修　士　論　文

染色体画像の検索システムの研究

指導教官　　佐藤 賢二　助教授

北陸先端科学技術大学院大学
知識科学研究科知識システム基礎学専攻

850023　川口 昌宏

審査委員：　佐藤　賢二　助教授（主査）
小長谷　明彦　教授
橋本　敬　助教授

2000 年 2 月

目　次

図　目　次

第　1　章
序論

1.1　背景と目的

従来、生物学研究の現場では、個々の研究者が実験の成果であるデータを独自に整理・分類していた。これは、研究者の間に、研究成果を手元に置き、独自の研究データを持つことによって独創性のある成果を得ようとする意識が強い、といった理由がある。そのため、同じ研究室の中であっても、知識を共有し効率よく共同作業を進めるという点において非常に遅れている。この問題を解決するためには、民間企業で導入されているグループウェア環境のような、研究者にとって使いやすく優れたシステムの開発が必要である[1]。

特に近年、多くの研究室でFISH法やGISH法を扱うようになってからは、染色体画像データを扱う研究者が増えている。これらは生物細胞内の遺伝子などを視覚的にとらえるための手法である。具体的には、特定の遺伝子配列を蛍光色素で標識し、それを染色体上で結合させることによって、遺伝子の物理的な位置や分布を視覚化する。さらに染色体の識別・同定を行う核型分析を合わせて行うことによって、全ゲノム配列決定を待たずとも同種の生物もしくは近縁な生物間での遺伝的差異を短時間で容易に発見することができる。そのため、実験の即時性が求められる医学的、育種学的研究にも有効な手法である。

これらの実験は動植物を問わず、生物種を選ばずに行うことができる。そのうえ少人数・短期間で行うことができるため、２～３人規模の研究室であっても、年間

数千枚単位の画像データを蓄積することになる。このような大量のデータを人手で整理・分類することは、大変な労力を要し、非効率的である。さらに、蓄積したデータの有効利用という点を考えると、染色体画像集合の中から必要な画像を容易に取得できるようにする必要がある。分子生物学の隆盛により、研究室で産出されるデータの量は近年飛躍的に増大している。中でも大量の画像データを取り扱うための負担は大きく、これを効率的に整理し検索するためのソフトウェアには大きな需要がある。

そこで本研究では、染色体顕微鏡写真を例として取り上げ、この種の画像の整理・分類および検索を支援するシステムについて、実験現場における実用性を重視した研究開発を行う。利便性を考え、個々の実験手法のデータを改変なしに直接利用でき、なおかつデータベースシステムに慣れていない実験現場の研究者にも容易に利用できるようなシステムの開発を行う。

1.2　本論文の構成

本論文は次のように構成される。第２章では、生物学実験手法のFISH法と核型分析について触れる。第３章では画像処理技術について述べる。第４章では、前章を受けて、画像情報の検索システムの設計と実装手法について述べる。第５章では、作成したシステムの設定と操作方法について詳しく述べる。最後に第６章で本論文のまとめおよび考察を行う。

第　２　章
生物学実験手法

2.1　生物細胞における染色体の働き

近年の分子生物学の発展により、医学・農学などの分野で細胞遺伝学が注目されている[2]。生物学の実験室では多種多様な生物が実験材料として利用されており、特に動植物の大部分を占める真核生物の細胞は一般的な生体材料の一つである。細胞は生物の基本構成単位であり、非常に小さいが高度に組織化されている。真核細胞には、膜で仕切られた様々な内部構造が存在する。その中で最大のものが、生体における遺伝に特に重要な構造体、細胞核である。

核は細胞内の構造体のうち最初に発見されたものであり、細胞の総体積の10%を占める。情報の司令塔といわれる核には、２つの大きな役割がある。それは生命活動を維持するために様々な指令を行う制御の役割と、遺伝情報を複製し次世代の細胞に伝える生殖の役割である。そのため、真核細胞のDNA（デオキシリボ核酸）のほとんどが局在している。DNAは化学構造的に非常に安定な物質であり、ワトソン・クリックの２重らせん構造で良く知られている。そして、生物にとってもっとも重要な物質であるタンパク質の設計図、つまり遺伝子をコーディングしている。このDNAを保持・複製することによって、遺伝的形質が細胞から細胞へ、また親から子孫へと受け継がれている。

細胞は、成長に必要な細胞の供給と、子孫を残すのに必要な生殖を行うために、分裂する必要がある。このとき、遺伝子を保持するDNAは正確に２倍に増殖し、等

しく分割されなければならない。しかし、細胞分裂間期の細胞核中においては、DNAは非常に細い糸状の構造を持っており、細胞核として一つの構造体に収まっている。そこで、DNAは、細胞が２つに分裂する際に凝縮し、Ｘ字型の染色体構造[図2.1]をとることが知られている。真核生物の有糸分裂中期では染色体はもっとも凝縮しており、しかも一平面上に存在しているので、観察するのに適している。

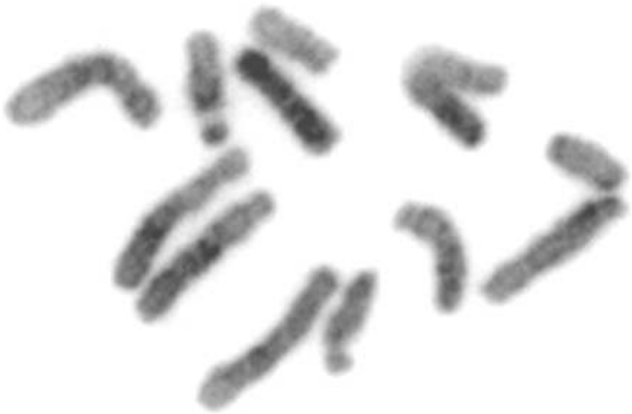

図2.1　真核生物の染色体

ソラマメ野生種(*Vicia Angustifolia*)

細胞遺伝学において、生物が生活していくのに必要な最小限の遺伝子を含む、１組の染色体セットのことをゲノムという。一般的に真核生物は倍数体であり、２つ以上のゲノムを持ち、２ｎ＝ｘでその数が表される。

2.2　ＦＩＳＨ法と染色体マッピング

本来、生体内の遺伝子を観察することはきわめて困難であるが、遺伝子の染色体上の位置を明らかにする試み（遺伝子マッピング）に多くの労力が費やされてきた。その中で、代表的な実験観察手法としてISH（*in situ* Hybridization）法が使われていた[3]。この手法は、染色体、組織、細胞をそのまま酢酸メタノール溶液で固定し、それらの形態を壊すことなく、DNAが存在する本来の部位を観察することができるという利点がある。しかし一方で、放射性同位体を目的遺伝子のプローブ(相補的DNA塩基の一本鎖)に標識し、染色体のDNA塩基配列上に固着(Hybridize)させ、銀

塩フィルムなどに数日から数ヶ月露光させる必要がある大がかりなものでもある。そのため、実験の実施に大変な手間がかかるものであった。また、長期間の露光を行っても十分な結果が出ないことがあったり、得られたシグナルの分布や数を統計処理する必要があったりするため、実験効率や即時性、また生産性などの面で優れているとは言い難かった。

このようなISH法の欠点を補うために改良された実験法が、FISH（Fluorescence *in situ* Hybridization）法[4]である。FISH法では標識としてビオチンやジゴキシゲニンなどの蛍光色素を用い、その蛍光色素に対応した励起フィルタと蛍光顕微鏡を用いるよう改良された。その結果露光時間が大幅に短縮され、実験後すぐにDNA上の遺伝子などを視覚的にとらえることが可能となった。また、遺伝子シグナルの増強効果が良好である上に、拡散の無い状態で染色体を観察できるといった利点もある。そして特に利用価値が高い改良点として、同時に数種類の蛍光色素を使用し、複数の異なったシグナルを観察するマルチカラーFISH法を行うことが可能となったことが挙げられる[5][図2.2]。以上に述べた拡張により、実験現場での生産性が大幅に向上し、FISH法は分子細胞遺伝学における実験手法として欠かせない、非常に有用なものとなった。

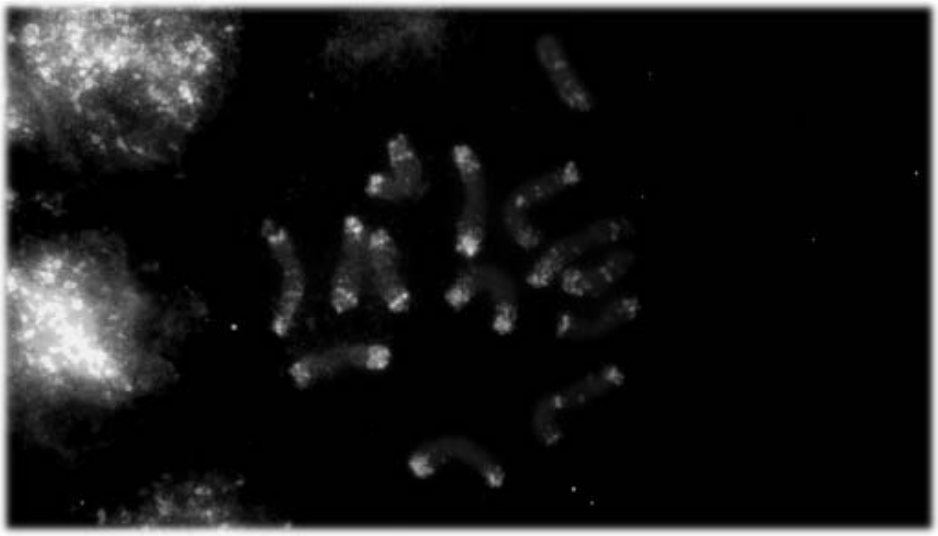

図2.2　マルチカラーFISH法による蛍光染色体画像

タルホコムギ(*Aegilops squarrosa*)

FISH法と同様の蛍光染色を用いる実験手法として、GISH法(Genomeic *in situ* Hybridization)[2]も挙げられる。FISH法と異なり、分析種の全DNAをプローブとして利用し、倍数種の相同な染色体DNAと分子交雑させ、分析種に存在して倍数種に存在する部位、染色体を識別する手法である[図2.3]。これは染色体彩色ともいわれ、倍数種の染色体を、分析種というクレヨンで塗り分けるようなものである。

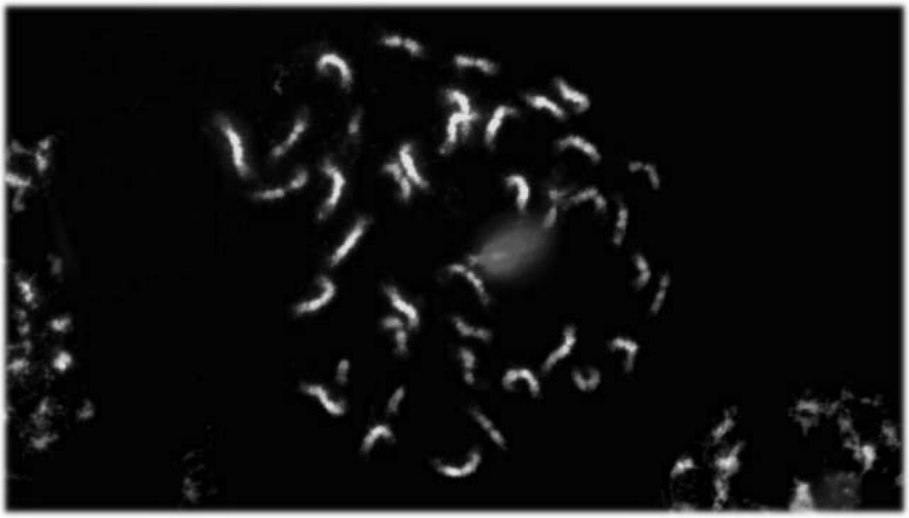

図2.3　GISH法による蛍光染色体画像

アラビカコーヒー(*Coffea arabica*)とリベリカコーヒー(*Coffea liberica*)の間では
ゲノムの半分が同じであり、共通である染色体が蛍光色素により光っている。

通常、実験結果が良好であれば、観察を行った後にできる限り結果を忠実に保存、もしくは記録を残す必要がある。先に述べた通り、ISH法においては元々銀塩フィルムによる露光方式が用いられていた。これは、放射性同位体が放出する非常に微弱な放射線を、時間をかけてフィルムに露光させる手法である。この場合実験結果がフィルムに残るため、そのまま保存することが可能である。一方、FISH法に用いられる蛍光色素は一般的に外光に弱く、強い光にさらされると瞬く間に崩壊・減衰を起こしてしまうため、実験後に素早く試料を記録・保存する必要がある。FISH法では、蛍光顕微鏡によって染色体を直接検鏡可能であるため、初期には、カメラフィルムでカラー写真に収めていた。しかしながら、近年、染色体画像解析が頻繁に

行われるようになり、得られるデータ量が多くなったことや、高解像度で検鏡写真を撮ることができる冷却CCDカメラが普及したことにより、直接計算機上に画像を取り込む方式が主流となった。

2.3　核型分析(Karyotype Analysis)

FISH法による実験で取り込んだ染色体画像を用いて、様々な画像解析を行うことができる。その中でもっとも頻繁に行われる手法の中に、核型分析（Karyotype Analysis）がある[図2.4]。これは、細胞遺伝学の最も基本的な情報である細胞内の染色体セット、つまりゲノムセットを染色体画像から得る画像解析である。

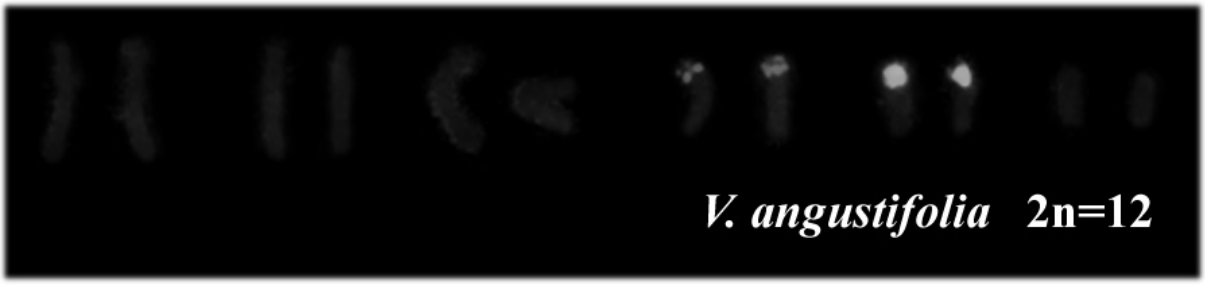

図2.4　核型分析

ソラマメ野生種(*Vicia angustifolia*)

染色体や、その集合としてのゲノムを操作していく上でまず必要なことは、目的とする染色体あるいはゲノムの識別・同定である。核型分析は、有糸分裂中期の細胞内の全染色体を、それぞれの相対長や染色体短腕と長腕の比率、凝縮のパターンなどを元に染色体番号順に識別する手法である。特にFISH法によって染色体マッピングを行った場合、個々の種に特徴的なシグナルがそれぞれ検出されるため、識別が正確に行われる。また、生物種に特異的な遺伝子分布や転座など、解析によってより多くの知見を得ることもできる。

核型分析を用いた研究の一例として、生物種間の遺伝的な近縁関係の同定がある。[6][7]では、植物を材料としたFISH（Fluorescence *in situ* Hybridization）法を用

いて染色体上でのリボゾーム遺伝子のマッピングを行い、*Vicia*属の核型分析について分子細胞生物学的に取り組んでいる。核リボゾームRNA遺伝子(rDNA)は、核の仁形成領域に示される、18S、5.8S、26Sの３種のrRNAをコードした反復配列遺伝子である。これらは生体活動に重要な遺伝子であり、他の主要なDNAと塩基配列がはっきりと異なるため、遺伝的に高い保存性を持っており、容易に識別することができる。また、個々の種について染色体マッピングを行った場合、種・亜種それぞれに特徴的なシグナルが検出される。なぜならリボゾーム遺伝子のシグナルが作り出すパターンは遺伝的保存性が高く、密接な関係のある個体群や種間で高い同一性があるからである。それゆえにリボゾームDNAの多様性は、遺伝子マーカーとして有効であり、特に栽培種とその野生種との遺伝的関係の分析において利用価値が高い。そこで一般に、リボゾームRNA遺伝子をプローブとして利用し、各種の核型上にそれぞれ異なった蛍光色素で検出し、マッピングを行う手法が良く取られる[8]。

これら核型分析には、計算機と画像処理ソフトを用いて、蛍光シグナルを効率よく画像解析処理することが必要である。これらの実験は、短期間・少人数で行うことができるので、２～３人規模の小規模な研究室でも、年間数千枚規模で画像ファイルが貯まることになる。しかし、従来の生物学関連のデータベースは、遺伝子情報などの文字や数値を扱う大規模なものが中心であり、画像の加工や解析を行うようなソフトウェアは存在しても、画像の蓄積や検索を適切に行うシステムが無かった。そのため、多くの場合研究者自らの手で画像の処理や加工を行う必要があり、また、実験現場では本来の研究以外のデータ整理という仕事が発生し、研究者に対する負担が大きくなりがちである。このことは、特に２～３人といった小規模な研究室において、効率の良い研究を行う上で大きな妨げとなっていた。そこで、大量の写真や図を分類・加工・管理するための、使いやすいデータ管理・検索システムを開発することが求められている。

第　3　章
画像処理手法

3.1　既存のアプリケーション

前章で述べた通り、画像の加工・解析を行うことができるアプリケーションソフトウェア自体は既に存在している。それぞれの特徴や利点、欠点を検討することにより、生物学の実験現場で必要とされるシステムの機能を検討する。

既存の代表的な画像解析ソフトウェアとしては、以下のようなものがある。

染色体画像解析システムCHIAS III [9]

農林水産省北陸農業試験場で開発された、通常のパソコンで利用できる染色体画像解析システムCHIASの第三世代システムである。元々、生物学の研究室では、実験設備とセットになった専用の画像処理装置を用いるのが主流であった。CHIAS IIIは初期のCHIASと異なり、専用の画像処理プロセサや画像入力装置を必要とせず、汎用のパーソナルコンピュータと汎用画像入力装置の組み合わせによって植物染色体の画像解析を可能としている。特に染色体の凝縮パターンを画像解析し、ヒストグラムやイディオグラムを得ることによって、染色体の識別同定を容易にしている。システム自体は単独のソフトウェアではなく、Apple Macintosh上のパブリックドメインソフトウェアであるNIH Image[10]上で動作するマクロプログラムである。また、同じくMacintosh用の表計算ソフトMicrosoft Excelとの連携機能もマクロプログラムとして持っている。しかし、入力画像はグレイスケールのものに限られており、マルチカラーFISH法やGISH法など、多色蛍光染色画像の解析には利用でき

ない。

IPLab Spectrum [11]

IPLabは、Scanalytics社の多機能画像解析ソフトウェアであり、生物学の他にも医学、薬学、化学、物理学など自然科学分野全般をカバーする汎用性の高いソフトウェアである。特徴として、多数の拡張モジュールが用意されており、実験設備や目的に応じて柔軟にシステムを構築することができる点が挙げられる。FISH法など、蛍光顕微鏡画像の解析には、Multi-Fluorescenceモジュールが用意されており、画像の合成、ノイズ除去、解析などを行うことができる。IPLabは商用ソフトウェアであり、Apple MacintoshやMicrosoft Windowsなど、汎用のパーソナルコンピュータで動作可能である。しかしその反面、本体・拡張モジュール共に高価である。通常、蛍光顕微鏡と冷却CCDカメラが接続されたコンピュータとセットで少数のみ、実験設備として導入されている。そのため、小規模な研究室ではコストパフォーマンスの面で導入自体が難しく、大規模な研究室においても大量に導入してグループで分散処理を行うことは困難である。

Image-Pro Plus [12]

Media Cybernetics社によって開発・販売されている汎用の画像解析・計測ソフトウェアである。特に粒子解析に優れており、この機能を応用することによって核型分析なども行うことができる。また、画像解析のみに留まらず、実験器具のコントロールをコンピュータ上で行うことも想定されており、接続可能な機器が多い。上記IPLab同様に拡張モジュール形式を採用しており、蛍光画像イメージングモジュールFloro-Proで蛍光顕微鏡写真の操作をサポートしている。しかし、特に染色体画像に特化した機能を持っているわけではない。商用ソフトウェアであるため、本体・拡張モジュール共に高価な点もIPLabと同様である。

以上のように、各アプリケーションソフトウェアにはそれぞれ利点、欠点が存在するが、画像解析のみをとっても、多色FISH法やGISH法に特化したソフトウェアは存在しない。また、蛍光染色体画像が解析可能であるものでも自然科学一般で利用できるような汎用的なアプリケーションであるため、価格が非常に高価であると

いうデメリットが存在する。また、検索・保存に関しては考慮されていないため、研究者のデータ整理に対する負担が大きい。以上をふまえて、FISH法・GISH法に必要な機能を備えた画像解析ソフトウェアについて検討する。

3.2　画像の入力と合成

現在、一般的なFISH実験によって生産される画像データは次のような形式である。染色体画像は、各波長の励起フィルタを通して蛍光顕微鏡で観察された後、冷却CCDカメラによって計算機上に取り込まれる。蛍光色素を３種類使ったならば３枚、２種類ならば２枚のグレイスケール画像が得られ、それぞれが一種類の蛍光色素に相当する。それらをＲ、Ｇ、Ｂ各プレーンに割り振り、画像を合成する[図3.1]。ただし、蛍光顕微鏡での画像取込の際、Ｒ、Ｇ、Ｂ各プレーンがそれぞれ数ドットずれることがある。これは、フィルタが一体型になっている蛍光顕微鏡において、各フィルタの光軸にズレがあることや、フィルタ交換型においては励起フィルタを交換する際の衝撃で、プレパラートおよび観察試料がわずかに動くことが原因と考えられる。そのため、画像を合成する際にＲ、Ｇ、Ｂのずれを修正する機構が必要となる。

また、取り込んだ染色体画像には、多くの場合バックグラウンドにノイズやゴミが固着している[図3.2]。これは、染色体試料を作る際、細胞の押しつぶし法を用いることが多いため、細胞内の染色体以外の構造物や細胞質、細胞分裂中期以外の細胞核などが試料内に含まれるからである。そのため、実験前に試料となるプレパラートの洗浄を行っているにも関わらず、蛍光色素がそれら不純物に固着してしまい、蛍光顕微鏡での検鏡時にノイズとなって現れてしまう。

そこで、後述の染色体切り出しに付随する輪郭抽出を正確に行うためにも、画像全体からのノイズ除去を行う必要がある。本システムでは、しきい値を用いた画像全体の平滑化によるノイズ除去[13]を行うこととし、Ｒ、Ｇ、Ｂ各プレーン独立で行うことができるようにした。しきい値以下の値を持つ画素に対しては、無条件で値を０にするように実装を行う。

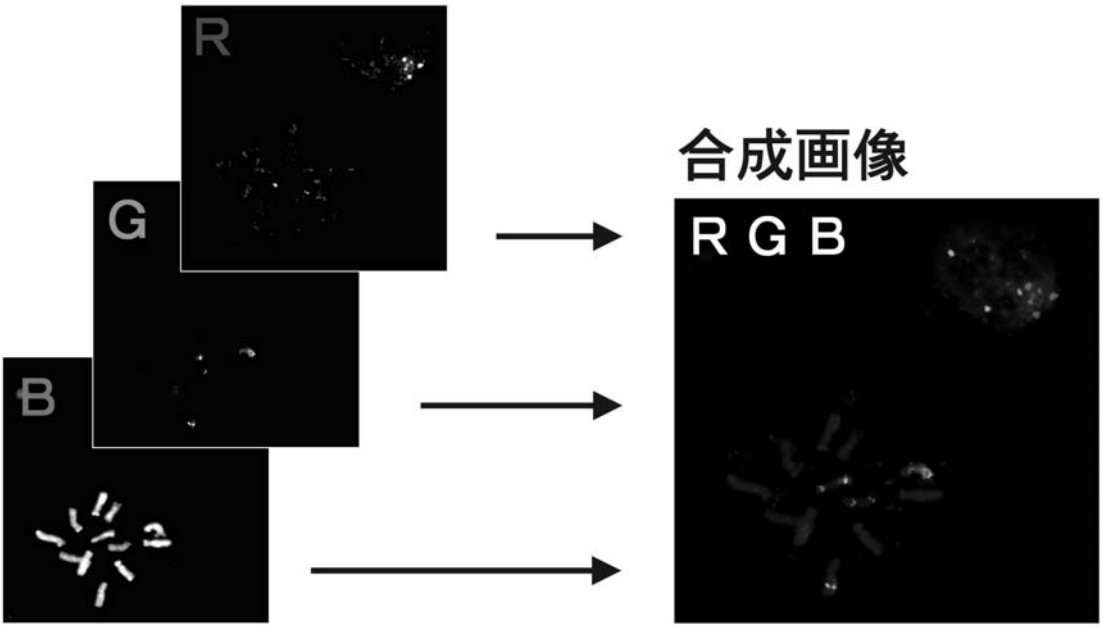

図3.1 RGB各プレーンの合成

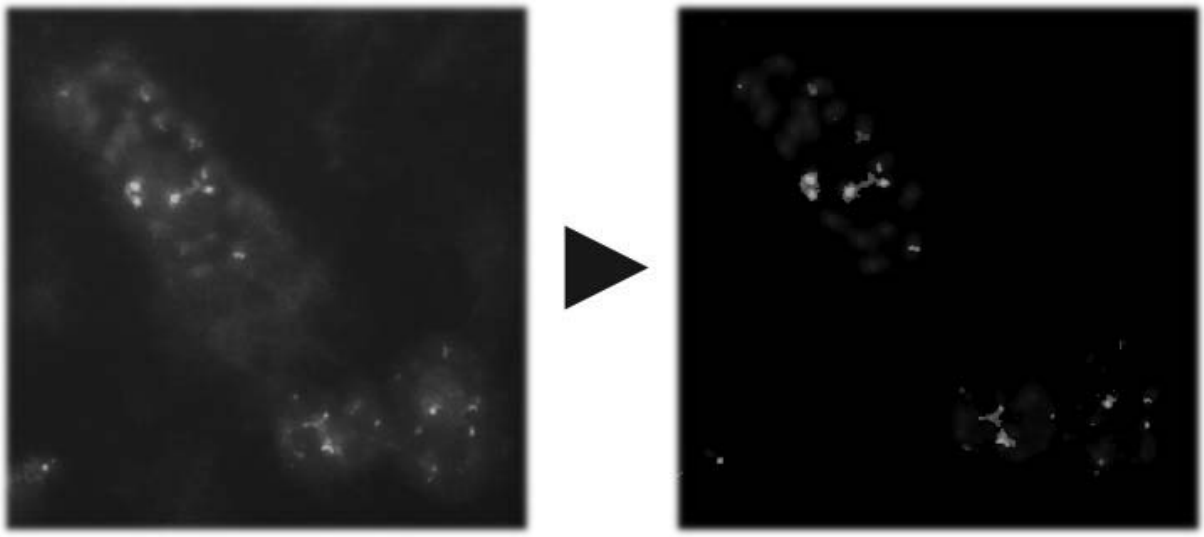

図3.2　RGBの位置合わせとノイズ除去

染色体上のシグナルの位置合わせとノイズ除去が必要

3.3　Ｓｎａｋｅｓによる染色体抽出

核型分析を行うためには、元となるＲ、Ｇ、Ｂの合成画像から、染色体を一本一本切り出す必要がある。従来、このような作業はフォトレタッチソフトウェアを用いて人力で行うか、実験装置と同様に高価な専用解析ソフトウェアを用いる必要があった[5][8]。しかし、これらの手段を用いることは、研究者と研究室、ともに大きな負担をかけるため、効率の良い染色体抽出法が求められている。そこで画像処理技術からのアプローチとして、代表的な輪郭抽出のモデル化技法であるSnakes[14]を用いた染色体切り出しを考えた。

一般に、閉曲線に対して対象領域の輪郭付近で最小となるようなエネルギー関数を定義し、これを最小化することで領域抽出を行う手法は、動的輪郭モデルと定義される。その代表的な手法であるSnakesは、輪郭抽出問題を直接モデル化する手法の一つとして提案された。数学的には、画像の輪郭線を $v(s)$ とし、この上で定義される内部エネルギーと画像エネルギーとの線形和であるエネルギー関数の最小化問題として定式化される。画像上での繰り返し演算の過程が、輪郭に蛇が寄っていくような動きを示す[図3.3]ため、Snakesと呼ばれている。

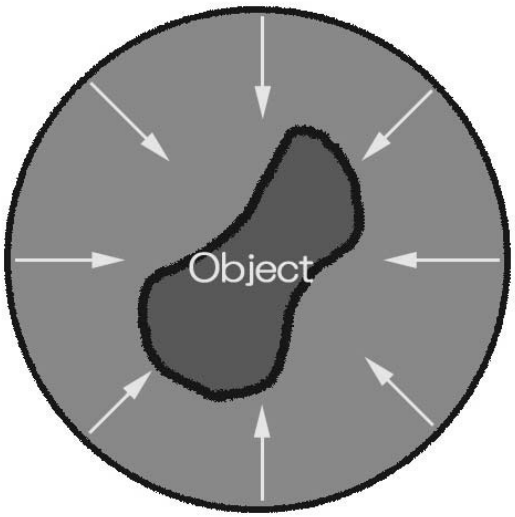

図3.3　動的輪郭モデル

v (s)の内部エネルギーは、輪郭線が一点に収束すれば最小となる関数である。また、画像エネルギーは画像の濃度変化が大きい方向に輪郭線が変化すれば小さくなる関数である。よって、エネルギー関数E_{snakes} [式3.1]は、輪郭の滑らかさを表す内部エネルギー項E_{int} [式3.2]と画像濃度エネルギー項E_{image} [式3.3]の和を、閉曲線に沿って積分する形で定義する。このエネルギー関数に対し、最急下降法を用いて最小化を行う。輪郭抽出にSnakesを利用する利点としては、はじめから閉じた輪郭線をモデル化しているため、閉じた領域を安定して抽出することができることが挙げられる。特に、ノイズの少ない画像においては大変良好な輪郭抽出結果が得られることで知られている[15]。

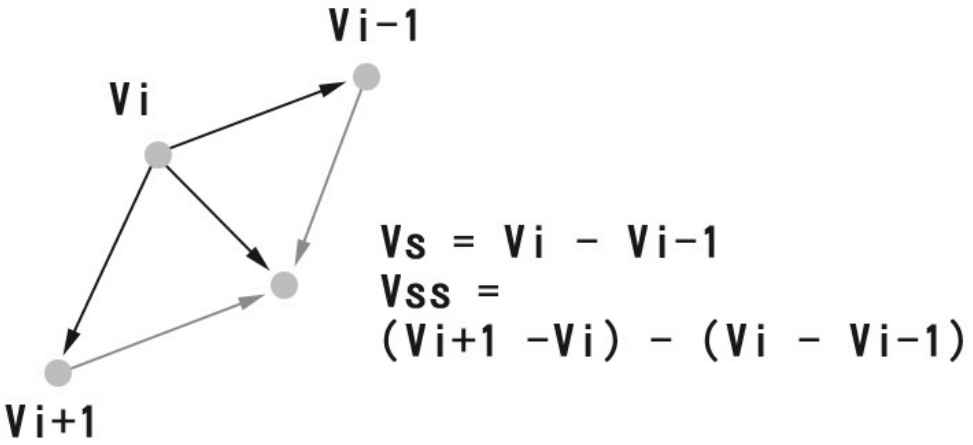

図3.4　節点の移動

$$3.1) \quad E_{Snakes} = \int_0^1 \{ E_{int}(v(s)) + E_{image}(v(s)) \} ds$$

$$3.2) \quad E_{int}(v(s)) = \frac{1}{2} \{ \alpha |v_s(s)|^2 + \beta |v_{ss}(s)|^2 \}$$

$$3.3) \quad E_{image}(v(s)) = -\frac{1}{2} \gamma |\nabla I(v(s))|^2$$

また、Snakesは情報科学分野において広く研究が行われており、様々な拡張を行った手法が提案されている。現在提案されている改良型Snakesには、以下のようなものがある[16]。

１．圧力エネルギーを導入したもの

凹型形状でなめらかな輪郭線抽出に適したSnakesである。対象物体に重心を設定し、重心方向に圧力エネルギーを設定することによって、輪郭線が凹型の場所に入り込む。しかし全体の形状が複雑になるため、全体のエネルギー項E_{Snakes}において振動がおきやすい。その結果安定した輪郭線抽出が難しいため、圧力エネルギーの重み係数付けをうまく設定する必要がある。

２．面積エネルギーを導入したもの

輪郭線内の面積に比例したエネルギーを導入することによって、より正確な凹型輪郭線を抽出する。エネルギー項E_{Snakes}に対して閉鎖域の面積を縮小するようにエネルギーが働く。１に比べ正確な抽出が可能だが、計算量は非常に大きくなる。

３．並列化など計算効率を重視したもの

S-ACM(Sampled Active Contour Model)は、E_{Snakes}の最小化問題を解かずに、エネルギー項をベクトルに置き換えることで、処理の単純化と離散化をはかっている。ノイズに強い、処理時間が短い、並列化が可能などといった特徴がある反面、凹型図形の抽出は困難であり、精度にも欠ける。

４．交差判定により分裂するもの

これまでのSnakesは一つの抽出対象に対して一つの初期閉曲線を設定する必要があった。C-Snakesは、交差判定により自動的に輪郭線が分裂していくため、初期輪郭の設定を細かく行う必要が無い[17]。そのため、画像全体の輪郭線を初期値として一度与えるだけで輪郭抽出を行うことができる。上記２.の面積エネルギー項を取り込んでおり、比較的正確に初期輪郭の抽出が可能である。このような利点が存在する反面、計算量が膨大なものとなり、処理時間がかかり過ぎるという欠点を持つ。

本システムにおいては、対象画像がFISH法により取り込まれた染色体画像に固定されている。この場合、別段拡張を行っていないSnakesでもノイズ状態が良好であれば、FISH画像に対しての染色体抽出は十分可能であり、複雑なアルゴリズムを用いて処理速度を低下させるより、シンプルな実装が有効であると考えられる。そこ

で本システム上に実装したSnakesでは、以下の処理を連続的に行うことにする。はじめに、物体の外部に円周を描き、その上に基準となる節点を作る。次にSnakesの移動エネルギー(ε)を計算する。まず、節点の画素濃度と、両隣の節点の画素濃度から、画素間の位置ベクトルを計算する[図3.4]。これらのベクトルを利用して、節点のつながりを滑らかにするためのエネルギーと、節点間の距離を等しくするためのエネルギーを、移動エネルギーとして算出する。次に、先の２つのエネルギーから、節点を内向けに移動させるためのエネルギーを計算する。最後に対象節点の近傍８点と元の点を合わせた計９点のうち、総エネルギー量が低いものを選び、新たな節点とする。これらのフェーズを一定回数繰り返すことにより、節点は濃度勾配が大きな輪郭上に集まる。この節点を繋ぎ合わせることによって、画像の輪郭を抽出することができる。システムが想定する蛍光顕微鏡画像は、染色体一つの画像が小さいため、染色体一本に対して一回のSnakesを行うように実装すれば、計算量が比較的少なく済む。Ｃ言語によるアルゴリズムとして、[18]を参考にした。

3.4　染色体画像の整列

切り出した染色体画像は、染色体番号順に整列させる必要がある[図2.2]。染色体番号は、２本１組で染色体の長さ順に付けられている。染色体を整列させることにより、異種性物間の遺伝的差異や、遺伝子の転座の様子などをわかりやすく観察・考察することができる。

本システムでは、染色体並び替えについて、直感的にわかりやすい操作体系として染色体のアイコンを利用したドラッグアンドドロップ方式を採用した。大量の画像を扱っても混同しないように、切り出した染色体画像集合をＣ＋＋言語における同一モジュール内で管理するような設計を行う。

第　4　章
情報検索手法

4.1　他分野における情報検索システム

第２章で述べた通り、既存のシステムには、染色体画像、FISH 法などに特化した情報保存・検索システムは存在しない。そこで、生物学以外の他分野での情報科学手法を用いた保存・検索システムの動向をふまえ、実験生物学における使いやすいシステムの設計について検討する。

FISH 法における実験によって得られるデータは、染色体の画像データである。一般に画像データは文字データに比べてその情報量が非常に多いために、画像データを効率的に管理・運用するシステムが必要となる。過去には写真やフィルムをファイルに綴じて保存していた時代もあるが、必要なデータを一枚一枚手作業で探し出すのは大変な手間であった。検索効率が悪いだけではなく、データを保管するスペースや保存媒体の劣化などの問題も発生した。このような理由により、画像データの管理を計算機上で行いたいという欲求が生まれ、画像を電子化して管理・運用するための画像データベースが登場した[19]。画像データベースは、文字・文献データベースとは次のような点で違いが存在する。

- **一つのデータが持つ情報量が非常に大きい**
- **多種多様な画像が存在する**
- **検索のキーとして指定される画像の内容が多種多様である**
- **検索結果の善し悪しの判断が人によって異なる可能性がある**

以上のような問題に対応するためには、画像圧縮やユーザインタフェースなどの様々な技術が必要となる。検索効率の良い画像データベースを構築するには、特に索引生成方法が重要になる。索引生成の方法により、画像データベースはテキスト型画像データベースと内容型画像データベースの２つに大別できる。テキスト型画像データベースでは、画像の登録時にキーワードや画像の内容などを記述した文字情報を画像に付加しておく。検索するときは、文字情報であるキーワードを入力することによって欲しい画像データを指定する方式である、キーワード検索を用いる。一方、内容型画像データベースでは、登録する画像に対して、画像から色、輪郭、テクスチャ、画像から得る印象などを抽出し、索引付けを行う。検索するときはサンプル画像を用いるなどして、指定した画像に類似した画像を返す方式を利用する。これを類似画像検索と呼ぶ。

ここで画像データベースが一般に利用されている例をいくつか挙げる。

絵画画像データベース [20]

絵画など、美術品の分野においては、写真や取込画像など、なんらかのフィルタを通したものよりも、はるかにオリジナルの作品自体が重視される。そのため研究材料を扱ったデータベースとしては成立しにくいが、美術研究の支援や、一般に対する商品としての絵画データベースには需要があり、研究が行われている。その結果、現在では主に CD-ROM としての販売や､インターネットを通した Web 美術館の公開といった形で絵画データベースが提供されていることが多い。多くの場合、テキスト型画像データベースが用いられるが、オリジナルの絵画の劣化・異常を調べるという観点から、画像の特徴を抽出し、微細な変化や劣化の程度を把握する目的で補助的に利用している例もある。火事や天災など、不慮の事故でオリジナル作品が失われた場合にも、作品を復元するために画像データが用いられることがある。

電子図書館データベース [21]

電子図書館で扱う情報は、図書や雑誌文献のみならず、写真や地図、オーディオ CD やビデオテープ、DVD など多様である。そのため、基本的に全文テキストページ

と画像イメージページの双方が利用される。テキストにはデータサイズが小さく、検索用途に向いているという利点があるが、文書をデータベースに登録する際、テキスト自動読みとりのための OCR が用意されていない言語や、古文書など資料の入力が難しいものは、テキストベースの資料であっても画像イメージで保存することになる。特に画像イメージデータに関しては、ビデオ画像シーン切り出しやライブラリ化が行われており、膨大なイメージライブラリが実現可能である。しかし画像イメージデータの相互利用性を高めるためには、付随するテキストデータや、ライブラリ全体に対するメタデータなどの整備が課題となっている。

原生生物と日本産アリ類の広域画像データベース [22]

日本産アリ類カラー画像データベースは、1995 年からサービスが開始された Web 上の生物学画像データベースの草分け的存在である。日本蟻類研究会と生物情報広域データベース化研究グループによって現在もメンテナンスされている。非常に大規模なプロジェクトであり、ディレクトリサービスとして、アリ類の標本画像を提供している。マトリックス法と呼ばれる、2 次元ディレクトリサービスの一種も行っている。原生生物情報サーバは多少趣が異なっており、素材情報データベースとしての役割を担っている。つまり研究者に対して広く公開し、論文などに自由に利用してもらう目的がある。両グループのデータベースは、画像を広く公開し、アクセスを分散させるために、全国５カ所にミラーサーバを公開している。また、情報検索は手作業によるディレクトリ検索のみであるため、おおざっぱなデータを一度に入手するような用途には向いていない。

他分野における画像データベースシステムの動向をふまえて本システムの構成を考えると、蛍光染色体画像における画像データは、テキスト型画像データベースと内容型画像データベースの２つの機能を合わせて持つことが有用であると考えられる。染色体画像は、それぞれが非常に似通っていることが多いため、画像処理技術によって得られたデータも類似する可能性が高い。そこで、類似画像検索を行う際に補助的なデータとして、テキスト検索データを利用することが必要であり、データフォーマット構造を明確にする必要がある。そのため、これら画像データとテキストデータ両方の情報を効率よく管理するために、フォーマット化されたデータの管理に適したシ

ステムを構築する必要がある。

4.2　ＤＢＭＳの利用

FISH法による蛍光顕微鏡写真や核型分析によって整列した染色体画像を、効率よく管理・蓄積および検索するために、DBMS（データベース・マネジメント・システム）をシステムのバックエンドサーバとして利用する[23]。DBMSはデータ、データ操作、データ定義、データベース管理機能を含んだシステムであり、フォーマット化されたデータの共有・検索を行うことができる。

一般にDBMSを利用することには以下のような利点がある。

- **処理の基盤となる確固としたデータモデルが存在する。**
- **構造を変えても以前のアプリケーションがそのまま動く（データの独立性）**
- **データ操作が迅速に行える**
- **同時に複数のユーザが利用しても正常に動作する（並行処理制御とトランザクション機能）**
- **データベースが正常な状態か否か（一貫性を保っているか）チェックできる**
- **ユーザのセキュリティ確保のためアクセス権管理ができる**
- **障害が発生しても少ない手間で障害前のデータベース状態に戻せる**

DBMSは、データ操作が正しく行えるということが何よりも重要である。そのため、DBMSはデータモデルという数学的に定式化されたモデルに基づいてデータ操作を行い、そのデータモデルに沿った形でデータベースにデータを格納していく。データモデルでは次のことが決められている。

- **データが、データベース上でどのようなデータ構造をしているか**
- **データに対してどのようなデータ操作ができるか**
- **データを正しい状態に保つためにどのようなルールを定義するか**

代表的なデータ構造は、木構造、ネットワーク構造、表形式の３種類である。DBMSのうちもっとも広く利用されているRDBMSは、リレーショナル・モデルに基づいた DBMS である。リレーショナル・モデルとは、データをすべて2次元の表形式で表現するモデルであり、SQL と呼ばれる問い合わせ言語による操作が可能である。SQLは、データ操作文（DML: Data Management Language）およびデータ定義文（DDL: Data Definition Language）の両方を含む言語である。リレーショナル・モデルでは、すべてのデータは2次元の表（table）で表す。表は複数の属性（attribute）から構成され、その属性値によってその特性を表している。属性はフィールド（field）とも呼ばれる。このような2次元の表に対して、画像イメージとそれに付随する情報を割り付け、SQL検索で利用できるようにする。

4.3　ＯＤＢＣ(Open Database Connectivity)

ODBC（Open Database Connectivity）は、米国Microsoft社が中心となって策定された、Ｃ言語でDBMSに接続し運用するためのインタフェースである[24]。特に計算機上でのDBMSクライアントアプリケーションを作成する際、API(Application Programing Interface)として広く使われている。

ODBCは、相互運用性を最大限に活用することを目標に設計されている。従来のクライアント／サーバモデルにおいては、[図4.1]の通り、一種類のDBMSに対して、そのDBMS専用のクライアントアプリケーションを作成する必要があった。これは、各DBMSベンダ間での取り決めが存在せず、それぞれのDBMSに接続して利用する際の手続きや、利用されているデータフォーマット、返値などが統一されていなかったためである。そのため、アプリケーションベンダは各DBMSに特化した実装を行わなければならなかった。しかし、計算機の利用が進み、より多くのハードウェアとソフトウェアが利用できるようになると、DBMSにも安いもの、最高速のもの、知名度が高いもの、市場で最新のもの、スタンドアロン環境で最良のものなど、様々な特徴を持ったものが現れた。それに伴い、専用のアプリケーションが大量に作成されることになり、互換性の無いさまざまなデータが、多種多様なシステムによってアクセスされ運用されることになった。また、アプリケーションの開発、保守と

いった観点から見ても、使用するDBMSごとに別個のデータアクセスルーチンを作成する必要があり、無駄が多かった。

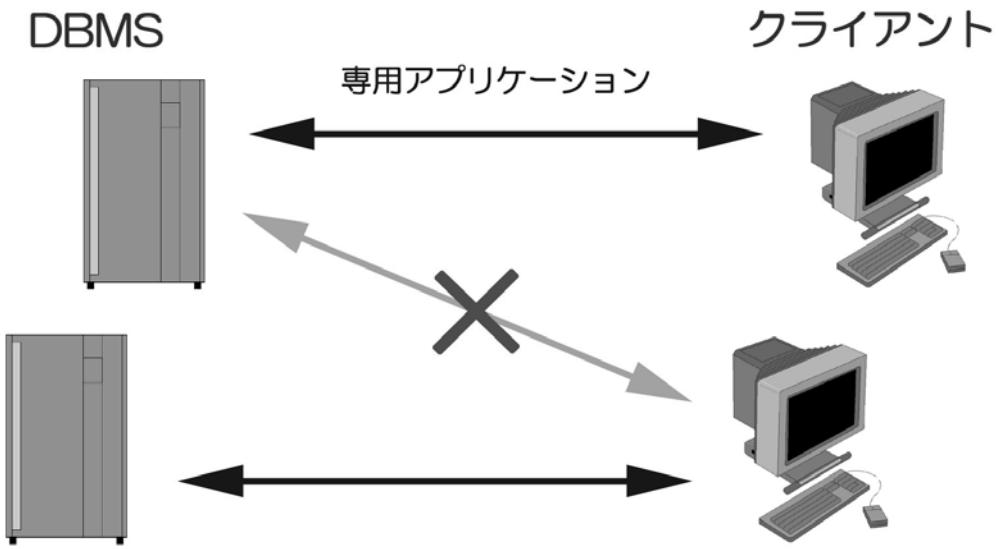

図4.1　従来のDBMSのモデル

それぞれのサーバ(DBMS)には専用アプリケーションしかアクセスできない

このような理由により一つのアプリケーションで複数のDBMSのデータをマージする方法、つまりDBMSに依存しない単一のアプリケーションを作成できる方法が必要とされていた。それに合致する解が、APIとしてのODBCである。アプリケーションをODBC準拠で開発することにより、[図4.2]のようにODBC準拠DBMS間で、クライアントを相互運用することが可能になった。ODBCはデータベースAPIの一つであり、個々のDBMSやOS、言語などに依存しない。現在の最新バージョンはODBC3.0であり、X/OpenやISO/IECといった国際規格標準に完全に準拠している。Microsoft Windows OSはもとより、Mac OSやUNIX等、OSを選ばないのも特徴であるが、特にWindows OSでは標準で各種DBMS対応ドライバが組み込まれているため、利用価値が高い。

現在、主なODBC対応DBMSとして次のようなものがある。商用DBMSとしてはオラクル社のORACLE、IBM社のDB2、Microsoft社のSQL Server、サイベース社のSybaseが代表的なものとして挙げられる。また、フリーソフトウェアの

PostgreSQL[25]が近年大きく支持を集めている。

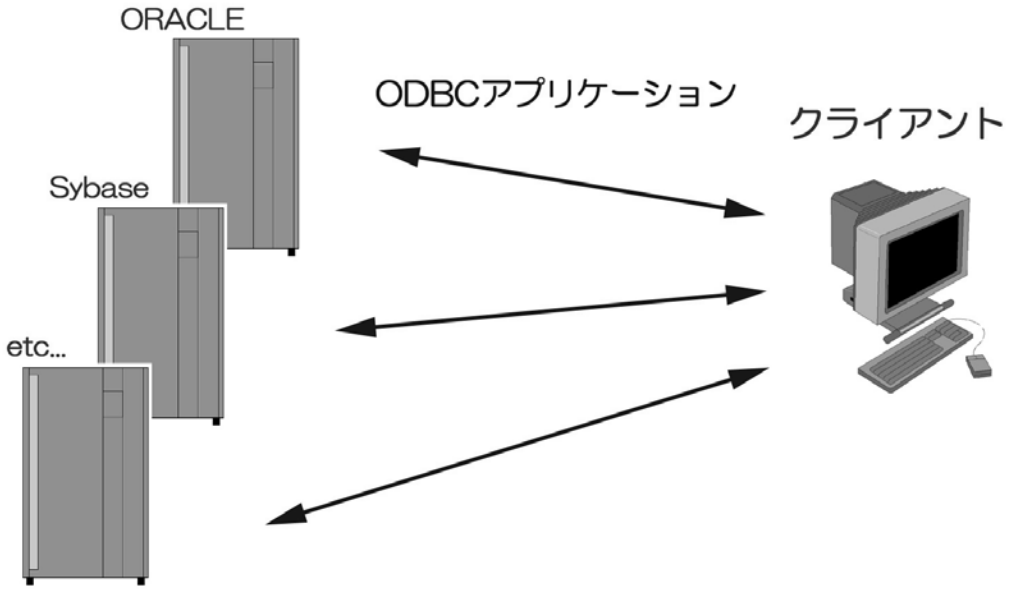

図4.2　ODBC対応DBMSのモデル

一つのアプリケーションで多数のDBMSに接続ができる

4.4　特徴量抽出(Feature Extraction)

上記のようなSQLによる検索は、テキスト型検索に分類される。画像データベースにテキスト型検索を利用した場合、登録されている画像の内容を把握していないと効率の良い検索を行うことができないという欠点がある。そこで、画像内容に基づく類似画像検索を合わせて行うことによって、初めてシステムに触れる研究者にも直感的でわかりやすい検索を行うことができる。

画像処理的なアプローチとしての類似画像検索は、画像の特徴を抽出するフェーズと、画像の特徴を用いて実際に画像を検索するフェーズに大別することができる。特徴量抽出の代表的な手法として、マスク画像を用いた特徴量抽出[26][27]と、Waveletを用いた特徴量抽出[28][29]を挙げることができる。マスク画像を用いた特徴量抽出では画像の特徴量をベクトルで抽出するのに対し、Waveletでは波形で抽出するという違いがある。Wavelet方式は比較的正確な特徴量が抽出できるとされ

ているが、計算量が膨大であることが欠点である。染色体画像で類似画像検索を行う場合、検索対象と考えられる母数が多く、計算量が大きい場合実用的では無くなると考えられる。本システムでは、特徴量抽出の手法として大津らが[26]で提案したＮ次自己相関マスクによる積分型特徴量を用いた。

ここで用いる積分型特徴量[30]とは、画像の一部分に注目してその特徴を数値ベクトル化し、その特徴ベクトルの画像全域における値を足し合わせたものをその画像の特徴量とする手法である。画像の一部分γから抽出される特徴量をχ（γ）とすると、画像全域に割り当てられる特徴量は以下の[式4.1]で表される。

$$4.1) \quad \chi = \sum_{\gamma} \chi(\gamma)$$

積分型特徴量では、大局的な空間構造を反映することが原理的に困難である。しかし、アルゴリズムが比較的単純であり、少ない計算量で実装できるため、現実的なシステム構築の要素技術として有用である。また、部分特徴量の抽出方法の変更によって、多種多様な画像群の検索に柔軟に対応可能であり、無作為な画像データを取り扱うという観点からも有益である。

特徴を抽出するために、まずγに対してＮ次数からなるマスクを適用し、それぞれの画素から数値を得る必要がある。ここでＮ次自己相関マスクに必要なドット数は、次数＋１点である。マスクのドット数を３×３、次数を２次までとすると、０次（ドット数１）では１個、１次（ドット数２）では４個、２次（ドット数３）では２０個、合計２５個の平行移動に対して不変なマスクが得られる[図4.3]。これを[31]にならい大津完全セットと定義する。対象画像全体に対してそれぞれのマスクを掛け、ドットの値を積算し、画像全体において和算した値がそのマスクにおける特徴ベクトルとなる。そのため、３×３、２次の自己相関マスクにおいては、２５次の独立な特徴ベクトルが得られる。２次以上のマスクをランダムに算出する手法[31]などもあるが、これら高次マスクは結局のところ大津完全セットのマスク同士の組み合わせで表現できるため、２５次の特徴ベクトルは必要十分なものである。

よって、大津完全セットを利用することにより、最小の計算量で画像の特徴量を過不足無く抽出することができる[32]。

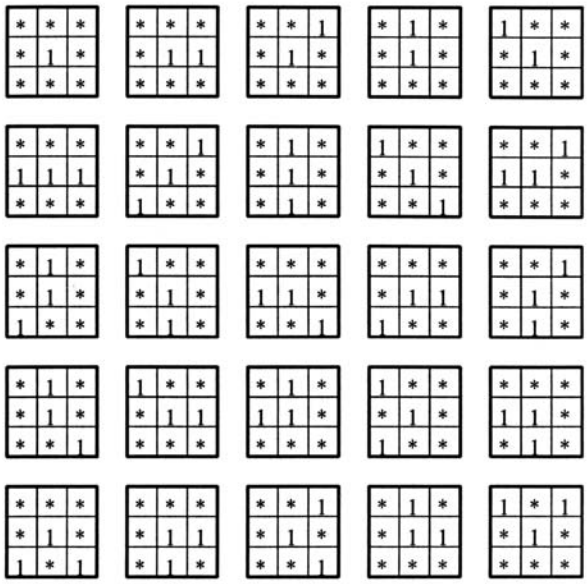

図4.3　25個の平行移動に対して不変なマスク(大津完全セット)

画像間の類似性を計る手段としては、画像の特徴ベクトル間での内積を利用し、それを各ベクトルの長さで正規化したものを用いた。これによって、特徴ベクトル同士の成す角度が求められる。たとえば、画像ａと画像ｂの特徴ベクトル同士が成す角度は、以下の[式4.2]によって与えられる。

$$4.2)\quad \cos\theta = \frac{< a \mid b >}{\| a \| \| b \|}$$

しかし、実際のインプリメントにおいては、計算上角度を正確に算出する必要はないため、分子となるθの値をそのまま比較すれば良い。このとき、同一画像において類似度θが１となり、１に近似しているほど特徴量が類似した画像であるといえる。そのため複数画像での類似度を比較し、降り順にソートすれば、画像が類似している順に画像集合をソートすることができる。

先に述べたように、情報科学分野における類似画像検索手法では、電子図書館や美術館画像分類のように、多様な画像集合を前提として、個々の画像から特徴となる要素を抽出して画像検索のキーとすることが多い。これに対し、染色体画像は生物種に関わらず非常に似通った偏った画像集合となるため、このような手法は効率が悪く、また精度にも欠けることが予想される。このため、類似画像検索のみで検索システムを構築するのではなく、既出のSQLによる、ファイル作成日時や生物種などの情報も画像検索と合わせて利用することで、検索精度の向上を図るべきである。

第　5　章
ユーザインタフェースと操作

5.1　起動前の設定

この章では、実際に作成したシステムのインタフェースと操作を中心に解説する。検証に用いた計算機の仕様は、以下の通りである。

DBMSサーバ (IBM PC/AT互換機)

- OS　Vine Linux 1.1 (Kernel 2.2.14)
- プロセサ　Intel P6系 400MHz Dual
- 主記憶　256MB SDRAM

ODBCクライアント (IBM PC/AT互換機)

- OS　Microsoft Windows98
- プロセサ　Intel P6系 400MHz
- 主記憶　128MB SDRAM

サーバとクライアントは、構内LANに100Base-TXで接続されている。また、クライアントのアプリケーションは、インプライズ社のBorland C++ Builder4で作成された。Microsoft ODBC 3.0を使用している。

次に[図5.1]にシステムの構成図を示す。システムのクライアントとサーバは、前章で説明したODBCを介して接続される。それぞれの機能は明確に切り分けられており、

特にサーバ側の機能は、ODBC対応DBMSであれば、種類を選ばずに動作させることができる。

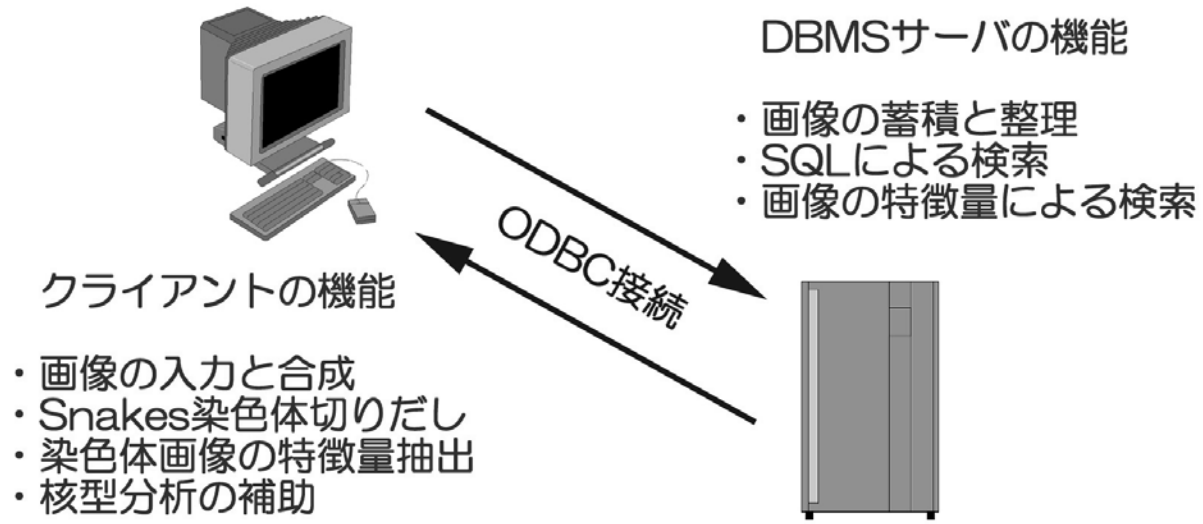

図 5.1　システム構成図

5.1.1　ＤＢＭＳ（ＰｏｓｇｒｅＳＱＬ）の設定

前章で触れた通り、DBMSには多種多様なプロダクトがあり、それぞれ一長一短の特徴を持っている。今回のシステムを構築するにあたり、DBMSとしてPostgreSQL[25]を利用した。PostgreSQLには以下のような特徴がある[33]。

- **SQL92のサポート**
- **無償で利用可能であり完全なソースコードを公開している**
- **UNIXを中心に数多くのプラットホームで稼働する**
- **多様なプログラミングインタフェースをサポートしている**
- **クライアント/サーバアーキテクチャ**
- **日本語化/国際化対応**
- **ODBC対応**

最新版のPostgreSQLは以下のURLから入手できる。

- ftp://ftp.jaist.ac.jp/pub/dbms/PostgreSQL/

まず、PostgreSQL管理用に、Linuxのroot権限でユーザアカウントpostgresを作成する。このLinuxユーザであるpostgresは、PostgreSQLのスーパーユーザとなる。次に、ディレクトリ/usr/local/pgsqlをオーナーpostgresで作成する。そしてソースコードを入手し、postgresアカウントで展開・コンパイル後にmake installすると、その実体は/usr/local/pgsqlに置かれる。その後コマンドサーチパスと環境変数を登録するため、postgresユーザアカウントの.bashrcに下記の記述を挿入する。

```
PATH="$PATH":/usr/local/pgsql/binexport POSTGRES_HOME=/usr/local/pgsql
export PGLIB=$POSTGRES_HOME/lib
export PGDATA=$POSTGRES_HOME/data
export MANPATH="$MANPATH":$POSTGRES_HOME/man
export LD_LIBRARY_PATH="$LD_LIBRARY_PATH":"$PGLIB"
```

変更を反映するために、source ~/.bashrcを実行して、ユーザアカウントの設定は終了である。その後、データベースを初期化するために、initdbを実行する。初期状態では他ホストから接続できないため、サーバとして利用するために/usr/local/pgsql/data/pg_hba.confに以下の記述を追加する。

```
host          all          0.0.0.0         0.0.0.0           trust
```

これは全てのアクセスを許容する設定であるが、IPアドレス毎に設定することも可能である。最後にpostmaster -S -i -o "-F"を起動して、PostgreSQLを起動する。サーバが再起動を行った際に自動的にpostmasterを起動するためには、/etc/rc.d/rc.localに以下の記述を追加する。

```
POSTGRESDIR=/usr/local/pgsql
if [ -x $POSTGRESDIR/bin/postmaster -a -d $POSTGRESDIR/data ]; then
        rm -f /tmp/.s.PGSQL.5432
        su - postgres -c "postmaster -S -i -o ¥"-F¥""
```

```
        echo -n 'postmaster '
fi
```

以上でDBMSの初期設定は終了である。

以上の設定を行った後、今回のシステムで利用するデータベース構築のため、postgresアカウントでcreateuserコマンドを使い、postgresの利用ユーザを設定した。利用ユーザのアカウント名とパスワードは、以降の設定全般に利用するため、重要である。利用ユーザ名でログインした後、createdb“データベース名”コマンドでデータベースを作成、コマンドライン上でのPostgreSQLフロントエンドであるpsqlで作成したデータベースにアクセスする。そしてシステムで利用するスキーマ定義を、以下のようにする。

- FISH画像とそれに付随する実験データに関するスキーマ設定。テーブル名odbca。filename：染色体画像名，date：作業日時，chromosomes：染色体数，height：画像の高さ，width：画像の幅，species：生物種，method：実験手法，rfilename：Rフレーム画像名，gfilename：Gフレーム画像名，bfilename：Bフレーム画像名，image：染色体画像バイナリ.
- 個々の染色体に対するスキーマ設定。テーブル名odbcb。

filename：染色体切りだし元画像名，place：染色体順番，feature0～24：特徴量ベクトル，edge0～47：染色体輪郭座標，iamge：染色体切り出し画像バイナリ，mask：染色体輪郭画像バイナリ.

以上をデータベースに作成すれば、サーバ側となるDBMSの設定は終了である。

5.1.2 ＯＤＢＣクライアントの設定

まず、Windows98に対してPostgreSQL用ODBCドライバを導入する。最新版のPostgreSQL ODBC Driverは以下のURLから入手できる。

- ftp://ftp.jaist.ac.jp/pub/dbms/PostgreSQL/odbc/

Microsoft Windows98では、ODBCに関する設定をコントロールパネルから行うことができる[図5.2]。コントロールパネルのODBCデータソースを開き、ユーザDNSの項目で追加のボタンを押し、PostgreSQLを選択、追加する。そして前章で設定した利用ユーザのアカウントとパスワード、データベース名を設定し、DataSourceのReadOnlyのチェックボックスを外す。

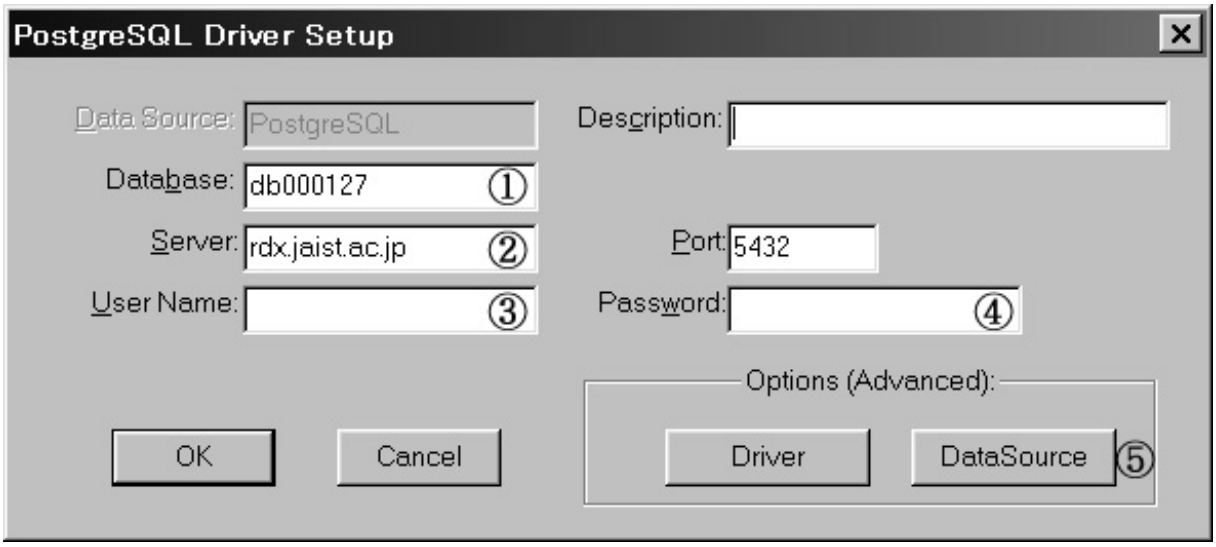

図5.2 PostgreSQL ODBC Driver 設定画面

①データベース名　②サーバ名　③PostgreSQLユーザ名(空欄可)

④ユーザパスワード(空欄可)　⑤オプション欄(ReadOnlyをoffに)

ODBCの設定終了後、クライアントアプリケーションをインストールする。以上でWindows側の設定は終了である。全ての作業終了後、インストールディレクトリからクライアントアプリケーションを起動する。

5.2　核型分析の操作

5.2.1　ユーザインタフェース

起動した直後のアプリケーション画面は次の通りである[図5.3]。基本的にメインウィンドウと３つのサブウィンドウから構成されており、それぞれRGB Control

Window、ODBC Control Window、Karyotype Analysis Windowである。それぞれ標準的なWindowsアプリケーションと同様、マウスとキーボードによる操作体系を持つ。前章でPostgreSQLユーザ名とユーザパスワードを空欄にした場合、起動時にそれらの入力が要求される。

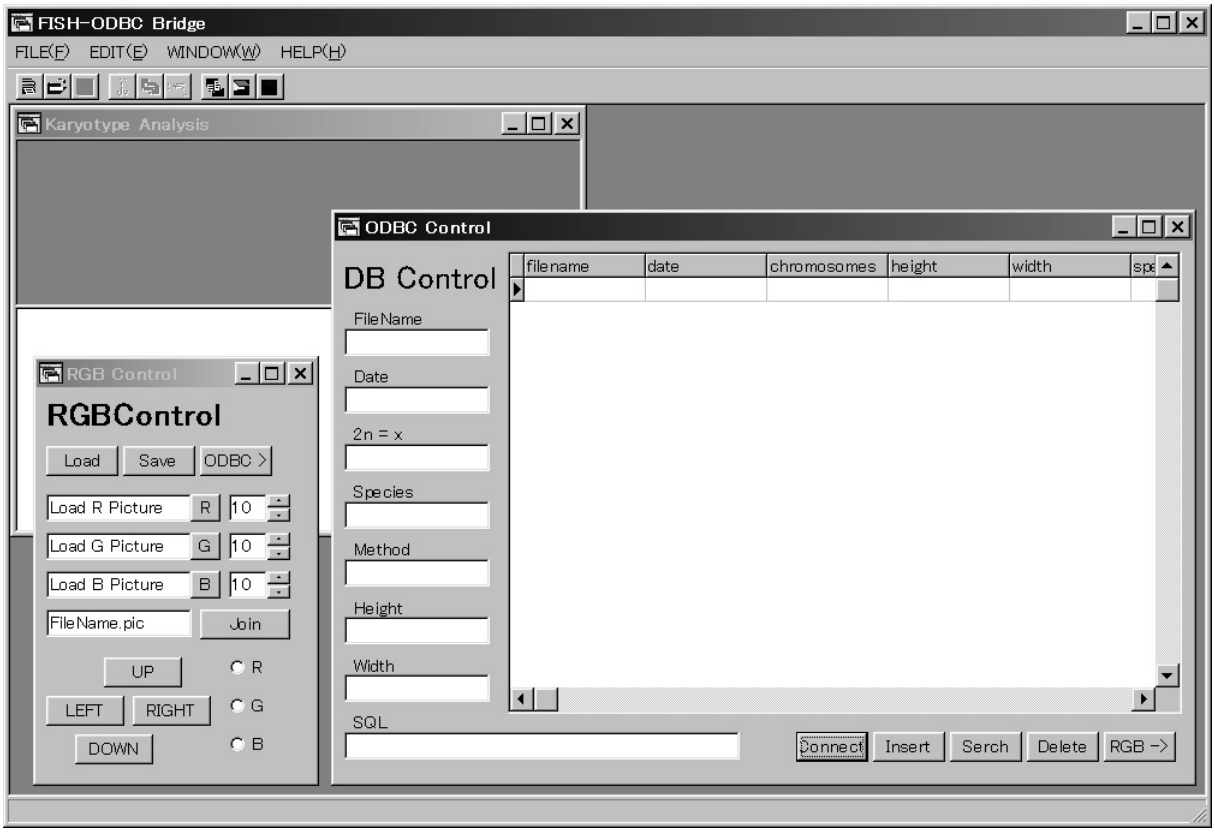

図5.3　起動直後のクライアント

5.2.2　RGB 合成

RGBの合成は、RGB Control Windowで行う[図5.4]。Load x Pictureと表示されているボックスの右にあるR、G、B各ボタンを押し、それぞれ読み込むべきグレイスケール画像を決定する。次にJoinボタンを押すと、合成した結果が新しいウィンドウに表示される。しかし、第２章で解説したように、試料作成時に細胞質がゴミとなって残った場合、バックグラウンドにそれらが残ってしまう。これを取り除くには、R、G、B各ボタンの右にある数値を0～256の範囲で変化させる。このオプションは、しきい値を用いたノイズ除去を行い、入力した数値以下の値を持つ画素を

切り落とす。また、蛍光顕微鏡からの画像取込時にRGBがずれ、合成時に染色体と蛍光シグナルがうまく重ならないような場合がある。そのときは、ずれを補正したいプレーンを下部右にあるラジオボタンで選択し、十字ボタンで上下左右に修正する。

また、すでに合成済みのフルカラー画像を使用したいときは、画面上部の Load ボタンを使用することで、新規ウィンドウに既存画像を呼び出し、同様に以降の処理を行うことができる。逆に、合成画像をこの時点で保存したい場合には、Save ボタンを使用すれば、クライアント側に保存することができる。

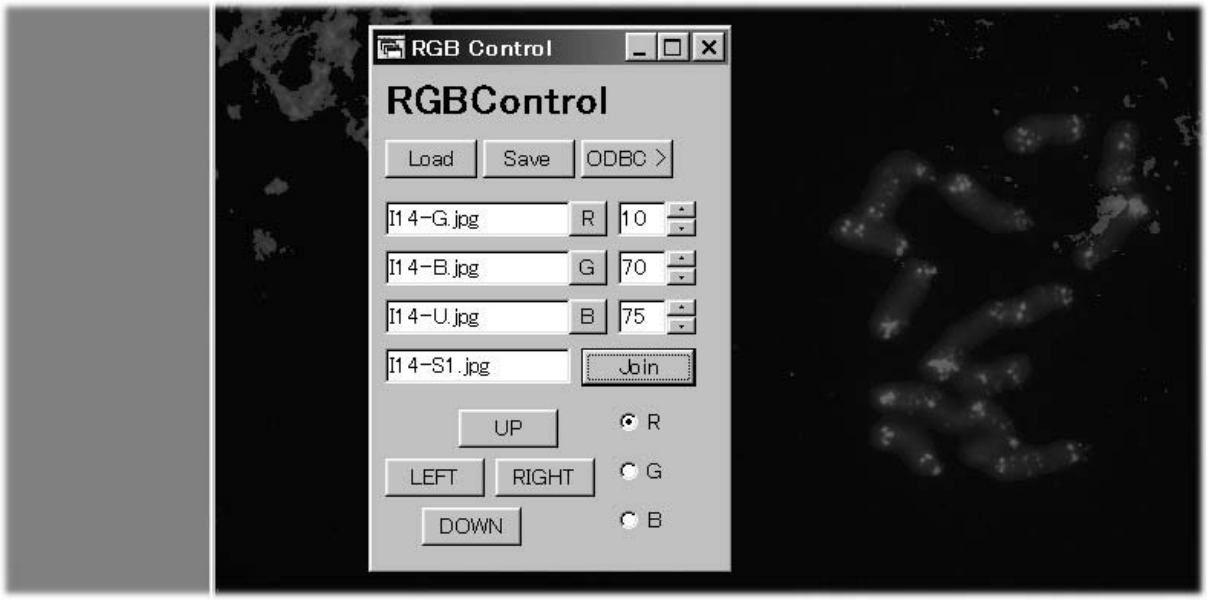

図5.4　RGB Control Windowによる画像合成と加工

5.2.3　染色体切り出しと並び替え

染色体画像の読み出し、合成が終われば、次は染色体を一本一本切り出す作業に入る。切り出したい染色体の中央にマウスカーソルを置き、マウスの右ボタンを押すことで、染色体の周りに初期閉曲線が円状に描画される。初期閉曲線はマウスの動きに合わせて収縮するので、染色体の大きさに合わせてボタンを離すと、座標位置と作業開始のダイアログが表示される。はい(Y)を選択すれば、Snakesによる輪郭線の抽出が行われる[図5.5]。また、染色体画像の切り出しと同時に、特徴量検索で利用する、染色体の特徴量ベクトルも抽出している。切り出しおよび特徴量抽出

のの進行状況は、メインウインドウのタイトルバーにパーセンテージで表示される。

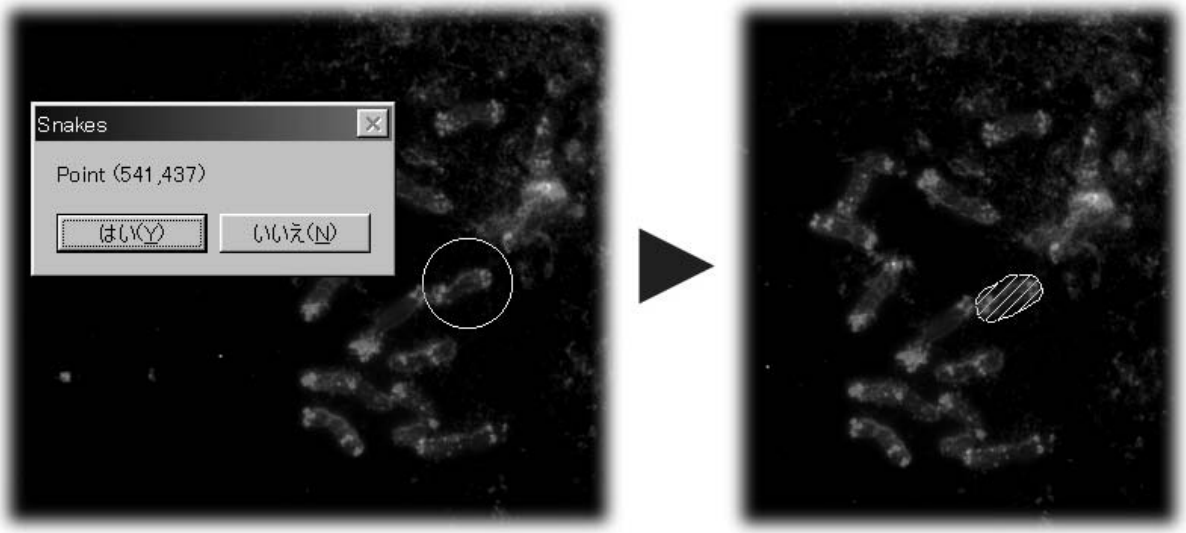

図5.5　Snakesによる染色体切り出し

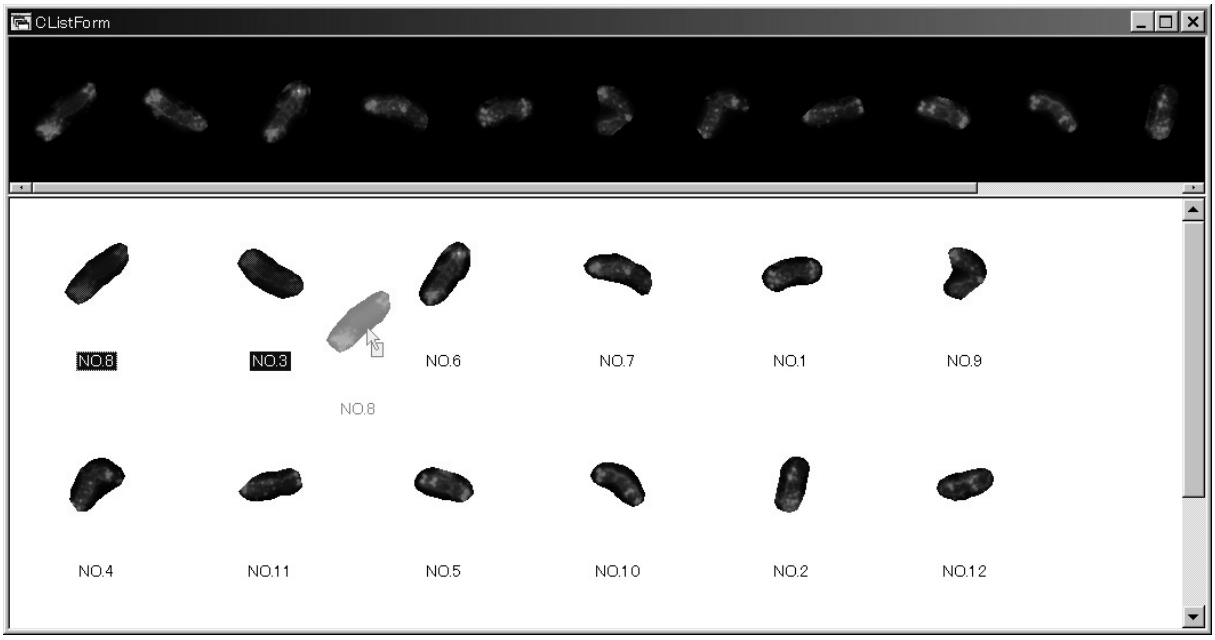

図5.6　Karyotype Analysis Windowでの染色体の並び替え

上部に並び替えた染色体が表示される。下部は並び替えのためのアイコン群。

これを染色体の本数分繰り返すことで、全ての染色体を切り出す。切り出した染色体は、Karyotype Analysis Windowに表示される[図5.6]。Karyotype Analysis Windowでは、全ての染色体に対してその座標、特徴量、切り出した順番がクライアント側に保持されているため、染色体抽出がうまくいかなかった場合でも、左クリックでキャンセル可能である。再試行の回数の制限などは存在せず、最初まで戻ってやり直すこともできる。切り出した染色体は、染色体の長さ順に並び替える。下部の作業スペースにアイコン化された染色体が並ぶ。それをドラッグ＆ドロップ方式で入れ替える。入れ替えた結果は即時にKaryotype Analysis Window上部に反映される。

染色体の入れ替えが終われば、RGB Control WindowのODBCボタンを押して、ODBC Control Windowを呼び出す。

5.2.4　項目設定とアップロード

ODBC Control Windowに移ると、左側部分のボックスに、画像から自動的に取得された情報があらかじめ入力されている[図5.7]。このため、修正の必要がある項目のみ、ボックス内を書き換えることができる。その後、下部のInsertボタンを押せば、ネットワーク経由でDBMSサーバにデータが送られ、蓄積される。

蓄積されたデータはODBC Control Window内の右部の表[図5.7の⑨]に表示される。表の項目は、DBMSが保持しているデータと同一であり、表の値を改変することで、直接データを編集することもできる。データベースからのデータ消去は、表の項目を選択し、Deleteボタンを押すことによって行われ、染色体画像などがすべてサーバから消去される。

また、表の項目を選択・ダブルクリックすることにより、登録されているデータをクライアント側に取得することができる。このとき、画像や入力されている項目、染色体並び順や座標情報などは、アンドゥ用の情報も含めすべてRGB Control WindowおよびKaryotype Analysis Windowに書き戻される。そのため、一度登録したデータを読み戻したときに、染色体抽出や核型分析をもう一度やり直し、再登

録をすることも可能である。

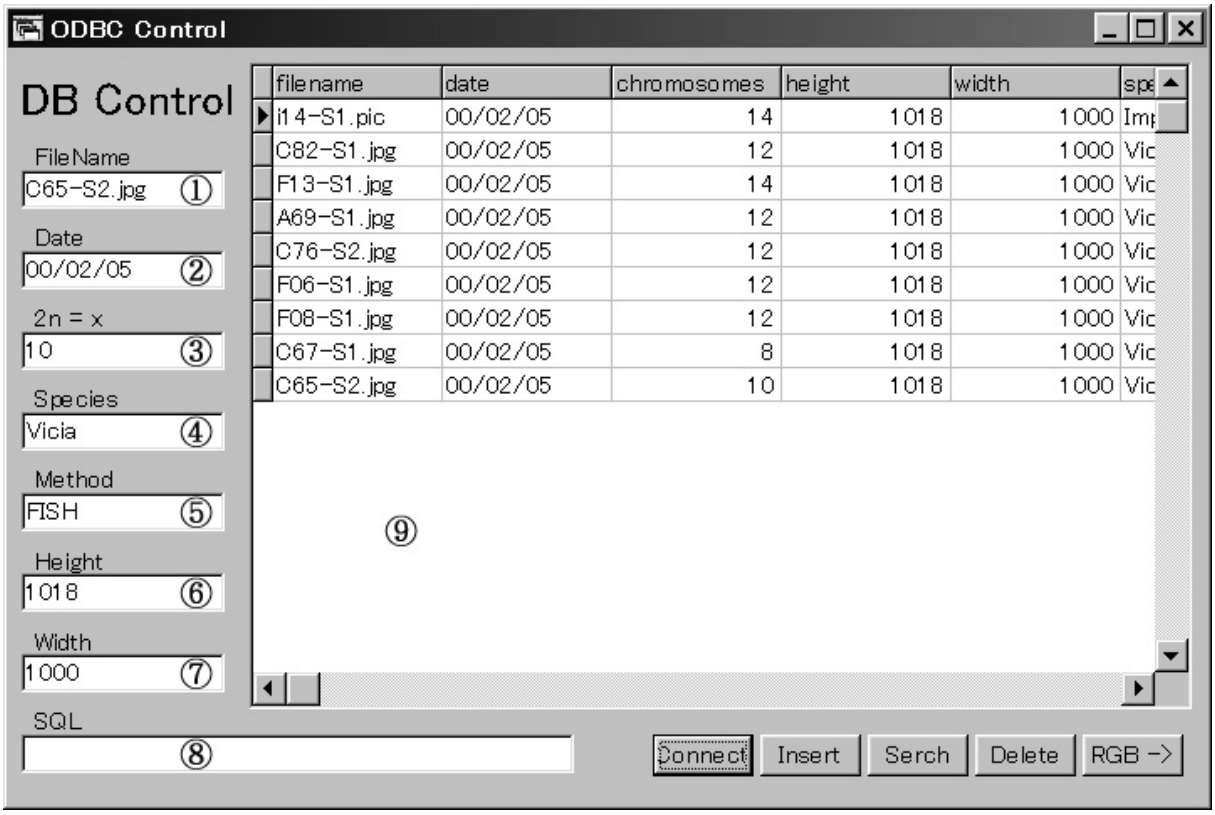

図5.7　ODBC Control Window

①画像ファイル名 ②登録日時 ③染色体数 ④生物種名 ⑤実験手法

⑥画像高さ ⑦画像幅 ⑧SQL直接入力(検索時のみ) ⑨結果表示窓

5.3　情報検索

5.3.1　ＳＱＬを利用した検索

ODBC Control Window[図5.7]は、画像情報の登録用インタフェースとしての役割の他に、本システムの検索機能を使うためのSQL検索インタフェースとしての機能も持つ。登録時に使用したボックスに、検索条件を書き入れ、Searchボタンを押すと、SQLによる検索が行われる。検索された結果は、右の表に表示される。検索条件は、画像ファイル名や生物種名などの文字項目については完全一致、登録日時や染色体数などの数値項目については範囲指定と完全一致の両方が利用できる。そ

れぞれの項目はand条件である。また、SQLボックスに直接SQL問い合わせ文を入力することにより、柔軟な検索を行うことができる。

5.3.2　画像特徴量を利用した検索

画像を処理していく過程で、染色体を切り出したときに、過去の類似した染色体と比較検討する必要にせまられることがある。そのような場合、あらかじめ抽出している特徴量から、類似画像を検索することが有効である。類似した染色体を探したい場合、Karyotype Analysis Windowにおいて、下部の染色体アイコンをダブルクリックする。DBMSに登録されている染色体画像集合から、画像特徴量的が近似しているものを探すことができる。検索結果はODBC Control Windowの表として表示される[図5.8]。

Filename	Feature	Place
i14-S1.pic	0.99999999999274	10
i14-S1.pic	0.999999999992289	0
i14-S1.pic	0.999999999992104	11
i14-S1.pic	0.99999999999164	3
i14-S1.pic	0.999999999991383	9
F13-S1.jpg	0.999999999991225	5
i14-S1.pic	0.999999999990212	12
i14-S1.pic	0.999999999989967	13
F13-S1.jpg	0.99999999998953	4
A73-S2.jpg	0.999999999987981	7
A69-S1.jpg	0.999999999985625	7
i14-S1.pic	0.999999999985548	1
i14-S1.pic	0.99999999998539	2
A73-S2.jpg	0.999999999984687	3
A73-S2.jpg	0.999999999984404	2
i14-S1.pic	0.999999999984093	4
F13-S1.jpg	0.999999999984091	3

図 5.8　特徴量検索

Filename：画像ファイル名　Feature：特徴量近似値　Place：染色体順番

表の上から類似していると考えられる順にソートされている。特徴量近似値が、大きいほど類似した画像である。画像ファイル名と染色体順番から類似した染色体を含む画像ファイルを特定し、通常の SQL 検索で呼び出すことによって、クライアント

側に必要な画像を取り込むことができる。特徴量検索は全ての染色体について特徴量の近似を計算するため、母集団の大きさによって所要時間が変化する。

第　6　章
まとめ

6.1　既存システムとの比較

本システムを実際に運用し、既存の画像解析システムと比較するために、サンプルとしてFISH法の染色体画像を100枚、システムに投入した。検討の結果、本システムには次のような利点があると考えられる。

- **FISH法などで得られた染色体画像の操作に適応している**

既存のアプリケーションを用いた場合、染色体解析、もしくは蛍光染色解析のどちらかに対応したアプリケーションは存在してもFISH法やGISH法のような、染色体の多色蛍光染色画像に特化したアプリケーションは存在していなかった。

- **画像の投入・加工から蓄積・検索まで、流れ作業的に行うことができる。**

ユーザとなる研究者は、全ての作業をクライアント側のアプリケーションから行う。そのため、バックエンドとなるサーバの存在を気にする必要はない。また、データの蓄積に関してもサーバ側で管理されるため、データファイル管理や、それに付随する実験データ情報などの整理からも解放される。

- **ODBC対応により、柔軟なシステム構成を取ることができる。**

ODBCに対応することによって、クライアントのアプリケーションとバックエンドのサーバは、完全に独立した構成となっている。そのため、データベースの規模や利用状況によって、柔軟なシステムを組むことができる。また、DBMSを利用す

ることにより、画像データベースとしての堅牢性が確保されている。

- **特徴量検索を行い、直感的に類似染色体を探すことができる。**

特徴量検索を行うことにより、データベース内部に蓄積されたデータを把握していなくても、直感的な検索を行うことが可能である。これは、初めての利用者に対してもわかりやすい検索法であり、システム導入や切り替えに際して、手間が省けることにもなる。本システムを利用する研究室内での知識共有にとって非常に有効である。

- **レスポンスが速く、スケーラビリティが高い。**

半手動登録でデータ登録を行ったため、染色体画像を100枚入力するのに約20分ほどかかった。しかし、クライアント側の処理もサーバ側の処理もレスポンスが速く、ストレスを感じることは全くなかった。また、データをDBMSに格納することにより、画像の枚数が増加してもレスポンスが落ちるようなことはなかった。よって高いスケーラビリティが期待できる。

- **フリーソフトウェアであるため、導入コストを考えないで済む。**

既存の画像処理アプリケーションは、高価な実験機材と同時に納入されることが多く、同時に価格も非常に高いものであった。本システムはフリーソフトウェアであるため、コストを気にせずに必要な数導入することができる。また近年になって、蛍光顕微鏡やCCDカメラなどの実験機材に安価なものが出現してきているため、本システムと組み合わせることにより、非常に安価な実験系を作り出すことが可能である。

6.2 検索結果についての考察

本システムにおいては、テキスト型検索であるSQLを利用した検索と、画像内容に基づく検索である類似画像検索の２種類の検索方式が実装されている。これら２方式を用いた場合の検索結果について考察を行う。

SQLでの検索については、一つのテーブルに対してSQLのSELECT文を送出している。ネットワーク経由であっても、クライアント側が送出するデータはSELECT文のみであり、後はDBMS側で検索を行った後、検索結果であるテーブルを受け取るだけであるから、トラフィックの量は非常に小さい。また、PostgreSQLは画像バイナリをテーブル構造の外に持つため、検索にかかる時間はテキスト・数値のみで構成されたテーブルと同等である。従って登録画像の量が大きくなっても、実用的な時間内で応答が返ってくるため、DBMSの負荷も小さいと思われる。よって現状のシステムで十分実用的であると考えられる。

画像検索に関しては、前章5.3.2のように行う。しかしながら、検索結果は必ずしも有効とはいえなかった。たとえば[図5.8]の場合、キーとして用いた染色体と同一画像i-14S1.pic内の他の染色体が優先されてしまい、他の原画像内のより似た染色体よりも、同一画像内のあまり似ていない染色体を取ってきてしまう傾向があることがわかった。このような事例に対して、原因を検討する。

最も考えられるのは、特徴量抽出に関して特定のファクターが強く効いているということである。元々、FISH法の場合、同じ実験によって得られた画像であっても、蛍光染色の染まり具合やノイズとなるゴミの残留量、蛍光顕微鏡の光量によって、色の濃さが大きく変化する。そのため、どうしても同一の染色体画像から切り出した染色体画像は、特徴量が似通ってしまう。加えて、今回利用したアルゴリズムは、元来テキスト形状などの２値画像を対象として発展してきたアルゴリズムである。２値画像から特徴量抽出をする場合、画像の形状の他に、画像に画素が存在するかどうかを重視する。つまり対象画像内の面積に対して、大きなファクターを持つという特徴がある。それをそのまま多色画像に当てはめた場合、画像形状よりも画像面積に対する比重が大きく取り上げられてしまう、という結果になる。このような理由により、画面全体の輝度が類似している同一画像由来の染色体の類似度が上がってしまうことが考えられる。

以上のように、誰の目から見ても正確に類似した画像を提示するといった使い方は難しい。しかしながら、数多くの染色体画像の中から大まかに類似している画像

を探すという点においては役に立つため、類似画像検索を利用する価値は存在する。

6.3 今後の展望

本システムを作成することによって、当初の目的である染色体画像の整理・分類および検索を支援するシステムを提案することができた。本システムをさらに使いやすいものに発展させるためには、以下のような機能の改良・追加を行うことが考えられる。これにより、細胞生物学の研究者により受け入れやすいものとなることが期待される。

- **核型分析機能の強化**

本システムでは、核型分析のための染色体切り出しに際して、Snakesを利用している。Snakesは輪郭抽出を確実に行うことができるアルゴリズムであるが、実験材料の状態が悪く、染色体画像に大きなノイズが含まれていたり、染色体が複雑に絡み合っている場合にはうまく働かない。そのため、手動で切り出す機能や、他のアルゴリズムを合わせて用いるなどの改良が望ましい。また、切り出し後の染色体並び替えについて、中心線抽出や染色体の長さ、短腕と長腕の比率などを自動検出し、ある程度自動的に並び替えを行うことを考えなければならない。また、並び替えた染色体を、短腕を上部に回転させ、比較検討しやすいようにする工夫も必要である。その他、核型分析後の画像をプリンタなどに出力可能なフォーマットに整形するといった工夫も考えられる。

- **類似画像検索機能の強化**

類似画像検索については、現状では、アルゴリズムに起因する精度の問題がある。これを改善するためには、以下のような方策が考えられる。まず、現状では画像の輝度値から特徴量抽出を行っているので、これをRGBの各値から求めて特徴量の値を３セット持たせる手法がある。また、検索を行う際に、R、G、Bそれぞれのプレーンついて、ベクトルの正規化を行うことも有効と考えられる。その他、複数の特徴量抽出アルゴリズムを利用して、別の種類の特徴量で２重化することも考えられる。しかし、これらの手法を用いても、一回のみの画像検索で常に最良の結果が得

られるとは限らない。なぜなら、大量のデータを扱うような場合には、特徴量は類似していても視覚的に類似していない、例外と考えられる画像が紛れ込んでしまう可能性があるためである。そこで、１度目の検索で大まかに分けた後に、２度目の検索を別個の特徴ベクトルで行うことにより、画像検索の精度を高める手法が望ましい。

これらの改良により、染色体解析の効率化がより進み、細胞生物学に携わっている研究者の負担を軽減することにつながると考えられる。

謝 辞

本研究を進めるにあたり、終始暖かく御指導をいただきました北陸先端科学技術大学院大学 遺伝子知識システム論講座 佐藤 賢二 助教授に厚く御礼申し上げます。

また、さまざまな面で御教授いただきました北陸先端科学技術大学院大学 遺伝子知識システム論講座 小長谷 明彦 教授に深く感謝いたします。

北陸先端科学技術大学院大学 知識システム構築論講座 櫻井 彰人 教授には、サブテーマで熱心に御指導をいただき深く感謝いたします。

また、日頃よりお世話になりました当研究室のみなさまに心より感謝いたします。

参 考 文 献

[1] A. Kawaguchi and K. Satou, A Database System for the Management and Karyotype Analysis of Chromosome Image Data, Genome Informatics 1999, No.10, pp-272-273, 1999.

[2] 米沢勝衛，向井康比己，福井希一，植物の遺伝と育種，新農学シリーズ，朝倉書店，1997.

[3] 船津高志，生命科学を拓く新しい光技術，シリーズ光が拓く生命科学７，共立出版，1999.

[4] Y Mukai and K Fukui, Condensation pattern as a new image parameter for identification of small chromosomes in plants, Jpn. J. Genet. 63, pp.359-336, 1988.

[5] 岡本美奈，FISH 法による小麦における有用遺伝子のマッピング，大阪教育大学卒業論文，pp.1-33，1997.

[6] S. N. Raina, K Yamamoto and M Murakami, Intraspecific hybridization and its bearing on chromosome evolution in *Vicia narbonensis*(*Fabaceae*), Plant Systematics and Evolution 167, pp.201-217, 1988.

[7] S. N. Raina and Y Ogihara, Ribosomal DNA repeat unit polymorphism in 49 *Vicia* species, Theor Appl Genet 90, pp.477-486, 1994.

[8] 川口昌宏，FISH 法と画像解析を用いた *Vicia* 属におけるリボゾーム RNA 遺伝子のフィジカルマッピング，大阪教育大学卒業論文，pp.1-18，1998.

[9] 加藤成二，廣瀬玉紀，秋山征夫，Carmel M. O'NEILL，福井希一，染色体画像解析システム CHIAS III 操作マニュアル，北陸農業試験場，北陸農業研究資料，No.36，1997.

[10] NIH Image Home Page, http://rsb.info.nih.gov/nih-image/

[11] IPLab Homepage，http://www.iplab.com/sos/product/gen/IPLab.html

[12] Image-Pro Plus Introduction，http://www.planetron.co.jp/ippguide.htm

[13] 大津展之，栗田多喜夫，関田巌，パターン認識 －理論と応用－，行動計量学シリーズ，朝倉書店，1996.

[14] 美濃導彦, 天野晃, Snakes: 現在・過去・未来, 信学技報, PRMU87-184, pp.81-88, 1997.

[15] 松澤悠樹，複数の動的輪郭モデルの競合による領域抽出に関する研究，北陸先端科学技術大学院大学修士論文，pp1-47，1999.

[16] 金田丘，動的輪郭モデルを用いた多方向に移動する複数物体の追跡に関する研究，北陸先端科学技術大学院大学修士論文，pp1-54，1997.

[17] 荒木昭一，横矢直和，岩佐英彦，竹村治雄，複数物体の抽出を目的とした交差判定により分裂する動的輪郭モデル，信学論 D-II，Vol.J79-D-II，No.10，pp.1704-1711，1996.

[18] 分散画像処理環境 VIOS，http://mars.elcom.nitech.ac.jp/vios/

[19] 小早川倫広，星守，画像内容に基づいた画像検索システム，ｂｉｔ，Vol.31，No.10，pp23-34，1999.

[20] 上田修一，絵画データベース －索引法と検索法を中心に－，情報処理，Vol.38，No.5，pp401-404，1997.

[21] 杉本重雄，デジタル図書館実現のための要素技術と環境要素，情報処理，Vol.37，No.9，pp820-825，1996.

[22] 原生生物と日本産アリ類の広域画像データベース －生物分類情報の広域データベース化とそのネットワークにおける利用システムの開発－，http://taxa.soken.ac.jp/WWW/Science_Internet/report96/menu.html

[23] 田中克己，ネットワーク社会とマルチメディアベース，情報処理，Vol.38，No.1，pp24-29，1997.

[24] Microsoft Corporation，Microsoft ODBC 3.0 プログラマーズリファレンス＆SDK，アスキー，1997.

[25] PostgreSQL Homepage，http://www.postgresql.org/

[26] 大津展之，島田俊之，森俊二，N次自己相関マスクによる図形の特徴抽出，信学技報，PRL78-31，pp.81-90，1978.

[27] N. Otsu and T. Kurita, A new scheme for practical flexible and intelligent vision systems, Proc. IAPR Workshop on Computer Vision, Tokyo, pp431-435, 1988.

[28] Elif Albus, Erturk Kocalar and Ashfaq A. Khokhar, Scalable Image Indexing and Retrieval Using Wavelets , SCAPAL Technical Report , http://www.scapal.ece.udel.edu/imagebas.htm, 1998.

[29] Apostol Natsev, Rajeev Rastogi and Kyuseok Shim, WALRUS: A Similarity Retrieval Algorithm for Image Databases, SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, pp395-406, 1999.

[30] 広池敦, 森靖英, 櫻井彰人, 積分型特徴量による形態識別, 信学論 D-II, Vol.J80-D-II, No.1, pp.81-91, 1997.

[31] 森靖英, 廣池敦, 櫻井彰人, ランダムテンプレートを用いた積分特徴量による画像分類手法, 信学論 D-II, Vol.J80-D-II, No.6, pp.1370-1378, 1997.

[32] T. Kurita, N. Otsu and T. Sato, A face recognition method using higher order local autocorrelation and multivariate analysis, Proc. of 11th International Conference on Pattern Recognition, Aug.30-Sep.3, The Hague, Vol.I, pp.530-533, 1992.

[33] 石井達夫, PC UNIX ユーザのための PostgreSQL 完全攻略ガイド, 技術評論社, 1999.

研 究 業 績

[1] A. Kawaguchi and K. Satou, A Database System for the Management and Karyotype Analysis of Chromosome Image Data, The Tenth Workshop on Genome Informatics, Tokyo, 1999.