Metodi statistici per l'interpretazione del dato molecolare

Leonardo Ventura

ISPO, Firenze

l.ventura@ispo.toscana.it

ISTITUTO PER LO STUDIO
E LA PREVENZIONE ONCOLOGICA

Obiettivi

- Problema di classificazione
  - Identificazione di nuovi classi tumorali usando profili genetici
  - Classificazione di lesioni maligne in classi predefinite
  - Identificazione di biomarcatori che caratterizzano le diverse classi tumorali

Approcci statistici per la classificazione

- Analisi discriminante
- Alberi di classificazione
- Reti neurali
- Modelli di regressione logistica
- Cluster analysis
- ……

World J Urol (2010) 28:681–686
DOI 10.1007/s00345-010-0583-x

ORIGINAL ARTICLE

Use of a combination of biomarkers in serum and urine to improve detection of prostate cancer

Celia Prior · Francisco Guillen-Grima · Jose E. Robles · David Rosell · Jose M. Fernandez-Montero · Xabier Agirre · Raúl Catena · Alfonso Calvo

Conclusion We conclude that analysis of this biomarker combination in body fluids improves very significantly the diagnosis of PCa compared to the PSA test.

BMC Bioinformatics

BioMed Central

Methodology article

Open Access

A simple method to combine multiple molecular biomarkers for dichotomous diagnostic classification

Manju R Mamtani[†1], Tushar P Thakre[*1,2], Mrunal Y Kalkonde[1], Manik A Amin[1], Yogeshwar V Kalkonde[1], Amit P Amin[1] and Hemant Kulkarni[†1]

Results: Our algorithm comprises of three steps: estimating the area under the receiver operating characteristic curve for each biomarker, identifying a subset of biomarkers using linear regression and combining the chosen biomarkers using linear discriminant function analysis. Combining these

10

Statistical Methods for Identifying Differentially Expressed Gene Combinations

Yen-Yi Ho, Leslie Cope, Marcel Dettling, and Giovanni Parmigiani

Summary

Identification of coordinate gene expression changes across phenotypes or biological conditions is the basis of the ability to decode the role of gene expression regulatory networks. Statistically, the identification of these changes can be viewed as a search for groups (most typically pairs) of genes whose expression provides better phenotype discrimination when considered jointly than when considered individually. Such groups are defined as being jointly differentially expressed. In this chapter several approaches for identifying jointly differentially expressed groups of genes are reviewed of compared on a set of simulations.

Analisi discriminante

- L'obiettivo è stabilire l'appartenenza di una nuova unità statistica ad uno dei *g* gruppi in base al valore che essa assume su *p* variabili quantitative.

- Fu introdotta da Fisher nel 1936 per risolvere il problema della classificazione di piante di iris appartenenti a tre specie diverse.

- Quindi avremo:
  - Una variabile definisce i gruppi
  - Una o più variabili quantitative

- Sulla base di queste informazioni, si costruisce una regola di classificazione (funzione di classificazione)

Analisi discriminante lineare

- Nell'analisi discriminante lineare la funzione di classificazione è data dalla combinazione lineare delle variabili quantitative

$$Y = a_0+a_1X_1+a_2X_2+\ldots+a_pX_p$$

- Dove Y rappresenta il punteggio discriminante e i coefficienti ***a*** sono scelti in modo da rendere massima la separazione tra i gruppi

- I coefficienti ***a*** vengono stimati massimizzando la differenza tra i gruppi ovvero massimizzando la matrice di varianza e covarianza tra i gruppi rispetto a quella entro i gruppi

Scomposizione delle devianze

- Si può dimostrare che la devianza totale dei dati può essere scomposta come segue

- $D_{tot} = D_b + D_w$
- D_b è la devianza tra le medie di gruppo (between)
- D_w è la devianza entro i gruppi (within)

$$D_{tot} = \sum_{g=1}^{G} n_g (\bar{y}_g - \bar{y})^2 + \sum_{g=1}^{G} \sum_{n=1}^{N} (y_{ng} - \bar{y}_g)^2$$

Between Within

- Massimizzare la separazione tra i gruppi significa massimizzare il rapporto fra la devianza between e la devianza within

Scomposizione della matrice di varianze-covarianze

- Passando al caso multivariato la scomposizione non sarò più delle sole devianze ma dovrò tener conto anche delle covarianze

- Analogamente al caso univariato la matrice di varianza-covarianza totale dei dati può essere scomposta come segue

- $C_{tot} = C_b + C_w$
- C_b è la matrice di varianza-covarianza tra le medie di gruppo (between)
- C_w è la matrice di varianza-covarianza entro i gruppi (within)

Matrice varianze-covarianze della funzione discriminante lineare

- La matrice di varianze-covarianze della funzione discriminante lineare sarà
- $C_Y = a^t C\, a = a^t C_b\, a + a^t C_w\, a$
- La massima separazione tra i gruppi si ottiene massimizzando il rapporto fra le matrici di varianze-covarianze between e within della funzione discriminante lineare, cioè
- $a^t C_b\, a \,/\, a^t C_w\, a$
- La soluzione di massimo si ottiene dagli autovalori e autovettori della matrice $(C_b)^{-1} C_w$

Missing data

- Metodi di classificazione capaci di lavorare con dati mancanti (CART…)
- Metodi che richiedono la completezza delle informazioni (Analisi Discriminante)
- Imputazione dei dati mancanti…

Valutazione della capacità discriminante

- La capacità discriminante di una analisi discriminante lineare è solitamente basata sul confronto fra la classificazione vera e quella derivante dall'analisi

- Se utilizziamo per il confronto gli stessi dati utilizzati per costruire il modello, la valutazione della performance sarà troppo ottimistica

- Uno dei metodi utilizzati per ovviare a questo problema è il sample split
  - Si suddividono le unità statistiche in un insieme di stima (training set) che viene utilizzato per costruire la funzione discriminante ed un insieme di validazione (testing set) che servirà a valutare la capacità discriminante

BMC Bioinformatics

BioMed Central

Methodology article

Open Access

A simple method to combine multiple molecular biomarkers for dichotomous diagnostic classification

Manju R Mamtani[†1], Tushar P Thakre[*1,2], Mrunal Y Kalkonde[1], Manik A Amin[1], Yogeshwar V Kalkonde[1], Amit P Amin[1] and Hemant Kulkarni[†1]

ISTITUTO PER LO STUDIO
E LA PREVENZIONE ONCOLOGICA

CRC screening Firenze

- Fra il 2008 e il 2011 sono stati arruolati soggetti partecipanti allo screening colorettale del programma Fiorentino.
- Sono stati selezionati soggetti FIT positivi in soggetti di età 50-70 anni che si erano sottoposti a colonscopia di approfondimento
- Sono stati esclusi soggetti con pregressa neoplasia colorettale e che presentavano la sindrome di inflammatory bowel disease
- I partecipanti hanno fornito un campione fecale raccolto prima della preparazione intestinale

Biomarkers

- Human DNA quantification (RT PCR)
- LONG DNA quantification (RT PCR)
- K-ras codon 12 and 13 gene mutation (pyrosequensing)
- p 53 exon: 5, 6, 7 e 8 gene mutation (pyrosequensing)
- APC exon 15 codon : 876, 1306, 1309, 1312, 1367p1, 1378p1, 1379, 1450p1, 1465 gene mutation (pyrosequensing)
- Valutazione della instabilità genomica mediante analisi delle alterazioni dei micro satelliti Bat26, Bat25 e Bat 40 mediante analisi con AB PRISM 310 Genetic Analyzer e GeneScan software.
- Valutazione dello stato di metilazione dei geni: Vimentina, SPRF2, MGMT e HTLF in pyrosequenziamento

- In tutto sono stati analizzati 32 biomarcatori

Cancri e Adenomi

esito	Freq.	Percent	Cum.
Cancro	17	8.54	8.54
Adenoma HG	102	51.26	59.80
Adenoma LG	36	18.09	77.89
Negativo	44	22.11	100.00
Total	199	100.00	

Step seguiti per l'analisi

1. Curve ROC su tutti i 32 marcatori
2. Selezionare solo i marcatori con AUC significativa
3. Verificare se ci sono dei marcatori molto correlati fra loro
4. Verificare la percentuale di valori missing fra i marcatori selezionati
5. Eseguire l'analisi discriminante su tutte le possibili combinazioni di marcatori (a due, a tre, a quattro...)
6. Selezionare la miglior combinazione in termini di AUC

AUC

- Calcolo dell'AUC su tutti i 32 biomarcatori
- Selezione dei soli biomarcatori con AUC significativa
- DNA, LONG-DNA, KRAS12-g, APC1450-a, BAT40-12, KRAS12-act

Marker	AUC	95% CI	
dna	0.59	0.52	0.67
long_dna	0.66	0.58	0.74
kras12_g	0.62	0.54	0.70
apc1450_a	0.59	0.51	0.67
bat40_12	0.63	0.53	0.73
kras12_act	0.62	0.54	0.70

Correlazioni e missing

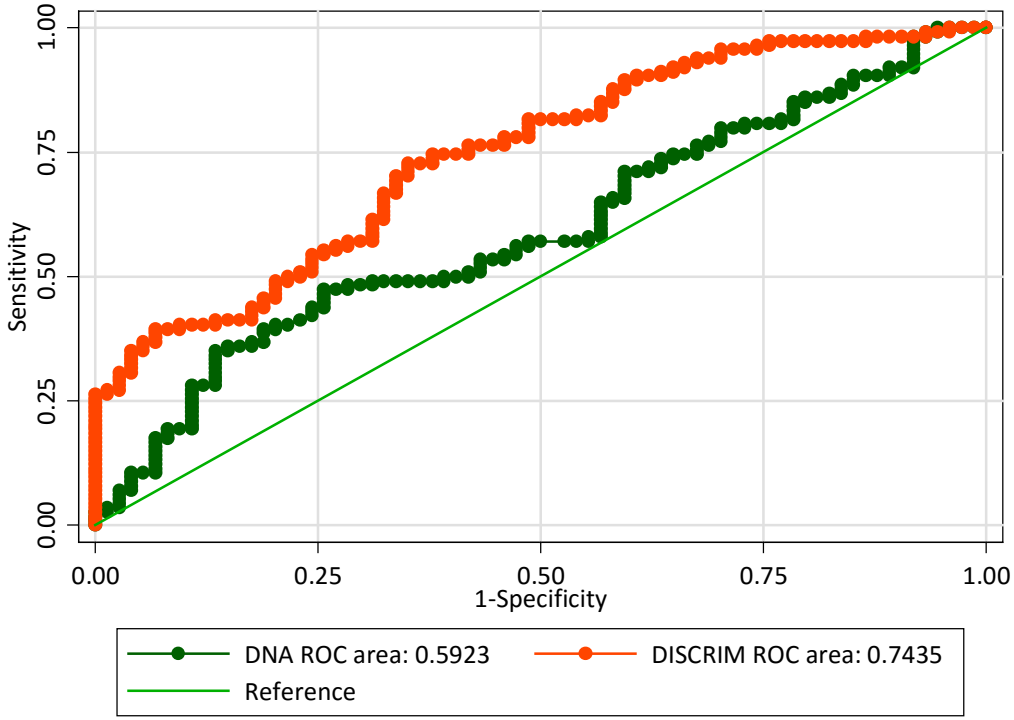
- KRAS12-g e KRAS12-act risultavano molto correlate
  - Si è quindi dato la precedenza al biomarcatore che con AUC maggiore, KRAS12-g
- BAT40-12 presentava una percentuale di missing troppo alta per poter essere considerato nell'analisi (35%)
  - E' stato eliminato dall'analisi

Analisi discriminante

- E’ stata condotta un’analisi discriminante su tutte le possibili combinazioni dei quattro biomarcatori rimasti (coppie, triplette, quartine…) aggiustando per sesso, età
- La miglior combinazione in termini di AUC è risultata quella composta da
  - DNA, LONG-DNA, KRAS12_G, APC1450

	Obs	Area	Std. Err.	[95% Conf.	Interval]
DNA	188	0.59	0.04	0.51	0.67
LONG-DNA	188	0.66	0.04	0.57	0.74
KRAS12-g	188	0.63	0.04	0.54	0.71
APC1450-a	188	0.59	0.04	0.51	0.68
DISCRIM	**188**	**0.74**	**0.04**	**0.67**	**0.81**

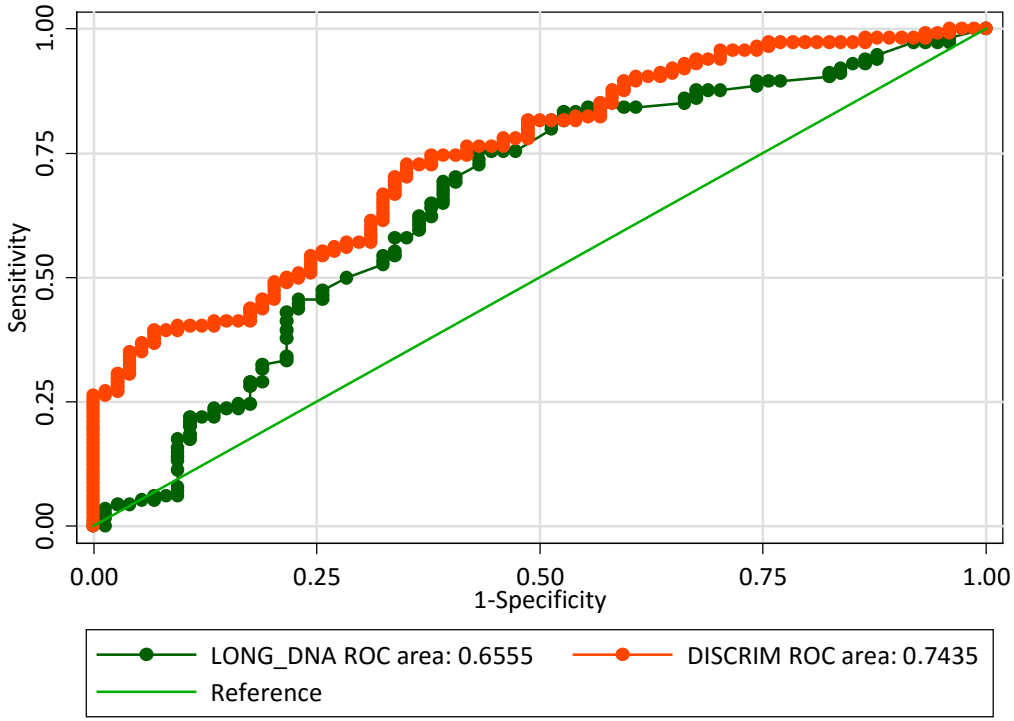
DNA vs DISCRIM



	Obs	Area	Std. Err.	[95% Conf.	Interval]
DNA	188	0.5923	0.0419	0.51031	0.67438
DISCRIM	188	0.7435	0.0358	0.67337	0.81359

Ho: area(DNA) = area(DISCRIM)
chi2(1) = 10.54 Prob>chi2 = 0.0012

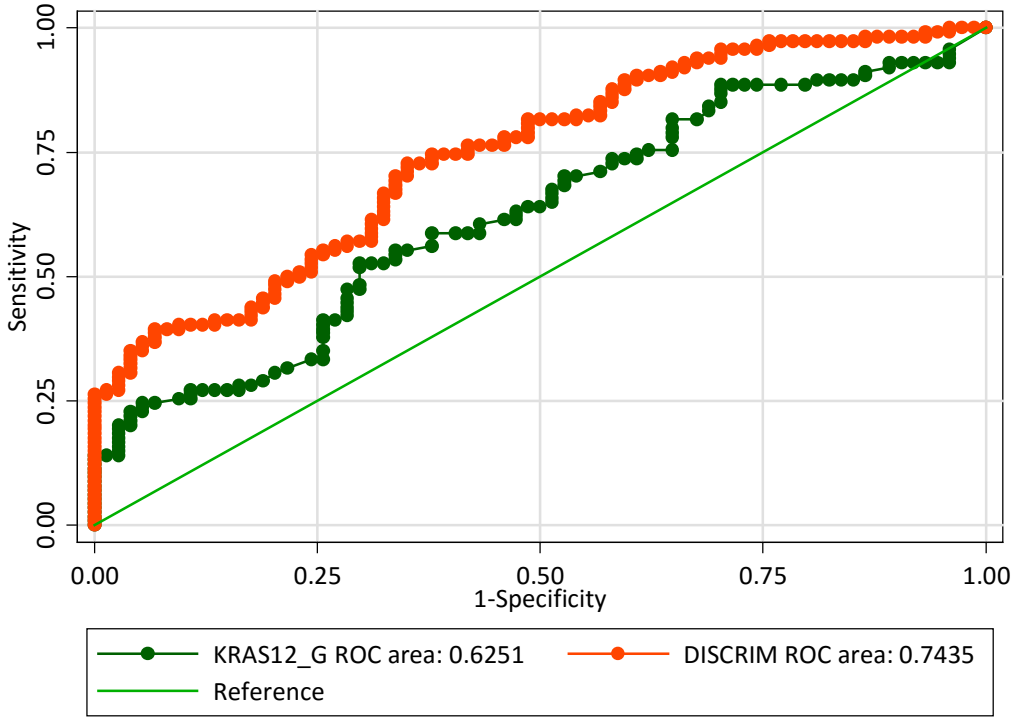
LONG-DNA vs DISCRIM



	Obs	Area	Std. Err.	[95% Conf.	Interval]
LONG_DNA	188	0.6555	0.0423	0.57268	0.73837
DISCRIM	188	0.7435	0.0358	0.67337	0.81359

Ho: area(LONG_DNA) = area(DISCRIM)
chi2(1) = 4.46 Prob>chi2 = 0.0347

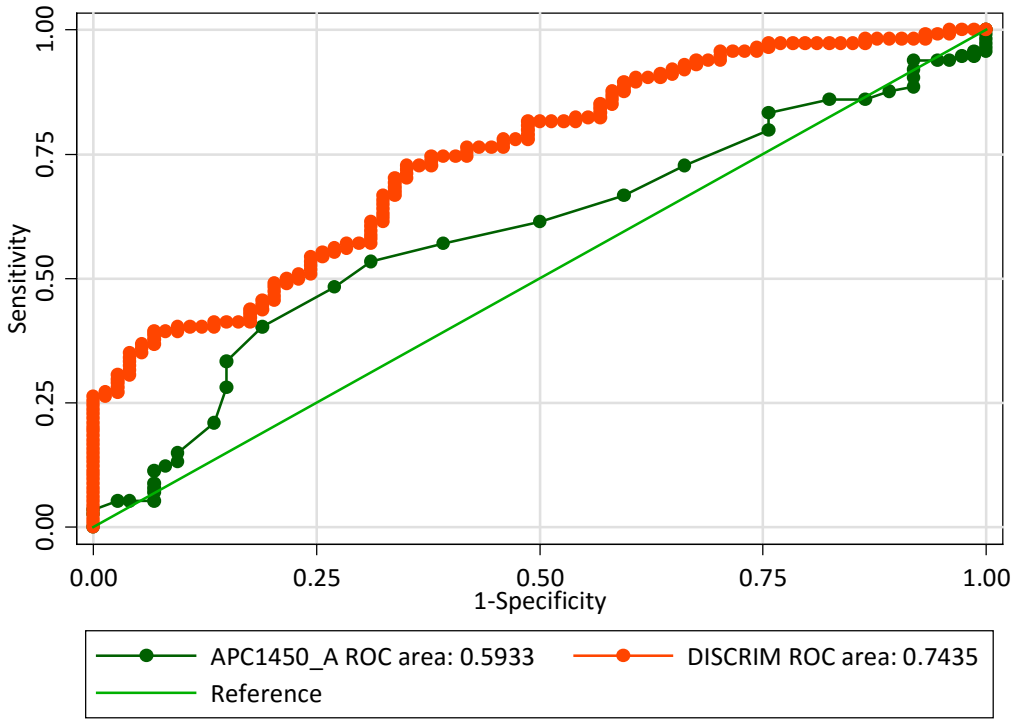
KRAS12-g vs DISCRIM



	Obs	Area	Std. Err.	[95% Conf.	Interval]
KRAS12_G	188	0.6251	0.0410	0.54461	0.70550
DISCRIM	188	0.7435	0.0358	0.67337	0.81359

Ho: area(KRAS12_G) = area(DISCRIM)
chi2(1) = 14.31 Prob>chi2 = 0.0002

APC1450-a vs DISCRIM



	Obs	Area	Std. Err.	[95% Conf.	Interval]
APC1450_A	188	0.5933	0.0420	0.51107	0.67552
DISCRIM	188	0.7435	0.0358	0.67337	0.81359

Ho: area(APC1450_A) = area(DISCRIM)
chi2(1) = 8.38 Prob>chi2 = 0.0038

Ulteriori considerazioni

- Miglior cut-off dello score discriminante
  - Migliore in termini di Se e Sp
  - Tale da massimizzare la Se per cancro (in ottica di triage)
- Validazione del risultato ottenuto nel training set sul testing set
- In un'ottica di triage (100% Se per cancro) possiamo andare a vedere quante colonscopie di approfondimento risparmieremmo a fronte di lesioni perse