論文内容の要旨

論文題目

A method based on multiple SVMs for a comprehensive prediction of protein-protein interactions

複数の SVMs に基づく網羅的タンパク質間相互作用予測手法

氏名

道菅　紳介

生物を統合化されたシステムとして理解する上で、タンパク質間の物理的な相互作用（Protein-Protein Interaction, PPI）を網羅的に調べることは重要である。しかしながら現状では、ゲノムが明らかにされた全ての生物種について網羅的 PPI 検出実験を行うことは費用、時間や労働力の制約上困難である。また、これまで明らかにされている PPI は生体内で起こるもののごく一部にすぎず、また偽陽性も多いため、それらに基づく解析は生体システムの誤った理解を招く可能性が指摘されている。そこで本研究では、より効率的な PPI の発見を可能にし、生体システムに関する確かな理解を導く目的で、Support Vector Machines（SVMs）に基づく網羅的 PPI 予測手法を開発した。

近年の PPI 予測手法は、主に酵母をターゲットとし、タンパク質ドメインを予測指標とすることで大幅な性能改善に成功した。しかしながら、これらの手法には共通して、更なる予測性能の向上を妨げる 2 つの欠点があった。1 つは異なるドメインの 3 つ以上の組み合わせが関与する PPI を考慮できない点で、もう 1 つはドメイン以外の予測指標を追加利用することが困難な点である。提案手法はこれら 2 つの欠点を克服した点に特徴がある。

我々はまず、比較的豊富な相互作用データが蓄積されている酵母を用いて交叉検定を行った。その結果、組み合わせを考慮したドメイン、アミノ酸組成、及び細胞内局在の情報が予測に有効であることが明らかとなった。また、提案手法は既存手法を凌ぐ F 値（精度と感度の調和平均）0.788 を達成することができた。相互作用の有無が未知のタンパク質ペアに対し相互作用予測を行うと、SVM のスコアが高いほど、タンパク質の機能の観点から、その相互作用はもっともらしい事が確認された。また、偽陽性を多く含むとされる PPI 実験データに対し予測を試みたところ、中でも信頼性が高いと考えられるデータのうち

58.6%を正しく予測することができた。これらのことは、本手法がもっともらしい PPI を新規に予測できるのみならず、エラーを含む実験データの信頼性評価にも利用可能であることを示している。次に、我々は哺乳類のタンパク質間相互作用予測を試みた。これまで、ヒト以外の哺乳類については予測器の訓練に十分な PPI データがなく、PPI 予測は困難であった。我々は、マウス等ヒトと進化的に近い生物種についてはヒトのデータで訓練した SVM が有効であることを明らかにし、ヒトについては 0.776、マウスについては 0.765 という高い F 値を得た。この結果は、本手法が酵母のみならず哺乳類の PPI 予測にも適用可能である事を示している。

ある生物がもつタンパク質の全組み合わせを入力データとし、全 PPI（PPI マップ）を予測する問題は、上記のような比較的優れた手法をもってしても膨大な偽陽性が発生することが予想される。そこでまず、我々は本問題が従来考えられていたよりも困難であることを定量的に示し、更なる手法の開発が必要であることを明らかにした。その上で、従来の研究で用いられている負例（相互作用しないタンパク質ペア）は考慮すべき全負例を代表できないとの考えから、複数の SVMs を用いた手法を開発した。酵母とヒトの PPI データを用い予測性能の検証を行ったところ、用いる SVMs の数、及び 1 つの SVM 当たりに使用する負例の数を増やすほど予測性能は改善された。また 1 つ以上の CPU が利用可能なハードウェア環境においては、本手法は予測性能の向上のみならず SVM の訓練時間の削減にも有効であることが明らかとなった。本手法により予測された PPI 及び PPI マップは、タンパク質の機能同定、疾患機構の解明や創薬ための重要なリソースとなることが期待される。