

Name : 志賀元紀
Title : 疾患遺伝子同定のためのデータマイニングの開発
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: Department of Bioengineering, Boston University
Duration: 6月15日～9月2日
Report: ボストンでの滞在中、ボストン大学の Department of Bioengineering および Bioinformatics Program に所属する Simon Kasif 教授の研究室 Computational Genomics Laboratory にお世話になった。 Kasif 教授の研究室は、ボストン大学構内のライフサイエンスビルディング（右の写真）にあり、その基本的なメンバーは、ポスドク3人と学生2人である。研究室のセミナーは、週1回の頻度で定期的に行われ、各週一人ずつ、研究室メンバーが持ち回りで最近の研究成果を発表する形式であった。私は、この研究室セミナーに参加させていただき、メンバーの研究内容をディスカッションしながら交流を図った。セミナーで発表された内容は、PNAS のようにインパクトファクターが高い雑誌に採択された目覚ましい研究成果のみならず、これから立ち上げるプロジェクトに関するものまで様々な話題であった。発表に対する聴講メンバーの態度は、内容に関わらず実に積極的であり、手厳しい質問や反論のみではなく、基本的な質問などもあり、おとなしい研究者が多い日本とは異なる雰囲気があり、一方、質問者に対する発表者の応答も真摯であった。海外のセミナーに参加するたびに毎回感じる事であるが、こうしたディスカッションを有効に利用する姿勢の大事さを改めて確認できた。 その他に、私の研究のために、基本的に週1回の頻度で、Kasif 教授と研究のディスカッションの機会を設けて頂き、さらに、必要な時に他のメンバーと逐次ディスカッションした。研究の内容は、疾患遺伝子群を同定するデータマイニング法の開発である。具体的な目的は、遺伝子発現量とタンパク質相互作用ネットワークを用いて、正常細胞の遺伝子発現量と疾患細胞の発現量が著しく異なる遺伝子群を抽出することである。研究に使用するデータセットには、論文として出版されている誰もが利用できるものを紹介して頂いただけではなく、Kasif 教授が共同研究しているウェット研究室が取得した質の高い最新のデータも提供していただいた。また、研究のディスカッションにおいて、Kasif 教授の研究室で疾患関連遺伝子の同定法 Gene Set Enrichment Analysis(GSEA)の研究を手がけた



Report (Continued) :

Manway Liu に参加してもらった。Liu らの提案した手法は、まず、正常細胞と疾患細胞の遺伝子発現量差をスコアリングして、スコアが平均的に高いタンパク質相互作用の部分ネットワークを抽出し、そして、部分ネットワークと関連する遺伝子集合 (Gene Ontology term など) を同定する手法である。したがって、私の研究目的が重なっている部分が多く、また、Liu は実際のデータを解析した経験があるので、私の研究をする際に手助けになった。実際、帰国後も、データの詳細に関して気軽に質問できるために非常に助かり、今後も質問して助けられる機会が多いのではないかと考えている。また、彼の方が生物データの解析に詳しいのに対して、私の方データマイニングの研究に詳しいので、今後も発展した共同研究を期待している。

次に、滞在期間中に行った研究内容を具体的に述べる。最初に行った事は、疾患遺伝子群の同定のために、統計的モデルによるデータセット (遺伝子発現量とタンパク質相互作用) のモデリングである。統計モデルとして、最初に、我々が ISMB/ECCB 2007 で発表したモデル Hidden Modular Random Fields の使用を考慮したが、確率分布として定義するための正規化定数の計算が困難であるという結論に至り、新しいモデルの定義とその学習アルゴリズムの導出を行った。新しいモデルは、Hofman (Phys. Rev. Lett. 2008) が提案しているネットワークの統計的モデルと遺伝子発現量の統計的モデルを組み合わせたモデルである。学習アルゴリズムとして、クラスタ数を最適化できる変分ベイズ法によるアルゴリズムを導出した。

人工データを生成して学習法の性能をチェックしたところ、我々の手法が上手く機能していることを確認できた。しかしながら、実際のデータに試したところ、計算コストが無視できないほど大きいことが明らかになった。使用しているデータセットの遺伝子数は約 2 万であり、重要と思われる遺伝子だけを予め選択したとしても、約 1 万の遺伝子が解析対象という規模の大きなデータである。さらに、最適化されたクラスタ数が非常に大きいことが計算コストを増大させる。現在の実験では、最適なクラスタ数は明確になっていないが、最適化するコストを曲線にして描いたところ、クラスタ数が 20 以上であるということがわかっている。学習のプロセスは、クラスタ数ごとに独立しているため、現在の手法ではかなり計算に時間を要することがわかった。現在では、計算コストの問題を解消するために、ノンパラメトリックベイズ法などの異なる学習法など、他の手法の適用を考えている

学習法とは別に、疾患関連遺伝子の抽出のためのスコアリング方法、2 つを考案した。一つ目は、正常細胞のデータセットから学習した統計モデルと疾患細胞のデータセットから学習した統計モデルの間の対数尤度比によりスコアリングする手法である。2 つ目は、学習された統計モデルを用いて、発現量の差の期待値によるスコアリングである。これらのスコアリングの優劣を調べるために、人工データセットを使用した。人工データで仮定した正常細胞のデータセットと疾患細胞のデータセットは、ほとんどが同じ統計的モデルを仮定しているが、一部分の遺伝子集合は、疾患細胞において正常細胞のモデルよりも過剰発現するような統計的モデルを考えた。そして、過剰発現する遺伝子群を抽出できれば、実行した手法が上手く機能したとみなす。上述した数値実験を用いて、スコアリングの比較したところ、2 つ目のスコアリング (発現量の差) の方が上手く働くようである。1 つ目のスコアリング (尤度比) は、統計的モデルの学習結果の微妙な違いに影響され易く、つまり、不安定な手法であると考えられる。

Plan (Continued)

以上で述べた研究を、帰国後においても、Kasif 教授らとコンタクトを取って継続し、実データの解析に耐えうる手法を開発してゆきたい。

ボストンの滞在期間中には、ボストン大学のバイオインフォマティクスプログラムにて、本国際交流プログラムに関わるワークショップ、The 9th Annual International Workshop on Bioinformatics and Systems Biology (IWBSB) が開催され、それに参加した。IWBSB では、参加のみではなく研究プレゼンテーションの機会を頂いた。その発表内容は、遺伝子発現量と遺伝子ネットワークを組み合わせる統計的モデルによるクラスタリングとクラスタリング結果を用いた遺伝子機能の予測である。発表スライドを情報学の研究者よりの視点に偏りすぎていたという失敗があったものの、発表に対してコメントを頂けたりして、交流のきっかけになった。また、ワークショップの期間中は、他の発表者とのディスカッションなどを通して、ボストンおよびドイツの参加者達と交流できた。ワークショップは、今後も毎年開催される予定なので、できれば参加をして研究仲間としての関係を継続してゆきたい。

最後に、本国際交流プログラムの代表者であり貴重な機会を与えて頂いた金久實教授、滞在の計画段階からお世話して頂いた馬見塚拓教授、また、滞在においてお世話になった Simon Kasif 教授（左下の写真）、Manway Liu さん（右下の写真）、ならび、同研究室のメンバーに心から御礼申し上げます。



Simon Kasif 教授

ディスカッション後に記念撮影をした。



Manway Liu さん

休日にボストンを案内して頂いた。写真は、ボストン湾の近くにあるクインシーマーケットである。