posizioni [77].

Infine, il tool microInspector consente di predire siti di legame per miRNA su un dato mRNA e si basa sulla ricerca di sequenze complementari a quelle dei miRNA [78]. microInspector, una volta individuate le sequenze candidate, utilizza il software per la predizione di strutture secondarie MFold per la predizione dei duplex miRNA/target, scartando quei duplex la cui energia è al di sopra di una certa soglia.

2.3 Il tool miRiam

miRiam è un tool computazionale per la predizione di target per miRNA, progettato e sviluppato nell'ambito di questa tesi. In particolare, obiettivo di miRiam è la predizione di interazioni tra un dato mRNA ed un database di miRNA. Il metodo usa un approccio molto semplice basato su caratteristiche termodinamiche e regole empiriche inferite da interazioni miRNA/target note.

2.3.1 Accessibilità del target e regole empiriche

Uno dei fattori chiave nell'efficienza del silenziamento genico post-trascrizionale è la struttura secondaria del target [71, 79, 80]. In particolare, la qualità della repressione dipende dall'accessibilità del sito di legame, data dalla propensione delle basi a formare appaiamenti intramolecolari stabili, e correlata all'energia libera locale ΔG: tanto minore è l'energia, tanto più stabile, e di conseguenza meno accessibile, è il sito.

Inoltre l'accuratezza delle predizioni dipende da osservazioni empiriche inferite dalle interazioni miRNA/mRNA note. E' ormai noto come la complementarietà perfetta della regione *seed* del miRNA (tipicamente i nucleotidi 2-8) con il suo target, sia una caratteristica comune a quasi tutte le interazioni conosciute [1]. Conseguentemente, l'energia libera ΔG del duplex seed/target contribuisce più dell'energia totale alla specificità e all'efficienza del silenziamento. In tutti quei casi nei quali la complementarietà del seed non è perfetta, è osservato solitamente un effetto compensatorio nella regione 3' [72].

I wobbles G:U, che sono ricorrenti nei duplex RNA/RNA stabili, sembrano invece sfavorire la funzione dei miRNA, soprattutto quando occorrono nella regione del *seed* [81].

Infine, la gran parte dei siti di legame per miRNA validati sperimentalmente, sono localizzati nelle regioni 3' UTR dei loro target, e la presenza di siti multipli, osservata frequentemente, sembra giocare un ruolo importante nell'efficienza del silenziamento [82].

```
miRNA   3'  C-UUUGGGUC-GUCUGUUACAUCGA     5'
            | |||  :|  : || ||||||||
mRNA    5'  GUAAA--UA-UU-GA-AAUGUAGC-     3'
```

Fig. 2.3.1 – **Esempio di appaiamento miRNA/target.** I trattini indicano appaiamenti canonici di Watson/Crick, i due punti indicano i GU wobble.

2.3.2 L'approccio di miRiam

Come già premesso, miRiam fa uso sia delle proprietà termodinamiche delle interazioni RNA-RNA, che delle regole empiriche dedotte dalle evidenze sperimentali. Per l'identificazione delle regioni accessibili, viene utilizzata una funzione di *scoring* semplice e computazionalmente leggera, basata sulle probabilità di appaiamento dei nucleotidi di un potenziale mRNA target. Le regole empiriche sono successivamente usate sia per guidare l'allineamento delle regioni candidate ai miRNA, sia per valutare e filtrare gli allineamenti prodotti. Questo metodo è abbastanza veloce da permettere la scansione di grandi database di miRNA, quale l'intero miRBase, alla ricerca di possibili interazioni, in un tempo accettabile.

2.3.3 L'algoritmo di miRiam

L'algoritmo di miRiam si compone dei seguenti passi: filtering dell'mRNA, calcolo delle interazioni miRNA/mRNA, filtro post-allineamento e valutazione dell'energia libera dei duplex.

- ***Filtering dell'mRNA***. Dato in input un dato mRNA target, le probabilità di appaiamento delle basi sono calcolate sull'intera sequenza, utilizzando la libreria di funzioni RNAlib del pacchetto Vienna RNA [83, 84]. Sia *(i,j)* una coppia di nucleotidi: la probabilità di appaiamento *p(i,j)* è la probabilità che *i* e *j* siano appaiati nella struttura secondaria a minima energia libera. Definiamo *w-binding-region* come una sottosequenza del target di lunghezza *w*, considerata come sito di legame candidato per un miRNA. Consideriamo solo *w-binding-region* situate nelle regioni 3' UTR dei target, essendo questa una caratteristica osservata in quasi tutta la totalità dei casi noti. Questo permette di ridurre lo spazio di ricerca e di velocizzare i calcoli. Comunque, viene data all'utente la possibilità di ricercare siti di legame sull'intera sequenza dell'mRNA. Il valore di *w* varia tra 28 e 32, in accordo con quanto dedotto dalle coppie miRNA/mRNA validate sperimentalmente. Per l'identificazione di tutte le *w-binding-region* occorre valutare l'accessibilità locale della molecola di mRNA target. Questo è eseguito in maniera semplice e veloce nel modo seguente. Inizialmente, la probabilità

di appaiamento media, per ogni nucleotide, viene calcolata attraverso la seguente formula:

$$\overline{p(i)} = \frac{\sum_{j \in mRNA} p(i,j)}{mRNA\_length - 1} \qquad (1)$$

Questo valore dà una stima del grado di appaiamento di ogni singolo nucleotide con ciascuna altra base nella sequenza di mRNA. Scorrendo il target, viene calcolato lo score di appaiamento medio per ogni possibile *w-binding-region*:

$$\overline{p(w)} = \frac{\sum_{i \in w-region} \overline{p(i)}}{w} \qquad (2)$$

Tutte le *w-binding-region* con score di appaiamento al di sopra di una certa soglia *t* sono scartate. Le regioni rimanenti vengono considerate potenzialmente accessibili per l'interazione con un miRNA. Questa fase di *filtering* velocizza il calcolo delle interazioni miRNA/mRNA, riducendo in maniera significativa il numero di allineamenti da calcolare. Gli utenti possono fissare un valore soglia *t* arbitrario, o scegliere uno fra tre valori dipendenti dalla sequenza e chiamati low, medium ed high.

$$\begin{aligned} t_{low} &= \frac{\max \overline{p(w)}}{4} \\ t_{medium} &= \frac{\max \overline{p(w)}}{2} \\ t_{high} &= \frac{3 \cdot \max \overline{p(w)}}{4} \end{aligned} \qquad (3)$$

- ***Calcolo delle interazioni miRNA/mRNA***. Per prima cosa, viene specificato un database di miRNA, tipicamente un sottoinsieme di miRBase, contenente miRNA di uno stesso organismo. Quindi, viene eseguito l'allineamento delle coppie di basi complementari tra i miRNA e le regioni accessibili calcolate al passo precedente. Tale allineamento è una variante del classico allineamento *pairwise* di sequenze nel quale sono ammessi solo match tra

coppie di basi complementari. Viene utilizzato il classico algoritmo di programmazione dinamica di Needleman-Wunsch con *affine-gap-penalties* [85], utilizzando la seguente funzione di *scoring*:

$$\delta(x,y) = \begin{cases} +5, \text{ se } (x,y) \text{ è una coppia di Watson/Crick;} \\ +1, \text{ se } (x,y) \text{ è un wobble G:U;} \\ -15, \text{ per i mismatch.} \end{cases} \quad (4)$$

Inoltre, uno schema di peso posizionale viene utilizzato per privilegiare l'appaiamento delle basi nella regione 5' del miRNA, in accordo con le regole empiriche descritte precedentemente. In particolare, in maniera simile a miRanda, lo score della coppia *(x, y)* è moltiplicato per 2 se *x* è la prima base del miRNA, per 3 se *x* è compreso tra le basi 2 e 10, per 1 altrimenti. In questo modo viene favorito l'appaiamento della regione *seed*.

Infine, lo schema di penalità *affine-gap*viene usato per ottenere allineamenti più compatti. In particolare il GOP, la penalità per l'apertura di un gap, è fissata ad 8, mentre il GEP, la penalità per l'estensione di un gap già esistente è fissata a 2.

- ***Filtro post-allineamento***. Un secondo filtro è utilizzato per scartare tutti gli allineamenti che non soddisfano le regole empiriche di interazione miRNA/mRNA. Sono implementate tre varianti. Il filtro *strict* consente un massimo di 2 mismatch e non più di 1 wobble G:U nella regione 2-8 del duplex. Il filtro *relaxed* permette fino a 3 mismatch e non più di 2 wobble G:U nella regione 2-10. Infine, uno schema libero consente all'utente di scegliere il massimo numero di mismatch e wobble insieme alla sottosequenza dell'mRNA alla quale applicare il filtro.
- ***Valutazione dell'energia libera dei duplex***. Per ogni allineamento che passa il filtro illustrato al passo precedente, viene calcolata l'energia libera della regione 5'. Tutti gli allineamenti con valori di energia al di sopra di una certa soglia sono scartati. I restanti allineamenti vengono restituiti all'utente, ordinati in base al valore di energia.

2.4 miRiam: risultati ottenuti

Le prestazioni di miRiam sono state valutate su un campione di 58 coppie miRNA/target validate sperimentalmente, per le quali è riportato almeno un sito di legame. Sono stati considerati 47 diversi target in *Homo sapiens* per un totale di 101 siti di legame, ottenuti da TarBase [68]. E' stato utilizzato il database completo di miRNA umani proveniente da miRBase, ed i test sono stati eseguiti utilizzando i parametri di default, soglia di accessibilità: *medium*, filtro post-allineamento: *strict*. Il confronto diretto è stato fatto con microInspector, l'unico tool disponibile che esegue lo stesso *task* di miRiam. E' stata testata la versione on-line con i parametri di default (Temperatura di ibridazione: 37°C, Soglia di ΔG: -20 Kcal/mol). miRiam è stato in grado di identificare correttamente 88 dei 101 siti di legame, mentre microInspector ne ha individuati solo 29. Inoltre, innalzando la soglia di accessibilità di miRiam da *medium* ad *high*, il numero di match corretti è salito a 94. Dei restanti sette siti di legame, tre non sono stati ritrovati, in quanto valutati strutturalmente inaccessibili. Gli altri quattro non soddisfacevano le regole imposte dal filtro post-allineamento *medium* a causa della loro complementarietà imperfetta con la regione *seed* del miRNA. La tabella 2.2 riporta alcuni dei risultati ottenuti.

miR/Gene	# BS	miRiam	microInsp.	miR/Gene	# BS	miRiam	microInsp.
miR-145/PARP8	2	1	1	mir-1/BDNF	3	3	0
mir-101/MYCN	2	1	0	mir-1/G6PD	3	2	0
mir-103/FBXW11	1	1	0	mir-1/HAND2	1	0	0
mir-124a/MTPN	1	1	0	mir-15/DMTF1	1	1	1
mir-130a/CSF1	2	2	0	miR-23a/HES1	3	3	0
mir-133a/SRF	2	2	1	let-7b/KRAS	8	6	2
mir-141/CLOCK	1	1	1	mir-1/HDAC4	2	2	0
mir-143/MAPK7	1	0	0	miR-1/TMSB4X	1	0	0
mir-196a/HOXB8	1	1	1	miR-23a/CXCL12	2	2	0
mir-24/MAPK14	1	1	1	let-7a/NRAS	8	6	0
mir-15/BCL2	1	1	0	miR-132/RICS/p250GAP	1	0	0
mir-196a/HOXA7	4	4	0	miR-10a/HOXA1	1	1	0
mir-34/DLL1	1	1	1	miR-155/AGTR1	1	1	0
mir-199/LAMC2	1	1	1	miR-127/BCL6	1	1	0
mir-19a/PTEN	3	3	0	miR-140/HDAC4	1	1	0
mir-16/BCL2	2	2	2	miR-23a/C6orf134	1	1	1
mir-223/NFIA	1	1	0	mir-20/E2F1	1	1	1
mir-23/POU4F2	3	3	0	miR-206/Fstl1	1	1	1
miR-17-5p/AIB1	1	1	1	miR-206/Utrn	1	1	0
mir-26a/SMAD1	2	2	0	miR-206/GJA1	2	2	0
mir-26b/SMAD1	2	2	0	miR-1/GJA1	2	2	0
mir-34/NOTCH1	4	3	3	miR-189/SLITRK1	1	1	0
let-7b/LIN28	2	2	1	miR-125a/Lin28	1	1	0
miR-127/Rtl1/Peg11	1	1	1	mir-16/CG38	1	1	1
miR-433/Rtl1/Peg11	1	1	1	mir-125b/LIN28	1	1	0
mir-221/KIT	1	1	0	let-7e/SMC1A	1	1	1
miR-222/KIT	1	1	0	mir-17-5p/E2F1	1	1	1
miR-431/Rtl1/Peg11	2	2	2	miR-27b/CYP1B1	1	1	1
mir-375/MTPN	1	1	0	mir-101/EZH2	1	0	1

Tabella 2.2 – Confonto tra miRiam e microInspector. I tool sono stati confrontati sul numero di siti di legame validati individuati correttamente. # BS rappresenta il numero di siti di legame validati nell'mRNA target.

Lo step successivo è stato il confronto di miRiam con gli altri tool di predizione più popolari, utilizzando i dati pre-calcolati consultabili sui loro siti web. Per questo confronto sono stati considerati TargetScan, PicTar, miRanda ed RNA22. miRiam ha ottenuto i risultati migliori nei confronti di tutti e quattro i tool, ed i risultati sono mostrati nella tabella 2.3.

Tool	Correct Binding Sites
MicroInspector	29/101
miRanda	44/101
miRiam (*medium-strict*)	**88/101**
miRiam (*medium-relaxed*)	**92/101**
miRiam (*high-strict*)	**94/101**
miRiam (*high-relaxed*)	**98/101**
PicTar	41/101
RNA22	11/101
TargetScan 3.1	56/101

Tabella 2.3 – Confonto tra miRiam e altri tool. La tabella mostra il confronto tra le prestazioni dei tool di target prediction, valutate sul numero di siti di legame validati sperimentalmente individuati.

miRiam è stato anche confrontato con due tool del pacchetto Vienna RNA (versione 1.6): RNAcofold ed RNAduplex [86]. Questi programmi non effettuano scansioni di database ma accettano in input due sequenze di RNA (una breve, ad es. un miRNA, ed una più lunga, ad es. un mRNA target), e restituiscono la struttura del duplex a minima energia libera (mfe). Entrambi i software sono basati su approcci termodinamici. In particolare, RNAcofold calcola la mfe di dimeri di RNA, considerando la struttura secondaria di entrambe le sequenze e consentendo appaiamenti di basi intra- ed intermolecolari. In RNAduplex invece sono ammessi solo appaiamenti intermolecolari. Questi tool non sono progettati specificamente per risolvere il problema del targeting dei miRNA, quindi possono produrre duplex inconsistenti, ad esempio coppie nelle quali il seed del miRNA è completamente spaiato. Inoltre, essi restituiscono solo un dimero per ogni input. Questo può costituire una limitazione importante, dato che molti miRNA legano i loro target su siti multipli.

Il pacchetto Vienna RNA contiene infine un tool chiamato RNAup che consente di valutare in maniera accurata l'accessibilità strutturale del target [87]. Sebbene il tool sia in grado di trovare buone regioni candidate per miRNA, è molto costoso in termini di tempo e di conseguenza inutilizzabile per i nostri scopi. Ad esempio, RNAup è stato in grado di determinare un sito di legame corretto per miR-15a su BCL2 in circa 14 ore. miRiam ha restituito lo stesso sito, insieme a quelli relativi a tutti gli altri potenziali miRNA regolatori di BCL2 in circa 90 minuti, sulla stessa macchina.

2.5 miRiam: valutazione delle prestazioni

miRiam permette la scansione di grandi database di miRNA per la predizione di interazioni potenziali con un dato mRNA e mostra buone prestazioni in termini di flessibilità, affidabilità e velocità.

Flessibilità: miRiam opera su sequenze di mRNA specificate dall'utente, permettendo la predizione di interazioni con database arbitrari di miRNA. Mutazioni nel 3' UTR dei geni possono influire drammaticamente sul silenziamento ad opera dei miRNA, in quanto possono rimuovere o creare siti di legame. E' stato riportato ad esempio, che una mutazione puntuale sul 3' UTR del gene della miostatina (GDF8), contribuisce all'ipertrofia muscolare a causa dell'introduzione di un sito di legame illegittimo per miR-1 e miR-206, altamente espressi nel muscolo scheletrico [88]. Inoltre, è plausibile che mutazioni in altre parti dell'mRNA, inclusi il 5' UTR e la sequenza codificante, possano modificare la struttura secondaria, rivelando o nascondendo potenziali siti di legame per miRNA. Tali eventi possono essere analizzati e predetti usando miRiam sulle sequenze di mRNA mutate ottenute dai database o dal sequenziamento customizzato (analisi funzionale di polimorfismi). Questa è una delle caratteristiche più importanti di miRiam, dato che i tool di predizione che rilasciano dati consultabili su web, escludono dai loro calcoli le sequenze non presenti nei database pubblici.

Un'altra proprietà importante di miRiam, è la sua capacità di investigare potenziali interazioni inter-specie. Ad esempio, è ormai noto come i miRNA virali siano in grado di regolare i trascritti della cellula ospite, contribuendo alla persistenza del virus in forma latente. Casi del genere possono essere predetti usando miRiam, eseguendo lo screening del database dei miRNA virali alla ricerca di interazioni con dati target cellulari. Analogamente, i trascritti virali possono essere regolati da miRNA cellulari durante le infezioni. Infatti, sembra ad esempio che i miRNA umani miR-29a, miR-29b, miR-149, miR-378 e miR-324-5p siano in grado di regolare geni di HIV-1 [89]. Un altro studio mostra come un miRNA specifico del fegato, miR-122, interagisca con una regione non codificante del genoma di HCV, modulando la replicazione del virus [90]. Tali casi possono essere studiati con miRiam, analizzando i trascritti virali alla ricerca di interazioni potenziali con i miRNA cellulari.

Affidabilità: I test eseguiti sul benchmark di coppie miRNA/target validate sperimentalmente, mostrano una qualità di predizione tre volte superiore a quella del suo diretto competitore, microInspector. Inoltre, il confronto con i tool più popolari dimostra come miRiam sia in grado di produrre predizioni di alta qualità ed il confronto con i tool del pacchetto Vienna RNA indica come le proprietà termodinamiche da sole non siano sufficienti a produrre predizioni coerenti. Infatti, i risultati ottenuti usando RNAcofold, RNAduplex ed RNAup non sono sempre consistenti con le coppie validate, dato che a volte coinvolgono regioni quali la CDS ed il 5' UTR e non soddisfano i vincoli di appaiamento richiesti.

Infine, la figura 2.6 mostra come la maggior parte dei siti di legame validati sperimentalmente, siano considerati accessibili dal filtro di miRiam.

Velocità: Il tempo di esecuzione di miRiam non può essere confrontato direttamente con quello di microInspector, dato che quest'ultimo è utilizzabile solo su web e non è disponibile per l'utilizzo in locale. Comunque, i test effettuati utilizzando la versione web, hanno mostrato che i due tool hanno tempi di

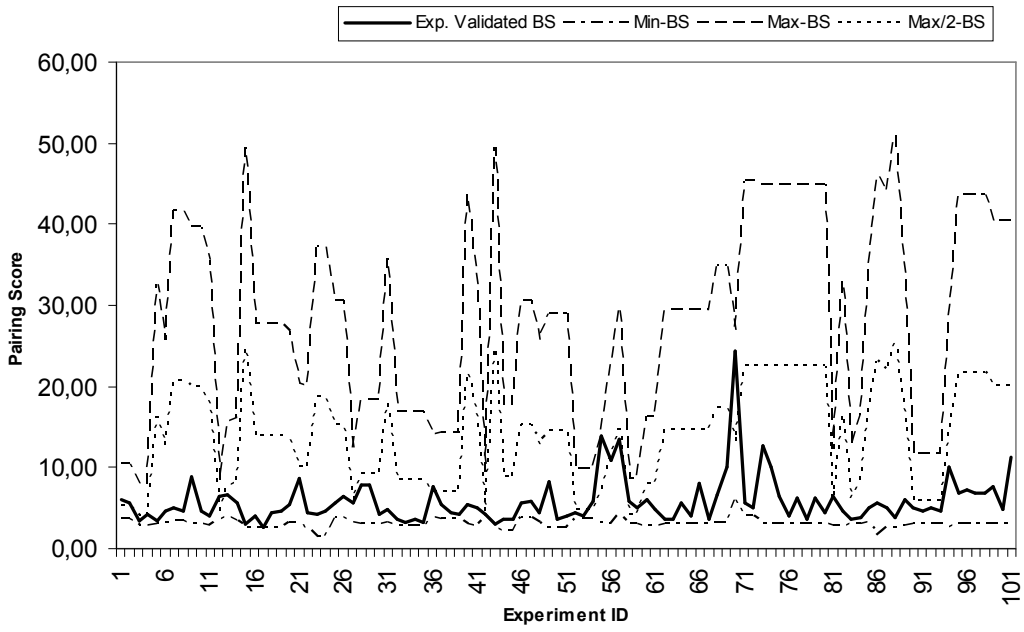


Fig. 2.6 – Valori di appaiamento dei siti di legame. *Exp. Validated BS* si riferisce ai siti di legame validati sperimentalmente. *Min-BS* (*Max-BS*) corrisponde alle regioni meno (più) appaiate. *Max/2-BS* corrisponde alla soglia di default.

esecuzione confrontabili, che non hanno in nessun caso superato i 90 minuti. Gli esperimenti su miRiam sono stati effettuati su un Centrino Dual Core 2GHz con 2 GB di RAM e Sistema Operativo Linux. Gli altri tool di predizione forniscono solo predizioni pre-calcolate e di conseguenza non è possibile valutarne la velocità di calcolo. RNA22 offre un servizio di predizione on-line ma non effettua scansioni di interi database. Per lo stesso motivo non è stato possibile utilizzare i tool del pacchetto Vienna. Ad ogni modo, il più accurato di essi è riuscito a predire un sito di legame corretto per miR-15a su BCL-2 in 14 ore, contro i 90 minuti necessari a miRiam per effettuare la scansione dell'intero database, sulla stessa macchina.

2.6 Problematiche e sviluppi futuri

I risultati ottenuti da miRiam sul dataset di coppie miRNA/target validate sperimentalmente sono pienamente soddisfacenti. Tuttavia, essendo le informazioni strutturali determinanti nell'individuazione dei siti di legame più probabili, si è ritenuto indispensabile provare il tool sulle stesse sequenze utilizzate nella validazione sperimentale delle interazioni, piuttosto che sulle sequenze originali di mRNA presenti in banca dati. Gli esperimenti di validazione infatti, prevedono l'utilizzo di geni *reporter*, solitamente luciferasi o GFP, nei quali vengono inseriti gli UTR dei geni originali contenenti i siti di legame candidati. Va dunque tenuta in considerazione la possibilità che isiti di legame originali, nel contesto del gene *reporter*, cambino la loro accessibilità a causa della differente struttura secondaria della molecola.

Inoltre, durante lo sviluppo di miRiam, sono stati pubblicati due tool di predizione di target per miRNA, StarMir e PITA, che utilizzano funzioni sofisticate per il calcolo dell'accessibilità strutturale delle molecole target. In particolare, StarMir modella l'interazione tra un miRNA ed il suo target come un processo di ibridazione a due fasi: la nucleazione su un sito accessibile, seguita dall'allungamento dell'ibrido per disfare la struttura secondaria locale del target e formare il duplex miRNA/target completo [91]. PITA invece è basato sul calcolo della differenza tra l'energia libera guadagnata dalla formazione del duplex miRNA/target e il costo energetico dell'apertura del sito di legame per l'appaiamento con il miRNA [92].

Si è pertanto deciso di effettuare una sperimentazione dei filtri di accessibilità di miRiam sul set di sequenze di geni reporter GFP di *C. elegans* pubblicate da Didiano ed Hobert, e di confrontare le prestazioni di miRiam con quelle di StarMir e PITA. Il set è composto da 13 potenziali target che contengono diversi livelli di complementarietà al *seed* del miRNA lsy-6 [93]. L'unico target che ha mostrato una regolazione efficiente è stata quella del gene cog-1, già target validato di lsy-6, mentre le restanti 12 sequenze non hanno mostrato variazioni significative nella loro espressione. I risultati ottenuti da miRiam su questo set, hanno rivelato una capacità discriminante tra siti di legame veri e falsi non soddisfacente, paragonabile comunque a quella di PITA (vedi Tabella). Il tool StarMir si è rivelato invece in grado di predire correttamente 12 casi su 13. Tuttavia, un secondo lavoro di Didiano ed Hobert, sempre su sequenze di C. elegans regolate da lsy-6, mostra prestazioni non soddisfacenti sia di StarMir che di PITA, nel predire siti di legame dimostrati sperimentalmente.

Appare dunque evidente la necessità di analizzare con maggiore accuratezza i dati sperimentali, alla ricerca di ulteriori regole per il raffinamento dei predittori, e di perfezionare i filtri di accessibilità strutturale. A tal fine sono in fase di studio ed implementazione le seguenti modifiche al tool miRiam.

Raffinamento del filtro di accessibilità. Un recente lavoro di Tafer e colleghi, sostiene la località dell'accessibilità strutturale, confortata da dati sperimentali, proponendo un modello semplificato nel quale il grado di appaiamento delle basi è calcolato su finestre scorrevoli del target (lunghe al più 40 nt), piuttosto che sull'intera sequenza [94]. Gli autori, infatti, sostengono che la presenza di strutture secondarie che coinvolgono basi molto distanti tra loro nella sequenza sia altamente improbabile, in quanto tali strutture verrebbero disfatte al passaggio dei ribosomi durante la traduzione e impiegherebbero troppo tempo per riformarsi. Tale approccio consente una notevole riduzione dei tempi di calcolo ed implica un miglioramento dell'accuratezza delle predizioni della struttura, il cui calcolo è notoriamente più affidabile su sequenze brevi. Si è scelto dunque di implementare tale filtro di accessibilità in miRiam.

Raffinamento delle regole di appaiamento. E' inoltre in corso l'implementazione di nuove regole empiriche di appaiamento miRNA/target suggerite dalle evidenze pubblicate da Grimson, Nielsen e Didiano/Hobert [72, 95, 96]. Tali regole

riguardano la composizione nucleotidica dei siti di legame, la posizione di tali siti negli UTR e le loro distanze reciproche.

Infine, è in fase di studio la possibilità di implementare un modulo per il riconoscimento di siti di legame per proteine sugli RNA target. Infatti, uno dei problemi relativi all'individuazione dei siti di legame per i miRNA è proprio la presenza di *RNA binding protein*, sulle quali non si hanno molte informazioni e che potrebbero legare eventuali siti apparentemente disponibili per l'interazione con un miRNA. A tal fine si è deciso di analizzare i dati presenti nel database RsiteDB, che riporta informazioni strutturali relative a coppie RNA/proteine dalle quali è possibile determinare i siti di legame di tali proteine sugli RNA [97]. Obiettivo è l'individuazione di regolarità di sequenza e di struttura da poter utilizzare per l'addestramento di filtri che permettano di scartare, o comunque mettere in evidenza, siti candidati al legame con una proteina e, di conseguenza, potenzialmente inaccessibili per l'appaiamento con un miRNA.

3 ANNOTAZIONE FUNZIONALE DEI MIRNA

In questo capitolo viene presentata la caratteristica principale del sistema web miRò per l'annotazione funzionale di miRNA. Il sistema, come verrà descritto ampiamente più avanti, è una vera e propria base di conoscenza, con interfaccia web che fornisce agli utenti associazioni miRNA/fenotipo nell'uomo. miRò integra dati da diverse fonti on-line, quali i database di miRNA, ontologie, malattie e target, in un unico ambiente dotato di un sistema di query flessibile ed intuitivo, e funzionalità di data mining. Obiettivi principali di miRò sono l'implementazione di una base di conoscenza che permetta analisi non banali attraverso tecniche di mining sofisticate, e l'introduzione di un nuovo livello di associazione tra geni e fenotipi, inferito dalle annotazioni dei miRNA.

3.1 Introduzione

Come già discusso nel capitolo 2, l'abbondanza di miRNA individuati in molte specie e il grande numero di geni soggetti a regolazione post-trascrizionale, implicano un coinvolgimento significativo di tali molecole in molti processi biologici. La presenza di siti di legame per un singolo miRNA su più trascritti, la co-regolazione di uno stesso trascritto da parte di più miRNA e l'esistenza di *feedback loop* negativi, secondo i quali un miRNA può essere regolato dai suoi stessi target,rivelano la grande complessità dei meccanismi di regolazione basati sui miRNA e il loro inevitabile coinvolgimento nella patogenesi di molte malattie.

Ad oggi, studi diversi hanno documentato correlazioni, più o meno significative, tra uno o più miRNA e processi fisiologici e patologici, ma l'associazione precisa tra miRNA e fenotipo è stata dimostrata solo per pochi casi. Molto di più, invece, si conosce sui geni. Ad esempio, il database Gene Ontology (GO) fornisce annotazioni circa i processi e le funzioni nei quali i geni sono coinvolti. Inoltre, esiste una vasta letteratura che documenta il ruolo dei geni nelle malattie. E' dunque possibile annotare i miRNA con le informazioni inerenti i loro target validati o predetti. Ad esempio, la correlazione tra la sotto-espressione di miR-15a e miR-16 e la sovra-espressione del gene anti-apoptotico BCL-2 nei pazienti di leucemia linfocitica

cronica a cellule B (CLL), permette di associare funzionalmente tali miRNA all'apoptosi e alla CLL [60].

Un approccio comune, nello studio di malattie e processi biologici che coinvolgono i miRNA, richiede l'estrazione di dati da diverse fonti indipendenti, quali i database di predizioni miRNA/target, le annotazioni funzionali dei geni, i profili di espressione e la letteratura biomedica. Quindi, è importante poter disporre di sistemi che integrino dati da fonti eterogenee in ambienti unici, estendibili ed aggiornabili. Tali sistemi dovrebbero inoltre essere provvisti di algoritmi di data mining in grado di inferire nuova conoscenza dai dati raccolti.

Il primo sistema per l'annotazione funzionale dei miRNA è miRGator, un database che integra dati da diverse fonti, quali i tool di target prediction e Gene Ontology, e li mette a disposizione degli utenti attraverso un insieme di query standard [132]. Sebbene miRGator rappresenti un primo tentativo nell'integrazione di tali dati, esso soffre di parecchie limitazioni. Ad esempio, non fornisce alcuna informazione sulle malattie, né implementa query flessibili o funzionalità di data mining.

Nell'ambito di questa tesi è stato sviluppato miRò, un nuovo sistema per l'annotazione funzionale dei miRNA nell'uomo. miRò è un ambiente web che permette l'esecuzione di ricerche semplici e di sofisticate query di data mining. Obiettivo principale del sistema è di mettere a disposizione degli utenti potenti tool per la scoperta di associazioni non banali tra dati eterogenei e permettere di conseguenza l'identificazione di relazioni tra geni, processi, funzioni e malattie a livello dei miRNA.

3.2 Specificità delle associazioni miRNA/fenotipo

Ad ogni coppia miRNA/processo o miRNA/malattia, in ogni sottoinsieme, viene attribuito uno score, calcolato attraverso una funzione di specificità. Questo permette di valutare le relazioni tra i miRNA e le loro annotazioni (processi e malattie). La specificità di un miRNA m_k per un processo p_j è definita come segue:

$$S_{m_k,p_j} = \frac{\left|G_{m_k,p_j}\right|}{\left|G_{m_k}\right|} \cdot \frac{\sum_{g_i \in G_{m_k,p_j}} S_{g_i}}{\left|G_{m_k,p_j}\right|} = \frac{\sum_{g_i \in G_{m_k,p_j}} S_{g_i}}{\left|G_{m_k}\right|}$$

dove $G_{mk,pj}$ è l'insieme dei geni target del miRNA m_k coinvolti nel processo p_j, e G_{mk} èl'insieme di tutti i target del gene m_k. La specificità di un gene S_{gi} è inversamente proporzionale al numero di processi nei quali il gene è coinvolto:

$$S_{g_i} = \frac{1}{\left|P_{g_i}\right|}$$

dove P_{gi} è l'insieme dei processi nei quali g_i è coinvolto.

Intuitivamente, un gene associato a pochi processi è più focalizzato su di essi. La specificità di un miRNA per un processo si basa dunque sul numero dei suoi target e sulla loro specificità nei confronti di tale processo. Questa funzione è stata applicata all'insieme delle interazioni miRNA/target validate.

I sottoinsiemi di miRNA associati frequentemente sono visualizzati su tabelle che mostrano i miRNA ed i processi o le malattie ai quali sono associati, con i relativi score di specificità. Le righe delle tabelle sono colorate in base ai valori di specificità, secondo un intervallo che va dal blu (valori bassi) fino al rosso (valori alti) (Vedi Fig. 3.2).

Processes	(1) miR-124	(2) miR-137
G1 phase of mitotic cell cycle	0.000297	0.018519
negative regulation of transcription from RNA polymerase II promoter	0.000789	0.027778
positive regulation of cell-matrix adhesion	0.000297	0.018519
regulation of transcription, DNA-dependent	0.021306	0.027778
protein amino acid phosphorylation	0.004036	0.018519
cell cycle	0.004978	0.018519
signal transduction	0.00744	0.027778
regulation of gene expression	0.000297	0.018519
hemopoiesis	0.000297	0.018519
gliogenesis	0.000297	0.018519
cell dedifferentiation	0.000297	0.018519
regulation of erythrocyte differentiation	0.000297	0.018519
negative regulation of osteoblast differentiation	0.001061	0.018519
positive regulation of transcription from RNA polymerase II promoter	0.00126	0.027778
positive regulation of fibroblast proliferation	0.000297	0.018519
negative regulation of epithelial cell proliferation	0.000297	0.018519
cell division	0.002744	0.018519
Max (cluster)	0.021306	0.027778
Max (global)	0.075306	0.166667

Fig. 3.2 -Esempio di un sottoinsieme contenente 2 miRNA (miR-124 e miR-137), entrambi coinvolti in 17 processi. Per ogni associazione è mostrato lo score di specificità. Le caselle sono colorate in base a questo score: le caselle rosse indicano alti valori di specificità all'interno del sottoinsieme, mentre quelle blu indicano valori bassi. In questo caso, le associazioni più rilevanti nel sottoinsieme sono quelle tra miR-137 ed i quattro processi evidenziati in rosso (*negative regulation of transcription from RNA polymerase II promoter*, *regulation of transcription DNA-dependent*, *signal transduction* e *positive regulation of transcription from RNA polymerase II promoter*). Questo può suggerire un ruolo specifico di miR-137 in tali processi.

3.2.1 Casi d'uso e validazione

Il sistema è stato validato sulla base di alcuni casi noti presenti in letteratura.

3.3 Il cluster miR-17-92

Il ruolo cruciale del cluster miR-17-92 nello sviluppo e in diverse malattie è stato ampiamente dimostrato. L'espressione di questi miRNA promuove la proliferazione cellulare, sopprime l'apoptosi delle cellule tumorali ed induce l'angiogenesi dei tumori. In particolare, tali miRNA sono coinvolti nel linfoma, nel melanoma e in altri tipi di cancro (mammella, colon-retto, polmone, ovaio, pancreas, prostata e stomaco). Il cluster miR-17-92 gioca inoltre un ruolo essenziale durante lo sviluppo normale di cuore, polmoni e sistema immunitario [41].

Eseguendo una ricerca avanzata in miRò, per le malattie correlate ai miRNA del cluster, si trova che quattro di essi (miR-17, miR-19a, miR-19b e miR-92a) sono associati ai tumori sopra citati insieme ad altre patologie. Inoltre, la ricerca

avanzata per i processi che coinvolgono il cluster, restituisce, tra gli altri, l'angiogenesi, l'apoptosi, il ciclo cellulare, la crescita e la proliferazione cellulare e lo sviluppo di cuore e polmoni, confermando quanto già riportato in letteratura.

Altri casi

La ricerca avanzata per i processi associati ai miRNA miR-1, miR-206 e miR-133a, il cui coinvolgimento nell'attività muscolare è ben documentato, restituisce la loro correlazione al processo *contrazione muscolare*. Inoltre l'analisi di data mining rileva un'alta correlazione di miR-1 e miR-206, frequentemente associati in termini di processi e patologie comuni [138].

Analogamente, l'analisi di miR-124 e miR-137, coinvolti nel glioblastoma, mostra la loro associazione in termini di diversi processi, tra i quali la biogenesi [139].

3.4 Validazione della funzione di specificità

La funzione di specificità, introdotta nel paragrafo 3, ha l'obiettivo di valutare le annotazioni dei miRNA, per consentire l'identificazione delle associazioni più significative.

Tra le associazioni miRNA/malattia e miRNA/processo con score più alto, vi sono casi riportati in letteratura. Ad esempio, l'associazione miRNA/malattia più forte è quella che collega miR-433 al morbo di Parkinson. Questo risultato è confermato da uno studio nel quale viene mostrato come la delezione del sito di legame di miR-433 nel trascritto del gene FGF20, aumenta il rischio dell'insorgenza della malattia. Infatti, la sovra-espressione di FGF20 è correlata alla sovra-espressione dell'alfa-sinucleina, per la quale è documentata una diretta correlazione con la malattia di Parkinson [140].

Analogamente, l'associazione tra miR-224 e l'apoptosi è tra le associazioni miRNA/processo con score maggiore. Questa è supportata da uno studio nel quale si mostra la correlazione tra miR-224, sovra-espresso nel carcinoma epatocellulare, e l'aumento della morte cellulare attraverso il targeting dell'inibitore dell'apoptosi API-5 [141].

4 LE BANCHE DATI BIOLOGICHE

4.1 Introduzione

Una Banca Dati Biologica comprende un archivio di dati biologici, un'organizzazione logica di queste informazioni ed una serie di strumenti per accedere a queste ultime. Numerosi sono oggi i siti biologici esistenti, ciò in relazione allo sviluppo delle biotecnologie molecolari che hanno portato alla produzione diversificata di enormi quantità di dati biologici. Attualmente, le banche dati delle sequenze nucleotidiche contengono 16 x 109 basi (abbreviato in *16 Gbp*). La maggior parte delle banche dati biologiche disponibili risulta essere ben strutturata e di notevole supporto alla moderna ricerca biomolecolare, per cui in questo capitolo si intende offrire una panoramica generale di esse distinguendole per grandi categorie. Prima di passare alla loro descrizione, risulta tuttavia necessario definire alcuni elementi fondamentali che sono di supporto alla comprensione dei contenuti e dell'organizzazione delle banche dati. Una banca dati biologica raccoglie informazioni e dati derivati dalla letteratura e da analisi effettuate sia in laboratorio sia attraverso l'applicazione di analisi bioinformatiche. Ogni banca dati biologica è caratterizzata da un ***elemento biologico centrale*** che costituisce l'oggetto principale intorno al quale viene costruita la ***entry*** della banca dati. Nel caso delle banche dati di sequenze di acidi nucleici l'elemento centrale è la sequenza nucleotidica di *DNA* o *RNA* a cui vengono associate annotazioni classificanti l'elemento stesso quali: il nome della specie, la funzione e le referenze bibliografiche. Ciascuna *entry* raccoglie quindi le informazioni caratterizzanti l'elemento centrale, ossia i suoi attributi. Nell'organizzare una banca dati si devono definire i tipi di attributi da annotare e il formato con cui tali informazioni vengono organizzate; ciò costituisce la ***struttura della banca dati***. Nei primi anni di sviluppo della bioinformatica le banche dati biologiche erano organizzate solo in formato ***flat-file****,* cioè un file sequenziale nel quale ogni classe di informazione è riportata su una o più linee consecutive identificate da un codice a sinistra che definisce gli attributi annotati nella linea stessa. Esso è leggibile come un qualsiasi testo descrittivo di fatti biologici e analizzabile mediante programmi che mirano ad estrarre dalla banca dati informazioni specifiche (come vedremo in seguito).

Successivamente, a seguito della notevole crescita dei dati e della complessità delle relazioni fra i dati stessi, si è reso necessario adottare *DataBase Management Systems* (*DBMS*), indispensabili per il disegno di banche dati complesse che permettono una completa integrazione fra i dati. In relazione ai dati forniti dalle banche dati, nasce l'esigenza di averne un accesso immediato e trasparente per ottenere informazioni distribuite anche fra banche dati eterogenee nei contenuti e nella struttura, e localizzate su siti distanti tra loro. A tale scopo sono stati creati flat-file aventi linee di *crossreferencing* (riferimento incrociato). Tali linee, che nella maggior parte delle banche dati biologiche organizzate in formato flat-file *EMBL* (*European Molecular Biology Laboratory*) vengono individuate dal codice *DR*, pongono in relazione dati annotati in una entry di una specifica banca dati con dati presenti in altre entry di altre banche dati. Attraverso l'implementazione di interfacce grafiche su internet per dati in relazione fra essi, ne è consentita la loro visibilità mediante l'uso dell'*hypertext link* disposto per ciascuna linea di cross-referencing. Con il susseguirsi degli sviluppi, a questi sistemi di integrazione dei dati se ne sono aggiunti degli altri più complessi come ad esempio quello per il rilascio della banca dati in formato ***XML*** (*e**X**tensible **M**arkup **L**anguage*) che è un linguaggio analogo al linguaggio *HTML* e consente di formattare, in un file di testo, sia la struttura della banca dati sia la semantica con la quale i dati sono organizzati. Come in un testo HTML, i dati di rilevante importanza sono marcati. Ciò rende una banca dati strutturata in XML facilmente integrabile via web in qualsivoglia sistema. Prerequisito fondamentale alla funzionalità di tale sistema è la standardizzazione dei nomi da assegnare a ciascun elemento descrittivo della struttura della banca dati. Pertanto, una banca dati biologica comprende un archivio di dati, un'organizzazione logica di queste informazioni ed una serie di strumenti per accedere alle stesse.

I siti biologici oggi esistenti sono distinguibili in due macro-categorie:

1. *banche dati **primarie*** relative a sequenze nucleotidiche e amminoacidiche; queste contengono informazioni molto generiche, ovvero quel minimo di informazioni associate alla sequenza per identificarla dal punto di vista *specie-funzione*;
2. *banche dati **specializzate*** relative a domini e motivi proteici, strutture proteiche, geni, trascrittoma, profili di espressione, pathways metaboliche e

altro. Queste banche dati derivano dalle primarie mediante l'estrazione di un sottoinsieme funzionale di dati che sono a loro volta caratterizzati dall'aggiunta di informazioni supplementari. Le tre più importanti banche dati primarie che operano al servizio della scienza biologica sono:

- **GenBank** di **NCBI**, *National Center for Biotechnology Information*;
- **DDBJ**, *DNA Database of Japan*;
- **EMBL-NSD**, *European Molecular Biology Laboratory – Nucleotide Sequence Database*;

Questi istituti cooperano tra loro al fine di condividere e rendere pubblicamente disponibili tutti i dati di cui dispongono e differiscono tra loro solamente per il loro formato:

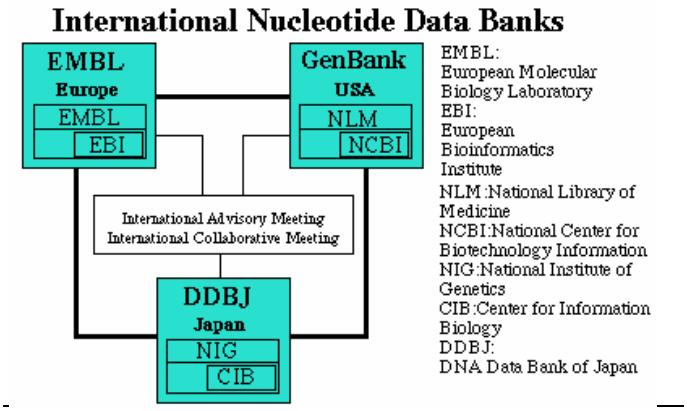


Figura 4.1 I tre database biologici

4.2.1 NCBI

Nato nel 1988 come risorsa nazionale per la biologia molecolare, NCBI implementa e gestisce database ad accesso libero, induce attività di ricerca scientifica nel campo della biologia computazionale e sviluppa software per l'analisi genomica dei dati. Tutto il suo lavoro è mirato allo scopo di fornire un valido supporto alla comunità scientifica internazionale. NCBI è il database americano del *NIH* (*National Institute of Health*) che rispetto agli altri database pubblici offre anche la possibilità di effettuare ricerche di tipo bibliografico e, soprattutto, di ospitare e gestire varie banche dati attraverso un sistema di ricerca integrato chiamato ***Entrez***. Questo è un potente motore di ricerca che permette una ricerca

contemporanea su differenti database biomedici, tra cui ***PubMed***, ***OMIM***, ***Gene***, ***Nucleotide and Protein***, ***Protein Structures***, ***Taxonomy***, ed altri che vedremo in seguito. Gli utenti possono avere accesso a numerose informazioni quali: sequenze nucleotidiche o proteiche, mappe, informazioni tassonomiche, dati strutturali di macromolecole. Tuttavia Entrez è una shell chiusa in quanto non è possibile scaricarsi via internet o ottenere in qualche modo il software che gestisce l'intero sistema e quindi non è possibile duplicare su altri computer il sito Entrez, ne è possibile installare in Entrez banche dati personali. Tra i programmi sviluppati e accessibili si trova ***BLAST*** (*Basic Local Alignment Search Tool*), un algoritmo che è in grado di eseguire confronti fra coppie di sequenze alla ricerca di regioni di similarità, o di confrontare tra loro due sequenze esterne immesse dal ricercatore.

4.2.2 DDBJ

Il database DDBJ nasce nel 1986 al *NIG* (*National Institute of Genetics*) allo scopo di ospitare attività di interrogazione su banche dati di sequenze di DNA, ed è il frutto della collaborazione con le altre due grandi famiglie, EMBL e NCBI. Il database virtuale unificato prende il nome di ***INSD (International Nucleotide Sequence Database)*.** La DDBJ e l'unica banca dati di sequenze nucleotidiche in Asia e riceve i suoi dati principalmente da ricercatori nipponici.

4.2.3 EMBL-NSD

EMBL-NSD conosciuto anche come ***EMBL-Bank*** o ***EMBL- EBI***, è nato dalla collaborazione di *GenBank* (USA) e di *DDBJ*. E' ospitato all'EBI (*European Bioinformatics Institute*) che ne è il diretto responsabile. L'EMBL è stato fondato nel 1974 e annovera fra i suoi membri numerosi paesi europei tra i quali anche l'Italia. L'EMBL promuove la collaborazione europea nell'ambito della ricerca fondamentale condotta in Biologia Molecolare, mette a disposizione l'infrastruttura necessaria e partecipa al continuo sviluppo di strumenti di alta qualità tra cui ***Ensembl Genome Browser***. *Ensembl* è un progetto open source che organizza e gestisce informazioni biologiche sui principali genomi eucariotici.

Questo framework integra qualsiasi informazione biologica che può essere rappresentata come caratteristica di una sequenza genomica. Ensembl è gestito in modo coordinato dal *Wellcome Trust Sanger Institute* e dallo *European*

Bioinformatics Institute ed è disponibile attraverso un sito web o come un insieme di flat-file oppure ancora sottoforma di un sistema software open source di genomi [157]. In questo modo sia il software che i dati immagazzinati possono essere utilizzati in maniera del tutto libera e gratuita. Attualmente Ensembl riporta i genomi della maggior parte dei vertebrati e di diversi organismi modello.

4.3 Banche dati Specializzate.

Oltre alle banche dati primarie descritte nei paragrafi precedenti, sono numerosissime le banche dati *specializzate*. Quest'ultime raccolgono insiemi di dati omogenei dal punto di vista tassonomico e/o funzionale disponibili nelle banche dati primarie e/o in letteratura, o derivanti da vari approcci sperimentali, rivisti e annotati con informazioni di valore aggiunto. Vediamo di seguito i più importanti tra essi, concentrandoci sui database di microRNA.

4.3.1 Database di MicroRNA: MirBase

MirBase è un progetto sviluppato dal *Wellcome Trust Sanger Institute* e fornisce un'interfaccia semplice e integrata di informazioni sui miRNA, sui geni target predetti e sulle relative annotazioni. E' disponibile all'indirizzo http://microrna.sanger.ac.uk/.

Principalmente è costituito da tre parti:

- ***miRBase Registry***, è considerata come un arbitro della nomenclatura dei miRNA; tramite esso infatti vengono assegnati i nomi alle nuove sequenze miRNA appena pubblicate;
- ***miRBase Sequences***, è la banca dati primaria contenente tutte le sequenze dei miRNA, le posizioni genomiche e le relative annotazioni;
- ***miRBase Targets***, è una banca dati contenente geni target predetti attraverso miRNA appartenenti a varie specie. Le sequenze miRNA sono ottenute dal database miRBase Sequences, mentre le sequenze genomiche da Ensembl.

Lo schema dei nomi adottato da MiRBase Registry consiste nel distinguere l'identificatore del miRNA in tre parti:

1. la prima parte è formata generalmente da 3 o 4 lettere e rappresenta la ***specie di appartenenza***. Ad esempio un identificatore in *Homo sapiens* (uomo) può essere del tipo hsa-miR-101, mentre in *Mus musculus* (topo) può essere del tipo mmu- miR-101;
2. la parte centrale contiene l'etichetta *miR*, per identificare la ***sequenza matura*****,** oppure *mir* per la ***sequenza precursore***;
3. la parte finale è costituita da un numero che può contenere ***suffissi letterali o numerici***; quando più sequenze mature differiscono di una o due posizioni, i numeri dei due identificatori sono seguiti da suffissi letterali, come ad esempio *mmu-miR-10a* e *mmu-miR-10b*; quando distinti precursori forniscono lo stesso miRNA maturo i loro identificatori contengono suffissi numerici, come ad esempio *dme-mir-281-1* e *dme-mir-281-2* in *Drosophila melanogaster.*

L'assegnamento dei nomi diviene più complicato quando differenti miRNA maturi sono estratti da bracci opposti dello stesso precursore a forcina ed in questo caso gli identificatori hanno suffissi del tipo *5p* per indicare che il braccio usato è 5' e *3p* per il braccio 3'.

Lo scopo è quello di associare all'identificatore del miRNA un'informazione significativa semplificata che indica l'organismo di appartenenza, il tipo di sequenza e le relazioni funzionali tra i miRNA maturi, come ad esempio *hsa-miR-101* e *mmu-miR-101* che sono ortologhi. Il database MiRBase Sequences contiene le sequenze di tutti i miRNA maturi pubblicati e quelle dei loro precursori a forcina predetti, le annotazioni corrispondenti alla loro scoperta, la loro struttura e la loro funzione. Il database è in continua crescita e attualmente contiene 4361 entries di precursori a forcina, 4167 miRNA maturi espressi nei primati, roditori, uccelli, pesci, vermi, insetti, piante e virus (Release 9.0 ottobre 2006). Poiché i nomi dei miRNA potrebbero cambiare nel corso tempo, riflettendo così le nuove relazioni fra le sequenze, il database fornisce degli accession numbers stabili che vengono assegnati sia ai miRNA precursori, ad esempio MI0000015, sia ai miRNA maturi, ad esempio MIMAT0000029, ognuno rappresentante l'entità sequenza. Il database ricava i miRNA principalmente da due fonti: quella dei miRNA *sperimentalmente verificati*, forniti da articoli provenienti dalla letteratura e dai metodi sperimentali

usati durante la scoperta, e quella dei miRNA non verificati, che sono sequenze predette omologhe ai miRNA verificati in un organismo affine. Ad esempio 223 su 313 miRNA maturi distinti nell'uomo (circa il 71%), sono provati sperimentalmente nell'uomo stesso, mentre il rimanente è costituito da omologhi identificabili da miRNA verificati nel topo (Mus musculus), nel ratto (Rattus norvegius) e nel pesce zebra (Danio rerio). Gli omologhi sono predetti secondo la similarità tra le sequenze, le caratteristiche del precursore a forcina, l'analisi e la conservazione del miRNA maturo [158]. In seguito è illustrata una pagina web ottenibile digitando nel form di ricerca di miRBase Sequences l'identificativo hsa-mir-25. Le tre sezioni della pagina descrivono il precursore a forcina predetto, la sequenza matura e gli articoli principali. Le coordinate genomiche e le altre informazioni si collegano al database Ensembl. Ogni sequenza matura contiene un campo evidence e vari links alle pagine dei target.

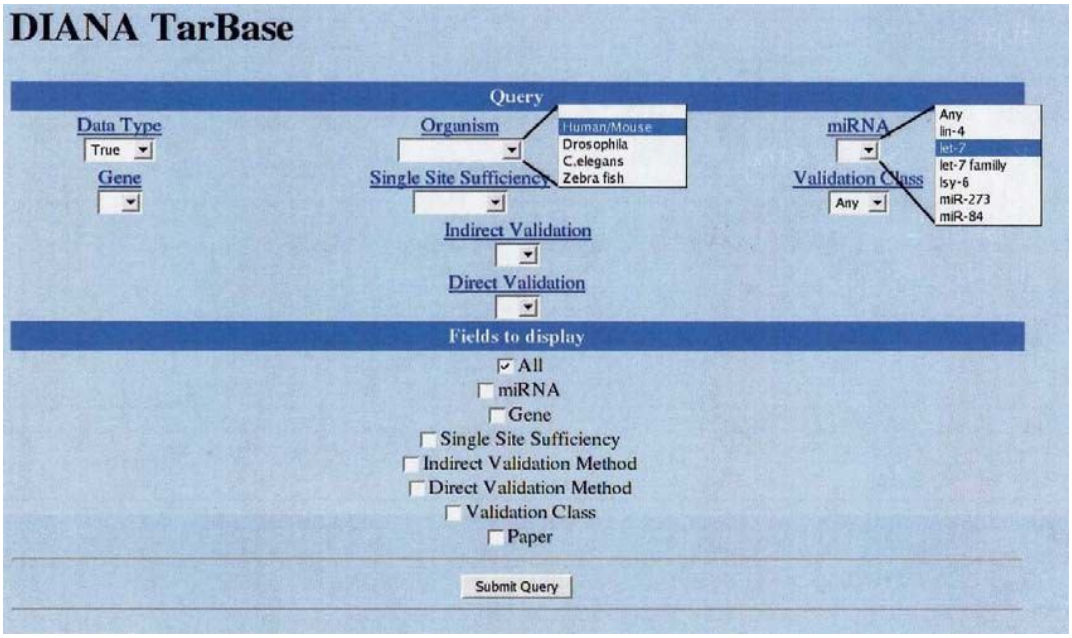
Mirbase

4.3.2 TarBase

TarBase è una banca dati che colleziona *target* per miRNA validati sperimentalmente nell'uomo, nel topo, negli insetti della frutta, nei vermi e nel pesce zebra, distinguendo quelli verificati positivamente (*true*) da quelli verificati negativamente (*false*). Ogni sito target positivo trovato è descritto secondo il miRNA ad esso collegato, il gene nel quale occorre, dalla natura degli esperimenti condotti durante il test, dalla sufficienza del sito a indurre l'inibizione traduzionale e/o l'operazione di taglio dell'mRNA e dal relativo articolo che contiene tutte queste informazioni [159]. TarBase è liberamente disponibile on-line all'indirizzo http://www.diana.pcbi.upenn.edu/tarbase e principalmente descrive due tipi di geni:

- ***translationally repressed miRNA targets*** ovvero geni target che reprimono la traduzione;
- ***cleaved miRNA targets*** ovvero geni che tagliano l'RNA messaggero.

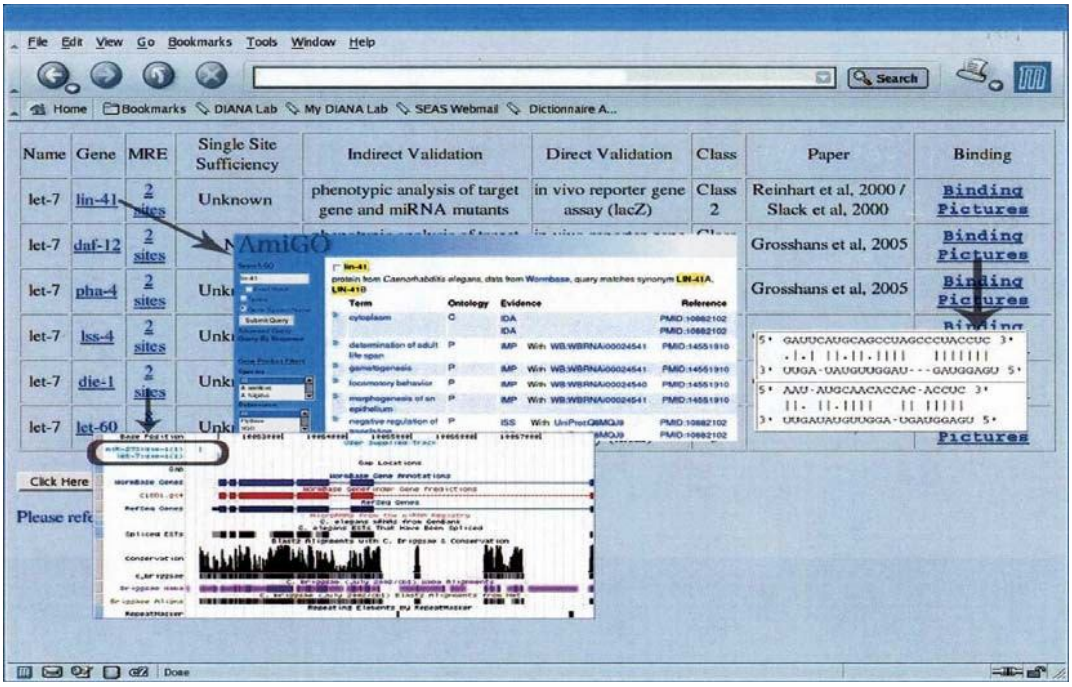
Ai primi appartengono almeno 45 geni dell'uomo e del topo, 28 geni degli insetti della frutta, 7 geni dei vermi e un gene del pesce zebra, mentre i secondi descrivono circa 350 geni dell'uomo e del topo e 3 geni del verme. Il numero complessivo dei siti target fornito da TarBase nell'uomo, nel topo, negli insetti della frutta, nei vermi e nel pesce zebra supera circa 550 geni che rappresentano un valore abbastanza grande da tenere in considerazione nei successivi programmi di predizione dei target. L'interfaccia grafica di ricerca al database messa a disposizione dell'utente è fornita nella figura di seguito. L'utente può effettuare *query* al database considerando singole informazioni o una combinazione di esse, quali l'organismo di appartenenza, il gene nel quale occorre, il miRNA ad esso collegato, le tecniche usate durante il test e altro. Inoltre l'utente può scegliere quali campi visualizzare.



L'interfaccia grafica messa a disposizione all'utente per interrogare il database TarBase

TarBase è funzionalmente collegato con altri database utili quali *Gene Ontology* (GO, che analizzeremo in seguito) e *UCSC Genome Browser*: un gene target sperimentalmente validato e linkato a GO per fornire un elenco chiaro dei processi biologici regolati; i siti target specifici all'interno del gene target sono linkati a UCSC Genome Browser per ottenere le loro locazioni genetiche e le loro sequenze; infine ogni sito target è linkato ad un database interno che contiene tutti gli allineamenti *miRNA- gene tar*get.

Di seguito viene riportata la pagina di risultato ottenuta dopo la ricerca effettuata precedentemente per mostrare le varie funzionalità del database TarBase. I geni target sono linkati a Gene Ontology; gli elementi riconosciuti del miRNA (miRNA Recognition Elements – MRE) sono linkati a UCSC Genome Browser attraverso le usuali tracce e gli allineamenti sono visibili in una nuova pagina.



Risultato della ricerca su TarBase

Gli utenti interessati possono scaricare per intero tutti i contenuti del database sotto forma di un file "*.tar*" che viene continuamente aggiornato e possono pubblicare nuovi target appena identificati, aiutando in questo modo la comunità scientifica di ricerca sui miRNA [159]. Tramite TarBase il ricercatore ha dunque la possibilità di andare alla ricerca di dati biologici, di filtrarli, di manipolarli e di sottoporli a tanti tools di elaborazione biologica.

4.4 Banche dati di motivi e domini proteici

Le banche dati sono utili sia per effettuare ricerche testuali, in modo tale da estrarre informazioni specifiche relativamente a un dato argomento, sia per la comparazione al fine di individuare, in nuove sequenze, caratteristiche strutturali e funzionali già riscontrate in altre sequenze ed annotate in specifiche banche dati. Dal punto di vista bioinformatico la più importante è la comparazione, che può essere effettuata attraverso l'applicazione di tecniche di ricerca di similarità (*metodi euristici di allineamento*) o, quando la ricerca di similarità non evidenzia sequenze simili a quelle in oggetto, attraverso l'applicazione di tecniche di ricerca di segnali (*pattern recognition*) basate su algoritmi più o meno complessi; questo secondo approccio consente di ritrovare segnali, motivi o domini strutturali e funzionali che si conservano nel tempo. Numerose banche dati specializzate che annotano informazioni relative a motivi e domini funzionali sono state integrate in ***InterPRO***, una risorsa bioinformatica, sviluppata all'EBI, che consente di ricercare contemporaneamente informazioni funzionali e strutturali relative a una proteina o a una famiglia di proteine su più banche dati, distribuite su calcolatori diversi e strutturate in modo differente. Attraverso il software InterPROscan è possibile ricercare motivi strutturali e funzionali annotati nelle banche dati integrate in InterPRO al fine di caratterizzare dal punto di vista funzionale nuove proteine derivate da progetti di sequenziamento genomico. Di seguito vengono elencate le più importanti banche dati integrate in InterPRO.

PROSITE

Sviluppata e mantenuta dal gruppo *Amos Bairoch*, questa banca dati annota patterns amminoacidici individuati in set di sequenze proteiche determinati sperimentalmente in una o più proteine. PROSITE contiene motivi codificati in due modi diversi: i *pattern* e le *matrici*. I pattern sono motivi definiti con una sintassi riconducibile a espressioni regolari mentre le matrici sono definite facendo ricorso alle matrici posizionali di peso. ***ProDOM*** è un database che raccoglie dati relativi alle famiglie di proteine generate automaticamente dall'applicazione di *PSI-BLAST* che, partendo dal confronto di una sequenza proteica (*sequenza sonda*) contro un database di proteine, raccoglie in un multi-allineamento tutte le sequenze proteiche per le quali BLAST ha determinato uno score più alto di una certa soglia

prestabilita (*threshold*); il multi-allineamento risultante genera un profilo che viene utilizzato per rilanciare il BLAST contro tutto il database di proteine allo scopo di individuare nuove sequenze correlate a quelle già allineate, che vengono aggiunte al multi-allineamento per ottimizzare ulteriormente il profilo. Questa procedura viene ripetuta un numero prefissato di volte o fino a quando non si raggiunge una certa convergenza.

Pfam

E' una banca dati di famiglie di proteine accomunate da elementi strutturali e funzionali. Ogni entry in Pfam è caratterizzato da un tipo che può essere *"famiglia"*, *"dominio"*, *"repeat"* o *"motivo"*. Il tipo *famiglia* consente di raggruppare sequenze proteiche accomunate dagli stessi *domini*; il tipo *dominio* definisce un'unità strutturale che può comunque essere presente in *famiglie* differenti; il tipo *repeat* raggruppa elementi funzionali attivi e presenti in copie multiple in proteine globulari; il tipo *motivi* include patterns componenti blocchi strutturali non associati a proteine globulari.

La definizione dei limiti di ciascun dominio annotato in Pfam è ottenuta dal database *SCOP*. Le famiglie di proteine non classificabili secondo i criteri sopradescritti ma che comunque sono state prodotte automaticamente attraverso l'applicazione di PSI-BLAST e quindi annotate in ProDOM, sono annotate in Pfam nel sottoinsieme *Pfam-B*; un database meno accurato ma comunque di supporto all'analisi proteomica.

PRINTS

Questo database raccoglie sequenze proteiche in *clusters* definiti da un comune *fingerprint*, dove per fingerprint si intende l'insieme di più motivi conservati e dedotti dall'osservazione di un multi-allineamento ottenuto applicando algoritmi per la ricerca di similarità locali. I clusters sono classificati in una forma gerarchia che definisce le *superfamiglie*, le *famiglie* e le *sottofamiglie* e associa famiglie correlate nella funzionalità. Il numero di famiglie annotate in PRINTS è ridotto rispetto a Pfam e ProDOM in quanto i dati, preliminarmente prodotti in modo

automatico, successivamente sono rivisti manualmente e annotati con dati biologici derivati dalla letteratura e da ulteriori analisi.

SMART

Acronimo di *Simple Modular Architecture Research Tool*, è una risorsa che raccoglie dati relativi a domini proteici e consente la ricerca di domini in nuove sequenze proteiche. SMART annota, per ogni famiglia di proteine associate a un dominio, informazioni sulla funzione, sulla localizzazione cellulare, sulla struttura terziaria in cui è coinvolto il dominio e su relazioni filogenetiche fra le specie da cui sono derivate le proteine componenti la famiglia. SMART cura particolarmente domini associati a elementi mobili presenti nei genomi eucariotici.

TIGRFAMs

E' una collezione di famiglie di proteine prodotta mediante annotazione biologica di semplici multi-allineamenti. TIGFRAMs è sviluppata presso il *TIGR* (*The Institute for Genomic Research*).

4.5 Banche dati di strutture proteiche

La conoscenza di motivi strutturali delle proteine è di grande importanza per la comprensione funzionale delle biosequenze. Per dati strutturali di una proteina si intende la distribuzione spaziale degli atomi componenti gli amminoacidi e quindi degli amminoacidi stessi.

PDB

La PDB (*Protein Data Bank*) è l'unica banca dati che raccoglie informazioni relative alle proteine e a dicembre 2002 riportava più di 16000 strutture proteiche. Fra le altre strutture sono da annoverare i complessi di proteine con acidi nucleici, le strutture di acidi nucleici (DNA e RNA) e carboidrati e modelli di strutture derivati con modelli computazionali. Tale banca dati è un riferimento unico per tutti gli studi strutturali. Ogni file della banca dati è identificato da 4 caratteri: un numero e

tre caratteri alfa-numerici. Il numero identifica la versione del file e i caratteri restanti sono scelti, ove possibile, in modo da ricordare i nomi delle strutture descritte nel file stesso. Ogni file inoltre è diviso in due parti: la prima parte contiene la descrizione della molecola contenuta e la risoluzione della sua struttura; la seconda parte contiene le coordinate atomiche. Le strutture descritte nei files PDB possono essere visualizzate con un programma di grafica molecolare.

MMDB

La MMDB (*Molecular Modeling DataBase*) contiene strutture di biomolecole ricavate dal PDB escludendo i modelli teorici, e le archivia in un diverso formato (ASN.1, *Abstract Syntax Notation One*). Le strutture ivi presenti contengono valore aggiunto consistente in informazioni derivate da diverse procedure di validazione, definizione uniforme delle strutture secondarie e la sequenza di ogni catena come derivata delle coordinate.

DSSP

La DSSP (*Dictionary of Protein Secondary Structure*) contiene informazioni sulle strutture secondarie di ogni entry del PDB. In particolare, per ogni file di coordinate del PDB, esiste un file DSSP con lo stesso codice.

HSSP

La HSSP (*Homology derived Secondary Structure of Proteins*) è ricavata dai files del PDB e contiene informazioni utilissime per costruire modelli di proteine a struttura *non nota* con proteine a struttura *nota*.

FSSP

La FSPP (*Fold classification based on Structure-Structure allignment of Proteins*) presenta a tutte le catene delle proteine del PDB i risultati dell'applicazione *DALI*, un programma che paragona mappe di contatti tra i carboni alfa di coppie di proteine. Ogni file della FSSP include l'allineamento con le proteine di struttura simile e riporta i residui che sono equivalenti nelle strutture.

SCOP

SCOP (*Structural Classification Of Proteins*) organizza le strutture proteiche gerarchicamente seguendo criteri evolutivi e di similarità strutturale. Si basa su domini e li raggruppa in famiglie di domini simili.

CATH

E' un database di classificazione gerarchica di domini di strutture proteiche del PDB. Presenta una classificazione simile al quella offerta da SCOP, ma di strutture con risoluzione migliore di 4.0 angstrom.

4.6 Banche dati biologiche per il sistema immunitario

L'immunologia è una branca della moderna ricerca biomedica che si basa, tra le altre cose, sullo studio funzionale e strutturale delle macromolecole biologiche e sull'analisi di variabilità molecolare associata alle risposte immunitarie. Vediamo di seguito una rapida carrellata dei più rilevanti database del settore immunologico.

IMGT

IMGT è il database internazionale di *ImmunoGenetica* e accoglie dati relativi alle Immunoglobuline, ai recettori delle cellule *T (TCR)* e al maggiore complesso di istocompatibilità (*MHC*) di classe I e II. Il database riporta dati relativi alle sequenze, ai genomi, alle strutture e alla variabilità delle macromolecole immunologiche umane e di altri vertebrati. Il sito di IMGT consente di accedere al database per effettuare ricerca di dati, ricerca di similarità e altre specifiche analisi in silico.

MHCpep

E' un database che annota i dati di sequenza dei peptidi che legano molecole di MHC (molecole costituenti il complesso di maggiore istocompatibilita) di uomo,

topo e in minima parte anche di ratto e di altri primati. Ad ogni entry è associato un determinato peptide che lega uno specifico allele MHC. Sono annotate anche informazioni sull'attività di legame e sui metodi con cui i peptidi sono stati determinati.

FIMM

E' un database di antigeni, molecole MHC, peptidi associati alle molecole MHC e dati correlati a patologie. A differenza di MHCpep che è un database disponibile in formato flat-file, FIMM è strutturato in un pacchetto chiuso, secondo gli schemi delle cosiddette *data-warehouse* che consentono la ricerca e l'analisi dei dati esclusivamente secondo percorsi pre-progettati dal produttore del pacchetto stesso.

MPID

L'MPID (*MHC Peptide Interactions DB*) annota informazioni relative alle correlazioni *sequenza-struttura-funzione* per i peptidi che legano MHC. MPID riporta in particolare tutte le strutture delle proteine contenenti peptidi che legano i complessi MHC e informazioni sulla caratterizzazione strutturale delle interazioni complesso-peptidi.

4.7 Banche dati di geni

Numerose banche dati di geni sono state sviluppate a partire prevalentemente da dati genomici o comunque da dati annotati nelle banche dati primarie.

LocusLink

E' uno dei database sviluppati da NCBI. Vengono annotati, per ogni *locus genetico* (ogni elemento funzionale di un genoma), il nome ufficiale ed eventuali sinonimi, il codice della classificazione internazionale degli enzimi, se trattiamo degli enzimi, il link a OMIM *(Online Mendelian Inheritance in Man*), gli *accession numbers* delle sequenze nucleotidiche associate al locus e annotate nelle banche dati primarie e il link alle banche dati RefSeq e UniGene.

Gene Ontology

Gene Ontology (*GO*) [161] è una banca dati che raccoglie le descrizioni funzionali dei geni. Gli autori hanno sviluppato tre strutture, note come *ontologie*, che descrivono le caratteristiche dei geni di una determinata ***COGs***. Questo database riporta una compilazione di geni ortologhi codificanti proteine relativi a organismi completamente sequenziati oppure clusters di geni paraloghi conservati in almeno 3 organismi differenti e significativamente distanti fra loro.

GENES

GENES annota le informazioni relative a tutti i geni identificati sui genomi completi sia di procarioti sia di eucarioti. Il grande vantaggio rispetto ai database primari e la certezza di ritrovare tutti i geni noti, anche quelli recentemente identificati attraverso l'analisi *in silico*.

EuGENES

E' una banca dati di geni e genomi relativi a 7 organismi eucariotici e descrive circa 150.000 geni noti, predetti o non classificati. Sono in oltre annotate informazioni sulle mappe genomiche degli organismi considerati.

4.8 Banche dati di pattern nucleotidici

Complementari alle banche dati dei geni sono le banche dati di *patterns nucleotidici* o di *regioni funzionali del gene* associati a specifiche funzioni regolatorie e di controllo. Le più influenti sono illustrate di seguito.

EPD

EPD (*Eukaryotic Promoter Database*) è una delle prime banche dati specializzate progettata. Essa annota le info bibliografiche e sperimentali sui promotori eucariotici (nome della entry EMBL, localizzazione del sito di inizio della trascrizione e la descrizione del tipo di sito).

TRANSFAC

TRANSFAC è la banca dati dei fattori di trascrizione che annota dati sui fattori proteici e sui corrispondenti siti di legame sul DNA coinvolto nell'attivazione e la regolazione della trascrizione. Dal suo sito è possibile ottenere una scheda in formato flat-file con le caratteristiche dell'elemento e contenente anche cross-referencing ad altre banche dati fra cui PROSITE.

UTRdb

Essa svolge un ruolo importante poiché annota tutte le sequenze non tradotte dei messaggeri eucariotici derivate dalla banca dati primaria EMBL con ulteriori informazioni relative alla descrizione dei siti funzionali caratterizzati su tali regioni.

TRANSTERM

TRASTERM è la banca dati degli elementi che regolano la traduzione e le modificazioni post-traduzionali. Gli elementi sono classificati dal punto di vista funzionale e strutturale, raggruppando gli elementi in categorie. Il database è generato a partire dalla banca dati GenBank dalla quale vengono estratte le varie regioni annotate nelle features tables e quindi riviste per eliminare ridondanze ed aggiungerci valore.

TRANScompel

TRANScompel è la banca dati degli elementi compositi coinvolti nella regolazione della trascrizione. Elementi regolatori compositi (CE) annotano due siti di legame situati in posizioni vicine nella unità trascrizionale, che legano due distinti fattori di trascrizione controllando in modo combinato la regolazione della trascrizione.

4.9 Banche dati del trascrittoma

Nell'evoluzione dei progetti genomici si sta sempre più diffondendo la tendenza a raggruppare le categorie di dati biologici in *omics* e in tale contesto rientra il database del trascrittoma costituito dall'insieme di tutti i trascritti di un dato organismo ottenuti attraverso il sequenziamento delle *EST* (*Expressed Sequenze Tags*) o dei cDNA completi. Qui di seguito alcune delle più importanti.

dbEST

In questo database vengono raccolti tutti i dati relativi alle EST (Expressed Sequence Tags), ottenute tramite il sequenziamento parziale di cloni di cDNA. La quantità di EST prodotte per ciascun gene e generalmente direttamente proporzionale al suo livello di espressione in relazione anche al fenotipo o al tessuto-specificità del gene. Il database dbEST è anch'esso sviluppato all'NCBI.

FANTOMdb

Questo database raccoglie dati di cDNA del topo, favoriti dall'evolversi delle tecnologie di sequenziamento.

UniGENE

Raggruppa sequenze geniche trascritte, dedotte da sequenziamento di cDNA o di EST di uomo, topo, ratto, *Drosophila*, *Anopheles*, *danio renio*, *Arabidopsis* e altri organismi modello, in clusters teoricamente corrispondenti ad un singolo gene, attraverso criteri di similarità o provenienza da uno stesso clone. Per ogni gene UniGENE riporta la sua localizzazione cromosomiale in ciascun organismo e tessuto in cui esso si esprime.

4.9.1 Banche dati di profili di espressione

La tecnologia dei *microarrays* permette in un solo esperimento di quantificare i trascritti di un intero genoma (il trascrittoma) e quindi di confrontare la variabilità di espressione di ciascun gene in tessuti diversi, in individui diversi, in stati patologici diversi (in altre parole consente di associare il livello di espressione di un gene al corrispondente fenotipo). Al mondo vi sono numerose banche dati o siti web in cui sono raccolti i profili di espressione con i dati relativi agli esperimenti. Analizziamo di seguito alcuni modelli molto utili.

Mouse GXD resource

E' un database integrato costituito da *Gene Expression Database* (GXD, *Anatomy Database* e *3D Atlas*).

PEDB

Il PEDB (*Prostate Expression Database*) è un database che raccoglie dati sull'espressione di geni connessi con studi delle patologie prostatiche dell'uomo e del topo.

GEO

Molte delle risorse dei profili di espressione sono prodotte in modo non coordinato. Solo recentemente sono stati realizzati progetti aventi l'obiettivo di coordinare meglio la raccolta dei dati relativi alle espressioni dei geni. Uno dei progetti prende il nome di GEO (*Gene Expression Omnibus*) ed è sviluppato da NCBI. Esso è una risorsa eterogenea per la sottomissione e il retrieval di dati correlati a esperimenti basati sulla tecnologia dei microarrays e preposti allo studio di espressione di geni e di ibridizzazione fra genomi. I dati sono classificati in tre categorie: *platform* (dati su tutte le sonde molecolari identificative di ciascuno spot per l'allestimento di un microarray), *samples* (dati sulle molecole che devono essere analizzate) e *series* (tutti i dati relativi a un esperimento).

ArrayExpress

ArrayExpress è l'equivalente europeo di GEO e raccoglie dati eterogenei su profili di espressione. E' strutturato utilizzando il *DMBS Oracle* secondo una definizione a oggetti. Riporta tutti i dati su interi esperimenti e anche le immagini non elaborate del profilo come viene prodotto dall'esperimento. Il database può essere interrogato attraverso un sistema semplice di ricerca testuale ed è interfacciato al sistema *Expression Profiler* che consente di analizzare i profili di espressione e di effettuare confronti tra differenti esperimenti.

KEGG/Expression

E' un database che raccoglie dati sui profili di espressione ottenuti con la tecnica dei microarrays in vari laboratori giapponesi. Presenti i profili di espressione relativi ai genomi di *Synechocytis* e *Bacullus subtilis*.

4.9.2 Banche dati di polimorfismi e mutazioni

Lo studio di un solo genoma di un unico organismo non è esaustivo in quanto non consente la valutazione della variabilità genica all'interno della specie, dovute a mutazioni o polimorfismi. Il termine ***mutazione*** indica la differenza *puntuale* evinta in un campione rispetto al genoma di riferimento a causa di disfunzioni di un gene e quindi di manifestazioni di fenotipi patologici. Il termine ***polimorfismo*** invece indica l'evento che lascia inalterata la funzionalità del gene. Una variazione che in una popolazione si riscontra con una frequenza superiore all'1% e considerata polimorfismo. Recentemente è stato introdotto un nuovo termine o meglio acronimo: *SNP* (*Single Nucleotdice Polymorphism*) e che dovrebbe indicare tutti i polimorfismi associati al cambiamento di un solo nucleotide.

HGMD

L'HGMD (*Human Gene Mutation Database*) raccoglie dati sulle mutazioni riportate come causa di alterazioni e disfunzioni dei geni nucleari in malattie ereditarie. Non vengono annotate mutazioni somatiche o del DNA mitocondriale; inoltre sono annotate solo mutazioni sperimentalmente determinate sul DNA e non sulla proteina. Per evitare confusioni tra mutazioni frequenti e ereditarie, ogni

mutazione è annotata una sola volta nella banca dati. Questo impedisce però di effettuare valutazioni statistiche di variabilità in base ai dati annotati in HGMD.

OMIM

OMIM *(Online Mendeliam Inheritance in Man*) raccoglie informazioni sulle malattie genetiche di origine *mendeliana*. In essa sono raccolti dati non solo sulle malattie genetiche di origine autosomica ma anche sulle malattie associate ad alterazioni dei cromosomi X e Y del mitocondrio.

Genes and Diseases

Sviluppata all'NCBI, Genes and Disease è una risorsa di dati sviluppata in base alla patologia, da cui si arriva al gene e ad informazioni correlate annotate in altre banche dati fra cui OMIM.

GAD

Genetic Association Database (*GAD*) [160] è una banca dati che colleziona le informazioni riguardanti le associazioni tra le malattie e i geni coinvolti nell'uomo rendendole accessibili alla comunità scientifica mediante una semplice interfaccia web. Lo scopo del GAD è quello di archiviare i risultati ottenuti dagli studi scientifici pubblicati e fornire uno standard sulla nomenclatura. Ogni record del database, che mette in evidenza la relazione tra il gene e le malattie connesse, è caratterizzato da links verso altri database, quali ad esempio quelli contenenti gli articoli pubblicati, quelli che si riferiscono agli studi molecolari nei dettagli e altri ancora. GAD, inoltre, mette a disposizione degli utenti alcuni tools per aggiungere, modificare e scaricare i dati da e verso il database.

Pharmacogenetics

Pharmacogenetics è una risorsa creata da una rete di laboratori di ricerca per la raccolta integrata di dati genomici, clinici e descrittivi del fenotipo. Prende vita dalla necessità di valutare sul singolo paziente la specifica risposta a un dato farmaco in relazione al proprio genoma.

4.9.3 Banche dati di pathways metabolici

Questi tipi di banche dati studiano i processi metabolici tramite network di dati biologici nei quali sono annotati i processi di interazione fra le molecole, per favorire la comprensione dei processi di regolazione dell'espressione genica e i processi post-traduzione relativi al trasporto e al metabolismo delle proteine.

ENZYME

ENZYME riporta in una struttura gerarchica la classificazione internazionale degli enzimi. Ogni entry riporta un "id" corrispondente all'*EC number*, il nome dell'enzima e i suoi sinonimi, l'attività catalitica, gli eventuali cofattori, il cross-referencing alla banca dati delle proteine e alla banca dati OMIM. La fonte primaria dei dati è la classificazione internazionale IUMB degli enzimi.

BRENDA

Come ENZYME annota le informazioni funzionali e molecolari degli enzimi. Associa ad ogni entry anche le informazioni dell'organismo da cui sono stati derivati i dati molecolari dell'enzima.

EcoCyc

EcoCyc è un database di un organismo modello, l'*Escherichia coli*, che annota dati non solo genomici e proteomici, ma anche quelli relativi ai processi metabolici, al trasporto e alla regolazione dell'espressione dei geni di *Escherichia coli*. In questa banca dati vengono annotati una grande quantità di geni la cui funzione è stata determinata sperimentalmente, quindi è un'ottima risorsa per predire nuovi geni in genomi di altri organismi microbici.

KEGG

KEGG è l'enciclopedia di Kyoto di geni e genomi ed è una risorsa integrata di banche dati correlate ai genomi completamente sequenziati o in fase di completamento. Lo scopo di tale banca dati è creare una rete tra le varie classi di

dati per la comprensione dei meccanismi preposti alla funzionalità delle cellule e degli organismi a partire dai dati genomici.

SSDB

SSDB (*Sequence Similarity Database*) è una banca dati che raccoglie i dati relativi all'applicazione del programma *SSEARCH* su tutte le sequenze amminoacidiche dei proteomi disponibili. Per ogni gene annotato in SSDB sono anche riportati i risultati della ricerca di patterns e domini presenti nelle sequenze in oggetto e dedotti dalle annotazioni in PROSITE e Pfam.

PATHWAYS

E' il database che annota i dati relativi alle interazioni fra le proteine. Tali interazioni sono: relazioni fra due enzimi in un processo metabolico, interazioni fra due proteine e interazioni a livello di espressioni dei geni.

MetaCyc

E' una banca dati che descrive per 158 organismi, 445 processi metabolici in cui sono coinvolti 1115 enzimi (dati luglio 2002). Tutti i processi sono determinati sperimentalmente.

4.9.4 Banche dati mitocondriali

Gli organismi eucariotici contengono nel citoplasma delle loro cellule, organuli di vario tipo fra i quali i mitocondri, che giocano un ruolo di assoluta importanza in moltissimi processi metabolici e di funzionalità della cellula.

La caratteristica principale di questo organulo è la presenza al suo interno di un proprio genoma che presenta dimensioni ridotte, forma circolare (nella maggior parte dei casi) mappa genomica conservata all'interno dei vertebrati e variabile negli altri animali. Numerose sono le informazioni disponibili tramite le banche dati che seguitiamo ad analizzare rapidamente.

GOBASE

GOBASE (*Organelle Genome Database*) è una risorsa genomica che raccoglie dati sui genomi di cloroplasti e mitocondri. I nomi dei geni sono annotati secondo un vocabolario controllato definito da esperti. I dati di sequenze nucleotidiche e proteiche sono estratti attraverso ENTREZ.

MITOMAP

MITOMAP (*Human Mitochondrial Genome Database*) è un report aggiornato ai dati pubblicati di tutte le variazioni riscontrate sul DNA mitocondriale di soggetti affetti da patologie e su soggetti i cui campioni sono stati prelevati per studi di genetica di popolazione. I dati sono annotati in tabelle e possono essere estratti attraverso l'utilizzo di un sistema di interrogazione semplice.

Human MitBASE

Human MitBASE è una banca dati nata per raccogliere in un'unica risorsa integrata i dati sul mitocondrio di tutti gli organismi eucariotici. I dati sono organizzati in base ad ogni individuo, alla sua origine geografica e alla sua descrizione dei dati clinici associati. Ogni entry raccoglie moltissime informazioni associate all'individuo ciò implica un notevole dispendio di risorse umane e una difficoltà di mantenimento della banca dati stessa, che risulta meno aggiornata rispetto a MITOMAP.

HvrBase

HrvBase è una banca dati che raccoglie i multi-allineamenti delle sequenze relative alle regioni di controllo del genoma mitocondriale dei primati. Per ogni sequenza è riportata l'informazione sull'origine del campione da cui la sequenza è stata ottenuta.

Mitop

MITOP raccoglie informazioni su geni correlati alla funzionalità del mitocondrio di uomo, topo, lievito, *Caenorhabditis elegans* e *Neurospora crassa*. Ogni entry è associata ad una proteina della quale sono annotate la classe funzionale, il codice dell'enzima, il complesso proteico di appartenenza della proteina, il peso molecolare, il punto isolettrico, correlazioni con patologie, processi metabolici e informazioni su eventuali geni ortologhi.

MitoNuc

MitoNuc è una banca dati di geni nucleari di *metazoi* per il mitocondrio. I dati sono estratti da SWISSPROT come sequenze mitocondriali di metazoi e vengono quindi accuratamente controllati e annotati con informazioni specifiche. Per quanto riguarda le proteine umane, la localizzazione del gene sul genoma umano è ottenuta attraverso le analisi effettuate su Ensembl.

AmmtDB

AMmtDB è la banca dati dei multi-allineamenti di geni codificati da genomi mitocondriali di Metazoi. Ogni entry corrisponde ad un gene e ad una classe-tassonomica specifica.

MITOCHONDRIOME

MITOCHONDRIOME è un sito web che raccoglie banche dati mitocondriali e informazioni correlate. Attraverso tale sito si accede alle banche dati Human_MitBase, MITONUC e AMmtDB oltre a dati ottenuti dall'analisi di variabilità e complessità di geni e genomi mitocondriali di metazoi.

PLMitRA

PLMitRNA è una banca dati di molecole e geni di tRNA identificati nei mitocondri di tutte le piante verdi. Informazioni caratterizzanti il gene o la molecola sono annotate e possono essere utilizzate per la ricerca dei dati. I tRNA possono essere selezionati per nome della specie o per raggruppamento tassonomico ed inoltre è anche disponibile il multi-allineamento di ciascun cluster di tRNA omologhi.

4.9.5 Risorse genomiche

L'avanzamento dei progetti genomici ha dato grande impulso allo sviluppo di risorse genomiche accessibili in modo più o meno libero sulla rete. Le risorse genomiche sono siti contenenti dati relativi al mappaggio e al sequenziamento genomico.

Le risorse sono:

- *Risorse integrate*, dove sono disponibili dati relativi a tutti i genomi attualmente in fase di studio (***Entrez_Genomes*** o ***EBI_Genome***);
- *Risorse relative ai genomi* di determinate categorie di organismi;
- *Risorse organismo specifiche* (***GadFly*** e ***FlyBASE***): permettono di scaricare sul proprio PC la sequenza dell'intero genoma (o parti di esse) individuate dalla localizzazione cromosomiale o da un marker;

5 PROGETTAZIONE E REALIZZAZIONE DEL SISTEMA MIR-ONTOLOGY

L'idea di sviluppare un sistema così articolato e completo come miRò, nasce dall'esigenza di avere uno strumento in grado di evidenziare le relazioni esistenti tra i microRNA, le caratteristiche ontologiche e le patologie ad essi appositamente correlate. Al momento esistono siti web di bioinformatica che forniscono agli utenti (medici, biologi, ricercatori, ...) sia informazioni riguardanti l'interazione tra miRNA e geni bersaglio sia quelle relative alle relazioni tra geni, malattie e ontologie. L'obiettivo di miRò è la progettazione, la realizzazione e l'interrogazione via web di un Database MySQL capace di riorganizzare e contenere un "set" di informazioni riguardanti i microRNA e i relativi geni bersaglio, in modo da focalizzare la connessione esistente tra un microRNA, le funzioni, i processi biologici da esso regolati e le patologie. Questi dati, raccolti opportunamente in apposite tabelle di MySQL attraverso un sistema automatico denominato BioXmlBuilder che verrà illustrato in seguito, vengono poi analizzati e processati secondo le tipologie di query più importanti che il sistema mette a disposizione.

5.1 Il sistema

Il sito web di miRò integra dati da diverse fonti, come illustrato dalla figura 5.1. Nelle sotto-sezioni successive saranno discussi i dettagli relativi all'integrazione dei dati e all'interfaccia web.

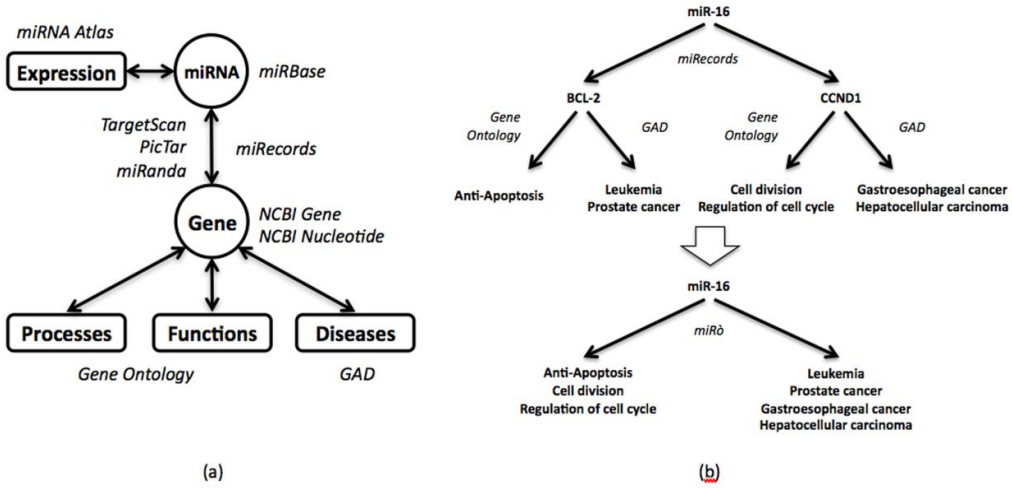


Fig. 5.1 - Lo schema di miRò. (a) I miRNA sono annotati con le loro informazioni provenienti da miRBase ed i loro profili di espressione ottenuti dal miRNA Atlas. Essi sono collegati a processi, funzioni e malattie attraverso i loro target predetti (da TargetScan, PicTar e miRanda) o validati (miRecords). (b) In questo caso, miR-16 ha due target validati, BCL2 e CCND1, tra gli altri. Questi geni sono annotati con dei termini GO (anti-apoptosis, cell division, regulation of cell cycle) e delle malattie (leukemia, prostate cancer, gastroesophageal cancer, hepatocellular carcinoma), di conseguenza miR-16, in miRò, eredita tali annotazioni.

5.2.1 Integrazione dei dati

I miRNA sono annotati con le informazioni sulle sequenze dei loro precursori e dei trascritti maturi, provenienti da miRBase, e con i profili di espressione ottenuti dal Mammalian microRNA Atlas [67, 133]. Questo atlante contiene profili di espressione di sequenze di pre-miRNA e miRNA maturi in diversi tipi di tessuti, sia normali che patologici. I miRNA sono inoltre associati ai termini di Gene Ontology (GO) e alle malattie attraverso i loro target: ogni miRNA eredita infatti tutte le annotazioni dei suoi geni target.

Le coppie miRNA/target supportate sperimentalmente provengono da miRecords, mentre i target predetti sono ottenuti da TargetScan, PicTar e miRanda [70, 73, 74]. I record relativi ai geni target sono arricchiti con informazioni generali quali il contesto genomico e le sequenze dei trascritti, provenienti dai database Gene e Nucleotide di NCBI. I termini ontologici con i quali i geni target sono annotati (processi e funzioni), sono ottenuti dal database Gene Ontology [134]. Infine, le relazioni gene-malattia provengono dal Genetic Association Database (GAD), una banca dati che contiene associazioni genetiche a disturbi e malattie complesse [135].

Tutti i dati sono raccolti e mantenuti aggiornati in un database mySQL. In particolare, i dati più rilevanti inerenti i miRNA ed i loro target sono memorizzati nel database per una disponibilità immediata, mentre gli altri dettagli sono raggiungibili attraverso i collegamenti ai siti originali. I dati sono generalmente ottenuti dalle fonti web originali sotto forma di *flat file*, ad eccezione di GO, che è scaricabile come database mySQL, e dell'atlante dei miRNA, che è distribuito come raccolta di file Excel.

Inizialmente, vengono memorizzate nel database le informazioni sui miRNA ricavate dai file di miRBase. Quindi, tutti i dati relativi alle predizioni sono esaminati e memorizzati, insieme alle informazioni principali inerenti i geni target, ricavate dai file dei database Gene e Nucleotide di NCBI. In questa fase vengono considerati anche gli alias dei nomi dei geni, in modo da facilitare le integrazioni successive. Nei file delle predizioni, i geni sono identificati attraverso i loro ID di NCBI, mentre i miRNA sono identificati o attraverso i loro accession number (miRanda) o i loro ID (TargetScan e PicTar). Quindi, i geni vengono annotati con i loro termini ontologici, provenienti dal database GO, e con le malattie ad essi associate, dedotte da GAD. In entrambi i casi, i geni sono identificati attraverso i loro nomi per cui, in questa fase, gli alias sono spesso determinanti per la loro corretta identificazione. Infine, vengono integrati i profili di espressione dei miRNA. Per ogni miRNA, individuato attraverso il suo ID, vengono memorizzati tutti i tessuti nei quali è espresso, insieme ai livelli di espressione. In questa fase, viene anche calcolata e memorizzata la distribuzione percentuale dei cloni dei miRNA nei vari tessuti, per una maggiore efficienza in fase di interrogazione.

Eventuali inconsistenze nell'integrazione dei dati sono prevenute attraverso un processo semi-automatico: il sistema rileva automaticamente gli errori e li salva su un file di *log*, per una successiva analisi da parte dell'operatore. Ad esempio, tutti i nomi dei miRNA contenuti nei file delle predizioni (PicTar e TargetScan), che non trovano corrispondenze in miRBase (per es. nel caso in cui i loro ID sono stati modificati rispetto alla versione precedente), vengono riportati nel file di log, assieme a tutti i nomi simili trovati all'interno del database. Questi vengono quindi controllati dall'operatore che procede, talvolta con l'ausilio delle sequenze, all'identificazione dei nomi corretti.

5.3 MySQL

In una prima fase di progettazione del sistema miRò, è stata eseguita un'analisi per la scelta oculata del Database utile per ospite e organizzare opportunamente tutti i dati raccolti dalle varie fonti biologiche. Tra le varie possibilità, è stato individuato il DMBS MySQL come database a supporto del progetto per le sue caratteristiche che di seguito verranno descritte. Una delle peculiarità che hanno orientato la scelta su questo DMBS è la sua disponibilità per la maggior parte dei sistemi operativi, oltre il fatto che è Open Source, è abbastanza veloce, scalabile ed esistono numerose librerie che permettono un suo utilizzo ad alto livello attraverso i vari linguaggi di scripting più usati nello sviluppo di applicazioni web. A tal proposito, esistono dei client web Open Source come ad esempio phpmyadmin, scritto interamente in PHP, che ne permettono una gestione completa del DB ad alto livello. MySQL supporta il modello relazionale dei dati che si basa sul concetto algebrico di relazione. Una caratteristica fondamentale del modello relazionale è la possibilità di definire "indici" sui singoli attributi o combinazioni di attributi. Gli indici sono strutture per l'accesso veloce al contenuto degli attributi e possono migliorare notevolmente le prestazioni del DBMS nel caso di interrogazioni al database. In MySQL una tabella può essere di diversi tipi (o *storage engine*). Ogni tipo di tabella presenta proprietà e caratteristiche differenti (transazionale o meno, migliori prestazioni, diverse strategie di locking, funzioni particolari, ecc). Esiste poi un'API che si può utilizzare per creare in modo relativamente facile un nuovo tipo di tabella, che poi si può installare senza dover ricompilare o riavviare il server. Tra i più importanti storage engine si evidenziano:

- *MyISAM*: lo storage engine di default maggiormente adoperato nel Web, all'interno di data warehousing, e altri ambient di sviluppo.
- *InnoDB*: questo storage rappresenta il modello transazionale (ACID compliant) con le relative istruzioni di commit, rollback, e crash-recovery per garantire l'integrità dei dati.
- *Memory* (conosciuto come Heap): immagazzina tutti i dati nella RAM per velocizzare l'accesso ai dati, utile in ambienti che richiedono un'estrema velocità di recupero delle informazioni a discapito di avere meno funzioni di ottimizzazione che invece si ritrovano in altri engine.
- *Merge*: questo engine tipicamente viene usato nei data warehousing per la sua caratteristica di consentire il raggruppamento di tabelle MyISAM viste come un unico oggetto.

- NDB o ClusterDB (introdotta nella versione 5.0)

Questi storage engine ottimizzano differentemente la memorizzazione dei dati, secondo le tipologie di query più frequenti (SELECT rispetto a UID); la categoria scelta per rappresentare nel modo migliore le tabelle del database di miRò è quella transazionale, utilizzando l'engine InnoDB. Questo engine immagazina i dati attraverso indici clusterizzati in modo da ridurre le operazioni di Input/Output per le query più frequenti che basano la ricerca sulla chiave primaria. Per garantire i vincoli di integrità referenziale, InnoDB supporta anche le Foreign Key. Attraverso i vincoli di integrità referenziale è possibile verificare automaticamente quando i valori della tabella madre vengono modificati o eliminati in modo da impedire queste modifiche o, viceversa, modificare di conseguenza anche i valori sulla tabella dipendente (tabella figlia). Inoltre non è possibile inserire nella tabella figlia valori che non hanno un corrispondente nella tabella madre. Questa opzione è di notevole importanza in quanto consente di risolvere quei problemi riguardanti l'inconsistenza dei dati. La struttura del database è costituita da 22 tabelle, molte delle quali rappresentanti le entità del modello dei dati (mature_mirna, gene, disease, ontology, …) e rappresentanti le relazioni che intercorrono tra esse (gene_disease, gene_ontology, gene_mirna,_anticorr, …). In seguito viene fornita una descrizione più dettagliata di ogni tabella.

5.4 Struttura del database di miRò

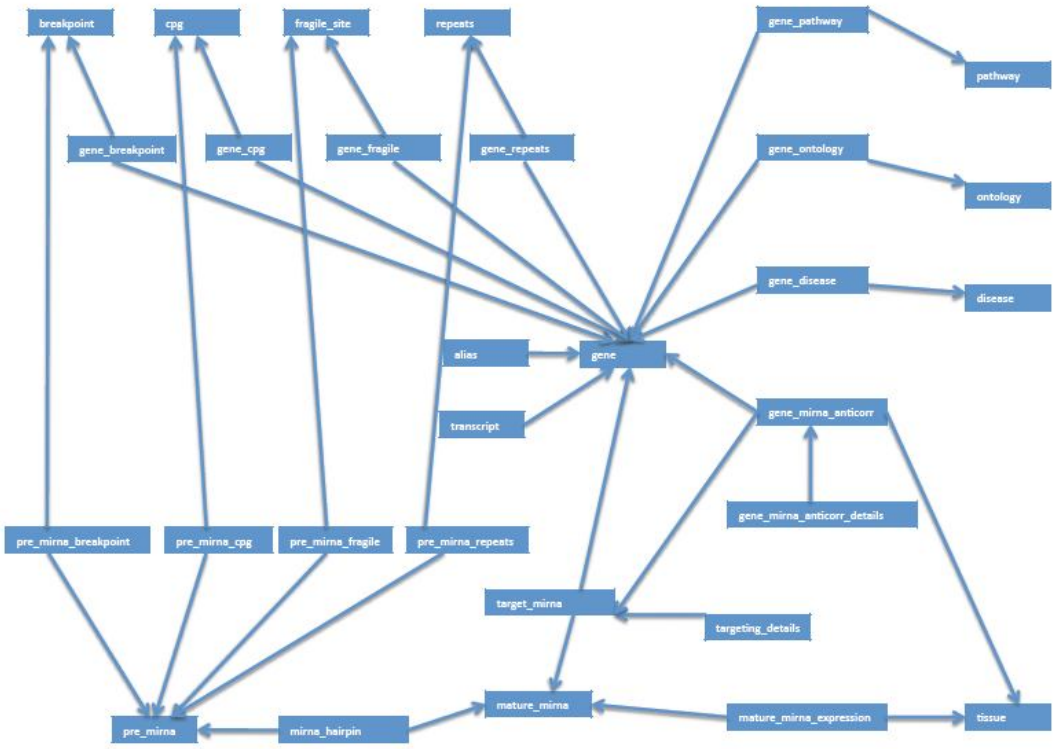


Diagramma E-R

Di seguito è riportato uno screenshot di *phpmyadmin* relativo alle tabelle di miRò:

	alias						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	breakpoint						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	cpg						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	disease						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	fragile_site						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	gene						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	gene_disease						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	gene_mirna_anticorr						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	gene_ontology						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	gene_pathway						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	mature_mirna						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	mature_mirna_expression						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	mirna_hairpin						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	ontology						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	pathway						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	pre_mirna						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	pre_mirna_expression						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	repeats						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	targeting_details						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	target_mirna						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	tissue						0	MyISAM	latin1_swedish_ci	1,0 KiB	-
	transcript						0	MyISAM	latin1_swedish_ci	1,0 KiB	-

Tabelle del database di miRò

Alias: questa tabella contiene tutti gli alias di ogni gene, ovvero, i nomi attraverso i quali uno stesso gene può essere identificato.

Breakpoint: questa tabella contiene la locazione genomica dei breakpoint noti in letteratura, ovvero, quelle zone del DNA maggiormente predisposte ad andare incontro a rotture e traslocazioni. Tali informazioni vengono recuperate dal database di NCBI.

CpG: questa tabella contiene le isole di CpG, ovvero, delle regioni genomiche che contengono un'alta concentrazione di citosina e guanina separate da un fosfato che lega i due nucleotidi nella sequenza di DNA. In molti mammiferi è stata riscontrata una forte correlazione tra le zone di CpG e l'inizio di un gene. Come conseguenza, la presenza di una CpG Island aiuta a capire i sistemi di predizione e annotazione dei geni. Un'alta concentrazione di CpG potrebbe essere associata con l'abbassamento della metilazione di citosina che spesso si è osservata in corrispondenza di queste regioni. Tale abbassamento favorisce una maggiore sopravvivenza delle CpG contro la vulnerabilità alle continue mutazioni che esse tendono a subire.

Disease: in questa tabella sono contenute tutte le malattie con componente genetica, recuperate dai database biologici KEGG, GAD e OMIM. In questa tabella sono presenti sia la classe di appartenenza che il nome della malattia.

Fragile_site: questa contiene il nome e la locazione genomica delle zone del DNA ritenute "fragili", ovvero, quelle zone che tendono a rompersi facilmente quando le cellule sono sottoposte a condizioni di stress a causa di continue replicazioni. Si tratta di zone che vengono ereditate nella duplicazione cellulare e, a seconda della frequenza con la quale queste rotture si verificano, i siti fragili sono stati classificati in comuni o rari. Al momento sono stati identificati più di 120 siti fragili nel genoma umano.

Gene: questa è una fra le più importanti tabelle del database e contiene le principali caratteristiche dei geni come l'identificatore geneid, il simbolo ufficiale del gene, una descrizione, il nome completo e il contesto genomico (cromosoma e coordinata di inizio e di fine);

Gene_Disease: questa è una tabella di correlazione molti a molti tra la tabella dei geni e quella delle malattie in modo da poter legare più geni a più malattie.

Gene_Mirna_Anticorr: anche questa tabella è di correlazione nella quale sono presenti i campi che legano la tabella dei geni con quella dei mature mirna, riportando anche il tessuto in cui è stata riscontrata tale correlazione e il p-value che verrà descritto più avanti. In particolare, attraverso questa tabella è possibile correlare un gene con molti mature_mirna (cardinalità uno a molti).

Gene_Ontology: questa tabella correla più geni con più ontologie (cardinalità molti a molti). Poiché un gene può svolgere una precisa funzione o appartenere ad un processo o costituire una componente cellulare, in questa tabella viene riportata questa informazione sottoforma di un codice che ne identifica uno dei suddetti tipi.

Gene_Pathway: in questa tabella vengono conservate le relazioni tra i geni e le pathway nelle quali essi sono contenuti. Questa tabella risolve una correlazione di tipo molti a molti.

Mature_Mirna: in questa tabella sono conservati tutti i mirna maturi recuperati dal database di miRBase, insieme all'accession number di miRBase, la sua sequenza e un campo contenente il numero di copie totali trovate.

Mature Mirna Expression: in questa tabella vengono correlati i mature mirna con i tessuti (relazione molti a molti) e in un apposito campo viene riportato il valore di espressione del mature mirna per ogni tessuto.

Mirna_Hairpin: in questa tabella è contenuta la relazione fra un pre-mirna e i suoi due possibili mature mirna, riportando anche le locazioni genomiche dei singoli mature mirna.

Ontology: questa tabella contiene tutte le informazioni sulle funzioni e sui processi biologici.

Pathway: in questa tabella sono riportati i nome delle pathway così come sono contenuti nel database biologico KEGG.

Pre_Mirna: contiene le caratteristiche dei precursori miRNA come ad esempio l'accession number di miRBase, il nome identificativo, la sequenza e i campi relativi alla posizione genomica (cromosoma, coordinata di inizio e di fine e il filamento di appartenenza) e alla famiglia di appartenenza.

Pre_Mirna_Expression: questa è una tabella di correlazione tra i pre-mirna e i tessuti. In particolare, in questa tabella vengono riportati i valori di espressione di ciascun pre-mirna per ogni tessuto.

Repeats: in questa tabella sono riportate le aree cromosomiche di repeats, ovvero, aree genomiche contenenti una sequenza di nucleotidi che si ripetono frequentemente.

Targeting_Details: in questa tabella vengono riportati i dettagli relativi alla relazione tra gene e miRNA. In particolare, viene specificato la sorgente dalla quale è stata recuperata la predizione di interazione (TargetScan, PicTar, miRanda, miRTarget2, PITA, RNA22, miRecords), l'id del gene, del trascritto, la sua posizione genomica ed uno score associato a ciascuna predizione che indicativamente ne permette di valutare la veridicità.

Target_Mirna: in questa tabella sono riportate le interazioni tra i miRNA e i geni bersaglio. I dettagli relativi alle sorgenti dalle quali sono state lette tali interazioni sono state specificate nella tabella precedente targeting_details.

Tissues: questa tabella contiene l'elenco di tutti i tessuti definiti nel miRNA Atlas, con una breve descrizione per ciascuno di essi, l'apparato in cui è presente, la libreria di appartenenza, il typecode e il tipo di tessuto (sistema nervoso, vasi sanguigni, …) .

Transcript: questa tabella contiene le sequenze e i dettagli dei trascritti associati ai geni, ricavati da NCBI.

La compilazione delle tabelle appena descritte, avviene per mezzo di apposite stored procedure scritte utilizzando il linguaggio di PL/SQL di MySQL. Queste procedure vengono invocate in automatico dalla web application denominata BioXMLBuilder che verrà descritta nel paragrafo seguente. Questa applicazione web, opportunamente configurata, si preoccupa di recuperare i dati dalle varie fonti biologiche, convertirli in un formato standard XML e di inserirli in apposite tabelle temporanee. Successivamente, ciascuna stored procedure, provvede a leggere i dati dalle tabelle temporanee per poi trasferirli opportunamente nelle tabelle di miRò, rispettando i vincoli di integrità referenziale dei dati.

Di seguito vengono riportate una serie di tabelle riportanti il mapping tra i dati contenuti nei file delle sorgenti biologiche e i campi dele relative tabelle del database di miRò.

5.4.1 Fonti biologiche usate per miRò

Source: miRBase

URL: www.mirbase.org

File: hsa.gff

Type: flat file

DB Tables: pre_mirna

FILE	DB	EXAMPLE
Col1: Chromosome	pre_mirna.chromosome	1
Col3: Start	pre_mirna.chr_start	30366
Col4: Stop	pre_mirna.chr_stop	30503
Col5: Strand	pre_mirna.strand	+
Col6: Accession	pre_mirna.mirbase_acc	MI0006363
Col7: miRNA ID	pre_mirna.mirbase_id	hsa-mir-1302-2

File: miRNA.xls

Type: Excel file

DB Tables: pre_mirna, mature_mirna

FILE	DB	EXAMPLE
ColA: Accession	pre_mirna.mirbase_acc	MI0000060
ColB: miRNA ID	pre_mirna.mirbase_id	hsa-let-7a-1

FILE	DB	EXAMPLE
ColD: sequence	pre_mirna.sequence	UGGGAUGAGGUAGUAGGU...
ColE: mature1_acc	mature_mirna.mirbase_acc	MIMAT0000062
ColF: mature1_ID	mature_mirna.mirbase_id	hsa-let-7a
ColG: mature1_seq	mature_mirna.sequence	UGAGGUAGUAGGUUGUAUAGUU
ColH: mature2_acc	mature_mirna.mirbase_acc	MIMAT0004481
ColI: mature2_ID	mature_mirna.mirbase_id	hsa-let-7a*
ColJ: mature2_seq	mature_mirna.sequence	CUAUACAAUCUACUGUCUUUC

File: miFam.dat

Type: flat file

DB Tables: pre_mirna

FILE	DB	EXAMPLE
Field: ID	pre_mirna.family	mir-17

File: miRNA.str

Type: text file

DB Tables: mirna_hairpin

FILE	DB	EXAMPLE
energy	mirna_hairpin.energy	-35,60
mature1_start	mirna_hairpin.mature1_start	6
mature1_stop	mirna_hairpin.mature1_stop	27
mature2_start	mirna_hairpin.mature2_start	57
mature2_stop	mirna_hairpin.mature2_stop	77
line1	mirna_hairpin.line1	U GU uuagggucacac
line2	mirna_hairpin.line2	uggga GAG AGUAGGUUGUAUAGUU c
line3	mirna_hairpin.line3	||||| ||| ||||||||||||| c
line4	mirna_hairpin.line4	aucCU UUC UCAUCUAACAUAUCaa a
line5	mirna_hairpin.line5	- UG uagagggucacc

Source: NCBI

URL: **ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/**

File: Homo_sapiens.gene_info.gz

Type: CSV

DB Tables: gene, alias

FILE	DB	EXAMPLE
Col2: GeneID	gene.gene_id	1
Col3: Symbol	gene.symbol	A1BG
Col5: Aliases *	alias.symbol	A1B|ABG|DKFZp686F0970|GAB|HYST2477
Col7: Chromosome	gene.chromosome	19
Col8: Map_location	gene.map_location	19q13.4
Col9: Description	gene.description	alpha-1-B glycoprotein

Source: NCBI

URL: ftp://ftp.ncbi.nih.gov/gene/DATA/

File: gene2refseq.gz

Type: CSV

DB Tables: gene, transcript

FILE	DB	EXAMPLE
Col4: RNA_nucleotide_accession.version	(transcript.transcript_acc, transcript.version)	NM_130786.3 => (NM_130785, 3)
Col6: protein_accession.version	transcript.protein_acc	NP_570602.2
Col10: Start	gene.chr_start	58858171
Col11: Stop	gene.chr_stop	58864864
Col12: Strand	gene.strand	-

Source: NCBI

URL: ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/

File: human.rna.fna.gz

Filetype: fasta

DB tables: transcript

FILE	DB	EXAMPLE
Sequence	transcript.sequence	...

Source: TargetScan

URL:http://www.targetscan.org/cgi-in/targetscan/data_download.cgi?db=vert_50

File: Conserved_Site_Context_Scores.txt, Nonconserved_Site_Context_Scores.txt

Type: flat file

DB Tables: target_mirna, targeting_details

FILE	DB	EXAMPLE
Col6: UTR_start	targeting_details.start_pos	226
Col7: UTR_end	targeting_details.end_pos	232
Col11: context_score	targeting_details.score	-0,4562

Source: miRanda

URL: http://www.microrna.org/microrna/getDownloads.do

File: hg19_predictions_S_C_aug2010.txt, hg19_predictions_S_0_aug2010.txt, hg19_predictions_0_C_aug2010.txt, hg19_predictions_0_0_aug2010.txt

Type: CSV

DB Tables: target_mirna, targeting_details

FILE	DB	EXAMPLE
Col12: Gene_start	targeting_details.start_pos	495
Col13: Gene_end	targeting_details.end_pos	516
Col19: svr_score	targeting_details.score	-0,1156

Source: PicTar

URL: http://pictar.mdc-berlin.de/

File: pictar4way.txt

Type: CSV

Da scaricare MANUALMENTE da UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables?command=start)

May 2004 Genome Assembly (hg17)

Group: Regulation - Track: PicTar miRNA - Pictar 4way

DB Tables: target_mirna, targeting_details

FILE	DB	EXAMPLE
Col3: chromStart	targeting_details.start_pos	65404922
Col4: chromEnd	targeting_details.end_pos	65404929
Col6: Score	targeting_details.score	63

Source: miRTarget2

URL: http://mirdb.org/miRDB/download.html

File: human_predictions_sept2008.txt

Type: CSV

DB Tables: target_mirna, targeting_details

FILE	DB	EXAMPLE
Col3: score	targeting_details.score	506.672

Source: PITA

URL: http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html

File: PITA_sites_hg18_0_0_ALL.tab

Type: CSV

DB Tables: mirna_target, targeting_details

FILE	DB	EXAMPLE
Col8: score	targeting_details.score	0,42
Col11: start	targeting_details.start_pos	9111585
Col12: end	targeting_details.end_pos	9111590

Source: miRecords

URL: http://mirecords.biolead.org/download.php

File: miRecords_version2.xls

Type: Excel file

DB Tables: target_mirna, targeting_details

FILE	DB	EXAMPLE
Col R: target site position	targeting_details.start_pos	801

Source: NCBI

URL: ftp://ftp.ncbi.nih.gov/genomes/MapView/Homo_sapiens/non_sequence/

File: mitelman.md.gz

Type: CSV

DB Tables: breakpoint, gene_breakpoint, mirna_breakpoint (? - NO map_location per miRNA)

FILE	DB	EXAMPLE
Col2: Chromosome	breakpoint.chromosome	1
Col3: Start	breakpoint.map_start	1p10
Col4: Stop	breakpoint.map_stop	1p10
Col5: featureName	breakpoint.feature_name	add(1)(p10)
Col6: featureID	breakpoint.feature_id	na:1

Source: NCBI

URL:

ftp://ftp.ncbi.nih.gov/genomes/MapView/Homo_sapiens/sequence/BUILD.37.1/initial_release/

File: seq_repeat.md.gz

Type: CSV

DB Tables: repeats, gene_repeats, premirna_repeats

FILE	DB	EXAMPLE
Col2: Chromosome	repeats.chromosome	1
Col3: chr_start	repeats.chr_start	1134

FILE	DB	EXAMPLE
Col4: chr_stop	repeats.chr_stop	1205
Col5: strand	repeats.strand	-
Col10: featureName	repeats.feature_name	tRNA-Gln-CAA_
Col11: featureID	repeats.feature_id	na:13895188

Source:NCBI

URL:

ftp://ftp.ncbi.nih.gov/genomes/MapView/Homo_sapiens/sequence/BUILD.37.1/initial_release/

File: seq_cpg_islands.md.gz

Type: CSV

DB Tables: cpg, gene_cpg, premirna_cpg

FILE	DB	EXAMPLE
Col2: Chromosome	cpg.chromosome	1
Col3: chr_start	cpg.chr_start	2895
Col4: chr_stop	cpg.chr_stop	2669

Source: NCBI

File: fragile_site.xls

Type: Excel

DB Tables: fragile_site, gene_fragile, premirna_fragile

FILE	DB	EXAMPLE
ColA: HGNC ID	fragile_site.hgnc_id	HGNC:3868
ColB: Approved Symbol	fragile_site.symbol	FRA1A
ColC: Approved Name	fragile_site.name	fragile site, aphidicolin type, ...
ColD: Map Location	fragile_site.map_location	1p36
ColE: Chr Start	fragile_site.chr_start	1
ColF: Chr End	fragile_site.chr_stop	28000000

Source: Gene Ontology

URL: http://www.geneontology.org/GO.downloads.ontology.shtml

File: OBO v1.2 (gene_ontology_ext.obo.txt)

Type: text

DB Tables: ontology

FILE	DB	EXAMPLE
id	ontology.go_id	GO:0000001
name	ontology.name	mitochondrion inheritance
namespace	ontology.type	biological_process

Source: Gene Ontology

URL: http://www.geneontology.org/GO.downloads.annotations.shtml

File: OBO v1.2 (gene_association.goa_human)

Type: text

DB Tables: ontology, gene_ontology

Source: KEGG pathway

URL: ftp://ftp.genome.jp/pub/kegg/xml/kgml/non-metabolic/organisms/hsa/
ftp://ftp.genome.jp/pub/kegg/xml/kgml/metabolic/organisms/hsa/

File: the entire directories.

Type: xml

DB Tables: pathway

FILE	DB	EXAMPLE
pathway name	pathway.kegg_id	path:hsa02010
title	pathway.name	ABC transporters

Source: KEGG pathway

URL ftp://ftp.genome.jp//pub/kegg//pathway/organisms/hsa/

File: hsa.list

Type: CSV

DB Tables: pathway, gene_pathway

Source: Pathway Commons

URL: http://www.pathwaycommons.org/pc-snapshot/gsea/by_species/

File: homo-sapiens-entrez-gene-id.gmt.zip

Type: flat file

DB Tables: pathway, gene_pathway

FILE	DB	EXAMPLE
Col1	pathway.name	Glucocorticoid biosynthesis

Source: KEGG disease

URL: ftp://ftp.genome.jp//pub/kegg//medicus/

File: disease

Type: flat file

DB Tables: disease, gene_disease

FILE	DB	EXAMPLE
Entry	disease.source_id	H00003
Name	disease.name	Acute myeloid leukemia (AML)
Category	disease.class	Cancer

Source: NCBI MIM (phenotypes)

URL: ftp://ftp.ncbi.nih.gov/repository/OMIM/

File: omim.txt.Z

Type: flat file

DB Tables: disease, disease_alias

FILE	DB	EXAMPLE
FIELD NO	disease.source_id	100070
FIELD TI	disease.name	AORTIC ANEURYSM, FAM...

Source: NCBI MIM (phenotypes)

URL ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/

File: mim2gene

Type: CSV

DB Tables: disease, gene_disease

Source: miRNA Atlas - Tissues

File: mmc2.xls

Type: Excel file

DB Tables: tissue

FILE	DB	EXAMPLE
Col A (Solo se grassetto)	tissue.apparatus	Neuronal tissues/cell lines
Col A (No grassetto)	tissue.tissue	Brain_normal adult
Col B	tissue.library	hsa_Cerebellum-adult
Col C	tissue.description	cerebellar cortex from...
Col D	tissue.typecode	9.110
Col E	tissue.tissuetype	nervous system, ...
Col F	tissue.malignancy	1

La codifica di tissue.malignancy è la seguente:

1	normal tissue
2	cancer tissue
3	non-cancer-derived cells/cell line
4	cancer-derived cells/cell line

Source: miRNA Atlas

File: mmc13.xls

Type: Excel file

DB Tables: tissue, mature_mirna, mature_mirna_expression

Source: GAD (Genetic Association Database)

URL: http://geneticassociationdb.nih.gov/cgi-bin/download.cgi (not automatic)

File: diseaselist.txt

Type: text

DB Tables: disease

Source: GAD (Genetic Association Database)

URL: http://geneticassociationdb.nih.gov/cgi-bin/download.cgi (not automatic)

File: all.txt

Type: CSV

DB Tables: disease, gene_disease

5.5 BioXML-Builder

BioXML-Builder è costituito da un front-end, sviluppato in Ruby on Rails, che gestisce l'interfaccia web e il servizio di configurazione per la conversione in XML dei file appartenenti alle fonti biologiche scelte dall'utente, e da un back-end, costituito da metodi scritti con il linguaggio Ruby, che utilizza il database MySQL per la gestione del meccanismo di conversione in XML dei dati ivi contenuti a basso livello. Il front-end si presenta con la struttura software classica di tutte le applicazioni web in Ruby on Rails: un interfaccia per l'elenco dei dati contenuti in una tabella e un'interfaccia per la loro modifica (modifica, inserimento e cancellazione). Per quanto concerne lo sviluppo, tutte le applicazioni sviluppate con Rails hanno una peculiarità, ovvero, sono tutte organizzate secondo una struttura gerarchica comune. Attraverso il comando *rails*, con il quale si creano nuove applicazioni all'interno del frame work, vengono generate automaticamente una serie di directory e file che forniscono una certa linea guida nello sviluppo, una linea che se rispettata permette a Rails di effettuare molte cose automaticamente (ad esempio caricare i file dell'applicazione, generarli ed individuarli a runtime e molto altro ancora). Questa struttura comune permette anche di comprendere con semplicità il codice dei progetti realizzati da altri, poiché risulteranno organizzati tutti nella stessa maniera. Vediamo di seguito la logica delle varie directory e dei file che esse contengono, prendendone in esame alcune nello specifico in relazione allo sviluppo di BioXML-Builder.

APP: questa directory costituisce il cuore dell'applicazione in quanto contiene il codice specializzato per l'applicazione web. All'interno di essa esistono quattro sottodirectory: *controllers*, *models*, *views* e *helpers*. Ciascuna di queste cartelle contiene un file per ogni elemento concettuale dell'applicazione, per cui vi saranno tanti file quanti sono i modelli (gli oggetti) facente parte dell'applicazione (Persone, Malattie, Geni, mRNA, …), tanti file quante sono le viste, rappresentate dalle singole pagine HTML, tanti quanti sono i controller, costituiti dalle classi in Ruby che interagiscono con le viste e infine tanti file quanti sono gli helper costituiti da metodi che aiutano a creare le viste.
Analizziamo di seguito le singole sottodirectory di *app* in BioXML-Builder:

- i *controller* presenti contengono la cosiddetta *business logic* dell'applicazione: come in tutte le applicazioni in Rails, vi è una classe controller principale denominata *application_controller*, che viene ereditata da tutti gli altri controller e di conseguenza, tutti i metodi definiti in questa classe sono disponibili a tutti gli altri controller, come ad esempio il modulo di libreria che gestisce il meccanismo di autenticazione fornito dal plugin corrispondente e che viene incluso in questo controller (questo è approfondito nella sezione riguardante la directory *vendor*); seguono i controller relativi ad ognuno dei modelli presenti nell'applicazione e generati attraverso il comando di script *scaffold* (impalcatura) che, come suggerisce la parola stessa, fornisce il minimo codice necessario per permettere all'applicazione di interagire e integrare in se il modello in questione. Ogni modello infatti è rappresentato da una classe Ruby contenente dei metodi che riflettono le caratteristiche del modello stesso; a tale classe viene associata una tabella sul database i cui campi corrispondono agli attributi dell'oggetto e i record rappresentano tutte le istanze della classe modello (si approfondirà questo concetto nella paragrafo dedicato al back-end); a questa classe modello si aggiunge una relativa classe controller che ne include i metodi "amministrativi", ossia quei metodi che permettono l'interazione con la relativa tabella nel database dell'applicazione; infine, vi sono i controller che gestiscono determinate funzionalità dell'applicazione, come ad esempio il *viewer_controller* che gestisce una visualizzazione delle pagine dell'applicazione distinguendole tra pubbliche e amministrative, vietando la visione di quest'ultime nel caso in cui l'utente non si sia qualificato come amministratore, e il *wizard_controller* che tiene conto dei file delle fonti biologiche selezionati dall'utente e dei relativi campi, gestendo attraverso l'azione *file_conversioni*, tutte le operazioni necessarie per la conversione dei dati selezionati in formato XML attraverso la libreria *Nokogiri* che possiede delle C-extension per lo sviluppo software con XML.
- i *modelli*, come già detto, hanno una corrispondenza con le tabelle del database e rivestono ruoli ben precisi nell'applicazione; in BioXML-Builder i modelli principali sono costituiti da:

- ***pagine pubbliche***, il cui contenuto html è conservato nei record della corrispondente tabella nel database;
- ***utenti amministratori***, con controlli di validità per i campi nome, login, password, conferma password ed email;
- ***file delle fonti dei dati biologici*** che contengono tutte le informazioni biologiche che poi vengono convertite nel formato standard XML;
- ***nomi delle fonti dei dati biologici*** che possiedono l'indirizzo web attraverso il quale l'applicativo BioXML-Builder è in grado di accedervi per il recupero dei file (questi siti vengono periodicamente visitati in automatico per valutare gli eventuali aggiornamenti dei file in essi contenuti; in questi casi il sistema provvederà nuovamente a scaricarli aggiornando le relative tabelle del database);

- le *viste* o *views*, costituite da file aventi estensione *.erb*, vengono generate in modo dinamico a seconda dell'azione definita in un determinato controller e si occupano di rendere visibile all'utente la componente relativa dell'applicazione. Non sempre ad ogni azione corrisponde necessariamente una view.

Components: questa directory contiene le componenti di alto livello riutilizzabili da più applicazioni come ad esempio quelle utili per gestire il login. In realtà i componenti vengono usati molto raramente perché Rails fornisce meccanismi migliori per rendere il codice riutilizzabile per cui risultano oramai deprecati.

Config: questa directory contiene le informazioni relative agli environment, ai dati di connessione al database e alle route. Gli ***environment*** sono una caratteristica di Rails particolarmente utile, in quanto permettono di utilizzare una stessa applicazione in tre modalità differenti: *development*, *production* e *test*. Questo perché un'applicazione Rails dovrebbe tipicamente essere sviluppata usando un database per lo sviluppo (development), uno per i test ed uno invece per quando l'applicazione è pronta ad essere utilizzata dagli utenti finali (production). Le opzioni di **accesso al database** vengono controllate da un singolo file, config/database.yml che è scritto in formato *YAML* (Yaml Ain't Markup Language −

Yaml Non è un Linguaggio di Markup), un linguaggio per la rappresentazione delle informazioni molto utilizzato nel mondo Ruby soprattutto per la costruzione di file di configurazione. Le informazioni relative alle route, infine, sono contenute nel file *routes.rb*; in esso ci sono le associazioni tra un URL ed una determinata azione, secondo il RESTful Routing System citato precedentemente.

Db: in questa directory si trovano quei file contenenti informazioni sul database, sui listati SQL e sul codice Ruby relativo alle *migration*, ovvero, quella funzionalità di Rails tramite la quale è possibile apportare modifiche incrementali al proprio database facendolo evolvere col tempo, con la possibilità aggiuntiva di poter tornare indietro se lo si desidera. Si approfondiranno tali aspetti nel paragrafo dedicato al back-end.

Doc: questa directory contiene tutta la documentazione relativa al progetto; inizialmente vi sarà solo il file di default *README*.

Lib: tutte le funzionalità non prettamente legate al lato web dell'applicazione saranno presenti in questa directory, come ad esempio i moduli per convertire dei dati o per effettuare calcoli o per interagire con sistemi esterni. Nel caso di BioXML-Builder, questa directory contiene gli script in Ruby per l'aggiornamento automatico del database contenente tutti i dati biologici reperiti dalla varie fonti online. In generale, in questa tabella sarà presente tutto il codice che non corrisponde direttamente ad un modello, ad un controller, ad un helper o ad una vista.

Log: in questa directory sono contenuti i log generati dal web server; nel nostro caso, il webserver utilizzato è Apache 2.2.

Public: questa è la web root, ovvero il posto in cui vanno messi i file HTML statici, le immagini, codici in JavaScript e fogli di stile CSS. Per questi ultimi tre esistono delle specifiche sottodirectory a dimostrazione del fatto che tutte le applicazioni Rails si prestano ad una manutenzione del codice semplice e chiara.

Script: in questa directory sono presenti alcuni applicativi che permettono di fare molte operazioni utili. Tra queste, vi è ad esempio l'applicativo ***WEBrick*** che è il Web server integrato in Rails, utile per iniziare a sviluppare immediatamente senza doversi preoccupare di particolari configurazioni. Tra gli altri script hanno particolare importanza ***generate*** e ***console***. L'applicativo *generate* consente di generare gli scheletri di alcuni file: se si decide di creare un modello per gli *User* (gli utenti dell'applicativo) si potrebbe usare il comando *generate model User* in modo da ottenere immediatamente un file per il modello *User*, una nuova *migration*, un file dove scrivere i test per la classe ed uno dove inserire dei dati predefiniti. Il fatto che questi file vengano generati in automatico, così come anche le directory che li contengono, rende lo sviluppo molto più veloce nelle fasi iniziali. L'applicativo ***console*** è invece uno strumento molto utile per interagire direttamente con l'applicazione senza l'uso di particolari costrutti, consentendo di far interpretare del codice scritte tramite console.

Test: questa directory contiene i test eseguiti con l'applicazione. In particolare essa conterrà sia i test relativi ad ogni parte dell'applicazione eseguita che i test che attraversano tutti gli strati (controller, modelli, viste).
Inoltre in questa directory si potranno mantenere dei dati di prova utili a far girare i test (le cosi dette *fixture*).

Tmp: in questa cartella ritroviamo i file temporanei, come quelli relativi alle sessioni o ai cookie, utilizzati in BioXML-Builder dal wizard_controller per gestire le scelte dell'utente; questi dati vengono comunque conservati in una tabella del database anziché in questa directory per motivi di maggiore usabilità dal lato applicativo. Queste informazioni infatti vengono utilizzate per tenere conto delle scelte medie fatte da ciascun utente e tornano utili ai fini statistici (ad esempio è possibile tener traccia della maggior parte delle fonti che vengono attinte più frequentemente e, tra queste, quelle in relazione a quali altre).

Vendor: in questa directory vengono inserite tutte le librerie di terze parti che possono tornare utili per la risoluzione di problematiche riguardanti il proprio progetto. Nello specifico, per creare HTML da semplice testo, è stato utile utilizzare

la libreria ***RedCloth***. Tra le librerie di terze parti ci sono anche i plugin, ovvero delle piccole librerie che contengono funzionalità che estendono Rails e che possono essere installate automaticamente tramite l'applicativo *plugin* contenuto nella directory Script descritta precedentemente. La lista dei plugin esistenti per Rails è particolarmente grande: contiene plugin per sistemi di autenticazione, per effetti grafici, per l'integrazione con sistemi esterni, etc. Ogni plugin può essere aggiornato indipendentemente dal resto dell'applicazione. Anche per lo sviluppo di BioXML-Builder sono stati utilizzati alcuni plugin e librerie:

- il plugin ***RESTful Authentication***, largamente utilizzato per fornire alla propria applicazione web una base utile per la gestione sicura dell'autenticazione degli utenti; nel caso di BioXMLBuilder, è stato utile l'utilizzo di questo plugin dovendo gestire l'autenticazione degli utenti come amministratore rispetto alle altre tipologie di autenticazione dell'applicativo
  [http://agilewebdevelopment.com/plugins/restful_authentication]
- il plugin ***acts_as_textiled***, che permette attraverso un semplice linguaggio di markup, ***textile***, di modificare il contenuto delle pagine di BioXML-Builder non amministrative senza dover scrivere codice HTML; questo plugin si appoggia proprio sulla libreria *RedCloth*;
- il plugin ***in_place_editor***, che permette di modificare le pagine web non amministrative di BioXML-Builder direttamente attraverso un'interfaccia grafica senza dover ricorrere alla modifica del codice e alla programmazione, garantendo una notevole comodità nella gestione e modifica di tali pagine; questa è una caratteristica comune ormai nelle moderne applicazioni web che si appoggia sulle librerie Javascript *Prototype* e *Scriptaculous* ;
- la libreria *gem33* di ***Nokogiri***, veloce e performante, scritta da Aaron Patterson e Mike Dalessio, è utile per parserizzare e generare HTML e XML a partire dai dati dei modelli: questa è la componente software che permette la rappresentazione in XML dei dati contenuti nel database di BioXML-Builder, poichè all'interno di ogni sua classe modello si è implementato un metodo di classe (e quindi valido per tutti gli oggetti di una particolare classe modello a prescindere dall'istanziazione degli oggetti

in se) denominato *printxml()* che accetta in input un sottoinsieme di oggetti di un modello (record della corrispondente tabella) e genera un file XML contenente tutti gli attributi degli oggetti in questione specificati in tale metodo; ogni modello ha una sua versione di printxml().

XML: questa è l'ultima directory utilizzata nell'applicazione di BioXML-Builder ed è quella che contiene l'XML generato dai file di fonti biologiche selezionati dall'utente, che funge da "contenitore" temporaneo di questi file che vengono poi raccolti in un archivio compresso e mandati all'utente.

5.5.1 Back-end

La parte del back-end di BioXML-Builder è costituita da una serie di script server-side che si occupano del reperimento dei dati biologici dalle varie banche dati online, del monitoraggio degli stessi (qualora questi vengano aggiornati) , di tutta la gestione degli account amministrativi e dell'interazione con il database. Attualmente, vengono riconosciuti tutti i file di dati biologici nel formato CSV, FLAT file e XLS. Di seguito si approfondiranno tutti i concetti legati alla struttura delle tabelle dell'applicativo, sia nel contesto del framework Rails, e di come questo gestisce il database, sia nel contesto della gestione effettiva dei dati biologici e delle fonti online.

5.5.2 ActiveRecord e i DBMS

Grazie allo sviluppo della tecnica di programmazione *ORM* (Object-Relational Model) è possibile avere l'integrazione di sistemi software aderenti al paradigma della programmazione orientata agli oggetti con i sistemi RDBMS, attraverso raffinati automatismi per la gestione della persistenza degli oggetti con un alto livello di astrazione nell'accesso ai database. Nel framework di Rails questo compito viene svolto da ***ActiveRecord***, una libreria scritta in Ruby dall'ideatore di Rails, David Heinemeier Hansson, per l'accesso alle basi di dati. Più correttamente si tratta di un design pattern originariamente pubblicato da Martin Fowler nel suo libro *Pattern of Enterprise Application Architecture*, edito da Addison-Wesley nel 2002, di cui il creatore di Rails ha fornito una implementazione in Ruby,

rendendola poi parte integrante del core di Rails. ActiveRecord stesso implementa al suo interno un altro pattern molto utile: il *Single Table Inheritance. Il pattern STI mappa tutti i campi di tutte le classi di una struttura di eredità in una singola tabella».* ActiveRecord è uno dei principali motivi per cui si è scelto Ruby on Rails come framework per lo sviluppo di BioXML-Builder e più in generale per la seconda edizione di miRò. Ciò non solo perché si presenta come un ORM, sgravando moltissimo dal lavoro di sviluppo l'onere di dover lavorare con codice SQL, ma soprattutto perché presenta dei vantaggi caratterizzati dal concetto di ***configurazione guidata da convenzioni*** (caratteristica principale del framework di Ruby on Rails) in modo da fare delle assunzioni di default senza la necessità di dover specificare le convenzioni da utilizzare, di scrivere file di configurazione o ancora dettagliare il mapping. Riportiamo qualche esempio in tal senso:

- tutte la tabelle presentano una chiave primaria chiamata ***id*** per garantire un unico nome alle chiavi primarie;
- il nome della tabella corrisponde alla forma plurale del nome della classe modello;
- gli attributi degli oggetti hanno una corrispondenza diretta con i nomi dei campi delle tabelle le cui righe ne rappresentano le istanze; il binding viene eseguito in automatico.

ActiveRecord è stato principalmente designato per eseguire le operazioni tipiche che vanno sotto l'acronimo di ***CRUD*** (Create-Read-Update-Delete), in modo molto semplice, veloce e senza richiedere necessariamente di scrivere codice SQL.
Riassumiamo alcuni dei vantaggi derivanti dall'uso di ActiveRecord:

- semplificazione della configurazione;
- binding automatico tra tabelle e classi e tra colonne ed attributi;
- validazione dei dati;
- logging;
- migrations;
- aggregazione.

Un altro vantaggio è che ActiveRecord è stato scritto in Ruby, quindi, tutto ciò che

è possibile fare con gli oggetti in Ruby, risulta fattibile anche con ActiveRecord. Tramite ActiveRecord è possibile connettersi in maniera semplice ed intuitiva al database dell'applicativo, mettendo a disposizione per questo numerosissimi metodi per che permettono l'interrogazione e l'aggiornamento delle tabelle, semplificando la gestione delle relazioni tramite l'uso delle funzioni di associazione. ActiveRecord gestisce anche le diverse tipologie di associazioni esistenti tra più tabelle (uno a uno, uno a molti e molti a molti), offrendo così una liberta ed un controllo totale delle nostre classi modello. Il codice relativo alle associazioni è inserito all'interno della classe modello stessa. Inoltre, ActiveRecord fornisce un meccanismo di correttezza semantica dei dati semplicemente applicando una serie di regole di validazione. Ovviamente la validazione viene eseguita prima dell'esecuzione delle istruzioni di INSERT e UPDATE sulla tabella. Il codice relativo alla validazione viene inserito all'interno dei file che definiscono il *model* nel framework MVC.

Come già discusso precedentemente, ActiveRecord possiede anche una feature molto interessante chiamata ***migrations***. Si tratta di una soluzione in puro codice Ruby per la gestione, per la creazione e per l'evoluzione di uno schema database. Un set di istruzioni per astrarre le differenze tra i diversi database e per gestire i cambiamenti all'interno dello stesso e delle sue tabelle. I vantaggi di lavorare con le migrations sono diversi, vediamone alcuni:

- i cambiamenti al nostro database avvengono tramite script, rendendo possibile la ricostruzione da zero dell'intera base di dati eseguendo gli script in sequenza e ricorrendo all'esecuzione del comando di rollback, se supportato dall'engine del db scelto, per il ripristino delle modifiche ;
- per tenere traccia degli aggiornamenti eseguiti su di una tabella, la stessa dovrebbe avere dei campi chiamati created_at e updated_at: quando si esegue una migration, ActiveRecord aggiorna automaticamente questi due campi con un timestamp;
- i campi di una tabella vengono popolati automaticamente e, se per un campo non viene specificato un valore, ActiveRecord gli assegna automaticamente il valore **NIL**.

Questa serie di caratteristiche e peculiarità rendono Rails un framework particolarmente adatto per lo sviluppo di applicazioni web supportate da database,

garantendo un ottimizzazione nei tempi di sviluppo, esponendo il meno possibile gli sviluppatori ai rischi di compiere errori di scrittura grazie alla gestione automatica del codice SQL e permettendo una facile manutenzione e leggibilità del codice grazie alla strutturazione automatica dello stesso.

5.5.3 Struttura delle tabelle del database di BioXml-Builder

Il database di BioXML-Builder è costituito dalle seguenti tabelle:

tabella **sources**: che, come indica il nome stesso, contiene i dati relativi a tutti i file presenti nei portali web delle banche dati biologiche e che BioXML-Builder riconosce e ingloba nel proprio sistema; ogni file è un oggetto - a cui corrisponde un relativo record attraverso il binding di ActiveRecord − con i relativi attributi:

- il nome del file (che funge da chiave primaria);
- il nome delle categoria di fonte (ai fini della visualizzazione per la scelta dell'utente);
- tutte le informazioni relative al reperimento del particolare file (l'*url della fonte di dati* e il protocollo da usare a seconda se il collegamento è in http o in ftp);
- l'*indirizzo di connessione* al server nel quale è contenuto il file;
- la directory *ftp* del file nel caso si trattasse di una fonte ftp;
- gli attributi relativi all'aggiornamento del file come la data di ultima modifica (che indica implicitamente la "versione" del file) e la data dell'ultimo download eseguito da BioXML-Builder;
- il tipo di file (text/xls);
- l'estensione nel caso in cui sia *compresso* (.zip, .rar, .gzip, …);
- una grammatica specifica ideata per raccogliere in un unico attributo due componenti essenziali relativi alla struttura del file nel caso in cui il formato sia CSV (il separatore di campo e quello di riga) che vengono poi opportunamente estrapolati in fase di importazione dei dati e il numero di linee iniziali da ignorare nel file di fonte per individuare le righe contenenti i dati veri e propri;

- gli attributi relativi al modello specifico che viene creato per ogni file ossia il nome della classe modello specifica (ai fini di attivarne il metodo statico printxml());
- l'elenco dei campi del particolare file di dati biologici (al fine di permetterne la selezione in visualizzazione);

serie di tabelle per ogni file delle fonti: sono state create tante tabelle quanti sono i file delle fonti biologiche, ciascuna contenente gli effettivi dati biologici che, nelle fasi successive, vengono manipolati dal meccanismo di conversione in XML; ad ogni tabella è stato aggiunto il campo *id*, posto come ultima colonna della tabella, per permetterne l'utilizzo del sistema di ActiveRecord. Ogni tabella corrisponde dunque ad un file e ad un modello ben preciso che contiene in se il metodo di classe (statico) *printxml()* che esegue la conversione in XML dei dati del rispettivo file;

tabella **pagine pubbliche**: comprende un campo *id* (chiave primaria), il *nome*, il *titolo* da visualizzare sulla relativa barra del browser, il campo contenente il *contenuto* (body) html della pagina stessa, e i campi dei *timestamp* assegnati da ActiveRecord;

tabella **utenti amministratori**: comprende un campo *id* (chiave primaria), il *nome*, la *login*, la *password* criptata e i campi dei *timestamp* assegnati da ActiveRecord;

tabella **sessioni**: contenente i dati delle varie sessioni per ogni utente che utilizza il sistema, che constano di un *session id*, dei *dati* stessi (che contengono le selezioni dei singoli utenti) ei *timestamp*;

tabella **schema migrations**: contenente le versioni di schema del database per ognuna delle migrations presenti nell'applicazione.

Le tabelle corrispondenti ai singoli file dei dati biologici vengono aggiornati periodicamente grazie ad una procedura scritta in Ruby che si trova nella directory *lib* sottoforma di script di Rake o *rakefile* che si illustra di seguito.

5.6 Aggiornamento del Database

Il database di BioXML-Builder appena esposto viene aggiornato periodicamente grazie ad uno script di Rake denominato ***bioupdate.rake***. Al suo interno, esso contiene una procedura (o *task*) che analizza la data e l'ora della modifica delle fonti remote in modo che, rilevando una recente loro modifica, provvederà a scaricarli nuovamente aggiornando e allineando il database locale di BioXML-Builder. Questa componente dell'applicazione è di fondamentale importanza e costituisce un punto di forza dell'applicazione in quanto garantisce che i dati contenuti nel database di BioXML-Builder siano sempre aggiornati con i dati pubblicati dalle fonti, e questo monitoraggio avviene più volte al giorno in maniera assolutamente automatica.

Questo aspetto rende tale applicazione utile anche in altri contesti, al di fuori di miRò, per allineare, più in generale, sistemi locali con dati remoti.

Cos'è Rake?

Rake è un'utility scritta da Jim Weirich in Ruby per eseguire comandi predefiniti. Questa libreria è universalmente riconosciuta come la versione Ruby del più conosciuto comando *make*, anche se si distingue da quest'ultimo per diversi fattori. Rake permette di creare task (procedure) completamente definiti in Ruby per eseguire qualsiasi tipo di comando: dalla compilazione di un programma alla manutenzione di un filesystem. Rails utilizza Rake in numerosissimi ambiti e numerosi task sono già incorporati nel framework per gestire varie funzioni comuni: funzioni per processare le modifiche sul database da un file di migration (rake db:migrate), per ripulire i file delle sessions, per esportare e importare lo schema del database, per generare test per l'applicativo, per creare documentazione per l'applicazione, per installare gems e molto altro.

Rake utilizza i blocchi di funzioni anonime di Ruby per definire le varie procedure all'interno del rakefile (lo script che raccoglie tutte le task).

Esiste una libreria che contiene funzioni per la manipolazione di file comuni e una libreria per rimuovere file compilati (la "clean" task). Così come fa make, anche Rake riesce a sintetizzare task basati su pattern, come ad esempio la costruzione

automatica di task per il confronto dei file basandosi sui pattern dei loro nomi. Rake fa parte ormai della libreria standard di Ruby versione 1.9 e presenta le seguenti caratteristiche:

- i Rakefile sono interamente definiti con la sintassi standard di Ruby, senza alcuna sintassi Makefile di cui preoccuparsi;
- gli utenti possono specificare task con determinati prerequisiti (come caricare l'environment dell'applicativo per manipolare i suoi modelli e aver accesso al suo codice);
- Rake supporta pattern di regole per sintetizzare task impliciti;
- possiede una libreria di task pre-create per facilitare la costruzione dei rakefile.

Alla luce di quanto esposto, procediamo con l'analisi della task di aggiornamento del database di BioXML-Builder.

5.6.2 La procedura update_db

La procedura di aggiornamento del database di BioXML-Builder si chiama ***update_db*** e si trova all'interno di un rakefile, *bioupdate.rake*, insieme ad altre task di test per l'applicativo. update_db carica come dipendenza l'environment dell'applicazione poiché necessita di manipolarne i modelli corrispondenti ai dati biologici. Attraverso la tabella *sources*, che contiene tutte le informazioni necessarie per reperire i file delle fonti di dati biologici online, la task controlla singolarmente i file, distinguendo opportunamente il tipo di fonte remota e controllando *da remoto* per ognuno di essi che la data di ultima modifica corrisponda a quella correntemente conservata nel relativo record della tabella sources. In caso di differenza tra le due date, la task procede all'aggiornamento del file, scaricandolo nuovamente, scompattandolo nel caso sia compresso e, a seconda del tipo di file (text/xls), richiama la rispettiva funzione di conversione ad un formato CSV convenientemente gestibile per caricare i dati attraverso la funzione messa a disposizione dal database MySql invocata con il comando *LOAD_DATA_INFILE*. In questa operazione di conversione, le funzioni coinvolte (contenute nel file *csv_generator.rb*) sono state concepite in modo tale da far fronte al problema dell'enorme quantità di record che i file di dati biologici contengono e

che potrebbero saturare la RAM se non fosse per l'opportuno utilizzo del *garbage collector* di Rails che periodicamente viene invocato durante la conversione. Inoltre, la task tiene conto della grammatica del file in questione (recuperata dalla tabella sources), che indica la struttura interna del file su come sono organizzati i dati internamente affinché la funzione di conversione possa eseguire il suo lavoro in maniera corretta. Una volta ottenuta la conversione nel formato CSV del file, questo viene dato in input al comando di MySQL, *LOAD_DATA_INFILE*, che in blocco esporta il contenuto nella rispettiva tabella avendo prima eseguito uno svuotamento dei dati obsoleti. Tale task viene automatizzato mediante l'uso del processo *cron* di linux che periodicamente lo invoca. Questo viene effettuato per tutti i file contenuti nella tabella sources. In tal modo, BioXML-Builder può accogliere file di dati biologici sempre nuovi, semplicemente inserendone le caratteristiche in tale tabella e creando un modello nel contesto dell'applicazione che tenga conto della particolare struttura dei dati per la rappresentazione in XML.

5.7 Interfaccia web di BioXml-Builder

Il front-end di BioXml-Builder è costituito da una serie di pagine web che guidano l'utente attraverso un wizard nella scelta oculata fra le opzioni messe a disposizione. L'utente può navigare tra le varie categorie di fonti di dati biologici che corrispondono alle varie banche dati libere presenti online e che sono state illustrate precedentemente. Per ogni categoria vengono elencati tutti i file che essa contiene e che sono inseriti nel sistema di BioXML-Builder e, per ognuno di questi, vi è la possibilità di selezionare i singoli campi di ciascun file, a seconda delle proprie esigenze, in modo da filtrare i soli dati di interesse e ottenere una versione dei dati nel formato XML in alternativa all'esportazione integrale che solitamente viene proposta come default. Tutte le selezioni vengono poi visualizzate in una pagina di riepilogo (equiparabile al "carrello" della spesa in un sistema di shopping on-line) che permette di deselezionare i file ivi visualizzati o confermare le scelte fatte. Se viene confermato il tutto, per ciascun file scelto viene eseguita l'operazione di esportazione nel formato XML secondo le esigenze espresse dall'utente. L'esportazione prevede una lettura del contenuto dei file che sono stati preventivamente caricati nel database con un formato CSV mediante il comando

LOAD_DATA_INFILE descritto precedentemente. Le righe corrispondenti al contenuto di ogni file vengono lette in successione ciclicamente e non in un'unica operazione per motivi di performance e scalabilità, dato che si tratta di tabelle contenenti centinaia di migliaia di record. Attraverso l'azione *file_conversion* del wizard_controller viene attivato il metodo di classe *printxml()* della classe del particolare modello a cui corrisponde il file e viene effettuata la conversione nel formato XML. Si ricorda che essendo ActiveRecord un ORM, ogni record considerato dalla tabella viene caricato in RAM come oggetto del modello, occupando memoria e comportando un peso prestazionale non indifferente. Per tale motivo i record vengono prelevati a piccoli blocchi, attraverso *find_in _batches* ("trova a blocchi") che è una funzione predefinita in rails che fa parte della categoria dei *finder dinamici* operante sulle tabelle del database per prelevare le informazioni a blocchi di record, anziché un record alla volta. In tal modo si ottimizza l'utilizzo della RAM durante lo svolgimento di tali operazioni. Tutti i file in formato XML vengono poi raggruppati in un unico archivio e mandati all'utente in risposta alla sua richiesta.

5.8 Sezione amministrativa

BioXML-Builder ha una sezione amministrativa progettata come un CMS, che si occupa degli account degli utenti amministratori nonché dell'amministrazione dell'aspetto visuale e funzionale dell'applicativo. La pagina amministrativa permette l'amministrazione di tre aspetti importanti: gli *account amministratori*, le *pagine pubbliche* e le *fonti di dati biologici*.

Attraverso un'interfaccia semplice ed intuitiva, è possibile creare, modificare e cancellare gli account degli utenti amministratori e proibire la visualizzazione della sezione amministrativa agli utenti non loggati. Tutto ciò è reso possibile dal plugin RESTful Authentication, discusso precedentemente. Le pagine pubbliche sono concepite come modelli nell'applicazione, per cui il loro contenuto è visto come attributo di tali modelli e quindi opportunamente conservato nel database in formato html, anche se, grazie al plugin *acts_as_textiled*, viene visualizzato come textile e quindi risulta più facilmente modificabile. Inoltre, a tutti gli utenti loggati come amministratori, è permesso modificare *in loco*, ossia senza dover accedere

alla sezione amministrativa, il contenuto delle pagine pubbliche dando fluidità alla gestione dell'aspetto visuale. Infine, è possibile aggiungere, modificare e cancellare i vari file di dati biologici, ampliando sempre più la gamma di banche dati online che BioXML-Builder attualmente gestisce. Questo permette uno sviluppo nel tempo e una sempre maggiore capacità di essere un punto di riferimento per la comunitaà scientifica bioinformatica.

5.9 L'interfaccia web di miRò

L'interfaccia web di miRò permette l'esecuzione di quattro diversi tipi di ricerche: la ricerca semplice, la ricerca avanzata, la ricerca personalizzata e il Data Mining.

Con la ricerca semplice si possono ottenere informazioni su un singolo oggetto quale un miRNA, un gene, un processo, una funzione, una malattia o un tessuto. Ad esempio, è possibile specificare un miRNA, o sceglierne uno dall'elenco a tendina, per ottenere la lista di tutte le malattie e i termini GO (processi e funzioni) che possono essere associati a quel miRNA attraverso i suoi target. I risultati sono ordinati in base al numero dei tool che predicono le coppie miRNA/target corrispondenti, e le associazioni validate sperimentalmente sono segnalate per prime. E' inoltre possibile specificare quali tool utilizzare come fonte delle predizioni, ponendoli in AND o in OR. Utilizzando l'AND per esempio, è possibile selezionare solo i termini associati a quei target predetti da tutti i tool selezionati. Questo può consentire l'identificazione delle associazioni maggiormente supportate e la riduzione dei falsi positivi. Analogamente, l'utente può ricercare tutti i miRNA associati ad un certo gene, malattia, processo o funzione, ed ottenere la lista di tutti i miRNA espressi in un certo tessuto con i relativi livelli di espressione.

La ricerca personalizzata permette di estendere la base di conoscenza con un proprio set di coppie miRNA/target. Queste coppie verranno memorizzate temporaneamente ed utilizzate in tutte le query della sessione. Questa funzionalità può essere utile per l'analisi di nuovi dati non ancora disponibili sui database ufficiali.

5.9.1 Interrogazione del database: ricerca semplice

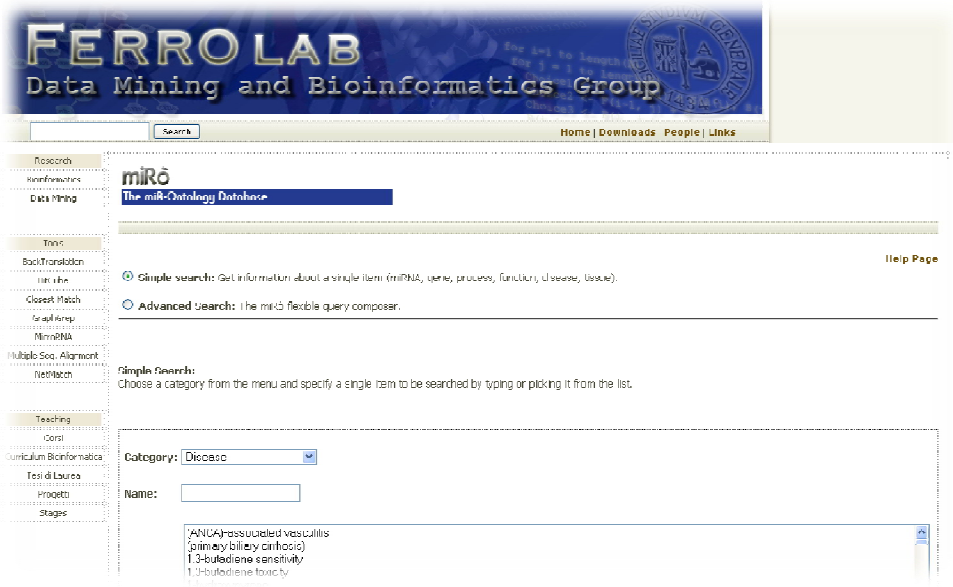
Il Sistema miR-Ontology interagisce con l'utente mediante un'applicazione web che nella prima versione è stata realizzata in PHP (linguaggio di scripting lato server, multipiattaforma45, incorporato nell'HTML). L'interfaccia web, sviluppata mediante pagine dinamiche HTML, permette di effettuare cinque tipi di ricerche semplici:

- **by Gene** che, parametrizzata con il simbolo del gene target (ufficiale o meno), restituisce tre sezioni. La prima visualizza la tabella che descrive le associazioni tra il gene e i miRNA mettendo in evidenza le fonti di provenienza, cioè specifica per ogni interazione se è sperimentalmente verificata o predetta dagli algoritmi di predizione. Gli elementi della tabella sono links che consentono all'utente di raggiungere pagine ancora più dettagliate, quali ad esempio quella che descrive le caratteristiche del miRNA e quelle che specificano gli allineamenti provenienti dai database esterni. La seconda sezione visualizza la tabella delle funzioni molecolari eseguite dal gene e la tabella dei processi biologici in cui il gene è coinvolto: ogni riga di tale tabella è un link a pagine presenti nel database GeneOntology. Infine l'ultima sezione mette in evidenza le patologie connesse, visualizzando, oltre al nome, anche il tipo di malattia e i links ai relativi articoli pubblicati su OMIM e PubMed;
- **by miRNA** che, parametrizzata con il nome del miRNA, restituisce 2 sezioni. La prima consiste nel visualizzare la tabella delle relazioni tra il miRNA e le patologie in cui essi sono parzialmente coinvolti attraverso i loro geni target. Per ogni entry si specificano i database da cui sono state estratte le associazioni tra i miRNA, i target e i links che consentono di raggiungere pagine contenenti informazioni ancora più dettagliate. Le successive due tabelle appartengono alla sezione miR-Ontology che mette in evidenza le funzioni molecolari e i processi biologici in cui è coinvolto il microRNA attraverso i suoi geni bersaglio e i database da cui sono stati prelevati i dati. Le colonne, che indicano i nomi delle funzioni e dei processi, sono link alle rispettive pagine presenti nel database GeneOntology;

- **by disease** che, parametrizzata con il nome della patologia, restituisce la tabella delle associazioni miRNA e gene ad essa connesse. Si specificano per ogni entry i database esterni che spiegano in dettaglio tali associazioni;
- **by ontology** che, parametrizzata con il nome della ontologia (funzione molecolare o processo biologico), visualizza una tabella di tutte le associazioni miRNA e gene ad essa correlate. Per ogni entry si forniscono i links a pagine contenenti in dettaglio le informazioni sui miRNA e sui geni regolati, gli allineamenti provenienti da TarBase se l'associazione è stata sperimentalmente verificata o dagli algoritmi di predizione trattati in questo lavoro di tesi se l'associazione è stata predetta. Per rendere le ricerche più semplici, l'applicazione web propone all'utente degli elenchi completi di tutti i simboli ufficiali dei microRNA e dei geni target presenti nel database. E' possibile visualizzare tutte le ontologie e le patologie trattate con la loro classe di appartenenza ed effettuare un ordinamento per nome o per tipo in base alle necessità. In seguito (Figura 5.6) è illustrata l'interfaccia web del sistema miR-Ontology.
- **by tissue** che, parametrizzata con il nome del tessuto, visualizza una tabella di tutte le associazioni tra i pre-miRNA, i mature-miRNA e i tessuti nei quali essi sono presenti. Nella tabella restituita come risultato, è possibile vedere la libreria alla quale appartiene il tessuto scelto e l'espressione dei pre-mirna o mature-mirna (a seconda della tipologia scelta) in quel tessuto.

Di seguito sono state inserite alcune pagine relative alle ricerche semplici descritte.

Ricerca semplice per miRNA

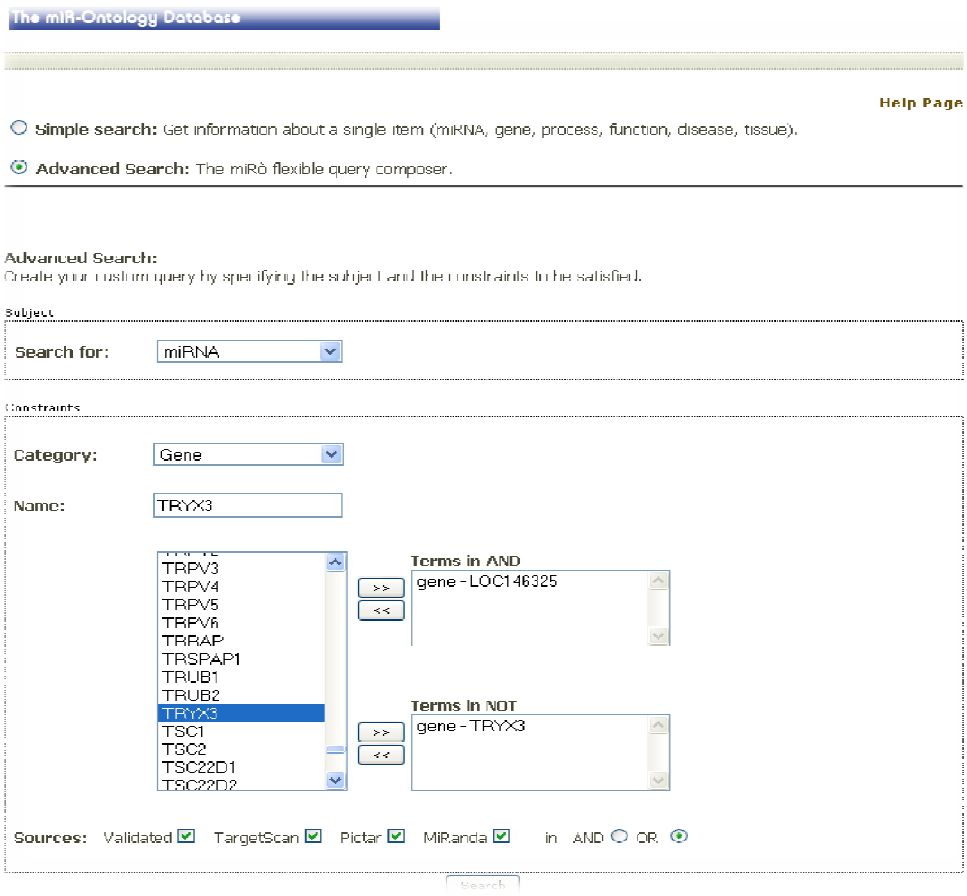


Ricerca semplice per malattia

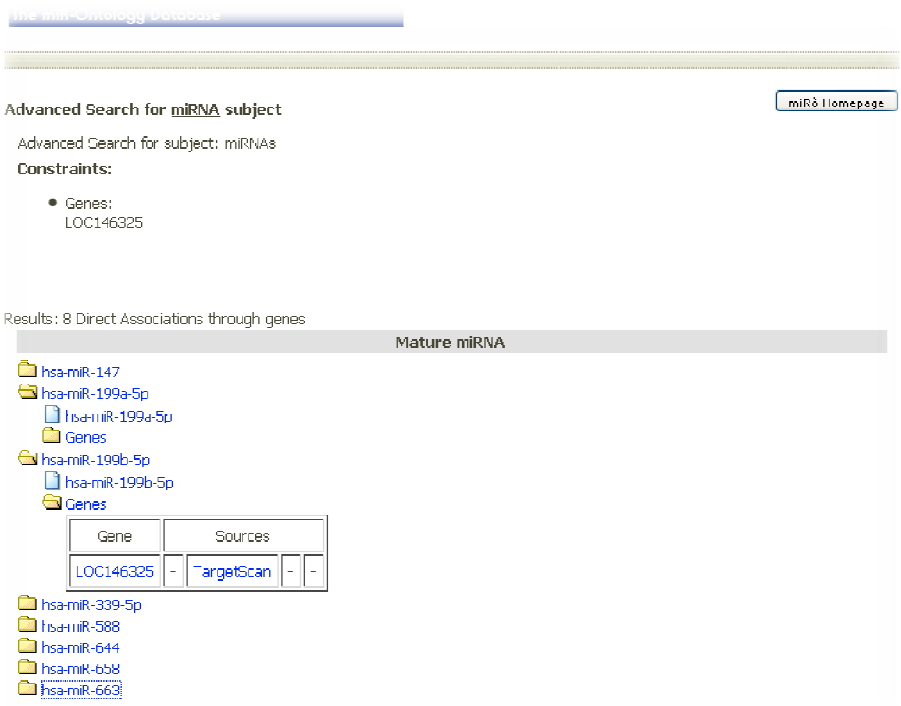
5.9.2 Interrogazione del database: ricerca avanzata

Uno degli aspetti più significativi dell'application web miRò, è la capacità di trovare le correlazioni tra i miRNA e tutte le altre tipologie di dati (geni, malattie, ontologie e tessuti) attraverso un filtro avanzato che consente di escludere o aggiungere delle clausole durante il criterio di ricerca, attraverso i connettori logici AND e OR. Grazie a questa feature, l'utente è in grado di effettuare delle ricerche più specifiche, escludendo possibili correlazioni che non devono essere tenute in considerazione in quanto non sono di interesse nella ricerca.

La ricerca avanzata permette di eseguire query più sofisticate. L'utente può scegliere un soggetto tra miRNA, gene, malattia, processo o funzione, e specificare una lista di vincoli che tale soggetto deve soddisfare. Ad esempio, è possibile chiedere al sistema di mostrate tutti i miRNA associati ai termini *heart failure*, *RNA binding* ed *apoptosis*, ma non al termine *congenital heart disease*. L'utente può inoltre scegliere le fonti delle coppie miRNA/target. Questo consente di restringere o rilassare le condizioni della query, in modo da ottenere un output più piccolo o più grande, rispettivamente. Il sistema mostrerà la lista di tutti i miRNA che soddisfano i vincoli specificati, con i dettagli relativi ai target coinvolti (vedi Fig. 6.2). Questo tool di interrogazione collega oggetti mediante associazioni basate su miRNA. Ad esempio, una malattia *d*ed un processo *p* che non sono correlati attraverso alcun gene comune, potrebbero essere associati attraverso un miRNA che regola un gene g_d, coinvolto in *d*, ed un gene g_p, coinvolto in *p*. Questo introduce un nuovo livello di associazione tra geni e fenotipi, basato sulle annotazioni dei miRNA. Tali associazioni sono restituite dalla ricerca avanzata, quando il soggetto della query è un termine GO o una malattia.



Ricerca avanzata per Gene



Risultato della ricerca avanzata per Gene

Dalla figura precedente è possibile vedere la sottotabella contenuta in corrispondenza del miRNA hsa-miR-199b-5p che riporta tutti i geni associati ad esso; nello specifico è stato ottenuto un solo risultato riportante il gene LOC1463325 che era stato impostato attraverso il filtro di ricerca.

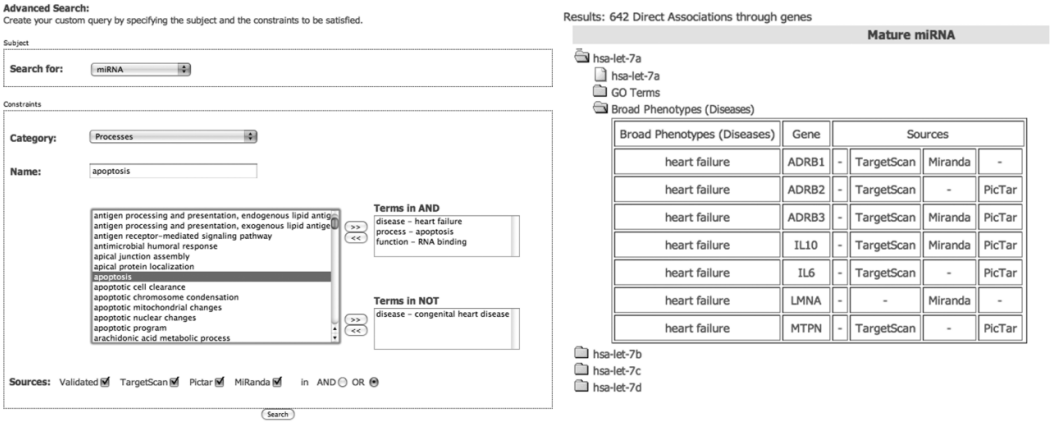
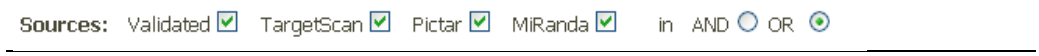


Fig. 6.2 - Un esempio di esecuzione di query avanzata. (a) Il modulo della ricerca avanzata con i vincoli selezionati: saranno restituiti i miRNA relativi ai termini*heart failure*, *apoptosis* e *RNA binding*, ma non a*congenital heart disease*. (b) Un sottoinsieme dei risultati corrispondenti. Ad esempio, il miRNA let-7a soddisfa i requisiti specificati. In particolare, let-7a ha 7 target predetti, coinvolti nella malattia *heart failure*. I dettagli relativi agli altri vincoli (*apoptosis* e *RNA binding*) sono mostrati nella cartella*GO Terms*.

Oltre ai geni correlati, la sottotabella mostra per ciascuno di essi anche la sorgente che ha predetto tale correlazione (nello specifico si tratta di TargetScan). Tramite il sistema del filtraggio avanzato dei dati, è possibile escludere o prendere in considerazione una sorgente, combinando tale scelta con l'esclusione o meno simultanea delle altre sorgenti.



Scelta in OR e AND delle sorgenti

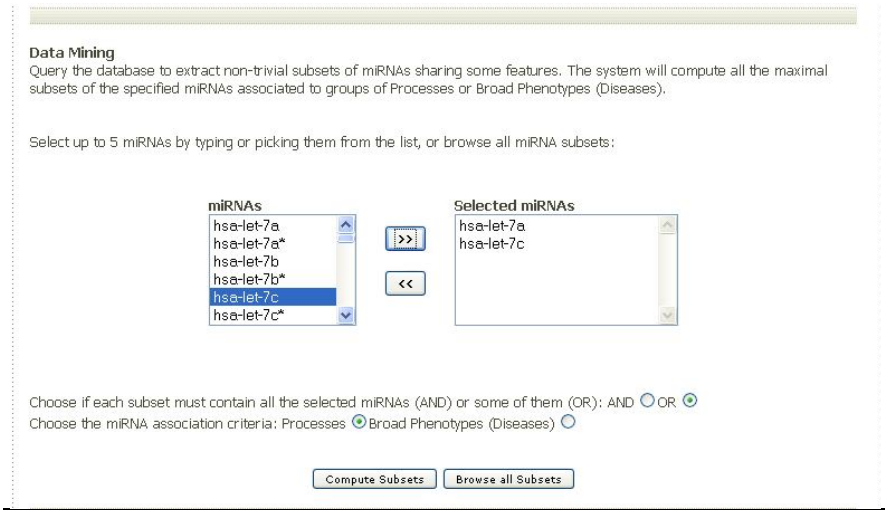
5.9.3 Data mining in miRò

L'associazione dei miRNA ai processi e alle malattie, permette di raggrupparli in base ai termini comuni. A tal fine, miRò è stato dotato di un modulo di data mining, basato sul calcolo degli insiemi frequenti massimali [136, 137]. Tale funzionalità offre all'utente la possibilità di interrogare il database per l'estrazione di sottoinsiemi non banali di miRNA che condividono certe caratteristiche. L'analisi è eseguita utilizzando diverse soglie di supporto: una soglia alta consente di ottenere un piccolo numero di sottoinsiemi di miRNA associati ad un grande numero di termini, mentre soglie più basse restituiscono un numero maggiore di sottoinsiemi associati a pochi termini. Tutti i sottoinsiemi sono pre-calcolati sul dataset delle coppie miRNA/target validate.

Nell'interfaccia di miRò, l'utente può scegliere fino ad un numero massimo stabilito di miRNA insieme ad un criterio di associazione (ad es. processo o malattia). Il sistema estrae tutti i sottoinsiemi contenenti i miRNA selezionati e i processi e le malattie ai quali essi sono più frequentemente associati. Questo tipo di analisi può permettere la formulazione di ipotesi circa l'azione cooperativa di un insieme di miRNA nell'espletamento di certe funzioni biologiche. Ad ogni associazione miRNA/termine viene inoltre attribuito uno score di specificità, che permette di evidenziare le associazioni più significative, come discusso nel prossimo sotto-paragrafo.

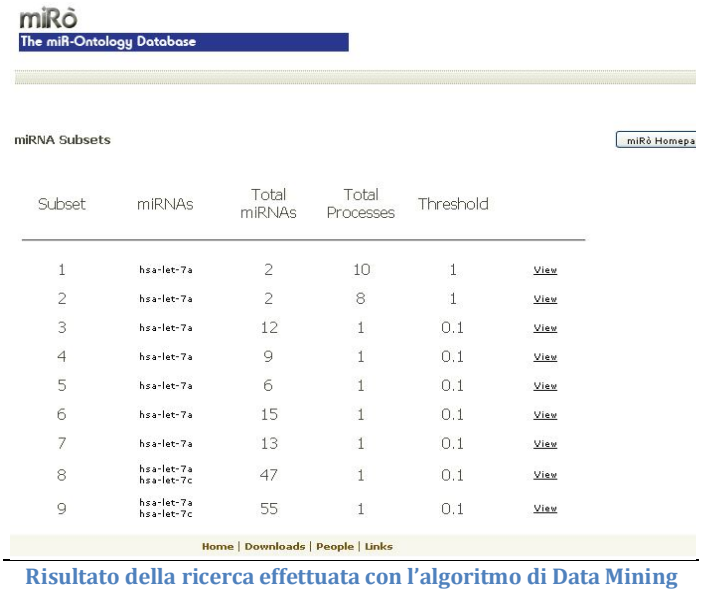
5.9.4 Interrogazione del database: Datamining

Un'altra interessante caratteristica di miRò è la parte di ricerca che utilizza l'algoritmo di Data Mining A-priori per inferire quali miRNA sono coinvolti contemporaneamente in una o più malattie o processi biologici. Il sistema, attraverso una semplice interfaccia, consente di scegliere uno o più miRNA dei quali si vuole conoscere la possibile correlazione in una o più malattie (o in uno o più processi).



Datamining area

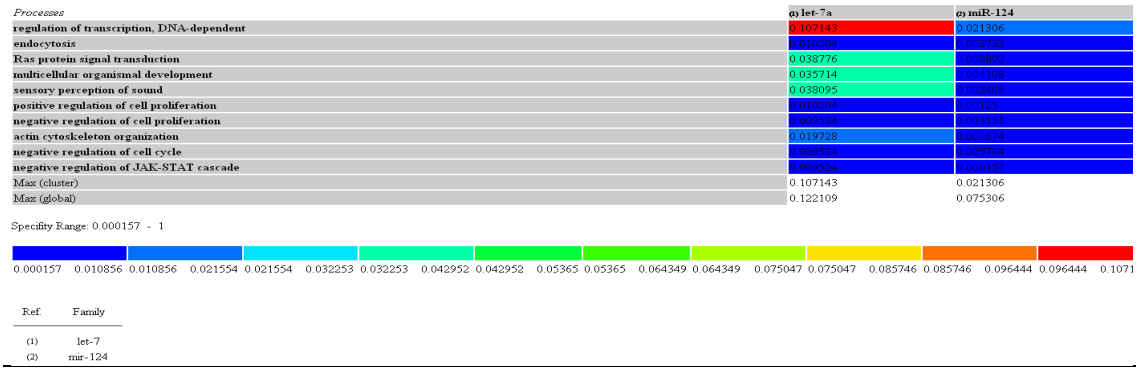
Come risultato della ricerca, si otterrà una tabella riportante l'identificatore del subset, il miRNA o i miRNA correlati, il numero dei miRNA coinvolti, il numero di malattie o processi in cui i miRNA sono coinvolti e la soglia utilizzata per ottenere il relativo risultato.



Risultato della ricerca effettuata con l'algoritmo di Data Mining

Scegliendo di visualizzare uno dei risultati ottenuti, cliccando sul link *view*, si vedrà una tabella di riepilogo che mostra i processi (o le malattie) correlati ai miRNA e il valore di specificità del miRNA per ogni processo (o malattia) caratterizzato da un

colore attraverso una mappa dei colori che va dal blu al rosso, secondo il valore se è più vicino allo zero o al valore 1.



Processes	ω let-7a	ω miR-124
regulation of transcription, DNA-dependent	0.107143	0.021306
endocytosis	0.010204	0.002793
Ras protein signal transduction	0.038776	0.008202
multicellular organismal development	0.035714	0.004198
sensory perception of sound	0.038095	0.002406
positive regulation of cell proliferation	0.010834	0.00125
negative regulation of cell proliferation	0.009524	0.003151
actin cytoskeleton organization	0.019728	0.002674
negative regulation of cell cycle	0.009524	0.005704
negative regulation of JAK-STAT cascade	0.009524	0.000157
Max (cluster)	0.107143	0.021306
Max (global)	0.122109	0.075306

Specifity Range: 0.000157 - 1

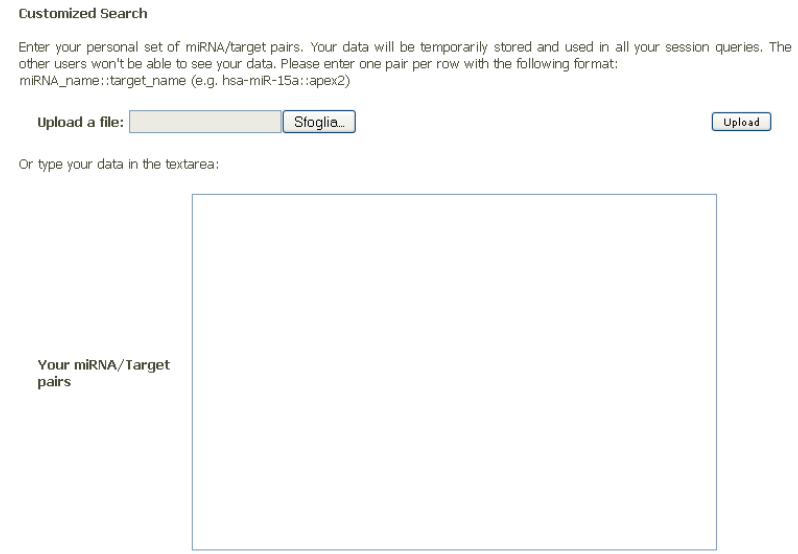
0.000157 0.010856 0.010856 0.021554 0.021554 0.032253 0.032253 0.042952 0.042952 0.05365 0.05365 0.064349 0.064349 0.075047 0.075047 0.085746 0.085746 0.096444 0.096444 0.1071

Ref.	Family
(1)	let-7
(2)	mir-124

Mappa dei colori relativa alla specificità di un miRNA per ogni processo

5.9.5 Interrogazione del database: Customized search

Ultima caratteristica di miRò è la possibilità di utilizzare un archivio personale di associazioni miRNA e geni target contenuti in un flat file, forniti al sistema attraverso un meccanismo di upload. Il file che verrà dato in input al sistema dovrà rispettare la seguente sintassi: *miRNA_name::target_name* (ad es. hsa-miR-15a::apex2). L'alternativa a ciò, è data dalla compilazione manuale dell'area di testo con le relative associazioni. Una volta inviata la richiesta di elaborazione, il sistema utilizzerà l'archivio inviato come dataset sperimentale temporaneo sul quale saranno eseguite tutte le ricerche messe a disposizione dal sistema descritte precedentemente.



Customized Search

6 CONCLUSIONI E SVILUPPI FUTURI

Questa tesi di dottorato racchiude al suo interno una serie di attività svolte al fine di realizzare un sistema web integrato di qualità, in grado di costituire un unico strumento valido ai fini della ricerca nel campo della biologia molecolare; l'obiettivo è quello di evitare il disperdersi delle energie per ottenere gli stessi risultati dovendo utilizzare contemporaneamente più strumenti e fonti biologiche presenti sul web.

Il lavoro eseguito per la realizzazione di quest'applicazione web è stato fatto secondo un criterio di modularizzazione del codice che ne permette un veloce assemblaggio per facilitare lo sviluppo di nuove funzioni. L'analisi che ha preceduto lo sviluppo è stata condotta con scrupolosità proprio per garantire una veloce integrazione con altri moduli web-oriented. Uno dei moduli che è già in fase di sviluppo è quello basato sui web services, in grado di permettere ad applicazioni web esterne, di effettuare le richieste allo stesso modo in cui vengono fatte dagli utenti, attraverso dei semplici script invocati via URL. Grazie a questa funzione, sarà possibile automatizzare le procedure di recupero, conversione e analisi dei dati biologici, permettendo all'applicazione di divenire un elemento fondamentale all'interno dei più diffusi workflow biologici oggi presenti. Un altro aspetto che si sta curando, è quello relativo all'applicazione di algoritmi di datamining sulle attività svolte dagli utenti attraverso miRò, in modo da ottimizzare tutte le operazioni che più comunemente vengono richieste dagli stessi. Questa serie di caratteristiche, ci permettono di pensare che l'applicativo miRò offrirà vantaggi in termini di qualità dell'informazione e velocità nel reperimento tali da renderlo unico e di riferimento fra gli strumenti di ricerca e sviluppo utili per tutta la comunità scientifica. Attualmente miRò si è dimostrato utilissimo per la ricerca biologica di base e, soprattutto, biomedica di diversi gruppi internazionali operanti sia negli USA che in Europa ed in Israele. Il lavoro svolto è stato pubblicato sulla rivista "Database: The Journal of Biological Databases and Curation" della Oxford University Press, una delle più prestigiose riviste internazionali nel settore dei database biologici. I risultati ottenuti, mostrano dunque l'importanza dell'impiego di metodi computazionali efficienti per l'analisi dei dati biologici, per la ricerca di pattern significativi e per la predizione di meccanismi e funzioni, e gettano le basi per ulteriori suoi sviluppi futuri.

Bibliografia

[1] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004;**116**:281–297.

[2] Fire A. et al. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 1998; **391**(6669): 806-11.

[3] Tomari Y and Zamore PD. Perspective: machines for RNAi. Genes Dev 2005;**19**: 517–529. [4] Meister G and Tuschl T. Mechanisms of gene silencing by double-stranded RNA. Nature 2004; **431**: 343–349.

[5] Zhang H, Kolb F, Jaskiewicz L, Westhof E, Filipowicz W. Single processing center models for human Dicer and bacterial RNase III. Cell 2004; **118**: 57–68.

[6] Macrae IJ, Zhou K, Li F, Repic A, Brooks AN, Cande WZ, Adams PD, Doudna JA. Structural basis for double-stranded RNA processing by Dicer. Science 2006; **311**: 195–198.

[7] Yigit E et al. Analysis of the C. elegans Argonaute family reveals that distinct Argonautes act sequentially during RNAi. Cell 2006; **127**: 747–757.

[8] Parker JS et al. Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. EMBO J 2004; **23**: 4727–4737.

[9] Song JJ et al. Crystal structure of Argonaute and its implications for RISC slicer activity. Science 2004; **305**: 1434–1437.

[10] Talmor-Neiman M et al. Identification of trans-acting siRNAs in moss and an RNA-dependent RNA polymerase required for their biogenesis. Plant J. 2006; **48**(4): 511-21.

[11] Fahlgren Net al. Regulation of AUXIN RESPONSE FACTOR3 by TAS3 ta-siRNA affects developmental timing and patterning in Arabidopsis. Curr Biol. 2006; **16**(9): 939-44.

[12] Montgomery TA et al. Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. Cell. 2008; **133**(1): 128-41.

[13] Mette MF et al. Transcriptional silencing and promoter methylation triggered by double stranded RNA. EMBO J. 2000; **19**(19): 5194-201.

[14] Vagin VV et al. Adistinct small RNA pathway silences selfish genetic elements in the germline. Science. 2006; **313**(5785): 320-4.

[15] Mochizuki K and Gorovsky MA. Conjugation-specific small RNAs in tetrahymena have predicted properties of scan (scn) RNAs involved in genome rearrangement. Genes Dev. 2004; **18**(17): 2068-73.

[16] Katiyar-Agarwal S et al. A novel class of bacteria-induced small RNAs in Arabidopsis. Genes Dev. 2007; **21**(23): 3123-34.

[17] Matranga C et al. Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. Cell 2005; **123**: 607–620.

[18] Miyoshi K et al Slicer function of Drosophila Argonautes and its involvement in RISC formation. Genes Dev 2005; **19**:2837–2848.

[19] Rand TA et al. Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. Cell 2005; **123**:621–629.

[20] Gregory RI et al. Human RISC couples microRNA biogenesis and post-transcriptional gene silencing. Cell 2005; **123**:631–640.

[21] Maniataki E and Mourelatos Z. A human, ATP-independent, RISC assembly machine fueled by pre-miRNA. Genes Dev 2005; **19**:2979–2990.

[22] Kanellopoulou C et al. Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. Genes Dev 2005; **19**: 489–501.

[23] Murchison EP et al. Characterization of Dicer-deficient murine embryonic stem cells. Proc. Natl. Acad. Sci. USA 2005; **102**: 12135–12140.

[24] Kim VN. MicroRNA biogenesis: coordinated cropping and dicing. Nat. Rev. Mol. Cell Biol 2005; **6**: 376–385.

[25] Denli AM et al. Processing of primary microRNAs by the Microprocessor complex. Nature 2004; **432**: 231–235.

[26] Okamura K et al. The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell 2007; **130**: 89–100.

[27] Ruby JG et al. Intronic microRNA precursors that bypass Drosha processing. Nature 2007; **448**: 83–86.

[28] Li X and Carthew RW. A microRNA mediates EGF receptor signaling and promotes photoreceptor differentiation in the Drosophila eye. Cell 2005; **123**: 1267–1277.

[29] Seggerson K et al. Two genetic circuits repress the Caenorhabditis elegans heterochronic gene lin-28 after translation initiation. Dev. Biol 2002; **243**: 215–225.

[30] Okamura K et al. The regulatory activity of microRNA* species has substantial influence on microRNA and 3′ UTR evolution. Nat. Struct. Mol. Biol 2008; **15**: 354–363.

[31] Bagga S et al. Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. Cell 2005; **122**: 553–563.

[32] Behm-Ansmant I et al. mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. Genes Dev 2006; **20**: 1885–1898.

[33] Giraldez AJ et al. ZebrafishMiR-430 promotes deadenylation and clearance of maternal mRNAs. Science 2006; **312**: 75–79.

[34] Wu L et al. MicroRNAs direct rapid deadenylation of mRNA. Proc. Natl. Acad. Sci. USA 2006; **103**: 4034–4039.

[35] Aleman LM et al. Comparison of siRNA-induced off-target RNA and protein effects. RNA 2007; **13**: 385–395.

[36] Bernstein E et al. Dicer is essential for mouse development. Nat Genet 2003; **35**: 215–217.

[37] Harris KS et al. Dicer function is essential for lung epithelium morphogenesis. Proc Natl Acad Sci U S A 2006; **103**: 2208-13.

[38] Harfe BD et al. The RNaseIII enzyme Dicer is required for morphogenesis but not patterning of the vertebrate limb. Proc Natl Acad Sci U S A 2005; **102**: 10898-903.

[39] O'Rourke JR et al. Essential role for Dicer during skeletal muscle development. Dev Biol 2007; **311**: 359-68.

[40] Muljo SA et al. Aberrant T cell differentiation in the absence of Dicer. J Exp Med 2005; **202**: 261-9.

[41] Ventura A et al. Targeted Deletion Reveals Essential and Overlapping Functions of the miR-17-92 Family of miRNA Clusters. Cell 2008; **132**: 875–886.

[42] Chen C et al. MicroRNAs modulate hematopoietic lineage differentiation. Science 2004; **303**: 83–86.

[43] Lu Jet al. MicroRNA-mediated control of cell fate in megakaryocyte-erythrocyte progenitors. Dev Cell 2008; **14**: 843–853.

[44] Sempere L et al. Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. Genome Biol 2004; **5**: R13.

[45] Sokol N and Ambros V. Mesodermally expressed DrosophilamicroRNA-1 is regulated by Twist and is required in muscles during larval growth. Genes Dev 2005; **19**: 2343–2354.

[46] Callis T et al. MicroRNAs in skeletal and cardiac muscle development. DNA Cell Biol 2007; **26**: 219–225.

[47] Wong C and Tellam R. MicroRNA-26a targets the histone methyltransferase Enhancer of Zeste homolog 2 during myogenesis. J Biol Chem 2008; **283**: 9836–9843.

[48] Boutz P et al. MicroRNAs regulate the expression of the alternative splicing factor nPTB during muscle development. Genes Dev 2007; **21**: 71–84.

[49] Makeyev E et al. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. Mol Cell 2007; **27**: 435–448.

[50] Wang WX et al. The expression of microRNA miR-107 decreases early in Alzheimer's disease and may accelerate disease progression through regulation of beta-site amyloid precursor protein-cleaving enzyme 1. J Neurosci 2008; **28**:1213-1223.

[51] Kim J et al. A MicroRNA feedback circuit in midbrain dopamine neurons. Science 2007; **317**: 1220-1224.

[52] Bilen J et al.A new role for microRNA pathways: modulation of degeneration induced by pathogenic human disease proteins. Cell Cycle 2006; **5**: 2835-2838.

[53] Lin SL et al. First in vivo evidence of microRNA induced fragile X mental retardation syndrome. Mol Psychiatry 2006; **11**: 616-617.

[54] Perkins DO et al. microRNA expression in the prefrontal cortex of individuals with Current Opinion in Pharmacology 2008; **8**: 661–667.

[55]Chang J et al. miR-122, a mammalian liver-specific microRNA, is processed from hcr mRNA and may downregulate the high affinity cationic amino acid transporter CAT-1. RNA Biol 2004; **1**: 106-113.

[56] Pedersen IM et al. Interferon modulation of cellular microRNAs as an antiviral mechanism. Nature 2007; **449**: 919-922.

[57] Eisenberg I et al. Distinctive patterns of microRNA expression in primary muscular disorders. Proc Natl Acad Sci U S A 2007; **104**: 17016-17021.

[58] Care` A et al. MicroRNA-133 controls cardiac hypertrophy. Nat Med 2007; **13**: 613-618.

[59] Yang B et al. The muscle-specific microRNA miR-1 regulates cardiac arrhythmogenic potential by targeting GJA1 and KCNJ2. Nat Med 2007; **13**: 486-491.

[60] Cimmino A et al. miR-15 and miR-16 induce apoptosis by targeting BCL2. Proc Natl Acad Sci USA 2005; **102**(39): 13944-13949.

[61] Johnson SM et al. RAS is regulated by the let-7 microRNA family. Cell 2005; **120**: 635-647.

[62] Kumar MS et al. Suppression of non-small cell lung tumor development by the let-7 microRNA family. Proc Natl Acad Sci USA 2008; **105**: 3903-3908.

[63] Costinean S et al. Pre-B cell proliferation and lymphoblastic leukemia/high-grade lymphoma in E(mu)-miR155 transgenic mice. Proc Natl Acad Sci USA 2006; **103**: 7024-7029.

[64] Petrocca F et al. E2F1-regulated microRNAs impair TGFbeta-dependent cell-cycle arrest and apoptosis in gastric cancer. Cancer Cell 2008; **13**: 272-286.

[65] Eisenberg I et al. Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. Nature 2007; **449**: 682-688.

[66] Tavazoie SF et al. Endogenous human microRNAs that suppress breast cancer metastasis. Nature 2008; **451**: 147-152.

[67] Griffiths-Jones S et al. miRBase: tools for microRNA genomics. Nucleic Acids Res 2008; **36**(Database issue): D154-8.

[68] Sethupathy, P et al. TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA 2006; **12**: 192-197.

[69] Xiao F et al. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. 2009; **37**(Database issue): D105-10.

[70] Lewis B et al. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. Cell 2005; **120**: 15-20.

[71] Hofacker IL. How microRNAs choose their targets. Nat Genetics 2007; **39**(10): 1191-92.

[72] Grimson A et al. MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. Mol Cell 2007; **27**: 91-105.

[73] John B et al. Human MicroRNA targets. PLoS Biology 2004; **2**(11): 1862-79.

[74] Krek A et al. Combinatorial microRNA target predictions. Nat Genetics 2005; **37**(5): 495-500.

[75] Kiriakidou M et al. A combined computational-experimental approach predicts human microRNA targets. Genes Dev 2004; **18**(10):1165-78.

[76] Miranda KC et al. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell 2006; **126**: 1203-1217.

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

[78] Rusinov V et al. MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. Nucleic Acid Res. 2005; **33**(Web Server Issue): W696-W700.

[79] Schubert S et al. Local RNA target structure influences siRNA efficacy: systematic analysis of intentionally designed binding regions. J. Mol. Biol. 2005; **348**: 883-893.

[80] Robins H et al. Incorporating structure to predict microRNA targets Proc Natl Acad Sci USA 2005; **102**: 4006-4009.

[81] Doench JG and Sharp PA. Specificity of microRNA target selection in translational repression. Genes Dev. 2004; **18**: 504-511.

[82] Vella MC et al. Architecture of a validated microRNA: target interaction. Chem. Biol. 2004; **11**: 1619-1623.

[83] Vienna RNA Package. http://www.tbi.univie.ac.at/RNA/

[84] McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 1990; **29**: 1105-1119.

[85] Needleman SB and Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970; **48**(3): 443-453.

[86] Bernhart S. et al. Partition function and base pairing probabilities of RNA heterodimers. Algorithms Mol Biol2006;**1**(1):3.

[87] Mückstein U. et al. Thermodynamics of RNA-RNA binding. Bioinformatics2006; **22**(10): 1177-1182.

[88] Clop A et al. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nature Genetics2006: **38**(7): 813-818.

[89] Hariharan M et al. Targets for human encoded microRNAs in HIV genes. Biochem Biophys Res Commun 2005; **337**(4): 1214-1218.

[90] Jopling CL et al. Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. Science2005; **309**(5740): 1577-1581.

[91] Long D et al. Potent effect of target structure on microRNA function. Nature Struct and Mol Biol 2007; **14**(4):287-94.

[92] Kertesz M. et al. The role of site accessibilità in microRNA target recognition. Nature Genetics 2007; **39**(10): 1278-84.

[93] Didiano D and Hobert O. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. Nature Struct and Mol Biol 2006; **13**(9): 849-51.

[94] Tafer H et al. The impact of target site accessibilità on the design of effective siRNAs. Nature Biotechnology 2008; **26**(5). 578-83.

[95] Nielsen CB et al. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. RNA 2007; **13**: 1894-1910.

[96] Didiano D and Hobert O. Molecular architecture of a miRNA-regulated 3' UTR. RNA 2008; **14**: 1297-1317.

[97] Shulman-Peleg A et al. RsiteDB: a database of protein binding pockets that interact with RNA nucleotide bases. Nucleic Acids Res. 2009; **37**(Database issue): D369-73.

[98] Nair V and Zavolan M.Virus-encoded microRNAs: novel regulators of gene expression. Trends Microbiol. 2006; **14**(4): 169-175.

[99] Cullen BR.Viruses and microRNAs. Nature Genetics 2006; **38** Suppl.:S25-30.

[100] Sullivan CS et al.SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. Nature 2005; **435**(7042): 682-686.

[101] Klase Z et al. HIV-1 TAR miRNA protects against apoptosis by altering cellular gene expression. Retrovirology 2009; **6**: 18.

[102] Mendez E et al. Caspases mediate processing of the capsid precursor and cell release of human astroviruses. J Virology 2004; **78**(16): 8601-8608.

[103] Grimm T et al. EBV latent membrane protein-1 protects B cells from apoptosis by inhibition of BAX. Blood 2004: **105**(8):3263-3269.

[104] Damania B. DNA tumor viruses and human cancer. Trends in Microbiol. 2007; **15**(1): 38-44.

[105] Si H and Robertson ES. Kaposi's sarcoma-associated herpesvirus-encoded latency-associated nuclear antigen induces chromosomal instability through inhibition of p53 function. *J Virology* 2006; **80**(2): 697-709.

[106] Pfeffer S et al. Identification of virus-encoded microRNAs. Science 2004; **304**: 734-736.

[107] Cai X et al. Epstein-Barr virus microRNAs are evolutionarily conserved and differentially expressed. PLoS Pathog.2006; **2**(3): e23.

[108] Kim do N et al. Expression of viral microRNAs in Epstein-Barr virus-associated gastric carcinoma. J Virology 2007; **81**(2): 1033-1036.

[109] Pfeffer S et al. Identification of microRNAs of the herpesvirus family. Nature methods 2005; **2**(4): 269,276.

[110] Marshall V et al. Conservation of virally encoded microRNAs in Kaposi sarcoma-associated herpesvirus in primary effusion lymphoma cell lines and in patients with Kaposi sarcoma or multicentric Castleman disease. J Infect Dis 2007; **195**(5): 645-659.

[111] Pampin M et al. Cross talk between PML and p53 during poliovirus infection: implications for antiviral defense. J Virology 2006; **80**(17): 8582-8592.

[112] Royds JA et al. p53 promotes adenoviral replication and increases late viral gene expression. Oncogene 2006; **25**(10): 1509-1520.

[113] Pfeffer S and Voinnet O. Viruses, microRNAs and cancer. Oncogene 2006; **25**: 6211-6219.

[114] Luo M et al. RNAi of neuro peptide Y (NPY) for neuropathic pain. Society for Neuroscience Abstracts 2005.

[115] Filleur S et al. SiRNA-mediated inhibition of vascular endothelila growth factor severly limits tumor resi stance to antiangiogenic thrombospondin-1 and slows tumor vascularization and growth. Cancer Research 2003; **63**(14): 3919-22.

[116] Zhang Y et al. Intravenous RNA interference gene therapy targeting the human epidermal growth factor receptor prolongs survival in intracranial brain cancer. Clin Cancer Research 2004; **10**(11): 3667-77.

[117] Gaudilliere B et al. RNA interference reveals a requirement for MEF2A in activity-dependent neuronal survival. Journal of Biol Chemistry 2002; **277**(48): 46442-6.

[118] Gonzalez-Alegre P et al. Silencing primary dystonia: lentiviral-mediated RNA interference therapy for DYT1 dystonia. Journal of Neuroscience 2005; **25**(45): 10502-9.

[119] Brummelkamp TR et al. A system for stable expression of short interfering RNAs in mammalian cells. Science 2002; **296**: 550-3.

[120] Martinez LA et al. Synthetic small inhibiting RNAs: efficient tools to inactivate oncogenic mutations and restore p53 pathways. Proc Natl Acad Sci USA 2002; **99**(14): 849-54.

[121] Choudhury A. et al. Small interfering RNA (siRNA) inhibits the expression of the Her2/neu gene, upregulates HLA class I and induces apoptosis of Her2/neu positive tumor cell lines. Int J Cancer 2004; **108**: 71-7.

[122] Scherr M et al. Specific inhibition of bcr-abl gene expression by small interfering RNA. Blood 2003; **101**: 1566-9.

[123] Zhang X et al. Enhancement of hypoxia-induced tumor cell death in vitro and radiation therapy in vivo by use of small interfering RNA targeted to hypoxia-inducible factor-1α. Cancer Res 2004; **64**: 8139-42.

[124] Song E et al. Antibody mediated in vivo delivery of small interfering RNAs via cell-surface receptors. Nature Biotechnology 2005; **23**: 709-17.

[125] Tsuda N et al. Synthetic MicroRNA designed to target Glioma-Associated Antigen 1 transcription factor inhibits division and induces late apoptosis in pancreatic tumor cells. Clin Cancer Res 2006; **12**(21): 6557-64.

[126] Kasper et al. GLI transcription factors: mediators of oncogenic Hedgehog signalling. Eur J Cancer 2006; **42**(4): 437-45.

[127] Thayer SP et al. Hedgehog is an early and late mediator of pancreatic cancer tumorigenesis. Nature 2003; **425**(6960): 851-6.

[128] Sanchez P et al. Inhibition of prostate cancer proliferation by interference with SONIC HEDGEHOG-GLI1 signaling. Proc Natl Acad Sci USA 2004; **101**(34): 12561-6.

[129] Bernhart SH et al. Local RNA base pairing probabilities in large sequenze. Bioinformatics 2006; **22**: 614-15.

[130] Istituto tecnologie biomediche, sezione di Bari: http://www.ba.itb.cnr.it/bighome/

[131] Chang K et al. Lessons from nature: microRNA-based shRNA libraries. Nature Met 2006; **3**(9): 707-14.

[132] Nam S et al. miRGator: an integrated system for functional annotation of microRNAs. Nucleic Acids Res. 2007; **36**(Database issue): D159-64.

[133] Landgraf P et al. A mammalian microRNA expression atlas based on small RNA library sequencing. Cell 2007; **129**(7): 1401-14.

[134] Ashburner M et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genetics 2000; **25**(1): 25-29.

[135] Becker KG et al. The genetic association database. Nat Genetics 2004; **36**(5): 431-432.

[136] Agrawal R et al. Fast Algorithms for Mining Association Rules. Proc. 20th Int. Conf. Very Large Data Bases, VLDB 1994; 487-499.

[137] Burdick D et al. MAFIA: a maximal frequent itemset algorithm for transactional databases. Proceedings of the 17th International Conference on Data Engineering 2001: 443-452.

[138] Rao PK et al. Myogenic factors that regulate expression of muscle-specific microRNAs. Proc Natl Acad Sci U S A. 2006 Jun 6; **103**(23): 8721-6.

[139] Silber J et al. miR-124 and miR-137 inhibit proliferation of glioblastoma multiforme cells and induce differentiation of brain tumor stem cells. BMC Med. 2008; **6**: 14.

[140] Wang G et al. Variation in the miRNA-433 binding site of FGF20 confers risk for Parkinson disease by overexpression of alpha-synuclein. Am J Hum Genet. 2008; **82**(2): 283-9.

[141] Wang Y et al. Profiling microRNA expression in hepatocellular carcinoma reveals microRNA-224 up-regulation and apoptosis inhibitor-5 as a microRNA-224-specific target. J Biol Chem. 2008; **283**(19): 13205-15.

[142] Fields BN and Knipe DM. Virology. Retroviridae and their replication. Vol. 3rd edition 1996. J.M. Coffin, editor. Raven Press, New York. 1437-1500.

[143] Purcell DF and Martin MA. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. J. Virol. 1993; **67**: 6365-6378.

[144] Schwartz S et al. Cloning and functional analysis of multiply spliced mRNA species of human immunodeficiency virus type1. J. Virol.199; **64**: 2519-2529.

[145] Malim MH et al. The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viralmRNA. Nature 1989; **338**: 254-257.

[146]Zapp ML and Green MR. Sequence-specific RNA binding by the HIV-1 Rev protein. Nature 1989; **342**: 714-716.

[147] Pollard VW and Malim MH. The HIV-1 Rev protein. Annu Rev Microbiol. 1998; **52**: 491-532.

[148] Henderson BR and Percipalle P. Interactions between HIV Rev and nuclear import and export factors: the Rev nuclear localisation signal mediates specific binding to human importin-beta. J. Mol. Biol. 1997; **274**: 693–707.

[149] Heapy S et al. HIV-1 regulator of virion expression (Rev) protein binds to an RNA stem-loop structure located within the Rev response element region. Cell 1990; **60**: 685-693.

[150] Malim MH et al. The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. Nature 1989; **338**: 254-257.

[151] Nameki D et al. Mutations conferring resistance to human immunodeficiency virus type 1 fusion inhibitors are restricted by gp41 and Rev-responsive element functions. J. Virol. 2005; **79**: 764-770.

[152] Reeves JD et al. Enfuvirtide resistance mutations: impact on human immunodeficiency virus envelope functionentry inhibitor sensitivity, and virus neutralization. J. Virol. 2005; **79**: 4991–4999.

[153] Mathews DH et al. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure *J. Mol. Biol*1999; **288**: 911-940.

[154] Parisien M and Major F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature 2008; **452**(7183): 51-5.

[155] Svicher V et al. Specific Enfuvirtide-Associated Mutational Pathways in HIV-1 Gp41 Are Significantly Correlated With an Increase in CD4(+) Cell Count, Despite Virological Failure. J. Infect. Dis. 2008; **197**: 1408-1418.

[156] Edgcomb SP. Protein structure and oligomerization are important for the formation of export-competent HIV-1 Rev-RRE complexes. Protein Sci. 2008; **17**: 420-430.

[157] Ensembl – web site (2007).

[158] Sam Griffiths-Jones, Russel J. Grocock, Stijn van Dongen, Alex Bateman and Anton J. Enright – **miRBase: microRNA sequences, targets and gene nomenclature** – Nucleic Acid Research 34, D140-D144 (2006).

[159] PRAVEEN SETHUPATHY, BENOIT CORDA, and ARTEMIS G. HATZIGEORGIOU - **TarBase: A comprehensive database of experimentally supported animal microRNA targets** - RNAjournal 12, 192-197 (2005).

[160] The Genetic Association Database - Nature Genetics 36, 431 - 432 (2004).

[161] GeneOntology – About GO – GeneOntology web site (2007).