Molecular classification of cancer types from microarray data using the combination of Genetic Algorithms and Support Vector Machine

Sihua Penga, Qianghua Xub, Xuefeng Bruce Lingc, Xiaoning Pengd, Wei Dua, Liangbiao Chenb;

Published by Elsevier on behalf of the Federation of European Biochemical Societies, 2003 Dec 4, 555(2):358-362

Obiettivi

- Creare un algoritmo in grado di effettuare una classificazione molecolare simultanea di campioni provenienti da diverse tipologie di tumore sulla base dell'analisi d'espressione, tramite microarrays, di migliaia di geni.

- Identificare il miglior set di geni predittivi della patologia, da utilizzare sia per poter effettuare diagnosi da affiancare al consueto approccio terapeutico, che per una più approfondita comprensione dei meccanismi patologici della malattia.

Introduzione

- L’analisi di dati d’espressione da Microarray fa parte delle tecnologie definite di “livello genomico”, con le quali è possibile studiare contemporaneamente l’espressione di migliaia di geni.

- Diversi studi hanno creato algoritmi in grado di effettuare su questi dati una classificazione di tipo binario (es. sano-tumore).

- Più difficile si è rivelata invece la classificazione multiclasse, (dove più tipologie di tumore sono confrontate contemporaneamente), che è importante target di ricerca a causa della eterogenicità morfologica e di decorso clinico presente nelle neoplasie.

Casistica utilizzata

4 data set provenienti da precedenti studi:

1. 58 campioni di linee cellulari tumorali (9 tipi) dal National Cancer Institute (NCI60), di cui è stata analizzata l'espressione di circa 6831 geni (casistica multiclasse).

2. 308 campioni (218 da 14 differenti tumori, 90 da tessuto sano) dal GCM data set, 16063 geni (casistica multiclasse).

3. 38 campioni di leucemie, 27 di AML (acute myeloid leukemia), 11 di ALL (acute lymphoblastic leukemia), 6817 geni (casistica binaria).

4. 40 campioni di tumore al colon e 20 da tessuto normale, 6500 geni (casistica binaria).

Step

a) Acquisizione dei dati d'espressione provenienti da microarray dei data set dei 4 studi.

b) Analisi dei data set con una combinazione di algoritmi per suddividere i campioni nelle classi tumorali.

c) Identificazione di un set di geni predittivo per la patologia.

d) Clustering sui campioni con il set genico predittivo.

e) Confronto della validità del risultato ottenuto con altri precedenti studi tramite LOOCV.

f) Valutazione dei meccanismi molecolari dei set di geni identificati come predittori per le diverse patologie.

Algoritmi utilizzati

1) Support Vector Machine: per la classificazione binaria

2) Genetic Algorithm: Per identificare il miglior set genico predittore

3) Recursive Feature Elimination: per eliminare i set genici con bassa fitness

4) Leave-One-Out Cross-Validation: per testare e confrontare il risultato

Strategia d'analisi dei dati

1. Pre-filtraggio nelle due casistiche multiclasse (NCI60 e GCM) per eliminare i geni con bassa differenza d'espressione.

2a. Classificazione binaria dei campioni di ogni casistica mediante AP/SVM.

2b. "Voting scheme" per passare dal dato binario ad una suddivisione multiclasse basato sui risultati della fase precedente.

3. Ricerca dei set genici predittivi mediante GA.

4. Eliminazione attraverso l'uso di RFE dei geni non predittivi dal dat set ottenuto con GA.

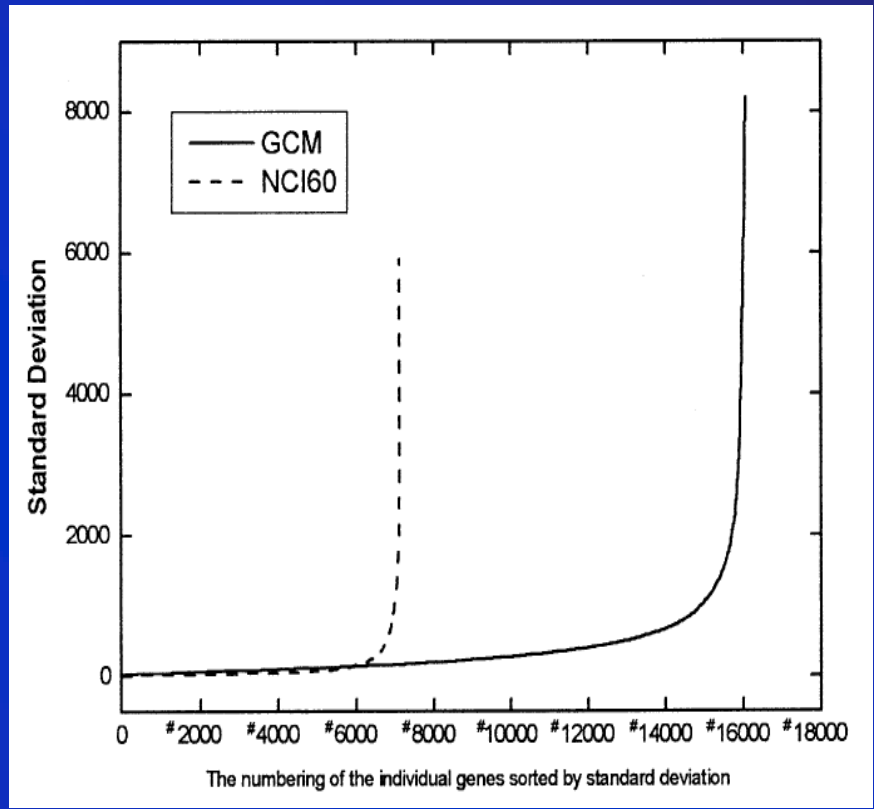
Fase 1: pre-filtraggio

Analisi della deviazione standard dei geni all'interno delle casistiche GCM e NCI60.

Eliminazione dei geni con bassa DS tra i diversi tipi di tumore.

Passano il filtro

- 1994 geni per NCI60
- 200 geni per GCM



Numero dei geni dei due data set rapportato alla DS

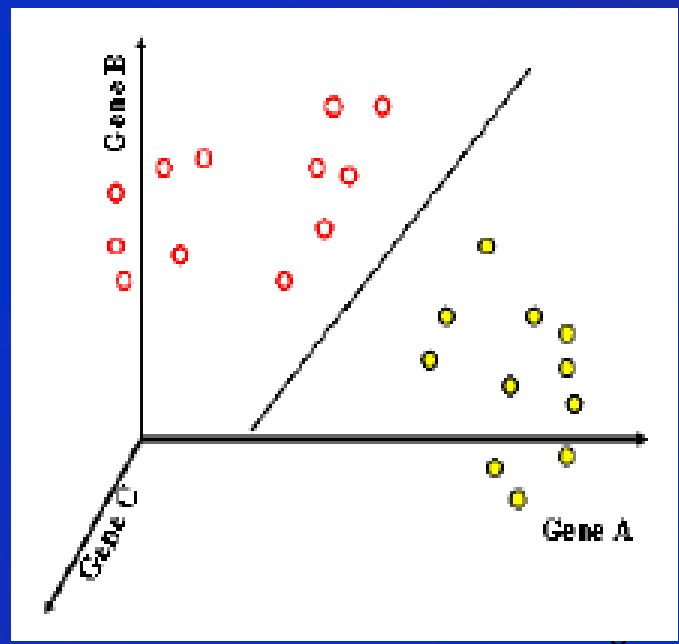
Support Vector Machine

- E’ un efficace metodo di tipo “supervised” utilizzato per la classificazione binaria dei campioni.

- I campioni vengono distribuiti su uno spazio multidimensionale e viene ricercato un “iperpiano” lineare in grado di separare efficacemente i campioni nelle 2 classi.

- Dove non è possibile una separazione lineare, la SVM può utilizzare diverse funzioni fino a trovare la più efficace.

- Più ampio sarà il margine tra i due gruppi di campioni, migliore sarà l’iperpiano (o classificatore) scelto.



Fase 2a: classificazione binaria

- Viene fatto un approccio definito All Paired: per ogni data set con la SVM vengono confrontati a coppie tutti le classi di tumore.

- Vengono costruiti per ogni data set un totale di *K(K-1)/2* classificatori (con *K* = numero di tipi di tumore della specifica casistica).

- Sono testati tre differenti tipi di funzione per identificare il classificatore migliore: lineare, polinomiale e sigmoide.

- Quello polinomiale è risultato il più efficace (di 4° grado).

Fase 2b: Voting Scheme e Multiclasse

- E’ stato utilizzato un sistema di punteggi per passare dalla classificazione binaria a quella multiclasse (per GCM e NCI60).

- Ognuna delle classificazioni binarie equivale per ogni campione ad un voto da attibuire alle classi tumorali (n. di classificazione per ogni data set = k(k-1)/2).

- Ognuna delle classi (e dei rispettivi campioni) viene analizzata un numero pari a k-1 volte.

- Il campione viene infine attribuito alla classe per la quale ha ottenuto il punteggio più alto.

BIOINFORMATICS
Vol. 19 no. 1 2003
Pages 37–44

Genetic algorithms applied to multi-class prediction for the analysis of gene expression data

C.H. Ooi[1] *and Patrick Tan*[2,*]

- Viene utilizzata una strategia chiamata GA/MLHD (Maximum Likelihood) dove gli algoritmi genetici sono utilizzati per trovare i set genici predittivi e il maximum likelihood come classificatore

- Il primo algoritmo trova un set di geni *R* in grado di classificare i campioni, mentre il secondo viene utilizzato per ottimizzare il processo di classificazione.

- Le casistiche analizzati sono sempre NCI60 e GCM, ma utilizzando in entrambi i casi solo i 1000 geni con la maggior DS (indicati con numeri progressivi da 1 a 1000).

Algoritmo Genetico

- Ogni stringa S_i è composta da R_{max}+1 elementi, dove il primo elemento indica il n. di geni che compongono la stringa e gli altri elementi (g_1 to g_{Rmax}) sono numeri random tra 1 e 1000 (i geni).

$$[R \quad g_1 \quad g_2 \ldots g_{R_{\max}}].$$

- R_{max} e R_{min} sono il range il numero di geni per il quale sono state valutate numerose combinazioni ([5,10], [11,15], [16,20], [21,25], [26,30]), trovando la 2ª e la 3ª combinazione come le più efficaci.

- La fitness di ogni stringa è valuta con la formula

$$f(S_i) = 200 - (E_C + E_I)$$

dove *Ec* = cross validation error rate, *El* = indipendent test error rate, che viene calcolato utilizzando i geni presenti nella stringa come variabili per classificare i campioni secondo l'algoritmo MHLD.

Algoritmo Genetico

- Innanzitutto viene creata la popolazione iniziale da un numero *N* (specificato dall'operatore) di stringhe random che rappresentano i set genici potenzialmente predittivi.

- La generazione successiva viene creata incrociando casualmente tutti gli individui della vasca: la selezione degli individui è stata testata sia con metodo SUS che RWS, trovando come più efficace il primo.

- Il crossover è stato fatto prendendo due individui a caso della popolazione con probabilità $P_c > 0,8$ (range testato 0,7-1,0). Sono stati considerati due tipi di crossover: one-point e uniforme, hanno verificato che il primo è più efficace nei range estremi di lunghezza delle stringhe, e il secondo (usato poi nella soluzione finale) in quelli intermedi.

Algoritmo Genetico

- Agli individui della generazione successiva è stata inoltre applicata una mutazione con probabilità P_m, testata tra 0,0005 e 0,01, è stata poi settata a 2×10^{-3} .

- Tutte le possibili variazioni dei parametri:
  i. Pc
  ii. Metodo di crossover
  iii. Pm
  iv. Metodo di selezione
  v. Range Rmin e Rmax per ogni stringa
  vi. Data set troncato (1000 geni) Vs. completo (6167 geni)

  sono state testate variando singolarmente ognuno dei parametri, ottenendo un totale di 96 differenti combinazioni.

Algoritmo Genetico

- Alla fine di tutte le generazioni viene confrontata la migliore stringa di ogni singola generazione e quella con la massima fitness tra tutte le generazioni viene presa come soluzione, quindi non per forza quella migliore dell'ultima generazione.

- Il data set genico finale è composto da 13 geni; utlizzato nell'analisi di clustering è stato in grado di suddividere i campioni di entrambe le casisticche nella corretta classe.

Fase 3: Genetic Algorithm

- Selezione degli individui dalla vasca, testati due diversi metodi:
  - Stochastic Universal Sampling (SUS)
  - Roulette Wheel Selection (RWS)

RWS ha dato risultati migliori.
(Con questo metodo gli individui con fitness maggiore hanno una probabilità più alta di passare alla generazione successiva).

- Cross-Over: hanno verificato come migliore quello uniforme rispetto a quello single-point. Si sono ottenuti risultati migliori con valori Pc >0.8.

- Tasso di mutazione: testati diversi valori compresi tra 0.0005 e 0.02. Con valori >0.01, la fitness diventava molto bassa. Con valori <0.001, la fitness non migliorava anche dopo molte generazioni. Valori ottimali tra 0.004 e 0.006.

Fase 3: Genetic Algorithm

- Il numero di generazioni è stato testato a partire da 100.000, aggiustato mediante LOOCV e sono stati ottenuti diversi valori per i 4 data set:
  - 31527 per i tumori del colon
  - 1219 per le AML e ALL
  - 6729 per NCI60
  - 12765 per GCM

  } Classificazione binaria

  } Classificazione multiclasse

- Testate diverse dimensioni della popolazione (tra 6 e 40): utilizzati 12 individui per la classificazione binaria, 30 per la multiclasse.

- Dimensione dei cromosomi utilizzata: tra 36 e 40 geni.

LOOCV: cross validazione dei dati dove ogni singola osservazione viene utilizzata come validazione dei dati, mentre tutte le rimanenti sono usate come training set. L'operazione viene ripetuta fino all'uso di tutte le singole osservazioni come validation data.

Fase 4: RFE post-processing

- RFE viene utilizzato per performare i set genici definiti come predittivi dall'Algoritmo Genetico.

- In ogni set genico vengono eliminati i geni con la minor predittività nel data set stesso, calcolando dopo ogni iterazione il nuovo valore di LOOCV.

- Se l'analisi con LOOCV dimostra un miglioramento, il gene viene eliminato dal set di geni predittivi.

- Questa fase viene ripetuta iterativamente fino ad identificare il minore data set genico predittivo.

Recursive Feature Elimination

Multiclass cancer diagnosis using tumor gene expression signatures

Sridhar Ramaswamy, Pablo Tamayo, Ryan Rifkin, Sayan Mukherjee, Chen-Hsiang Yeang, Michael Angelo, Christine Ladd, Michael Reich, Eva Latulippe, Jill P. Mesirov, Tomaso Poggio, William Gerald, Massimo Loda, Eric S. Lander, and Todd R. Golub

PNAS 2001;98;15149-15154; originally published online Dec 11, 2001;
doi:10.1073/pnas.211566398

- Creatori del GCM data set, progettano un algoritmo (SVM-OVA, one-versus-all) per la classificazione multiclasse dei tumori.

- RFE viene utilizzata dopo l'analisi con SVM dei campioni, per riuscire a definire attraverso più iterazioni il miglior numero di geni da utilizzare nella classificazione dei campioni stessi.

Riassunto dei risultati

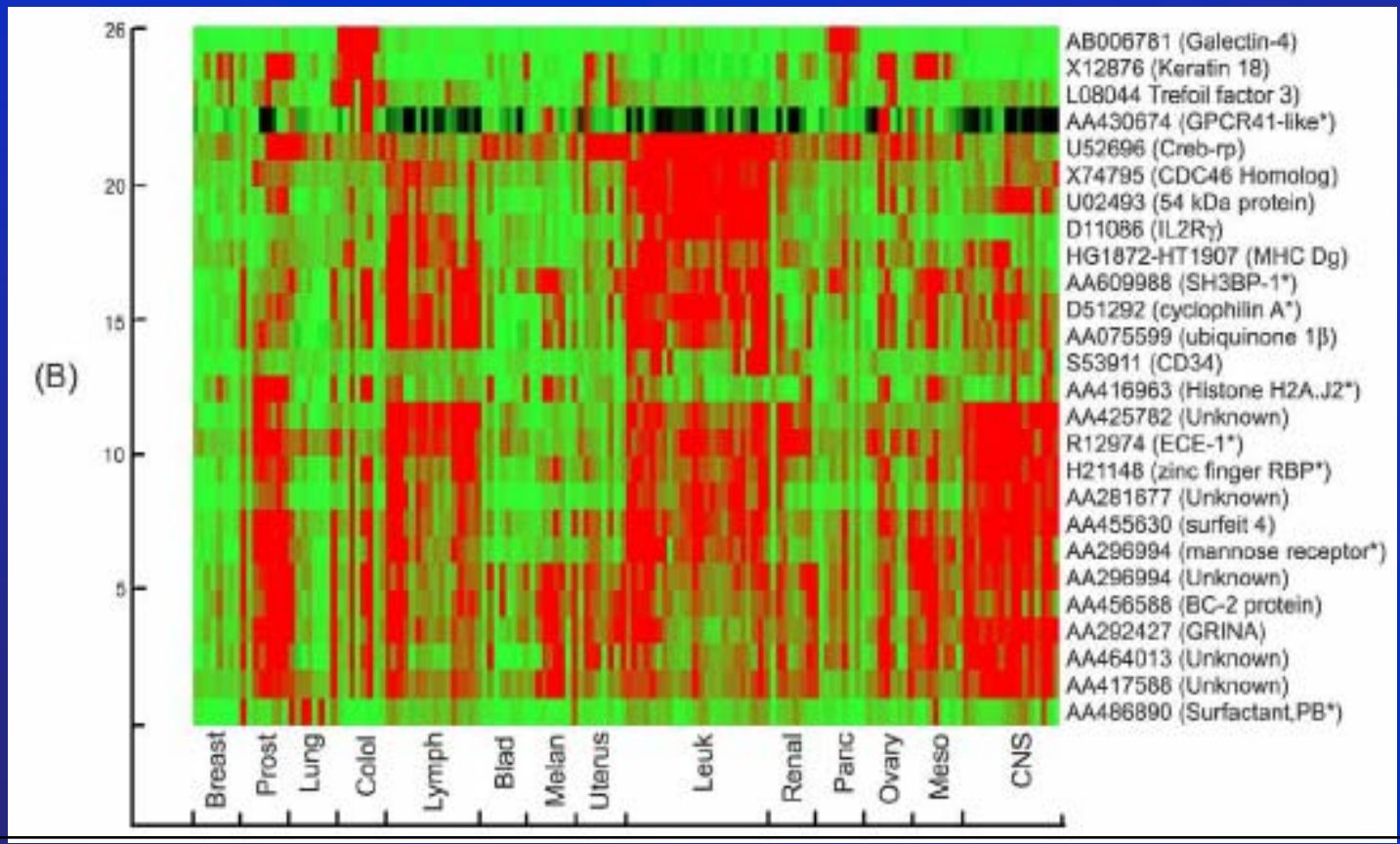
Table 1
The parameters and outcome of the GA/SVM in the four data sets

Data set		GA	SVM	Predictive Gene Set	
				Number of features	LOOCV (%)
Leukemia	Uniform	$P_c = 1$ $P_m = 0.005$	Poly	6	100.0
	RWS	Popsize = 12 $n = 1219$	Degree = 4		
Colon	Uniform	$P_c = 1$ $P_m = 0.006$	Poly	12	93.55
	RWS	Popsize = 30 $n = 31527$	Degree = 4		
NCI60	Uniform	$P_c = 1$ $P_m = 0.005$	Poly	27	87.93
	RWS	Popsize = 30 $n = 6729$	Degree = 4		
GCM	Uniform	$P_c = 0.98$ $P_m = 0.002$	Poly	26	85.19
	RWS	Popsize = 30 $n = 12765$	Degree = 4		

Pc: probabilità di crossover
Pm. Probabilità di mutazione
n: numero di generazioni

Clustering

- Col set finale di geni predittori è stato infine effettuato un clustering gerarchico dei campioni tumorali sulle due casistiche multiclasse (programmi usati: Gene Cluster 3.0 e Java TreeView).

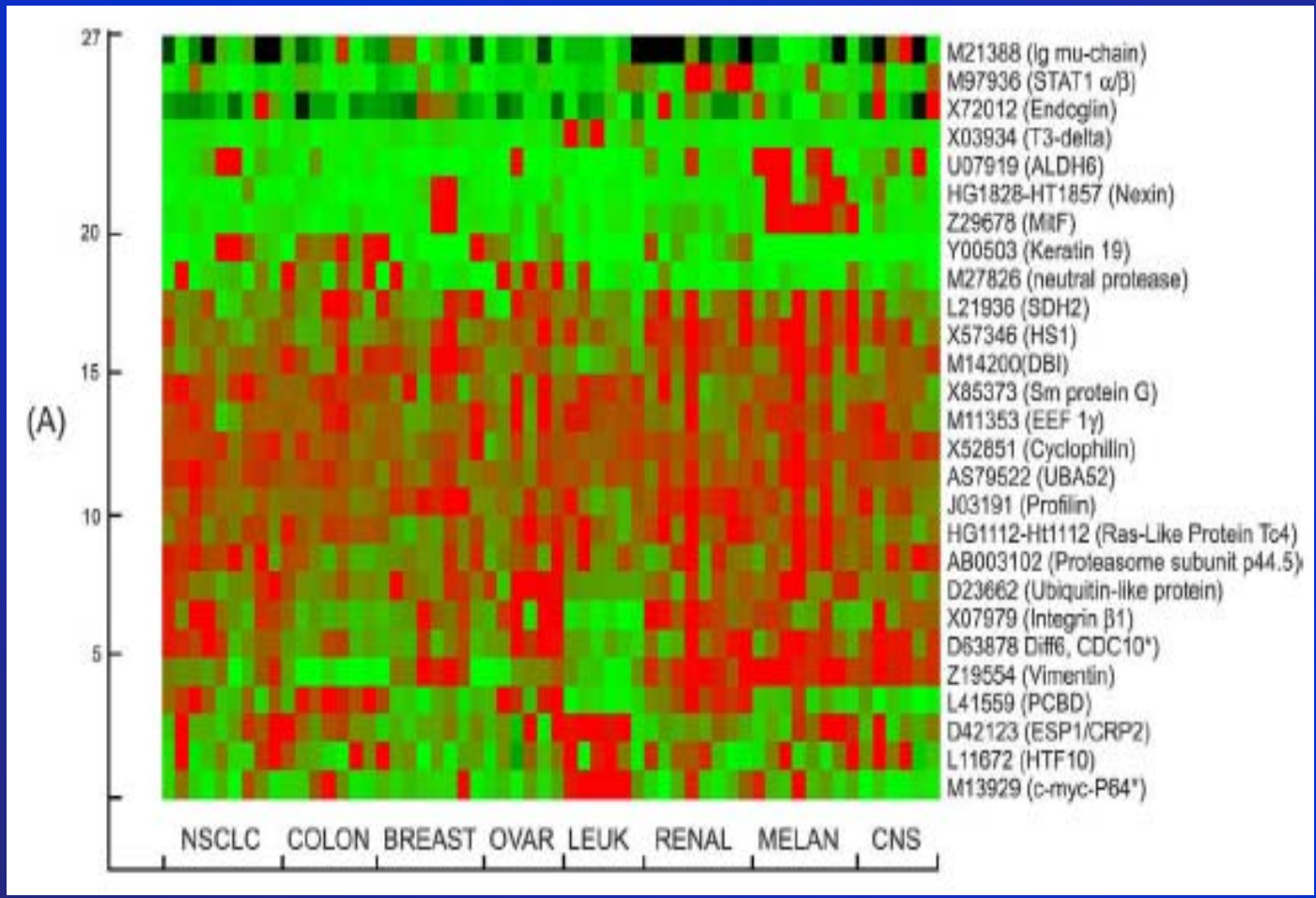


- Rosso: up-regolazione
- Nero: down-regolazione
- Verde: senza variazioni

GCM data set:

In ascissa i 218 campioni tumorali, in ordinata i 26 geni predittori

Clustering



- Rosso: up-regolazione
- Nero: down-regolazione
- Verde: senza variazioni

NCI60 data set:

In ascissa i 90 campioni tumorali, in ordinata i 27 geni predittori

Confronto dei risultati

- Gli stessi 4 data set sono stati analizzati anche in altri studi: un confronto del valore di LOOCV indica risultati comparabili o migliori.

- Il set di geni è molto compatto → Eliminazione della ridondanza.

Table 2
Results comparison of GA/SVM with some other algorithms

Classification Method	NCI60 data set		GCM data set		Reference
	LOOCV (%)	Number of features (genes)	LOOCV (%)	Number of features (genes)	
Hierarchical clustering	81	6831	–	–	[4]
OVA/SVM	–	–	78	16063	[8]
OVA/SVM	–	–	81.25	16063	[9]
OVA/KNN	–	–	72.92	16063	[9]
GA/MLHD	85.37	13	79.33	32	[10]
GA/SVM/RFE	87.93	27	85.19	26	This study

- Sovrapposizione tra i geni identificati negli studi: pressochè nessuna. Anche il confronto con i risultati dell'algoritmo GA/MLHD ha identificato 3 geni comuni in GCM e nessuno in NCI60.

Gene Ontology

Numerosi dei geni identificati sono importanti nello sviluppo della cancerogenesi e sono specifici per alcune forme tumorali.

A. NCI60:

- Up-regolazione nei melanomi di: MitF, ruolo cruciale nella sopravvivenza e proliferazione dei melanociti, e ALDH, coinvolto nello sviluppo di metastasi anche in altri tumori.
- Integrina B1 e Vimentina, non espressi solo nel tumore al colon e leucemie.
- Endoglina, agisce nell'adesione cellulare e motilità nelle cellule prostatiche.
- Stat1 ruolo fondamentale nell'arresto del ciclo cellulare e nell'apoptosi.
- M13929, presunto gene c-Myc (P64).

B. GCM (presente una più netta distinzione dei pattern genici)

- 7 geni sono fortemente correlati e up-regolati nella loro espressione in glioblastomi, leucemie, linfomi e cancro alla prostata.
- Gene AA430674, presunto recettore per una proteina G (identifcato tramite confronto di sequenza).

Conclusioni

- L’algoritmo utilizzato è stato in grado di identificare un pattern d’espressione genica in grado di effettuare un clustering dei campioni nella corretta classe tumorale in distinte casistiche.

- Il numero di geni individuato è ridotto, dando la possibilità di concentrare l’attenzione sui possibili meccanismi molecolari coinvolti.

- Il confronto dei risultati indica che l’uso combinato degli Algoritmi Genetici con la Support Vector Machine conferisce maggior efficacia all’analisi con microarray.

Conclusioni

- I precedenti algoritmi hanno la necessità di dover utilizzare i dati d'espressione di tutti i geni per riuscire a classificare correttamente i campioni.

- Con l'uso della RFE combinata agli AG si è in grado non solo di specificare i geni che appartengono ad un set predittivo, ma anche di trovare le dimensioni migliori del set genico stesso.