UNIVERSITA' DEGLI STUDI DI BOLOGNA

DIPARTIMENTO DI BIOCHIMICA

SETTORE DISCIPLINARE (MED/03)

TESI DI DOTTORATO IN BIOCHIMICA

(XIX CICLO)

IDENTIFICAZIONE DI NUOVI GENI ASSOCIATI AL FENOTIPO DI HIRSCHSPRUNG IN C.ELEGANS E LORO CONTROPARTE UMANA

Relatore:

Chiar.mo Prof. GIOVANNI ROMEO

Coordinatore:

Chiar.mo Prof. GIORGIO LENAZ

Correlatore:

Chiar.mo Ing. DIEGO DI BERNARDO

Candidato:

ALBERTO GOLDONI

ANNO ACCADEMICO 2005-2006

ai miei genitori…

Indice

1. INTRODUZIONE

La malattia di Hirschsprung (HSCR)

La Malattia di Hirschsprung (HSCR) (OMIM 142623), o megacolon congenito, è una malformazione caratterizzata dal mancato sviluppo delle cellule gangliari enteriche dei plessi sottomucosi (di Meissner) e mioenterici (Auerbach), con conseguente severa costipazione, distensione addominale e ostruzione intestinale durante il primo anno di vita.

La sua incidenza è di 1 su 5000 nati vivi (Amiel J, 2001) con rapporto di incidenza interfamiglia tra maschi e femmine di quattro a uno..

La malattia, generalmente sporadica nell'80-85% dei casi, è caratterizzata da penetranza incompleta, multifattorialità ed espressività variabile.

La malattia di HSCR è causata da anomalie della migrazione delle cellule delle creste neurali che derivano dalle regioni vagale, del tronco e sacrale del tubo neurale, ovvero delle cellule che contribuiscono alla formazione del sistema nervoso enterico.

HSCR si può manifestare in forma più o meno grave a seconda dell'estensione intestinale, vale a dire forma lunga o corta. La variabilità del fenotipo anatomopatologico e clinico è in relazione alla lunghezza del segmento agangliare, che nelle forme classiche di HSCR è confinato al retto-sigma e al colon discendente, nelle forme ultralunghe può invece estendersi a tutto il colon e parte dell'intestino tenue.

Da un punto di vista fisiopatologico il segmento agangliare è caratterizzato da un alterata motilità intestinale ed in particolare da una incapacità di rilasciamento e progressione del bolo fecale. Le alterazioni innervative del segmento agangliare rendono ragione del principio cardine della strategia chirurgica della HSCR, che è quello di resecare il segmento distale che presenta un sistema nervoso enterico senza cellule gangliari.

Il sistema nervoso enterico (SNE) ha origine dalla cresta neurale (CN) e le cellule della CN che contribuiscono alla formazione del SNE hanno origine dalle regioni vagali, del tronco e sacrali del tubo neurale (Manié S., 2001).

Le cellule che hanno origine dai derivati della regione vagale della cresta neurale (figura 1.1, in rosso), derivano dalla porzione dorsale del tubo neurale e sono localizzate fra i somiti uno e cinque, colonizzano l'intero intestino e danno origine al ganglio superiore cervicale (il ganglio posto più rostralmente nella catena simpatica, SGC) e alle cellule C della tiroide. Queste cellule della CN costituiscono la linea enterica-simpatica, co-esprimono RET e GDNFa-1 e dipendendono dall'interazione con il ligando GDNF durante la migrazione o al sito di gangliogenesi.

GDNF è espresso nel mesenchima della parete intestinale. Le cellule del tronco della cresta neurale che emergono dai somiti sei e sette (chiamata linea simpatico-adrenergica, in verde) popolano l'intestino prossimale e la catena posteriore simpatica della SCG e sono indipendenti da RET/GDNF. Le cellule della CN sacrale (in blu) colonizzano la parete intestinale più distale. Le frecce in figura 1.1. mostrano la direzione della migrazione.

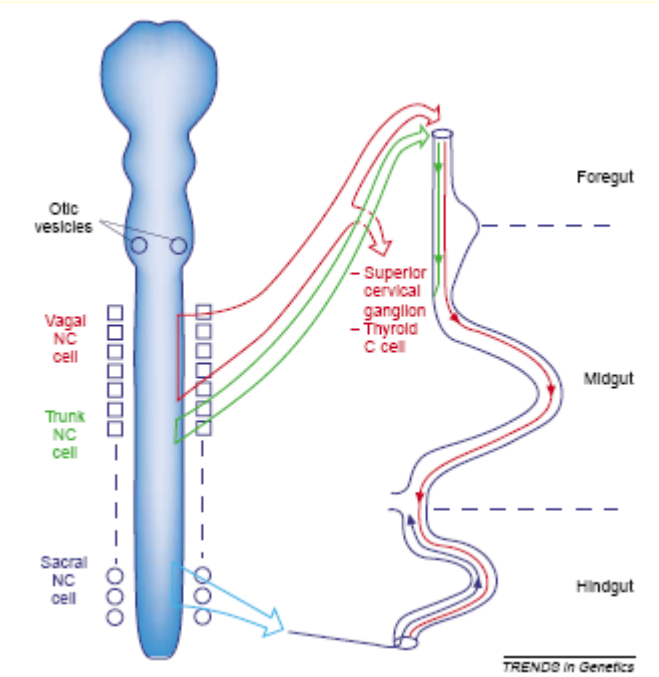


Figura 1.1: le differenti sottopopolazioni delle cellule della cresta neurale contribuiscono alla formazione del sistema nervosa

Geni coinvolti nella patogenesi della malattia di Hirschsprung

Mutazioni nel gene *RET* sono associate con circa la metà dei casi familiari e una piccola frazione di casi sporadici di malattia di HSCR (Mulligan LM, 1994; Bolk S., 1995; Amiel J., 1996).

Sono state individuate in pazienti con l'HSCR mutazioni nei geni che codificano per GFL (GDNF e NRTN), per componenti del complesso recettoriale dell'endotelina (EDN) di tipo B (EDNRB, EDN3, ECE-1), per fattori di trascrizione, come la proteina-1 che interagisce con SOX10 e SMAD (SIP1) e PHOX2B (Parisi MA 2000; Cacheux V 2001; Wakamatsu N 2001). Alcuni di questi geni sono responsabili di forme sindromiche di HSCR (figura 1.2).

Gene	Map position	Phenotype	Inheritance
RET	10q11.2	Non-syndromic Hirschsprung-MEN2A/FMTC	Dominant, incomplete penetrance
GDNF	5p13	Non-syndromic	Non-Mendelian
Neurturin	19p13	Non-syndromic	Non-Mendelian
EDNRB	13q22	Shah–Waardenburg Non-syndromic	Recessive Dominant (*de novo* in 80%)
EDN3	20q13	Shah-Waardenburg Non-syndromic	Recessive Dominant, incomplete penetrance
SOX10	22q13	Shah-Waardenburg	Dominant (*de novo* in 75%)
ECE-1	1p36	Congenital heart malformation	*De novo* dominant
ZFHX1B	2q22	Mowat-Wilson	*De novo* dominant
KIAA1279	10q21.3-q22.1	Goldberg-Shprintzen	Recessive
PMX2b	4p12	Congenital central hypoventilation syndrome (CCHS)	Dominant (*de novo* in 90%)
Unknown	3p21	Non-syndromic, S-HSCR	Non-Mendelian
Unknown	9q31	Non-Syndromic, L-HSCR	Dominant, incomplete penetrance
Unknown	19q12	Non-syndromic, S-HSCR	Non-Mendelian
Unknown	16q23	Shah-Waardenburg	Non-Mendelian

*Reviewed in Ref. [20].

Figura 1.2: geni coinvolti nella patogenesi dell'HSCR isolato e sindromico.

Mutazioni nei singoli loci non sono né necessarie né sufficienti a causare il fenotipo clinico. E' interessante notare a questo riguardo, che sono stati identificati numerosi polimorfismi di *RET* associati a HSCR (Borrego S, 1999; Fitze G, 1999; Borrego S, 2000; Gath R, 2001; Fitze G, 2002; Borrego S, 2003; Fitze G, 2003).

Inoltre recentemente sono stati descritti alcuni polimorfismi e aplotipi comprendenti l'introne 1 in associazione con la malattia suggerendo un ruolo di RET nell'eziopatogenesi di questa malattia, anche in assenza di mutazioni nella porzione codificante (Emison ES, 2005; Lantieri F, 2006).

Analisi di genetica molecolare hanno sottolineato che le interazioni tra mutazioni nei geni che codificano per RET e EDNRB hanno un ruolo centrale nella patogenesi di HSCR (Carrasquillo MM, 2002).

In accordo con quanto detto fino ad ora, topi transgenici *Ret* $^{+/-}$ non mostrano aganglionosi intestinale e topi omozigoti nulli per EdnrB presentano un'aganglionosi solo nella parte distale del colon. Quando topi *Ret*$^{+/-}$ vengono incrociati con topi che rechino diverse combinazioni dell'allele nullo EdnrB, il calo del dosaggio di EDNRB aumenta drammaticamente la lunghezza dell'aganglionosi (McCallion AS, 2003). Da questo si evince che è necessaria una collaborazione tra le vie di segnale di EDNRB e di RET sia per un normale sviluppo del sistema nervoso enterico che per l'insorgere della malattia di Hirschsprung. Un possibile meccanismo di "cross-talk" fra i patways di RET e delle endoteline (figura 1.3) consiste nell'attivazione del recettore B delle endoteline (EDNRB) da parte dell endotelina 3 (ET-3) che inattiva la protein chinasi A (PKA) e può danneggiare l'N-terminale Rac/Jun (Rac/JNK) attraverso la fosforilazione della serina 696 in RET (Fukuda T, Asai N, Enomoto A, Takahashi M, 2005).

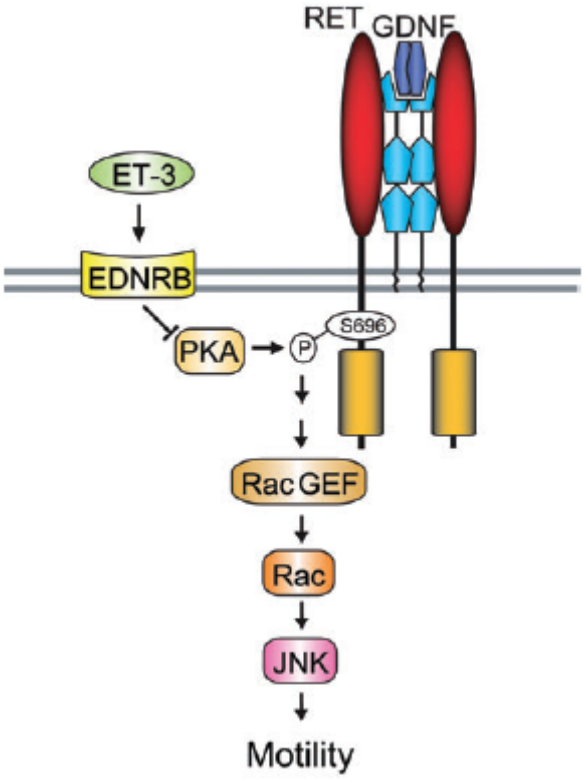


Figura 1.3: cross-talk fra RET ed il pathway delle endoteline nei neuroni enterici.

Mutazioni di *RET* sono distribuite lungo la sequenza codificante ed includono delezioni, inserzioni, frameshift, mutazioni non senso, e missenso (Romeo G, 1994; Eng C, 1997; Parisi MA, 2000; Griseri P, 2002). La maggior parte di queste mutazioni causa sia una diminuzione del dosaggio della proteina sia la perdita di funzione di RET (Parisi MA, 2000; Iwashita T, 2001), suggerendo che HSCR risulta da un'aploinsufficienza di RET. Le conseguenze funzionali di mutazioni HSCR sono correlate con la loro posizione nella sequenza codificante e sono state classificate in quattro gruppi (figura 1.4) (Pelet A, 1998; Iwashita T, 2001; Manié S, 2001).

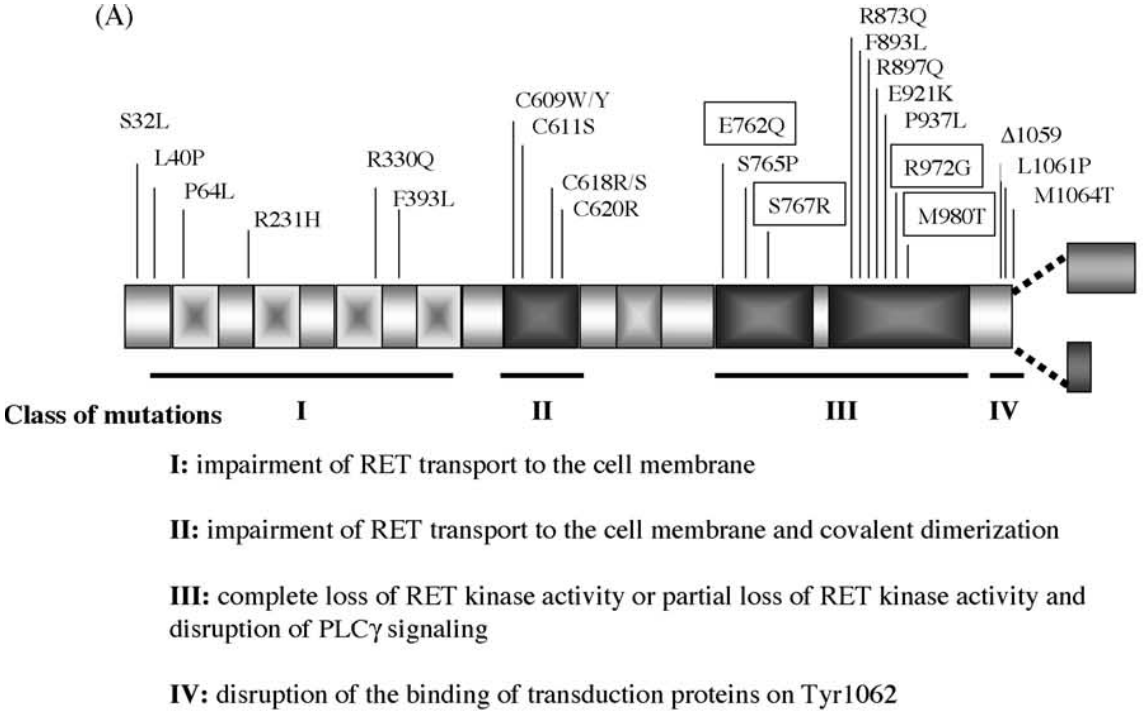


Figura 1.4: classificazione delle mutazioni coinvolte nell'HSCR a seconda delle conseguenze a livello molecolare sulla funzione di RET. Le mutazioni evidenziate nel riquadro portano ad una parziale perdita di funzione dell'attività chinasica di RET.

Le mutazioni di Classe I, localizzate nel dominio extracitoplasmatico, interrompono la maturazione di RET e inibiscono la sua traslocazione nella membrana plasmatica (Carlomagno F, 1996; Iwashita T, 1996).

Le mutazioni di Classe II causano la sostituzione di uno dei quattro residui di cisteina extracitoplasmatici (Cys609, 611, 618 e 620) con un altro amminoacido. Mutazioni nella linea germinale di residui di cisteina molto simili (figura 1.5) si ritrovano anche in MEN2A (Neoplasia Endocrina Multipla di tipo 2A) e FMTC (Carcinoma Midollare Familiare della Tiroide) e in alcuni individui MEN2A/FMTC che cosegregano con HSCR (Mulligan LM, 1994; Romeo G, 1998). In alcune rare famiglie in cui è presente una di queste mutazioni, il fenotipo HSCR cosegrega col fenotipo tumorale infatti queste mutazioni portano paradossalmente ad una contemporanea perdita di funzione (HSCR) e acquisizione di funzione (MEN2A/FMTC) (figura 1.5).

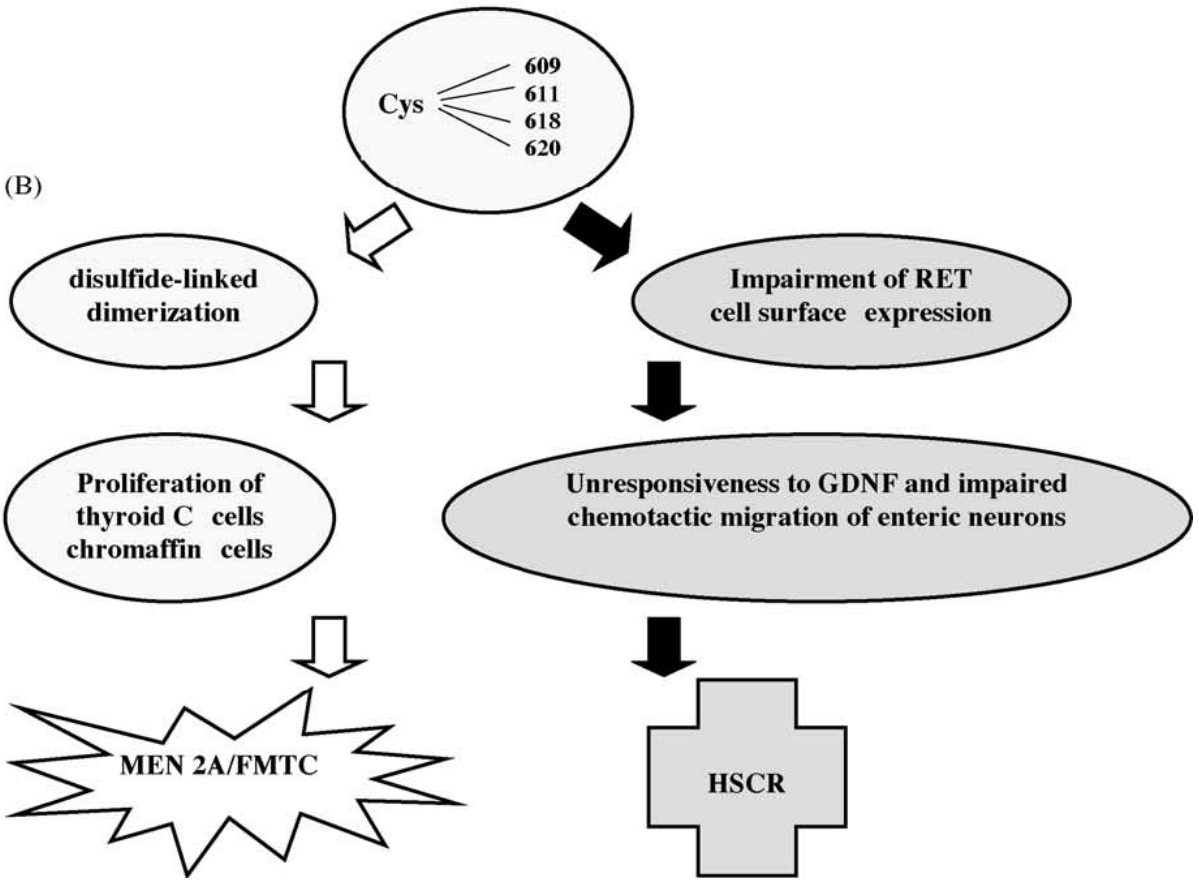


Figura 1.5: il paradosso MEN 2/HSCR in cui le mutazioni di Classe II sono associate sia con HSCR che con MEN 2A/FMTC.

Mutazioni di Classe III colpiscono il dominio tirosino chinasico e ostacolano l'attività catalitica (Takahashi M 1999; Iwashita T 2001).

Mutazioni di Classe IV non diminuiscono l'attività tirosino chinasica ma interferiscono specificatamente col legame dei componenti della segnalazione di RET come Shc, PLC?, FRS2 e IRS-1 (Lorenzo MJ, 1995; Carlomagno F, 1996; Pelet A, 1998; Iwashita T, 2001). E' stato proposto inoltre che alcune mutazioni missenso coinvolte nell'insorgenza dell'HSCR abbiano un effetto dominante negativo (figura 1.5) (Pasini B, 1995; Cosma MP, 1998).

RET: il gene e la proteina

Il proto-oncogene *RET* (REarranged during Transfection), identificato nel 1985 in un saggio di trasformazione di fibroblasti murini NIH3T3 con DNA di linfoma umano a cellule T, mappa sul braccio lungo del cromosoma 10 in posizione 11.2 (figura 1.6) (Ishizaka Y, 1989).

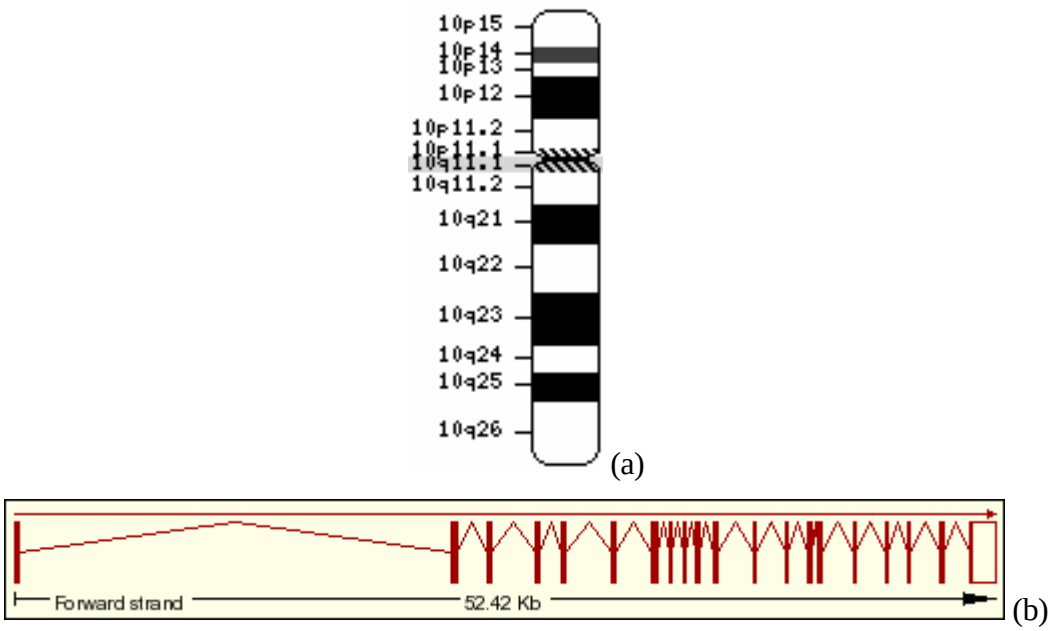


Figura 1.6: (a) posizione del gene RET all'interno del cromosoma 10
(b) struttura del trascritto del gene RET.

E' costituito da 21 esoni che codificano per un recettore di membrana a funzione tirosino chinasica (TK). Il gene RET è fondamentale per lo sviluppo di tessuti derivanti dalle creste neurali ed è coinvolto in diverse neurocristopatie, quali le Neoplasie Endocrine Multiple di tipo2A e 2B (MEN2A, MEN2B), il Carcinoma Midollare della Tiroide nella forma sporadica e familiare (MTC e FMTC) e la malattia di Hirschsprung.

I trascritti riportati in letteratura si distinguono per l'uso di differenti siti di poliadenilazione e per processamenti alternativi del trascritto primario (Xing S, 1994; Lorenzo MJ, 1995; Myers SM, 1995).

Le principali isoforme proteiche sono RET9 (1072 aa), RET51 (1114 aa) e RET43 (1106 aa), che differiscono tra loro nella porzione carbossi-terminale (figura 1.7). Bisogna però ricordare che le due isoforme RET51 e RET9 sono le più importanti e sono altamente conservate in svariate specie. Il fatto che queste due isoforme differiscano nella posizione carbossi-terminale intracitoplasmatica deputata alla trasduzione del segnale suggerisce che le due isoforme possono esercitare differenti funzioni fisiologiche di RET. Il turnover di RET51 e RET9 è mediato dal reclutamento differenziale della ligasi Cbl dell'ubiquitina (Scott RP, Eketjall S, Aineskog H, Ibanez CF, 2005) Si è notato, inoltre, che i topi deficitati dell'isoforma lunga (RET51) sono normali, mentre se mancano della isoforma corta (RET9) mostrano malformazioni renali e difetti dell'innervazione enterica, come per i topi che possiedono una forma di RET non funzionale.

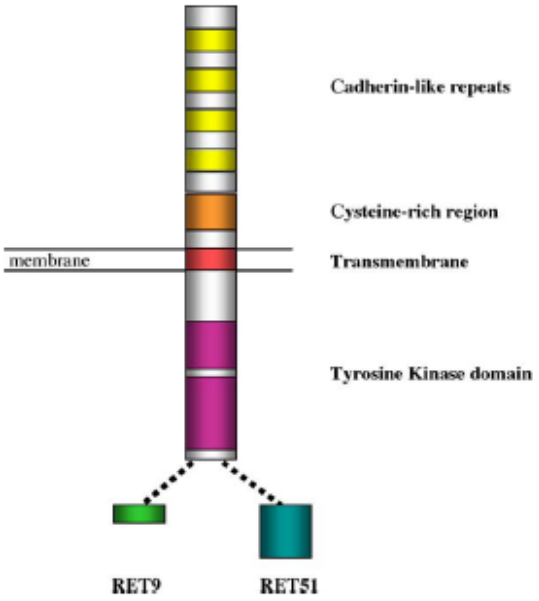


Figura 1.7: struttura schematica della proteina RET che mostra le quattro ripetizioni extracellulari di caderina, la regione ricca di cisterna e i domini di transmembrana e i due diversi domini tirosin-chinasici.

La regione extracellulare responsabile dell'ancoraggio del ligando, presenta caratteristiche condivise dalla famiglia delle caderine, proteine implicate nell'adesione omofilica Ca^{++}-dipendente. Di solito tra i dominii contenuti sia nella sequenza di *RET* che nel dominio extracellulare delle caderine, è presente un sito di legame per il Ca^{++}, il quale induce linearizzazione e rigidità della parte extracellulare, proteggendo la proteina dalla degradazione proteolitica e garantendo l'adesione cellulare. Per queste caratteristiche strutturali è stato ipotizzato che *RET* possa aver preso origine dalla ricombinazione di una *caderina* ancestrale con un gene tirosino chinasico (Anders J 2001). Tale regione contiene diversi inoltre siti di glicosilazione (Takahashi M, 1991).

Nella porzione extracellulare adiacente al dominio transmembrana, è presente una regione di circa 120 residui amminoacidici caratterizzata dalla presenza di 14 cisteine altamente conservate. Tale dominio sembra avere un ruolo critico nel formare corretti legami disolfuro intramolecolari, garantendo la corretta struttura terziaria della proteina (Ponder BA, 1999; Anders J, 2001).

Il dominio transmembrana è codificato da parte dell'esone 11, ed è seguito dalla porzione citoplasmatica con attività tirosino chinasica. Quest'ultima contiene due moduli ad attività TK, interrotti da 27 residui rappresentanti il dominio interchinasico e altamente conservati in altri recettori della stessa famiglia.

RET è l'elemento trasduttore di un complesso multimerico formato da una componente recettoriale rappresentata dai co-recettori di membrana GFRa 1-4 (GDNF family receptor-a) (figura 1.7) e da ligandi (GFL, GDNF family ligand) appartenenti alla famiglia del fattore neurotrofico

derivante dalla linea cellulare gliale (Glial cell line-Derived Neurotrophic Factor, GDNF) (Durbec P, 1996; Ibanez C F, 1998). Ad oggi sono stati identificati quattro ligandi naturali in grado di attivare la funzione TK di RET a seguito del legame con la componente corecettoriale del complesso: GDNF, Neurturina (NTN), Persefina (PSPN) e Artemina (ARTN) (Kotzbauer PT, 1996; Milbrandt J, 1998; Baloh RH, 1998).

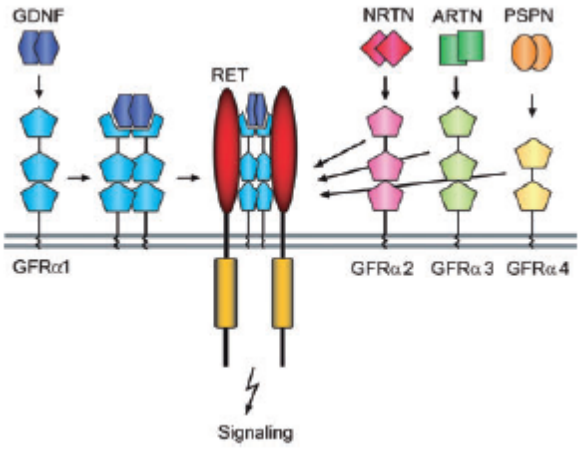


Figura 1.7: interazione del ligando GDNF con il recettore proteico GFRa e la TK di RET. GDNF, neurturina (NRTN), artemina (ARTN) and persefina (PSPN) attivano RET attraverso il legame con GFRa1, GFRa2, GFRa3 and GFRa4, rispettivamente.

I GFRa sono proteine estrinseche di membrana, legate alla superficie cellulare mediante una molecola di Glicosil Fosfatidil Inositolo (GPI), ne sono note 4 (1-4) ognuna specifica per un singolo ligando.

Una volta legato il ligando la componente corecettoriale GFR, ha lo scopo di reclutare due monomeri di RET determinandone la dimerizzazione e la successiva autofosforilazione su specifici residui tirosinici a livello del domini TK.

Sono conosciuti due modelli di attivazione per RET: una in *cis* ed una in *trans*. Nelle cellule che co-esprimono RET e GFRα1, nella sua forma inattiva RET è fuori dall'isola lipidica di membrana, e solo dopo la stimolazione di GDNF, viene reclutato da GFRα (figura 1.8A).

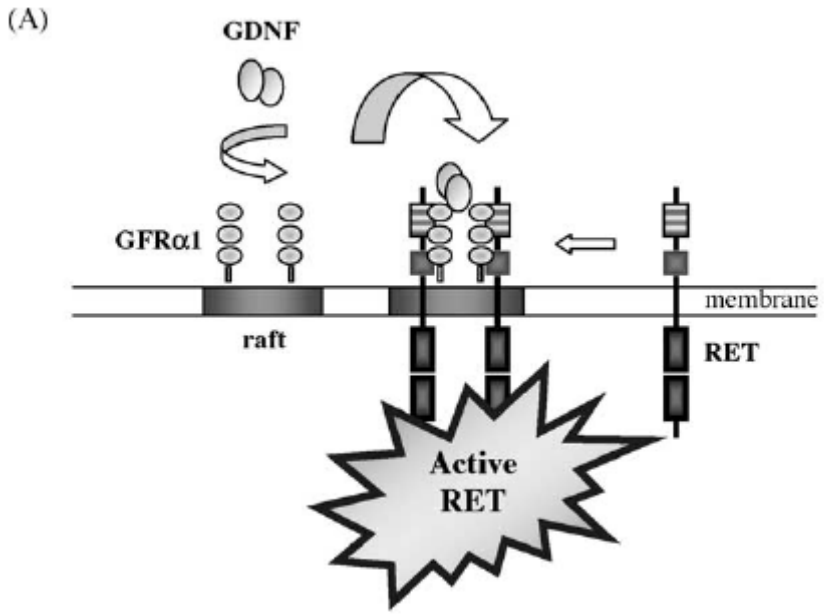


Figura 1.8A: modello di attivazione in *cis* di RET

I GFL inducono la dimerizzazione di GFRα, l'alta affinità del complesso GFL-GFRα per RET lo trascina nella piattaforma lipidica e ne promuove la dimerizzazione, indipendentemente dalla sua attività tirosino chinasica.

Un modello alternativo suggerisce un'azione dei GFRα nella loro forma solubile (Paratcha G. 2001). GFRα può catturare dei ligandi sciolti dallo spazio della matrice extracellulare e quindi presentare tali fattori *in trans* alle cellule che esprimono RET. La forma solubile, biologicamente attiva di GFRα, ottenuta dal clivaggio della porzione GPI, è stata ritrovata in un terreno condizionato di colture neuronale e gliali. La stimolazione di RET *in trans*, particolarmente dipendente dalla sua attività catalitica (Paratcha G, 2001), suggerisce la partecipazione di eventi intracellulari e porta ad una attivazione più duratura del complesso di segnalazione a valle e potenzia gli effetti biologici di GDNF (figura 1.8B) (Manié S, 2001).

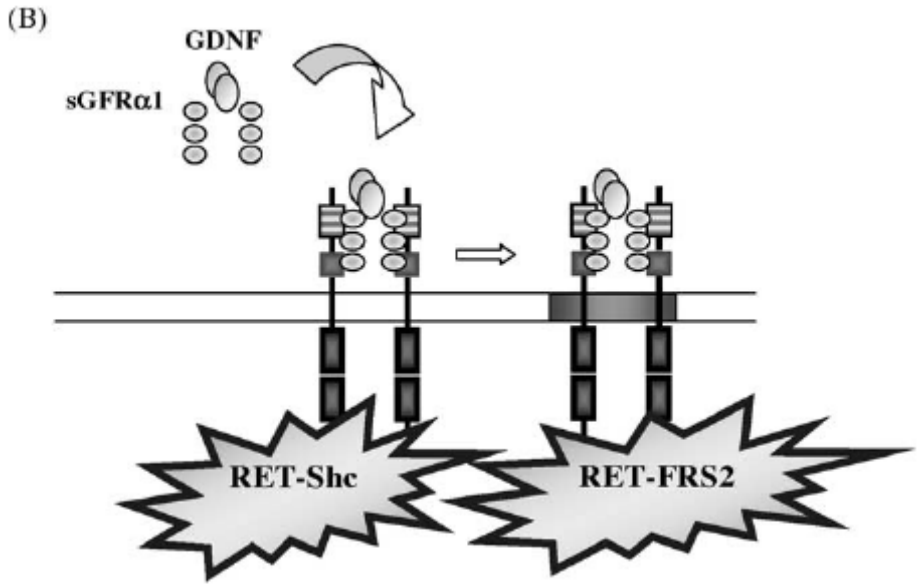


Figura 1.8B: modello di attivazione in *trans* di RET

Il legame con GFL permette la formazione di ponti disolfuro e fa sì che fra le due parti intracellulari dell'omodimero di RET si generi una stretta interazione, in modo tale da avere un'autofosforilazione delle tirosine. Conseguenza di tale fenomeno è l'attivazione di RET che determina la trasduzione del segnale attraverso effettori che riconoscono ed interagiscono con la forma fosforilata del RTK. Sebbene le vie di trasmissione del segnale possano essere condivise da diversi recettori, l'interazione ligando-recettore è molto specifica (Manié S, 2001).

RET contiene 18 residui tirosinici e, a seguito della fosforilazione, quattro di questi (Tyr905, Tyr1015, Tyr1062 e Tyr 1096) diventano siti d'ancoraggio per proteine che contengono specifici moduli di interazione quali SH2 (Src homology 2) e PTB (phosphotyrosine-binding).

Proteine che contengono SH_2, GRB7 (growth factor receptor bound protein), GRB10, PLC-γ (phospholipase C) e GRB2 legano le tirosine 905, 1015 e 1096 fosforilate (figura 1.9). Effettori trasduzionali, quali SHC, FRS2 (fibroblast growth factor receptor substrate 2) e IRS1 (Insulin receptor substrate 1**)** riconoscono la Tyr1062 fosforilata attraverso il dominio PTB, ed inoltre SHC può anche interagire con la Tyr1062 di RET9 attraverso il suo modulo SH2.

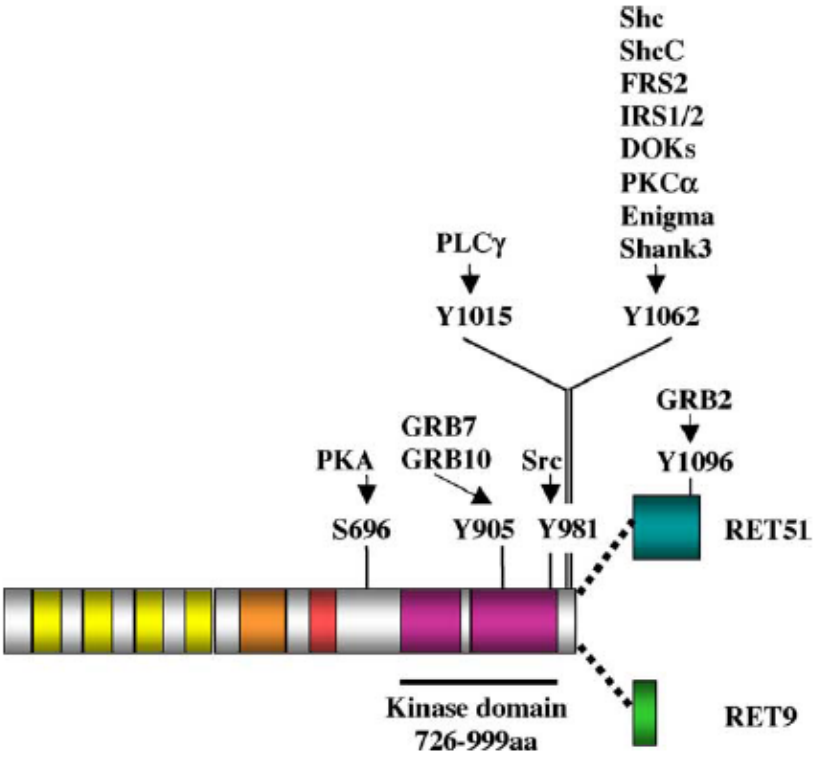


Figura 1.9: trasduttori del segnale a livello intracellulare della forma di RET attivata. Il legame PKCa a Tyr1062 è indiretto. Enigma e Shank3 si legano specificamente alla Tyr1062 di RET9. Le sequenze di RET9 e RET51 sono identiche fino alla Gly1063.

IRS attiva la PI3K (Phosphatidylinositol-3-Kinase), che a sua volta stimola l'azione di RAS (rat sarcoma) e il successivo coinvolgimento nella migrazione cellulare.
Inoltre, la proteina ENIGMA che mostra un dominio PD2 amino-terminale e tre domini LIM al carbossi-terminale (domini zinc-finger ricchi in cisteine), lega la Tyr1062 in modo indipendente dal suo stato di fosforilazione.

GDNF stimola l'attivazione del pathway della PI3K (phosphatidylinositol3-kinase) e della cascata delle MAP chinasi (mitogen-activated protein kinase), l'ultima promuove le ERKs (extracellular-regulated kinase), JNKs (c-jun amino-terminal protein kinase), la MAP chinasi p38 e la BMK1 (big MAP kinase) ERK5 (Jhiang 2000; Hansford 2000; van Weering, and Bos 1998b; Hayashi 2000). La stimolazione del pathway di PI3K porta all'attivazione a più livelli della proteina serina/treonina chinasi AKT/PKB, un effettore centrale di PI3K.

La trasmissione di un segnale PI3K-dipendente, risulta nella fosforilazione contemporanea delle proteine associate alle adesioni focali - pFAK (focal adhesion kinase), p130 CAS, e Paxillina - e provoca anche l'attivazione di NFkB, con un effetto sulla proliferazione e sopravvivenza cellulare.

Di recente è stata messa in evidenza la relazione di RET con l'apoptosi. È stato suggerito, infatti, che RET appartenga ad una nuova famiglia di recettori funzionalmente correlati: i recettori a dipendenza. Questi recettori condividono la capacità di trasdurre due differenti segnali intracellulari: in presenza di ligando tasducono un segnale positivo di sopravvivenza, differenziamento o migrazione; in assenza di ligando invece danno inizio o amplificano un segnale di apoptosi.

La mancata dimerizzazione del recettore RET determina l'attivazione di un segnale di morte cellulare, che è probabilmente dovuta all'azione delle caspasi. Queste proteine coinvolte nell'apoptosi, sembrano attivarsi degradando RET inattivo ed è stato osservato che i frammenti di RET così prodotti sono in grado di attivare altre caspasi, generando una cascata d'attivazione di queste, che determina la morte cellulare (Bordeaux MC, 2000).

Scopo del progetto

Il gruppo del Professor Giovanni Romeo da anni si occupa di identificare nuovi geni che potrebbero essere coinvolti nella patogenesi della malattia di Hirschsprung. Lo scopo del lavoro di questa tesi rappresenta la necessità di identificare nuovi geni che permettano di far luce sui pathways coinvolti nella patologia. Per raggiungere questo scopo, sono state utilizzate due diverse strategie: strategia knock-out di *C.elegans* (figura 1.10) ed utilizzo dello Split Ubiquitin Membrane Yeast Two Hybrid.

Figura 1.10: immagine del nematode *C.elegans*

Oltre ai modelli murini, ampiamente utilizzati, la scelta è caduta proprio su *C.elegans* perché rappresenta un buon modello per lo studio di malattie genetiche umane, come descritto in alcuni esempi (Bessou C, 1998; Wittenburg N, 2000; Satyal S H, 2000; Baumeister R, 2002).

Dal momento che non è stato identificato l'omologo del gene RET umano in *C.elegans*, come alternativa si sono studiate le funzioni del gene omologo ad ECE-1 nel nematode.

ECE-1 (endothelin converting enzyme 1) è un gene coinvolto nella migrazione delle cellule derivate dalla cresta neurale. Esistono altri geni coinvolti in questo processo ed una parte di essi appartiene ad un pathway comune attivato dall'interazione di differenti ligandi coi propri recettori.

L'importanza del gene ECE-1 nella via di segnale mediata da RET e EDNRB è evidenziata dalla presenza di mutazioni in ECE-1 in alcuni pazienti affetti da HSCR (Hofstra RM, 1999). Il gene omologo al gene RET umano non e stato ancora identificato a livello di sequenza codificante in *C.elegans*, tuttavia la via di segnale è conservata. E quindi interessante analizzare la funzione di un omologo di ECE-1 in *C.elegans*.

In particolare la via di segnalazione mediata dall'Endotelina 3/EDNRB sembra ben conservata in *C.elegans* indicando l'importanza di questo pathway e l'identificazione dei meccanismi che regolano l'espressione di RET e la sua interazione con molecole effettrici diviene fondamentale per la comprensione dei meccanismi che guidano lo sviluppo del sistema nervoso enterico.

Studi recenti svolti sia nell'uomo che nel topo hanno dimostrato una possibile interazione genetica a livello tissutale tra le vie di RET ed EDNRB che porterebbe allo sviluppo di HSCR (Carrasquillo MM, 2002; McCallion AS, 2003; Barlow A, 2003). E' stato quindi proposto che, analogamente a quanto descritto per altri recettori, le vie di RET ed EDNRB potrebbero interagire attraverso un effettore attivato da EDNRB che agisca sulla proteina RET (Carrasquillo MM, 2002).

Quindi si può capire come l'identificazione dei geni appartenenti al pathway delle endoteline possa rappresentare uno step fondamentale per poter meglio capire il meccanismo patogenetico alla base dell'HSCR.

Lo studio del ruolo funzionale dei geni coinvolti nel differenziamento e nella crescita richiedono un appropriato modello sperimentale e *C.elegans* rappresenta un modello facile da manipolare e meno costoso in termini di spese, spazio, e crescita rispetto ai modelli murini o di Drosophila. Inoltre, questo organismo è stato completamente caratterizzato a livello genomico (finito di sequenziare nel 1998).

Esiste una mappa del destino differenziativo di ogni cellula ed uno schema della connessione dei suoi 302 neuroni. Il gene ECE-1 codifica per un enzima proteolitico (figura 1.11) ed è coinvolto nella via di attivazione mediata da Endotelina-3 (funzione di conversione dalla forma

inattiva delle endoteline in quella attiva) e dal suo recettore EDNRB, che contribuisce alla regolazione della migrazione delle cellule derivate dalla cresta neurale.

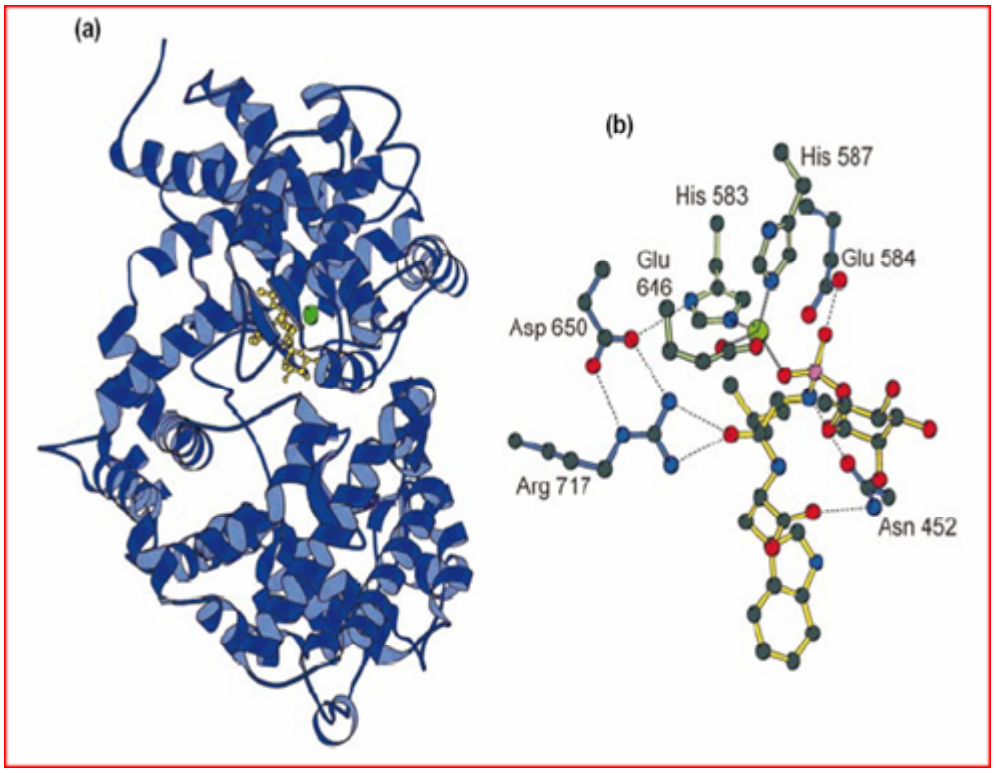


Figura 1.11: schema tridimensionale della proteina codificata dal gene ECE1.

Alcuni pazienti con HSCR sono portatori di mutazioni di ECE-1 (Hofstra RM, 1999), e, oltre a queste, sono state identificate anche altre mutazioni in due geni correlati ad ECE-1 (EDNRB e EDN3) in pazienti HSCR (Tanaka H, 1998; Svensson P J, 1999). Il Laboratorio di Neurogenetica Molecolare situato a Friburgo, diretto dal Prof. Baumeister, ha sviluppato una libreria di knock-out che virtualmente contiene mutanti per ognuno dei 19253 geni di *C.elegans.* All'interno di questa libreria è stato trovato un ceppo knock-out per il gene omologo a ECE-1 (gene denominato ZK20.6) (figura 1.12).

Figura 1.12: omologia di sequenza fra ECE1 ed il gene omologo in *C.elegans*.

Il fatto che ECE-1 mutato sia stato trovato in pazienti affetti e che il topo knock-out per ECE-1 sia stato definitivamente associato al fenotipo HSCR giustificano la scelta del suddetto gene.

In particolare è stato confrontato il trascrittoma (totalità dei geni espressi in un organismo) dei seguenti ceppi di *C.elegans*:

- wild-type
- animali knock-out per l'omologo di ECE-1

Questo esperimento è stato svolto in due stadi evolutivi, quello adulto e quello larvale L3. Studi di espressione estesi su tutto il genoma, come quelli che sono resi possibili dai microarrays, possono essere utilizzati per identificare sia i geni downregolati (che potrebbero essere elementi downstream del pathway), sia i geni upregolati (che potrebbero rappresentare un tentativo di compensare l'assenza del sistema delle endoteline). Entrambi i tipi di geni vengono considerati come geni candidati coinvolti nella sopravvivenza e nella migrazione delle cellule che derivano dalla cresta neurale. È stato utilizzato il GeneChip *C.elegans* Genome Array (Affymetrix) che rappresenta i 22.500 trascritti unici di *C.elegans*. Successivamente è stato estratto l'RNA totale utilizzando il kit Qiagen RNeasy partendo con una quantità di 10 µg di RNA totale per ogni ceppo. Avvalendoci della libreria di knock-out di *C.elegans* disponibile presso il Laboratorio di Neurogenetica di Friburgo, siamo stati in grado di recuperare i ceppi "null mutant" per ogni gene di interesse ed abbiamo verificato la presenza di un chiaro fenotipo neurologico.

Un risultato positivo ha indicato la presenza di una omologia funzionale tra i geni di *C.elegans* e quelli umani. I primi due stadi del progetto, e cioè gli studi di espressione, dovrebbero

aiutarci a costituire una mappa del "signalling network" che include i geni che appartengono al sistema delle endoteline e del protooncogene RET.

Per identificare le nuove proteine che interagiscono con RET è stata usata la tecnica "split-ubiquitin membrane two hybrid" (MbYTH) che rappresenta un perfezionamento del sistema "yeast two hybrid" che permette di analizzare lo stato di interazione fra due proteine di membrana.

L'abilità del sistema MbYTH aumenta la potenza nello studio dell'interazione delle proteine superando la limitazione dello "yeast two hybrid" nello studiare legami di membrana fra proteina-proteina.

RET è presente in due isoforme principali: quella lunga (RET51) e quella corta (RET9) che differisco per la porzione intra-citoplasmatica. Queste due isoforme sono entrambe usate nello "split-ubiquitin membrane two hybrid" come esche per testare possibili interazioni con la libreria di espressione umana del cervello. In questo modo, diventa possibile distinguere fra i complessi proteici che interagiscono direttamente con la isoforma lunga (RET51) e separatamente, con l'isoforma corta (RET9).

In via preliminare, i risultati hanno mostrato che dieci geni prodotti, alcuni dei quali non sono stati completamente caratterizzati, interagiscono direttamente con RET51. I risultati sono stati successivamente validati attraverso esperimenti di espressione transiente e di co-immunoprecipitazione con RET. Dal momento che queste interazioni sono validate, il ruolo delle proteine che interagiscono con RET sarà valutato nel contesto dei networks molecolari che regolano la patogenesi dell'HSCR.

Lo scopo della mia ricerca consiste nel verificare se ed in che modo questi geni interagiscono tra di loro utilizzando la tecnica del reverse engineering. Ciò implica l'utilizzo di programmi specifici che permettono di sondare i collegamenti più probabili tra i geni di interesse, partendo dai dati di microarrays, al fine di poter ricostruire il network di geni che maggiormente rappresenta il modello reale.

2. SYSTEM BIOLOGY

La System Biology, in generale, tratta lo studio di processi che avvengono nell'organismo umano (fenomeni e sistemi biologici), attraverso un approccio meccanicistico, che permette di analizzare i sistemi in esame dal punto di vista della loro complessità, in particolare trattando a differenti livelli di profondità tali sistemi, ad esempio studiando i sottoelementi che li compongono, con lo scopo finale di comprendere correlazioni e relazioni di causalità tra fenomeni e proprietà relativi a tali processi biologici.

In particolar modo la System Biology è orientata allo studio dei processi che avvengono nell'organismo attraverso un approccio tipico della teoria dei sistemi, in modo da descrivere il comportamento dei geni in determinate condizioni ed in risposta a determinati stimoli. La *biologia* ha fornito finora una grossa quantità di essenziali informazioni sulla struttura e sul ruolo, nell'organismo, delle diverse cellule, degli elementi che le compongono e di varie molecole che si possono trovare nell'organismo. In particolare, di grande importanza, sono stati gli studi sul codice genetico, su come esso venga codificato, e su come l'informazione che viene conservata nei geni, che concretamente sono costituiti da acido desossiribonucleico presente fisicamente nel nucleo della cellula, venga decodificata e come tale informazione venga utilizzata ai fini della sintesi delle proteine.

Pur essendo, al giorno d'oggi, chiaro quale siano i ruoli dei singoli componenti della cellula all'interno della stessa e dell'intero organismo, risultano essere poco chiari quali siano i fenomeni che avvengono in essa che coinvolgono catene di corpuscoli e di eventi, e che sostanzialmente modificano la sua struttura, il suo comportamento, e la sua funzione, a seguito di modificazioni ambientali e di segnali (ad esempio *ingressi* costituiti da ormoni che vengono a contatto con recettori situati in corrispondenza della membrana citoplasmatica) che dall'esterno della cellula stessa innescano tale concatenazioni di eventi.

Ciò che ci si propone di ottenere con la System Biology sono le suddette catene o reti di eventi che comprendono i *signaling pathways* o *signaling networks*, i *transcriptional pathways* e i *metabolic pathways* e i modelli *a scatola grigia* che da considerazioni biologiche e biochimiche se ne possono evincere. In particolare lo studio di tali catene di eventi cerca di estrarre da esse delle proprietà generali dei modelli che le descrivono e cosa più importante cerca di identificare i possibili collegamenti dei geni conosciuti nonché la presenza di nuovi geni che possono eventualmente essere coinvolti nel signaling pathways della malattia studiata (T.S. Gardner, J.J. Faith, 2005).

Il “reverse-engineering” si fonda su di un modello nel quale il trascritto si comporta come un segnale regolatorio che controlla la percentuale di sintesi di altri RNAs. Questo tipo di modello è spesso chiamato *“gene regulatory network”* o *“gene network”* e l’algoritmo di reverse-engineering non usa e non modellizza dati di proteine e metaboliti. Siccome la concentrazione di molti fattori di trascrizione dell’RNAs non correlano con l’attività dei loro prodotti proteici, un modello di gene network non descrive di solito la relazione fisica fra regolatori e trascritti. Tuttavia, i modelli di gene network danno una visuale globale della regolazione genica che non è ristretta all’interazione TF/promotore. Possono catturare implicitamente i fattori proteici e metabolici che possono influenzare l’espressione genica. Questi modelli sono inoltre importanti nel predire la risposta trascrizionale a nuovi trattamenti cellulari.

In generale, i modelli di gene network si possono rappresentare come un grafo (figura 2.1)

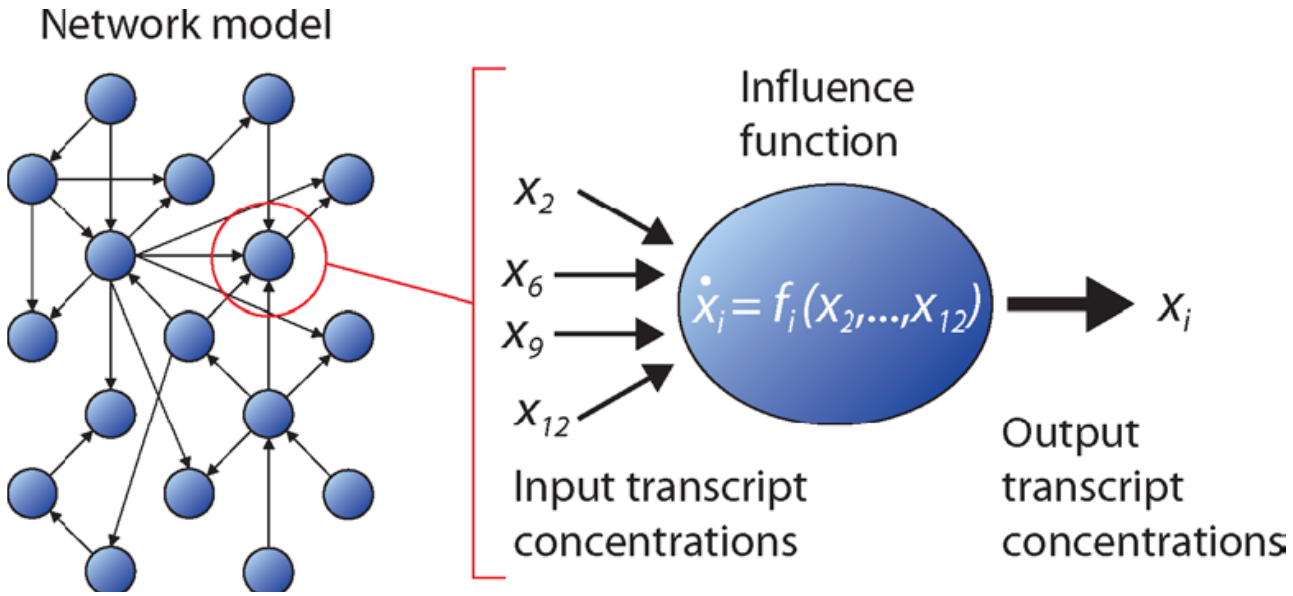


Figura 2.1: schema di un grafo che schematizza un modello di gene network basandosi sulle relazioni della funzione di influenza.

A seconda dell’approccio di reverse-engineering usato, si possono descrivere questi grafi matematicamente in modo differente.

Esempio di un semplice procedimento per modellizzare la trascrizione dell’RNA

Un semplice modello di trascrizione può essere rappresentato considerando la sintesi (trascrizione) dell’RNA, la quale è controllata dall’attività dell’RNA polimerasi (RNAP) che

rappresenta il complesso proteico che legge il DNA e lo trasforma in RNA. La trascrizione del DNA inizia quando l'RNAP riconosce e si lega ad un promotore (figura 2.2:).

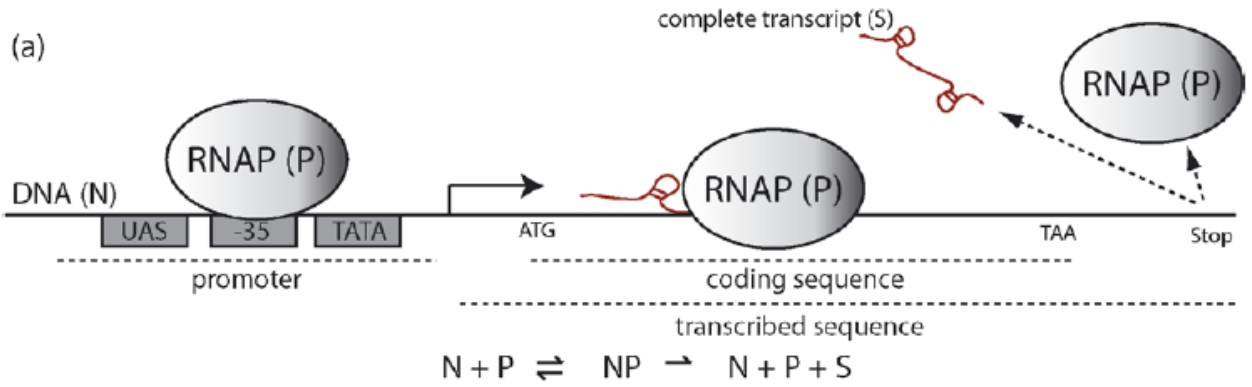


Figura 2.2: schema di trascrizione dell'RNA mediato dall'RNA polimerasi. I due strands di DNA sono separati e l'RNAP si muove lungo il DNA, trascrivendo una copia dell'RNA finché incontra uno stop codon sul DNA.

Dopo essersi legato, RNAP apre la doppia elica del DNA e scivola lungo la sequenza nucleotidica allungando il filamento dell'RNA mediante l'aggiunta di ribonucleotidi. L'allungamento del trascritto procede finché l'RNAP incontra una sequenza di stop nel DNA. I fattori che si legano direttamente al complesso RNAP possono modulare la percentuale di binding, la sua specificità, l'efficienza nel processo di allungamento dell'RNA e la fine dell'allungamento stesso.

Per poter quindi rappresentare il tutto come una formula che rappresenti la cinetica di reazione possiamo indicare:

- RNA polimerasi (RNAP) (P),
- promotore del DNA (N),
- trascritto dell'RNA (S)

come specie chimiche in un ambiente ben amalgamato.

Applicando quindi i principi della modellizzazione delle cinetiche molecolari, otteniamo:

$$P + N \underset{k_{-1}}{\overset{k_{+1}}{\rightleftarrows}} PN \xrightarrow{k_{+2}} P + N + S, \tag{1}$$

Dove k_{+1} e k_{-1} descrivono le percentuali positive e negative dell'affinità di legame dell'RNAP e k_{+2} riflette la percentuale di allungamento del trascritto. Sotto l'assunzione che k_{+2} sia piccolo o assumendo che la concentrazione del promotore (N) sia molto più piccola della concentrazione dell'RNAP (P), si può scrivere l'equazione differenziale per il trascritto come:

$$\frac{ds}{dt} = \frac{V_m p}{p + K_m} - \delta s, \qquad (2)$$

dove le lettere minuscole denotano la concentrazione delle rispettive specie, $V_m = k_{+2}n$ è la percentuale massima di sintesi, $K_m = (k_{-1} + k_{+2})/k_{+1}$ è la costante di Michaelis e d rappresenta una percentuale costante di degradazione del trascritto dell'RNA. Più formalmente, l'assunzione appena formulata indica la presenza di scale di tempi che possono essere lenti e veloci nella reazione e l'eq. (2) descrive le dinamiche lente. K_m descrive la soglia di attivazione per *p*. Per concentrazioni di *p* sopra K_m, la percentuale di sintesi inizia a saturare al suo massimo valore, V_m.

Dato il piccolo numero di molecole reagenti, l'omissione di svariate altre specie coinvolte nella trascrizione e la natura non omogenea del contenuto cellulare porta l'eq (2) a rappresentare un modello imperfetto. Comunque risulta efficace nel carpire le caratteristiche qualitative e quantitative dell'espressione del gene.

In questo modello è anche possibile introdurre il controllo della regolazione del trascritto. Per esempio, RNAP non riconosce alcuni promotori senza l'aiuto di un TF attivato che prima si leghi al promotore (figura 2.3).

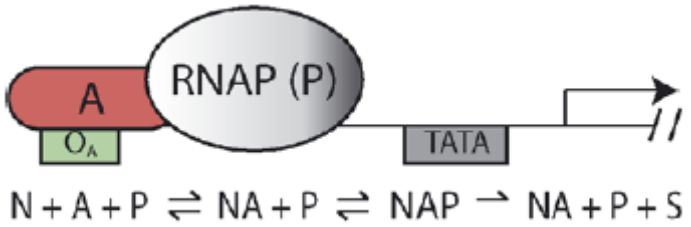


Figura 2.3: alcuni geni hanno una proteina che si lega ad un motivo (OA) vicino al promotore. Questa proteina ha la funzione di aumentare l'affinità di RNAP al promotore.

Questo schema della reazione porta alla seguente forma modificata dell'eq. (2):

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \left(\frac{a^{\alpha}}{a^{\alpha} + K_a}\right)\frac{V_{\mathrm{m}}p}{p + K_{\mathrm{m}}} - \delta s, \quad (3)$$

dove *a* rappresenta la concentrazione di *A*, l'attivatore e K_a è la soglia di attivazione per *A*. L'attivatore modula semplicemente la percentuale massima del trascritto. Il parametro *a* nell'equazione è un esponente di co-operatività ed è determinato dal numero di copie del sito di binding per *A* nel promotore. Se M_a è il numero di siti di binding che devono essere occupati da *A* per il reclutamento RNAP, allora $a = M_a$.

Il trascritto dai promotori potrebbe essere inoltre inibito dal repressore TFs. Sono possibili svariate forme di inibizione. Una delle più comuni è quella competitiva, che indica che il legame del repressore TF e dell'RNAP sono mutuamente esclusive (figura 2.4).

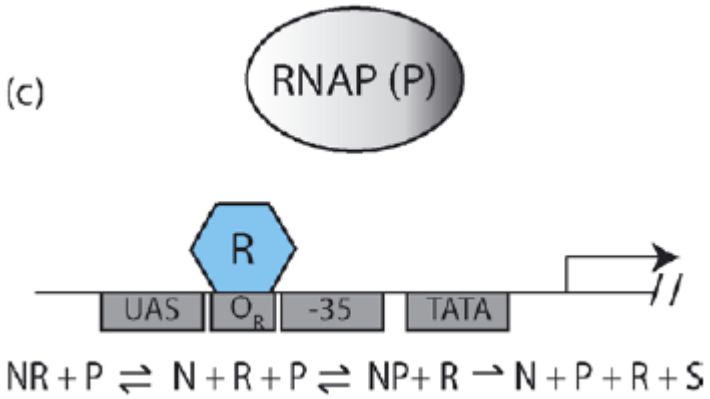


Figura 2.4: altri geni hanno delle proteine con funzione di repressori che si legano ai motivi (OR) nei promotori. Le proteine con funzione di repressori possono agire bloccando l'accesso di RNAP ad importanti regioni del promotore.

Per una inibizione competitiva l'eq (2) deve essere ulteriormente modificata nel seguente modo:

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \frac{V_{\mathrm{m}}p}{p + K_{\mathrm{m}}(1 + r^{\beta}/K_r)} - \delta s, \quad (4)$$

dove *r* indica la concentrazione del repressore, *R* e K_r è la soglia di inibizione per *R*. Il repressore può essere visto come la modulazione della soglia della costante di attivazione, K_m, del promotore da parte di RNAP. Il parametro *ß*, è similare ad *a* ed è determinato da M_b che rappresenta il numero di copie del sito di binding per *R* nel promotore.

Bayesian network models

I networks Bayesiani (J. Pearl, 1988) sono in grado di rappresentare la struttura con le relative dipendenze fra quantità multiple interagenti (es: livelli di espressione di geni differenti). Questi modelli permettono di tenere conto del rumore e della presenza di dati limitati che possono presentarsi negli studi di espressione ed inoltre conservano la logica combinatoriale della regolazione trascrizionale. Sono quindi uno strumento molto promettente per descrivere processi costituiti da componenti locali che interagiscono dove il valore di ogni componente dipende direttamente dai valori del piccolo numero di componenti. Questo tipo di approccio è stato ampiamente applicato in svariati campi con successo. Tutto ciò porta i modelli Bayesiani a rappresentare matematicamente le connessioni di un network di geni come un modello di influenza causale diretta (D. Heckerman, C. Meek, and G. Cooper. 1997, J. Pearl and T. S. Verma. 1991, P. Spirtes, C. Glymour, and R. Scheines, 1993).

Un network Bayesiano è rappresentato dalla distribuzione della probabilità congiunta. Questa rappresentazione consiste di due componenti, la prima G, rappresenta un grafo aciclico diretto i cui vertici corrispondono alle variabili random $X_1,\ldots..,X_n$. La seconda componente descrive una distribuzione di probabilità condizionale per ogni variabile data dal precursore di G. Insieme queste due componenti rappresentano una distribuzione unica su $X_1,\ldots..,X_n$.

Il grafo G rappresenta l'assunzione di indipendenza condizionale che permette alla distribuzione congiunta di essere decomposta risparmiando nel numero dei parametri e di conseguenza nella complessità di calcolo.

Applicando la regola a catena delle probabilità e la proprietà dell'indipendenza condizionale, ogni distribuzione congiunta che soddisfa l'assunzione che ogni variabile X_i è solo influenzata dal valore dello stato che lo precede direttamente (assunzione di Markov) in modo da ridurre la complessità nel pianificare una sequenza di azioni in quanto permette al diagramma di rimuovere gli stati dopo un certo periodo di tempo assumendo che lo stato dopo un certo lasso di tempo non abbia più nessun effetto su quello corrente e per questo motivo la formula può essere decomposta nel seguente modo:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \mathbf{Pa}(X_i)), \tag{5}$$

dove $Pa(X_i)$ rappresenta il set di precursori di X_i in G, mentre la seconda componente descrive la distribuzione $P(x_i|Pa(X_i))$ per ogni possibile valore x_i di X_i.

Nel caso di variabili con valori finiti si possono rappresentare queste distribuzioni condizionali come tabelle.

Generalmente i networks Bayesiani sono flessibili e possono essere adattati a molte forme di distribuzione condizionale includendo svariati modelli continui.

Dato un network Bayesiano, si può dare una risposta a molte domande che coinvolgono la probabilità congiunta (es: quale è la probabilità che X = x se si conoscono altre variabili) o l'indipendenza nel dominio (es: se X e Y sono indipendenti una volta che conosciamo Z) (F. V. Jensen, 1996, J. Pearl, 1988).

Per permettere che i network Bayesiani possano lavorare sia con variabili discrete che con variabili continue bisogna modificare l'eq (5) nel seguente modo fattorizzando nella parte discreta e in quella mista:

$$p(x) = p(i, y) = \prod_{\delta \in \Delta} p\left(i_\delta | i_{\mathrm{pa}(\delta)}\right) \prod_{\gamma \in \Gamma} p\left(y_\gamma | i_{\mathrm{pa}(\gamma)}, y_{\mathrm{pa}(\gamma)}\right). \tag{6}$$

dove ? e G sono un set di nodi discreti e continui rispettivamente, da cui il set di variabili X può essere indicato come $X = (X_v)_{veV} = (I,Y) = ((I_d)_{de?}, (Y_?)_{?eG})$, dove I e Y sono i sets di variabili discrete e continue rispettivamente.

Definizione della probabilità di distribuzione

La probabilità congiunta delle variabili random può essere formulata per quanto riguarda i nodi discreti nel seguente modo:

$$\theta_{i_\delta | i_{\mathrm{pa}(\delta)}} = p\left(i_\delta | i_{\mathrm{pa}(\delta)}, \theta_{\delta | i_{\mathrm{pa}(\delta)}}\right), \tag{7}$$

dove:

$$\theta_{\delta|i_{\mathrm{pa}(\delta)}} = (\theta_{i_\delta|i_{\mathrm{pa}(\delta)}})_{i_\delta \in \mathcal{I}_\delta}$$

ed i parametri soddisfano:

$$\sum_{i_\delta \in \mathcal{I}_\delta} \theta_{i_\delta|i_{\mathrm{pa}(\delta)}} = 1 \quad \text{e} \quad 0 \leq \theta_{i_\delta|i_{\mathrm{pa}(\delta)}} \leq 1$$

Per i nodi continui dove la probabilità di distribuzione è rappresentata da una regressione lineare Gaussiana sui nodi genitori continui con parametri dipendenti dalla configurazione dei nodi genitori discreti:

$$\theta_{\gamma|i_{\mathrm{pa}(\gamma)}} = \left(m_{\gamma|i_{\mathrm{pa}(\gamma)}}, \beta_{\gamma|i_{\mathrm{pa}(\gamma)}}, \sigma^2_{\gamma|i_{\mathrm{pa}(\gamma)}}\right), \quad (8)$$

così che si ottiene:

$$\left(Y_\gamma|i_{\mathrm{pa}(\gamma)}, y_{\mathrm{pa}(\gamma)}, \theta_{\gamma|i_{\mathrm{pa}(\gamma)}}\right) \sim \mathcal{N}\left(m_{\gamma|i_{\mathrm{pa}(\gamma)}} + y_{\mathrm{pa}(\gamma)}\beta_{\gamma|i_{\mathrm{pa}(\gamma)}}, \sigma^2_{\gamma|i_{\mathrm{pa}(\gamma)}}\right). \quad (10)$$

per la variabile discreta d la distribuzione di probabilità locale suggerita:

$$p(i_\delta|i_{\mathrm{pa}(\delta)})$$

è assunta essere uniforme per tutti i livelli di ogni configurazione:

$$p(i_\delta|i_{\mathrm{pa}(\delta)}) = 1/\mathcal{I}_\delta.$$

definendo

$$z_{\mathrm{pa}(\gamma)} = (1, y_{\mathrm{pa}(\gamma)})$$

e ponendo

$$\eta_{\gamma|i_{\mathrm{pa}(\gamma)}} = (m_{\gamma|i_{\mathrm{pa}(\gamma)}}, \beta_{\gamma|i_{\mathrm{pa}(\gamma)}}),$$

dove

$$m_{\gamma|i_{\mathrm{pa}(\gamma)}}$$

rappresenta l'intercetta e

$$\beta_{\gamma|i_{\mathrm{pa}(\gamma)}}$$

è il vettore dei coefficienti.
Per una variabile continua ?, la distribuzione di probabilità locare suggerita:

$$\mathcal{N}(z_{\mathrm{pa}(\gamma)}\eta_{\gamma|i_{\mathrm{pa}(\gamma)}}, \sigma^2_{\gamma|i_{\mathrm{pa}(\gamma)}}) \quad (11)$$

è determinata come una regressione dei nodi genitori continui per ogni configurazione dei nodi parenti discreti.

La distribuzione congiunta

Si può ricavare la probabilità congiunta di un network dalla probabilità di distribuzione locale. Per la parte discreta del network, la distribuzione congiunta di probabilità può essere rappresentata nel seguente modo:

$$p(i) = \prod_{\delta \in \Delta} p\left(i_\delta | i_{\mathrm{pa}(\delta)}\right). \quad (12)$$

Per variabili continue la distribuzione di probabilità

$$\mathcal{N}(M_i, \Sigma_i)$$

è determinata per ogni configurazione delle variabili discrete applicando un algoritmo sequenziale sviluppato da Shachter e Kenley (1989) (R.D. Shachter and C.R. Kenley, 1989).

Parameter learning

Per individuare i parametri nel network, si una un approccio Bayesiano. Si realizza l'incertezza riguardo ai parametri ? in una distribuzione preesistente p(?), usando il dato d per aggiornare la distribuzione ed ottenere la distribuzione finale p(?|d) mediante il teorema di Bayes:

$$p(\theta|d) = \frac{p(d|\theta)p(\theta)}{p(d)}, \qquad \theta \in \Theta. \tag{13}$$

dove T è il parametro spaziale, d è un campione random dalla probabilità di distribuzione p(x|?) e p(d|?) è la probabilità di distribuzione congiunta di d, che è chiamata la probabilità di ?. Questo in gergo viene definito *"parameter learning"*.

Assumendo che i parametri associati con una variabile sono indipendenti dai parametri associati con altre variabili ed in aggiunta che i parametri sono indipendenti per ogni configurazione dei nodi genitori discreti si può formulare la relazione precedente nel seguente modo:

$$p(\theta) = \prod_{\delta \in \Delta} \prod_{i_{\mathrm{pa}(\delta)} \in \mathcal{I}_{\mathrm{pa}(\delta)}} p(\theta_{\delta | i_{\mathrm{pa}(\delta)}}) \prod_{\gamma \in \Gamma} \prod_{i_{\mathrm{pa}(\gamma)} \in \mathcal{I}_{\mathrm{pa}(\gamma)}} p(\theta_{\gamma | i_{\mathrm{pa}(\gamma)}}), \tag{14}$$

che viene chiamata *indipendenza dei parametri* (Bøttcher, 2001).

Come distribuzione locale dei parametri si usa la distribuzione di Dirichlet per le variabili discrete e quella Gaussiana "inverse-Gamma" per quelle continue.

Procedura del precursore dominante

La procedura iniziale per costruire un network Bayesiano consiste di tre passi principali:

1. specificare un network Bayesiano di partenza (es: un DAG iniziale) ed una prima distribuzione di probabilità locale e successivamente calcolare una prima distribuzione congiunta.

2. da questa prima distribuzione congiunta si passa alla distribuzione marginale di tutti i parametri nella famiglia consistente di tutti i nodi e dei nodi precursori che possono essere determinati. Ciò viene chiamato in gergo *"master prior"*.
3. la parametrizzazione dei precursori locali è ora determinata attraverso il condizionamento nella distribuzione del *"master prior"*.

Questa procedura assicura l'indipendenza dei parametri ed ulteriormente ha la proprietà che se un nodo ha lo stesso set di parametri in due differenti networks, allora il parametro locale precedente per quel nodo sarà lo stesso nei due networks. Inoltre, si deve solo dedurre il parametro locale precedente per un nodo dando lo stesso set di precursori una volta soltanto. Questa particolarità viene chiamata *"modularità di parametrizzazione"*.

Se si parla di precursore dominante per nodi discreti si deve porre:

$$\Psi = (\Psi_i)_{i \in \mathcal{I}}$$

Come i parametri per la distribuzione congiunta delle variabili discrete. La distribuzione congiunta del parametro precursore è assunta essere una distribuzione di Dirichlet:

$$p(\Psi) \sim \mathcal{D}(\alpha),$$

che ha come iperparametro

$$\alpha = (\alpha_i)_{i \in \mathcal{I}}$$

Per specificare questa distribuzione di Dirichlet si deve specificare questo iperparametro. Considerando la seguente relazione per la distribuzione di Dirichlet:

$$p(i) = \mathbb{E}(\Psi_i) = \frac{\alpha_i}{N},$$

con

$$N = \sum_{i \in \mathcal{I}} \alpha_i$$

si può usare la probabilità nel network precursore come una stima di

$$\mathbb{E}(\Psi_i)$$

così che si necessita solo di determinare N per calcolare il parametro a_i.

Si può quindi determinare N usando il concetto di un database immaginario. Si può, infatti, immaginare di possedere un database di casi dai quali si possa aggiornare la distribuzione di ? . La dimensione del campione immaginario di questo tipo di database è così N. Esso esprime quanta confidenza si ha nell'indipendenza espressa nel network di partenza (Heckrman D, 1995).

Si usa quindi questa distribuzione congiunta per dedurre la distribuzione del precursore dominante della famiglia $A = d \cup pa(d)$.

Otteniamo quindi:

$$\alpha_{i_A} = \sum_{j:j_A=i_A} \alpha_j,$$

si pone

$$\alpha_A = (\alpha_{i_A})_{i_A \in \mathcal{I}_A}.$$

quindi la distribuzione marginale di $?_A$ è rappresentata da Dirichlet:

$$p(\Psi_A) \sim \mathcal{D}(\alpha_A)$$

Questo è il precursore dominante nel caso discreto. La parametrizzazione dei precursori locali può ora essere affrontata con il condizionamento di queste distribuzioni dei precursori dominanti.

Se si parla di precursore dominante per nodi continui ci si deve rifare a Bøttcher (2001) che ha derivato la procedura per i casi misti. Per una configurazione i delle variabili discrete si pone:

$$\nu_i = \rho_i = \alpha_i,$$

dove a_i è stato determinato precedentemente. Inoltre $\Phi_i = (v_i - 1)\sum_i$.

La parametrizzazione dei precursori è assunta essere distribuita come:

$$\begin{aligned} p(M_i|\Sigma_i) &= \mathcal{N}\left(\mu_i, \frac{1}{\nu_i}\Sigma_i\right) \\ p(\Sigma_i) &= \mathcal{IW}(\rho_i, \Phi_i). \end{aligned}$$

Non si può usare queste distribuzioni per derivare precursori per altri networks, così che si usa il database immaginario per derivare i precursori dominanti locali.

Si definisce la notazione:

$$\rho_{i_{A\cap\Delta}} = \sum_{j:j_{A\cap\Delta}=i_{A\cap\Delta}} \rho_j$$

e similarmente per $v_{i_{A\cap\Delta}}$ e $\Phi_{i_{A\cap\Delta}}$.

Per la famiglia $A = g \cup pa(g)$, il precursore dominante locale è formulato in questo modo:

$$\begin{aligned} \Sigma_{A\cap\Gamma|i_{A\cap\Delta}} &\sim \mathcal{IW}\left(\rho_{i_{A\cap\Delta}}, \tilde{\Phi}_{A\cap\Gamma|i_{A\cap\Delta}}\right) \\ M_{A\cap\Gamma|i_{A\cap\Delta}}|\Sigma_{A\cap\Gamma|i_{A\cap\Delta}} &\sim \mathcal{N}\left(\bar{\mu}_{A\cap\Gamma|i_{A\cap\Delta}}, \frac{1}{\nu_{i_{A\cap\Delta}}}\Sigma_{A\cap\Gamma|i_{A\cap\Delta}}\right), \end{aligned}$$

dove

$$\begin{aligned} \bar{\mu}_{i_{A\cap\Delta}} &= \frac{\sum_{j:j_{A\cap\Delta}=i_{A\cap\Delta}} \mu_j\nu_j}{\nu_{i_{A\cap\Delta}}} \\ \tilde{\Phi}_{A\cap\Gamma|i_{A\cap\Delta}} &= \Phi_{i_{A\cap\Delta}} + \sum_{j:j_{A\cap\Delta}=i_{A\cap\Delta}} \nu_j(\mu_j - \bar{\mu}_{i_{A\cap\Delta}})(\mu_j - \bar{\mu}_{i_{A\cap\Delta}})^\top. \end{aligned}$$

come prima, la parametrizzazione dei precursori locali si basa sul condizionamento di questi precursori dominanti locali.

L'apprendimento dei network Bayesiani

Il problema di istruire un network Bayesiano può essere schematizzato nel seguente modo. Dato un *training set* D = {x^1,…,x^N} di istanze indipendenti di ?, trovare un network B = <G,T> che

meglio rappresenta D. Sarebbe meglio indicare non un network, ma una classe di networks che meglio si adattano a D. L'approccio più comune a questo problema consiste nell'introdurre una funzione detta di *"score"* che valuta quanto bene il modello si avvicina ai dati di partenza e che, successivamente, fra vari modelli seleziona quello migliore.

Una funzione di score fra le più usate é quella Bayesiana (G. F. Cooper, E. Herskovits, 1992, D. Heckerman, D. Geiger, and D. M. Chickering, 1995):

$$S(D) = p(D,d) = p(d|D)p(D) \quad (16)$$

che rappresenta lo *score del network*.

Questo score fattorizza in una parte discreta e una parte mista:

$$S(D) = \prod_{\delta \in \Delta} \prod_{i_{\mathrm{pa}(\delta)} \in \mathcal{I}_{\mathrm{pa}(\delta)}} S_{\delta | i_{\mathrm{pa}(\delta)}}(D) \prod_{\gamma \in \Gamma} \prod_{i_{\mathrm{pa}(\gamma)} \in \mathcal{I}_{\mathrm{pa}(\gamma)}} S_{\gamma | i_{\mathrm{pa}(\gamma)}}(D). \quad (17)$$

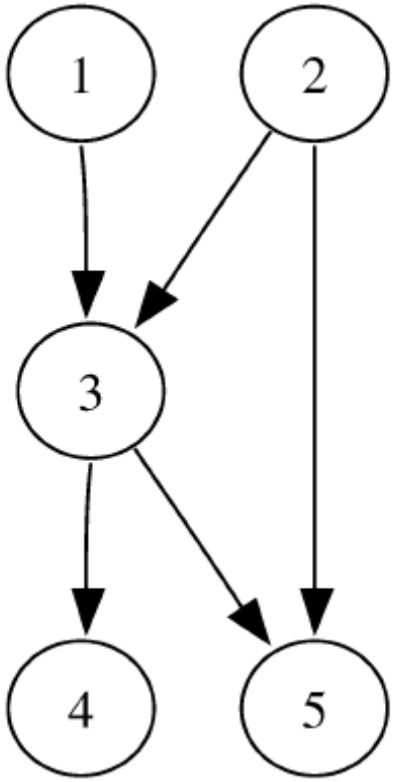
In questo modo questa metrica di valutazione è asintoticamente consistente, infatti dato un sufficientemente largo numero di campioni, le strutture di grafi che catturano esattamente tutte le dipendenze nella distribuzione riceveranno, con alta probabilità, un punteggio più alto rispetto agli altri grafi. Questo indica che dato un numero di istanze sufficientemente largo in un ampio data set, la procedura di learning può indicare l'esatta struttura di un network fino ad ottenere la corretta classe di equivalenza.

I networks Bayesiani possono rappresentare lo stato di un trascritto come una variabile random, X_i. La variabile random è specificata da una funzione, f_i, che descrive la probabilità di distribuzione che è dipendente da un set di trascritti regolatori, X_j. Possiamo identificarli in questo modo:

$$\mathrm{Prob}(X_i = x_i \,|X_j = x_j) = f_i(x_i \,|x_j) \quad (18)$$

dove j = 1,…,N, e j ? i, e x denota un particolare valore di X. Per semplicità possiamo scrivere la parte sinistra dell'equazione come $P(X_i \,|X_j)$. Questa distribuzione condizionale ha una ulteriore restrizione, un trascritto potrebbe essere regolatorio di ogni altro trascritto, a condizione che il network non contenga cicli (i.e., no feedback loops). Questa restrizione rappresenta la principale limitazione dei modelli che sfruttano gli algoritmi Bayesiani. Ciò deriva dalla funzione di distribuzione della probabilità per tutti i trascritti, *f*, dalla quale si può derivare l'eq. (18).

Per esempio, lo stato dei trascritti nel network sottostante:



è dato dalla seguente distribuzione:

$$\text{Prob}(X_1 = x_1, \ldots, X_5 = x_5) = f(x_1, \ldots, x_5)$$

Si può scrivere la parte sinistra più semplicemente come. $P(X_1, \ldots, X_5)$ ed a questo punto la distribuzione di probabilità congiunta si ottiene come il prodotto delle probabilità condizionali:

$$P(X_1, \ldots, X_5) = P(X_5|X_4, \ldots, X_1)P(X_4|X_3,X_2,X_1)P(X_3|X_2,X_1)P(X_2|X_1)P(X_1)$$

e questo ci riporta all'eq. (18) dimostrando che la distribuzione congiunta non può essere soddisfatta da un network la cui topologia contiene cicli.

Se si assume che la probabilità di ogni trascritto dipenda solo dallo stato dei trascritti, allora l'equazione sopra può essere ulteriormente semplificata:

$$P(X_1, \ldots, X_5) = P(X_5|X_3,X_2)P(X_4|X_3)P(X_3|X_2,X_1)P(X_2)P(X_1)$$

Per ottenere il reverse-engineer di un modello di network Bayesiano di un network di trascrizione, si devono ottenere due data sets di parametri: la topologia del modello (es., i regolatori di ogni trascritto) e la funzione di probabilità condizionale che lega lo stato dei regolatori allo stato dei trascritti. Se il network è ampiamente connesso, ogni trascritto avrà pochi regolatori, i quali aiuteranno a minimizzare il numero di parametri nel modello (Neapolitan RE; 2004).

Più i modelli sono realisti, tanto più sarà presente un numero maggiore di parametri, che richiederanno un maggior numero di dati sperimentali e un maggior sforzo computazionale per essere risolti.

Tipicamente gli algoritmi usati riguardano gli stati stazionari, ma non permettono di modellizzare un network dinamico. Per fare ciò è necessario usare i Dynamic Bayesian networks (van Berlo RJP, van Someren EP, Reinders MJT, 2003, Nachman I, Regev A, Friedman N, 2004) insieme con i dati del tipo "time-series". Questi modelli, inoltre, offrono la possibilità di catturare i "feedback loops", anche se possono aumentare la complessità dei dati e quindi lo sforzo computazionale.

Proprio per questi motivi i ricercatori negli ultimi anni hanno rivolto considerevole attenzione nell'uso dei Bayesian network per il reverse-engineering (Segal E, et al, 2003; Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. 2002; Nachman I, Regev A, Friedman N. 2004; Segal E, Taskar B, Gasch A, Friedman N, Koller D. 2001; Friedman N, Linial M, Nachman I, Pe'er D. 2000;).

3. TECNOLOGIA DEI MICROARRAYS

Un DNA microarray (comunemente conosciuto come gene chip, DNA chip, o biochip) è costituito da una collezione di microscopiche sonde di DNA attaccate ad una superficie solida come vetro, plastica, o chip siliconici formanti un array. Tali array sono usati per la esaminare il profilo d'espressione di un gene o per identificare la presenza di un gene o di una breve sequenza in miscela di migliaia (spesso anche tutto il patrimonio genetico di un individuo umano o non).

I microarrays sfruttano una tecnica di ibridazione inversa, consiste cioè nel fissare tutti i probe su un supporto e nel marcare, invece, l'acido nucleico target. È una tecnica che è stata sviluppata negli anni '90, oggi permette l'analisi dell'espressione genica monitorando in una sola volta gli RNA prodotti da migliaia di geni. Per studiare gli mRNA, essi vengono prima estratti dalle cellule, convertiti in cDna, con l'uso di un enzima chiamato transcriptasi inversa e allo stesso momento marcati con una sonda fluorescente. Quando si fa avvenire l'ibridazione fra la sonda presente sulla matrice e il cDna target, quest'ultimo rimarrà legato alla sonda e può essere identificato semplicemente rilevando la posizione dove è rimasto legato. Il segmento di DNA legato al supporto solido è noto come probe.

Migliaia di probe sono usati contemporaneamente in un array (figura 3.1). Questa tecnologia è nata da una tecnica più semplice nota come Southern blotting, dove frammenti di DNA attaccati ad un substrato sono testati da sonde nucleotidiche aventi sequenze conosciute. La misura dell'espressione genica mediante microarrays ha un notevole interesse sia nel campo della ricerca di base che nella diagnostica medica, in particolare di malattie a base genetica, dove l'espressione genetica di cellule sane viene comparata con quella di cellule affette dalla malattia in esame.

Le principali applicazioni dei microarrays sono l'analisi dei polimorfismi SNP, il confronto di popolazioni di RNA di cellule diverse e l'utilizzo per nuove metodologie di sequenziamento del DNA, nonché per lo screening di sequenze senso e antisenso nella ricerca degli oligonucleotidi usati in campo farmaceutico.

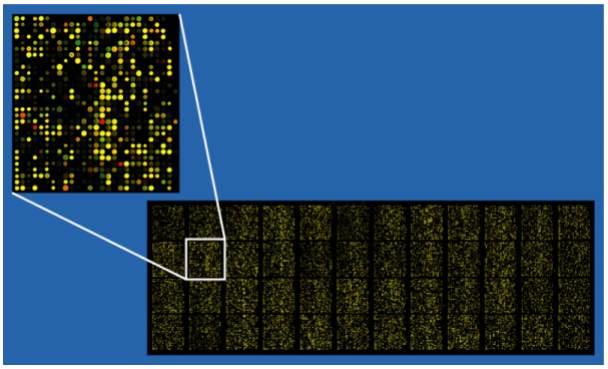


Figura 3.1: schema di un microarray che mette in evidenza il numero e la complessità dei probe.

Origini e tecnologia dei microarrays

Il primo lavoro sui microarrays è stato pubblicato nel 1995 da Mark Schena e collaboratori (Schena M., 1995) dell'università di Stanford e il primo genoma eucariotico completato con analisi di microarrays fu quello del Saccharomyces cerevisiae nel 1997 (Science).

L'idea ebbe origine dalla necessità di studiare l'espressione genica delle piante attraverso la caratterizzazione dei loro fattori di trascrizione: la difficoltà dovuta all'assenza di adeguati strumenti di analisi fece avanzare la proposta di sviluppare degli appositi chip di vetro come dispositivi utili allo studio dei trascritti.

Il Davis Laboratory e il dipartimento di biochimica di Stanford realizzarono microscopici array (microarray) contenenti sequenze geniche di piante bloccate su un substrato di vetro; i microarrays furono poi utilizzati per misurare l'espressione genica di tali piante in esperimenti di ibridazione con campioni di mRNA (RNA messaggero) marcati in fluorescenza.

In condizioni sperimentali appropriate, i segnali fluorescenti sulla superficie del vetrino producono una misura dell'espressione di ogni gene rappresentato sul microarray: dalla quantificazione di tale fluorescenza è possibile risalire al livello di espressione di ciascun gene.

Il laboratorio di Stanford utilizzò tecniche fotolitografiche, ink jetting e contact printing per creare i microarrays, mutuando tre approcci tradotti in realtà solo negli ultimi ventanni in ambiente microelettronico per la costruzione di circuiti microlavorati (MEMS): si può, quindi, comprendere il motivo di un tale divario temporale fra la comprensione del paradigma biologico alla base dei microarrays e la loro effettiva realizzazione.

La tecnologia alla base dei microarrays a DNA

La tecnologia dei microarrays a DNA si basa sulla capacità di ibridizzazione degli acidi nucleici, secondo cui due filamenti di DNA ibridizzano tra di loro se sono complementari l'uno all'altro. Questa complementarità riflette la regola di Watson e Crick secondo la quale l'adenina si lega alla timina e la citosina si lega alla guanina. Uno o entrambi i filamenti di DNA ibridizzati possono essere sostituiti con RNA che, pur differendo per la presenza dell'uracile al posto della timina, va incontro ugualmente al fenomeno dell'ibridizzazione.

L'ibridizzazione è stata per decenni utilizzata in biologia molecolare come principio base di metodiche quali il Southern blotting e il Northern blotting ed i microarrays a DNA sono una massiccia parallelizzazione di queste tecniche poiché sono in grado di analizzare migliaia di geni contemporaneamente (figura 3.2).

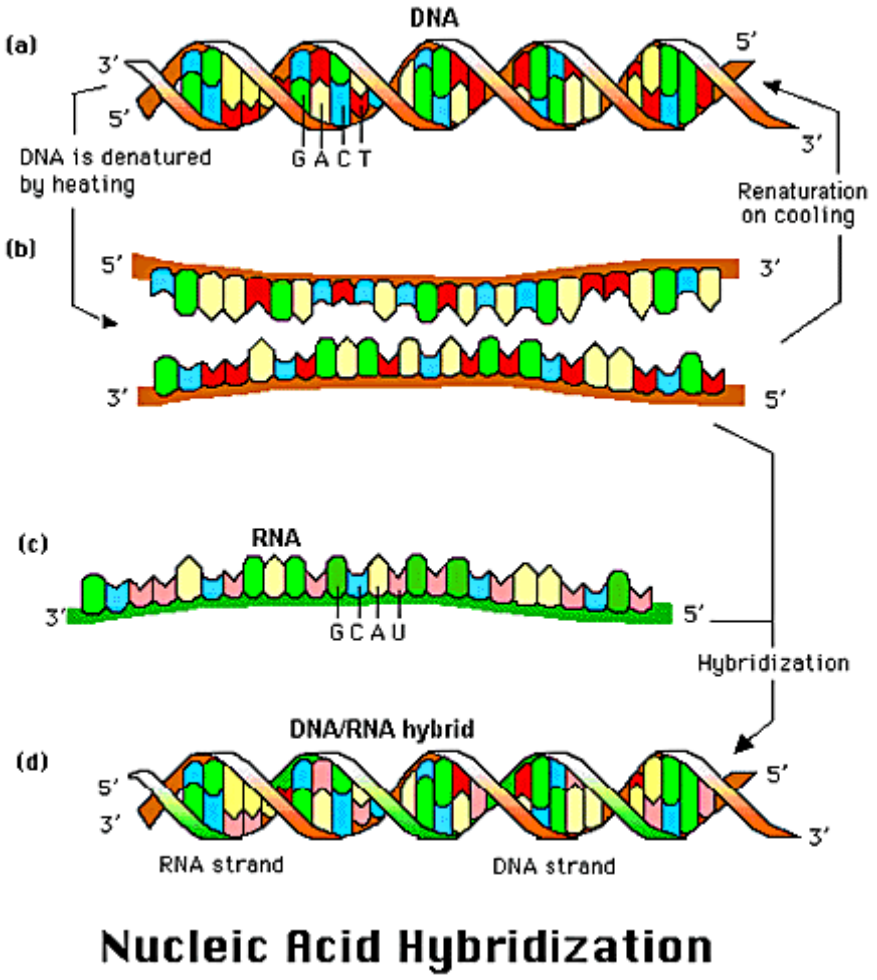


Figura 3.2: ibridazione di DNA e RNA.

Nel caso dei microarrays invece di distribuire le sonde oligonucleotidiche su un gel che contiene i campioni di RNA o DNA, esse vengono bloccate su una superficie di vetro. Sonde diverse possono essere posizionate alla distanza di qualche micron l'una dall'altra in modo da disporne un numero molto elevato in pochi centimetri quadrati. Il campione in studio viene marcato con fluorocromi e lasciato ibridizzare con le sonde presenti sul microarray. Dopo aver lavato l'eccesso di materiale non ibridizzato, i fluorocromi legati al campione ibridizzato vengono eccitati con un laser di opportuna lunghezza d'onda che scandisce la superficie del chip. Poiché la posizione delle sonde è individuabile grazie ad uno schema a mappa cartesiana, è possibile quantificare l'ammontare di campione ibridizzato a partire all'immagine generata con lo scanner.

La concentrazione di un particolare mRNA è il risultato dell'espressione del gene da cui esso viene trascritto (figura 3.3); per questo motivo le applicazioni che fanno uso di microarrays a cDNA vengono spesso denominate analisi dell'espressione genica.

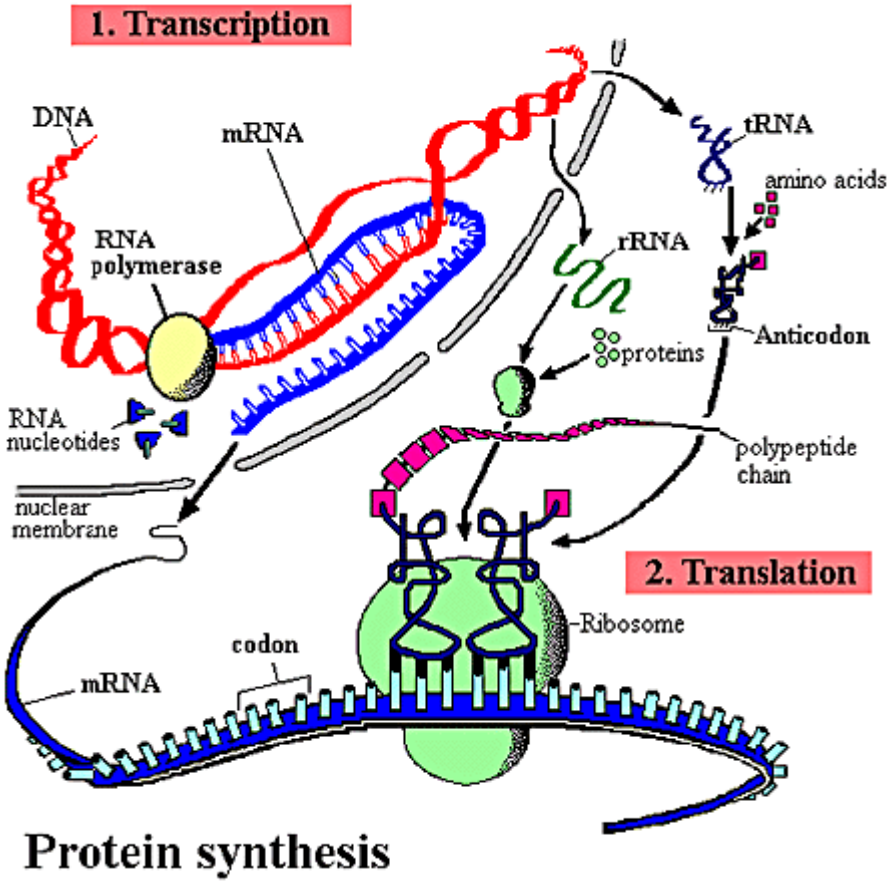


Figura 3.3: processo di sintesi delle proteine.

Quando si vuole evidenziare la differente risposta di un gene alla sua esposizione a trattamenti diversi o osservare la sua espressione in momenti diversi si dice che si sta generando un profilo di espressione.

Un'altra applicazione tipica dei microarrays è la rilevazione di polimorfismi in geni specifici: la peculiare struttura parallela dei microarrays consente di rilevare simultaneamente numerosi polimorfismi genetici in più geni, permettendo in questo modo di fare una genotipizzazione.

Esistono diversi tipi di microarrays, catalogati, a seconda del materiale che viene utilizzato come sonde:

- Microarray a cDNA, con sonde di lunghezza maggiore di 200 basi ottenute per retrotrascrizione da mRNA, frammentate, amplificate con PCR e depositate su un supporto di vetro o di nylon;
- Microarray ad oligonucleotidi, con sonde di lunghezza fra 25 e 80 basi ottenute da materiale biologico o per via artificiale e depositate su un supporto di vetro;

- Microarray ad oligonucleotidi, con sonde di lunghezza fra 25 e 30 basi sintetizzate in situ con tecniche fotolitografiche su wafer di silicio (figura 3.4).

Figura 3.4: microarrays di oligonucleotidi da cui si possono notare le dimensioni estremamente ridotte dal confronto con un fiammifero. Questo risultato è stato possibile grazie alla continua miniaturizzazione.

Per l'analisi dell'espressione sono presenti sul mercato due tecnologie dominanti: Affymetrix, Inc. GeneChip e quella degli "spotted" array a cDNA.

Tecnologia Affymetrix GeneChip

Affymetrix utilizza attrezzature simili a quelle che servono a realizzare i chip di silicio per i computer, che consentono di avere una produzione massiva di chip ad un costo ragionevole. Così come i chip per computer sono fatti utilizzando maschere che controllano il processo di deposizione e rimozione del silicio dalla superficie del chip, analogamente Affymetrix usa maschere di controllo della sintesi degli oligonucleotidi sul microarray. Il risultato di questo processo è la produzione di alcune centinaia di migliaia di oligonucleotidi differenti, ciascuno dei quali presente in milioni di copie sul vetrino (figura 3.5).

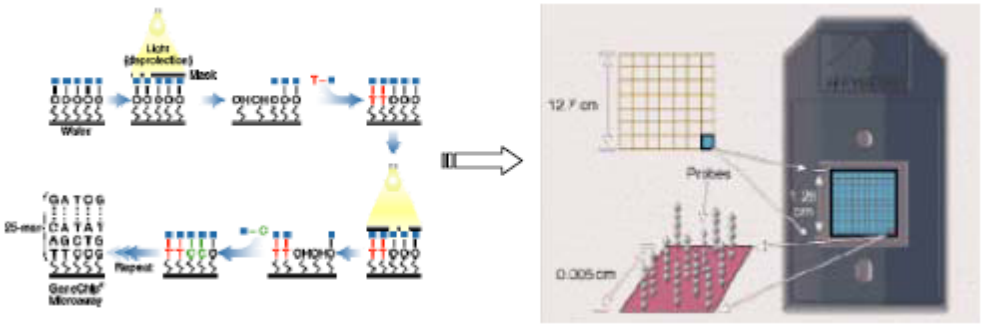


Figura 3.5: microarray Affymetrix

Per l'analisi di espressione sono utilizzati gruppi di sonde di almeno 40 oligonucleotidi per gene; Affymetrix ha selezionato, per ogni gene, una regione con la minor omologia con altri geni. A partire da questa regione vengono disegnati da 11 a 20 oligonucleotidi rappresentativi del perfect match (PM), cioè della perfetta complementarità con l'mRNA bersaglio, e 11-20 oligonucleotidi identici ai precedenti tranne che per il nucleotide centrale, utili per rilevare il mismatch (MM), cioè la non perfetta complementarità (figura 3.6).

Affimetrix afferma che gli oligonucleotidi MM sono capaci di mettere in evidenza la presenza di segnali aspecifici permettendo di rilevare con maggior sicurezza i segnali deboli.

L'ibridizzazione di ogni oligonucleotide con il proprio complementare dipende dalla sequenza specifica e poiché si è interessati alla misura del cambiamento di espressione di un gene è necessario ottenere un dato cumulativo da tutte le sonde che identificano quel gene. Affymetrix calcola questo dato cumulativo facendo una media della differenza fra sonde PM e MM dello stesso gene:

$$AvgDiff = \frac{\sum_N (PM - MM)}{N}$$

dove N è il numero di sequenze specifiche che identificano un gene. Se il numero che si ottiene da questo calcolo è negativo o molto piccolo significa che il cDNA bersaglio è assente o che si è verificata un'ibridizzazione non specifica.

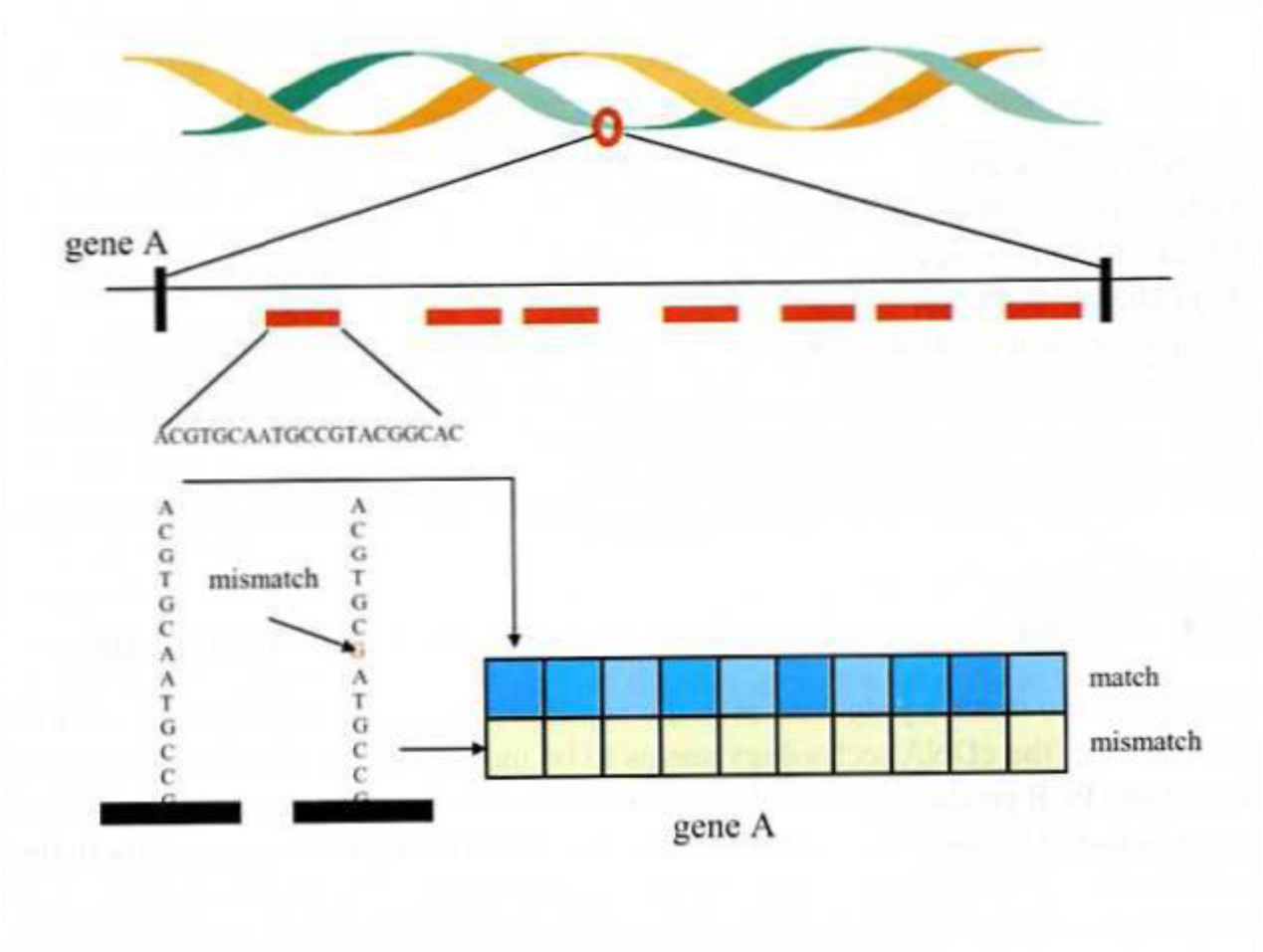


Figura 3.6: il principio della tecnologia Affymetrix. La struttura del probe set nei microarrays di oligonucleotidi. Nella celletta superiore del probe pair è presente un frammento della sequenza originale del trascritto (PM), nella celletta inferiore la sequenza MM alterata nella settima base. Ogni probe pair rappresenta una porzione specifica dell'mRNA. Ogni probe set rappresenta un trascritto unico.

Una volta che il microarray è stato costruito o comprato e il campione di acidi nucleici da analizzare è stato isolato si fa avvenire la reazione di ibridazione, che permette la formazione degli eteroduplex. Per ottenere dei buoni microarrays è essenziale difenderli dall'umidità (se l'ambiente è secco la soluzione evapora, se invece è umido si deposita dell'acqua) e dalla polvere (ogni spot è grande circa 50 micron,un granello di polvere e più grande di 50 micron, per cui può coprire vari spot), per questo motivo esistono delle camere apposite per l'ibridazione dei microarrays che vengono sigillate.

Dopo l'ibridazione il microarray viene lavato per rimuovere il cDNA che non si è legato. Generalmente il DNA fluorescente dei campioni sperimentali è mescolato con un DNA di un soggetto di controllo marcato con un colorante fluorescente diverso. Per i microarrays si usano solitamente Cy3 (che emette una lunghezza d'onda nel campo del verde) e Cy5 (che emette nel

campo del rosso). In questo modo se la quantità di RNA espressa da un gene nelle cellule di interesse è aumentata (up regolata) rispetto a quella del campione di riferimento, lo spot che ne risulta sarà del colore del primo fluorescente. Viceversa se l'espressione del gene è diminuita (down regolata) rispetto al campione di riferimento lo spot sarà colorato dal secondo fluorescente.

La fluorescenza è rilevata poi grazie ad uno scanner a laser, grazie al quale si acquisisce un'immagine per ogni fluoroforo. Poi vengono usati dei software appositi per convertire i segnali in una gamma di colori dipendente dalla loro intensità. Il segnale rilevato dallo scanner viene poi sottoposto ad altri algoritmi di filtrazione e di pulizia e convertito in valori numerici.

Tutti gli algoritmi che riguardano la rilevazione di ibridizzazione sul chip, la generazione del dato cumulativo e la sua elaborazione sono protetti dalla tecnologia proprietaria Affymetrix che, per altro, si riserva di modificarli senza renderli noti.

Le fasi di un esperimento di analisi dell'espressione genica che fa uso di chip Affymetrix sono:

- Estrazione dell'RNA totale dal campione;
- Separazione dell'mRNA dall'RNA totale utilizzando colonnine con code di poly-T;
- Conversione dell'mRNA in cDNA utilizzando la trascrittasi inversa e i primer poly-T;
- Amplificazione del cDNA utilizzando T7 RNA polimerasi in presenza di biotina-UTP e biotina-CTP in modo da ottenere da 50 a 100 copie di cDNA marcato;
- Incubazione del cDNA a 94°C in un buffer di frammentazione per produrre frammenti di lunghezza tra 35 e 200 nucleotidi;
- Ibridizzazione sul chip e successivi lavaggi;
- Marcatura del cDNA ibridizzato con Streptavin-Phycoerythrin e successivi lavaggi;
- Acquisizione dell'immagine del chip con scanner laser;
- Analisi dell'immagine per l'estrapolazione dei dati.

"Spotted" array

L'altra tecnologia largamente utilizzata per produrre microarrays è quella degli "spotted" array; in questo caso viene utilizzato un robot che preleva una piccola quantità di sonda in soluzione da una piastra da microtitolazione e la deposita sulla superficie del microarray. La sonda può essere cDNA, prodotto mediante PCR od oligonucleotidi; ogni sonda è complementare ad un unico gene.

Esistono diversi metodi per fissare le sonde alla superficie del vetrino; il più utilizzato consiste nel ricoprire il supporto con uno strato di poli-lisina che determina la formazione di legami aspecifici con le sonde.

Il processo di “spotting” di questi microarrays può essere schematizzato come segue:

- Copertura del vetrino con poli-lisina;
- Preparazione delle sonde in una piastra da microtitolazione;
- Programmazione del robot per le operazioni di “spotting” mediante pin e ugelli ink-jet;
- Deposizione delle sonde in blocchi ordinati seguendo la mappa programmata per stabilire la posizione e la concentrazione di ogni spot;
- Saturazione delle aree non stampate con anidride succinica per sfavorire legami aspecifici fra il cDNA bersaglio e il supporto;
- Denaturazione delle sonde ad alta temperatura in modo che siano a singolo filamento.

Una volta realizzato il microarray si può procedere alla preparazione del campione e alla sua ibridizzazione (figura 3.7) come segue:

- Estrazione dell’RNA totale dalle cellule;
- Isolamento (opzionale) dell’mRNA grazie alla presenza delle code di poly-A;
- Retrotrascrizione dell’RNA in cDNA in presenza di amino-allil-dUTP (AA-dUTP);
- Marcatura dei filamenti di cDNA con i fluorocromi Cy3 e Cy5, che si legano all’AA-dUTP;
- Ibridizzazione del cDNA marcato con le sequenze presenti sul vetrino;
- Asportazione mediante lavaggi del materiale non ibridizzato;
- Acquisizione dell’immagine del vetrino con scanner laser;
- Analisi dell’immagine per l’estrapolazione dei dati.

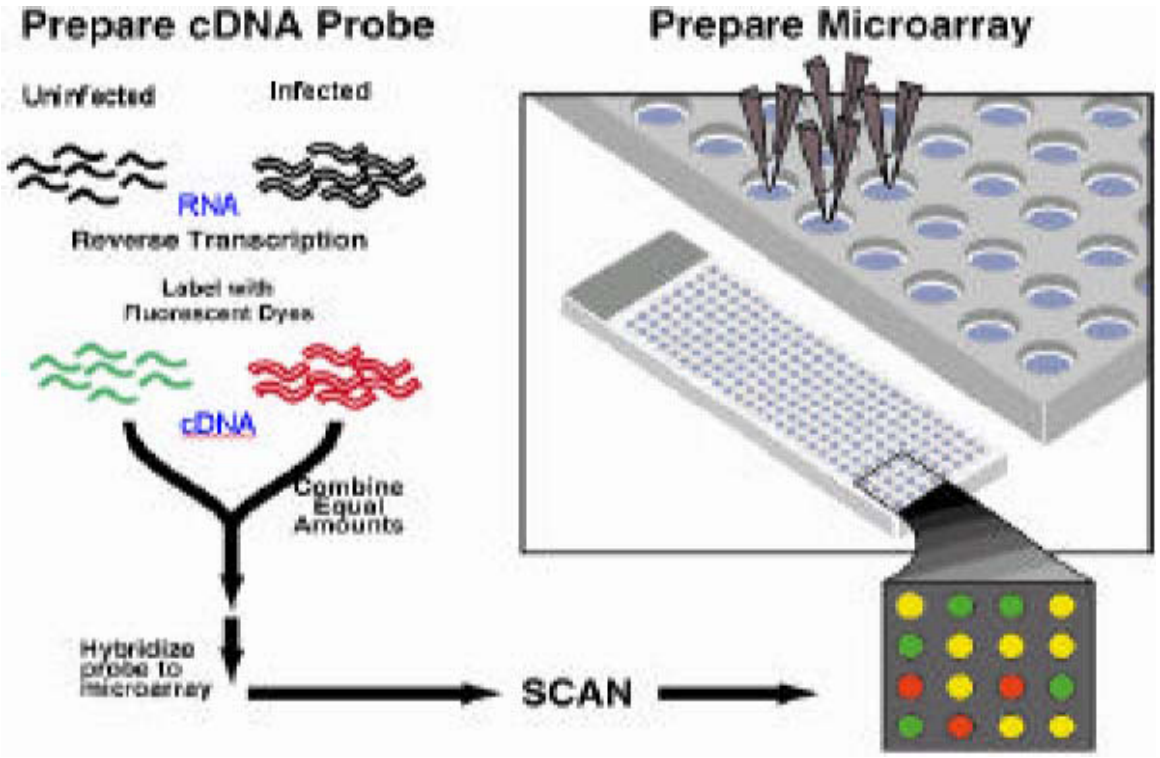


Figura 3.7: processo di spotting dei microarrays e ibridazione del campione.

Rispetto alla tecnologia Affymetrix, negli "spotted" array l'irregolarità nell'operazione di deposizione delle sonde si può ripercuotere sulla corretta estrazione del dato. Inoltre, la presenza di sonde PM e MM sui vetrini Affymetrix, conferisce a questi microarrays una maggiore affidabilità nella rilevazione di segnali di ibridizzazione aspecifica.

Il vantaggio principale degli "spotted" array, invece, consiste nella possibilità che ogni laboratorio ha di disegnare le sonde da utilizzare nello "spotting" e nella maggiore flessibilità di questa tecnologia rispetto ad Affymetrix, i cui dati spesso non sono analizzabili con gli innumerevoli software per l'elaborazione di dati disponibili.

Caratteristiche di un microarray

Un microarray può essere definito come una matrice ordinata di elementi microscopici su un substrato planare che consente il legame specifico di geni o di prodotti di geni. La parola microarray deriva dal greco mikro, che significa piccolo, e dal francese arayer, che significa arrangiare; i microarrays, anche conosciuti come biochip, DNA chip e gene chip, contengono, infatti, collezioni di microscopici elementi, spot, disposti in righe e colonne.

Ogni riga di elementi deve essere disposta sul substrato lungo una linea orizzontale e ogni colonna deve formare una linea verticale perpendicolare alla riga. Gli elementi ordinati devono avere uguale dimensione, uniforme spaziatura e posizione unica sul substrato (figura 3.8).

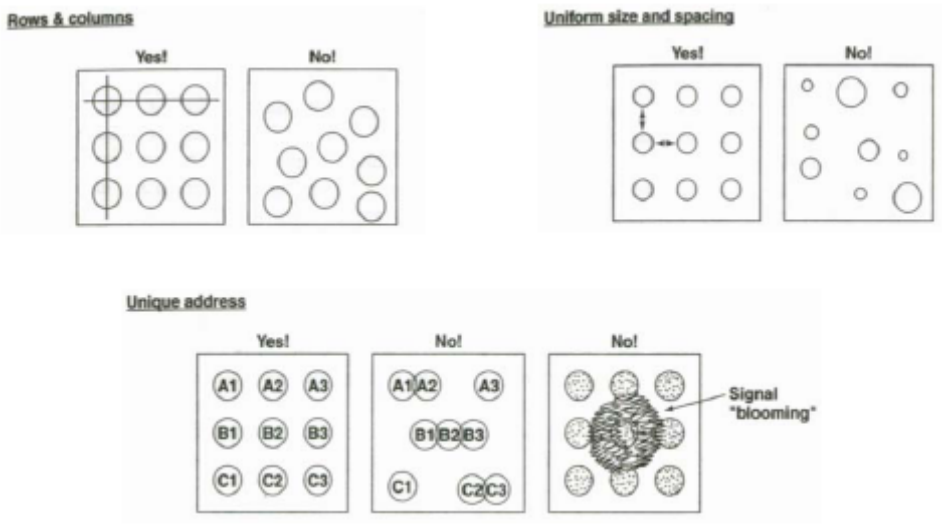


Figura 3.8: microarray ordinato.

Su un singolo substrato planare possono essere combinati diversi microarrays e ciò è utile sia dal punto di vista dell'analisi successiva, sia per i processi di realizzazione in parallelo di tali dispositivi (figura 3.9).

L'ordinamento in righe e colonne degli elementi è un grande vantaggio per l'analisi dei microarrays, poiché questo tipo di disposizione consente una rapida deposizione, individuazione e quantificazione degli spot.

La disposizione degli spot in righe e colonne può essere ottenuta utilizzando tecnologie standard di motion control, come attuatori lineari ed encoder, e ciò permette un abbattimento dei costi di produzione, in quanto i microarrays possono essere stampati in modalità rapida e completamente automatizzata, con una velocità e una precisione che non sarebbero possibili con formati irregolari.

La regolarità della disposizione degli spot, inoltre, favorisce il processo di quantificazione, poiché i software di elaborazione fanno uso di griglie ordinate per l'estrazione del dato numerico e di una "mappa cartesiana" per assegnare allo spot l'identificativo del gene che rappresenta.

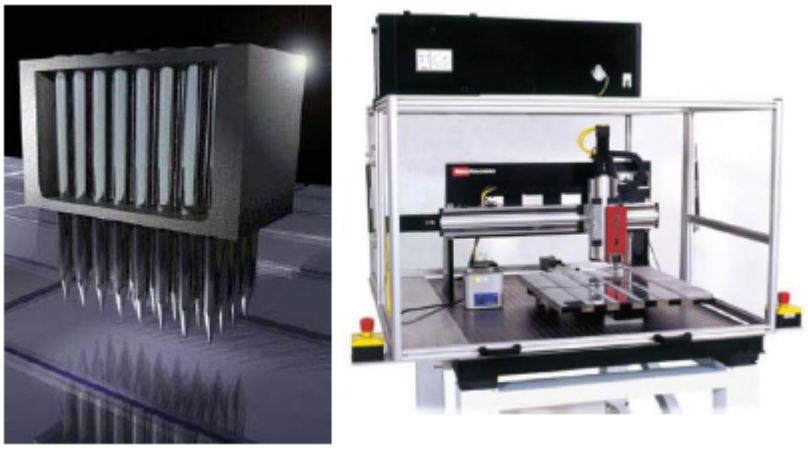


Figura 3.9: printer-head di un robot per spotting di microarrays e camera di printing.

Un tipico spot contiene approssimativamente 109 molecole bloccate sul substrato di vetro. Queste molecole sonda possono essere DNA genomico, cDNA, mRNA, proteine, tessuti o altri tipi di molecole che necessitano di un'analisi quantitativa. Oligonucleotidi sintetici, cioè piccole molecole di DNA a singolo filamento sintetizzate chimicamente possono costituire un tipo eccellente di sonda.

I vantaggi di avere elementi microscopici sono:

- alta densità degli spot (> 5000 elementi/cm^2);
- rapida cinetica di reazione;
- possibilità di analizzare interi genomi su un singolo vetrino.

Gli esperimenti che esaminano tutti i geni di un genoma su un singolo substrato procurano una visione globale del fenomeno biologico, impossibile da ottenere con tecnologie limitate a sottoinsiemi di geni.

Per substrato si intende un supporto parallelo e piatto sul quale viene configurato un microarray. Uno dei materiali più utilizzati è il vetro per la sua capacità ideale di consentire il legame con le molecole sonda, ma possono essere utilizzati anche materiali plastici, silicio, filtri di nylon e nitrocellulosa.

Per essere utilizzato per la costruzione di un microarray il substrato deve essere planare: tutti i materiali planari sono solidi, ma non tutti quelli solidi sono planari.

Il vantaggio di avere un substrato piatto su tutta la superficie si ripercuote sull'automatizzazione della procedura di stampa mediante pin e ugelli ink-jet o sulla precisione delle fotomaschere per la fotolitografia. I materiali planari consentono anche un accurato "scanning" del microarray, grazie alla precisa individuazione della distanza fra gli elementi ottici dello scanner e la superficie del microarray (distanza del fuoco ottico).

I materiali planari, inoltre, tendono ad essere impermeabili ai liquidi, consentono di realizzare piccoli spot e di minimizzare il volume di reazione durante l'ibridizzazione.

Microarrays e bioinformatica

Analisi statistica

L'analisi di DNA microarrays propone numerosi problemi di carattere statistica, compresa la normalizzazione dei dati. Analizzando i microarrays pare evidente che, il grande numero di geni presenti in un singolo array pone lo sperimentatore davanti ad un problema di test multiplo: anche se è estremamente raro e casuale ogni gene può dare un risultato falso positivo, un test effettuato su più geni è più sicuro che mostri un andamento scientificamente più probante. Una differenza fondamentale tra i microarrays e gli altri metodi di analisi biomedici tradizionali sta nella dimensione dei dati. Studi che contengono 100 analisi per paziente per 1000 pazienti possono essere considerati vasti studi clinici. Uno studio microarrays di media vastità comprende diversi migliaia dati per campione su centinaia di campioni diversi.

Standardizzazione

La mancanza di standardizzazione negli arrays presenta un problema interoperativo nella bioinformatica, che non può far prescindere dallo scambio di dati ottenuti con tale tecnica. Diversi progetti open-source si prefiggono di facilitare l'interscambio di dati ottenuti da arrays. Il "Minimum Information About a Microarray Experiment" (MIAME) XML standard base per la descrizione di esperimenti di microarrays è stato adottato da molti giornali scientifici come standard richiesto per l'accettazione di lavori che contengono risultati ottenuti attraverso analisi di microarrays.

Nonostante negli ultimi anni, comunque, siano stati fatti molti progressi questo problema rimane tuttora e causa difficoltà nel confronto di dati; inoltre, se oggi con questa tecnica è possibile analizzare i livelli di espressione di un singolo gene ottenendo degli ottimi risultati, la combinazione dello studio di molte migliaia di geni risulta molto complicato e può portare spesso a dei falsi positivi, questo accade anche a causa del fatto che alcuni cDNA possono cross-ibridare altre sonde (che avrebbero dovuto rilevare altri geni).

Un altro problema è presentato dai fluorofori, che nonostante siano molto simili fra loro presentano delle differenze problematiche. Esiste una diversa efficienza di fluorescenza tra Cy3 e

Cy5 che deve essere standardizzata dai software di rilevazione, inoltre poiché Cy3 è più piccolo di Cy5, c'è un diverso livello di incorporazione dei due fluorofori, in quanto la polimerasi presenta più difficoltà a inserire il nucleotide marcato con Cy5 a causa dell'ingombro sterico; come se non bastasse Cy5 si presenta più labile di Cy3, quindi una prima scansione di Cy3 con il laser potrebbe ridurre la fluorescenza di Cy5. Per ovviare a tutte queste problematiche e per creare un minimo di standardizzazione si effettua il Die swap: consiste nel effettuare un secondo microarray scambiando l'uso dei fluorofori. Se nel primo microarray Cy3 è stato usato per marcare il cDNA sperimentale, nel secondo microarray si userà Cy3 per marcare il cDNA del soggetto di controllo, e viceversa per Cy5.

Applicazioni dei microarrays

I microarrays si stanno rivelando degli strumenti efficaci in differenti campi di indagine. I primi esperimenti hanno fatto uso di questi supporti per verificare ipotesi formulate in studi precedenti.

Ultimamente la tendenza si è invertita e i microarrays vengono utilizzati come dispositivi di indagine primaria, capaci di fornire risposte robuste, ma anche di porre nuovi quesiti al ricercatore.

Da questo punto di vista il grosso vantaggio dei microarrays, oltre all'estesa potenza di calcolo parallelo, sta nella possibilità di poter coinvolgere nella reazione di ibridizzazione molti geni sullo stesso supporto. Questo dà un contributo alla possibilità di ricreare pathway di co-regolazione e di analizzare le inter-relazioni tra geni diversi.

Di seguito sono descritti alcuni esempi dei settori di applicazione dei microarrays.

Tassonomia di tessuti

Cellule appartenenti allo stesso organismo possiedono lo stesso genoma anche se differiscono per forma e funzione. La differenziazione di ogni cellula in un tipo o in un altro si realizza grazie ad una programmazione genetica ben definita che modifica nel corso dello sviluppo l'insieme dei geni espressi.

Esaminando il pattern di espressione genica su scala genomica con i microarrays è possibile catalogare i differenti tessuti in modo da costituire un database di espressione. Uno degli scopi degli studi di questo tipo è la comprensione dei meccanismi che stanno alla base dello sviluppo e della differenziazione cellulare, che, una volta alterati, possono determinare l'insorgenza di malattie.

Identificazione delle basi molecolari delle malattie

Conoscere le basi molecolari di una malattia può aiutare a comprenderne la trasmissione genetica, la modalità d'insorgenza e la prognosi al fine di poter fare una diagnosi precoce o di agire con terapie mirate. Il confronto dell'espressione genica tra tessuti sani e malati mediante microarrays può essere un valido strumento per l'identificazione di quei geni che sono coinvolti nello sviluppo di una patologia, o come geni causativi o semplicemente come fattori di rischio predisponenti.

Diversi gruppi di ricerca stanno facendo uso dei microarrays per creare una carta d'identità dettagliata dei vari tipi di tumore al fine di costituire una vasta raccolta di profili di espressione genica da utilizzare a fini diagnostici.

Si possono ricordare in questo ambito lo studio di Ross (Ross DT, 2000) su sessanta linee cellulari differenti di cancro, denominato NCI60, o gli studi estensivi sui linfomi a cellule B giganti di Alizadeh (Alizadeh AA, 2000), entrambi dell'Università di Stanford.

Riuscire ad effettuare una classificazione così dettagliata si riflette sulla possibilità di riconoscere la malattia fin dai primi stadi di sviluppo, in modo da poter programmare terapie più mirate.

Analisi del meccanismo di azione dei farmaci

I farmaci funzionano legandosi a specifiche molecole bersaglio e il risultato di questa interazione può essere l'alterazione dell'espressione di geni. E' possibile utilizzare i microarrays per individuare quei geni la cui espressione viene modificata dall'impiego di farmaci, sia in studi in vitro su linee cellulari trattate a confronto con le stesse cellule non trattate, sia in trial clinici in cui si generano profili di espressione in pazienti sottoposti a trattamento farmacologico. Il profilo di espressione in seguito a trattamento farmacologico può essere utile anche per identificare l'alterazione nell'espressione di geni che provocano effetti collaterali.

Un approccio di questo tipo può ridurre i costi di sviluppo dei farmaci e produrre medicine più efficaci e con meno effetti collaterali.

Un'altra applicazione dei microarrays in questo ambito è la genotipizzazione dei pazienti, in modo da suddividere la popolazione in soggetti farmaco-sensibili e farmaco-resistenti allo scopo di definire una terapia più mirata.

Perchè usare i microarrays?

I microarrays rappresentano un sistema di analisi in parallelo, che velocizza considerevolmente l'esplorazione genomica: permettono, infatti, di esaminare contemporaneamente l'espressione di migliaia di geni o un ampio numero di polimorfismi genetici. Un altro vantaggio è dato dai costi relativamente contenuti se rapportati al numero di geni o polimorfismi analizzabili per esperimento.

4. R: IL PROGRAMMA UTILIZZATO PER L'ANALISI DEI MICROARRAYS

R è un programma per la manipolazione dei dati, per l'analisi statistica e grafica che permette di essere programmato per soddisfare qualsiasi esigenza (high level graphics, interfaces to other languages and debugging facilities, etc…) (W. N. Venables, D. M. Smith, 2006).
Le sue caratteristiche principali sono:

• un programma per la gestione dei dati ed il loro utilizzo,

• una suite di operazioni sugli arrays, in particolare matrici,

• una grande, coerente ed integrata collezione di tools per l'analisi dei dati,

• una serie di "graphical facilities" per l'analisi e la visualizzazione dei dati analizzati direttamente sul computer,

• un linguaggio ben sviluppato, semplice ed efficace che include variabili, loops, funzioni ricorsive definite dall'utente e svariate opzioni di visualizzazione dei risultati.

Il termine "environment" è inteso per caratterizzare un sistema perfettamente studiato e coerente, piuttosto che un insieme di specifici ed inflessibili tools, come accade con altri software di analisi di dati.

R rappresenta attualmente il modello per sviluppare nuovi metodi di analisi interattiva. Si è sviluppato rapidamente ed è stato successivamente ampliato mediante l'aggiunta di una larga collezione di pacchetti. Comunque la maggior parte dei pacchetti scritti in R sono essenzialmente scritti per una specifica funzione.

Questo linguaggio è un fork derivato da un altro linguaggio sviluppato per l'analisi statistica chiamato "S", il quale fu sviluppato negli anni 80s nei laboratori della Bell da parte di John Chambers and Allan Wilks e negli anni ha avuto grande successo (R.A. Becker, J.M. Chambers and A.R. Wilks, 1988). L'evoluzione del linguaggio S è descritto nei quattro libri di John Chambers che rimane il suo progettista principale ed al quale fu riconosciuto nel 1998 l'ACM Software Systems Award (J.M. Chambers and T.J. Hastie eds, 1992).

La sintassi ha una certa similarità, anche se a livello superficiale, con quella del linguaggio C, ma la semantica relativa al linguaggio funzionale della programmazione presenta una certa varietà con grandi affinità con il Lisp e l'APL (J.M. Chambers, 1998). In particolare permette quello che in gergo prende il nome di "computing on the language", che a sua volta rende possibile la costruzione di funzioni personalizzabili dall'utente (D. M. Bates and D. G. Watts, 1988; S. D. Silvey, 1970; A.J. Dobson, 1990), proprietà che risulta molto utile nella modellazione statistica e nella visualizzazione grafica.

Il programma permette sia di sfruttare certe applicazioni già preconfigurate per essere "user-friendly" sia di utilizzare la riga di commando eseguendo semplici espressioni (C. Davison and D. V. Hinkley, 1997). Non è necessario raggiungere questo livello di esperienza per usare R, ma alcuni utenti più smaliziati che necessitano di personalizzare le funzioni già esistenti o crearne di nuove "ad hoc" al fine di sistematizzare lavori ripetitivi vanno incontro all'esigenza di creare dei pacchetti che permettano di soddisfare le loro necessità (P.McCullagh and J.A. Nelder, 1989).

Fra R ed S ci sono delle sostanziali differenze anche se R deriva direttamente da S. La filosofia con cui il linguaggio è stato costruito contiene un gran numero di punti in comune che potrebbero sorprendere l'utente. Ci sono un certo numero di utili scorciatoie da tastiera ed idiomi, i quali permettono di esprimere in modo succinto operazioni complicate. Molte di queste diventano naturali appena si familiarizza con questi concetti. In alcuni casi ci sono molteplici vie per raggiungere un obbiettivo, ma alcune tecniche richiedono un livello di conoscenza del programma avanzato in quanto lavorano ad un alto livello di astrazione.

R rappresenta un ambiente di programmazione dinamico nel quale si incontrano molte tecniche sia nuove che classiche di analisi statistica. Alcune di queste sono direttamente implementate all'interno del programma base, ma molte possono essere successivamente aggiunte mediante il caricamento di nuovi pacchetti a seconda delle necessità individuali (J.A. Rice, 1995).

La maggior parte delle tecniche di analisi statistica e delle ultime metodologie sono disponibili con R, ma gli utenti devono possedere una certa esperienza per poterci lavorare e per conoscerli a fondo.

Esiste un'importante differenza che sta dietro alla filosofia di S (e quindi R) rispetto agli altri principali sistemi di analisi statistica. In S un'analisi statistica è costituita normalmente come una serie di passi successivi, con risultati intermedi salvati in gruppi, così mentre SAS e SPSS daranno un output copioso ed abbondante da una analisi ottenuta da una regressione o da un discriminante, R darà un risultato minimale ed immagazzinerà i risultati in un susseguirsi di svariati output ed in una serie di oggetti che potranno essere interrogati e sfruttati da ulteriori funzioni di R.

5 BANJO: PROGRAMMA PER I GENE NETWORKS

Banjo è un'applicazione software per lo studio dei Bayesian networks statici e dinamici, che permette di analizzare data sets anche di grandi dimensioni ed essendo stato sviluppato con tecnologia Java, è sia facile da utilizzare che al tempo stesso potente.

Banjo si focalizza sulla previsione della struttura dei modelli utilizzando gli algoritmi Bayesiani. Attualmente prevede l'analisi di modelli di network sia statici che dinamici e come strategie di ricerca include il "simulated annealing" e il "greedy search" che possono essere accoppiati con la valutazione di un "single random local move" o con la ricerca globale "all local moves" ad ogni passo. L'algoritmo di ricerca di Banjo consiste in un set di componenti:

• "Proposing a new network (or networks)" gestito da un componente "proposer"
• "Checking the proposed network(s) for cycles" gestito da un componente "cycle checker"
• "Computing the score(s) of the proposed network(s)" gestito da un componente "evaluator"
• "Deciding whether to accept a proposed network" gestito da un componente "decider"

Questi sono i componenti principali (figura 5.1) che caratterizzano il cuore del programma e che permettono di predire il network basandosi sul dataset.

L'algoritmo è ottimizzato per variabili discrete, ma esiste un sistema di discretizzazione che ne rende possibile l'utilizzo anche con variabili continue sfruttando il metodo dei quantili o degli intervalli. Il programma, quindi, permette di ottenere i valori di "influence score" che descrivono il peso statistico dei singoli nodi e può oltremodo generare un file formattato che tramite altri programmi di visualizzazione del layout come "dot" può essere trasformato in un formato grafico facilmente comprensibile a tutti.

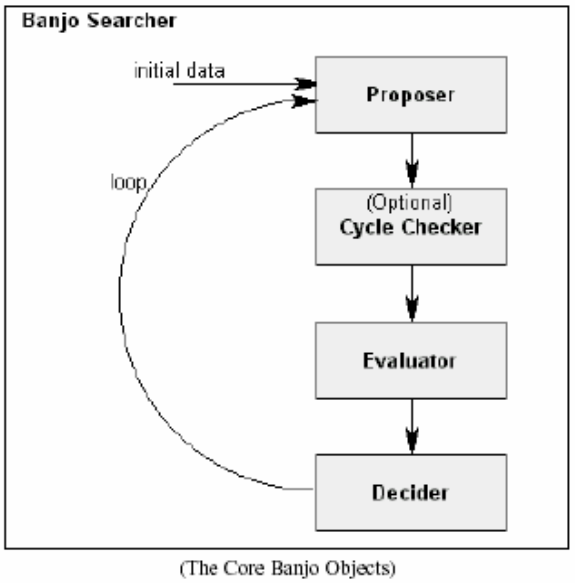


Figura 5.1: schema delle componenti chiave del programma Banjo.

Per decidere quale strategia di ricerca utilizzare si possono usare sia il "Greedy" che il "Simulated Annealing" che possono essere applicati attraverso l'appropriato "Searchers". Ognuno di questi algoritmi può essere usato con un approccio di tipo "random local move" o "all local moves" specificando l'appropppriato "Proposer". L'approccio di tipo "Greedy" usa un "greedy Decider", il quale accetta solo networks con il miglior score, mentre l'approccio "simulated annealing" accetta networks basati su di un "Decider" stocastico (che implementa l'algoritmo "Metropolis-Hastings"). La tabella sottostante mostra come si possono selezionare le differenti componenti per svolgere la ricerca:

Searcher Options	Dependent core objects	Choices for the dependent core objects	Explanation
SimAnneal (for simulated annealing search)	Proposer	**RandomLocalMove**	Addition, deletion, or reversal of an edge in the current network, selected at random.
		AllLocalMoves	All changes arising from a single addition, deletion, or reversal of an edge in the current network.
	Decider	**Metropolis**	A Metropolis-Hastings stochastic decision mechanism, where any network with a higher score is accepted, and any with a lower score is accepted with a probability based on a system parameter known as the "temperature".
Greedy (for "greedy" search)	Proposer	**RandomLocalMove**	Addition, deletion, or reversal of an edge in the current network, selected at random.
		AllLocalMoves	All changes arising from a single addition, deletion, or reversal of an edge in the current network.
	Decider	**Greedy**	A network is accepted if and only if its score is better than or equal to that of the current network. In the case of AllLocalMoves, the best of the local moves is considered.

6. RISULTATI SPERIMENTALI

Ipotesi di lavoro su C.elegans:

- Analisi preliminare della qualità dei chip
- Confronto dell'espressione genica tra il ceppo WT e KO di tutti i microarrays utilizzati
- Identificare i geni up/down regolati la cui espressione è influenzata dal gene ECE-1
- Chiarire le possibili interazioni fra i geni coinvolti

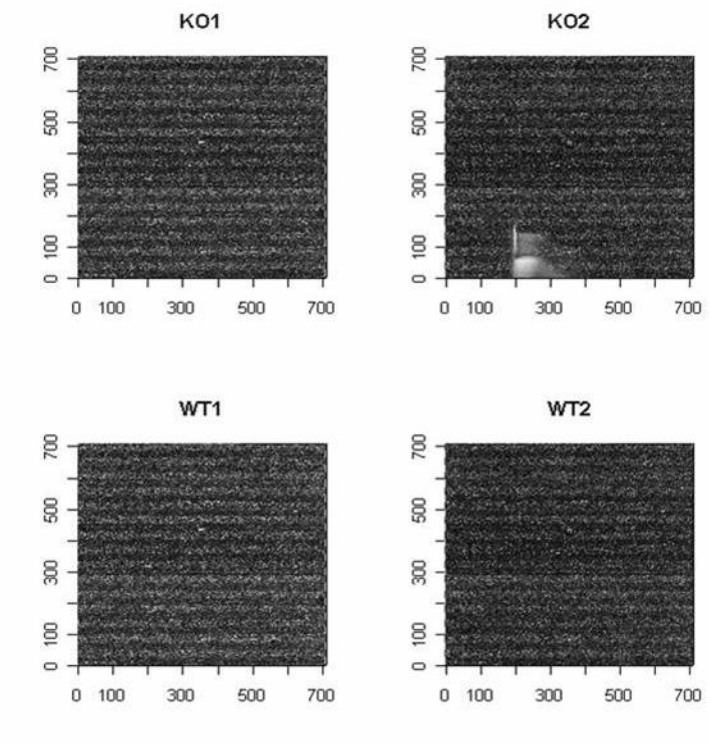
In via preliminare sono stati analizzati 4 microarrays Affymetrix dei quali due relativi al *C.elegans* nello stadio adulto e duo allo stadio larvale L3. Questo ha permesso di analizzare i microarrays individuando un cluster di geni up/down regolati probabilmente coinvolti nell'HSCR.

Lo scopo di questo procedimento consiste nell'identificare fra il ceppo adulto ed il ceppo larvale (L3) quali geni sono up o down regolati.

Dati analizzati:

wild type (wt) e knock out (ko) stadio larvale L3

wild type (wt) e knock out (ko) stadio adulto



Adulto Larvale L3

Figura 6.1: identificazione per via grafica della qualità dei microarrays.

Qualità

Dal grafico della qualità si nota come il chip relativo allo stadio larvale (knock-out) abbia un difetto probabilmente dovuto al procedimento di ibridazione, ma come vedremo in seguito non dannoso in quanto, siccome i probe set relativi ad un gene sono sparsi su tutto il chip, solo qualche probe sarà danneggiato. Da queste considerazioni si giunge alla conclusione che questo problema non influenza più di tanto il risultato finale perchè gli altri probe relativi allo stesso gene (figura 6.1) sono stati correttamente ibridizzati e sono sicuramente indicativi. Nonostante ciò è consigliato trattare con cura i dati relativi allo stadio larvale knock out (KO2).

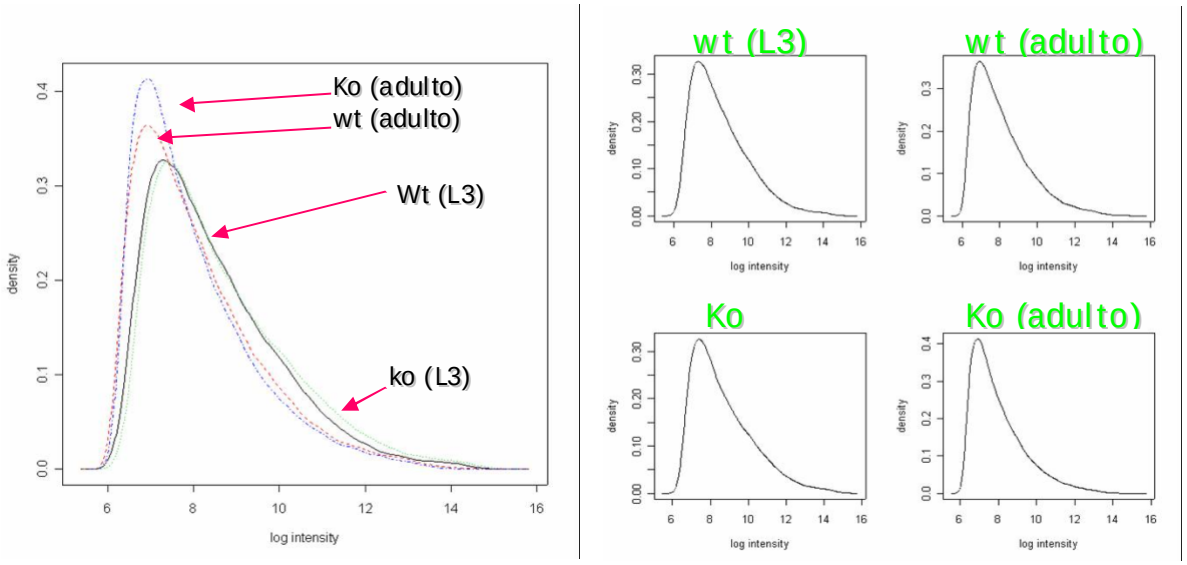


Figura 6.2: grafico relativo alla densità.

Anche dai grafici relativi alla densità in funzione del logaritmo dell'intensità si nota come gli andamenti delle curve siano identiche, il che è indice di una buona qualità sperimentale (figura 6.2).

I chip sono poi stati sottoposti al controllo sulla degradazione dell'RNA (figura 6.3) mediante il grafico "RNA digestion plot" il quale mette in evidenza il grado d'intensità medio in funzione della posizione 3'-5'. Questo fenomeno si verifica in quanto il processo di trascrizione dell'RNA mediamente non sempre ha la stessa efficacia. Quindi il processo parte da 3' per finire in 5' ed è per questo motivo che, mentre per tutti i segmenti trascritti la parte iniziale è sempre la stessa, quella finale non sempre coincide dipendendo appunto dall'efficienza del processo.

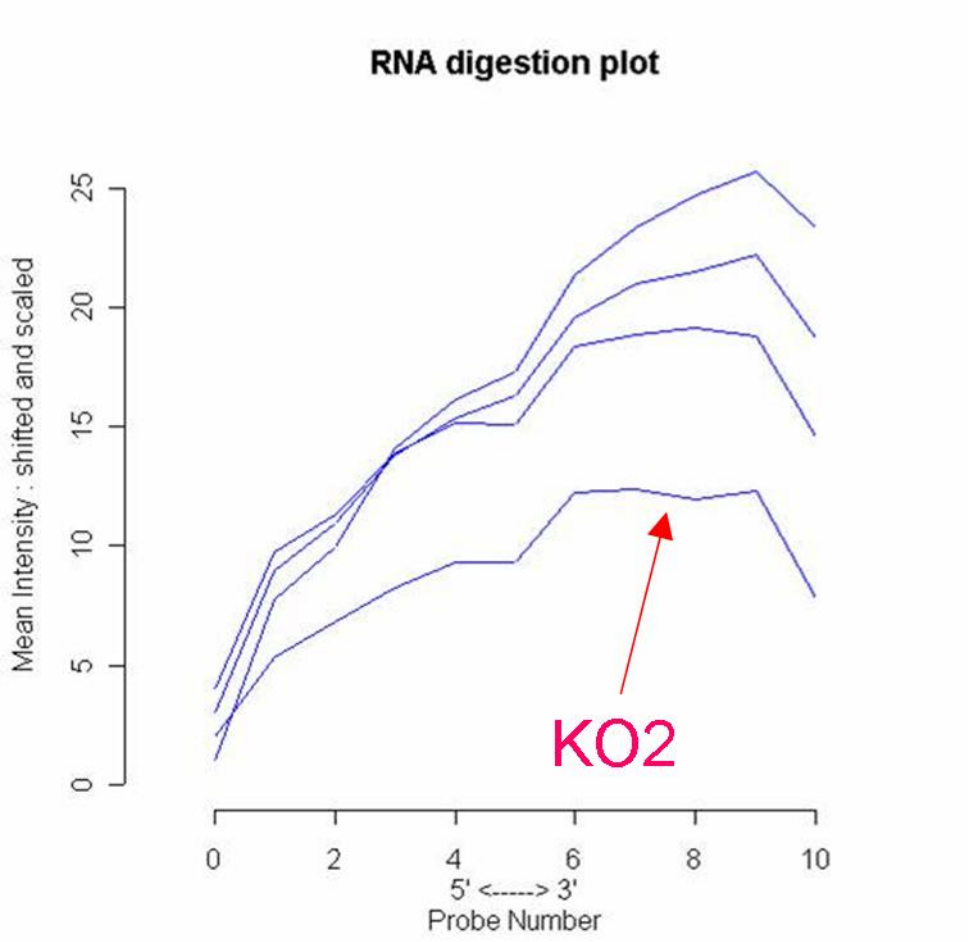


Figura 6.3: andamento della degradazione dell'RNA.

Dal grafico (figura 6.3) si nota che il chip danneggiato ha delle intensità più basse in 5' rispetto a 3' probabilmente dovute ad un segnale più degradato per ko (L3). Questo fenomeno è avvenuto in quanto essendoci dei probe sets danneggiati, questi non hanno risposto all'intensità del segnale dello scanner inficiando tutto l'andamento del grafico. Però come descritto prima il grafico descrive una degradazione parziale dell'RNA che comunque non dovrebbe pregiudicare l'esito dei risultati sperimentali.

Successivamente i nostri dati sono stati normalizzati per renderli paragonabili e per poter fare i confronti incrociati. Questo procedimento è di importanza vitale per estrarre informazioni corrette e consiste nello sfruttare alcuni probe sets che non contengono dati "geneticamente utili", in quanto l'intensità emessa serve solo per confronto. In questi probe sets l'intensità viene paragonata e normalizzata per avere un segnale uniforme fra i vari microarrays e solo dopo questo procedimento si può passare all'estrazione dei dati veri e propri (figura 6.4).

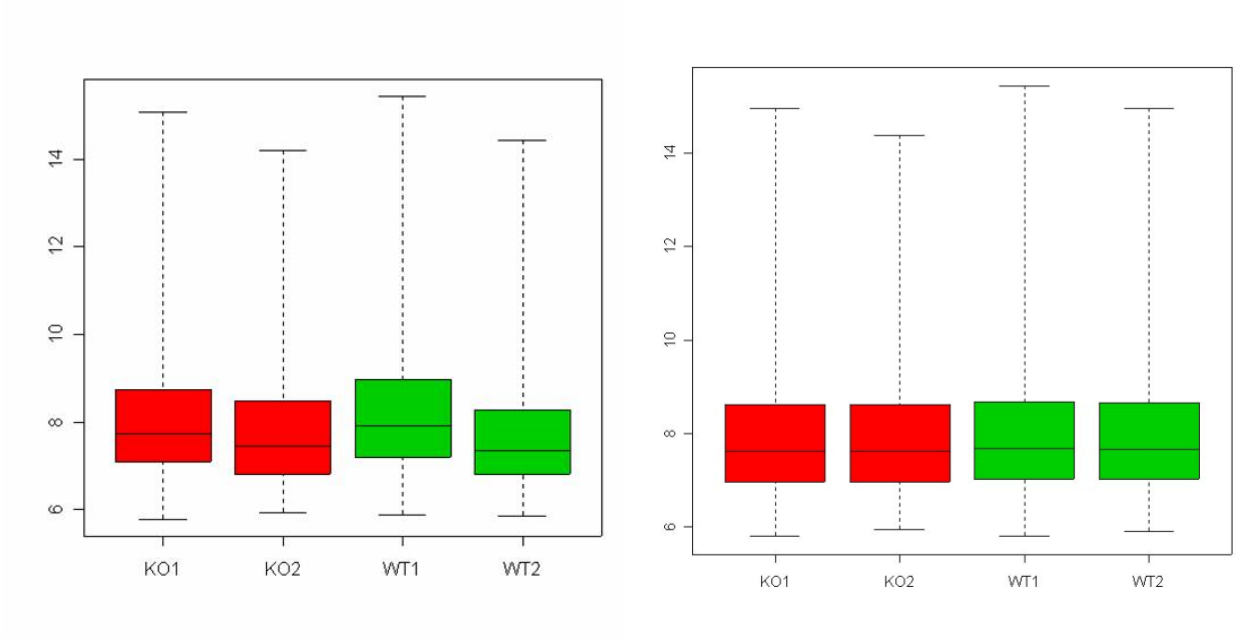


Figura 6.4: grafico denominato "box-plot" che individua le intensità pre e post normalizzazione.

Il grafico MVA (log ratio versus average log intensity) indica quanto si discosta il logaritmo dell'intensità rispetto a quello dell'intensità media, definendo quanto i segnali sono uniformi e quindi confrontabili. Questi dati sono stati presi prima e dopo la normalizzazione che mette in evidenza quanto il procedimento del confronto fra i probe sets di controllo sia fondamentale, non solo per individuare possibili errori nell'intensità del chip, ma soprattutto per poter successivamente confrontare i nostri dati sperimentali ed estrarre informazioni "statisticamente significative" (figura 6.5 e 6.6).

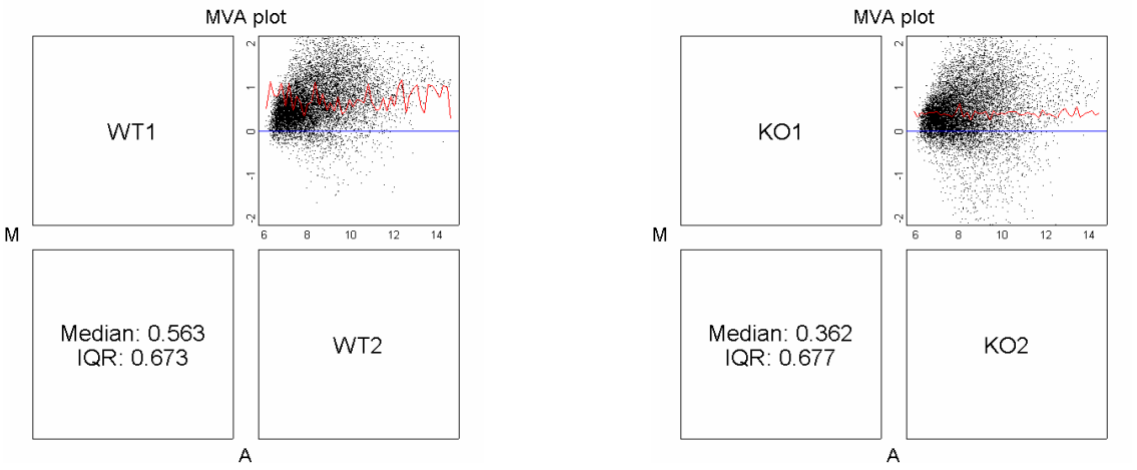


Figura 6.5: grafico MVA prima della normalizzazione che mette in evidenza quanto i dati dei singoli probe sets si discostano dalla media (linea rossa).

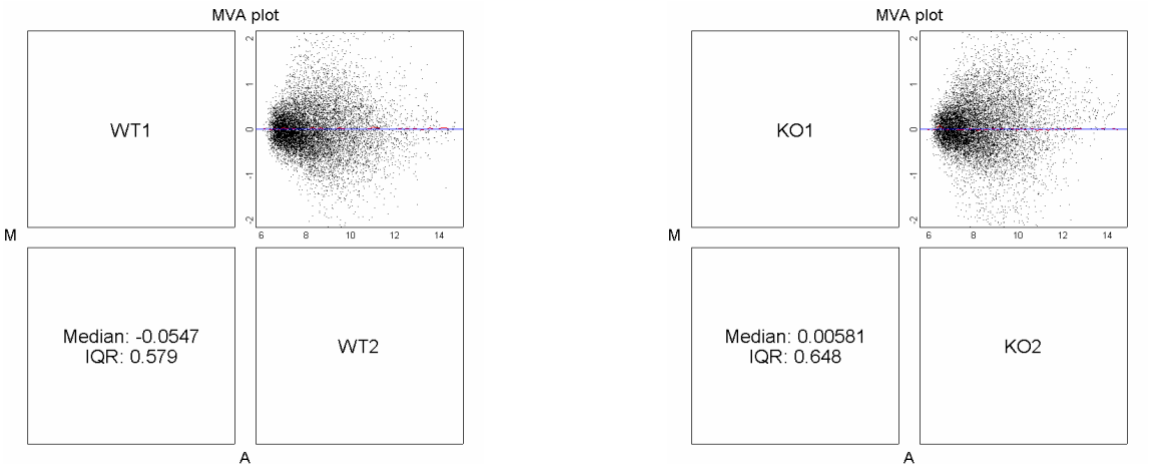


Figura 6.6: grafico MVA dopo la normalizzazione che mette in evidenza quanto i dati dei singoli probe sets siano in linea con il logaritmo dell'intensità media (linea rossa).

A questo punto i dati sono normalizzati e sono pronti per essere sottoposti alla procedura di analisi vera e propria che porterà ad avere dati informativi.

Dobbiamo definire, quindi, la soglia di falsi positivi, cosa molto importante in quanto, se la stringenza è troppo elevata si potrebbe verificare una perdita di geni significativi, mentre se è troppo bassa molti geni non significativi potrebbero essere inclusi nella selezione. Questo procedimento si basa sull'utilizzo del test di Wilcoxon (noto pure come test U di Mann-Whitney) che rappresenta uno dei più potenti test non parametrici per verificare, in presenza di valori ordinali provenienti da una distribuzione continua, se due campioni statistici provengono dalla stessa popolazione e che nel nostro caso ci permette di individuare i geni up/down regolati.

Per quanto riguarda il confronto nello stadio adulto sono stati individuati 8349 geni su 22625 probe sets differenzialmente espressi, mentre per quanto riguarda lo stadio L3 abbiamo individuato 879 geni differenzialmente espressi. Questo indica la presenza di un certo numero di geni che potrebbero essere "implicati" nella malattia dell'HSCR. Per poter però avere informazioni maggiormente significative è necessario restringere i risultati ad un solo probeset per gene, questo permette di evitare il problema della "over-rappresentazione" delle categorie funzionali fra geni regolati mediante l'analisi su un subset con un solo probeset per Swissprot ID. Ciò porta ad una riduzione drastica dei probesets da 22625 a 16920 ed in particolar modo da 8349 a 6025 differenzialmente regolati per lo stadio adulto e da 879 a 627 differenzialmente regolati per lo stadio L3.

L'analisi delle categorie funzionali ha permesso, grazie all'utilizzo della Gene Ontology (GO), di ottenere informazioni sul ruolo dei geni e delle proteine ed attraverso gli PFAM domains

(Protein Families Database of Alignments) di ricavare le famiglie proteiche ottenute mediante allineamento nei database.

NEP/WT (adulto)

Property Name	Property Size	Selection Prop	p-Value
Col_cuticle_N	109	63	8.27E-07
Proteasome	11	11	9.57E-06
Collagen	120	65	1.15E-05
Skp1_POZ	19	15	0.000113
Skp1	19	15	0.000113
PDZ	23	17	0.00016
FKBP_C	8	8	0.000224
PCI	10	9	0.000538
FARP	6	6	0.001837
FA_desaturase	8	7	0.003566
Motile_Sperm	68	35	0.00374
Aa_trans	10	8	0.004812
EMP24_GP25L	5	5	0.005252
Ligase_CoA	5	5	0.005252
Lipocalin	5	5	0.005252
MCM	5	5	0.005252
Pro_isomerase	12	9	0.005595
7tm_1	54	28	0.007949
LSM	7	6	0.009004
AAA	23	14	0.009985

Figura 6.7: categorie funzionali nel confronto ottenuto dallo stadio adulto.

NEP/WT (L3)

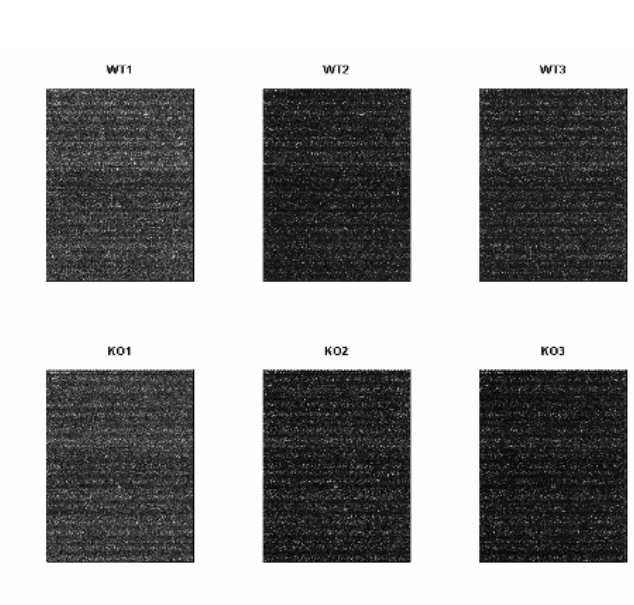
Property Name	Property Size	Selection Prop	p-Value
Transposase_1	15	7	5.35E-07
GLTT	7	5	1.53E-06
GETHR	4	4	2.07E-06
Collagen	120	14	0.000172
Hint	4	3	0.000213
Col_cuticle_N	109	13	0.000235
DB	12	4	0.000802
FAD-oxidase_C	2	2	0.001449
FATC	2	2	0.001449
FAT	2	2	0.001449
ACOX	7	3	0.001711
AMP-binding	16	4	0.002615
Acyl-CoA_dh	16	4	0.002615
ABC_tran	39	6	0.003279
Cyclin_C	3	2	0.004236
UDPGT	42	6	0.004793
FAD_binding_4	4	2	0.00826

Figura 6.8: categorie funzionali nel confronto ottenuto dallo stadio larvale L3.

Discussione dei risultati ottenuti

Dalle due tabelle (figura 6.7 e 6.8) risulta evidente una diminuzione dell'espressione di un subset di geni appartenenti alla famiglia dei collageni sia nell'adulto che nello stadio larvale. Considerando che la corretta migrazione dei precursori delle cellule della cresta neurale è influenzata sia da fattori genetici che ambientali (ad es. la matrice extracellulare), si può ipotizzare che una disregolazione dei geni, che codificano per i collageni, possa alterare la corretta migrazione delle cellule che daranno origine ai gangli nell'intestino.

Comunque questi esperimenti necessitano di repliche a causa della intrinseca variabilità del sistema e per questo sono stati analizzati altri 6 microarrays affimetrix per *C.elegans* in formazione di 3 wild type contro 3 knock out. L'RNA è stato estratto nel laboratorio di Friburgo del Professor Baumainster e successivamente l'ibridazione è avvenuta nel laboratorio di Modena del Professor Ferrari dove è presente anche lo scanner per i chip. Il dato grezzo è stato analizzato e sono stati ripetuti i test visti in precedenza.



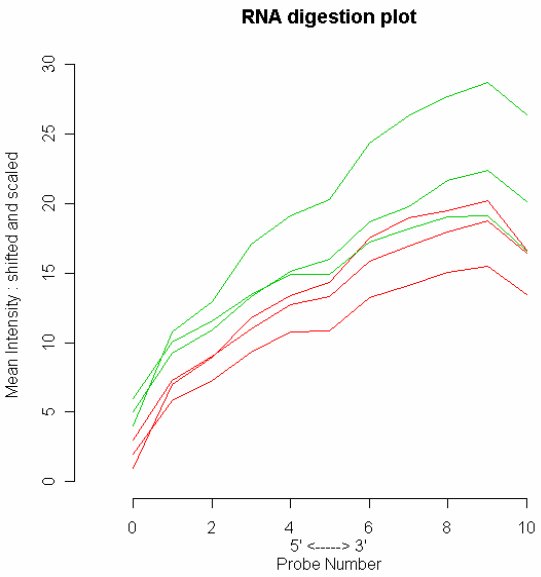


Figura 6.9: immagine della qualità del chip.

Figura 6.10: degradazione dell'RNA.

L'analisi della qualità ha messo in evidenza che i chip sono stati ibridizzati nel modo corretto senza sbavature o segni (figura 6.9) che potrebbero inficiare la qualità dei dati sperimentali. Inoltre il grafico relativo alla degradazione dell'RNA (figura 6.10) mostra come l'andamento dell'intensità media sia simile in tutti e sei i chip: altro dato che conferma la buona qualità dei nostri dati.

Successivamente si è provveduto a verificare l'andamento della normalizzazione che grazie al grafico di tipo boxplot (figura 6.11) conferma che ora i nostri dati sono confrontabili.

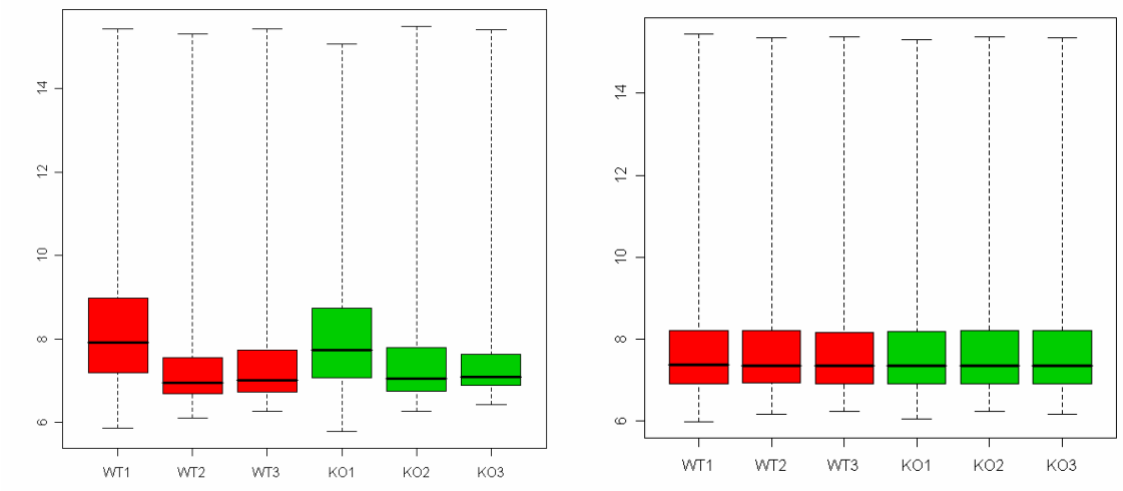


Figura 6.11: grafico di tipo "boxplot" prima e dopo la normalizzazione.

Sono state ripetute le analisi MVA prima e dopo la normalizzazione (figura .6.12 e 6.13). L'analisi ottenuta mediante MVA (log ratio versus average log intensity) anche in questo caso ha mostrato che i segnali sono uniformi e quindi confrontabili.

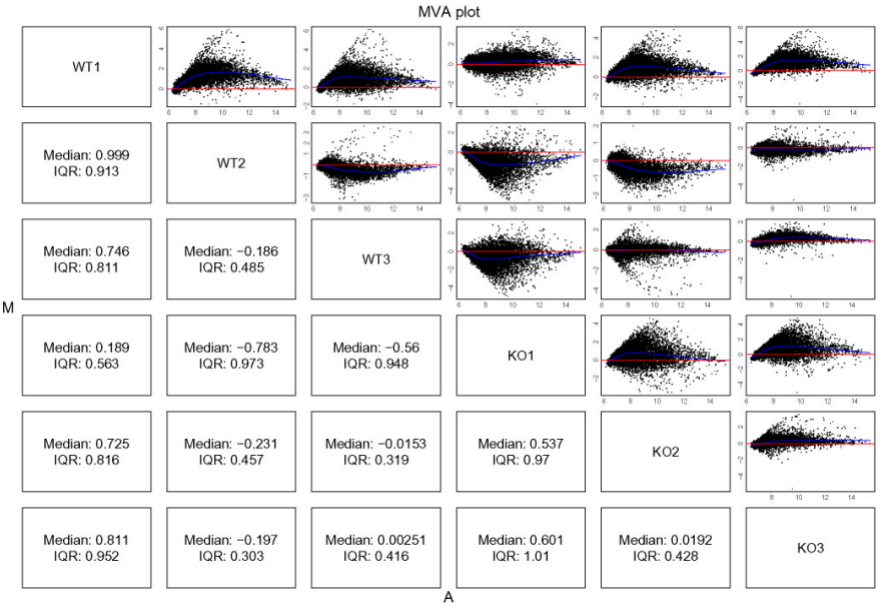


Figura 6.12: grafico MVA prima della normalizzazione che mette in evidenza quanto i dati dei singoli probe sets si discostano dalla media (linea rossa).

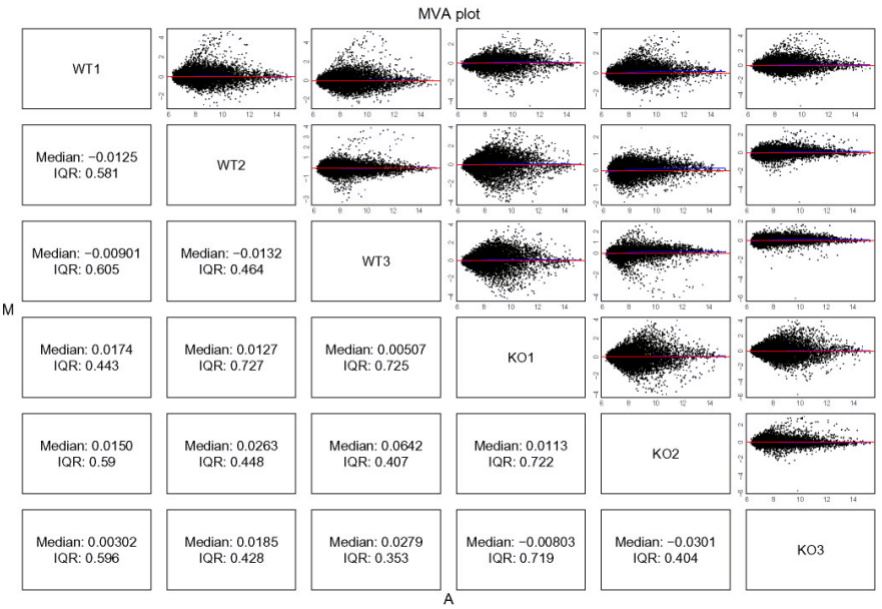


Figura 6.13: grafico MVA dopo la normalizzazione che mette in evidenza quanto i dati dei singoli probe sets siano in linea con il logaritmo dell'intensità media (linea rossa).

L'analisi dei dati relativi all'esperimento dello stadio adulto ha permesso di evidenziare 6 geni upregolati (tabella 6.1) e 7 geni downregolati (tabella 6.2).

GENI UPREGOLATI (tabella 6.1)

	Probe set ID	Gene ID	Gene title	GO Biological Process Description	GO Molecular Function Description	Human homolog	Position
1	172741_x_at (t)	F41F3.3	cuticlin			KRT10 NM_000421	17q21
2	1777I 47_at (t)	F58G6.2	Serpentine Receptor	copper ion transport	G-protein coupled receptor activity	SLC31A1: solute carrier family 31 (copper transporters), member 1 NM_001859	9q31-q32
3	178262_at (t)	F01D5.5				Metallotionein 1A; MT1A NM_005949	16q13
4	182091_s_at (t)	T04A11.4		biosynthesis	catalytic activity	MAWD binding protein NM_022129	10pter-q25.3
5	183028_s_at (t)	Y22F5A.5	Lys-2		lysozymal function immune function		
6	191541_at (t)	C54D10.1	glutathione S-transferase	metabolism	glutathione transferase activity	C6orf168 NM_032511	6q16.3

GENI DOWNREGOLATI (tabella 6.2)

	Probe set ID	Gene ID	Gene title	GO Biological Process Description	GO Molecular Function Description	Human homolog	Position
1	176872_at	F59D8.E	Vitellogenin	Lipid transport	Lipid transporter activity	APOB-100 NM_000384	2p24-p23
2	173341_at	Y48G8AL.11	haf-6	transport	nucleotide binding, ATP-binding cassette (ABC) transporter activity ATP binding	NM_007188	7q36
3	186611_at	Y45G12B.3		electron transport	oxidoreductase activity	C14orf160 NM_024884	14q21.3
4	173701_s_at	T03D3.5				Hypothetical protein TR:Q96HF8 - NEFH	22q12
5	177899_at	T03D8.7		electron transport	electron transporter activity	AB081338	17q12-21
6	179659_at	R06B9.3		signal transduction, sensory perception	Thioredoxin binding	NM_015683	19p13.11
7	189163_at	C36A4.2	cytochrome P450	electron transport	monooxygenase activity	NM_017460	7q21.1

Tra i geni differenzialmente espressi la nostra attenzione si è soffermata sul gene che codifica per la vitellogenina 3 (vit-3). Questo gene è risultato essere un buon candidato per il fatto di essere espresso in maniera sesso-specifica, stadio-specifica e tessuto-specifica. Infatti vit-3 è espresso solo negli ermafroditi (nell'uomo la probabilità di sviluppare la malattia nei maschi e nelle femmine è in rapporto di 4:1); è espresso dal tardo stadio larvale L4 fino allo stadio adulto (la malattia di HSCR è causata da difetti dello sviluppo delle cellule della cresta neurale); ed infine vit-3 è espresso solamente nell'intestino di *C.elegans* (come già ampiamente descritto

nell'introduzione: la malattia di HSCR è caratterizzata dall'aganglionosi di tratti di lunghezza variabile nell'intestino).

Il gene vit-3 appartiene alla famiglia delle vitellogenine, proteine strutturali fondamentali per lo sviluppo dell'embrione. Queste proteine inoltre sono coinvolte nel trasporto del colesterolo negli oociti ed hanno anche un ruolo importante nello sviluppo degli spermatozoi (informazione molto interessante considerato il comprovato ruolo di RET nella spermatogenesi (Jain S, 2004)).

La downregolazione di vit-3 allo stadio adulto è stata poi verificata mediante la tecnica di PCR quantitativa.

Utilizzando il data base WormBase di *C.elegans* (www.wormbase.org) abbiamo identificato la controparte umana del gene vit-3: l'apolipoproteina B-100. Questa proteina è codificata dal gene APOB e costituisce la maggior componente dei chilomicroni, LDL e VLDL, adibiti al trasporto del colesterolo e dei trigliceridi nel sangue.

Esperimenti condotti sui topi mostrano che il topo apob-/- presenta mortalità a livello embrionale, mentre il topo apob+/- mostra l'incompleta chiusura del tubo neurale ed un'aumentata predisposizione alla sterilità dei maschi. (Huang LS,1995; Huang LS,1996).

Il problema che ci si pone di fronte rappresenta quello di capire come questi geni possano essere implicati nel fenotipo di HSCR, come possano contribuire alla patogenesi della malattia e soprattutto come possano far parte di un possibile pathway coinvolto nell'HSCR.

Ciò che ci proponiamo di fare, consiste nell'individuare come una serie di geni possano essere connessi e soprattutto capire quanti di questi siano implicati più o meno direttamente per formare un "network" coinvolto, in qualche modo, nella patogenesi della malattia.

L'idea di come poter risolvere questo problema, ossia di poter identificare le possibili interazioni fra i geni coinvolti, è nata grazie ad una collaborazione con l'Ing. Diego di Bernardo del TIGEM di Napoli che grazie all'utilizzo della System Biology ed in particolar modo degli algoritmi Bayesiani ha reso possibile l'identificazione delle possibili connessioni fra questi geni e le relazioni che li legano, basandoci sui dati di microarrays da noi svolti e da quelli liberamente scaricabili dai data base presenti in rete.

La nostra strategia operativa può essere schematizzata nel seguente modo (figura 6.16):

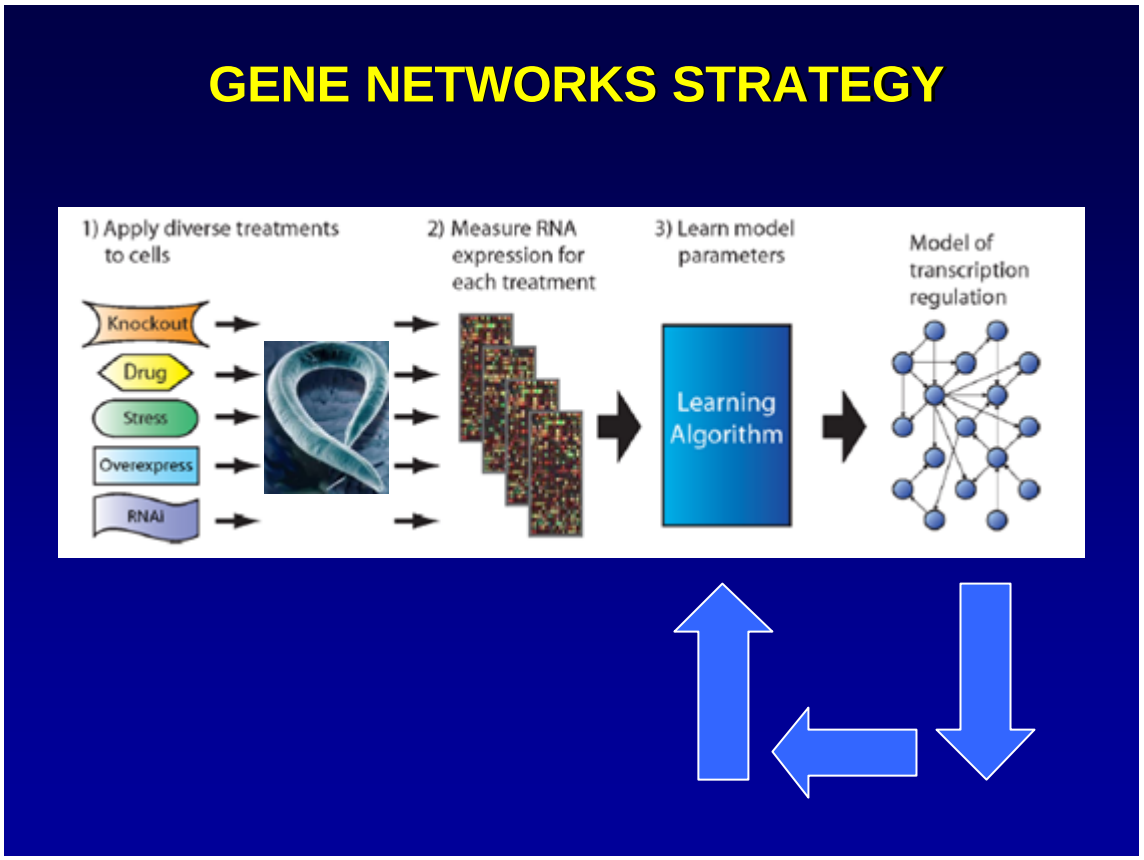


Figura 6.16: strategia utilizzata per identificare le relazioni che intercorrono fra i geni individuati.

Sul nostro organismo modello (*C.elegans* in questo caso) vengono applicate una serie di perturbazioni di vario genere e successivamente si misura l'espressione ad ogni trattamento. I dati così ottenuti sono sottoposti ad un algoritmo detto "LEARNING ALGORITHM" il quale non fa altro che analizzare i dati sperimentali e costruire un modello teorico del network. Questo procedimento viene poi ripetuto varie volte in modo da ottenere quello statisticamente più significativo.

I programmi di gene networks per poter funzionare correttamente o meglio per poter dare dei risultati "statisticamente significativi" hanno però bisogno di un numero molto elevato di dati sperimentali e quelli in nostro possesso non erano sufficienti quindi, per ovviare a questo problema abbiamo utilizzato quelli liberamente scaricabili nelle banche dati come quella di Stanford (figura 6.17) che contiene qualche centinaia di esperimenti su *C.elegans* che aumentano di giorno in giorno in quanto non rappresenta un data base statico perché viene aggiornato in continuazione da chi decide di inserire i propri dati sperimentali.

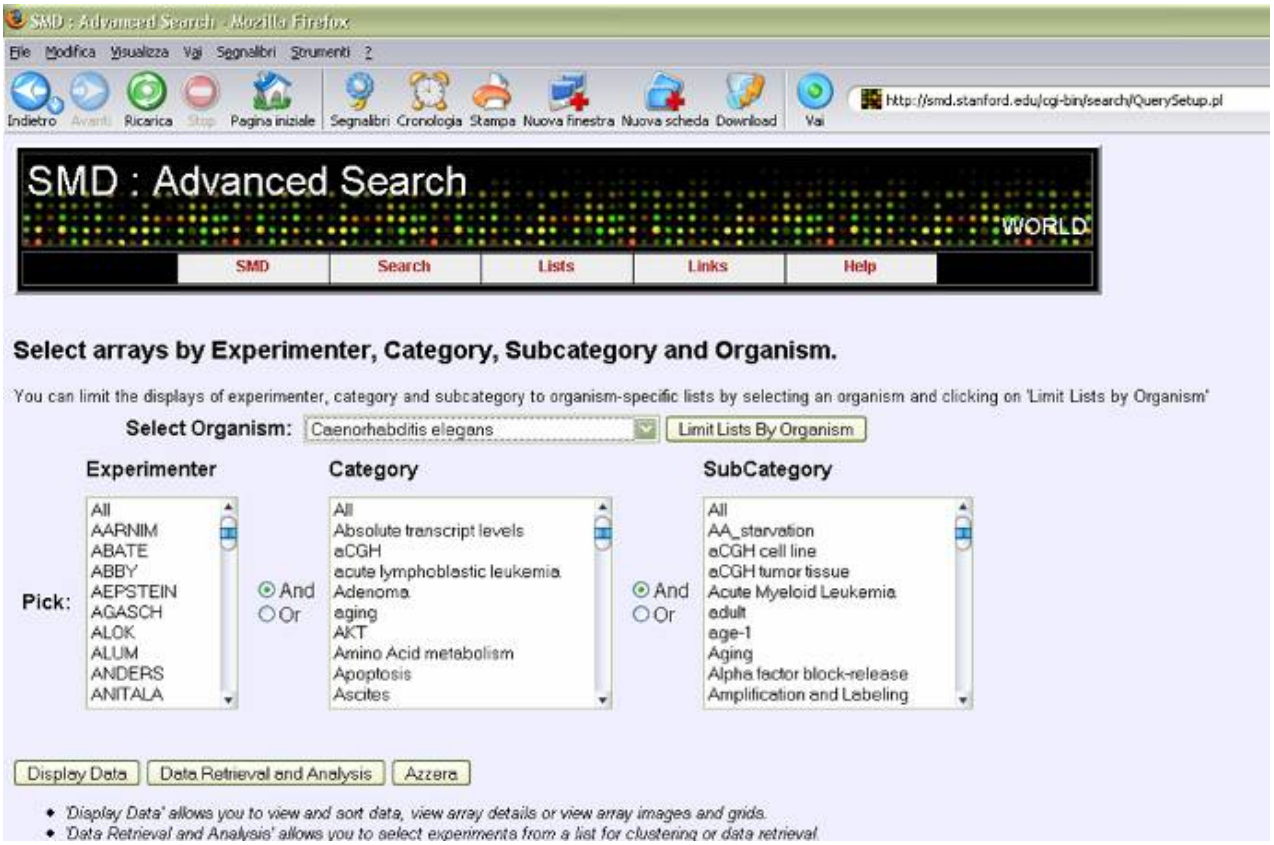


Figura 6.17: interfaccia utente della schermata del database di Stanford in cui agevolmente si può selezionare l'organismo modello di cui scaricarsi i dati di espressione e successivamente si può anche scegliere l'esperimento in base allo sperimentatore che lo ha svolto, la categoria di appartenenza e la sottocategoria.

Network identificato e discussione dei risultati

Utilizzando questi dati ed inserendoli in un programma di gene-networks abbiamo ottenuto per *C.elegans* il seguente grafo (figura 6.18).

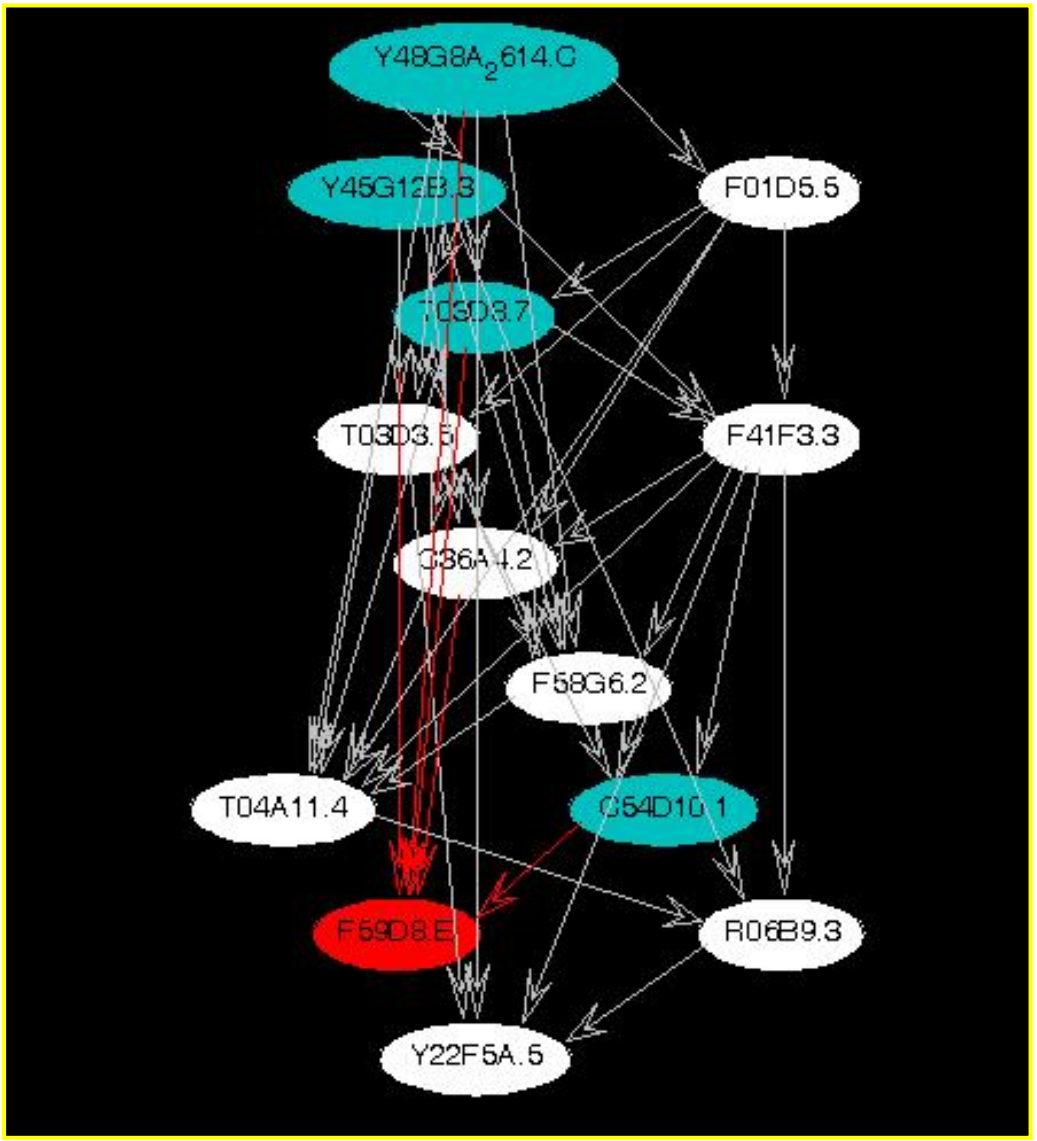


Figura 6.18: grafo che rappresenta il gene network dell'organismo modello ottenuto basandoci sui dati sperimentali scaricati dal database di Stanford.

Dove si possono notare in rosso il gene della vitellogenina ed in blu i geni che il programma ritiene aver una maggior probabilità di essere connessi.

I vari geni connessi hanno svariate funzioni quali il trasporto di elettroni attraverso la membrana cellulare o il trasporto dei lipidi nonché l'attività glutatione-transferasica ed il legame di nucleotidi (tablella.6.3).

Gene ID	Gene Title	GO Biological Process	GO Molecular Function	Human Hmolog	Position
C54D10.1	glutathione S-transferase	metabolism	glutathione transferase activity	C6orf168 NM_032511	6q16.3
F59D8.E	Vitellogenin	Lipid transport	Lipid transporter activity	APOB-100 NM_000384	2p24-p23
Y48G8AL.11	haf-6	transport	nucleotide binding, ATP-binding cassette (ABC) transporter activity ATP binding	NM_007188	7q36
Y45G12B.3		electron transport	oxidoreductase activity	C14orf160 NM_024884	14q21.3
T03D8.7		electron transport	electron transporter activity	AB081338	17q12-21

Tablella 6.3: geni reputati interagire nel modello sperimentale di *C.elegans*.

Il modello di *C.elegans* ha rappresentato per il nostro progetto un modello semplice e facile da studiare, ma piuttosto riduttivo in quanto presenta molti inconvenienti quali l'assenza di un numero significativo di geni ortologi nell'uomo, pochi datasets disponibili in rete nonché per l'assenza dei geni maggiormente coinvolti (RET). Di conseguenza abbiamo deciso di passare al modello umano ed abbandonare il nematode che comunque ha rappresentato la nostra base di partenza per poter comprendere la strategia da percorrere al fine di ottenere dei risultati informativi. È proprio grazie all'esperienza accumulata studiando il modello sperimentale di *C.elegans* che abbiamo individuato la tecnica da applicare a organismi più complessi come l'uomo.

Ipotesi di lavoro sull'uomo

- Individuazione del database da cui scaricare i dati
- Utilizzare i dati dell'esperimento di "split-ubiquitin membrane yeast two hybrid" per individuare i geni da analizzare in base ai dati scaricati dal database
- Identificare i geni che possono interagire con RET
- Confermare sperimentalmente i nodi del network appena ottenuto
- Chiarire le possibili interazioni fra i geni coinvolti

Abbiamo quindi ripetuto le stesse operazioni sull'uomo (figura 6.19):

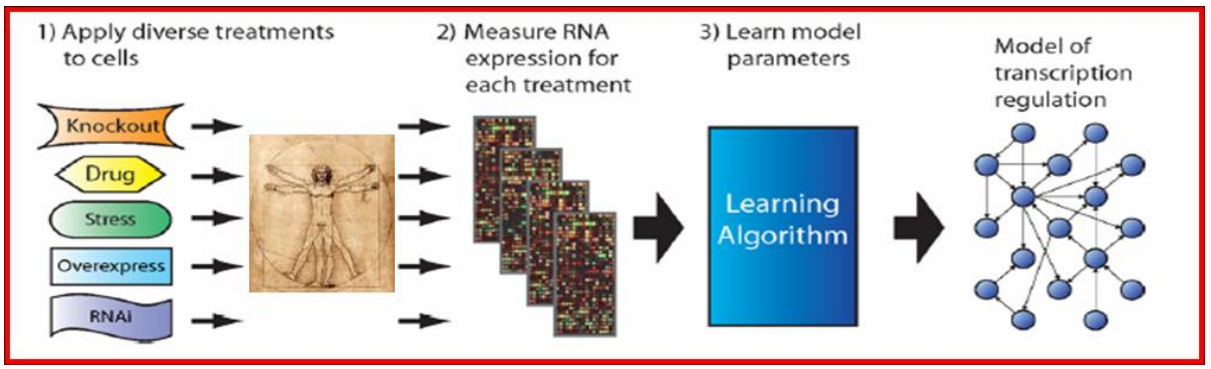


Figura 6.19: strategia utilizzata nel modello umano per identificare le relazioni che intercorrono fra i geni individuati.

e per poter disporre di un numero sufficiente di dati sperimentali abbiamo fatto ricorso ai dati presenti nel database di GEO (figura 6.20) che rappresenta uno strumento formidabile dell'NCBI.

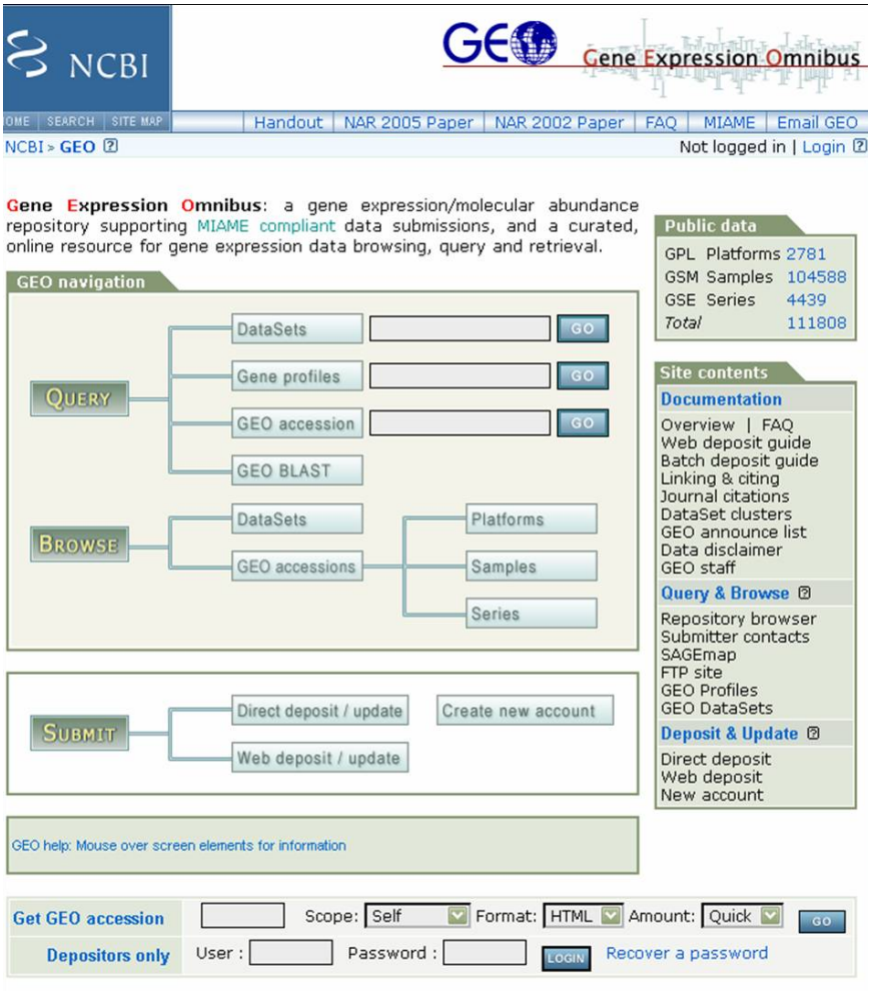


Figura 6.20: home page di GEO che permette all'utente di individuare velocemente gli esperimenti presenti in rete mediante ricerche incrociate per "data sets", "gene profiles" e per "GEO accession number".

Dati analizzati

La nostra attenzione si è focalizzata sull'esperimento numero GSE2841 riguardante l'expression profiling del feocromocitoma. Questa neoplasia ha origine dalle cellule della cresta neurale da mutazioni sia sporadiche che famigliari in almeno sei geni indipendenti quali RET, VHL, NF1 e le subunità B, C, D della deidrogenasi (SDH). Il feocromocitoma è una rara forma di tumore in cui le cellule neoplastiche derivano da particolari cellule chiamate *cellule cromaffini.* La maggior parte dei feocromocitomi origina nella ghiandola surrenalica (midollare del surrene) dove sono localizzate per la maggior parte le cellule cromaffini. Ci sono due ghiandole surrenali, localizzate sopra i reni nel retro dell'addome superiore ed hanno la funzione di produrre importanti ormoni che aiutano l'organismo a funzionare in modo corretto.

Normalmente il feocromocitoma origina da una sola ghiandola surrenalica; a volte può avere origine anche in altre parti del corpo, come la zona che circonda il cuore o la vescica.

La maggior parte dei tumori che originano dalle cellule cromaffini non sono maligni e non si diffondono ad altre parti dell'organismo: si parla in questo caso di *tumori benigni.*

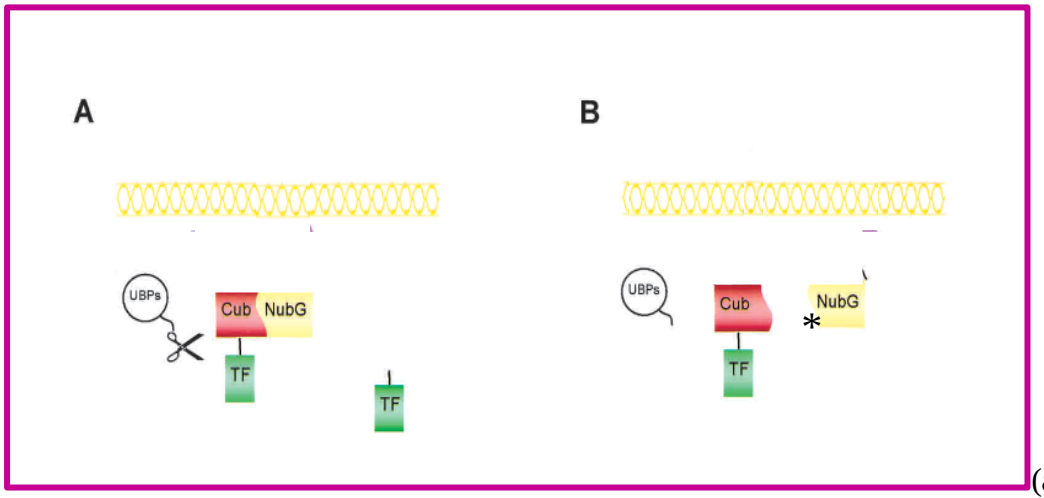
I feocromocitomi spesso fanno sì che le ghiandole surrenali producano troppi ormoni chiamati *catecolamine.* L'eccesso di catecolamine provoca pressione sanguigna elevata (ipertensione), che può a sua volta determinare mal di testa, sudorazione, battito cardiaco affannoso, dolore al torace e senso di ansia.

Il feocromocitoma in alcuni casi si inquadra all'interno di una sindrome particolare definita MEN (multiple endocrine neoplasia), caratterizzata dalla presenza di altri tumori (per esempio tumore della tiroide) ed altri problemi ormonali.

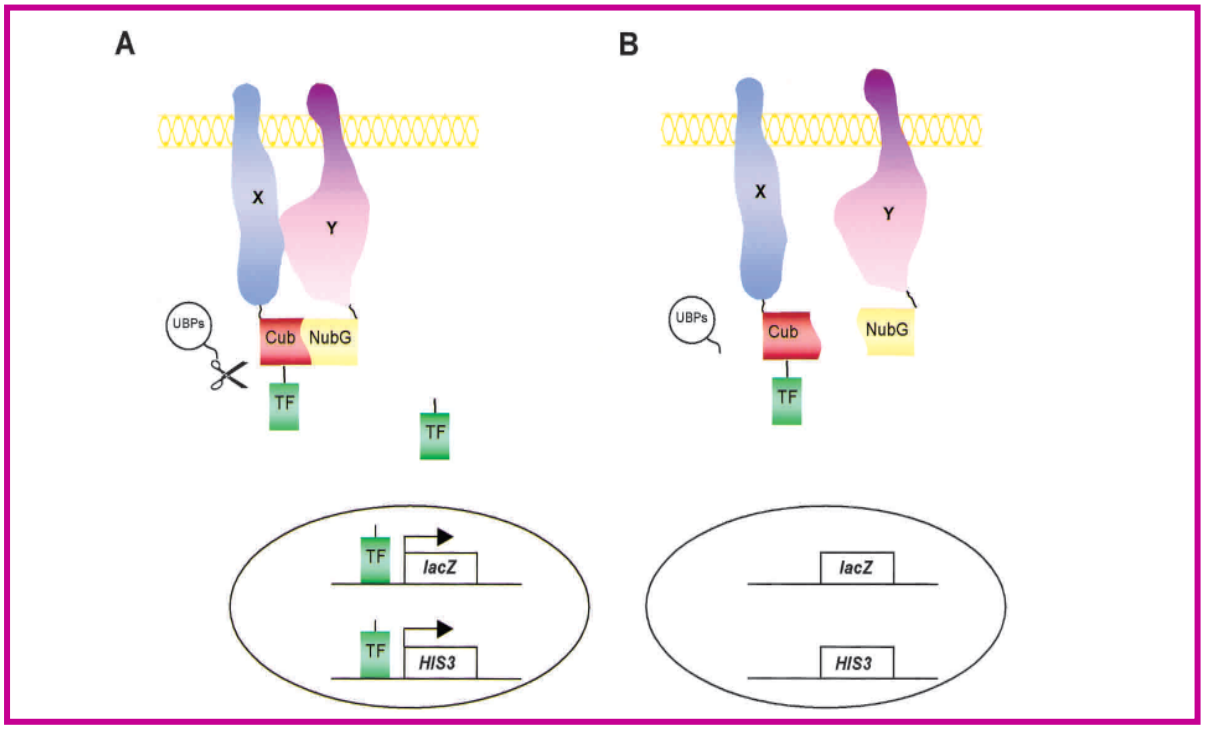
Le proteine codificate dai geni coinvolti nella malattia sono coinvolte in molteplici funzioni. E' quindi stata svolta un'analisi comparativa del profilo di espressione dei 76 microarrays che costituiscono l'esperimento presente nel data base.

A questo punto i dati così ottenuti sono stati inseriti nel nostro programma di gene networks (banjo) e mediante l'algoritmo di learning sono stati elaborati al fine di ottenere un modello di network da valutare e confermare sperimentalmente.

Per identificare i geni da inserire nel programma di gene network e dei quali vogliamo identificare le possibili connessioni abbiamo sfruttato i risultati sperimentali ottenuti nel nostro laboratorio di genetica medica mediante l'esperimento di "Split-ubiquitin membrane yeast two hybrid" per identificare proteine che legano il recettore tirosin chinasico RET.



(a)



(b)

Figure 6.21a 6.21b: rappresentazione grafica del meccanismo alla base dello ”split-ubiquitin membrane yeast two hybrid” (MbYTH).

Allo scopo di identificare nuove proteine che interagiscono con l’isoforma lunga di RET (RET-51), è stato impiegato un metodo modificato del sistema del doppio ibrido, lo ”split-ubiquitin membrane yeast two hybrid” (MbYTH) (figura 6.21a e 6.21b).

Questo metodo sfrutta la capacità dell’ubiquitina, una piccola proteina citoplasmatica, di legare le proteine e indirizzarle rapidamente verso la degradazione mediata dal sistema proteosoma ubiquitina dipendente. In particolare le proteasi ubiquitina specifiche (UBPs) tagliano l’ubiquitina dalla proteina target permettendo di riciclare l’ubiquitina nel citoplasma. Quando il frammento C-

terminale (Cub) viene espresso in lievito come prodotto di fusione con una proteina reporter, quest'ultima verrà riconosciuta e tagliata dalle proteasi ubiquitina specifiche solo se all'interno della stessa cellula viene espressa anche la porzione N-terminale wild type (Nub). Nel sistema dello "split-ubiquitin membrane yeast two hybrid" il frammento N-terminale è mutagenizzato (NubG) in modo da impedirne la spontanea riassociazione con la porzione C-terminale. Se due proteine che interagiscono tra loro vengono fuse ai due frammenti NubG e Cub, l'interazione tra le due proteine porterà alla parziale riassociazione dei due frammenti e al conseguente rilascio della proteina reporter da parte delle proteasi ubiquitina specifiche. Quando le due proteine X e Y interagiscono, i due frammenti NubG e Cub si riassociano e le proteasi ubiquitina specifiche rilasciano la proteina reporter nel citoplasma. La proteina reporter solitamente è un fattore di trascrizione che, una volta veicolato nel nucleo, attiva la trascrizione di un gene reporter, come LacZ o His3.

In questo esperimento è stata utilizzata come bait la porzione citoplasmatica di RET-51 contro una libreria di cDNA relativa al cervello umano. Dall'esperimento sono risultate interagire con l'isoforma lunga di RET dieci proteine.

L'espressione dei geni codificanti per queste dieci proteine è stata analizzata in due linee di neuroblastoma umano, SHSY-5Y e IMR32 e in linee di origine non neurale, che comprendono la linea di carcinoma cervicale HeLa e di rene embrionale HEK 293.

La presenza di tutti i trascritti attesi è stata riscontrata solo in una linea cellulare di neuroblastoma, SHSY-5Y, e in una linea cellulare di origine non neurale, HeLa.

I cDNA full-lenght codificanti per queste proteine sono stati clonati dalla linea SHSY-5Y e poi inseriti nel vettore pcDNA 3.1/myc-His che ne permette l'espressione in cellule di mammifero e che, grazie alla presenza di un tag C-terminale codificante per un epitopo di myc, permette di testare l'espressione della proteina ricombinante.

I costrutti così generati sono stati trasfettati nella linea cellulare umana di rene embrionale umano HEK293 che non esprime RET.

Grazie a questo esperimento sono stati identificati dieci geni che potrebbero interagire (tabella 6.4):

Protein ID	Funzione	Pos
1) **ARF-1** (ADP-ribosylation factor 1)	Superfamiglia delle G-protein. Coinvolto nel trasporto vescicolare nel Golgi (sia vescicole rivestite di clatrina sia COPI)	1q42-13
2) **BAB70795**	/	3p21
3) **NR2E1** (nuclear receptor subfamily 2, group E, member 1)	Recettore orfano che lega il DNA come monomero agli elementi di risposta agli ormoni (HRE). Potrebbe essere coinvolto nella regolazione dello sviluppo cerebrale e della retina (By similarity-GO)	6q21
4) **AIP** (aryl hydrocarbon receptor interacting protein)	Immunophillin-like protein Componente del complesso recettoriale AHR-hsp90-AIP	11q13-3
5) **KIAA0363**		7p13
6) **C15orf21** D-PCa-2 protein isoform b	Sconosciuta. Regione omologa a high-mobility-group nucleosomal bindig protein 2. Due NLS. Esclusivamente espressa nella prostata.	15q21-1
7) **MATR3** (matrin 3)	Componente della matrice della membrana interna nucleare. Interagisce con la PKC probailmente trslocandola nel nucleo. Coinvolta nella regolazione della trascrizione, dello splicing e dell'editing dell'mRNA.	5q31-2
8) **IRF-2BP2B** (interferon regulatory factor-2 binding protein 2B)	Co-repressore trascrizionale che interagisce con IRF-2 regolando negativamente molti geni responsivi all'IFN. Metabolismo dei composti aromatici (GO)	1q42-1-43
9) **FKBP4** (FK506-binding protein 52)	Immunofillina Componente del complesso recettoriale GR-hsp90-FKBP52	12p13-33
10) **GSTP1** (Glutathione transferase)	Catalizza attacco del glutatione ridotto (GSH) a composti apolari	11q13-qter

Tabella 6.4: tabella indicante le dieci proteine ed i relativi geni che le codificano individuate mediante l'esperimento dello "split-ubiquitin membrane yeast two hybrid" (MbYTH).

Individuazione del network "statisticamente più significativo"

Partendo da questa tabella di dieci geni abbiamo cercato di costruire il nostro network (figura 6.22) utilizzando i dati sperimentali ottenuti dai microarrays scaricati dal database di GEO.

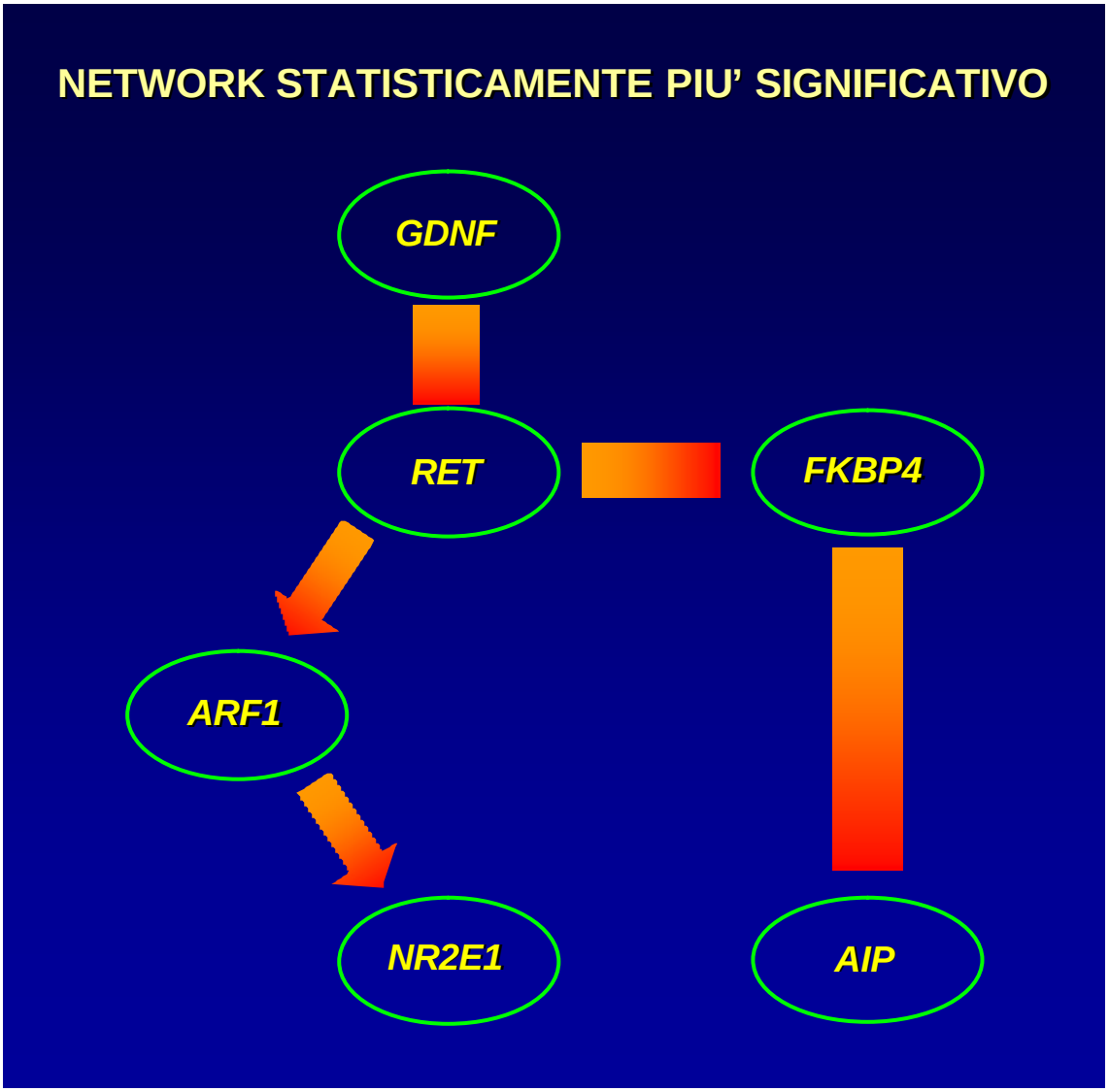


Figura 6.22: network ottenuto con il programma banjo, analizzando 10 variabili (geni) su 75 esperimenti osservati. Il programma ha individuato questo network come quello "statisticamente più significativo" fra i 13197000 networks possibili valutati. Si può notare come il programma includa nel network solo 4 delle 10 variabili osservate, questo indica che l'algoritmo ritiene possibili solo queste 4 interazioni con il gene RET e di conseguenza non include gli altri 6 geni in base ai dati di espressione scaricati dal database di GEO.

Da questo network possiamo subito vedere che dei dieci geni che codificano per le 10 proteine identificate con lo Split Ubiquitin solo quattro geni (figura 6.23) vengono inclusi nel network e quindi ritenuti di poter interagire con il gene RET.

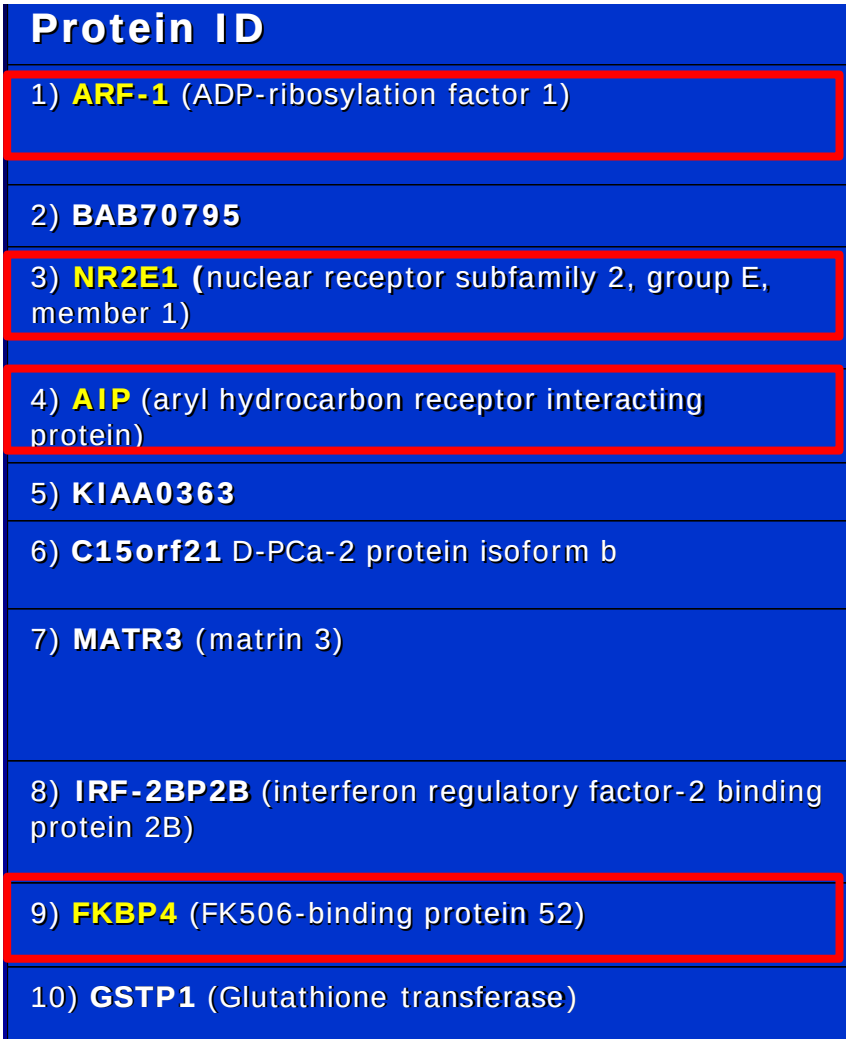


Figura 6.23: elenco dei 4 geni inclusi nel network dal programma ed i restanti 6 geni che sono stati scartati perché ritenuti non interagenti.

Per poter validare le interazioni con il recettore tirosin-chinasico RET ottenute basandoci sugli algoritmi di gene-networks abbiamo utilizzato metodi di coimmunoprecipitazione e colocalizzazione. Questi si basano sull'utilizzo della strategia di clonaggio del cDNA full lenght delle proteine in analisi in un vettore di espressione con un tag da utilizzare in co-trasfezioni insieme ad un costrutto contenente RET51 in una linea cellulare RET negativa.

Mediante questa tecnica abbiamo quindi verificato che RET interagisce in vitro con i 4 geni (AIP, ARF1, NR2E1 e FKBP4). Queste interazioni sono state confermate (figura 6.24), il che porta

alla conclusione che i geni codificanti per queste proteine sono ottimi candidati come geni modificatori nelle malattie in cui il proto-oncogene RET è coinvolto.

Inoltre stiamo effettuando uno screening su un ampia casistica di pazienti Hirschsprung ricercando l'eventuale presenza di mutazioni causative nei geni codificanti per queste proteine.

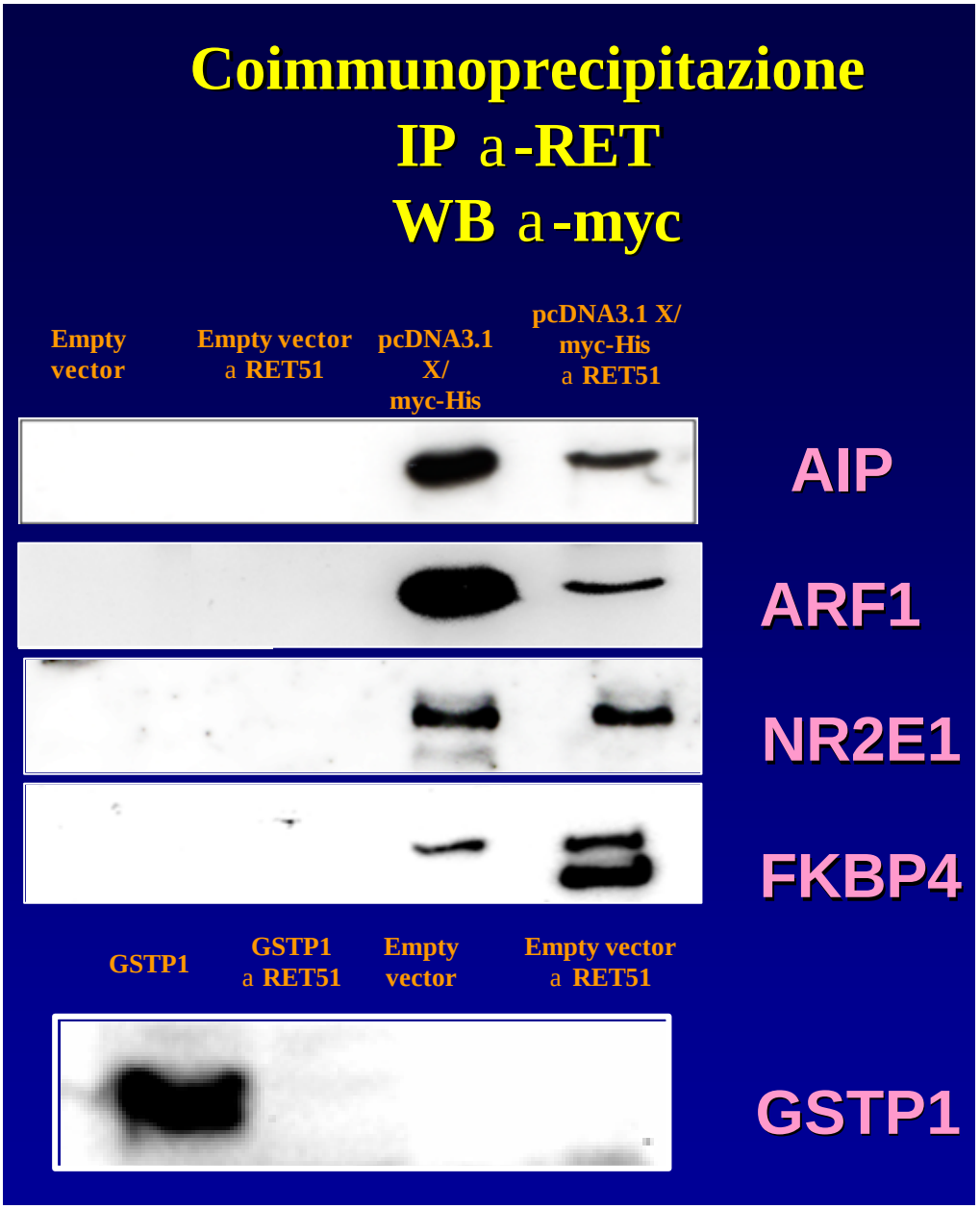


Figura 6.24: risultato dell'esperimento di coimmunoprecipitazione e colocalizzazione in cui si nota come i 4 geni (AIP, ARF1, NR2E1 e FKBP4), che il programma ha incluso nel network, effettivamente interagiscano con RET mentre come controprova il gene GSTP1 che non è stato incluso nel network non interagisce neppure nell'esperimento condotto in laboratorio.

Discussione dei risultati ottenuti sull'uomo

Questi geni codificano per proteine molto ben caratterizzate.

Il recettore NR2E1 (conosciuto come Mtll o Tlx) è un membro della famiglia dei recettori nucleari orfani (figura 6.25), recettori per i quali non è ad oggi conosciuto il ligando (figura 6.26) (Roy K., 2004).

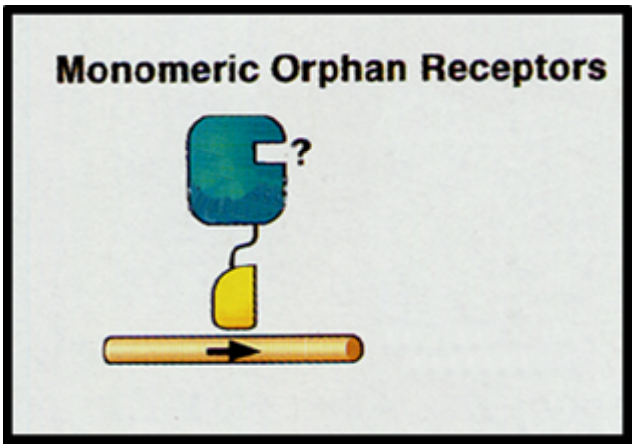


Figura 6.25: recettore nucleare orfano monomerico

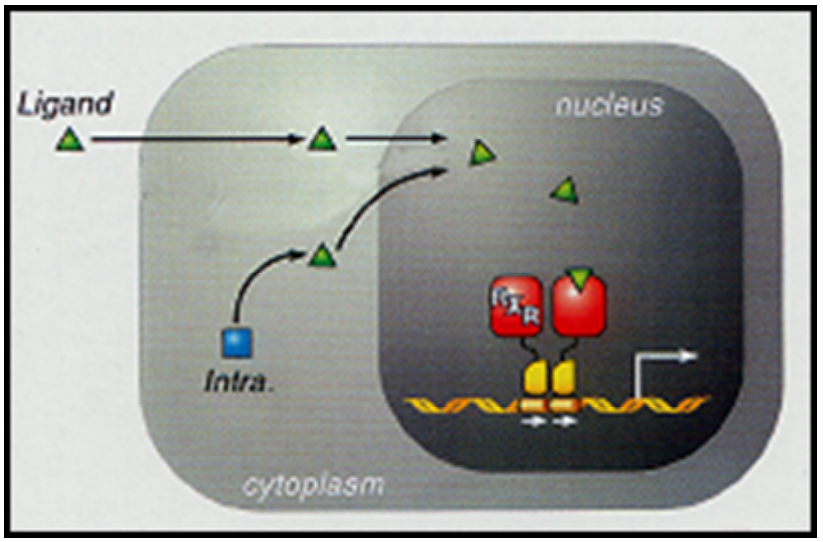


Figura 6.26: probabile meccanismo mediante il quale la famiglia dei recettori nucleari orfani si legano al ligando.

La famiglia dei recettori nucleari è altamente conservata in un range di specie che spaziano dalla Drosophila all'uomo (Yu RT, 1994; Jackson A., 1998; Abrahams BS., 2002).

Il gene che codifica per NR2E1 mappa sul cromosoma 6 (6q21) ed è costituito da 9 esoni (figura 6.27).

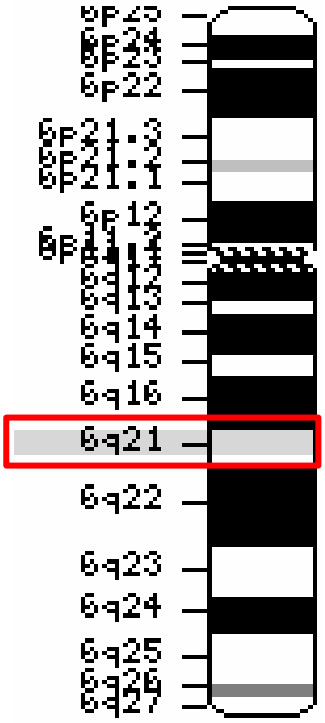


Figura 6.27: gene NR2E1 mappato sul cromosoma 6q21.

La proteina codificata da questo gene (figura 6.28) ha la funzione di legare il DNA come un monomero ad una sequenza detta "Hormone Responsing Element" (HRE) la cui sequenza 5'-AAGGTCA-3' determina un'alta specificità di questo recettore. Il recettore nucleare 2E1 (NR2E1) è espresso nel cervello del feto umano e dell'adulto ed il suo ruolo nel cervello umano e nello sviluppo è sconosciuto.

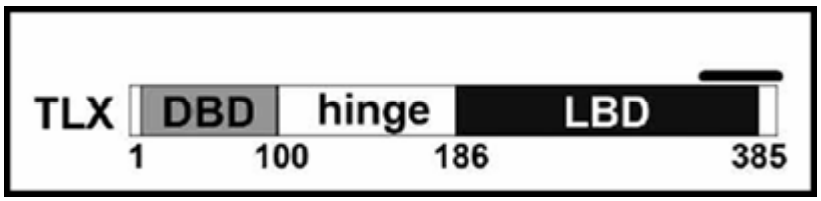


Figura 6.28: proteina codificata dal gene NR2E1.

L'immunofilina ad alto peso molecolare FKBP52 è un bersaglio del farmaco immunosoppressivo FK506 (figura 6.29). FKBP52 esibisce una attività peptidil-prolil *cis–trans* isomerasi (PPIase) che viene inibita dal legame con FK506, proprietà che condivide con la più piccola, ma meglio studiata immunofilina FKBP12 (Davies TH, Sanchez ER., 2005).

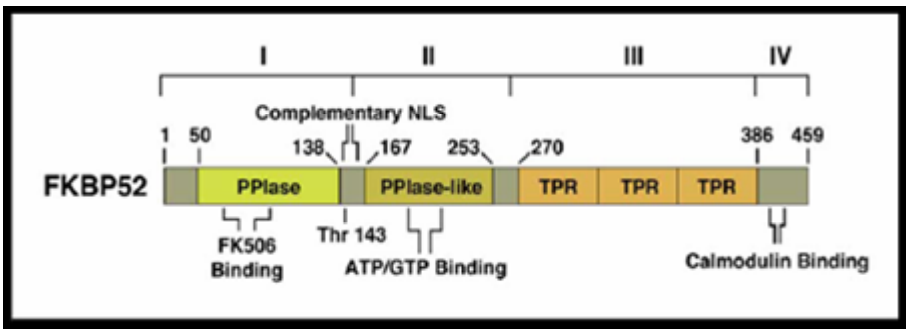


Figura 6.29: immunofilina-target di FK506 codificata dal gene FKBP52.

Caratteristiche:

- espressione ubiquitaria
- espressa nel citoplasma e negli assoni di neuroni mesencefalici e di materia grigia profonda
- mediatore dell'azione neurotrofica di FK506

A differenza di FKBP12, però, FKBP52 non media l'azione immunosoppressiva di FK506 ma, grazie alla sua maggior dimensione, contiene numerosi domini funzionali aggiuntivi. Essa presenta una serie di ripetizioni di domini di tetrapeptidici (TPR) che fungono da siti di binding per lo chaperon molecolare ubiquitario Hsp90. Grazie al legame con questo chaperon, FKBP52 espleta la funzione per la quale è stata meglio caratterizzata, cioè quella di trasportare dal citoplasma al nucleo diversi recettori per ormoni steroidei (Kimmins S, MacRae TH. 2000; Avramut M, Achim CL., 2002), quali il recettore per gli androgeni e per il progesterone (figura 6.30). Sebbene il ruolo maggiormente riconosciuto di FKBP52 sia proprio questa funzione di shuttle, c'è ancora molto da imparare su questa proteina.

Due fenomeni scoperti recentemente potrebbero far luce su alcune funzioni sconosciute di FKBP52, come un suo probabile ruolo nell'allungamento assonale nei neuroni (Galigniana MD, Radanyi C, Renoir JM, Housley PR, Pratt WB, 2001) che può essere stimolato da FK506. Nei topi Fkbp4 -/- si sono riscontrati difetti nel signaling per il recettore degli androgeni e conseguenti difetti nel sistema riproduttivo maschile.

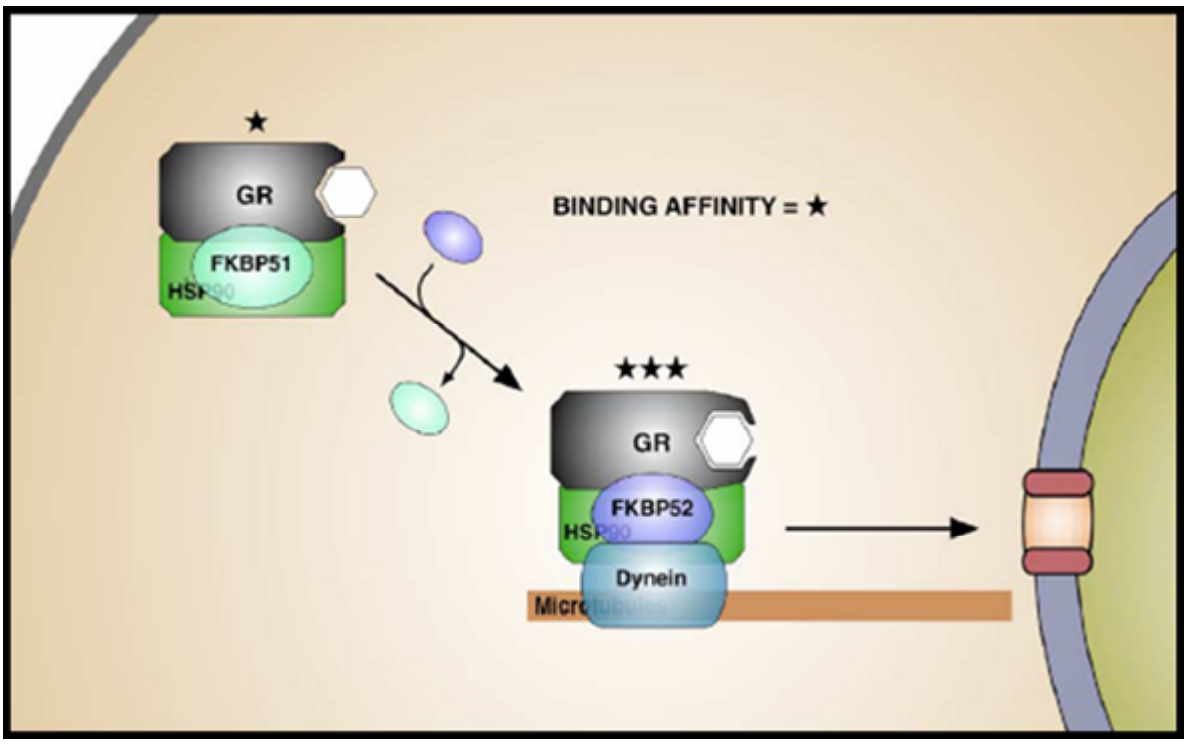


Figura 6.30: componente del complesso recettoriale GR-hsp90-FKBP52, probabilmente coinvolta nella traslocazione di molti recettori per gli ormoni steroidei.

Poiché anche RET51 è coinvolto nel mantenimento del trofismo neuronale e nella promozione dell'allungamento assonale, stiamo cercando di comprendere se RET51 e FKBP52 agiscano sinergicamente nell'espletare questa funzione.

Lo split ubiquitin yeast two hybrid system ha mostrato inoltre l'interazione di RET51 con un'altra immunofilina conosciuta come AIP, XAP2, o Ara9 (Carver LA, 1998; LaPres JJ, 2000). Il gene che codifica per questa proteina è mappato sul cromosoma 11q13.3 (figura 6.31). Questa proteina (figura 6.32) interagisce con il recettore per gli idrocarburi aromatici AhR in maniera simile a ciò che fa FKBP52 con i recettori per gli ormoni steroidei.

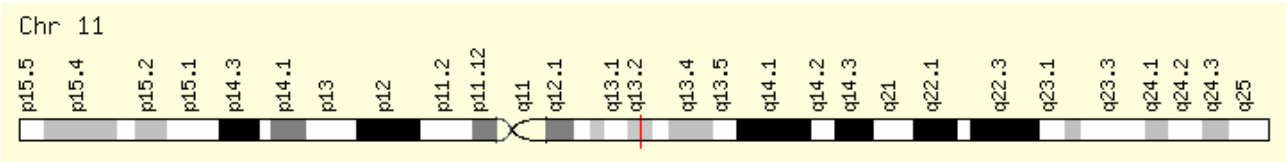


Figura 6.31: gene AIP mappato sul cromosoma 11q13.3.

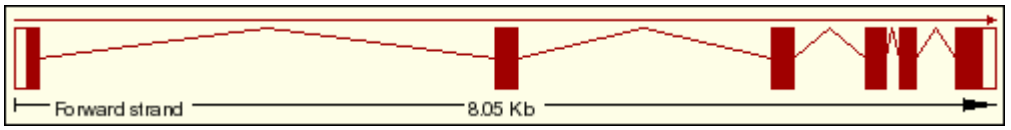


Figura 6.32: proteina codificata dal gene AIP mappato sul cromosoma 11q13.3.

Sembra che AIP accresca la capacità di AhR di legare i propri ligandi e di traslocare nel nucleo. Comunque non è ancora chiaro come AIP accresca la funzionalità di AhR. Ad oggi non sono note in letteratura altre proteine legate da AIP. Poiché RET, in quanto recettore a dipendenza, in assenza di ligando rilascia un frammento pro-apoptotico che espleta la sua funzione in modo ancora sconosciuto, stiamo valutando un eventuale coinvolgimento di AIP nello svolgimento di questo ruolo.

ARF1 è un membro della famiglia dei geni ARF (ADP-ribosylation factor) che codificano per proteine (figura 6.33) capaci di stimolare l'attività dell'ADP-ribosiltransferasi tossina colerica. Queste proteine giocano un ruolo fondamentale nell'assemblamento delle vescicole e nella regolazione del trasporto vescicolare (D'Souza-Schorey, 2006).

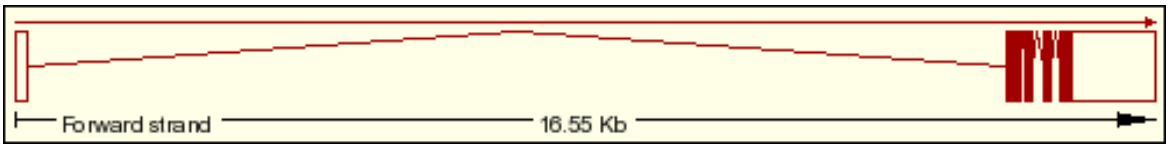


Figura 6.33: trascritto del gene ARF1.

ARF1 in particolare è coinvolto nella regolazione dell'assemblamento sia delle vescicole ricoperte da clatrina che coatomeri di tipo 1 (Traub LM, 2005; Kartberg F, 2005; Bonifacino JS, 2004; Bonifacino JS, 2003) e regola sia il trasporto anterogrado che retrogrado nell'apparato del Golgi (figura 6.34).

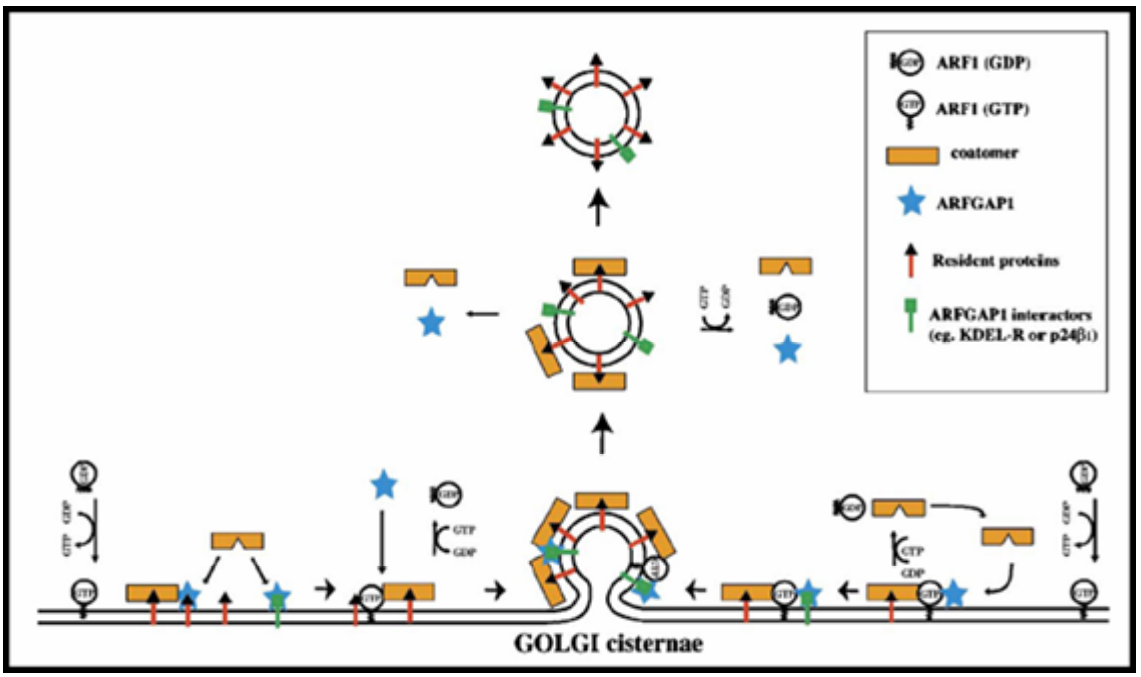


Figura 6.34: ARF1 gioca un ruolo fondamentale nell'assemblamento delle vescicole e nella regolazione del trasporto vescicolare ed in particolare nell'apparato del Golgi.

Inoltre sembra che sia fondamentale nella formazione delle vescicole endocitiche. Il gene che codifica per questa proteina è stato mappato sul cromosoma 1 (1q42.13) e contiene 5 esoni (figura 6.35).

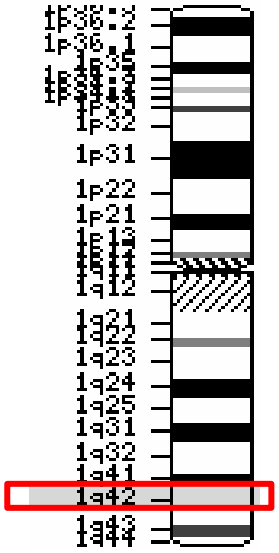


Figura 6.35: gene ARF1 mappato sul cromosoma 1q42.13.

Recentemente si è scoperto che RET, in seguito al legame con GDNF, viene internalizzato in vescicole endocitiche (D S Richardson, A Z Lai and L M Mulligane, 2006) e questo meccanismo è alla base della trasduzione di alcuni segnali mediati da RET (figura 6.36). Attualmente stiamo cercando di verificare se ARF1 è coinvolto nell'internalizzazione di RET.

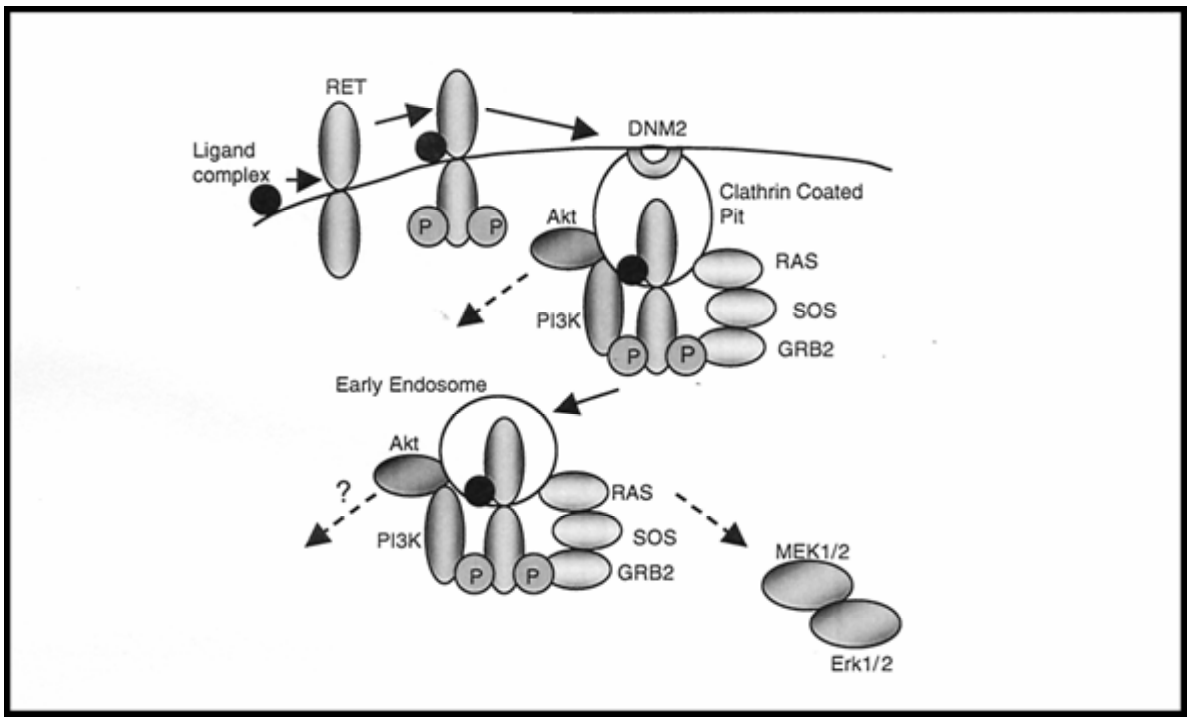


Figura 6.36: internalizzazione di RET indotta dal ligando e sue conseguenze nel downstream signaling.

Tutti questi geni si legano ad un ligando e ne mediano il trasporto nel nucleo.

RET è un recettore a dipendenza ossia quando è in assenza del ligando diventa un substrato delle caspasi e viene rilasciato un frammento pro-apoptotico il quale induce la morte cellulare programmata.

È possibile che il frammento pro-apoptotico esplichi la sua azione agendo sull'attività trascrizionale di altri geni a valle e perciò stiamo cercando di verificare il ruolo di FKBP4, AIP e NR2E1 nell'eventuale traslocazione del frammento pro-apoptotico.

7. SVILUPPI FUTURI

Ciò che ci proponiamo di verificare nel futuro consiste nel ripetere l'esperimento di split ubiquitin utilizzando come esche il frammento pro-apoptotico di RET e la sua isoforma corta (RET9). Tutto ciò per verificare la specificità delle proteine legate da RET51, ossia per verificare se le proteine legate da RET9 e dal frammento pro-apoptotico di RET siano effettivamente diverse da quelle legate da RET51. Questo permetterebbe di affermare che l'interazione è implicata in un qualche fenomeno specifico per una delle due isoforme: RET51 (trofismo dei neuroni simpatici), RET9 (sviluppo embrionale).

Con il programma di gene networks si potrebbe ripetere il calcolo del network statisticamente più probabile per identificare le possibili connessioni con altri geni identificati dal nuovo esperimento di split ubiquitin, che se confermate dalla coimmunoprecipitazione e colocalizzazione, permetterebbero di aggiungere nuove informazioni utili a chiarire ulteriormente il meccanismo alla base del pathway dell'HSCR.

8. BIBLIOGRAFIA

Abrahams BS, Mak GM, Berry ML, Palmquist DL, Saionz JR, Tay A, Tan YH, Brenner S, Simpson EM, Venkatesh B, (2002) Novel vertebrate genes and putative regulatory elements identified at kidney disease and NR2E1/fierce loci; Genomics. Jul;80(1):45-53.

Alizadeh AA (2000), Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000 Feb 3;403(6769):503-11.

Amiel J, L. S. (2001). "Hirschsprung disease, associated syndromes, and genetics: a review." J Med Genet 38: 729-739.

Anders J, K. S., Ibanez CF. (2001). "Molecular modeling of the extracellular domain of the RET receptor tyrosine kinase reveals multiple cadherin-like domains and a calcium-binding site." J Biol Chem. Epub 276:35808-17.

Angrist, M., S. Bolk, et al. (1995). "Mutation analysis of the RET receptor tyrosine kinase in Hirschsprung disease." Hum Mol Genet 4(5): 821-30.

Avramut M, Achim CL. Immunophilins and their ligands: insights into survival and growth of human neurons. Physiol Behav. 2002 Dec;77(4-5):463-8. Review.

Bates D. M. and D. G. Watts (1988), Nonlinear Regression Analysis and Its Applications. John Wiley & Sons, New York.

Attie, T., J. Amiel, et al. (1996). "Genetics of Hirschsprung disease". Ann Chir 50(7): 538-41.

Baumeister R, Ge L. (2002) The worm in us - Caenorhabditis elegans as a model of human disease. Trends Biotechnol Apr;20(4):147-8

Becker R.A., John M. Chambers and Allan R. Wilks (1988), The New S Language. Chapman & Hall, New York. This book is often called the “Blue Book”.

Bessou C, Giugia JB, Franks CJ, Holden-Dye L, Segalat L. Mutations in the Caenorhabditis elegans dystrophin-like gene dys-1 lead to hyperactivity and suggest a link with cholinergic transmission. Neurogenetics. 1998 2(1):61-72.

Bonifacino JS and Lippincott-Schwartz J, (2003), Coat proteins: shaping membrane transport, Nature Reviews, Molecular Cell Biology; vol 4, May 2003 - 409-414.

Bonifacino JS and Lippincott-Schwartz J, (2004), The Mechanisms of Vesicle Review Budding and Fusion Cell, Vol. 116, 153–166, January 23, 2004, Copyright 2004 by Cell Press.

Bordeaux MC, F. C., Granger L, Corset V, Bidaud C, Billaud M, Bredesen DE, Edery P, Mehlen P. (2000). "The RET proto-oncogene induces apoptosis: a novel mechanism for Hirschsprung disease." EMBO J 19: 4056-4063.

Borrego S, R. A., Sáez ME, Gimm O, Gao X, López-Alonso M, Hernández A, Wright FA, Antiñolo G, Eng C. (2000). "RET genotypes comprising specific haplotypes of polymorphic variants predispose to isolated Hirschsprung disease." J Med Genet 37: 572-578.

Borrego S, S. M., Ruiz A, Gimm O, López-Alonso M, Antiñolo G, Eng C. (1999). "Specific polymorphisms in the RET proto-oncogene are overrepresented in patients with Hirschsprung disease and may represent loci modifying phenotypic expression." J Med Genet 36: 771-774.

Borrego S, W. F., Fernández RM, Williams N, López-Alonso M, Davuluri R, Antiñolo G, Eng C. (2003). "A founding locus within the RET proto-oncogene may account for a large proportion of apparently sporadic Hirschsprung disease and a subset of cases of sporadic medullary thyroid carcinoma." Am J Hum Genet 72(88-100).

Carrasquillo MM, McCallion AS, Puffenberger EG, Kashuk CS, Nouri N, Chakravarti A (2002) Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. Nat Genet 32: 237-44

Cacheux V, D.-L. M. F., Kääriäinen H, Bondurand N, Rintala R, Boissier B, Wilson M, Mowat D, Goossens M. (2001). "Loss-of-function mutations in SIP1 Smad interacting protein 1 result in a syndromic Hirschsprung disease." Hum Mol Genet 10: 1503-1510.

Carlomagno F, D. V. G., Berlingieri MT, de Franciscis V, Melillo RM, Colantuoni V, Kraus MH, Di Fiore PP, Fusco A, Santoro M. (1996). "Molecular heterogeneity of RET loss of function in Hirschsprung's disease." EMBO J 15: 2717-2725.

Carver LA, LaPres JJ, Jain S, Dunham EE, Bradfield CA. Characterization of the Ah receptor-associated protein, ARA9. J Biol Chem. 1998 Dec 11;273(50):33580-7.

Chambers J.M. and Trevor J. Hastie eds. (1992), Statistical Models in S. Chapman & Hall, New York. This is also called the “White Book”. ù

Chambers J.M. (1998) Programming with Data. Springer, New York. This is also called the “Green Book”.

Cosma MP, C. M., Carlomagno F, Colantuoni V. (1998). "Mutations in the extracellular domain cause RET loss of function by a dominant negative mechanism." Mol Cell Biol 18: 3321-3329.

D’Souza-Schorey C and Chavrier P, (2006), ARF proteins: roles in membrane traffic and beyond; Nature Reviews; Molecular Cell Biology, vol 7 May 2006, 347-358.

Davies TH, Sanchez ER. FKBP52. Int J Biochem Cell Biol. 2005 Jan;37(1):42-7. Review.

Davison C. and D. V. Hinkley (1997), Bootstrap Methods and Their Applications, Cambridge University Press.

Dobson A.J. (1990), An Introduction to Generalized Linear Models, Chapman and Hall, London.

Eng C, M. L. (1997). "Mutations of the RET proto-oncogene in the multiple endocrine neoplasia type 2 syndromes, related sporadic tumours, and hirschsprung disease." Hum Mutat 9: 97-109.

Fitze G, C. J., Ziegler A, Schierz M, Schreiber M, Kuhlisch E, Roesner D, Schackert HK. (2002). "Association between c135G/A genotype and RET proto-oncogene germline mutations and phenotype of Hirschsprung's disease." Lancet 359: 1200-1205.

Fitze G, S. M., Kuhlisch E, Schackert HK, Roesner D. (1999). "Association of RET protooncogene codon 45 polymorphism with Hirschsprung disease." Am J Hum Genet 65: 1469-1473.

Fitze G, S. M., Kuhlisch E, Schreiber M, Ziegler A, Roesner D, Schackert HK. (2003). "Novel intronic polymorphisms in the RET proto-oncogene and their association with Hirschsprung disease." Hum Mutat 22: 177.

Fukuda T, Asai N, Enomoto A, Takahashi M. Activation of c-Jun amino-terminal kinase by GDNF induces G2/M cell cycle delay linked with actin reorganization. Genes Cells 2005; 10:655–63

Galigniana MD, Radanyi C, Renoir JM, Housley PR, Pratt WB. (2001), Evidence that the peptidylprolyl isomerase domain of the hsp90-binding immunophilin FKBP52 is involved in both dynein interaction and glucocorticoid receptor movement to the nucleus. J Biol Chem. 2001 May 4;276(18):14884-9. Epub 2001 Feb 13.

Gardner T. S., Faith J.J. / Physics of Life Reviews 2 (2005) 65–88

Gath R, G. A., Keller KM, Koletzko S, Coerdt W, Muntefering H, Wirth S, Hofstra RM, Mulligan L, Eng C, von Deimling A. (2001). "Analysis of the RET, GDNF, EDN3, and EDNRB genes in patients with intestinal neuronal dysplasia and Hirschsprung disease." Gut 48: 671-675.

Giovanni Romeo, P. R., Yin Luo, Virginia Barone, Marco Seri, Isabella Ceccherini, Barbara Pasini, Renata Bocciardi, Margherita Lerone, Helena Kääriäinen* & Giuseppe Martucciello (1994). "Point mutations affecting the tyrosine kinase domain of the RET proto-oncogene in Hirschsprung's disease." Nature 367: 377-378.

Griseri P, P. B., Patrone G, Osinga J, Puppo F, Sancandi M, Hofstra R, Romeo G, Ravazzolo R, Devoto M, Ceccherini I. (2002). "A rare haplotype of the RET proto-oncogene is a risk-modifying allele in hirschsprung disease." Am J Hum Genet 71(4): 969-74.

Heckerman D, C. Meek, and G. Cooper. A Bayesian approach to causal discovery. Technical report, 1997. Technical Report MSR-TR-97-05, Microsoft Research.

Hofstra RM, Valdenaire O, Arch E, Osinga J, Kroes H, Loffler BM, Hamosh A, Meijers C, Buys CH (1999) A loss-of-function mutation in the endothelin-converting enzyme 1 (ECE-1) associated with Hirschsprung disease, cardiac defects, and autonomic dysfunction. Am J Hum Genet 64: 304-8

Huang, L.-S.; Voyiaziakis, E.; Chen, H. L.; Rubin, E. M.; Gordon, J. W.: A novel functional role for apolipoprotein B in male infertility in heterozygous apolipoprotein B knockout mice. Proc. Nat. Acad. Sci. 93: 10903-10907, 1996.

Huang, L.-S.; Voyiaziakis, E.; Markenson, D. F.; Sokol, K. A.; Hayek, T.; Breslow, J. L.: apo B gene knockout in mice results in embryonic lethality in homozygotes and neural tube defects, male infertility, and reduced HDL cholesterol ester and apo A-1 transport rates in heterozygotes. J. Clin. Invest. 96: 2152-2161, 1995.

Ishizaka Y, I. F., Tahira T, Ikeda I, Sugimura T, Tucker J, Fertitta A, Carrano AV, Nagao M. (1989). "Human ret proto-oncogene mapped to chromosome 10q11.2." Oncogene 4: 1519-21.

Iwashita T, K. K., Qiao S, Murakami H, Asai N, Kawai K, Hashimoto M, Watanabe T, Ichihara M, Takahashi M. (2001). "Functional analysis of RET with Hirschsprung mutations affecting its kinase domain." Gastroenterology 121: 24-33.

Iwashita T, M. H., Asai N, Takahashi M. (1996). "Mechanism of ret dysfunction by Hirschsprung mutations affecting its extracellular domain." Hum Mol Genet 5: 1577-1580.

Jackson A, Panayiotidis P, Foroni L., (1998) The human homologue of the Drosophila tailless gene (TLX): characterization and mapping to a region of common deletion in human lymphoid leukemia on chromosome 6q21. Genomics. May 15;50(1):34-43.

Jain S, Naughton CK, Yang M, Strickland A, Vij K, Encinas M, Golden J, Gupta A, Heuckeroth R, Johnson EM Jr, Milbrandt J., (2004),Mice expressing a dominant-negative Ret mutation phenocopy human Hirschsprung disease and delineate a direct role of Ret in spermatogenesis. Development. 2004 Nov;131(21):5503-13. Epub 2004 Oct 6.

Kartberg F, Elsner M, Fröderberg L, Asp L, Nilsson T, (2005), Commuting between Golgi cisternae—Mind the GAP!, Biochimica et Biophysica Acta 1744 (2005) 351 – 363.

Kimmins S, MacRae TH. Maturation of steroid receptors: an example of functional cooperation among molecular chaperones and their associated proteins. Cell Stress Chaperones. 2000 Apr;5(2):76-86. Review.

LaPres JJ, Glover E, Dunham EE, Bunger MK, Bradfield CA. ARA9 modifies agonist signaling through an increase in cytosolic aryl hydrocarbon receptor. J Biol Chem. 2000 Mar 3;275(9):6153-9.

Lorenzo MJ, E. C., Mulligan LM, Stonehouse TJ, Healey CS, Ponder BA, Smith DP. (1995). "Multiple mRNA isoforms of the human RET proto-oncogene generated by alternate splicing." Oncogene 10: 1377-83.

Manié S, S. M., Fusco A, Billaud M. (2001). "The RET receptor: function in development and dysfunction in congenital malformation." Trends Genet 17: 580-589.

McCallion AS, S. E., Conlon RA, Chakravarti A. (2003). "Phenotype variation in two-locus mouse models of Hirschsprung disease: tissue-specific interaction between Ret and Ednrb." Proc Natl Acad Sci U S A 100: 1826-1831.

McCullagh P. and John A. Nelder (1989), Generalized Linear Models. Second edition, Chapman and Hall, London.

Mulligan LM, E. C., Attié T, Lyonnet S, Marsh DJ, Hyland VJ, Robinson BG, Frilling A, Verellen-Dumoulin C, Safar A. (1994). "Diverse phenotypes associated with exon 10 mutations of the RET proto-oncogene." Hum Mol Genet 3: 2163-2167.

Myers SM, E. C., Ponder BA, Mulligan LM. and O. 6:297-301. (1995.). "Characterization of RET proto-oncogene 3' splicing variants and polyadenylation sites: a novel C-terminus for RET." Oncogene 11: 2039-45.

Paratcha G., L. F., Baar L., Coulpier M., Besset V., Anders J., Scott R., Ibanez CF. (2001). "Released GFRalpha1 potentiates downstream signaling, neural survival, and differentiation via a novel mechanism of recruitment of c-Ret to lipid rafts." Neuron 29: 171-184.

Parisi MA, K. R. (2000). "Genetics of Hirschsprung disease." Curr Opin Pediatr 12: 610-617.

Pasini B, B. M., Greco A, Bongarzone I, Luo Y, Mondellini P, Alberti L, Miranda C, Arighi E, Bocciardi R, Seri M, Barone V, Radice MT, Romeo G, Pierotti MA. (1995). "Loss of function effect of RET mutations causing Hirschsprung disease." Nat Genet 10: 35-40.

Pelet A, G. O., Edery P, Pasini A, Chappuis S, Atti T, Munnich A, Lenoir G, Lyonnet S, Billaud M. (1998). "Various mechanisms cause RET-mediated signaling defects in Hirschsprung's disease." J Clin Invest 101: 1415-1423.

Pearl J, Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Francisco, Calif., 1988.

Pearl. and T. S. Verma. A theory of inferred causation. In Principles of Knowledge Representation and Reasoning: Proc. Second International Conference (KR '91), pp. 441–452. 1991.

Rice J.A. (1995), Mathematical Statistics and Data Analysis. Second edition. Duxbury Press, Belmont, CA.

Richardson DS, Lai AZ, Mulligan LM. RET ligand-induced internalization and its consequences for downstream signaling. Oncogene. 2006 May 25;25(22):3206-11.

Romeo G, C. I., Celli J, Priolo M, Betsos N, Bonardi G, Seri M, Yin L, Lerone M, Jasonni V, Martucciello G. (1998). "Association of multiple endocrine neoplasia type 2 and Hirschsprung disease." J Intern Med. 243(6): 515-20.

Roy K, Kuznicki K, Wu Q, Sun Z, Bock D, Schutz G, Vranich N, Monaghan AP. The Tlx gene regulates the timing of neurogenesis in the cortex. J Neurosci. 2004 Sep 22;24(38):8333-45.

Ross DT, Molecular portraits of human breast tumours. Nature. 2000 Aug 17;406(6797):747-52.

Satyal SH, Schmidt E, Kitagawa K, Sondheimer N, Lindquist S, Kramer JM, Morimoto RI. Polyglutamine aggregates alter protein folding homeostasis in Caenorhabditis elegans. Proc Natl Acad Sci U S A. 2000 97(11):5750-5.

Schena M, Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. (1995) Oct 20;270(5235):467-70.

Silvey S. D. (1970), Statistical Inference. Penguin, London.

Spirtes P, C. Glymour, and R. Scheines. Causation, prediction, and search. Springer-Verlag, 1993.

Svensson PJ, Von Tell D, Molander ML, Anvret M, Nordenskjold A. (1999) A heterozygous frameshift mutation in the endothelin-3 (EDN-3) gene in isolated Hirschsprung's disease. Pediatr Res May; (Pt1): 714-7.

Takahashi M, B. Y., Taniguchi M. (1991). "Identification of the ret protooncogene products in neuroblastoma and leukemia cells." Oncogene 6: 297-301.

Takahashi M, I. T., Santoro M, Lyonnet S, Lenoir GM, Billaud M. (1999). "Cosegregation of MEN2 and Hirschsprung's disease: the same mutation ofRET with both gain and loss-of-function?" Hum Mutat 13: 331-336.

Tanaka H, Moroi K, Iwai J, Takahashi H, Ohnuma N, Hori S, Takimoto M, Nishiyama M, Masaki T, Yanagisawa M, Sekiya S, Kimura S. (1998) Novel mutations of the endothelin B receptor gene in patients with Hirschsprung's disease and their characterization. J Biol Chem. 273: 11378-83.

Traub LM, (2005), Common principles in clathrin-mediated sorting at the Golgi and the plasma membrane; Biochimica et Biophysica Acta 1744 (2005) 415 – 437.

Venables W. N., D. M. Smith and the R Development Core Team, (2006) An Introduction to R

Wakamatsu N, Y. Y., Yamada K, Ono T, Nomura N, Taniguchi H, Kitoh H, Mutoh N, Yamanaka T, Mushiake K, Kato K, Sonta S, Nagaya M. (2001). "Mutations in SIP1, encoding Smad interacting protein-1, cause a form of Hirschsprung disease." Nat Genet 27: 369-370.

Wittenburg N, Eimer S, Lakowski B, Rohrig S, Rudolph C, Baumeister R. Presenilin is required for proper morphology and function of neurons in C. elegans. Nature. 2000 406(6793):306-9.

Xing S, T. Q., Suzuki T, Jhiang SM. (1994). "Alternative splicing of the ret proto-oncogene at intron 4." Biochem Biophys Res Commun. 205: 1526-32.

Yu RT, (1994), Relationship between Drosophila gap gene tailless and a vertebrate nuclear receptor Tlx, Nature. Aug 4;370(6488):375-9.

RINGRAZIAMENTI

Ecco, ci siamo finalmente arrivati! Se sto scrivendo i ringraziamenti è perché sono riuscito, in un modo o nell'altro, ad arrivarci in fondo. Per me, comunque, rimane la parte più difficile da scrivere per tanti motivi: timidezza, difficoltà ad esprimere i miei sentimenti o forse, perché no, l'ora tarda nella quale li sto scrivendo e la stanchezza soprattutto.

Veniamo a noi, comunque, vorrei innanzitutto ringraziare i mie genitori che, come al solito, mi hanno aiutato anche in questa avventura chiamata "dottorato", facendomi sentire la loro presenza ed il loro affetto. Un grazie particolare al Professor Giovanni Romeo che fin dall'inizio ha creduto in me, permettendomi di partecipare ai tre anni più intensi della mia vita, sia dal punto di vista professionale, che da quello umano. Un bacio grosso ad Alessandra che mi ha sopportato durante tutti quei periodi in cui ero nervoso e stanco (cosa non facile visto il mio carattere!!!).

Un ringraziamento a tutti i miei amici e compagni di questo viaggio, che non cito uno ad uno perché sarebbero troppi (non riuscirei mai citarli tutti)…ragazzi favolosi che da soli valgono il periodo di dottorato, senza i quali il mio percorso sarebbe stato sicuramente molto, molto, molto più complicato.