第47回
インシリコ・メガバンク研究会開催のお知らせ

2014 2.28[金]　開場 16:45　開演 17:00〜18:30

[会　場]　東北メディカル・メガバンク機構2階会議室1

[演　題]　Discovering Combinatorial Gene Interactions in High-Dimensional Data

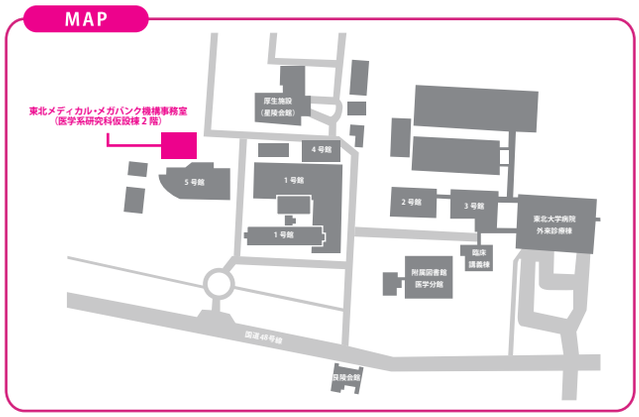
[講　師]　**David duVerle**
（産業技術総合研究所生命情報工学研究センター）

[問合せ]　東北大学 東北メディカル・メガバンク機構
ゲノム解析部門 長﨑研究室事務
TEL 022-273-6051/ E-mail nagalab-jimu@csml.org

In the past decade or so, new technologies in biotech have meant an explosion in the availability of high-dimensional genomic data (microarrays, SNP data, RNA-Seq...) : their dimension and noise levels making it necessary to rely on machine learning techniques and statistical models to extract meaningful signal and narrow down the field for further experimental research. In this presentation, I will try to give a very general overview of some of the statistical techniques commonly used to treat high-dimensional data, as well as a more detailed illustration of the technique we developed to identify combinatorial interaction effects in such data. A crucial aspect of machine-learning when dealing with high-dimensional data, is the concept of sparsity: how much of the input' s variables find their way in the model. By using regularisation techniques (the addition of a tailored penalisation component), it is possible to ensure certain properties of the statistical model (size, elimination of collinear variables …). Another, is the fitting of complex statistical models that cannot be solved analytically, usually requiring optimising a non-linear objective function (e.g to maximise likelihood or minimise empirical error). While relatively simple in application, both techniques require some understanding of the underlying statistical assumption and information theoretic implications, in order to obtain satisfying results. After giving a brief overview of regularisation techniques and their use in typical regression problems encountered in bioinformatics, I will introduce our recent work, which combines them with data-mining (itemset mining) and fractional programming techniques to fit complex statistical models over (non-linear) combinations of heterogeneous input variables, allowing for example to identify sets of genes (up- or down-regulated) that drive complex phenotypes or clinical observations This work was in particular successfully applied to a combination of cDNA microarray and gene mutation copy number data paired with (right-censored) survival data, to identify interactions (and potential synthetic lethals) playing a role in neuroblastoma and breast-cancer.

世話人： 長﨑 正朗

MAP



※本セミナーは医学系研究科系統講義コース科目の授業としての振替可能なセミナーです。