プロジェクト・バイオ

バイオ画像自動分類ソフトウェアCARTAの開発

（東京大学大学院新領域創成科学研究科，エルピクセル株式会社）朽名　夏麿

近年，撮像技術や可視化手法の研究開発が盛んであり，時空間分解能の改善，多波長化，高次元化，自動撮像，そして撮像のスループット向上が進んでいる．その結果，研究現場で得られる画像データの次元，枚数，種類，サイズはいずれも増大を続けている．これは塩基配列や発現解析データと同様に「ビッグデータ」としてバイオ画像を取り扱う研究の可能性が拓けつつあることを意味する．こうした研究では多量に得られる画像のすべてを目で見ることは困難であり，大規模画像群からのデータマイニングには定量的で信頼性の高い画像解析が必須であろう．画像解析における計算機支援のニーズは高まっている．しかしながら，バイオ画像の特色とも言える多様性と多目的性ゆえ，画像の解析をサポートするソフトウェア環境の普及は遅れている．

機械学習による画像の自動分類

バイオ画像の自動分類は，膨大なデータの解析に伴う負担とコストを軽減し，バイアスやミスを減らす上で鍵となる要素技術である．当初，バイオ画像の自動分類システムの作成は，研究者自身が分類に用いる基準をソフトウェア上でも再現するというアプローチが採用されていた．このアプローチでは研究者の高次な知見をコンピュータ上の演算のレベルに還元するために分類規則を定式化していく必要がある．この開発には膨大な時間を要し，さらに，精度を高くすることは困難であった．それゆえ，生命科学分野での研究用途の利用は一部に留まっていた[1]．

この状況を打開しつつあるのがコンピュータによる「学習」，機械学習である（図1）．なかでも教師付き学習によるバイオ画像の自動解析は応用が進んでいる．教師付き学習とは，最初に研究者が分類対象となるデータの一部（教師画像）に対して分類結果をアノテーション（注釈）として付与し，これをコンピュータの学習における手本とする．そして教師画像とアノテーションという手本から，コンピュータは自動的に分類基準を探し出す（図1A）．教師付き学習により，アノテーションを付与した教師画像を用意すれば自動的に分類ソフトウェアを開発できることになる．分類システムの精度は教師画像の質や量に依存するため，高精度の分類システムを作るには多くの画像に正確なアノテーションを付与しなければならない．

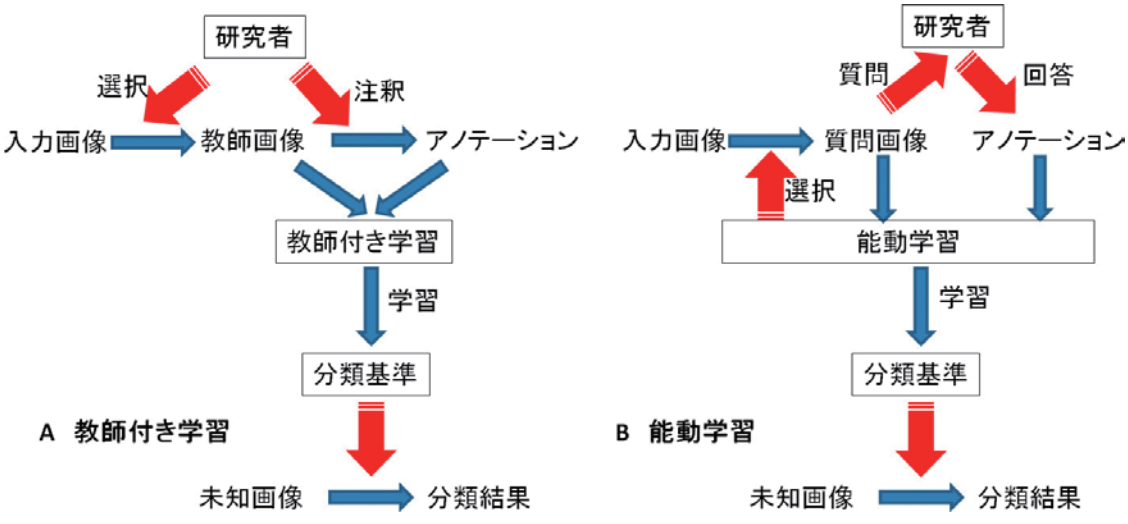

図1．学習プロセスの例．A：教師付き学習による開発のプロセス．研究者は教師画像を選び，アノテーション（注釈）を付与する．教師画像とアノテーションから自動的に分類基準が作成される．これにより未知の画像について自動分類が可能となる．B：能動学習による開発のプロセス．入力された画像群から，研究者がアノテーションすべき画像である質問画像が自動的に選ばれる．研究者は質問に答えることでアノテーションを付与する．この工程を繰り返すことでアノテーション情報を収集し，分類基準を効率よく作成する．

著者連絡先　E-mail: kutsuna@lpixel.net　http://lpixel.net/

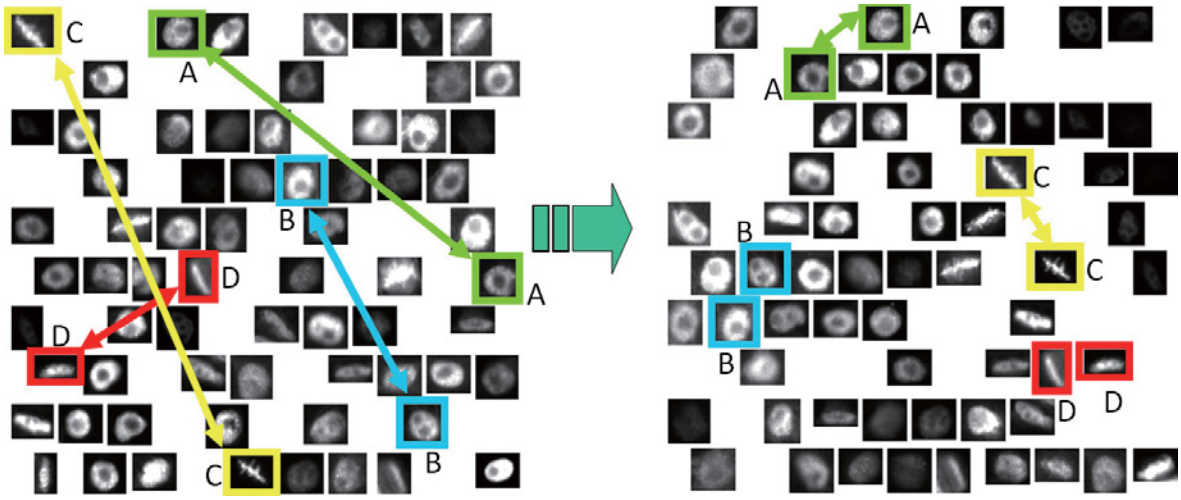

図2．CARTAによる能動学習工程の例．培養細胞の核・染色体の蛍光画像を入力画像群とした．左：学習初期のクラスタリング解析の結果．入力画像を100クラスターに分け，各クラスターから最大1枚ずつ画像を選び表示した．研究者は表示された質問画像の一部に対してアノテーション（注釈）を行う．ここでは間期（A），分裂前期（B），分裂中期（C），分裂終期（D）の核・染色体画像，合計8枚に対しアノテーションを加えた．右：アノテーションを反映して生成したクラスタリング結果．CARTAによるクラスタリング条件の調整により，同じ分裂期の各フェーズとしてアノテーションされた画像が図左に比べて近付いた．このとき，アノテーションを施していない画像についても分裂期の各フェーズに従って分布していることがわかる．

能動学習による効率の良い学習

アノテーション作業を効率よく進める環境を整えることで，同じ時間と労力で，さらに高精度な分類システムを構築できるはずである．こうしたアイディアのもとで，我々はさまざまな生物画像と目的に適用できる汎用性を備えた，適応的な画像分類システムとしてclustering-aided rapid training agent（CARTA）の研究開発を進めてきた[2,3]．CARTAは対話的に研究者の知識を収集し，バイオ画像から抽出可能な多様な評価尺度の組合せの中から，各自の目的に相応しい分類基準を探し出すことを目指した能動学習のシステムである．能動学習[2]は名前の通り，コンピュータがアノテーション作業にあたる研究者に対し能動的に質問することで学習を進める（図1B）．従来の教師付き学習と異なり，アノテーション作業を効率化の対象とした点が能動学習の特長である．細胞周期を例にとると，コンピュータ側から質問画像が示され，それに研究者は「G1期」「M期」といった答を入力する．すると次に別の質問画像が表示され，これに研究者はまた答える……．この対話的なやりとりによりアノテーションが進む．

この能動学習では，学習が効率的に進むように「良い質問」を研究者に尋ねる必要がある．たとえば分類の境界に近い，判断の難しい画像は質問画像にふさわしい[4]．アノテーション済みで答がわかっている画像に似た画像より，すでにアノテーションした画像とは様子の異なる画像について研究者に質問した方が，得られる情報は多い．質問の方法としてCARTAでは，次のようなクラスタリング解析を採用した．まず，画像に対するアノテーション情報やアノテーションを付与したか否かに関わりなく，画像相互の類似性によって入力画像は複数のクラスターに分けられる[2]．次に，各クラスターから1枚ずつ画像を抜き出して並べ，これを質問画像群として研究者に示す（図2左）．研究者は質問画像群の一部にのみ，アノテーションを付与すればよい．CARTAはアノテーション画像が互いに近付くよう，クラスタリング解析に用いる画像相互の類似性の尺度を調整し，次の質問画像群を生成する（図2右）．これにより分類目的にあった，アノテーション作業のしやすい質問画像群を研究者に呈示できる．また，入力画像群が多数ある場合，アノテーション済みの画像とはかけ離れた画像が，質問画像群として優先的に選ばれる．これらの性質はいずれも，短時間のアノテーション作業で多くの知識を研究者から引き出すことに効果的に働く．

しかも，CARTAを使って作成した分類システムは，従来の教師付き学習で作成した分類システムと比べ，複数のバイオ画像でより高精度となった[3]．この結果は効率的なアノテーション作業の整備が，開発コストを下げるとともに分類性能の向上をもたらしたことを示す．今後，このような分類基準の効率的な学習の枠組みが，撮像機器のポテンシャルを引出す[5]とともに，幅広い研究分野で活用される基盤技術へと成熟していくことが期待される．

文　献

1) Gambe, A. E. *et al.*: *Cytometry A*, **71A**, 286 (2007).
2) Kutsuna, N. *et al.*: *Nat. Commun.*, **3**, 1032 (2012).
3) 松永幸大：生物工学，**91**, 33 (2013).
4) Balcan, M. F. *et al.*: *Mach. Learn.*, **80**, 111 (2010).
5) Homeyer, A. *et al.*: *J. Pathol. Inform.*, **2**, S11 (2011).