科学研究費補助金「重点領域研究」

平成３年度発足重点領域申請書

ゲノム解析に伴う大量知識情報処理の研究

申請代表者　金　久　　實

(京都大学化学研究所教授)

目次

１．申請領域の研究の必要性

生物の全遺伝情報を構成するゲノムを解析することが実験的に可能になり、ヒトを始めとする各種生物ゲノムの研究が開始されている。その結果、これまでにも飛躍的な勢いで蓄積されてきたＤＮＡの塩基配列という一次元の文字列データが、さらに爆発的に増加すると考えられる。本来ここには生命の設計図、たとえば生物が生命活動を営むうえで必須の分子に関する情報と、それらがいつどのような状況の下で発現するべきかといった制御情報などが含まれているが、ＤＮＡの文字列だけからこれら高次の情報を解読する方法はほとんど確立されていない。一方、情報科学の分野では人工知能の研究をベースとし、知識処理と呼ばれる新しいコンピュータの技術が実現してきた。配列データの生物的な意味を理解することは基本的に知識処理の問題であり、情報科学の分野でも新しい応用問題としてとらえることができるだろう。バイオサイエンスとコンピュータサイエンスの新たな展開の中で、長期的な研究計画の下に、ゲノム情報解析に必要な大量知識情報処理の基礎的な方法論を確立することが、本申請領域研究の目的である。

知識処理のイメージを明確にするために、もう少し具体的な例を挙げてみよう。ＤＮＡやタンパク質の配列データからその生物的な意味を解釈するために、現在のところ最も有力とされている方法は、類似配列から推定する方法である。例えば新たに決定されたアミノ酸配列をもつタンパク質の機能を推定するために、データベースから類似配列を検索するホモロジーサーチが盛んに行われている。それは一次構造の類似性が進化上の関連を示し、立体構造や機能の類似性をも示していることが経験的に確かめられているからである。しかしながら、ホモロジーサーチで検出されるのは、あくまでも一次構造の類似関係であり、それが意味する生物的な類似関係の解釈については、専門家がその知識をもとに何らかの方法で推論を行っている。知識処理とは後者の推論の部分をも含めてコンピュータで処理するものであり、ホモロジーの生物的意味などをコンピュータ処理可能な知識として体系化することにより、情報科学の分野で行われてきた知識処理の一般的な枠組みを適用することができるのである。

さらに知識処理の立場は、ホモロジーサーチの現状での大きな問題点、すなわち検索時間と精度の問題についても、根本的な解決法となる。ホモロジ

ーサーチは一次情報のデータをそのまま検索の対象としているため、ゲノム研究の進展とともに急激に増大している膨大な量のデータに対しては、スーパーコンピュータを使っても何時間もかかることになるだろう。一方、知識処理の立場では一次情報から集約された知識を検索の対象とする。知識の獲得のためにはスーパーコンピュータが必要かもしれないが、知識を利用する段階になれば、一般ユーザーの検索はワークステーションで可能になる。ホモロジーサーチとは結局のところスーパーファミリーのような進化上関連のあるグループを探していることなので、あらかじめ個々のグループの特徴を知識として集積しておけば、その知識ベースから同等のことをより高速に得ることができる。さらに知識ベースにホモロジーでは分からない別の情報も蓄積しておけば、新たなレベルの知見も得られることになり、ホモロジーサーチの精度の限界も乗り越えることができる。

バイオサイエンス研究全体の中でコンピュータを用いた情報解析は、すでに実験研究者にとっても不可欠の手段として定着している。しかしながら、現状でのＤＮＡ・タンパク質データベースのデザイン、あるいはホモロジーサーチを始めとする解析ソフトウェアは、過去の研究成果に基づくもので、その連続的な変更・改良だけでは、新たなバイオサイエンスの展開には対応できない。遺伝子クローニングとＤＮＡ塩基配列決定法に代表される革新的な実験技術が、これまでバイオサイエンスにおける情報量を不連続的に増大させてきた。国際的なゲノム研究の推進とともに、今後も革新的な実験技術が開発され、情報量はさらに飛躍的な増加を示すことが予想される。従って、情報解析の分野でも既存の考え方にとらわれない革新的な研究を緊急に開始しなければならない。本重点領域研究は、知識処理という新しい考え方でバイオサイエンスの情報解析の問題にアプローチするのであり、その研究成果は例えば知識ベースやエキスパートシステムの形にまとめられ、ゲノム研究全体の整合性のある発展に大きく寄与するものと思われる。

コンピュータサイエンスの中での知識処理は、これまで例えば自然言語処理のように、人間の知識が比較的よく整理された分野で、知識の利用という立場から行われてきた。バイオサイエンスの分野では、まず知識をどのように獲得し、どのように表現するかが大問題あり、これまでのコンピュータサイエンスとしては未踏の分野である。本重点領域研究はバイオサイエンスとコンピュータサイエンスの接点に位置し、その研究はコンピュータサイエン

スに対しても大きな波及効果をもたらすであろう。

なお、本申請領域に最も関連の深い重点領域研究としては、平成元年度に発足した「大腸菌ゲノムの全体像」がある。大量知識情報処理に関する理論的研究が、実際のゲノム解析の実験研究の中でどれだけの意義をもつかのテストケースとなるであろう。

２．申請領域の国内・外の研究状況

分子生物学の分野に人工知能など新しいコンピュータ科学の手法を取り入れる最初の試みは、すでに1970年代の中頃より米国スタンフォード大学で行われていた。SUMEX (Stanford University Medical EXperimental) Computer Project の中の MOLGEN と呼ばれるプロジェクトである。ＤＮＡ塩基配列決定法が確立し、米国でＤＮＡデータバンク設立の機運が高まってきた1979年頃には、配列解析と分子生物学者のためのコンピュータネットワークがこのプロジェクトの中心となっている。MOLGEN プロジェクトの主要メンバーはその後 IntelliGenetics 社を設立し、NIH の支援の下に BIONET プロジェクトを開始したが、充分な研究成果がなかったため、1989年に打ち切りとなっている。もともとスタンフォード大学では、Feigenbaum の知識工学の考え方に基づき、人工知能の方法を現実の問題に適用すること、とくに専門家の知識を非専門家に提供するための、エキスパートシステムの作成に重点が置かれていた。MOLGEN および BIONET も同様の方向をとっていたが、分子生物学者に大きなインパクトを与えることができなかったのである。分子生物学では知識の利用よりも知識の獲得が問題だったからであろう。

国内では分子生物学と人工知能を融合した研究はまだほとんどなく、人的交流も少ないが、それぞれの分野では活発な研究活動が行われてきた。人工知能に関しては、とくに通産省が1982年より始めた第五世代コンピュータプロジェクトが、知識情報処理指向の新しいコンピュータ技術を確立することを目指している。そのために設立された(財)新世代コンピュータ技術開発機構では、分子生物学への応用は一切なされてこなかった。しかし、次節に述べるように、この状況は変わりつつある。

一方、分子生物学での情報解析はこれまで、実験データをただ単に貯えた

だけのデータベースを検索することが中心であったが、データベースから新しい知識を獲得するという視点をもった研究も始まってきた。例えば、類似の機能をもつ配列データのグループから、特徴的なパターンをコンセンサス配列あるいはモチーフとして定義した例が数多く報告されている。またモチーフを自動的に検出するアルゴリズムは、コンピュータによる知識獲得の試みである。さらにタンパク質の立体構造予測の分野でも、これまでのエネルギー計算など原理的な方法の限界から、構造パターンを知識ベース化するという経験的な方法に重点が移りつつある。すなわち、あらかじめ特定の物理化学のモデルなど仮定せずに、生物のデータをそのまま素直に眺めようというもので、知識処理は生物学の本来の姿を表わしていると考えられる。

米国を中心にヒトを始めとする各種生物ゲノムの研究が進行する中で、そのための情報解析の分野が新たな展開を迫られている。情報解析にはデータベース作成を中心とした研究支援の部分と、配列データの解釈に関する理論的研究の部分があるが、本研究領域はこのうちの後者に焦点を当てたものである。しかしながら、ゲノム研究に関連のあるデータベースをほとんどすべて欧米からの輸入に頼っている日本の現状では、前者の状況にも触れておかなければならない。これまで、データベースはほとんど欧米で先に標準化され、日本は作業分担の形で国際協力を行ってきた。国立遺伝研のＤＮＡデータベース活動（GenBank/EMBLに協力）や、東京理科大のタンパク質データベース活動（NBRF-PIRに協力）がその代表例である。別の形の国際協力は日本で独自の、しかしすでに欧米にあるものとは相補的なデータベースを作成することであるが、現実問題として、実験データをそのまま蓄積した一次情報のデータベースに関しては、今後相補的に必要になるものもすべて欧米で標準化され、整備されていく可能性が高い。従って、一次情報から知識を集約した知識ベースが、日本として独自に貢献できる領域であると考えられる。

３．申請までの準備研究・調査の状況等

学術審議会の建議「大学等におけるヒト・ゲノムプログラムの推進について」（添付参考資料）に基づき、平成元年度より文部省科学研究費総合研究

(A)「ヒトゲノム・プログラムの推進に関する研究」（代表者・大阪大学教授・松原謙一）が発足した。この研究班では (1) ヒトゲノム解析の実践、(2) cDNA ライブラリーの作成、(3) DNA 解析技術の開発、(4) 大量情報処理系の開発、(5) 各種生物ゲノムの解析について準備的研究を行ってきた。さらに、研究班の下に設置されたタスクグループ「ヒトゲノム情報小委員会」（世話人・京都大学教授・金久實）では、ゲノム研究推進の中でデータベースやコンピュータによる情報解析をどのように位置づけるかを明確にさせ、そのための研究体制および研究支援体制について議論し、将来に向けた提案を作成しつつある。その中では、将来必要となるデータベース、知識ベース、およびソフトウェア環境を実現するために、現段階で早急に基礎的な研究を開始することが重要であると確認されている。とくに配列データの大量知識情報処理は、従来からのバイオサイエンスの枠にとらわれず、新しいコンピュータサイエンスとの接点として発展させなければならないので、重点領域研究として推進することが最もふさわしいと考えられた。

平成元年１２月１１～１２日に総合研究(A)の研究班が主催し東京で開催された公開ワークショップ「ヒト・ゲノム研究の現状と展望」では、９８件中２０件が情報関係の口演で、この分野への関心の高さが示された。また、平成２年１月８日のヒトゲノム情報小委員会では、各種生物ゲノムの研究者と合同会議を開き、大腸菌、酵母、線虫、ショウジョウバエ、マウス、イネなどのゲノム研究状況を調査し、大量知識情報処理が重要な課題であることを確認した。

一方、第五世代コンピュータプロジェクトの中心機構である(財)新世代コンピュータ技術開発機構（ＩＣＯＴ）でも、平成元年度より「遺伝子情報処理ワーキングループ」（主査・東京大学教授・米澤明憲）が設置された。並列推論マシンを遺伝子やタンパク質などの分子生物学の分野に応用するための課題の検討とシステムイメージの確立を目的とし、分子生物学の専門家とコンピュータの専門家による比較的少人数の会合を定期的に開いてきた。第五世代コンピュータプロジェクトは、前期３年、中期４年が終わって一応のマシンが完成し、後期３年が始まった現在、応用問題として分子生物学を取りあげたことになる。ワーキンググループその他の研究交流で、ゲノム情報解析のニーズとＩＣＯＴがもつ知識処理技術のシーズが、自然な形で融合することが明らかになった。

４．重点領域研究を推進するに当たっての基本的考え方

主要研究項目

従来のバイオサイエンスおよびコンピュータサイエンスの分野の枠を広げて、その接点に新しい研究領域を構築するという基本的な考え方から、まず２つの主要研究項目を設定した。主要研究項目Ａ「ゲノム言語と並列処理」はコンピュータサイエンスを出発点とし、主要研究項目Ｂ「ゲノム情報の知識ベース」はバイオサイエンスを出発点として、知識情報処理の研究を行う。さらに、ゲノム解析を実践するバイオサイエンスの実験研究者との接点のために、主要研究項目Ｃ「ゲノム解析のソフトウェア環境」では大量情報処理の研究を行う。主要研究項目の間の相互関係は次頁の概念図に示されている。

研究組織

それぞれの主要研究項目は、まず５人の研究者で組織する計画研究班でスタートさせる。これらは既存の分野にない新しい研究領域であり、新たな分野として確立させるために若手の精鋭の研究者を中心とする。また公募研究を重視し、研究者層の拡大・充実をはかる。さらに総括班には、わが国のバイオサイエンスおよびコンピュータサイエンス研究で指導的立場にある研究者を集め、全体の推進を図りつつ、これをゲノム研究と密接に結びつけるべく、情報解析の方向づけを行う。

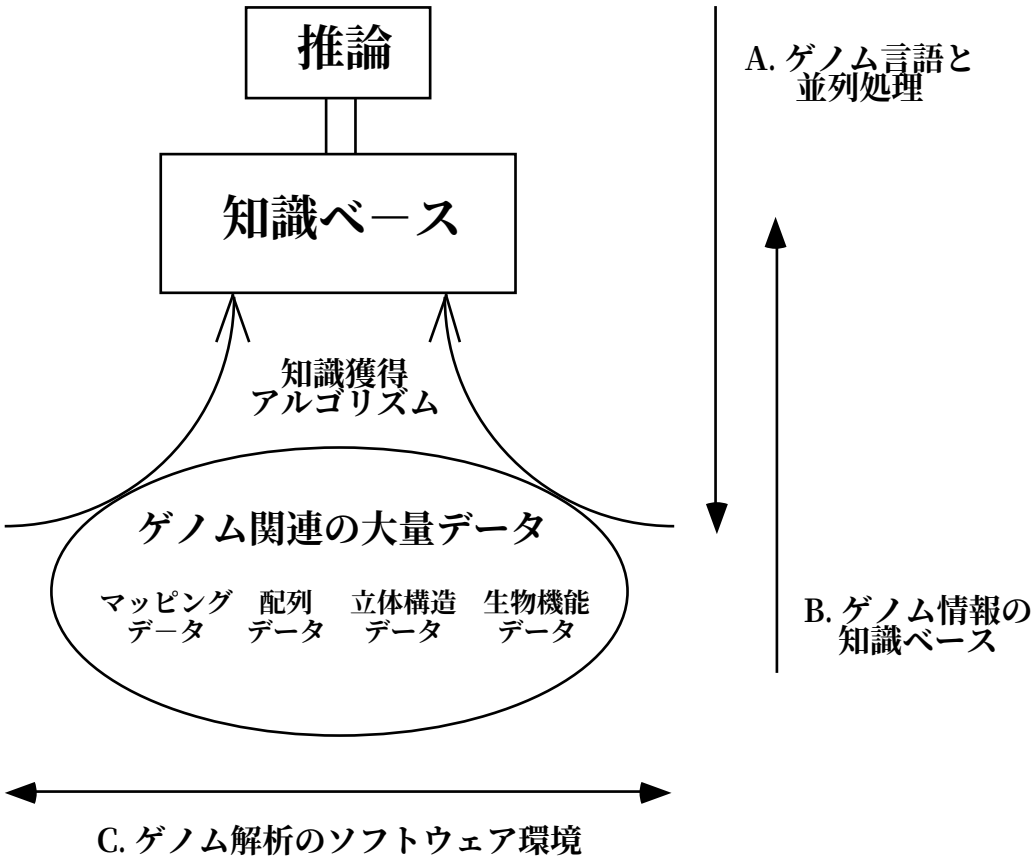
研究期間

研究期間は５年とするが、各種生物ゲノム研究の進展状況に対応するため、３年目に見直しを行う。また、５年目には公募研究は行わない。

研究経費

設備備品はワークステーションを中心とし、研究者間のコンピュータネットワークを構築する。

主要研究項目の概念図


推論
知識ベース
知識獲得
アルゴリズム
ゲノム関連の大量データ
マッピング
データ
配列
データ
立体構造
データ
生物機能
データ
A. ゲノム言語と
並列処理
B. ゲノム情報の
知識ベース
C. ゲノム解析のソフトウェア環境

５．重点領域研究の内容

(1) 申請領域の研究の具体的な内容

本研究では３つの主要研究項目を設定した。

申請領域名　ゲノム解析に伴う大量知識情報処理の研究

主要研究項目Ａ　ゲノム言語と並列処理
主要研究項目Ｂ　ゲノム情報の知識ベース
主要研究項目Ｃ　ゲノム解析のソフトウェア環境

それぞれの具体的な研究内容は以下の通りである。

主要研究項目Ａ「ゲノム言語と並列処理」

異なる分野の研究者がそれぞれの問題意識の延長上にお互いの接点を見いだすため、単純なアナロジーであるが、ゲノム情報解析の問題を言語処理の問題として考えてみる。すなわち、ＤＮＡ塩基配列やタンパク質アミノ酸配列の文字列をゲノム言語と呼ぶことにする。自然言語との大きな違いは、まず単語に相当するものがどこにあるかよく分からない。しかし、例えばタンパク質や核酸の機能部位を構成する一次構造には、コンセンサス配列あるいはモチーフなどと呼ばれる保存パターンがよく見られることが経験的に知られている。一般に複数個の保存パターンから成るモチーフでは、配列上離れた所にあるパターンが立体構造上では１箇所に集まって、機能部位を構成する。従って、ゲノム言語解釈の立場からは、個々の保存パターンを１つの単語と定義し、その集まり方の規則で生物的な意味が発生すると考える。

そこで、本研究項目ではモチーフの知識処理により、配列データから高次の生物的意味を推論する方法論の研究を行う。その手順は以下の通りである。

１） 配列データを何らかの意味でのグループに分類する。
２）個々のグループを特徴づけるモチーフを発見する。
３）モチーフとその意味を知識ベースにする。
４）知識ベースをもとに意味の推論を行う。

グループ化の基準となる配列データの意味としては、タンパク質や核酸が持つ、機能・構造・進化という３つの側面に分けて考えることにする。もちろん、それぞれは相互に深く関連しており、一般にホモロジー（進化の関連）があれば構造や機能の類似性もあるが、逆は必ずしも真ではない。また、機能が同じなら立体構造も同じ可能性が高いが、局所的に立体構造が似ていても機能的には異なる場合もある。

１）ではまず進化上の関連によるグループ化として、ホモロジー（配列データ全体としての類似性）で分類されたスーパーファミリーを利用する。また立体構造パターンの解析から、局所立体構造によるグループ化を行う。さらに実験データに基づき、機能部位によるグループ化も行う。

２）のモチーフの発見が最も困難な知識獲得の段階であり、データベースからモチーフを自動的に抽出するアルゴリズムの研究が非常に重要である。従来からの方法としては、複数配列アライメント法、多変量解析法、ニューラルネットワーク法などがある。また、モチーフとはそのパターンが特定グループを特徴づけているのであるから、データベース全体の中でパターンのユニーク性に着目した方法も考えられている。

３）の中心はモチーフ辞書の作成である。モチーフ辞書のデータからどのような高次情報が得られるのかの検討と、高次情報の仮説生成のシステマティックな方法の研究、仮説の間の制約条件の解析などが必要である。また、分子生物学の一般的知識の抽出と知識ベースの作成、知識表現と推論方式の検討も行う。

４）は第五世代コンピュータの逐次型推論マシンおよび並列推論マシンで意味解析支援システムのプロトタイプを作成し、並列処理による高速化の効果を検討する。最終的には、UNIX 環境でこの機能が利用できるように整備することが望ましい。

主要研究項目Ｂ「ゲノム情報の知識ベース」

主要研究項目Ａではモチーフを自動的に検出するアルゴリズムに主眼があったが、ここでは実験事実をモチーフその他の知識として集大成し、新しい研究活動を可能にする知識ベースを構築することに主眼がある。

第１のタイプの知識ベースは実験データから情報を集約することで、とくにＤＮＡの遺伝子発現制御情報、およびタンパク質の構造・機能情報に着目

し、知識ベースの構築を行う。具体例としては、（財）蛋白質研究奨励会の文献検索をもとに、配列データと高次情報の関連を原論文より抽出する。また、タンパク質立体構造データベースのパターン解析により、立体構造要素の知識ベースを作成する。さらに、タンパク質の辞書化、例えば酵素に関して、その反応、基質、生成物、補酵素などを一次構造、立体構造とともに蓄積する。このような新しい知識は、主要研究項目Ａで獲得された知識と統合し、推論システムに取り入れる。

第２のタイプの知識ベースは、既存の公共データベースの再編成ないしは相互索引づけに関するものである。例えば、アミノ酸配列データベースを中心とし、立体構造情報、エキソン・イントロン構造情報、スーパーファミリー分類などとの相互関係を明確にする。具体的には NBRF-PIR に PDB および GenBank からの情報を付加し、配列間比較による類似関係の情報も蓄積しておく。このような相互索引は、主要研究項目Ａで行うモチーフに関する知識獲得のための基礎データとなる。

知識ベースの構築には当面はリレーショナルデータベースを採用する。将来的には、第五世代コンピュータの逐次型推論マシンで、非正規関係モデルに基づく知識ベース管理システムを考慮する。非正規関係モデルはリレーショナルデータベースの限界を打破し、知識をより自然に表現できるモデルである。

また、知識ベースを利用した解析・設計の新しい研究を行う。とくにタンパク質の辞書などには、グラフィックデータを始め、さまざまなタイプのデータが混在するので、オブジェクト指向のソフトウェアが必要である。

主要研究項目Ｃ「ゲノム解析のソフトウェア環境」

ゲノム研究には異なるレベルでのデータが必要である。遺伝子地図、染色体物理地図、整列クローンライブラリー、制限酵素地図などのマッピングデータ、塩基配列およびアミノ酸配列のシーケンスデータ、さらに核酸・タンパク質の立体構造のデータや機能データなど、それぞれ大量のデータが関与している。ゲノム研究に必要なデータをどのように蓄積し、どのような形でユーザーに提供するか、そのためにどのようなアルゴリズムが必要かはまだ研究開発段階にあるといえるだろう。ゲノムデータベースとは、これまで個別に作成されてきた各種マッピング関連のデータベースを統合し、配列関連

のデータを取り入れたものであると考えられる。ゲノムを異なる見地、異なる解像度で眺めたデータを統合するのであるから、それを利用するソフトウェアのデザインが重要である。

データベースの利用形態には、センターのホスト計算機にアクセスして使うやり方と、個々の研究者がパソコンなどにデータベースのコピーをもらって使うやり方が考えられる。それぞれに利点があるが、現状では両者は全く分離された形で存在している。ゲノム解析を実践していく上で、データベースやネットワークの環境をどのように整備していくかも、大きな研究課題である。ゲノム解析が扱う大量のデータはセンターのデータベースでしか管理できないが、各研究者にとってはあたかも手元にあるデータベースのように利用できるようにするためには、分散処理の環境を整備する必要がある。具体的には UNIX をベースとし、センターのスーパーコンピュータと各研究者の所有するワークステーションを統合して利用できる環境を実現することが望ましい。そのようなコンピュータネットワークは、電子メールや電子掲示板など、ゲノム研究者の情報交換の場としても利用することができる。

(2) 主要研究項目の研究内容と研究組織

主要研究項目A　ゲノム言語と並列処理　（班長：金久　實）

金久　實	京都大学化学研究所・ 教授・生物物理学
田中 穂積	東京工業大学工学部・ 教授・情報科学
米澤 明憲	東京大学理学部・ 教授・情報科学
小山 照夫	学術情報ｾﾝﾀｰ ・助教授・情報科学
新田 克己	(財)新世代ｺﾝﾋﾟｭｰﾀ技術開発機構 ・主任研究員・情報科学

主要研究項目B　ゲノム情報の知識ベース　（班長：西岡 孝明）

西岡 孝明	京都大学化学研究所 ・助教授・情報化学
高橋 由雅	豊橋技術科学大学・ 助教授・情報化学
陶山　明	長岡技術科学大学・ 助教授・生物物理学
美宅 成樹	東京農工大学工学部・ 助教授・生物物理学
瀬戸 保彦	(財)蛋白質研究奨励会・ 主任研究員・情報化学

主要研究項目C　ゲノム解析のソフトウェア環境　（班長：五條堀 孝）

五條堀 孝	国立遺伝学研究所・ 助教授・分子生物学
久原　哲	九州大学農学研究科・ 助教授・分子生物学
伊藤　彬	(財)癌研究会癌研究所・ 部長・医療情報学
後藤　修	埼玉県立がんｾﾝﾀｰ ・研究員・生物物理学
蓑島 伸生	慶應義塾大学医学部・ 助手・分子生物学

(3) 総括班の組織と役割

研究代表者：

金久　實	京都大学化学研究所・ 教授・生物物理学

班員：

石濱　明	国立遺伝学研究所・ 教授・分子生物学
磯野 克己	神戸大学理学部 ・教授・分子生物学
内田 俊一	(財)新世代ｺﾝﾋﾟｭｰﾀ技術開発機構・ 室長・情報科学
佐々木慎一	豊橋技術科学大学・ 副学長・情報化学
清水 信義	慶應義塾大学医学部 ・教授・分子生物学
関口 睦夫	九州大学医学部・ 教授・分子生物学
高浪　滿	京都大学化学研究所・ 教授・分子生物学
田中 穂積	東京工業大学工学部・ 教授・情報科学
伏見　譲	埼玉大学工学部・ 教授・生物物理学
松原 謙一	大阪大学細胞工学ｾﾝﾀｰ・ 教授・分子生物学
宮澤 三造	国立遺伝学研究所・助教授・生物物理学
吉川　寛	大阪大学医学部・ 教授・分子生物学

総括班はバイオサイエンスとコンピュータサイエンスの研究で指導的立場にある研究者で構成し、若手の研究者が中心の各研究班の方向づけを行う。また、実際のゲノム研究と密接に結びつけるべく、ワークショップやシンポジウムの開催、その他の活動を行う。

６．研究期間

本研究は、バイオサイエンスとコンピュータサイエンスの接点を確立することが最終目標であり、長期的な研究計画が必要である。一方、各種生物ゲノム研究の進展に機敏に対応するために、３年目に研究計画の見直しを行う。また、５年目には公募研究は行わない。