UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E GESTIONE DELLE IMPRESE

TESI DI LAUREA

Test alternativi per la selezione di geni differenzialmente espressi: uno studio di simulazione.

Relatore: Ch.ma Prof.ssa Chiogna Monica

Laureando: Cavallin Francesco

455219 – SGI

ANNO ACCADEMICO 2004-2005

Indice.

Introduzione.

Da tempo, lo studio del DNA umano sta assumendo sempre più importanza e diffusione in quanto,dopo lunghe analisi,si è ipotizzato che gran parte delle malattie derivino da alterazioni del codice genetico.
Una tecnologia utile per identificare mutazioni presenti nei geni è quella del DNA microarray; questa può essere inoltre utilizzata per comprendere, tramite l'analisi simultanea di migliaia di geni, la patogenesi di malattie genetiche e quella di malattie multifattoriali (come diabete, arteriosclerosi, osteoporosi).
In particolare, essa rappresenta un ottimo strumento per identificare i geni con livelli di espressione diversa (in diverse condizioni o soggetti), per identificare i gruppi di geni che con buona probabilità sono correlati tra loro, per caratterizzare le cellule malate tramite la classificazione di campi biologici, per individuare i geni marcatori (geni il cui livello di espressione è utile in campo biologico per determinare un gruppo o fenotipo).
Le applicazioni che ne derivano sono numerose e di notevole importanza: vanno dallo sviluppo di nuovi farmaci ad un miglior utilizzo di quelli già esistenti, dalla prevenzione di malattie genetiche (l'identificazione di mutazioni è fondamentale per la diagnosi precoce dei tumori) all'identificazione di ceppi virali in microbiologia, senza dimenticare che un'accurata distinzione dei diversi tipi di tumori permette un trattamento più mirato e una minore esposizione dei pazienti alla tossicità delle terapie. Infatti, i metodi convenzionali di classificazione (con basi cliniche e morfologiche) delle forme tumorali presentavano limitazioni, specialmente in campo diagnostico; poiché le diverse classi tumorali presentano diversi decorsi clinici e diverse caratteristiche molecolari, è stata suggerita una differenziazione dei trattamenti in relazione alla differenziazione delle tipologie tumorali che potrebbe aumentare l'efficacia della terapia.

In questo elaborato è presentato uno studio di simulazione in cui si analizzano alcuni metodi per l'individuazione di geni differenzialmente espressi.

La letteratura presenta l'applicazione di diversi metodi per la distinzione di diverse forme di cancro, tuttavia quando questi sono applicati a dati di espressione genica insorgono alcuni problemi di tipo statistico: questi trovano spiegazione nella diversità di tali dati rispetto ai dati di cui normalmente lo statistico si occupa.

Questi problemi sono stati esposti da Gordon K. Smith et al. Rif[1] e sono di seguito brevemente riassunti.

In primo luogo, a causa della natura stessa dei dati, vi è una massiccia presenza di rumore biologico o tecnico, che si ripercuote sull'attendibilità dei dati e che non è sempre possibile isolare, data la difficoltà di valutazione.

Inoltre, un'ulteriore difficoltà è data dal dataset, che presenta una grande dimensionalità a cui si contrappongono campioni limitati (qualche migliaio di geni contro campioni di un centinaio di unità). Questo problema è noto in letteratura con l'espressione "grande p e piccolo n". Le sue conseguenze sono tempi di elaborazione lunghi e rischio di sovraparametrizzazione (dovuta all'alta dimensionalità e alla ridotta numerosità campionaria).

Infine, la maggior parte dei geni nel dataset costituisce un rumore dal momento che i geni rilevanti per la classificazione sono una piccola parte dei geni considerati; questo aumenta la difficoltà di classificazione e i tempi di calcolo. I metodi di classificazione esistenti risultano inadeguati, proprio perché non sono stati concepiti per dati di questo tipo, quindi alcuni studiosi hanno proposto di raggruppare i geni in classi omogenee come operazione preliminare, al fine di ridurre la dimensionalità e i tempi di calcolo tramite l'eliminazione dei geni irrilevanti (i quali riducono l'accuratezza della classificazione).

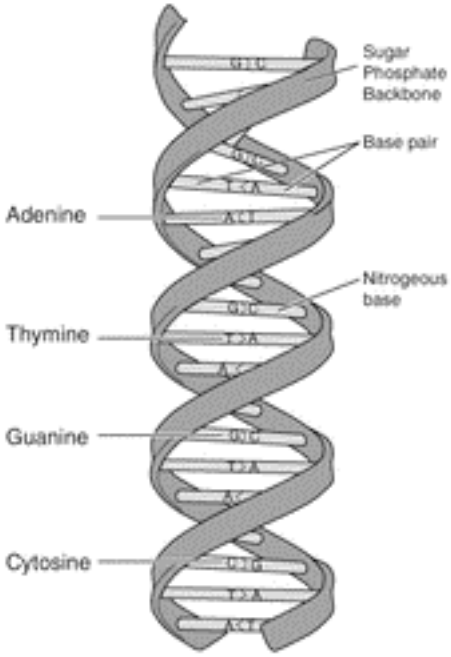
Altri problemi esulano poi dal contesto statistico e riguardano invece il contesto biologico e l'importanza dei risultati dal punto di vista medico; a titolo di esempio, un problema riguarda la corrispondenza tra rilevanza biologica e statistica di uno stesso gene differenzialmente espresso come classificatore: la rilevanza biologica va tenuta

in grande considerazione perché ogni informazione ricavata dall'analisi può essere utile per ulteriori sviluppi, come la scoperta di interazioni tra geni, l'individuazione di geni marcatori e la scoperta delle funzioni specifiche di un determinato gene.

CAPITOLO I

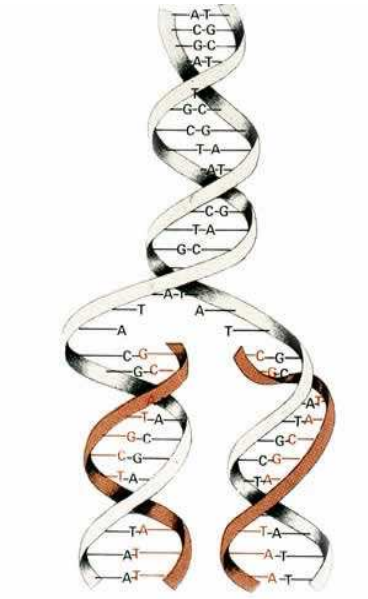
Richiami di biologia.

1.1 Il DNA.



Il DNA, sigla di *acido deossiribonucleico*, è una lunga molecola composta da due filamenti avvolti in modo elicoidale l'uno sull'altro e costituiti da subunità ripetute di un gruppo fosfato e dello zucchero deossiribosio a cinque atomi di carbonio; i due filamenti sono uniti da ponti infinitesimali, detti *ponti idrogeno*, formati da una coppia di basi azotate.

Vi sono quattro tipi di basi: *adenina, timina, citosina e guanina*; queste possono accoppiarsi solo in un determinato modo, adenina con timina e citosina con guanina. A seconda di come si presentano le triplette, si ha la formazione di un particolare gene, il quale è un segmento di DNA che ha la funzione di trasmettere informazioni per la sintesi proteica.



Nella duplicazione del DNA, infatti, si rompono i legami ad idrogeno e i due filamenti si staccano, formando ciascuno con le sue basi azotate uno stampo per la formazione di un nuovo filamento complementare; in questo modo l'informazione ereditaria si trasmette dalla cellula madre a quella figlia (*duplicazione semiconservativa).*

Uno zucchero deossiribosio, un gruppo fosfato e una base azotata formano un *nucleotide*; a loro volta, una sequenza di tre nucleotidi codificano un aminoacido (necessario per la sintesi proteica).

Il processo di traduzione del DNA in proteine è costituito da due fasi, trascrizione e traduzione.
La prima avviene nel nucleo: su un singolo filamento di *RNA messaggero* (mRNA) viene trascritta l'informazione da un filamento di DNA, grazie al fatto che l'RNA messaggero è simile al DNA con la sola differenza che al posto della timina vi è l'uracile.
Completata la trascrizione, l'RNA messaggero esce del nucleo e si sposta sui *ribosomi*, costituiti da *RNA ribosomiale* (rRNA) e proteine, dove avviene la fase successiva; qui interviene un terzo tipo di RNA, l'*RNA di trasporto* (tRNA), che è costituito da una tripletta di basi azotate (*anticodone*) specifica per l'aminoacido che trasporta. Durante la sintesi, a ciascuna tripletta (*codone*) del mRNA si attacca una molecola di tRNa avente un anticodone adeguato, apportando così un determinato aminoacido; in tal modo, gli aminoacidi vengono allineati secondo la sequenza dettata inizialmente dal DNA e formano la catena polipeptidica (la *proteina*).

Le mutazioni sono cambiamenti nella sequenza o nel numero dei nucleotidi presenti nell'acido nucleico della cellula, dovuti all'aggiunta, alla mancanza o alla sostituzione di un nucleotide con un altro; ciò provoca mutazioni nei geni che codificano alcune proteine e quindi la mancanza o inattività di quest'ultime, causa dell'insorgere di molte malattie genetiche.
È importante, per la comprensione della tecnica dei microarray, aggiungere che nella fase di trascrizione la cellula produce RNA solo per quei geni che sono attivi in quel momento: questo implica che l'analisi dell'RNA prodotto dalla cellula offre un modo per determinare quali geni sono attivi e quali inattivi in un determinato istante. Questa è l'intuizione da cui si sviluppa la tecnica dei microarray di DNA.

1.2 Le cellule.

Le cellule sono le unità strutturali biologiche di base.
Sono circondate da una membrana che le separa dall'ambiente esterno ed ha la funzione di conservarne l'integrità funzionale e di regolare il passaggio delle sostanze dall'interno all'esterno e viceversa.
L'interno è costituito dal *citoplasma* (una soluzione acquosa concentrata contenente enzimi, ioni, molecole disciolte ed organuli) e dal *reticolo endoplasmatico* (un sistema di membrane); tra questi organuli vi sono i *ribosomi*, dove ha luogo la sintesi proteica, che possono trovarsi sia sul reticolo sia liberi nel citoplasma. Sempre nel citoplasma si trovano anche l'*apparato di Golgi* (dove vengono immagazzinate le molecole sintetizzate nella cellula), i *mitocondri* (in cui avvengono le reazioni chimiche che forniscono energia alla cellula), i *lisosomi* e i *perossisomi* (dove le molecole vengono scomposte in elementi più semplici che vengono utilizzati dalla cellula o eliminati).
La forma della cellula è determinata dal *citoscheletro*, che inoltre consente alla cellula di muoversi e fissa i suoi organuli.
La parte più importante è comunque il *nucleo* che regola le attività cellulari; è separato dal citoplasma da una doppia membrana, al cui interno vi è il *nucleolo* dove si formano le subunità ribosomiali e la *cromatina* (sostanza, formata da DNA e proteine, che costituisce i cromosomi e prende il nome di nucleo quando si trova in forma disciolta).

1.3 Il percorso degli studi sul DNA.

Le caratteristiche di ciascun individuo sono specificate da quella che viene chiamata *informazione biologica*; quest'ultima è organizzata in unità, i geni, che determinano i caratteri e sono ereditati dai genitori: per questo si parla caratteri ereditari, che vengono trasferiti da una generazione all'altra tramite la riproduzione.

Gli studi su evoluzione ed ereditarietà cominciarono quando ancora non si aveva nessuna conoscenza della struttura del DNA e quindi procedettero in varie direzioni teoriche, talvolta portando a soluzioni curiose.

La prima teoria è quella di Lamarck. Nel 1797 lo scienziato Lamarck aveva suggerito l'ipotesi che i caratteri acquisiti in vita fossero ereditabili; famoso è l'esempio della giraffa, che allunga il collo per brucare le foglie dei rami più alti e trasmette così alla generazione successiva questa caratteristica sviluppata in vita. Tuttavia è palese il fatto che molti caratteri acquisiti in vita dal genitore non siano poi riscontrati anche nel figlio ed è proprio questa critica quella più frequentemente rivolta allo scienziato dai suoi colleghi.

Il primo ad occuparsi del problema in modo scientifico fu l'austriaco Gregor Mendel nel 1909; egli utilizzò delle piante di piselli odorosi con due importanti caratteristiche: capacità di autofecondazione e cicli vitali non troppo lunghi. I risultati degli esperimenti da lui compiuti gli permisero di individuare alcuni caratteri che ricomparivano regolarmente nelle popolazioni, anche se i legami che mettevano in connessione questi eventi non erano ancora chiari. Forse proprio a causa di questa incapacità di individuare delle conclusioni chiare, le sue ricerche non furono prese in considerazione dai contemporanei, nonostante gettassero le basi della genetica e delineassero la direzione giusta per trovare le risposte alle domande che si era posto.

Bisogna far trascorrere molti anni per giungere alla prima prova decisiva che il DNA è il depositario dell'informazione biologica: nel 1952, infatti, Harshey e Chase dimostrarono che i batteriofagi iniettano una molecola di DNA nel batterio ospite per introdurvi il loro materiale ereditario.

Questa scoperta suscitò l'interesse della comunità scientifica sulla struttura di tale molecola e proprio in questa direzione, qualche anno dopo, James Watson e Francis Crick iniziarono a lavorare insieme al Cavendish Laboratory, in Inghilterra. Il loro lavoro non si basò su esperimenti veri e propri, ma fu piuttosto teso ad analizzare ed organizzare in modo logico tutti i dati allora noti sul DNA. Le informazioni principali a loro disposizione erano la sua struttura lunga e filiforme e le sue grosse dimensioni; inoltre era stato suggerito un comportamento simile a quello delle molecole di proteine, analogia derivante dalla scoperta (qualche anno prima) che le proteine hanno una forma elicoidale mantenuta da legami ad idrogeno che si formano tra le spire adiacenti. In seguito, venne dimostrata ai raggi X la forma ad elica del DNA (Walkins e Frankling) e fu verificata l'impossibilità di legare chimicamente tra loro due basi purine o due basi pirimidine, rispettivamente a due ed a un anello (Chargaff); questo portò ad affermare l'esclusività dei legami timina-adenina e citosina-guanina.

Il risultato di tutte queste scoperte è la forma strutturale della molecola di DNA attualmente nota: una doppia elica formata da due lunghi filamenti costituiti da molecole alternate di zucchero e fosfato; i due filamenti sono avvolti a spirale e tenuti insieme da legami covalenti tra le coppie di basi azotate.

Nel corso degli anni si è poi arrivati a scoprire molti aspetti del trasferimento dell'informazione biologica e molte particolarità del DNA, tramite lo studio e la classificazione dei suoi componenti, i geni.

L'ambito applicativo più interessante è lo studio delle malattie, dato che da tempo molti studiosi concordano sul fatto che alcune patologie derivino da alterazioni del codice genetico; sono proprio alcune differenze nell'espressione dei geni (differenze che riguardano sia le proteine generate sia le modalità di utilizzo di questi geni) che distinguono un individuo malato da uno sano.

In quest'ambito, il ricercatore biologo è supportato dalla disciplina della *biostatistica*, che fornisce assistenza per quanto riguarda il disegno e la valutazione probabilistica della variazione di espressione dei geni: lo scopo è quello di capire,a livello genetico,

ciò che differenzia una cellula malata da una sana e poter quindi riconoscere un paziente come sano o malato una volta indagato il suo profilo genetico.

CAPITOLO II

2.1 La tecnologia DNA microarray.

Descritta per la prima volta nel 1995 e messa sul mercato nel 1996, questa tecnologia è un importante strumento delle *nanotecnologie* e trova utilità nello studio dell'espressione genica; per questo motivo è di grande interesse per i ricercatori che si occupano di studiare le basi molecolari del cancro, nonché in campo farmacologico per l'individuazione di nuovi farmaci.

Un'importante passo in avanti permesso dai microarray consiste nel fatto che essi consentono di analizzare contemporaneamente l'attività di decine di migliaia di geni, mentre prima si poteva analizzare solo un gene alla volta (si diceva infatti "un gene, una vita").

Attualmente si distinguono due tipi:

1. *microarray di cloni di DNA micropipettati*: vengono depositati dei singoli geni sui vetrini trattati con agenti chimici che favoriscano il legame del DNA;
2. *microarray di oligonucleotidi sintetizzati in situ*: vengono sintetizzati i nucleotidi che rappresentano la sequenza bersaglio direttamente su chip di silicio.

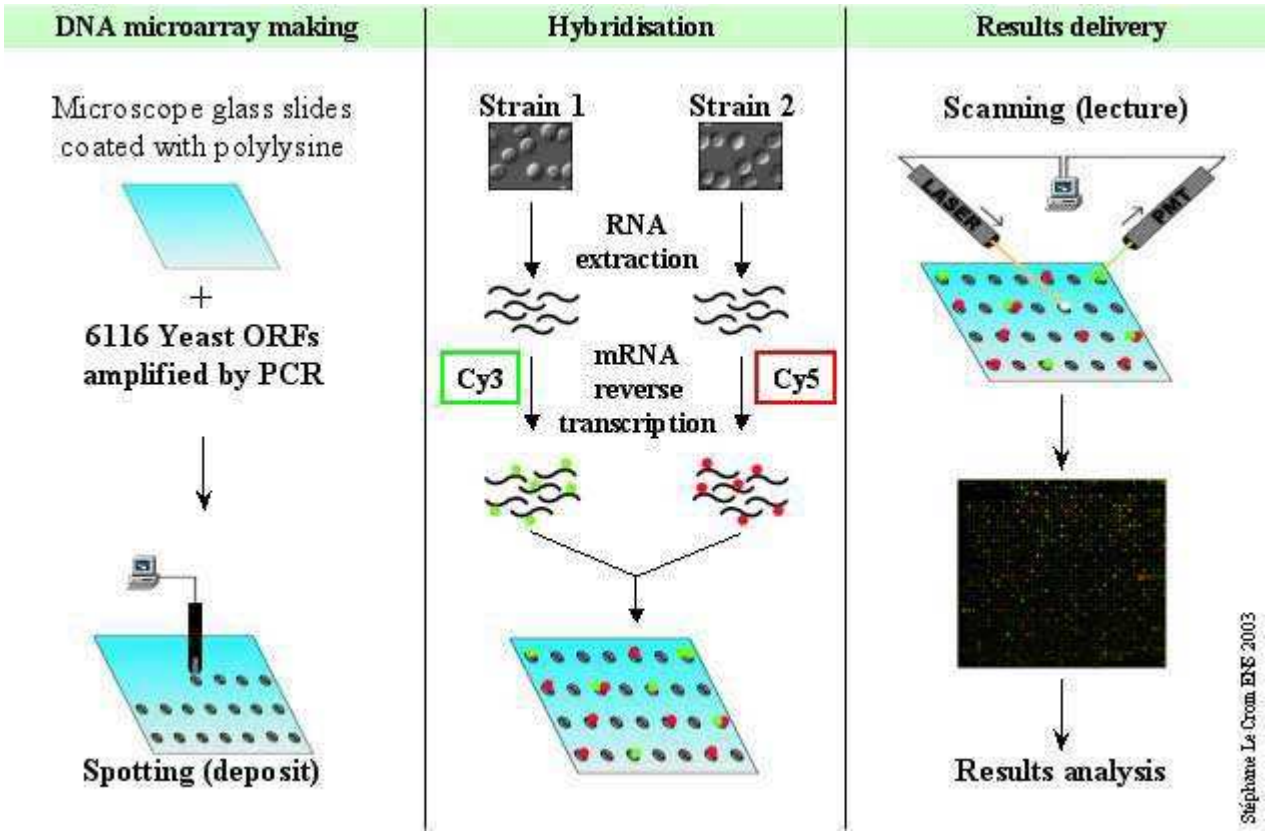
I vantaggi presentati sono, rispettivamente:

1. alta automazione nella fase sperimentale, non è necessario conoscere la sequenza di Dna da stampare, basso costo (che favorisce l'alta riproducibilità);
2. si può adattare alle varie esigenze sperimentali, grazie al fatto che si può disegnare la sequenza bersaglio.

Con l'utilizzo dei microarray, detti anche *DNA chip*, è possibile verificare quali geni sono attivi in un tessuto o in un tipo di cellula, qual è il loro profilo di espressione e quali variazioni si riscontrano in caso di malattia; l'importanza di questi risultati è evidenziata dalle possibili applicazioni pratiche:

- identificazione dei geni potenzialmente oncogeni che sono attivi nelle cellule tumorali di un paziente e non in un altro;
- identificazione dei geni che determinano la differenza tra tumore primario e relativa metastasi;
- supporto alla classificazione tradizionale di tipo isto-morfologico dei tumori, tramite una classificazione complementare basata sul profilo di espressione genica.

2.2 L'esperimento.



Preparazione dei microarray.

Segmenti di cDNA sintetico (detti *probe* o *sonde*), che riproducono i geni che si vogliono indagare, vengono fissati su vetrini da microscopio ricoperti precendentemente con *polylysine*, il cui scopo è assicurare la fissazione del DNA attraverso interazione elettrostatica. Infine, viene bloccata la *polylysine* non fissata al DNA per impedire che fissi invece il target.

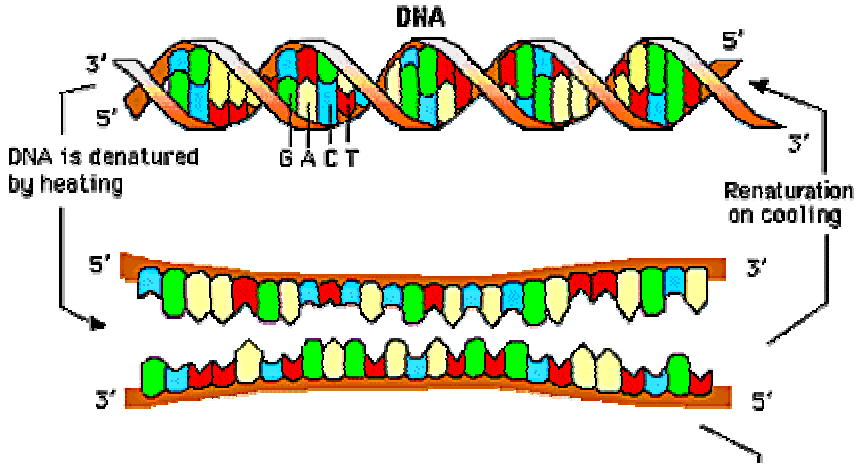
Preparazione del target.

Si estrae l'mRNA prodotto dalle due cellule di cui si vuole comparare il livello di espressione e tramite una reazione biochimica (*reverse transcription*) esso è trasformato in cDNA, data la sua maggior stabilità rispetto all'mRNA.

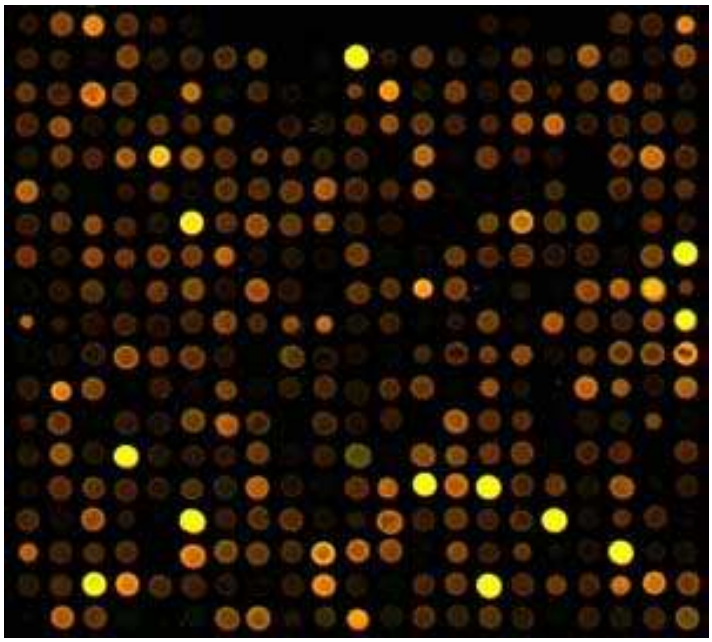
In questa fase, vengono introdotti nel cDNA particolari molecole (i *recettori*) aventi la proprietà di legarsi a sostanze fluorescenti, allo scopo di permettere la successiva marcatura del cDNA delle due cellule. Infatti, ai recettori vanno a legarsi i marcatori fluorescenti Cy3 (verde) e Cy5 (rosso), per segnare rispettivamente le cellule sane e quelle malate.

Ibridizzazione.

Il cDNA delle due cellule viene mischiato assieme (prendendo il nome di *target*) e depositato sull'array affinché si ibridi con le sonde; il chip viene poi incubato una notte a 60 gradi, temperatura a cui i segmenti di DNA riconoscono le sonde complementari e vi si legano.



Scanning del vetrino.



Il chip viene eccitato con un laser affinché i marcatori fluorescenti emettano un segnale luminoso e l'emissione fluorescente è quindi analizzata tramite un *foto-moltiplicatore* (*PMT*) accoppiato con un microscopio: l'intensità degli *spot* (ciascun puntino che rappresenta un gene) verdi misura la quantità di cDNA segnata con Cy3 e quella degli spot rossi la quantità di cDNA segnata con Cy5; ciò indica le quantità di

mRNA prodotto, rispettivamente, da cellule sane e malate. Si ottengono così due immagini monocromatiche in verde e in rosso, le quali vengono sovrapposte per ottenere una visione finale dove ad ogni spot corrisponde un gene e il suo colore è indicatore della sua condizione nella cellula sana e in quella malata.
Il colore verde corrisponde ad un gene attivo nella cellula sana e inattivo in quella malata, il rosso indica il caso contrario, il giallo indica un gene attivo in entrambe le cellule e il nero un gene inattivo in entrambe le cellule.
Queste misure vanno aggiustate per considerare un eventuale disturbo di fondo causato da contaminazione del target, alta concentrazione di sale o detergente durante la fase di ibridizzazione oppure qualcuno degli altri problemi che si verificano involontariamente durante l'esecuzione degli esperimenti di laboratorio.

Analisi dei dati.

I valori di intensità luminosa ottenuti devono ora essere tradotti in un coefficiente numerico: tramite un'operazione di *grigliatura* vengono ritrovati sull'immagine gli spot corrispondenti alle sonde (operazione non particolarmente difficile poiché già si conosce la posizione degli spot sul microarray), poi si calcola l'intensità del verde, del rosso e del background e infine si separa il segnale dei marcatori fluorescenti da quello del background per isolare le quantità di interesse.
In quest'ultima fase, può risultare qualche valore negativo: la spiegazione sta nella maggior intensità del background rispetto agli spot, che in questo caso vengono trascurati e si attribuisce loro un valore piccolo e positivo a piacere.

Distorsioni.

Come precedentemente accennato, durante la preparazione ed esecuzione di un esperimento di laboratorio possono verificarsi involontariamente delle distorsioni sistematiche. Queste vanno considerate al momento del trattamento statistico dei dati e rimosse tramite normalizzazione, per rendere compatibili i risultati ottenuti su array diversi.

Vediamo ora i più noti effetti distorsivi:

1. Il *dye-effect* (effetto colore) è la diversa intensità di fluorescenza dei due marcatori: l'emissione di fluorescenza del rosso è infatti più intensa di quella del verde. Una soluzione sarebbe quella di ripetere due volte l'esperimento scambiando l'assegnazione dei marcatori trai due target; il principale inconveniente è tuttavia l'aumento dei costi dell'esperimento.
2. Il *print-tip* (deposito irregolare) è causato dalle microscopiche differenze delle puntine del robot che stampa l'array e consiste nella diversità di materiale genetico depositato sul vetrino.
3. L'*array-effect* (effetto intensità) consiste in differenze di intensità tra un array e l'altro, dovute alle diverse condizioni in cui si svolgono le varie fasi dell'esperimento.

Per risolvere i problemi suddetti, si ricorre alla normalizzazione e si calcolano fattori di standardizzazione per ciascun effetto distorsivo: al segnale viene sottratto una media generale di array, la differenza tra le medie degli spot stampati da ciascun print-tip e la media generale, e la differenza tra la media delle intensità con fluorescenza verde e rossa.
Per la scelta dei geni da usare nella standardizzazione, sono utilizzati tre approcci.
Il primo si basa sull'ipotesi che la parte di geni differenzialmente espressi sia piccola, per cui si possono utilizzare i restanti geni (aventi quindi livello di espressione costante) come indicatori dell'intensità relativa ai due colori, ovvero per la normalizzazione. Tuttavia, è molto difficile individuare nella pratica un gruppo di geni con segnale costante da cui ricavare un fattore di correzione; pertanto si preferisce usare tutti gli spot dell'array nella fase di normalizzazione, qualora il numero di geni differenzialmente espressi sia limitato rispetto al numero totale di geni considerati.

Il secondo approccio si basa sull'ipotesi opposta, ovvero che siano i geni con livello di espressione costante ad essere una piccola frazione del totale considerato; ma proprio il loro piccolo numero rappresenta il limite del loro utilizzo, dal momento che li rende poco rappresentativi nonché difficili da identificare.
Il terzo approccio prevede la realizzazione di un microarray per un singolo campione di mRNA, diviso in due parti uguali, ciascuna marcata con un colore diverso: lo scopo è sfruttare il fatto che si dovrebbe avere, dopo l'ibridizzazione, la medesima intensità degli spot per il rosso e per il verde, dato che il campione di provenienza è lo stesso; eventuali differenze possono essere utilizzate come fattore di standardizzazione.

Oltre ai procedimenti volti ad eliminare gli effetti distorsivi sopra descritti, viene solitamente effettuato un altro trattamento preliminare dei dati: la *filtrazione*. La sua funzione è ridurre la variabilità (tramite l'eliminazione dei geni con misura non sufficientemente accurata) e la dimensionalità dei dati (tramite l'imitazione dei geni con livello di espressione negativo o molto piccolo). Nella pratica vengono eliminati o sostituiti con un valore arbitrariamente piccolo tutti gli spot con differenza di intensità tra foreground e background inferiore a 1.4 fold; questa soglia trova giustificazione nell'evidenza empirica che livelli di espressione inferiori sono di solito dovuti ad errori di misura. Ovviamente, ogni operazione di filtrazione è soggetta ad arbitrarietà nella scelta delle soglie limite che regolano l'accettazione dei valori.

CAPITOLO III

La simulazione.

I dati derivanti da esperimenti di microarray possono essere analizzati utilizzando diverse tecniche statistiche; essi devono tuttavia essere calibrati a seconda del contesto.

In questo studio di simulazione ci si è occupati dell'identificazione dei geni differenzialmente espressi e si è scelta la verifica di ipotesi come metodologia di analisi. In particolare, sono state confrontate le prestazioni offerte da tre test statistici, ovvero il test t-Student, un test semiparametrico proposto dal prof. Adimari e dalla prof.ssa Chiogna e il test di Wilcoxon-Mann-Whitney.

Per ogni gene si è simulato un campione di livelli di espressione per m casi e di un campione di livelli di espressione di n controlli (nello studio si è supposto $n=m$).

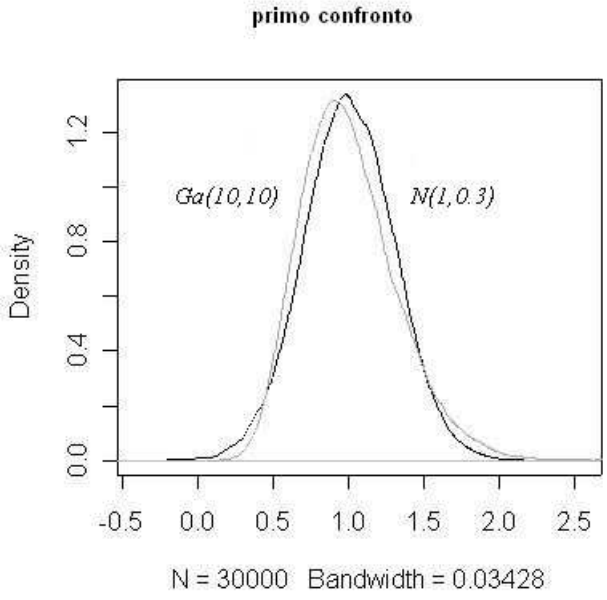
Nella simulazione sono stati considerati:

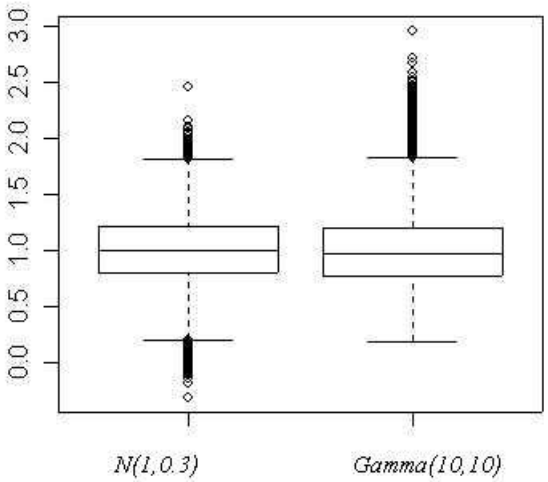
- un microarray di 2000 geni, di cui 100 differenzialmente espressi;
- 4 numerosità: 15, 25, 35, 100;
- 3 casi di distribuzione dei geni non differenzialmente espressi e di quelli differenzialmente espressi : *N(0,1)* e *N(1,1)* , *N(1,0.3)* e *N(2.3,0.3)* , *Gamma(10,10)* e *Gamma(10,4.35)*.

Per ogni caso ed ogni numerosità, è stato confrontato tramite la verifica d'ipotesi il primo array avente 2000 geni non differenzialmente espressi con il secondo array avente i primi 1900 non differenzialmente espressi e solo gli ultimi 100 differenzialmente espressi.

La distribuzione gamma è stata utilizzata nel caso 3 con l'intenzione di estendere la simulazione anche a distribuzioni diverse dalla normale.

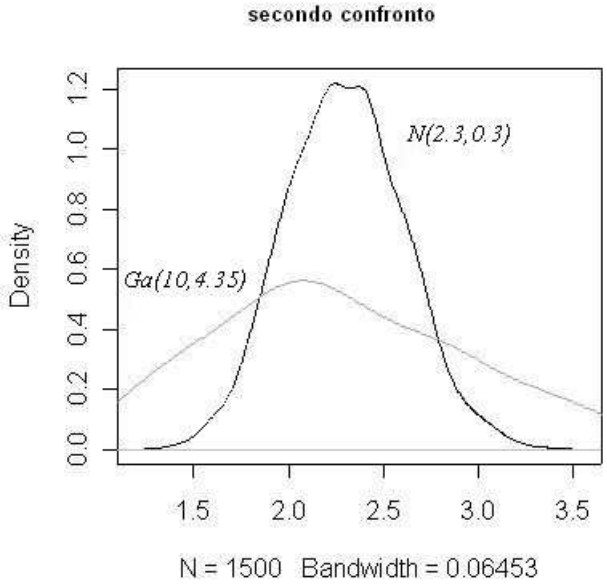
La funzione di distribuzione dei 2000 geni non differenzialmente espressi simulati dalla *Gamma(10,10)* è simile a quella dei 2000 geni non differenzialmente espressi simulati dalla *N(1,0.3)*:

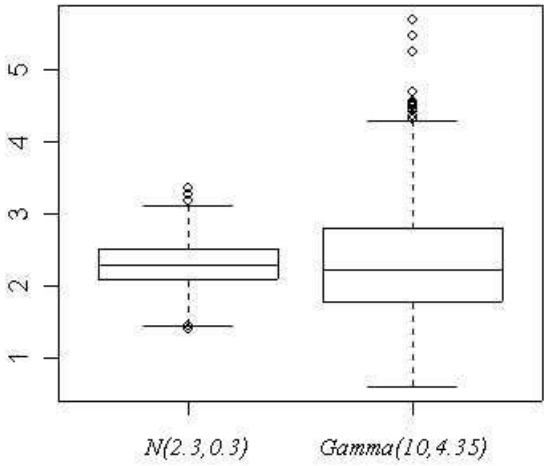




.

La funzione di distribuzione dei 100 geni differenzialmente espressi simulati dalla *Gamma(10,4.35)* ha un andamento sostanzialmente simile, anche se più variabile, rispetto a quella dei 100 geni differenzialmente espressi simulati dalla *N(2.3,0.3)* :





.

Per generare i dati ed effettuare le elaborazioni è stato utilizzato il programma R; i codici utilizzati sono riportati in appendice.

3.1 Test statistici utilizzati.

3.1.1 Il test t-Student.

Sia, per ogni gene, $y=(y_1,\ldots,y_m)$ il campione di livelli di espressione per gli m casi e $x=(x_1,\ldots,x_n)$ il campione relativo ai controlli; si suppone che y sia un campione casuale semplice da una distribuzione normale con parametri μ_y e σ^2_y e che x sia un campione casuale semplice da una distribuzione normale con parametri μ_x e σ^2_x .
L'ipotesi da verificare è:

$$\begin{cases} H_0 : \mu_x = \mu_y \\ H_1 : \mu_x \neq \mu_y \end{cases}$$

con cui si vuole accertare se i livelli di espressione medi differiscono o meno.
La statistica test utilizzata è:

$$t = \frac{\bar{y} - \bar{x}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

con $\bar{y} = \frac{\sum_{i=1}^{m} y_i}{m}$, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ e $s = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \bar{y}) + \sum_{i=1}^{n}(x_i - \bar{x})}{n+m-2}}$.

La statistica test si distribuisce sotto H_o come una t-Student con $(n+m-2)$ gradi di libertà, per cui il suo valore osservato viene confrontato con i quantili di una t-Student: si rifiuta H_0 (e quindi vi è differenza significativa tra le medie) se il valore osservato è maggiore di quello del corrispondente quantile, si accetta H_o (e quindi non vi è differenza significativa tra le medie) se è invece inferiore.

3.1.2 Il modello semiparametrico.

Come sopra, si dispone di due campioni x e y di numerosità rispettivamente m e n; tuttavia, mentre si suppone che Y abbia distribuzione $F_y(Y,\theta)$ nota (e precisamente

una distribuzione normale), sulla variabile X non è fatta alcuna assunzione distributiva.

L'ipotesi da verificare è:

$$\begin{cases} H_0 : \mu_x = \mu_y \\ H_1 : \mu_x \neq \mu_y \end{cases}$$

che corrisponde a

$$\begin{cases} H_0 : \Pr[X > Y; \vartheta] = \Pr[X < Y; \vartheta] \\ H_1 : \Pr[X > Y; \vartheta] \neq \Pr[X < Y; \vartheta] \end{cases}$$

la quale, ponendo $\Pr[X > Y; \vartheta] = \rho$, diventa

$$\begin{cases} H_0 : \rho = \rho_0 = 0.5 \\ H_1 : \rho \neq \rho_0 \end{cases} .$$

Per la stima dei parametri ϑ e ρ si utilizza la stima di massima verosimiglianza:

$$\hat{\vartheta} = (\hat{\mu}_x, \hat{\sigma}_x) = (\frac{1}{m}\sum_{i=1}^{m} y_i, \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{\mu}_y)^2)$$

$$\hat{\rho} = \frac{1}{n}\sum_{i=1}^{n} S(x_i; \vartheta) = \frac{1}{n}\sum_{i=1}^{n}(1 - F_Y(x_i; \vartheta)) = \frac{1}{n}\sum_{i=1}^{n}(1 - \phi(\frac{x_i - y_i}{\sigma_y}))$$

dove $\phi(.)$ è la funzione di ripartizione della normale standard.

La statistica test è

$$t = \frac{\hat{\rho} - \rho_0}{\varpi / \sqrt{n}}$$

che sotto H_0 segue una distribuzione normale standard, poiché $\sqrt{n}(\hat{\rho} - \rho_0) \sim N(0, \varpi^2)$.

Per la stima di ϖ^2 si ha

$$\hat{\varpi}^2 = \hat{\varpi}_s^2 + \frac{n}{m}\hat{\beta}^T \Omega \hat{\beta}$$

dove $\hat{\varpi}_s^2 = \sum_{i=1}^{n}(S(x_i; \vartheta))^2$, Ω è la matrice di varianze e covarianze di $\sqrt{n}(\hat{\rho} - \rho_0)$ (in questo caso è $\Omega = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$) e $\hat{\beta} = \begin{bmatrix} \frac{1}{n}\hat{\sigma}_y \sum_{i=1}^{n} \phi(\frac{x_i - \hat{\mu}_y}{\hat{\sigma}_y}) \\ \frac{1}{2n\hat{\sigma}_y^2}\sum_{i=1}^{n} \phi(\frac{x_i - \hat{\mu}_y}{\hat{\sigma}_y}) \end{bmatrix}$.

A questo punto si calcolano i limiti superiori e inferiori dell'intervallo di confidenza per ρ relativi al livello di significatività scelto; tuttavia, il metodo precedentemente descritto può fornire dei limiti superiori o inferiori esterni all'intervallo [0,1] , a cui appartiene ρ essendo definito come $\Pr[X > Y]$. Per ovviare a ciò, si utilizza una trasformazione. Nel caso di una trasformazione di tipo logit, ad esempio, si ottiene $\tau = \log(\frac{\rho}{1-\rho})$ e il sistema diventa così $\begin{cases} H_0 : \tau = \tau_0 \\ H_1 : \tau \neq \tau_0 \end{cases}$; sotto H_0 $\hat{\tau} = \log(\frac{\hat{\rho}}{1-\hat{\rho}})$ si distribuisce asintoticamente come una normale $N(0, \frac{\varpi^2}{np^2(1-p)^2})$ e tramite la standardizzazione si può ottenere una distribuzione normale standard.

Dai limiti superiore e inferiore dell'intervallo di confidenza per τ si ricavano, applicando la funzione inversa, quelli per ρ ed essi appartengono per costruzione all'intervallo [0,1].

3.1.3 Il test di Wilcoxon-Mann-Whitney.

È uno dei test non parametrici più potenti ed è una valida alternativa al test parametrico t-Student nel caso in cui si vogliano evitare le ipotesi di quest'ultimo: infatti la sua potenza-efficienza è del 95% su dati che potrebbero essere analizzati in modo più idoneo con il test t, sia per campioni di dimensioni elevate sia per campioni di dimensioni modeste.

Il test di Wilcoxon serve a verificare se due campioni indipendenti provengono dalla stessa popolazione.

Vediamo nel dettaglio: *x* e *y* sono quindi due campioni indipendenti e identicamente distribuiti (i.i.d.) delle variabili *X* e *Y*, hanno dimensione rispettivamente *n* e *m* (e inoltre $N=n+m$) e l'ipotesi nulla è che *X* e *Y* abbiano la stessa distribuzione.

Prima di applicare il test, le osservazioni di entrambi i gruppi vanno ordinate in ordine di grandezza crescente ed ogni osservazione viene poi sostituita con il rango ricevuto nella fase di ordinamento; le somme dei ranghi delle osservazioni dei due

gruppi sono definite W_x e W_y e la loro ulteriore somma è pari alla somma dei primi N numeri interi: $W_x + W_y = \frac{N(N+1)}{2}$.

Se l'ipotesi nulla fosse vera si dovrebbe osservare in media l'uguaglianza W_x tra W_y ; se invece la somma dei ranghi di un gruppo è molto grande (o molto piccola) allora è legittimo pensare che i campioni non appartengano alla stessa popolazione.

Consideriamo W_x : la sua distribuzione campionaria sotto H_0 è nota e inoltre se $n>10$ o $m>10$, per n e m crescenti essa tende alla distribuzione normale con media $\mu_{Wx} = \frac{n(N+1)}{2}$ e varianza $\sigma^2_{Wx} = \frac{nm(N+1)}{12}$; per determinare la significatività di un valore osservato di W_x si usa la statistica test $z_x = \frac{W_x \pm 0.5 - \mu_{W_x}}{\sigma^2_{W_x}}$ che ha distribuzione asintotica $N(0,1)$ e può quindi essere confrontato con i valori del tabulato della normale standard.

Analoghe considerazioni valgono per W_y .

3.1.4 Ranghi ripetuti nel test di Wilcoxon.

Il test di Wilcoxon presuppone che i valori del campione derivino da una distribuzione continua.

Se le misure di una variabile continua sono effettuate con precisione, è quasi nulla la probabilità di ottenere due valori uguali, detti "ties"; questo però non si verifica in campo sperimentale, dove le misurazioni sono un po' sommarie o le differenze tra due valori sono così piccole da non essere colte dagli strumenti, per cui si possono avere valori uguali.

Se i ranghi ripetuti si verificano tra osservazioni appartenenti allo stesso gruppo, il valore della W del gruppo non cambia; se invece i ranghi ripetuti interessano osservazioni di entrambi i gruppi, si determinano cambiamenti nel valore delle W e più precisamente nella variabilità della serie dei ranghi.

Anche se l'effetto è di solito trascurabile, è consigliata comunque la correzione per i ranghi ripetuti: la varianza della distribuzione campionaria di W_x (analogamente per W_y) diventa così $\sigma^2_{W_x} = \frac{nm}{N(N-1)}(\frac{N^3 - N}{12} - \sum_{j=1}^{g} \frac{t_j^3 - t_j}{12})$, dove $N=n+m$, g è il numero di raggruppamenti di ranghi replicati e t_j è il numero dei ranghi uguali nel raggruppamento *j-esimo*.

La statistica test si modifica di conseguenza ma, nel caso di assenza di ranghi ripetuti, la sua espressione si riduce a quella del paragrafo precedente per i ranghi non ripetuti. La correzione aumenta leggermente il valore della statistica test, così la probabilità associata ai dati osservati e non corretti risulta leggermente più grande rispetto a quanto si riscontra effettuando la correzione.

3.2 Risultati.

Per meglio rappresentare i risultati delle elaborazioni, sono di seguito riportate le tabelle Test/Realtà, che riportano il numero di veri positivi, veri negativi, falsi positivi e falsi negativi:

		Realtà	
		H_0	H_1
Test	H_0	Veri negativi =”d”	Falsi nagativi =”b”
	H_1	Falsi positivi =”c”	Veri positivi =”a”

Lo scopo del test è individuare i geni differenzialmente espressi e quindi si ha un “positivo” quando il test porta ad accettare l'ipotesi alternativa di non uguaglianza in media. Analogamente, si ha un “negativo” quando viene accettata l'ipotesi nulla. I risultati ottenuti dal test sono poi confrontati con i valori reali e perciò distinti in “veri” e “falsi”.

La sensibilità del test è data da $\frac{a}{a+c}$ e la specificità è data da $\frac{d}{d+b}$.

Caso 1: *N(0,1)* e *N(1,1)* :

Numerosità 15:

	H_0	H_1
H_0 t-Student	1806	20
H_0 test semiparametrico	1813	22
H_0 test Wilcoxon	1819	23
H_1 t-Student	94	80
H_1 test semiparametrico	87	78
H_1 test Wilcoxon	81	77

Numerosità 25:

	H_0	H_1
H_0 t-Student	1820	8
H_0 test semiparametrico	1820	9
H_0 test Wilcoxon	1818	12
H_1 t-Student	80	92
H_1 test semiparametrico	80	91
H_1 test Wilcoxon	82	88

Numerosità 35:

	H_0	H_1
H_0 t-Student	1802	1
H_0 test semiparametrico	1802	1
H_0 test Wilcoxon	1800	2
H_1 t-Student	98	99
H_1 test semiparametrico	98	99
H_1 test Wilcoxon	100	98

Numerosità 100:

	H_0	H_1
H_0 t-Student	1808	0
H_0 test semiparametrico	1804	0
H_0 test Wilcoxon	1809	0
H_1 t-Student	92	100
H_1 test semiparametrico	96	100
H_1 test Wilcoxon	91	100

Caso 2 : *N(1,0.3)* e *N(2.3,0.3)* :

Numerosità 15:

	H_0	H_1
H_0 t-Student	1812	0
H_0 test semiparametrico	1814	0
H_0 test Wilcoxon	1815	0
H_1 t-Student	88	100
H_1 test semiparametrico	86	100
H_1 test Wilcoxon	85	100

Numerosità 25:

	H_0	H_1
H_0 t-Student	1810	0
H_0 test semiparametrico	1825	0
H_0 test Wilcoxon	1820	0
H_1 t-Student	90	100
H_1 test semiparametrico	75	100
H_1 test Wilcoxon	80	100

Numerosità 35:

	H_0	H_1
H_0 t-Student	1801	0
H_0 test semiparametrico	1805	0
H_0 test Wilcoxon	1814	0
H_1 t-Student	99	100
H_1 test semiparametrico	95	100
H_1 test Wilcoxon	86	100

Numerosità 100:

	H_0	H_1
H_0 t-Student	1803	0
H_0 test semiparametrico	1803	0
H_0 test Wilcoxon	1798	0
H_1 t-Student	97	100
H_1 test semiparametrico	97	100
H_1 test Wilcoxon	102	100

Caso 3 : *Gamma(10,10)* e *Gamma(10,4.35)* :

Numerosità 15:

	H_0	H_1
H_0 t-Student	1817	0
H_0 test semiparametrico	1821	0
H_0 test Wilcoxon	1820	0
H_1 t-Student	83	100

H_1 test semiparametrico	79	100
H_1 test Wilcoxon	80	100

Numerosità 25:

	H_0	H_1
H_0 t-Student	1801	0
H_0 test semiparametrico	1809	0
H_0 test Wilcoxon	1815	0
H_1 t-Student	99	100
H_1 test semiparametrico	91	100
H_1 test Wilcoxon	85	100

Numerosità 35:

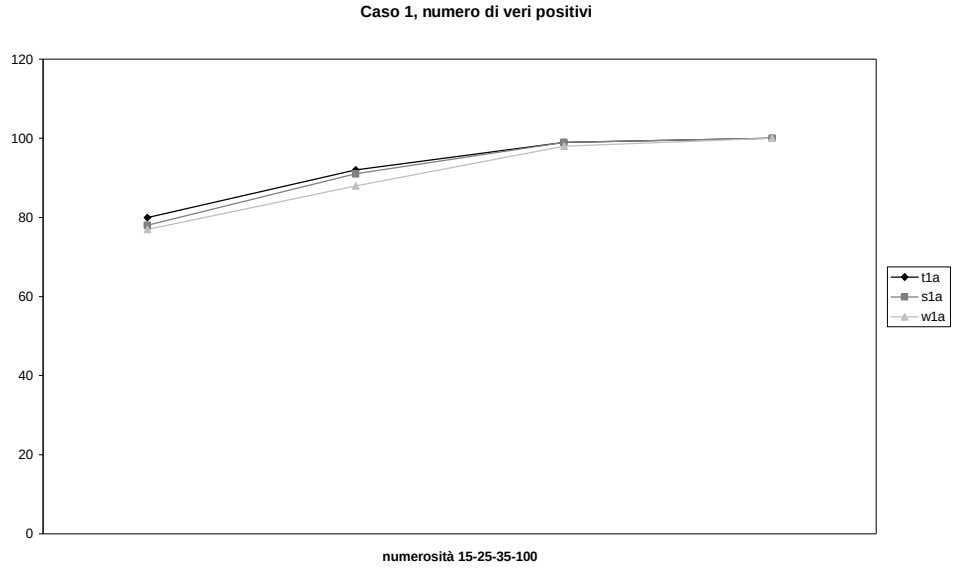
	H_0	H_1
H_0 t-Student	1812	0
H_0 test semiparametrico	1801	0
H_0 test Wilcoxon	1798	0
H_1 t-Student	88	100
H_1 test semiparametrico	99	100
H_1 test Wilcoxon	102	100

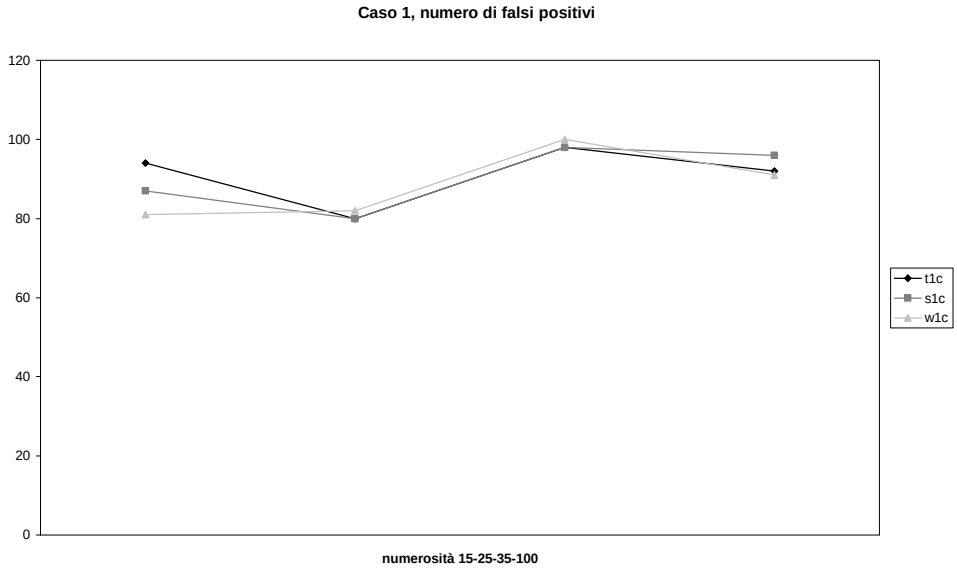
Numerosità 100:

	H_0	H_1
H_0 t-Student	1789	0
H_0 test semiparametrico	1778	0
H_0 test Wilcoxon	1796	0
H_1 t-Student	111	100
H_1 test semiparametrico	122	100

H_1 test Wilcoxon	104	100

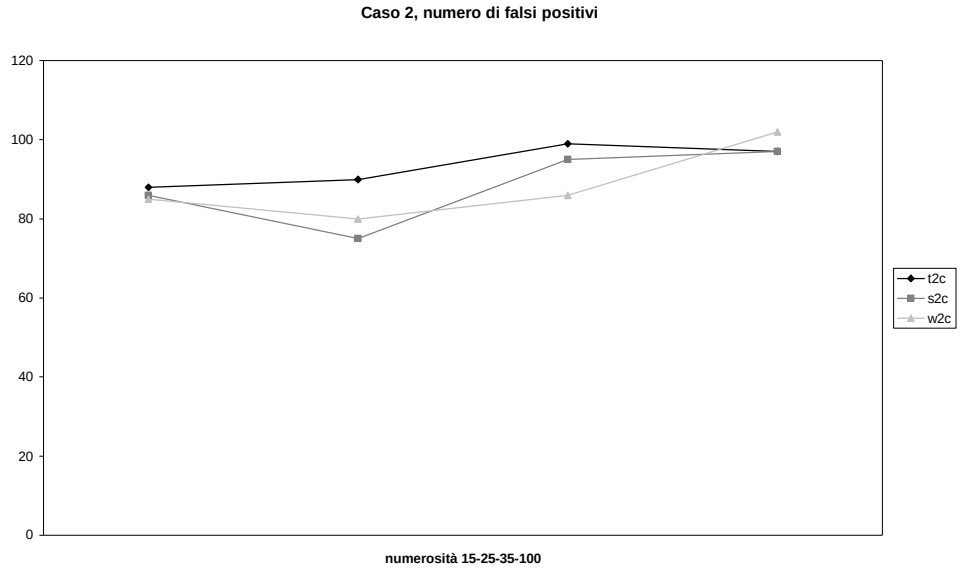
Nel caso 1, tutti i test aumentano il numero di veri positivi all'aumentare della numerosità, si comportano in maniera simile e convergono all'aumentare della numerosità; per quanto riguarda il numero di falsi positivi, hanno comportamento simile anche se c'è maggior divergenza per basse numerosità (n=15); non appare una sostanziale direzione evolutiva.



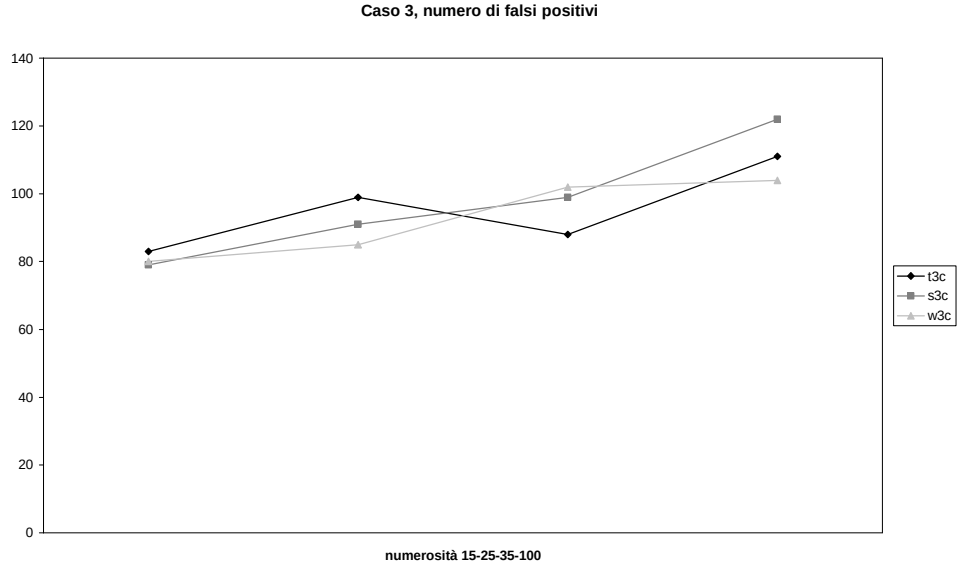


Nel caso 2 e 3, tutti i test hanno numero di falsi negativi pari a zero.

Nel caso 2, per quanto riguarda il numero di falsi positivi, il test semiparametrico e il test di Wilcoxon si comportano in maniera simile e presentano valori più bassi rispetto al test t (a parte per n=100 dove il test di Wilcoxon ha il valore più alto); si nota una leggera tendenza crescente.



Nel caso 3, per quanto riguarda il numero di falsi positivi, si riscontra tra i risultati dei test un andamento simile e una tendenza crescente; per basse numerosità (n=15) sono vicini (meglio i test semiparametrico e di Wilcoxon anche se di poco) poi divergono e il test semiparametrico presenta il valore peggiore e il test di Wilcoxon il migliore.



Di seguito sono riportati i valori di sensibilità e specificità:

Caso 1: *N(0,1)* e *N(1,1)* :

	n=15		n=25		N=35		N=100	
	Sensib.	Specifi.	Sensib.	Specifi.	Sensib.	Specifi.	Sensib.	Specifi.
t-Student	0.45977	0.98904	0.53488	0.99562	0.50253	0.99944	0.52083	1
Semiparametrico	0.47272	0.98801	0.53216	0.99507	0.50253	0.99944	0.51020	1
Wilcoxon	0.48734	0.98751	0.51764	0.99344	0.49494	0.99889	0.52356	1

Caso 2 : *N(1,0.3)* e *N(2.3,0.3)* :

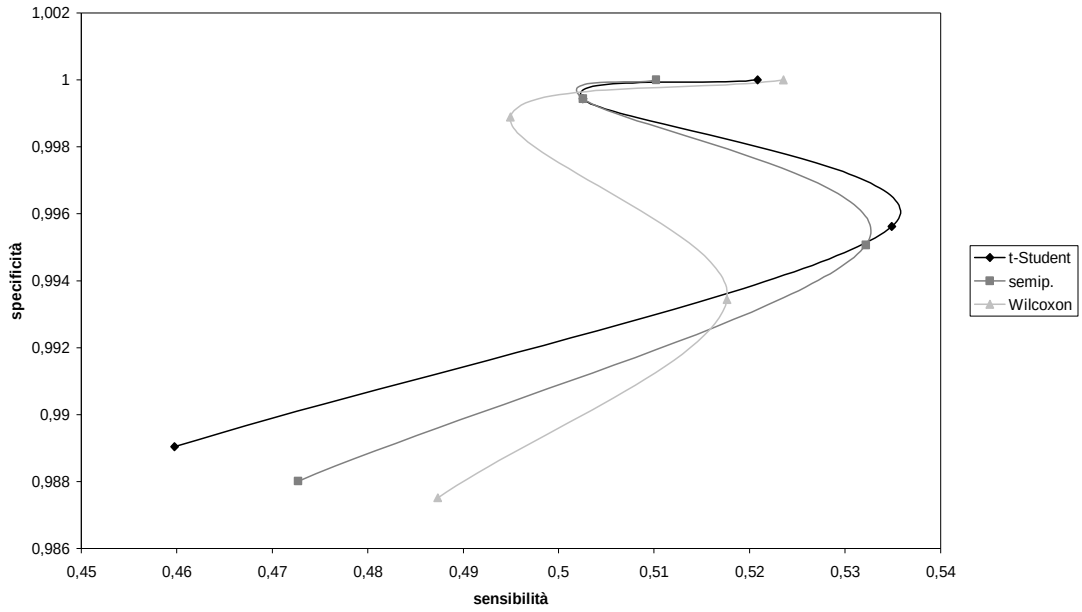
	n=15		n=25		N=35		N=100	
	Sensib.	Specifi.	Sensib.	Specifc.	Sensib.	Specifi.	Sensib.	Specifi.
t-Student	0.53191	1	0.52631	1	0.50251	1	0.50761	1
Semiparametrico	0.53763	1	0.57142	1	0.51282	1	0.50761	1
Wilcoxon	0.54054	1	0.55555	1	0.53763	1	0.49504	1

Caso 3 : *Gamma(10,10)* e *Gamma(10,4.35)* :

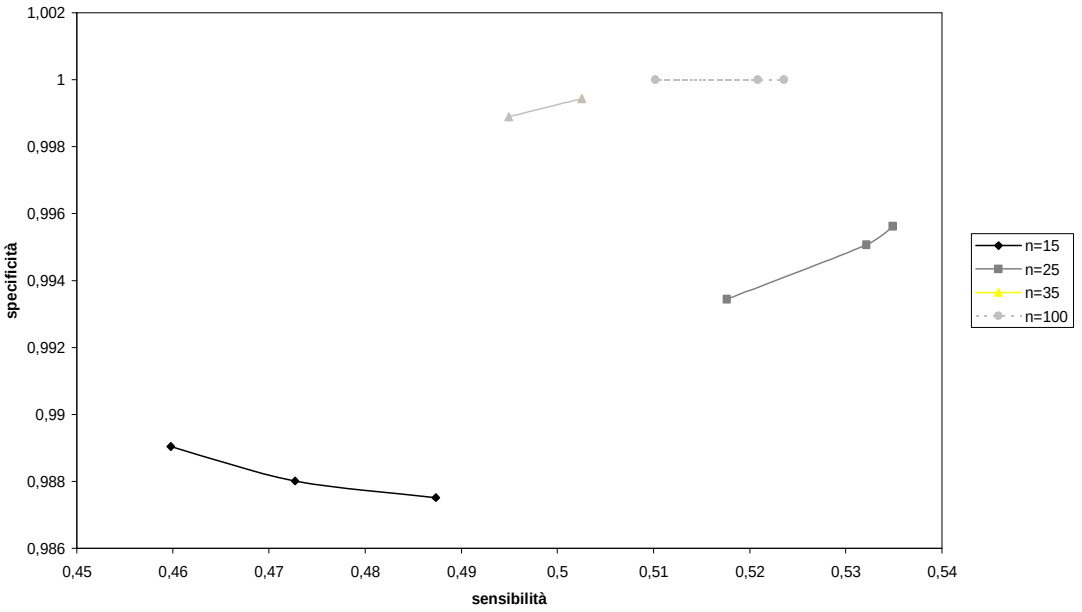
	n=15		n=25		N=35		N=100	
	Sensib.	Specifi.	Sensib.	Specifi.	Sensib.	Specifi.	Sensib.	Specifi.
t-Student	0.54644	1	0.50251	1	0.53191	1	0.47393	1
Semiparametrico	0.55865	1	0.52356	1	0.50251	1	0.45045	1
Wilcoxon	0.55555	1	0.54054	1	0.49504	1	0.49019	1

Nel primo caso, all'aumentare della numerosità da n=15 a n=25 si nota per tutti i test un aumento sia in sensibilità che in specificità; nel passaggio da n=25 a n=35 si osserva invece un'inversione di tendenza, per cui aumenta la specificità ma diminuisce la sensibilità; nel successivo aumento della numerosità da n=35 a n=100 la specificità si stabilizza sul valore massimo e la sensibilità torna ad aumentare. Complessivamente, i valori della specificità sono alti e quelli della sensibilità rimangono compresi nell'intervallo tra 0.46 e 0.54 ; si osserva una maggior variabilità tra gli andamenti per basse numerosità.

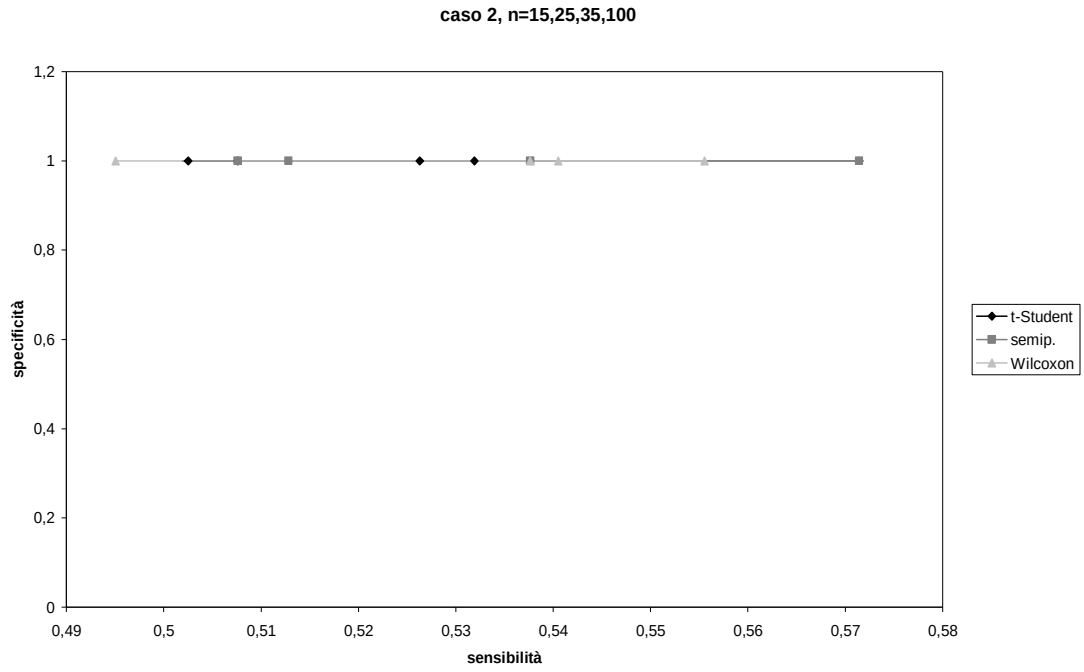
caso 1, n=15,25,35,100

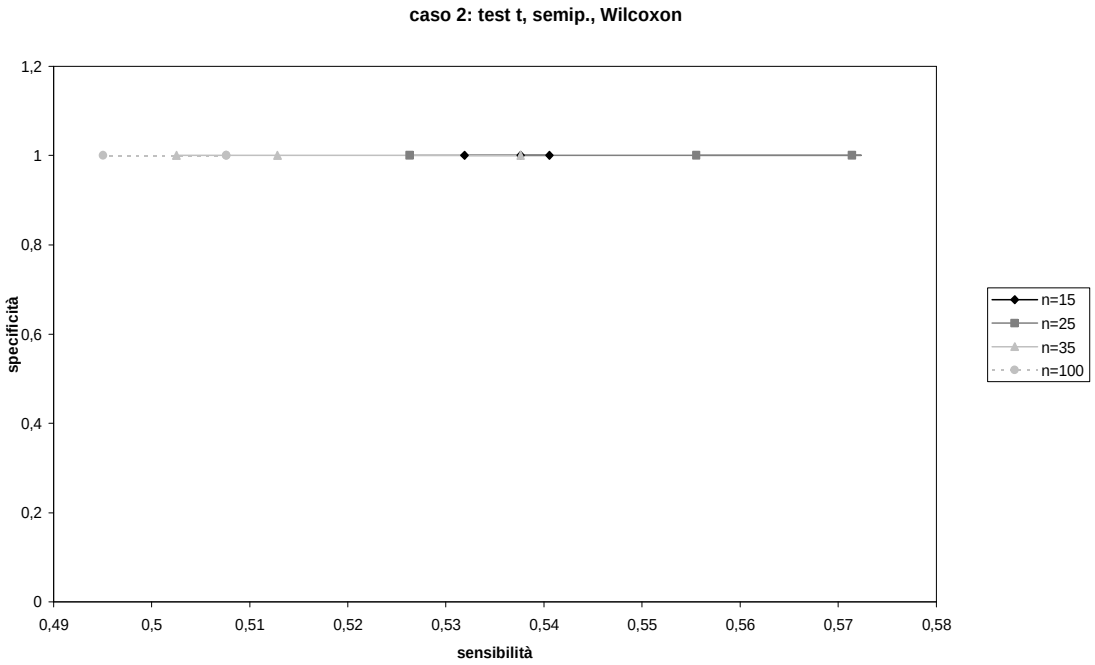


caso 1: test t, semip., Wilcoxon

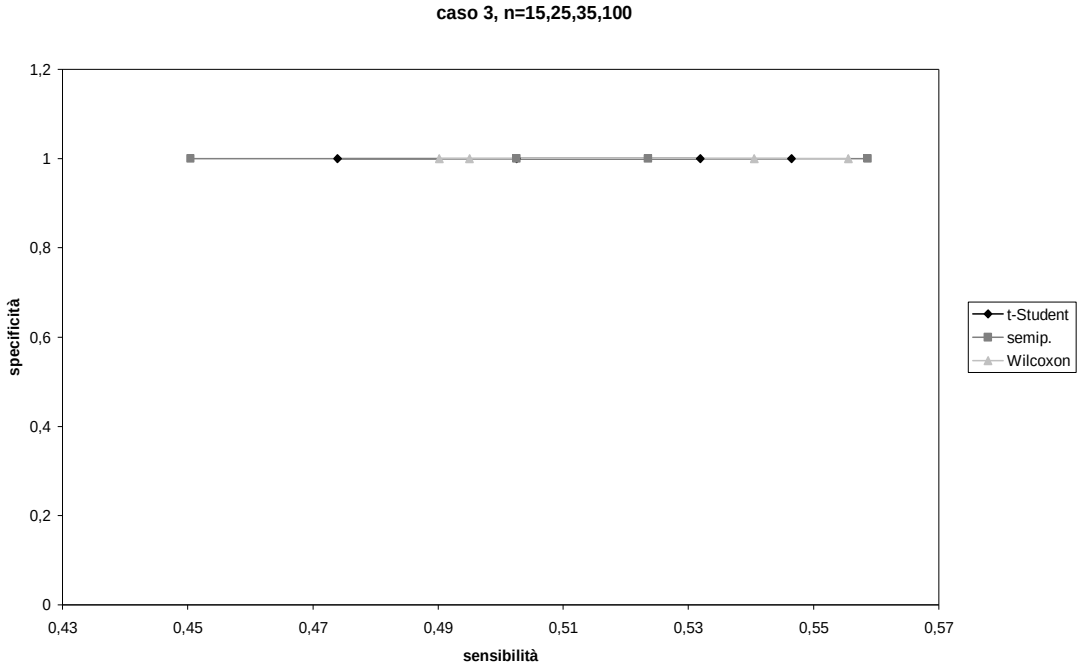


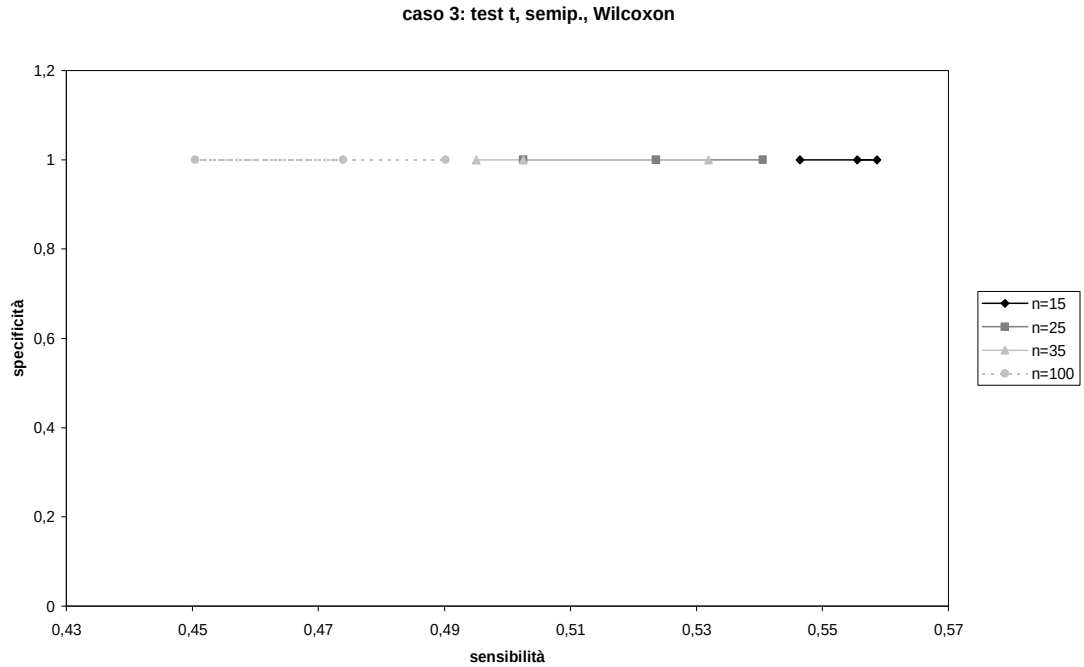
Nel secondo caso, per tutti i test la specificità rimane al livello massimo mentre la sensibilità tende a diminuire; i valori di sensibilità sono compresi nell'intervallo tra 0.49 e 0.58.





Nel terzo caso, i valori della specificità e della sensibilità sono simili a quelli riscontrati nel caso precedente.





RIFERIMENTI BIBLIOGRAFICI.

Gordon K. Smyth, Yee Hwa Yang and Terry Speed , “Statistical issues in cDNA microarray data analysis”, (2002).

David m: Rocke and Blythe Dubrin, “A model for measured for gene expression arrays”, (2001).

B. Lausen, “Statistical analysis of genetic distance data”.

T.R. Golub, D.K. Slomin et al, “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring”, (1999).

J.M. Clavarie, “Computational methods for identification of differential coordinated gene expression data”, (1999).

J. Platt, “Fast training of support vector machine using sequential minimal optimization”, (1999).

M. P. S. Brown, W.N. Grundy et al, “Knowledge-based analysis of microarray gene expression data by using support vector machines”, (2000).

N. Friedman et al, “Tissue classification with gene expression profiles”,(2000).

A. A. Alizade, M. B. Elisen, R. E. Davis et al ”Distinct types of diffuse large B-cell lymphoma identified by gene expression profile”, (2000).

A.Keller, M. Shummer, L. Hood, W. Ruzzo, “Bayesian classification of DNA array expression data”, Technical report, University of Washington (Aug2000).

Gen Hori, Masato Inoue, Shin-ichi Nishimura and Hiroyuki Nakahara, “Bild gene classification based on ICA of microarray data”.

Sidney Siegel et al, “Statistica non parametrica”, (1992).

Peter Armitage and Geoffrey Berry, “Statistica medica”, (1996).

Harsko Ervin, “Identificazione di geni differenzialmente espressi: uno studio di simulazione”, tesi di laurea in STI, Università di Padova, A.A. 2004-2005

www.transcriptome.ens.fr/sgdb/presentation/principle.php

www.genelex.com

www.sanger.ac.uk/Teams/Team52/currentprojects/microarrayer.shtml

APPENDICE

Funzione “semiparametric-test” :

```
semiparametric.test<-function(x,y,level=0.05)
{
#salva i quantili a seconda del livello di significativita' epresso in "level"
pval<-level/2
z1<-qnorm(pval)
z2<-qnorm(1-pval)

#se le matrici hanno i pazienti sulle righe,le traspone
#salva in "n" e "m" il num. di pazienti di cuascuna patologia
n1<-nrow(x)
n2<-ncol(x)
m1<-nrow(y)
m2<-ncol(y)
if(n1<n2){
x<-t(x)
n<-ncol(x)}
if(n1>=n2){
n<-n2}
if(m1<m2){
y<-t(y)
m<-ncol(y)}
if(m1>=m2){
m<-m2}
ngen<-nrow(y)

#fa corrispondere a "x" la matrice con un magg.num. di pazienti e ad "y" quella
#con min.num.
if(m>n){
z<-x
x<-y
y<-z
rm(z)
n<-ncol(x)
m<-ncol(y)}

#calcolo di media,var,sd(y),rho
#mean,sd,var,rho sono vettori p-dim. ,S matice pxn
mean<-apply(y,1,mean)
var<-apply(y,1,var)
var<-((m-1)/m)*var
sd<-sqrt(var)
S<-1-(apply(x,2,pnorm,mean=mean,sd=sd))
rho<-1/n*(apply(S,1,sum))
```

```
#calcolo di sd(rho):
##calcolo di w^2 esse
g<-(S-rho)**2
w2s<-(1/(n-1))*(apply(g,1,sum))
##matrice "omega"
omega1<-var
omega2<-2*(var^2)
##calcolo primo elemento del vettore beta
zi<-(x-mean)/sd
densxi<-apply(zi,2,dnorm)
beta1<-(1/(n*sd))*apply(densxi,1,sum)
##calcolo secondo elemento del vettore beta
si<-zi*densxi
beta2<-(1/(2*n*var))*apply(si,1,sum)
##infine sd(rho)
p1<-omega1*(beta1^2)
p2<-omega2*(beta2^2)
p<-(n/m)*(p1+p2)
var.rho<-(w2s+p)

#passaggio al logit
tau<-log(rho/(1-rho))
var.tau<-(var.rho)/(rho^2*(1-rho)^2*n)
toss.tau<-(tau)/sqrt(var.tau)
tau.inf<-z1*sqrt(var.tau)
tau.sup<-z2*sqrt(var.tau)
rho<-exp(tau)/(exp(tau)+1)
rho.inf<-exp(tau.inf)/(exp(tau.inf)+1)
rho.sup<-exp(tau.sup)/(exp(tau.sup)+1)
test<-toss.tau
var.test<-var.tau

#calcolo valori p e conta dei geni significativi
test<-toss.tau
non.agg<-2*(1-pnorm(abs(test)))
pvalue<-non.agg

list(rho=rho, rho.inf=rho.inf, rho.sup=rho.sup,   test=test, var.test=var.test,
pvalue=pvalue)
}
```

Codice con cui sono stati generati i dati ed effettuati le verifiche di ipotesi:

per i casi 1 e 2, con distribuzione normale:

```
library(sm)
```

```
# ini.  variabili:
f<-n*1900
o<-n*100
test<-rep(0,2000)
prob<-rep(0,2000)
wiltest<-NULL
wilprob<-NULL

x<-rnorm(f,m,s)
x<-matrix(x,nrow=1900,ncol=n)

h<-rnorm(o,m,s)
h<-matrix(h,nrow=100,ncol=n)

L<-rnorm(f,m,s)
L<-matrix(L,nrow=1900,ncol=n)

G<-rnorm(o,mw,sw)
G<-matrix(G,nrow=100,ncol=n)

# matrici con i dati:
w<-rbind(x,h)
p<-rbind(L,G)

# t.test tra w[i] e p[i] :
for(i in 1:2000){
test[i]<-t.test(w[i,],p[i,],var.equal=T)$statistic
prob[i]<-t.test(w[i,],p[i,],var.equal=T)$p.value
}

# matrici dei risultati del ciclo for precedente:
tau<-matrix(test,ncol=1,byrow=T)
tai<-matrix(prob,ncol=1,byrow=T)

# wilcox test tra w[i] e p[i] :
for(i in 1:2000){
wiltest[i]<-wilcox.test(w[i,],p[i,])$statistic
wilprob[i]<-wilcox.test(w[i,],p[i,])$p.value
}

# test semiparametrico
sem<-semiparametric.test(w,p)
c<-sem$test
r<-sem$pvalue
c<-matrix(c,ncol=1,byrow=T)
r<-matrix(r,ncol=1,byrow=T)
```

```
# salva dati in files
write.table(round(tau,6),"c:\\dati-tesi\\ttest.txt",sep="
",row.names=F,col.names=F)
write.table(round(tai,6),"c:\\dati-tesi\\tprob.txt",sep="
",row.names=F,col.names=F)
write.table(round(w,6),"c:\\dati-tesi\\dati0101.txt",sep="
",row.names=F,col.names=F)
write.table(round(p,6),"c:\\dati-tesi\\dati0111.txt",sep="
",row.names=F,col.names=F)
write.table(round(G,6),"c:\\dati-tesi\\dati--11.txt",sep="
",row.names=F,col.names=F)
write.table(round(wiltest,6),"c:\\dati-tesi\\wiltest.txt",sep="
",row.names=F,col.names=F)
write.table(round(wilprob,6),"c:\\dati-tesi\\wilprob.txt",sep="
",row.names=F,col.names=F)
write.table(round(c,6),"c:\\dati-tesi\\semtest.txt",sep="
",row.names=F,col.names=F)
write.table(round(r,6),"c:\\dati-tesi\\semprob.txt",sep="
",row.names=F,col.names=F)
```

per il caso 3 con distribuzione gamma:

```
library(sm)

# ini.  variabili:
f<-n*1900
o<-n*100
test<-rep(0,2000)
prob<-rep(0,2000)
wiltest<-NULL
wilprob<-NULL

x<-rgamma(f,m,s)
x<-matrix(x,nrow=1900,ncol=n)

h<-rgamma(o,m,s)
h<-matrix(h,nrow=100,ncol=n)

L<-rgamma(f,m,s)
L<-matrix(L,nrow=1900,ncol=n)

G<-rgamma(o,mw,sw)
G<-matrix(G,nrow=100,ncol=n)

# matrici con i dati:
w<-rbind(x,h)
p<-rbind(L,G)
```

```
# t.test tra w[i] e p[i] :
for(i in 1:2000){
test[i]<-t.test(w[i,],p[i,],var.equal=T)$statistic
prob[i]<-t.test(w[i,],p[i,],var.equal=T)$p.value
}

# matrici dei risultati del ciclo for precedente:
tau<-matrix(test,ncol=1,byrow=T)
tai<-matrix(prob,ncol=1,byrow=T)

# wilcox test tra w[i] e p[i] :
for(i in 1:2000){
wiltest[i]<-wilcox.test(w[i,],p[i,])$statistic
wilprob[i]<-wilcox.test(w[i,],p[i,])$p.value
}

# test semiparametrico
sem<-semiparametric.test(w,p)
c<-sem$test
r<-sem$pvalue
c<-matrix(c,ncol=1,byrow=T)
r<-matrix(r,ncol=1,byrow=T)

# salva dati in files
write.table(round(tau,6),"c:\\dati-tesi\\ttest.txt",sep="
",row.names=F,col.names=F)
write.table(round(tai,6),"c:\\dati-tesi\\tprob.txt",sep="
",row.names=F,col.names=F)
write.table(round(w,6),"c:\\dati-tesi\\dati0101.txt",sep="
",row.names=F,col.names=F)
write.table(round(p,6),"c:\\dati-tesi\\dati0111.txt",sep="
",row.names=F,col.names=F)
write.table(round(G,6),"c:\\dati-tesi\\dati--11.txt",sep="
",row.names=F,col.names=F)
write.table(round(wiltest,6),"c:\\dati-tesi\\wiltest.txt",sep="
",row.names=F,col.names=F)
write.table(round(wilprob,6),"c:\\dati-tesi\\wilprob.txt",sep="
",row.names=F,col.names=F)
write.table(round(c,6),"c:\\dati-tesi\\semtest.txt",sep="
",row.names=F,col.names=F)
write.table(round(r,6),"c:\\dati-tesi\\semprob.txt",sep="
",row.names=F,col.names=F)
```

Funzione “pat”, con cui si ottiene il numero di falsi negativi e veri negativi:

```
pat<-function(TestPValue){
r1<-TestPValue<0.05     # mi da' true(rifiuto H0) o false
```

```
r2<-r1[1901:2000]  #prendo gli ultimi 100
r1<-r1[1:1900]    #prendo gli altri
mode(r1)<-"numeric"
mode(r2)<-"numeric"
#false=0,true=1
c<-sum(r1)
a<-sum(r2)
list(c,a)
}

c(pat(tai),pat(r),pat(wilprob))
```

Dati iniziali:

Caso 1:	Caso 2:	Caso 3(gamma):
n=15,25,35,100	n=15,25,35,100	n=15,25,35,100
m=0	m=1	m=10
s=1	s=0.3	s=10
mw=1	mw=2.3	mw=10
sw=1	sw=0.3	sw=4.35

Codice con cui sono stati ottenuti i grafici per confrontare la N(1,0.3) con la Gamma(10,10) e la N(2.3,0.3) con la Gamma(10,4.35):

Primo confronto:

```
#legge i dati
y<-scan("c:\\dati-tesi\\Ga1010-15\\dati0101.txt")
x<-scan("c:\\dati-tesi\\N103-15\\dati0101.txt")
##(legge i dati,che nel file sono in matrice,riga per riga
##per cui gli ultimi n*100 valori sono le ultime 100 righe)
#fa i grafici a colori
plot(density(x),col=2)
lines(density(y),col=4)
#bianco e nero
plot(density(x),col="black")
lines(density(y),col="dark grey")
```

Secondo confronto:

```
#legge i dati
a<-scan("c:\\dati-tesi\\N103-15\\dati--11.txt")
b<-scan("c:\\dati-tesi\\Ga1010-15\\dati--11.txt")
#fa i grafici a colori
```

```
plot(density(a),col=2)
lines(density(b),col=4)
#bianco e nero
plot(density(a),col="black")
lines(density(b),col="dark grey")
```

Boxplot:

```
boxplot(x,y,names=c("x","y"))
boxplot(a,b,names=c("a","b"))
```