実験データから生物学的知識へ：機械学習によるアプローチ

Jean-Francois (Kenichi) Pessiot
細胞機能設計チーム
産総研特別研究員

私の研究の興味は、実際の生物学的問題を解決することに基づいています。目標としては、生物学的または臨床試験からの実際のデータを利用して、効率的なデータ解析アルゴリズムを設計・実装し、検証することです。こうしたアルゴリズムは、データの基礎となる隠れた生物学的過程を解読・理解することを支援し、新たな生物医学的・生物学的洞察を提供するはずです。

計算生物学という観点から、こうした一般的な手法を2つの異なる問題に適用しています。1つ目は、次世代シーケンス・データを活用して、転写因子結合モチーフ（TFBM）を発見するためにPeakRegressorという新規手法を開発しました。PeakRegressorは、ChIP-Seqデータを使って、STAT1およびRNAポリメラーゼIIのTFBMを同定するのに成功し、更に、迅速かつ容易に解釈できます。

図1

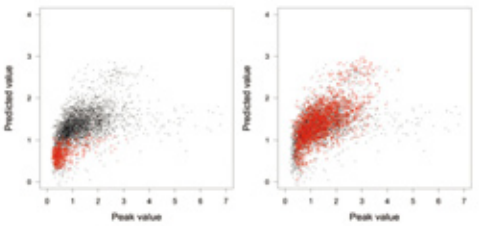

図2

Filtering method	#Peaks (STAT1/Pol II)	STAT1	Pol II
None	36998/24739	0.50	0.44
Promoter proximity	3,907/9,094	0.41	0.53
Q-value $<10^{-3}$	16639/17580	**0.65**	**0.66**

The correlation coefficients are averaged in 30-fold cross-validation.
doi:10.1371/journal.pone.0011881.t001

図1（左図）2種類のフィルタリング手法と組み合わせたSTAT1回帰の結果：Q値（右）、プロモーターへの近接（左）。
図2（右図）テスト・データセットでのピーク値と予測値との間の相関係数に対するピーク・フィルタリング手法の影響。

2つ目の問題は、プラットフォームを越えてマイクロアレイ・データを解析することです。私が開発したペアワイズ・ランキング成分分析（Pairwise Ranking Component Analysis；PARCA）は、誘導多能性幹細胞（iPS細胞）と胚性幹細胞（ES細胞）を分析するための新たな距離学習方法を取り入れています。PARCAは、iPS細胞、ES細胞とがん細胞の異なるサブタイプを識別するのに成功しています。

図3

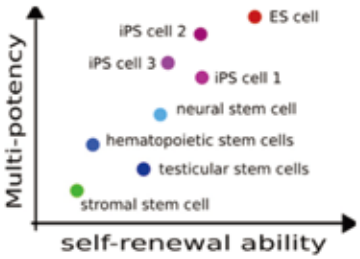

図4

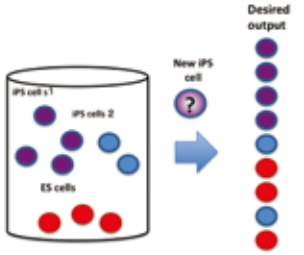

図3（左図）幹細胞とiPS細胞は、様々なレベルの多能性と自己再生能力があります。
図4（右図）iPS細胞とES細胞のデータベースを利用して、PARCAは未知のiPS細胞を識別することができます。

私は現在、遺伝子ネットワークの解析・推定に取り組んでいます。遺伝子ネットワークは、遺伝子とその相互作用の集積です。遺伝子とその相互作用は、共に、遺伝子産物の量を支配し、メカニズムが動作している細胞において主要な役割を果たしています。実験による測定から、こうした遺伝子ネットワークを推定することは、それらのメカニズムや調節異常に関わる疾患をより理解するのに重要な一歩となります。

そこで、現在、私は、遺伝子発現の時系列から遺伝子制御ネットワークを推定するために、ReverSimという新しい方法を開発しています。ReverSimは、利用可能な遺伝子相互作用情報が少ない場合に、利用可能な遺伝子相互作用と時系列の類似性を活用して、遺伝子ネットワークを再構築できます。長期的な目標は、細胞分化、代謝、細胞周期などの生命の重要な過程を理解することです。

図5

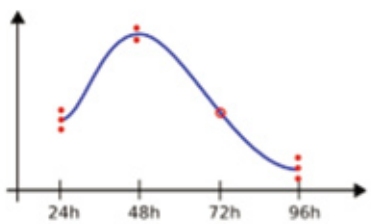

図6

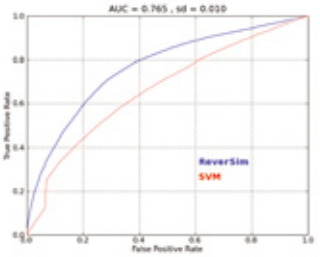

図5（左図）ReverSimを活用して、遺伝子発現の時系列が多項式曲線で近似できます。
図6（右図）初歩的な結果ですが、ReverSimが、機械学習法のSVMを用いた方法と比べても引けをとらないことが分かります。（原文英語）