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：鎌田　真由美
Title：タンパク質の立体構造揺らぎの時系列解析と機械学習による化合物分類に関する研究
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: Macromolecular Modeling Group (Knapp Laboratory),　Freie Universität Berlin
Duration:　2010 年 9 月 27 日～12 月 20 日
Report: 9 月の末より約 3 カ月間、ドイツのベルリン自由大学の Macromolecular Modeling Group，Knapp 研究室に滞在させて頂いた。Free University of Berlin は、ベルリンの中心地から少し離れた緑に囲まれた住宅街あり、Knapp 研究室は無機化学研究所近くのコンテナーにある。学生部屋のひとつに私の机も用意してくださり、毎日複数の学生と共に研究をさせて頂いた。ベルリンに到着した初日には研究室のメンバーの方が空港まで迎えに来てくださり、翌日は大学の案内やメンバーの紹介をしてくださった。Knapp 研究室の基本的なメンバーは PhD の学生が 7 人とポスドクが 3 人で、ドイツだけでなくイタリアやメキシコ・アルメニアなど世界各国から、物理や化学・情報など様々な学位を持ったメンバーが集まっている。Knapp 研究室では、週に 1 度のペースで各々の研究内容やその成果を発表するセミナーがあり、時に他の研究室と合同で行われることもあった。セミナーでは発表に対し多くの質問や意見が飛び交い、積極的な議論が行われていた。また、既述のようにメンバーは各々異なるバックグラウンドを持つ為、その研究テーマは分子動力学シミュレーションなどの物理化学的なものから計算機の並列計算の様な情報学に関するものなど多岐にわたっており、セミナーを通して多様な研究発表を聞くことが出来た。滞在中、私も自身の研究について発表する機会をいただき、大変有意義なコメントや質問を頂くことが出来た。  (上)Free University of Berlin 構内と最寄駅 (下)Knapp 研究室 研究室では教授・ポスドク・学生が同じ目線で活発に意見を交換し議論しているのがとても印象的で、各学生の学生という立場に甘んじることの無い研究者としての意識の高さを感じた。12 月には研究室でクリスマスパーティーも開催され、Knapp 教授が装飾してくださったテーブルで、皆が持ち寄ったドリンクやスナックをつまみながら会話を楽しんだ。この様なとても温かく、また自身を奮起させてもらえる雰囲気の中で有意義な時間を過ごし、今後の自身の研究活動においてとても有用で貴重な経験をさせて頂くことが出来た。 Knapp 研究室では機械学習による Drug Classification 手法の開発も行っている。今回の滞在では、

X-mas party (上)と研究風景(下)

機械学習の研究を行っている Özgür Demir に機械学習について基本的なことから教えて頂きながら、回帰問題に関する研究を行った。このテーマは、現在の私の研究内容との関連性は低いが、今後の研究発展に機械学習を用いたいと以前から考えていた為、今回の滞在中このように基本的な事から学ぶまとまった時間を持てたことはとても嬉しかった。

研究の内容としては、Özgür の行っていた線形回帰モデルを用いたクラスタ分類[1]の回帰問題に着目し、その学習における正則化部分の改良と、正則化における特徴選択について様々な検証を行った。正則化には、主に正則項として重みの絶対値の和を用いる L1 正則化と 2 乗和を用いる L2 正則化がある。L1 正則化では多くの不要な重みはゼロとなりスパースな重みベクトルを得ることが出来る。また、L2 と L1 を比較した実験では、L1 を用いた場合 L2 のほうがやや高いものの、ほぼ同程度の精度が L2 の 1/10 から 1/100 の特徴数で得られることが報告されている[2]。そこでまず、L2 正則化を用いていた既存モデルに対して L1 正則化を適用した。そして予測精度を検証し、僅かな特徴数で L2 と同程度、もしくはそれよりも高い精度が得られることを確認した。

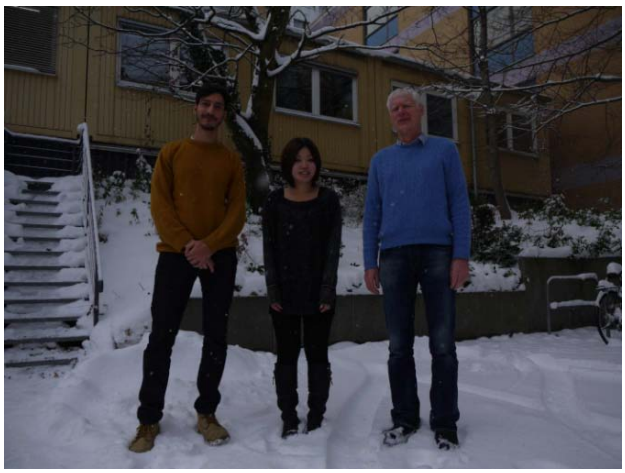
次に L1 と L2 を段階的に用いることを試みた。これは、まず L1 正則化を第 1 段階の学習に用い特徴数の制約、つまり特徴選択を行い、残った特徴のみで再度 L2 正則化を用いた学習を行い、最終的な予測を行うというものである。対象データセットには CoEPrA2006 と呼ばれる分類と回帰に関する大会で用意された 4 つのタスクを用い、各タスクの 1 位、2 位、3 位のチームの結果(予測精度)との比較を行った。第 1 段階の学習において、特徴数はすべてのタスクで元の特徴数の約 1%以下に制約された。そして 2 段階学習後の予測値は全タスクにおいて 3 位以内に入る高い精度値を得ることが出来た。また段階ごとの予測精度の結果を検証したところ、ほとんどのタスクにおいて 2 段階目の適用によって精度の向上が見られた。

さらに特徴選択の過程や特徴数と予測精度の関係性についても確認を行った。選ばれる特徴のセットは、必ずしも段階的に選択されていくのではなく、正則項のパラメータ値ごとに異なる特徴が選択されている。そして、特徴数と予測結果は比例しておらず、特徴数を多く取った場合よりも少ない特徴数の方が良い結果を得る場合がある、という興味深い結果も得られた。今後、今回得られた上記の結果をまとめると共に、さらに学んだことを自身の研究にも活かしていきたい。

また、上記のプロジェクトとは別に、自身の研究であるタンパク質の動的挙動の時系列解析に対して、長時間の MD データの適用を行いたいと考えていたところ、分子動力学シミュレーションを行っている Knapp 研究室の方と今後共に研究を行う機会を頂くことが出来た。今後もコンタクトを取りながらこちらの研究についても進めていきたい。

参考文献：
[1] O Demir-Kavuk et al. Bioinformatics. 2010 Mar 126(5):603-9.
Exploring classification strategies with the CoEPrA 2006 contest.
[2] J Gao et al. ACL 2007
A comparative study of parameter estimation methods for statistical natural language processing.

謝辞：
最後に、本国際交流プログラムの代表者であり今回この様に貴重な機会を与えて下さった金久實教授、滞在の計画からサポートしてくださった馬見塚拓教授と阿久津達也教授、そして滞在先において温かな配慮で面倒を見てくださったWalter-Knapp教授、Özgür Demirさん、ならびに同研究室のメンバーに心から感謝申し上げます。



Knapp教授(右)とÖzgürさん(左)とラボの前で。

ベルリンの街。12月には至る所でクリスマスマーケットが開催されていた(下)。