Deploying a Pharmacoinformatics Grid for Integrative Biomedical Researches

Jung-Hsin Lin (林榮信)
School of Pharmacy, National Taiwan University & Institute of Biomedical Sciences, Academia Sinica
http://rx.mc.ntu.edu.tw/~jlin/
Pharmacoinformatics integrates Bioinformatics and Chemoinformatics for Drug Discovery
Identifying Drug Targets using Microarray

Hierarchical clustering of expression profiles and drug response. Independent clustering of a set of cell lines according to (A) gene expression level (green = low expression, red = high) and (B) sensitivity to a panel of drugs (blue = low sensitivity, yellow = high) yields clusters. (C) Computing the correlation between gene expression level and drug sensitivity across the cell lines may identify candidate genes that are implicated in the drug response. (D) An example from Scherf et al. (2000) shows how high levels of expression of the gene asparagine synthetase in a set of 60 cancer cell lines affords some protection against the drug L-asparaginase. The correlation computed for a subset of six leukemia cell lines (orange circles) was reported to be significant ($R = -0.98$, $P < 0.01$).

The compendium approach. This technique is used to identify genes and treatments that act on similar pathways. Profiling a large number of mutations or treatment conditions identifies clusters of genes that are co-regulated across a range of conditions; for example, here genes in clusters A1 and A2 have opposite effects, while those in B and C display specific features of interest. Simultaneous hierarchical clustering by treatment (across the top of the figure) also groups conditions that lead to similar overall transcription profiles, allowing researchers to generate hypotheses as to the functions of the mutant genes, drugs, or other environmental agents that led to these perturbations in gene expression.
GenePathway Viewer

GenePathway Viewer is a web server that can be used to visualize gene expression levels and correlated genes on the maps of the biological pathways. GenePathway Viewer, different from other existent similar pathway viewers, is facilitated directly by the web service provided by the KEGG API and will acquire the most up-to-date pathway information in the KEGG database. Web service is an emergent powerful technology identified by a Universal Resource Identifier (URI), whose public interfaces and bindings are defined by XML. A distinct feature of web service is that the constituent software components, which are communicated via the Simple Object Access Protocol (SOAP), can be loosely coupled, in contrast to the more traditional client-server models that are very tightly coupled. On the other hand, GenePathway Viewer is also a meta-server that can combine various resources on gene identification and gene annotation information. With these integrated features, GenePathway Viewer will help expedite the understanding of gene functions and their intercorrelation or causal relationships upon different medicinal treatments.
GenePathway Viewer

<<Step 1>> Upload file.

Show Tips

Upload File: C:\Documents and Setting

submit Reset
<<Step 2>> Fill out the table.

- Choice the species of the genes:
  - Mouse (Mus musculus)
  - Human (Homo sapiens)
  - Rat (Rattus norvegicus)
  - Dano rerto
  - Drosophila melanogaster
  - Caenorhabditis elegans
  - Arabidopsis thaliana

Or input the three-letter KEGG species code: 
You can check the species code from here.

- Set threshold to find a pathway: 

- Connect to other database (Swissprot and pir): 

- E-Mail: jin@ntu.edu.tw, jin@ha.mc.ntu.edu.tw

If you want to send the result to more than one user, use comma (,) to separate address strings.
TETRACHLOROETHENE DEGRADATION

Cluster Color Schema

Pathway:
Tetrachloroethene degradation - Mus musculus (mouse)

Condition:
MAS
Animate
1 Sec.

FL: Fold log ratio (base 2)
FL ≤ 8
6 ≤ FL < 8
4 ≤ FL < 6
2 ≤ FL < 4
1 ≤ FL < 2
1 < FL ≤ -1
-2 < FL ≤ -1
-4 < FL ≤ -2
-6 < FL ≤ -4
-8 < FL ≤ -6
FL ≤ -8

S: Swissprot
P: Pir

Acetyle
11.1
Ethylene oxide
11.1
Ethylene glycol
42.1
Acetyl-CoA
Pyruvate metabolism
Glutamate and dicarboxylic metabolism

08625 380482
Constructing Biological Pathways and Networks
Mathematical Modeling of Signaling Pathways

Coupled Differential Equations for Biological Pathways

\[
\begin{align*}
\frac{dx_2}{dt} &= k_{-1}x_3 + k_{-3}[PTP]x_5 - k_x x_2 + k_{-4}x_6 - k_4 x_2 \\
\frac{dx_3}{dt} &= k_1 x_1 x_2 - k_{-1} x_3 - k_3 x_3 \\
\frac{dx_4}{dt} &= k_2 x_1 x_5 + k_{-2} x_4 - k_{-4} x_7 + k_4 x_4 \\
\frac{dx_5}{dt} &= k_3 x_3 + k_{-2} x_4 - k_2 x_1 x_5 + k_{-3}[PTP]x_5 - k_{-4} x_8 - k_4 x_5 \\
\frac{dx_6}{dt} &= k_5 + k_{-5} x_6 + k_6[PTP](x_7 + x_8) + k_4 x_2 - k_{-4} x_6 \\
\frac{dx_7}{dt} &= k_4 x_4 - k_{-4} x_7 - k_6[PTP]x_7 \\
\frac{dx_8}{dt} &= k_4 x_5 - k_{-4} x_8 - k_6[PTP]x_8 \\
\frac{dx_9}{dt} &= k_{-7}[PTP]x_{10} - k_{7} x_9 (x_4 + x_5)/(IR_p) \\
\frac{dx_{10}}{dt} &= k_7 x_9 (x_4 + x_5)/(IR_p) + k_{-8} x_{12} - (k_{-7}[PTP] + k_8 x_{11})x_{10} \\
\frac{dx_{11}}{dt} &= k_{-8}x_{12} - k_8 x_{10}x_{11} \\
\frac{dx_{12}}{dt} &= k_8 x_{10} x_{11} - k_{-8} x_{12} \\
\frac{dx_{13}}{dt} &= k_9 x_{14} + k_{10} x_{15} - (k_{-9}[PTEN] + k_{10}[SHIP])x_{13} \\
\frac{dx_{14}}{dt} &= k_{-9}[PTEN]x_{13} - k_9 x_{14} \\
\frac{dx_{15}}{dt} &= k_{-10}[SHIP]x_{13} - k_{10} x_{15} \\
\frac{dx_{16}}{dt} &= k_{-11} x_{17} - k_{11} x_{16} \\
\frac{dx_{17}}{dt} &= k_{11} x_{16} - k_{-11} x_{17} \\
\frac{dx_{18}}{dt} &= k_{-12} x_{19} - k_{12} x_{18} \\
\frac{dx_{19}}{dt} &= k_{12} x_{18} - k_{-12} x_{19} \\
\frac{dx_{20}}{dt} &= k_{-13} x_{21} - (k_{13} + k_{13'})x_{20} + k_{14} - k_{-14} x_{20} \\
\frac{dx_{21}}{dt} &= (k_{13} + k_{13'})x_{20} - k_{-13} x_{21}
\end{align*}
\]
Virtual Screening after Target Identified

1. Start with crystal coordinates of target receptor and locate the active site

2. Generate the molecular surface for the receptor

3. Search for the optimal position and location based on some scoring function

4. Pick up the conformations (or compounds) with best scores
State Vector in the Flexible Docking Problem

\[(x_{CM}, y_{CM}, z_{CM}, \phi, \theta, \psi, \chi_1, \chi_2, \cdots, \chi_k)\]
Characteristics of Biological Complex Problems

• The potential energy function is extremely rugged.
• The potential energy surface is usually highly asymmetric.
• The true global minimum is often surrounded by many deceptive local minima.
• The biological complex problems are mostly in the space of high dimensionality.
How to explore the phase space?
(Or, how to find a needle in a haystack?)
---Importance sampling

- We should only explore the important region of the phase space, not the entire phase space.
- Stochastic methods usually outperform deterministic approaches in higher dimensional space.
Genetic Algorithm

1. **[Start]** Generate random population of \( n \) chromosomes (suitable solutions for the problem)
2. **[Fitness]** Evaluate the fitness \( f(x) \) of each chromosome \( x \) in the population
3. **[New population]** Create a new population by repeating following steps until the new population is complete
   a. **[Selection]** Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)
   b. **[Crossover]** With a crossover probability cross over the parents to form new offspring (children). If no crossover was performed, offspring is the exact copy of parents.
   c. **[Mutation]** With a mutation probability mutate new offspring at each locus (position in chromosome).
   d. **[Accepting]** Place new offspring in the new population
4. **[Replace]** Use new generated population for a further run of the algorithm
5. **[Test]** If the end condition is satisfied, **stop**, and return the best solution in current population
6. **[Loop]** Go to step 2
Chromosomes for GA Docking

Crossover operation

Lamarckian Genetic Algorithm

- LGA is a hybrid of the Genetic Algorithm with the adaptive local search method.
- As in the GA scheme, energy is regarded as the phenotype, and the compound conformation and location are regarded as the genotype.
- In the LGA scheme, phenotype is modified by the local searcher, and then the genotype is modified by the locally optimized phenotype.
- In AutoDock, the so-called Solis-Wet algorithm is used (basically energy-based random move).
A Maximum Entropy Evolutionary Algorithm for the Docking Problem


- $n$ individuals, denoted by $s_1, s_2, \ldots, s_n$, are generated. Each $s_i$ is a vector corresponding to a point in the domain of the objective function $f$. In order to achieve a scale-free representation, each component of $s_i$ is linearly mapped to the numerical range of $[0,1]$.

- The individuals in each generation of population are then sorted in the ascending order based on the values of the energy function on evaluated on these individuals. Let $t_1, t_2, \ldots t_n$ denote the ordered individuals and we have $f(t_1) < f(t_2) < f(t_n)$.

- $n$ Gaussian distributions, denoted by $G_1, G_2, \ldots G_n$, are generated before the new generation of population is created. The center of each Gaussian distribution is selected randomly and independently from $t_1, t_2, \ldots t_n$, where the probability is not uniform but instead follows a discrete diminishing distribution, $n : n-1 : \ldots : 1$.

$$p(s'_i) = \left( \frac{1}{\sqrt{2\pi} \cdot \sigma_i} \right) \exp \left( -\frac{(s'_i - \mu_i)^2}{2\sigma_i^2} \right)$$

$$\sigma_i^2 = \alpha + \frac{(\beta - \alpha)i}{n - 1}$$
LGA versus ME

- The ME algorithm avoids the "purification" effect inherent in the genetic algorithm and its derivatives, and therefore reduce the over-compression of information in the searching process.
Figure 2. Molecular graphics display of the four benchmark cases: (a) HIV-II protease complexed with its inhibitor L-735,524 (PDB ID: 1HSH); (b) FKBP-FK506, an immunophilin-immunosuppressant complex (PDB ID: 1FKP); (c) Complex formed between phospholipase A2 and aspirin (PDB ID: 1OXR); and (d) TATA-box binding protein (YTBP) complexed with DNA containing a TATA-box (PDB ID: 1YTB).
Figure 3. Number of runs to reach convergence versus the number of energy evaluations consumed (in units of $10^7$): blue, MEDock results; magenta, LGA results (with parameters tuned); green, LGA results (with default parameters). (a) HIV-II protease complexed with its inhibitor L-735,524 (PDB ID: 1HSH); (b) FKBP-FK506, an immunophilin-immunosuppressant complex (PDB ID: 1FKF); (c) complex formed between phospholipase A2 and aspirin (PDB ID: 10XR); and (d) TATA-box binding protein (YTBP) complexed with DNA containing a TATA-box (PDB ID: 1YTB).
MEDock: a Maximum-Entropy based docking web server for efficient prediction of ligand binding sites

The MEDock (Maximum-Entropy based Docking) web server is aimed at providing an efficient utility for prediction of ligand binding site. A major distinction in the design of MEDock is that its global search mechanism is based on a novel optimization algorithm that exploits the maximum entropy property of the Gaussian distribution.

Quick Start

There are three ways to use MEDock. If this is your first time to use MEDock, you need to set up the following parameters:

2. Three-dimensional structures of interacting proteins.
3. Ligand(s) of interest.

MEDock at NTU
MEDock at NTU Medicine
MEDock at Y2U

Overexpression of P-glycoprotein is the major cause for multidrug resistance problems in cancer chemotherapies

- Multidrug resistance (MDR) has posed a serious clinical problem in cancer chemotherapy.
- MDR will cause the reduction of bioavailability of drugs.
- P-gp, product of the \textit{mdr1} gene in humans, localization on chromosome 7q21, is a member of the large ATP binding cassette (ABC) family of proteins.
- P-gp is 1280 amino acids long and is very dynamic inside membranes.
Structure Prediction the MDR Protein Pgp

1. Predicting the structure of Pgp using homology modeling
2. Molecular dynamics simulation in a lipid bilayer
Paclitaxel (Taxol)

- THR-199 [TM3]
- PHE-303, TYR-307, PHE-314 [TM5]
- SER-344, VAL-345, GLN-347 [TM6]
Summary

- Grid is the ideal computing architecture that enables integrative biomedical and pharmaceutical researches, which often require access to heterogeneous computing resources.
- The GenePathway Viewer allows conversion of microarray data into colored pathway information, which uses the Web Service technology that can update the pathway database from KEGG in a real-time and automatic fashion.
- Molecular dynamics simulations of biomacromolecules usually generate huge amount of data in a very high speed, and therefore good archiving facilities like DataGrid is important to ensure data security and integrity.
- The ME algorithm is intrinsically parallel, and therefore straightforward to be implemented on the Grid architecture.
Acknowledgement

Computer Science and Information Engineering, NTU
• Prof. Yen-Jen Oyang
• Tien-Hao Chang

School of Pharmacy, NTU
• Pei-Hua Lo
• Hui-Hsuan Tu

National Science Council of Taiwan