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

In recent years, the advancement of high-throughput technologies, combined with the constant decrease of the data-storage costs, has led to the production of large amounts of data from different experiments that characterise the same entities of interest. This information may relate to specific aspects of a phenotypic entity (e.g. Gene expression), or can include the comprehensive and parallel measurement of multiple molecular events (e.g., DNA modifications, RNA transcription and protein translation) in the same samples.

Exploiting such complex and rich data is needed in the frame of systems biology for building global models able to explain complex phenotypes. For example, the use of genome-wide data in cancer research, for the identification of groups of patients with similar molecular characteristics, has become a standard approach for applications in therapy-response, prognosis-prediction, and drug-development.ÂăMoreover, the integration of gene expression data regarding cell treatment by drugs, and information regarding chemical structure of the drugs allowed scientist to perform more accurate drug repositioning tasks.

Unfortunately, there is a big gap between the amount of information and the knowledge in which it is translated. Moreover, there is a huge need of computational methods able to integrate and analyse data to fill this gap.

Current researches in this area are following two different integrative methods: one uses the complementary information of different measurements for the

study of complex phenotypes on the same samples (multi-view learning); the other tends to infer knowledge about the phenotype of interest by integrating and comparing the experiments relating to it with respect to those of different phenotypes already known through comparative methods (meta-analysis). Meta-analysis can be thought as an integrative study of previous results, usually performed aggregating the summary statistics from different studies. Due to its nature, meta-analysis usually involves homogeneous data. On the other hand, multi-view learning is a more flexible approach that considers the fusion of different data sources to get more stable and reliable estimates. Based on the type of data and the stage of integration, new methodologies have been developed spanning a landscape of techniques comprising graph theory, machine learning and statistics. Depending on the nature of the data and on the statistical problem to address, the integration of heterogeneous data can be performed at different levels: early, intermediate and late. Early integration consists in concatenating data from different views in a single feature space. Intermediate integration consists in transforming all the data sources in a common feature space before combining them. In the late integration methodologies, each view is analysed separately and the results are then combined.

The purpose of this thesis is twofold: the former objective is the definition of a data integration methodology for patient sub-typing (MVDA) and the latter is the development of a tool for phenotypic characterisation of nanomaterials (INSIdEnano). In this PhD thesis, I present the methodologies and the results of my research.

MVDA is a multi-view methodology that aims to discover new statistically relevant patient sub-classes. Identify patient subtypes of a specific diseases is a challenging task especially in the early diagnosis. This is a crucial point for the treatment, because not all the patients affected by the same disease will have the same prognosis or need the same drug treatment. This problem is usually solved by using transcriptomic data to identify groups of patients that share the same gene patterns. The main idea underlying this research work is that to combine more omics data for the same patients to obtain a better characterisation of their disease profile. The proposed methodology is a late integration approach

based on clustering. It works by evaluating the patient clusters in each single view and then combining the clustering results of all the views by factorising the membership matrices in a late integration manner. The effectiveness and the performance of our method was evaluated on six multi-view cancer datasets related to breast cancer, glioblastoma, prostate and ovarian cancer. The omics data used for the experiment are gene and miRNA expression, RNASeq and miRNASeq, Protein Expression and Copy Number Variation.

In all the cases, patient sub-classes with statistical significance were found, identifying novel sub-groups previously not emphasised in literature. The experiments were also conducted by using prior information, as a new view in the integration process, to obtain higher accuracy in patients' classification. The method outperformed the single view clustering on all the datasets; moreover, it performs better when compared with other multi-view clustering algorithms and, unlike other existing methods, it can quantify the contribution of single views in the results. The method has also shown to be stable when perturbation is applied to the datasets by removing one patient at a time and evaluating the normalized mutual information between all the resulting clusterings. These observations suggest that integration of prior information with genomic features in sub-typing analysis is an effective strategy in identifying disease subgroups.

INSIdE nano (Integrated Network of Systems bIology Effects of nanomaterials) is a novel tool for the systematic contextualisation of the effects of engineered nanomaterials (ENMs) in the biomedical context. In the recent years, omics technologies have been increasingly used to thoroughly characterise the ENMs molecular mode of action. It is possible to contextualise the molecular effects of different types of perturbations by comparing their patterns of alterations. While this approach has been successfully used for drug repositioning, it is still missing to date a comprehensive contextualisation of the ENM mode of action. The idea behind the tool is to use analytical strategies to contextualise or position the ENM with the respect to relevant phenotypes that have been studied in literature, (such as diseases, drug treatments, and other chemical exposures) by comparing their patterns of molecular alteration. This could greatly increase the knowledge on the ENM molecular effects and in turn

contribute to the definition of relevant pathways of toxicity as well as help in predicting the potential involvement of ENM in pathogenetic events or in novel therapeutic strategies. The main hypothesis is that suggestive patterns of similarity between sets of phenotypes could be an indication of a biological association to be further tested in toxicological or therapeutic frames. Based on the expression signature, associated to each phenotype, the strength of similarity between each pair of perturbations has been evaluated and used to build a large network of phenotypes. To ensure the usability of INSIdE nano, a robust and scalable computational infrastructure has been developed, to scan this large phenotypic network and a web-based effective graphic user interface has been built. Particularly, INSIdE nano was scanned to search for clique sub-networks, quadruplet structures of heterogeneous nodes (a disease, a drug, a chemical and a nanomaterial) completely interconnected by strong patterns of similarity (or anti-similarity). The predictions have been evaluated for a set of known associations between diseases and drugs, based on drug indications in clinical practice, and between diseases and chemical, based on literature-based causal exposure evidence, and focused on the possible involvement of nanomaterials in the most robust cliques. The evaluation of INSIdE nano confirmed that it highlights known disease-drug and disease-chemical connections. Moreover, disease similarities agree with the information based on their clinical features, as well as drugs and chemicals, mirroring their resemblance based on the chemical structure. Altogether, the results suggest that INSIdE nano can also be successfully used to contextualise the molecular effects of ENMs and infer their connections to other better studied phenotypes, speeding up their safety assessment as well as opening new perspectives concerning their usefulness in biomedicine.