

N48-02-0068

(K)

平成9年度 先導研究報告書

NEDO-PR-9703

ゲノム情報解読利用技術の調査研究

RECEIVED
AUG 13 1998
OSTI

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED
FOREIGN SALES PROHIBITED
NT

平成10年3月

新エネルギー・産業技術総合開発機構
委託先 財団法人バイオインダストリー協会

DISCLAIMER

Portions of this document may be illegible electronic image products. Images are produced from the best available original document.

(調査項目) ゲノム情報解読利用技術の調査研究

(調査委託先名) 財団法人バイオインダストリー協会

(報告書作成年月日) 平成10年3月 (頁数) 149

(調査目的)

ゲノムに存在する遺伝子の機能と発現制御機構、そしてそれらの相互作用ネットワークを解明することにより細胞機能の全体像を明らかにし、得られる有用な細胞やタンパク質機能を産業面に応用するため、生物学と情報科学の両分野を含めた視点から調査することを目的とした。

ゲノム情報解読利用技術の調査研究調査報告書

(平成9年度)

目 次

	執筆担当 (敬称略)	頁数
まえがき／委員会名簿		
I. 概要 本田皓一	
英文		[1]
和文		[12]
II. 調査研究の詳細		
第1章 はじめに 榊佳之	1
第2章 変異・多型部位の検出・取得技術	神野英毅	5
2-1 DNA チップ 狩野恭一	5
2-1-1 目的と役割		5
2-1-2 技術開発		5
2-1-3 企業・機関情報		6
2-1-4 DNA チップの課題		16
2-2 ゲノム全体の比較解析法 有澤章, 縣直樹	19
2-2-1 Restriction Landmark Genomic Scanning (RLGS) 法		19
2-2-2 Arbitrarily Primed PCR (AP-PCR) 法		21
2-2-3 In-Gel Competitive Reassociation (IGCR) 法		23
2-2-4 Representational Difference Analysis (RDA) 法		25
2-2-5 ミスマッチ特異的結合タンパク質 MutS を用いた方法		27
2-3 特定領域遺伝子の比較解析 大島淳	31
2-3-1 序文		31
2-3-2 PCR を基にした様々な解析法		31
2-3-3 課題と将来性		34
2-4 遺伝子増幅法の多様化 神野英毅	34

2-4-1	遺伝子增幅法の現状	34
2-4-2	臨床検査で利用されている遺伝子検出法	34
2-4-3	各方法論の比較検討	39
2-4-4	遺伝子診断の今後の展開	40
第3章	特異的発現様式解析技術 榎佳之・伊藤隆司(東京大学) 41
3-1	DNAチップによるもの 今村亨・長谷川幸雄・ 41
3-1-1	Oligonucleotide array	後藤雅式(アマシャム 41
3-1-2	cDNA array	ファルマシアバイオテク) 42
3-1-3	その他の microarray	43
3-1-4	今後の展望	46
3-2	cDNA Sequencing によるもの 扇谷悟 48
3-2-1	ボディマップ	49
3-2-2	SAGE法	52
3-2-3	MPSS法	54
3-3	メッセージディスプレイによるもの	56
3-3-1	Differential Display法 藤原力 56
3-3-2	Molecular Index法 今村亨・長谷川幸雄・後藤雅式 58
第4章	発現調節ネットワーク解析技術 町田雅之 60
4-1	転写制御シグナルの探索・解析技術	61
4-1-1	シークエンス比較解析技術 町田雅之 62
4-1-2	転写制御領域の実験的解析技術 町田雅之 64
4-1-3	特定転写制御因子の制御領域分取・解析技術	66
	 丹羽真一郎
4-2	転写制御因子の探索・解析技術	68
4-2-1	DNA-タンパク質間相互作用を用いた解析技術	69
	 丹羽真一郎・町田雅之
4-2-2	タンパク質-タンパク質間相互作用を利用した解析技術	73
		... 丹羽真一郎

第5章	コンピュータモデル解析技術	85
5-1	遺伝子領域・機能予測技術 …… 浅井潔	85
5-1-1	原核生物の遺伝子領域予測技術 …… 市吉伸行	85
5-1-2	真核生物の遺伝子領域予測技術	87
5-1-3	コード領域の配列情報に基づいた遺伝子機能予測技術 …… 松浦幸男	89
5-1-4	制御領域の配列情報に基づいた遺伝子機能予測技術 …… 市吉伸行	90
5-2	遺伝子発現制御シミュレーション …… 浅井潔	95
5-3	データベース・クラスタリング技術 …… 永井啓一	97
5-3-1	データベース技術	97
5-3-2	クラスタリング技術	104
第6章	国際動向	
6-1	国際シンポジウム報告 …… 岡村和彦	107
6-2	ゲノム関連バイオベンチャーの動向 …… 大滝義博 (株)ジャフコ	115
6-2-1	はじめに	115
6-2-2	第1世代ゲノム・ベンチャー企業の動向	115
6-2-3	第2世代ゲノム関連ベンチャー企業の動向	117
6-2-4	バイオ・インフォーマティクス	118
6-2-5	動物系を用いた遺伝子機能解析	119
6-2-6	ツールの開発	119
6-2-7	Pharmacogenetics	120
6-2-8	まとめ (ポスト・ゲノムの創薬)	120
6-3	ゲノム情報解析とコンピュータ構造解析 —欧米の動向— …… 八尾徹 (三菱化学株)	122
6-3-1	はじめに	122
6-3-2	欧米の構造生物学の施策	123
6-3-3	国際学会・シンポジウムでの話題	123
6-3-4	欧米研究機関の動向	124

第7章 まとめと今後の進め方への提言 高木利久 127

第8章 むすび 岡村和彦 131

III. 付録（参考資料）

通商産業省 工業技術院 生命工学工業技術研究所 平成9年度研究報告 132

通商産業省 工業技術院 電子技術総合研究所 平成9年度研究報告 135

通商産業省 製品評価技術センター バイオテクノロジーセンター 平成9年度調査報告 138

以下に示す。

本調査研究における点、多くは現在まで遺伝子の機能を表現制御機構、そしてその相互作用について多くの説明を行う。特に、機能的DNA塩基配列における機能分子の作用を目的とした生物学的調査研究を行った。得られた有用な知識により、より質機能を座標面に応用する上での全体像を明らかにするため、生物化学・遺伝学・情報科学の面分野を含めた複数の研究を行った。特徴的な機能的DNA塩基配列が一つの基礎となるが、この機能的DNA塩基配列が分子生物学・生物化学・遺伝学等の各分野で重要な役割を果たす。また、分子生物学・生物化学・遺伝学等の各分野で重要な役割を果たす。

發現DNA的解旋的第一步就是DNA的複基配列。醫學上重要的是遺傳子的差異為分子生物學。發現DNA的複基配列之後，生物學家們開始研究DNA的複基配列的性質。這些性質包括DNA的複基配列的對稱性、DNA的複基配列的穩定性和DNA的複基配列的可變性。DNA的複基配列的對稱性是指DNA的複基配列在兩端是對稱的。DNA的複基配列的穩定性是指DNA的複基配列在溫度和酸鹼度的變化下不會改變。DNA的複基配列的可變性是指DNA的複基配列在某些化學反應中會發生改變。

本報告書は、新工具等一・座業技術総合開発機構（NEDO）が主に委託したもの、財
バーチナルデータベース（JBA）及び、昨年度刊行された、平成9年度実施ルート／人情
報解説利用技術の調査研究」の成果を取まとめたものです。

能になることを期待したい。

最後に、本調査にご協力いただいた多くの関係者の方々に、心より御礼申し上げたい。

平成10年3月

(財) バイオインダストリー協会

ゲノム情報解読利用技術の調査研究平成9年度調査委員会名簿

委員長：榎 佳之 東京大学医科学研究所 ヒトゲノム解析センター教授
委員：本田 翔一 生命工学工業技術研究所 分子生物部長
町田 雅之 生命工学工業技術研究所 分子生物部遺伝子工学研究室 町田雅之グループ
今村 亨 生命工学工業技術研究所 生体情報部今村亨グループ グループリーダー¹
浅井 潔 電子技術総合研究所 知能情報部遺伝子情報ラボ 主任研究官
高木 利久 東京大学医科学研究所 ヒトゲノム解析センター教授
永井 啓一 (株) 日立製作所中央研究所バイオシステム研究部 主任研究員
松浦 幸男 富士通(株) 科学システム統括部 バイオ担当部長
有澤 章 メルシャン(株) 生産技術センター 研究主任
縣 直樹 メルシャン(株) 中央研究所第二研究室 室長
市吉 伸行 (株) 三菱総合研究所 システム科学部第二室 室長 主任研究員
扇谷 悟 北海道工業技術研究所 低温生物化学部生物化学研究室主任研究官
大島 淳 宝酒造(株) 遺伝子解析センター
小林 利克 三菱化学生命科学研究所 副所長
柳沢 一向 オリンパス光学工業(株) 技術開発本部 複合精密技術部部長
狩野 恭一 オリンパス光学工業(株) バイオメディカルリサーチセンター 所長
具嶋 弘 山之内製薬(株) 創薬研究本部研究推進部 理事
藤原 力 大塚製薬(株) GEN研究所 室長
神野 英毅 三菱化学(株) 診断事業部部長

調査委員会関係者：

大窪 道章 通産省工業技術院研究開発室 生体機能応用技術企画官
大西 啓仁 通産省工業技術院研究開発室 研究開発専門官
新聞 陽一 通産省工業技術院研究開発室 研究官

青柳 昌宏 通産省機械情報産業局 電子機器課解析技術課長
舟橋 友道 通産省製品評価技術センター バイオテクノロジーセンター解析
技術課長
長田 敏 通産省製品評価技術センター バイオテクノロジーセンター
吉澤多喜男 通産省製品評価技術センター バイオテクノロジーセンター情報
業務課長
吉田修一郎 新エネルギー・産業技術総合開発機構 産業技術研究開発部
研究開発課総括
宮山 哲夫 新エネルギー・産業技術総合開発機構 産業技術研究開発部
研究開発課主査
堅尾 和夫 通産省基礎産業局 生物化学産業課長
岡田 武 通産省基礎産業局 生物化学産業課
本田 真也 通産省基礎産業局 生物化学産業課
大塚 理絵 通産省基礎産業局 生物化学産業課

オブザーバー：

松尾 次雄 バイオテクノロジー開発技術研究組合 専務理事
越石喜代三 バイオテクノロジー開発技術研究組合 担当部長
長谷川幸雄 アマシャムファルマシアバイオテク（株） 研究開発室長
丹羽真一郎 住友電気工業（株）大阪研究所 バイオメディカル研究部 部長
伊藤 隆司 東京大学医科学研究所ヒトゲノム解析センター助手
平野 昌彦 東レリサーチセンター生物科学研究所第1研究室室長
佐藤 恒雄 旭化成工業（株）ライフサイエンス基礎研究所
三沢 典彦 キリンビール（株）基盤技術研究所
平野 憲一 浜松ホトニクス
陶山 明 東京大学大学院総合文化研究科広域科学専攻生命環境科学系物理
学
嵯峨 直悟 東海大学海洋学部海洋科学科教授

委員会事務局：(財) バイオインダストリー協会

調査実施責任者：(財) バイオインダストリー協会 常務理事 木幡 守

調査実施研究員：(財) バイオインダストリー協会 技術部長 岡村 和彦
(財) バイオインダストリー協会 学術技術部 矢田 美恵子

I. 概 要

Technologies and Techniques for Analysis and Use of Genome Information, 1997

Summary

Thanks to successful integration of different DNA sequencing technologies in a most efficient process flow, large – scale genomic DNA sequencing has been under way. The complete genomic DNA sequences of some bacteria such as Escherichia coli and Bacillus subtilis, as well as baker's yeast, have already been determined. By the summer of 1998, the complete DNA sequence of the nematode genome will have also been determined. The Human Genome Project is making satisfactory progress and genome biology that elucidates the phenomena of life on the basis of genomic information is developing. Under the circumstances, our team conducted a comprehensive review of international trends of the state – of – the – art hardware and software technologies for efficiently and effectively extracting target genome function data from enormous amount of genomic DNA sequence data. Special emphasis was placed on the fields of comparative genome, gene expression profile and gene – gene (or protein – protein) interaction analyses. It is the purpose of this study to provide an important input for the efforts to improve and maintain Japan's competitiveness in this area.

Chapter 1 : Now that the international efforts of genomic DNA sequencing are picking up speed, faster DNA sequencing and more efficient functional analysis of sequenced genes through computer data processing have become more and more important. Also needed is the acceleration of biological analysis of genomic DNA sequences by linking them to phenotypes and functions that are expressed. To this end, Chapter 1 calls for the development of : (1) the resources required for analysis and the techniques

to prepare a variety of species samples ; (2) the tools to analyze the differences in DNA sequences and in the kinetics of their expressions ; and (3) large - scale DNA database that accommodate a rapidly increasing number of DNA information and data processing technologies. Chapter 1 also calls for the development of advanced analytical techniques that ensure the quality and quantity of genome information, the compilation of useful information according to the objectives of use, and development of computer and information science technologies to make use of such genome information in order to improve and maintain the competitiveness of Japan's bioindustry.

Chapter 2 reviews the technologies for detecting and extracting the mutation and polymorphic sites on genome. These technologies are expected to be used in gene diagnosis and gene function analysis. Chapter 2 discusses there technologies in the three major areas, DNA chips, comparative analysis of genome as a whole, and comparative analysis of genes in particular regions.

(1) A DNA chip has a variety of oligonucleotides integrated on it. Each of them hybridizes a particular DNA sequence. The pattern of hybridization is analyzed to detect a particular mutation or polymorphic site. Thus, the trends in manufacturing, pattern analysis and microfabrication technologies were reviewed. Also reviewed were the latest information about the technologies developed by Affimertrix and other organizations who are the technology leaders in this field. The results were compiled in a technology map. This section also discussed such issues as cost reduction and automation of microtechnologies.

(2) As a possible approach toward comparative analysis of entire genomes, we selected the separation of particular DNA fragments by gel

electrophoresis. Among those applicable even the genomic DNA sequence is unknown, this section discusses the restriction landmark genomic scanning (RLGS) and the arbitrarily primed PCR (AP - PCR) methods that analyze genomic DNA sequences as fingerprints, the representational difference analysis (RDA) and the in - gel competitive reassociation (IGCR) methods that detect a particular DNA by subtracting one set of genome from another, as well as those based on a mismatch repair enzyme (MutS). Their developmental stages, principles, possible applications, trends and related technologies, and challenges are reported.

(3) Four methods for comparative analysis of genes in particular regions were selected : the single strand conformation polymorphism (SSCP) ; the denaturing gradient gel electrophoresis (DGGE) ; the restriction fragment length polymorphism (RFLP) ; and the cleavase fragment length polymorphism (GFLP) methods. They are all based on the PCR amplification of target genes and the analysis of gel electrophoresis separation patterns. They are also superior to other methods in terms of lower cost, easier handling, and higher productivity. Their outlines, prospects for automation, and reproducibility are discussed.

(4) As alternative gene amplification methods that do not infringe the PCR patent held by Roche and is currently used in clinical testing, four methods were selected and their technologies are outlined. They are the gap ligase chain reaction (Gap LCR) method based on DNA ligase, the transcription mediated amplification (TAM) method for RNA amplification, the nucleic acid sequence - based amplification (NASBA) method, and the cycling probe technology (CPT) method which use the RNase degradation of chemically synthesized DNA - RNA - DNA sequences. Alternative technologies including non - gene amplification technologies that are available on the gene diagnosis market are compared and their

developers are positioned on the market.

Chapter 3 reviewed the technologies for analyzing specific expression modes that are needed to clarify the profiles of gene expressions. The technologies are divided into three groups : (1) DNA chip – based ; (2) cDNA sequencing – based ; and (3) message display – based.

(1) Selected DNA chip – based technologies include the oligonucleotide array to analyze gene expressions that was not mentioned in Chapter 2, the cDNA array with synthesized cDNA placed on a chip, the one based on an electric field, and the atlas array of low cost and high efficiency. Discussed are the trends in the development activities, the moves of Affimetrix and other leading companies toward standardizing array technologies, the attempts by related U.S. firms of controlling the genome industry through M & A and business tie – ups and the need for Japan's original technologies, the trends of protein chip developments.

(2) Concerning cDNA sequencing – based technologies, and as an example of research activities aimed at analyzing the profiles of gene expressions through identification of the sequence of expressed mRNA molecules and the analysis of appearance frequency, the research project of Osaka University to build a database of mRNA expressed in human organs (body map). Also discussed in terms of technical outlines and future prospects are : the serial analysis of gene expression (SAGE) method that improves determination of gene expression profile and identification of mRNA molecules by modifying a technique of cDNA sequence analysis ; the mini – fragment analysis of gene expression (MAGE) method which is an improved version of the SAGE method ; and the massive parallel signature sequencing (MPSS) method which is a new technique to elucidate mRNA profiles on the basis of rapid base sequence

analysis.

(3) As examples of message display – based approaches, the differential display (DD) method are presented together with its improved version developed by the author team. These methods analyze specific mRNA messages through simultaneous comparison of multiple and comparative mRNA molecules. The molecular index (MI) method is outlined as a promising method to identify entire genes that are expressed in a particular tissue or cell.

Chapter 4 discusses, in detail ,the techniques to analyze expression regulation networks, dividing them into two groups : (1) the techniques to search and analyze transcription regulatory signals ; and (2) the techniques to search and analyze transcription regulatory factors. These techniques are required to elucidate the mechanism of transcription regulation, which is the basic foundation of gene expression regulation, at the molecular level.

(1) Concerning the techniques to search and analyze transcription regulatory signals, a method to search promoter regions based on the prediction of structural gene region (gene finding) and on the oligocapping method for cloning 5' – terminal, as an example of techniques to compare and analyze sequences. Collecting basic information about the detailed recognition regions and regulatory functions for each transcription regulatory factor is critical in order to establish a method to predict the structural gene regions on the basis of base sequence data. The features of representative reporter gene methods, as an experimental technology to analyze transcription regulatory regions, are tabulated. Representative in vitro analytical methods that use the property of transcription factors to bind specifically the transcription regulatory

regions, are presented. As techniques to separate and analyze the regulatory region for a particular transcription regulatory factor, the genomic binding site cloning technique, which is based on filter - separation of factor - DNA complex, and the casting method, which separates the complex using the gel shift or immune precipitation.

(2) Among the techniques to search and analyze transcription regulatory factors, the southwestern method based on lambda phages, the one - hybrid method based on yeasts, and the phage display method which can obtain both a protein and its coding gene at the same time, are outlined as examples of analytical techniques based on the DNA - protein interactions. These methods are also discussed in terms of their future prospects as techniques to acquire transcription regulatory factors. Another group of analytical techniques based on the protein - protein interactions are also discussed in terms of their principles and pitfalls. They are the far western method that replaces the probe DNA with a labeled protein, the protein - immobilized column method which is an affinity chromatography that uses transcription regulatory factors immobilized on beads as ligands, and the tagged protein techniques based on the transcription regulatory factors to which affinity - bonding tag sequences are bound. Furthermore, in the discussion of other techniques to search transcription regulatory factors, this section calls for the need to clarify the relationship between the transcription regulation of cell - specific genes (differentiation marker genes) and the cell functions in the manner of multivariate analysis including time and space axes, taking the transcription factor cascades in the cell differentiation of the mesodermal system. This section also emphasize the importance of modifying reactions such as phosphorylation, transcription - mediating factors (cofactors), as well as selective splicing which also regulates gene expression in multicellular organisms outside the transcription process.

Chapter 5 focuses on the computer – based model analysis techniques that predicts particular gene regions and functions, and even gene networks based exclusively on base sequence data and without relying on experiments. It reviews the current status of techniques for (1) gene region /function prediction, (2) the computer simulation of gene expression regulation, and (3) database clustering. Pending problems of this post – genome sequence technologies are also discussed.

(1) Gene region/function prediction : As techniques to predict the gene regions of prokaryotes, GeneMark.hmm and GeneMark – Genesis which use the framework of the hidden Markov model (HMM) to resolve the problems that conventional techniques have faced, GeneHacker which incorporates the profile of the ribosomal binding site (RBS), and GLMMER which features high speed processing due to use of the Markov model with unspecified order, are presented. GENSCAN which applies generalized hidden Markov model (GHMM) has been used to predict the gene regions of eukaryotes. VEIL, HMMgene and GeneDecoder are based on the HMM. A new method is intended to predict the promoter regions of eukaryotes. Improvements are proposed for such conventional techniques as MORGAN based on the decision tree, Genie based on GHMM, GRAIL which applies the artificial neural network (ANN), and FGENEH based on linear discrimination. Future potential of a new prediction method called AMELIE is examined. This method does not rely on coding potential data such as the bias in frequencies of codon appearance or the periodicity of particular bases that appear in coding regions. The third group of techniques are intended to predict gene functions based on the DNA sequence data in the coding regions. Presented as a new technique of this group that does not rely on homology search, was the one that predicts the sites where particular proteins are localized based on the

information about the amino acid sequence within a signal peptide using an intelligent classifier that learns. It also predicts the enzymatic functions of proteins based on the physical, chemical and statistic values calculated from amino acid sequences. Additionally, the techniques to predict gene functions based on the information about the DNA sequence within the regulatory region, is presented. These techniques predict the expression conditions for downstream genes based on the combination of sigma factor dependencies of the promoter sequences from prokaryotic microorganisms and the preserved sequence fragments from eukaryote promoter. The results of the validity tests of these techniques conducted on computers is presented.

(2) Computer simulation of gene expression regulation: Even if a huge amount of information about gene functions and expressions is accumulated in the near future through proteome analysis and others, one has to take into account the interactions between thousands of genes in order to understand the network of gene functions as a system. This means that cell models must be developed on computers. Thus, we examined the experience of modeling and simulation on computers of fruit fly's segment formation and nematode's development process, as well as the expression of mycoplasma's entire genes. Although there are some problems with modeling and computation time, we tried to predict the possibility of establishing the cell simulation technology and hypothetical *in silico* experiment.

(3) Database clustering: The current status and pending problems of clustering techniques are discussed here. These techniques are intended to bring about, on a certain criteria, useful information from the databases and accumulated information that use the technologies and techniques we already mentioned in Chaper 2 – 4.

As a database on genetic mutation sites, we presented the Human Gene Mutation Database (HGMD) that integrates a variety of mutation databases for individual locus. As examples of the databases for the analysis of specific expression modes, we included the BodyMap and the Expressed Gene Anatomy Database (EGAD). These mRNA expression frequency databases were derived from large – scale human EST analysis. Also included is a database derived from the analysis of mRNA expression profiles by DNA chips. Among the databases on expression regulation networks, a database on expression regulatory factors called the Eukaryotic Promoter Database that holds the information about the promoters that are recognized by the RNA polymerase II of higher eukaryotes.

Other examples of expression regulatory factor databases are the Transcription Factor Database (TFD) which retains the data on the sequence and interrelations concerning the transcription of many eukaryotes, and TRANSFAC which is a database of eukaryote DNA elements and factors. As the database on gene networks, the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the Encyclopedia of *E. coli* Genes and Metabolism (EcoCye) which specialize in metabolic systems, and the Signaling Pathway Database (SPAD) which specialize in signal transformation paths.

As examples of the clustering databases based on the sequence information, we included the BodyMap and EGAD which compile information about mRNA expression frequencies according to individual tissues after clustering the EST sequences that are from the same genes but are registered in the different parts of the databases. Also presented is the UniGene database which offers unique sets of genes after clustering the human mRNA sequences and the EST sequences that are registered in the latest GenBank and dbEST.

Comparison of genome sequence between different organisms necessary to elucidate the phenomena of life, is mainly performed by homology – based clustering at present. Thus, the MicroBial Genome Database for comparative analysis (MBGD) is discussed in detail. This is a database workbench system operated by the Human Genome Center Institute of Medical Science, University of Tokyo, for mutual comparison of all genome sequences that have been determined by the institute.

Chapter 6 reviews the international trends in genome research activities. It reports the discussions at international symposiums held as part of leading study “Technologies and Techniques for Analysis and Use of Genome Information.” Chapter 6 also review the U.S. and European trends in the analysis of genomic information and structural analysis on computers.

At the symposium, nine domestic and foreign scientists gave lectures on the current status of genome interpretation, and techniques for efficient sequencing, and analyzing gene networks and genome information under a title “Genome Research Opens a New World to Bioindustry ? New Developments in Genome Informatics Technologies.” The outlines of their lectures are presented.

In the review of international trends, the report points out that the U.S. and European decision – makers, academia and other research organizations recognize the effectiveness of computers in the analysis of three – dimensional structures and prediction involving an enormous amount genome information. More protein analysts are entering the field of genome information analysis, more conventional protein analyzing systems are automated, and researchers are applying computer technologies to its full to build databases for an increasing amount of genomic DNA information.

A new focus is placed on "comparative genomics" which compares the genomes from different species and it is already yielding results. Charts illustrate the flow of computer screening of genome data and interrelationship among genome information analysis, structural biology and bioinformatics.

ゲノム情報解読利用技術の調査研究報告書(平成9年度)概要

DNAシークエンシング技術のシステム化、ライン化の成功により、大量ゲノムシークエンスデータ生産が可能となり、大腸菌、枯草菌等のバクテリア、及びパン酵母の全ゲノムシークエンスが決定され、続いて線虫ゲノムも今夏には全シークエンス決定が終了する。ヒトゲノム計画も順調に進展しており、ゲノム情報に基づいて生命現象を解明するゲノム生物学が発展しつつある。このような状況下にあって、大量ゲノムシークエンスデータからゲノム機能情報を効率的かつ効果的に抽出するための先端技術、特に比較ゲノム解析、発現プロファイル解析、遺伝子間（タンパク質間）相互作用解析に関するハード及びソフト両面についての国際的な動向の調査研究を行った。本先導研究における調査研究が、我が国この分野における国際競争力を強化を促すための一助になれば幸いである。

第1章では、大量ゲノムシークエンス決定時代に突入し、基本となるシークエンスデータ生産とともに、コンピュータ情報処理によるゲノム機能の解析の重要性が益々高まり、ゲノム配列と生物の示す表現型や機能との対応付けによる生物学的な解析データの取得の加速的推進が要求されること、そのためには、1) 解析リソース及び多様な生物種サンプルの整備手法、2) 配列の相違や遺伝子発現の動態の比較解析用ツール、3) 大量取得データのデータベース化と情報処理技術等の開発が必要であり、データの質と量を保証できる高度解析手法の開発と目的別有用データの蓄積及びそれを活用するための情報科学技術の開発が、我が国のバイオ産業の競争力を高める上で重要であることを提言した。

第2章では、遺伝子診断や遺伝子機能解析への応用が期待されているゲノム上の変異・多型部位の検出・取得技術の調査を行い、①DNAチップに関する技術情報、②ゲノム全体の比較解析法、③特定領域遺伝子の比較解析法、④遺伝子增幅法の多様化、についてまとめた。

①DNAチップは、チップ上に集積された多種多様のオリゴヌクレチドとのハイブリダイゼーションのパターン解析を原理とすることから、製造技術、パターン解析技術、マイクロファブリケーション技術の3項目に関する技術開発動向と現在市場をリードしているAffimetrix社をはじめ代表的な企業・機関の最新技術情報を調査するとともに技術マップにまとめ、ローコスト化やマイクロ自動化等の課題について論じた。

②ゲノム全体の比較解析法では、ゲル電気泳動によるDNA断片の分離を原理とし、ゲノム配列が未知でも適用可能な手法として、ゲノムDNAをフィンガープリントとして解析する Restriction Landmark Genomic Scanning (RLGS) 法及び Arbitrarily Primed PCR (AP-PCR) 法、またゲノム間のサブトラクションにより特異的なDNAを検出する Representative Difference Analysis (RDA) 法及び In-Gel Competitive Reassociation (IGCR) 法について、開発・原理、応用、動向・関連技術、課題を紹介しミスマッチ修復酵素 (MutS) を用いる方法についても言及した。

③特定領域遺伝子の比較解析法では、PCRによるゲノムからのターゲット遺伝子の増幅とゲル電気泳動による分離パターンに基づいた解析法として、コスト、簡便性、処理能力に優れた4手法、Single Strand Conformation Polymorphism (SSCP) 法、Denaturing Gradient Gel Electrophoresis (DGGE) 法、Restriction Fragment Length Polymorphism (RFLP) 法、Cleavase Fragment Length Polymorphism (GFLP) 法について概要を紹介し、自動化や再現性等の将来課題について言及した。

④遺伝子増幅法の多様化では、Roche社のPCR法と特許的に抵触せず、臨床検査等で利用されている遺伝子増殖・検出法として、DNA ligaseを用いる Gap Ligase Chain Reaction (Gap LCR) 法、RNA 増殖法である Transcription Mediated Amplification (TAM) 法及び Nucleic Acid Sequence Based Amplification (NASBA) 法、化学合成プローブである DNA - RNA - DNA 配列の RNase 分解を利用する Cycling Probe Technology (CPT) 法の技術概要を紹介し、遺伝子診断製品市場における非増幅法も含めた各方法論の位置づけを開発企業とともに表にまとめた。

第3章では、遺伝子発現プロファイルを知るために必要な特異的発現様式解析技術について調査を行い、①DNAチップによるもの、②cDNA sequencingによるもの、③メッセージディスプレイによるもの、に分類して技術動向をまとめた。

①DNAチップによるものでは、ハイブリダイゼーションを基本とした比較的新しい microarray 技術として、第2章では触れなかった遺伝子発現解析用 Oligonucleotide

array、合成したcDNAをチップ上にスポットしたcDNA array、電場を利用したチップや安価で効率のよいAtlas array等、の開発動向を紹介し、Affimetrix社等の先行企業によるarray技術の規格化や米国の関連ベンチャー企業間の吸収合併、業務提携によるゲノム産業支配と日本独自の技術開発の必要性、Protein Chip開発の動き等を今後の展望としてまとめた。

②cDNA Sequencingによるものでは、発現しているmRNA分子種のシーケンスによる同定と出現頻度解析から発現プロファイルを求める方法の代表として、ヒトの各種臓器におけるmRNA発現データベースの作成を目的とした阪大グループのボディマップ法、cDNAシーケンス解析手法の工夫により発現プロファイル決定とmRNA分子種同定を効率化したSerial Analysis of Gene Expression (SAGE) 法とその改良版のMini-fragment Analysis of Gene Expression (MAGE) 法、高速塩基配列解析に基づいたmRNAプロファイルの新解析手法であるMassive Parallel Signature Sequencing (MPSS) 法、について概要と展望をまとめた。

③メッセージディスプレイによるものでは、複数のmRNA同等分子種の同時比較によりmRNAの特異的メッセージを解析するDifferential Display (DD) 法を執筆者グループによる改良法を含めて解説し、mRNAの全集団を576のグループに分割してPCR産物を電気泳動解析するMolecular Index (MI) 法を特定の組織あるいは細胞における発現全遺伝子の同定を従来技術で行える有力な方法として概説した。

第4章では、遺伝子発現調節制御の基本である転写制御機構を分子レベルで解明するために必要な発現調節ネットワーク解析技術について、①転写制御シグナルの探索・解析技術、②転写制御因子の探索・解析技術、に分けて詳細に解説した。

①転写制御シグナルの探索・解析技術では、シークエンス比較解析技術として、構造遺伝子領域の予測(ジーンファインディング)と5'末端のクローニング法であるオリゴキャッピング法に基づくプロモーター領域の探索法を紹介し、転写制御因子毎の詳細な認識配列と制御機能に関する基礎的情報の収集が塩基配列情報からの制御領域推定には不可欠であることを述べた。また、転写制御領域の実験的解析技術として、レポーター遺伝子を用い

る主な方法と特徴を概説して表にまとめ、転写制御領域に配列特異的に転写因子が結合する性質を利用した試験管内解析法であるフットプリント法等について述べた。さらに、特定転写制御因子の制御領域分取・解析技術として、因子-DNA結合体をフィルターで分離する Genomic Binding Site クローニング技術とゲルシフトまたは免疫沈降で分離する Casting 法の概要を紹介した。

②転写制御因子の探索・解析技術では、DNA - タンパク質間相互作用を用いた解析技術として、λ ファージを利用した South Western 法、酵母を利用した One - hybrid 法、タンパク質とそれをコードする遺伝子とを同時に入手できる Phage Display 法を概説し、転写制御因子取得技術としての展望を述べた。また、タンパク質 - タンパク質間相互作用を利用した解析技術として、South Western 法のプローブ DNA をラベル化されたタンパク質に置き換えた Far Western 法、ビーズに固定した転写制御因子をリガンドとするアフィニティクロマトグラフィーであるタンパク質固定化カラム法、アフィニティ結合可能なタグ配列を結合させた転写制御因子を利用するタグ付きタンパク質応用技術の原理及び問題点を論じた。さらに、他の転写制御因子の探索技術においては、中胚葉系の細胞分化における転写因子カスケードを例に、細胞種特異的遺伝子（分化マーカー遺伝子）群の転写制御と細胞機能との関係を時間軸や空間軸を含めて多変量解析的に解明する必要性、リン酸化等の修飾反応や転写介在因子（コファクター）の重要性、多細胞生物における転写過程以外での遺伝子発現制御である選択的スプライシングの重要性を論じた。

第5章では、塩基配列の情報だけから、実験によらず遺伝子の領域や機能さらには遺伝子ネットワークを予測することを目指したコンピュータモデル解析技術について、①遺伝子領域・機能予測技術、②遺伝子発現制御シュミレーション、③データベース・クラスタリング技術の現状を調査し、ポストゲノムシーケンスの中核技術としての問題点を把握した。

①遺伝子領域・機能予測技術では、原核生物の遺伝子領域予測技術として、高精度化のために、隠れマルコフモデル（hidden Markov model : HMM）の枠組みを取り入れた GeneMark.hmm や GeneMark - Genesis、ribosomal binding site (RBS) のプロファイルを組み込んだ GeneHacker、次数を固定しない Markov model に基づく高速処理

が特徴のGLMMERを紹介した。また、真核生物の遺伝子領域予測技術として、一般化隠れマルコフモデル (generalized hidden Markov model : GHMM) を応用したGENSCAN、HMMを応用したVEIL や HMMgene や GeneDecoder 等の方法、プロモーター領域を予測する新手法、及び決定木を応用したMORGAN や GHMM を応用したGenie、ニューラルネットワーク (artificial neural network : ANN) を応用したGRAIL、線形判別法に基づいたFGENEH 等の予測精度改善技術を紹介し、コドンの出現頻度の偏りやコード領域における出現塩基の周期性等のコーディングポテンシャル情報を使用しない新予測法 AMELIE の将来性を論じた。次に、コード領域の配列情報に基づいた遺伝子機能予測技術として、機械学習によって構築された分類器により、シグナルペプチドのアミノ酸配列情報からタンパク質の局在化部位を、またアミノ酸配列から算出される物理化学量や統計量に基づいてタンパク質の酵素機能を予測する、ホモロジー検索に頼らない新しい予測技術を報告した。さらに、制御領域の配列情報に基づいた遺伝子機能予測技術として、原核微生物はプロモータ配列のシグマ因子依存性、また真核生物はプロモータの保存配列断片のセットから、それぞれ下流遺伝子群の発現条件を予測する手法と計算機実験によるこれらの妥当性の評価結果を解説した。

②遺伝子発現制御シミュレーションでは、プロテオーム解析等により近い将来遺伝子機能や遺伝子発現に関する莫大なデータが蓄積されても、遺伝子機能のネットワークをシステムとして理解するためには、数千、数万の遺伝子の相互作用を考慮しなければならず、コンピュータによる細胞のモデル化が必要不可欠であるとの観点から、ショウジョウバエの体節形成や線虫の発生過程、マイコプラズマの全遺伝子発現のモデル化とシミュレーションの実例を検証し、モデル構築や計算時間等の問題はあるが細胞のシミュレーション技術の確立により計算機上 (in silico) での仮想実験が可能になることを予測した。

③データベース・クラスタリング技術では、第2章から第4章までの各章で報告された技術に関連するデータベース (DB) 及び蓄積したデータを活用して有用な情報を何らかの指標を基に引き出すためのクラスタリング技術の現状と課題をまとめた。

遺伝子変異部位に関するDBとしては、個々の遺伝子座ごとの変異DBを統合したDBであるHGMD (The Human Gene Mutation Database) を、特異的発現様式解析に関する

るDBとしては、ヒトESTの大規模解析によるmRNA発現頻度DBであるBodyMapとEGAD (The Expressed Gene Anatomy Database) 及びDNAチップを用いた酵母mRNAの発現プロファイル解析DBを紹介した。発現調節ネットワークに関するDBのうち発現調節因子に関するDBとしては、高等真核生物RNAポリメラーゼIIにより認識されるプロモータ情報が蓄積されているEPD (Eukaryotic Promotor Database) や多くの真核生物転写に関する配列情報や相関情報が蓄積されているTFD (Transcription Factor Database)、真核生物のDNAエレメントと因子に関するDBであるTRANSFACを、また遺伝子ネットワークに関するDBとしては、代謝系を対象としたKEGG (Kyoto Encyclopedia of Genes and Genomes) やEcoCyc (Encyclopedia of E.Coli Genes and Metabolism)、シグナル変換経路を対象としたSPAD (The Signaling Pathway Database)を取り上げて概説した。

配列情報をベースにしたクラスタリングの代表例として、別々に登録されている同一遺伝子由来のEST配列をクラスタリングして各種組織におけるmRNAの発現頻度情報を整理したBodyMapやEGAD、最新のGenBank及びdbESTに登録されているヒトmRNA配列等とEST配列をクラスタリングしてユニークな遺伝子セットを提供しているUniGene等を示した。また、生命現象の原理解明上必須の生物種間のゲノム配列比較はホモロジーによるクラスタリングが基本となることから、東大医科研ヒトゲノム解析センターの全配列決定ゲノムの相互比較用データベース・ワークベンチシステムMBGD (MicroBial Genome Database for comparative analysis) の機能を詳細に解説した。

第6章では、ゲノム関連研究の国際動向の観点から、先導研究「ゲノム情報解読利用技術」の一環として開催された国際シンポジウム及びゲノム情報解析とコンピュータ構造解析に関する欧米の動向を報告した。

上記シンポジウムでは、「バイオインダストリーに新しい世界を開くゲノム研究—ゲノムインフォマティクス技術の新展開」と題して、ゲノム解読の現状、効率的配列決定技術、遺伝子ネットワーク解析技術、ゲノム情報解析技術等について内外の講師により発表された最新の研究成果9件の要旨をまとめた。

上記動向調査では、欧米の施策・学会・研究機関は、膨大なゲノム情報の解析にはコンピュータによる立体構造解析・予測技術が非常に有効であるとの認識から、タンパク質解析研究者のゲノム情報解析への参入、既存のタンパク質解析システムの自動化等による対応、次々と解読されてくる各生物種のゲノムデータベースへの本予測手法の徹底的適用を進めており、生物種間の「比較ゲノミックス」が新たな分野としてクローズアップされ、すでに成果も出始めていることを紹介した。また、ゲノム情報のコンピュータスクリーニングの流れ及びゲノム情報解析と構造生物学とバイオインフォマティクスとの相関を図としてまとめた。

II. 調査研究の詳細

第1章 はじめに

ゲノム解析研究はDNAシークエンシング技術のシステム化、ライン化の成功により、大量ゲノムシークエンス決定時代に突入した。表1-1に示すように大腸菌、枯草菌などの数多くのバクテリア、及び単細胞真核生物であるパン酵母の全ゲノムシークエンスが決定され、続いて多細胞真核生物である線虫ゲノムも今夏には全シークエンス決定が終了する。このような大量シークエンステータ生産の流れは、ゲノム解析の基本として今後もヒトゲノムを中心に様々な生物に対して続けられるであろう。

そしてこのような大量シークエンステータをもとに、そこに書き込まれた情報を解読するポストシークエンス研究、すなわちゲノムシークエンスから発見される大量の新規遺伝子の機能の解明や、更にはそれら遺伝子が全体として相互に関連しながら形成する遺伝情報のネットワークの解明などが求められている。そこではコンピュータによる情報処理がきわめて重要な役割を演することは間違いない。そこでは、相同性などをもとに機能既知遺伝子との関連性を推定することが重要なポイントとなる。しかしこの解析を大々的に進めるには、まだ基礎データの絶対量が不足しているのが現状である。従って、当面はゲノム機能を解析するために必要なデータの組織的な取得を進めることが必要になる。ゲノム配列（遺伝子）と生物の示す表現型や機能との対応付けを行うには様々なアプローチがあるが、ひとつの有力な手段は遺伝学的手法である。単純な遺伝形式を示すケースでは連鎖解析などが有効であるが、より一般的には、互いに異なる表現型を示す個体間（患者と健常者、変異株と野生株など）、あるいは種間でのゲノム配列の相違の比較解析が重要となる。また、遺伝子破壊株のような遺伝背景の違い、あるいは発生・分化の過程のような生理条件の違いに対応した遺伝子発現の変化の系統的な解析も重要である。これらを通して表現型と遺伝子との関連が明らかにされると期待される。もちろん、遺伝子は単独ではなく遺伝子間あるいは蛋白間の相互作用を通して様々な生体反応を成り立たせている。従って、ゲノムの担う機能を真に理解するには、各遺伝子（その産物のタンパク質）の相互作用やそれらが生み出すパスウェイを解明し、ゲノムという設計図を通して生物をシステムとして理解してゆくことが必要になる。

このようなポストシークエンスのゲノムの機能解析の推進には、1)もちろん各々の目的や目標の設定とそれに必要なリソース、すなわち患者サンプルや変異株（自然に作られたケース以外に、遺伝子破壊等によって人為的に作成されたものも含む）、及び多様な生物

種のサンプルの整備が不可欠である。2) そしてそれらを解析するツール、すなわちゲノム配列の相違や遺伝子発現の動態を大量かつシステムティックに比較解析するためのツールの開発、3) 更にそれらのツールを用いて大量に取得されたデータの適切なデータベース化と、そこから目的に応じた有用な情報を抽出するための情報処理技術の開発などが要求される。ゲノム分野では、これからは物だけではなく情報が非常に高価な商品となる時代である。従って、最終的に取得されるデータの質と量の重要性を再度認識しておく必要がある。我が国の製薬企業の少なからずが、欧米ベンチャーが独自に開発したデータベースへのアクセス権を高額を支払って取得していると言われている。このようなデータの欧米依存体質から脱却することが重要となる。今後は高度な解析手法の開発とともに、それを用いた、目的に応じたあるいは他にユニークな有用データの蓄積とそれを有効に利用する情報科学技術の開発を自ら行うことが、我が国のバイオ産業の競争力を高める上で重要である。

本先導研究ではこのような背景を念頭に、ポストシークエンス時代に要求されるゲノム情報の取得とその処理の先端技術、特に比較ゲノム解析、発現プロファイル解析、遺伝子間(蛋白間)相互作用解析に関するハード及びソフト両面についての国際的な動向の調査研究を行った。この調査研究が我が国この分野における国際的競争力の強化を考える際の一助になれば幸いである。

表1-1 Recent Progress of Sequencing of Small Genomes (Feb, 1998)

Name	Japanese Name	Genome Size(Mb)	State	Presenter or Institute
Eubacteria				
<i>Staphylococcus epidermidis</i>	表皮ブドウ球菌	2.30	N	William J. Kimmerly (Glaxo Wellcome Inc.)
<i>Staphylococcus aureus</i>	黄色ブドウ球菌	2.80	N	GTC
<i>Streptococcus pyogenes</i>	溶血連鎖球菌	1.92	N	Bruce A. Roe(U Oklahoma)
<i>Streptococcus pneumoniae</i>	肺炎球菌	2.05	N	Brian A. Dougherty(TIGR)
<i>Enterococcus faecalis</i>	腸球菌	3.00	C	TIGR, HGS Inc
<i>Neisseria gonorrhoeae</i>	りん菌	2.05	N	Bruce A. Roe(U Oklahoma)
<i>Neisseria meningitidis</i>	脑膜炎菌	2.30	P	TIGR & Oxford U
<i>Pseudomonas aeruginosa</i>	綠濃菌	5.80	P	GTC
<i>Legionella pneumophila</i>	レジオネラ	4.10	P	TIGR
<i>Escherichia coli</i>	大腸菌	4.64	C	H. Mori(Nara Ins ST) et al Fred Blattner(U Wisconsin)
<i>Caulobacter crescentus</i>		3.80	P	TIGR
<i>Rhodobacter capsulatus</i>		4.00	P	Chicago
<i>Rhodobacter sphaeroides</i>		3.80	P	Texas
<i>Sphingomonas aromaticivorans</i>		—	P	PNL
<i>Ehrlichia sp.</i>		1.40	P	TIGR
<i>Salmonella typhimurium</i>	ネズミチフス菌	4.50	P	TIGR
<i>Helicobacter pylori</i>	ピロリ菌	1.67	C	Jean-Francois Tomb(TIGR)
<i>Vibrio cholerae</i>	コレラ菌	2.50	P	TIGR
<i>Haemophilus influenzae</i>	インフルエンザ菌	1.80	C	TIGR
<i>Clostridium acetobutylicum</i>		4.10	P	Douglas R. Smith (Genome Therapeutics Corp.)
<i>Bacillus subtilis</i>	枯草菌	4.20	C	N. Ogasawara(Nara Ins ST) & EU sequencing groups
<i>Mycobacterium tuberculosis</i>	結核菌	4.40	N	Robert D. Fleischmann(TIGR)
<i>Mycobacterium avium</i>		4.70	P	TIGR
<i>Mycobacterium leprae</i>	らい菌	2.80	P	Sanger Center, GTC
<i>Actinobacillus actinomycetemcomitans</i>		2.20	P	U Oklahoma
<i>Aquifex aeolicus</i>	(嫌気性菌)	1.55	N	Ronald V. Swanson (Recombinant Biocatalysis Inc.)
<i>Deinococcus radiodurans</i>		1.05	P	TIGR
<i>Porphyromonas gingivalis</i>		2.20	P	TIGR
<i>Shewanella putrefaciens</i>		4.50	P	TIGR
<i>Thermotoga maritima</i>		1.80	P	TIGR
<i>Treponema pallidum</i>	梅毒スピロヘータ	1.05	C	Claire Fraser(TIGR)
<i>Treponema denticola</i>	(スピロヘータ)	3.00	P	TIGR & U Texas
<i>Borrelia burgdorferi</i>	(スピロヘータ)	1.30	C	Claire Fraser(TIGR)
<i>Rickettsia prowazekii</i>	癆瘧チフスリケッチャ	1.10	N	Siv G. E. Anderson(Uppsala U)
<i>Chlamydia trachomatis</i>	トラコーマクラミジア	1.05	C	UC Berkeley & Stanford, GENSET
<i>Chlamydia pneumoniae</i>		1.23	C	GENSET
<i>Mycoplasma capricolum</i>		1.10	P	George Mason
<i>Mycoplasma genitalium</i>	(マイコプラズマ)	0.58	C	Claire Fraser(TIGR)
<i>Mycoplasma pneumoniae</i>	肺炎マイコプラズマ	0.80	C	R. Himmelreich(U Heidelberg)
<i>Ureaplasma urealyticum</i>	(マイコプラズマ)	0.76	N	Elliot J. Lefkowitz(U Alabama)
<i>Synechocystis sp.</i>	(らんそう)	3.57	C	S. Tabata(KDRI)
Archaea				
<i>Halobacterium salinarium</i>	(好塩菌)	4.00	P	MPG(Germany)
<i>Halobacterium halobium</i>	(好塩菌)	—	P	Massachusetts
<i>Methanococcus jannaschii</i>	(メタン產生菌)	1.66	C	C.J. Bult(U Illinois)
<i>Methanobacterium thermoautotrophicum</i>	(メタン產生菌)	1.70	C	Douglas R. Smith (Genome Therapeutics Corp.)
<i>Archaeoglobus fulgidus</i>		2.20	C	Hans-Peter Klenk(TIGR)
<i>Pyrobaculum aerophilum</i>	(好熱菌)	1.80	N	Sorel Fitz-Gibbon(U California)
<i>Pyrococcus furiosus</i>	(好熱菌)	2.10	P	Robert B. Weiss(U Utah)
<i>Pyrococcus horikoshii</i>	(好熱菌)	2.00	P	MITI & U Tokyo
<i>Sulfolobus solfataricus</i>		3.00	P	Christoph W. Sensen(Institute for Marine Biosciences, Canada)
<i>Thermoplasma acidophilum</i>		1.70	P	Max-Planck-Ins. for Biochemistry

Eucaryotes				
<i>Saccharomyces cerevisiae</i>	パン酵母	12	C	RIKEN & EU sequencing groups
<i>Shizosaccharomyces pombe</i>	分裂酵母	14	P	Sanger Center
<i>Aspergillus nidulans</i>		31	P	Oklahoma
<i>Neurospora crassa</i>		47	P	New Mexico
<i>Plasmodium falciparum</i>	マラリア原虫	30	P	S. L. Hoffmann(Naval Medical Res. Institute)
<i>Caenorhabditis elegans</i>	線虫	100	N	Steven J.M. Jones(Sanger Center) & Washington U
<i>Caenorhabditis briggsae</i>		—	P	Washington U
<i>Drosophila melanogaster</i>		100	P	Berkeley
<i>Arabidopsis thaliana</i>	シロイヌナズナ	100	P	W. R. McCombie(Washington U), EU & Japanese sequencing groups
<i>Oryza sativa</i>	イネ	450	P	MAFF JAPAN
<i>Fugu rubripes</i>	フグ	400	P	MRC

C; Complete, N; Nealy Complete, P; in Progress

第2章 変異・多型部位の検出・取得技術

2-1 DNAチップ

2-1-1 目的と役割

DNAチップとは、形態的に小型のシリコン又はガラス基板上に多種・多数のDNAオリゴ鎖が集積されたものであり、半導体のリソグラフィー技術やマイクロドッティング技術を駆使している点に特徴がある。応用としては、医薬品の開発、農産物の管理などの幅広い分野に加えて、臨床における遺伝子検査、特に癌や感染症、遺伝子疾患などの早期・確定診断を目的に開発が進められている。最近ではポストゲノムである遺伝子機能解析への応用が検討されており、周辺機器の小型化に伴い将来の臨床の場において大きな役割を期待されている。

DNAチップは当初、ヒトゲノムプロジェクトにおいてDNAシーケンシングの高速マイクロ化を実現する技術として期待された。しかし、チップ上に固相するオリゴヌクレオチドの長さが限定されることやリピート配列の解析が困難であること、さらにハイブリダイゼーション条件の設定が難しい点から、現在では遺伝子診断や発現遺伝子解析を目的とした応用開発へと重点が移ってきている。

2-1-2 技術開発

DNAチップを用いた遺伝子検査システムの技術開発において (1) DNAチップ製法技術、(2) パターン解析技術、(3) マイクロファブリケーション技術の3つが大きな開発項目として上げられる。

(1) DNAチップ製法技術

Affymetrix社は光反応とフォトリソグラフィー技術を利用してチップ上でオリゴヌクレオチドを固相合成した“GeneChip”を開発し、 1.28cm^2 のチップに約10万個のオリゴマーを固定しアレーを形成することが可能である。この方法はチップの大量生産には向いているが、30merを超える長鎖オリゴマーの固相は難しい。そこで最近ではあらかじめ合成したオリゴマーまたはcDNAを基板上に固相する方法を利用したDNAチップが各社・各機関で相次いで開発されている。さらに、2次元のチップ構造がハイブリダイゼーション条件の設定を困難にしていることから、効率良い反応を行えるチップ構造、

オリゴマーの固相方法、ミスマッチの少ないアッセイ方法の改良も進められている。

(2) パターン解析技術

小型マイクロチップ上におけるオリゴマー固相法の開発が進みオリゴマー間の分解能が増すにつれて、ハイブリダイゼーション反応パターンを読み取る技術の進歩も欠かせないものとなる。高感度検出システムとしてはアルゴンレーザー励起光をチップ上でスキャニングさせてハイブリダイゼーションの結果得られる蛍光を高感度コンフォーカル顕微鏡で検出する仕組みが一般的である。さらにシステム全体のマイクロ化が進む中では検出器自体の小型化も重要な開発項目となり、Genometrix社が開発しているCCDによる直接検出方法はその一例である。

(3) マイクロファブリケーション技術

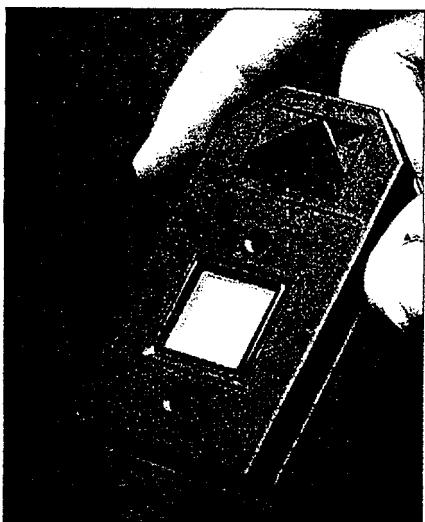
DNAチップ本来のメリットである微量サンプル処理を有効に行うためには、チップ自体の小型化のみならず周辺機器のマイクロ化が必要となる。同時にトータルアッセイシステムとして、サンプル前処理装置、遺伝子增幅装置、DNAチップ、検出器などの集積化 (μ TAS: μ Total Analysis SystemあるいはLab On A Chip) が各機関で進められている。このように微量サンプル検査システムが自動化されることにより、検査コストの低下および市場の拡大が期待できる。

2-1-3 企業・機関情報

現在DNAチップ技術および市場をリードしているAffymetrix社をはじめ、代表的な企業・機関の最新技術情報¹⁾を報告する。

Affymetrix社 (<http://www.affymetrix.com/>) ではフォトリソグラフィー技術を用いてチップ上でのオリゴスクレオチドの固相合成を実現し (GeneChip: 図2-1)、BRCA1²⁾などの点変異検査では基板上で15-20 mer程度のオリゴマーを合成し、1.28 cm² のチップに約10万個のオリゴマーを固定しアレーを形成したものを使用している。

フォトリソグラフィーを利用したオリゴマーチップの製法では、基板上においてマスキングにより限定された部位で光照射により目的の塩基を反応させる操作を繰り返すことによりマイクロアレーが形成される³⁾ (図2-2)。解析装置はHewlett Packard社と共同開発しており、4mWアルゴンレーザーによりチップ上でスキャン励起され、得られた蛍光はGeneChip ソフトウェアによりパターン解析されるしくみになっている (図2-3)。



High-density oligonucleotide probe array package.

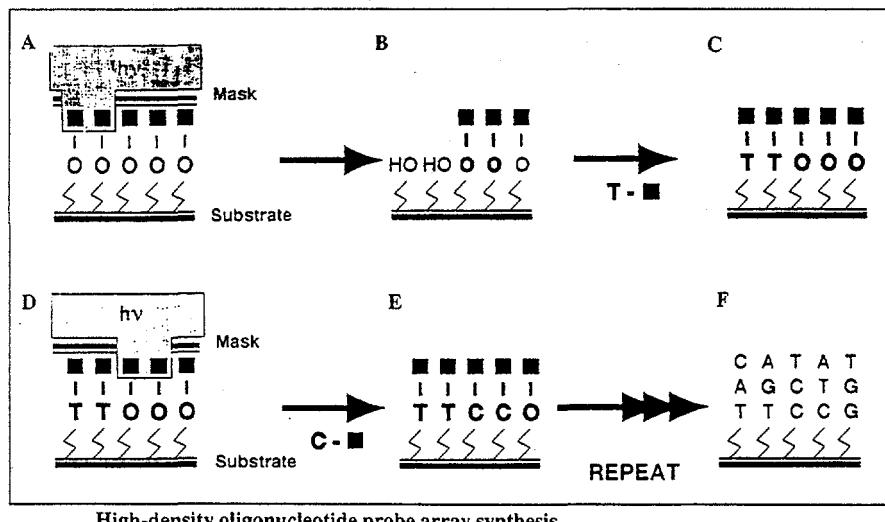
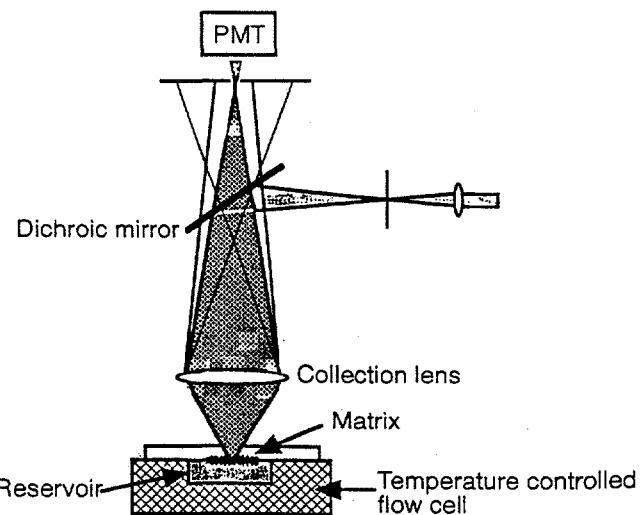


図2-1 Affymetrix社開発、GeneChip

図2-2 チップ上でのオリゴマー合成法

Fluorescence Detection

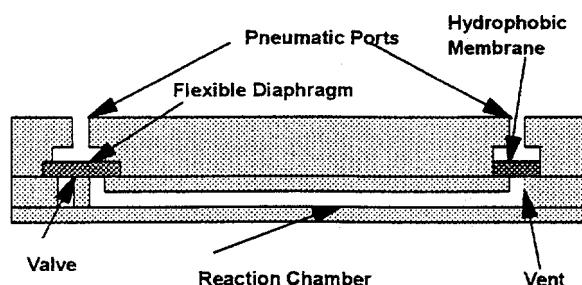


Detection format. Probe-target hybridization is detected by epi-fluorescence confocal scanning.

図2-3 ハイブリダイゼーションパターン解析装置模式図

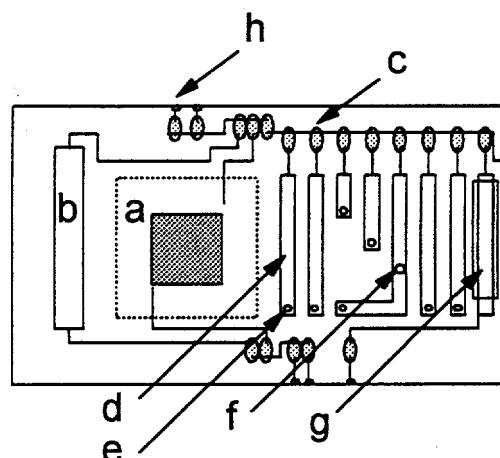
臨床における応用を目的にAffymetrix社は他企業・組織と積極的に共同研究開発を進めている。中でも、97年度には英国Glaxo Wellcome社とHIV遺伝子データベースを設立し、Roche社と共同で遺伝子発現モニターシステムの開発を進め、Oncormed社とはBRCA1,2点変異検出システムを共同開発し、ごく最近（97年11月）ではHewlett-Packard社とcytochromeP450多型性検出システムの開発に乗り出すことが報道された。

さらに、Affymetrixではマイクロファブリケーション技術を用いて、Microfluidic Biochemical Analysis SystemつまりDNA抽出からDNAチップ上でのハイブリダイゼーションを小型カートリッジ内にて行うことを可能にした（図2-4-1,2）。応用例としてHIVウイルスの変異検査では10 μL の少量サンプルにて実現することを示した。⁴⁾



Cross section diagram of chamber with diaphragm valves and porous hydrophobic vents for sensorless fluid positioning. The lower wall of the reaction chamber is 40 to 250 μm thick for thermal addressing.

図2-4-1 Microfluidic Biochemical Analysis System チェンバー断面図



Top view of integrated cartridge system measuring 8.5 x 40 x 70 mm. Elements include a) GeneChip and hybridization chamber, b) auxiliary hybridization chamber, c) diaphragm valves, d) chambers for reagent storage, metering, and reactions, e) hydrophobic vents, f) linking chamber, g) extraction chamber with glass ceiling, and h) fluidic ports. Reaction chambers are 1.5 mm wide by 13 mm long by 500 μm deep with 10 μL volume

図2-4-2 集積カートリッジシステム上面図

Affymetrix社がチップ上でオリゴヌクレオチドを固相合成するのに対して、あらかじめ合成したオリゴマーあるいはcDNAを基板上に固相する方法を利用したDNAチップも開発されている。Synteni社 (<http://www.synteni.com/>) は米国スタンフォード大学

のDr.Schenaからライセンスを獲得し、マイクロアレー製作用ロボットを用いて数1000遺伝子の発現を決定できるDNAマイクロアレイ技術(GEM: Gene Expression Microarray)の開発を進めている。図2-5は製品化されたDNAチップである。

実際のアッセイでは2種蛍光ラベルmRNAをターゲットにし、ノーマルと検体とで競合ハイブリダイゼーションを行わせてその蛍光比パターンを検出することにより、発現パターンを比較している^{5,6)}(図2-6)。

図2-5 Synteni社開発、DNAチップ

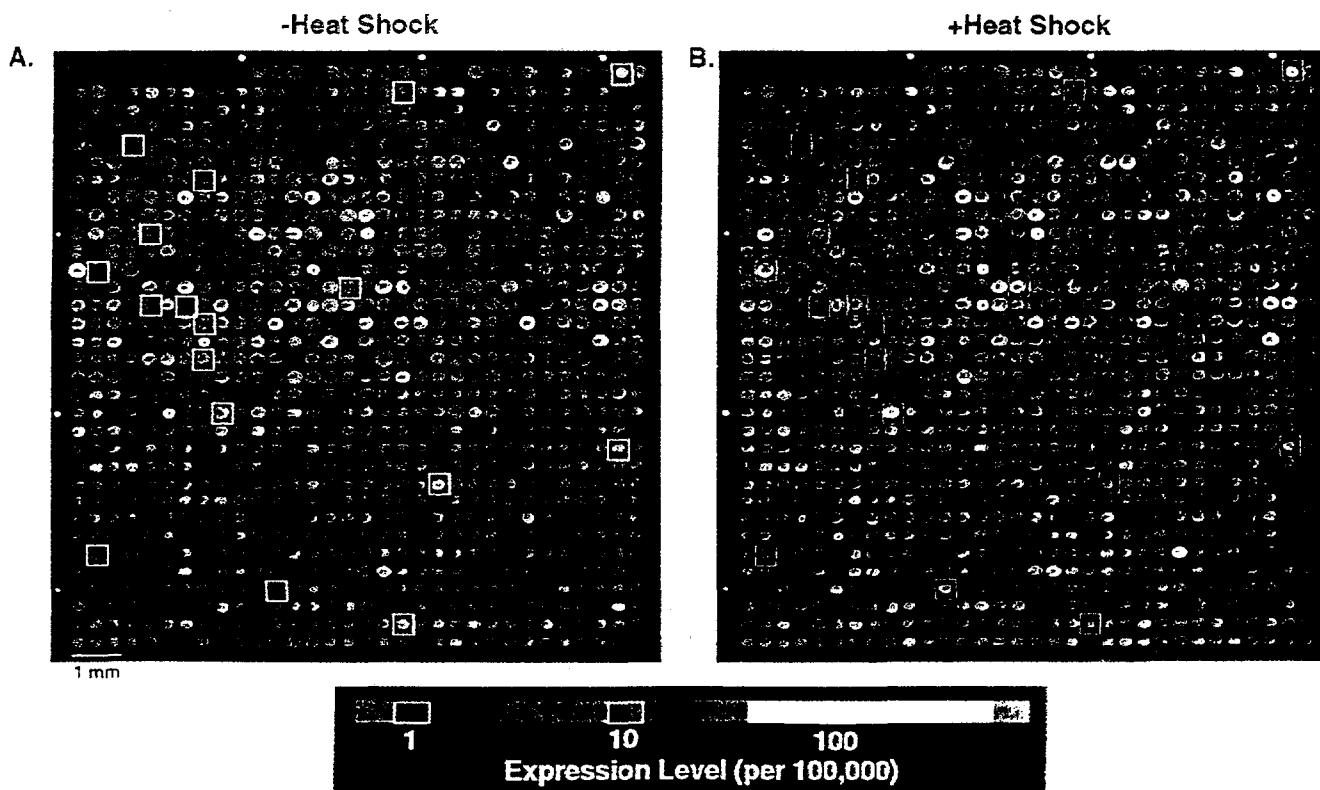
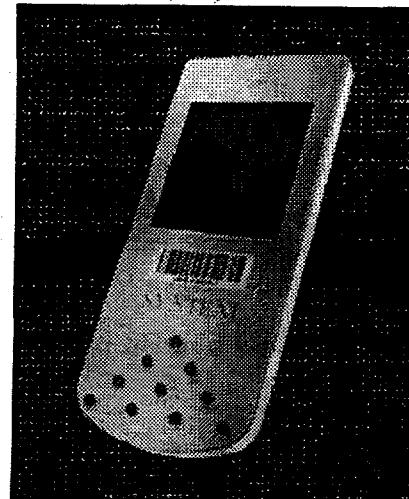


FIG.1. Human gene expression monitored on a microarray. Fluorescent scans represented in a pseudocolor scale correspond to expression levels. The array contains 10 *Arabidopsis* controls (upper left corner, elements 1-10) and 1046 human peripheral blood cDNAs. Fluorescent probes were prepared by labeling mRNA from Jurkat cells grown at 37°C (-Heat Shock, A) or 43°C (+Heat Shock, B). Array elements that display altered fluorescence intensity (white boxes) corresponded to genes activated (red boxes) or repressed (green boxes) by heat shock. The color bar was calibrated in separate experiments using known quantities (wt/wt) of *Arabidopsis* control mRNAs added to the labeling reaction. Microarray rows (at left) and columns (at the top) are demarcated at 10 element increments (white circles). (Bar = 1 mm.)

図2-6 cDNAチップにおける発現パターンの比較

Hyseq社は既知のDNA配列に基づいたDNAプローブを使って塩基配列を決定する技術(SBH: Sequencing by Hybridization)を開発し、基本特許を所有している。これによりHyseq社は10塩基のプローブをシリコン表面に結合させ100万塩基の解析が可能なDNAチップを開発し、「97年3月には7塩基のプローブで1万6000塩基を解析できることを明らかにしている。これを利用してHIVゲノム全体を一度に調べ、HIVの既知の配列を検出できることを確認している。

Vysis社 (<http://www.vysis.com/>) はゲノムを解析するためのDNAプローブを基礎にFISHやCGH(Comparative Genomic Hybridization)⁷などを利用した診断薬や機器の開発、販売を行っている企業である。最近ではCGHを拡張して、2cm²のグラスチップ上における2色蛍光DNAの競合ハイブリダイゼーション法(GCGH: Genosensor-based CGH)を開発している。

ゲノム情報が調べられていくにつれそのデータ解析技術も高度化し、データベースの重要性が認識される中、Incyte Pharmaceuticals社 (<http://www.incyte.com/>) はゲノム情報の商業化をリードするゲノム・ベンチャーであり、ゲノム情報をデータベース化した「LifeSeq」へのアクセス権を製薬企業などに非独占的に販売している。Affymetrix社とはDNAチップを使った遺伝子発現解析のための共同技術開発を進めており、「LifeChip」と称して前立腺ガン、乳ガン、炎症疾患、G蛋白共役受容体の分野に適用を試みている。

Molecular Dynamics社 (<http://www.mdyn.com/>) はレーザースキャニングシステムにより高速に遺伝子発現を検出するDNAマイクロ・アレー技術により、Amersham Pharmacia Biotech社と共同で遺伝子機能解析ツールの応用開発を進めている。将来は10,000 spots/slideのDNAチップ開発を計画している。97年10月にはDuPont社と提携して農産物の遺伝子発現への応用検討を開始した。

DNAチップを含めた周辺機器の小型化は将来の実用性の鍵を握っていると思われるが、Nanogen社 (<http://www.nanogen.com/>) はマイクロリソグラフィとマイクロマシン技術を用いたデバイス開発により、サンプル調製、ハイブリダイゼーション、検出、データ解析などのステップを集積し、電気や光コネクションを備えた高度のマイクロ自動検査

システム (APEM: Automated Programmable Electronic Matrix) を実現した。図 2-7 はシリコンチップの構造を図示したものである⁸⁾。

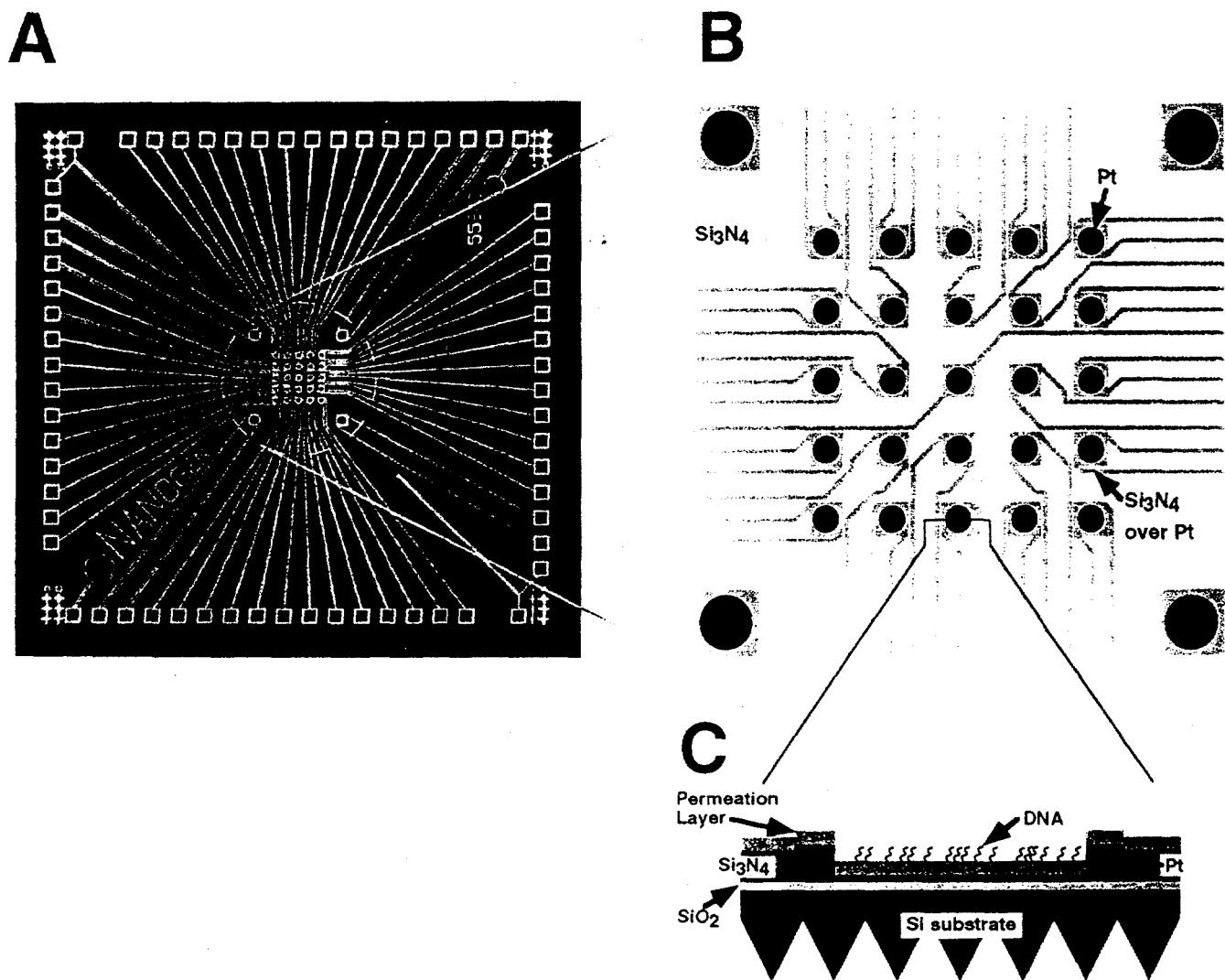


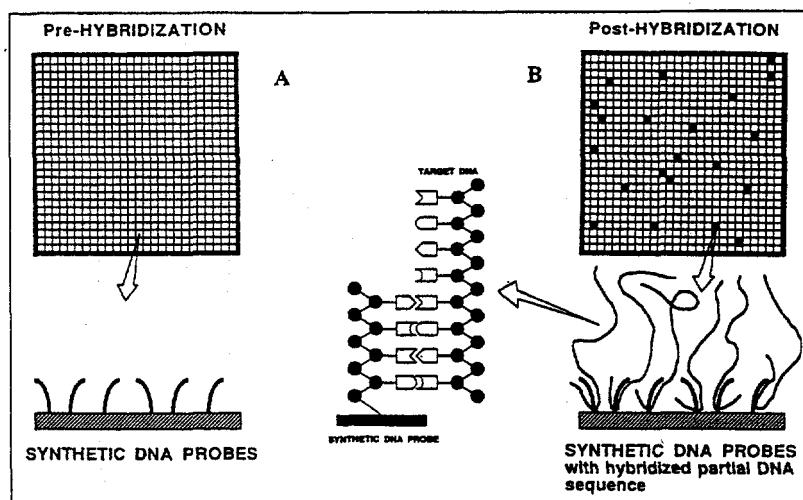
FIG. 1. Silicon chip with an array of electrodes. (A) Overview of the microfabricated device. Chip dimension was 1 cm square. Light squares along the perimeter are exposed platinum contact electrodes for connections to power supply. Light lines are platinum leads insulated with dielectric, connecting contact electrodes to the exposed platinum electrodes. (B) Electrode array region of the chip. The central 1 × 1 mm test site array region consists of 4 large (160-μm diameter) corner electrodes and 25 central 80-μm electrodes. Pt, exposed Pt test sites; Si₃N₄, dielectric; Si₃N₄ over Pt, Pt insulated by Si₃N₄. (C) Cross section of an electrode test site. Location of section is indicated by lines extending from B. Pt, Si₃N₄, Pt insulated by Si₃N₄; SiO₂, dielectric layer; Si substrate, wafer material; Permeation Layer, agarose layer containing streptavidin; DNA, biotinylated oligonucleotides bound to streptavidin. DNA binding was not limited to the surface.

図 2-7 Nanogen社開発、シリコンチップ構造

さらに電場内にてハイブリダイゼーションを行うことにより、効率的にDNAの移動や濃縮、非結合DNA部の除去を行わせることを可能としているのが特徴である。

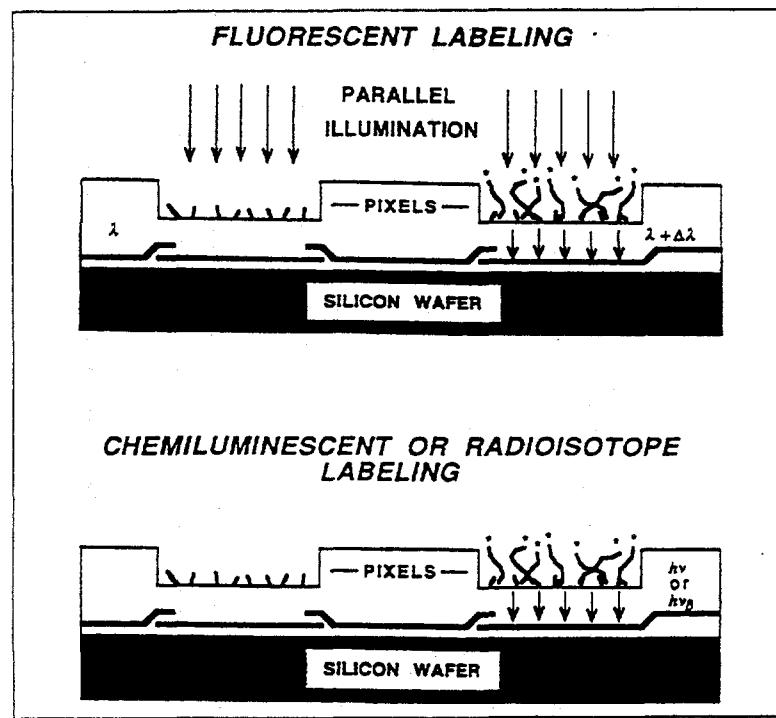
臨床応用としては、97年5月から米国Becton Dickinson社のDNA増幅システムをマイクロシステムに組み込んだ感染症の診断システムを共同開発している

Genometrix社では検出部のスペースを最小限にするために、インクジェット方式で固相されたDNAチップの下に直接設置されたCCDがハイブリダイゼーションの結果を検出し解析できるシステムを開発した⁹⁾(図2-8-1,2)。



Molecular binding assay A) before and B) after probe/target hybridization.

図2-8-1 チップ上におけるハイブリダイゼーション前後の反応模式図



Modes of direct CCD molecule detection.

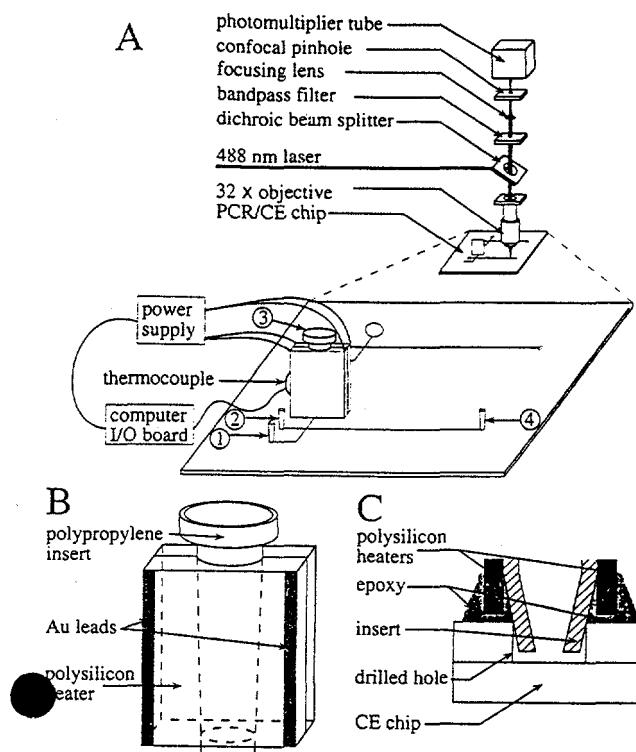
図2-8-2 Direct CCDシステムチップ断面図

Molecular Tool社はオリゴヌクレオチドの固相開発を行ってきたが、GBA (Genetic Bit Analysis) と称して95年に全自動BRCA1シークエンスシステムを開発し、チップ上でのシステム化を可能とした。

Caliper Technologies社 (<http://www.caliper.com/>) においても “Lab On A Chip” を提唱し、マイクロチップ上にてマイクロフルイデックス技術によりトータルシステムを開発している。

一方、大学や研究機関においてもマイクロファブリケーション技術の進歩は著しく、DNAチップを取り込んだμ TAS (or Lab On A Chip) の開発が急速に進められている。

BerkeleyのMathiesらはCAE (Capillary Array Electrophoresis) チップ上にて高電圧による電気泳動を行いDNA断片を分離し、レーザー励起してコンフォーカル蛍光顕微鏡で検出解析するシステムを開発した。さらにLawrence Livermore National LaboratoryとUniversity of California, Berkeleyとの共同研究では、PCR部分をマ



Schematic of the integrated PCR-CE microdevice. (A) Laser-excited confocal fluorescence detection apparatus and an integrated PCR-CE microdevice. (B) Expanded view of the micro-fabricated PCR chamber. (C) Expanded cross-sectional view of the junction between the PCR and CE devices. The size of the epoxy-filled gaps is exaggerated for viewing clarity.

図2-9 PCR-キャピラリー電気泳動マイクロデバイス模式図

イクロファブリケーション技術により機能集積した装置が開発された^{10, 11, 12)} (図2-9)。また、University of Pennsylvaniaではシリコンガラスチップに形成したPCR反応容器の表面処理と反応効率の検討¹³⁾、及びLCR (Ligase Chain Reaction)への応用¹⁴⁾を試みた。これらの研究はDNAチップにおける反応の前処理段階のマイクロ化、高速化として期待がもたれる。

その他、Patric Brown's Laboratory (<http://cmgm.stanford.edu/pbrown/>)では、図2-10のように3つのサーボモータでコントロールされるリニアレールテーブルを設計し、プリントチップによりDNAを固相する“アレーシステム”を開発し、マイクロチップの大量生産に向けた技術開発を促進している。

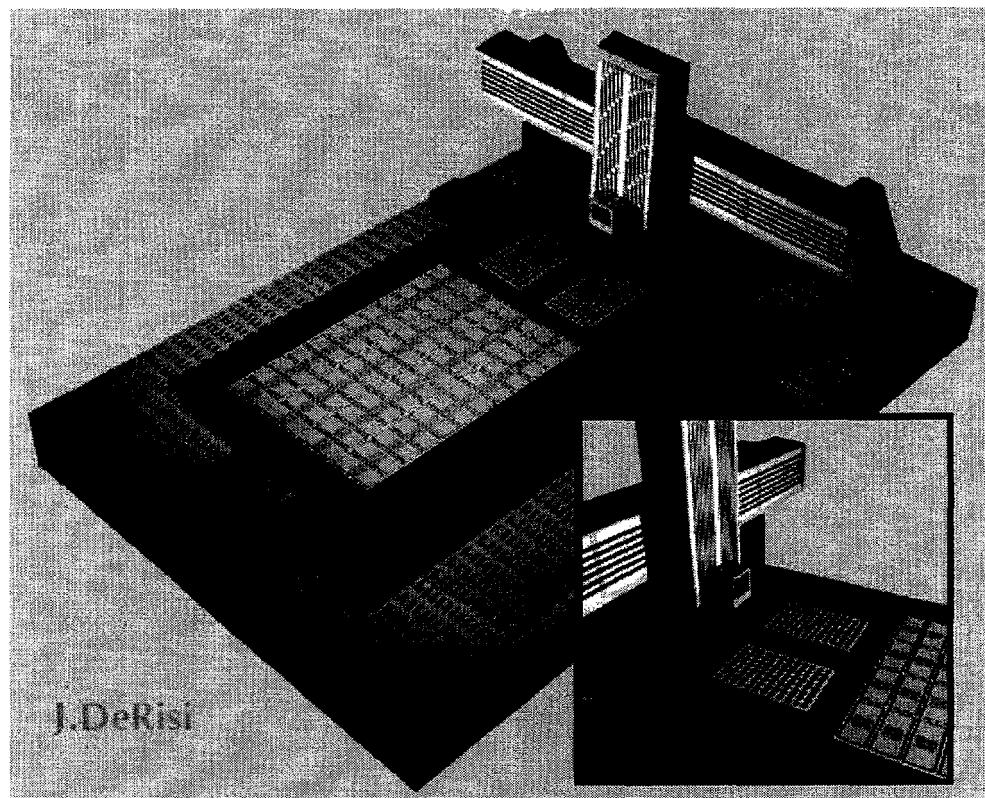


図2-10 Patric Brown's Laboratory開発アレーシステム

表2-1に前述の各企業、研究所の動きをまとめた。特許に関してはPATORIS、WPATによるキヤワード検索およびUS PATENT（ホームページ）での検索結果である。

Affymetrix社は特許出願件数においても抜きんでており、DNAチップの制作法、検出、応用と幅広い出願を行っている。

表2-1 DNAチップ技術マップ

企業・研究機関	技術内容	応用	提携・共同開発組織	特許(1998.1.6現在)
Affymetrix, Inc.	GeneChip (商品名) ・フォトリソグラフィー技術 ・マイクロファブリケーション技術	・チトクロームP450多型性検出('97) ・BRCA1.2 タイピング ('97) ・p53突然変異検出システム ('97) ・遺伝子発現モニタ一検討 ('97) ・HIV 遺伝子データベース設立 ('97) ・細菌の遺伝子発現量検出 ('96) ・BRCA1点変異検出 ('96) ・悪性黒色腫由来のスポットcDNA ('96) ・疾患遺伝子データベース ('96) ・囊胞性纖維症変異検査 ('96) ・DNA診断用デバイス ('94) ・DNA解析システム ('94) ・ヒト遺伝子の機能解析技術 ('94)	Hewlett-Packard社 Oncormed社 Roche社 英国 Glaxo Wellcome社 Roche社 米国 NIH Genetics Institute社、Stanford大学 米国 Incyte Pharmaceuticals社 Roche社 米国 Molecular Dynamics社 米国 Hewlett-Packard社 米国 Genetics Institute社	出願件数: Affymetrix社から49件、 Affymetrix社から5件の計54件。 内容 ・SBH法によるシーケンシング ・フォトリソグラフィによるチップ作製方法 ・光学検出法 ・システム化のためのマイクロバルブ等
Synteni	・GEM (Gene Expression Micro-array) ・Stanford大学(Dr.Sohenaら)からライセンス獲得 ・cDNAプローブをチップに固相 ・2種蛍光ラベルmRNAをターゲット	・遺伝子発現パターン検出		出願件数: 0件
Hyseq, Inc.	HyChip (Super Chip) ・SBH(Sequencing by Hybridization)	・HIV ゲノム配列検出 ('97) ・植物、微生物DNA検査 ('97)		出願件数: 3件 ・SBH法によるシーケンシング ・Affymetrix社に対して提訴
Vysis, Inc.	・GCCH (Genosensor-based Comparative Genomic Hybridization)			出願件数: 2件 ・染色体検出用プローブ
Incyte Pharmaceuticals, Inc.	LifeChip LifeSeq: cDNAゲノム情報データベース	・遺伝子発現解析 ('96より)	Affymetrix社	出願件数: DNAチップに関してはなし。
Molecular Dynamics, Inc.	DNA診断用デバイス	・農産物の遺伝子発現への応用検討 ('97) ・遺伝子機能解析ツールの応用開発 ・DNA診断用デバイス開発	DuPont社 Amersham Pharmacia Biotech Affymetrix社	出願件数: DNAチップに関してはなし。
Nanogen, Inc.	APEM (Automated Programmable Electronic Matrix) ・マイクロエレクトロニクス技術を用いて電気的バイオ アッセイを開発	・感染症診断システム ('97) ・高感度システム開発 ('97)	Beeton Dickinson社 米国 ProImx社投資	出願件数: 3件 ・微細電極アレーによる反応制御等。
Genometrix, Inc.	・インジェット方式によるオリゴDNAプローブ固相 ・ハイブリダイゼーション結果はチップ下に据えたCCD により直接検出。			出願件数: 0
Molecular Tool, Inc.	GBA (Genetic Bit Analysis) Biochip 12-30mer オリゴを96ウェルプレートに固相する方法(AnalBiochem,1995)	・全自动BRCA1シークエンスシステム ('95)		出願件数: 3件 ・ポリスチレン、またはガラス基板へのオリゴスクレオドの固相化方法等。
Caliper Technologies	キャビラリ電気泳動(Lab On A Chip) マイクロファブリケーション技術			出願件数: 0件
Lawrence Livermore National Laboratory	マイクロファブリケーション技術 (PCR部分)	・マイクロPCRによる感染症診断システム ・PCRとキャビラリ電気泳動を組み合わせた分析システム	UC Berkeley の研究者がLLNLへ移動	出願件数: 4件 (UC Berkeleyから出願) ・マイクロPCRの構成
University of California at Berkeley	CAE (Capillary Array Electrophoresis)			
University of Pennsylvania	マイクロPCR 及び LCR	・ポイントミューテーションの検出		出願件数: 2件 ・PCRの構成
Patric Brown's Laboratory	アレーシステム	・マイクロチップの大量生産システム		

2-1-4 DNAチップの課題

DNAチップは次世代の遺伝子検査・診断に役立つものと注目されている一方、現時点での技術的限界を無視した過度の期待もしばしば聞かれる。しかし実際に臨床の場において広く一般に利用されるようになるにはまだ多くの課題が残されている。

DNAチップ及びその応用についての課題を論じた総説はいくつか見られるが^{15, 16)}、共通して取り上げられている重要課題としては、やはりチップのコストが如何に下げられるかということであろう。そのためにはチップそのものの材質改良をはじめ、オリゴプローブの効率的固相法の開発、チップ製法の自動化などがコスト低下につながると考えられる。コスト低下は研究分野から臨床検査に向けた市場の広がりをもたらし、さらには周辺機器の小型化とともにベッドサイド検査・Point of Careシステムとして大いにその役割を増すものと期待される。

DNAの固相技術については、Affymetrixに代表されるチップ上でのオリゴマー固相合成法とあらかじめ合成したオリゴマーまたはcDNAを固相する方法の2通りがある。前者の方法では作成においてその微細加工と大量生産には向いているが30 mer以上の長鎖オリゴプローブの合成は困難である。一方、後者の方法では長鎖オリゴマーの固相が容易であることから自由度の高いチップが作成出来るが、欠点としては数多くのDNA鎖をあらかじめ準備する必要があるために調製に要する時間とコストが問題となる。

また、チップ上での多種プローブの分解能が高まるにつれてそれに対応したハイブリダイゼーション条件の設定が困難となる。ミスマッチを最小限に抑えるためにも、有効なハイブリダイゼーションを実現する上で必要な固相プローブの配置や密度の検討が必要であり、そのために適切な基板形状のアイデア開発も望まれる。例として、キャピラリー型やフロー型のチップが検討されており、Nanogen社のパテントではフローシステム上でサンプルを流すことによりハイブリダイゼーションの効率を上げている。

さらに、プローブ自身の設計もミスマッチを防ぐ重要なファクターとなり、ハイブリダイゼーションを含んだアッセイ系の検討も重要課題である。Heinrich¹⁷⁾らはDNAの代わりにPNA (Peptide Nucleic Acid) をチップ上に固相し、ハイブリダイゼーションしたターゲットDNAのみに含まれるリン酸値を測定する系を開発し、その測定には高感度のSIRIMP / SIMS (Sputter Initiated Resonance Ionization Microprobe / Secondary Ion Mass Spectrometry) を利用している。

微量サンプルの解析にはDNAチップのみならず周辺機器の小型化が求められ、さらに

それらを集積したマイクロ・トータルシステムの開発がNanogen社を代表として各社・各機関で進められているが、まだ実用には時間がかかりそうである。チップ上のハイブリダイゼーション反応とそのパターン検出・解析のみならず、チップにサンプルを導入するまでの前処理ステップ（サンプル処理、遺伝子抽出、PCRなど）においてもマイクロ自動化が必要となる。このためにマイクロファブリケーション技術の進歩にも期待がかけられている。

DNAチップの応用を考えた場合、当初のシークエンス解析よりも遺伝子変異検査、DNAマッピングそして遺伝子発現解析において効果を発揮できるものとして応用開発が進められている。特に遺伝子変異検査の応用では、BRCA1,2やp53などの突然変異検出システムが開発されているが、将来、さらに多くの遺伝子変異が検査項目として選択された場合、DNAチップ上の多項目検査が可能となると考えられる。

現在、ヒトゲノムプロジェクトにより遺伝子のデジタル情報が明らかになりつつある一方、ポストゲノムである遺伝子機能解析の重要性が問われており、ヒトの細胞や臓器において遺伝子の発現・機能を解析し、遺伝子発現プロファイルの作製も試みられている¹⁸⁾。このような目的でDNAチップは遺伝子機能の解析ツールとしてもその役割が期待されている。そのためにはmRNAに対するプローブを固相したチップも開発される必要があり、発現量を検討する上で定量化が重視されることからハイブリダイゼーションの再現性、チップの均一性、検出感度およびダイナミックレンジの向上が開発の課題となる。

文献

- 1) 97/98世界のバイオ企業2200社、日経バイオテク
- 2) Hacia,J.G., Brody,L.C., Chee,M.S., Fodor,S.P.A. and Collins,F.S. : *Nature Genetics*, 14, December, 441 - 447 (1996)
- 3) Lipshutz,R.J., Morris,D., Chee,M., Hubblell,E., Kozal,M.J., Shah,N., Shen,N., Yang,R. and Fodor,S.P.A. : *BioTechniques*, 19, 3, 442 - 447 (1995)
- 4) Anderson,R.C., Bogdan,G.J., Barniv,Z., Dawes,T.D., Winkler,J. and Roy,K. : International Conference on Solid - State Sensors and Actuators, Chicago, June 16 - 19 (1997)
- 5) Schena,M., Shalon,D., Heller,R., Chai,A., Brown,P.O. and Davis,R.W. : *Proc.Natl.Acad.Sci.*, 93, 10614 - 10619 (1996)

- 6) Heller,R., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J., Woolley, D. and Davis, R. : *Proc.Natl.Acad.Sci.*, 94, 2150 – 2155 (1997)
- 7) Morrison,L.E. and Legator,M.S. : *Cytometry*, 1, 27 (4), 314 – 326 (1997)
- 8) Sosnowski,R.G., Tu,E., Butler,W.F., O'Connell,J.P. and Heller,M.J. : *Proc. Natl.Acad.Sci.*, 94 (4), 1119 – 1123 (1997)
- 9) Eggers,M., Hogan,M., Reich,R.K., Lamture,J., Ehrlich,D., Hollis,M., Kosicki,B., Powdrill,T., Beattie,K., Smith,S., Varma,R., Gangadharan,R., Mallik,A., Burke,B. and Wallace,D. : *BioTechniques*, 17, 3, 516 – 524 (1994)
- 10) Woolley,A.T.,Hadley,D.,Landre,P.,deMello,A.J., Mathies,R.A. and Northrup, M.A. : *Analytical Chemistry*, 68, 23, 4081 – 4086 (1996)
- 11) Woolley,A.T., Sensabaugh,G.F. and Mathies,R.A. : *Analytical Chemistry*, 69, 11, 2181 – 2186 (1997)
- 12) Colyer,C.,Tang,T.,Cheim,N. and Harrison, J. : *Electrophoresis*, 18, 1733 – 1741 (1997)
- 13) Shoffner,M.,Cheng,J.,Hvichina,G.,Kricka,L. and Wilding,P. : *Nucleic Acids Research*, 24, 2, 375 – 379 (1996)
- 14) Cheng,J.,Shoffner,M., Mitchelson,K.,Kricka,L. and Wilding,P. : *Journal of Chromatography A*, 732, 151 – 158 (1996)
- 15) Castellino,A.M. : *Genome Research*, 7, 943 – 946 (1997)
- 16) The Genesis Report/Dx, 6, 1, 1 – 6 (1996)
- 17) Arlinghaus.,H.F., Kwoka,M.N. and Jacobson,K.B. : *Analytical Chemistry*, 69,18, 3747 – 3753 (1997)
- 18) Okubo,K. and Matsubara,K. : *FEBS Lett.*, 24, 403 (3), 225 – 229 (1997)

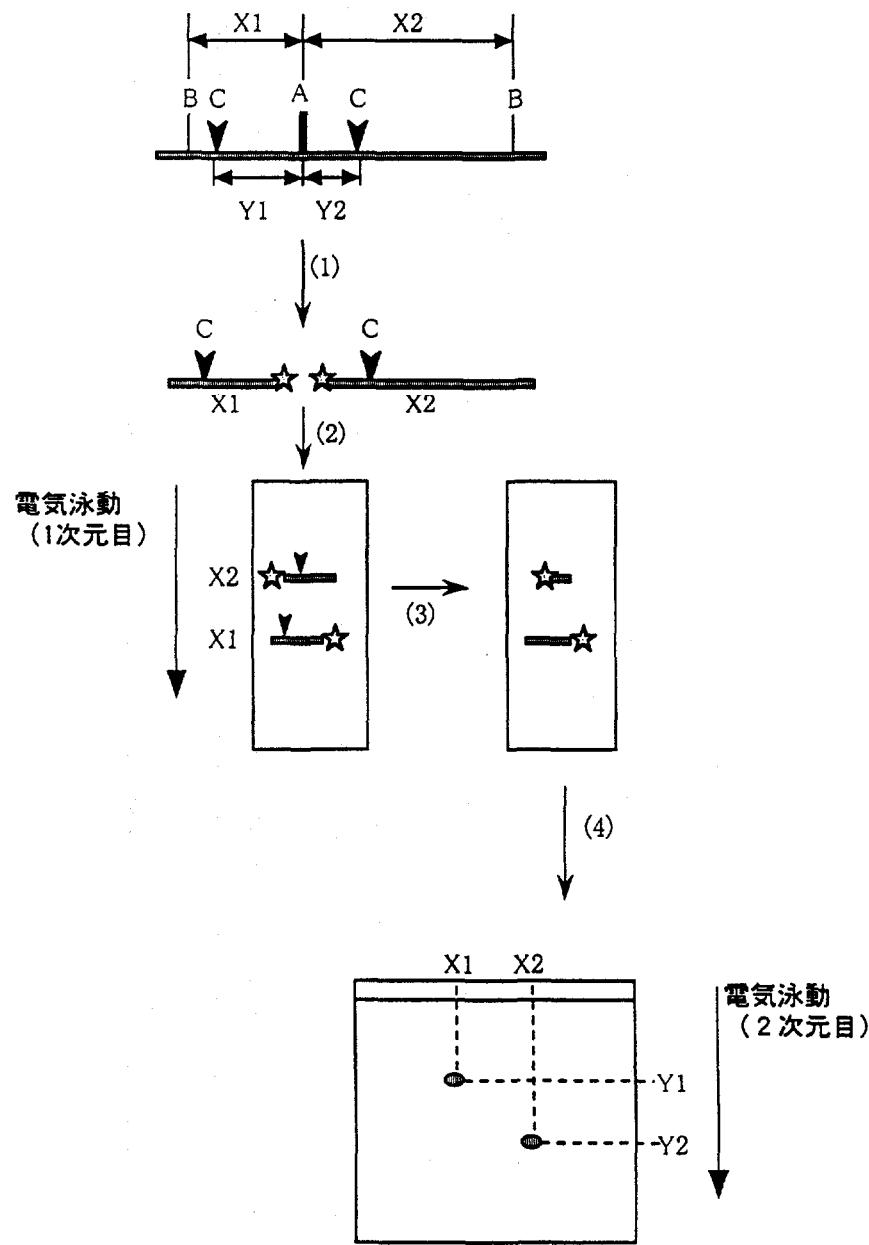
2-2 ゲノム全体の比較解析法

マクロなレベルで検出できる染色体の変化と特定の疾患を関連づけてその原因遺伝子を究明してゆく作業は以前より行われていたが、正確な座位の決定およびその配列の決定に至るまでには多大な労力を必要としてきた。研究対象となる生物の全ゲノム配列が決定されていれば、これを対照配列としてゲノムDNA上に高密度に設定したマーカー配列の有無を調べるだけで原因領域を特定することが理論上可能となる。しかしそのために塩基配列決定の高速化や多数のマーカーシグナルを検出処理するためのDNAチップ化等の解決すべき技術上の課題は少なくない。現在までに、ゲノムDNA間の異なる領域を特異的に検出し解析できるいくつかの方法が考案されているが、いずれも対象となる生物のゲノムDNAの塩基配列が未知であっても適用可能であり、ヒトのようにゲノム配列がすべて解っていない場合には、比較ゲノム解析の有力な手段となっている。ここでは、ゲノムDNAをフィンガープリントして解析する Restriction Landmark Genomic Scanning (RLGS) 法および Arbitrarily Primed PCR (AP-PCR) 法と、ゲノム間のサブトラクションにより特異的なDNAを検出する Representational Difference Analysis (RDA) 法および In-Gel Competitive Reassociation (IGCR) 法について述べる。このほか、ミスマッチ修復酵素 (MutS) を用いる方法についても紹介する。

2-2-1 Restriction Landmark Genomic Scanning (RLGS) 法

a 開発および原理 RLGS法はゲノムスキャニングすなわちゲノムDNA上の多数の座位由来のシグナルを一度に全体にわたってチェックする方法の一つとして1991年畠田らにより開発された¹⁾。ゲノムDNA上の特定 (NotIなどの低頻度制限酵素) の制限酵素サイトをランドマークとして、制限酵素断片の二次元電気泳動による分離パターンを比較する方法である(図2-11)。まず、ランドマークとするサイト以外の箇所、例えば調製されたゲノムDNA中のニックや非特異的切断末端が、標識されるのを防ぐためブロッキングを行う。ついで、ゲノムDNAをランドマークとする制限酵素で消化し、末端標識したのち電気泳動する。こうして一次元的に分画されたDNAを含むゲルを短冊状に切り出し、別の制限酵素を含むバッファー中で処理することにより、制限酵素消化断片をさらに細分化する。このゲル切片を新たに調製したポリアクリルアミドゲルの一辺に密着させ二次元目の電気泳動を行った後、オートラジオグラムを撮る。得られた数千のスポットパターンを同様に処理して得た対照のパターン

と比較することにより、ゲノムDNA上の変化を検出する。強度の異なるスポットのDNAをPCRで増幅してクローニングする手法が取られるが、スポットの違いを検出するまで、すなわち本法の本質的な部分はPCRを使わないという点で他の方法と差別化できる。したがって、他の方法でPCRの非特異的増幅等の弊害を受け、目的とするものが得られない場合には有効な手段となり得る。



- (1) ブロッキングしたDNAを制限酵素Aで切り、末端をラベルする。
断片の短小化が必要であれば制限酵素Bで切る。
- (2) アガロースゲル電気泳動をおこなう。
- (3) ゲルを短冊状に切り出し、制限酵素Cで切る。
- (4) ポリアクリルアミドゲルで2次元目の電気泳動の後、オートラジオグラフィーを行う。

図2-11 RLGS法

- b 応用 本法は感度が高くバクテリアから高等生物に至るあらゆるゲノムDNAに応用可能である。一度にスキャンニングできる染色体上の座位は数千種の restriction pattern に限られるが、異なる制限酵素部位をランドマークとして、スキャンニング可能な範囲を拡大できる。スポットの強度によりハプロイド、ディプロイドあるいはそれ以上の増幅度を判定することができるのも本法の特徴である。また、メチル化感受性酵素 (NotI, BssH IIなど) を用いればDNAのメチル化部位の違いを検出することも可能である (RLGS-M法)。操作が比較的煩雑である割には本法を応用した報告は少なくない。本法により human glioma や前立腺癌などの腫瘍細胞で遺伝子の欠失・増幅やDNAのメチル化が検出されている^{2~5)}。バクテリアのゲノム解析にも応用されており、Fujimura らは本法によりメチシリソ耐性の増強した *Staphylococcus aureus* において溶菌酵素類似の産物をコードする遺伝子の欠失を見出している⁶⁾。
- c 動向・関連技術 この方法自体あるいはこの方法に関連した特許は出されていない。
- d 課題 フィンガープリントの検出にラジオアイソトープを用いることは検出感度を高め、低いレベルの遺伝子増幅現象も解析できる利点があるが、操作の安全性・迅速性の点で問題がある。今後、検出系に蛍光法を応用することによりこれらの点が改善され解析効率が向上するであろう。また、copy number reference をとらなければならない非特異的シグナルにより false positiveを取り易いなどの問題点が挙げられる。

2-2-2 Arbitrarily Primed PCR (AP-PCR) 法

- a 開発および原理 本法は1990年に Welsh らによって開発された⁷⁾。任意の塩基配列を有するプライマーを用いて緩やかなアニーリング条件下 (40°C以下) で非特異的なPCRを行い、変性ポリアクリルアミド電気泳動で展開して各PCR産物を比較することによりゲノムDNAの欠失・増幅を検出する方法である (図2-12)。

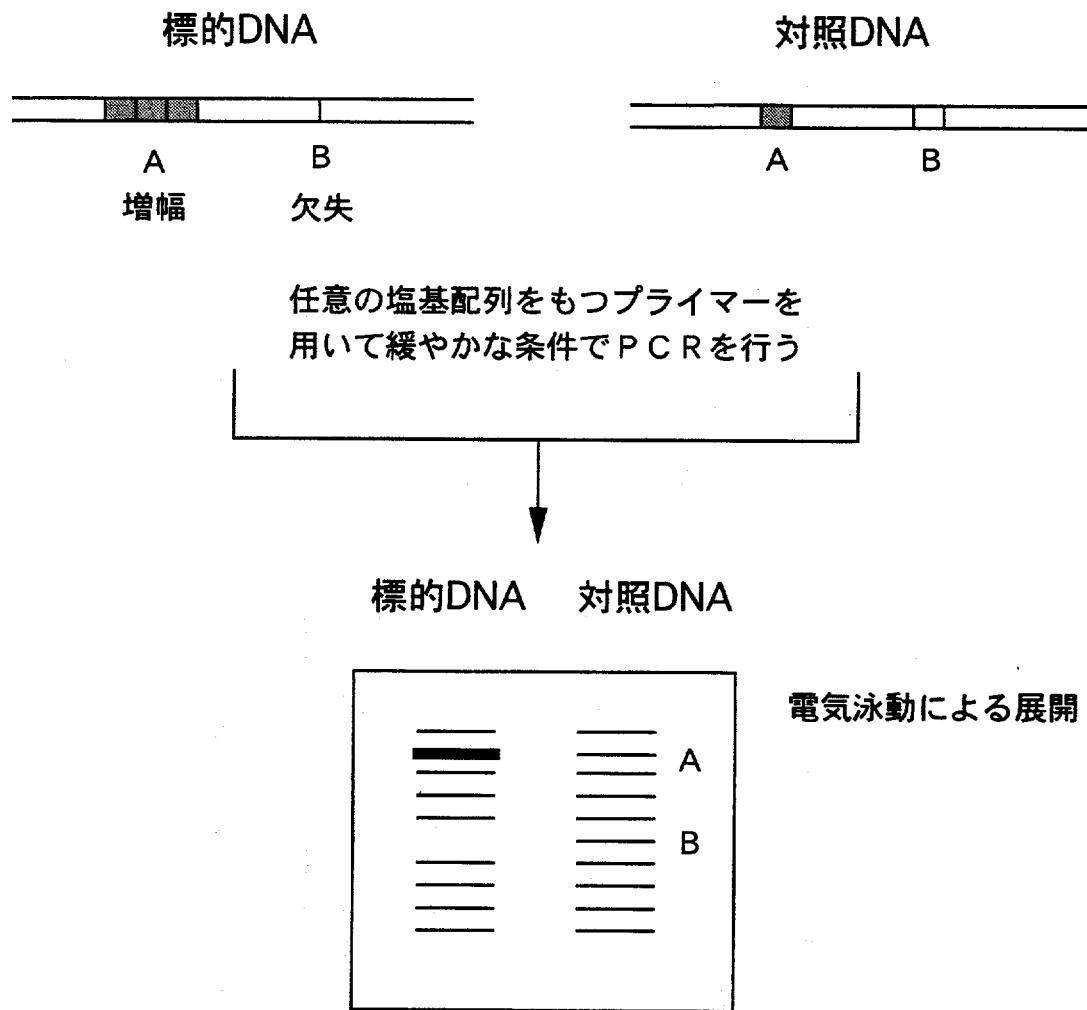


図2-12 AP-PCR法

b 応用 原法ではバクテリアのstrainの比較解析に応用されているが、高等生物を含むあらゆるゲノムDNAに応用可能である。本法はPCRを用いた方法であり、必要な検体DNA量が少量ですむため、臨床サンプルの診断的解析への応用も期待されている⁸⁾。また、検出されたバンドからのDNAのクローニング化が容易にできるのも本法の利点の一つである。AP-PCR法の原理をRT-PCR法に応用したのが Differential

Display (DD) 法⁹⁾ であり、この DD 法は発現パターンを指標にした遺伝子クローニング法として現在広く活用されている。

- c 動向・関連技術 AP - PCR 法を応用した生物種の同定法・分類法 (STRATAGENE 社) や腫瘍細胞の同定法 (CALIFORNIA INST. BIOLOGICAL RES.) などに関する特許が出願されている。
- d 課題 プライマーを変えることで多数の DNA 領域について染色体の欠失・増幅を検出することが理論的に可能であるが、実際には数百 Kb 以内でホモ欠失している領域を検出するためにはかなり多数のプライマーセットが必要となるようである。また、数少ない塩基の欠失や挿入、染色体の転座などは検出が困難である。したがって、遺伝子異常を幅広くスキャニングするためには、プライマーの配列や組み合わせなど工夫すべき点が多い。

2 - 2 - 3 In - Gel Competitive Reassociation (IGCR) 法

- a 開発および原理 ゲノムサブトラクション法は、2つのゲノム DNA を混合、変性、再会合して、片方にのみ存在する DNA 断片を濃縮する方法で、原理的には cDNA サブトラクション法と同様である。1987年に筋ジストロフィーの一つであるDMDに関連した X 染色体上の DNA を分離された報告¹⁰⁾ が最初である。しかしながら、この方法は DNA 断片の濃縮効率が低くゲノム DNA のサブトラクションがうまくいかないという欠点を有していた。この欠点を克服するため、PCR 法を応用した IGCR 法や後で述べる Representational Difference Analysis (RDA) 法などが開発された。IGCR 法は 1990 年に横田らにより開発された方法¹¹⁾ で、その後木山らにより改良され、高等生物のゲノム解析に適用しやすくなった¹²⁾。まず、Biotin 標識した一方のゲノム DNA 断片と大過剰の他方の脱リン酸化した未標識ゲノム DNA 断片を混合後、各 DNA 断片の実効濃度を上げるために電気泳動により各々の DNA 断片をサイズ分画する。ゲル中で DNA 断片を変性させた後、再会合させて DNA を回収し、アダプターを接着後、avidin - trap で目的とする DNA 断片を濃縮して、特異的に PCR で増幅する (図 2 - 13)。

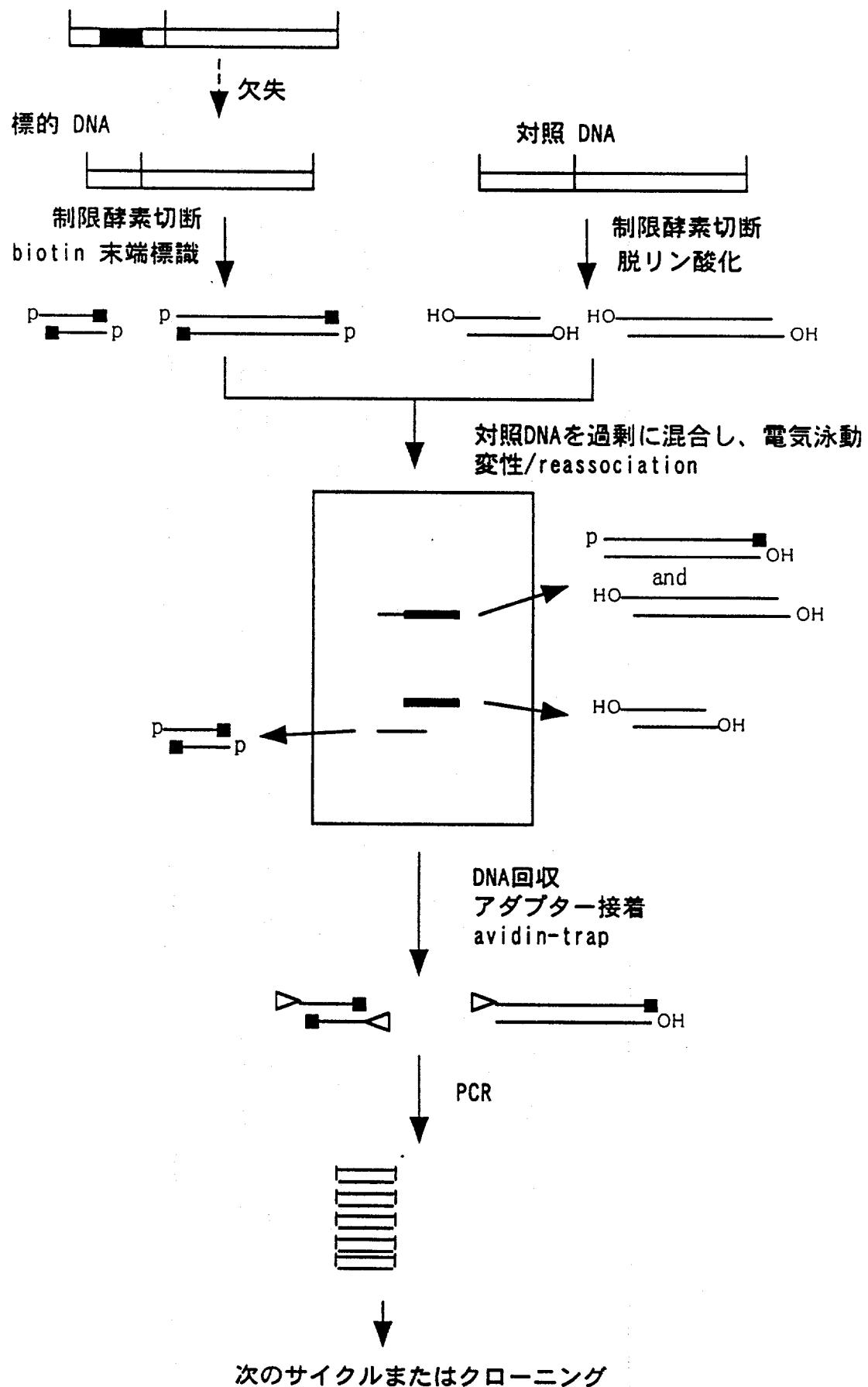
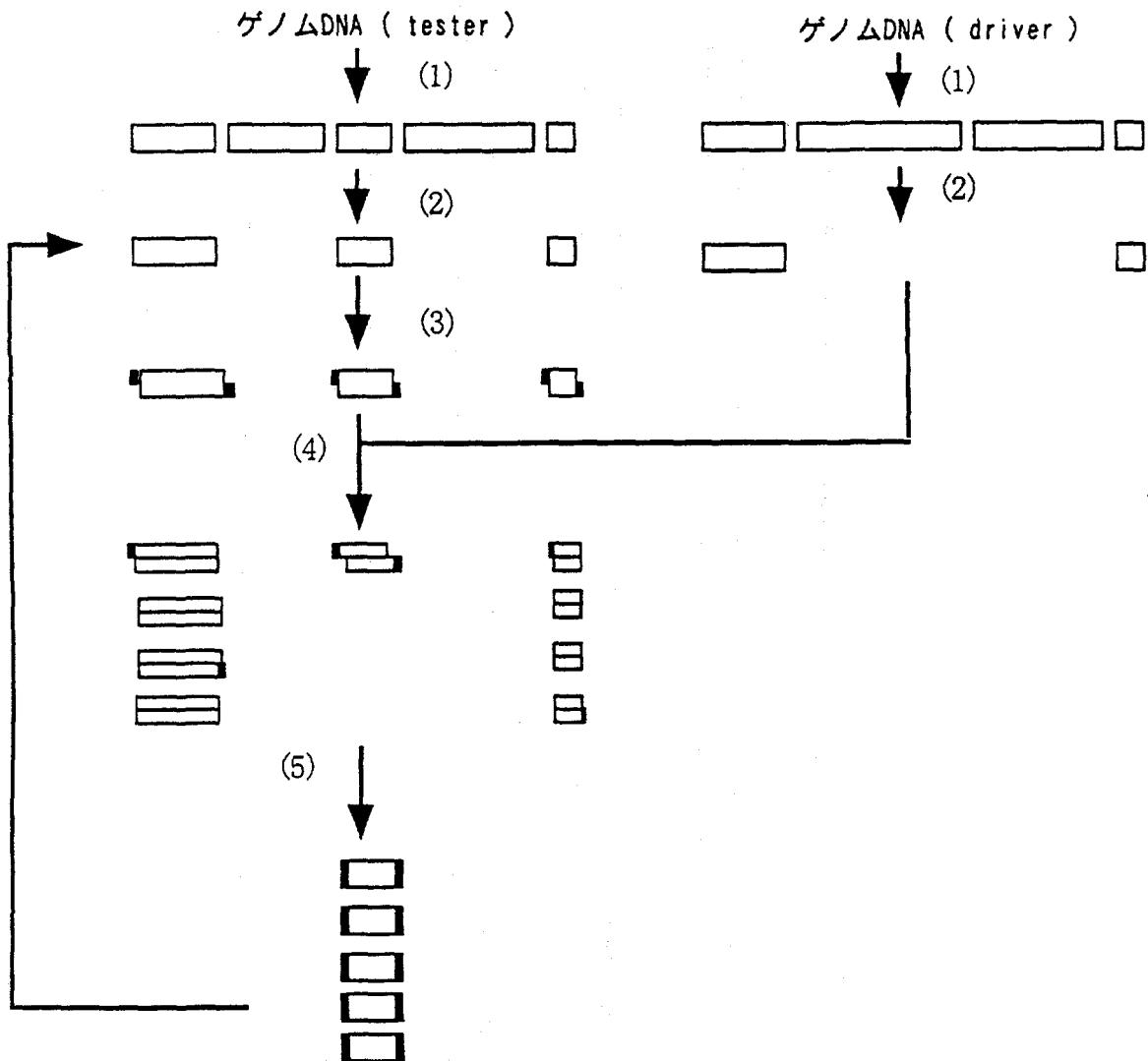


図 2-13 IGCR 法

- b 応用 主にヒトおよびマウスゲノムを対象として利用されている。多型マークーの濃縮・分離に有効でヒトおよびマウスで多型マークーのライブラリーを構築したとの報告がある¹³⁾。また、メチル化DNA認識・非認識制限酵素の使い分けでメチル化部位をもつ断片を濃縮したり、メチル化部位の違いに注目した subtraction ライブラリーの構築が行なわれた¹⁴⁾。
- c 動向・関連技術 この方法自体あるいはこの方法に関連した特許は出されていない。
- d 課題・将来性 電気泳動による分画でDNA断片の実効濃度を高めるところが本法の特徴である。したがって、RLGS法と同様に、電気泳動およびゲル処理の操作性に問題がありながらも、原理的に必須のステップであるので簡略化は難しい。DNA断片の再会合を効率よく行わせるために cetyltrimethylammonium bromide (CTAB)などを用いた改良研究が報告されている¹²⁾。

2-2-4 Representational Difference Analysis (RDA) 法

- a 開発および原理 本法は1993年、Lisitsyn ら¹⁵⁾により考案された。ゲノムDNAの制限酵素消化断片にアダプターを接着し、アダプターの塩基配列でPCRを行うことにより、200~2000bpほどのサイズのDNA断片のみを増幅してアンプリコンを作製する。このようなアンプリコンは全ゲノムの1/10~1/100の複雑度であり、アンプリコンを用いることにより競合的ハイブリダイゼーションを、ゲノム全体を使用した場合に比べ、効率よく再現性よく行えるようになった。同様の操作により得られた対照アンプリコンと混合した後、変性・ハイブリダイゼーション・PCRを行うことで、電気泳動（分画操作）を行わずに最終的にゲノムの欠失・再編成領域等を特異的に増幅し検出する（図2-14）。



- (1) 制限酵素で切断する。
- (2) アダプターを接着後PCRを行ってampliconを調製する。アダプターは除去する。
- (3) tester DNAのみに新しいアダプターを接着する。
- (4) driver DNA 過剰で競合的ハイブリダイゼーションを行う。
- (5) tester DNA 特異的に 存在するampliconすなわち両端にアダプターをもつ断片をPCRを行って増幅する。
- (6) その後アダプターを除去し、(3)から(5)のサイクルを繰り返す。

図2-14 RDA法

- b 応用 多型マーカーの分離（ヒト、ラット、マウス）^{15~17)}、疾病関連遺伝子の構造変化の解析（マウス肺臓癌、腎臓癌・大腸癌培養細胞）^{18~20)}、メチル化の検出（マウス肝癌）²¹⁾、ウイルスの検出（ヘルペスウイルス等）²²⁾など、高等生物を中心に多数の応用例が報告されている。
- c 動向・関連技術 Cold Spring Harbor 研究所から、二つのゲノムDNAから異なる配列を検出するためのプローブ製造法として本法の特許が公開されている。
- d 課題・将来性 本法は、一回の試行で全ゲノムの1/10 – 1/100のDNA断片を対象とするためすべての変異を検出することはできないが、電気泳動（分画操作）を行わず直接目的の断片を獲得できる点で操作性に優れている。最近、ゲノム中の繰り返し配列の primer²³⁾ や arbitrarily primer²⁴⁾ を使って amplicon の調製を行う方法が報告された。また、特定の遺伝的多型あるいは表現形質に注目して集めた固体のDNAをプールして RDA 法を行い、注目した染色体座位あるいは表現形質を支配する座位に選択的にマーカーを分離する方法²⁵⁾ も開発されるなど本法の優れた応用性、将来性の一端をうかがわせる。

2-2-5 ミスマッチ特異的結合タンパク質 MutS を用いた方法

MutSは大腸菌 (*E. coli*) のDNA複製時に起こるミスマッチを修復するための酵素系の一つで、DNA上のミスマッチ部位、例えばsingle base pair mismatch や short deletion によるgap に、*in vivo* および *in vitro* で特異的に結合する性質をもっている²⁶⁾。この性質を利用して特定の領域における変異を検出する試みがなされているが、ゲノムDNAレベルでの応用の報告例はなく、まだ精度の点で問題があると考えられる。応用例として H-ras, K-ras, cystic fibrosis 原因遺伝子等の変異検出が挙げられる^{27, 28)}。この他に改良研究として Geschwind らは MutS に biotin 化したペプチドを融合した蛋白質 (MutSb) を作製し、avidin-coated magnetic beads を使ってミスマッチを含むDNA断片を効率良く濃縮することに成功している²⁹⁾。MutSに関連した特許として、

- 1) 点突然変異の検出のための MutS 融合蛋白質 (SRL KK)
- 2) 耐熱性 MutS (New York 大学)
- 3) ヒト由来のミスマッチ修復酵素 (Johns Hopkins 大学、De La Chapelle)
- 4) 固定化 MutS (Gene Check Inc.)³⁰⁾
- 5) 診断への応用 (Upstate Biotechnology Inc.)

6) genomic mismatch scanning による Genetic mapping (Stanford 大学)
等が出願されている。

今後、これらの特許で支持されているような応用価値を実現化していくことに課題が置かれる。MutS蛋白のミスマッチ部位への結合親和性の向上や固定化などにより変異検出キットや遺伝子診断キットへの応用が期待される。一方で、DNAのreassociationの効率や、ノイズの改善が実用化に向けて要求される。現状では、特定の遺伝子などの限定された範囲の変異検出には有効であるが、ゲノム全体が対象範囲になった場合には変異箇所の濃縮分離は困難であろう。

文献

- 1) Hatada, I., Hayashizaki, Y., Hirotsume, S. and Komatsubara, H. : *Proc. Natl. Acad. Sci. USA*, 88, 9523 – 9527 (1991)
- 2) Nakamura, M. Konishi, N., Tsunoda, S., Hiasa, Y., Tsuzuki, T., Aoki, H., Kobitsu, K., Nagai, H. and Sakaki, T. : *J. Neurooncol.*, 35, 113 – 120 (1997)
- 3) Konishi, N., Hiasa, Y., Nakamura, M., Kitahori, Y., Matsubara, K. and Nagai, H. : *Am. J. Pathol.*, 150, 305 – 314 (1997)
- 4) Kromitsu, J., Kataoka, H., Yamashita, H., Muramatsu, M., Furuichi, Y., Sekine, T. and Hayashizaki, Y. : *DNA Res.*, 2, 263 – 267 (1995)
- 5) Miwa, W., Yashima, K., Sekine, T. and Sekiya, T. : *Electrophoresis*, 16, 227 – 232 (1995)
- 6) Fujimura, T. and Murakami, K. : *J. Bacteriol.*, 179, 6294 – 6301 (1997)
- 7) Welsh, J. and MaClelland, M. : *Nucleic Acids Res.*, 18, 7213 – 7218 (1990)
- 8) Arribas, R., Capella, G., Tortola, S., Masramon, L., Grizzle, W.E., Perucho, M. and Peinado, M.A. : *J. Clin. Oncol.*, 15, 3230 – 3240 (1997)
- 9) Liang, P. and Pardee, A.B. : *Science*, 257, 967 – 971 (1992)
- 10) Smith, T.J., Wilson, L., Kenrick, S.J., Forest, S.M., Speer, A., Coutelle, C. and Davis, K.E. : *Nucleic Acids Res.*, 15, 2167 – 2174 (1987)
- 11) Yokota, H. and Oishi, M. : *Proc. Natl. Acad. Sci. USA*, 87, 6398 – 6402

(1990)

- 12) Kiyama, R., Inoue, S., Ohki, R., Kikuya, E., Yokota, H. and Oishi, M. : *Adv. Biophys.*, 31, 151 – 161 (1995)
- 13) Inoue, S., Kiyama, R. and Oishi, M. : *Genomics*, 31, 271 – 276 (1996)
- 14) Mizuguchi, R., Inoue, S., Kiyama, R. and Ohishi, M. : *DNA Res.*, 2, 219 – 223 (1995)
- 15) Lisitsyn, N., Lisitsyn, N. and Wigler, M. : *Science*, 259, 946 – 951 (1993)
- 16) MacCarthy, L., Hunter, K., Schalkwyk, L., Riba, L., Anson, S., Mott, R., Newell, W., Bruley, C., Bar, I. and Ramu, E. : *Proc. Natl. Acad. Sci. USA*, 92, 5302 – 5306 (1995)
- 17) Toyota, M., Canzian, F., Ushijima, T., Hosoya, Y., Kramoto, T., Serikawa, T., Imai, K., Sugimura, T. and Nagao, M. : *Proc. Natl. Acad. Sci. USA*, 93, 3914 – 3919 (1996)
- 18) Schutte, M., da Costa, L. T., Hahn, S. A., Moskaluk, C., Hoque, A. T., Rozenblum, E., Weinstein, C. L., Bittner, M., Meltzer, P. S. and Trent, J. M. : *Proc. Natl. Acad. Sci. USA*, 92, 5950 – 5954 (1995)
- 19) Lisitsyn, N. A., Lisitsyn, N. M., Dalbagni, G., Barker, P., Sanchez, C. A., Gnarra, J., Linehan, W. M., Reid, B. J. and Wigler, M. H. : *Proc. Natl. Acad. Sci. USA*, 92, 151 – 155 (1995)
- 20) Kastury, K., Baffa, R., Druck, T., Ohta, M., Cotticelli, M. G., Inoue, H., Negrini, M., Rugge, M., Huang, D., Croce, C. M., Palazzo, J. and Huebner, K. : *Cancer Res.*, 56, 978 – 983 (1996)
- 21) Ushijima, T., Morimura, K., Hosoya, Y., Okonogi, H., Tatematsu, M., Sugimura, T., Nagao, M. : *Proc. Natl. Acad. Sci. USA*, 94, 2284 – 2289 (1997)
- 22) Challoner, P. B., Smith, K. T., Parker, J. D., MacLeod, D. L. and Coulter, S. N., Rose, T. M., Schultz, E. R., Bennett, J. L., Garber, R. L. and Chang, M. : *Proc. Natl. Acad. Sci. USA*, 92, 7440 – 7444 (1995)
- 23) Elango, R., Riba, L., Housman, D. and Hunter, K. : *Mamm. Genome*, 7, 340 – 343 (1996)

- 24) 吉田幸成, 牛島俊和, 今井浩三, 杉村隆, 長尾美奈子, 第20回日本分子生物学会年会
講演要旨集, 3 - 509 - P - 553, (1997)
- 25) Lisitsyn, N. A., Segre, J.A., Kusumi, K., Lisitsyn, N. M., Nadeau, J.H.,
Frankel, W.N., Wigler, M. H., and Lander, E.S. : *Nature Genet.*, 6, 57 -
63 (1994)
- 26) Bellanne - Chantelot, C., Beaufils, S., Hourdel, V., Lesage, S., Morel, V.,
Dessinais, N., Le Gall, I., Cohen, D. and Dausset, J. : *Mutat Res.*, 382, 35
- 43 (1997)
- 27) Parsons, B. L. and Heflich, R. H., : *Mutat Res.* 374, 277 - 285 (1997)
- 28) Lishanski A, Ostrander, E. A. and Rine, J. : *Proc. Natl. Acad. Sci. USA*,
91, 2674 - 2678 (1994)
- 29) Geschwind D. H., Rhee, R. and Nelson, S. F., : *Genet. Anal.*, 13, 105 - 111
(1996)
- 30) Wagner, R., Debbie, P. and Radman, M. : *Nucleic Acids Res.*, 23, 3944 -
3948 (1995)

2-3 特定領域遺伝子の比較解析

2-3-1 序文

特定領域・遺伝子の比較解析を迅速に行なうために様々な手法が開発されてきたが、そのほとんどがPCRによるゲノムからのターゲット遺伝子の増幅が前提となっている。PCRプロダクトを直接シークエンスするのが一番確実で情報も多く得られるが、コスト、簡便性、処理能力を目指すということで以下4つの手法を紹介する。

2-3-2 PCRを基にした様々な解析法

(1) SSCP (Single Stranded Conformation Polymorphism) 法

たとえばPCRによって増幅された2本鎖DNAを1本鎖に加熱変性し、通常の（尿素などの変性剤を含まない）ポリアクリルアミドゲル中での電気泳動によって分離する方法である。ゲル中において1本鎖DNAは分子内の複雑な相互作用によって塩基配列に依存した高次構造をとり変異配列は対照となる正常配列とは異なった三次元構造をもつ。これが原因で両者はゲル電気泳動中に異なった移動度を示し、そのバンドの位置をることにより変異の有無を判定する。この方法は操作が簡単ではあるが未知の変異が存在しないことの証明に用いることはできない（図2-15）。

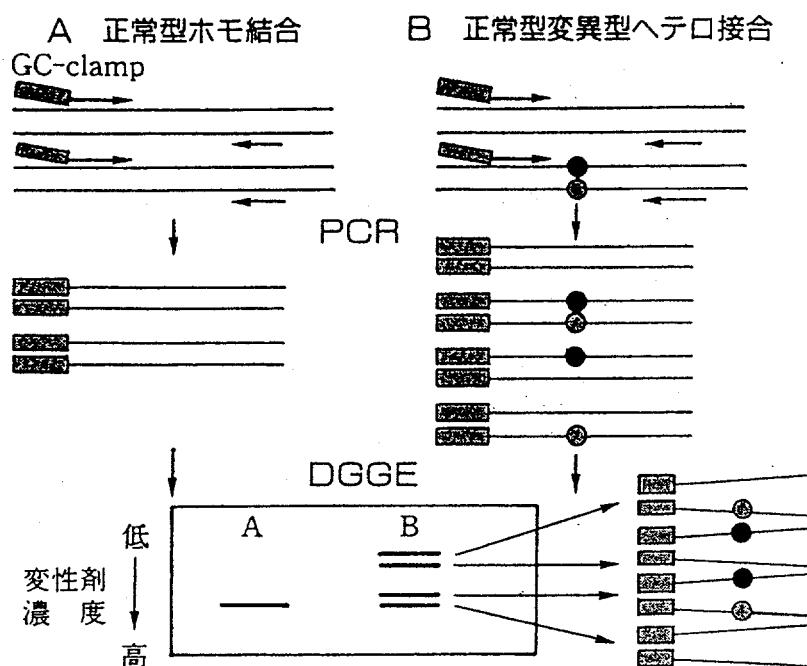


図2-15

(2) DGGE (Denaturing Gradient Gel Electrophoresis) 法

まずGCクランプのついたプライマーでPCRを行ない、増幅産物の片側に非常に安定な領域を付加する。そしてこの2本鎖DNAがポリクリルアミドゲル電気泳動中に温度及び変性剤によって変性し、Y字構造をとるが、その際移動度が非常に小さくなる。変性剤の濃度勾配をもつゲルにて電気泳動することにより標的配列に1塩基でも変異があれば変性の始まる濃度が異なり電気泳動的に正常配列と変異配列とを分離することができる。これを利用して配列の違い、ミスマッチの存在を検出する。

(3) RFLP (Restriction Fragment Length Polymorphism) 法

標的配列をPCR増幅後、制限酵素によって切断し、そのパターンの違いにより変異を検出する方法であるが、変異の位置が塩基配列上、制限酵素のサイトに一致していなければ検出できない欠点がある。

(4) CFLP (Cleavase Fragment Length Polymorphism) 法

最近開発された標的遺伝子の変異解析法である。これはPCR法によって特定領域を増幅し、熱変性したのち冷却する。すると同一分子内で2次構造を形成し、特異的なヘアピン構造をもつ1本鎖DNAとなる。その構造はDNAの配列に依存し、点変異が1ヶ所あるだけでも2次構造は変化する。この1本鎖DNAをCleavaseIで部分消化する。CleavaseIはヘアピン構造やループ構造を形成している1本鎖DNAの1本鎖部分と2本鎖部分を特異的に認識し、その5'側を切断する。(図2-16) 切断後、電気泳動を行うと正常型と変異型では異なるパターンを示すことになり両者を比較することにより変異の検出やタイピングを行なうことができる。(図2-17) 特長としては100~2000bpのPCR productの変異の検出が可能であり変異のおおよその位置もわかる。

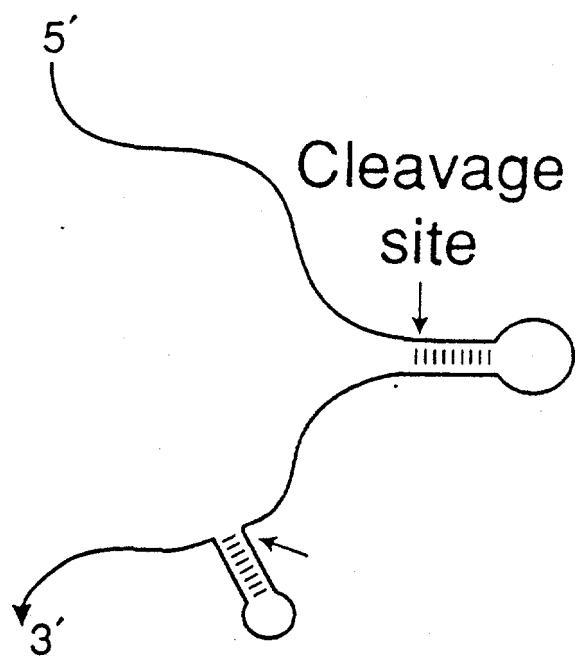


図2-16

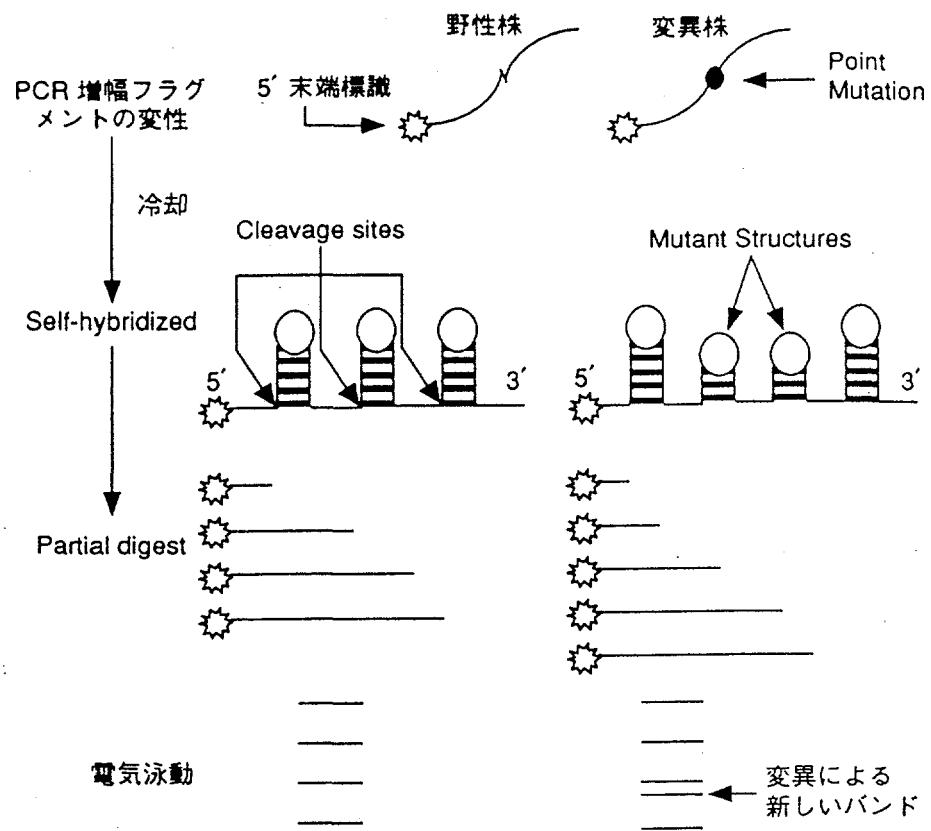


図2-17

2-3-3 課題と将来性

以上4つの手法を概説したが、他にも様々な方法が報告されている。しかし現在実績のある特定領域の変異検出法としては上記4つの方法に代表されると思われる。いずれの方法もコストダウンのための自動化、そして再現性の良い結果を得るために技術開発が必要である。そのため研究レベルとしては広く用いられているが、医学の分野で不特定多数の患者を対象に、上記手法を用いて遺伝子診断がルーチンに行なわれている現状ではない。

2-4 遺伝子増幅法の多様化

2-4-1 遺伝子増幅法の現状

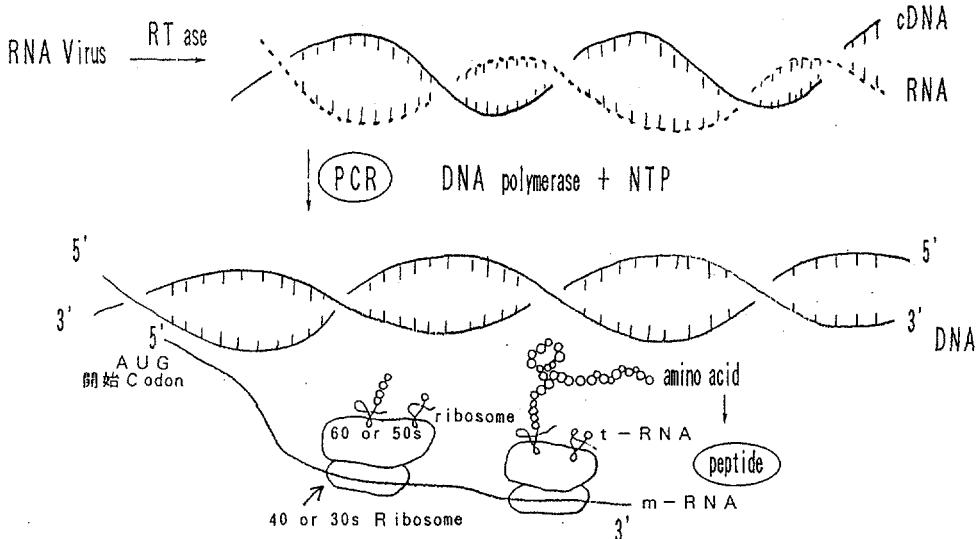
近年、耐熱性polymeraseが応用され、polymerase chain reaction (PCR) による遺伝子の増幅法が開発され、遺伝子解析技術となり、thermocyclerによる遺伝子増幅法はPCR法として基礎的にも応用実験にも使用された。遺伝子レベルでの研究や分子生物学的研究においても技術の大半がPCR法により成果が上げられてきており、その技術的寄与は計り知れないものがある。このPCR技術の臨床的応用も著しく進歩し、特に臨床検査の分野において、このPCR技術を応用して、疾病の原因となる遺伝子の検出を行ない、その原因の解明や治療の指標として多くの知見を得ている。

2-4-2 臨床検査で利用されている遺伝子検出法

(1) PCR法

これらのPCRを基本技術とした臨床検査機器システムはAmplicore[®]等であり細菌やウイルス感染症、癌遺伝子等の検出・定量に用いられている。しかし、このPCR法は特許的にRoche社におさえられており、後より追随して発明されている技術はPCRとは異なるメカニズムであるが、しかし当然ながらその原理はすべてDNA情報による機能蛋白質の生合成、いわゆるセントラルドグマに起因している。その例として、RNAウイルスがDNAを複製し、各種機能性タンパク分子を合成するRNAウイルスの複製機構にが挙げられる（図2-18）。

PCR 法は DNA polymerase を使用するため、複製すべき鑄型 DNA 分子とその primer 及び核酸分子と耐熱性 polymerase によって構成され、thermocycler を使用することにより annealing と hybridization を繰り返し、合成される検出すべき “鑄型” DNA の 2^n 乗的スピードによる遺伝子の増幅を伴なう (図 2-19)。



RNA ウィルス → タンパク合成 (ウィルス構築)

DNA / RNA probe はこれらの process のどこかの DNA 又は RNA を利用して検査されている

図 2-18

技術比較

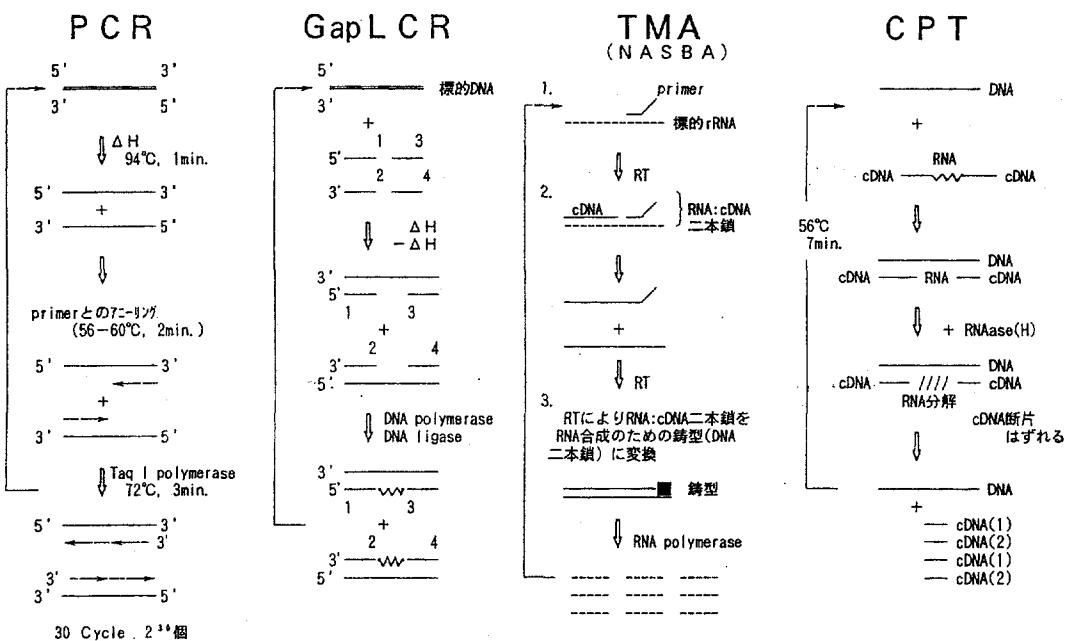


図 2-19

(2) GapLCR法

これに対し、第2のPCRに対抗すべき方法として、DNA ligaseを用いてのAbbott社のGap ligase chain reactionは検出すべき遺伝子のsequenceに相補性のある各鎖のDNAを相補的sequenceで2blockにして合成をしておき、目的の検出すべきDNAにhybridizeして、その間をDNA ligaseで合成して埋めていき合成したcDNAをannealingすることにより各一本鎖DNAとして更に、ligaseにより連結部分をつなぎ1本の完成された相補的DNAを合成し、この方法も又、 2^n 乗的にDNAが増幅されて検出されるべきDNAのsequenceが増加する。

(3) TMA/NASBA法

第3の方法は、増幅されるRNAを検出する方法であり、前2者のDNAbaseのものより増幅効率の良いRNAを検出するものであり、耐熱性RNA polymeraseを使用するtranscription mediated amplification (TMA)法と呼ばれるものであり、Gene probe社のTMAやOrganoteknica社のNucleic Acid Sequence Based Amplification (NASBA)法がこの分類に入る。これらのRNA増幅法は、DNAよりRNAのほうが分子数が多いため、感度がよいとされている。これらの方法は、RNAをreverse transcriptase (RT)により、primer-1をのばしDNAに読み替え、cDNA-RNAcomplexを作成し、更にprimer-2でDNA-DNA二重ラセンをreverse transcriptase (RT)で合成し、合成されたDNA二本鎖よりT7 RNApolymeraseで多くのRNA断片を合成させ、そのRNAを増幅することにより標的RNAを感度良く検出する。

(4) CPT法

前述の遺伝子診断法がその増幅法においてすべてDNAもしくはRNAをDNA polymeraseやRNA polymeraseによって、核酸やデオキシリボ核酸をDNA情報に従って酵素的に増幅する方法になっている。従って、もし、外的因子やあるいは増幅条件で誤ったsequenceができると増幅はその誤った方向に合成されてしまう。これを一種の読み取りミスとする。特に遺伝子を臨床検査等のように他のタンパクや遺伝子が多く存在する中で増幅する場合が多い例の時は大きな問題となる。これらの誤読をさせない方法として次に述べるCRT (Cycling Probe Technology)のような増幅法は臨床検

査における遺伝子診断法として重要である。CRT法ではあらかじめ、既知の遺伝子情報data baseに従って化学合成的に遺伝子を合成しておく。更にsequenceとして、DNA各10残基程度とRNA5残基を、DNA - RNA - DNAの順に合成しておき、このsequenceは目的とする標的遺伝子に相補的なsequenceでありこれをhybridizeさせて、すぐに共存する耐熱性酵素RNaseで処理することによりRNAの部分を切る。両端に残るDNA部分は小断片となり結合力を失い遊離する。この断片化したDNAのfragmentを検出し、その増幅法とする。あらかじめsequenceを化学合成し、その分解反応を利用して検出をする方法である。この方法は次の図のごとく行なわれ、その断片化したfragmentは電気泳動法にてラベル化法で検出する(図2-20, 表2-2)。

遺伝子診断 PCR と CPT の 技術比較

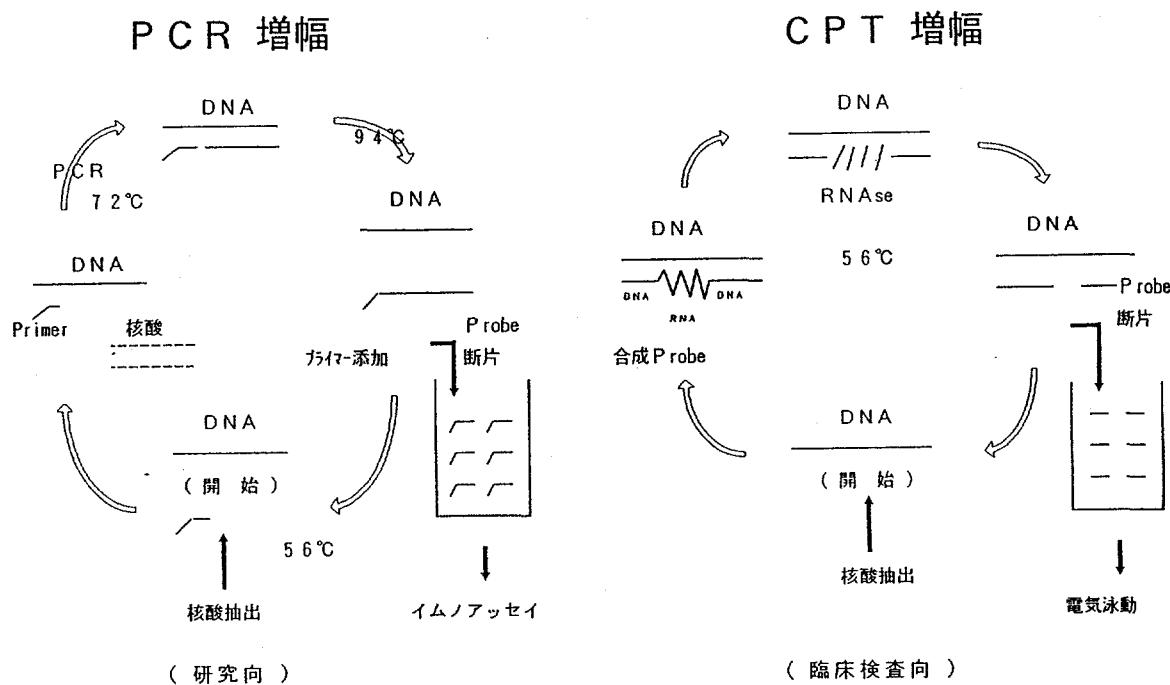


図2-20

表2-2

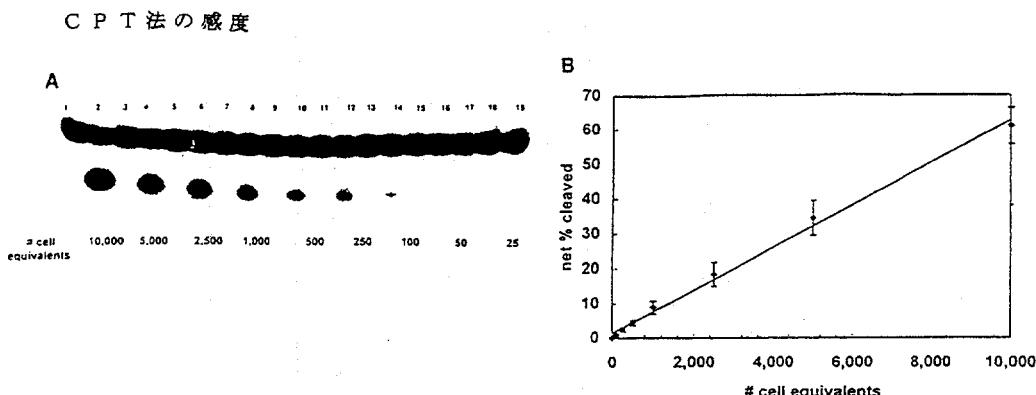


FIG. 6. (A) Sensitivity of the cycling probe reaction. Lanes 2, 4, 6, 8, 10, 12, 14, 16, and 18 show cycling reactions performed with DNA isolated from *M. tuberculosis* cells and diluted to the number of cell equivalents indicated. Lanes 3, 5, 7, 9, 11, 13, 15, 17, and 19 similarly show reactions with *M. gordonae* DNA. Lane 1 shows a control reaction with probe and RNase H, but no target DNA. The upper band represents intact probe, and the lower band represents cleaved probe fragments. (B) Linearity of the cycling probe reaction. The percent probe cleaved at each of the number of cell equivalents for *M. tuberculosis* in panel A was corrected for the background cleavage in reactions containing the same amounts of *M. gordonae* DNA to give the net percentage cleaved of cut probe. The average of seven independent experiments is plotted, and the standard deviation between experiments is indicated by the error bars.

J. Clin. Microbiology, 34. 12. 2985. より引用

表2-3

D N A 診断製品市場

用途 遺伝子 增幅	未知 D N A 遺伝子	既知 D N A 遺伝子
有	Roche PCR Abbott LCR Becton Dickinson SDA Organo technika NASBA	ID Biomedical CPT (臨床検査向)
無	3 r d Wave	Gene Probe Chiron Digene (市場開発中)

2-4-3 各方法論の比較検討

遺伝子診断製品として、特に臨床検査法に関する限り遺伝子の增幅を非増幅法を含めて、現在市場にでているのは、Becton Dickinson社のSDAや、Organotechnika社のNASBA法のごとく現在臨床応用検討中のものを含めて、表2-3に示すものである。

市場的には、日本において昨年度調査において55~60億円くらいの試薬の売上げ高と報告されている。この他にこれらの遺伝子診断をサポートする測定機器やその周辺機器を含めて75億円くらいのビジネスとされている。しかし、これらの総市場がこれから臨床検査における遺伝子診断のウェートが大きくなるにつれて急激に市場規模を拡大していくものと考えられる。また、これらの製品で言えることは、遺伝子の検出に項目や目的によっては必ずしもプロセス的に難のある、コスト高につながる遺伝子増幅法をとらなくても良いものもある。

特に感染症領域において単に感染源を検出する目的の場合は（例えばC型肝炎等）なにも増幅法に依らなくても、その検出法にケミルミ法等の高感度検出を利用すれば感染の有無は判定できる。その代表的なものにGene Probe社やChiron社のChistomas tree法が挙げられ、更にDigene社のHybrid Capture法が挙げられる。hybrid capture法は、主にhuman papiloma virus等に有力な手法となっている。また、Third wave社も特殊なNucleaseによりDNAのfragmentation patternによりpoint mutationやbacteriaやvirusの変更株を同定することができるものがある。

DNAやRNAの増幅において表-1に示したものがあるがRoche PCR法が現在一番多く普及しており商品名Amplicoreとして販売されている。また、更にAbbott社のLCRも特にSexual Transmitted Disease関連で臨床的に重要であるChlamydia trachomatisや、Neisseria gonorrhoeaeらの製品が出されている¹⁾。

また、Gene Probe社のTMA法は特に培養の困難な項目で結核菌の検査としての製品が出されている。

これらの製品が日本における遺伝子検査市場の拡大を狙って、製品群の開発を加速している。また、PCR等の特許に抵触しない他のグループとして、Becton Dickinson社SDA、Organotechnika社NASBAのように現在治験中のものも将来市場にでてくるものと考えられる²⁾。

更に、これらの方法はその技術的な精度、感度の高さとその上に臨床検査法としてのコストの安さが比較検討されなければならない。技術的には今後市場に登場すると見られる

ID Biomedical 社のCPT 製品等は技術的な面よりこれらのコスト面からは有利な立場にある製品と考えられる。

2-4-4 遺伝子診断の今後の展開

遺伝子診断は今後の展開としては現在のルーチン臨床検査に実施されている肝機能等に代表される GOT や GPT の生化学検査、感染症や癌マーカー等の高感度、特異的な免疫的検査の他に全く別の方法論である遺伝子解析分野が拡大するのは必至であり、これらに関する製品開発も急ピッチで進んでいる。現在は PCR やその他の技術をベースとした方法論の開発が急ピッチであるが、更にこれらの技術に立脚した新しいフォーマット、検出法や情報の拡張化が順次進んでいる。例を挙げれば、遺伝子断片の microtip 化、更により簡便なフォーマットによるコストエフェクティブな迅速検査法等が考えられる。前者では DNA 断片をウェハー上に固定化し各 space に地番をふってその検査による hybridization の状況のデータをとったりその検出を mass spectrometry に検出する方法が報告されている³⁾。

更に、 microparticle や金コロイド担体を使用し homogeneous に反応し、検出をする粒子固定型遺伝子検査が考えられる⁴⁾。

今後は近い将来に human genome の full sequence の解明がされると考えられ、個人の疾患状態や疾患の予知を考える上で非常に重要な情報となることは自明であり、これらの情報が予防医学につながれば医療経済に対するプラス効果につながるものと考えられ遺伝子検査の重要度はますます増大すると予想される。しかし、遺伝子情報は個人の秘密に属する情報でありたとえ医療診断といっても個人との了解の上で情報処理、インフォームドコンセントの重要さも増してくるものと考えられる。

文 献

- 1) 能澤 一他.西日本泌尿器科. 58, 5, 600 (1996)
- 2) B. van Gemen *et al.* : *Medical Virology*, 5, 205 (1995)
- 3) J. Monforte *et al.* : High - through put DNA analysis by time - of - flight mass spctrometry, *Nature Medicine*, 3, 3, 360 (1997)
- 4) R. F. Service : *Science*, 277, 22, 1036 (1997)

第3章 特異的発現様式解析技術

3-1 DNAチップによるもの

遺伝子の特異的発現様式解析技術には大別して2つの技術がある。1つは3-2章、3-3章に述べるように、電気泳動やシーケンシング等の従来技術を用いて発現遺伝子を分類、同定する方法で、代表的なものとしてはボディマップ、SAGE、Differential Display等があげられる。一方、ハイブリダイゼーションを基本とした比較的新しい技術として、DNAチップがある。この技術はAffymetrix社を筆頭に、アメリカの大学やベンチャー企業を中心開発が進められており、現在最も注目されている分野のうちの一つである。第2章に代表的なDNAチップの基本概念や遺伝子診断への応用例について述べたが、この章では遺伝子発現様式解析や第2章では触れなかった新技術にターゲットを絞り、各種 microarray の開発状況について述べる。

3-1-1 Oligonucleotide array

Oligonucleotide array を開発している代表的な企業は Affymetrix 社である¹⁻³⁾。Chip上で、任意の配列のオリゴヌクレオチドを半導体作製に用いられているフォトリソグラフィー技術を用いて合成するもので、Affymetrix社副社長 Thomas R. Gingeras の言葉を借りれば、小さな硬貨ほどの面積に 450,000 種類のオリゴヌクレオチドを合成することができるということである。これにより、技術的にはヒトの各遺伝子に特異的なプローブを全て固定化できることになる。しかし、全ての遺伝子の発現用のプローブを固定化するには特許の問題が生じてくる。他社により特許出願されている遺伝子の一部をプローブとする場合にはライセンス料を支払う必要があり、場合によっては非常に高価なチップになる可能性もある。一方、2章にも述べた遺伝子診断関連では、現在のところ HIV、p53、p450、BRCA1, 2 遺伝子の診断用チップが開発されており、この種類は増加していくと思われる⁴⁾。Affymetrix社と共同で開発を進めている Incyte 社は、免疫病理学、G-蛋白質関連のレセプターや転写因子解析用チップ (LifeChip) のほか、Signalling & neoplastic growth、ラットの毒物に関する遺伝子解析用チップの開発も行っている。また、Hyseq 社も独自に 10 塩基のオリゴヌクレオチドを固定化した HyChip を開発しており、100 万塩基の解析が可能で、第1世代の HyChip は間もなく市場に出るとのことである。Hyseq 社はもともと SBH を提唱し特許も出願しているためか (Affymetrix 社と競争

中である)、現在もDNAシーケンシングの可能性を探っているが、遺伝子発現用のアレイ開発も行っている。

3-1-2 cDNA array

一方、Molecular Dynamics社やStanford大学が開発しているのがcDNA arrayである。cDNA arrayは合成したcDNAをchip上にスポットするもので、Shalonらが開発したchip作製ロボットの概略を図3-1に示す。

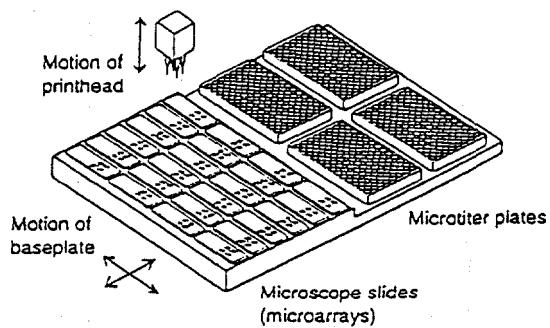


図3-1 Shalonらが開発したcDNAチップ作成ロボットの概略図

このロボットは9mm間隔の4本のピンセット型printing tipsを持つprintheadと、スライドグラスや96穴プレートを含むbaseplateおよびこれらをx,y,z軸に動かす駆動部等からなる⁵⁾。マイクロプレート中のPCR産物をprinting tipsで吸い上げ、poly-L-lysineをコートしたスライドグラスに約5nlずつスポットすることによりcDNA arrayを作製する。この装置により、1チップ上に6000種類のcDNAを固定化し、酵母の全遺伝子の発現を解析した例もある(詳細については第2章参照)⁶⁾。さらに細密化することにより、ヒト遺伝子発現解析への応用も可能になると考えられる。Synteni社も同様の固定化技術により2色蛍光色素を利用したcDNA array(Gene Expression Micro-array; GEM)の開発を行っている。Molecular Dynamics社はスポットから検出装置まで総合的にサポートすることを目指している(図3-2参照)⁷⁾。cDNA arrayの特長は、塩基配列の明らかになっていない遺伝子の探索に適していること、プローブが長いのでハイブリダイゼーションの特異性が高いことがあげられる。しかし、フォトリソグラフィーで構築したOligonucleotide arrayに比べ、1プローブの面積が広くなるため、チップ自体も大きくなるといった欠点もある。

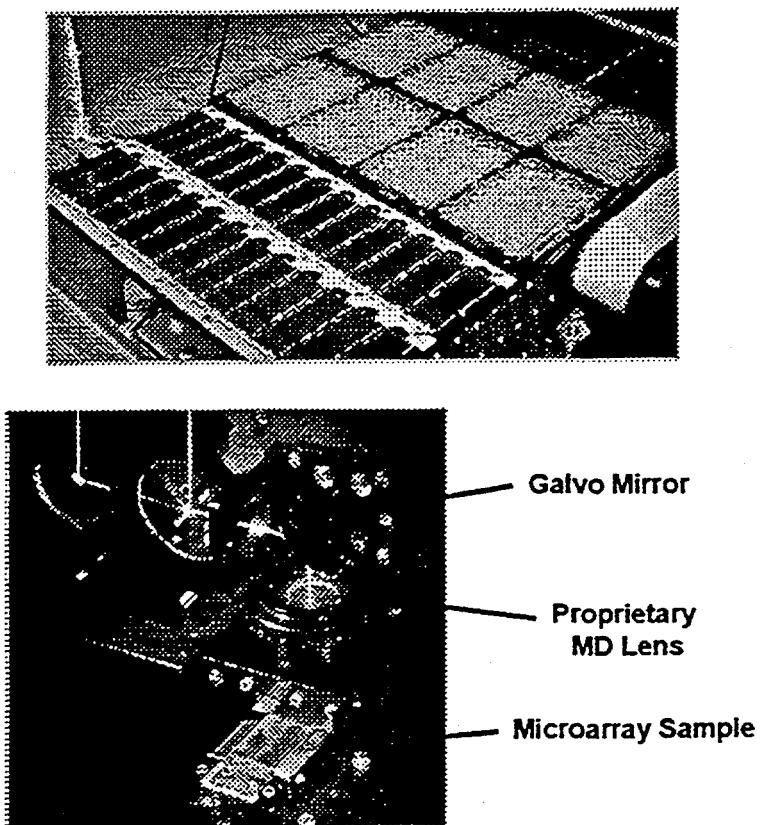


図3-2 Molecular Dynamics社が開発している第2世代アレイスポット（上）とアレイスキャナ（下）

3-1-3 その他のmicroarray

また、ユニークな microarray としては Nanogen 社が開発した電場を利用したチップがある⁸⁾。電場中でハイブリダイゼーションを行うことにより DNA の移動、濃縮、結合、解離が促進され、効率よくミスマッチ検出ができる。6mm × 6mmのチップ (APEX) の中心には100種類のオリゴヌクレオチドプローブが固定化されており、プローブの周囲には wire が張り巡らされている (図3-3参照)⁹⁾。このチップを含むカートリッジを図3-4に示した。カートリッジには、チップのほかにサンプルやバッファーの流路、気泡検出器、電気系統へのコネクタが存在しており、全体の大きさも年々小型化されている。Nanogen社は血液や組織サンプルからDNAを抽出し、ハイブリダイゼーションによる遺伝子診断や発現遺伝子解析までのトータルシステムの開発を目指しており、この装置はその中核をなすものである。現在は100種類のプローブしか固定化することはできないが、1998年には1,000、1999年以降には10,000種類のプローブを固定化できるようなシステムを構築することが目標である。

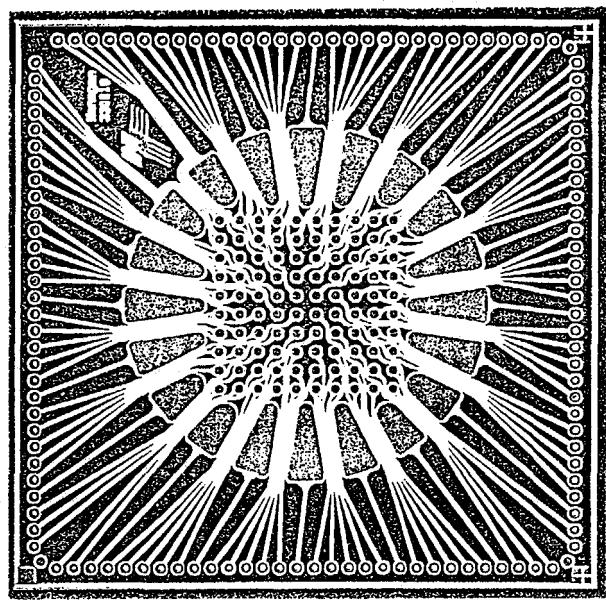


図3-3 Nanogen社が開発している電場を利用したチップ (APEX 1010S) 6mm × 6mm のチップに100種類のプローブが固定できる。

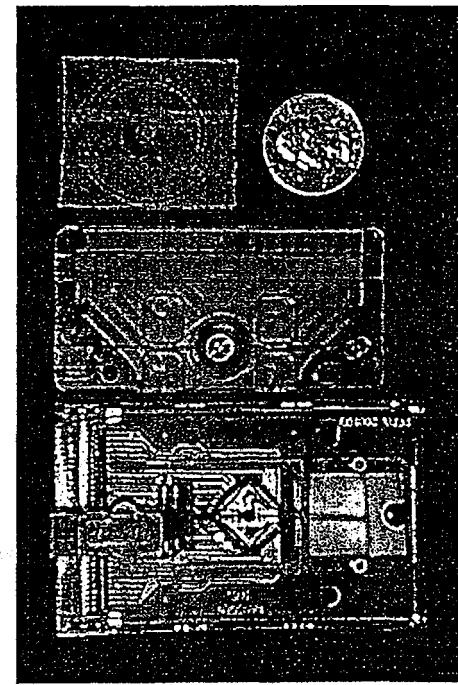


図3-4 Nanogen社が開発した年々小型化されたるカートリッジ

CLONTECH社は、安価で効率のよいAtlas Arrayを提案している¹⁰⁾。通常のメンブレンにプローブを固定化し、ラジオアイソトープ等で標識したDNAをハイブリダイズし、結果をコンピュータ処理することにより、発現プロファイル情報を解析するシステムである。図3-5にはAtlasの操作法、図3-6にはヒト Blioblastoma細胞にBCNU処理を施したものと、処理していない細胞の遺伝子発現を解析した例を示す。比較的簡便な装置で解析ができるが、一方ではRIを使うなど原始的とも言えるシステムである。

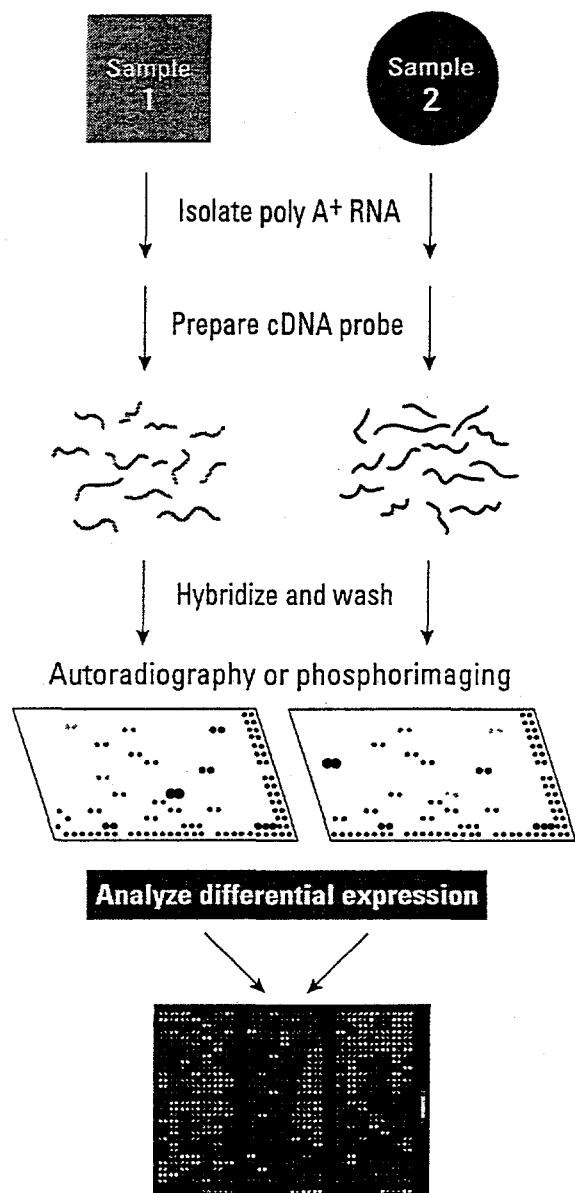


図3-5 CLONTECH社が提案しているAtlasによるハイスループット発現遺伝子解析法

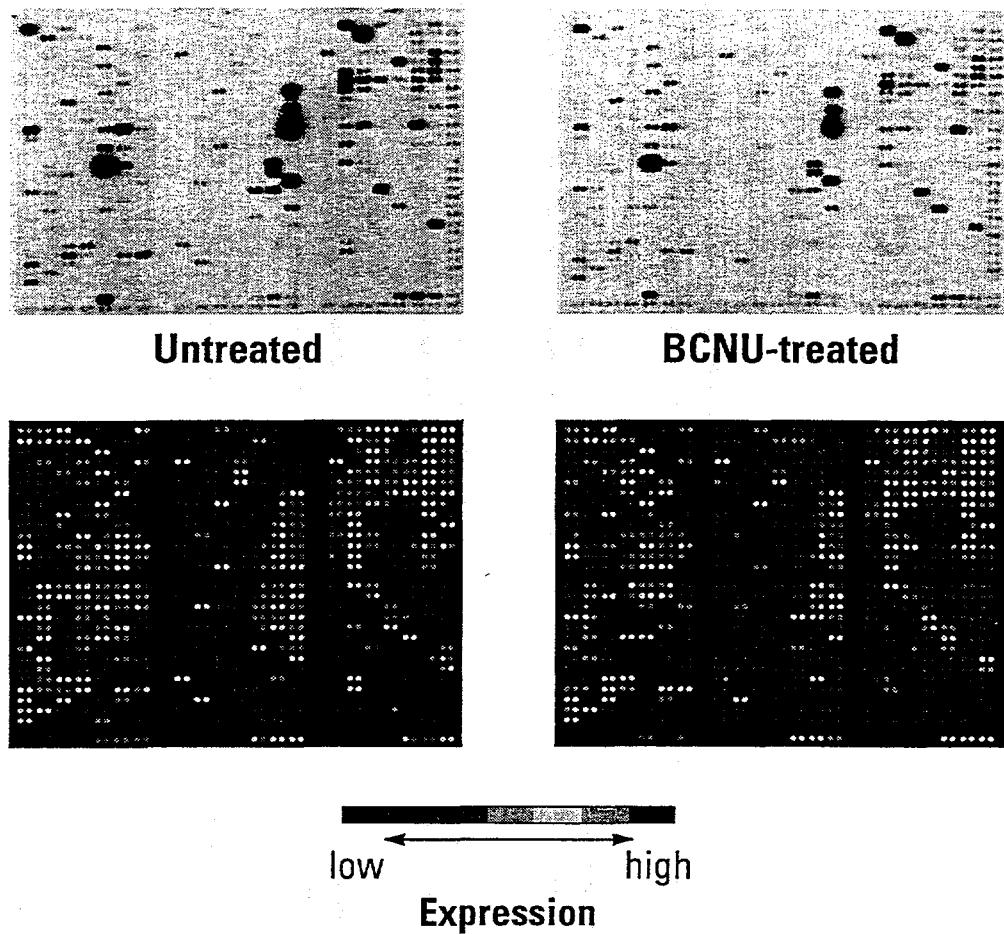


図3-6 ヒト Glioblastoma 細胞にたいする BCNU 処理の影響をモデルにした Atlas 解析例

遺伝子発現パターン解析法READSを開発したGene Logic社は、arrayの表面積を増加させるために3次元チップを考案し、開発を行っている。とはいえ、流路を立体的にしてチップ全体を小型化することはできるが、蛍光検出を行う以上array部分の面積を小さくすることはできないもようである。Protogene社はプリンターのインクジェット技術を応用してDNAをチップ上に固定化する技術の開発を行っている。

3-1-4 今後の展望

以上、簡単にmicroarray技術についての研究開発動向を述べた。今後の課題としては、ハイブリダイゼーションや検出に要する時間を短くすること(数時間をする)、蛍光物質を用いた検出系のダイナミックレンジが、発現遺伝子の量を測定するうえで十分かどうかを見極める必要がある。ちなみに、Affymetrix社では $10^3 \sim 10^4$ の範囲で蛍光強度に直線性が見られるとのことである。さらに、ハイブリダイゼーションには非特異反応が多い

ことが予想される。各研究機関とも、いかにミスマッチをもつDNAの結合を抑制するかに工夫を凝らしているが、さらに効率を高めることも必要である。また、技術とは直接的には関係ないが、Affymetrix社とMolecular Dynamics社の間で遺伝子解析技術協定(GATC)が締結され、array技術を用いた治療、診断、病気の管理を目的とする製品を規格化していくものと思われる。具体的には、arrayの形状や方法、検出、解析を統一することにより試薬や装置、解析ソフトを共用できるため、研究者の負担が軽減される。ゲノム関連情報企業のIncyte社もベンチャー企業のSynteni社の吸収合併を発表した。Incyte社はこれまでヒト、動物、植物、微生物の遺伝子配列、発現情報、マッピング情報や各種解析ソフト、試薬類、Affymetrix社の技術を導入したoligonucleotide microarray(LifeChip)の販売をしてきた。さらに、Synteni社のcDNA microarray技術とIncyte社が保有するヒトの遺伝子クローンを組み合わせることにより、ヒト全遺伝子の発現解析チップ開発に拍車がかかるものと思われる。

この様に、アメリカのmicroarray関連のベンチャー企業間で吸収、合併、業務提携が頻繁に行われるようになりつつある。これは今まで技術中心に各社が開発を行ってきたが、各社の事業をさらに拡張させるとともに、ゲノム産業を発展させるべく将来を見据えたときに、現在の独立した手法、装置では研究者間のコンセンサスが得られず科学の進歩に遅延が生じると考えたためであろう。日本におけるmicroarrayを中心とするゲノム情報研究は残念ながら非常に遅れている。今後、我々がこの分野で生き残ることができるとすれば、全く新しい技術を打ち立てるか、独創性のある技術で、なおかつ先行するアメリカの製品規格を意識したものを作る必要があるかもしれない。

最後にDNAチップからは離れるが、アメリカではProteomics研究の成果発表も行われはじめている。今後は、従来の2次元電気泳動やMS解析だけではなく、Protein Chipの技術開発が鍵になることは間違いない。

文献

- 1) Lipshutz, R.J. et al. : *BioTechniques*, 19, 442 - 447 (1995)
- 2) McGall, G. et al. : *Proc.Natl.Acad.Sci.USA*, 93, 13555 - 13560 (1996)
- 3) Kozal, M. J. et al. : *Nature Medicine*, 2, 753 - 759 (1996)
- 4) Hacia, J.G. et al. : *Nature Genetics*, 14, 441 - 447 (1996)
- 5) Shalon, D. et al. : *Genome Res.*, 6, 639 - 645 (1996)

- 6) DeRisi, J.L. et al. : *Science*, 278, 680 – 686 (1997)
- 7) Barker, D.L. : Fifth annual Human Genome Project 講演要旨集 (1998)
- 8) Ronald, G. et al. : *Proc.Natl.Acad.Sci.USA*, 94, 1119 – 1123 (1997)
- 9) Heller, M.J. : Fifth annual Human Genome Project 講演要旨集 (1998)
- 10) Siebert, P.D. : Fifth annual Human Genome Project 講演要旨集 (1998)

3 – 2 cDNA Sequencing によるもの

ある単細胞生物あるいは多細胞生物のある組織において、どのようなタンパク質が発現しているかを調べることは、その生物や組織の機能を知る上で非常に重要な情報となる。mRNA の発現プロファイルの解析は、どのようなタンパク質が発現しているのかを調べる直接的な手段であると共に、ゲノムプロジェクトにおいて解析された Open reading frame (ORF) が実際に転写されている遺伝子であるかどうかを検証するための手段となる。ゲノムプロジェクトにより全ての ORF が正しく同定されたという仮定の下では、mRNA の発現プロファイルを解析するには、前項に記述した cDNA マイクロアレイなどを用いた方法^{1, 2)} が簡便性と迅速性の面で最も適していると考えられる。しかしながら、多くの生物ではこのような検出方法が現在提供されていない。また、ゲノムプロジェクトが完了した生物においても、ORF としての条件を満たさず見過ごされた遺伝子の発現については情報を得られない。このように既知の塩基配列情報を利用して発現プロファイルを解析する手法では解析できない、あるいは解析が不完全になるケースも多々ある。そこでこのようなハイブリダイゼーションを原理とする方法とは対照的に、発現している mRNA 分子種の塩基配列を数多くシーケンスしてその同定を行い、それぞれの mRNA 分子種ごとに出現頻度をカウントして発現プロファイルを解析する方法が開発されている。現在その代表となっているのがボディマップと SAGE (Serial analysis of gene expression Serial analysis of gene expression) 法である。これらの方法は未知の遺伝子も含めて発現している mRNA 分子種を網羅できるという共通の特長を有するが、一方で一回の解析にかかる時間が膨大に長く、発現量は出現回数すなわち整数値で表現されるため有効数字の桁が小さく、発現量が極めて低い分子種は検出されにくいという欠点があげられている。このような欠点により現時点ではこれらの方法がルーチン的な発現プロファイル解析に用いられる可能性は低いが、発現プロファイルに関して情報が少ない生物の解析に対しては強力なツールとなっていくと考えられる。また、将来的に MPSS

(Massively parallel signature sequencing) 法などのようにさらに改良が進むと思われる。

3-2-1 ボディマップ

ボディマップ^{3, 4)} は、ヒトの各臓器で発現している mRNA 分子種を次々にシーケンスし、同定・分類することによりヒトの各種臓器における発現データベースを作成することを目的として、我が国の大久保らによって進められている研究である。

特定の組織などの mRNA から作成した cDNA を無作為にその両端のいずれかよりシーケンスすることによって、発現している mRNA 分子種の部分配列のデータが蓄積される。このようにして得られた配列データは EST (expressed sequence tag)⁵⁾ と呼ばれている。これまでにさまざまな生物の EST データが多数の研究室において解析され、DNA データベースの一つのカテゴリ (dbEST) に膨大なエントリーが蓄積されている。しかしながら、これらの EST 配列データは多くの研究室において独自に、また様々な手法により調製したサンプルを用いて得られたものであるために、同じ分子種に対して多くのエントリーが存在し、またその EST 配列も cDNA 中のさまざまな位置に由来している。このため、EST 配列データからは、ある臓器においてその分子種が発現していることは明らかとなるが、ある臓器における様々な分子種の発現量の比に関する情報は得られない。そこで、ボディマップの研究においては、cDNA を *Mbo* I で切断して得られる最も 3' 末の cDNA 断片について、*Mbo* I 切断部位から 3' 末に向けての塩基配列を解析するというストラテジーを採用している。このような工夫によって同じ cDNA 分子種からは同じ塩基配列データ（それぞれの分子種を識別できるという意味で gene signature (GS) と命名されている）⁴⁾ が得られるため、それら塩基配列をカタログ化し、それぞれをカウントすることによってそれらの分子種数を代表するデータ、すなわち発現プロファイルが得られる。

1996年末までにヒトの様々な臓器・細胞から単離された 47,727 クローンが解析・カタログ化され、12,984 種類の GS クラスターが登録されている⁶⁾。現在、ボディマップの World Wide Web page (<http://www.imcb.osaka-u.ac.jp/bodymap/welcome.html>) には、1995年末までに収集された 39 種の細胞由来の 10,896 種の GS が公開されているようである (図 3-7, 3-8)。同 WWW site では、In - silico Northern ハイブリダイゼーションを試すことが出来、目的とする配列を有する mRNA がヒトの各臓器・細胞

でどの程度発現しているのかをボディマップのデータを基に調べることができる。



BodyMap: Anatomical Expression Database of Human Genes

BodyMap is a data bank of expression information of human genes, novel or known, in various tissues or cell types. It is created by random sequencing of clones in 3'-directed cDNA libraries. Since these libraries were made so as to faithfully represent the composition of material RNA, redundancy of the same sequence reflects the quantitative aspect of gene expression in various tissues. This project has been conducted by Kousaku Okubo and Kenichi Matsubara since 1991.

Note: This server utilizes some features of HTML3.0 (e.g., tables, background coloring). Use appropriate browsers such as Netscape 1.1N or later.

- Brief Description of BodyMap
- Gene Expression Profiles of Cells and Tissues
- Cells and Tissues Analyzed
- Expression Patterns of Selected Genes
- Library Information
- *In-silico* Northern Hybridization of Your Sequence
- BodyMap Entry Retrieval

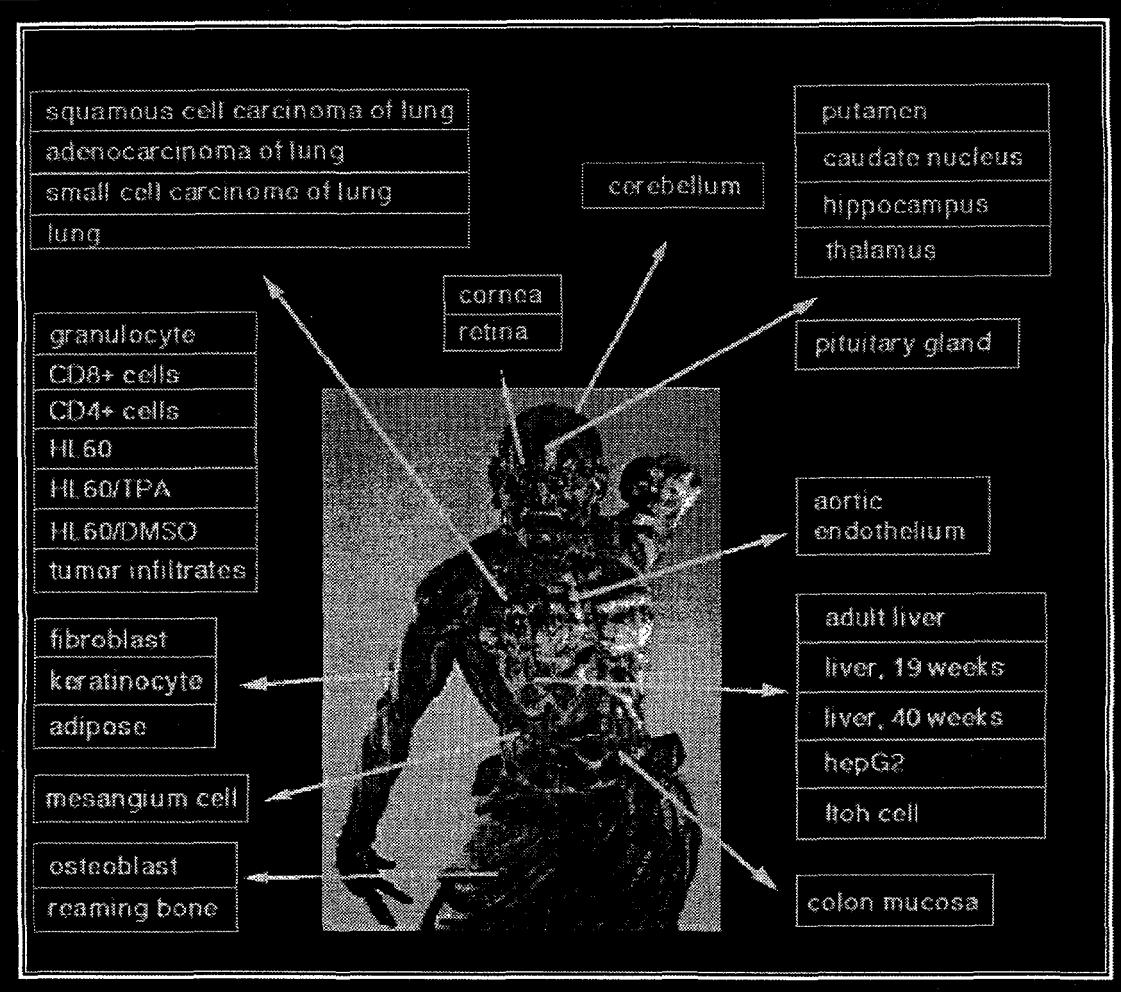
- Version Information
- Further Description
- Slide Archive
- Summary Statistics
- Bibliography
- List of Contributors
- Anonymous FTP
- Summary of Human EST

This BodyMap WWW server was constructed by K.Nakai

Last update: Aug 7, 1996
kousaku@imcb.osaka-u.ac.jp

図3-7 ボディマップのホームページ

BodyMap Gallery



This is a clickable map. Clicking any blue rectangle leads you to the graphical page of anatomical explanation for the cell types whose expression profile has been obtained. If you continue to click the picture you encounter, you will finally get to the page of its expression profile.

This image was created by K. Okubo based on a figure ...

[Go to "Gene Expression Profiles of Cells and Tissues"](#)
[Go to "BodyMap Library Information"](#)
[Go to "BodyMap Home Page"](#)

Last update: April 7, 1996
kousaku@imcb.osaka-u.ac.jp

図3-8 ボディマップにより解析されている臓器・細胞

ボディマップの手法は解析しようとする臓器・細胞における発現プロファイルを調べる上でそのmRNA分子種そのものを一つずつ調べていくものであり、したがって分類・同定という観点から見ると確実性があり、また他のEST解析手法と比べると発現量を見積もる上での定量性に優れる。このため、発現している分子種の情報が極めて少ない状況下においては、有用な手法であると考えられる。しかし一方で、発現プロファイルを得るまでに膨大な数のcDNAの単離とシーケンスを必要とするため、各研究者が自分の目的とする細胞の発現プロファイルを得るための手段としては、解析のための時間と労力という大きな問題を抱えている。したがって、ボディマップの手法は、発現プロファイルを得るために他の情報源が限られているために、本手法が極めて有効な手段である場合に利用されると考えられる。また、既に解析されたボディマップデータを利用して、自分の目的とする分子種の発現情報を得るという場合に有用であると思われる。

3-2-2 SAGE法

SAGE法は、1995年にVelculescuら⁷⁾によって報告された方法であり、基本的にESTやボディマップと同じようにサンプル中のcDNAを無作為にシーケンスして同定・分類を行い、その出現頻度によって発現プロファイルを解析する方法である。SAGE法では、cDNAを同定するためのタグを9 bpの短い配列とし、それらをタンデムに繋げてからシーケンスすることによって一回のシーケンスで数十クローンが一度に解析でき、解析速度を大幅に向上させている。実験手法(図3-9)としては、全mRNAをcDNAに変換後ビーズに固定し、次いで4 bp認識の制限酵素(*Nla* IIIなど)で切断し、最も3'側の断片を単離する。これを2つのフラクションに分け、後にPCRによる増幅を行うためのアダプター(別々のもの)を連結したのちに、非対称DNA切断をするクラスIIS制限酵素(*Bsm*F Iなど)によってmRNA由来の9 bpタグを得、2つのフラクション間でダイマーを形成させた後にPCRで増幅し、さらにコンカテマー(共重合体)の連結によって得る。これをシーケンスし、周期的に出現する9 bpタグの配列を読みとり、その統計を取ることによってmRNA分子種のプロファイルを得る。このようにして得られたmRNA分子種プロファイルについては、ハイブリダイゼーションによって得られる結果と相関することが一部のcDNAについては調べられ⁷⁾、基本的に比較的発現量の高いmRNA分子種については定量性があると考えられる。このため長所としては、(1) 発現プロファイルの取得とmRNA分子種の同定が同時に出来る、(2) 実験手法として容易、(3) 比較的短時間に解

析できる、などがある。一方で、欠点としては、(1) 発現量の低いmRNAは連結やシーケンスのステップにおいて取りこぼす可能性がある、(2) 最初の制限酵素切断部位を含まないmRNAは検出できない、(3) 同一の9 bp タグを与えるmRNA分子種同士は区別できない、(4) 統計的に信頼性のあるプロファイルデータを得るには数万のタグを調べる必要がある、などが指摘されている。

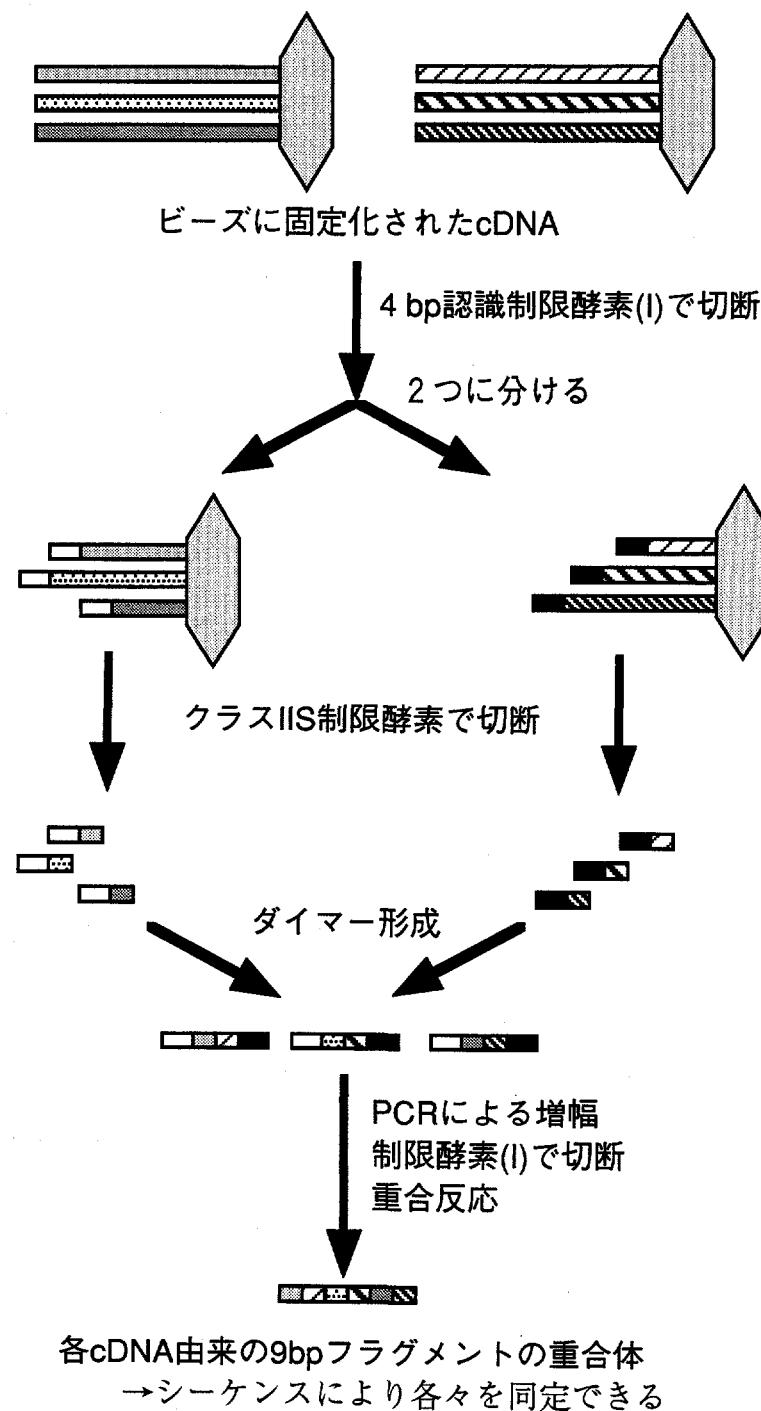


図3-9 SAGE法の概略

これまでにヒト肺臓⁷⁾、ラット胚線維芽細胞 (p53依存発現プロファイルの変化)⁸⁾、酵母 (細胞周期による変化)⁹⁾、ヒト線維芽細胞 (放射線誘導)¹⁰⁾に、ヒト大腸および肺臓癌細胞¹¹⁾について、SAGE法により mRNA 分子種のプロファイルが解析されている。特に、ゲノムプロジェクトの終了した酵母の論文⁹⁾では、タンパク質の二次元電気泳動によって得られる発現タンパク質の一組を表す「proteome」^{12)、13)}という言葉に習って、この SAGE 法によって特定した発現 mRNA の一組を「transcriptome」と命名している。また、SAGE 法と基本原理を同じくし、制限酵素の組み合わせなどの改良により特異性などを向上させた mini - fragment analysis of gene expression (MAGE) 法¹⁴⁾が東京医科歯科大学のグループにより開発されている。

解析時間の問題はあるものの、SAGE 法は、複数の生理的条件下における mRNA 分子種のプロファイルの変化を網羅的に調べるために従来の subtraction 法、differential display 法、2D PAGE 法などに比較して優れている。特にゲノムプロジェクトによりゲノムの全塩基配列が決定されている生物においては、9 bp タグによる遺伝子の特定が容易であるために強力なツールとなると考えられる。

3-2-3 MPSS 法

アメリカ国の LYNX THERAPEUTICS 社により mRNA プロファイルを解析するための新しい方法、Massively parallel signature sequencing (MPSS) が開発された。同社は高速塩基配列解析法を開発した^{15)、16)}。これは、オリゴヌクレオチドタグを目的とする DNA に結合させた後、タグ付きの DNA をラベル化し (このときに目的 DNA の末端の 1 または数塩基の配列を認識するようにする)、次いでラベル化されたタグをアレイに供して検出する。この方法により DNA 末端から 1 あるいは数塩基の塩基配列を同時に数サンプル分検出できる。詳細は不明であるが、MPSS 法では、クラス IIS 制限酵素認識配列を有するタグを利用し、上記高速塩基配列解析法で目的 DNA の末端塩基を解析した後、クラス IIS 酵素で切断する (このとき、タグの隣すなわち解析された末端塩基が切り離されるようにタグをデザインしてある)。このような末端塩基の決定と切り離しを繰り返すことにより、目的の DNA を末端から順次高速に決定できる。この方法は自動化と複数の DNA の同時解析が可能な方法であり、SAGE 法と同様な塩基配列を分類タグとする発現プロファイルの解析が高速化できる。

文献

- 1) DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A. and Trent, J.M. : *Nature Genet.*, 14, 457 – 460 (1996).
- 2) Heller, R.A., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J., Woolley, D.E., Davis, R.W. : *Proc. Natl. Acad. Sci. USA*, 94, 2150 – 2155 (1997).
- 3) Okubo, K., Hori, N., Matoba, R., Niiyama, T. and Matsubara, K. : *DNA seq.*, 2, 137 – 144 (1991).
- 4) Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y. and Matsubara, K. ; *Nature Genet.*, 2, 173 – 179 (1992).
- 5) Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., et al. : *Science*, 252, 1651 – 1656 (1991).
- 6) 大久保公策 : 蛋白質核酸酵素, 42, 2822 – 2829 (1997).
- 7) Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. : *Science*, 270, 484 – 487 (1995).
- 8) Madden, S.L., Galella, E.A., Zhu, J., Bertelsen, A.H. and Beaudry, G.A. : *Oncogene*, 15, 1079 – 1085.
- 9) Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett Jr., D.E., Hieter, P., Vogelstein, B. and Kinzler, K.W. : *Cell*, 88, 243 – 251 (1997).
- 10) 原 健、難波裕幸、山下俊一 : 長崎医学会雑誌, 71, 344 – 347 (1997).
- 11) Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B. and Kinzler K.W. : *Science*, 276, 1268 – 1272 (1997).
- 12) Wilkins, M.R., Sanchez, J.C., Gooley, A.A., Appel, R.D., Humphery – Smith, I., Hochstrasser, D.F. and Williams, K.L. : *Biotechnol. Genet. Eng. Rev.*, 13, 19 – 50 (1996).
- 13) Wilkins, M.R., Sanchez, J.C., Williams, K.L. and Hochstrasser, D.F. : *Electrophoresis*, 17, 830 – 838 (1996).
- 14) 梁 明秀、羽田亮之、山本直樹、山本三毅夫 : 生化学, 69, 701 (1997).

- 15) Patent No. WO9612014 – A1 (96.04.25)
- 16) Patent No. WO9612039 – A1 (96.04.25)
- 17) Patent No. WO9732999 – A1 (97.09.12)

3-3 メッセージディスプレイによるもの

3-3-1 Differential Display法

Differential Display (DD) 法はmRNA 発現様式を調べるアプローチとしては、やや古典的な方法となりつつあるが、別の言い方をすれば応用面が充実してきた方法であるとも言える。1992年8月、Liang と Pardee¹⁾ によりサイエンス誌上に発表されて以来、このDD法は複数のmRNA 種を同時に比較することにより、点で見ていた細胞の状態を線でみることができる非常に有用な方法として多くの研究グループにより応用され、多くの成果をあげてきた。medline により最近の論文発表状況を検索したところ、過去5年間で約700件もの発表が認められ、1992年以来、年々その件数が著しく増加してきた。それらの中には例えば、薬物処理前後あるいは時間を追って mRNA メッセージの量的変化²⁾ 、あるいは疾病患者の患部、非患部組織でのメッセージの差異³⁾ 、あるいは細胞の分化、細胞周期や発生段階によるメッセージの差異⁴⁾ 、また、機能未知遺伝子（特に、新規の転写因子様のもの）の機能解析は、遺伝子産物のネットワークの解析により可能となると思われるが、その一環として機能未知遺伝子の導入もしくは欠損させたときに発現量の変化する遺伝子群の同定および解析⁵⁾ においても DD 法は大きな力を発揮してきた。また、最近では非放射能による検出方法（例えば、銀染色法や蛍光標識法等）が開発されつつあり、特に、前回の本調査報告書にも記載されている伊藤らの方法は 3' – anchored primer の 5' 端に fluorescein isothiocyanate の蛍光物質を導入した方法で感度、迅速性において非常に優れた方法である^{6, 7)} 。

私たちのグループにおいても疾患原因あるいは関連遺伝子を効率的に単離し、遺伝子の機能解析により治療への指針にする目的で、従来法より再現性（同一の Display DNA 断片が検出できるか）と確実性（Display DNA 断片が示す特異性がノーザンの結果と一致するか）と感度（発現頻度の低い cDNA を発見できるか）の点で改良を加えて、これまでに多数の組織特異的に発現する遺伝子の単離を行ってきた^{8, 9)} 。ヒトの遺伝子は約 10 万種類あると言われているが生命活動を維持していく上で多くの遺伝子が発生段階特異的あるいは組織特異的に発現制御をうけて正常個体へと発生が進められていく。この際、組織

特異的遺伝子はその組織が組織として機能するために必須となる遺伝子が多く、そのような遺伝子の変異により疾患が引き起こされると考えられる。また、組織特異的遺伝子の機能解析はノックアウトマウスを用いた個体レベルの解析において、致死性を起こしにくく、表現型の現れるものが効率よく出現することが予想され、しいては新規遺伝子の機能の解明が効率よく行われるものと考えられる。また、一方では、組織特異的遺伝子のプロモーター等の調節領域の解析から遺伝子治療用のベクターの開発にとってもこのアプローチは重要な役割を担うものと考えている。

具体的には14種類のヒト組織mRNAより、逆転写酵素と3' – anchored primer (GT₁₅ MAあるいはGT₁₅ MCあるいはGT₁₅ MGあるいはGT₁₅ MT； M : A,C,Gのミックス) を用いてcDNAを合成した後、³³ Pで標識した3' – anchored primerおよび10塩基からなる5' – arbitrary primerにより30~40サイクルのPCRを行った。増幅されたmRNAの3' 末端領域(100~500塩基)のPCR産物をシーケンスゲルで電気泳動し、ゲルを乾燥後、オートラジオグラフィーを行い、組織特異的バンドを乾燥ゲルより切り出した。バンドより抽出したDNAを、再度、同一PCR条件下でサイクル数を少なくして2nd PCRを行い、電気泳動後、複数のバンドが出現する場合は正確に同じ位置のバンドを切り出し、pUCベクター等にサブクローニングし配列決定を行った。2nd PCRを行うことにより目的のバンドを効率よく単離することが可能となった。このシステムでは1回の電気泳動で約10 クローンの組織特異的遺伝子断片を得ることができ、それらのDNA断片をプローブとしてノーザンプロットで確認したところ、90 %以上のDNA断片で組織特異性が確認できた。図3-10-Aには9種類の組織を用いて行ったDisplayの一例と図3-9-Bには断片a及びbが実際にどのような組織分布をするのかをこれらのDNA断片をプローブとして16種類の組織を用いてノーザンプロット解析を行った。感度についても、各組織のcDNAライブラリー100万pfuをまき、ラークハイブリダイゼーションにより、ポジティブクローニングをカウントすることで0.0001%以下のもの(平均0.0001~0.001%)をとらえることができた。

今後のアプローチとしては、Display DNA断片より効率よく完全長cDNAの単離と構造解析を行うとともに、ノックアウトマウスを用いた個体レベルの解析と組換え蛋白質を用いた蛋白質工学的手法によりこれらの遺伝子の機能解析をすすめ、疾患遺伝子の同定とそのネットワークの解析から、遺伝子診断、治療薬開発に進むことが期待できる。

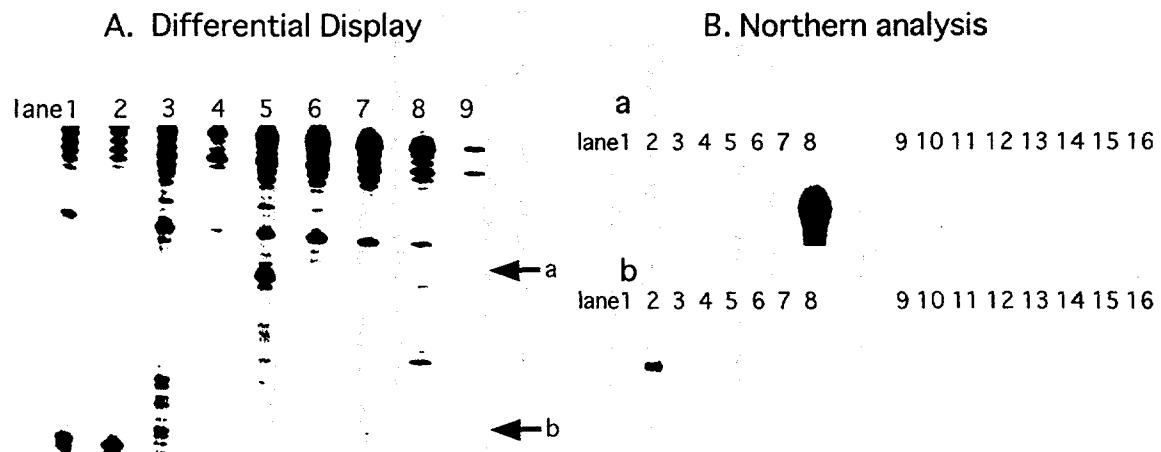


図3-10

文献

- 1) Liang, P. and Pardee, A.B. : *Science*, 257, 967 – 971 (1992)
- 2) White, J.A. et al. : *J. Biol. Chem.*, 271, 29922 – 29927 (1996)
- 3) Lim, J. et al. : *Cancer Res.*, 57, 921 – 925 (1997)
- 4) Zimmermann, J.W. and Schultz, R.M. : *Proc.Natl.Acad.Sci.U.S.A*, 91, 5456 – 5460 (1994)
- 5) 石川晋之ほか : 実験医学 15, 76 – 79 (1997)
- 6) Ito, T. et al. : *FEBS Lett.*, 351, 231 – 236 (1994)
- 7) Kito, K. et al. : *Gene*, 184, 73 – 81 (1997)
- 8) Ozaki,K. et al. : *Genomics*, 36, 316 – 319 (1996)
- 9) Ozaki,K. et al. : *Genes, Chrom. and Cancer* : in press (1997)

3-3-2 Molecular Index法

1995年にKatoによって開発されたMolecular Index法は極微量のものを除き、発現量に関係なく、重複も少なく表示されるため、解析の感度がよいといえる^{1,2)}。この方法は3種類のClassIIIS制限酵素、64種類のアダプター、3種類のアンカープライマーを用いmRNAの全集団を576のグループに分割しPCR産物を電気泳動する方法で、Differential Display や RNA Arbitrary – Primed PCR に比べ、(1) PCRのアニーリング温度の変化による影響を受けにくい、(2) 増幅される遺伝子群の重複がほとんどない、などの点において優れているばかりでなく、各グループの電気泳動の1つのバンドが1つの遺伝子

に対応するので、電気泳動のパターンにおける各バンドの有無がそのまま各遺伝子の発現を表すという特徴を持つ。そのため各グループの各バンドと遺伝子とを対応づけることより、どの細胞においてどの遺伝子が発現しているのかを電気泳動のパターンで表すことができる。以上よりこの方法は特定の組織あるいは細胞において発現している全遺伝子の同定に最も適していると考えられる。その一方で、(1) 3種類のClassIIIS制限酵素による切断部位のないcDNAの認識（しかし加藤菊也先生の話によると、この方法で99%の遺伝子をカバーできるそうである）、(2) 操作の簡便化、(3) グループ間での重複をさらに少なくする、(4) 発現量の少ない遺伝子を検出する、等という点において検討の余地がある。また、各グループの各バンドと遺伝子との対応づけをするため、cDNAの3末端のPCR断片である各グループの各バンドを単離し、塩基配列を決定しその配列より対応する遺伝子のデータベースよりの検索、あるいは遺伝子の単離と塩基配列の決定を行う必要がある。しかしながらこの一連の実験はかなりの時間と労力を必要とすることより、これを簡便に行える技術の開発や自動化の検討も必要とされる。

Molecular Index法を用いた最近の研究としては、1997年の分子生物学会年会でヘルックス研究所グループが3T3-L1細胞ならびにNT-2細胞の分化で誘導される遺伝子群の解析した結果を発表した。彼らはMolecular Index法による解析結果をデータベース化し、公開することを目的としている。また、従来576グループに分類していたものを192グループに減らしても解析可能とのことで、特別な設備なしでできる技術としては、最も効率の良い方法のうちの1つである。

文献

- 1) Kato, K. : *Nucleic Acids Res.*, 23, 3685 - 3690 (1995)
- 2) Kato, K. : *Nucleic Acids Res.*, 24, 394 - 395 (1996)

第4章 発現調節ネットワーク解析技術

生物は、様々な外因性あるいは内因性の刺激に対し、細胞や組織が特異的な反応を示す。細胞外の情報は、情報伝達系を利用したり直接核内にはいることにより、転写因子などに作用して遺伝子発現を変化させ、細胞を分化、増殖、あるいは死滅といった方向に向かわせる。このように、転写制御は、生物が有する多数の遺伝子を、様々な状況で巧みに使い分けることによって、極めて複雑で柔軟に対応するための、最も基本となる制御系であり、多数の研究者が精力的に研究を進めてきた。

転写制御の研究は、組換えDNA技術、DNA導入技術、レポーターベクターの開発、無細胞転写系の確立、タンパク質の分離・精製などの、基盤的技術の上に成り立っている。近年では、実験解析技術の進歩により、転写制御領域および転写制御因子など、個々の遺伝子の転写制御に関わる解析法も充実してきた。また、近年では、ヌクレオソーム構造やクロマチン構造などのDNAの高次構造と転写制御との関連も研究することも可能になってきており、細胞の発生・分化に関わる遺伝子の不活性化、染色体上の位置によるグローバルな遺伝子の転写制御など、より生命現象を正確に反映した転写制御の解析が可能となってきた。

しかし、全ゲノム解析に見られるように、生物の持つ全ての遺伝子を、包括的な観点から解析するためには、依然として膨大な労力と時間を必要とする。これによって、生体内の遺伝子発現の調節機構が解明できるだけでなく、組織特異性が高く、安全で、正確に発現調節の可能なプロモーターを設計できる技術の開発と遺伝子治療への応用など、産業面での利用価値も計り知れない。

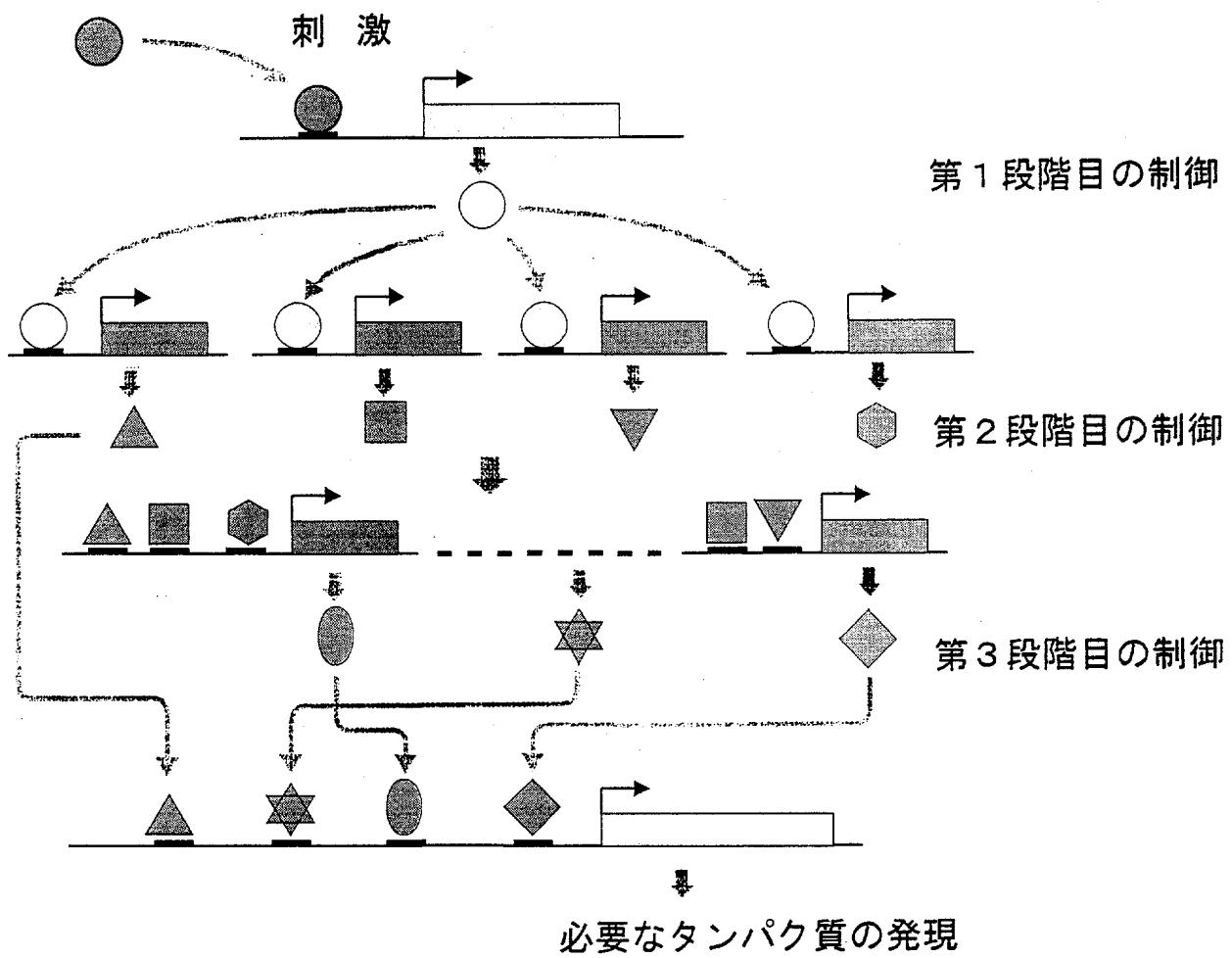


図4-1 転写制御ネットワークの模式図

4-1 転写制御シグナルの探索・解析技術

転写制御は、主に転写を誘導的に制御するエンハンサーによって為されている。エンハンサーは、基本的に位置・方向性・距離に無関係に働くが、エンハンサーによっては基本的転写領域からの距離を大きく変えると効果がほとんど見られなくなるものもある。酵母などの下等真核生物のエンハンサー（UAS）では、転写開始点の上游数100塩基程度の位置に存在することが普通であり、高等なヒトの場合でも、上游1,000塩基対以内に、80%以上の遺伝子の転写制御領域が存在すると考えられている。さらに、エンハンサーの近傍には、転写を抑制的に制御するサイレンサーが存在することも多く、これらの誘導的、抑制的な転写制御領域によって、複雑な転写制御が為されていると考えられる。

4-1-1 シークエンス比較解析技術

全ての遺伝子は、それぞれの遺伝子に存在する転写制御領域の塩基配列に、特異的な転写制御因子が結合することによって、転写発現の調節がなされている。従って、転写制御領域を網羅的に解析し、転写制御領域に必要に必要な塩基配列と、その転写制御領域の制御機能に関する情報が得ることは、生物の最も基本的な制御システムを理解するうえで重要である。

(1) プロモーター領域の探索法

全ゲノム情報が明らかになっても、遺伝子の制御領域の位置とその調節機能を予測することは容易ではない。一般的に、転写制御領域は遺伝子の転写開始点の上流に位置することが多く、そうでない場合でも遺伝子の内部か近傍に位置することから、構造遺伝子領域（タンパク質コード領域）を予測するいわゆるジーンファインディングと組合わせることによって、おおまかな転写制御領域の位置を予測することが可能である。しかし、転写制御領域へ塩基配列特異的に結合する転写制御因子の認識塩基配列は、制限酵素などに比較してはるかに曖昧であることから、転写制御領域の位置と制御機能を正確に予測することは困難である。プロモーター領域や転写制御領域の予測に比較して、構造遺伝子領域の予測はより容易であることから、転写制御領域の予測の第一歩として、ジーンファインディングは重要である。近年では、高等真核生物の場合でもジーンファイディングの予測精度がかなり良くなっているが、全ゲノム塩基配列が解析されている場合でも、ジーンファイディングの結果や転写開始部位の確定のために、完全長cDNA合成に見られるような、5'末端のクローニング技術は重要である。

5'末端のクローニング法としては、オリゴキャッピング法¹⁾が頻用されている。この方法は、3段階の酵素反応を用い、5'末端に特異的に存在するcap構造を、合成オリゴに選択的に置換することによって、mRNAの5'末端を特異的にクローニングする方法である（図4-2）。この様にして得られた情報から、直ちにその領域が構造遺伝子領域であることが導かれ、さらに、転写開始部位5'末端に特異的なcDNAタグを用いることにより、その上流にある制御領域をPCR用いたDNAウォーキングなどによってクローニングし、塩基配列を解析することが可能となる。

フランスのGensetでは、cDNAの5'末端のクローニング法に関して、化学的な方法を用いる特許を有しており、ヒトcDNAの5'末端タグの塩基配列を大量に解析してい

る。さらに、この塩基配列に基づいて5' 上流領域をクローニングして塩基配列を解析し、最終的に情報科学的に転写制御領域とその制御機能を予測する試みを行っている。

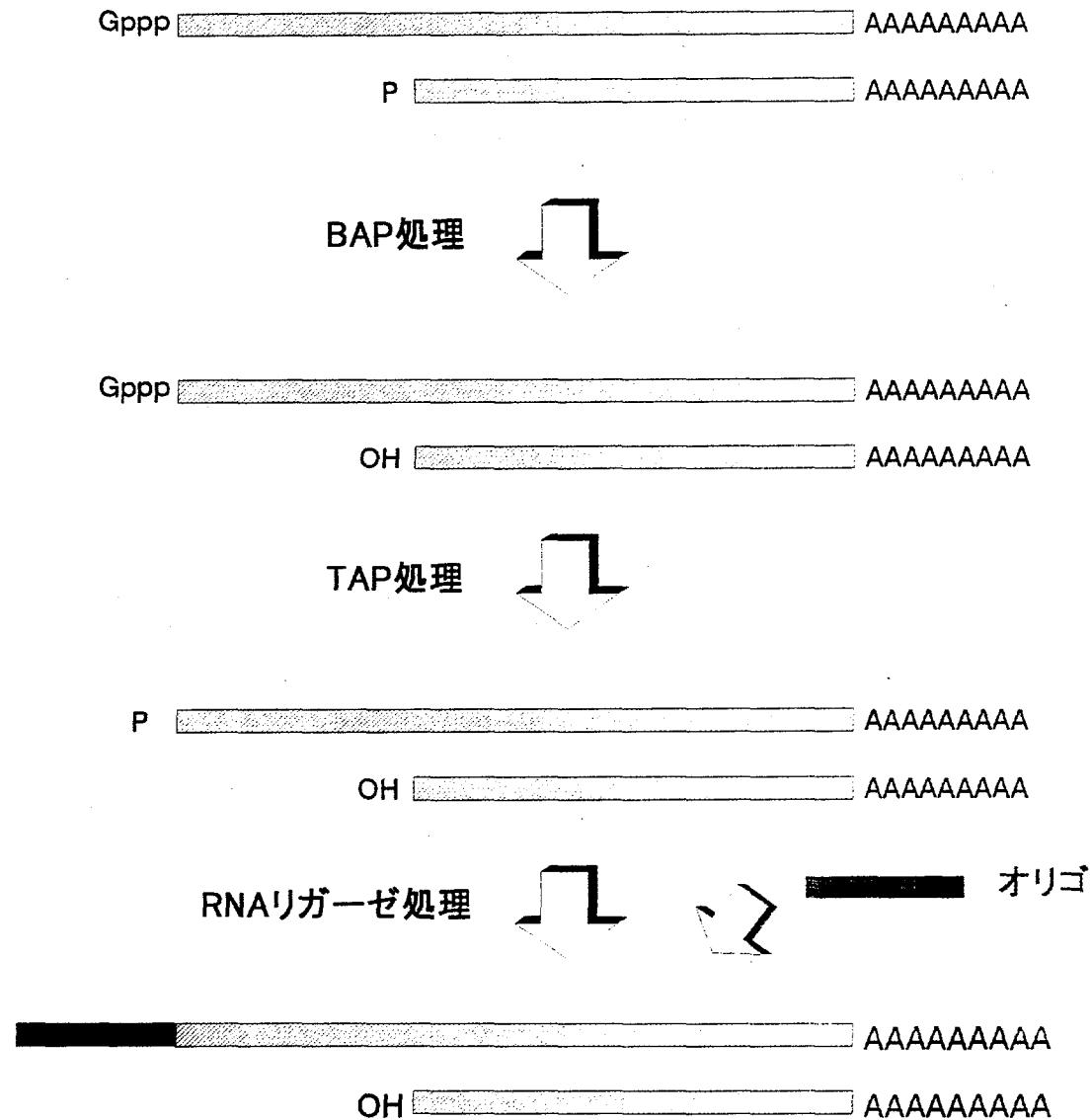


図4-2 オリゴキャピング法による完全長cDNAの合成

キャップ構造を持たない切断されたmRNAは、5' リン酸(P)が存在し、アルカリリフォスファターゼ(BAP)処理によって、除去されて水酸基(OH)となる。一方、キャップ構造(Gppp)はBAPでは除去されず、次のステップでタバコ酸性ピロフォスファターゼ(TAP)処理で除去され、5' 末端にリン酸を生成する。次に、適当な配列を有するオリゴDNAを、T4 RNAリガーゼを用いて5' 末端に結合させるが、元々キャップ構造を持たない切断されたmRNAは5' 末端が水酸基のため、オリゴの結合は起こらない。そこで、5' 末端のオリゴがついたものだけが、cDNAに合成されることによって、完全長のcDNAだけを合成することが可能となる。

(1) 乙未一夕一遺作于法

4-1-2 軟件與制圖領域的美學的解釋技術

(2) 墓基配列情報&ひづり推定法

直接測定することができることから、非常に簡便である。また、GFP タンパク質への人工的な変異の導入によって、蛍光波長、励起波長とともに、いくつかのバリエーションも作り出されており、多重蛍光検出による複数の遺伝子発現の同時検出も可能となってきた。

表4-1 主なレポーター遺伝子と特徴

レポーター遺伝子	長 所	短 所
クロラムフェニコール アセチルトランスフェ ラーゼ (CAT)	最も一般的 高感度 真核細胞は内在活性が無い	放射線設備が必要
β -ガラクトシダーゼ (β -gal)	簡便 (生細胞での検出可能)	低感度
β -グルクロニダーゼ (GUS)	簡便 (生細胞での検出可能)	動物細胞では、細胞自体が 有する活性のため不適
ルシフェラーゼ	簡便 高感度 (CATの30~1000倍)	光検出の設備が必要
緑色蛍光タンパク質 (GFP)	簡便 (生細胞での検出可能) 基質が不要	低感度 若干の検出設備が必要

実際の転写制御領域の解析では、転写制御領域が含まれると考えられる領域の上流あるいは下流からDNAを徐々に削っていき、転写量に低下が見られた所が転写制御に重要な領域を含むと考えられる。さらに詳細に転写制御領域を解析するために、DNA断片の内部に欠失を作成したり、特定の塩基に変異を導入して解析することもある。

(2) 試験管内解析法

1980年の初頭にHeLa細胞抽出液を用いて試験管内転写実験が可能となり、以降それらの実験系を用いて、数多くの基本転写因子が同定された。さらに、1987年にはCornbergらによって、出芽酵母を用いた試験管内転写実験系が可能となった。これら

の試験管内転写実験系は、主として基本転写因子の探索・解析に極めて重要な役割を果たした。

一方、転写制御領域の解析に関する試験管内実験系は、転写制御領域であるエンハンサーやサイレンサーには配列特異的に転写制御因子が結合するという性質を利用している。これらの実験系には、フットプリント法⁶⁾、ゲルシフト法⁷⁾、メチル化干渉法⁸⁾、UVクロスリンク法⁹⁾などがある。フットプリント法では、DNAの切断反応にDNase IやExoIIIなどの酵素を用いる方法や、ハイドロキシル・ラジカルやDMSなどの化学物質を用いる方法があり、特にハイドロキシル・ラジカルを用いた場合には、転写制御因子の単なる結合部位を明らかにするだけでなく、その結合に重要な部位を塩基レベルで解析することが可能である。

4-1-3 特定転写制御因子の制御領域分取・解析技術

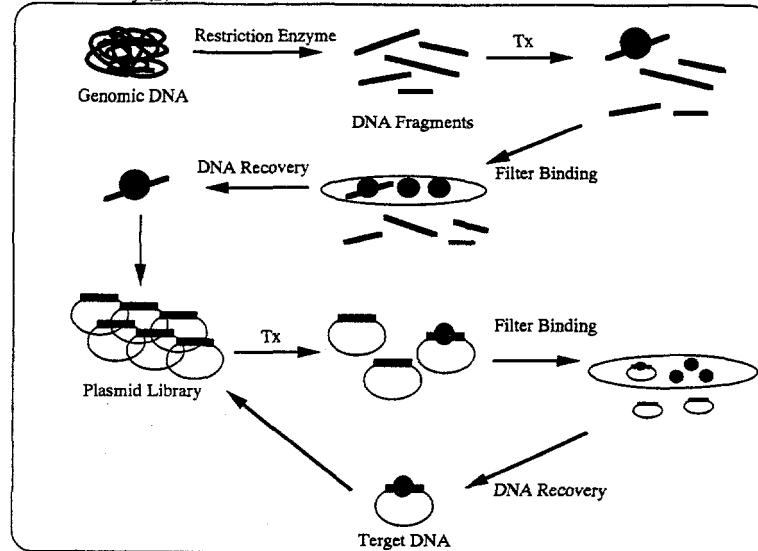
DNA結合能を有する転写制御因子は、特定遺伝子（複数の場合も多々ある）をターゲットとしてその転写制御領域に結合し、直接・間接にターゲット遺伝子の発現量の調節を行っていると考えられている。そのため転写制御因子を中心とした発現制御ネットワークを知る上で、特定の転写制御因子のDNA結合領域（配列）を分取・解析することは非常に重要なことであると考えられる。これまでにゲノム配列上での転写制御因子の結合配列、多様なDNA配列中からの結合アフィニティーの高い配列の取得を目的とした技術が以下に記述するように幾つか開発されてきており、今後の技術改良も期待される。しかしながら近年の解析より個々のDNA結合性転写制御因子の結合配列には、曖昧さが存在する上、転写制御因子の複合体化（他のタンパク質との相互作用）やリン酸化等の修飾によりDNA結合性（結合能、結合配列）が変化することが明らかとなってきた。今後、実際の細胞中の結合配列の同定技術の開発やDNA結合複合体の同定・取得・解析といった技術開発が望まれる。

(1) Genomic Binding Site クローニング技術

この技術は、特定の転写制御因子の標的遺伝子（結合DNA領域を有する遺伝子）を効率的に選択・取得する技術である。また日本において開発された技術として特筆すべきものであると考えられる¹⁰⁾。実際の技術の内容は、i) ゲノムDNAを適当な長さに切断し、ii) プラスミドライブライ化またはPCR用のアダプター（キャッチリンク）を

結合する、iii) *in vitro* で転写制御因子タンパク質と混合し、iv) タンパク質-DNA複合体をニトロセルロースフィルターに吸着させ選択する、v) 得られた複合体のDNAを回収し増幅する、vi) iii) - v) の工程を数回繰り返す。これにより得られた遺伝子の例として、転写制御因子としてエストロゲンレセプターを用いて *efp* 遺伝子が取得されている。この技術を実際に応用するために、ゲノムサイズの大きいヒトゲノムでは、まず一度DNA断片の状態で転写制御因子と混合して選択し、残ったDNAに対してプラスミドライブラー化やアダプター結合を行うという工夫がなされている¹¹⁾。また、より生体内での反応に近づけるために、クロマチン構造を持つゲノムDNAに対して検討を行うという応用も考えられる。しかしながら、具体的な技術の改良・応用の例は未だ報告されていない。

1) Plasmid Library 法



2) PCR 法

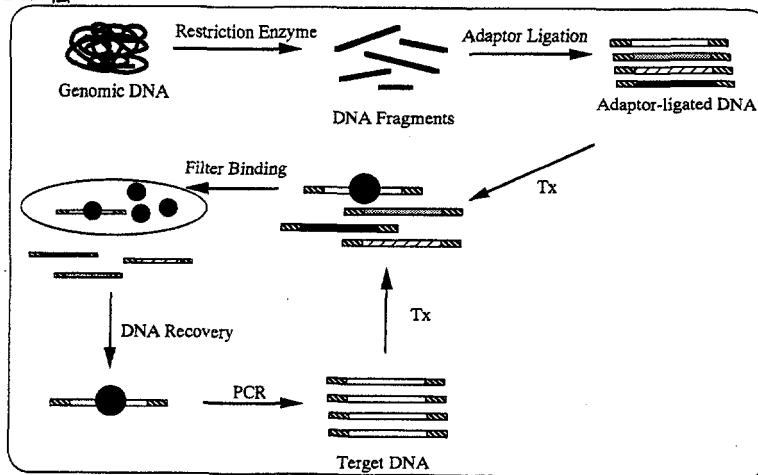


図4-3 Genomic Biotyping Site クローニング技術

(2) Casting 法

この技術は、特定の転写制御因子の結合配列を多様な配列の中から選択・取得する技術で、ゲルシフト法^{12, 13)} および免疫沈降法¹⁴⁾ をベースとした2種類の主な方法が開発されている¹⁵⁾。この技術の原理は、i) PCR用アダプターを両末端に結合したランダムなDNA配列(15–30mer程度)を合成し、ii) *in vitro* で転写制御因子と複合体を形成させ、iii) この複合体をゲルシフト(ポリアクリルアミド電気泳動)または、転写制御因子の抗体による免疫沈降にて分離・回収し、iv) 回収された複合体よりDNAを回収した後、v) アダプター配列を利用したPCRによりDNAを増幅し、vi) ii) – v) の工程を数度繰り返す、という操作により目的とする転写制御因子に高い親和性で結合するDNA配列を取得・解析することが可能となる。しかしこの方法で回収されるDNA配列のコンセンサスは、6–8merと短いことや曖昧さを含んでいることも多く、その配列から直接ゲノム中のターゲット遺伝子を同定するまでには至れないこと、実際の生体条件での結合配列との差異は不明である点等に課題を含んでいる。しかしPCRを活用することで以前に比べ格段に転写制御因子の結合可能なDNA配列の絞り込みが容易になった点で非常に高く評価できる技術である^{16, 17)}。

4–2 転写制御因子の探索・解析技術

転写制御因子は、全ての遺伝子の発現を調節しており、ゲノムDNAから遺伝子情報が発現するために必須のものである。このため転写調節因子を取得し、そのネットワークを明らかとすることはゲノムDNAの遺伝子情報のみならず生命現象の発現機序を知る上で非常に重要なことである。また転写制御因子は、数千種類存在すると考えられており、タンパク質のカテゴリーの中では非常に大きな部分を占めるものの一つである。このため転写制御因子は、未だに未知のものが多く新しい効率的な取得技術の確立が非常に望まれている。特に近年、転写制御因子は従来のDNA結合性を示すものばかりでなく、DNA結合性転写制御因子とRNAポリメラーゼを主とする転写装置本体の間に介在し、タンパク質–タンパク質間相互作用により転写のレベル、時期、特異性等を微妙に制御する非DNA結合性転写制御因子群の存在がクローズアップされてきた。そこで本節ではDNA–タンパク質間相互作用(DNA結合性転写制御因子)、タンパク質–タンパク質間相互作用(非DNA結合性転写制御因子)を利用して転写制御因子取得技術について述べる。

4-2-1 DNA-タンパク質間相互作用を用いた解析技術

DNA-タンパク質間相互作用を利用したDNA結合性転写制御因子の取得解析技術としては以下のものが開発・利用されている。

(1) South Western 法

この技術は、従来の生化学的な方法の代表的なものであるDNAアフィニティクロマトグラフィーに存在する技術的な制約、大量の核抽出液が必要、目的とする転写制御因子にたいする *in vitro* 転写系¹⁸⁾ の確立が必須、等がネックとなり適用できない系に用いられる技術として開発されてきたものである^{19, 20, 21)}。この技術は、転写制御因子のDNA結合能を指標としたスクリーニング技術であり、転写制御活性そのものは選択指標としていない。従来より用いられている遺伝子取得技術である入ファージを利用した発現クローニングの応用技術である。技術の原理は以下の通りである。i) 大腸菌の発現ベクター (λ gt11, λ ZAP 等) にcDNAを組み込んだcDNAライブラリーを調製し、ii) プレート上でplaques中にβガラクトシダーゼ等との融合タンパク質を発現させ、iii) タンパク質を発現したplaquesをニトロセルロース膜に転写し、iv) 目的とする結合DNA配列を含むラベル化されたDNAを結合させ、v) DNAの結合したplaquesを回収する。これにより融合タンパク質として、用いたDNAに結合するタンパク質を発現しているplaquesを回収でき、このplaquesから直接目的とするcDNAが得られる。しかしながらこの技術では、i) 原理的にヘテロダイマー以上の複数種のタンパク質からなる複合体に対しては適用できること（1種類のcDNA由来のタンパク質のスクリーニングであるため）、ii) タンパク質の修飾（リン酸化、切断、アセチル化等）により初めてDNA結合できるタンパク質は取得できない、iii) non-specific DNA結合活性をもつタンパク質によるバックグランドがあり目的とする転写制御因子の取得効率も高いとはいえない。またこの方法で得られた転写制御因子が目的のものであるかどうかの判定には、実際にタンパク質を発現させゲルシフト法によりDNA結合性を確認すること、また転写制御活性については、cDNAを用いたCAT assay、Luciferase assay等による転写制御能の確認が必要である。

(2) One-hybrid 法

この技術は Two-hybrid 法の変法として開発されたもので、酵母を用いて目的

DNA配列に結合する転写制御因子を探索・同定・取得を行ったり、逆に転写制御因子の結合配列を特定する（ターゲットゲノム遺伝子を探索する）ことに使用する技術である^{22, 23)}。この技術の原理は以下の通りである。i) His3, LacZ 等のレポーター遺伝子のプロモーター領域に目的DNA配列を導入したレポータープラスミドを作製し酵母に形質転換する、ii) Gal4, VP16 等の転写制御因子の転写活性化領域にcDNAを結合させたcDNAライブラリーを作製する、iii) ライブラリーをi)で作製した酵母に形質転換し、iv) His3, LacZ 等のマーカーが活性化された（導入したcDNAによる融合タンパク質が転写制御因子として目的DNA配列に作用した）酵母をスクリーニングする。この技術により、融合タンパク質遺伝子として、目的DNA配列に結合するcDNAが直接に得られる。この技術は、先に報告したSouth Western 法で問題となる、i) 原理的にヘテロダイマー以上の複数種のタンパク質からなる複合体に対しては適用できないこと（1種類のcDNA由来のタンパク質のスクリーニングであるため）、ii) タンパク質の修飾（リン酸化、切断、アセチル化等）により初めてDNA結合できるタンパク質は取得できない、に関して、i) ではヘテロダイマーを形成する相手が判っている場合、現状では一つの遺伝子に関して事前に酵母に導入することで問題を回避し新規転写制御因子取得に成功した例がある²⁴⁾。またii) の問題に関しては、酵母におけるタンパク質修飾は動物細胞に近い部分があるので解決の可能性が十分にあると考えられる。しかしながら、転写制御複合体を相手とするにはまだまだ技術革新の余地が大きく存在し、酵母という扱い易さ、動物細胞と似た細胞内制御機構の存在といった利点が非常に大きいことから今後の新たな応用技術の開発が望まれる。本技術の変法として、特定転写制御因子を導入した酵母に対し、ゲノムDNAライブラリーをレポーター遺伝子と融合させたものを導入することで、転写制御因子のターゲットゲノムDNAが得られるという方法が考案・活用されている。

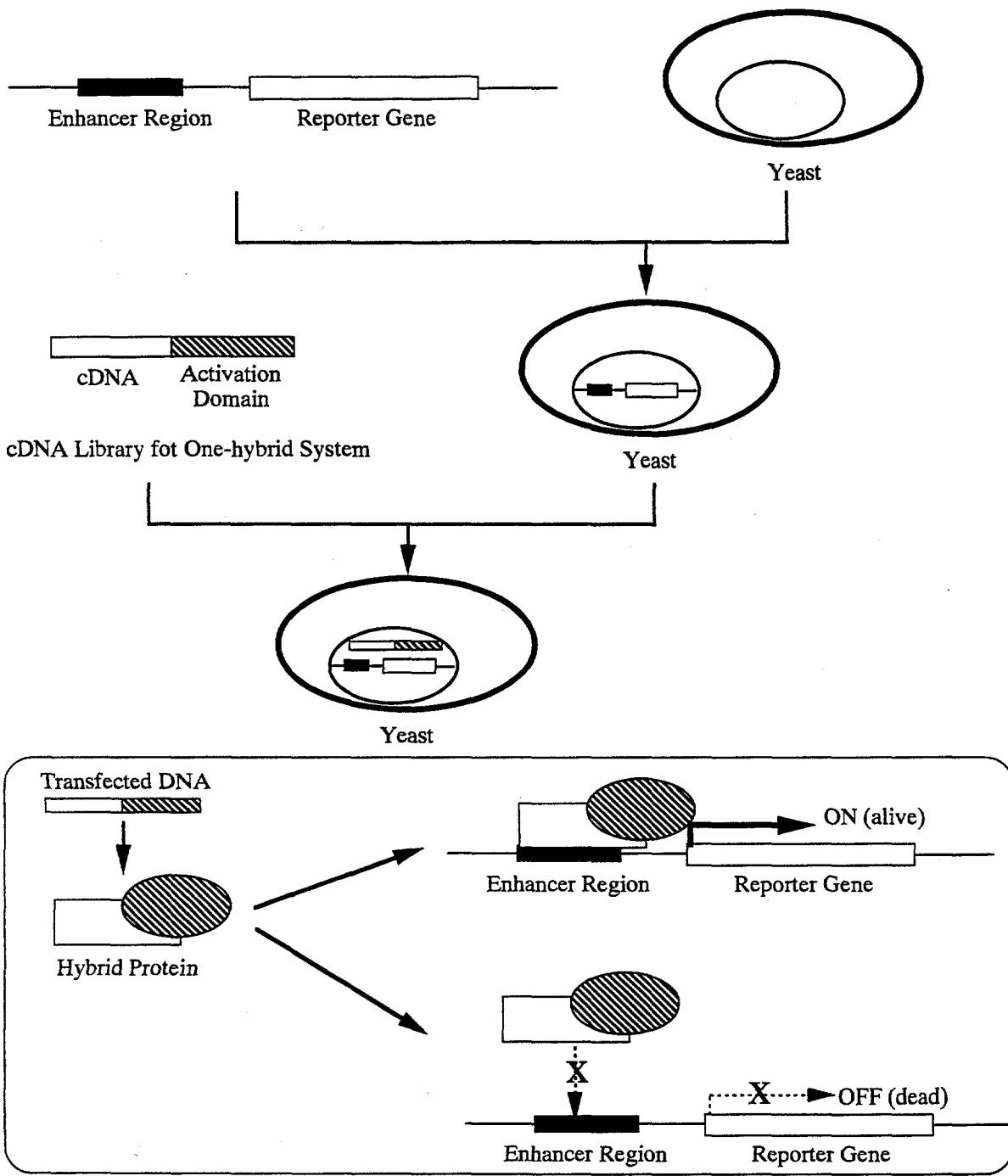


図4-4 One-hybrid法

(3) Phage display 法

ファージディスプレイの特徴は、タンパク質とそれをコードする遺伝子とが、1対1の関係で結び付いた分子を形成していることにある。すなわち、極めて多岐な性質を有するタンパク質をその特有の性質に従って分離した時、そのタンパク質をコードする遺伝子を同時に手に入れることができることから、タンパク質自体を解析するよりもはるかに効率的に、その1次構造を解析することが可能になる。さらに、ファージの増殖性を利用して、タンパク質と遺伝子を同時に無尽蔵に増幅することが可能である。同様の性質を有する実験系としては、ウイルスや細胞によるディスプレイの系が考えられる。ファージが大腸菌などの原核生物を宿主として増殖するのに対して、ウイルスは真核細胞を宿主として増殖することから、提示されたタンパク質が真核生物に特有の翻訳後修飾を受ける可能性がある。従って、ヒトなどの真核生物由来のタンパク質を解析する時には、ウイルスの系の方がより生体内に近い結果を得られる可能性が高い。しかし、大腸菌などの細菌に比較して、宿主となる細胞の増殖に手間がかかり、増殖速度が遅いことから、広く用いられるには至っていない。また、細胞表面への提示は、提示されたタンパク質のスクリーニングという観点から見た場合、ファージよりはるかに大きなマスによって、同じ実験規模でスクリーニングできる数が大幅に減ってしまう。最近では、ポリソームによる翻訳過程で生じた mRNA とポリペプチドとの結合体を利用することによって、タンパク質と遺伝子とを対応付けた分子の生成にも成功しており、ポリソームディスプレイと呼ばれている。しかし、それ自体には増殖性はなく、合成されたペプチドは mRNA がコードしている全長であるとは限らない。

ファージディスプレイは、以上のような性質から、膨大な数の異なる性質を持ったペプチドあるいはタンパク質から、目的の物質と特異的に相互作用するペプチドやタンパク質を迅速に探索するために最適である。さらに、酵母の One - hybrid 法や Two - hybrid 法とは違って、相互作用による結合の過程を試験管内で行うことから、相互作用の条件を厳密にコントロールすることが可能で、自動化にも向いている。しかし、ファージ粒子表面に提示されたタンパク質によって、一般的にファージの感染や増殖速度に影響を及ぼすことが観察され、特に完全長 cDNA を利用したような完全なタンパク質を提示させることは容易ではない。また、従来から広く使われている纖維状ファージ (M13 phage)^{25, 26)} の場合には、提示されるタンパク質が大腸菌の膜を通過して分泌される必要があり、提示されるタンパク質によっては分泌が不可能であったり、効率が

大幅に低下する。

そこで、近年では上記の分泌に関する問題点などを克服するために、溶菌性のバクテリオファージである、 λ ファージ^{27, 28)} や T7 ファージ²⁹⁾ を利用した、ファージディスプレイの系が開発されている。これらのファージでは、100 kDaを越すような大きさのタンパク質の提示が可能であり、さらに、 λ 系のファージでは4次構造の構築が可能で、全体で 465 kDa にもおよぶタンパク質の提示にも成功している²⁷⁾。この様な、タンパク質全体の提示が可能になることにより、本来のタンパク質が有する正確な機能を反映した選択が可能となり、さらに、従来の機能ドメイン単位での提示に比較して、はるかに少ない数による効率的なスクリーニングが可能になると考えられる。

今後、さらに大きなタンパク質を安定して提示するための工夫や、真核生物型の翻訳後修飾をするための改良などが期待される。

表4-2 ファージディスプレイ法の比較

	M13 (pJuFo)	T7 (T7-Select)	λ (λ foo / λ fooDc)
提示最大分子量 (実績)	72 kDa	c.a. 110 kDa	116 kDa (465 kDa,tetramer)
提示分子数	c.a. 1	0.1-1	> 0.5 / > 34
提示分子数の制御機構	無し	転写／翻訳レベル	サブレッサー
融合化に用いる遺伝子	gene 3 (Jun-pIII,Fos-prt.)	gene 10 (capsid)	gene V / gene D (tail) (capsid)
提示分子の分泌	必要	不要	不要
サブユニット構造構築	可能	不可能	可能

4-2-2 タンパク質-タンパク質間相互作用を利用した解析技術

転写制御因子による転写制御は、そのDNA-タンパク質間相互作用を除くほとんどがタンパク質-タンパク質間相互作用であり、転写複合体は最低でも数千kdaの巨大な複合

体で数十種類のタンパク質からなっていると考えられている。この転写複合体は、個々の遺伝子制御に於いて、その構成因子を一部交換、修飾することにより多様な遺伝子の微妙な制御を行っていると考えられている。このため転写制御のネットワークを解明する上で、転写制御複合体の構成因子を取得・解析することは非常に重要なものであり、そのための技術開発は緊急度の高い必須のものである。本節では、これまでに報告されているタンパク質-タンパク質間相互作用を利用した転写制御因子の取得・解析技術について代表的なものを以下に解説する。

(1) Far Western 法

この技術は、未知のタンパク質-タンパク質間相互作用を検出する技術として開発された技術である。2種類のタンパク質-タンパク質間相互作用を物理化学的に検出する方法として、ゲル滻過クロマトグラフィー、密度勾配遠心分離法³⁰⁾、架橋剤によるタンパク質間相互作用の共有結合化、等が挙げられる。Far Western 法は、先の South - Western 法のプローブDNAをラベル化されたタンパク質に置き換えたもので、他に Western blot メンブレンを用いた方法も数多く使用されている³¹⁾。この技術も South - Western 法同様に直接遺伝子取得につながる技術として効果的な方法と考えられる。具体的な原理は South - Western 法と同様であるが、タンパク質（転写制御因子等）を RI (³²P, ³⁵S, ¹²⁵I等) もしくは、ビオチン化や抗体等によりラベル化し検出に用いる点がDNAにはない部分で、またこの部分がSouth - Western 法より煩雑で難しい操作となっている。この技術の問題点は、i) バックグランドが高いこと（特にブラークスクリーニングの場合）、ii) 擬陽性の遺伝子を取得しやすいこと、iii) プローブがタンパク質であるため実験条件が個々に異なってくることが挙げられる。特に iii) に関しては、事前に Western blot メンブレンを用い目的のタンパク質とプローブのタンパク質が結合する条件を検討しておくことが非常に重要である。この技術は、記載した問題点により実際に目的とする転写制御因子が得られたかが判定が難しいため、一度試してみると当ても目的的な技術にとどまっているのが現状である。

(2) タンパク質固定化カラム法

この技術は、原理的には非常に単純で、アフィニティーコロマトグラフィーのリガンドに手元にある転写制御因子を用い、そのカラムにタンパク質溶液を流し、結合したタ

ンパク質を溶出・回収するというものである³²⁾。近年の遺伝子工学技術の進歩により高純度の均質な組み換えタンパク質を大量に調製できるようになり発達してきた技術である。i) 調製した組み換え転写制御因子タンパク質を共有結合、抗体等を用いてビーズの表面に固定し、ii) そこに核抽出液（場合によっては細胞・組織抽出液）を流し、iii) 十分にビーズを洗浄した後、iv) 塩濃度やpH等を変化させるなどして吸着したタンパク質を溶出し、v) 電気泳動により溶出タンパク質を分画し、vi) 必要に応じてタンパク質のアミノ酸配列を決定し遺伝子を取得する、という原理からなる技術である。しかしながら、転写制御因子の組み換えタンパク質は非常にaggregationが起こりやすいため大量に調製することが困難なものが多く、また細胞毒性のあるタンパク質も多くタンパク質発現が非常に難しい因子も多い、しかも核抽出液を大量に調製することが大変な量力を要すること等から極限られた対象でしか適用が難しい技術である。しかし、タンパク質レベルで機能的に結合する因子を同定できるため非常に魅力的な技術であり、今後の改良が待たれる。

(3) タグ付きタンパク質応用技術

この技術は、アフィニティー結合可能なタグ配列を結合させた転写制御因子タンパク質を用いたものである。従来より、タグ付きの因子を細胞中で発現させ、タグを用いてその因子を含む複合体を分離することは細胞質タンパク質を主として行われてきたが、転写制御因子等の核内因子では因子の分析目的では報告があるが取得目的では動物細胞中では核内での高レベルのタンパク質発現がうまくいかない、タグがうまく機能しない、等の理由よりほとんど例がない。しかし極最近になって、日本独自の技術として*in vitro* ではあるがタグ付きタンパク質を利用した転写制御因子複合体の取得技術が萌芽的レベルで構築・報告された^{33, 34)}。その技術の原理は、i) Hisタグ付き転写制御因子（この場合TBP）を大腸菌で発現・精製し、ii) Hisタグ付き転写制御因子をNi-アフィニティーゲルに固定し、iii) 予備精製した核抽出液と混合、iv) ゲルをよく洗浄し、v) イミダゾールでタグ付き転写制御因子ごと溶出し、vi) 電気泳動で溶出タンパク質を分離し、vii) アミノ酸配列を決定し、遺伝子取得を行う、というものである。この技術では最低でも8種類のTBP結合性タンパク質が取得され、その中に2-3種類の新規タンパク質が含まれていることが判明し、遺伝子取得に至っている。この技術は、大腸菌を用いて純度の高い転写制御因子を調製し、大量に高純度な核抽出液を調製できる肝

臓組織より核出液を調製することで、課題となりうる部分を克服している点が特筆される。しかし、この技術でははじめから用いる転写制御因子に強固に結合している因子は、原理的にタグ付き転写制御因子に結合しなおす可能性が低いため取得が難しいこと、適用範囲に制限が存在することなどの問題が存在する。しかし、タグ付き転写制御因子はタグを対象とした精製・取得技術であるため普遍性が高く、精製等の技術レベルも高く設定できることから今後の発展が大きく期待される。

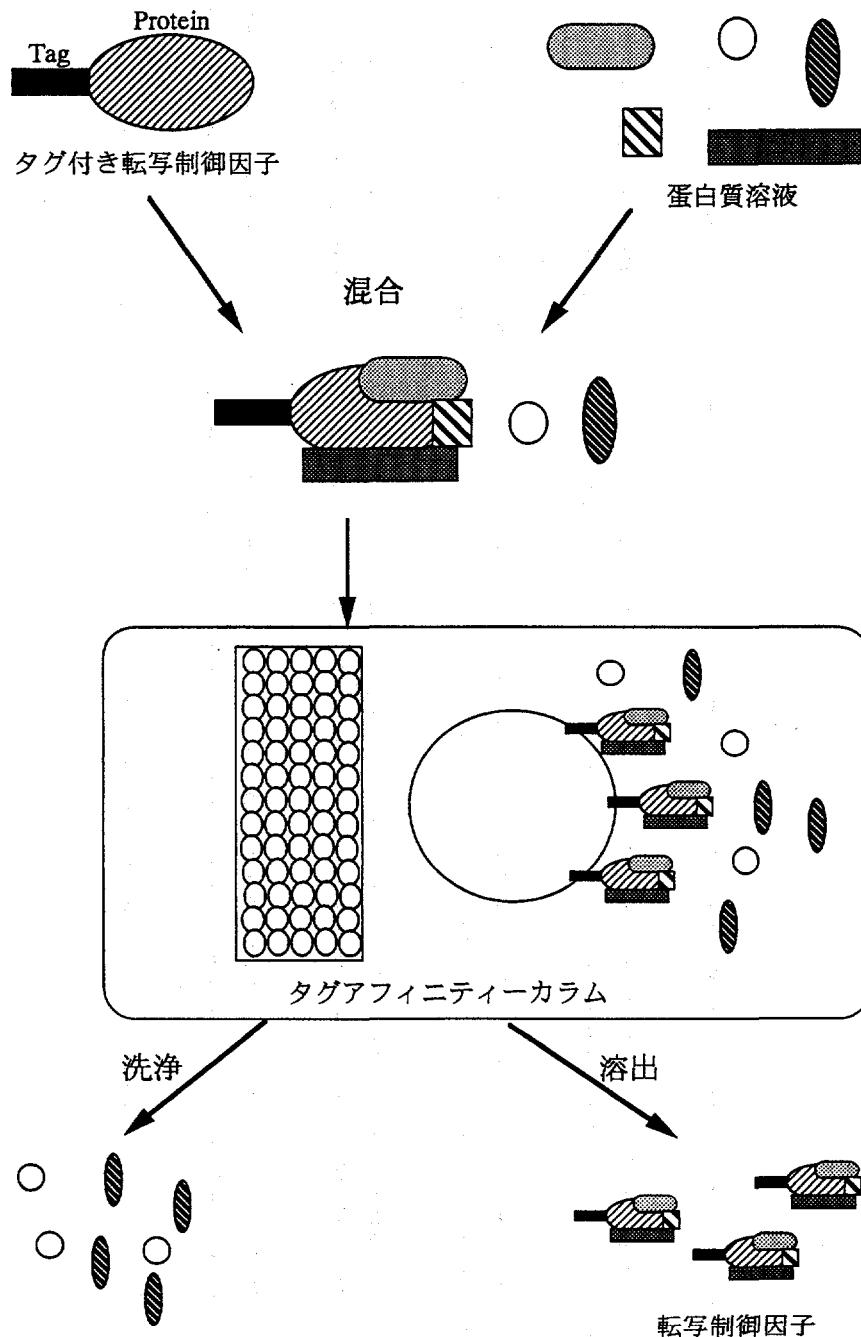


図4-5 タグ付きタンパク質による転写制御因子の解析法

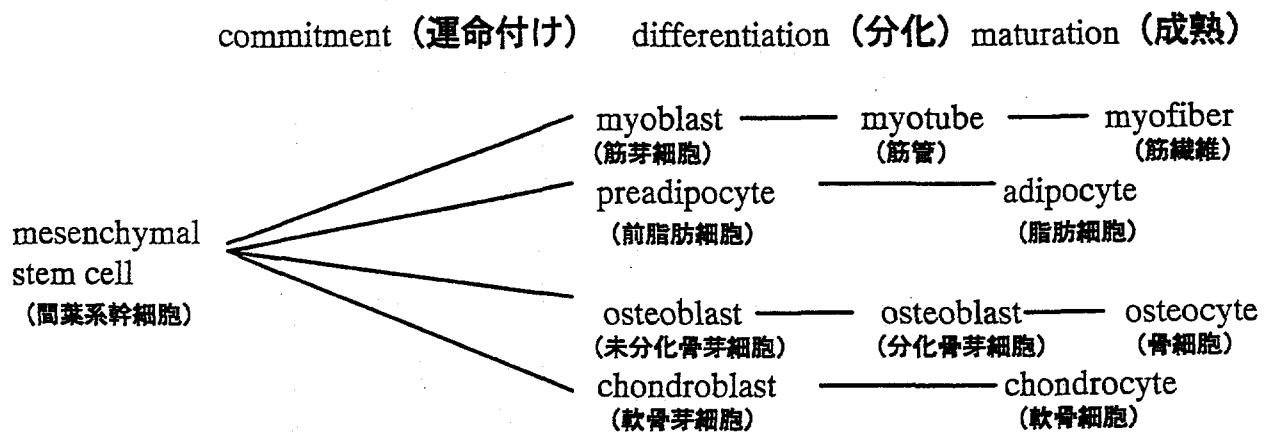
4-2-3 その他の転写制御因子の探索技術

ゲノム情報の解析が進み個々の遺伝子のcDNA配列のみならず、プロモーターやエンハンサーに関する情報が急速に蓄積しつつある。その結果、多数の転写因子が共同して個々の遺伝子の発現調節を行っている、いわば発現調節ネットワークが存在することが判明しつつある。ヒトゲノムシーケンス解析が終了し、7万とも十万ともいわれる遺伝子のほとんどが明らかになった時点では、DNAチップを用いたMicro arrayなどにより全遺伝子についてその発現レベルが測定できるようになり、細胞分化を含め様々な生物学的現象について時間を追って全遺伝子の発現レベルを測定することで、遺伝子群カスケードを解析することが可能になると考えられる。

(1) 細胞分化に伴う遺伝子発現の変動とその制御

分化した細胞は、細胞種毎に形態や機能を異にし、それぞれ固有の形質を示す。このような細胞種毎の形質の違いは、遺伝子発現の質的量的違いによるものであり、細胞種特異的な遺伝子（分化マーカー遺伝子）群が固有の転写制御により固有のバランスで発現することで細胞種固有の形質を示すと理解されている。例えば筋分化を例にとると、MyoDファミリーと総称される一連の転写因子群（MyoD, Myf5, Myogenin, MRF4）は、骨格筋特異的に発現し、筋分化を促進的に制御しているが、これら筋分化制御転写因子群自身が筋分化過程で厳密に設定されたプログラムにより発現制御されていることが分かってきた（図4-6）。すなわち、MyoDとMyf5は、筋芽細胞の決定維持に機能していると考えられ、機能的に重複しており、Myogeninの発現を正に制御している。Myogeninは、筋管形成に必須の働きをしており、筋特異的遺伝子の発現制御を通して、筋細胞が固有の形質を表すための鍵分子として働く。また、これらMyoDファミリー遺伝子群の発現制御は、階層構造やフィードバックによる制御を受け、筋分化に抑制的に転写因子（Id）の存在も明らかにされている。このような細胞分化に伴う発現調節ネットワークの存在は、筋分化以外にも一般化され得る。事実、脂肪細胞や骨芽細胞においてもそれぞれC/EBP α 、PPAR γ 2やCbfa1などの特異的転写制御因子（群）が、細胞分化を規定していることが明らかになってきた。これらの細胞において、未だ細胞分化の機構を転写制御ネットワークで示せるほどの情報の蓄積はないが、研究の進展に伴いそれぞれの細胞分化を制御する新たな転写因子群が見いだされなければ、筋分化で明らかにされたのと同じように転写制御ネットワークの存在が明らかになっていく

ものと期待される。細胞分化と遺伝子発現制御の連関を中胚葉系の細胞分化を例に述べたが、基本的な転写制御の概念は、血球細胞や神経細胞の分化・発生過程においても同様と考えられ、個々の系で働く因子の同定が進むに連れて、発現調節ネットワークの理解が深まっていくものと思われる。



細胞分化を規定する核内受容体・転写因子

筋分化：正の制御因子（myf 5, Myo D, E2A, Myogenin, MRF4）

負の制御因子（Id, 増殖因子）

脂肪細胞分化：PPAR γ 2, ADD1, C/EBP α , C/EBP β , C/EBP δ

骨芽細胞分化：ADD1, Cbfa1

軟骨細胞分化：不明

図 4 - 6 間葉系細胞の細胞分化における転写因子カスケード

(2) 分化マーカー遺伝子の利用

発現調節ネットワークの理解が進むにつれて、様々な分野で遺伝子発現情報が利用されるようになろう。現在も、細胞機能の異常を分化マーカー遺伝子の発現をモニターすることにより捉えることが出来るが、こうした遺伝子発現情報をさらに有効に利用する必要がある。インスリン抵抗性糖尿病を例にとると、II型糖尿病においては脂肪細胞特異的分化マーカー遺伝子の発現抑制が認められるが、このような分化マーカー遺伝子の発現量に関する情報は、転写制御ネットワークが明らかになるにつれて、病態の理解や治療方法の改善に資するようになりつつある。すなわち、TNF α のノックアウトマウスが、インスリン抵抗性になりにくいという事実から、上記したインスリン抵抗性は、脂肪細胞におけるTNF - α の産生亢進による脂肪細胞特異的分化マーカー遺伝子の発現抑制に起因すると考えられる。脂肪細胞の分化を規定するリガンド誘導性転写因子(核内受容体PPAR γ 2)のリガンドをII型糖尿病患者に投与することでインスリン抵抗性が改善することから、薬物による転写制御で脂肪細胞における分化マーカー遺伝子の発現を制御することで疾患の治療が可能であることを示している。疾患病変部における遺伝子発現パターンが、正常人のそれとどのように異なり、また、どのような遺伝子発現制御を行えば、正常人の発現パターンに近づけることが出来るかが明らかになれば、治療薬創製を目指した研究が可能になってくるものと考えられる。即ち、細胞分化の遺伝子発現制御機構を明らかにすることで、単に生物学的現象の理解が進むだけでなく、治療薬の探索という応用研究が飛躍的に進むものと思われる。逆にいえば、細胞分化の遺伝子発現制御機構を理解せずに、徒に治療薬の探索を行えば、世界的な創薬競争に敗北は必至となろう。

(3) 培養幹細胞と分化

現在の培養細胞系のデータから、“細胞分化機構”という概念を再定義すると、“均一な細胞集団が不均衡を生み出す仕組み”と言えるかも知れない。細胞集団が何らかの刺激に、一時的かつ一様に、形態変化を起こすような現象と“分化”とは一線を画す必要がある。すなわち、細胞分化とは“不等分割を伴う細胞増殖過程”ではないかと想像される。こうした概念は、分化能を保持した“培養幹細胞系”を利用したデータから樹立されつつある。筋細胞、脂肪細胞、骨芽細胞及び軟骨細胞は、ともに同一の細胞系列から分化することが知られているが、実際、図4-7に示したような幹細胞を用いた培養系

で、培養条件を選択することにより、分化の方向性を決めて細胞分化を研究することが可能と考えられ、事実、一部については、可能になっている。このような培養細胞系において発現遺伝子の変動をモニターすることにより、細胞分化を規定する転写制御過程の解明が可能となるはずである。既知転写制御因子について分化した細胞での発現を確認し、発現しているもの全てについて分化過程における発現変動を測定し、発現時期を特定することにより、現在未同定の分化制御因子に関する情報も得られるようになる。

筋分化に関与する制御因子

正の制御因子：myf5, MyoD, E2A, Myogenin, MRF4

負の制御因子：Id, 増殖因子

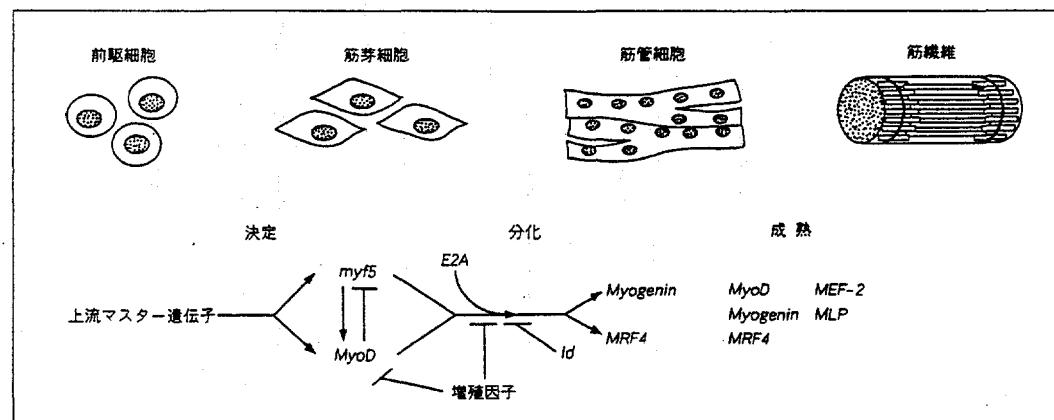


図4-7 細胞分化における転写因子カスケードの例

(4) 転写制御ネットワークモデルの構築

このように、転写制御ネットワークを理解することにより、細胞分化など根幹的生命現象を解明できるものと思われるが、その際に注意すべき点は、単に個々の遺伝子発現をON/OFFのようにとらえるのではなく、多数の遺伝子発現量を定量化し、時間軸や空間軸を考慮に入れた多変量解析モデルを構築する必要があるということである。従って、転写制御ネットワークの解明は、多数の遺伝子発現量を定量的にモニターする新しい解析系や、定量化した膨大なデータを処理し利用するためのコンピュータ解析系の充実と密接にリンクしている。詳しくは、第5章を参照していただきたい。

(5) 転写制御因子の概念の拡張について

近年の転写調節研究の成果として、直接DNAに結合する転写因子以外の転写制御因子の重要性がクローズアップされつつある。例えば、転写因子CREBやNF-kBではリン酸化による蛋白修飾反応が、その活性制御に重要で、種々のリン酸化酵素や脱リン酸化酵素の活性レベルに関する情報も、転写制御ネットワークの理解には不可欠となっている。また、CREB結合蛋白としてクローニングされたCBPは、転写介在因子（コファクター）として多くの転写因子に結合し、基本因子群とも複合体を作ることが明らかとなった。転写介在因子として、CBPの他に、P300、P/CAF、RAc-1、NcoR、SMRT、SRC-1など多数がクローニングされ、こうした転写介在因子の多様性が、転写制御ネットワークによる転写制御の多様性を支えているものと思われる。さらに、転写介在因子はコアクチベーターと呼ばれる活性化因子の他に、HDAC1などコリプレッサーと呼ばれるものが存在することが明らかとなった。興味深いことには、コアクチベーターにはヒストンアセチル化酵素（HAT）活性が、コリプレッサーにはヒストン脱アセチル化酵素活性が証明され、クロマチン蛋白のアセチル化状態を調べることが転写制御ネットワークを解明する上で必要となる。さらに、HATは現在のところ、ヒストンを基質にして調べられているが、他の核蛋白のアセチル化が転写制御に重要な役割を果たしている可能性も否定できない。蛋白アセチル化反応は蛋白リン酸化反応と同様、遺伝子発現を調節する重要な細胞内蛋白修飾反応として位置づけられる可能性がある。

(6) 転写過程以外での遺伝子発現制御

前段まで、“転写調節=遺伝子発現調節”のような書き方をしてきたが、これは実は正確ではない。というのは、一つの遺伝子より転写されたmRNAが選択的スプライシングを受けることにより、異なった蛋白が産生される（alternative splicing）ため、遺伝子発現調節機構の全容を知るために選択的スプライシングを考慮する必要がある。選択的スプライシングは多細胞生物に特有の現象で、これにより生命体は一層の多様性を獲得しているものと思われる。*C.elegans*でゲノム解析が進んでいるが、選択的スプライシングがあるため、そのデータを元にORFを予測しても、実際にどのような蛋白が発現するかは、酵母のようにゲノム情報からだけでは予想できない。現在知られている alternative splicingのパターンは図4-8に示したように7種類に大別できるが、個々の遺伝子発現でその制御がどのようになされているかについて、ほとんど解明されてお

らず、選択的スプライシングの分子機構については、転写制御過程以上に不明の点が多いといえる。最近、AG/GU … CAG/G以外の配列でスプライスされる irregular splicing の例も報告されている。ヒトが *C.elegans* 以上に多様なスプライシングを行っていることを考えると、ヒトゲノム情報の利用においても、選択的スプライシングの問題は大変な重要性を秘めていると言わざるを得ない。

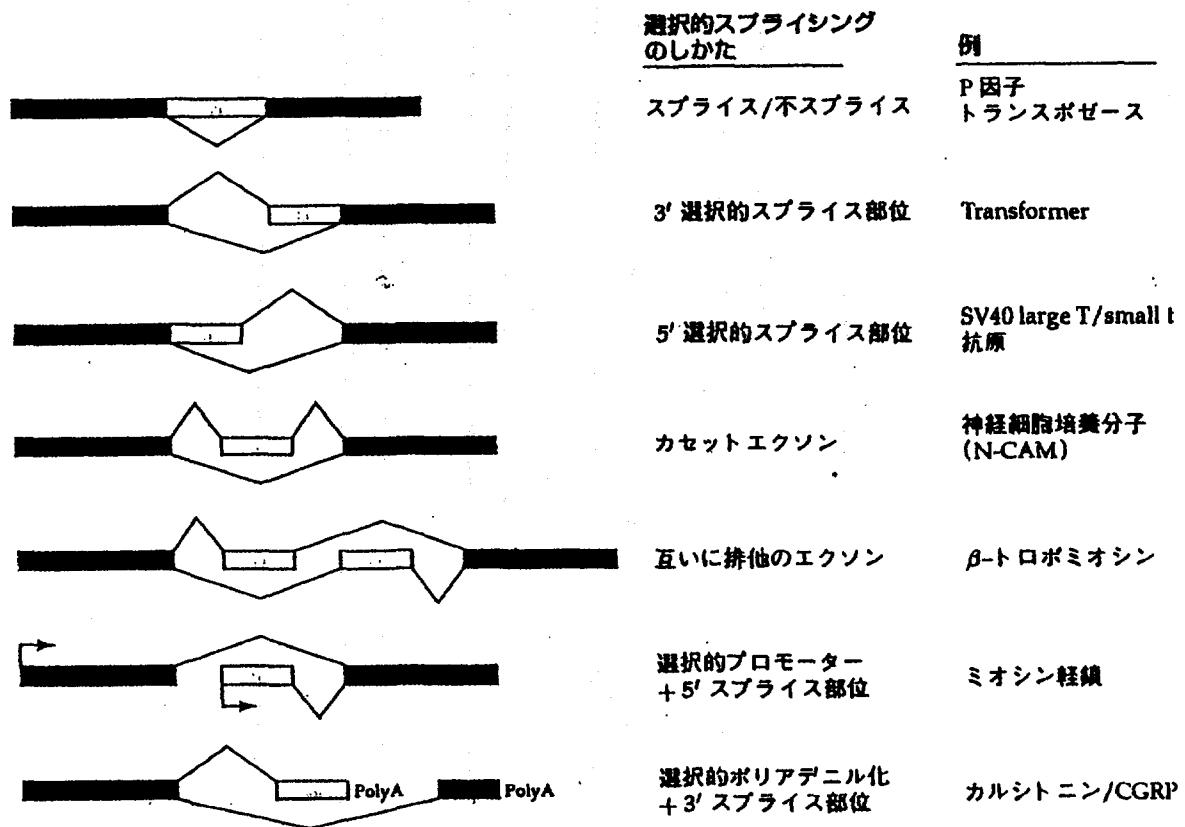


図4-8 選択的スプライシングの諸パターン

文献

- 1) Suzuki, Y., Yoshitomo - Nakagawa, K., Maruyama, K., Suyama, A. and Sugano, S. : *Gene*, 200, 149 - 156 (1997)
- 2) 矢田哲士、石川幹人、田中秀俊、浅井潔：情報処理学会論文誌、37, 1117 - 1129 (1996)
- 3) Yada, T., Sazuka, T. and Hirosawa, M. : *DNA Res.* 4, 1 - 7 (1997)
- 4) Fondrat, C. and Kalogeropoulos, A. : *CABIOS*, 12, 363 - 374 (1996)
- 5) Prasher, A.P., Eckenrode, V.K., Ward, W.W., Prendergast, F.G. and Mormier, M.J. : *Gene*, 111, 229 - 233 (1992)
- 6) Galas, D. and Schmitz, A. : *Nucleic Acids Res.*, 5, 3157 - 3170 (1978)
- 7) Garner, M.M. and Revzin, A. : *BioTechniques*, 7, 346 - 355 (1989)
- 8) Li, Q. and Wrangle, O. : *Methods*, 12, 96 - 104 (1997)
- 9) Walter, J., Dever, C.A. and Biggin, M.D. : *Genes & Dev.*, 8, 1678 - 1692
- 10) Inoue, S. et al. : *Proc. Natl. Acad. Sci. USA*, 90, 11117 - 11121 (1993)
- 11) Watanabe, T. et al. : *Mol. Cell. Biol.*, 18, 442 - 449 (1996)
- 12) Kishimoto, T. et al. : *Biochem. Biophys. Res. Commun.* 223, 746 - 751 (1996)
- 13) Konishi, Y. et al. : *J. Neurochem.*, 69, 476 - 484 (1997)
- 14) Comai, L. et al. : *Cell*, 68, 965 - 976 (1992)
- 15) Wright, W. E. and Funk, W. D. : *TIBS*, 18, 77 - 80 (1993)
- 16) Clauss, I. M. et al. : *Nucleic Acids Res.*, 24, 1855 - 1864 (1996)
- 17) Tao, Y. et al. : *Mol. Cell. Biol.*, 17, 6994 - 7007 (1997)
- 18) Tamura, T. et al. : *Methods: A companion to methods in enzymology*, 10, 312 - 319 (1996)
- 19) Singh, H. et al. : *Cell*, 52, 415 - 429 (1988)
- 20) Vinson, C. R. et al. : *Genes Dev.*, 2, 801 - 806 (1988)
- 21) Jupin, I. and Chua, N. H. : *EMBO J.*, 15, 5679 - 5689 (1996)
- 22) Wang, M. M. and Reed, R. R. : *Nature*, 364, 121 - 126 (1993)
- 23) Brachmann, R. K. and Boeke, J. D. : *Curr. Op. Biotech.*, 8, 561 - 568 (1997)

- 24) Naya, F. J. *et al.* : *Genes Dev.*, 9, 1009 – 1019 (1995)
- 25) Fowlkes, D.M., Adams, M.D. Fowler, V.A. and Kay, B.K. : *BioTechniques* 13, 422 – 428, (1992)
- 26) Kay, B.K. Adey, N.B., He, Y.S., Manfredi, J.P., Mataragnon, A.H. and Fowlkes, D.M. : *Gene*, 128, 59 – 65 (1993)
- 27) Maruyama, I.N., Maruyama, H.I. and Brenner, S. : *Proc. Natl. Acad. Sci. USA*, 91, 8273 – 8277 (1994)
- 28) Sternberg, N. and Hoess, R.H. : *Proc. Natl. Acad. Sci. USA*, 92, 1609 – 1613 (1995)
- 29) Rosenberg, A., Griffin, K., Studier, F.W., McCormick, M., Berg, J., Novy, R. and Mierendorf, R. : *inNovations*, 6, 1 – 11 (1996)
- 30) Kato, K. *et al.* : *Nucl. Acids Res.*, 22, 1179 – 1185 (1994)
- 31) Makino, Y. *et al.* : *Biochem. Biophys. Res. Commun.*, 220, 1049 – 1054 (1996)
- 32) Hughe, P. and Baldacci, G. : *Nuclic Acids Res.*, 25, 3881 – 3888 (1997)
- 33) Yogosawa, S. *et al.* : *Biochem. Biophys. Res. Commun.*, 229, 612 – 617 (1996)
- 34) Makino, Y. *et al.* : *J. Biol. Chem.*, 272, 23201 – 23205 (1997)

第5章 コンピュータモデル解析技術

5-1 遺伝子領域・機能予測技術

シークエンシング技術の進歩により、ゲノムの塩基配列が高速に決定されるようになつたために、塩基配列の情報だけから実験によらず遺伝子の領域や機能を予測する技術の重要性が増してきている。また、情報学的な立場からも、複雑メカニズムを持つ生物の遺伝子の機能の記述とその制御がどのように行なわれているのかを知ることは興味のあるところである。

遺伝子領域の予測技術は、主にコード領域の予測についてその精度が急速に向上去っており、比較的容易な原核生物だけでなく、より複雑なモデル化が必要な真核生物に対しても、信頼性の高いシステムが公開されるようになってきている。さらに、近年は単に遺伝子領域を予測するだけでなく、遺伝子の機能も併せて予測し、塩基配列データによる自動アノテーションを行なうことを目指す動きが注目を集めている。

5-1-1 原核生物の遺伝子領域予測技術

現在の原核生物の遺伝子領域予測技術は、新たな段階への移行時期と捉えることができる。原核生物ではさまざまな種をモデルとしたシークエンシングプロジェクトが推し進められ、いくつかの種ではすでに全ゲノム配列が決定されている。これまでの遺伝子領域の予測技術でも、比較的高い精度でコード領域を予測することが可能であった。しかし、予測結果を詳細に検討してみると以下のようないくつかの問題点が浮き彫りにされてきた。

問題（1） 翻訳開始点の予測精度が低い

問題（2） 短いコード領域の予測精度が低い

問題（3） 膨大なアノテーション情報が予測モデルの構築に必要

これらの問題点に対して、既存の予測技術の改良や新しい予測技術の研究開発が盛んに行なわれている。

問題（1）を解決するために、代表的な原核生物の遺伝子領域予測プログラムである GeneMark¹⁾ は隠れマルコフモデル（hidden Markov model : HMM）の枠組を取り

り入れた (GeneMark.hmm 図5-1)²⁾。ここでは、遺伝子間領域 ==> 開始コドン ==> コード領域モデル ==> 終始コドンのサイクルがモデル化され、さらに相補鎖に存在する遺伝子のモデル (shadow gene model) が付与されている。コード領域モデルには統計量の異なる2種類のモデルが定義され、これらが従来の GeneMark に相当する。この改良によって、*E.coli*ゲノムにおいて 68 % であった翻訳開始点の予測精度が 79 % に改良されたことが報告された。また、コード領域の予測精度は 95 % で、若干の改良が観察された。しかし、HMM を導入したことによって、従来は予測可能であった overlapping gene が検出できなくなった。このため、GeneMark の予測結果も反映させて最終的な予測結果としている。また、より正確な翻訳開始点の予測のために、ribosomal binding site (RBS) のプロファイルを利用することが検討されている³⁾。

さらに、問題 (3) を解決するために、クラスタリングを繰り返しながら GeneMark を構築する GeneMark - Genesis と呼ばれるプログラムが開発された⁴⁾。

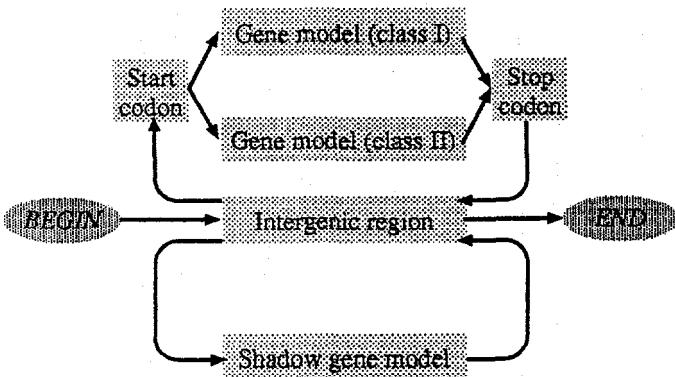


図5-1 GeneMark.hmm に採用された遺伝子領域モデルの概要

HMM を応用した原核生物の遺伝子領域予測プログラムである GeneHacker⁵⁾ は、問題 (1) を解決するために、RBS のプロファイルをモデルに組み込んだ⁶⁾。

その結果、*Synechocystis*ゲノムにおける翻訳開始点の予測精度が 83.3 % に改善されたことが報告された。また、コード領域の予測精度は GeneMark をはるかに上回るものであった。さらに、GeneHacker は問題 (2) に対してロバストであることが示された⁶⁾。 *Synechocystis*ゲノムにおける 150nt 以下のコード領域について、GeneMark の予測精度が 21.4 % であるのに対して、GeneHacker の予測精度は 47.4 % であった。さらに、問

題(3)を解決するために、アノテーション情報なしのゲノム配列のみから、遺伝子領域の予測とGeneHackerのコード領域のモデルを構築する自己学習アルゴリズムが開発された⁷⁾。このアルゴリズムによって学習されたGeneHackerは、アノテーション情報ありのゲノム配列から作成されたモデルと同等の予測精度を獲得することが示された。

新しい原核生物の遺伝子領域予測プログラムとしてGLIMMERが開発された⁸⁾。GeneMarkは次数を固定したMarkov modelに基づく遺伝子領域予測プログラムであるのに対して、GLIMMERはinterpolated Markov model (IMM)と呼ばれる次数を固定しないMarkov modelに基づく遺伝子領域予測プログラムである。

*H. pylori*や*H. influenzae*ゲノムにおける遺伝子領域予測の結果、GLIMMERはGeneHackerと同じ程度の高いsensitivityとspecificityを示した。また、短いコード領域に関する予測精度もGeneHackerと同程度であった。GLIMMERの大きな特徴は、モデルの学習と認識の処理速度が高速であることである。

5-1-2 真核生物の遺伝子領域予測技術

最近、真核生物の遺伝子領域予測技術を取り巻く状況は大きく変化した。ヒトをはじめとする真核生物のシークエンシングプロジェクトは順調にデータを産出している。これまでに蓄積されたデータを加えると、より細かなデータの特徴を記述した遺伝子領域の予測モデルを構築することが可能になった。このため、真核生物の遺伝子領域の予測精度は大きく改善されることとなった。しかしながら、この予測技術は遺伝情報の転写やスプライシングや翻訳のメカニズムを十分に取り込んでいるわけではないので、新しい視点に立った予測技術の研究が望まれている。真核生物の遺伝子領域予測技術に関する状況は以下のように整理することができる。

- ・従来の予測精度を大きく上回る予測技術の出現
- ・予測精度の改善を目指した従来の予測技術の改良
- ・新しい視点に立った予測技術の研究

真核生物の遺伝子領域予測技術は、実用性と予測精度の面でGENSCAN⁹⁾が画期的な予測率を示した。GENSCANは、一般化隠れマルコフモデル(GHMM)を応用して真核

生物の遺伝子構造をモデル化している。このため、予測結果の確からしさは確率によって数値化されている。また、GENSCAN は、ゲノム配列に含まれる複数の遺伝子や部分的な遺伝子を相補鎖も含めて予測することができる。さらに、ゲノムデータを細かく分類して統計量を算出することによって、従来の手法をはるかに上回る予測精度を実現している。

GENSCAN の出現に触発されて、新しい真核生物の遺伝子領域予測プログラムが次々と報告された。VEIL¹⁰⁾ や HMMgene¹¹⁾ や GeneDecoder¹²⁾ といった HMM を応用した予測プログラムが提案されたが、実用性や予測精度の面で GENSCAN を上回ることはできていない。

また、転写因子の組合せに着目してプロモータ領域を予測する手法が提案された¹³⁾。

一方、従来の遺伝子領域予測プログラムでは、予測精度の改善を目指してさまざまな改良が行なわれている。決定木を応用した真核生物の遺伝子領域予測プログラム MORGAN¹⁴⁾ では、塩基の 2 連性に着目したスプライス部位と転写開始点の重み行列 (positionalweight matrix : PWM) を開発し¹⁵⁾、予測精度の改善を試みている。GHMM を応用した真核生物の遺伝子領域予測プログラム Genie¹⁶⁾ では、塩基の 2 連性を学習させたスプライス部位のニューラルネットワーク (artificial neural network : ANN) を開発し¹⁷⁾、予測精度の改善を試みている。ANN を応用した真核生物の遺伝子領域予測プログラム GRAIL¹⁸⁾ では、データベースに大量に蓄積された EST (expressed sequence tag) をホモロジー検索の対象とすることによって¹⁹⁾、スプライス部位や転写開始・終結点などの予測精度の改善を試みている。線形判別法に基づいた真核生物の遺伝子領域予測プログラム FGENEH²⁰⁾ では、転写終結点の予測モデルを開発し²¹⁾、予測精度の改善を試みている²²⁾。

以上の予測技術は、コドンの出現頻度の偏りやコード領域における出現塩基の周期性などのコーディングポテンシャルを利用して真核生物の遺伝子領域を予測している。一方、遺伝情報の転写やスプライシングや翻訳に関与する生体分子が、これらのポテンシャルを認識しているとは考えにくい。遺伝子領域の予測プログラムは生体内で行なわれる現象を理解するために研究・開発されると考えると、理想的なプログラムは生体内で認識されるのと同じシグナルを解釈して転写・スプライシング・翻訳を行うべきである。この視点に

立って、コーディングポテンシャルの情報を使わずに遺伝子領域を予測するプログラム AMELIEが開発された²³⁾。AMELIEにはいくつかの制約が設けられているものの、その予測精度は GRAIL と同じ程度であった。AMELIEには改良の余地が多く残されているが、より本質的な真核生物の遺伝子領域予測技術のブレイクスルにつながるアプローチとして、今後の研究展開が大変興味深い。

5-1-3 コード領域の配列情報に基づいた遺伝子機能予測技術

現在、微生物ゲノムプロジェクトでは、シークエンシングプロジェクトに引き続いで機能解析プロジェクトが立ち上がりうとしている。一般に、遺伝子機能の類推は、コード領域のホモロジー検索によって行なわれる。しかしながら、シークエンシングプロジェクトの結果、データベースにホログが登録されていない遺伝子が数多く報告されている。これらの遺伝子では、コード領域のホモロジー検索による機能の類推が困難であるため、新しいタイプの遺伝子の機能予測技術の確立が望まれている。一方で、ホモロジー検索で得られるスコアとタンパク質の EC (enzyme commission) クラスとの関連を網羅的に評価し、ホモロジー検索による機能予測の精度を改善させようとする試みも研究されている²⁴⁾。

タンパク質の部分配列で、機能と深い関わりのあることが知られているものを、モチーフと呼んでいる。モチーフは、タンパク質のアミノ酸配列から機能予測に用いられている。

モチーフのデータベースとして、Bairoch らの PROSITE や Henikoff らの BLOCKS が昨年度の調査で報告されている。

配列情報からのモチーフ抽出については、多くの試みがなされている。昨年度の調査報告で、統計的解析を行う K-tuple 解析や、隠れマルコフモデルなどの確率モデルや確率的探索アルゴリズムを用いる確率的モチーフ抽出が報告されている。

国立遺伝学研究所のグループでは、進化系統樹を用いた塩基配列データベース (DDBJ) の網羅的な解析によってモチーフの抽出を行なっている^{25, 26)}。ここでは、相同な遺伝子間に共通な遺伝子領域をモチーフとして抽出するために、約14万のアミノ酸配列データを相同性に基づいて約5万のファミリーに分類し、2個以上の遺伝子からなる約2万ファミリーからマルチプルアラインメントによって約13万個のモチーフ領域の候補を

抽出している。

ホモロジー検索に頼らない遺伝子機能の予測技術として、アミノ酸配列のシグナルペプチドの配列情報に基づいたタンパク質の局在化部位予測が挙げられる²⁷⁾。この予測技術では、モデルパラメータの最適化とあわせて、さまざまな機械学習アルゴリズムに基づく分類器の開発が行なわれている。現在では、k nearest neighbor 法²⁸⁾や HMM²⁹⁾による予測技術が開発され、予測精度の改善が報告されている。

新しいタイプの遺伝子機能予測技術として、アミノ酸配列から算出されるさまざまな物理化学量や統計量に基づいてタンパク質の機能を予測する手法が報告された³⁰⁾。

ここでは、機械学習によって構築された分類器を用いて、タンパク質のアミノ酸配列長、アミノ酸組成、分子量、電荷、等電点、消衰係数、2次構造情報などから、(1) そのタンパク質は酵素か否か (2) もし酵素なら、そのECクラスは何か、を識別する。計算機実験の結果、この分類器は、質問 (1) について 74%、質問 (2) について 68% の識別精度を示した。現在のところ、この分類器は酵素を識別する能力しかないため、より網羅的な遺伝子産物のクラスを識別する分類器の開発が望まれる。

5-1-4 制御領域の配列情報に基づいた遺伝子機能予測技術

遺伝子の発現制御の第1段階は転写の調節である。よって、同時に転写されるべき遺伝子のプロモータは、同じ転写因子のセットに制御されていると考えられるため、同じような配列パターンを示すことが期待される。このため、プロモータの配列パターンは下流に存在する遺伝子の発現条件を類推する手がかりを提供する³¹⁾。

微生物では、複数のシグマ因子を使い分けることで、いろいろな状況に適した遺伝子の発現調節を行っている。このことを利用して、プロモータ配列のシグマ因子依存性を予測することで下流の遺伝子群の発現条件を予測する手法が提案された³²⁾。ここでは、枯草菌について、シグマ因子依存性が既知のプロモータ配列を文献から収集し、それをアラインメントして、各位置における塩基の保存度をカイ二乗検定で判定する。その結果に基づいて、各シグマ因子のプロモータ認識能力をシミュレートする HMM を構築した。

クロス検定の結果、これらのモデルは 75.5% の精度でプロモータのシグマ因子依存性を

識別することが明らかになった。また、これらのモデルを実際のゲノム配列中のORF上流配列に適用して、高いスコアを示す遺伝子を調べたところ、学習データに用いたものや機能未知のものが多い中で、そのアミノ酸配列から考えて、正しい予測ができていると思われる遺伝子がいくつか見つかった。これらの計算機実験の結果は、プロモータのシグマ因子依存性に基づく微生物遺伝子の機能予測手法が妥当であることを示している。

真核生物では、プロモータの保存配列断片のセットから下流の遺伝子の組織特異的な発現を予測する手法が提案された³³⁾。ここでは、肝臓で特異的に発現する遺伝子のプロモータ群を用意し、マルコフ鎖と二項分布の仮定に従って統計的に有意に出現する部分配列群を取り出す。部分配列群は、アライメントされたのちに各位置における情報量を計算することで配列長が調節される。このようにして取り出された部分配列群を線形判別に適用することによって、肝臓で特異的に発現する遺伝子のプロモータを識別する。計算機実験の結果、この手法は、肝臓で特異的に発現する遺伝子のプロモータを62.9%の精度で識別することが明らかになった。この時、ハウスキーピング遺伝子のプロモータに関する識別精度は77.6%であった。現在のところ、この予測手法は遺伝子の肝臓特異的な発現を予測できるだけである。より広範囲の組織について、特異的な発現を予測する手法の開発が望まれる。

文献

- 1) Borodovsky, M., Rudd, K.E. and Koonin, E.V. : Intrinsic and extrinsic approaches for detecting genes in a bacterial genome : *Nucleic Acids Res.*, Vol.22, pp. 4756 – 4767 (1994).
- 2) School of Biology, Georgia Institute of Technology : GeneMark : Gene Recognition Software, <http://genemark.biology.gatech.edu/GeneMark/> (1998).
- 3) Hayes, S.W. and Borodovsky, M. : Deriving Ribosomal Binding Site (RBS) statistical models from unannotated DNA sequences and the use of the RBS model for N – terminal prediction, *PSB98, World Scientific*, Vol.3, pp. 279 – 290 (1998).
- 4) Hayes, S.W. and Borodovsky, M. : How to interpret an anonymous

- genome? : Learning Markov models for gene identification in parallel with the sequence annotation, *Genome Res.* (In submission).
- 5) Yada, T. and Hirosawa, M. : Gene Recognition in Cyanobacterium Genomic Sequence Data Using the Hidden Markov Model, *ISMB96*, Vol. 4, Menlo Park, CA, AAAI Press, pp. 252 – 260 (1996).
 - 6) 矢田哲士：隠れマルコフモデルによる原核生物の遺伝子同定, *蛋白質・核酸・酵素*, Vol.42, pp. 2993 – 3000 (1997).
 - 7) Asai, K., Itoh, K., Y., U. and Yada, T. : Recongition of Human Genes by Stochastic Parsing, *PSB98, World Scientific*, Vol.3, pp. 228 – 239 (1998).
 - 8) Salzberg, S., Delcher, A.L., Kasif, S. and White, O. : Microbial gene identification using interpolated Markov models, *Nucleic Acids Res.*, Vol.26, pp. 544 – 548 (1998).
 - 9) Burge, C. and Karlin, S. : Prediction of Complete Gene Structures in Human Genomic DNA, *J. Mol. Biol.*, Vol.268, pp. 78 – 94 (1997).
 - 10) Henderson, J., Salzberg, S. and Fasman, K.M. : Finding Genes in DNA with a Hidden Markov Model, *J. Comput. Biol.*, Vol.4, pp. 127 – 141 (1997).
 - 11) Krogh, A. : Two method for improving performance of an HMM and their application for gene finding, *ISMB97, Menlo Park, CA, AAAI Press*, Vol.5, pp. 179 – 186 (1997).
 - 12) Chen, Q.K., Hertz, G.Z. and Stormo, G.D. : PromFD 1.0 : a computer program that predicts eukaryotic pol II promoters using strings and IMD matrices, *Comput. Appl. Biosci.*, Vol.12, pp. 71 – 80 (1996).
 - 13) Tsunoda, T. and Takagi, T. : Automatic Extraction of Position Specific Cooccurrence of Transcription Factor Bindings on Promoters, *PSB98, World Scientific*, Vol.3, pp. 252 – 263 (1998).
 - 14) Salzberg, S., Chen, X., Henderson, J. and Fasman, K. : Finding genes in DNA using decision trees and dynamic programming, *ISMB96, Menlo Park, CA, AAAI Press*, Vol.4, pp. 201 – 210 (1996).
 - 15) Salzberg, S. : A method for identifying splice sites and translational

- start sites in eukaryotic mRNA, *Comput. Appl. Biosci.*, Vol.13, pp. 365 – 376 (1997).
- 16) Kulp, D., Haussler, D., Reese, M.G. and Eeckman, F.H. : A generalized hidden Markov model for the recognition of human genes in DNA, *ISMB96*, Menlo Park, CA, AAAI Press, Vol.4, pp. 134 – 142 (1996).
 - 17) Reese, M.G., Eeckman, F.H., Kulp, D. and Haussler, D. : Improved Splice Site Detection in Genie, *J. Comput. Biol.*, Vol.4, pp. 311 – 323 (1997).
 - 18) Uberbacher, E.C. and Mural, R.J. : Locating protein – coding region on human DNA sequences by a multiple sensor – neural network approach, *Proc. Natl. Acad. Sci. U.S.A.*, Vol.88, pp. 11261 – 11265 (1991).
 - 19) Xu, Y. and Uberbacher, E.C. : Automated Gene Identification in Large – Scale Genomic Sequences, *J. Comput. Biol.*, Vol.4, pp. 325 – 338 (1997).
 - 20) Solovyev, V.V., Salamov, A.A. and Lawrence, C.B. : Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames, *Nucleic Acids Res.*, Vol.22, pp. 5156 – 5163 (1994).
 - 21) Salamov, A.A. and Solovyev, V.V. : Recognition of 3' – processing sites of human mRNA precursors, *Comput. Appl. Biosci.*, Vol.13, pp. 23 – 28 (1997).
 - 22) Solovyev, V.V. and Salamov, A. : The Gene – Finder computer tools for analysis of human and model organisms genome sequences, ISMB97, Menlo Park, CA, AAAI Press, Vol.5, pp. 294 – 302 (1997).
 - 23) Vignal, L. and Lisacek, F. : A Multi – Agent System for Exon Prediction in Human Sequences, *Genome Informatics Workshop 97*, Vol. 8, pp. 156 – 165 (1997).
 - 24) Shah, I. and Hunter, L. : Predicting Enzyme Function from Sequence : A Systematic Appraisal, *ISMB97*, Menlo Park, CA, AAAI Press, Vol.5, pp. 276 – 283 (1997).
 - 25) Y.Tateno, K.Ikeo, T.Imanishi, H.Watanabe, T.Endo, Y.Yamaguchi, Y.

- Suzuki, K.Takahashi, K.Tunoyama, M.Kawai, Y.Kawanishi, K.Naitou, T. Gojobori : "Evolutionary motif and its biological and structural significance", *J.Mol.Evol.*, 44 (Suppl 1) S38 – S43 (1997).
- 26) Toshinori Endo, Kazuho Ikeo, and Takashi Gojobari : "Large – scale search for genes on which positive selection may operate" *Mol.Biol. Evol.*, 13 (5), pp. 685 – 690 (1996).
- 27) Nakai, K. and Kanehisa, M. : A knowledge base for predicting protein localization sites in eukaryotic cells, *Genomics*, Vol.14, pp. 897 – 911 (1992).
- 28) Horton, P. and Nakai, K. : Better Prediction of Protein Cellular Localization Sites with the k Nearest Neighbors Classifier, *ISMB97, Menlo Park, CA, AAAI Press*, Vol.5, pp. 147 – 152 (1997).
- 29) Fujiwara, Y., Asogawa, M. and Nakai, K. : Prediction of Mitochondrial Targeting Signals Using Hidden Markov Models, *Genome Informatics Workshop 97*, Vol.8, pp. 53 – 60 (1997).
- 30) desJardins, M., Karp, P.D., Krummenacker, M., Lee, T.J. and Ouzounis, C. A. : Prediction of Enzyme Classification from Protein Sequence without the use of Sequence Similarity, *ISMB97, Menlo Park, CA, AAAI Press*, Vol.5, pp. 92 – 99 (1997).
- 31) Bucher, P., Fickett, J.W. and Hatzigeorgiou, A. : Computational analysis of transcriptional regulatory elements : a field in flux, *Comput. Appl. Biosci.*, Vol.12, pp. 361 – 362 (1996).
- 32) Yada, T., Totoki, Y., Ishii, T. and Nakai, K. : Functional prediction of *Bacillus* genes using hidden Markov model, *ISMB97, Menlo Park, CA, AAAI Press*, Vol.5, pp. 354 – 357 (1997).
- 33) Fujibuchi, W. and Kanehisa, M. : Prediction fo Gene Expression Specificity by Promoter Sequence Patterns, *DNA Res.*, Vol.4, pp. 81 – 90 (1997).

5-2 遺伝子発現制御シミュレーション

ゲノムに書かれた遺伝情報は、最終的にタンパク質として発現したり、転写その他のメカニズムを制御したりしてその役割を発揮している。ゲノムの中に隠された遺伝情報を「解読」するためには、単に個々の遺伝子、タンパク質の機能を分析するだけではなく、遺伝情報がどのような形でゲノムの中に記述され、それらが相互にどのような関係で機能しているのか、それら遺伝子機能のネットワークをシステムとして理解する必要がある。

遺伝子の機能を解明する実験では、遺伝子およびその転写産物の特定とそれらのカスケードの決定に重点を置いて行なわれているが、遺伝子群の全体としての機能を解明を目指した、遺伝子（タンパク質）機能の網羅的な解析も行なわれ始めている。遺伝子をシステムティックに一つ一つ破壊して表現型を見たり、様々な条件での遺伝子発現パターンを網羅的に分析する、プロテオーム解析とよばれる手法は、ゲノム配列のデータベース化や、DNAチップなどのハイブリダイゼーション技術の急激な進歩により可能となった。すでに、マイコプラズマ菌はアメリカ、酵母菌はヨーロッパで本格的なプロテオーム解析がスタートしており、近い将来、遺伝子機能や遺伝子発現に関する莫大なデータが蓄積されるものとの予想される。

しかし、細胞全体としての振舞いを把握するために、遺伝子の集合体が全体としてどのように機能するのかを解析するには、数千、数万の遺伝子の相互作用を考慮しなければならず、コンピュータによる細胞のモデル化が必要不可欠である。そのようなモデル化によって、特定の遺伝子をノックアウト/強制発現したり、培地成分を変化させたりした時に、細胞全体がどのような振舞いをするか、シミュレーションによって予測することができれば、生物学者が様々な状況における細胞活動の計算機上 (in silico) での仮想実験が可能となる¹⁾。

東京大学の萩谷昌己らは、ショウジョウバエの体節形成における関係遺伝子の相互作用についてシミュレーションを行なうシステムを開発した²⁾。このシステムでは、実験によって確かめられた遺伝子発現のパターンがシミュレーションによって再現されるような遺伝子間の相互作用のルールを自動的に生成できることが示された。

ソニー CSL 研究所の北野宏明らは、*C. elegans* の発生過程を、細胞同士の力学的均衡や移動を考慮して細胞の3次元的な視覚化を行なうシステムを作成し、細胞間のシグナルと遺伝的カスケードのモデル化を目指している³⁾。また、ショウジョウバエの発生過程における初期胚発生の体節形成、複眼形成、成虫原器の発生に関するモデル化、human fibroblast の老化機構のモデル化等に取り組んでいる。

慶應大学の富田勝らは、細胞内の各遺伝子の機能のコンピュータ上でのモデル化とソフトウェアを開発を行い、全DNA配列が判明しているマイコプラズマ菌でシミュレーションを行なっている⁴⁾。マイコプラズマ菌は、知られている生物のなかで最小のゲノム(0.58Mb)と最小の遺伝子セット(約470)を持ち、システムティックな遺伝子破壊によるプロテオーム解析が現在米国 TIGR 研究所において進行中である。この研究では、細胞の振舞いを予測することを目的とし、シグナル伝達、細胞周期、転写制御、発生分化といったような、細胞代謝の一部だけでなく、全遺伝子のシミュレーションを行なっている点が注目される。ただし、量的なシミュレーションは現状では困難なので、質的なシミュレーションに標的を絞っている。

このようなシミュレーションを大規模に行なうことは、モデル構築の複雑・困難さ、計算時間等の問題があるが、今後さらに進展するものと期待される。対象となる生物としては、日本に大きな研究グループのある大腸菌、96年にかずさ DNA 研究所によって全ゲノム配列が同定され、光合成を行なう細胞として重要なモデル生物であるシアノバクテリア、96年に全ゲノム配列が同定されて以来、DNA チップなどを用いた大規模な遺伝子機能解析がヨーロッパを中心に行なわれ、イネやヒトなどの高等生物のモデリングへの通過点として、重要なモデル真核生物である酵母菌などがあげられる。このような研究によって細胞のシミュレーション技術が確立されると、計算機上の仮想実験が可能になり、すべての薬に対して影響を調べたりすることなど、実験室の実験では時間的に不可能な作業をコンピュータ上で高速にシミュレーションすることが期待できる。

文献

- 1) Galper, A. and Brutlag, D. : "Computational Simulations of Viological Systems," In Smith, D., ed., Biocomputing : Informatics and Genome Projects. San Diego, CA : Academic Press (1993).

- 2) Arita, M., Hagiya, M. and Shiratori, T. : "GEISHA SYSTEM : An Environment for Simulating Protein Interaction", *Genome Informatics* 1994, Vol. 5, pp.81 – 89 (1994).
- 3) Kitano, H. and Imai, S. : "The Virtual Cell Laboratory Predicts The Molecular Mechanism of Cellular Senescence – A Computer – Aided Approach to Biology of Aging," *Molecular Biology of the Cell*, Vol. 7, Supplement 533a (1996).
- 4) Tomita, M. et al. : "E – CELL : Software Environment for Whole Cell Simulation," *Genome Informatics* 1997, Vol. 8, pp.147 – 155 (1997).

5 – 3 データベース・クラスタリング技術

5 – 3 – 1 データベース技術

本項では第2章から第4章までの各章で報告された技術に関するデータベースの現状について報告する。

(1) 遺伝子変異部位に関するデータベース

遺伝子変異部位に関するデータベースとしては、個々の遺伝子座ごとの変異データベースと、これらを統合したものとがある。後者の代表ともいべきものが、ウエールズ大学の Krawczak と Cooper を中心に構築されている、HGMDTM (The Human Gene MutationDatabase ; <http://www.cf.ac.uk/uwcm/mg/hgmd0.html>) である¹³⁾。このデータベースは、元来ヒト遺伝子の変異のメカニズムを研究するために作られたものであるが、今日ではより広範な遺伝子変異の参照用データベースとしての性格をもつに至っている。本データベースには、遺伝病の原因となるヒトの核内遺伝子のタンパク質コード領域内の変異で、明らかな表現型の変化を示すものを中心に登録されている。データは、マニュアルあるいはコンピュータ化された検索手段により、250種類の学術誌から抽出されている。1998年1月現在、681の遺伝子に関する12,380の変異が登録されている。その内訳の一例を表5 – 1に示す。また、本データベースから個々の研究機関で構築されている個別遺伝子座毎の変異データベースへのリンクが張られており、遺伝子座の重複を含めて60種類程度のデータベースへのアクセスが可能である(表5 – 2参照)。これらを参照することにより、個々の遺伝子毎のより詳細な変異情報

を得ることが可能となる。

表5-1 HGMD™に登録されている遺伝子変異の分類

Mutation type	No. of entries
<u>Micro-lesions -</u>	
Missense/no	7882
Splicing	1186
Regulatory	106
Small deletions	2053
Small insertions	723
Small indels	95
<u>Gross lesions -</u>	
Repeat expansions	15
Gross insertions & duplications	71
Complex rearrangements (including inversions)	83
<u>Gross deletions</u>	166
Total	12380

表5-2 個別遺伝子座ごとの遺伝子変異データベース

Gene(s)	Database	Institution
ALDOB	Hereditary Fructose Intolerant	Boston University, USA
APC	Adenomatous polyposis coli	Mayo Clinic, USA
AR	Androgen receptor	McGill University (Quebec), Canada
ATM	Ataxia-Telangiectasia	Virginia Mason Research Center (Seattle), USA
ATM	Ataxia-Telangiectasia	A-T Project (Florida), USA
BRCA1	Breast cancer	National Human Genome Research Inst., USA
BTK	X-linked agammaglobulinaemia	University of Helsinki, Finland
CDKN2A	Cyclin-dependent kinase inhibitor 2A	University of Oslo, Norway
CD40LG	CD40L defect database	University of Geneva, Switzerland
CFTR	Cystic fibrosis	Toronto Hospital for Sick Children, Canada
COL1A1, COL1A2, COL3A1	Collagen (type I)	Leeds University, UK
CYBB	X-linked chronic granulomatous disease	University of Helsinki, Finland
DMD	Duchenne/B\u00e4cker muscular dystrophy	Leiden University, Holland
EMD	Emery-Dreifuss muscular dystrophy	Cambridge University, UK
FAC\A, FACC	Fanconi anaemia mutations	Rockefeller University (New York), USA
FBN1	Marfan database	H\u00fcpital Necker-Enfants Malades (Paris), France
F7	Factor VII	Hammersmith Hospital (London), UK
F8C	Haemophilia A	Hammersmith Hospital (London), UK
F9	Haemophilia B	Guy's Hospital (London), UK
G6PD	Favism research papers	Scripps Research Institute (California), USA
GAA	Acid alpha-glucosidase	Erasmus University (Rotterdam), Holland
GCH1	GTP cyclohydrolase I deficiency	University Children's Hospital (Zurich), Switzerland
GRL	Glucocorticoid receptor resource db	Georgetown University (Washington), USA
HEXA	Hexosaminidase A	McGill University (Quebec), Canada
HPRT1	Lesch-Nyhan syndrome	University of North Carolina, USA
IL2RG	IL2RGbase	National Human Genome Research Inst., USA
L1CAM	L1 cell adhesion molecule	Antwerp University, Belgium
LDLR	Familial hypercholesterolaemia	University College London, UK
LDLR	Familial hypercholesterolaemia	H\u00fcpital Necker-Enfants Malades (Paris), France
MLH1	International Group on Hereditary Non-Polyposis Colorectal Cancer	Amsterdam, Holland
MSH2	International Group on Hereditary Non-Polyposis Colorectal Cancer	Amsterdam, Holland
NF1	Neurofibromatosis type I consortium	NF1 Consortium, USA
OTC	Ornithine transcarbamylase deficiency	University of Minnesota, USA
PAH	Phenylketonuria	McGill University (Quebec), Canada
PAX6	PAX6	MRC Unit (Edinburgh), UK
PCBD	Pterin-4-a-carbinolamine dehydratase deficiency	University Children's Hospital (Zurich), Switzerland
PKD1	Polycystic kidney research foundation	Genome Database, USA
PTS	6-Pyruvoyl-tetrahydropterin synthase deficiency	University Children's Hospital (Zurich), Switzerland
QDPR	dihydropteridine reductase deficiency	University Children's Hospital (Zurich), Switzerland
RDS	Retinal degeneration, slow database	University of Iowa, USA
RHO	Rhodopsin	University of Iowa, USA
TP53	p53	Mayo Clinic, USA
TP53	p53	Human Genome Center (Tokyo), Japan
TP53	p53	Institut Curie (Paris), France
TSC2	Tuberous Sclerosis mutation database	University of Wales College of Medicine (Cardiff), UK
VWF	Von Willebrand disease	University of Michigan, USA
VHL	Von Hippel-Lindau disease	H\u00fcpital Necker-Enfants Malades (Paris), France
WT1	Wilms Tumour database	H\u00fcpital Necker-Enfants Malades (Paris), France
	Globin gene server	Pennsylvania University, USA
	M6P/IGF2R	Duke University (North Carolina), USA
	Mitochondrial	Emory University (Atlanta), USA
	Skin disease mutation Database	University of North Carolina, USA
	G-Protein coupled receptors	Tromso, Norway
	G-Protein coupled receptors	National Institute of Health (Bethesda), USA
	G-Protein coupled receptors	EMBL (Heidelberg), Germany
	Mutation spectra database for bacterial and mammalian genes	Yale University (Connecticut), USA
	Mucopolysaccharidoses disorders	University of Minnesota, USA
	Ion channel mutations	Iowa University, USA
	Thalassemia database	Prince of Songkla University, Thailand
	Chaperonin (GroES, GroEL)	University of Texas (San Antonio), USA

また、遺伝子座ごとの変異データベースとその統合化に関する HUGO の取り組みを知るには、HUGO Mutation Database Website (http://ariel.ucs.unimelb.edu.au:80/~cotton/mut_database.htm) が便利である。現状は、個別遺伝子ごとに各研究室単位で膨大な情報を集積しているという段階で、データベースの中身も文字データを表形式にまとめるに留まっている。今後、増加するデータへの対応と個別データベースの統合化が必要ではあることは言うまでもないが、さらに膨大なデータを使った解析を支援するためのツールの開発も重要である。

(2) 特異的発現様式解析に関するデータベース

ヒト EST の大規模解析による、mRNA 発現頻度に関するデータベースとしては、Okubo と Matsubara による BodyMap (<http://www.imcb.osaka-u.ac.jp/bodymap/>)²⁾ と TIGR による EGAD (The Expressed Gene Anatomy Database; <http://www.tigr.org/tdb/egad/egad.html>) が代表的である。BodyMap では、mRNA の 3' 端のポリ A テールから制限酵素 *Mbo*I の切断部位までの塩基配列をマーカとして、mRNA のカタログ化を行い、組織あるいはセルタイプ毎の 1000 クローン程度のシーケンシングにより発現頻度の解析を行っている。現在、Web サイトには 39 種類の組織を対象に得られた、トータル 42,122 クローンの配列解析結果と、10,896 のユニーク配列を元に解析された発現頻度情報が登録されている。図 5-2 にデータの一部を示す。ここでは、アクチンに関連した 7 種類の遺伝子の各組織における発現の様子を表形式で見ることができる。

actin

		blood cells										liver		connective tissue							
GS num	Total	#lib	mm	pm	mp	gr	fr	et	ti	a	f1	'lf	al	ae	fc	ob	bx	ap	vf	it	an
1507	7	6	.	1	.	1	2	.	1	1
3142	15	7	2	2	.	1	.	7	.
244	68	23	.	11	1	4	.	.	3	2	.	.	.	1	8	2	1	.	1	2	.
114	62	23	1	3	3	1	1	.	.	1	8	3	4	.	1	.	5
1688	13	8	.	.	1	3	1	.
1178Q	1	1
3304	1	1	1

図 5-2 BodyMap における特定遺伝子の発現パターンの例

EGADでは、GenBankから抽出した、冗長性のないヒトの転写配列をカタログ化している。各配列には遺伝子との対応関係がつけられており、そのそれぞれについて対応するESTデータとそれが由来するライブラリーの情報が添付されている。

また、ヒト以外のモデル生物でも、その全ゲノム配列が解析された酵母を対象に、DNAチップを用いたmRNAの発現プロファイル解析が報告されているが³⁾、その実験データがデータベース化され、インターネットでアクセス可能となっている (<http://cmgm.stanford.edu/pbrown/explore/index.html>)。

今後、メッセージディスプレイ法やDNAチップを用いた、mRNA発現プロファイルデータの集積は、ヒトおよびモデル生物で急速に進むものと予測される。現在、多くの実験手法が開発され、その有効性が試されている状況から見て、将来的には、異なる手法で得られた実験データを統一的に扱えるデータベース構造の検討が必要である。また、莫大な数の発現プロファイルデータから、相互に関連しあう遺伝子を抽出し、その相互作用を効率的かつ精度よく解析できる情報科学的手法の開発も重要である。

(3) 発現調節ネットワークに関するデータベース

a 発現調節因子に関するデータベース

転写制御因子に関するデータベースについては昨年度の報告書の3—2節にも記述したので、ここでは各データベースの概略についてのみ報告する。現在インターネットを通じてアクセス可能な公共データベースには、真核生物遺伝子のプロモータ領域、種々の転写因子、ならびにDNAのシグナル配列に関する情報が蓄積されている。

Eukaryotic Promotor Database (EPD)⁴⁾ は Weizmann Institute of Science in Rehovot (イスラエル) で設計、開発され、現在、ISREC (スイス) で一つのファイル (303kB, Release 48, 1996年10月) として管理されている (<ftp://expasy.hcuge.ch/databases/epd>)。EPDはEMBLデータライブラリの特殊な注釈データベースで、EMBLから入手可能な真核生物プロモータに関する情報を提供する。EPDには高等真核生物RNAポリメラーゼIIにより認識される（実験的に決定された）プロモータが蓄積されている。全ての情報は直接に科学的文献から抽出されている。

Transcription Factor Database (TFD)⁵⁾ は、多くの真核生物転写に関する配列情報を蓄積し、それら情報間の関連づけにより構成されている。David Ghosh (ghoshd@rockvax.rockefeller.edu) によって管理されている。遺伝子発現の異なる段階における

る転写因子の機能と相互作用に関する異なる型の情報が、1件対1件あるいは1件対複数件の対応で関連づけられている。データベースは、真核生物転写機構に関する情報の主たるクラスを全て含むように構成され、情報全体は相互に関連するテーブルに分けられている。

1993年から Mizushima らはTFDを元に転写因子データベース (TFDB) を開発し始めた。現在、文献データベース MEDLINE を検索することにより更新されている⁶⁾。

TRANSFACは、真核生物のcis – acting 制御DNA エレメントとtrans – acting 因子に関するデータベースであり、酵母からヒトまでの全領域をカバーする。TRANSFACは1988年から印刷物⁷⁾としてスタートし、1990年に電子化された (<http://transfacv.gbfraun-schweig.de>)。TRANSFAC データは主として原論文から抽出されているが、場合により他の適切な文献からも抽出される。

転写制御因子に係わるデータベースはまだ発展段階にある。今後、mRNA 発現頻度解析、ツーハイブリッド法などの実験データの集積とその解析が急速に進むものと期待できる。それらデータの蓄積、整理、構造化及び統合化が今後の課題である。

b 遺伝子ネットワークに関するデータベース

ゲノム解析からポストゲノム解析に向かう中で、遺伝子間の相互作用を実験的に解析していくと同時に遺伝子間のネットワークをシステムティックに解析していくための情報科学的取り組みが重要になってきている。

遺伝子間のネットワークに関するデータベースとしては、京大化研で開発中のKEGG (Kyoto Encyclopedia of Genes and Genomes ; <http://www.genome.ad.jp/kegg/>)、SRI の EcoCyc (Encyclopedia of E. Coli Genes and Metabolism ; <http://www.ai.sri.com/ecocyc/ecocyc.html>)、九大のSPAD (The Signaling Pathway Database ; <http://kintaro.grt.kyushu-u.ac.jp/eny-doc/spad.html>) などが代表的なものである。前二者が主に代謝系を対象としたもので、最後の SPAD が増殖因子やサイトカイン関連のシグナル変換経路を対象にしたものである。以下、KEGG を例に説明する。

KEGG プロジェクトのターゲットは、第一に分子細胞生物学の広範な知識を分子のパスウェイ情報としてコンピュータ化して、生命系の機能カタログを構築すること。第二にあらゆる生物種で、ゲノム解析がもたらす遺伝子カタログの各部品（遺伝子産物）を

機能力タログ上に対応づけること。第三に相互作用データから可能なパスウェイを計算したり、ポストゲノム解析に伴う新しい情報処理技術を開発すること、である。

KEGGの具体例を図5-3に示す。本図はリジンの合成系に関する生化学の知識をコンピュータ化したものである。図中の四角で示されたものは、遺伝子産物である酵素を示し、中に対応するEC番号が記載されている。図中の四角のうち背景がグレーのものは、大腸菌のゲノムから見い出された遺伝子に対応している。パスがきちんとつながって合成系ができているかをみるとことにより、各遺伝子の機能予測の適切さを判断できる。また、この例のようにパスが切れていれば、ゲノムにある遺伝子を見落としているか、あるいは別の酵素による反応経路が存在する可能性がでてくる。KEGGではそれらの可能性を計算により調べることができるようになっている。

現在、KEGGで公開されているのは代謝系が中心であるが、今後種々の制御系についても公開されていく予定になっている。遺伝子間の相互作用データのコンピュータ化、知識ベース化とそこから有用な情報を引き出すための解析ツールの開発は、ポストゲノム解析時代の大きな課題の一つと考える。

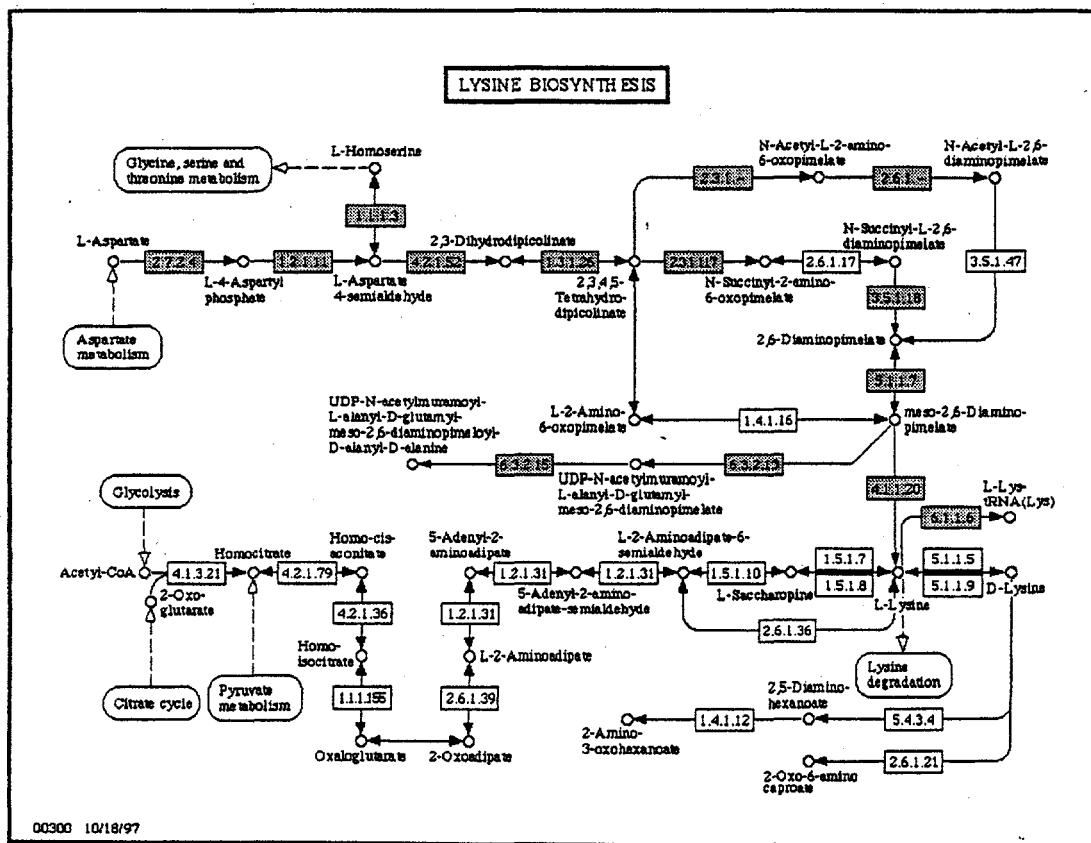


図5-3 KEGGにおける代謝マップの例

5-3-2 クラスタリング技術

前項で報告した種々のデータベースや、より一般的なDNAあるいはタンパク質の配列データベース、立体構造データベースなど、現在数多くのゲノム関係のデータベースが構築されており、その中には莫大な量の貴重な情報が蓄積されている。それらデータの種類と量は、今後とも急速に拡大していくものと予測される。ただし、現状ではデータベースの登録データに冗長性があったり、個別データ間の関係について十分な解析がなされているわけではない。今後、蓄積したデータを活用して有用な情報を引き出して行くためには、各種データベースの個々のデータ間の比較を何らかの指標をベースに行い、データ間の関係を整理して、目に見える形にしていく必要がある。そのための基本となるのがクラスタリング技術である。以下、配列情報をベースにしたクラスタリングの代表例について報告する。

(1) 同一遺伝子由来のシーケンスベースのクラスタリング

現在、GenBankを始めとするDNA配列の公共データベースには、ヒト遺伝子由来の配列データが大量に集積されている。特にEST配列は、エントリー数でデータベース全体の70%、塩基数で同じく40%を占めている。これらの中には同一遺伝子に由来する複数の配列が別々に登録されているケースも多い。前項(2)で説明した、BodyMapやEGADも同一遺伝子由来の配列をクラスタリングすることにより、各種組織におけるmRNAの発現頻度に関する情報を整理している。NCBI(National Center for Biotechnology Information)のUniGene(<http://www.ncbi.nlm.nih.gov/UniGene/index.html>)では、最新のGenBank及びdbESTに登録されているヒトのmRNA配列あるいは遺伝子コード配列とEST配列をクラスタリングして、ユニークな遺伝子のセットを提供している。1998年1月現在、603,307配列から39,472のクラスタを得ている。UniGeneに対しては、キーワード、染色体、ライブラリなどによる検索が可能となっている。

(2) ホモジーベースのクラスタリング

微生物ゲノムプロジェクトの進展により、真正細菌、古細菌、真核生物の幅広い種に渡って、その全ゲノム配列決定されている。これらのゲノム配列を相互に比較することは、各生物の全体像を遺伝情報の観点から再構築する際の助けになるだけでなく、幅広い生命現象の根底にある普遍性と多様性の原理に迫ることにもつながるものと考えられ

る。その際、ホモロジーによるクラスタリングが基本技術となる。

東大医科研ヒトゲノム解析センターでは、全配列が決定されたゲノムを相互に比較するためのデータベース・ワークベンチシステムMBGD (MicroBial Genome Database for comparative analysis) が構築され⁸⁾、インターネット上で公開されている (<http://mbgd.genome.ad.jp>)。MBGDには各ゲノム中に同定された全ORFに関する情報（推定された機能、マップ位置、配列中に含まれるモチーフなど）とそれらをホモロジー検索により相互に比較した結果とが蓄積されており、これらのデータを駆使してゲノムの比較解析を進めるための様々な機能が実装されている。

MBGDの機能の中心は指定された複数の生物種について、与えられたホモロジー、オーソロジーの基準に基づいてクラスタリングを行い、生物種ごとに対応遺伝子をまとめた表を自動作成する機能にある。オーソロジーの対応づけはゲノムの比較を行う上で必須の作業であるが、1対1の対応づけが常に行えるわけではなく、一般に遠縁の生物種間になるほど対応づけが困難または不可能なケースが増加するため、定義によって変化する要素を持っているともいえる。MBGDではカットオフパラメータ等の指定や生物種の限定を行うことによって、いくつかの異なる立場から対応表を作成することができるようになっている。

MBGDには、こうして得られたクラスタ表を基にしてゲノムの比較解析を進めるために、以下のような機能がある。得られたクラスタに対して、どのゲノムに含まれ、どのゲノムに含まれないかといった観点や、クラスタを構成する遺伝子の機能カテゴリの観点などから分類して整理したクラスタマップの作成機能。クラスタ表に対するキーワード検索などの検索機能。各クラスタを構成する遺伝子集合について、ゲノム中の周辺遺伝子の並びを比較する機能、およびゲノム全体における位置の分布を比較する機能。各クラスタを構成するタンパク質配列のマルチプルアライメント、およびそれを基にした系統樹の作成機能。ユーザの配列群に対してホモロジー検索を行い、MBGDが作成するクラスタ表と重ね合せる機能等である。

本項で述べた例のほかに、古くはタンパク質配列データベースPIRのスーパーファミリー分類や最近のSCOP⁹⁾などによるタンパク質立体構造の分類もクラスタリングの代表例である。また、モチーフや種々のシグナル配列はクラスタリングを行う際の有効な指標である。今後、蓄積された多様かつ莫大なデータから、ヒトを初めとする各種生物のゲノムの基本原理を理解して、有用な産業応用に結び付けていくためには、配列、構

造、その他様々な指標から各種データのクラスタリングを行い、それを総合的に整理して、提示していく技術の開発が重要と考えられる。

文 献

- 1) Krawczak, M. and Cooper, D. N. : *Trends Genet.*, 13, 121 – 122 (1997).
- 2) Okubo, K., Hori, N., Matoba, R., Niizuma, T., Fukushima, A., Kojima, Y. and Matsubara, K. : *Nature Genet.*, 2, 173 – 179 (1992).
- 3) DeRisi, J. L., Iyer, V. R. and Brown, P. O. : *Science*, 278, 680 – 686 (1997).
- 4) Buchner, P. and Trifonov, E. N. : *Nucleic Acids Res.*, 14, 10009 – 10026 (1986).
- 5) Ghosh, D. : *Nucleic Acids Res.*, 18, 1749 – 1756 (1990).
- 6) Ghosh, D. : *Nucleic Acids Res.*, 20, 2091 – 2093 (1992).
- 7) Ozaki, T. et al. : Genome Informatics Workshop 1996, 218 – 219 (1996).
- 8) 内山郁夫、高木利久：微生物ゲノムワークショップ、1997年12月（東京）。
- 9) Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. : *J. Mol. Biol.*, 247, 536 – 540 (1995).

第6章 国際動向

6-1 國際シンポジウム報告

「バイオインダストリーに新しい世界を開くゲノム研究

—ゲノムインフォマティクス技術の新展開—」

10月20日、笹川記念会館において標記シンポジウムが開かれた。これは産業科学技術研究開発制度の先導研究「ゲノム情報解読利用技術」の一環として行われたものである。参加者は予定を上回る200人以上とこの分野の関心の高さを示した。海外講師3人を含む計9人の講師（同時通訳付）により最新のテーマについて熱氣あるシンポジウムとなった。角田周一部長（NEDO）、地崎修専務理事（JBA）の挨拶に続いてコーディネータ榎佳之教授（東大医科学研究所）より挨拶があり、シンポジウムに入った。以下に本シンポジウムの内容の要旨をまとめた。

1. 微生物ゲノム解析の最前線

小笠直樹毅（奈良先端大・バイオ）

現在、微生物について遺伝子の塩基配列が世界的に爆発的勢いで決定されつつある。発現ネットワーク解析、病原微生物の感染関与遺伝子を見つける等新しい研究が始まっている。全遺伝子の決定後研究はどのような方向に行くべきか等問題が多い。個々の遺伝子の解析の積み重ねで全体像に迫れるのではないかと考えている。このような観点に基づき演者の仕事を紹介しつつ、インフォマティクス等新しい方法論が必要になっている背景について議論された。

微生物では蛋白質をコードする領域の特定は割合容易であり、細菌では1kbに1個、酵母では2kbに1個の割合で存在する。機能については約1/2が一応分かっている。ヘモフィリス・インフルエンザの糖の利用性は8種の糖を利用できることが遺伝子から分かった。これからは制御系ネットワークの解析が必要である。枯草菌の全配列は日本が1/3 (1.3Mb) を分担した。この中に約1300個の遺伝子がありその中1/4の機能が分かった。未知の3/4について機能を調べる実験を始めており1000個のノックアウトを目標に行っており500個が終了した（1997.10現在）。細胞に必要な遺伝子については発現を人為的に制御して機能解析をしようとしている。枯草菌の複製には200個の遺伝子があり140個が未解析、120個は変異株、発現調節変異株をつくった。まだ始めたばかりで25%のphenotypeが分かったところである。期待する方法としてはRNAレベルでもっと直接的

に発現が見られれば良いと思っている。すべての遺伝子は4000個で全RNA分析からどの遺伝子がどんな条件で発現しているかが調べられれば迅速に次のステップの解析が行えるであろう。問題は細菌のmRNAの精製が難しく、リボソームRNA、安定RNAのノイズを除いて信頼性のあるデータを得ることである。これは今後の新しい解析技術として開発してみたいものでありプロジェクトを相補する有力な手段となるであろう。

2. "DNA配列決定" 高速化への試み—スタッキングプローブを用いたハイブリダイゼーション法— 湯川英明（三菱化学・筑波研）

SBH (Sequencing by hybridization) の開発の現状、可能性について述べた。SBHはオリゴヌクレオチドプローブを標的DNAにハイブリダイズすることにより迅速に配列を決めようとするものでメガベースの配列決定に大きな威力を発揮すると考えられる。8塩基のプローブを用いると約65000種、10塩基プローブで約100万個のプローブからなるチップが必要になる。現在半導体チップ製造の技術を用いて100万個レベルのプローブをつくることができる。10塩基プローブでは原理的には1kbの配列が瞬時に分かる。

応用微生物の立場からみると1個の遺伝子(1~1.5kb)の配列が完全に正確でなくてもよいから迅速に決定したい要望があるのでこれをプロジェクトの目標にしている。問題点はミスマッチでありミスマッチの種類は大きく分けて1) 中央部がハイブリダイズしない2) 末端がハイブリダイズしないがありこの差が性能に大きく影響する。完全マッチを100とすると末端ミスマッチは25、中間ミスマッチは1となる。25位であると完全マッチと区別できないのでミスマッチの程度は1~2に押さえる必要がある。演者らはこの解決法として末端ミスマッチを内側に移し、末端にスタッキング効果のあるユニバーサル塩基(イノシン、5-ニトロインドールまたは3-ニトロピロール)を4個付けて識別能を高めた本法は前処理が少なく自動化が可能で瞬時に長い配列を決定できる長所をもっており、実用化に向けた研究が進行している。

3. マッピング、配列決定のためのDNAファイバーマッピング技術

H-U.Weier (ローレンスバークレー国立研)

米国エネルギー省からの基金により研究している。大規模DNA配列決定研究のために高解像度物理地図の構築と minimal tiling path の決定が必要である。ショットガン配列決定プロジェクトの完成にはギャップの物理的位置、サイズの情報が必要となる。演

者は個々の伸長させたDNA分子に対する蛍光標識プローブを用いたハイブリダイゼーションによるマッピング法として高分解能のある光学的手法を開発した。本法によれば、あるDNAクローンと他のクローンの位置関係が可視できる方法でありマップ構築を容易にするものである。その原理は、DNA分子を単離し分子の一端を固相基板（マイカ、スライドグラス）に結合しDNA分子を引き伸ばし（カバーガラスを引き上げることによって水の半月状界面に引き伸ばされる）DNAファイバーを形成させる。これを蛍光標識プローブでハイブリダイズし、結合したイメージをCCDカメラを装備した蛍光顕微鏡で記録するものである。この方法をQDFM (Quantitative DNA fiber mapping) と名付けた。QDFMはマップアセンブリのみでなく minimal tiling path の確立にも重要な機能を果たす。現在ヒトゲノムクローンは再構築されたもので真のヒトゲノムを表していないと考えられ、QDFMによりゲノムの欠失部位が容易に検出できる。QDFMの高感度、高精度を利用して高分子量DNAから調製したDNAファイバー上に個々のエキソンを迅速にマップすることができるし、cDNAマッピングによって種間のゲノムを比較も容易にできる。マッピング距離は数百塩基で500kb位の分離ができる。

4. ゲル電気泳動を用いた高速・大容量DNA分析技術

神原秀記（日立製作所・中研）

高速、高スループットDNA分析法の開発についての現状、将来について述べられた。この10年間でDNA分析技術は、ゲノムプロジェクトのスタート、蛍光検出技術、シーケンサー機器の進歩に伴って飛躍的に進んだ。対象もマッピングからシーケンス決定へと進展してきた。最初にシーケンサーが開発された1986年に比べて10倍のスループットを持つようになり数年先にはさらに10倍のものが実用化されるであろう。手法としてはスラブゲルからキャピラリアレー、マイクロファブリケーションを用いた電気泳動へと発展し、反応部まで含んだシステムオンチップまである。さらに今後は質量分析、DNAチップ等目的に応じた組み合わせで使う時代になるであろう。高電圧化、発熱防止、高速化を目的にスラブゲルから毛細管に変わり、ゲル作成条件もポリマーゲルからクロスリンクゲルへと発展し歩留まりも100%近くに向かっている。検出装置も数々の改良が重ねられ一つのレーザーですべての泳動路を照射し、CCDカメラ、2次元検出器を組み合わせて4色の蛍光体を正確に検出分析する装置の開発にも成功した。分離時間、分解能を上げる努力もなされ温度(35~40°C)、ゲル濃度、電解強度、泳動路等の検討により短時間で1塩基の違

いを分離できるようになった。例えば200と201塩基は100V/cm、ゲル長3.3cmで約8分間で分離分析できる。700と701個の識別も可能である。96本のキャピラリー装置で500塩基は1hrで解析できる性能をもつ。応用面では制限酵素分解物や発現プロフィル分析に末端塩基の種類によりグループ分けする方法もある。

これから必要な分析システムは発現情報を分析する装置と必要に応じてDNA断片をとってくる装置が必要になってくる。電気泳動装置と試料調製装置が一体化し、データベースも活用したトータルシステムの時代になるであろう。

5. 遺伝子発現及び相互作用のゲノムワイドスクリーニングを目指して

伊藤隆司（東大・医科研）

ゲノム配列、解析が進んで新しい遺伝子が数多く見出されている。これらの機能を明らかにすることが次の10年間のゲノム研究の大きなターゲットである。クローナ化された遺伝子はまずホモジサーチを行って似ている遺伝子の裏にかくれた機能から新しく見出された遺伝子の機能を探ろうとする。しかしこれでも理解できないものはどうするか。これについて演者は未知遺伝子を他の遺伝子との関係性の中に位置づけて理解しようとしている。その方法として遺伝子と遺伝子の相互関係及び蛋白質と蛋白質との相互関係を調べる方法がある。対象として酵母の多剤耐性とPDRネットワークを取り上げフルオレンスディファレシャルディスプレー(FDD)法により解析しPDR5を見出した。染色体左腕231番目のORFの解析によりPDR1の標的として薬剤耐性関与の機能を見出した。FDD解析による予想外の発見として転写因子の活性と逆の相関を示すtRNAをコードする遺伝子も見つかった。このような挙動を示す断片を10個程見つけて解析している。単純な破壊だけでなく過剰生産から見ることも大切であるのでプロモータの置き換えによるアプローチも行っている。このときすべての遺伝子をヒットできるようなプライマーのデザインをインフォマティク技術でできないかと考えている。インフォマティクでデザインしたプライマーを用いてDDを行い、言わばin silicoのDDと実際のパターンを比較することによってシーケンスしないで断片の同定が可能になる。

次に蛋白と蛋白の相互反応であるが酵母のツーハイブリッドシステムを用いてお互いにinteractするペアを沢山集めたいと考えている。これにはdirectアプローチにより酵母の6000個余りの全遺伝子のマトリクスをつくり総当たりを行えばよいのでこの面からも進めている。モデルとして28個のSH3ドメインについてクローナ化し、DNA結合ドメイ

ンの融合プラスミド中に入れプローブを作つてスクリーニングを行つてゐる。pgal69株を用いレポータを勝手に活性化しないものだけを解析に用いる。SH3でキチンと調べられてゐるのは6個位しかないので、その他のものの機能を調べる。希望として蓄積したデータの中から重要な生物学的知識を抜き出したり interact データからモチーフが分かれば良いと思う。AがBに結合し、BがCに結合するというような単純なバイナリ関係を積み重ねてもっと複雑な関係も見えてくるであろう。生物学的解釈には他の生物からの知識の動員も重要であるが、これは個人の力では不可能である。ある程度自動的にそこまでできるシステムがあればよいと思う。細胞内の分子的pathway、制御ネットワーク、機能未知の遺伝子のネットワーク、代謝 pathway 上の位置づけ等により生物学が進み生物学者を新しい分野に誘う契機になればゲノム研究の新しい展開になるであろう。

6. 高速DNAシーケンス解析技術について

D.Barker (モレキュラーダイナミクス)

インフィマティクスに新しい分析システムが要求されておりモレキュラーダイナミクス社のこの分野の機器開発状況について紹介された。開発研究が最終段階に入りまもなく上市される予定の「MegaBACE1000」は96本のキャピラリ電気泳動を同時に動かし2つのレーザー(488、532nm)で励起しサンプルローディングを自動的に行うことで600塩基のシーケンスが可能である。キャピラリ壁のコーティング技術の改良により100回の繰り返し使用が可能であり、交換は5分間でできる。検出システムは共焦点対物レンズがキャピラリ上をスキャンし蛍光を4個のチャンネルに分けて検出する仕組みになっている。1回の運転は15分間で済む。サンプル～サンプル間の所用時間は100分位である。分離マトリクスはヒドロキシエチルセルロースをベースとし6M尿素を使用している。ワントン大でテストされており好結果を得ている。蛍光剤は熱安定性エネルギーransferに基づいたものを開発している。

ソフトウェアーアーキテクチャー：オープンでransfer可能なデータとしWindows NTのプラットフォームを使って開発した。すべてのアレイシステムは内部のメガソフトアクセスマネージャーにリンクされている。DNAシーケンスをオペレータが直接データ入力できる領域を設けている。

DNAマイクロアレイ：マクラーレン法ということで現在開発初期段階である。このアレイは何千もの遺伝子発現を並行して分析できる。システム構成は1nl以下の試料を

ガラス板に添加し、アレイスキャナーは2色の検出で蛍光を用いている。画像解析、データベース解析できるものもある。次世代のものは1万スポット／slideを目指している。大量のデータを収集しつつありかなりのソフトが必要である。インタラクティブソフトウェアを目指している。

今後のインフォマティクス環境は遺伝子発現のデータベース、バンクのようなもの遺伝子クローニング、マッピング、発現、変異体、クローニング品質管理等をカバーするものとなろう。マイクロアレイからのデータを解析し発現データベースに入れる。インフォマティクスシステムによってこの間の通信を計る。外部のデータベース、公共データベース、内部データベースから取り入れることが必須になる。本シンポジウムの課題は医薬品ターゲット開発、治療用新薬開発などの分野に参考になる示唆に富んだもので有効であった。

7. 転写制御ネットワーク解析技術の開発

町田雅之（生命研）

転写ネットワークは何千もの遺伝子を制御するもので細胞にとって重要である。共通の制御因子が関与するネットワークの詳細が分からなければ遺伝子機能も分からぬであろう。全ての遺伝子機能を塩基配列だけから推定することは困難であり、全体の半数以上の推定不可能な遺伝子機能を明らかにすることが重要であるという観点から演者らが開発中の解析技術について解説された。進行中の技術は赤外蛍光を利用したDNAと蛋白質の相互作用解析とファージディスプレーを利用した転写制御因子の解析である。赤外蛍光DNAシーケンサーを利用して3fmol以下のRIに匹敵する感度で真核生物の転写制御領域を一度に解析できる技術を開発した。ロングレンジフットプリントリングでは900bpが一度にスキャンでき、これは世界最高レベルである。ファージディスプレーでは高分子量でサブユニット構造をもった蛋白質の提示が可能な入系のシステムを用い完全長のcDNAを導入したライブラリーの構築と転写制御因子の解析を実施中である。これは新しい方法で、酵母と*A. oryzae*について行っているがすべての生物の転写制御因子の解析に応用可能である。*A. oryzae*では解糖系遺伝子、プロモータ領域のクローニングを行い、バイオインフォマティクスの技術を用いてモチーフのサーチ等をやっているが今後mRNA発現プロフィール解析にインフォマティックs技術の力をかりて多くの情報をを集め転写ネットワークのモデル化シミュレーションが可能となるであろう。

8. 確率モデルによるゲノムの構文解析

浅井 漢（電総研）

隠れマルコフモデル（HMM）を使った遺伝子領域発見システムについて話された。このシステムは確率モデルに基づいて、原核、真核生物のDNA配列から遺伝子領域発見の有効な手段になっている。ゲノム情報はDNA上に書かれている訳であるが、単に翻訳機能のみでなく蛋白質での機能が重要である。塩基配列の構文解析をするためには、単に転写、翻訳等左側からの情報だけでなく蛋白質に関する機能の情報も得たいと考えている。ゲノム解析には遺伝子の同定のみならず全体像をつかむことが必要である。遺伝子領域発見の定義はコンピュータを使用しての塩基配列からの遺伝子を同定することと定義したい。遺伝子領域発見の鍵には相同性、シグナル発見、コード同定がある。相同性が最も信頼性高く応用も広くかつ簡単であろう。真核生物ではスプライシングがあるのでギャップ位置も明らかにしなければならない。シグナルには開始コドン、終止コドン、ドナー、アクセプターサイト、転写因子、結合部位等がある。これらには何らかの生物学的要素が関わっている。HMMを用いてこれらシグナル検出を行うが、まだ完璧なものではない。必ず擬陽性がある。コードの可能性は蛋白質の塩基配列の特徴を見る訳であるがスプライシング、翻訳との関連性は余りない。フーリエ変換等ではランダムな配列ではある程度識別できる。GC含量も特徴となるがcoding potentialもエラーが多く、多くのものを組み合わせてもコード領域を検出することは困難である。遺伝子領域の発見は単なるシグナル同定、coding potential、相同性によるのではなくこれらの情報を統合することである。コンポーネントモデルはプロモータ、開始、終止コドン等をコンポーネント化し統合することによって評価因子により推定を行う。最も良いparseを確率的に見出していくわけである。通常のホモジーサーチとの違いは左から右のみでなく、グローバルルートがあり、複数の遺伝子あるいは複数のスタート、カットがあってもどのサブシーケンスからもこのモデルから発見できるということである。確率モデルを使うと、1個づつ満足のいくまでどんどん複雑にしていける利点がある。最近の遺伝子領域発見システムはすべて確率モデルであろう。確率によって統合化が可能である。演者らのシステムは他のシステムと競合しており国際レベルに近づきつつある。実際には単なるコード領域遺伝子を予測するだけではなく機能も明らかにしたい。コンピュータによる遺伝子発見は必要であり可能になってきた。これには確率モデルが有望で機能予測することが重要である。

9. ゲノム解析の複雑性—ウイルス及び動く遺伝子—

M.A.McClure (ネバダ大)

アレルギー感染症の NIH 国立センターからの研究資金によって行っている研究を中心として話された。遺伝子配列解析決定を行う際には変化のタイプとして点突然変異、挿入、欠失、モチーフサブセットの獲得、進化、スプライシング、転移、転座、逆位等があり複雑性の原因となっている。欠失にも種々のタイプがあり、数個の塩基変化によっても ORF が影響を受けない場合やフレームシフトを起こして翻訳に支障を起こす場合もある。新しいフレームスイッチによりもとの機能が失われて新しい機能が獲得されることもある。この複雑なゲノム解析のフレームワークは我々のゲノム解析プロセスがより明らかになるにつれてアプローチもフレキシブルに変化し広がる。演者は RNA 研究が専門であり各種レトロウイルスの逆転写酵素 (RT) を隠れマルコフモデルで解析し系統樹を作成した。古細菌に RT はあるが配列は未だ決定されていない。セグメント進化は互いに単一系統発生樹ではない。レトロウイルス間にも見出され MoMLV、MMTV のレトロウイルス系統の祖先は ENN 遺伝子交換を行ったことが演者により見出されている。

結論として、複雑なゲノム解析のフレームワークは遺伝子とゲノムの進化において可能なシナリオを探求する戦略を提供する。in silico で大量のゲノムデータを解析する能力は我々に遺伝子とゲノム構築の実際の進化ネットワークを調べ、記述を可能にする。そのような分析データと解釈はアカデミックな興味を呼び起こすとともにバイオインダストリーにドラッグデザイン戦略に有効な情報を与えるものである。レトロイドエレメントとその宿主の共同進化の複雑性はすべての細胞活性の無差別の排除が宿主に重大な結果を与えることを明らかにするであろう。

最後に本田皓一博士（生命工学工業技術研究所部長）から本先導研究に統いて来年度から開始の予定されている産業科学技術研究開発「ゲノムインフォマティクス技術」について産・学・官の協力をお願いするとの挨拶で閉会、その後レセプションにおいて情報交換がなされた。

6-2 ゲノム関連バイオベンチャーの動向

6-2-1 はじめに

これまでの数年間、大量の塩基配列解析に関する情報蓄積が進み、新規遺伝子の発見が相次いでいる。米国やヨーロッパでは次の段階としてポスト・ゲノム・シーケンスとも称される開発研究に注目が集まっている。すなわち、蓄積されたDNAの塩基配列情報から、如何にして遺伝子の機能解析を行うかと言うことである。世界中がこの課題に如何に関心を持っているかは、ゲノム関連のベンチャー企業に対して各国の大手企業が競って提携を結んでいると言う事実に表れている。例えば、9社のゲノム・ベンチャー企業にこの数年間に支払われた提携料は2,131百万ドルにものぼっているのである。

ゲノム関連ベンチャー企業群は便宜的に第1世代および第2世代に分けられる。

Human Genome Sciences社に始まる第1世代ゲノム・ベンチャー企業がとった戦略は大量塩基配列解析あるいはポジショナル・クローニング法であった。これら第1世代企業のうち3社について設立時から現在までに、どの様な成果を挙げてきたかを最初に概観する。

6-2-2 第1世代ゲノム・ベンチャー企業の動向

(1) Progenitor社

当社は、Developmental Biology、Genomic Technology、Disease Geneticsの3技術を柱として、新規遺伝子の機能を調べてきた。当社はLeptinのレセプター、赤血球のGrowth Factor、Angiogenesis/Osteogenesis遺伝子、Hemochromatosis遺伝子、Epilepsy遺伝子などを発見して権利化している。機能解析ではAffymetrixと提携してDNAチップを用いるアプローチをとっている。

(2) Sequana Therapeutics社

当社は、ポジショナル・クローニングを用いて新規遺伝子の探索を続けてきた。97年5月にはTristan da Cuhna地方に喘息が多く、しかも遺伝子的なバックグラウンドがはっきりしていることに着目し、これら住民の血液を採取して喘息の遺伝子を発見した。その他いくつかの新しい遺伝子を発見し権利化を行っている。また、上海のShanghai First People Hospitalと提携し中国の家系を対象にして疾患に関係する遺伝子の探索を行っている。

(3) Myriad社

当社は、がん、心臓病、中枢神経系疾患などに関する遺伝子に絞って研究を進めてきた。すでに7つの遺伝子を発見している。Schering – Plough, Novartis, Bayer, Eli Lillyと提携しているが、当社自体は診断の方に主力を注いでいる。1997年には、がんにおけるMKK4の役割を発見し、それに続いて8つの主要な遺伝子の領域を同定した。また、タンパク質–タンパク質間の相互作用解析に力を入れている。これまでに発見した主な遺伝子は①BRACA – 1,2 (肺がんと卵巣がん)、②AGT (高血圧)、③MTS1 (メラノーマ) ④MMAC1 (進行性がん) などでそれぞれ特許を取得、あるいは申請している。

このような流れの中で彼らが目指しているところは、診断–モニタリング–治療の流れであるが、将来的には疾患原因遺伝子を解明することによって各個人の遺伝的なリスクプロファイルを知ることができ、これによって予防的なアクションをとる流れに変わると予想している。これにより、早期診断–最適な治療が可能となり、生存率が劇的に改善されるはずと言う信念を持って開発を進めている。ゲノム配列情報から機能解明に持っていく際に、多くの企業が試みているDifferential Displayを指標として遺伝子を発見しようとする方法がある。しかし、この方法では塩基配列の差異がかなり大きくなければ見つからないこと、情報が多く得られないこと、1次的な原因遺伝子であるのか2次的な原因遺伝子であるのか分からぬなどの欠点を持つ。彼らは、試行錯誤の結果ProNetという方法を考え出した。そのポイントは、遺伝子として機能が分かっている場合は、その機能が分かれているタンパク質を利用して、相互作用を決定していく方法をとる。また、機能が全く分かっていない新しい遺伝子の場合は、よく機能が分かっているタンパク質を利用して、相互作用をとにかく決めていくという形で、pathwayを調べていくことがタンパク質機能を決定していく上で効率的であるとしている。全てのタンパク質とその相互作用のパートナーを調べてデータベースを構築すると、それによってSpecific Pathwayにおける各タンパク質がデータベース化される。その結果、重要な疾患のPathwayが明らかになるととともにそのPathwayの中で新しいタンパク質が同定され、その機能も類推されることになる。彼らは2001年までに8万から10万のタンパク質を同定し、最終的には40万のタンパク質の同定を目指している。

以上紹介したのは第1世代といわれるゲノム・ベンチャー企業の中の3社であるが、彼らは単にシーケンスの決定だけでなく、機能についても研究しており、Positional Cloning、ESTの蓄積、Protein – Protein Interactionなどと方法は異なっているが、各

社ともそれぞれ新しい開発の方向に向かいつつある。

6-2-3 第2世代ゲノム関連ベンチャー企業の動向

第2世代ゲノム関連ベンチャー企業は、まだ、その規模が小さいこともあり、第1世代企業のように1社で塩基配列解析から機能解析までを行うというよりも、機能解析・支援技術の研究・開発に焦点を絞る企業が多い。支援技術の中には以下のものがある。

(1) バイオチップ技術

Affymetrix 社や Synteni 社の DNA チップが脚光を浴びたが、最近 RosettaBioSystems 社というベンチャー企業が現われ、Affymetrix 社と共同開発を行っている。ガラス板の基板上に 10 万個から 15 万個のオリゴヌクレオチドを結合させておき、正常細胞と異常な細胞を使い、そこから mRNA をとり、cDNA をとって、それをハイブリダイズさせる。このように各種の DNA チップを使うことによって、どの遺伝子が活性化され、どの遺伝子の発現が抑えられているかを調べ、ゲノムの機能を調べる糸口としている。

もう一つは、プロテイン・チップで、Ciphergen 社が開発している。サンプルをマトリックスの中に埋め込み、レーザー光を当てると切断されたタンパク質分子が飛び出し、これを質量分析するという原理による。タンパク質については分子量 1 万以上が一度に測れる。この原理を用いて、チップの形にしてタンパク質をマススペクトルで調べようとするもので、非常に高速で操作ができる。このシステムの条件をいろいろと変えることによってマススペクトルの変化からタンパク質の機能を推定しようというものである。さらに発展してレセプター・チップ、シーケンス・チップなどの構想がある。

(2) 物理的・化学的方法による遺伝子機能解析

1. Time - of Flight Mass Spectrometry

DNA をマススペクトルで調べようとするものである。Genetrace Systems 社では、チップのマトリックス中に埋もれた DNA をパルス・レーザーで飛びださせて、縦型のマススペクトル分析装置で質量によって検出する方法をとっている。非常に短時間で機能解析ができるという利点がある。この機械を用いた遺伝子同定で、High - Throughput、Genotyping、High - Throughput cDNA screening ができる。また、この機械を用いて Mass Signature Technology と呼ぶ手法で、ラベリングしたものを、そのラベルを外した後、マススペクトル分析をするものがある。その他、細胞培養の後、Liquid

Northernを行ってからマススペクトル分析をして、遺伝子発現のパターンが全て分かるという手軽な手法もある。SEQUENOM社では、同じ原理で横形のマススペクトル分析機を使用している。チップはスペクトロ・チップと呼び、BiomassPROBE、BiomassSEQUENCEなどの手法を開発している。これらを組み合わせることによって、医薬品開発のターゲットを同定し、最終的に治療薬の開発に結びつけようというものである。

(3) タンパク質合成

遺伝子の機能解析のための1つの鍵は目的とする遺伝子が規定するタンパク質を実際に合成し、これを用いて機能解析するということがある。Gryphon Sciences社は短時間に大量のタンパク質を合成するシステムを開発している。この技術で未知のタンパク質を大量に合成して機能を調べるという意図を持っている。これは化学的合成であるが、このような技術も今後ゲノムに関連していくと考えられる。

(4) Transcription Factor

ヒト・ゲノム情報を利用し、これに対するDNABinding Proteinを設計することによって、どんな遺伝子でも認識できるというUNIVERSAL GENE RECOGNITIONを考えているのがSangamo社である。これにより、いわゆるFunctional Genomicsや診断、医薬品による治療に応用しようというものである。当社は、DNAの塩基を認識して結合するいわゆるZinc Finger DNA Binding Proteinsを用いて結合部位の対応を全て調べた。次にZinc Fing側のアミノ酸を置換して特定の塩基配列を認識して特異的に結合するZinc Finger Proteinのグループをつくった。これを使うことによってTranscriptionに関係するタンパク質を調べることができるほか、治療薬や診断薬の開発にもつながるとしている。

6-2-4 バイオ・インフォーマティックス

バイオ・インフォーマティックスだけで機能を全て調べることはできないが、いろいろなデータを統合することによって機能を絞ることができる。Pangea、Molecular Application Group、NetGenicsといった企業がこの分野の開発を行っている。いろいろな機能を推定するような新規のアルゴリズムを開発する会社も次々と設立されている。Synomics、British Bioinformatics、StructualBioinformaticsの3社は、タンパク質の高次構造の中で重要なバインディングサイトなどをコンピュータでデータ化し、そ

れとシーケンスとの整合性のデータを蓄積している。それによってある特定の遺伝子の配列が、タンパク質としてどの様な構造をとるかを知り、ゲノムのデータからタンパク質の三次元構造を再現するソフトを開発しようとしている。NetGenics社はIncyte社と提携して医薬品創製のための統合化ソフトを開発している。ここではターゲットの同定と、CharacterizationをするBioinformaticsの他、Cheminformaticsをつくり、化合物のスクリーニングを最適化をコンピュータの上でやろうとしている。さらに前臨床、臨床から最終的な経済効果の測定に至るまでのソフトウェアを考えている。彼らの基本的な考えは、既存のデータベースにアクセスして、自社の独自のデータベースを構築するというフレキシブルな発想を持っている。MolecularApplication Group社はGeneMine Discovery Engineと呼ぶものをつくり、ホモロジー、モティーフ、推定構造などのデータを全て統合して、効率よく医薬品の開発に結びつけるプログラムを開発している。

6-2-5 動物系を用いた遺伝子機能解析

動物を用いて遺伝子の機能を解析しようという1群のベンチャー企業群がある。その1つは、ノックアウト・マウスの開発を進めていくという会社である。Lexicon社はES細胞から各種ノックアウト・マウスの受精卵をつくって凍結し、凍結卵の状態で供給する。Deltagene社はノックアウトマウスをつくる期間を従来の約51週から18週と短くした。これに対し、ヒト以外の各種動物の遺伝子の配列を次々と調べて、それらの比較の中からモチーフを探していくという戦略をとる会社群がある。Exelixis社は線虫とショウジョウバエを用いて新しい遺伝子を導入したり、ノックアウトして機能解析を行っている。また、Gene/Networks社はJackson Laboratories社等と提携して各種系統のマウスの遺伝子を全て解析し、その結果からデータベースをつくる。その遺伝子型、表現型を調べ、最終的に遺伝子発現のプロファイルのデータを構築して、医薬品の開発につなげようと計画している。

6-2-6 ツールの開発

Firefly社はポーラスなシリコン層を用い、これに光を与えると増幅が起きて非常に感度よく物質を検出できる装置を開発している。これを用いると物質をラベルすることなく反応を追跡することができる。例えば、DNAのハイブリダイゼーションの前後や、タンパク質と抗体の結合によるディフェレンシャル・スペクトルを定量的に正確に測ることがで

きる。次の世代の機能解析装置として注目される。

6-2-7 Pharmacogenetics

現在のDNAの機能解析から、次の世代は、集積された遺伝情報を使って有効な医薬品を開発するためにPharmacogeneticsが必要であるといわれている。例えば、開発した医薬品に関して、Response genesやMetabolism genesについて解析して、効果と副作用の情報を確認したり、遺伝子測定の結果によって個々の患者に適合した治療薬を調合あるいは創製するという方向に向かうとされている。

6-2-8 まとめ（ポスト・ゲノムの創薬）

従来の医薬品の開発の流れと全く異なった、ポストゲノム時代の創薬の流れが、かなり明確になってきた。薬のターゲットとなる分子の同定に関しては、GenomicsとProteomicsによるデータの蓄積が進み、Functional Assessmentがひとつのブームとなっている。一方、High - Throughput ScreeningさらにはUltraHigh - Throughput Screening技術の開発も行われており、このようなシステムからリード化合物が見つけられ、有効性、毒性を調べて臨床試験に進められる（表6-1）。ゲノム・ベンチャーは各社固有の技術を駆使して開発を進めている（表6-2）。これらのデータをすべてコーディネートする形でデータベースが構築されて、最終的に医薬品が開発される。さらに、臨床実験のプロトコールのコンピュータによる作成や、製品の経済性の試算を含めたソフトの開発が進められている。今後、このような方向で創薬の時代が始まるものと思われる。

表6-1 ゲノミックス技術の創薬および開発へのインパクト

Technology	Target Identification	Target Validation	Assay Development	Preclinical (Toxicology)	Clinical (Pharmacogenomics)
Genome mapping	●				●
Positional cloning	●				
Random sequencing	●				
Genetics of model organisms	●	●	●		
Two-hybrid assays	●		●		
Hybridization on a chip	●			●	●
Differential display	●			●	
Gene knockouts		●			
Cellular assays			●		
Massively parallel sequencing	●			●	●
Yeast reporter array				●	
2-D gel electrophoresis	●				

出典 : H & Q Industry Report

表6-2 ゲノム関連ベンチャー企業と、その主要技術

Company	Genome Mapping	Positional Cloning	Random Sequencing	Genetics of Model Organisms	Two-hybrid Assay	Hybridization on a Chip	Differential Display	Gene Knockout	Cellular Assays	Massively Parallel Sequencing	Yeast Reporter Array	2-D gel Electrophoresis
Acacia												
Affymetrix						●						
Aurora									●			
Cadus				●					●			
Chirosciences	●			●								
Duragen					●		●		●			
Deltagen							●					
Digital Gene												
Evotec										●		
Exelixis				●								
Gene Logic							●	●				
Gene Networks				●								
Gene Trace			●					●				
Genome Therapeutics	●	●	●									
Genset SA	●	●	●									
Hexagen				●								
Human Genome Sciences			●									
Hyseq						●						
Incyte	●		●			●	●					
Large Scale Biology											●	
Lexicon				●					●			
Lynx								●		●		
Millennium	●			●				●				
Myriad	●				●							
Nanogen						●						
OncorMed	●						●					
Oxford Glycosciences											●	
Progenitor		●										
Proteome											●	
Protogene						●						
Rosetta						●						
Sequana	●	●		●								
Synteni							●					

出典 : H & Q Industry Report を改変

6-3 ゲノム情報解析とコンピュータ構造解析

一欧米の動向一

6-3-1 はじめに

タンパク質などの生体分子の機能をその立体構造を介して深く理解しようとする「構造生物学」は、最近次々と解読されて来ている「ゲノム情報」の機能解明に非常に重要な役割を果たし始めている。

欧米ではこの構造生物学の重要性を早くから認識し、特にここ数年その推進策が強力に進められている。構造生物学の具体的な手段としては、X線結晶構造解析、NMR溶液構造解析、更には極低温電子顕微鏡、AFMなどがあるが、これらに加えてコンピュータによる立体構造予測、モデリング、フォールディングシミュレーション等も有力な手段になっていっている。

特に、膨大なゲノム情報の解析に当たっては、コンピュータによる立体構造解析・予測技術が非常に有効である。これまでに開発された各種解析技術を自動化して、膨大なゲノム情報のコンピュータスクリーニングをしようとする動きも出ている。(図6-1)

以下では、まず欧米の構造生物学に対する施策を紹介し、次にタンパク質関連の学会でのゲノム情報解析関連の話題を述べ、最後に各研究機関での活動や自動化の動きを紹介する。

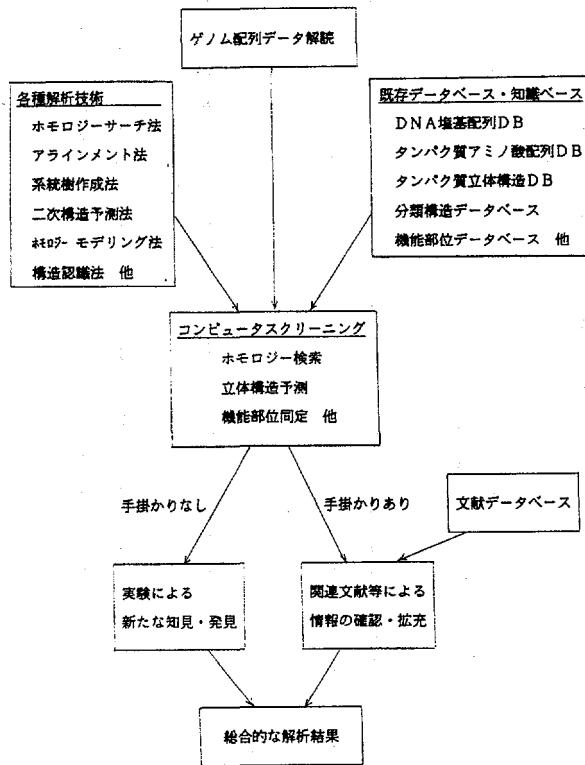


図6-1 ゲノム情報のコンピュータスクリーニング

6-3-2 欧米の構造生物学の施策

米国では、1992年から始まった政府の21世紀に向けたグランドチャレンジプログラムの中で構造生物学が一つの重要テーマとして取り上げられて来た。NIH（国立保健機構）、DOE（エネルギー省）などの各種のグラントでその推進が図られ、多くの重要なタンパク質の立体構造解析が進んだ。又、コンピュータによるタンパク質の折れたたみ過程のシミュレーションや立体構造の分類なども進んだ。更に、NSF（全米科学財団）傘下のサンジエゴスーパーコンピュータセンターは、1997年10月から計算構造生物学の推進を標榜して拡大した。

米国ではこのような政府の強力な支援の下に、各所に構造解析に強い研究グループがあり、その対象も、生体内の転写機構・情報伝達機構などの解明の科学研究と、医薬・診断、環境、植物、バイオプロセス、エレクトロニクス等広範囲な応用を視野に入れている。

欧州では、1995年からEU15ヶ国の構造生物学の調査が始まられ、1997年7月その報告書が出された。これには各国の構造生物学の現状と、今後の課題、特にEUとしてまとまって取り組むべき課題が挙げられている。これらについてのプロジェクト化がすでにいくつか始まっている。更に、1997年10月に構造生物学の成果を産業界に移転を促進するための機構として「SBIP」(Structural Biology Industrial Partner Ship) が結成された。

6-3-3 国際学会・シンポジウムでの話題

最近、欧米のタンパク質関連の国際学会・シンポジウムでは、ゲノム情報解析とのつながりに関する話題が急速に増えて来た。特に、コンピュータによるデータベース解析が強力な武器になること、なかんずく立体構造までを考慮したデータベース解析技術が注目を集めている。この背景には、タンパク質立体構造予測技術の中で、特にホモロジーモデリング法(配列ホモロジーのあるタンパク質の既知立体構造を基に推定構造を構築する方法¹⁾と構造認識法(これまで分かっている立体構造パターンのどれに近いかを予測する方法²⁾の近年の急速な進歩がある。

- 1996年3月にフランスで行われた欧州蛋白工学国際シンポジウムでは、主題が「タンパク質の構造から機能へ」で、その中のトップテーマが「バイオインフォマティックス」であった。EMBL/EBIのC.Sanderが、酵母ゲノムデータに対してタンパク質解析の各種の手法(ホモロジーサーチ、アラインメント、系統樹解析、二次構造予測、内外性予測、ホ

モロジーモデリング、構造認識法、機能予測法など) を駆使して、実験のみでは得られない多くの情報を得たことを報告した。彼はインフルエンザ菌でも成果を挙げている。³⁾

- ・1997年3月に米国で行われた生物物理学会シンポジウムでは、「ゲノム配列情報の立体構造ベース解析」という特別セッションが開かれた。アルゴンヌ国立研究所の総合的ゲノムデータベース解析システムMAGPIE、S.Bryant (NCBI/NIH) から構造認識法の進歩と配列データ解析とのリンク、更に A.Sali (Lockefeller) から自動ホモロジーモデリングシステムMODELERのゲノム情報への適用について紹介があった。A.SALiはホモロジーモデリングはゲノム配列の約20%に使え、NMR以上の精度が得られることを強調した。
- ・1997年6月に英国で開かれた欧州蛋白工学国際シンポジウムでは、主題の一つが、「ゲノム情報からタンパク質構造・機能へ」であった。特に、C.Venter (TIGR) から「比較ゲノム解析」から多くの情報が得られること、更に遺伝子ノックアウトなどによる「ゲノム工学」により、代謝メカニズムや必要最少遺伝子セットの解明が可能になりつつあることが紹介された。
- ・1997年7月の米国タンパク質学会シンポジウムでは、タンパク質の構造変化と病気との関連、タンパク質フォールディング解析に加えて、ゲノム遺伝子のネットワーク解析が議論された。A.Arkin (Stanford) からゲノム遺伝子の発現の時間的・場所的変動をネットワークシミュレーションによって解明しようとするアプローチが紹介された。

6-3-4 欧米研究機関の動向

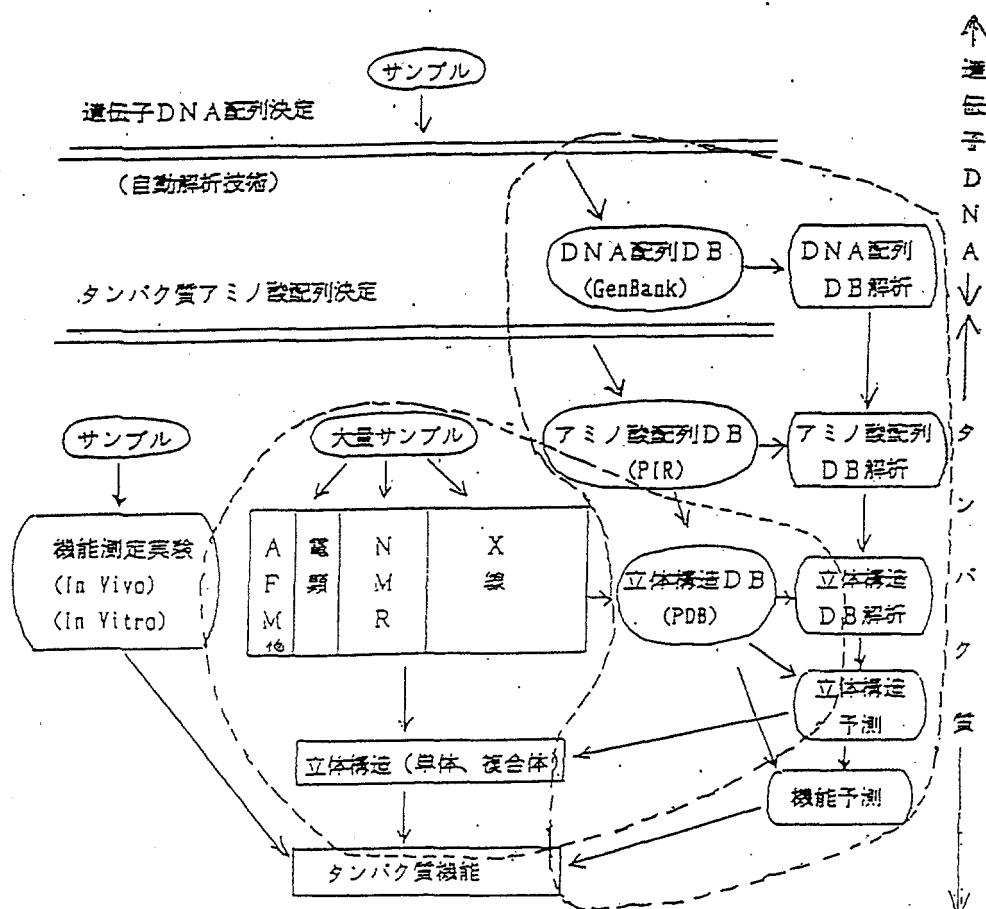
ここ2~3年に訪問した研究機関の中から、ゲノム情報解析に関係あるいくつかの研究機関の状況をご報告する。最近、タンパク質の立体構造データベース解析に詳しい研究者の多くがゲノム情報解析分野に参入している。

- ・欧州バイオインフォマティックス研究所 (EBI) は、ケンブリッジのサンガーセンターに隣接して数年前に設立された。ここには、データベースの構築・維持・提供などを行うグループ以外に、解析研究グループがある。28名いて、そのリーダーにはタンパク質の理論解析・データベース解析で有名な C.Sander や S.Wodak がいる。ここでも、「ゲノム情報から立体構造を通して機能へ」が研究の方向になっており、立体構造分類の自動化 (DALI) が進められている。
- ・ロンドンカレッジ (UCL) の中に、J.Thornton が率いる構造モデリングユニットがある。ここでもタンパク質立体構造分類データベースCATHが構築されており、立体構造と

ゲノム情報の橋渡しのためのシステムSASがある。更に、立体構造予測法の一つである構造認識法プログラム THREADER を、最近ゲノム情報解析用に自動化を進めている。THorntonは、通常ゲノム情報の配列解析のみでは機能は40%以下しか分からぬが、立体構造データの解析技術を駆使すれば更に20~40%向上出来ると断言した。尚、Thorntonは、上記EBIも兼務し、立体構造データベースの面で協力することになった。

- ・米国国立バイオテクノロジー情報センター (NCBI/NIH) の中に、D.Lipman が率いる計算生物学 (Computational Biology) の部門があり、ここでも立体構造分類 (VAST) の自動化が進められている。更に、配列および立体構造のモチーフ抽出が行われている。又、最近ゲノム情報からのタンパク質ファミリー分類の成果を発表している。⁴⁾

ゲノム情報解析—構造生物学—バイオインフォマティックス



6-3-5 終わりに

以上、欧米の施策・学会・研究機関について、いくつかの動きを紹介した。タンパク質の解析技術、特に立体構造データベース解析技術がゲノム情報解析に非常に重要であるとの認識が高い。そのためにタンパク質解析研究者がゲノム情報解析に参入している。又、既存のタンパク質解析システムを自動化して膨大なゲノム情報解析に備えていっている。次々と解読されてくる各生物種のゲノムデータベースに対し、これらの手法の迅速且つ網羅的な適用がなされている。更に、生物種間の「比較ゲノミックス」が新たな分野としてクローズアップし、すでに成果も出始めている。コンピュータ解析力を高める必要性を痛感する。

文献

- 1) Lee,R. : Protein Model Building using Structural Homology. *Nature*,356, 543 - 544 (1992) Sanchez,R.and Sali,A. : Advances in Comparative Protein Structure Modelling. *Current Opinion in Structural Biology*,7,206 - 241 (1997)
- 2) Bowie,J.,Luthy,R. & Eisenberg,D. : A method to identify protein sequences that fold into a known three-dimensional structure. *Science*,253,164 - 170 (1991) Jones,D.,Talor,W.& Thornton,J. : A New Approach to Protein Fold Recognition. *Nature*,358,86 - 89 (1992) Jones, D. : Progress in Protein Structure Prediction. *Current Opinion in Structural Biology*,7,377 - 387 (1997)
- 3) Sander,C. et al. : Challenging Times for Bioinformatics. *Nature*,367,647 - 648 (1995)
- 4) Tatusov,R.Koonin,E. & Lipman,D. : A Genomic Perspective on Protein Families. *Science*,278,631 - 637 (1997)

第7章 まとめと今後の進め方への提言

1980年代の終わりから1990年代初頭に始まった各種生物のゲノム計画は、計画当初の予想を上回る急速な勢いで進行し、単に「染色体地図作成や遺伝子配列決定を行い、その情報を他の生命研究に提供する」という枠を越え、最近は生命研究そのものに変革とも呼べるような、新たな研究アプローチ、研究スタイルを生み出しつつある。今日ゲノム生物学や情報生物学の名で呼ばれているものがそれである。この変革の波は、創薬を始めとしたバイオ産業にも急速に及びつつあり、バイオ系企業が近年一斉に「ゲノムインフォマティックス」を担当する部署の整備・拡充に乗り出すというような状況も生まれている。

さて、ゲノム計画は、ヒトゲノムについて言えば、マッピングと呼ばれる染色体地図・遺伝的 地図の作成段階（第1期）を終え、現在シークエンシングと呼ばれるゲノム配列を大規模に決定して行く段階（第2期）を迎えている。2003年から2005年にはヒトの全ゲノム配列が決定される予定である。一方、微生物ゲノムについては、大腸菌、枯草菌を中心としてすでに10を越える生物種について全ゲノム配列が決定され、現在、遺伝子の同定やその機能の解明が進められている。

このようなゲノム計画の進展状況の中で、冒頭に述べたように、DNAの塩基配列を始めとするゲノムの情報が生命研究やバイオ産業にとって有用かつ重要であるということが認識かつ実証された一方で、ゲノムの配列情報の限界、つまり、配列情報だけから遺伝子の機能を予測することの困難さも見えてきた。例えば、枯草菌を例にとれば全体で約4.2 Mbpのゲノム配列中に約4,000個の遺伝子があると推定されているが、そのうちホモジニ検索と呼ばれる配列解析手法を用いて機能の推測がつくものは、高々その6割程度の遺伝子に過ぎない。また、たとえ、遺伝子の単体の機能を明らかに出来たとしても、多数の遺伝子の集まりが全体としてどのような機能を果たしているのかは配列情報を眺めていただけでは分からぬ。

当然のことながら、ゲノム解析はその配列を決めることが目的ではない。ゲノム配列から多様で複雑な生命現象を読み解き、さらにそれを医学を始めとする生命研究やバイオ産業に生かせるようにすることが最終的な目標である。遺伝子の機能の十分な理解はそれをゲノム全体の機能の中に位置付けることによってはじめて可能となる。つまり、遺伝子を分子のレベルでとらえるだけでなく、細胞レベル、個体レベルの文脈の中に位置付けることによってはじめてその機能を理解することができる。ゲノム全体（細胞レベル、個体レ

ベル) における遺伝子やゲノム配列の機能を理解するには、ゲノム配列情報や染色体地図情報あるいはタンパク質の構造や機能といった情報の他に、細胞内におけるDNA やタンパク質などの分子間相互作用、空間特異的あるいは時間特異的な遺伝子発現様式、遺伝子発現調節ネットワーク、遺伝子やゲノムの変異・多型と疾患その他の表現型との関係、生物種間での遺伝子やゲノムの対応関係、などの広い意味でのゲノム情報（ここではこれらの情報を高次ゲノム情報と呼ぶ）を計算機にデータベースとして蓄えるとともに、その中から有用な生物知識を抽出し遺伝子制御ネットワークなどのモデルを構築し、各種のシミュレーションを行うことが不可欠である。このようにして始めて、ゲノムのトータルな理解に近づくことができるるのである。

このような高次のゲノム情報を網羅的に収集しデータベース化することの重要性は以前から認識されてはいたが、現実には、上記のような情報を系統的にかつ効率良く产生する実験計測技術がなかったり、細胞レベル個体レベルの情報を扱える情報処理技術が未熟だったり、といった事情などから、具体的にそのようなデータベースを構築しようにも実際に実現することは困難であった。

しかしながら、ここ1、2年の間に状況は一変した。これは、DD (Differential Display) 法やTwo-hybrid法を始めとする、上記の情報を取得するためのゲノム解析手法の開発およびそれらの解析を高効率で可能とするキャピラリー電気泳動装置やDNAチップなどの実験計測技術の大幅な進歩による。また、先に述べたように、そのような情報の取得の重要性が基礎研究のみならずバイオ産業においても認識され、需要が一気に高まったこともその大きな要因である。

そこで、現在、世界のいくつかの研究機関やベンチャー企業において、高次ゲノム情報を系統的に高精度で効率良く取得する技術や装置の研究開発が積極的に組織的に進められている。また、一方で、高次ゲノム情報のデータベース化やシミュレーション技術の研究開発も急ピッチで進められている。我が国においても、もちろん、このような観点での研究開発は行われてはいるが、それらは一研究室一企業のレベルに留まっており、多数のベンチャー企業がしのぎを削っている欧米に比べると質量ともに遅れをとりつつあることは否めない。

高次ゲノム情報の高速高効率取得技術、データベース化技術、情報解析技術は生命研究およびバイオ産業にとって必要不可欠の知的基盤であるが、大規模な装置開発を必要とすることから、また、生物系、計測系、情報系にまたがる技術や人材を必要とすることから、

個人あるいは少人数のグループの努力では対処することはもはや不可能になりつつある。日本においても、生物系、計測系、情報系が一体となって、かつ、産学官挙げて取り組むことが必要である。

以下に今後早急に取り組むべき課題を簡単にまとめる。詳細については、本報告書の各章にも、また、昨年度の報告書にも記載されているので、ここでは項目を列挙する程度に止める。

- (1) 遺伝子発現の状態とその時間的空間的变化を網羅的にとらえるための解析手法と計測技術を開発すること。また、それを高精度でかつ効率良く行うための装置を開発すること。
- (2) 遺伝子発現を制御している、とくに転写を制御している因子とそれが結合するゲノム配列の領域を系統的に同定・取得するための解析手法と計測技術を開発すること。また、それを高精度でかつ効率良く行うための装置を開発すること。
- (3) 上記(1)(2)から得られる実験データをデータベース化し、さらにゲノム配列データやタンパク質データ等の各種生命情報との統合化を図り、それらの情報から遺伝子制御ネットワークのモデルを構築するための情報処理技術を開発すること。
- (4) ゲノムの配列情報から遺伝子領域とくに制御領域をその機能を含めて予測するための情報処理技術を開発すること。また、遺伝子制御ネットワークのモデルからそれを構成する遺伝子の振る舞いをシミュレートするための情報処理技術を開発すること。
- (5) ゲノムを比較して、変異や多型を高速にかつ精度よく検出する計測技術とそのための装置を開発すること。この課題は遺伝子制御ネットワークなどの高次のゲノム情報とは直接的な関りはないが、ゲノム情報の産業応用を考えるうえで欠かせないものである。

上記のプロジェクトに取り組むに当たっては、以下のことに留意すべきであると考える。

- (A) 先にも述べたが、上記の研究開発を進めるに当たっては、生物系、計測系、情報系が研究者レベルにおいても企業レベルにおいても一体となって取り組むことが肝要である。どんなに高密度のDNAチップを開発しても、それを計測する技術がなければ、意味がない。また、計測できてもそれを解析するソフトウェアやデータベースがなければ、膨大なデータはゴミと化してしまう。このような状況に陥らないように、技術や装置の開発に際しては、それが生み出すデータが最終的にどのように役立てるかを常に意識して進めることが大切である。
- (B) 上の考えをさらに推し進めれば、従来どちらかといえば「さきに実験・計測ありき」で出てきたデータをどのように（情報）処理するかはあとで考えるというスタイルであったが、いまにもまして情報の洪水が今後予想される状況では、出てきたデータをどのように料理できるかということ、つまり情報処理を先に考えてから、実験手法や装置をそれに合致するように組み上げるという発想も成り立つ。現実的には中々困難であろうが、このようにしてプロジェクトを立案するのも一つの方法であろう。
- (C) 上で生物系、計測系、情報系一体になってプロジェクトを推進すべきであると述べたが、これを実現するのはそれほど容易なことではない。一体となるためには、これらの三つの分野の研究者や技術者が出来るだけ多くの時間を共有することが不可欠である。プロジェクトマネジャーは、そのようなことを可能にするための種々の方策を講じることが必要である。
- (D) ゲノム研究およびそれへの産業応用は、これまで当初の予想をはるかに超えるスピードで展開してきた。今後このスピードは速まることはあっても、遅くなることはない。プロジェクトの遂行に当たっては、状況の変化に機動的に柔軟に対応できるよう、常に見直しを図るべきである。
- (E) 日本ではゲノムインフォマティックス分野の人材が非常に不足している。人材の養成は一朝一夕に行えないため、当面のプロジェクトの遂行には間に合わないが、将来のことを考えて、プロジェクトを進めるに当たっては、人材の育成を図るという視点も考慮しながらその実施を図るべきである。

第8章 むすび

インターネットウェブサイトでの遺伝子塩基配列のデータ公開数が刻々と増え、またネット上での遺伝子解析も可能となり、いよいよポストゲノムシーケンス時代の到来がひしめきと感じられる。諸学会の動きを見ても配列から機能解析への研究方向が強くなりつつあり、各分野において精力的な研究が進められている。インフォマティクス技術は配列から遺伝子の機能予測、同定を行う重要な要素技術として地位を固めつつあり、生物学、情報工学、機器開発の各分野において新しいアプローチと実績の積み重ねが行われている。このような状況の中で本先導研究「ゲノム情報解読利用技術の調査研究」は2年目を迎えた。調査範囲は1) 変異・多型部位の検出・取得技術 2) 特異的発現様式解析技術 3) 発現調節ネットワーク解析技術 4) コンピュータモデル解析技術 5) 国際動向 に分けて調査し、各項目毎に要素技術の検討、研究開発ターゲットの考察を行った。これらの内容はゲノム比較解析法、遺伝子検出分析法、発現検出法、発現調節解析法、遺伝子間の相互作用解析技術、コンピュータモデルによる遺伝子領域予測技術、遺伝子発現制御シミュレーション技術等調査委員の研究及び文献調査内容を盛り込んで行われた。国際動向としては、今年度は特に我が国のゲノムインフォマティクス技術の水準を世界に照らし合わせてどのような位置にあるか、これからプロジェクトはどうあるべきかを考える資料するために国際シンポジウムが開催された。生物学分野、情報解析分野、機器開発分野の専門家を招いて討論が行われた。各分野の連携、協同研究の必要性が再認識される意義あるシンポジウムとなった。今年度は上記項目に関して内外の特許検索調査も行われ、各調査項目の中で触れられた。

以上の調査結果を踏まえプロジェクトの進め方への提言が行われた。欧米では高次ゲノム情報のデータベース化、解析技術がベンチャーも加わって激しい競争のもとに活発に行われているのに対して、我が国では質量ともに遅れているのが現状である。この分野は生物学分野、情報工学分野、機器開発分野の密接な協力のもとで行う必要がある。一企業、一研究所レベルでは不可能であり産・官・学の一体化した研究開発体制の必要性が強調された。

III. 付録 (参考資料)

工業技術院生命工学工業技術研究所

平成9年度研究報告

1. DNA シークエンサーによる高感度ゲル・シフト法

DNA とタンパク質との相互作用の解析法には、フットプリント法やフィルター結合法など、様々な方法が開発されている。平成8年度の本先導研究においては、DNA シークエンサーを用い、あらゆる真核生物について、転写制御領域に結合するタンパク質の検出と、その結合部位の高速・高感度解析を可能にする、長距離・高感度フットプリント法を開発した。平成9年度では、転写制御領域に結合するタンパク質をより簡便かつ高感度で検出することにより、タンパク質の精製やフットプリント法を用いる前段階での予備的な解析を可能にすることによって、解析システムとしてより完成した物とすることを目的として、DNA シークエンサーを用いた高感度ゲル・シフト法を開発した。

評価系としては、出芽酵母 *Saccharomyces cerevisiae* のエノラーゼ1 遺伝子 (*ENO1*) の転写制御領域 (*ENO1* - UAS) と、この領域に塩基配列特異的に結合する転写制御因子 (RAP1) の結合を用いた。まず、*ENO1* - UAS の塩基配列を有する 36bp からなり、5' 末端を赤外蛍光色素である IRD41 で標識した单鎖オリゴDNA と、相補的な非蛍光標識の单鎖オリゴDNA をアニーリングし、36bp からなる 2重鎖DNA プローブを作成した。このプローブに対して、*S.cerevisiae* 細胞を破碎して調製した微量の全細胞抽出液を添加し、室温で約 15 分間結合反応を行わせた後、ゲル長 18cm のポリアクリルアミド非変性ゲルで電気泳動を行った。検出は、LI - COR Model4000L DNA シークエンサーを用い、イメージデータは Photoshop を用いて、明度やコントラストの調整および縦方向の圧縮を行った。

この結果、約 5 μg のタンパク質を含む酵母全細胞抽出液より、RAP1 と考えられる位置に、泳動度が著しく減少したバンドが高感度で検出された。このバンドの形成は、プローブDNA と同じ配列を有する非蛍光標識 2 本鎖DNA で特異的に阻害されることから、確かに、RAP1 による塩基配列特異的結合を検出していることが確認された。また、この時の検出感度は、³²P で標識された DNA プローブを用いた時と、ほぼ同程度であった。

2. ファージディスプレイによる転写制御因子の高速クローニングの評価

ファージディスプレイは、外来タンパク質あるいはペプチドをファージコートタンパク質の融合タンパク質としてファージ表面に発現させ、特異的な相互作用により、目的のタ

ンパク質やペプチドの遺伝子を単離する技術である。従来より、纖維状ファージM13を利用したファージディスプレイ系が中心として利用されているが、この系では原理的に大きな分子量を有するタンパク質などの提示は難しい。そこで本研究では、cDNAなどのライブラリーからタンパク質あるいはDNAと特異的に結合する因子の選択と機能解析を目的として、 λ 系ファージディスプレイを用い、Gal4p DNA結合ドメインをモデルタンパク質としてDNA結合タンパク質選択系を検討した。

まず、出芽酵母 *Saccharomyces cerevisiae* 由来の *GAL4* 遺伝子を、 λ ファージディスプレイベクター (λ foo) の V 遺伝子の下流にアンバー終止コドンで発現量がコントロールされるように挿入し、 λ foo-galを作成した。 λ foo-galは、アンバーサプレッサーを有する大腸菌 XL1-Blue 菌体に感染させて増幅し、アンチ Gal4p 抗体を用いたウェスタンブロッティングにより、ファージ表面への提示を確認した。こうして得られた λ foo-gal と、Gal4p DNA 結合ドメインの挿入断片を持たない λ foo の混合体に対して、Dynabeads を用いた DNA アフィニティー精製により、 λ foo-gal を特異的に 10^2 以上に濃縮することに成功した。以上のことから、 λ 系のファージディスプレイを用いることにより、細胞内の全 mRNA に由来する cDNA を挿入断片とするライブラリーを作成し、DNAとの相互作用を用いてスクリーニングすることによって、転写制御因子を高速に解析することが可能であることが明らかとなった。

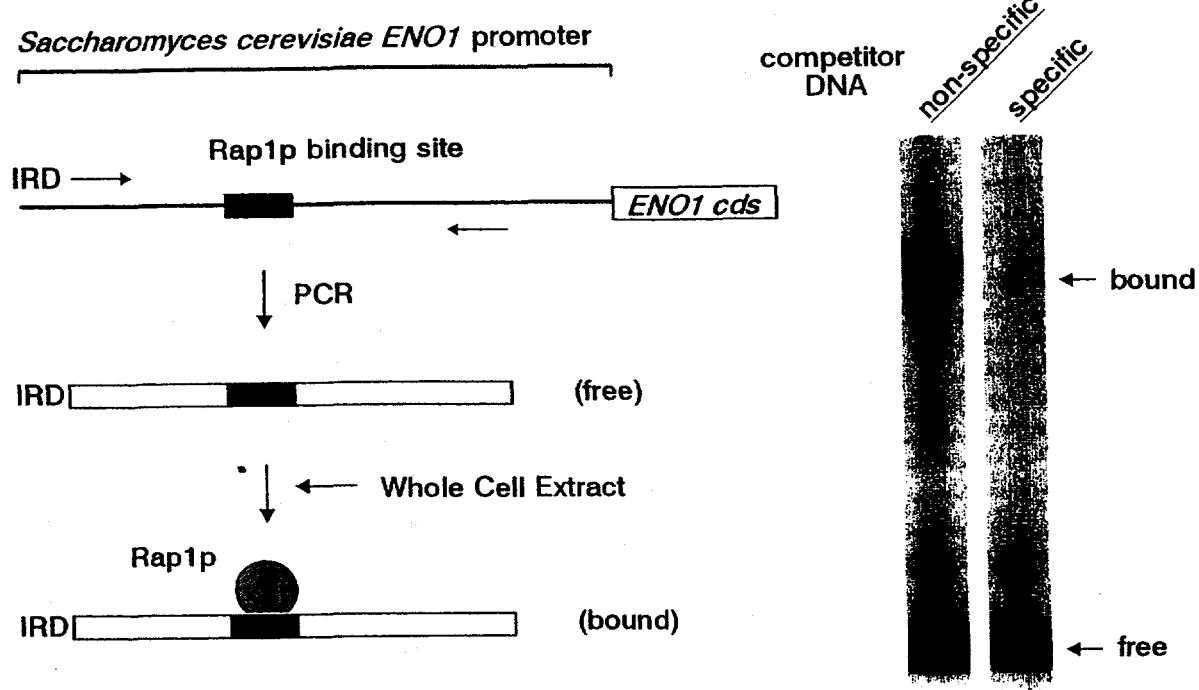


図1 DNA シークエンサーによる高感度ゲル・シフト法

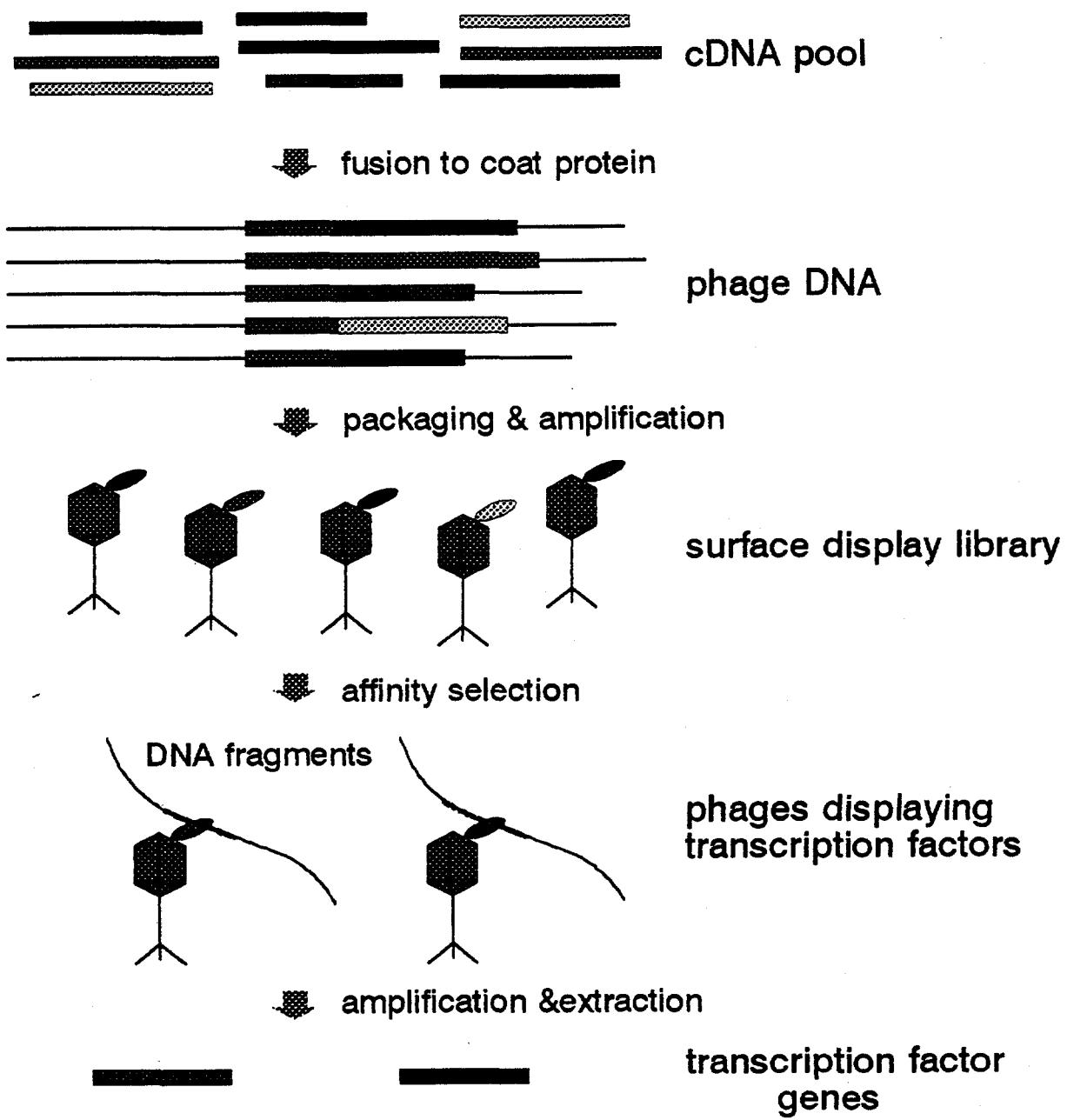


図2 ファージディスプレイによる転写制御因子の高速解析法

工業技術院電子技術総合研究所

平成9年度研究報告

遺伝子機能指向型情報技術

遺伝情報はタンパク質として機能を発揮するまでに複雑な情報処理を生じるから、転写機構、スプライシングの仕組み、タンパク質の立体構造と機能に至るまでの情報をいかに配列データの情報構造の理解につなげるかが重要である。

遺伝情報の情報構造を明らかにし、構造・機能の関係を解明するため、配列データの統計情報やタンパク質の機能情報を確率モデルを用いて表現し、遺伝子構造のモデルを構築するための研究を行った。

また、配列データの統計情報を隠れマルコフモデル等の確率モデルで表現し、タンパク質の機能情報をモチーフ辞書の形で組み合わせることによって遺伝子構造をモデル化する手法を開発した。

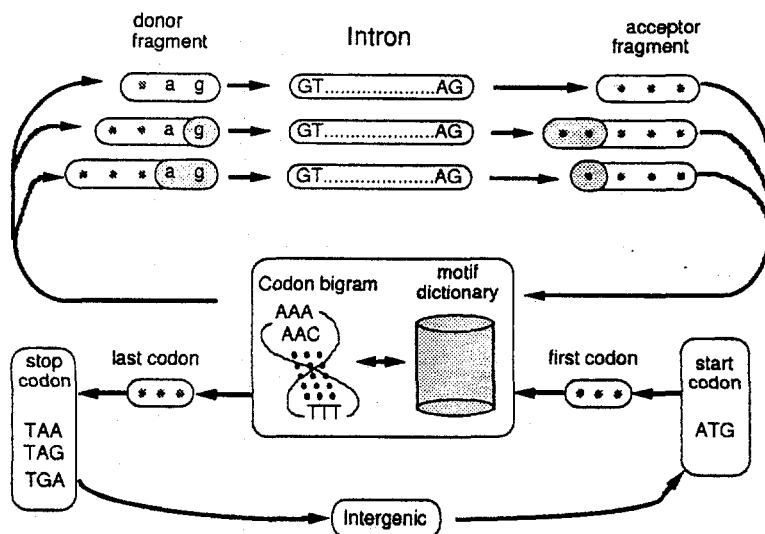


図1 確率モデルによる真核生物の遺伝子構造のモデル化

この手法では、開始コドン、終始コドン、内部コドンの bigram (2次のマルコフモデル)、非コード領域の確率分布などを隠れマルコフモデルを用いてモデル化し、それらの出現の順序関係を文法で記述することにより、構文解析を行っている。

また、タンパク質の配列・機能情報を遺伝子領域の認識に利用するため、タンパク質の配列モチーフを辞書として用意し、構文解析に用いている。これは、アミノ酸の標準パターンで表現されたモチーフを単語とする辞書を作り、その音素に当たるアミノ酸やその

標準パターンを隠れマルコフモデルでモデル化し、コード領域の中に現れる単語パターンの検索に用いることによって、より強力にコード領域を認識するためのものである。

真核生物の遺伝子発見の評価用データとして広く用いられている Burset/Guigo の脊椎動物のDNA配列を用いて認識実験を行った。モチーフ辞書としては、PROSITE（タンパク質モチーフデータベース）の933個のタンパク質モチーフの中から、複雑性の尺度を用いて一定以上の選別性を持つモチーフを選び出して用いた。

Burset/Guigo のデータは、570種類の脊椎動物の遺伝子のDNA配列のデータで、モチーフ辞書のモチーフをコード領域中に97種類、延べ241回含んでいる。このシステムは、そのうち167個を正しく発見した。エキソン、イントロンの認識率を表1に示す。

表1 Burset/Guigoによる脊椎動物570配列に対する遺伝子領域確認率

Program	Accuracy per nucleotide				Accuracy per exon				
	Sn	Sp	AC	CC	Sn	Sp	Avg.	ME	WE
本手法	0.87	0.82	0.81	0.81	0.62	0.51	0.57	0.13	0.11
GENSCAN	0.93	0.93	0.91	0.92	0.78	0.81	0.80	0.09	0.05
HMMgene	0.88	0.94	n/a	n/a	0.74	0.78	n/a	0.13	0.08
Genie	0.87	0.88	0.85	n/a	0.69	0.70	0.69	0.10	0.15
GeneID	0.63	0.81	0.67	0.65	0.44	0.46	0.45	0.28	0.24
GenLang	0.72	0.79	0.69	0.71	0.51	0.52	0.52	0.21	0.22
GeneParser2	0.66	0.79	0.67	0.65	0.35	0.40	0.37	0.34	0.17
GRAIL2	0.72	0.87	0.75	0.76	0.36	0.43	0.40	0.25	0.11
SORFIND	0.71	0.85	0.73	0.72	0.42	0.47	0.45	0.24	0.14
Xpound	0.61	0.87	0.68	0.69	0.15	0.18	0.17	0.33	0.13
GeneID+	0.91	0.91	0.88	0.88	0.73	0.70	0.71	0.07	0.13
GeneParser3	0.86	0.91	0.86	0.85	0.56	0.58	0.57	0.14	0.09

隠れマルコフモデルによる確率的な構文解析手法は、原核生物だけでなく、真核生物の遺伝情報の認識にも有用であることが判明した。イントロン／エクソンの構造を持つ真核生物では、アミノ酸配列モチーフは一般に一つのエクソンの中に閉じて存在するとは限らないので、エクソン中に保存されて現れるパターンから、新たにモチーフ辞書を作る必要があると考えられる。

論文発表

1. Asai,K. ; Ueno,Y. ; Itou,K. and Yada,T. : "Automatic Gene Recognition without Using Training Data," Genome Informatics, No.8, pp.15 – 24 (1997).
2. Asai,K. ; Itou,K. ; Ueno,Y. and Yada,T. : "Recognition of Human Genes by Stochastic Parsing," Pacific Symposium on Biocomputing '98, Vol.3, pp.228 – 239 (1998).
3. Ueno,Y. and Asai,K. : "A Hight – throughput Graphics Library Designed for Portable Molecular Structure Viewer , " Pacific Symposium on Biocomputing '98, Vol.3, pp.201 – 212 (1998).

製品評価技術センターバイオテクノロジーセンター

平成9年度調査報告

1. 塩基配列解析の効率的手法に関する調査

1.1 塩基配列解析の効率的手法としてのショットガンライブラリーの作成方法について

コスミドクローンの解析は一般的にショットガンライブラリーを作成し、リダンダンス(シーケンスの重複)が5以上になるようにシーケンスを行う。この際、シーケンスの数はショットガンライブラリーの質に左右され、再現性・精度が良く、効率の高いライブラリーの作成技術が要求される。

通常ショットガンライブラリーの作成はプラントエンドライゲーションによって行われているため、複数のインサートDNAのコライゲーション、ベクターのセルフライゲーションが生じるため、ライブラリーの質が安定しにくい。PovinelliとGibbs等¹⁾は、アダプターライゲーションストラテジーによるラージスケールシーケンスのための高効率ショットガンライブラリー作成方法を発表している。

これは、ショットガンライブラリーを作成する過程でシーケンスペクターとインサートDNAの粘着末端を加工することにより、理論的に1組のベクターとインサートDNAのみがライゲーション可能となるストラテジーであり、高精度・高効率なショットガンライブラリーを比較的容易に作成することができるものである。

当該研究調査によるDNA塩基配列解析開始当初、ショットガンライブラリーの効率は65~70%程度であったため、前述した方法により効率化を図ることとした。

また、最終的に得られたコンセンサスシーケンスの確認は、通常、DNAを制限酵素により切断し、電気泳動による切断パターンとシミュレートパターンとの比較、あるいはLong-PCRによる合成DNAとゲノムDNAの電気泳動パターンの比較及びシミュレートを行うが、今回はコスミドDNAのダイレクトシーケンスを試みた。

1.2 解析対象物

種名：分裂酵母 *Schizosaccharomyces pombe*

株名：972h- (Mizukami Library)

整列クローン提供者：柳田充弘（京都大学理学部生物物理学教室教授）

1.3 ショットガンライブラリー作成方法の検討

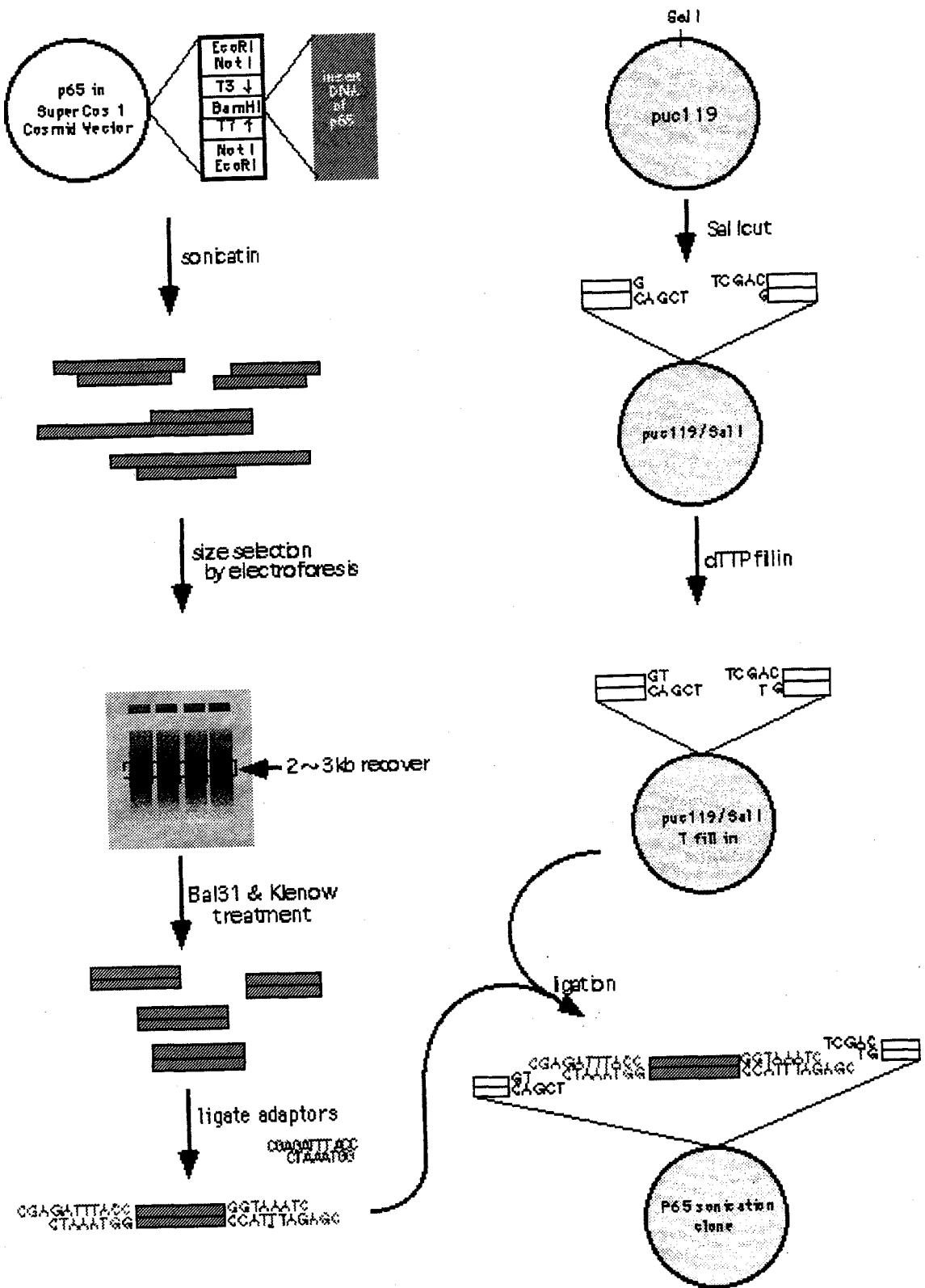
通常、ショットガンライブラリーの作成は、コスミドDNAをソニケーションし、シーケンスに適したサイズフラクションを取り、フラグメント末端をプラント化し、シーケンス用ベクターとプラントエンドライゲーションを行い、これをトランスフォーメーションしている。我々も当初この方法に基づき、ショットガンライブラリーを作成したが、その結果、インサートDNAの挿入がされていないにも関わらずホワイトコロニーとなる（false positive）割合が30～35%程度と比較的多く、ライブラリーの効率が良くなかった。

このため、PovinelliとGibbs等のアダプターライゲーションストラテジーによるラージスケールシーケンスライブラリー作成の文献を参考に再現性・精度が良く、効率の高いショットガンライブラリーの作成方法を検討した。

この方法のポイントは、プラント化したフラグメント末端にアダプターを結合させ、シーケンスベクターと粘着末端によるライゲーションを行うところにある。

シーケンスベクターはpUC119をSalI cutした後、Taq polymeraseによりチミン(T)をFill inし、BAP処理を行うためベクターのセルフライゲーションができない。また、インサートDNAの両端はアダプターと結合しているため、インサートDNA同士のコライゲーションができなくなる。

以上のことから、理論的にライゲーションが可能な組合せは、1組のインサートDNAとベクターのみであり、精度・効率が高いショットガンライブラリーの作成が可能である。



アダプターライゲーションストラテジーによるショットガンライブライマー作成方法

1.3.1 コスミドDNAの抽出、精製

コスミドクローンのcultureからアルカリ-SDS法によりコスミドDNAを抽出し、フェノール・クロロホルム抽出、塩化セシウム密度勾配法による精製を行ったコスミドDNA $40\ \mu g$ を $100\ \mu l$ のTE Bufferに溶かしサンプルとした。

1.3.2 超音波処理の最適化

コスミドDNAをソニケーター（島津社製UPS600）により超音波処理し $2\sim 3\text{ kbp}$ のフラグメントが最大量になるように電気泳動で確認し、超音波処理の条件を最適化した。

1.3.3 ショットガンライブラリーの作成

ショットガンライブラリーの作成は以下の手順により行った。

(1) インサートDNAの調整

- ①コスミドDNAの超音波処理（4・1・3で最適化した条件による）
- ②電気泳動による $2\sim 3\text{ kb}$ のフラグメントサイズの分画・抽出・精製
- ③DNA末端の平末化（BAL31-S NucleaseとKlenowを使用）
- ④DNA両端へアダプターを結合

(2) シーケンス用ベクターの調整

- ①pUC119プラスミドをSalIで切断
- ②Taq polymeraseにより $5'$ 末端のチミンを挿入

(3) クローニング

- ①インサートDNAとシーケンス用ベクターのライゲーション
- ②エレクトロポレーション法による大腸菌への形質導入
- ③抗生物質（ampicillin）を入れたプレートに培養しカラーセレクションによりホワイトコロニーを選択。

(4) ショットガンライブラリーのインサートDNA保持率の確認

作成されたショットガンライブラリーから約100クローンのDNAを抽出し、ベクターとインサートDNAを制限酵素で切断し電気泳動によりインサートDNAと保持率の確認を行い、高効率なショットガンライブラリーの作成方法を確立する。

1.4 コスミドクローン6個のDNA塩基配列の解析

1.3により確立された条件により6個のコスミドクローンについてショットガンライブラリーを作成し、以下の手順により塩基配列の解析を行った。

1.4.1 素データの収集

- ①コスミドクローン1個につき500個のショットガンライブラリーをそれぞれ培養し、DNAをDNA抽出ロボット（クラボウ社製PI-100 Σ）により抽出
- ②正逆両方向についてダイプライマー法によるシーケンシング
- ③編集ソフトウェア（Sequencher）を用い、ベクター配列及び信頼性が薄い部分の削除、塩基配列の結合編集

1.4.2 コンセンサスシーケンスの決定

- ①全体をカバーするシーケンスが得られなかったものは、コンティグの端の部分からGCGのPrimeによりプライマーを設計し、ダイターミネーター法によりコスミドDNAのダイレクトシーケンスを行った。
- ②リダンダンスが不足している部分は同様にデータを追加
- ③以上により得られたデータ（1コスミドクローンあたり800～900シーケンス）からコンセンサスシーケンスを決定

1.4.3 コンセンサスシーケンスの確認

- ①電気泳動による制限酵素切断パターンとコンピュータによりシミュレートした制限酵素切断パターンにより確認
- ②コンセンサスシーケンスの300～400bpおきにプライマーを作成し、コスミドクローンのダイレクトシーケンスを行い、クローン全長をカバーするシーケンスを確認

1.5 結果

1.5.1 ショットガンライブラリーの作成

ショットガンライブラリーの作成方法について検討を行ったところ、従来法により作成したショットガンライブラリーのインサートDNA保持率は65～70%であったが、確立したショットガンライブラリーの作成方法においては95%以上の高い効率を得た。また、

シーケンスからアダプター配列が容易に確認できることから、インサートDNAの配列を確認したところ、キメラ配列は存在しなかった。

1.5.2 コスミドクローン6個のDNA塩基配列の解析

1.3により確立された条件により、6個のコスミドクローンについてショットガンライブラリーを作成し、塩基配列の解析を行った。この結果、6個のコスミドクローンから連続した塩基配列189,165bpを確定した。また、コスミドクローンのダイレクトシーケンスによりアッセンブルを行ったコンセンサスシーケンスの確認を行い、矛盾の無いことが確認された。

1.6 考察

塩基配列解析の効率的手法に関する調査については、今回のDNA解析により解析手法の要であるショットガンライブラリーの作成手法について、95%以上という非常に効率的で精度の高い手法を確立した。また、コスミドのダイレクトシーケンスによりコンセンサスシーケンスの確認が容易に行われ、シーケンスに関する効率的な手法について一定の成果が得られた。今後のDNA塩基配列解析に十分活用できるものと確信する。

2. 遺伝子機能推定技術開発のための塩基配列データからの機能遺伝子等の推定に関する調査

2.1 塩基配列データからの機能遺伝子等の推定について

平成8年度は、分裂酵母 (*Schizosaccharomyces pombe*) の約40Kbp (1コスミドクローン) の領域について遺伝子機能等の調査を行ったが、平成9年度は、約190Kbp (6コスミドクローン分) の塩基配列に対して、ORF検索、motif検索、ホモジジー検索及び統計的な検索等の手法を用い、遺伝子領域の推定及び機能遺伝子の推定の可能性についての調査を行った。

2.2 遺伝子領域の推定

①分裂酵母は遺伝子中にイントロンを含むため、分裂酵母における遺伝子領域開始・終止コドン及びイントロンとその前後の特徴的な配列を検出し、イントロンを考慮した遺伝子領域の推定を行うINTRON.PLOT (Coldspring HarborのDr.Michael Zhangの開発

した分裂酵母専用のGene Finding ソフトウェア。以下「分裂酵母専用型ソフトウェア」という。) を用いた遺伝子領域の推定を行った。

②同様に、開始コドン (ATG/GTG) から終止コドン (TAA/TAG/TGA) までが 100 アミノ酸以上の遺伝子領域を検索する汎用型の Gene Finding ソフトウェア (以下「汎用型ソフトウェア」という。) によりイントロンを考慮しない遺伝子領域の推定を行い、①で得られた結果との比較を行った。

2.3 遺伝子機能の推定

遺伝子領域の推定の結果得られた推定遺伝子領域について、DNA 及びタンパク質データベース (DDBJ、GENBANK、SWISS PROT) とのホモロジー検索 (BlastX、BlastN、Smith – Waterman) を行うことによって機能の推定を行った。

2.4 結果

2.4.1 遺伝子領域の推定

分裂酵母専用型ソフトウェアによって 80 個の遺伝子領域が推定され、約 70 % が既知の遺伝子と相同性が高く、その中には細胞分裂を制御する遺伝子、細胞の生死を制御する遺伝子があった。また、分裂酵母専用型ソフトウェアによって推定した 80 個の推定遺伝子領域の半分はイントロンを持つ遺伝子であった。

汎用型ソフトウェアによって 112 個の遺伝子領域が推定された。分裂酵母専用型ソフトウェアによって推定された 80 個の遺伝子領域との比較を行った結果は以下のとおりである。

- ①分裂酵母専用型ソフトウェアによる推定遺伝子領域と全く同じ領域を推定した数は 30 個
- ②分裂酵母専用型ソフトウェアによる推定遺伝子領域の全領域をカバーする領域の数は 9 個
- ③分裂酵母専用型ソフトウェアによる推定遺伝子領域の部分領域をカバーする領域の数は 22 個
- ④分裂酵母専用型ソフトウェアのみで推定された推定遺伝子領域の数は 19 個

2.4.2 遺伝子機能の推定

2.4.1で推定された推定遺伝子領域についてDNA及びタンパク質データベース(DD BJ、GENBANK、SWISS PROT、OWL)とのホモロジー検索(Smith-Waterman)を行い、遺伝子機能を推定した結果を別表に示す。

推定された遺伝子機能のほとんどは、分裂酵母専用型ソフトウェアによる推定遺伝子領域と汎用型ソフトウェアによる推定遺伝子領域で同様な結果を示しているが、分裂酵母専用型ソフトウェアによる推定遺伝子領域において機能推定されたものが若干多かった。

2.5 考察

塩基配列解析の効率的手法の開発によって大量の塩基配列のデータは得られるようになってきた。このような大量の塩基配列データから遺伝子情報を解析するため、Gene Findingを行うトライアルが行われている。

今回、遺伝子領域の推定においては、2つのソフトウェアによって試みを行った。

INTRON.PLOTは、分裂酵母用に最適化されたGene Findingソフトウェアであるため既知の遺伝子と相同性の高い遺伝子領域を推定できたが、これは先人の分裂酵母の研究により得られた遺伝子及びイントロンの特徴的な配列に関する情報を組み込んだソフトウェアを活用した成果である。

しかし、DNA解析を行う生物種に関する研究を生かしたこのようなソフトウェアが常に存在するわけではない。また、INTRON.PLOT以外のGene Findingソフトウェアもあらゆる生物種に対応した万能なものではない。

大量の塩基配列のデータを産業利用していくために、遺伝子の推定はこれから的研究開発においてますます重要になると思われる。

このようなことから、今後、遺伝子領域の推定をより高精度に行うために、解析した塩基配列に対し一定の遺伝子領域を推定し、その領域に対する同種の生物のホモロジー検索、motif検索の結果を行うとともに、生物種及び遺伝子種に表れる特徴的な配列を考慮する必要がある。

さらに、解析対象生物に近縁の生物種のDNA解析データを考慮するとともに、遺伝子領域の推定だけでなく推定した遺伝子領域の機能を同時に解析することが可能なGene Finding systemの構築が必要である。

文献

- 1) Christine M. Povinelli and Richard , ANALYTICAL BIOCHEMISTRY,
210,16 - 26 (1993)

別表 開始コドン・終止コドン及びINTRON.PLOTによる遺伝子領域の推定結果と機能予測

StartStop	Product	INTRON.PLOT	Product
orf001	HISTONE H2A VARIANT.	pi001	HISTONE H2A VARIANT
orf002	CELL DIVISION CONTROL PROTEIN 2 (EC 2.7.1.-) (P34 PROTEIN KINASE).	pi002	CELL DIVISION CONTROL PROTEIN 2
orf003			
orf004			
orf005		pi003	
orf006			
orf007	HYPOTHETICAL 45.0 KD PROTEIN IN NOT1/CDC39-HMR INTERGENIC REGION.	pi004	HYPOTHETICAL 45.0 KD PROTEIN IN NOT1/CDC39-HMRINTERGENIC REGION
orf008		pi005	
orf009			
orf010		pi006	
orf011		pi007	
orf012		pi008	
orf013			
orf014			
orf015	HISTIDINOL-PHOSPHATE AMINOTRANSFERASE	pi009	HISTIDINOL-PHOSPHATE AMINOTRANSFERASE
orf016		pi010	Glycosyltransferase
orf017			
orf018	GLYCOSYLTRANSFERASE ALG2 (EC 2.4.1.-)		
orf019	HYPOTHETICAL PROTEIN C23D3.15 IN CHROMOSOME I (FRAGMENT).	pi011	HYPOTHETICAL PROTEIN C23D3.15 IN CHROMOSOME I
orf020			
orf021			
orf022		pi012	ACTIN
orf023		pi013	HNF-3/fork-head transcription factor homolog
orf024			
orf025		pi014	cdc2 kinase homologue
orf026			
orf027		pi015	
orf028			
orf029	HYPOTHETICAL 61.4 KD PROTEIN IN SGA1-THS1 INTERGENIC REGION.	pi016	HYPOTHETICAL 61.4KD PROTEIN IN SGA1-THS1 INTERGENIC REGION
orf030	Schizosaccharomyces pombe G protein beta subunit Gpb1 gene, complete cds.	pi017	G protein beta subunit
orf031		pi018	
orf032	hypothetical protein YDR339c - yeast (Saccharomyces cerevisiae)	pi019	hypothetical protein YDR339c
orf033	BEM46 PROTEIN (FRAGMENT).	pi020	BEM46 PROTEIN
orf034			
orf035			
orf036			
orf037		pi021	
orf038			
orf039			
orf040			
orf041			
orf042		pi022	
		pi023	
orf043	PRL1 protein - Arabidopsis thaliana	pi024	PRL1
orf044		pi025	HYPOTHETICAL 27.7 KD PROTEIN IN CPT1-SPC98 INTERGENIC REGION
orf045	TAT-BINDING HOMOLOG 7.	pi026	TAT-BINDING HOMOLOG 7.
orf046		pi027	
orf047		pi028	

orf048	hypothetical protein YPR112c	pi029	hypothetical protein YPR112c
orf049			
orf050		pi030	HYPOTHETICAL 105.9 KD PROTEIN IN AAC3-RFC5 INTERGENIC REGION.
orf051			
orf052	ATP-DEPENDENT RNA HELICASE MSS116 PRECURSOR.	pi031	ATP-DEPENDENT RNA HELICASE MSS116 PRECURSOR
orf053	ALPHA-ADAPTIN	pi033	ALPHA-ADAPTIN
orf054		pi034	
orf055		pi035	
orf056		pi036	PAS4 Protein
orf057	LIPOATE-PROTEIN LIGASE A (EC 6.-.-.-).	pi037	LIPOATE-PROTEIN LIGASE A
		pi038	HYPOTHETICAL 47.4KD PROTEIN IN SHP1-SEC17 INTERGENIC REGION
orf058	glutathione reductase (pgr1+)	pi039	glutathione reductase
orf059		pi040	PROTEIN-TYROSINE PHOSPHATASE YVH1
orf060			
orf061		pi041	HLJ1 PROTEIN
orf062	METHIONYL-tRNA SYNTHETASE, CYTOPLASMIC (EC 6.1.1.10)	pi042	methionyl-tRNA synthetase
orf063			
orf064			
orf065		pi043	HYPOTHETICAL 32.8KD PROTEIN IN NCE3-HHT2 INTERGENIC REGION
orf066	hypothetical protein YPL063w	pi044	hypothetical protein YPL063w
orf067		pi045	WEB1 PROTEIN
orf068		pi046	MEIOSIS INDUCTION PROTEIN KINASE SME1/IME2
orf069			
orf070	ADENOSYLHOMOCYSTEINASE (EC 3.3.1.1)	pi047	ADENOSYL HOMOCYS TEINASE
orf071			
orf072		pi048	ALPHA-1,2-GALACTOSYLTRANSFERASE
orf073		pi049	
orf074	LIPOIC ACID SYNTHETASE PRECURSOR (LIP-SYN).	pi050	LIPOIC ACID SYNTHETASE PRECURSOR(LIP-SYN)
orf075	syntaxin 5 - rat	pi051	syntaxin 5
orf076		pi052	HYPOTHETICAL 59.2KD PROTEIN IN PFK26-SGAI INTERGENIC REGION
orf077		pi053	HYPOTHETICAL 31.7 KD PROTEIN IN VMA7-RPS31A INTERGENIC REGION
orf078		pi054	
orf079	protein arginine N-methyltransferase (PRMT1)mRNA	pi055	protein arginine N-methyltransferase
		pi056	spliceosomal protein
orf080		pi057	
orf081			
orf082	ISOLEUCYL-tRNA SYNTHETASE, CYTOPLASMIC (EC 6.1.1.5)	pi058	ISOLEUCYL-tRNA SYNTHETASE, CYTOPLASMIC
orf083			
orf084			
orf085			
orf086		pi059	
orf087			
orf088		pi060	histone H3.1
orf089	HISTONE H3.1/H3.2.		
orf090	HISTONE H4.	pi061	histone H4
		pi062	
		pi063	probable membrane protein YOL129w
orf091		pi064	note
orf092	MITOCHONDRIAL IMPORT RECEPTOR SUBUNIT TOM40	pi065	MITOCHONDRIAL IMPORT RECEPTOR SUBUNIT TOM40

orf093	MITOCHONDRIAL IMPORT RECEPTOR SUBUNIT TOM40	pi065A	MITOCHONDRIAL IMPORT RECEPTOR SUBUNIT TOM40
orf094	probable membrane protein YOL130w	pi066	probable membrane protein YOL130w
orf095			
orf096			
orf097			
orf098		pi067	
		pi068	
orf099	MITOCHONDRIAL FAD CARRIER PROTEIN FLX1.	pi069	MITOCHONDRIAL FAD CARRIER PROTEIN FLX1.
orf100		pi070	HYPOTHETICAL 229.9KD PROTEIN IN NUC1-PRP21 INTERGENIC REGION
orf101			
orf102		pi071	ORF YLL031c
orf103	probable membrane protein YLL031c	pi072	
orf104		pi073	
orf105		pi074	
orf106	C-5 STEROL DESATURASE (EC 1. -.-.-)	pi075	C-5 STEROL DESATURASE
orf107			
orf108		pi076	MITOCHONDRIAL OUTER MEMBRANE PROTEIN MMM1
orf109		pi077	
orf110		pi078	basic transcription factor 2.35KD subunit
orf111	HYPOTHETICAL 75.4 KD PROTEIN IN HAP2-ADE5_6 INTERGENIC REGION	pi079	HYPOTHETICAL 75.4KD PROTEIN IN HAP2-ADE 5_6 INTERGENIC REGION
orf112	CELLULAR APOPTOSIS SUSCEPTIBILITY PROTEIN.	pi080	CELLULAR APOTOSIS SUSCEPTIBILITY PROTEIN

本報告書の内容を公表する際はあらかじめ
新エネルギー・産業技術総合開発機構
産業技術研究開発部の許可を受けてください。

電話 03-3987-9355
FAX 03-3981-1536