Application of Deep Learning to Drug Discovery

Hiroshi Tanaka
Tohoku Medical Megabank Organization, Tohoku University
Current Situation of Drug Discovery

- **Rapid increase of R&D expenditure**
  - More than 1B $ for one marketed drug
- **Decrease of success rate**
  - now about 1/20,000 ~ 1/30,000
  - Remarkable Drop Between non-clinical and clinical test (**phase II attrition**)
- **Clinical Predictability**
  - At as early as possible stage, estimation of clinical efficacy and toxicity
- **Efficient measures**
  - Use Disease-specific iPS cell
  - Use of Human Bio Big Data in early stage
Biomolecular Profiling
DrugDiscovery/DR
Relation among Drug, Disease and Target

Main causes of Disease

Disease-related molecules

Drug: acting through Drug Target to exert influences on Disease-related genes

Target Molecule of Drug
Receptor, Enzyme, Channel

Human biological system(network)

Disease-related molecules

disease

Drug → Target

receptor

enzyme

channel
Biomolecular profiling Drug Discovery

New approach to computational drug discovery

Computational drug discovery so far
in silico drug discovery

Molecule Centric
- Structure-based rational drug design
- Computational Molecular Docking Design between Target M (receptor, enzyme, channel) and Drug (ligand)
- Molecular Orbital Method, Molecular Mechanics, Molecular Dynamics (MD method)
  ex. influenza drug; zanamivir (relenza)
- Optimization of lead compounds

Quantitative structure-activity relationship
- QSAR: bioactivity and molecular structure
- Between drug and response, biosystem exists
Biomolecular profiling Drug Discovery/DR

New computational Drug Discovery/DR (Biomolecular Profiling)

Gene Expression Profile at Disease Contraction
Induced by Disease-related gene’s activity
Disease-specific gene expression profile change

Gene Expression Profile at Drug Prescription
Genome-wide change of gene expression
Induced by junction between drug and target
Drug-specific gene expression profile change

Comprehensive molecular profile change
⇒ molecular network change of total biosystem

<drug-target molecule docking> exerts an influence to <diseased state>
From the genome-wide viewpoint of total biosystem

Utilize big data DB connecting compounds, target molecules, disease
Basic structure of profile-based computational drug discovery

**Framework of Triple-layer disease and drug network**

- **Drug C** has therapeutic effect to disease **D**

**Profile comparison-based Drug Discovery/DR**

**Molecular network-based Drug Discovery/DR**

**DR**: Drug Repositioning: is the application of known drugs (compounds) to treat new indications (i.e., new diseases)
Common Platform of DrugDiscovery/DR
Protein-Protein interaction network (PPIN)

- Common Platform bionetwork: mediating disease and drug action
- **Protein-protein interaction network (PPIN)** as common platform
- **Disease**: Scaffolding in PPIN: Disease-related protein (gene)
- **Drug**: Scaffolding in PPIN: Drug Target protein
- Based on the **distance (proximity)** between Disease-related protein and target protein, the impact of the drug is measured
Relative Proximity Index $d_c$:
① Distance between Target and the nearest protein among disease-related protein module
② The distance is normalized among the distance of the molecules in same context
  $z < -0.15 \Rightarrow$ proximate
③ Closest measure $d_c$ : best index to measure efficacy
Need for Learning

- We are **still missing in understanding** of the necessary conditions for molecule to be effective
- We should find these conditions by learning from the **succeeded <disease-drug-target molecule> combinations**
- **Artificial Intelligence** (AI), specially **Deep Learning** is now the most powerful method
Artificial Intelligence based Drug Discovery/DR
Artificial Intelligence revolution by Deep Learning

- **Limitation of Machine Learning**
  - “Supervised learning”
    - Construct AI by providing the feature and answer
- **Deep Learning revolution**
  - “Unsupervised learning”
    - Learn the features by itself

Conventional Machine Learning

Data $\xrightarrow{\text{Extracted by operator}}$ feature$^1$, feature$^2$, feature$^3$, feature$^4$ $\xrightarrow{\text{AI}}$ DL learning $\xrightarrow{\text{Learn by provided answer}}$ answer

Deep Learning

Data $\xrightarrow{\text{DL learning}}$ feature$^1$, feature$^2$, feature$^3$, feature$^4$ $\xrightarrow{\text{answer}}$
• Principle of autoencoder: Learn specific intrinsic features of the big data
• Restore the node values of input layer from the node values of next layer where the number of nodes is decreasing compared with input layer. 
  → Intrinsic features should be explored so that the input layer to be recovered as same as possible 
  → discover intrinsic features
Revolutionary point of DL Autoencoder 2

- Consisting of multiple layers of autoencoders
  - stacked autoencoder
- Consisting of input layers (encoder) and output layers (decoder) which are contrasted
  - deep autoencoder
Revolutionary point of DL Autoencoder

- Any number of layers can be used to construct the DL neuronetwork by applying autoencoder to each layer → “autoencoder” for each layer (stacked autoencoder)
- Since the next layer is created using feature quantities learned in the previous layer, higher-order features are created through the hierarchy
- "Supervised learning" is finally necessary to combine feature expressions and human concepts/intention
- Overcome the limits of previous learning methods by automatic feature extraction
  → Structural understanding by intrinsic features
  → e.g. AlphaGo which exceed human ability
Problem: number of attribute value \((p)\) \(\gg\) number of sample \((n)\)

- \(p\): may be several hundred million
- \(n\): At most tens of thousands, normally thousands

If all these huge number of attribute variables are independent, structural analysis of big data is impossible.

Big data is composed of a large number of independent components but less than the number of attribute values as its basis.
Contracting method for multidimensional network by Deep Learning

- Application to big data of medicine and drug discovery
- Increase of “multidimensional network information structure”
  - Genome medicine: <comprehensive molecular information
    – clinical phenotype information>
  - Genome cohort: <gene – environment information>
    (lifestyle)
- Deep Learning-based Network Contraction
  Multidimensional network information structure
  ⇒ Contact to be composed of a few network bases
- Not linear decomposition.
- Project to be composed of intrinsic bases by nonlinear decomposition.
Decomposition to intrinsic network basis

Contracting to sum of intrinsic network bases

DL is applied to iterative presentation of element NW, where all network links are viewed as a sum of element NW (partial NW)

Sum of intrinsic network basis that reproduces all networks

PPI Network
Deep Learning Architecture

input layer

hidden layers

output layer

Basal cell carcinomas
- Epidermal benign
- Epidermal malignant
- Melanocytic benign
- Melanocytic malignant

Squamous cell carcinomas

Melanomas

Nevi

Seborrhoeic keratoses
Our Approach

• By using deep learning and genome-wide protein interaction network,
• We build a computational framework to predict potential Drug Target genes and
• Repositionable drugs for Alzheimer’s disease.
Our computational workflow

**Step 1: Input data**

- Genome-wide PIN
- Drugs and their targets information

**Step 2: Feature Engineering**

- Feature engineering by "deep autoencoder" and a state-of-the-art feature selection algorithm
  - Dimensional reduction by "deep autoencoder"
  - Feature selection by Relief algorithm
Step 3: Classifier model

A binary classifier model to target prioritization by state-of-the-art machine learning algorithms

SMOTE algorithm to build a training data

Xgboost algorithm to build a binary classifier

Step 4: Target prioritization

Scores for potential targets

<table>
<thead>
<tr>
<th>Gene</th>
<th>Score (mean probability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRASP</td>
<td>0.982971499</td>
</tr>
<tr>
<td>PGRMC1</td>
<td>0.98234516</td>
</tr>
<tr>
<td>GPM6A</td>
<td>0.98234516</td>
</tr>
<tr>
<td>NRP2</td>
<td>0.975193546</td>
</tr>
<tr>
<td>PFKM</td>
<td>0.972127568</td>
</tr>
<tr>
<td>DLGAP2</td>
<td>0.953659343</td>
</tr>
<tr>
<td>CD81</td>
<td>0.941095327</td>
</tr>
<tr>
<td>IQGAP1</td>
<td>0.926867425</td>
</tr>
<tr>
<td>TROVE2</td>
<td>0.916866333</td>
</tr>
<tr>
<td>TOP3B</td>
<td>0.915745951</td>
</tr>
<tr>
<td>TJP1</td>
<td>0.914564961</td>
</tr>
<tr>
<td>PDGFB</td>
<td>0.914082375</td>
</tr>
<tr>
<td>SETD2</td>
<td>0.905462331</td>
</tr>
<tr>
<td>CFLAR</td>
<td>0.900456615</td>
</tr>
<tr>
<td>PROS1</td>
<td>0.883435477</td>
</tr>
<tr>
<td>SIT1</td>
<td>0.879989294</td>
</tr>
<tr>
<td>SIGLEC7</td>
<td>0.879989294</td>
</tr>
<tr>
<td>SHC2</td>
<td>0.879989294</td>
</tr>
</tbody>
</table>
By using the AI-based method, we successfully predict potential drug targets (more than 100 genes) for Alzheimer’s disease.
Example,

**SLC25A38 (APPOPTOSIN)**

SLC25A3 increases in the brain from Alzheimer’s disease patients as well as from infarct patients. Further, SLC25A38 downregulation is likely to inhibit apoptosis induced by Bax/BH3I and neuronal death induced by Aβ/glutamate.

**Appoptosin is a Novel Pro-Apoptotic Protein and Mediates Cell Death in Neurodegeneration**

Han Zhang, Yun-wu Zhang, Yaomin Chen, Xiumei Huang, Fangfang Zhou, Weiwei Wang, Bo Xian, Xian Zhang, Eliezer Masliah, Quan Chen, Jing-Dong J. Han, Guojun Bu, John C. Reed, Francesca-Fang Liao, Ye-Guang Chen, and Huaxi Xu

Possible DR for Alzheimer’s Disease

- Computational network based methods may be among the most promising approaches for computational drug repositioning.
- Especially, drug-disease-target network would be useful resources to investigate novel indications for existing drugs.
- We mapped predicted targets on the drug-disease-target network and analysed the network to investigate novel indications for existing drugs.
If predicted target for disease A is known drug-target of drug R for disease B, the drug R may be repositionable drug for disease A.
If predicted target for disease A is known drug-target of drug R for disease B, the drug R may be repositionable drug for disease A.
Potential (predicted) repositionable drugs for Alzheimer’s disease

<table>
<thead>
<tr>
<th>repositionable drug</th>
<th>target</th>
<th># of target category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tamoxifen</td>
<td>PRKCB PRKCE PRKCG ESRRG</td>
<td>4 Anti-Estrogens; Antineoplastic Agents; Antineoplastic Agents</td>
</tr>
<tr>
<td>Mianserin</td>
<td>SLC6A4 DRD3 OPRK1 ADRA1B</td>
<td>4 Adrenergic Agents; Adrenergic alpha-Antagonists; Adrenergic beta-Antagonists</td>
</tr>
<tr>
<td>Amitriptyline</td>
<td>SLC6A4 OPRK1 ADRA1B OPRM1</td>
<td>4</td>
</tr>
<tr>
<td>Dextromethorphan</td>
<td>SLC6A4 PGRMC1 OPRM1 OPRK1</td>
<td>4 Alkaloids; Antitussive Agents; Central Nervous System</td>
</tr>
<tr>
<td>Mirtazapine</td>
<td>OPRK1 ADRA1B DRD3 SLC6A4</td>
<td>4 Adrenergic Agents; Adrenergic alpha-Antagonists; Alkaloids</td>
</tr>
<tr>
<td>Tramadol</td>
<td>OPRM1 OPRK1 SLC6A4</td>
<td>3 Alcohol; Amines; Analgesics; Analgesics, Opioid; Central Nervous System</td>
</tr>
<tr>
<td>Zinc</td>
<td>MPG SERPINA1 SERPIN1D1</td>
<td>3 Acetates; Acetic Acid; Acids; Acids, Acyclic; Acids, Non-Cyclic; Acids, Nuclear</td>
</tr>
<tr>
<td>Amoxapine</td>
<td>SLC6A4 DRD3 ADRA1B</td>
<td>3 Adrenergic Agents; Adrenergic Uptake Inhibitors; Alkaloids</td>
</tr>
<tr>
<td>Etorphine</td>
<td>OPRM1 OPRK1 OPRL1</td>
<td>3 Analgesics; Analgesics, Opioid; Central Nervous System</td>
</tr>
<tr>
<td>Tapentadol</td>
<td>OPRM1 OPRK1 SLC6A4</td>
<td>3 Analgesics; Analgesics, Opioid; Benzene Derivatives</td>
</tr>
<tr>
<td>Loxapine</td>
<td>ADRA1B DRD3 SLC6A4</td>
<td>3 Antipsychotic Agents; Antipsychotic Agents (First Generation)</td>
</tr>
<tr>
<td>Pethidine</td>
<td>OPRK1 OPRM1 SLC6A4</td>
<td>3 Acids, Heterocyclic; Adjuvants; Adjuvants, Anesthetic; Analgesics</td>
</tr>
<tr>
<td>Talampanel</td>
<td>GRIA1</td>
<td>1 Benzazepines; Heterocyclic Compounds; Heterocyclic Compounds</td>
</tr>
<tr>
<td>Etanercept</td>
<td>FCGR3B</td>
<td>1 Amino Acids, Peptides, and Proteins; Analgesics; Anti-Inflammatory</td>
</tr>
<tr>
<td>Vitamin E</td>
<td>PRKCB</td>
<td>1 Antioxidants; Benzopyrans; Chemical Actions and Uses</td>
</tr>
<tr>
<td>N-(2R)-2-benzyl-4-(hydroxyamino)-4-LTA4H</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Adalimumab</td>
<td>FCGR3B</td>
<td>1 Amino Acids, Peptides, and Proteins; Anti-Inflammatory</td>
</tr>
<tr>
<td>ALPHA-HYDROXYFARNESYLPHOSPH FNTB</td>
<td>1</td>
<td>Alcohol; Fatty Alcohol; Hydrocarbons; Lipids; Organ</td>
</tr>
</tbody>
</table>
Example,
The two FDA-approved drugs, adalimumab and etanercept, may be most promising candidates, because they are inhibitors of TNF-alpha (a key cytokine to regulate immune response) and overexpression of TNF-alpha cause inflammation in various organs, especially in central nerve system.
Our computational method would be a promising approach to investigate novel drug targets and new indications for FDA-approved drugs.

We are now applying the method to investigate novel targets and repositionable drugs for various diseases including various types of cancers, rheumatoid, diabetes and etc.
Future strategies and trends

• Second stage of **genomic medicine and drug discovery**
• Contracting method for multidimensional network by Deep Learning
  – Apply to big data medicine
  – Correlative network structure of comprehensive molecular information – clinical phenotype in genome medicine
  – Disease onset and genetic – environment factor in biobank
• Framework of AI drug discovery has possibility to achieve
• Undertake the implementation of AI drug discovery at the end of this year. Otherwise all will be taken to the United States.
  – Establish the “Big data medicine/AI drug discovery consortium” to promote the project, coordinated by pharmaceutical company, IT company and medical institution