

Machine learning per la medicina di precisione: esempio applicativo

Salvatore Fasola, PhD
Assegnista Post-Doc
CNR-IRIB

EUROPEAN
BIOTECH
WEEK
INNOVATION IS IN OUR GENES



FEDERCHIMICA
ASSOBIOTEC

Associazione nazionale per lo sviluppo
delle biotecnologie



Coordinated by



General Assembly led by



National Research
Council of Italy

Scientific Committee led by



INCiPiT Italian Network for
Paediatric Clinical Trials

Indice

Article

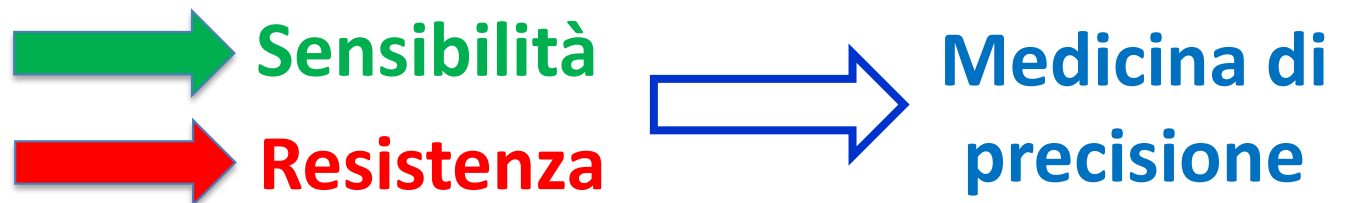
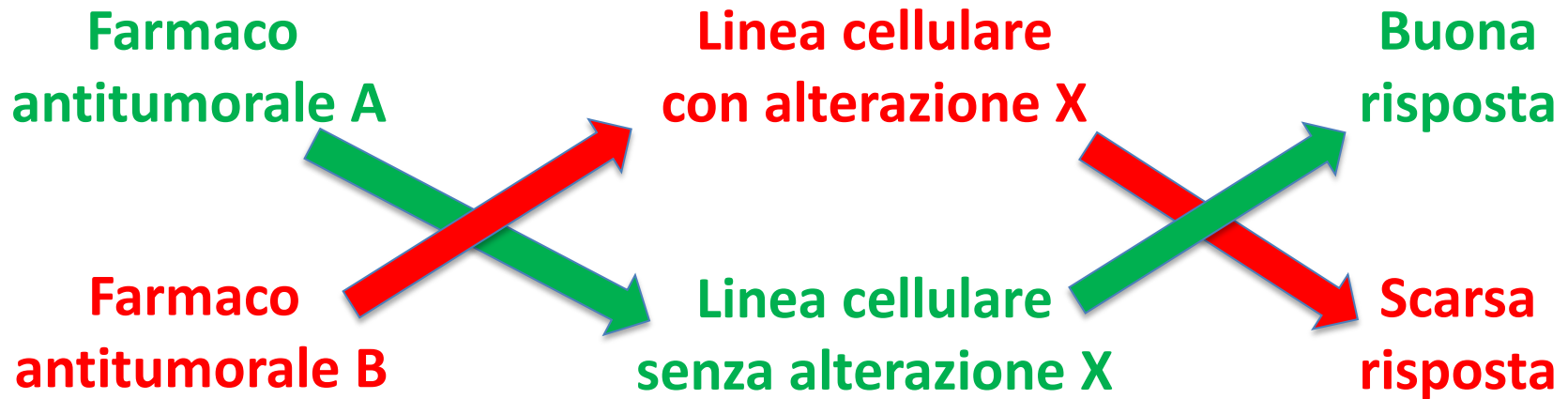
A Methodological Framework to Discover Pharmacogenomic Interactions Based on Random Forests

Salvatore Fasola^{1,*} , Giovanna Cilluffo¹ , Laura Montalbano¹, Velia Malizia¹, Giuliana Ferrante² 
and Stefania La Grutta¹ 

- Obiettivo
- Metodologia
- Risultati
- Conclusioni

Obiettivo

Proporre una procedura avanzata di analisi per l'identificazione di **interazioni farmacogenomiche...**



Metodologia

2 basi di dati

Database di alterazioni presenti o assenti (1 o 0)

- Progetto “Cancer Cell Line Encyclopedia” (CCLE)
- **48270** alterazioni x 648 linee cellulari tumorali

Database di risposte farmacologiche (AUC, tra 0 e 1)

- Progetto “Genomics of Drug Sensitivity in Cancer” (GDSC)
- 265 farmaci x 648 linee cellulari tumorali (le stesse)

Metodologia

Considerazioni preliminari

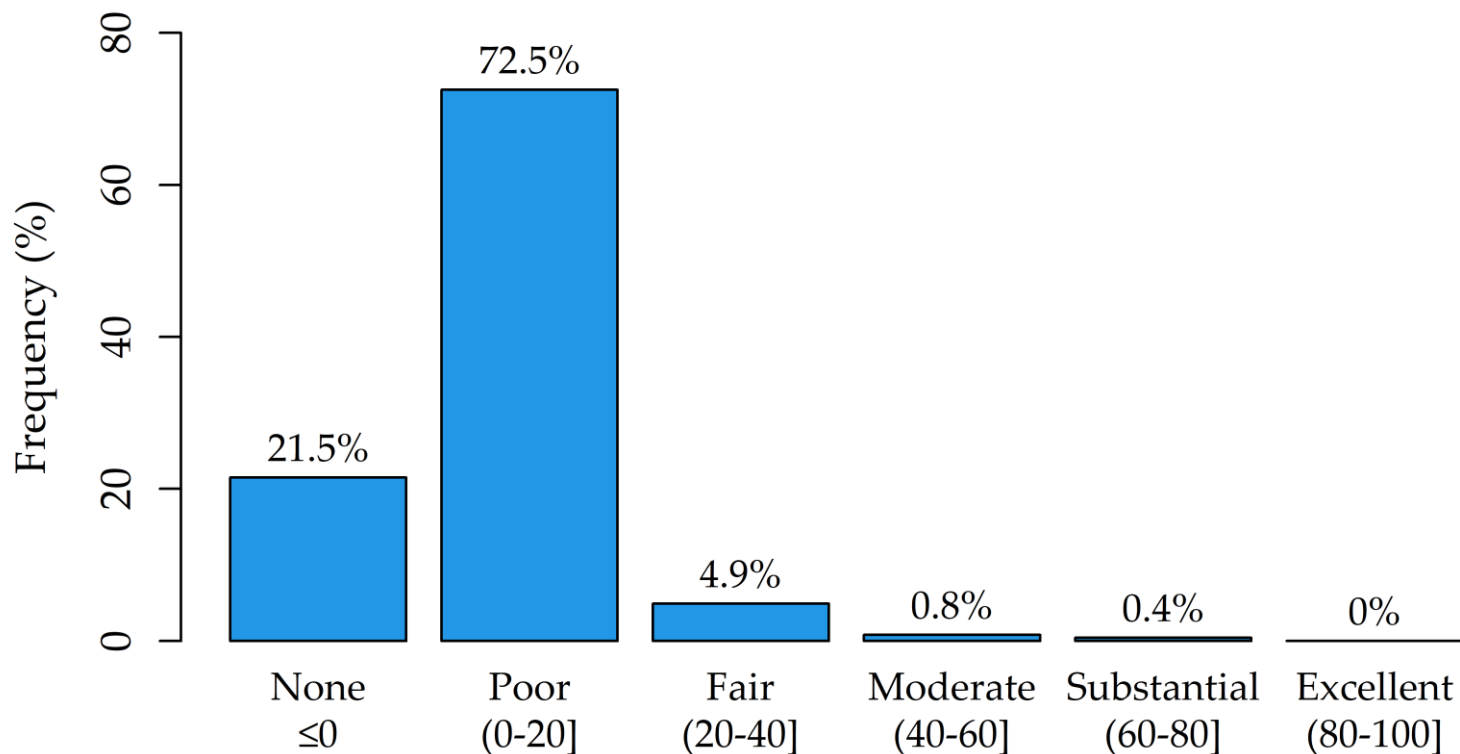
- vogliamo valutare se e quanto le alterazioni predicono le risposte;
- vogliamo identificare le alterazioni più importanti per la previsione;
- i predittori (alterazioni) sono 48270, quindi potenziale overfitting.



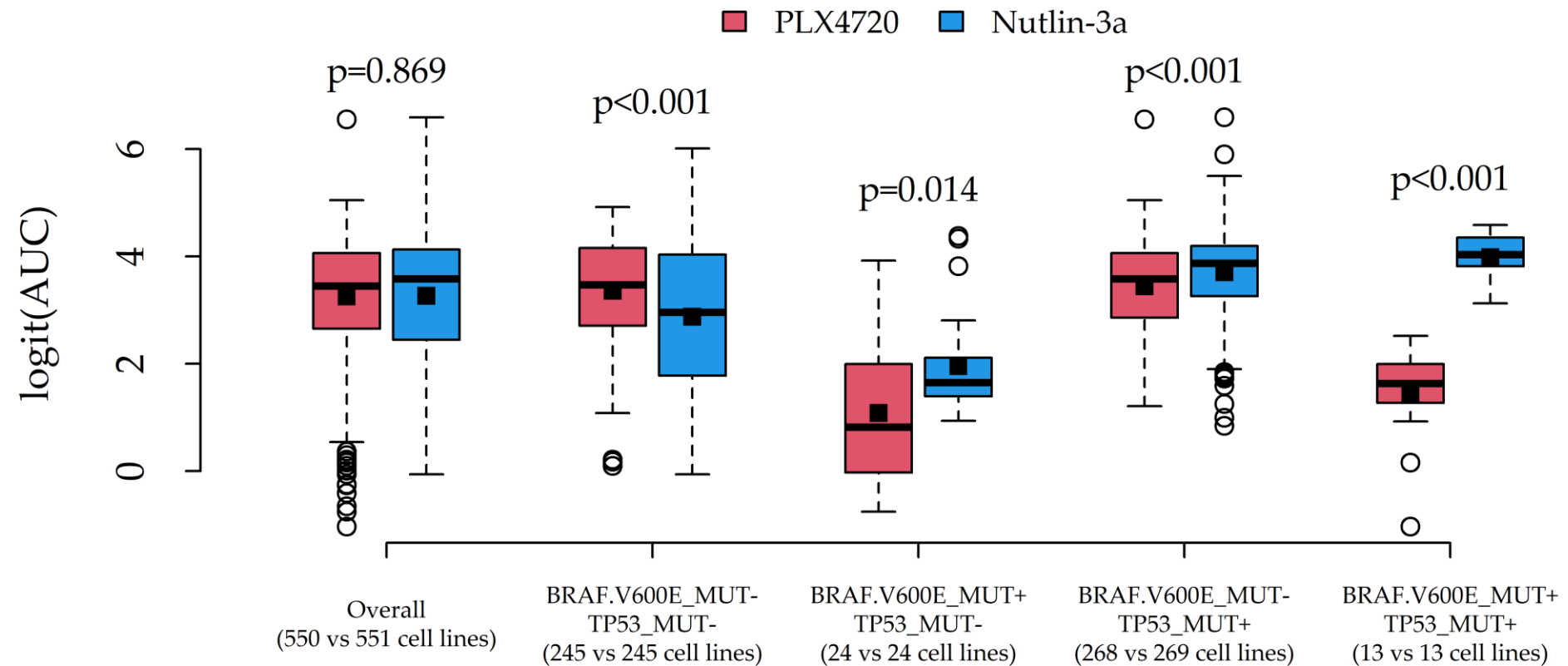
Modelli random forest (R software)

Risultati - 1

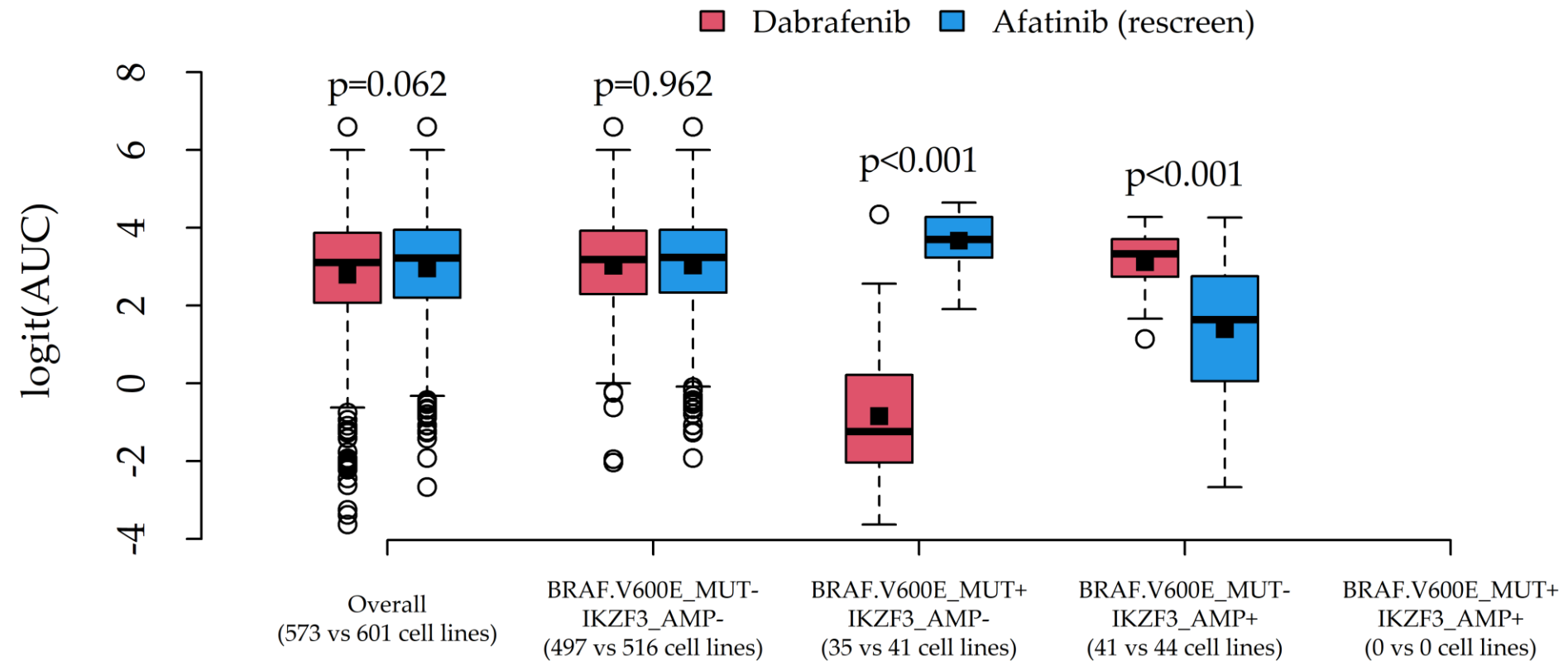
Distribuzione percentuale, sui 265 modelli random forest (uno per ciascun farmaco), di una misura sintetica del potere predittivo delle alterazioni nei confronti delle risposte farmacologiche:



Risultati – 2 – Interazione nota



Risultati – 3 – Interazione non nota



Conclusioni

- I modelli random forest sono molto adatti per l'identificazione di interazioni farmaco-gene;
- Sono state identificate associazioni farmaco-gene già note e potenziali nuove;
- I tempi di calcolo sono risultati ragionevoli (non giorni ma ore);
- E' stato fornito un codice R per l'implementazione della procedura di analisi proposta.