Data Integration in Early Drug Development Phase

The QSTAR Modeling Framework

Adetayo Kasim
Outline

• Background and Data Structure

• Gene X Bioassay X Fingerprint Analysis

• Co-clustering framework

• Biclustering Framework

• QSTAR Consortium
DRUG DISCOVERY STAGES

Drug Discovery Phases
DRUG DISCOVERY STAGES

• The tradition drug discovery approach relies mostly on chemical properties at early stage

• Potential side effects are often identified in later toxicity studies

• Less than 25% of success rate in Phase III trials

• Early identification of potential sides effects may prevent expensive Phase II & Phase III trials.

• Early toxicity detection can be facilitated by incorporating relevant biological data in the early stages of drug developments, particularly in lead optimisation.
QSTAR: FRAMEWORK
QSTAR: Data Structures

\[
Z_{M \times C} = \begin{pmatrix}
z_{11} & z_{12} & \cdots & z_{1C} \\
z_{21} & z_{22} & \cdots & z_{2C} \\
\vdots & \vdots & \ddots & \vdots \\
z_{M1} & z_{M2} & \cdots & z_{MC} 
\end{pmatrix}
\]

\[
Y_{N \times C} = \begin{pmatrix}
y_{11} & y_{12} & \cdots & y_{1C} \\
y_{21} & y_{22} & \cdots & y_{2C} \\
\vdots & \vdots & \ddots & \vdots \\
y_{N1} & y_{N2} & \cdots & y_{NC} 
\end{pmatrix}
\]

\[
X_{G \times C} = \begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1C} \\
x_{21} & x_{22} & \cdots & x_{2C} \\
\vdots & \vdots & \ddots & \vdots \\
x_{g1} & x_{g2} & \cdots & x_{gC} 
\end{pmatrix}
\]
QSTAR: Gene X Bioassay X Fingerprint Analysis

• To find a group of genes that are jointly associated with chemical structure and bioactivity data.

• To find a group of genes that are associated with a chemical structure, but that are conditionally independent of bioactivity data.
QSTAR: Joint Modelling

- Significant association between gene expression and fingerprint
- Significant association between bioactivity and fingerprint
- Significant association between gene expression and bioactivity data
• Significant association between gene expression and fingerprint
• Significant association between bioactivity and fingerprint
• **NO** association between gene expression and bioactivity data
\begin{align*}
\begin{pmatrix}
X_j \\
Y_i
\end{pmatrix}
\sim \mathcal{N}
\begin{pmatrix}
\mu_{jk} + \alpha_{jk} \times Z_{jk} \\
\mu_{ik} + \beta_{ik} \times Z_{ik},
\end{pmatrix}
\Sigma_{i,j}
\end{align*}

Where,

\begin{align*}
\Sigma_{i,j} &=
\begin{pmatrix}
\sigma_{X_j}^2 & \sigma_{X_jY_i} \\
\sigma_{X_jY_i} & \sigma_{Y_i}^2
\end{pmatrix}
\end{align*}

\begin{itemize}
\item $\beta_{ik}$ denotes the effect of the $k^{th}$ fingerprint on the $i^{th}$ bioassay
\item $\alpha_{jk}$ denotes the effect of the $k^{th}$ fingerprint on the $j^{th}$ gene.
\end{itemize}
QSTAR: Joint Modelling

- Examples of genes with significant association with fingerprint and bioactivity data
QSTAR: Joint Modelling

- Examples of genes with significant association with fingerprint, but conditionally independent of bioactivity
QSTAR: Path Analysis

- To find genes with direct and indirect effect of chemical structures on bioactivity data
- To find genes with only direct effect of chemical structure
QSTAR: Path Analysis

\[ Y_i = \vartheta_{zyj} Z_k + \varepsilon_{2j} \]

\[ X_j = \vartheta_{xyj} Y_i + \vartheta_{zxj} Z_k + \varepsilon_{1j} \]

- \( \vartheta_{zyj} \) is the total effect of the \( k^{th} \) fingerprint on the \( i^{th} \) bioassay
- \( \vartheta_{zxj} \) is the Direct effect the fingerprint on the \( j^{th} \) gene
- \( \omega_{zxj} = \vartheta_{xyj} \cdot \vartheta_{zyj} \) is the indirect effect of the fingerprint on gene expression given the bioactivity data
QSTAR: Conditional Model

\[
g\{E(X_j | Z_k)\} = \theta_{1i} + \theta_{2i}Z_k
\]

\[
g\{E(X_j | Z_k, Y_i)\} = \varphi_0 + \varphi_{1i}Y_i + \varphi_{2i}Z_k
\]

- \(\theta_{2i}\) is the unadjusted effect of fingerprint on gene expression data.

- \(\varphi_{2i}\) is the adjusted effect of fingerprint on gene expression data.

- This conditional modelling framework fits within the principle of generalised linear model.
Relevance to Radiation Study

• These modelling framework can potentially be used as a surrogate marker identification/validation in low radiation studies.
QSTAR: Co-clustering Framework

• To find a group of compounds with a similar chemical structures and bioactivity data and their corresponding discriminating gene signatures
There are several methods for simultaneous analysis of gene expression data to overcome the limitation of gene-by-gene analysis.

Some of the methods are variation of penalised-likelihood approach
- Elastic net
- LASSO

Clustering methods have also been considered for this purpose
- Hierarchical clustering method
- K-means

In co-clustering approach, we are interested in multi-view clustering of two data matrices.
QSTAR: Weighted Clustering

Structure (Z) 

Bioactivity (Y)

\[ S_N^W = \omega_Z S_N^Z + \omega_Y S_N^Y \]

Structure-based Clustering; \( C(Z, k) \)

Bioactivity-based Clustering; \( C(Y, k) \)

Weighted Similarity-based Clustering; \( C(W, k) \)

Cluster-level analysis

Cluster k vs other compounds outside Cluster k
**QSTAR: Weighted Clustering**

Chemical Structure

Bioactivity

$\omega_z = 0.45$

$\omega_z = 0.55$
QSTAR: Weighted Clustering

- Group of genes linked to the selected cluster
QSTAR: Biclustering Framework

- To find similar local patterns two matrices. Illustration with gene expression matrix and high content screening matrix
• Finding local patterns in gene expression data based on simultaneous clustering of genes and compounds
QSTAR: Multiview Biclustering

- Methods:
  - FABIA
  - Multiple Factor Analysis / Sparse Multiple Factor Analysis
QSTAR: Gene expression and High Content screening
QSTAR: Pathway Analysis

• Exploring potential biological pathways based on existing literatures and database

• Exploring target predictions based on existing literature and database
QSTAR: Pathway Analysis

• The are publicly available databases for gene annotations and functions
  • KEGG (http://www.genome.jp/kegg/)
  • GO (http://geneontology.org/page/go-database)

• The identified bioassay and chemical structures can also be explored further using:
  • chEMBL (https://www.ebi.ac.uk/chembl/)
  • drugBank (http://www.drugbank.ca/)
QSTAR: Pathway Analysis

- Example of MPL pathway analysis
QSTAR CONSORTIUM
(http://www.qstar-consortium.org/)

- Janssen Pharmaceuticals scientists from various therapeutic areas such as Oncology, Neurology, Infectious Diseases
- Academic collaborations: chemoinformatics, statistics, machine learning, platform-specific data preprocessing
- Further Janssen Pharmaceuticals team members from: Medicinal chemistry, chemoinformatics, systems biology, molecular profiling, statistics, IT, HTS, exploratory toxicology
Thank you