若手研究者インターナショナル・トレーニング・プログラム(ITP)

バイオインフォマティクスとシステムズバイオロジーの国際連携教育研究プログラム 応募書類

Name: 西村 陽介

Title: Statistical analysis of heterogeneous high-throughput data for cancer research

Institute: 京都大学化学研究所バイオインフォマティクスセンター

Partner institute of your choice : Centre for Computational Biology, Mines ParisTech

Duration of your choice: 2012 年 1 月 10 日 - 2012 年 3 月 30 日

Plan:

<u>滞在目的</u>

応募者はバイオインフォマティクスセンター・化学生命科学研究領域に所属する後期博士課程の学生であり、現在の研究課題として真核生物における miRNA とゲノム上において重複して存在している遺伝子との機能的関連性について生物種間比較を用いた解析を行なっている。また修士課程時には同所属において植物二次代謝化合物を基質とする糖転移酵素の系統解析を行った。これらの解析は化学生命科学研究領域によって開発されてきたKEGGデータベースの情報を用いて行われている。KEGG データベースは各生物種における代謝パスウェイや病気、医薬品等のシステム情報、遺伝子を中心とするゲノム情報、代謝物質や生体内化学反応等のケミカル情報を統合的に参照できるデータベースであり、化学生命科学研究領域及び応募者は KEGG データベースを中心とした実データ及び生物学的、生化学的知見を基礎に置き、様々な研究課題に対する計算機科学的な解析手法を蓄積してきている。

応募者が滞在を希望するパリ国立高等鉱業学校の CBIO(Centre for Computational Biology)においては、確率モデルや統計的機械学習を用いて、他の研究室との共同研究を通じて癌を中心とする様々な疾患の新規治療法の発見につながる手法の開発に取り組んでおり、具体的にはゲノム配列、タンパク質構造、遺伝子ネットワークを用いたオーダメイド医療や診断手法の開発、次世代シークエンサや細胞観察用顕微鏡画像等の新しい技術に関する情報処理手法、網羅的な化学物質情報を利用した新規創薬スクリーニング手法、またこれらの手法を支える基礎的な理論やアルゴリズムの開発を目標としている。近年では次世代シークエンサを中心とする high-throughput な実験手法が様々な分野で開発されておりそれに対応してデータベースの情報量も増加し大規模化が顕著である。生命現象を理解する上でこれらの大規模データから適切に情報を得ることの必要性は年々高まっておりこの点において統計学的、情報学的手法が非常に有用であることが知られている。

そこで応募者は CBIO の Dr. Jean-Philippe Vert 及び山西芳裕博士の指導の元、上記の手法についての理解を深めることにより化学生命科学研究領域で培った生物学的知見や実データから仮設を立案する方法やそれに関連する解析手法に加え、数理的土台となる統計学的、情報学的手法を習得することを通じて応募者の知識及び技術を広げることにより、バイオインフォマティクス分野の研究課題に関し幅広く理解し的確なアプローチを行える人材となることを目的とする。具体的には研究室のミーティングを始めとして情報学セミナーや論文紹介に参加し情報学的観点からの生物学的問題への取り組み方法を学ぶと共に現地の研究者と交流を深め、英語でのディスカッションやコミュニケーション能力を研鑽したい。

Plan:(続き)

研究計画

(背景)

CBIO に所属する山西芳裕博士のグループ等、複数の研究グループが既知の薬物−遺伝子相互作用情報から機械学習やカーネル関数を用いて新規の薬物−遺伝子相互作用を予測する手法を開発している。この相互作用情報は創薬においてリード化合物の探索や副作用予測、特定の遺伝子を標的とするプローブの開発等において重要であることが知られているが現状では化合物の網羅的な情報が収載されている PubChem 等の一般に解放されているデータベースにおいては実験的に既知の薬物−遺伝子相互作用情報は限られているため、これを既知の情報を用いて高精度で予測する手法は有用である。応募者はこれらの手法を応用して化合物・遺伝子・病気の三者の既知の相互作用から新規の化合物・遺伝子・病気の相互作用を予測する手法の開発を目標とする。病気と化合物・遺伝子の相互作用を新規探索することにより、疾患マーカー遺伝子の探索や特定の疾患における標的遺伝子の発見、また特定の疾患と関連する環境物質の解明等が期待される。

(手法)

上記下線部の三者の既知の相互作用情報については CTD(The Comparative Toxicogenomics Database)を利用する。このデータベースには 2003 年より複数の博士号を持つ研究員が論文から拾い集めた三者の相互作用情報が多数蓄積されており、テキストマイニング等の情報学的手法を用いて集められたデータよりも信頼性が高いため今回の研究課題に適していると考えられる。この情報を元に、カーネル関数を用いた予測手法を応用する。具体的には、

[1] 相互作用情報から、化合物間、遺伝子間、病気間の類似性について RBF カーネルを設計する。

[2] 化合物間の構造上の類似性、遺伝子間の配列上の類似性、病気間のオントロジー上の類似性からカーネルを設計し、[1]のカーネルと重みを付けて統合する。

[3] RLS 分類器を用いて最終的な相互作用スコアを算出し、cross-validation によって評価し、パラメータを再設定するなど適宜改良を重ねていく。

(参考文献)

Yamanishi, Y. *et. al.*, Prediction of drug-target interaction networks from the integration of chemical and genomic spaces, *Bioinformatics*, 24:i232-40, 2008

Bleakley, K. and Yamanishi, Y., Supervised prediction of drug-target interactions using bipartite local models, *Bioinformatics*, 25:2397-403, 2009

Laarhoven, T. V., *et. al.*, Gaussian interaction profile kernels for predicting drug-target interaction, *Bioinformatics*, 26:3036-43, 2011

Davis, A. P., et. al., The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res.*, 39:D1067-72, 2011