1．研究実施計画

課　題　名：ゲノム DNA 情報の構造生物学的解析
研究機関名：独立行政法人　産業技術総合研究所
任期付研究員氏名：舘野　賢

①研究の意義・目的・必要性

(a)意義

課題提案の時点においても，既に１ダース以上の生物種についてゲノム全体の塩基配列が決定されている。このように，サイズの小さいゲノムに関しては塩基配列の決定自身は確立された技術である。したがって，現在のゲノム科学の最先端のテーマは，塩基配列の決定というよりはそこに記録されている情報（遺伝子の同定など）の系統的，総合的解明にある。このための科学的・技術的開発なくしては，たとえヒトゲノムの全塩基配列が将来決定されたとしても，そこから生物学的情報を得ることは極めて困難である。

ゲノム DNA 塩基配列に含まる情報を総合的に解明するには，現在発展中である構造生物学を中心として，分子生物学や生化学などの膨大な複合領域にわたる広範な知識や先端技術を総合する必要があり，ゲノム情報科学にとどまらず，今後のポストゲノム時代の生物学をも左右する重要な課題である。その実現には適切な規模と広範囲にわたる研究者の集中的投入が必須であり，国立研究所が中心となって推進すべき重要かつ大規模な課題であると考える。

(b)目的

本研究では，ゲノム DNA 塩基配列内の遺伝子の同定と，同定された多数の遺伝子に関する情報の解析の２つを目的とする。現在の遺伝子同定法は信頼性に乏しく，生物学的に不十分なものである。そこで本研究では，転写・翻訳開始のシグナルを同定することによって，その下流に存在する遺伝子を同定するための新規のアルゴリズムを確立する。同定された遺伝子をもとにゲノムの全体像を再構成するには，多数の遺伝子がコードするタンパク質の立体構造や機能，動的構造などを系統的に解析することが必要である。そのために，分子の立体構造や動的構造に関する構造生物学的な情報を，ゲノム情報から理論構造生物学的な手法を用いて取得し，それらの間の相関を系統的に解明する。

本研究では，以上の解析をサイズの小さいより単純なゲノムに関して行うことによって，その過程で得られた知見を発展させ，その結果を一般化・総合することによって、より複雑で大規模なヒト等の真核生物や大腸菌等の真正細菌のゲノム解析にも応用可能な技術開発のための科学的基礎を確立することを目的とする。

(c)必要性

ヒトゲノムなどの高度に分化したシステムの全体像を理解するには，より単純な系の全体像をその DNA 塩基配列をもとに再構成できなければならない。そのためには，ヒトゲノムなどのよいモデル系を用いて，遺伝子同定法の基礎技術と，さらに同定された多数の遺伝子に関する系統的な情報（特に構造生物学的な情報）を取得するための技術とを確立する必要がある。古細菌ゲノムは，真核生物に極めて近い転写・翻訳システムをもちながらも，その全遺伝子数は約 2,000 個と少なく，ヒトゲノムのよいモデル系（サイズは約 1/1000）であると考えられ，本研究課題に最適の系である。

生命工学工業技術研究所では，ゲノム生物学と構造生物学の融合とを，将来に向けての大きな研究課題の一つと位置付け，古細菌ゲノム等を対象として重点的に研究資源の投資を行ってきた。本研究は，理論的（情報科学的，理論構造生物学的）アプローチによるゲノム研究をさらに補強し，

他の実験的研究分野とのより効果的な戦略的融合，発展形成を試みるものである。

②研究の概要

現在一般に使用されている，ゲノム内の遺伝子を同定する方法（開始コドンと終了コドンとの間に，ある個数以上のアミノ酸がコードされうるものを遺伝子とするなど）は，生物学的には信頼性が乏しく，多くの誤りを含んでいる。本研究では，TATA ボックス（転写シグナル）および SD 配列（翻訳シグナル）を同定することにより，転写ユニットを同定し，以ってその内部に存在する遺伝子を同定する方法を開発する。古細菌ゲノムの約 9 割は遺伝子のコード領域であり、残りの約 1 割の領域についてシグナルの検索を行えばよく，遺伝子同定法の開発に非常に適している。一方，ヒトゲノムでは比率が逆転しており，9 割以上が複雑な転写シグナルの記録に使われている可能性もある。また翻訳シグナルについては，古細菌に関する知見の蓄積はあるが，ヒトゲノムでは今のところよくわかっていない。古細菌ゲノムは，真核生物に極めて近い転写，翻訳システムをもちながら，その全遺伝子数は約 2,000 個と少なく，ヒトゲノムのよいモデル系と考えられる。

シグナルを精密に同定するためには，ゲノム DNA の塩基配列に対する統計学的な情報の抽出のみでは不十分であり，シグナルが転写因子等によってどのように認識されるのか，その分子認識機構を構造生物学的視点から原理的に理解することが必要である。そのために，転写因子と DNA との複合体の立体構造に対して，分子動力学計算などの理論構造生物学的な手法を駆使することによって，それらの分子間相互作用を解析する。この解析結果を再び遺伝子自動同定システムのアルゴリズムに統合することによって，より精密で精度の高いシグナルおよび遺伝子の同定を実現できるものと期待される。同定された古細菌ゲノムの遺伝子データベースは，工技院情報計算センターのネットワーク支援網により，一般に公開する。

③研究目標

(a)転写・翻訳シグナルの同定による遺伝子自動同定システムの構築

古細菌の転写・翻訳開始のシグナルを統計学的に解析し，遺伝子自動同定システムを構築する。

(b)シグナルの認識機構の構造生物学的解析

基本転写因子である TATA-box 結合タンパク質（TBP）は，ターゲット DNA 分子を大きく屈曲させることによって結合する。そこで，分子動力学計算によって両者の動的な相互作用過程にまで踏み込んで、その分子機構を解明する。

(c)同定された遺伝子に関する情報の導出

ゲノムがコードするタンパク質の立体構造を理論的に構築し，分子動力学計算などを駆使することによって、それらの分子の動的構造を明らかにする。得られたこれらの情報と機能との相関関係を系統的に解明するための基礎を確立する。

ゲノムDNA配列情報の構造生物学的解析

- **ゲノムとは？** —— 音楽カセットテープとの比較

古細菌ゲノムDNA（音楽カセットテープ）

転写・翻訳開始のシグナル（曲のヘッド出しシグナル）

- **ゲノム内の遺伝子を見つける** ⟶ ヘッド出しシグナルを探す

● 統計学的な解析 —— シグナルの探し方１

シグナルの塩基配列のパターン（頻度）

シグナルと開始コドン（遺伝子の開始点）との距離

⟶ これらの統計的なルールを抽出する ⟶ 遺伝子の自動同定

● 構造生物学的な解析 —— シグナルの探し方２

シグナルは転写因子（TBP、TFBなど）によって認識される ⟶ 転写因子によるDNA（シグナル）の認識機構を構造生物学的に解明

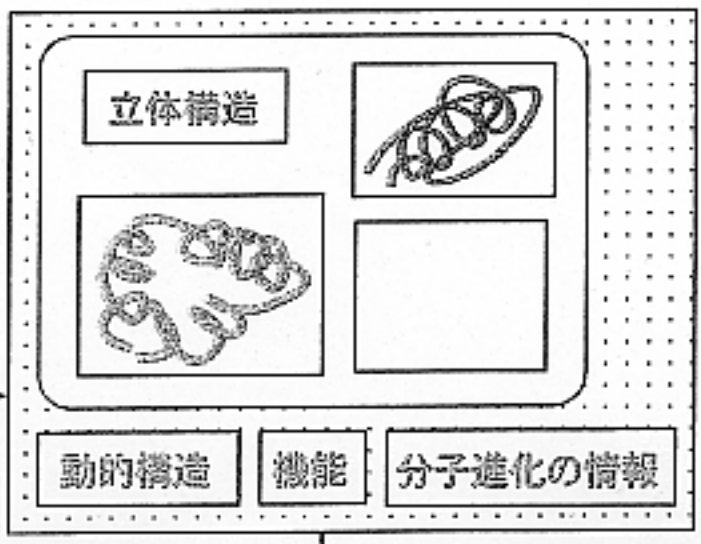
理論構造生物学の手法
分子の立体構造を構築する
分子の動的な構造を調べる

- **ゲノムの全体像をつかむ**

遺伝子 —— DNA塩基配列

遺伝子はタンパク質のアミノ酸配列をコードしている

タンパク質 —— アミノ酸配列



理論構造生物学の手法
分子の立体構造を構築する
分子の動的な構造を調べる

ゲノムの全体像を再構成

２．研究成果の概要

①研究成果

（１）遺伝子同定の計算アルゴリズムの開発と半自動解析のための情報システムの開発

-----ゲノム塩基配列情報の半自動的，高速かつ正確な解析

ゲノム DNA の塩基配列に対して，遺伝子を正確に同定するためのアルゴリズムを開発するには，ヒトなどの極めて複雑なゲノムを用いる以前に，より単純な生物種をモデルとして用いることが有効である。そこで，古細菌のゲノム DNA 塩基配列を例に，計算アルゴリズムおよびその情報システムの開発・構築を行った．ここで重要な点は，「ゲノムに内在する転写シグナルや翻訳シグナルなどは，全ての遺伝子が共通に持つ」と予想される，物理的な特徴（パターン）になっている点である。したがって，これらの物理特徴を用いれば，その直下に存在する遺伝子を発見することが可能であろう。

そこでまず，①古細菌のゲノム塩基配列内に見られる「少数の」既知遺伝子（＝生存に必須である遺伝子）のみに対して，これらの物理的特徴（各シグナル）を抽出し，それぞれのシグナルに内在する共通の統計的性質（統計的特徴）を抽出するのが，解析の第一段階である．次に，②古細菌ゲノム DNA の「全」塩基配列に対して，抽出された統計的特徴と合致する塩基配列（すなわち各シグナル）を探索する．これによって，（同定された）シグナルの近傍に存在する遺伝子（正確には ORF）を理論的に見出すことができる。

この統計・情報科学理論に基づき，古細菌ゲノム DNA 配列を半自動的に解析することの可能な情報システムを開発した。その結果，～2×10^6 塩基対のゲノムに対して約１週間という高効率で，遺伝子（正確には，ORF，オペロン，および偽遺伝子）を高精度に同定することに成功した。

（２）遺伝子のタイプ判別アルゴリズムによる古細菌ゲノム DNA 塩基配列の解析

-----ゲノム間の"のっとり"の痕跡を発見

こうして，ゲノム DNA の全塩基配列内に遺伝子を正確に同定するための情報科学的な基礎理論の確立と，それに基づいたアルゴリズムおよび情報システムの開発に成功した。そこで次に，このシステムを用いて古細菌ゲノム内の遺伝子を実際に解析・同定し，その結果を元に，ゲノム塩基配列の生物学的解析をさらに進めた。

まず上記の半自動ゲノム解析システムを用いて，古細菌 *Pyrococcus sp.* OT3 のゲノム DNA の全塩基配列から，すべての遺伝子を同定した（約 1800 個）。次に，これらの遺伝子に対する生物学的全体像を把握するために，"遺伝子産物（タンパク質および RNA）のそれぞれが，真核生物および真正細菌のいずれのタイプに近いのか（あるいは古細菌特有のものであるのか）"について，統計・数理学的な解析手法を開発した。これは例えば，「*Pyrococcus sp.* OT3 が保持するあるリボゾームタンパク質は，ヒト（真核生物）あるいはバクテリア（真性細菌）が保持するいずれのリボゾームタ

ンパク質により近いのか，または，それらいずれとも異なる（古細菌）固有のものであるのか」について，統計的に分類しようというものである（遺伝子のタイプ判別アルゴリズム）。これはひいては，"*Pyrococcus sp.* OT3 はヒトに近いのか，それともバクテリアに近いのか？"などの興味深い疑問にも関連するものである。

この解析のために，分類の生物学的基準を設定し，これを各遺伝子産物（タンパク質および RNA）に統計的に適用することにより，それぞれのタイプを識別するための情報科学的理論を開発した。この手法を適用して各遺伝子の型を決め，*Pyrococcus sp.* OT3 のゲノム内にマップした。その結果，真正細菌の遺伝子に近い *Pyrococcus sp.* OT3 の遺伝子群（真正細菌型遺伝子）は，そのゲノム全体に分散しているのに対して，真核生物の遺伝子に近い *Pyrococcus sp.* OT3 の一連の遺伝子群（真核生物型遺伝子）は，ゲノム内の一定の場所に集中して分布している傾向の強いことが明らかになった。これは進化の過程で，真核生物（の始源生物）のゲノムが，真正細菌のゲノムに挿入・融合されて，*Pyrococcus sp.* OT3 のゲノムが新たに創生された可能性を示唆しており，かつてゲノム間の"のっとり"が起こったことを意味するものである。

これをさらに支持する事実として，以下のような興味深い知見も見出した。生体（細胞）内においては，リボゾームがタンパク質の合成の場（タンパク質生成"工場"）であり，「リボゾームタンパク質」および「リボゾーム RNA」により構成される，一個の細胞内構築物である。ところが，これらリボゾームタンパク質およびリボゾーム RNA の遺伝子は，上記の遺伝子タイプ判別アルゴリズムによって，それぞれ真核細菌型および真正細菌型と識別され，互いに異なる型に属することが明らかになった。これはとりもなおさず，*Pyrococcus sp.* OT3 のリボゾームが，真正細菌型および真核細菌型の両遺伝子のキメラであることを示唆している（ゲノム間の"融合"が起こって，一方のリボゾームが他方にのっととられた！）。

（3）転写シグナルの認識機構に関する理論構造生物学的解析

-----生体高分子のダイナミクス（動的構造＝熱運動性）の重要性

前記の遺伝子同定アルゴリズム内で用いた転写（開始）シグナルは，実際の生体システムにおいてはタンパク質（転写因子）が DNA の塩基配列（＝転写開始シグナル）を認識することによって行われる。したがって，より精密かつ正確な転写シグナルの同定を実現するためには，転写因子と DNA との相互作用を原子解像度のレベルにおいて解析し，その分子機構を原理的かつ詳細に理解することが必須である。そのために本研究では，理論構造生物学的手法を駆使して，基本転写因子と DNA との相互作用を例に，原子分解能において解析した。

転写の開始点を決定する，DNA 上の重要なシグナルは「TATA-box」と呼ばれ，これは「TATA-box 結合タンパク質（TBP）」によって認識される。TBP による DNA（TATA-box）の認識機構を理解するためには，以下に記す理由により，それぞれの分子の運動性を調べることが必要である。しかし，実験的手法によってこれを行うことは，現在最先端の立体構造構造解析技術を用いても，極めて大きな困難を伴うか，または解析の限界に遭遇することは必至である。そこで本研究では，高い精度のコンピュータシミュレーション（分子動力学計算）を駆使することによって，これを実現した。

（a）TBP の熱運動における特徴

TBP は，互いに対称なふたつのドメインから構成されている（擬２回対称軸を持つ）。常温の水

溶液において TBP は熱的に振動しているが（熱振動），それぞれのドメインの立体構造は，その間も基本的にほぼ一定に保持されていることが，分子動力学計算によって明らかになった。ところが，TBP のふたつのドメイン間の相対的な位置関係については，その変位を理論的にさらに解析すると，熱ゆらぎによって変化し，それぞれのドメインは全体として熱運動してゆらいでいることが明らかになった。従来 TBP のこうしたドメイン間振動運動は，その有無自体に疑問もあったが，本研究によって，その運動性についての精密な定量化が始めて実現された。

（b）TATA-box に内在する熱運動の特徴

TBP が DNA に特異的に結合しその生物機能を発現する際には，以下に記す極めて興味深いふたつのポイントがある：① TBP および DNA のいずれの両分子も，基本的に 2 回対称な立体構造をもつにもかかわらず，TBP は DNA の一定の向きを認識する（逆向きに DNA に結合することはない。これは厳密には，真核生物由来の TBP に見られる特徴である）。②通常直線状である DNA 分子に対して，TBP はこれを大きく屈曲・変形させて結合する。

(a)で明らかになった TBP のドメイン間振動運動は，こうした DNA 屈曲をもたらす原動力となっている可能性がある。そこで次に，DNA（TATA-box）側の熱運動性についても解析するために，以下の計算を実行した。TBP によって大きく屈曲・変形された DNA 分子に対して，（TBP を除去した後に）単独で分子動力学計算を実行し，その形状の変化を追跡した。その結果，興味深いことに，TBP によって変形された屈曲 DNA のコンフォメーションは，通常 DNA が水溶液において保持する直線状の形状（B 型 DNA）へと，速やかに変化した。

ところが，この形状変化に伴う緩和過程を定量的かつ理論的に解析すると，この DNA 分子（TATA|AAAG）の左右ふたつの領域で，その緩和速度（＝直線状に戻っていく速度）が異なることが明らかになった。すなわち TATA 領域では，屈曲した形状の戻る速度が，AAAG 側に比較して非常に遅いのである。TATA-box のもつ，こうした熱運動の非対称性が，TBP による DNA 認識においてその結合の向きを定めるのに寄与している可能性があり，これによって転写の向き（＝ DNA の遺伝情報を読み取る向き）が定まるものと考えることができる。

同時にまた，DNA が保持するこうした分子のダイナミクスは，TATA-box においてのみ成立するのではなく，実際には DNA 分子一般に広くあてはまる結論であり，DNA 分子のコンフォメーションとそのエネルギー状態との対応関係（energy landscape）が明らかになったことを意味するものである。したがって，例えばタンパク質に結合した DNA がどのようなエネルギー状態にあるかなどを考察する場合などに，一般に非常に有益な結果である。このように，生物学における特定の系から出発しながらも，生体高分子が保持する一般的な法則性を導き得た研究例は，筆者の知る限りにおいてきわめて少なく，本手法の有効性を示すものである。最後に，これは物理学的には，東大理学部の鈴木増夫教授らが提唱する統計力学的解析手法「非平衡緩和法」の分子生物物理学（および分子動力学計算）への応用にもなっていることを付記しておく。

②波及効果、発展方向、改善点等

ゲノム内に遺伝子を同定するための新規アルゴリズム（「遺伝子同定アルゴリズム」）および「半自動遺伝子同定情報システム」は，真正細菌や真核生物などのより複雑な系に応用するために，現在さらに種々の改良を試みているところである。また，これらのシステムによって既に解析された

ゲノム情報は，産総研の研究情報公開のためのデータベース群（RIO-DB）のひとつとして，既に一般の利用に広く供している。また，遺伝子同定の検知精度をそれ自身さらに向上させるために，構造生物学的な解析（現在もなお解析は発展的に進行中である）の結果を，計算アルゴリズムのパターン認識過程に今後反映させる予定である。

さらに，生体高分子の原子解像度における熱運動性の解析では，新規の強力な手法である『非平衡緩和 MD 法』（MD：分子動力学計算）を，本研究における DNA での最初の成功を基礎にして，今後さらに多くの生体高分子系に応用を展開する予定である。例えば，浸透圧に応答し特定の物質を選択的に透過させるチャネルタンパク質やトランスポータなどに対しては，既にその応用を開始しているところである。

このように，本研究において新たに開発された配列解析（塩基配列およびアミノ酸配列）の手法，および立体構造・動的構造（熱運動性）の解析手法は，いずれも飛躍的な発展が強く期待される，極めて有効な手法であり，現在さらに発展の途上にある。今後さらに生物学上の具体的な実際問題への応用発展を精力的に推進したいと考えている。