第 3 回　情報生命科学特別講義　2011 年 10 月 21 日（金）

Pair-wise sequence comparison: old and new solutions

Martin Frith / マーティン　フリス
National Institute of Advanced Science and Technology (AIST)
Computational Biology Research Center (CBRC)
産業技術総合研究所　生命情報工学研究センター

Pair-wise sequence comparison is arguably the most fundamental task in computational biology. This lecture aims to explain the main issues. Although many of these issues were studied several decades ago, *all* of them involve problems that are poorly understood by most practitioners in 2011.

* The maximum-score approach to sequence comparison: advantages and disadvantages.
* X-drop algorithms: advantages and disadvantages.
* Choice of score parameters.
* The significance of a similarity: how frequently such a similarity would occur between random sequences.
* Avoiding false homology predictions due to low-complexity tracts (e.g. atatatatatatatat). The standard methods do not work!
* The (un)reliability of each column in an alignment.
* How to compare sequences with unusual composition, such as 80% A+T malaria genomes.
* Incorporating per-base error estimates for DNA sequences.
* Spaced seeds, subset seeds, and adaptive seeds.

After understanding these issues, we can reliably perform all kinds of sequence comparisons, including: large (multi-gigabase) datasets, long or short sequences, strong or weak similarities, etc.