京都大学	博士（薬科学）	氏名	David Alexander duVerle
論文題目	Building a Machine-Learning Framework for Protein Interactions: Calpain Cleavage Prediction and Gene Regulatory Network Inference （タンパク質相互作用に対する機械学習手法の構築：カルパイン基質切断部位の予測及び遺伝子制御ネットワークの推論）		

（論文内容の要旨）

This thesis presents two separate projects in the field of protein interaction. Our first project is split in three chapters, matching the sequential steps in the design and implementation of a new machine-learning framework for the analysis of calpain cleavage and its extension to general proteolysis prediction. In a first stage, we built a publically accessible database of curated calpain-related proteolytic events (I. 1), which then served as experimental support to the development of a calpain cleavage prediction algorithm based on the Multiple Kernel Learning machine-learning framework (I. 2). Finally, we extended the frame of our research on calpain cleavage to reach general conclusions on the prediction of proteolytic events (I. 3). During our second project, we built a new algorithmic framework to refine the prediction of certain motifs in gene regulatory network using high-throughput time-course gene expression data (II).

I. Machine-Learning Framework for Proteolytic Cleavage Prediction

Calpain is a proteolytic enzyme ubiquitously expressed in mammals and many other organisms, which plays a central role in many metabolic processes: by cleaving specific proteins, it helps regulate vital cellular functions such as cell motility or apoptosis. Its dysfunction in humans has been linked to muscular dystrophies, diabetes and tumorigenesis.

Despite its important roles, little is still known about the exact mechanism of calpain regulation and the nature and configuration of the protein sequences it can cleave. Experimentally searching for such substrates and their cleavage location is a time-consuming and costly process that cannot realistically be undertaken exhaustively, hence the need for efficient computational methods that can predict potential calpain substrates along with putative cleavage locations.

Our first task was to build a repository of calpain and calpain-related sequences, with full annotations and cross-references to other databases. A main section of the database was made of a number of curated, experimentally validated, substrate sequences, complemented by an extended set of candidate substrates obtained through standard sequence matching methods, as well as sequence data for different types of calpain proteins and calpastatin (an endogenous inhibitor specific to calpain).

In addition to standard analyses of the data in our repository, we sought to provide advanced computational tools that could be used by experimentalists to quickly screen and select potential calpain substrates of interest. While many algorithms have been developed to predict cleavage by other types of proteases (such as caspases or HIV proteases), they generally do not perform well with calpain, whose mode of action and preferences in substrate sequence seem to require more complex models. A central concept of statistical prediction of cleavage by a generic protease is that the exact amino acid sequence of the putative substrate is the main (possibly only) information needed to predict if and where cleavage will occur. The majority of statistical methods therefore attempt to gather statistically significant sequence features (such as the existence of a given amino acid at a specific position) from known substrates in order to predict new substrates. We found that these approaches might not be enough to yield satisfyingly accurate results on calpain and suggested a new statistical method that uses both sequence and a wider set of features to predict cleavage. Our method is based on "Multiple Kernel Learning" (MKL) a novel extension to the well-established framework of Support Vector Machines that allows us to combine multiple heterogeneous categories of input features into a unified machine-learning model. For instance, we were able to add hints about secondary structure (spatial configuration) and chemical properties of the candidate proteins, in addition to the traditional amino acid sequence information. We also combined different mathematical models in each components ("sub-kernels") of the predictor in order to match

different aspect of the underlying biological process.

Our novel approach has produced a cleavage predictor for calpain that outperforms all existing methods and has the potential to be applied to other difficult sequence prediction problems. In addition, it has lead to some new insight into the biological mechanism of calpain by highlighting subtle differences between substrate affinities across calpain sub-families (calpain 1 and 2), theretofore considered to have identical modes of action.

Our extensive work on cleavage prediction lead to an important amount of benchmark data comparing the strengths and weaknesses of the different machine-learning techniques currently used to treat cleavage prediction problems. We summed up our findings in an exhaustive overview of cleavage prediction methods, providing a formal basis for approaching this class of problems.

II. Gene Regulatory Network motif inference

Our second project focussed on a different set of problems linked to Gene Regulatory Network (GRN) inference. In this project, we aimed to use mathematical constraints based on biophysical models of gene interactions in order to identify characteristic regulatory motifs (such as feedback or feed-forward loops) that are not well predicted by existing methods. Decomposing gene expression time-series using Fourier series, we are able to compare the fitness of different models based on ordinary differential equations (ODE) for a given set of nodes in the GRN. Our approach provides a computationally efficient algorithm for refining the output of existing GRN inference methods, with the potential to be extended to a standalone method.

（続紙　２　）

（論文審査の結果の要旨）

データから内在する仮説やルールを計算機により抽出する「機械学習」と呼ばれる研究領域は、１９８０年頃に始まり１９９０年代に様々な技術が開発され、２１世紀に入り大きく成熟した、計算機科学で広範かつ活発に研究がなされている分野である。近年機械学習は、様々な応用領域で利用され、各応用に適した独自の進化をも遂げている。生命科学も応用の例外ではない。具体的には、遺伝子発現、タンパク質・基質相互作用、タンパク質相互作用といった多様で大量のデータは既に人手による解析がはるかに困難な規模に達しており、古典的な多変量解析や統計科学、そして現代に急速に進展してきた機械学習による解析が必要となるのは自然である。本論文は、生命科学データ、特にタンパク質（遺伝子）相互作用への機械学習技術の適用、および対象としたデータに沿った機械学習技術の検討を行い、生命科学データ、特にある種のタンパク質相互作用に特定して、効率的なデータ解析手法を構築・実装したものである。

本論文で着目したタンパク質相互作用は主に以下の２つに分けられる：１）まずカルパインと基質の相互作用、実際に構築した手法の目的はカルパインの基質切断部位の予測である。２）遺伝子発現制御ネットワーク、実際に構築した手法の目的は時系列遺伝子発現データからの遺伝子発現制御ネットワークの推定である。

上記１）は、本論文の１章から３章に渡っている。カルパインは、システインプロテアーゼの一つであり、タンパク質を基質とする。カルパインの大きな特徴は、プロテオリシス、すなわち基質であるタンパク質を切断により失活させるのではなく、基質であるタンパク質の活性を制御し、それにより信号伝達等の現象をも制御することにある。さらに、カルパインは、アルツハイマー疾患、糖尿病等多岐に渡る疾病との関連が指摘されており、プロテオリシスを行うプロテアーゼの中でも、特に興味深いタンパク質の一つである。著者は、このカルパインが対象とする基質の切断部位を機械学習により予測を行うために、まず、カルパイン研究者と協力し、基質及び切断部位を中心とした、カルパイン関連のデータ収集を行った。さらに、このデータを、１）カルパイン、２）基質、３）カルパイン阻害剤（カルパスタチン）の３種類に分け、CaMPDBという名前のデータベースとして構成し、誰もがアクセス可能なWWW上で公開(http://calpain.org)した。この結果については第１章にまとめられている。続いて、著者は、得られたデータに適用する機械学習技術を構築した。カルパイン基質の切断部位を予測する問題を、所与のアミノ酸配列（または断片）を切断するか否かという２値予測（分類）問題と設定し、切断配列（収集したカルパイン基質）と切断されない配列から、２値分類を行う手法を構築した。２値分類に対しては、サポートベクトルマシンと呼ばれる手法が機械学習では既に定評を得ている。ここで、入力となるデータはアミノ酸配列のみならず、アミノ酸の属性（疎水性、サイズ）やさらには二次構造なども入力可能である。そこで、著者は、サポートベクトルマシンの枠組みの中で、そのような複数の入力に重みを付け、予測能力により重みを変えるマルチプルカーネルラーニング（ＭＫＬ）

と呼ばれる手法を採用し、この方法に基づき予測手法を構築した。ＭＫＬによる予測性能は、交差検証法と呼ばれる評価手法で既存の全てのカルパイン基質を用い、AUCと呼ばれる評価基準で83%を達成した。この数字は、既存の機械学習手法、例えばＳＶM等、さらには実在するカルパイン基質切断部位予測手法の性能を統計的に有意に凌駕するものであった。構築したＭＫＬ手法は実装され、CaMPDB内で利用可能である。この結果については、２章にまとめられている。最後に、著者は、カルパインを中心としたシステインプロテアーゼ全般に関して、基質切断部位予測手法を調査するとともに、この予測問題に対してＳＶMのみならず様々な機械学習法の適用可能性を検討した。この結果については、３章にまとめられている。

上記２）については、時系列遺伝子発現データから遺伝子制御ネットワークを構築するために、まず、３つの遺伝子からなるネットワークモチーフを推定し、モチーフをベースにして遺伝子ネットワーク全体を構築する枠組みを設定した。ここで、ネットワークモチーフでの遺伝子同士の制御関係はフーリエ級数分解によりモデル化し、そのパラメータを時系列発現データより推定する手法を構築した。そして、この手法の有用性を人工データにより確認した。この結果については、４章にまとめられている。

以上、本研究は、タンパク質相互作用における重要な２つの問題、カルパイン基質切断部位予測および時系列発現データによる遺伝子制御ネットワークの構築いずれにおいても、有用性の高い新規手法を構築したものである。特に、カルパイン基質切断部位予測においては、基質を含む網羅的なデータベースを構築し切断部位予測手法を実装することにより、カルパイン研究の進展、さらには関連疾病の分子機構や治療に向けて有用な知見を提供するものと評価される。

よって本論文は博士（薬科学）の学位論文として価値あるものと認める。

さらに、平成２４年２月２３日論文内容とそれに関連した口頭試問を行った結果、合格と認めた。

論文内容の要旨及び審査の結果の要旨は、本学学術情報リポジトリに掲載し、公表とする。特許申請、雑誌掲載等の関係により、学位授与後即日公表することに支障がある場合は、以下に公表可能とする日付を記入すること。

要旨公開可能日：　平成　　年　　月　　日以降