教師あり学習による遺伝子ネットワーク推定エンジンの開発

Development of gene network inference engine based on supervised analysis

化学生命科学領域　　小寺正明

背景と目的

多くの生物学的機能は遺伝子間やタンパク質間の相互作用によるものであり、ゲノム科学はそのような相互作用による生物の振る舞いをシステムレベルで予測し実用に役立てることを目的の一つとしている。近年はトランスクリプトームやプロテオームなどの技術も発達し、遺伝子やタンパク質に関する大規模データも急速に蓄積してきている。これらの異種データは、分子ネットワークを予測するのに有用なリソースであり、効果的にデータ統合を行う方法の開発が望まれている。特に、実験分子生物学者にも使いやすい分子ネットワーク予測ツールの開発は、バイオインフォマティクス技術の実用化と普及という点で必要不可欠であると考えられる。このため本研究では、インターネットを介してウェブブラウザ上で扱える柔軟なネットワーク予測プログラム GENIES (GEne Network Inference Engine based on Supervised analysis) の開発をおこなった。

検討内容

本研究で重視したのはデータ形式および、アルゴリズムや教師データセットを柔軟に選択できるインターフェースである。タブ区切り形式のテキストファイルであればどのような種類のデータも入力に使うことができるよう設計した。例えば、遺伝子発現プロファイル、タンパク質局在プロファイル、生物系統プロファイル等は「profile」型のタブ区切りファイルとして入力可能であり、様々な定義による遺伝子間類似性行列は「kernel」型のタブ区切りファイルとして入力可能とした。教師付き学習を行うためのトレーニングデータセットとして、ユーザは KEGG PATHWAY にある既知の分子ネットワークを用いることも出来るし、ユーザ独自のネットワークデータを用いることもできるようにした。ネットワーク推定のアルゴリズムも、ユーザが選択でき、パラメータ調整やデータ統合時の重み付けもユーザが自由に設定できるようにした。

結果

ウェブサーバーGENIES を、ゲノムネットのサービスの一つ http://www.genome.jp/tools/genies/ として開発して公開した。GENIES は教師付きネットワーク推定によって遺伝子ネットワークを予測するウェブサーバーであり、既知のネットワーク情報と様々な種類のデータを統合し、カーネル法を用いた予測を行う(図1)。出力として、予測された遺伝子ペアのリストや、KEGG PATHWAY 上へのマッピングなどを行うことができるようにした。

考察

GENIES で用いているアルゴリズムの有用性は過去に発表しており (山西ら、2005)、例えば注目している生

物種の代謝経路中のミッシング酵素（存在すると思われているが遺伝子配列が明らかでない酵素）の推定などに用いることができると考えられ、実際に実験生物学者との共同研究によりミッシング酵素の同定に役立てた例がある（山西ら、2007）。本研究ではそれをウェブサーバーとすることで、インターネットが使えるパソコンであれば誰でもアクセスして計算を行えるようにした点で、実験分子生物学者にも比較的利用しやすいものになっている。今後は実験分子生物学者に広く宣伝し、実際に使ってもらって意見を収集し、より使いやすいものになるよう改良を進めて行く予定である。

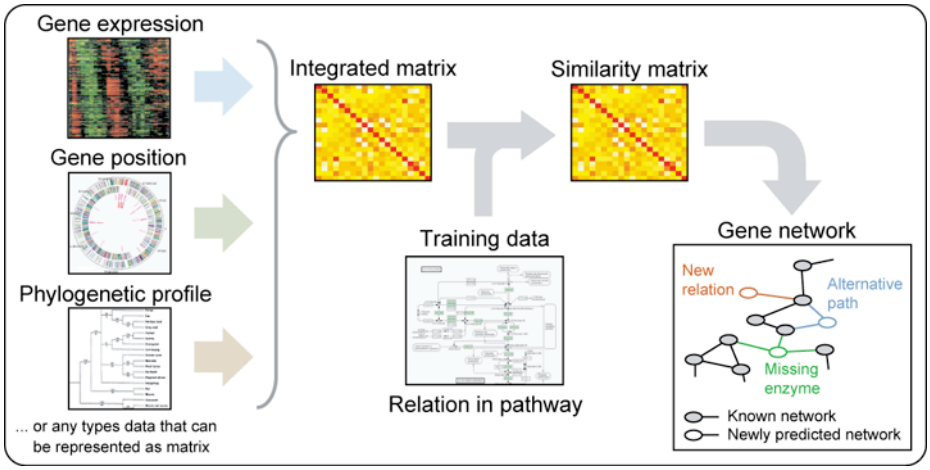

図1、GENIES の概略

発表論文

- Kotera, M., Yamanishi, Y., Moriya, Y., Kanehisa, M. and Goto S. (2012) GENIES: gene network inference engine based on supervised analysis. *Nucleic Acids Res.* 40 (Web Server issue):W162–167.

参考論文

- Yamanishi, Y., Vert, J. P. and Kanehisa, M. (2005) Supervised Enzyme Network Inference from the Integration of Genomic Data and Chemical Information. *Bioinformatics*, 21, i468–i477.
- Yamanishi, Y., Mihara, H., Osaki, M., Muramatsu, H., Esaki, N., Sato, T., Hizukuri, Y., Goto, S. and Kanehisa, M. (2007) Prediction of missing enzyme genes in a bacterial metabolic network. Reconstruction of the lysine–degradation pathway of Pseudomonas aeruginosa. *FEBS J.*, 274, 2262–2273.