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 報告書

| |
|---|
| 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: Centre for Computational Biology, Mines ParisTech |
| Duration: 2012 年 1 月 10 日 - 2012 年 3 月 30 日 |
| Report:<br><u>研究生活</u><br>ITP の支援によって 2012 年 1 月より 3 ヶ月間 Mines ParisTech の CBIO(Centre for Computational Biology)グループに滞在し、Jean-Philippe Vert 博士及び山西芳弘博士の元で充実した研究生活を送ることが出来た。研究室は学生街となっているパリ 5 区に所在しておりアパートを 5 区に借りて生活の拠点にしていたが、周囲は大学や学校、書店などが多く学生で溢れており、周辺の食料品店や飲食店の人々もフランス語をほとんど話すことのできない私に優しく接してくれ大変過ごしやすい環境であった。また冬にもかかわらず今年は幸運にも気候が穏やかであり、1 月末から 2 週間ほど極寒であっことを除いては日本よりも温暖で過ごしやすい気温であったが、フランスで頻発すると言われるアクシデントにも見舞われた。借りていたアパートが水漏れ事故の被害にあい、上階からの水により部屋が損害を受けた為、保険関係のフランス語の書類を何種類か準備し損害の査定等の手続きをする必要に迫られたが、研究室のメンバーに助けられて帰国までに各種手続きを終わらせることが出来た。海外生活ならではの事件でありこれも貴重な体験であると肯定的に捉えている。<br>CBIO のメンバーは私が滞在を始めた頃には PhD の学生が 6 名とポスドクが 1 名､指導教員は Dr. JP Vert､山西博士を含み 3 名であった。CBIO は Emmanuel Barillot 博士率いる Inserm U900 というユニットに所属しており全体として数十名のメンバーが主にバイオインフォマティクス分野の研究を行なっている。Inserm U900 は Mines ParisTech と Institut Curie の合同ユニットであり研究室は Institut Curie の敷地内にある。火曜の午後に行われるセミナーはこのユニットのメンバーによって開催されており週替わりの担当者が各自の研究発表を行う場となっている。またこの時間に外部の研究者を招いた公演も行われていた。<br>CBIO では統計的機械学習の理論的手法の開発や病気に関係するデータに対する適用を行なっている。併設の病院から得られる遺伝子発現等の臨床データを用いた研究や、実験化学者と共同で特定のタンパク質に結合する化合物について立体構造を用いたドッキングシミュレーションによって予測を行うなど研究対象は様々である。CBIO では火曜の午前にミーティングを行なっており毎週各自の進捗状況を報告するがメンバーは非常に活発に外部の研究者と議論や会議の場を持っており誰と会ってこういう内容の話をしたという報告が多く、外部の研究者との情報交換が自身の研究を進める上で役に立つのみならず、後の研究の広がりや人脈、キャリアパスを考える上でも重要であることを再認識した。このミーティングでは毎週一人が自分の研究発表をスライドを使って行うが、発表中にメンバーが質問やコメントを次々と発表者に投げかけ、また聞き手同士でも解釈の助けになるような議論を行なっていた。その為か短い時間の中で深い理解が全体で共有されており、メンバーの科学コミュニケーション能力や研究理解力が洗練されているという印象を受け、私にとって良い刺激であった。<br>また滞在期間中には、研究室のメンバーから機械学習の理論や応用の手法について多くのことを学 |

Report (Continued) :

ぶことが出来た。私が所属している研究室は KEGG データベースを中心とした実データから生物学的、生化学的な知見を得ることに重点が置かれているため、普段体験することのできない貴重な経験となった。特に機械学習を用いた研究を実際に進めていく上での問題や困難に対する解決策を、研究室のメンバーが行なっている研究についての議論や、自分の研究の問題点をメンバーに聞いてもらうことを通して学ぶことが出来た。一方で、研究に関する議論をその場で組み立てていくために必要な英語能力が現状では不十分であることも痛感した。滞在中にある程度の上達を実感できたのでこれを機に語彙や表現力を継続して鍛え直していきたい。

ミーティング後に研究所付近のレストランにて (左手前から時計回りに)Dr. Andrea Cavagnino, 西村、Dr. Thomas Walter、Pierre Chiche、Toby Dylan Hocking、Dr. Véronique Stoven、Matahi Moarii、Dr. Jean-Philippe Vert

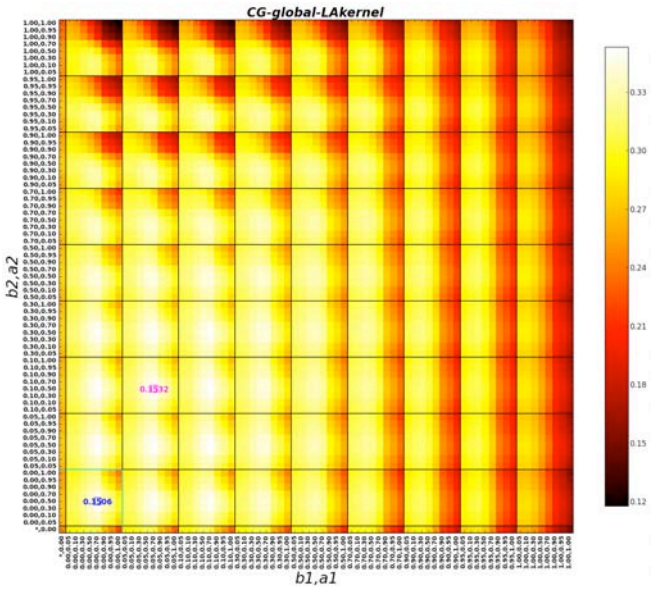
**研究成果**

山西博士のグループは、既知の化合物—遺伝子相互作用情報から機械学習を用いて新規の化合物—遺伝子相互作用を予測する手法を開発している[1][2]。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 等のパブリックなデータベースにおいて実験的に既知な化合物—遺伝子相互作用情報は限られているため、既知の情報を用いて実験的に検証されていない未知の相互作用を予測する手法は有用であると考えられる。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を開発した。病気と遺伝子の相互作用を新規探索する事により、特定の疾患におけるマーカー遺伝子の探索や標的遺伝子の発見が期待でき、病気と化合物の相互作用から特定の疾患と関連する環境物質の解明などが期待される。

これら三者の既知の相互作用情報については CTD(The Comparative Toxicogenomics Database)[3]を利用してデータセットを作成した。このデータベースには複数の博士研究員が論文から収集し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で自動収集された情報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。このデータセットを用い、Laarhoven *et. al.*[4]の手法を改良して三者の相互作用予測を行った。具体的には化合物—遺伝子、遺伝子—病気、病気—化合物の既知の相互作用情報を用いて GIP カーネルを作成し、これらのカーネルを線形結合させてから RLS 分類器を用いてこれら三者の相互作用を予測した。線形結合に関して 4 つのパラメータを設定し、最適になる組み合わせを計算したが、その際にデータセットのサイズが大きいことと 4 つのパラメータについて同時に調整する必要があることから試す必要のあるパラメータの組の数が 5,329 に達し、大型計算機を用いても予測スコアの計算において膨大な時間がかかることが分かったため、matlab、ruby、python において実装を行い、ベンチマークテストにおいて最速であった python を用

Report (Continued) :

いてパラメータ調整を行った。この作業過程においてこれまでに私が解析に用いたことがなかった matlab や python のコーディング技術を習得し、解析技術の幅を広げることが出来た。




[図 1]化合物—遺伝子相互作用予測における
パラメータ調整

手法の予測精度評価に関しては、クロスバリデーションを行い、precision-recall 曲線の下線部面積(AUPR)を用いた。三者の相互作用予測のうち、化合物—遺伝子相互作用予測においては先行研究の予測精度と比較することが可能であるが、我々の手法においては化合物—遺伝子の相互作用情報だけでなく、遺伝子—病気、病気—化合物の相互作用情報を加えて化合物—遺伝子の相互作用予測を行なっている。結果として予測結果が先行研究のものを少しではあるが上回っていたことに加えて、我々の手法によって遺伝子—病気、病気—化合物の相互作用を化合物—遺伝子の相互作用予測と同様に予測することが可能となった。さらに現在では GIP カーネルの代わりに他のカーネルを用いることによる予測精度の改良にも取り組んでおり、今回の結果について学会・論文での発表に向けて連携を取りながら解析を継続中である。

## 参考文献

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

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

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

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

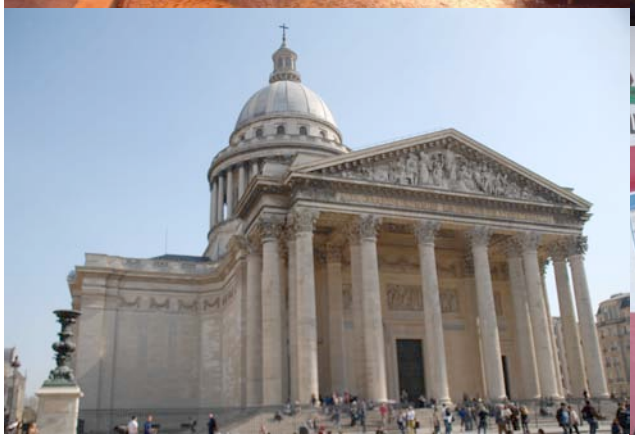
山西芳弘博士(左)と Toby Dylan Hocking(右)

Dr. Andrea Cavagnino(左)と Matahi Moarii(右)

Report (Continued) :

**謝辞**

本国際交流プログラムの事業実施専攻長で貴重な機会を与えて下さったのみならず、滞在の準備や計画、研究のアドバイス等全面的に協力して下さった馬見塚拓教授、滞在の計画をサポートして下さった五斗進准教授、また滞在を快く受け入れて下さった Jean-Philippe Vert 博士、山西芳弘博士、滞在期間中に事務手続きでお世話になった Kopf Katharina さん、水漏れ事故の際に解決に向けての手続きの大部分を助けて下さった Véronique Stoven 博士、Matahi Moarii さん、現地で非常に面倒を見て下さった Andrea Cavagnino 博士、CBIO のメンバーの方々に感謝致します。海外での研究生活においては様々な困難を予想していましたが、万事順調の運びとなり、皆様の御陰で非常に充実した研究生活を送ることができました。この場を借りて御礼申し上げます。

(左上) パリ 5 区の食料品、飲食店通りである Rue Mouffetard。週末は観光客で賑わう。様々な食材が手に入るため、毎日のようにお世話になった。

(右上) Institut Curie の研究室のある建物の屋上からの眺望。遠くにパンテオンが見える。時折、この屋上から景色を見ながら研究室のメンバーでランチを食べることも。

(左下) パンテオン。内部には Curie 夫妻や Joseph-Louis Lagrange、Jean-Jacques Rousseau などフランスを代表する偉人たちの墓や、Foucault の振り子がある。

(右下) Toby Dylan Hocking 邸でのホームパーティー。パリは賃料が高い為ルームシェアリングも多い。Toby(中央)は 4 人でシェアしていた。Anne-Claire Haury(左)、ルームメイト(右)と共に。