第 11 回 2014 年 1 月 10 日（金）14:50 ~ 15:40

Discovering Combinatorial Regulations from Life Science Data
生命科学データからの組合せ効果発見

Koji Tsuda / 津田　宏治

Computational Biology Research Center (CBRC)
National Institute of Advanced Industrial Science and Technology (AIST)
産業技術総合研究所　生命情報工学研究センター

More than three transcription factors often work together to enable cells to respond to various signals. The detection of combinatorial regulation by multiple transcription factors, however, is not only computationally nontrivial but also extremely unlikely because of multiple testing correction. The exponential growth in the number of tests forces us to set a strict limit on the maximum arity. Here, we propose an efficient branch-and-bound algorithm called the "limitless arity multiple-testing procedure" (LAMP) to count the exact number of testable combinations and calibrate the Bonferroni factor to the smallest possible value. LAMP lists significant combinations without any limit, whereas the family-wise error rate is rigorously controlled under the threshold. In the human breast cancer transcriptome, LAMP discovered statistically significant combinations of as many as eight binding motifs. This method may contribute to uncover pathways regulated in a coordinated fashion and find hidden associations in heterogeneous data.

Keywords: Multiple testing, Itemset mining, False discovery, Machine learning, Data mining

パーソナルゲノム時代に入り、生命科学で扱うデータは増加の一途であり、解析に困難をきたすことも難しくない。各個人のゲノム、エピゲノム、遺伝子発現、タンパク質発現、代謝物プロファイルなどが比較的安価で得られるようになり、Multi-omics 解析も可能になってきている。このような大量データ解析において、よくある誤解は、解析に必要な計算量はデータの量に比例するというものである。単一の種類のデータを解析する場合には、それで概ね正しいのであるが、生命科学においては、異なる種類のデータ間の関連を明らかにする統合解析が主なタスクであるので、状況はもっと悪い。例えば、100 万 SNP、１万発現量、１万 CNV の関連を明らかにしようとすると、100 兆回の評価値計算が必要になる。データの増加そのものが問題なのではなく、多様なデータが引き起こす組合せ爆発こそが最も深刻な問題なのである。組合せ効果を発見できるデータマイニング手法を生物学データに応用する試みは、情報分野では広く行われているにも関わらず、生命科学の論文で広く採用されるには至っていない。その理由は、マイニングで得られた結果に関して、統計的有意性が証明できないことにある。殆どのジャーナルでは、主要な結果に関しては統計的に有意であることを求めており、特に、検証実験ができない疫学の分野では、その傾向が顕著である。データマイニングでは、非常に大きな数の仮説の中から、データに合う仮説を選び出すということを行うので、多重検定の問題をクリアするのが難しい。本発表では、この問題を解決するための手法 LAMP について述べる。

キーワード: 多重検定、アイテムセットマイニング、偽発見、機械学習、データマイニング