創薬現場に向けたデータサイエンス
～データサイエンスを活かすには～

加藤　竜司*・蟹江　慧・清水　志保

データサイエンスがもたらすもの

明確な定義はないが，データサイエンスには『専門的な知識』『解析の実装能力』『データの分析能力』の三要素が多くあげられる場合が多い（筆者が意訳したベン図を図1に示す）．

『専門的な知識』とは，「domain knowledge」などと表現されるが，分析・解析の中心を担うデータを理解するための知識であり，解析目標を設定する専門性である．我々にとって，それは生物学・生化学・分子生物学などの基礎知識である．『解析の実装能力』とは，「エンジニアリング力」「コンピュータ技術」「hacking skills」などと表現されるが，簡単に言えば『プログラムを書けるか』ということである．『データの分析能力』とは，「数理・統計力」と表現されるが，簡単に言えば『客観的にデータを見ることができますか』ということである．データサイエンスは，この三つスキルを総合的に活用して実行される．

しかし実は，各要素をただ高めても，データサイエンスが成果を生むことはない．たとえば，生物学者とプログラマーと統計家をいきなり雇っても，チームは機能しない．人工知能を高額で導入しても，受託解析に分析を依頼しても，現場を改善できる可能性は低い．なぜなら，データサイエンスは1回やって終わりではなく，「活用して」こそ意味があるからである．

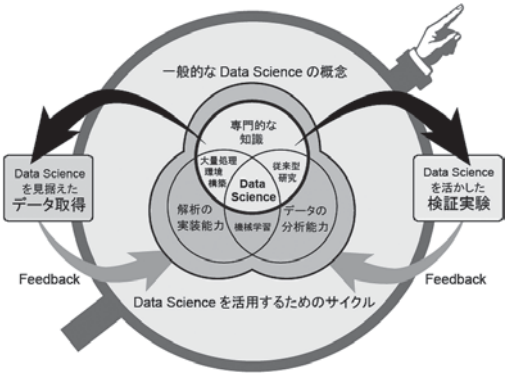

図1．データサイエンスの構成要素と活用サイクル

筆者は，データサイエンスの凄みとは，実は上記三要素の外側（図1の横に広がる矢印）にこそあると考えている．それが「データサイエンスを意識してサイクルを回すこと」である．

「データサイエンスを意識したデータ取得」とは，「大量の解析ができるようにデータをとる」「大量解析にふさわしいデータをとる」ことである．とりあえず実験を行ってから，「ではデータサイエンスを使おうか」ではないのだ．

「データサイエンスを活かして検証を進める」とは，既存概念や通説よりも解析結果を信じて，実験で実証を繰り返すパワーと勇気である．データサイエンスはそのための方位磁針のようなものである（図1中の指矢印）．

本稿では，このヒントとなる事例を二つ紹介したい．

データサイエンスを見据えたデータ取得とは

近年，創薬研究ではヒト細胞に対するニーズが急増している．幹細胞研究の進歩は，目的の病気を研究するための細胞を自在に扱える時代をもたらしたのである．結果，cell-based assayについて，創薬の基礎研究からスクリーニングに至るまでの研究と開発が過熱している．

創薬におけるcell-based assayの大前提は，大量データ取得とその解析である．最新解析技術がもたらす網羅性が，大量データから新しい発見を可能にしている．しかし現実には，cell-based assayのデータには，大量であるが故のノイズという問題が存在している．

筆者らはこれまで，cell-based assay用の評価技術として，培養中の細胞画像の解析技術を研究開発してきた[1,2]．しかし，せっかくデータを集めて解析しても，なかなか共通の傾向を見つけ出せない，という問題があり，その一つの原因に「作業者による細胞の撒きムラ」，という根本的な懸念事項があった．

顕微鏡写真は，容器全体からするときわめて一部分の情報である．細胞の播種状態にムラがあると，各画像は極端に様相が異なってしまう．しかし，細胞播種について定量的な操作指針を記した文献は存在していなかった．

そこで筆者らは，大量の画像解析に見合ったデータを作るために「なんとなくできている」と思っていたこの

*著者紹介　名古屋大学大学院創薬科学研究科基盤創薬学専攻（准教授）　E-mail: kato-r@ps.nagoya-u.ac.jp

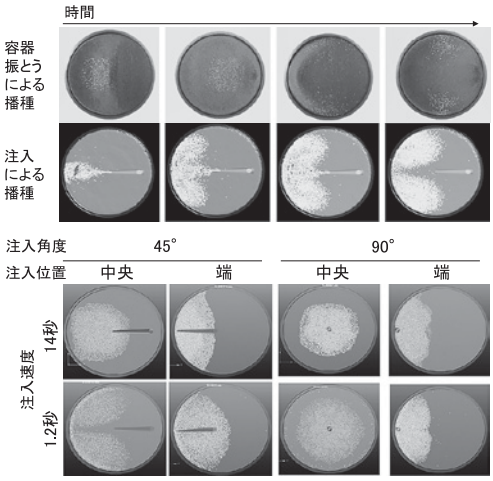

図2．流体シミュレーション（Particleworks）を用いた細胞播種条件の最適化

作業を見直そうと考えた．つまり，データサイエンスを意識して，「基本中の基本」を疑うことにしたのである．

筆者らは，流体シミュレーションソフトウェア（Particleworks・構造計画研究所）を導入し，容器の振とう条件や，ピペットの形状・角度・流速などをシミュレーションから網羅的に検証した（図2）．

結果，「通説どおりの播種ではうまくいかないこと」がわかり，同時に「確実に均質な播種に必用なパラメータ（容器の形，振幅，振とう回数など）」を得ることができた．また，創薬で多用される「ゆすっても攪拌できない小さなウェルプレート」においても，ムラなく細胞播種が行えるパラメータをも得ることができ，その後の細胞画像解析は飛躍的に安定化した．

これは，高度なデータサイエンスを積み重ねようとするのであれば，その土台となる「データそのもの」を見直す方が効果的かつ重要だという一つの例である．

データサイエンスを活かした検証実験とは

次に，データサイエンスを活用した事例として，生物を利用した医薬品安定生産のリスク対応への活用を紹介する．

近年，バイオ医薬品（抗体医薬品なども含む）は，その効果および副作用の少なさから世界的な売り上げ品目のトップを独占するまでに成長し，製薬企業における成長戦略の重要な柱となっている．製造の観点からも，バイオ医薬品の実現には高度に練り上げられた生産技術とノウハウを必要とし，高い品質の医薬品を患者さんに提供し続ける重要な意味合いを持つ．微生物による低分子医薬品の発酵生産は，正確にはバイオ医薬品の範疇ではないが，生物を用いた医薬品生産の複雑性のコントロールの観点から，製造データを活用するデータサイエンスの参考事例として紹介する．

免疫抑制剤であるプログラフ製品の原薬タクロリムス（FK506）は，1984年に新規に分離された放線菌 *Streptomyces tsukubensis* No. 9993の二次代謝産物で，その優れた免疫抑制機能から臓器移植時の拒絶反応の抑制に加え，自己免疫性疾患に起因する疾病の治療薬としても承認を得ている．2008年の特許満了（米国）以降も，全世界の患者様からのニーズは高く，発売以降約25年間安定的な商用製造を継続し，変わらぬ品質を提供している．

タクロリムス発酵培地には，天然物由来の原料が用いられる．天然物原料は安価で二次代謝物生産に効果的であるが，天候などによって品質が変動し，長期にわたる商用製造の生産性の変動の主たる原因であることが経験的にわかっていた．このように経験的な知見を，蓄積したデータ分析から再検証し，より深いプロセスの理解を進めることを目指した．具体的に筆者らは，長期間の商用製造で蓄積している多種類のインプロセスパラメータ［pH, DO（溶存酸素），排出ガス濃度など］のデータベースについてトレンド解析を行い，実際に原料の品質変動を培養で制御する方法へとフィードバックを行った．

トレンド解析とは，時系列的に得られた膨大なプロセス内のデータを，統計や多変量解析でさまざまな側面から分析するデータサイエンスである．筆者らはまず「手持ちのデータ」を分析し，過去のデータの整理と新しいデータの蓄積方法を検証するところに戻って実験を組みなおした．すなわち，データサイエンスから得られた「客観的な結果」を受けて，データサイエンスに見合うデータの準備に改めて約2年を費やしたのである．図1で言えば，手持ちのデータをデータサイエンスで検証したのち，改めてデータ取得からやりなおした（左のサイクルを回した）のである．

その後十分に吟味されたデータが蓄積された後，筆者らは2回目の解析（データサイエンス）として，先端的な機械学習やクラスタリング分析を行い，多面的に変動要因となるパラメータを探索した．

この解析でもっとも重要だったことは，どんなツールで解析したのかということよりも，「実験者と解析者が一丸となって何が原因となるかを考えた」ことであった．すなわち，実験者はこれまでの知見を一端捨て，冷静に分析結果から原因を考察した．また，分析者はデータをひたすら処理するのではなく，実験者の知見をもとに数値の羅列の中から「結果としての数値」と「制御結果が記

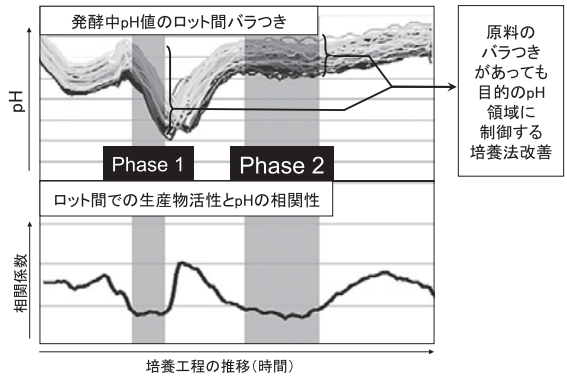

図3．培養プロセス中の計測値トレンド解析を活かした培養制御方法の改善

されただけの数値」を整理して分析に反映した．これはすなわち，図1の右のサイクルに向かって，データサインエンスの各要素を融合しようとした取組みと言える．

結果，得られた原因候補パラメータはきわめてシンプルであった．図3に示すように，培養中の「pHトレンド」にはタクロリムス生産性とある培養フェーズで強い相関がみられ，その影響は単因子の相関解析だけでなく，教師あり・教師なし学習のどの分析からも複合的な因子の一つであることが確認されたのである．

ここでまた本解析の大事なポイントは，複数の解析から得られる大量の解析データを並べながら，実験者と解析者が「発酵生産という専門知識」のもとで，注目すべき因子にさらに絞り込みをかけたことである．なぜならば，生産現場において「この因子が重要です」と解析で指摘されたとしても，実際の現場には「作業的に制御がしにくい」「規制・ルール的に制御できる幅が限定されている」などの現場の実情があるからである．

一般的に，解析者にとって，データはただの数字の塊に過ぎない場合が多い．すなわち，解析だけに集中してしまうと，このような現場の感覚･生産の実情という『専門的な知識』を通してデータを見ることができずにいることが多い．

本研究では，結果得られたpHの変動という因子に対して，これを制御しうる培養法を実験者が工夫を重ね，商用スケールの実証機で検証するまでに至った．結果，考案された新しい培養方法は，これまでの課題であった原料品質の劣化による生産性の変動を克服し，原薬の安定供給に継続的に貢献できるものであった．

すなわち，図1における右のサイクルが回ったことで，分析結果という「絵の中の餅」は，「本物の餅」へと生まれ変わることができたと言える．

まとめ

以上二つの実例の中には，「人工知能」や「オミクスデータ」のようなホットなキーワードは出てこない．また，高度なプログラムを書き上げた話や，何千・何万ものビッグデータを処理した話でもない．

しかし筆者らは，このような実例にこそデータサイエンスを活かす道があると感じている．事実，上述した例は現場を改善したデータサイエンスの一例であり，多くの解析でたどり着く「大量に解析した結果，ぼんやり現象がわかった」ような事例よりも，「実際に改善を生み出した」という点で，きわめて実用的な効果を生んでいる．

データサイエンスは最終目標ではなく，次に進むためのツールである．もしも従来の研究と違う部分があるとすれば，図1に示すように「大量処理を実現し」「機械学習などの新しいツール」を導入するによって，過去や古き知識にとらわれず，data-drivenに進もうという意識である．

まずは手持ちのデータから，まずは今制御できない課題から，一つひとつに対してできる限りの客観的なデータ解析を行い，その結果を「方位磁針として」進めばよいのである．

このような「データサイエンス」を軸にしたサイクルが回ることが，実はデータサイエンスの実用への近道であり，本稿がその一助となれば幸いである．

謝　辞

本研究において，流体シミュレーション解析の実施には構造計画研究所の山田剛史様，松岡毅様には多大なるご協力を頂いた．お二人の熱意と努力の結果，本来細胞分野では実績のなかった粒子法での流体シミュレーションは初めて実現したと言える．この場を借りて深く御礼申し上げたい．また，タクロリムス発酵生産の最適化研究では，アステラスファーマテック富山技術センターの岡崎宜恭様，竹下敏一様が膨大な実データの取得と整理，そして商用生産での実証検証にご尽力いただいた．またこのようなデータサイエンスを医薬品発酵生産の現場への導入を牽引し共同研究を主導してくださいました同センターの神田宗和様に，深く御礼申し上げたい．

文　献

1) 加藤竜司：生物工学，**96**, 121 (2018).
2) Kawai, S. *et al.*: *J. Biomol. Screen.*, **21**, 795 (2016).