膨大な配列データとの闘いを振り返って

大野 朋重　塩野義製薬（株）フロンティア医薬研究所

〔受賞論文〕
A Method for Isoform Prediction from RNA-Seq Data by Iterative Mapping
大野朋重，瀬尾茂人，竹中要一，松田秀雄（大阪大学大学院情報科学研究科）
IPSJ Transactions on Bioinformatics, Vol.5, pp.27-33（2012）

このたび，標記の論文で本会論文賞をいただくことになった．多数の論文の中から選出されたということで非常に光栄である．本研究はひとえに各方面からのサポートの賜物であり，研究会や査読，議論等で有用なコメントをくださった各位に心より御礼申し上げる．

アイソフォームとは選択的スプライシングにより同一遺伝子から転写される，互いに異なるRNAのことであり，多様な生命現象や疾患との関連からその全容解明が求められている．本研究は，次世代シーケンサ（NGS）でRNA配列を解読する技術であるRNA-Seqを用いて得られたデータから，約2万個の遺伝子が生成するアイソフォームを網羅的かつ効率的に推定する手法開発を目的としたものである．RNA-Seqが用いられるようになったのは2009年頃であるが，筆者が本研究を始めたのもそれから間もない時期である．技術の黎明期には避けて通れないことではあるが，解析の流れや用語すらもコンセンサスが取れておらず，文献調査や先行事例の理解に苦労した．NGS研究を始めてまず驚いたのはそのデータの膨大さである．NGSは従来のシーケンサとは文字通り桁違いに大量の短い塩基配列断片（リード）を出力する．そこで数十GBにも上るテキストデータから有用な情報を抽出する解析アルゴリズムが必須となる．

RNA-Seqデータ解析では，まずリードと参照ゲノム間で配列比較を行う．その際にゲノムが1対1対応しないリードが大量に出てくる（**図-1**参照）．原因の1つはシーケンスミスや微生物などの混合によりリードが参照ゲノムにない配列を持つこと，もう1つはリードの短さやゲノム配列の相同性によりリードの由来個所を1カ所に特定できないことである．前者はノイズであり破棄されるが，問題は後者であり，これをどう捉えるかについて研究開始当時にはさまざまな立場が混在していた．これもノイズとして破棄することもあれば，ランダムもしくは均等に割り振るものもあった．筆者はその情報量を活かすことが重要と考え，反復マッピングというアプローチを考案しアイソフォーム推定精度向上を達成した．

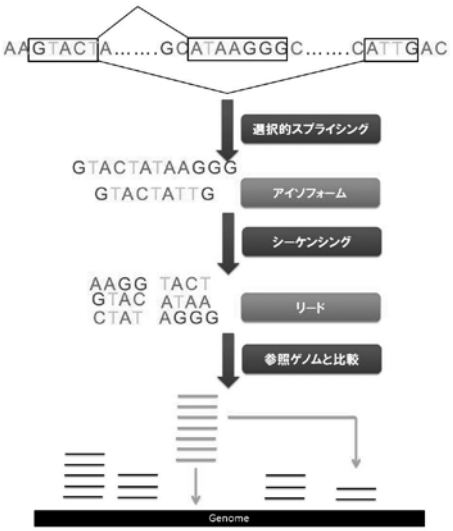

図-1　選択的スプライシングとシーケンシング

筆者は現在創薬研究に携わっているが，医療・製薬業界においてもNGS導入をはじめ，イメージングの普及や臨床データの増大など，情報科学の重要性が一層高まりつつあると感じる．今まで学んだ知識や研究の方法論を基に，薬という面から社会に貢献できるようさらなる精進に励む所存である．

（2014年5月15日受付）

大野 朋重（正会員）　tomoshige.ohno@shionogi.co.jp
2009年大阪大学基礎工学部卒業後，同大情報科学研究科に進学．2014年3月博士（情報科学）取得．同年4月より塩野義製薬（株）フロンティア医薬研究所にてバイオインフォマティシャンとして創薬研究に従事．