89. Kubisch CH, Gukovsky I, Lugea A, et al. Long-term ethanol consumption alters pancreatic gene expression in rats: a possible connection to pancreatic injury. *Pancreas* 2006; 33: 68–76.
90. Kodavanti UP, Schladweiler MC, Ledbetter AD, et al. The spontaneously hypertensive rat: an experimental model of sulfur dioxide-induced airways disease. *Toxicol Sci* 2006; 94: 193–205.
91. Bruder ED, Lee JJ, Widmaier EP, Raff H. Microarray and real-time PCR analysis of adrenal gland gene expression in the 7-day-old rat: effects of hypoxia from birth. *Physiol Genomics* 2007; 29: 193–200.
92. Almon RR, DuBois DC, Yao Z, Hoffman EP, Ghimbovschi S, Jusko WJ. Microarray analysis of the temporal response of skeletal muscle to methylprednisolone: comparative analysis of two dosing regimens. *Physiol Genomics* 2007; 30: 282–299.
93. Chan MM, Lu X, Merchant FM, Iglehart JD, Miron PL. Gene expression profiling of NMU-induced rat mammary tumors: cross species comparison with human breast cancer. *Carcinogenesis* 2005; 26: 1343–1353.
94. Kendziorski C, Irizarry RA, Chen KS, Haag JD, Gould MN. On the utility of pooling biological samples in microarray experiments. *Proc Natl Acad Sci USA* 2005; 102: 4252–4257.
95. Aplin AC, Gelati M, Fogel E, Carnevale E, Nicosia RF. Angiopoietin-1 and vascular endothelial growth factor induce expression of inflammatory cytokines before angiogenesis. *Physiol Genomics* 2006; 27: 20–28.
96. Rampil IJ, Moller DH, Bell AH. Isoflurane modulates genomic expression in rat amygdala. *Anesth Analg* 2006; 102: 1431–1438.
97. Collins JF. Gene chip analyses reveal differential genetic responses to iron deficiency in rat duodenum and jejunum. *Biol Res* 2006; 39: 25–37.
98. Guzelian J, Barwick JL, Hunter L, Phang TL, Quattrochi LC, Guzelian PS. Identification of genes controlled by the pregnane X receptor by microarray analysis of mRNAs from pregnenolone 16alpha-carbonitrile-treated rats. *Toxicol Sci* 2006; 94: 379–387.
99. Gebel S, Gerstmayer B, Kuhl P, Borlak J, Meurrens K, Muller T. The kinetics of transcriptomic changes induced by cigarette smoke in rat lungs reveals a specific program of defense, inflammation, and circadian clock gene expression. *Toxicol Sci* 2006; 93: 422–431.
100. Su Y, Simmen FA, Xiao R, Simmen RC. Expression profiling of rat mammary epithelial cells reveals candidate signaling pathways in dietary protection from mammary tumors. *Physiol Genomics* 2007; 30: 8–16.
101. Rowe WB, Blalock EM, Chen KC, et al. Hippocampal expression analyses reveal selective association of immediate-early, neuroenergetic, and myelinogenic pathways with cognitive impairment in aged rats. *J Neurosci* 2007; 27: 3098–3110.
102. Volpicelli F, Caiazzo M, Greco D, et al. Bdnf gene is a downstream target of Nurr1 transcription factor in rat midbrain neurons in vitro. *J Neurochem* 2007; 102: 441–453.
103. Stemmer K, Ellinger-Ziegelbauer H, Ahr HJ, Dietrich DR. Carcinogen-specific gene expression profiles in short-term treated Eker and wild-type rats indicative of pathways involved in renal tumorigenesis. *Cancer Res* 2007; 67: 4052–4068.
104. Impey S, McCorkle SR, Cha-Molstad H, et al. Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell* 2004; 119: 1041–1054.
105. Bush EW, Hood DB, Papst PJ, et al. Canonical transient receptor potential channels promote cardiomyocyte hypertrophy through activation of calcineurin signaling. *J Biol Chem* 2006; 281: 33487–33496.
106. Zhou Z, Cornelius CP, Eichner M, Bornemann A. Reinnervation-induced alterations in rat skeletal muscle. *Neurobiol Dis* 2006; 23: 595–602.
107. Bursztyn M, Gross ML, Goltser-Dubner T, et al. Adult hypertension in intrauterine growth-restricted offspring of hyperinsulinemic rats: evidence of subtle renal damage. *Hypertension* 2006; 48: 717–723.
108. Thomas H, Senkel S, Erdmann S, et al. Pattern of genes influenced by conditional expression of the transcription factors HNF6, HNF4alpha and HNF1beta in a pancreatic beta-cell line. *Nucleic Acids Res* 2004; 32: e150.
109. Schumann A, Nutten S, Donnicola D, et al. Neonatal antibiotic treatment alters gastrointestinal tract developmental gene expression and intestinal barrier transcriptome. *Physiol Genomics* 2005; 23: 235–245.
110. Roy S, Khanna S, Kuhn DE, et al. Transcriptome analysis of the ischemia-reperfused remodeling myocardium: temporal changes in inflammation and extracellular matrix. *Physiol Genomics* 2006; 25: 364–374.
111. Tugues S, Morales-Ruiz M, Fernandez-Varo G, et al. Microarray analysis of endothelial differentially expressed genes in liver of cirrhotic rats. *Gastroenterology* 2005; 129: 1686–1695.
112. Akavia UD, Shur I, Rechavi G, Benayahu D. Transcriptional profiling of mesenchymal stromal cells from young and old rats in response to Dexamethasone. *BMC Genomics* 2006; 7: 95.
113. Zhou M, Roma A, Magi-Galluzzi C. The usefulness of immunohistochemical markers in the differential diagnosis of renal neoplasms. *Clin Lab Med* 2005; 25: 247–257.
114. Schiffer D, Giordana MT, Mauro A, Migheli A, Germano I, Giaccone G. Immunohistochemical demonstration of vimentin in human cerebral tumors. *Acta Neuropathol* 1986; 70: 209–219.
115. Niehans GA, Manivel JC, Copland GT, Scheithauer BW, Wick MR. Immunohistochemistry of germ cell and trophoblastic neoplasms. *Cancer* 1988; 62: 1113–1123.
116. Iwakuma T, Lozano G. Crippling p53 activities via knock-in mutations in mouse models. *Oncogene* 2007; 26: 2177–2184.

117. Marine JC, Jochemsen AG. Mdmx as an essential regulator of p53 activity. *Biochem Biophys Res Commun* 2005; 331: 750–760.
118. Tang Y, Zhao W, Chen Y, Zhao Y, Gu W. Acetylation is indispensable for p53 activation. *Cell* 2008; 133: 612–626.
119. Gong X, Kole L, Iskander K, Jaiswal AK. NRH:quinone oxidoreductase 2 and NAD(P)H:quinone oxidoreductase 1 protect tumor suppressor p53 against 20s proteasomal degradation leading to stabilization and activation of p53. *Cancer Res* 2007; 67: 5380–5388.
120. Lai Z, Yang T, Kim YB, et al. Differentiation of Hdm2-mediated p53 ubiquitination and Hdm2 autoubiquitination activity by small molecular weight inhibitors. *Proc Natl Acad Sci USA* 2002; 99: 14734–14739.
121. Wang W, Ho WC, Dicker DT, et al. Acridine derivatives activate p53 and induce tumor cell death through Bax. *Cancer Biol Ther* 2005; 4: 893–898.
122. Kawata K, Yokoo H, Shimazaki R, Okabe S. Classification of heavy-metal toxicity by human DNA microarray analysis. *Environ Sci Technol* 2007; 41: 3769–3774.
123. Fry RC, Navasumrit P, Valiathan C, et al. Activation of inflammation/NF-kappaB signaling in infants born to arsenic-exposed mothers. *PLoS Genet* 2007; 3: e207.
124. Chang L, Zhou B, Hu S, et al. ATM-mediated serine 72 phosphorylation stabilizes ribonucleotide reductase small subunit p53R2 protein against MDM2 to DNA damage. *Proc Natl Acad Sci USA* 2008; 105: 18519–18524.
125. Kollberg G, Darin N, Benan K, et al. A novel homozygous RRM2B missense mutation in association with severe mtDNA depletion. *Neuromuscul Disord* 2009; 19: 147–150.
126. Liu X, Xue L, Yen Y. Redox property of ribonucleotide reductase small subunit M2 and p53R2. *Methods Mol Biol* 2008; 477: 195–206.
127. Spinazzola A, Invernizzi F, Carrara F, et al. Clinical and molecular features of mitochondrial DNA depletion syndromes. *J Inherit Metab Dis* 2009; 32: 143–158.
128. Tyynismaa H, Suomalainen A. Mouse models of mitochondrial DNA defects and their relevance for human disease. *EMBO Rep* 2009; 10: 137–143.
129. Ceryak S, Zingariello C, O'Brien T, Patierno SR. Induction of pro-apoptotic and cell cycle-inhibiting genes in chromium (VI)-treated human lung fibroblasts: lack of effect of ERK. *Mol Cell Biochem* 2004; 255: 139–149.
130. Fanzo JC, Reaves SK, Cui L, et al. Zinc status affects p53, gadd45, and c-fos expression and caspase-3 activity in human bronchial epithelial cells. *Am J Physiol Cell Physiol* 2001; 281: C751–C757.
131. Shih RS, Wong SH, Schoene NW, Lei KY. Suppression of Gadd45 alleviates the G2/M blockage and the enhanced phosphorylation of p53 and p38 in zinc supplemented normal human bronchial epithelial cells. *Exp Biol Med (Maywood)* 2008; 233: 317–327.
132. Toyoshiba H, Sone H, Yamanaka T, et al. Gene interaction network analysis suggests differences between high and low doses of acetaminophen. *Toxicol Appl Pharmacol* 2006; 215: 306–316.
133. Yamanaka T, Toyoshiba H, Sone H, Parham FM, Portier CJ. The TAO-Gen algorithm for identifying gene interaction networks with application to SOS repair in *E. coli*. *Environ Health Perspect* 2004; 112: 1614–1621.
134. Sone H, Imanishi S, Akanuma H, et al. Gene expression signatures of environmental chemicals in cancer and in developmental disorders. In: Zhao BDM, Cadeans E, ed. *The Roles of Free Radicals in Biology and Medicine*. Beijing: Medimond; 2009, p. 45–52.
135. Chua PJ, Yip GW, Bay BH. Cell cycle arrest induced by hydrogen peroxide is associated with modulation of oxidative stress related genes in breast cancer cells. *Exp Biol Med (Maywood)* 2009; 234: 1086–1094.

Author Query Form

AQ1: "16 molecules meet 44 molecules" is unclear. Please confirm or rewrite.
AQ2: Please provide full publication information for Ref. 50.
AQ3: Please provide full publication information for Ref. 75.

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〕

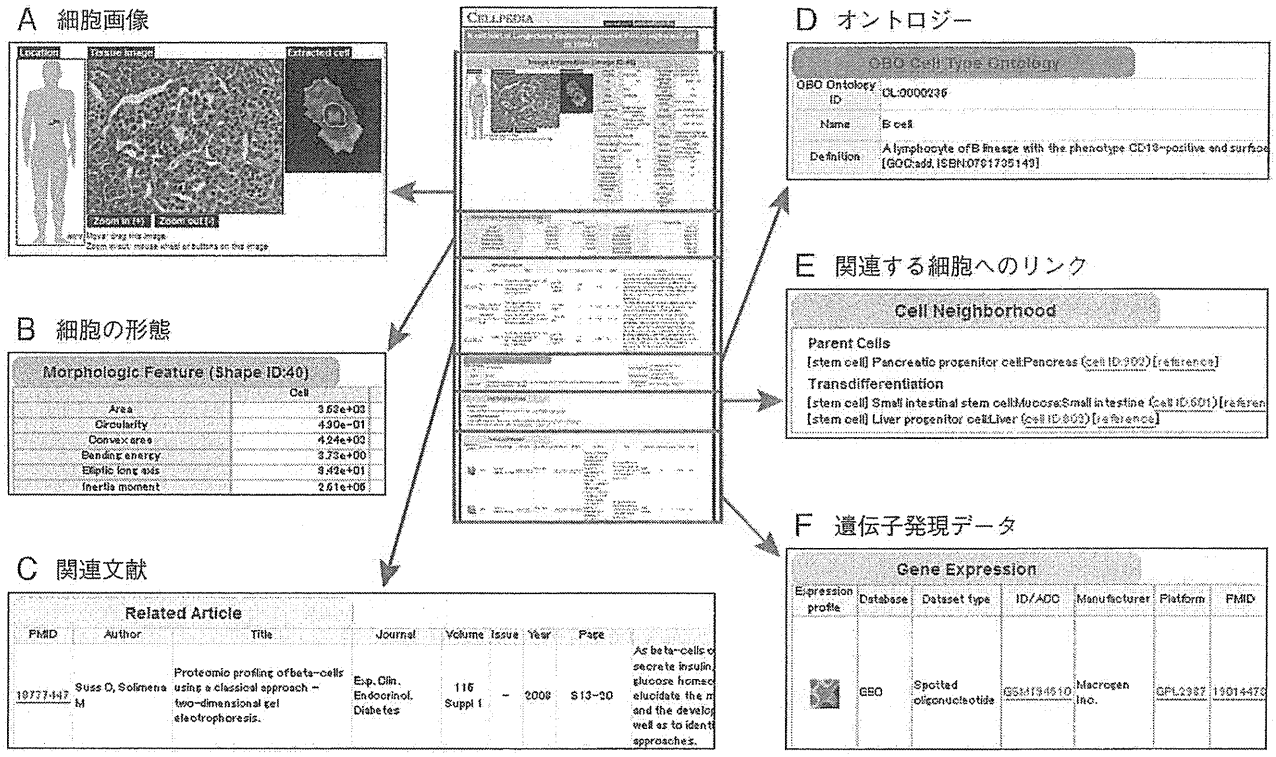


図1 CELLPEDIAに含まれる情報（例：膵β細胞）

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．

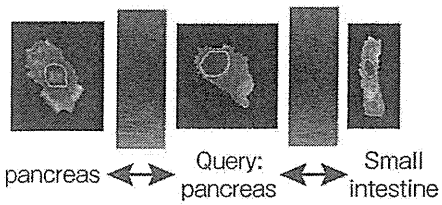
図2Bの例は，クエリーに膵臓の発現プロファイル，データベースにCELLPEDIAを指定して検索をかけた結果だ．上位には膵臓がヒットしているが（赤枠），5番目をみると小腸がヒットしている（青枠）．これらは比較的「近い」細胞であると考えられる．実際CELLPEDIAを参照すると，腸幹細胞が分化転換してインスリンを分泌するようになる例が報告されている（**図1E**）．

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

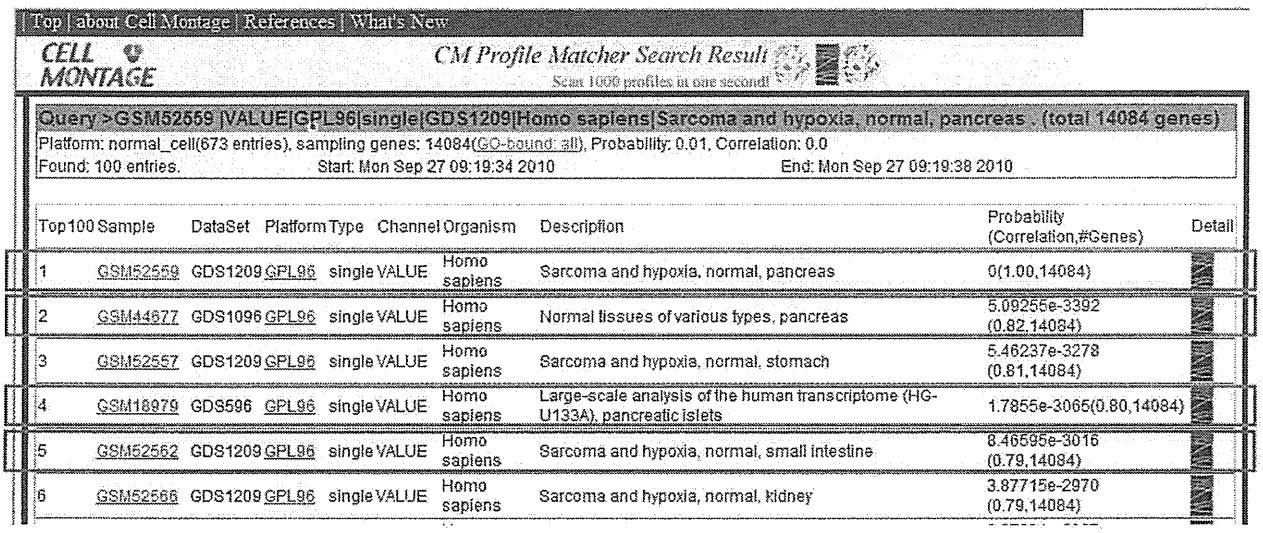
ここまで，発現プロファイル同士の比較について説明した．ではそうした発現プロファイル間の違いには，どのような遺伝子が関与しているのだろうか．SAMURAI[4]（http://samurai.cbrc.jp/）は，発現プロファイルを入力として，共通の遺伝子制御を受けている遺伝子群，すなわち遺伝子モジュールを網羅的に抽出するツールだ．特色は，似た遺伝子をまとめると同時に，細胞もグループ化するバイクラスタリング法である．このため，特定の細胞種で限定的に現れる遺伝子モジュールを捉えるこ

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



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	GSM44877	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

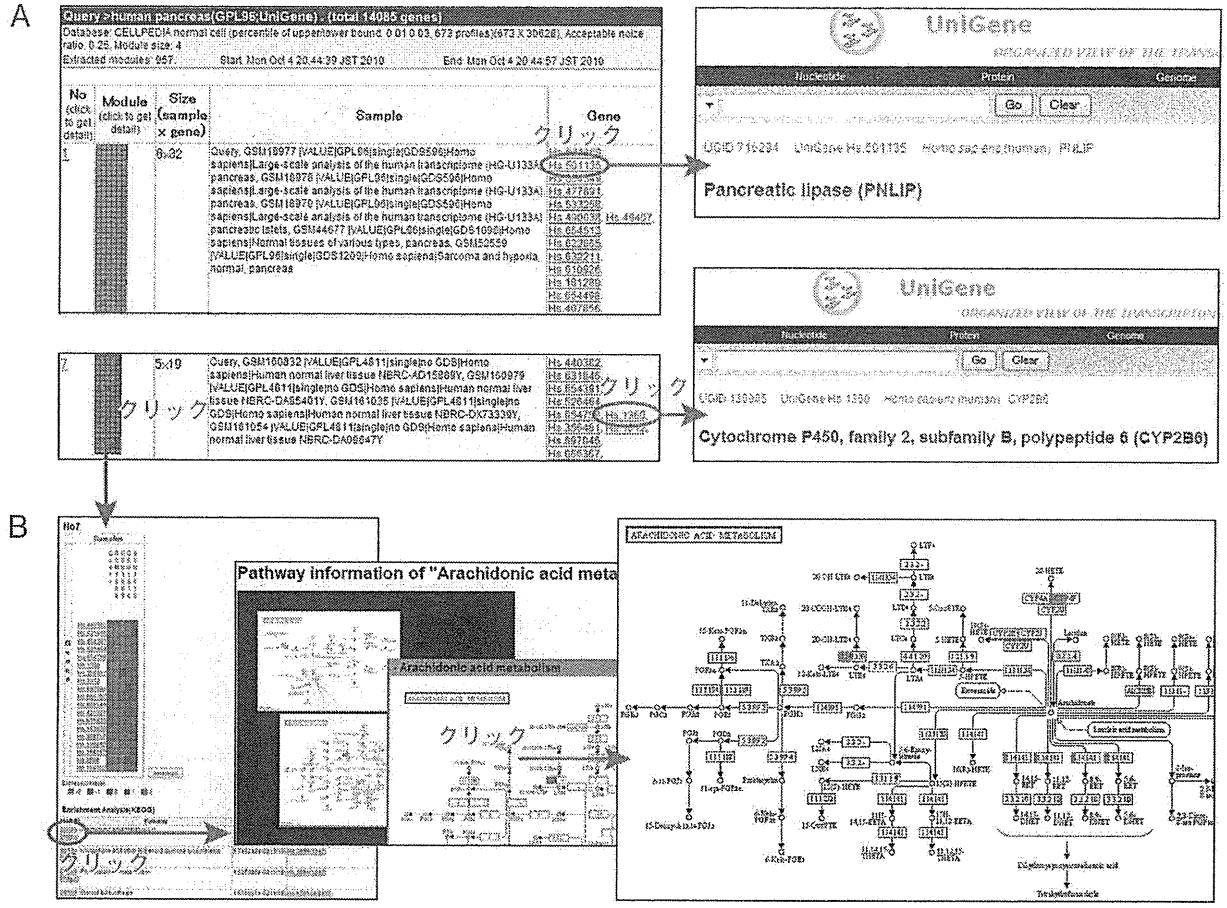


図3 SAMURAIにより抽出された遺伝子モジュール
A）遺伝子モジュールのリスト，B）遺伝子モジュールの詳細な情報

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．

9　シミュレーテッドアニーリングによる多重プライマー配列デザイン法 —細胞内mRNA絶対定量に向けて—

藤渕　航*

9.1　はじめに

細胞内には10^5種類とも言われる遺伝子のmRNA分子が溢れている。これらを網羅的に高速シーケンサーで測定するためには，計算機による厳密なターゲッティングによるプライマー配列の設計が必要である。プライマーの設計も様々なバリエーションがあり，単独プライマー以外にも複数プライマーを同時に溶液中に混合する多重プライマーの設計になると，他のプライマーの意図しないターゲットRNAへのクロスハイブリダイゼーションなどが問題となってくる。これらの問題を解消するには人の手による計算は不可能で，計算機による高速組み合わせ探索が必要である。組み合わせ探索は計算機が最も得意とする分野であり，本節ではシミュレーテッドアニーリングアルゴリズムによる最適化を紹介することにする。

9.2　細胞解析とパイロシーケンサー

文部科学省特定領域研究「ライフサーベイヤ」では1細胞の機能を完全に理解するために細胞内分子を網羅的に定量することが目標として掲げられてきた。その中で遺伝子をコードしているmRNAの網羅的測定は重要なゴールの一つとして掲げられている。特に高速シーケンサーでのmRNAの網羅的配列決定による定量は重要な技術と考えられている。

次世代シーケンサーとして開発された454パイロシーケンサーによるプライマーを用いた核酸配列の決定には，大まかに2種類の方法が存在する。1つは4～6mer程度の短いプライマー配列を用いたランダムシーケンシング法と，8merやそれ以上の長めのプライマーによる特定の遺伝子やエクソンなどの核酸配列を決定するターゲットシーケンシング法がある。ランダムシーケンシング法では，4merの全ての配列パターンを利用すると最大でも4^4=256通りになり，どの遺伝子も検出できるが，反対に必要としない配列であるオフターゲットにも結合してしまう不都合がある。一方，ターゲットシーケンシング法では利用するプライマー配列を限定し，積極的にターゲットの配列を検出するものである。

ターゲットシーケンシング法で目的遺伝子の配列を正確に読み取ることができるためには，mRNAやエクソン配列に対して的確に結合する良質なプライマーをデザインすることが必要である。しかし，実際には最適なプライマーセットをデザインすることは難しい問題であり，例え

* Wataru Fujibuchi　㈱産業技術総合研究所　生命情報工学研究センター
細胞機能設計チーム　研究チーム長

ば10 merなら候補になるプライマーの数が4^{10}などと多く，さらには，オフターゲット遺伝子へのクロスハイブリダイゼーションの最小化，T_m値によるプライマーの選択，多重プライマー設計時でのプライマー間相互作用をするプライマーの除去などを考慮する必要がある。これらの条件下で，「限られた少数のプライマーで最大限のターゲットmRNAを認識できる」プライマーセットを設計しなくてはならない。過去において幾つかの方法が提唱されてきたが，これらの手法はターゲットmRNAの数を最大化するが，多重プライマー設計においてクロスハイブリダイゼーションの数を最小化する様な方法ではなかった。

9.3　高速・高性能プライマー設計システム

高性能プライマーの設計には，大きく分けて，①候補配列の全ゲノムに対するユニーク性検索，②プライマーT_m値の計算，③多重プライマーセットの最適化の3段階が必要である。どの工程も膨大な計算を必要とするもので，それぞれについてできるだけ最新の手法を用いてプライマーを設計することで設計の高速化を実現し，さらにプライマーの特異性もできるだけターゲットと結合しオフターゲットと結合しないようにシステムを開発することが必要である。下にシス

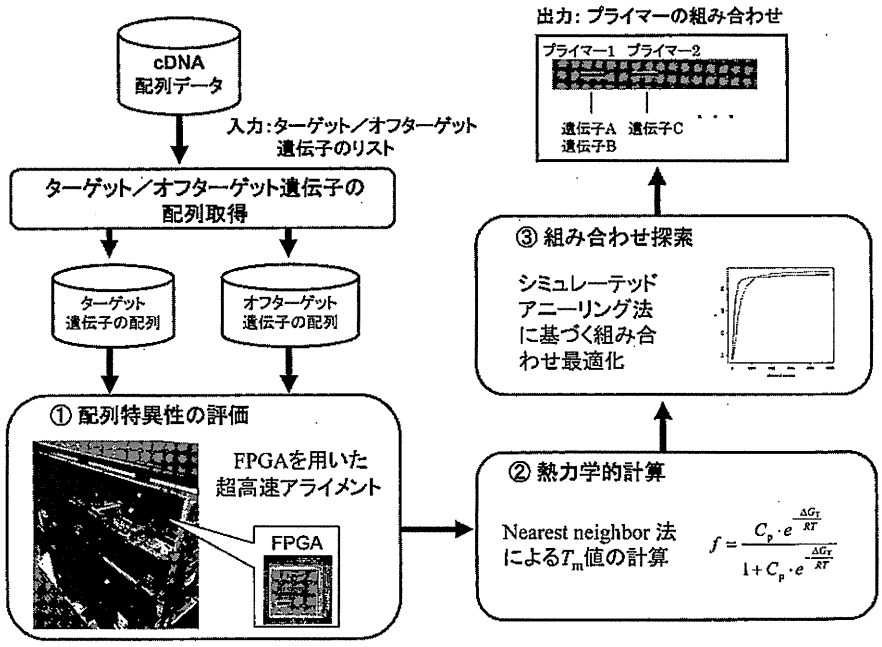


図1　多重プライマー配列設計システムのフロー概念図
3つの工程それぞれに膨大な計算リソースを必要とするが，特に時間のかかる①候補配列の特異性検索と③多重プライマーセットの組み合わせ最適化についてハード・ソフトの両面から工夫を行い，妥協を許さないシステムに仕上げている。

テム全体のフローを示してある（図１）。

これら３つの工程について実際に具体的な例を挙げながらさらに詳しく述べることにする。基本的にプライマー配列はゲノム上のどの配列に対しても設計可能であるが，ここでは各遺伝子mRNAについてのプライマー設計についてのみ説明している。

９.３.１　候補配列の特異性高速検索

通常の実験では細胞からの抽出物に含まれるmRNAをpoly-A配列の相補配列であるpoly-Tで捉えることが多いため３'側の配列の方がcDNAに合成されて増幅されやすい。そこでプライマー配列もpoly-A側で設計する方がよりcDNAを検出できることになるため設計はできるだけ３'側に近いところで行うのが慣習となっている。従って我々のシステムでも，poly-Aのすぐ上流配列1,000塩基以内からプライマーを設計する仕様とし，各遺伝子につき約1,000カ所の候補を生成する。これら各プライマーを約50,000種類もの全cDNAと比較してその配列特異性（結合できる遺伝子の数）を検証することが必要である。ここでcDNAの配列データはEnsemblデータベース（http://www.ensembl.org/）から取得した。

実際に配列特異性を検証する場合に，50,000個のcDNAを１つずつ取り出し，残りの全cDNAに対してギャップなしアライメントを取ると32 CPUでも１年もかかってしまう。我々は，Field Programmable Gate Array（FPGA）と呼ばれるプログラム可能な論理演算設計ボードを用いて，プライマーの特異性検出用の並列計算専用回路を設計した。この加速ボードを通常のPCに挿入して候補プライマーのアライメントを行ったところ，計算を約1,000倍に高速化することに成功

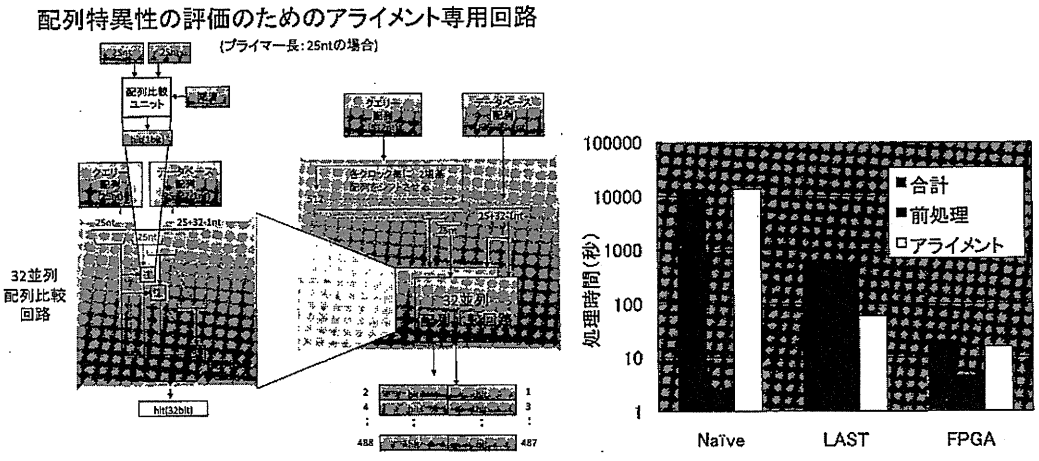


図２　FPGA加速ボードを用いた場合のプライマー配列の特異性検索に特化した専用回路（左）とFPGA回路を用いて計算に要する処理時間を単純法（Naïve）と高速ソフトウェア（LAST）と比較した結果（右）前処理を除く時間で比較すると，単純法に比較してLASTで約200倍，FPGAで約1,000倍の高速化が達成されている[1)]。

し，50,000遺伝子であっても１CPUで約10日程度の計算時間にまで短縮した（図２）。

9.3.2　プライマーのT_m値の計算

T_m値はプライマーの総量の半数が結合する温度である。この計算にはNearest neighbor法[2]が一般的に用いられる。T_m値の求め方の手順はイオン濃度なども考慮しながらプライマーの温度Tにおける結合自由エネルギー変化ΔG_Tを計算し，これから結合率fが丁度半分になる（f=0.5）温度を，式(1)を利用して求めるものである。ここで，C_pはプライマー濃度，Rは気体定数，Tは絶対温度である。１遺伝子の計算時間は僅かであるが，50,000遺伝子となるとやはり数ヶ月と言った膨大な時間がかかる。そこでこれを数日で計算できる様に既存のプログラムhybrid-min[3]を高速なバイナリーサーチアルゴリズム[4]と組み合わせてT_m値を絞り込むプログラムを作成した。バイナリーサーチアルゴリズムは，n個のデータがある場合の計算量が平均$log2n$回で目的の値を探すことができる手法で，nが増えても計算量はlogスケールでしか増加しない$O(log\ n)$の優れた手法である。さらに，クラスター計算機による10 CPUでの実装を行った。

$$f=\frac{C_p\cdot e^{\frac{\Delta G_T}{RT}}}{1+C_p\cdot e^{\frac{\Delta G_T}{RT}}} \tag{1}$$

また，オフターゲット遺伝子に対するクロスハイブリダイゼーションを避けるため，ハイブリダイゼーション率を計算し，ターゲット遺伝子に99%結合した場合の温度における５%以上の結合率を有するオフターゲット遺伝子をミスハイブリダイゼーションと定義し，このミスハイブリダイゼーション遺伝子数を許可されている数以下になるようにするシステムを開発した（図３）。

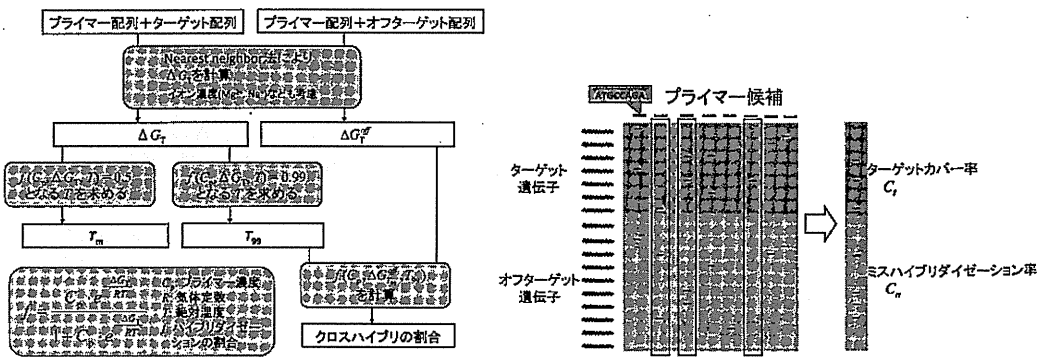


図３　プライマーのターゲット特異性の計算方法（左）とプライマーの組み合わせによるターゲッティングを行うシステムの概念図（右）

9.3.3　多重プライマーセット探索

多重プライマーセットの設計は，複数のプライマーの組み合わせを限られた条件で探索する問題であり，組み合わせ最適化問題に属するため，情報科学でよく用いられるシミュレーテッドアニーリングアルゴリズムを用いて準最適解を高速に求める方法を考案した。実際のシーケンサー解析では一度に加えることのできるプライマー数は少ない方がよいと考えられるため，ユーザーが与えたプライマー数で複数の遺伝子へ結合冗長性を許した組み合わせを解くプログラムを開発した。さらに問題を一般化し，「ターゲットにしているt個の遺伝子についてK個以下のプライマーで冗長性を許しながらn個のオフターゲット遺伝子との相対結合比を最大化する」問題として以下のプロトコルを確立した[5]。

プロトコル

①ターゲットとオフターゲット遺伝子に対するプライマー候補全てのハイブリダイゼーションポテンシャル行列（HP行列）を作成する。具体的には，各行にターゲット遺伝子とオフターゲット遺伝子を，各列にユーザーが与えたT_m値を満たすプライマー候補を持つ表を作成し，9.3.2項に定義したようにミスハイブリダイゼーションを計算して，結合を１とし非結合を０としておく（図３右)。

②初期プライマーセットとしてユーザーが与えた個数のプライマー列のみ上記①のHP行列から選択する。

③次ステップでのプライマーセットは，シミュレーテッドアニーリングアルゴリズムにより，最適化関数のスコアを最大化していくように選択される。この過程は時間を経るにつれて，ターゲットの結合率を増加させ，オフターゲットの結合率を減少させることになる。

多重プライマーセットの設計でもう一つ考慮しなくてはならないことにプライマー同士の会合による凝集がある。これを避けるため，プライマー間で会合するものに１を入れた表をもう１つ作り，これを衝突行列として利用することで，会合する可能性のあるプライマー同士は同時に選択できないことにする。従って，次のステップでは，

④選択したプライマー候補間で衝突を調べ，会合する可能性がある場合はそのプライマーセットは選択せずに捨てる。

ことが必要となる。

実際のシミュレーテッドアニーリングではステップ③と④を最適化関数がある閾値に達するまで何度も繰り返す。最適化関数も様々であるが，我々が幾つか検証した結果，線形式である

$s = C_t - C_n$

が良い精度を示す評価関数であることがわかった。ここで，C_tはターゲットのカバー率でC_nはオフターゲットへのミスハイブリダイゼーション率であり，最適化関数のsは次の様に最大化する。

①置き換える現在のプライマーを１つ選択し，時間tにおけるスコアs_tを計算する。

②新プライマー候補iについて時間$t+1$における$s_{t+1,i}$を計算する。

③スコア増加量Δs=min(１,exp{$(s_{t+1,i}-s_t)/T_s$}) に比例する確率でその新プライマー候補iを選択する。ここで，T_sはΔs＜１の時にT_s=T_s/1.05で減少する温度パラメータである。

シミュレーテッドアニーリングの性質に，「温度パラメータの冷却スケジューリングをうまく調整することで最適解を得る確率が１に近づく」ことが知られている。今回のスケジューリングパラメータも様々な試行錯誤の末に得られたものであることを補足しておく。

9.4　多重プライマーセットの（準）最適解探索結果

シミュレーテッドアニーリングによる効率的なプライマーセット探索と比較するため，常に最も高いスコア差である（$s_{t+1,i}-s_t$）を採択するものを貪欲法として同時に実験した（図４）。ここでは，人工データを作成し，70/100と５/400がそれぞれターゲットとオフターゲットに結合する最適プライマーセットの正解数とした（図４左）。さらに，より現実的なデータとして，Ensemblデータベースから1,000ずつのターゲットとオフターゲット（ハウスキーピング）遺伝子を抜き出して実験を行った（図４右）。なお，この実験ではHP行列作成の際に，プライマー

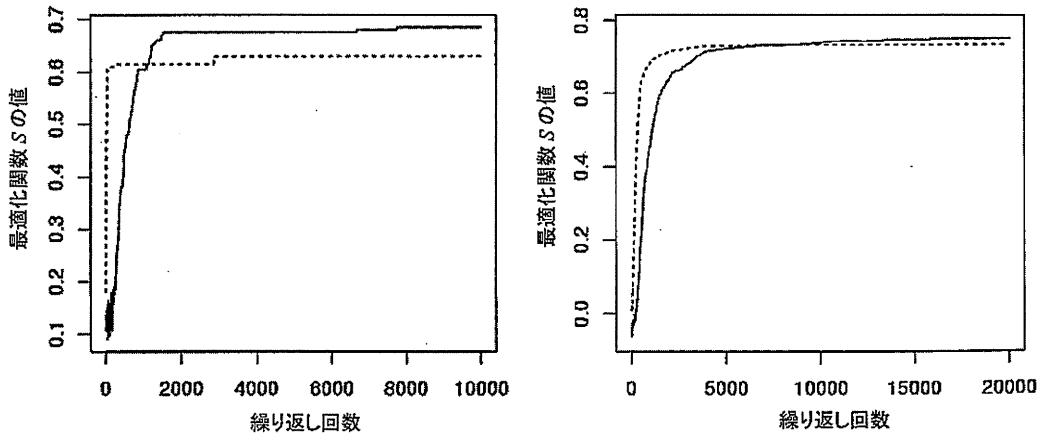


図４　人工データ（左）とEnsemblデータベースからの実際のデータを用いた場合（右）の最適化スコアの変化

実線がシミュレーテッドアニーリングによる探索で点線は対象とした貪欲法によるもの。

とオフターゲットが完全一致する場合を結合（1），それ以外を非結合（0）として計算を行った。

人工データでは7,000回の繰り返しで埋め込んでおいた最適解のs=70/100－5/400=0.6875を見つけた後，さらに良い解（s=0.6975）が存在しこれを発見した。Ensemblデータでは，長さ9 merのプライマー150個を用いて20℃≤T_m≤30℃の制約の下1,000ずつのターゲット／オフターゲット遺伝子を最適化させたところ，貪欲法ではs=817/1,000－83/1,000=0.734しか見つからなかったが，シミュレーテッドアニーリングだとさらに良い解のs=853/1,000－103/1,000=0.750が発見できた。

9.5　RT-PCR実験によるプライマーの検証

多重プライマーではないが我々の方法で厳密に設計した単独プライマーの性能を図るため，RT-PCR実験による検証を行った。ヒト乳がん由来MCF-7細胞およびマウス初期胚細胞を用いて，発現していると考えられる遺伝子を数個～10個程度選択し，プライマーを設計した。初めに実験を行ったところ，T_m値の依存度がかなり大きいことがわかり，再度プライマー設計システムを用いて配列のT_mをさらに絞り込み調整することで検出能力の高いプライマーが得られた。

図5では，ヒト乳がん細胞で発現していると考えられている遺伝子3種についてRT-PCRを行ったところ，目的とするmRNAのアンプリコンサイズのみが増幅されていることが検証された。この例では，58～68℃程度のアニーリング温度の幅で検出できている。また，マウス初期胚で重要と考えられる転写因子10種を選定し，このプライマーを設計してURR（Universal reference RNA）とマウス初期胚とで検出を比較したところ，どのURRに対しても強いシングルバンドが得られ，さらに実際の胚から抽出したRNAを用いて3転写因子で検証したところ，

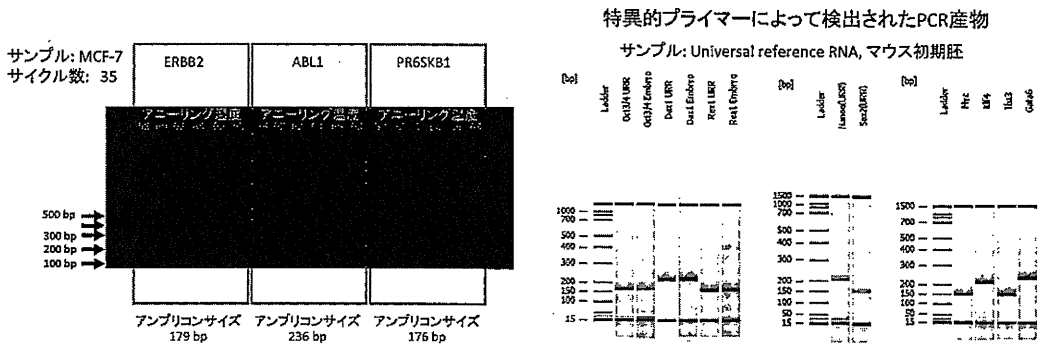


図5　ヒトMCF-7においてERBB2，ABL1，PR6SKB1についてPCRを行った結果（左）。目的としているアンプリコンサイズがはっきりとバンドで得られた。マウスの初期胚からのRNAとURRとで同じ転写因子RNA産物であるOct3/4，Dax1，Rex1を増幅した結果(右)。同じ位置にバンドが得られている。それ以外の転写因子ではURRでの結果になっている。

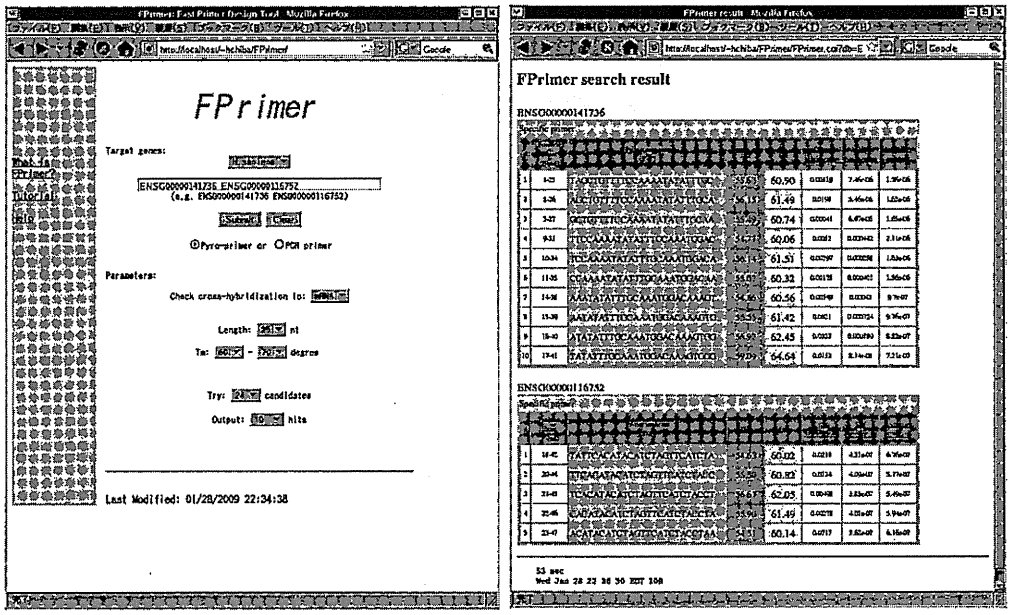


図6　遺伝子名（複数）とT_m値を入力するとFPGAボードを起動してユニークなプライマー配列セット候補のリストを返すWebシステム「FPrimer」のプロトタイプ

そのバンド位置もURRのものと一致した。

9.6　自動検索システム

このプライマー探索法については，簡易にユーザーに利用してもらえるようにWebインタラクティブ版のシステムを開発している（図6）。遺伝子名とプライマー長およびT_m値の条件を入力することで各cDNAの3'端から1,000 bp上流までのプライマー配列候補を作成し，FPGAボードを利用してそのユニーク性を検索し，さらにhybrid-minを起動してT_m値を計算して最適なものから順にソートしたものを表示することができる。

9.7　おわりに

本節では高速シーケンサーの高性能プライマー設計を迅速かつ効率的に行う方法を紹介したが，内容については中国で行われた国際ポストゲノムフォーラムで発表した[6]。本研究を行うに当たり㈱日立製作所　中央研究所の神原秀記先生，早稲田大学の竹山春子教授および岡村好子准教授に多大なご助言およびご助力をいただきましたこと，ここに感謝の意を述べたいと思います。本研究は文部科学省特定領域研究「ライフサーベイヤ」の助成を受けている。

文　　献

1) H. Chiba, S. Nakagawa, T. Taniguchi, W. Fujibuchi, *IPSJ SIG Technical Report*, **2008**, 1 (2008)
2) J. SantaLucia Jr., *Proc. Natl. Acad. Sci. USA*, **95**, 1460 (1998)
3) NR. Markham, M. Zuker, UNAFold: software for nucleic acid folding and hybridization in JM. Keith editor, Bioinformatics, Volume II. Structure, Functions and Applications (Methods in Molecular Biology), chapter 1, p.3 (2008)
4) S. Kirkpatrick, C. D. Gelatt Jr., M. P. Vecchi, *Science*, **220**, 671 (1983)
5) 知的財産特許：藤渕航，千葉啓和「プライマーセット探索装置，方法およびプログラム」特願2009-212703 (2009)
6) W. Fujibuchi, H. Chiba, H. Akiyama and H. Shiku, Proceedings of the 6'th International Forum on Post-genome Technologies, p.253, Beijing (2009)

14
Kernel Classification Methods for Cancer Microarray Data

Tsuyoshi Kato and Wataru Fujibuchi

14.1
Introduction

Cancer is one of the most malignant diseases affecting almost all tissues of all people of all ages and arising from a group of cells that grow uncontrollably from the normal state. More precisely, a group of cells that show only abnormal but controlled or limited growth is called benign tumor, while cancer refers to malignant tumor cells that show unlimited growth, usually invading other tissues directly or by spreading to distant locations in the body via lymph or blood. The spread of cancer cells is called metastasis and the cells are called metastatic cells, which are considered to be the worst malignancy, leading to high mortality rates. Thus, predicting the state of cancer, that is, whether it is metastatic or not, from specimens is one of the most important studies in cancer diagnosis.

Since the invention of gene expression microarrays in the mid-1990s, classification analyses based on gene expression data from distinct biological groups have become a fundamental approach in various cancer/tumor studies, such as tumor diagnosis [16, 34], anticancer drug response analysis [32, 44], and prognosis analysis [25, 50]. Among various classification methods, kernel-based methods [13] have played important roles in such disease analyses, especially when classifying data with support vector machines (SVMs) [51] by weighting feature or marker genes that are correlated with the characteristics of the groups. In most of those studies, only standard kernels, such as linear, polynomial, and RBF (radial basis function), which take vector data as input and basically convert them into inner-products between vectors, have been popularly used and are generally successful.

Most importantly, however, in microarray analysis designed for cancer study, one of the main issues that limit accurate and practical predictions is the lack of repeat experiments, often due to financial problems or rarity of specimens, such as minor diseases, as well as too much variability of cell types. Some gene expression databases contain disease microarray data (e.g., GEO [2], ArrayExpress [7], and Oncomine [36]) and the use of public or old data together with one's current data could solve this problem; many studies combining several microarray datasets have been

Medical Biostatistics for Complex Diseases. Edited by Frank Emmert-Streib and Matthias Dehmer

ISBN: 978-3-527-32585-6

performed [29, 35, 53]. Nevertheless, due to the insufficient amount of gene overlaps and consistencies between different datasets, kernels that use vector data as the primary input are often unsuccessful in classifying data from various datasets if naïvely integrated [53].

Instead of the above *vectorial data kernel* family, there is another family called *structured data kernel* family that has been studied in many other fields, including bioinformatics and machine learning [23, 26, 48, 49]. Among them, the synthetic distance-based kernels, or what we call *metrization kernels*, can take any distance data between sample vectors (or samples in short) as primary input without recognizing the original vectorial data from which the distance is calculated while holding positive semidefiniteness of kernel matrices, and is thus applicable to the Euclidean or other distance measures among sample vectors once converted into a distance relationship. Moreover, the metrization kernels have, unlike the RBF kernel, the special property of excluding arbitrary gene values in vectorial data when calculating the distances among samples. Hence, by ignoring only spurious gene values in distinct samples without deleting those genes entirely from a dataset, the metrization kernel can effectively utilize gene expression information in heterogeneous data containing mosaic-like missing or noisy values.

In this chapter, we first describe the general mechanisms of machine learning by kernel methods and SVMs, comparing the properties of standard and metrization kernels as well as referring to two noise handling methods in microarray data. Then, we demonstrate a few machine classification examples using kernel-SVM methods for cancer microarray data, together with different noise-reduction methods, to learn practical issues in handling disease datasets that are noisy and promiscuous. The proofs of Theorems 14.1 and 14.2 are given in Appendix 14.A

14.1.1 Notation

Vectors are denoted by boldface italic lower-case letters and matrices by boldface italic upper-case letters. The transposition of matrix A is denoted by A^T, and the inverse of A is denoted by A^{-1}. The $n \times n$ identity matrix is denoted by I_n. We use E_{ij} to denote a matrix in which (i,j) element is one and all the other elements are zero. The n-dimensional column vector all of whose elements are one is denoted by $\boldsymbol{1}_n$. We use $\mathbb{R}$ to denote a set of real numbers, $\mathbb{R}^n$ to denote a set of n-dimensional real column vectors, and $\mathbb{R}^{m\times n}$ to denote a set of $m \times n$ real matrices. The set of real non-negative numbers is denoted by $\mathbb{R}_+$, and the set of n-dimensional real non-negative vectors is denoted by $\mathbb{R}^n_+$. We use $\mathbb{S}^n$ to denote a set of symmetric $n \times n$ matrices, $\mathbb{S}^n_+$ to denote a set of symmetric *positive semidefinite* $n \times n$ matrices, and $\mathbb{S}^n_{++}$ to denote a set of symmetric *strictly positive definite* $n \times n$ matrices. We will define positive semidefiniteness and strictly positive definiteness later. $\mathbb{N}$ is a set of natural numbers. $\mathbb{N}_n$ is a subset of $\mathbb{N}$, and is defined by $\mathbb{N}_n \equiv \{i \in \mathbb{N} | i \leq n\}$. Symbols $\leq$ and $\geq$ are used to denote not only the standard inequalities between scalars but also the componentwise inequalities between vectors. Finally, $\langle \cdot, \cdot \rangle$ is the operator of inner-product among vectors.

14.2
Support Vector Machines and Kernels

This section reviews the support vector machine (SVM; e.g., [14]). Nowadays we can find lots of tutorials and introductions about SVM elsewhere [8, 9, 13, 17, 41, 52]. Most of the tutorials describe SVM as a *large margin classifier;* SVM finds a *hyperplane* with the largest margin between two *classes* to determine the classification boundary. Here we attempt to introduce SVM with a different explanation that is based on the literature [3, 12].

14.2.1
Support Vector Machines

SVM is basically a framework that automatically learns a linear classifier to distinguish a positive class from a negative class. For learning, we need a dataset. The dataset used for learning is called a training dataset. Each sample in the dataset has a binary label, $+1$ or -1. In the later section we discuss a case where SVM learns the classifier that discriminates the data of human kidney of normal tissues from those of renal clear carcinoma tissues. In this case, we assign the positive label $+1$ to the normal tissues, and the negative label -1 to the carcinoma tissues. Each sample is represented by a fixed-length vector $\boldsymbol{x}$ often called an input vector. In the case of classification of microarray data, an input vector typically consists of gene expression values. For example, if we use the expression data of d genes, the length of the input vector is d, that is, $\boldsymbol{x} \in \mathbb{R}^d$.

The SVM classifier is a score function of input data. After training the SVM classifier, we compute the score of the data with unknown labels. Unlabeled samples are classified by examining whether the score is greater than a threshold. The threshold is often set to zero. The boundary that distinguishes the positive class from the negative one is a hyperplane. This is because the score function of SVM is a linear function expressed as:

$$f(\boldsymbol{x}; \boldsymbol{w}, b) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b,$$

where $\boldsymbol{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are the model parameters of SVM. The score contains a confidence level; a larger value will be a confident prediction of being in a positive class.

Let us consider how to determine the parameters of the classification hyperplane, $(\boldsymbol{w}, b)$. To learn the normal vector automatically, we usually gather training samples first:

$$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell) \in \mathbb{R}^d \times \{\pm 1\}.$$

We then compute some statistics from the samples to determine $\boldsymbol{w}$. If we have computed the means of the two classes, $\boldsymbol{m}_+ \in \mathbb{R}^d$ and $\boldsymbol{m}_- \in \mathbb{R}^d$, one of the simplest approaches to classification is to classify a new sample $\boldsymbol{x}$ to the class whose mean is closer. The classification boundary of this approach is the hyperplane that is orthogonal to the line segment between $\boldsymbol{m}_+$ and $\boldsymbol{m}_-$ and bisects the line segment

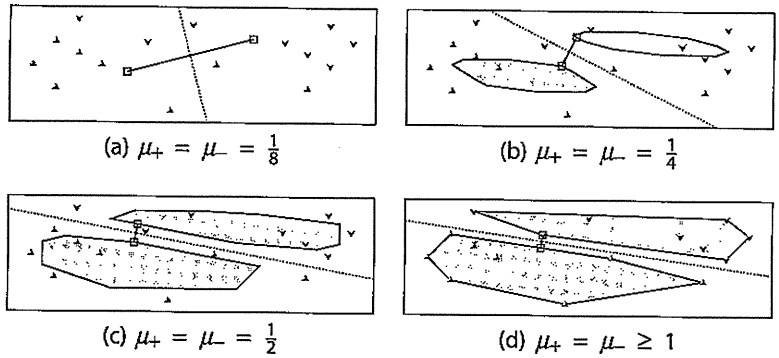


Figure 14.1 Geometrical interpretation of SVM. Positive and negative samples in $\mathbb{R}^2$ are plotted by upward- and downward-pointing triangles, respectively. (a) Shown is the classification boundary designed in a simple way; the boundary bisects the line segment between the means of two classes. (d) The closest points of the two convex hulls are depicted by squares. The convex hulls are obtained by setting $\mu_+ = \mu_- = 1$. Varying the values yields different convex sets and leads to different classification boundaries, as shown in (b)–(d). When $\mu_+ = \mu_-$, all the boundaries can also be produced by SVM with a suitable choice of C.

(Figure 14.1a). The normal vector is obtained from the difference between two points, m_+ and m_-:

$$w = m_+ - m_-$$

and the offset is computed by:

$$b = \frac{1}{2}(\|m_+\|^2 + \|m_-\|^2).$$

Note that the mean of the positive class is close to the new sample if and only if the score is positive. Although the approach is very simple and intuitive, it does not give consideration to how the samples are distributed. We now generalize the simple approach in order to consider the distribution. Let us consider two vector sets, $\mathcal{V}_+$ and $\mathcal{V}_-$, that include the mean of the corresponding class, respectively (i.e., $m_+ \in \mathcal{V}_+$ and $m_- \in \mathcal{V}_-$). We first find the geometrically closest points between the two sets, $v_+ \in \mathcal{V}_+$ and $v_- \in \mathcal{V}_-$, and then construct the hyperplane by:

$$w = v_+ - v_-, \qquad b = \frac{1}{2}(\|v_+\|^2 + \|v_-\|^2).$$

The hyperplane bisects the shortest line connecting the two sets. The simple approach we have introduced first is a special case of the second one in which $\mathcal{V}_+ = \{m_+\}$ and $\mathcal{V}_- = \{m_-\}$. The two closest points can be expressed as the solution of the following minimization problem:

$$\min \quad \|v_+ - v_-\|^2 \qquad \text{wrt} \quad v_+ \in \mathcal{V}_+ \quad \text{and} \quad v_- \in \mathcal{V}_-. \tag{14.1}$$

Let us denote the index sets of the positive training set and the negative training set by: