CBRC NEWS LETTER
Computational Biology Research Center
CBRC
No.2
December 2002
http://www.cbrc.jp/

独立行政法人　産業技術総合研究所　生命情報科学研究センター

配列情報解析の新たな発展

数理モデル･知識表現チーム長
（副研究センター長）
浅井　潔

バイオインフォマティクスでは、様々な生命情報を計算機で扱える形に表現し、有効な解析手法（アルゴリズム）を用いる必要があります。塩基配列やアミノ酸配列などの生物配列情報の解析には、動的計画法、形式言語理論、確率モデルなどの表現・解析手法が応用されてきました。BLASTによる相同性検索や隠れマルコフモデル（HMM）によるモチーフ表現や遺伝子領域予測など、手法は既に確立していると思われる方もいるかもしれません。また研究のトレンドが、配列情報そのものからタンパク質の構造・機能、遺伝子集団の振る舞いなど、より高次の現象に移りつつあり、新たな配列情報の解析手法の必要性はないという人さえいます。

しかし、生物配列にはゲノム構造、転写、スプライシング、タンパク質の立体構造、進化の歴史など、様々な情報が凝縮されて詰め込まれていて、その意味が単純な動的計画法による配列の比較だけで解明できないことは明らかです。むしろ遺伝子やタンパク質の発現、構造、遺伝子の集団としての機能、個体差と疾病との関係など高次の情報と、配列情報を結びつけるための新しい情報表現と配列解析の考え方が求められています。

CBRC数理モデル・知識表現チームでは、確率モデル上のMarginalized Kernelという、配列情報の新しい表現モデルを提案し、様々な問題に応用を開始しています。配列情報だけからは確定的に予測できないが、生体内のメカニズムや構造では実体を持つエキソン・イントロン構造や、RNA・タンパク質の二次構造などは、HMMや確率文脈自由文法（SCFG）などの確率モデルでは「隠れ状態」として表現されています。我々の手法は、確率モデルの隠れ状態を配列比較の柔軟なテンプレートと見なし、配列の中に隠された未知の立体構造・機能の高次相互作用を確率的に取り込んで特徴付けするもので、プロファイル型HMM（Pfamなど）の一般化になっています。単純な確率モデルによる近似的な表現と高次Kernelによる精密な特徴付けを用いて、従来計算が困難だった遠距離相互作用をもつ配列の比較や分類に可能性が開けてきました。90年代に花開いた確率モデルの生物配列情報への応用には、世紀を越えても、まだまだ新しい可能性があると期待しています。

【参考文献】
(1) K. Tsuda, T. Kin and K. Asai: "Marginalized kernels for biological sequences," Bioinformatics, Vol.18, Suppl. 1, S268 - S275 (2002).
(2) T. Kin, K. Tsuda and K. Asai: "Marginalized Kernels for RNA Sequence Data Analysis," GIW2002, to appear.