配列情報からの規則性の発見

●中井 謙太[1)]　◆坂内 英夫[1,2)]　◆宮野 悟[1)]（2000～2002年度）　◆丸山 修[3)]（2000～2002年度）

1) 東京大学医科学研究所ヒトゲノム解析センター　2) 現：九州大学大学院システム情報科学研究院　3) 九州大学大学院数理学研究院

〈研究の目的と進め方〉

本特定領域研究の実施期間の間にも、ヒトをはじめとする多くの代表的な生物種の全ゲノム塩基配列が決定されてきた。しかし、それらの配列中に書き込まれている生物学情報を読み出すための技術や背景知識は、まだまだ十分であるとはいえない。そこで本研究では、DNA、RNA、アミノ酸の各配列データに潜む種々の規則性（ルール）をコンピュータ解析によって明らかにし、ゲノム配列データからの情報解読に役立つ知見を得ることによって、従来は十分にコンピュータで読み取ることのできなかった情報を解釈するための方法論の開発を目的とした。具体的には、主に転写調節領域中にある種々の転写因子結合部位を抽出して、その調節情報との相関を探る研究と、組織特異的なRNAスプライシング（選択的スプライシング）を配列上で特徴づけるエレメントの探索、アミノ酸配列のN末端領域に存在する細胞内局在化シグナルの特徴抽出と、膜タンパク質において、低疎水性の膜貫通部位を生みだす要因をアミノ酸配列から探る研究などに取り組んだ。

本計画研究は最初の3年間は3名の研究者、すなわち生物学知識とその応用に詳しいゲノム情報学研究者（中井）と機械学習理論・発見科学などの情報科学のエキスパート（宮野・丸山）が協力して行った。基本的には、3人が週1回集まって議論をしたが、その他にも各班員の得意の領域については個別でも研究を進めた。後半の2年間は、ゲノム生物学領域に移籍し、中井と情報科学のエキスパート（坂内）が協力して研究を進めるとともに、実験科学者との交流も積極的に行うことを目指した。

〈研究開始時の研究計画〉

5年間の研究計画は以下のようなものであった

1) アミノ酸配列からの規則発見

1-1) 膜タンパク質のトポロジー決定シグナル
αヘリックス型の膜タンパク質では、基本的に疎水性のアミノ酸残基が集中した領域が膜貫通部位になり、その周りに存在する正電荷をもつアミノ酸（塩基性のアミノ酸）の分布によって、膜を貫通する向きが決まると言われている（positive-inside ルール）。しかし、実際には疎水性が十分でない領域が貫通部位になっていたり、逆に疎水性が十分高いにもかかわらず、貫通部位として使われていない領域も存在する。これらの存在をアミノ酸配列から説明するルールの発見を目指した。

1-2) タンパク質の細胞内局在化シグナル
アミノ酸配列中に書き込まれた細胞内局在化シグナルを検出して、その配列がコードするタンパク質の局在部位を予測する問題は、研究代表者によって1990年頃に開拓され、その後のゲノム時代・プロテオーム時代に配列解析の一分野として確立された。この研究は個々のシグナルをいかに発見し、うまくモデル化するかという前半の問題と、個々の既知シグナルの特徴を数値化したとき、未知配列の局在部位をその特徴ベクトルからいかに総合的に予測するかという後半のパターン認識問題に分けることができる。これらを両面から発展させる。

1-3) その他
近年、組織的な2ハイブリッド実験などによって、タンパク質間相互作用の網羅的データがいくつも発表されている。それらのデータを用いて、タンパク質の紀要予測をどの程度行うことができるのかを評価し、実用的な機能予測がどの程度可能かを検討しようとした。

2)　RNA 塩基配列からの規則発見

2-1) 長大なイントロンを認識するための配列シグナル
イントロンの両端には GU-AG ルールで代表されるコンセンサス配列が存在するが、それらの存在だけでは、ヒトなどの高等生物に多数存在する長大なイントロンがなぜ認識されるのかが説明できない。おそらく他にも存在するのであろうルールの発見を目指した。

2-2) スプライシング異常データからの規則発見
上記 2-1 の問題を解くために有力な手がかりを与えてくれそうなデータとして、遺伝病等にみられる異常スプライシングパターン情報を研究代表者は以前から収集してきた。このデータを用いて、たとえばエキソンスキップと代替クリプティック部位活性化のどちらが起こるかを決めている配列的要因を発見することを目指した。

2-3) 真核生物 mRNA の半減期に関わるモチーフ探索
近年、マイクロアレイ実験等によって、多数のmRNAの半減期を網羅的に測定したデータが発表されている。それらのデータと3' UTRなどの配列データを組み合わせて、半減期の長短にかかわるモチーフの探索を試みた。そのために新しい配列内のパターン探索アルゴリズムの開発に取り組んだ。特に、パターンがあるかないか（全か無か）ではなく、その存在確率が与えられた数値に相関するようなパターンの組み合わせを探索する問題に注目した。

3) DNA 塩基配列からの規則発見

3-1) 細菌ゲノムにおける遺伝子発見
同僚の矢田助教授（現：京大工）との共同研究で、以前に矢田氏らが開発した細菌遺伝子発見プログラムGeneHackerを改良して、世界一の性能を持たせることを目指した。

3-2) 細菌遺伝子の転写制御シグナル
当時、ゲノム配列の比較が十分に行えるのは細菌ゲノムだけであった。その強みを活かして、また従来から行ってきた枯草菌の既知転写因子結合部位の収集データを活かして、枯草菌や大腸菌をはじめとする細菌の転写制御シグナルの情報解析を目指した。たとえば、

共通の転写因子によって制御される遺伝子群（レギュロン）を、その上流に近縁種間で保存された共通配列を持つか否かに従って予測することを試みた。また、枯草菌の転写データベースを更新し、比較ゲノム学的視点から新機能を付加した。さらに、その情報をもとにして、枯草菌の既知レギュロンにおける未知のメンバー発見を組織的に行った。

3-3) 高等真核生物遺伝子の転写制御シグナル

同僚の菅野助教授（現：東大新領域教授）との共同研究で、彼らがオリゴキャップ法を用いて網羅的に決定したヒト遺伝子の転写開始点情報を整理し、正確な転写開始点情報にもとづく上流配列の解析を行った。たとえば、転写開始点がよくそろっている遺伝子とぶれが激しい遺伝子の違いは何かという問題にアプローチした。理研の林崎グループが発表しているマウス遺伝子転写開始点の情報と比較することにより、高等生物遺伝子のプロモーター比較研究も試みた。

4) 上記の発見を可能にするための基盤情報研究

4-1) 発見科学からのアプローチ

分担研究者の宮野・丸山らが従来から試みていた発見科学的研究を発展させて、本プロジェクトに役立たせる。たとえばHypothesisCreatorという一種のプログラムライブラリを用いて、さまざまな仮説形成と検証を網羅的に行うことを目指した。その他、情報科学的な基盤研究をいくつか試みた。

4-2) DNA 共通モチーフ抽出プログラムの改良

当初はDNA塩基配列データからのモチーフ抽出プログラムの性能評価に重点を置く計画であった。すなわち、共通の転写因子結合部位を含むことが期待される一群の塩基配列を与えられたときにそのモチーフを発見することが、どの程度プログラムで可能なのか、またパラメータ値を変化させることで結果が殿程度変わるのかを確かめることで、現状のアルゴリズムの問題の把握と新たな改良の糸口をつかむことを考えていた。

〈研究期間の成果〉

1) アミノ酸配列からの規則発見

1-1) 膜タンパク質のトポロジー決定シグナル

九州大学グループの実験では、複数貫通型膜蛋白質の膜貫通部位はそれぞれ多かれ少なかれ固有のトポロジー形成能力をもっていることが示唆された。この主張は、一番N末端側の貫通部位がトポロジーを決め、続く貫通部位は膜透過停止配列になるか、透過再開配列になるかが自動的に決まるとする従来の説と対立する。しかし、この考えを推し進めると、たとえば二つ並んだ貫通部位がどちらも同じトポロジー形成能力をもっていると、矛盾を解消するために比較的低疎水性の領域が両者の間で貫通部位になることが説明できる。そこで、我々は膜蛋白質データベースを使って、この説を検証してみた。ただし、各貫通部位のもつトポロジー形成能力の実験値がないので、一本貫通型膜蛋白質において詳しく調べられている正電荷内側ルール（正電荷が多い側のループ領域が細胞質側を向く）が、平均疎水性が比較的高い内部膜貫通部位のトポロジー形成能力を決めていると仮定した。

データとしては、トポロジー実験情報をもつ複数貫通型膜蛋白質の中で、オルガネラ蛋白質や、35%以上保存されているものを間引いたが、原核生物の膜蛋白質の多くは真核生物と同じSRP/トランスロコン類似系を通して膜に挿入されているらしいので残した。さらに、貫通部位の位置情報を原論文でチェックし、貫通範囲の基準を統一した。次に正電荷内側ルールと高疎水性貫通部位の向きの相関を調べたところ、86%であった。さらに低疎水性貫通部位を平均疎水性の閾値から定義すると、26%の複数貫通型膜蛋白質に存在した。最後に、低疎水性貫通部位と高疎水性のループセグメントについて、それぞれ両隣の貫通部位の推定トポロジー形成能との間に高い相関（7割以上）が見られることを確認した。その中には、低疎水性セグメントが九大グループによる想定とは逆の向きの場合もあった。この結果は、九大グループの主張をより一般化された立場から裏付けるものであると言える（Araki et al., 未発表）。

1-2) タンパク質の細胞内局在化シグナル

まずは実験的に知られているシグナルや局在部位（またはシグナル）予測法の情報を広く収集し、折にふれ、総説にまとめた（1, 5, 15）。

アミノ酸配列のN末端領域に存在する局在化シグナルである、シグナルペプチド、ミトコンドリア移行シグナル、葉緑体移行シグナルの三種類の存在を与えられたアミノ酸配列中でコンピュータを使って識別・検出する問題は、研究代表者も以前から研究してきた。中でも北欧のグループがいわゆるニューラルネット法を使って開発したTargetPというプログラムが有名である。しかし、彼らの用いたニューラルネット法は、一般に高いパターン認識能力を示すものの、ネットワークが実際に認識している特徴の解釈が難しいという問題がある。そこで、我々は、十分高い識別能力を持ちながらも、なるべく直感的に理解しやすい局在化シグナルの配列上の特徴を発見することを目指して、いろいろな特徴候補の中から、シグナルの識別能力を指標にして、判別ルールの形で質の良い特徴量を大規模に探索した。

直感的に理解しやすい特徴として、二種類を考えた。一つは、20種類のアミノ酸を何らかの性質を表す数値に変換して、ある部分配列においてその平均値を求めるというもので、この場合は、部分配列の範囲と、数値パラメーターの選択と、シグナルの存在を判断する閾値を探索する必要がある。パラメーターの候補には、研究代表者らがその構築を始めたAAindexデータベースの全データと、各アミノ酸の存在頻度を用いた。もう一つの特徴としては、かつて研究分担者らがBONSAIというシステムで用いたアミノ酸インデキシング法で表現される特徴を考えた。これは20種類のアミノ酸をたとえば3つのグループに分類し、それぞれ1, 2, 3とラベル付けした後、そのラベルで表現されたパターン（たとえば123123）の存在を調べるもので、ある程度挿入・欠失やミスマッチを許すこともある。この場合、探索すべき特徴は、主にグループ分けとモチーフ探しになり、膨大な組み合わせを探索することになる。そこで、比較的計算時間が少なくて済む場合は全探索を行い、そうでない場合は局所探索を行った。

その結果、以下の特徴を使うと、TargetPと比べてそれほど遜色のない成績が得られることを発見した。すなわち、シグナルペプチドでは最先端部を除いた20残基ほどの領域の平均疎水性値が高いことが重要で、ミトコンドリア・葉緑体移行シグナルでは共に酸性のア

ミノ酸が少ないことが重要であり、中でもミトコンドリア移行シグナルが、より塩基性の残基を多く含む傾向があった。さらに、ミトコンドリア・葉緑体移行シグナルは共にアルギニン残基が大体両親媒性αヘリックスを形成するようなパターンをもち、ミトコンドリア移行シグナルでは、リジンも含めた塩基性残基の両親媒性も重要であることが判明した。これらの特徴は、それほど新しい地検であるとは言えないが、これらの知識を翻訳した、図に示した程度の簡単なルールによって、高い識別能力（植物で84%、非植物で88%）が得られることは驚きである（11）。

結局のところ、ニューラルネット法によって検出されているシグナルの配列上の特徴もそれほど複雑なものではなく、比較的簡単な性質がシグナルを指定する上で本質的であるものと結論される。ただし、この場合、単に直感的に得られたルールを並べただけではだめで、時間をかけて良い特徴量を探索することで初めてここまでの精度が得られたことを強調しておきたい。最後に、得られた予測プログラムは、iPSORTという名前でインターネット上で公開した（）。

さらに、カナダのグループとの共同研究により、グラム陰性菌のためのタンパク質の細胞内局在部位予測システムPSORT-Bを発表した（18）。

1-3) その他

酵母のタンパク質間相互作用データを使って、未知タンパク質の機能を推定する簡単な枠組みを提案し、これを使ってどの程度まで機能予測が可能なのかを網羅的に調べた（3）。相互作用するタンパク質は関連した機能を持っていることが多いことが予想されるので、"Guilt by association"の原理に基づき、機能未知の対象タンパク質と相互作用するタンパク質の機能をリストアップして、その最大数を占める機能を対象タンパク質に割り当てようというのがもともとの考え方である。実際には、その考え方をさらに発展させて、直接相互作用している相手だけでなく、k-1個のタンパク質を媒介にした間接的相互作用をしているものまで含めて調査を行い、最適なk値を探索した。また、機能グループの大きさがときに大きく異なるため、多数決だと小さな機能グループには予測されにくくなってしまう問題を避けるため、いわゆるカイ二乗値を各機能グループについて計算した。予測を行うべき「機能」としては、YPDデータベースにおける３つの分類（細胞内局在、細胞内での役割、生化学的機能）を利用した。相互作用データには、MIPS、CuraGen、Itoのデータを用いた。相互作用するタンパク質なら、同じ局在部位に存在することは自明のようにも思えるが、いわゆる２ハイブリッド実験の条件下では、必ずしもそうではない。実際、調べてみると、細胞内局在、細胞内での役割、生化学的機能で、それぞれ72.7%、63.6%、52.7%の機能を予測できることがわかった。前二者では直接相互作用するものだけをみたk = 1が最適で、生化学的機能ではわずかにk = 2が最適であった。50%そこそこの予測能力といっても、分類が細かいことを考えると十分に意味のある情報を提供している。ちなみに、ランダムに機能を帰属させたデータでは10%以下しか正解は得られなかった。以上の解析結果をもとに、いわゆるホモロジー検索で機能が推定しにくい16の酵母ORFの機能を予測しておいた。これらの予測は将来の実験解析において検証されることを期待している。

2) RNA 塩基配列からの規則発見

2-1) 長大なイントロンを認識するための配列シグナル

当初の予定とは少し異なって、「長いイントロン」がなぜ正確に認識できるのかという問題に対して、いくつかの角度から検討した。「長いイントロン」とは、イントロン長の対数を横軸にとった頻度分布を作ったときに現れる二番目のピークの分布を指す（Burge らによって提唱された）。

イントラスプライシングのモデル提唱

イントラスプライシングと呼ぶ、長いイントロンのスプライシングに関する新しいモデルを提唱した。このモデルの基本的なアイディアは、長いイントロンがスプライシングを受ける前に、内部で短いイントロンのスプライシングが起こることによって、10 万塩基にも及ぶ長いイントロンの認識が容易になるかもしれないというものである。このモデルの可能性を探るために、コンピュータを用いたモデル計算を行い、イントラスプライシングが長いイントロンのモデルとして成立しうることを確認した（16）。

情報量的観点からの長いイントロンの研究

長いイントロンと短いイントロンの塩基組成をいくつかの生物種で調べてみたところ、ヒトとマウスでは、イントロンが長くなるほどU含量が増え、C含量が減る傾向が認められた。さらに、3'スプライス部位のもつ情報量も、イントロンが長くなるほど大きくなる傾向があった。これらの傾向はあくまで統計的なものであり、これでもって直ちに長いイントロンの認識メカニズムを説明できるわけではないが、ゲノムデータベースを駆使することで、ゲノムワイドなデータを用いた、偏りの比較的少ない結論であると思われ、ある種の進化的な最適化の一断面を観察しているものと思われる（Bannai et al., 未発表データ）

長いイントロンの 3' 末端特異的に出現するモチーフの探索

文字列と数値のペアからなるデータを与えられた時に、「文字列中にパターンがある/ない」と「数値」との間の相関が最も高いパターンを求める、という新しいパターン発見の問題を定義した。また、この問題を解く効率的な枝刈りアルゴリズムを開発した。このアルゴリズムはミスマッチを含むパターンや、任意文字長のワイルドカードを含むパターン等、様々なパターンの種類に対しても適用が可能で、最適解を求める事ができる。応用として、イントロンの3'スプライス部位の配列と、イントロン全体の長さのペアに対してこのアルゴリズムを適用した。その結果、比較的短いイントロンの3'部位にはあまり現われないが、長いイントロンの3'部位には多く現われるパターンを得る事ができた。この方法は、たとえば遺伝子の発現量と相関するプロモーター内のモチーフ発見などにも適用可能と思われる（12）。

2-2) スプライシング異常データからの規則発見

この課題については、若干試みてみたものの、上述のテーマの方がデータ量が多くて扱いやすいということになって、結果を出すところにまでは至らなかった。

2-3) 真核生物 mRNA の半減期に関わるモチーフ探索

接尾辞木 (suffix tree) を用いることで、与えられたスコア関数に対して最適解を与える二つの部分文字列パターン（モチーフ：存在しないことが意味をもつ負のモチーフも許す）の組み合わせを、用いた総配列長に対してO(N2) 時間で求めるアルゴリズムを開発した(25, 26, 27)。

このアルゴリズムで、酵母の mRNA の半減期のデータ（Wang et al., 2002）とその3’ UTR配列を探索してみると、UAAAAAUA or UGUAUAAというパターンが、半減期が10分以内の 393のうち55個に含まれていないのに対して、50分以上の379中に 7 個しか含まれていなかった。また、AUCCを持たず、UGUAを持つというパターンは、393のうちの240が該当し、379のうちでは52が該当することがわかった。UGUAというモチーフはPUFという因子の結合配列と思われ、第一のルールにも含まれるため、この結果は生物学的知見とも矛盾しない。ここで注意すべきは探索したのは、上述の両極端のデータだけではなく、データ全部における相関を調べたことである。

さらにヒトのデータ（Yang et al., 2003）に対しても適用してみたところ、UUAUUU or UGUAUAパターン、もしくは関連の深いUAUUUAU or UGUAUAUパターンが発見された。AUUUAというモチーフはAREというエレメントと思われ、UGUAは酵母のときと同じPUF結合配列と思われる（Bannai et al., 未発表データ）

3) DNA 塩基配列からの規則発見

3-1) 細菌ゲノムにおける遺伝子発見

細菌ゲノムにおける遺伝子発見はかなり高いレベルの信頼性を備えているものの、特に翻訳開始点の予測能力向上が大きな課題となっている。この場合の難しさは、予測そのものの難しさのほかに、信頼できる翻訳開始点のデータの不足によるところが大きい。そこで我々は、以前矢田らが構築した遺伝子発見システム GeneHacker (Yada and Hirosawa, 1996)を新たに全面的に書き直した新システムGeneHacker Plusを構築した(4)。他の多くのシステム同様、本システムも隠れマルコフモデル(HMM)に基づいている。また、コード領域を典型的なコドン使用のタイプと非典型的なタイプに分けているところも、米国のGeneMark.hmmと同じである。少し違うところは、HMMでゲノム全体をモデル化するのではなく、遺伝子（翻訳制御領域＋コード領域）をモデル化して、実際に遺伝子発見するときには、いわゆる局所アラインメントの場合のように、局所的にマッチした部分を切り出してくるところである。これによって、オーバーラップした遺伝子の処理も自然な形で行うことができる。さらに、プロテオーム実験などで翻訳開始点が知られているデータが多数入手できる生物種に対しては、そちらのデータを利用し、翻訳制御領域とスペーサー領域を条件付き確率を使って、精密にモデル化している。これによって、シアノバクテリアのように、通常とは少し異なるタイプの制御領域をもった生物種に対しても、高い翻訳開始点の予測能力（90%前後）を得ることができた。また、コード領域の予測能力も世界的に用いられているプログラムと比べて同等以上であると言える。

3-2) 細菌遺伝子の転写制御シグナル

ゲノム比較を用いた枯草菌レギュロン予測

共通の転写因子で制御される遺伝子群（共発現遺伝子）は原核生物ではレギュロンと呼ばれる。これらの遺伝子は細胞機能的に関連していることが多く、この制御関係を明らかにすることは転写制御ネットワークの解明の最初の一歩であると言える。これらの遺伝子は（オペロン構造を考えなければ）皆、上流に共通の転写因子結合配列をもっているはずなので、原理的にはそのような共通配列を探せばよいが、実際にはノイズが多くて難しい。そこで、我々は枯草菌と同属のB.halodurans とB.stearothermophilusのゲノム配列との保存情報を用いてレギュロン構造を予測した。同属といっても、明らかなオーソロガス遺伝子は1,568遺伝子しか得られなかったが、その上流から1,884の保存エレメントを抽出した。それらは既知の転写因子結合部位を多く含み、ランダムなアラインメントの結果などとの比較から、高い保存度を示すもには擬陽性は少ないものと推定した。。-35/-10ボックスやSD配列による悪影響も少なかった。共通エレメントをもつ遺伝子群には、新規メンバーを含む既知のレギュロンも含まれていた。また、共通のアテニュエーション因子によって翻訳後に制御される遺伝子群が予想外に多く含まれていたことも、本アプローチの有効性を裏付けるものであった。

他の細菌におけるレギュロン予測

枯草菌の近縁種で行ったレギュロン予測研究の方法論をいくつかの点で改良した。その方法を用いて、クラミジア、マイコプラズマ、ストレプトコッカス、マイコバクテリウム、ガンマプロテオバクテリアの一部、シアノバクテリアなどなどにおいても、上流の保存エレメントの網羅的な抽出と、それらの類似性に基づくレギュロン予測を行った（Makita et al. 未発表データ）。また、得られた保存エレメントの生物学的意味を確認する手段として、それらのエレメントが真正細菌全体にわたって同様に保持する種があるかどうかを網羅的に探索している。得られた結果は後述のようにDBTBSで公開したが、レギュロン予測については、たとえばストレス応答エレメントの保存などにおいて、いくつか興味深い結果が得られたが、いまだ論文にはできていない。

枯草菌転写データベースDBTBSの更新

1999年に発表した枯草菌の転写因子とその結合部位に関するレファレンスデータベースDBTBSを大幅にリニューアルした。基本的には実験によって確かめられ、論文として発表された転写因子の結合配列情報を収集したものである。直接参照している論文の数は291報から378報に増え、シグマ因子を含む114の転写因子、525遺伝子の633プロモーターの情報を備えるようになった。得られた情報から構築した転写因子認識配列の重み行列を入力配列に対して適用し、新規結合部位候補を探索する機能や、前述の予測レギュロンの表も追加した。さらに、まだ改良の余地は多いものの、50の真正細菌ゲノムに対して、枯草菌の転写因子の推定オーソログ遺伝子、被制御遺伝子の推定オーソログ遺伝子、ならびにその上流配列中における転写因子結合配列様モチーフの存在を表示する機能を付加した（20)。

枯草菌転写制御システムに関する網羅的予測研究

DBTBSを用いた応用研究として、前述の上流配列のゲノム比較に基づくレギュロン予測の他、マイクロアレイデータなども利用したシグマ因子依存性予測（22：宮野研との共同研究)、ターミネータ予測結果の比較ゲノム解

析（32）、ベイズ推定法を用いた転写因子結合部位の予測（31）など、一連の網羅的予測研究に取り組んだ。それぞれ簡単に紹介する。

細菌の遺伝子がどのシグマ因子によって転写を制御されているかを予測することは、その生物の転写制御メカニズムを知る第一歩である。今回、ベイジアンネットワーク法、微分方程式を用いた動的モデル、共発現情報の教師付き学習法、などによって、その予測を試みた。配列データとマイクロアレイによる発現情報をロジスティック回帰スコアを用いてうまく組み合わせることができた。結果としては、共発現情報の教師付き学習が最も効果的であった。いくつかの未知遺伝子についても高い信頼性が期待される予測結果が得られた。

続いて、シグマ因子が特定のリプレッサーやアクティベーターなどの転写因子と共役して用いられる現象を精度良く予測するために、DBTBSの既知データからそのようなケースをリストアップして、間の距離の統計分布、シグマ因子依存性スコア、などなどの情報をベイズ統計法を用いて統一的に扱う方法を定式化した。この方法によれば、転写因子やシグマ因子の依存性をそれぞれ単独で行ったときよりも高い予測精度が得られた。得られたスコア関数を枯草菌の全遺伝子に適用して、レギュロンの新メンバーを予測した。

さらに、ρ因子非依存性ターミネーターの網羅的な予測を行った。大腸菌と違って、枯草菌（と、おそらくファーミキューテス全体）ではρ因子は必須ではないので、こちらだけを調べても片手落ちとは思えない。463個の既知枯草菌ターミネーターをもとに得た予測プログラムを用いてオペロン構造を予測してみたところ、感度、特異性ともに94%程度と見積もられた。同じルールを他細菌にも適用したところ、57種のファーミキューテスについては枯草菌と同様の感度が得られたが、その他についてはかなり低い感度しか得られなかった。これは転写終結システムの系統進化的な違いを反映しているものと思われる。我々の結果から、ファーミキューテスにおいては、実験的に得られたオペロン情報がなくても、十分高い精度でオペロン構造を予測できることが推定された。予測結果はインターネット上で公開している。

3-3) 高等真核生物遺伝子の転写制御シグナル

転写開始点データベースDBTSSの構築と更新

菅野純夫氏（東大医科研、現：東大新領域）のグループとの共同研究で、彼らが決定したヒト遺伝子転写物5'末端配列をゲノム上にマップしたデータベースDBTSSを構築した（10）。通常の公共塩基配列データベースにあるcDNAの配列は多くが5'端の情報が欠けており、その部分にエキソン境界が存在したりすると、その上流で制御配列を解析しようとしても誤ってしまうことになる。そこで我々は菅野氏らが開発したオリゴキャップ法を応用して、ヒトなどの遺伝子のcDNAで完全な5'端をもつライブラリを作成し、網羅的に配列決定した。7,889遺伝子の111,382クローンが既知のcDNA配列とマッチした。得られた転写開始点情報とRefSeqを比較したところ、34%は5'端情報が欠けていると判定された。クローンデータをヒトゲノムにマップして、ヒト遺伝子の転写開始点情報を収めたデータベースとして整理した。

2004年に大幅に更新したバージョンを公開した（21）。主な変更点は、データ量の大幅な増加（RefSeqとゲノムの両方にマッチするクローンの数は111,382から190,964になり、11,234遺伝子をカバー）、dbSNPのSNP情報を付加して、制御領域における多型情報を閲覧できるようにしたこと、対応生物種にマウスとマラリアを追加したこと、である。対応生物種は今後も増やしていく予定である。さらに、マウスのデータは単独で閲覧できるだけでなく、ヒトプロモーターとの比較も行えるようにでき、既知の転写因子結合部位なども検索できるようになった（後述）。

転写開始点のゆらぎと選択的プロモーター

多数の転写開始点情報を用いて、ヒト遺伝子転写開始点付近の網羅的な配列解析を行った。転写開始点のゆらぎを各遺伝子ごとに調べてみると、ゆらぎがまったく見られない遺伝子群、比較的小さいゆらぎをもつ遺伝子群、大きなゆらぎをもつ遺伝子群の3つに分類されることがわかった。大きなゆらぎはエキソン・イントロン構造の違いをうみだしており、ほとんどは選択的プロモーターの存在によるものと思われる。TATA ボックスの存在ではゆらぎの有無を完全には説明することができないが、ゆらぎのない遺伝子群の80%にはTATAボックスが存在した。選択的プロモーターの網羅的探索、およびその進化的起源に関する情報は次期ゲノム特定領域研究において、我々の主要な研究テーマの一つとなり、ひき続き研究を続けている（たとえばKimura et al., Genome Res. 2006、Tsuritani et al.投稿準備中）。

転写開始点とCpGアイランドの関係

一般にCpGアイランドはハウスキーピング遺伝子の転写開始点付近に多く存在すると言われているが、正確な転写開始点との位置関係や、CpGを近くにもつ遺伝子の発現情報の網羅的統計解析はこれまで行われていなかった。そこで我々はヒトとマウスの数千のオーソログス遺伝子の正確な転写開始点のデータを用いて、CpGアイランドの存在は統計的に転写開始点のところで最大ピークとなっていることを見いだした。次に我々は、-100:+100領域にCpGアイランドがあるかないかで、ヒト（マウス）のプロモーターを6600 (2948)のCpG+と2619 (1830)のCpG-に分類し、そのハウスキーピング（組織非特異）性をUniGeneデータベースにおいて観測されたライブラリー数によって見積もった。その結果、ヒトでもマウスでも、CpGアイランドを転写開始点付近にもたない遺伝子は組織特異性発現をする傾向が強いことを見いだした。同様の傾向は遺伝子のGene Ontologyアノテーションの解析でも見られた。コントロール実験などの結果ともあわせ、CpGアイランドの効果は、遺伝子が存在するバンド構造上の位置等よりも、その発現により強く相関していると結論した。

TOP 遺伝子の網羅的予測

転写開始点にゆらぎのない遺伝子の転写開始点付近の配列を詳しく調べてみると、その中の一部はTOP (terminal oligo-pyrimidine-rich)遺伝子と呼ばれているグループで占められていた。TOP遺伝子はリボソームタンパク質遺伝子など、翻訳関係の遺伝子を多く含み、細胞にストレスなどをかけると、翻訳レベルで発現が抑えられるという特徴をもっている。また、名前にもあるように、転写開始点付近に特徴的な配列パターンをもっている。DBTSSの正確な転写開始点情報とこのパターンの情報を組み合わせることで、ゲノムワイドなTOP遺伝子予測を行ったところ、ヒトゲノム全体で1600以上もの遺伝子がTOP遺伝子であると予測された。その中には多くの（これまでTOP遺伝子として報告されていなかった）翻訳

関連遺伝子が含まれていた。ヒトでTOPと予測された遺伝子の32%はマウスでもTOP遺伝子のコンセンサス配列を転写開始点付近にもっていた（Yamashita et al.,未発表データ）。現在、実験によってこの予測を検証すべく準備を行っている。

ヒトとマウスの網羅的プロモーター比較

ヒトとマウスのオーソログガス遺伝子の上流配列をプロモーター候補として網羅的に解析した（23）。5 '端の位置は従来報告されていたそれよりも平均4kb上流にあった。3,324遺伝子の上流配列を比較してみたところ、平均して約45%の領域が保存されていた。22,793箇所の転写因子結合予測部位も保存されていた。すべてのデータはDBTSS上で公開している。

さらに、3,324対の遺伝子の転写開始点上流1 kbをアラインしてみたところ、保存領域は転写開始点付近から平均510 bp上流までのびて、その後はまったく類似性が見られないのが普通であることが分かった。すなわちこのように類似性が不連続的に失われるというブロック構造が約1/3のプロモーターに見られた。ブロックの内側と外側では、G+C含量やCpGの頻度が異なっていた。ブロック内部では、配列の一致度はブロック長にかかわらず均一に65%程度であった。既知の転写因子結合部位の90%はブロック内に存在した。46%のブロックはリピート配列で終わっていたので、これらはゲノムの再編成によって生じたのかもしれない。ブロック長が一番短い遺伝子は転写因子や脳特異的に発現していた。このことはこれらの遺伝子の進化では転写制御メカニズムの変化が重要だったことを意味しているのかもしれない（24）。

その他

伊藤隆司教授（金沢大、現：東大新領域）の酵母プロモーター解析に協力した（7）。酵母の多薬剤耐性遺伝子を制御する転写因子Pdr1pによって制御されていると思われる遺伝子上流に存在するコンセンサスモチーフの探索を行った。

また、DBTSSの研究で得られた4,870のヒトの5 'UTR領域を調べたところ、約半数は最長ORFが始まるATGより上流に別のATGをもっていた。それらは平均長31アミノ酸の短いORFを構成しており、今後なんらかの意味が見いだされる可能性がある（19）。

加藤菊也教授（奈良先端大、現：府立成人病センター）と協力して、いくるかのがん細胞における遺伝子発現情報などを収めたデータベース CGED (Cancer Gene Expression Database)を構築した（28）。

4) 上記の発見を可能にするための基盤情報研究

4-1) 発見科学からのアプローチ

分担者の丸山らは、情報科学の立場からDNAモチーフ配列の共起性に着目し、共起シグナル配列をとらえるモデルの定式化を行った。さらに、これらを探索するビット演算に基づくアルゴリズムの設計と計算機実験によるモデルの有効性の検証を行った（2, 8, 9）。

仮説モデル間の構造的関係性の研究：計算機を用いたゲノムデータから知識発見では，様々な仮説モデル(決定木, association rule, 論理式，線形判別, Support Vector Machine, neural network 等)が用いられている．しかしながら，これらの優劣や構造的関係性については明らかにされておらず，ユーザは経験的に仮説モデルを選択し試行錯誤的にモデルの最適化を行っているのが実状である．本研究では，仮説モデル間の類似度関数を定式化し，解析を行い，構造的関係性を明らかにした（13, 14）。

4-2) DNA 共通モチーフ抽出プログラムの改良

いろいろなモチーフ抽出プログラムの結果や、一つのプログラムに対していくつか異なるパラメータ指定をしたときの結果の違いをグラフィカルに表示するモチーフ抽出支援ツールMelinaを開発し、公開した（17）。現在のところ、マイクロアレイ実験などで示唆される共制御遺伝子候補の上流配列から共通のモチーフを抽出する場合に、あらゆる状況でベストなアルゴリズムは知られておらず、いくつかのプログラムを併用することが奨励されている。本ツールはその作業を支援し、モチーフの見逃しを防ぐ。また、パラメータ値をいろいろ変化させたときの結果の違いを見ることで、発見したモチーフの安定性を調べたり、モチーフ抽出条件を最適化したりできる。なお、Melinaは、現在さらにいろいろな機能とより使い易いインターフェースを備えた新しいバージョンMelina IIとして現在更新作業中である。

さらに、人工配列の中にモチーフを埋め込む計算機実験を網羅的に行うことで、代表的な塩基配列モチーフ抽出プログラムであるGibbs samplerとMEMEについて、その性能限界を調べたり、いろいろな条件における最適パラメーター値の探索を行った。それらをいわゆるパラメータ・ランドスケープとしてユーザの利用に供することを目指している。特にGibbs samplerとMEMEについて、そのパラメータ値の性能への影響を組織的に解析した。その結果、Gibbs samplerは比較的いくつかのパラメーター値の変化に敏感であり、これに対してMEMEはより安定であることを見いだした。Gibbs samplerについて、4つのモチーフ長のクラスにおいて、それぞれ最適と思われるパラメータ値を求めた。従って、あらかじめ探索すべきモチーフの長さがまったく不明である場合は、これらの4つの設定で実行して結果を比較することにより、正解を見逃す可能性を大きく減らすことが出来るものと思われる（30）。

〈国内外での成果の位置づけ〉

上述の4つの分野のそれぞれについての反響をまとめてみる。以下で参考に付した論文の引用数は2006年2月現在のGoogle scholarによるものである。

まず、アミノ酸配列解析関連では、iPSORTの論文（11）が世界的に評価され、論文引用数はすでに111を数えている。実際、この研究は当初の計画通り、この予測問題のエキスパートである研究代表者と、発見科学・アルゴリズム開発の分野で実績をあげてきた研究分担者の能力が、週1回のミーティングを続けたのちに最もよく発揮されたと感じている。また細胞内局在部位予測問題に関する総説も広く引用されていると言える（1が112回、5が25回など）。研究代表者はその後もしばしば総説の執筆を依頼されている。カナダのグループとの共同研究PSORT-B（18）も46回引用されている。さらに、タンパク質間相互作用データから、タンパク質の機能を推定する方法論に関する研究（3）は、地味なジャーナルに発表され、発表当初はほとんど反響がなかったにもかかわらず、その後徐々に評価を高め、この分野における先駆的な業績の一つとしてしばしば引用されている（引用数42）。

RNA スプライシングのインフォマティクス研究はおそらく国内で唯一であり、Intrasplicing というモデルに関

しても、国際会議で発表し、それなりの関心を呼んだが、まだまだ海外での知名度は低い。多くの研究者に受け入れられるためにはさらなる研究が必要である。後述のように、RNAの配列解析が当初の予定ほどには行えなかったのは、その後のこの分野の発展を見ても非常に残念である。

DNA配列の解析について。まず、矢田氏との共同研究による細菌の遺伝子発見プログラムGeneHacker Plusは今も非常に高い性能を誇っており、九大や奈良先端大のゲノム解析グループに利用されているものの、海外での知名度は芳しくなく、論文の引用数も13にとどまっている。これは、遺伝子発見問題のトピックがその後高等真核生物に移り、また遺伝子発見自体が下火になっていったことが主な原因であると考えている。

枯草菌を中心とする細菌遺伝子の転写制御領域に関する一連の研究は、先駆的かつ室の高いものになったと自負しているが、客観的にみて大きな反響を呼んだとは言えない。引用数では、DBTBSの最初の論文が18、本研究による更新版の論文（20）が23であり、比較ゲノムを用いたレギュロン予測（6）は世界的にみてもタイムリーな論文であったが13である。その他の論文もISMBやPLoS Comput. Biol.などの競争率の高い学会や雑誌に採択されている。データベースとしてのDBTBSは毎月平均約10万アクセス、ユニークビジター数が月平均約3,500と、世界的に認知されつつある。

菅野研究室との共同研究によるヒトを中心とした高等生物のプロモーター解析は非常にうまく進行していると言える。まず転写開始点とその上流プロモーターデータベースDBTSSはこの分野のスタンダードと見なされつつある。論文の引用数では2002年のもの（10）が91、2004年のもの（21）が37で、2006年のNAR Database issueにも採択されている。アクセス数は、アクセスページ合計が月平均40万程度、ユニークビジター数が2,000程度である。このデータを用いた我々自身の情報解析研究結果はまだそれほど引用されているとは言えないが、今後も続々と新しい成果が生み出されていく予定である。CpGアイランドに関する研究（29）は海外であまり知られていなかったため、2006年に非常に似た内容の論文がPNAS誌に発表されたが、我々の指摘で先行研究の存在を知った著者の一人Paul Bergは今後必ず我々の仕事を引用することを約束してくれた。

モチーフ抽出支援ツールMelinaは、月平均アクセスページ合計が約25,000、ユニークビジター数が約1,000であり、今後新バージョンが有名になればより利用者数が増加していくものと考えている。また、パラメータランドスケープ研究の論文（30）は、Regulatory Genomicsという国際学会で採択されている。

〈達成できなかったこと、予想外の困難、その理由〉

上述のように、本計画研究では5年間に多くの成果をあげてきた。その中には予想以上に進展したテーマもあるが、計画通りに進まなかった研究も多い。

まずアミノ酸配列解析では、膜タンパク質トポロジーに関する研究がいまだに論文としてまとめられていない。これはその研究を行っていた大学院生が卒業・就職してしまったことによる。その学生は、その後もこの研究を続けることを希望し、実際折にふれて論文化の努力を続けてきたものの、結局転職後、仕事が忙しくなり過ぎ、またデータも今ひとつ決定的な結果がでていないこともあって、まとめられていないのである。今後、このテーマに興味をもつ学生が出てくれば、引き継いでいくことを考えている。これまでのところはなぜか学生はアミノ酸配列解析よりも、塩基配列解析を好む傾向が強かった。本計画研究全般に言えることだが、学生の教育という観点からすると、ある程度本人の自主性を尊重すれば、どうしても論文としてまとめるのが遅くなるのは仕方がない面がある。また、タンパク3000プロジェクトにも参加させていただいたために、成果の切り分けが必要になった事情もあった。

次にRNAの配列解析は、当初本計画研究で最も重視していたが、結果としてはほとんどめぼしい成果を上げることができなかった。その理由はいくつか考えられる。第一に、まさに問題そのものの複雑さ・難しさが予想以上であったことがある。第二に情報科学のエキスパートは、大きくて複雑な問題に試行錯誤しながらアプローチしていくよりは、たとえ難しくても最初から明快に定式化された問題設定を好むことを感じた。生物学の問題の多くは問題の設定の仕方そのものが時間とともに変わっていくようなところがあるが、そのような問題に時間をかけて取り組んでもらうことはあまり歓迎されなかった。逆にiPSORTの研究には非常に高度なテクニックも用いられたが、問題そのものは過去の論文から学習データを丸ごと拝借したに過ぎず、このあたりに好みの違いを感じた。結局、RNA スプライシングに関しては、当初の計画のようには異常スプライシングの研究が進展せず、そこをステップにした選択的スプライシングの研究にも手がつけられなかった。選択的スプライシングは（タンパク3000プロジェクトにおいて）イソフォームの選択的局在化という別の観点から研究を進めた。第3に前半の分担者であった宮野グループの研究者が九州等に異動になり、また宮野教授の研究テーマがシステム生物学にシフトしたことも、十分に共同研究を極められなかった理由であった。

DNAにコードされた転写制御領域の研究は全般にまずまずの成果を収められたが、学生の研究がなかなか進まなくて苦労した面はある。微生物プロモーターの研究では、マイコバクテリウム菌で興味ある繰り返し配列をみつけて、その解析に時間を割いていたが、結局、その繰り返し配列が既知であることがその後判明するというお粗末もあった。

後半の2年間では、ゲノム生物学に移籍して、実験研究者との連携を図った。基本的には実験系の集まりである「グラム陽性細菌のゲノム生物学」という合宿形式のミーティングに積極的に参加・発表するなどの努力を行った。特に摂南大学の高松先生とはコンピュータ予測の結果を実験で検証するために共同研究を行った（論文準備中）。また、公募班員の知花先生とカンジダ酵母のゲノム解析で若干共同研究も行った。最後に、「統合ゲノム」の日下部先生と共同研究を始めるきっかけもできた（現在基盤研究Bが進行中）。その意味で、「ゲノム生物」の中で一定の寄与は行ったものと考えている。しかし、いずれも研究期間中に顕著な成果を上げるところまでは至っていない。実験研究者と情報研究者の共同研究の難しさについては以前指摘されていたほどの壁はなくなっているが、よほど意見交換を密にしないと、たとえば情報研究者が研究発表したとき、実験上の細かい点ばかり質問されたりするなどの問題に直面した。

〈まとめと今後の課題〉

5年間、「配列情報に潜む規則性の発見」と題して、DNA. RNA. アミノ酸の各配列に潜む種々の規則性をコンピュータ解析によって明らかにする研究に取り組んでき

た。前半３年間は「ゲノム情報科学」領域に属し、後半は「ゲノム生物」領域に移って、実験系研究者との連携を図ることにも努めてきた。それら配列情報に包含される規則性の世界は深遠で、５年間かけてもまだその上っ面をなでただけかもしれないが、これまで述べてきたように、着実に成果をあげ、研究成果の一般公開も十分行ってきたと考えている。

全般にみて、ゲノム情報科学という分野が世界的に注目されるにつれ、競争が激化している。専門誌などの敷居も年々高くなっている印象がある。ただ、この５年間のうちに、最初は転写制御領域の研究で世界的には無名に近い存在であった代表者たちが、近年は少しは名前を知られるようになり、それなりのジャーナルの論文の査読を依頼されるようになってきた。今後も着実に実績を積み重ねて、我々のグループが世界的な研究拠点の一つとなることを目指したい。そのために、維持している基盤となるデータベースの一層の拡充に努めたい。

より具体的には、細菌の転写制御配列解析では、膨大な数のゲノム配列をいかした比較ゲノム研究とDBTBSに蓄えられた知見をいかにうまく組み合わせていくかが鍵であると考えており、その線に沿った研究を現在進めている。高等真核生物のプロモーター解析では、目下の焦点は選択的プロモーターの問題であるが、興味深い研究テーマがいくらでもほりだしてこれそうな印象がある。良いテーマを探して、転写制御配列の設計原理のようなものを明らかにしていきたい。そのためにも細菌とヒトの中間のレベルの実験生物として、現在ホヤのプロモーター解析にも取り組んでいる。RNAについては、近年のmiRNAなどに関する知見の集積に鑑みて、新しい研究を始めつつある。タンパク質の細胞内局在に関しても、産総研CBRCのホートン氏のチームと共同研究を進めている。本研究は決してシステム生物学と相反するものではないが、比較的古くからの視点に立つ研究として、流行に流されず、地道でかつ質の高い研究を積み重ねていきたいと考えている。

〈研究期間の全成果公表リスト〉

１）論文／プロシーディング（査読付きのものに限る）

1. 0602130201
Nakai, K., Protein sorting signals and prediction of subcellular localization, Adv. Protein Chem., 54, 277-344, (2000).

2. 0602130157
Maruyama, O. and Miyano, S.: Design Aspects of Discovery Systems, IEICE Trans. Inf. and Sys., E83-D(1), 61-70, (2000).

3. 0202182245
Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., and Takagi, T., Assessment of prediction accuracy of protein function from protein-protein interaction data, Yeast, 18 (6), 523-531 (2001)

4. 0202182251
Yada, T., Totoki, Y., Takagi, T., and Nakai, K., A novel bacterial gene-finding system with top-class accuracy in locating start codons, DNA Research, 8 (3), 97-106 (2001)

5. 0202182256
Nakai, K., Prediction of in vivo fates of proteins in the era of genomics and proteomics, J. Struct. Biol., 134 (2/3), 103-116 (2001)

6. 0202182304
Terai, G., Takagi, T., and Nakai, K., Prediction of co-regulated genes in Bacillus subtilis on the basis of upstream elements conserved across three closely related species, Genome Biology, 2 (11), research0048.1-0048.12 (2001)

7. 0202182327
Miura, F., Yada, T., Nakai, K., Sakaki, Y., and Ito, T., Differential display analysis of mutants for the transcription factor Pdr1p regulating multidrug resistance in the budding yeast, FEBS Letters, 505, 103-108 (2001)

8. 0202182324
Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S., Views: fundamental building blocks in the process of knowledge discovery, 14th Int. FLAIRS Conf., 233-238, AAAI Press (2001)

9. 0602131034
Maruyama, O., Shoudai, T., Furuichi, E., Kuhara, S., and Miyano, S., Learning Conformation Rules. In Proceedings of the 4nd International Conference of Discovery Science (Lecture Notes in Artificial Intelligence, 2226), Springer-Verlag, 243–257, (2001).

10. 0202182312
Suzuki, Y., Yamashita, R., Nakai, K., and Sugano, S., DBTSS: database of human transcriptional start sites and full-length cDNAs, Nucl. Acids Res., 30(1), 328-331 (2002)

11. 0202182321
Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S., Extensive feature detection of N-terminal protein sorting signals, Bioinformatics, 18 (2), 298-305 (2002)

12. 0303310158
Bannai, H., Inenaga, S., Shinohara, A., Takeda, M., and Miyano, S., A string pattern regression algorithm and its application to pattern discovery in long introns, Genome Informaticss, 13, 3-11 (2002).

13. 0602130204
Maruyama, O., Bannai, H., Tamada, S., Kuhara, S., and Miyano, S., Fast Algorithm for Extracting Multiple Unordered Short Motifs Using Bit Operations. In Proceedings of Joint Conference on Information Sciences, 1180-1185 (2002).

14. 0303261723
Maruyama, O., Shoudai, T., and Miyano, S., Toward drawing an atlas of hypothesis classes; approximating a hypothesis via another hypothesis model, Proc. 5th Int. Conf. Discovery Sci. (Lecture Notes in Computer Science) 2534, 220-232 (2002).

15. 0304041357
Nakai, K., Signal peptides, (ed. U. Langel) Cell-Penetrating Peptides: Processes and Applications (CRC Press), 295-324 (2002).

16. 0303302252
Ott, S., Tamada, Y., Bannai, H., Nakai, K., and Miyano, S., Intrasplicing: analysis of long intron sequences, Pacific Symposium on Biocomputing, 8, 339-350 (2003).

17. 0303302237
Poluliakh, N., Takagi, T., and Nakai, K., MELINA: motif extraction from promoter regions of potentially co-regulated genes, Bioinformatics, 19(3), 423-424 (2003).

18. 0401092329
Gardy, J. L., Spencer, C., Wang, K., Ester, M., Tusnady, G. E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K., and Brinkman, F. S. L., PSORT-B: Improving protein subcelular localization prediction for Gram-negative bacteria, Nucl. Acids Res., 31(13), 3613-3617 (2003).
19. 0403261407
Yamashita, R. et al., Small open reading frames in 5' untranslated regions of mRNAs, C. R. Biol., 326(10-11), 987-991 (2003).
20. 0401092349
Makita, Y., Nakao, M., Ogasawara, N., and Nakai, K., DBTBS: Database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics, Nucl. Acids Res., 32, D75-D77 (2004)
21. 0401100001/0403261407
Suzuki, Y., Yamashita, R., Sugano, S., and Nakai, K., DBTSS (DataBase of Transcriptional Start Sites): Progress Report 2004, Nucl. Acids Res., 32, D78-D81 (2004).
22. 0602130210
De Hoon, M.J.L., Makita, Y., Imoto, S., Kobayashi, K., Ogasawara, N., Nakai, K., and Miyano, S., Predicting gene regulation by sigma factors in Bacilllus subtilis from genome-wide data, Bioinformatics, 20(Supp. 1), I101-I108 (2004)
23. 0602130215
Suzuki, Y., Yamashita, R., Shirota, M., Sakakibara, Y., Chiba, J., Mizushima-Sugano, J., Kel, A. E., Arakawa, T., Caminci, P., Kawai, J., Hayashizaki, Y., Takagi, T., Nakai, K., and Sugano, S., Large-scale collection and characterization of promoters of human and mouse genes, In silico Biol., 4, 0036 (2004).
24. 0602071730
Suzuki, Y., et al., Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions, Genome Res., 14, 1711-1718 (2004)
25. 0602131038
Bannai, H., Hyyro, H., Shinohara, A., Takeda, M., Nakai, K., and Miyano, S., Finding optimal pairs of patterns, Lecture Notes in Comp. Sci. (WABI 2004), 3240, 450-462 (2004).
26. 0602131041
Inenaga, S., Bannai, H., Hyyro, H., Shinohara, A., Takeda, M., Nakai, K., and Miyano, S., Finding optimal pairs of cooperative and competing patterns with bounded distance, Lecture Notes in Comp. Sci. (DS 2004),, 3245 32-46 (2004).
27. 0602131045
Bannai, H., Hyyro, H., Shinohara, A., Takeda, M., Nakai, K., and Miyano, S., An O(N^2) algorithm for discovering optimal boolean pattern pairs, IEEE/ACM Transactions on Computational Biology and Bioinformatics (special section on the Workshop on Algorithms in Bioinformatics), 1(4), 159-170 (2004).
28. 0602071730
Kato, K., Yamashita, R., Matoba, R., Monden, M., Noguchi, S., Takagi, T., and Nakai, K., Cancer gene expression database (CGED): a database for gene expression profiling with accompanying clinical information of human cancer tissues, Nucl. Acids Res., 33, D533-D536 (2005).
29. 0602131056
Yamashita, R., Suzuki, Y., Sugano, S., and Nakai, K., Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue-specificity, Gene, 350(2), 129-136 (2005)
30. 0602131105
Poluliakh, N., Konno, M., Horton, P., and Nakai, K., Parameter landscape analysis for common motif discovery programs, in Eskin, E. & Workman, C. (eds.), Regulatory Genomics, RECOMB 2004 International Workshop, RRG 2004, San Diego, CA, USA, March 26-27, 2004, Revised Selected Papers. Lecture Notes in Computer Science 3318, pp. 79-87, Springer (2005).
31. 0602131110
Makita, Y., De Hoon, M.J.L., Ogasawara, N., Miyano, S., and Nakai, K., Bayesian joint prediction of associated transcription factors in Bacillus subtilis, Pacific Symposium on Biocomputing 2005 (Altman et al ed.), 507-518, World Scientific (2005).
32. 0602180036
De Hoon, M.J.L., Makita, Y., Nakai, K., and Miyano, S., Prediction of transcriptional terminators in Bacillus subtilis and related species, PLoS Comput. Biol.1(3), e25 (2005).

2）データベース／ソフトウェア
a. 0202251630 (No.41) DBTSS
b. 0202251703 (No.43) Melina
c. 0202251712 (No.114) iPSORT
d. 0403271527 (No.85) Cancer Gene Expression Database (CGED)
e. 0602131207 (No.42) DBTBS

3）特許など
該当無し
4）その他顕著なもの、
特になし