UNIVERSITA' DEGLI STUDI DI SALERNO

DOTTORATO IN INFORMATICA E INGEGNERIA
DELL'INFORMAZIONE

CURRICULUM INFORMATICA

COORDINATORE: Ch.mo. Prof. Alfredo De Santis

Ciclo XV N.S.

Multi-View Learning and Data Integration for *omics* Data

Relatori
Ch.mo. Prof. Roberto Tagliaferri
Ch.mo. Prof. Dario Greco

Candidato
Angela Serra
Matr. 8888100002

ANNO ACCADEMICO 2015/2016

Abstract

L'avanzamento tecnologico delle tecnologie high-throughput, combinato con il costante decremento dei costi di memorizzazione, ha portato alla produzione di grandi quantità di dati provenienti da diversi esperimenti che caratterizzano le stesse entità di interesse. Queste informazioni possono essere relative a specifici aspetti fenotipici (per esempio l'espressione genica), o possono includere misure globali e parallele di diversi aspetti molecolari (per esempio modifiche del DNA, trascrizione dell'RNA e traduzione delle proteine) negli stessi campioni.

Analizzare tali dati complessi è utile nel campo della systems biology per costruire modelli capaci di spiegare fenotipi complessi. Ad esempio, l'uso di dati genome-wide nella ricerca legata al cancro, per l'identificazione di gruppi di pazienti con caratteristiche molecolari simili, è diventato un approccio standard per una prognosi precoce più accurata e per l'identificazione di terapie specifiche. Inoltre, l'integrazione di dati di espressione genica riguardanti il trattamento di cellule tramite farmaci ha permesso agli scienziati di ottenere accuratezze elevate per il drug repositioning.

Purtroppo, esiste un grosso divario tra i dati prodotti, in seguito ai numerosi esperimenti, e l'informazione in cui essi sono tradotti. Quindi la comunità scientifica ha una forte necessità di metodi computazionali per poter integrare e analizzate tali dati per riempire questo divario.

La ricerca nel campo delle analisi multi-view, segue due diversi metodi di

analisi integrative: uno usa le informazioni complementari di diverse misure per studiare fenotipi complessi su diversi campioni (multi-view learning); l'altro tende ad inferire conoscenza sul fenotipo di interesse di una entità confrontando gli esperimenti ad essi relativi con quelli di altre entità fenotipiche già note in letteratura (meta-analisi). La meta-analisi può essere pensata come uno studio comparativo dei risultati identificati in un particolare esperimento, rispetto a quelli di studi precedenti. A causa della sua natura, la meta-analisi solitamente coinvolge dati omogenei. D'altra parte, il multi-view learning è un approccio più flessibile che considera la fusione di diverse sorgenti di dati per ottenere stime più stabili e affidabili. In base al tipo di dati e al livello di integrazione, nuove metodologie sono state sviluppate a partire da tecniche basate sulla teoria dei grafi, machine learning e statistica. In base alla natura dei dati e al problema statistico da risolvere, l'integrazione di dati eterogenei può essere effettuata a diversi livelli: early, intermediate e late integration. Le tecniche di early integration consistono nella concatenazione dei dati delle diverse viste in un unico spazio delle feature. Le tecniche di intermediate integration consistono nella trasformazione di tutte le sorgenti dati in un unico spazio comune prima di combinarle. Nelle tecniche di late integration, ogni vista è analizzata separatamente e i risultati sono poi combinati.

Lo scopo di questa tesi è duplice: il primo obbiettivo è la definizione di una metodologia di integrazione dati per la sotto-tipizzazione dei pazienti (MVDA) e il secondo è lo sviluppo di un tool per la caratterizzazione fenotipica dei nanomateriali (INSIdEnano). In questa tesi di dottorato presento le metodologie e i risultati della mia ricerca.

MVDA è una tecnica multi-view con lo scopo di scoprire nuove sotto tipologie di pazienti statisticamente rilevanti. Identificare sottotipi di pazienti per una malattia specifica è un obbiettivo con alto rilievo nella pratica clinica, soprattutto per la diagnosi precoce delle malattie. Questo problema è generalmente risolto usando dati di trascrittomica per identificare i gruppi di pazienti che condividono gli stessi pattern di alterazione genica. L'idea principale alla base di questo lavoro di ricerca è quello di combinare più tipologie di dati omici per gli stessi pazienti per ottenere una migliore caratterizzazione del loro profilo.

La metodologia proposta è un approccio di tipo late integration basato sul clustering. Per ogni vista viene effettuato il clustering dei pazienti rappresentato sotto forma di matrici di membership. I risultati di tutte le viste vengono poi combinati tramite una tecnica di fattorizzazione di matrici per ottenere i metacluster finali multi-view. La fattibilità e le performance del nostro metodo sono stati valutati su sei dataset multi-view relativi al tumore al seno, glioblastoma, cancro alla prostata e alle ovarie. I dati omici usati per gli esperimenti sono relativi alla espressione dei geni, espressione dei mirna, RNASeq, miRNASeq, espressione delle proteine e della Copy Number Variation. In tutti i dataset sono state identificate sotto-tipologie di pazienti con rilevanza statistica, identificando nuovi sottogruppi precedentemente non noti in letteratura. Ulteriori esperimenti sono stati condotti utilizzando la conoscenza a priori relativa alle macro classi dei pazienti. Tale informazione è stata considerata come una ulteriore vista nel processo di integrazione per ottenere una accuratezza più elevata nella classificazione dei pazienti. Il metodo proposto ha performance migliori degli algoritmi di clustering clussici su tutti i dataset. MVDA ha ottenuto risultati migliori in confronto a altri algoritmi di integrazione di tipo ealry e intermediate integration. Inoltre il metodo è in grado di calcolare il contributo di ogni singola vista al risultato finale. I risultati mostrano, anche, che il metodo è stabile in caso di perturbazioni del dataset effettuate rimuovendo un paziente alla volta (leave-one-out). Queste osservazioni suggeriscono che l'integrazione di informazioni a priori e feature genomiche, da utilizzare congiuntamente durante l'analisi, è una strategia vincente nell'identificazione di sotto-tipologie di malattie.

INSIdE nano (Integrated Network of Systems bIology Effects of nanomaterials) è un tool innovativo per la contestualizzazione sistematica degli effetti delle nanoparticelle (ENMs) in contesti biomedici. Negli ultimi anni, le tecnologie omiche sono state ampiamente applicate per caratterizzare i nanomateriali a livello molecolare. E' possibile contestualizzare l'effetto a livello molecolare di diversi tipi di perturbazioni confrontando i loro pattern di alterazione genica. Mentre tale approccio è stato applicato con successo nel campo del drug repositioning, una contestualizzazione estensiva dell'effetto dei nanomateriali

sulle cellule è attualmente mancante. L'idea alla base del tool è quello di usare strategie comparative di analisi per contestualizzare o posizionare i nanomateriali in confronto a fenotipi rilevanti che sono stati studiati in letteratura (come ad esempio malattie dell'uomo, trattamenti farmacologici o esposizioni a sostanze chimiche) confrontando i loro pattern di alterazione molecolare. Questo potrebbe incrementare la conoscenza dell'effetto molecolare dei nanomateriali e contribuire alla definizione di nuovi pathway tossicologici oppure identificare eventuali coinvolgimenti dei nanomateriali in eventi patologici o in nuove strategie terapeutiche.

L'ipotesi alla base è che l'identificazione di pattern di similarità tra insiemi di fenotipi potrebbe essere una indicazione di una associazione biologica che deve essere successivamente testata in ambito tossicologico o terapeutico. Basandosi sulla firma di espressione genica, associata ad ogni fenotipo, la similarità tra ogni coppia di perturbazioni è stata valuta e usata per costruire una grande network di interazione tra fenotipi. Per assicurare l'utilizzo di INSIdE nano, è stata sviluppata una infrastruttura computazionale robusta e scalabile, allo scopo di analizzare tale network. Inoltre è stato realizzato un sito web che permettesse agli utenti di interrogare e visualizzare la network in modo semplice ed efficiente. In particolare, INSIdE nano è stato analizzato cercando tutte le possibili clique di quattro elementi eterogenei (un nanomateriale, un farmaco, una malattia e una sostanza chimica). Una clique è una sotto network completamente connessa, dove ogni elemento è collegato con tutti gli altri. Di tutte le clique, sono state considerate come significative solo quelle per le quali le associazioni tra farmaco e malattia e farmaco e sostanze chimiche sono note. Le connessioni note tra farmaci e malattie si basano sul fatto che il farmaco è prescritto per curare tale malattia. Le connessioni note tra malattia e sostanze chimiche si basano su evidenze presenti in letteratura del fatto che tali sostanze causano la malattia. Il focus è stato posto sul possibile coinvolgimento dei nanomateriali con le malattie presenti in tali clique. La valutazione di INSIdE nano ha confermato che esso mette in evidenza connessioni note tra malattie e farmaci e tra malattie e sostanze chimiche. Inoltre la similarità tra le malattie calcolata in base ai geni è conforme alle informazioni basate sulle loro informazioni cliniche.

Allo stesso modo le similarità tra farmaci e sostanze chimiche rispecchiano le loro similarità basate sulla struttura chimica.

Nell'insieme, i risultati suggeriscono che INSIdE nano può essere usato per contestualizzare l'effetto molecolare dei nanomateriali e inferirne le connessioni rispetto a fenotipi precedentemente studiati in letteratura. Questo metodo permette di velocizzare il processo di valutazione della loro tossicità e apre nuove prospettive per il loro utilizzo nella biomedicina.