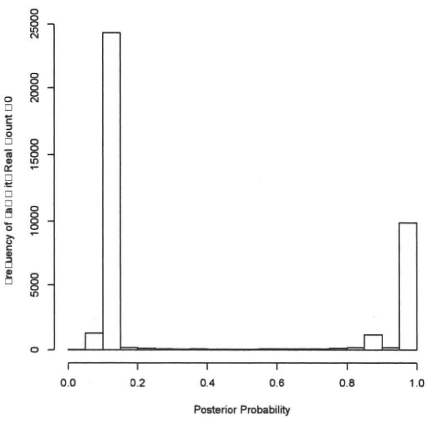


Figure 1. Posterior probability distribution of true sequences

In addition to the well known general trend for error probability to increase with position, there are some interesting and apparently non-random patterns regarding the position specific error rate. Presumably this reflects details of the particular experimental sample we used to train our simulator.

V. Conclusion

In this paper we have demonstrated that identification of reliable sequences from next generation sequencing reads can be done accurately by machine learning classification techniques using observed count, estimate true count from error correction tool (RECOUNT), log likelihood with entropy penalty, log likelihood ratio, expectation matching score, and specific correction coefficient (SCC) as features. We also show that by using SVM posterior probability ~0.12 the biologist can select reads with high precision and recall.

One advantage of our approach is that it does not rely on a reference genome. This is useful because in some important cases a reliable reference genome is not available; e.g. metagenomic analysis, transcriptomic analysis of cancer cells (with possible chromosomal rearrangements), and analysis of highly polymorphous species. Also, even if we have a reliable reference genome, the sequences may not be easily alignable if the sequences have been spliced or otherwise processed. Our result should provide some aid in further analysis of next generation data such as assembly or mapping.

Acknowledgment

This work was supported by Grant-in-Aid for Scientific Research on Innovative Areas (221S002).

References

[1] Akmaev, V.R. and Wang, C. J, Correction of sequence-based artifacts in serial analysis of gene expression, *Bioinformatics*, 20:1254-1263, 2004.

[2] Beißbarth, T., *et al.*, Statistical modeling of sequencing errors in SAGE libraries, *Bioinformatics*, 20:i31-i39, 2004.

[3] Bianchetti, L. *et al.*, SAGETTARIUS: a program to reduce the number of tag mapped to multiple transcripts and to plan SAGE sequencing tags, *Nucleic Acids Research*, 35(18):e122, 2007.

[4] Colinge, J. and Feger, G. Detecting impact of sequencing errors on SAGE data, *Bioinformatics*, 17(9):840-842, 2001.

[5] Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data using the EM algorithm. *Journal of Royal Statistical Society*, (39): 1-38, 1977.

[6] Dohm, J. C, *et al.* Substantial biases in ultra-short read data sets from high-throughput DNA sequencing, *Nucleic Acids Research*, 36(16):e105, 2008.

[7] Ewing. B, and Green, P., Base-calling of automated sequencer traces using Phred II error probabilities, *Genome Research*, (8):186-194, 1998.

[8] Frith, M.C, Wan, R. and Horton, P. Incorporating sequence quality data into alignment improves DNA read mapping, *Nucleic Acids Research*, 38(7):e100.

[9] Hildebrand, B. F., *Introduction to Numerical Analysis: 2nd edition*, Dover Publications, 1987.

[10] Ghildiyal, M. *et al.* Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells, *Science*, (5879):1077-81, 2008.
learning strategies, *Genome Biology*, 10(8):R83, 2009

[11] Qu, W., Hashimoto, S. and Morishita, S., Efficient frequency-based de novo short read clustering for error trimming in next-generation sequencing, *Genome Research*, (19):1309-1315, 2009.

[12] Yang X., Dorman K.S., Aluru S., Reptile: representative tiling for short read error correction, *Bioinformatics*, 26:2526-2533, 2010.

[13] Rougemont, J. *et al.*, Probabilistic base calling of Solexa sequencing data, *BMC Bioinformatics*, (9):431, 2008.

[14] Schröder J., Shröder H., Puglisi S., Sinha R., Schmidt B. SHREC: a short-read error correction method, *Bioinformatics*, 25:2157-2163, 2009.

[15] Vapnik, V. *The Nature of Statistical Learning Theory*, Springer, 1995.

[16] Velculescu, V.E. *et al.*, Analysis of human transcriptomes, *Nature Genetics*, (270):484-487, 1999.

[17] Weihs, C., Ligges, U., Luebke, K. and Raabe, N. klaR Analyzing German Business Cycles. In *Baier, D., Decker, R. and Schmidt-Thieme, L. (eds.). Data Analysis and Decision Support*, 335-343, Springer-Verlag, Berlin, 1996.

[18] Wijaya, E., Frith, M., Suzuki, Y., Horton, P., RECOUNT: Expectation maximization based error correction tool for next generation sequencing data, *Genome Informatics*,23:189-200, 2009.

本コーナーでは，情報共有によるバイオ研究の発展をめざして，Webコンテンツや研究リソースの提供機関，受託研究機関などバイオ研究者に有益な情報を，不定期連載としてご紹介いただきます．

情報共有と有効活用のための
バイオ研究耳よりツール

細胞情報解析に役立つツール

—幹細胞研究の進展とその創薬応用に向けて

千葉啓和，藤渕　航

はじめに

近年，幹細胞の研究が目覚しい進展を遂げており，これからは創薬への応用が見込まれている．こうした時代には，細胞情報の圧倒的な多様化と，大規模化に対応しなければならない．まず，**細胞の辞書が必要**になるだろう．次に，**その辞書を高速に探索する手段も必要**だ．CELLPEDIAは，細胞に関するさまざまな情報を統合したデータベースであり，手元にある細胞情報を調べるときの辞書として活用できるものである．CellMongateとSAMURAIは，発現データに潜む特徴を抽出するソフトウェアだ．手元の発現プロファイルはどういった細胞に近いのか，またそこではどのような遺伝子群が発現しているのかを高速に調べることができる．

1 ヒト細胞辞書：CELLPEDIA

手元のデータを正しく解釈するためには，信頼できる情報を参照することが必要だ．細胞情報を網羅的に収集し，利用しやすい形にまとめた，いわば細胞の電子辞書が必要である．CELLPEDIA[1]（http://cellpedia.cbrc.jp/）は，こうした情報をまとめ，加工して提供するものである．データベースの骨格は，ヒト体細胞および幹細胞を2,000種類以上に分類した表だ．分類された各細胞に対して詳細なアノテーション（注釈）が施されている．細胞分化の情報も入っている．例えばある種類の細胞から出発して，その「親」の細胞種，あるいは「子」の細胞種へとリンクをたどることができる．

CELLPEDIAは，ユーザーからのサブミッションによって拡大する．サブミットされた情報はまず1次情報として蓄えられる．キュレーション（検証と修正）によってそれらの情報が整理され，さらにソフトウェアを用いたデータマイニングが行われ，2次情報としてCELLPEDIAに登録される．登録されているデータの具体例は，**図1**を参照して欲しい．画像情報，発現データ，文献情報を中心に，さまざまな情報がまとめられている．細胞の形態は，Cytometricaというソフトウェアにより，細胞画像から自動的に抽出されたものだ．こうして得られた細胞ごとの発現プロファイルや，細胞形態データは，定量的な細胞解析の基盤を提供する．

2 発現データ検索：CellMotage

CellMontage[2]（http://cellmontage.cbrc.jp/）は，発現データを比較するツールだ．これによって，

Hirokazu Chiba/Wataru Fujibuchi：Cell Function Design Team, Computational Biology Research Center（CBRC）, National Institute of Advanced Industrial Science and Technology（AIST）〔産業技術総合研究所（AIST）生命情報工学研究センター（CBRC）細胞機能設計チーム〕

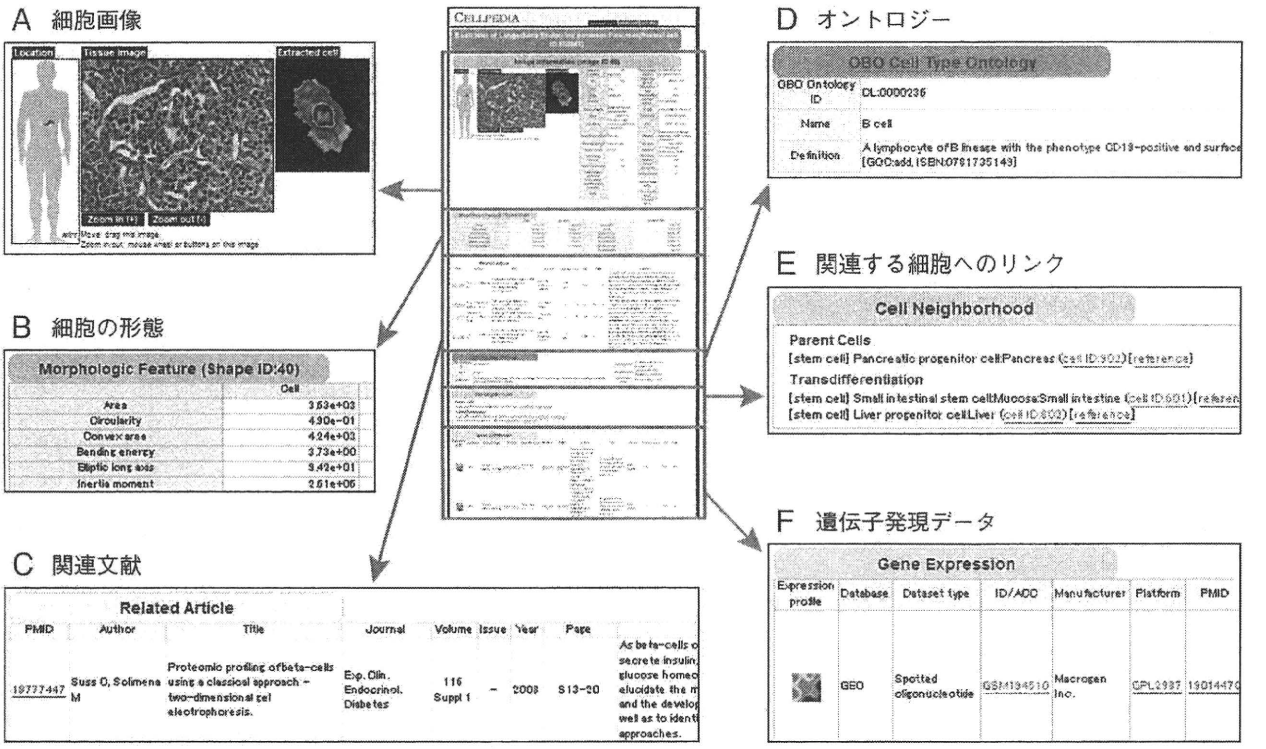


図1　CELLPEDIAに含まれる情報（例：膵β細胞）

手元の発現プロファイルが，どういった細胞のプロファイルに近いのかを知ることができる．問題は，大量のデータを処理するために，高速でなければならないということだ．CellMontageは，RaPiDSアルゴリズム[3]によって高速な検索を実現している．**図2A**に示すように，シンプルな重み付きの順位相関係数に基づいて類似度を評価する（他の重み付けの定義もある）．この手法を用いて，発現データベース中に埋もれている，よく似たプロファイルを瞬時に探し当てることができる．

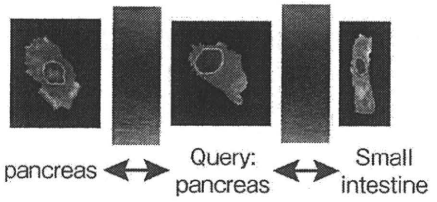
図2Bの例は，クエリーに膵臓の発現プロファイル，データベースにCELLPEDIAを指定して検索をかけた結果だ．上位には膵臓がヒットしているが（赤枠），5番目をみると小腸がヒットしている（青枠）．これらは比較的「近い」細胞であると考えられる．実際CELLPEDIAを参照すると，腸幹細胞が分化転換してインスリンを分泌するようになる例が報告されている（**図1E**）．

細胞分化研究の時代には，多様な細胞を構造化して扱うために，細胞の類似度を定量化するアプローチがますます重要になるだろう．こうした定量化は，細胞の分化誘導，分化転換の研究に対しても重要な示唆を与えると考えられる．今後は，iPS細胞等についても多種多様な細胞株が作製されると考えられる．そうしたケースでも，細胞の類似度を瞬時に測定することは，解析の重要な切り口となるだろう．

3 遺伝子モジュール抽出：SAMURAI

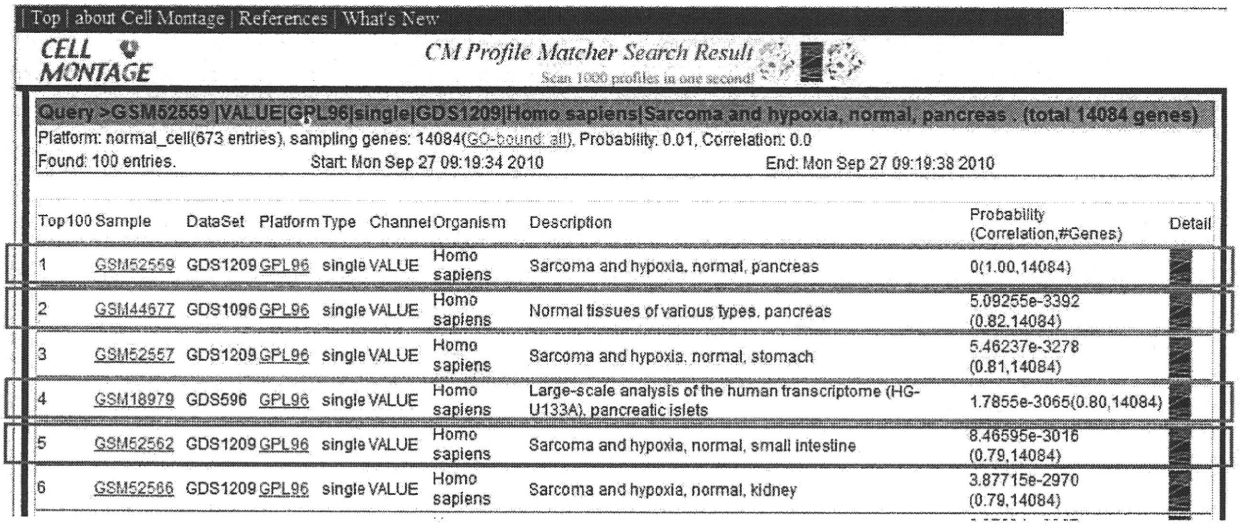
ここまで，発現プロファイル同士の比較について説明した．ではそうした発現プロファイル間の違いには，どのような遺伝子が関与しているのだろうか．SAMURAI[4]（http://samurai.cbrc.jp/）は，発現プロファイルを入力として，共通の遺伝子制御を受けている遺伝子群，すなわち遺伝子モジュールを網羅的に抽出するツールだ．特色は，似た遺伝子をまとめると同時に，細胞もグループ化するバイクラスタリング法である．このため，特定の細胞種で限定的に現れる遺伝子モジュールを捉えるこ

A



$$CorrW(x_1,x_2)=\frac{\sum_i^n (w_i r_1(g_i)-\bar{r}_{1w})(w_i r_2(g_i)-\bar{r}_{2w})}{\sqrt{\sum_i^n (w_i r_1(g_i)-\bar{r}_{1w})^2 \sum_i^n (w_i r_2(g_i)-\bar{r}_{2w})^2}},\ \bar{r}_{1w}=\frac{\sum_i^n w_i r_1(g_i)}{\sum_i^n w_i},\ \bar{r}_{2w}=\frac{\sum_i^n w_i r_2(g_i)}{\sum_i^n w_i}$$

B



Top | about Cell Montage | References | What's New

CELL MONTAGE — *CM Profile Matcher Search Result* — Scan 1000 profiles in one second!

Query >GSM52559 |VALUE|GPL96|single|GDS1209|Homo sapiens|Sarcoma and hypoxia, normal, pancreas , (total 14084 genes)

Platform: normal_cell(673 entries), sampling genes: 14084(GO-bound: all), Probability: 0.01, Correlation: 0.0

Found: 100 entries. Start: Mon Sep 27 09:19:34 2010 End: Mon Sep 27 09:19:38 2010

Top100	Sample	DataSet	Platform	Type	Channel	Organism	Description	Probability (Correlation,#Genes)	Detail
1	GSM52559	GDS1209	GPL96	single	VALUE	Homo sapiens	Sarcoma and hypoxia, normal, pancreas	0(1.00,14084)	
2	GSM44677	GDS1096	GPL96	single	VALUE	Homo sapiens	Normal tissues of various types, pancreas	5.09255e-3392 (0.82,14084)	
3	GSM52557	GDS1209	GPL96	single	VALUE	Homo sapiens	Sarcoma and hypoxia, normal, stomach	5.46237e-3278 (0.81,14084)	
4	GSM18979	GDS596	GPL96	single	VALUE	Homo sapiens	Large-scale analysis of the human transcriptome (HG-U133A), pancreatic islets	1.7855e-3065(0.80,14084)	
5	GSM52562	GDS1209	GPL96	single	VALUE	Homo sapiens	Sarcoma and hypoxia, normal, small intestine	8.46595e-3016 (0.79,14084)	
6	GSM52566	GDS1209	GPL96	single	VALUE	Homo sapiens	Sarcoma and hypoxia, normal, kidney	3.87715e-2970 (0.79,14084)	

図2 CellMontageによる発現比較

A）重み付き順位相関係数による発現の類似度の計算，B）膵臓の発現プロファイルをクエリーにして検索をかけた結果

とが可能となる．しかしここでもやはり問題になるのが計算量だ．SAMURAIは，LCM[5]とよばれるアルゴリズムを応用することで，非常に高速にモジュールを抜き出すことを可能にしている．

図3に示すのは，クエリーに膵臓のプロファイル，データベースにCELLPEDIAの発現データを指定し，遺伝子モジュールを抽出した結果だ．上位には膵臓特異的なモジュール（pancreatic lipaseを含む）がヒットしているが，7番目には肝臓で発現するモジュール（cytochrome P450を含む）がヒットしている（**図3A**）．ここでCELLPEDIAを参照してみると，興味深いことに肝前駆細胞から膵前駆細胞への分化転換が報告されている（**図1E**）．上のモジュール群はこうした分化転換のメカニズムを知る手がかりとなりうるだろう．各遺伝子モジュールについて詳細な情報が知りたければ，KEGG（http://www.genome.jp/kegg/）のパスウェイ中に位置付けて確認することができる（**図3B**）．SAMURAIのデモ版のプログラムについては，ダウンロードして手元のコンピュータ上で動かすことも可能である．

おわりに

CELLPEDIAは，ユーザー参加型で拡大するものになっているので，これからますます拡充され，将来的には，新たに出現する多様なデータ形式をも取り込んで発展していくだろう．今後多様化する細胞情報が構造化された状態で蓄積していくと期待される．また，これからあらゆる情報が大規模化すると予見される．ここで紹介したような高速化されたソフトウェアがこれからの時代には必須にな

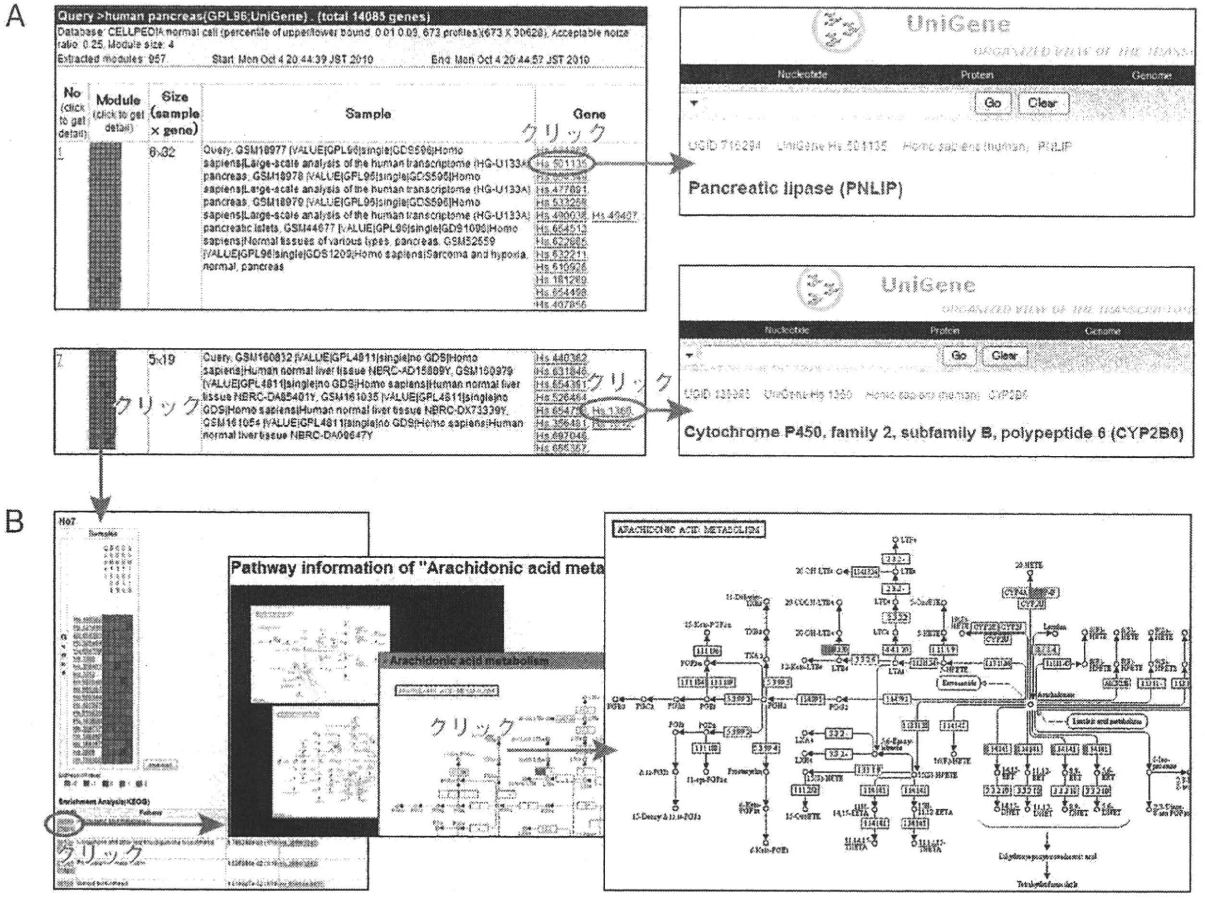


図3 SAMURAIにより抽出された遺伝子モジュール
A）遺伝子モジュールのリスト，B）遺伝子モジュールの詳細な情報

るだろう．また，ここで紹介したツールは，細胞情報からネットワークを構築する技術にも応用できる．われわれの所属する研究チームでは，細胞のプロファイル比較を応用して細胞の分化ネットワークを推定したり，抽出した遺伝子モジュールを初期値として利用して遺伝子の制御ネットワーク推定を高速化する研究が進められている．

文献

1）幡野晶子 他：IPSJ SIG BIO. 20：1-4, 2010
2）Fujibuchi, W. et al.：Bioinformatics, 23：3103-3104, 2007
3）Horton, P. et al.：Genome Informatics, 17-2：67-76, 2006
4）Fujibuchi, W. et al.：Methods in Molecular Biology, 577：55-65, 2009
5）Uno, T. et al.：Lecture Notes in Artificial Intelligence, 3245：16-31, 2004

筆頭著者プロフィール Profile

千葉啓和：東京大学大学院情報理工学系研究科修士課程修了．2008年から産業技術総合研究所生命情報工学研究センターに勤務し，細胞情報を扱う技術の研究開発に携わる．細胞情報計算における組合わせ最適化アルゴリズムやFPGAを用いた高速アラインメント回路等の設計に携わる一方，実験研究者と共同で細胞分化にかかわる新規遺伝子の探索等を行っている．計算機科学と実験生物学をどのように組み合わせれば効果的な発見ができるかに興味がある．