

Analisi di dati di espressione provenienti da esperimenti di Microarray su leucemie infantili (ALL,AML)

Partecipants

- Ing. Silvio Bicciato (Università degli studi di Padova)
 - D.ssa Truus Tekronnie (U. di Padova)
 - D.ssa Elda Rossi (CINECA)
 - Dott.Andrew Emenson (CINECA)
 - Dott. Giorgio Pedrazzi (CINECA)
 - Dott. Francesco Falciano (CINECA)
-

Introduction

Il progetto riguarda l'analisi di dati di espressione provenienti da esperimenti di Microarray su leucemie infantili (ALL,AML). Sara' sviluppato presso il Cineca principalmente da Francesco Falciano, come attivita' di ricerca relativa al suo Dottorato.

I dati saranno forniti dal Dipartimento di Pediatria dell'Università di Padova.

Il progetto si svolge in tre fasi principali:

1. **Data integration:** i dati di riferimento sono stati ottenuti con diversi modelli di chip Affimetrix (U95, U952, U133a) e necessitano perciò di essere uniformati in modo da poter essere analizzati in blocco. Per fare questo ci si propone di utilizzare tool già sviluppati all'Università di Padova eventualmente integrati con nuovi strumenti.
2. **Analisi di dati:** I dati così ottenuti dovranno essere analizzati ed elaborati con le metodiche standard di analisi di dati di espressione in modo da ottenere liste di geni correlate a fenotipi ALL, AML.
3. **Post-analisi:** Questa è la parte principale del progetto. Ipotizziamo una post-analisi su questi dati pre-elaborati per estrarne la massima informazione possibile mediante l'utilizzo di tecniche di mining e di knowledge discovery sia di tipo testuale (su fonti bibliografiche) che di tipo ontologico.

Queste tecniche verranno implementate in MedMOLE, uno strumento già in fase di sviluppo al CINECA (<http://medmole.cineca.it/>).

MedMOLE è una applicazione web che si basa sul text-mining delle informazioni bibliografiche

contenute nella banca dati Medline. MedMOLE permette di interrogare l'archivio di tutte le pubblicazioni scientifiche e di estrarre i lavori correlati con parole chiave di interesse (per esempio la lista dei geni ottenute dall'analisi di dati dei Microarrays). Questi ultimi vengono poi suddivisi in gruppi funzionali (con tecniche di Clustering) basati sull'interpretazione delle parole contenute negli abstract permettendo di raggruppare gli articoli che trattano dello stesso argomento.

È quindi possibile estrarre dai diversi gruppi i nomi dei geni in essi contenuti e quindi poter stabilire le diverse correlazioni che ne potrebbero risultare.

La migliore comprensione degli aspetti funzionali dei risultati degli esperimenti di espressione sarà inoltre perseguita mediante lo sviluppo di tecniche e strumenti per la Gene Ontology.

Un altro argomento di interesse comune riguarda **lo studio e la predisposizione di un archivio comune** a livello nazionale di dati di espressione genetica.

Questo archivio potrebbe interessare, almeno nella prima fase, solo i dati prodotti con piattaforma Affimetrix, coinvolgendo alcune delle circa 15 stazioni Affimetrix presenti oggi in Italia. Un modello di integrazione di diversi modelli e generazioni di chip dovrà essere previsto in modo analogo al punto sulla "Data integration" descritto precedentemente.

Il modello di archivio deve adottare gli standard più diffusi in modo da poter scalare facilmente a livello europeo ed internazionale. Inoltre deve implementare uno schema di sicurezza e riservatezza dei dati molto elevato: ad esempio solo i metadati (descrizione degli esperimenti) saranno pubblici per default, mentre i dati stessi lo saranno solo e quando il proprietario lo deciderà'.

Infine si dovrà prospettare un modello "distribuito" di distribuzione dei dati che permetta i singoli laboratori di gestire e mantenere i loro dati su piattaforme locali che all'occorrenza possano legarsi agli altri archivi analoghi per permettere una fruizione più ampia dell'informazione.